UNIVERSITY OF MINNESOTA

This is to certify that I have examined
this copy of a doctoral dissertation by

Chelsey Legacy

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Robert delMas	Andrew Zieffler
Name of Faculty Co-Advisor	Name of Faculty Co-Advisor
Signature of Faculty Co-Advisor	Signature of Faculty Co-Advisor
Date	Date

GRADUATE SCHOOL

Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Chelsey Legacy

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Robert delMas and
Andrew Zieffler, Advisors

August 2022

Acknowledgements

There are many people to thank and acknowledge for their role in helping me across this finish line. I will start with those closest to this project - my advisors. Bob, your infectious enthusiasm, and endless stream of ideas are enough to drive several lifetimes of careers in statistics education research. I'm thankful to be infused with that energy as I move forward in my career. Thank you for your encouragement, kind words, and support along this journey. Andy, I cannot thank you enough for the time you have taken to teach me the lay of the land in all things related to academia. We have spent countless hours discussing the latest in statistics education research, choosing color palettes for visualizations, and critiquing tv shows. All your time, encouragement, and advice are greatly appreciated. You both have provided mentorship that was above and beyond my expectations, and I take many life lessons from both of you. For example, you have taught me when to care more (e.g. proofreading) and when to care less (e.g. find hobbies outside of work). Many of the lessons seem simple but are invaluable.

Thank you to my committee members Erin Baldinger and Sashank Varma. I am thankful for your wisdom, guidance, and thoughtful feedback on this research. This study was also greatly improved by feedback from Suzanne Loch, Jonathan Brown, Vimal Rao, and Regina Lisinker. Analysis of data would not have been possible without patience from Vimal Rao while I figured out how to merge Windows and Mac versions of NVivo. Thank you to Suzanne Loch for allowing me to incorporate an entire unit of new material in your class. The in-class student participation was also greatly appreciated, particularly those students that volunteered to be interviewed or observed in class. Additionally, it would not have been possible to navigate all the steps and paperwork required to get through this degree without Lori Boucher and Sharon Sawyer. Thank you both for having all the answers.

It has been an invaluable experience to be a part of the thoughtful, welcoming statistics education program at the University of Minnesota. Special thanks to the founder, Joan Garfield, for shaping both this program and its graduates (past, present, and future). Additional thanks to former statistics education students that have taken time to discuss research and teaching with me these past few years: Laura Le, Matthew Beckman, Michael Huberty, and Elizabeth Fry. Endless thanks to Vimal Rao and Jonathan Brown for your companionship, brilliant meandering thoughts, advice, and encouragement along this journey.

Dr. Robin Lock and Dr. Ivan Ramler, I can officially say (after taking countless statistics courses) you will forever remain the best statistics teachers I have had. Your approachability, humor, and ability to engage students are qualities that have influenced my teaching practices and continue to inspire me.

I'm fortunate to have many friends to thank. Marissa Cook, Allison Bruso, and Marci Wood, thank you for your friendship, laughter, and support as we have navigated not only elementary, middle, and high school, but also college, graduate programs, and adulthood. Lindsey Wyatt, your encouragement, keen intellectual insights, and altruism are appreciated more than I can put into words. Thank you for always checking in and helping me think through many of life's decisions, both big and small. Manju Johny, we started our graduate school journeys together in Fall 2014, and we can finally close this chapter! I couldn't have done it without all the laughter, Panera Bread, and work sessions we shared at ISU. Your support and friendship mean the world to me. Eric Christianson, thank you for checking my papers for grammar and spelling, listening to my practice talks, and working through various activities I created for teaching and research. Most importantly, thank you for making me laugh every day. Your humor and positive nature were the sunshine of some gloomy graduate school days.

I'd be remiss to not mention how grateful I am for my family throughout this process. Thank you to my dad, sister, step-father, and grandfather for keeping me humble and reminding me that life is about the simple things. Finally, I would like to thank my mom for supporting all my decisions, answering every phone call, and teaching me to dream big. I truly could not have done it without you. I plan to repay you one Starbucks coffee at a time.

Dedication

This dissertation is dedicated to my mother, Virginia McLaughlin, and my father, Bryan Davis, in appreciation for their never-ending support and encouragement.

Abstract

Multivariate thinking is an increasingly recommended and important skill for developing statistical thinking. Currently, few studies have explored how students develop multivariate thinking. This study was conducted to learn more about developing this skill particularly when using visualization. It explored the following research questions: *(1) How does students' multivariate thinking develop as they take part in a series of activities designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester? (2) What challenges surrounding multivariate thinking persist after taking part in the intervention? Do any new challenges emerge after the completion of these activities?*

For this study, a unit on multivariate thinking was created for a data visualization course that consisted of ten activities and three assignments, implemented in Fall 2021. The students' responses on assignments were qualitatively analyzed for evidence of multivariate thinking pertaining to seven learning outcomes. Two students were observed from different sections of the course to gain insight into students' multivariate reasoning throughout the unit. Additionally, three students were interviewed at the end of the unit to provide rationale for their answers on the last assignment.

Results indicated that over the course of the multivariate thinking unit, students improved in their ability to create multivariate graphs using R. Overall students' reasoning with multiple variables improved throughout the unit, until the assignments and activities asked them to reason with more than three variables. At the end of the unit, most students still did not know if it was appropriate to make causal claims with their data. However, they remained consistently apt in their ability to create and update directed acyclic graphs, propose relationships among their variables of interest, and provide logical potential causal

variables.

Analysis of responses across the three assignments helped identify trends in the students' performance on each learning outcome and identified similar challenges as seen in the literature, such as confusion about observational data, making causal claims, and potential bias in responses due to the context of the data. Finally, the cognitive interviews provided insight into some challenges and misconception students held and gave a sense of their final multivariate reasoning skills at the end of this unit. Future work is needed to define the skills needed for multivariate thinking, the sequence of those skills for a learning trajectory, and to determine additional ways to support students' development of multivariate thinking.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Multivariate thinking is an important skill for developing statistical thinking. It requires consideration of the relationships among multiple variables (e.g., Committee, 2016; Horton, 2015; Mason & Young, 2004). However, introductory statistics courses often only give students experiences reasoning about properties of one variable or the association between two variables. The Guidelines for Assessment and Instruction in Statistics Education (GAISE) express the importance of students developing multivariate thinking skills to help make informed decisions and think critically with data. They suggest providing students' experiences with multivariate data such as visualizing different trends in disaggregated data and reasoning about relationships among variables including confounding and causal relationships (Committee, 2016).

In statistics education, multivariate thinking concepts are being newly implemented in both introductory and upper-level undergraduate courses. Past research in statistics education placed a focus on bivariate and covariational reasoning (e.g., Batanero, Estepa, & Godino, 1996; Batanero, Estepa, Godino, & Green, 1996; Cobb, McClain, & Gravemeijer, 2003; Gil & Gibbs, 2017; Moritz, 2004; Zieffler & Garfield, 2009). This research has given insight into the challenges that students face while trying to coordinate the associations between two variables. However, introducing a third variable brings in different challenges such as confounding. For this reason, only results from statistics education research focusing on students' thinking with three or more variables are presented in this paper.

Science education research has, more extensively than statistics education, covered multivariate reasoning and provides insight into some of the challenges students face trying to reason with more than two variables. Finally, cognition research on multivariate thinking provides insight into students' and adults' natural abilities in this area and some common misconceptions and challenges that persist, even after extended exposure to the material. The activities and assignments created were informed by current implementations of multivariate thinking in statistics education, science education, and cognitive psychology. The activities were designed to introduce them to multivariate thinking concepts, while the homework assignments provided insight into the growth of their multivariate thinking skills.

1.1 Description of the Study

The purpose of this study was to investigate the development of multivariate thinking skills in introductory level statistics students through a series of activities and assignments. The activities were designed to promote multivariate thinking with consideration of previous related research. Studies suggest students primarily consider bivariate relationships when faced with multivariate data, making it difficult to get them to reason about how the relationships among multiple variables may be working together to affect an outcome variable. Because GAISE promotes teaching students to consider that an association between two variables might be affected by other variables, these activities were designed to encourage and develop multivariate thinking in this way. The study aimed to answer the following research questions:

1. How does students' multivariate thinking develop as they take part in a series of activities designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester?

2. What challenges surrounding multivariate thinking persist after taking part in the *intervention? Do any new challenges emerge after the completion of these activities?*

To answer the first set of research questions, student homework assignments were evaluated for evidence of growth in multivariate reasoning. Each assignment contained at least one question that required the student to coordinate effects of multiple variables. The second set of research questions were answered through evaluation of students' work on a final homework assignment that asked questions pertaining to all they have learned about multivariate thinking. This short response assignment required students to answer questions targeted at multivariate thinking skills and common difficulties. After the students completed the assignment, three students were interviewed about their responses to provide further evidence of their multivariate thinking abilities.

1.2 Structure of the Dissertation

Chapter 2 summarizes relevant literature pertaining to the study. First, multivariate thinking coverage in statistics textbooks and research around data visualization in statistics education is detailed. Then, multivariate thinking literature from cognitive sciences and science education is discussed. A table summarizing this literature presents the common challenges students face while reasoning with multiple variables. Next, causal inference is discussed highlighting the use of directed acyclic graphs to propose and predict causal relationships among variables, along with introductory work using this method in statistics education. The chapter concludes with a summary and critique of the literature and the problem statement guiding this research.

Chapter 3 presents the methodology of the study. It describes the development of the in-class materials with a full description of the context and learning goals of each activity and assignment. Next, the results from four rounds of think-aloud interviews to refine the assignments are described, noting relevant changes made to the assignments after each interview. Then, the plans for collecting and analyzing data from student's assignments, cognitive interviews with students at the end of the unit, and the in-class observations are detailed.

Chapter 4 presents the results of the study. First the results from the class observations of two students are presented, highlighting the student's multivariate thinking over the course of the unit as it pertains to the defined learning outcomes. Then, the results from the class assignments are presented along with a table of the percentage correct for each learning outcome on each assignment and a discussion of these values. Additionally, this section provides a discussion of common challenges or notable themes that emerged during the coding. Finally, the results from the three interviews with students about their last assignment are presented, again highlighting their final reasoning as it pertained to the learning outcomes and other notable results.

Chapter 5 discusses the results of the study to answer each of the research questions. Then, the limitations and implications for teaching and research are discussed. All materials created and a codebook with examples resulting from the qualitative coding are in the Appendix A and Appendix B .

Chapter 2

Review of the Literature

This study focuses on the development of undergraduate students' multivariate thinking. The chapter begins by defining multivariate thinking and examining its relation to current statistics and data science education research. Then, background for the study is provided by reviewing literature related to scientific reasoning, as this is closely tied to multivariate reasoning. Similarly, multivariate reasoning often involves causal reasoning. Relevant statistical causal inference work will be presented as it pertains to this study. Research on methods for teaching multivariate thinking are also reviewed. The chapter concludes with a summary and discussion of the literature.

2.1 Multivariate Thinking in Statistics Education

Multivariate thinking is an important skill for developing statistical thinking. This section provides an overview of current research surrounding multivariate thinking in statistics education. Given that the terms "reasoning" and "thinking" have often been used interchangeably in the literature, they will also be used interchangeably in this paper (delMas, 2004). Multivariate thinking requires consideration of the relationships among multiple variables (Mason & Young, 2004). An in-depth description of multivariate thinking skills provided by Adams, Baller, Jonas, Joseph, & Cummiskey (2021) explains: "[m]ultivariable thinkers can employ an intuitive sense of concepts such as confounding, mediation, association, interaction, and causality to create a more complete understanding of relationships in their data" (p. S125). The skills described by Adams et al. are increasingly recommended as learning outcomes for introductory statistics courses (Committee, 2016; Horton, 2015).

The Guidelines for Assessment and Instruction in Statistics Education (GAISE) express the importance of students developing multivariate thinking skills to help make informed decisions and think critically with data. They suggest providing students' experiences with multivariate data such as visualizing different trends in disaggregated data and reasoning about relationships among variables including confounding and causal relationships (Committee, 2016).

However, introductory statistics courses often only give students experience in reasoning about properties of one variable or the association among two variables. Past research in statistics education placed a focus on bivariate and covariational reasoning (e.g., Batanero, Estepa, & Godino, 1996; Batanero, Estepa, Godino, & Green, 1996; Cobb et al., 2003; Gil & Gibbs, 2017; Moritz, 2004; Zieffler & Garfield, 2009). This research gave insight into challenges that students face while trying to coordinate the associations between two variables. However, introducing a third variable brings in different challenges such as understanding confounding. For this reason, only results from statistics education research focusing on students' thinking with three or more variables are presented in this paper. The next section will focus on the extent that multivariate thinking is incorporated into introductory statistics curriculums.

2.1.1 Multivariate Thinking Curriculum

In today's world, seemingly endless streams of data drive decision making about everything from politics and science to shopping. While this data is typically observational and multivariate, introductory statistics courses do not often give students the experience of working with this data in a meaningful way (Ridgway, Nicholson, & McCusker, 2009; Schield, 2004). However, some statistics courses that added activities featuring interactive visualizations and rich, multivariate data sets to give students more varied experiences with data are discussed in the next section.

2.1.2 Using Data Visualization to Support Multivariate Thinking

Given that multivariate thinking is not always explored in current introductory statistics courses, some instructors and researchers have successfully engaged students in exploring relationships in multivariate data using interactive visualizations (e.g., Ridgway, Nicholson, & McCusker, 2007; Sutherland & Ridgway, 2017; Valero-Mora & Ledesma, 2011). Prodromou (2014) posed a research question to 14–16 year-old students inquiring about where they would prefer to live, based on a multivariate data set. The students were given an interactive visualization to help them investigate the covariation of the data to help them make an informed decision. This study suggests that even without scaffolding, the easy-to-use interactive visualization helped the students arrive at an informed decision about the research question.

However, Gil & Gibbs (2017) study of covariational reasoning using multivariate data in interactive graph software indicated that students still had trouble considering more than two variables at a time. After completing a series of activities designed to promote reasoning about covariation, students' responses indicated they typically thought about only two variables at once. Gil and Gibbs suggest more scaffolding to promote more multivariate thinking, even in an easy-to-use interactive software environment. Both Prodromou's and Gill and Gibbs's studies proposed that students might reason better with three variables when the third variable was time. This indicates that time could be a reasonable third variable for students to consider when trying to promote thinking beyond two variables.

While the previous studies focused on easy-to-use software to simplify introducing multivariate thinking to students, other studies have used R packages that are less interactive and require computing skills. For example, Wang, Rush, & Horton (2017) described an activity used in the first week of class to get students started creating univariate, bivariate, and multivariate graphs using the MOSAIC package (Pruim, Kaplan, & Horton, 2017). They argue that this early introduction to multivariate graphs, though not without difficulty given the high bar of learning programming, gets students introduced to the ideas early on and builds their interest for future work in the course. Student reviews of the activity indicated that they enjoyed this immersive initial experience, but some struggled with describing the graphs or choosing which graphs were appropriate for their data. However, students showed potential to reason about and create graphs with multiple variables early in a course using this activity.

Similarly, Adams et al. (2021) encouraged getting students to work on computing to facilitate their multivariate thinking skills. Adam's research team recognized that there is a delicate balance between having enough computing for students to tackle interesting and relevant data problems with many variables and having too much computing causing students to feel overwhelmed. They recommended the Tidyverse package (Wickham et al., 2019b) to facilitate this in a user-friendly way by focusing on a limited number of functions. They argue this helps the students not get lost in the code. Additional recommendations for teaching computing to support multivariate thinking included include summarizing conditional distributions, visualizing multiple variables, and learning multiple regression.

2.1.3 Summary of Multivariate Thinking in Statistics Education

Multivariate thinking has been defined as a learning goal for introductory statistics students, though it has yet to be broadly incorporated into introductory statistics courses. The research on teaching multivariate thinking has largely focused on multiple regression or interactive visualizations using point and click software. Some instructors used programming languages to introduce students to early computation skills and multivariate data. These instructors have found this somewhat challenging, but not impossible, if both concepts are thoughtfully introduced using a limited number of functions through packages like Tidyverse and MOSAIC. Primarily, studies in statistics education research have not focused on student's multivariate thinking, but instead on pedagogical decisions related to possible introductions to multivariate graphs and analysis (e.g., software choice and content).

Though statistics education research has more recently delved into multivariate reasoning, science education and cognition research have, more extensively than statistics education, investigated multivariate reasoning. Studies in these fields provide some insight into the nature of reasoning with more than two variables.

2.2 Cognition and Science Education Literature

This section discusses the literature surrounding multivariate thinking as it pertains to scientific reasoning. The research describes students' and adults' natural abilities in this area and some common misconceptions and challenges that persist, even after extended exposure to the material.

2.2.1 Scientific Reasoning

Multivariate thinking is considered a core principle of scientific thinking (Kuhn, Iordanou, Pease, & Wirkala, 2008). There is a rich field of research around the development of scientific reasoning in cognitive psychology, developmental psychology, and science education research. To bridge the divide among these different research areas with overlapping interests, Zimmerman (2000) reviewed this literature with the goals of: 1) providing an overview of literature that is related simultaneously to developmental and cognitive psychology 2) noting the changes in the focus of this research over time, and 3) discussing how literature from cognitive and developmental psychology can be put into practice by science educators. In Zimmerman's article, the terms scientific thinking, scientific reasoning and scientific investigation appear to be used interchangeably.

Zimmerman describes two different approaches to studying scientific reasoning: domain specific and domain general. The domain specific approach is focused on scientific concepts in a particular field of science (e.g., chemistry, physics, biology). Because the studies in this area are tied so closely to their field of science, they are not related to the multivariate works being reviewed here. In contrast, the domain general approach focuses on the reasoning strategies involved as students apply methods, design and critique experiments, and evaluate evidence from experiments. Often these reasoning strategies require multivariate thinking.

Research in the domain general approaches to scientific reasoning studied the reasoning required to solve science problems across contexts aiming to study the reasoning skills that transferred across the discipline. To do this, researchers often asked participants to ignore the context of the problems they were working on to gauge general thinking instead of the participants' content knowledge. Zimmerman explained that recent research has found the context of the task cannot easily be ignored and will likely have some effect on the reasoning of the participants. Newer research in this area aims to study conceptual knowledge and strategies of participants as they work through scientific reasoning in a "simulated discovery context". In these studies, the participants explored a hands-on or virtual simulation of an experiment to "discover laws or generalities in the multivariate causal system through active experimentation" (p. 139). When this field of research focused on multiple variables, it was often through studying students' reasoning on experiments using control of variable techniques.

There has been extensive research on coordinating the effects of multiple variables through studying how students learn to control variables for experimentation (Chen & Klahr, 1999; Robert F. Lorch et al., 2017; Robert F. Lorch et al., 2019; Strand-Cary & Klahr, 2008; Wood, 2015; Zimmerman, 2000). The control of variable strategy (CVS) is a particular way of designing experiments so that only one variable is manipulated at a time to allow causal claims through unconfounded experiments. Chen & Klahr (1999) argued that being able to design these experiments and understanding the logic behind why these allow for causal claims to be made are both key skills in developing scientific reasoning.

Schwichow, Croker, Zimmerman, Höffler, & Härtig (2016) conducted a meta-analysis on the research surrounding CVS. They analyzed 72 studies using inclusion criteria that necessitated the study be written in either English or German, was science oriented, used a comparison or control group, excluded students with learning disabilities, reported only test results that were about CVS (no other science skills), and the final assessment required students go beyond stating the rules for controlling variables. The final assessments required students to design, correct, or choose a correct experimental design with confounding or unconfounding variables.

Schwichow et al. collected information related to the design of the study, age of students, type of instruction, and type of assessment used. Notable results from the study indicated that the age of the students did not have a significant impact on their CVS abilities. The authors argued that this implies there is no set age at which to optimally teach CVS. But they offer the disclaimer that because different teaching strategies were used across studies, we cannot generalize that all teaching strategies will be effective with all ages, indicating that more research is needed to determine if there are optimum strategies for teaching CVS at different ages.

Although the study did not find any evidence that a certain amount of scaffolding or support is needed for students to learn CVS, use of cognitive conflict did impact the level of CVS ability achieved. Typically, cognitive conflict is evoked by providing the students with new information at odds with their current thinking, resulting in student reassessment of their current thinking to adjust to the new information presented. In CVS studies, cognitive conflict often is induced when instructors present a confounded experiment or interpretation of results to students in hopes they notice that the claims made are not justified. Most often the strategies that successfully invoked cognitive conflict were introduced by instructors imploring the students to think about the design of the experiment. Teaching students to think through study design and conclusions based on the design is a worthwhile cause in promoting causal reasoning and multivariate thinking.

This area of research, though still full of questions, has been a fruitful focus of research on scientific reasoning thus far. However, Kuhn argued that there is more to scientific thinking that should be studied (Kuhn, 2007; Kuhn, Ramsey, & Arvidsson, 2015). Kuhn et al. (2008) identify three other skills that are key to developing scientific reasoning: the ability to argue with scientific evidence, an understanding that science "laws" are developed by humans, and the ability to reason causally about the effects of several variables on an outcome. A singular focus on reasoning about control of variables is limiting; only considering the effect of one variable on an outcome discounts the possibility of other variables interacting with the outcome variable (Kuhn et al., 2008; Kuhn et al., 2015). The next section focuses solely on multivariate thinking research from the scientific reasoning literature.

2.2.2 Multivariate Thinking as a part of Scientific Reasoning

Multiple studies done by Kuhn et al. (2015) focused on developing multivariate reasoning in both children and adults. This research suggests that reasoning with multiple variables, associated with an outcome variable, is difficult regardless of age. Abdelhadi (2016) confirmed these challenges in a study investigating high school chemistry students' ability to reason with multiple variables.

One study by Kuhn et al. (2008) indicated that students struggled to consistently associate causal variables with an outcome variable. Another finding from this study determined that the participants distinguished between the level of the variable and the variable itself when determining causality. In the example given in the 2008 paper, students did not point to the amount of snow as a causal variable affecting the chance of an avalanche; instead, they only selected heavy snow as a possible cause of an avalanche.

In Kuhn et al. (2015), middle school students showed improved reasoning with multiple variables when given thoughtful problem-based activities in both a short and an extended intervention. Students were given data to investigate research questions and participated in class discussions that prompted consideration of multiple variables. Students demonstrated progress toward implicating multiple variables in affecting the outcome. However, the study design did not provide proof that long-term retention of multivariate reasoning developed from these activities.

Abdelhadi (2016) studied high school chemistry students' multivariate reasoning abilities. The students were given a research question and investigated relationships among multiple variables to determine which variable caused a change in an outcome variable. Students struggled to come up with a valid justification for their choice of variable affecting the outcome, even if they identified the correct dependent variables. However, students provided a better justification for their choice if they spent time weighing the evidence for each variable. Similarly, adults in a study by Kuhn et al. (2015) had trouble considering the size of the effects of variables when investigating which variables most likely caused a change in an outcome variable.

Other common challenges identified in these studies are the persistence of a belief bias affecting the variables participants chose as significantly affecting the outcome and participants not considering additive effects of variables on the outcome. Though multivariate thinking was the focus of the previous studies, another recurring factor in students' reasoning was their determination to focus on only one variable as the sole cause of change in an outcome variable. This indicated causal reasoning abilities are intertwined with multivariate reasoning abilities, which has been discussed in other studies (e.g., Kuhn et al., 2015; Zimmerman, 2007). A summary of this literature is provided in Table 2.1.

Table 2.1
Challenges to Developing Multivariate Reasoning

Challenge	Literature
Difficulties with Respect to Conclusions	

(Table continues on next page)

Table 2.1
Summary of Questions Mapped to Learning Outcomes (continued)

Challenge	Literature
Students often attribute causality to a single variable	Abdelhadi, 2016; Casparo & Grulich, 2019; Kuhn, 2007; Kuhn et al., 2015; Ridgway et al., 2007
Students made bivariate conclusions even after the study	Gil and Gibbs, 2017
Students inconsistently attribute causality to variables	Abdelhadi, 2016; Kuhn, 2007; Kuhn 2008; Kuhn et al., 2015
Students confuse levels of variables with the variables themselves	Kuhn 2008
Difficulties with Respect to Size of Effect	
Students could more easily describe the effects when both independent variables made a noticeable difference, but struggled when there were smaller impacts from the independent variables	Ridgway et al., 2007
Adults did not consider size of effects of variables on outcomes	Kuhn et al., 2015
Difficulties with Respect to Context	
Students had trouble transferring their knowledge between tasks, even using the same formatting of items	Kuhn, 2007
Students used prior knowledge for conclusions in place of evidence at hand	Abdelhadi, 2016
Adults' responses likely included instances of belief bias from previous context knowledge	Kuhn et al., 2015

2.3 Causal Inferences

A primary goal of multivariate thinking as expressed by GAISE is to study the relationships among variables (Committee, 2016). In the previous section, when investigating relationships among multiple variables the studies in cognition and developmental psychology often investigated participants' reasoning with multiple variables to determine causality (i.e., determine which variables caused a change in an outcome variable). Their studies largely made use of control of variable strategies to allow causal inferences. Similarly, many introductory statistics courses teach students to make causal conclusions only if the study design used random allocation. Much like control of variables, the random allocation allows us to create identical groups (on average) to apply our treatment variable and measure the differences between groups. Thus, if we see a big enough difference in groups, we can infer we have enough evidence to conclude the difference is likely caused by the treatment and no other factors. See Fry (2017) for a full review of how this is typically taught in introductory statistics courses. Though this has long been the "gold standard" for causal claims, it is not the only way one might find evidence of causality.

As students in introductory statistics courses start working with multivariate data, much of it may be observational. Analysis of multivariate observational data at this level typically involves analyzing graphs and perhaps conducting multivariate regression in more advanced introductory courses. Many students coming into introductory statistics courses are likely familiar with the expression "correlation is not causation". But students entering the class might not exactly understand the meaning of this phrase without more explicit instruction (Fry, 2017). Research has shown that humans naturally want to make causal claims from observational data, particularly if they suspect there is a causal relationship before seeing any data (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Gopnik & Schulz, 2007; Kahneman, 2013; Kuhn, Amsel, & O'Loughlin, 1988).

"Correlation is not causation" may not be nuanced enough to capture the true nature of relationships among certain variables. When we see a correlation, it might mean that there is a causal relationship between two variables. The correlation itself does not prove the causation, but it does provide evidence that a causal relationship might exist between two variables and more study is needed to determine possible causation (Pearl & Mackenzie, 2018). Seeing that humans are inclined to think with causal structures and an overwhelming amount of data today is observational data, not experimental data, it is helpful to have a structure that makes investigating causal claims from observational data possible. There are differing philosophies among statisticians, computer scientists, and other researchers about how, and if, one can rightfully make causal claims with observational data alone. However, one method becoming more common in research from health, social sciences and economics focuses on causal inferences from observational data makes use of causal Bayesian networks. This method is discussed in the next section.

2.3.1 Causal Inferences with Observational Data

Judea Pearl is a computer scientist that has studied causal inference methods for work with artificial intelligence systems. One part of Pearl's research on artificial intelligence (AI) focuses on causal inference using observational data. His work is detailed not only in his academic papers (e.g., Pearl, 1995, 2009), but also through his book Causality Pearl (2000) and co-authored books such as Causal Inference in Statistics: A Primer (Pearl, Glymour, & Jewell, 2016) and The Book of Why (Pearl & Mackenzie, 2018). Given the utility of this work, there are growing fields of research advocating for and making use of these methods in health sciences and epidemiology (Greenland, Pearl, & Robins, 1999; M. A. Hernán, 2002; Miguel A. Hernán, Hsu, & Healy, 2019; Shrier & Platt, 2008; Suzuki, Shinozaki, & Yamamoto, 2020; Tu & Gilthorpe, 2012; Williams, Bach, Matthiesen, Henriksen, & Gagliardi, 2018) and textbooks developed for teaching his methods for social science and machine learning (Elwert, 2013; Morgan & Winship, n.d.; Peters, Janzing, & Scholkopf, 2017).

Pearl's method of causal inference involves identifying potential causal variables based on prior knowledge, identifying causal effects among the variables, and determining which variables should be controlled to guide making a claim with observational data. Pearl explains in *The Book of Why* that though this method has gained popularity through Pearl and his contemporaries, this strategy is built on work by Wright in 1920 (Pearl & Mackenzie, 2018). Wright studied genetics using path diagrams (also referred to as path analysis) to find path coefficients, where path coefficients signified the strength of causal effects of various guinea pig genes. Wright's path coefficients are typically described as the variability in an outcome variable explained by another variable.

These diagrams and analysis did not gain traction in the research applications until the 1960s when they were picked up by social scientists studying changes from policy implementation. Sociologists ultimately renamed path analysis structural equation modelling (SEM) and continued to use the path diagrams, however they left out the justification of causal claims made by Wright.

Throughout much of statistics' history, researchers have wanted to avoid causality in favor of letting the data speak for itself, claiming that since causality is a subjective construct, it cannot compete with the objectivity of data. But Pearl argues that we are limited if we do not consider making causal claims based on knowledge from context and subjective means.

2.3.1.1 Causal Diagrams.

Central to Pearl's framework for determining causality is the idea of using graphical models called directed acyclic graphs (DAGs) or causal diagrams. DAGs are used to portray the hypothetical relationships of the many variables acting on an outcome variable. They are representations of a system of multiple variables. Creation of these diagrams is often based on preexisting knowledge or theory about the causal mechanisms of the system under study. In the diagrams each variable is represented as a node and the causal relations are depicted by arrows or edges connecting the nodes. The structures of the arrows and nodes can be used to help determine which variables should and should not be controlled for when performing analysis and provide an efficient means of communicating relationships among variables (Gopnik & Schulz, 2007; Pearl, 2000).

Using DAGs as a method to aid determining which variables to control stands in stark contrast to traditional statistical analysis on observational data in which as many variables as possible should be measured and controlled for during analysis. Pearl argues that too often variables are controlled for that need not be controlled for in an analysis, leading to incorrect estimation of effects on the outcome variable. Using DAGs helps identify which variables should be controlled for to help determine cause, and which variables might not help us determine cause. Figure 1 depicts a possible DAG on the left side of the figure and a possible contextual instantiation of the causal model on the right. In this case, Y represents an outcome variable that is affected by variables X and Z. The arrows pointing from Z and X into Y indicate that they cause changes in Y. For example, we might be attempting to model a person's wealth (instantiation for Y). If we think a person's income (instantiation for X) and their Inheritance (instantiation for Z) might affect their wealth, we could model this system using the DAG on the right side of Figure 2.1.

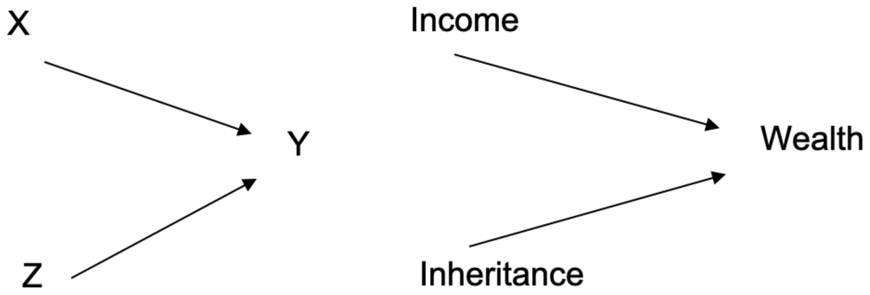


Figure 2.1. Example DAGSs

Undoubtedly, these graphs can be immensely more complicated. Suppose that we also want to consider the number of investments a person has and its effect on wealth. A DAG including this additional variable might look like Figure 2. We would expect the amount in investments (W in Figure 2.2) to be affected by both income and inheritance and this would also affect the final amount in wealth. As we can see adding more variables makes these diagrams more complex. Once all necessary variables are identified, the DAG can be used to determine which variables should be controlled for in analysis and which can be excluded.

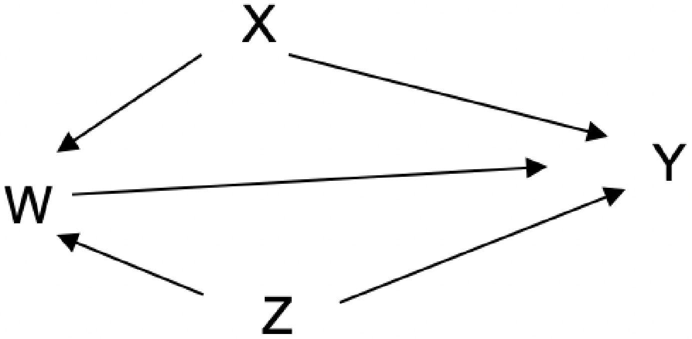


Figure 2.2. More complex example DAGS

DAGs can then further be used as more than just a visual reference for determining which variables to control for in an analysis. More advanced use of DAGs and causal inference can be done using probability calculations and *do calculus* (a way to determine a counterfactual probability given a complex DAG) to obtain insight into counterfactual claims. However, this method is beyond what one could hope to introduce in an introductory statistics course, and thus is beyond the scope of this project (for some further information see Elwert, 2013; Pearl, 1995, 2000; Suzuki et al., 2020)

2.3.1.2 Critiques.

As previously mentioned, there are other methods for causal inference analysis. Another common framework for causal analysis is the potential outcome framework from Rubin (Holland, 1986; Imbens, 2020; Rubin, 1974). Imbens (2020) provided a comparison of the potential outcome framework and Pearl's DAG method in response to the growing interest in causal inference. The author's main critique of DAGs is that the example studies provided by Pearl are often overly simplified with little contextual guidance or complexity. Imbens argues that there are not enough convincing empirical studies that use the DAGs method to justify a widespread adoption of it currently. Many more empirical examples and studies have been undertaken using the potential outcomes framework provided by Rubin, often making this the more appealing method to researchers (Imbens, 2020; Powell, 2018)

Another critique of Pearl's method is that the models used for Pearl's analysis require more assumptions and may be difficult when large numbers of variables are needed to answer a research question. However, Imbens (2020) argues that if the assumptions hold, Pearl's method would offer more insight into a variety of research questions, including those pertaining to counterfactuals, which other methods shy away from. Methodological and philosophical preference differences may always exist between researchers as new causal inference analysis methods emerge, but regardless of these differences, Pearl's methods using DAGs are increasingly used for research and being incorporated into curriculum for social, economic, and physical sciences. The next section talks about their implementation at the introductory statistics level.

2.3.1.3 DAGs in the Undergraduate Statistics Curriculum.

As discussed in a previous section, introductory statistics courses have limited, if any, introduction to multivariate thinking; they are similarly limited in their discussion of making causal claims. This discussion is often focused solely on random allocation of treatments to groups. However, some educators are expanding their course content to broaden the introduction to causal claims and make way for more multivariate thinking. This section discusses some of the most recent curriculum implementations.

Some implementations of this have made it into introductory statistics course books. Notably, Kaplan (2017) emphasizes the importance of considering causality in contexts outside of experiments. In the introduction to his book, *Statistical Modeling: A Fresh Approach*, he introduces the idea of counterfactuals and stresses that modern experiments are as close as we can get to counterfactuals but explains that experiments are not feasible in all disciplines. This makes it important to learn about the insights we can and cannot glean from our observational data. He argues that this allows us to make inferences and plan future work to fill in any gaps remaining in our analysis.

He introduces "hypothetical causal networks" with nodes and links (his version of the DAGs previously described) for proposing relationships among variables. He describes how the hypothetical nature of the diagrams encodes what we think the mechanisms are for the causal links among the variables. He then explains how to use these diagrams to help choose variables for a linear model for analysis of the relationships among the variables.

First published in 2009, this textbook provides a more extensive discussion and framing of causal thinking among multiple variables than other introductory statistics textbooks. Other implementations that use DAGs or causal inference have since been incorporated into various classroom activities in introductory statistics courses and are discussed next.

Cummiskey et al. (2020) used DAGs to facilitate multivariate thinking while teaching multiple regression. Following Pearl and MacKenzie's framework for causal inference (Pearl & Mackenzie, 2018), their approach used DAGs to help students reason and communicate about relationships among variables. They argue these graphs provide a useful visual that helps students identify confounding variables and explore the investigative data cycle. They found the diagrams helpful for engaging students in a discussion about how and why they drew the relationships among variables a certain way.

Lübke, Gehrke, Horst, & Szepannek (2020) describe another introduction to causal inference for use in an introductory statistics course. They used DAGs to explore various types of relationships among three variables using simulated data and multiple regression. They provided activities consisting of simple examples to demonstrate the effects of 1) unnecessarily adjusting for variables (that should not be adjusted for) which introduces bias, 2) not adjusting for variables (that need to be adjusted for) which introduces bias, and 3) exploring experiments. They argue that, though these are simplistic examples that do not use real data, they still represent real situations that provide a starting point for students to get some experience working with causal inferences.

2.3.2 Summary of Causal Inference

Pearl's causal inference methods using DAGs as a modelling tool to aid analysis visualization and to align theory and analysis are becoming more commonly used. As causal inference is used more in practice, it is important to teach these methods to researchers and those working with data. These methods are becoming increasingly common in AI, epidemiology, and social sciences. However, students in introductory statistics courses typically do not get much exposure to ideas of multivariate reasoning and causality.

Given arguments to include topics of multivariate reasoning and causality in introductory statistics courses, some have begun implementing some introductory versions of these ideas through textbooks and class activities. But with only a handful of recent studies implementing these methods at the introductory statistics level, there is not much empirical evidence showing what students understand about creating or analyzing DAGs and drawing conclusions from them. More research is needed to find out about students' reasoning around this idea. Additionally, information about how we can best teach students to consider multiple variables for analysis is needed. This is where we turn in the next section.

2.4 Supporting Multivariate Thinking

One way to promote multivariate thinking is through guided discovery learning (Brown & Campione, 1994). This type of learning helps students work through constructing their own knowledge of the subject, while still working within a guided framework that is attainable given their current level of knowledge (Committee on How People Learn II: The Science and Practice of Learning, Board on Behavioral, Cognitive, and Sensory Sciences, Board on

Science Education, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine, 2018). Activities to promote this type of learning include scaffolding for the students to build up their knowledge.

Scaffolding and active learning have been incorporated for teaching multivariate thinking in previous studies (e.g., Kuhn et al., 2015). One study by Caspari & Graulich (2019) created a single guided activity to develop student multivariate reasoning using scaffolding for a chemistry education research study. Their results indicated that the scaffolding helped students consider the effects of multiple variables and better justify their responses when determining which of multiple variables affected an outcome. The reported effectiveness suggests future research is needed to see if students can build up their multivariate thinking through this method over time.

2.5 Discussion

This literature review thus far has detailed some of the important work to date on multivariate thinking in statistics education, cognitive science, and developmental psychology. This review also covered the importance of including casual reasoning in the introductory statistics curriculum. The first part of this section provides a summary and critique of that work. The last part of this section discusses future directions for this research to motivate the study presented in Chapter 3.

2.5.1 Summary and Critique

Multivariate thinking has been identified as a necessary skill for living in our data rich world. Increasingly, students learning outcomes pertaining to multivariate thinking are being introduced into introductory statistics courses. However, we do not know much about how to effectively teach students these skills. Initial research in statistics education has pointed to some new ideas such as teaching with time as a third variable before introducing other more complex third variables, introducing students to multivariate plots, using interactive software to capture students' attention, and using less function heavy and more intuitive packages in R to help decrease the cognitive load required to create multivariate graphs. Though focusing on technology integration into curriculum is a worthwhile goal, creating the graphs themselves is not the only skill that students should have coming out of an introductory statistics course. It is also important to reason effectively and communicate what is in multivariate graphs. We do not have much statistics education research on students' reasoning with more than two variables at a time, but this type of thinking has been studied in science education.

From cognitive science and science education literature we know that reasoning with multiple variables poses many challenges for students. They confuse levels of a variable with the variable and often introduce belief bias into their analysis and conclusions. When given many variables to consider, students often only consider bivariate relationships, choosing to focus on one variable that they believed caused a change in the outcome variable. While students place a focus on the causal narratives they can derive from their variables, traditional statistics has limited ways of determining causality among variables.

Pearl's framework for determining causal relationships offers a way to answer casual questions using prior knowledge to one's advantage. Though most powerful in the hands of subject matter researchers and someone proficient in the methods needed to calculate the effects of each variable on the outcome, DAGs can provide a starting point to thinking about causal relationships and confounding variables. Some introductory statistics courses have started to incorporate these models into their classroom when teaching multivariate thinking and multiple regression as a method of variable selection and communication about relationships among variables. However, not all courses cover multiple regression or similar advanced computing methods, and we have yet to see any implementation of DAGs at a lower introductory level.

If multivariate thinking is going to be a learning outcome in introductory statistics, it is important to be thoughtful in incorporating it into the curriculum. More insight is needed into what students are thinking as they work through multivariate data problems. Though it may be easy to incorporate it into a unit on multiple regression, many introductory statistics courses do not cover multiple regression. As such, many students that do not take more statistics courses are not exposed to reasoning with more than two variables. Yet the research has shown both that it is possible to start developing this skill earlier in the curriculum and it takes time to develop. Further research is needed in this area that investigates introductory statistics students' reasoning with multiple variables that are not only categorical (as previous cognitive research has focused on) but also quantitative variables.

2.5.2 Problem Statement

Previous research suggests students primarily only consider bivariate relationships when faced with multivariate data, making it difficult to get them to reason about how the relationships among multiple variables may be working together to affect an outcome variable. Because GAISE promotes teaching students to consider that an association between two variables might be affected by other variables, research should be conducted to learn more about student's current thinking with multiple variables. No current studies in statistics education give insight into how students' learning develops around thinking with multiple variables and causal reasoning at an introductory level. However, to successfully implement this topic into courses we need to know more about students' reasoning with multiple variables.

Chapter 3

Methods

3.1 Introduction

The purpose of this study is to investigate undergraduate students' reasoning as they work through a series of activities and assignments designed around multivariate thinking. Specifically, it aims to answer the following research questions:

1. How does students' multivariate thinking develop as they take part in a series of activities designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester?

2. What challenges surrounding multivariate thinking persist after taking part in the intervention? Do any new challenges emerge after the completion of these activities?

This chapter begins with an introduction to the course in which the study took place. Then, the course materials are described with emphasis on how they address the student learning outcomes (SLOs). Next, think-aloud protocols and course implementation of the materials are outlined. Finally, data collection methods and analyses are described.

3.2 Overview

In-class activities and assignments were designed to promote multivariate thinking with consideration of the research outlined in Chapter 2. To answer the first set of research questions, all student homework assignments were qualitatively evaluated for evidence of growth or change in multivariate reasoning. Additionally, one student in each section was observed throughout the unit as they worked on the in-class activities. This data was qualitatively analyzed to provide insight into the students' reasoning as they worked through the unit. The second set of research questions were addressed through evaluation of students' work on a final homework assignment that contained questions pertaining to all SLOs covered in the multivariate thinking unit. After the students completed the assignment, some student volunteers were interviewed about their responses to provide further evidence of their multivariate thinking abilities. These interviews gave further insight into research question 2.

3.3 Course Structure

For this study, students from the two sections of an introductory course on communication and visualizations were recruited for participation in the Fall of 2021. Section 001 contained 18 students that met on Tuesdays and Thursdays from 9:45am to 11:00am. Section 002 met Mondays and Wednesdays from 9:45am to 11:00am and contained 25 students. Both sections were taught by the same instructor that has taught the course since its debut in Fall 2017. Until this project, the content and activities in the course were relatively unchanged from their original versions.

Given the continued proliferation of the COVID-19 pandemic, the course was taught in a flexible format that gave the instructor and students the ability to work in-person or remotely. Except for one entirely remote class, the instructor was always present in the classroom and over Zoom version 5.8 (Zoom Video Communications, 2022), a video conferencing platform, during class time. On average, approximately five to ten students met in the classroom on a given day, while there were another five to ten students working on Zoom for a portion of the class period. Other students did not come to class remotely or in person, but instead completed the class activities on their own time and occasionally asked questions via email or Zoom.

The course is aimed at undergraduates in their first and second year of university and allows them to earn a mathematical thinking credit needed for graduation. Though a few students in the class had previously taken a computer science course, typically students have no prior computer programming education and at most high school math experience. The course content centers around creating and communicating about visualizations using R's **ggplot2** package (Wickham, 2016). The course content starts with histograms and bar charts (one to two variables), followed by line plots and scatterplots (three or more variables), and finally with maps.

The course consists of students completing weekly readings, group discussions, in-class activities, and homework assignments. The readings and asynchronous class discussion boards provide background information about creating graphs, communicating about data, and working in RStudio. The students also use the discussion boards for building up "cheat sheets" for reference when coding. The readings and discussions always cover similar content to the activities and assignments but were not created as a part of the multivariate thinking unit and are not discussed in any further detail for this study.

In-class, self-guided activities allowed students to work in groups, but each student submits their own work. Homework assignments were completed individually, most often outside of class time. In a typical class, the instructor made announcements and gave a brief overview of the class activity for the day. The students spent the remainder of the class working on the activity and were free to leave once they had submitted their work.

3.4 Development of Multivariate Thinking Materials

Ten activities and three assignments make up the multivariate thinking unit created for this study (see Appendix A). Some course activities and assignments from previous semesters were updated to incorporate more multivariate thinking questions, while others were entirely new to the course. Older course materials had students create multivariate visualizations but were adapted by asking students to describe the multivariate relationships and potential causal or confounding relationships featured in the graphs. They were also updated to ask about the nature of the data (i.e., experimental or observational) and to have students create a directed acyclic graph (DAG) to hypothesize about the nature of relationships among variables in a dataset. New activities and assignments were also created with these same foci. This unit was implemented in weeks 5-10 of a 15 week course. Course material covered before the unit focused on introducing the basics of uploading data and creating histograms and bar graphs in R. After the unit ended the focus shifted to creating and describing choropleth and point maps.

GAISE provides a possible trajectory for developing multivariate thinking including key points such as identifying observational studies, learning to be wary of cause-and-effect conclusions, and learning to consider potential confounders. These topics were explored throughout the activities, while also keeping in mind other key GAISE recommendations: using open ended questions, real data, and complicated real-world questions to engage students. Though the literature review suggested using interactive graphs to introduce multivariate concepts to students, these activities focused on creating static graphs in R because that is the focus of this course.

The assignments and in-class activities were subjected to multiple rounds of feedback and revision. SLOs for the multivariate thinking unit and the development of activities and assignments are described next.

3.4.1 Development of In Class Activities

In-class, guided discovery activities were designed for students to complete in small groups. This structure and format of class work was familiar to the students, as it is typically the structure of all the course activities throughout the semester. Given this learning environment, the activities were created with a social constructivist theory of learning in mind (Vygotsky & Cole, 1978). Activities were designed to allow students to work together to construct knowledge about creating multivariate plots and reasoning about relationships among variables, while allowing them to build on knowledge from previous activities and their contextual knowledge. Because the context of activities has been known to influence multivariate reasoning in past studies (e.g., Abdelhadi, 2016; Kuhn et al., 2015), considerable thought went into providing familiar contexts for this diverse group of students.

Seven SLOs were created to align with GAISE recommendations for promoting multivariate thinking, while also encouraging hypothesizing about potential causal relationships through the creation of DAGs. The class materials were designed to cover the following SLOs:

At the end of the unit on multivariate thinking the students should be able to...

1. Create graphs displaying the relationships among three or four variables in one plot
2. Explain the relationships among three to four variables using graphs
3. Identify data as observational
4. State the limitations in making causal claims with observational data
5. Create DAGs to guide analysis of relationships among variables
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables

7. Develop a hypothesis about variables not investigated and their relation to an outcome variable

Once developed, the activities were given to two statistics education advisors and the current instructor of the course for feedback. Minor revisions were made to add more scaffolding for the students and adjustments were made by request of the instructor to ensure the material aligned with the course SLOs and sequencing of the material. The activity and its corresponding learning objective are described next.

3.4.1.1 Activity 1: Hexadecimals.

In this activity students created bar graphs, stacked bar graphs, and side-by side bar graphs using **ggplot2**. They were instructed to customize the plots by changing the *y-axis* labels, customizing the legend title of a graph, changing the width of a figure, and creating custom colors by converting RGB color codes into hexadecimals. Though the conversion of RGB to hexadecimals was the most time-consuming part of the activity, it was not included to facilitate multivariate thinking, but is an important course learning goal for students to earn their mathematical thinking credit.

This activity was only modified slightly from its original version, given that it contained many crucial course SLOs. Two questions were added after the creation of each graph that prompted the students to reason about the relationships among the variables in the graphs they created. For the first graph, students were asked to consider the relationship between cell phone operating systems (iPhone or Android) and music streaming preference. To investigate this relationship, they created a stacked bar chart displaying music streaming preferences colored by cell phone type. This is the first occurrence of the students' making inferences from a plot they created in the course.

Data from the Midlife in the U.S. survey (Ryff et al., 2019) was introduced next in the activity. This dataset contains variables pertaining to the behavior, psychological state, and health of middle-aged Americans. Students chose two variables to explore and then looked at the difference between the binary sex options (male/female) given in the survey. Then they made inferences about their chosen variables based on their plot.

3.4.1.2 Activity 2: Bar Charts from Summary Information.

In Activity 2, students created bar graphs from tables of counts for each level of a categorical variable. This requires different syntax than creating bar charts from data in the case-by-variable format that students have used thus far in the class. Both types of data structure were introduced to get students thinking about how the data structure affects how we use ggplot to create a graph. In this activity, students continued to practice using custom colors and converting them into hexadecimal format. They also learned how to change the font, size, and rotation of the *x-axis* labels. The first part of this activity did not change from the original format or content because these are all important course SLOs.

Additional tasks were added to the end of this activity using a Valentine's Day dataset. This dataset contains the percentage of Valentine's Day Gifts purchased over multiple years. To investigate the relationship between these variables the students created a bar graph with the percentage of people that gave certain gifts for Valentine's Day over three years. Once they had their final plot, they were given the same data plotted as a line graph. They were asked what aspects of the data each graph highlights best and to consider which they would choose to show the relationships among these variables. This part of the activity encouraged students to contemplate how we can use different types of plots for the same information, making some features harder or easier to extract. These skills were needed later in the course when students were given more open-ended prompts for creating graphs. This part of the activity also provided a transition into line plots, which are featured in the next activity.

3.4.1.3 Activity 3: Fan Cost Index.

Activity 3 focused on a Fan Cost Index dataset to study the relationship among three variables: season, fan cost index (FCI), and team. Though this activity was used in previous semesters, it was updated to include questions asking the students to reason about and interpret their multivariate visualization. The activity started with exploring the cost of attending games over time for sports teams in Minnesota in a scatterplot. Then they created their first line plot which shows how the FCI has changed over time for different sports teams in Minnesota. Students then used this graph to answer questions encouraging them to describe the variability of the change in FCI for teams over time by having them discuss overall trends and trends for specific teams.

Then the activity introduced them to another dataset that contains the FCI for all teams in the National Hockey League. They were asked to use this data set to create a second line plot to look at FCI over time with teams from the NHL, highlighting the Minnesota Wild team. To accomplish this, the students created another dataset and learned about more aesthetic features they can use in **ggplot2** for enhancing their plot to highlight this specific team. These plot aesthetic changes are all part of the SLOs from the course pertaining to data visualization aesthetics. The final question on this assignment, however, was meant to elicit multivariate thinking. This question asked for a comparison of how Minnesota's NHL team's FCI has changed over time compared to other NHL teams.

Line plots were originally introduced toward the end of the course, but some literature suggests that reasoning with "time" as a third variable might help provide a scaffold to help students reason with more than two variables (Gil & Gibbs, 2017; Ridgway et al., 2007). In updating the materials for this course, the sequence of introducing different graph types was considered. Given the suggestion from the literature, line plots were moved to be introduced before scatterplots. Technically, the students do create a scatterplot at the start of this activity, before creating lines based on the team variable, but they are only being asked to describe the general relationship in the plot, not discuss linearity, slope, or strength of relationship more formally. These features were explained in the activity introducing scatterplots, which were introduced after DAGs were brought into the unit in Activity 4: Directed Acyclic Graphs.

3.4.1.4 Activity 4: Directed Acyclic Graphs (DAGs).

In this activity the students considered potential causal relationships among multiple variables through drawing directed acyclic graphs. Until this point in the course, students were mainly focused on graph creation and R code, but in this activity, they used R only to submit the assignment and not to create any DAGs. This allowed students to focus only on this skill before they merged creating graphs in R and DAGs by hand in the next activity. This activity started with an example DAG in a simple context using plants. This example was discussed with the entire class in a five-minute lecture. Then the students broke into small groups to develop their own hypothesis about other variables that could affect plant growth and modeled those relationships using a DAG. A short class discussion followed in which students shared the DAGs they created and their reasoning behind them.

This process was repeated as students created DAGs in the context of social media followers and then one final DAG in the context of their choosing. The lesson ended with a larger group discussion about the complexities of the DAGs and how they could start to use them to help propose and model relationships among variables before looking at the data to investigate the relationships and refining the DAGs once they know more about the data.

3.4.1.5 Activity 5: Women in STEM.

In Activity 5: Women in STEM, students created and interpreted scatterplots by describing linearity, strength, and slope. The students started the activity by considering the relationship between women's income and the proportion of women in the career. The context and content of this activity were largely unchanged from their original version, except for the first part of the activity in which the students created and interpreted scatterplots. Their first scatterplot suggested that the higher the proportion of women in a particular career the less income they earn, on average. The students were asked to explore the claim that the type of STEM major that attracts women are the same majors that have lower income. To do this they added color to their scatterplot for the different types of stem majors (e.g., Physical Science, Health, Engineering, Computers & Mathematics, and Biology & Life Sciences). The students then reasoned about the clustering of colors they saw for different STEM majors. Finally, the students created a DAG to hypothesize about other factors that might affect the income of a woman in the STEM field.

The second part of this activity was left unchanged because it emphasized learning various aesthetics to make plots more attractive, such as adding text to the graph, changing the shape of the points, and altering the axis labels. The students also learned about adding annotations with arrows and using different themes for their plots.

3.4.1.6 Activity 6: High Peaks.

Activity 6: High Peaks was entirely new to the course and focused on creating and evaluating scatterplots using data on the High Peaks of the Adirondack Mountains. The activity explored four different variables (elevation, length of hike, time to hike, and difficulty) to determine what affects the difficulty rating of the hikes. Initially, the students made a prediction about which variables they thought would be related to the hike difficulty and they create a DAG modeling their prediction. Then they investigated two variables at once in a series of plots and make judgments about whether there is a relationship between the variables they are looking at. First, they created a scatterplot with time and difficulty, then with elevation and difficulty, and finally with length and difficulty. Of these pairs, it appeared that time and length have a relationship with difficulty, but the relationship between elevation and difficulty is less clear, which might contradict initial predictions the students make. Then the students created a scatterplot with difficulty, length, and time to investigate possible relationships among these three variables. Lastly, in the interest of time, they were given a scatterplot with ascent, length, and difficulty and asked about potential associations. The activity ended with considering a graph of difficulty, time, length, and ascent to consider the relationships among these variables. The students were tasked with drawing a final DAG but must only include variables they thought affect the difficulty of the hike based on the plots they created.

After they explored the relationships among the variables in this dataset they recreated a scatterplot with certain colors, themes, and titles to get more practice with these skills in R. This section of the activity was created to mimic a task from a previous course activity that had been removed to make room for this new activity.

3.4.1.7 Activity 7/8: World Data.

This activity was included in the previous course materials. However, it was updated to include more consideration of multivariate relationships and DAGs. The dataset used in this activity is from Gapminder ("Gapminder," n.d.) and contains information about different regions of the world. The activity started by asking the students to predict the relationships among three different variables and create a DAG to model those proposed relationships (life expectancy, fertility rate, and region of the world). Then, the students investigated the relationships among the variables using a scatterplot of life expectancy versus fertility rate colored by the region of the world. Next, they provided an updated DAG based on evidence from their plot to support the relationships they proposed in that DAG.

The activity continued to have them investigate a fourth variable for the first time. They added the population of the country to the plot by mapping the population to the size of the dot in the scatterplot. Using this plot, they answered a series of questions probing them to consider all the relationships displayed in this plot and determine which they think are potentially associated. From this logic, they created a final DAG.

In Part II of this activity, they learned new **ggplot2** functions to remove the title from the legend. Then they transitioned into working on the Gapminder website where this data set originated to explore the relationships among income, CO2 emissions, region, and population size. Their final task was to interpret any associations among these relationships and create a final DAG with justifications based on their plot.

3.4.1.8 Activity 9 Evaluation.

Similar to the previous activities, in Activity 9: Evaluation students explored factors that might affect the course evaluation scores of college instructors. This dataset contains many variables for the students to consider, however trying to look at them all at the same time makes it difficult to discern any relationships. This activity is unique in that it starts with critiquing a graph that displayed all five variables in the dataset at the same time. Students then explored the relationship among three variables at a time to determine which to include in their final polished graph. They also created a DAG with the final variables they thought were related to each other using evidence from their graph to justify their answers.

3.4.1.9 Activity 10 SAT.

This activity introduced Simpson's Paradox through visualization. To begin, students considered the relationship between SAT scores and teacher salary in each state. When these variables were plotted in a scatterplot, there appeared to be a negative relationship between them, which seemed counterintuitive to what the students often expected. Once they discussed potential reasons for this, they investigated further by adding in the percentage of students taking the SAT in each state. After adding this variable to the plot using color, the students were able to see that the relationship between SAT score and teacher salary then "changed direction" and appeared to be positive within certain ranges of students taking the exam. This allowed for more discussion about how stratifying on a third variable can reveal different relationships in the data and the implications of it.

3.4.2 Development of Assignments

As a part of typical coursework, students completed out of class homework assignments every other week. One of the previous assignments was updated to include multivariate thinking items. Two additional homework assignments were created to assess multivariate thinking in new contexts.

The third assignment was the final assignment for data collection in this study and was created to assess all learning goals in the unit. Assessments exist for statistical reasoning (e.g., Sabbag & Zieffler, 2015) and computer based assessments of multivariate reasoning (e.g., Ridgway et al., 2007), but none fit a visualization specific undergraduate course in statistics. For this reason, these assessments were not used in this study. This final assignment consisted of 15 short response questions for the students to demonstrate their knowledge on questions pertaining to multivariate reasoning. The items were written in accordance with constructed response item writing guidelines (Haladyna & Rodriguez, 2013). The questions were based on the SLOs specified above and targeted common challenges found in previous literature on multivariate reasoning.

The assignments went through multiple rounds of feedback and updates based on think-aloud interviews conducted with three statistics education graduate students and the course instructor. The final versions of the assignments are described next, and the changes made to them over time are discussed later in the chapter.

3.4.2.1 Assignment 1.

This assignment used data from the World Health Organization to look at tuberculosis deaths over time for different regions of the world. The students created several different line plots and manually created subsets of the data (in a similar way to Activity 3: Fan Cost Index) to meet the visualization requirements for the course.

To assess their multivariate thinking, they were asked about the nature of the data (observational) and the implications for studying that data (can't necessarily draw causal conclusions) in Questions 1 and 2. Three subsequent questions asked them to create line plots featuring the rate of tuberculosis deaths over a decade for each country in several different regions.

To assess the students' ability to consider three variables at once, they described their line plots in Questions 5 and 7. Students were probed to think further about variables not in the dataset in Question 6. For the first time on an assignment, they were asked to create a DAG in Question 7 and use their plot to justify their DAG in the final question. They also practiced other skills not related to multivariate thinking, such as creating detailed graphs with titles captions, special colors, and themes, and creating a dataset by taking a subset from another dataset. These skills likely required a similar amount of effort as the questions designed to assess their multivariate thinking.

3.4.2.2 Assignment 2.

This assignment focused on sample housing data from Zillow. Students created scatterplots using color and size and added a customized theme to their graph. To assess multivariate thinking, the students considered the nature of the data (it is observational), then created a DAG modeling which variables from the dataset they think affect the price of a house (number of bedrooms, age, and square footage). Then they drew a DAG to predict the relationships before creating a scatterplot to investigate further. Once they had created the scatterplot, they used it to assess the relationships among the variables and update their DAG.

From the graph, it appeared that age and square feet affected the price, but the number of bedrooms did not seem to have as strong of a relationship with price, which was likely surprising to students. There was also a potentially unexpected association between square footage and age of the house for the students to discover. These unexpected relationships might force them to think more about their DAG, update it, and provide thoughtful justification using evidence from their plot.

3.4.2.3 Assignment 3.

In Assignment 3 students investigated used car data from a study by Kuiper (2008). This was a more open-ended assessment than the previous two assignments. The students had more freedom in what variables they chose to explore. First, they looked at the relationship between price and mileage and then they considered this same relationship after faceting on type of car (i.e., sedan, convertible, etc.). Once faceted, the negative relationship between price and mileage, which was not clear in the first graph, became more distinct for each type of car. They were asked to describe how this relationship changed after faceting (if at all).

Then, students chose another variable they thought might be related to the price of a used car. They created a DAG to model the relationships among the variables they had chosen and price, mileage, and type. Then they created a plot with these four variables adding in the fourth variable in the manner of their choosing to create an aesthetically pleasing plot with titles, no NAs, using a new color palette, and theme.

Once they had their graph and interpretation of the relationships within it, they were asked to reimagine this graph and create a new graph with the same four variables but mapped to the aesthetics of the plot differently. They were asked to pick different themes, colors, labels, map new variables to the *x-axis*, color, shape, or facet differently, but to keep the price mapped to the *y-axis*. Next, they were asked to determine if their new plot or the plot they created previously displayed the relationships among the variables more clearly. This required them to use what they had learned about visualizations and communication so far to ensure their graph was highlighting the intended relationships.

Finally, the students used their graphs to help them answer a prediction question. This last question was an extension of the knowledge beyond what they had done in class previously. They were given a scenario where their friend was considering buying a used car (of the same year and manufacturer as those in this dataset) for $40,000. They had to give

Table 3.1

Summary of the schedule and activities and assignments

Week	Activity/Assignment	SLOs
5	Activity 1: Hexadecimals	1, 2
5	Activity 2: Bar Charts from Summary Information	1, 2
6	Activity 3: Fan Cost Index	1, 2
7	Activity 4: Introduction to Directed Acyclic Graphs	3, 4, 5, 6
7	Activity 5: Women in STEM	1, 2, 5, 6, 7
8	Assignment 1: Tuberculous	1, 2, 3, 5, 7
8	Activity 6: High Peaks	1, 2, 5, 6, 7
8	Activity 7: World	1, 2, 3, 4, 5, 6, 7
9	Activity 8: World part II	1, 2, 3, 4, 5, 6, 7
9	Activity 9: Evaluation	1, 2, 5, 6, 7
9	Assignment 2: House Prices	1, 2, 3, 5, 6, 7
10	Activity 10: SAT	1, 2, 5, 6, 7
11	Assignment 3: Cars	1, 2, 3, 4, 5, 6, 7

their friend guidance on what features a car at that price would ideally have to ensure they were not getting overcharged. There was a small subset of cars priced at (or over) $40,000 in the graph. Ideally, the student only mentioned variables that they found impacted the price, and that they found the levels within those variables that put the price at or above $40,000. Table 3.1 depicts the full list sequence of activities and assignments with their SLOs.

3.4.3 Think Aloud Sessions for Developing the Assignments

To gather validity evidence for the items on the assignments, think-aloud sessions were conducted with three statistics education graduate students and the instructor of the visualization course in Summer 2021. The participants were selected based on their knowledge of the course content, R, and statistics education research and teaching. They each individually participated in an interview for approximately one to two hours via Zoom. The participants were given consent forms and the interviews were audio and video recorded. During the interview participants completed each of the three assignments while reading each question aloud and providing a rationale for each answer they had given.

Results for each interview were used to make changes to the assignments before the next interview (notes and each version can be found in Appendix B. There were four rounds of updates made to the assignments. The updates often clarified the item to ensure it elicited the intended responses. Some updates consisted of adding sentence limits for open ended questions, minor clarifications in wording, and more plot customization was added to ensure assignments met course requirements for creating visualizations. Updates from each think-aloud interview are described in detail in the next section.

3.4.3.1 Round 1

In the first round of think-aloud interviews, changes were made to the overall structure and content of some assignments. This interview was conducted with the instructor of the course who provided input on the timing of the activities and the scaffolding needed to help students move through the activity. For example, in Assignment 1, there were only minor typos fixed, but the instructor thought the activity would be too short and did not take advantage of the graphing abilities students had worked on previously in the semester. To address this concern, Assignment 1 was updated to ask the students, not only to create a particular graph, but also to add a theme, color palette, and informative captions and titles. These same additions were made to Assignments 2 and 3 for creating the final plots. This interview helped align the assignments with the course material and ensure that it met the requirements for the course's SLOs.

3.4.3.2 Round 2

In the second round of think-aloud interviews, a statistics education graduate student was given the assignments, which had been updated from the previous feedback. Given that this student was not familiar with the course content or structure, their feedback was most helpful to clarify the language of the questions. For example, at least one question per activity asked if one variable "affects" another, but given that this automatically implies a causation, this language was changed to ask whether the students thought one variable was "associated with" another variable.

This interview also prompted adding a note to limit the sentence count for some of the open-ended questions in the assignments. This participant was concerned that the students might feel the need to discuss all possible relationships among three to four variables and write entire pages of information justifying their answer using outside information. However, it was not the intent to have the students complete other research or provide exhaustive answers to questions about potential relationships among variables. To address this concern the phrase "write no more than 5 sentences" was added to all open-ended description items.

3.4.3.3 Round 3

In round 3 of think-aloud interview feedback, fewer changes were suggested. Again, this interview was conducted with a statistics education graduate student using the updated materials from the previous two rounds of feedback. From this interview, a few more typos were corrected, and some items were clarified to address which variable was the outcome variable and to make clear that only associational claims could be made with this data. Another update in Assignment 3 was in a question where the students were asked to describe a set of faceted scatterplots and the relationships among all the variables (five total). However, the response to the item as written would have been quite lengthy and the students were already asked to demonstrate, in a previous question, that they can describe a set of plots with four variables. As a result, this item was cut in the interest of keeping the assignment from becoming significantly longer than previous assignments.

3.4.3.4 Round 4

In the final interview, only a few changes were made to improve plot aesthetics and clarity of items. In Assignment 2, students explore the relationships of variables pertaining to the price of a house through creating a scatterplot with price on the *y-axis*, square footage on the *x-axis*, age mapped to color and number of bedrooms mapped to size. However, all participants in the think-aloud interviews pointed out that the sizes of the points are so similar that it is difficult to discern any difference in the number of bedrooms the houses have. To fix this, the plot was updated to have the students map age to size (because there is a greater range of house ages, the sizes are more distinct with this mapping) and color to the number of bedrooms.

In Assignment 3, Question 6 also received an update for clarification. The item initially was written: "Choose another variable from the list of variables in the dataset. Create a DAG to propose relationships among the four variables." All participants expressed some confusion over how many variables should be put into the DAG because it was unclear whether "four variables" included the outcome variable. In response to combat this confusion, the item was updated to read: "Choose another variable from the list of variables in the dataset. Create a DAG to propose relationships among the four variables (3 possible causal variables and price)."

Additionally, Question 7 was updated to be more clearly defined. Originally this question asked for an interpretation of the relationships featured in the scatterplot they created, but students would have already described the relationships among three of the variables in a previous question, so there would be only need to describe the new (fourth) variable introduced. This update helped decrease redundancy in answers and made the lengthy activity a bit more manageable. Finally, for the last question in Assignment 3 updates were made to specify that students should mention at least two variables in their final description of variables they think affect the price of a car.

3.5 Data Collection and Analysis

To evaluate the development of students' multivariate reasoning after taking part in this study, data in the form of classroom activities and assignments were collected from the two sections of a communication and visualization course in Fall 2021. One student in each section was observed as they worked through the activities to attain further understanding of their thought process and any challenges as they worked through the materials. All students' assignments were qualitatively analyzed for evidence of multivariate reasoning. The final homework assignment was qualitatively analyzed to get a sense of any misunderstandings or misconceptions the students held after completing the series of activities. Additionally, think-aloud interviews with two volunteers from each section gave further insight into their reasoning while working through this last assignment. The methods for data collection and analysis for this part of the study are described next.

3.5.1 Observation of Individuals Procedures

One volunteer from each section was observed for the duration of the multivariate thinking unit. The researcher sat with the student during the unit and audio recorded their discussions throughout the class period, while taking observation notes. Recordings were collected and saved on the researcher's laptop in Google Drive via the Audio Recorder app and on the researcher's cell phone via the Voice Memos application. Dual recordings were saved in case one was lost or of unusable quality. Notes were taken via Google Docs and saved in the same Google Drive as the rest of the data collected throughout the study. The students in-class activities were downloaded from Canvas for analysis.

The student observed in Section 1 was recorded eight days out of 10. This student was in-person for seven days, online one day, and did not attend class for two days. When in class, this student worked alone, only receiving help on the assignment from the researcher when in-person and the instructor when working outside of class time. The student observed in Section 2 was recorded for all 10 days, with two of those days being recorded via Zoom. This student typically worked with a group of one to three students in the classroom. This group often discussed ideas, compared code, and helped each other with the activities. Observation notes, audio, and activities from these students were qualitatively analyzed for evidence of the students' reasoning related to the learning goals and evidence of a common trajectory for reasoning development as they worked through this unit.

3.5.2 Qualitative Analysis of Assignment Data

Students submitted class activities and assignments individually via Canvas. Only the in-class activities of students observed by the researcher were collected and analyzed, which is described in another section. Though submissions for class activities were made by the end of every class period, assignments were often due by 9:00 p.m. on the due date. Students could work in groups in class on the assignments, but they typically worked outside of class, so it is unknown how much help they received from outside sources.

The assignments for all students were downloaded from Canvas, stored in Google drive, and then de-identified. Files downloaded from Canvas contain the student's name as the file name, so to de-identify the assignments the files were given a numerical and informative label based on the assignment (i.e., 01-cars for the first submission of the cars assignment). Because there was variability in which students submitted each assignment, "01-cars" was not necessarily submitted by the same student that submitted "01-house". Students were given the option to request their work not be used in the study, but no student made such a request. There was a total of 38 submissions of Assignment 1, 37 submissions of Assignment 2, and 33 submissions of Assignment 3 out of 43 possible submissions from both sections.

3.5.3 Coding Selection

To begin the coding process, four submissions for each assignment were randomly selected to be coded by the researcher and a second coder. The `RANDBETWEEN` function from Google Sheets was used to randomly generate five numbers between 1 and 38. The same process was repeated for Assignment 2 and Assignment 3 to generate a random selection for those assignments using the total number of submissions for each assignment (37 and 33 respectively) for the last argument of the `RANDBETWEEN` function. Once the four assignments were selected, they were uploaded to NVivo (2020) into an initial sample folder for coding. Twelve of these assignments make up roughly 10 percent of the total number of submissions recommended for checking reliability (e.g., O'Connor & Joffe, 2020). An additional selected submission from each assignment (to make the total of five selected randomly) was used as an initial calibration for the two coders to discuss potential codes and how to apply the initial set of codes to the assignments.

3.5.4 Codes

To provide evidence for answering Research Question 1, answers on assignments were coded using an exploratory coding method (Miles, Huberman, & Saldaña, 2020; Saldaña, 2016). Using this method, provisional codes were created based on the SLOs, the literature reviewed in Chapter 2, and classroom observations. These codes were then expanded during content analysis. The provisional coding scheme is discussed next and the codes that emerged after initial coding are discussed in the results chapter. All codes with examples can be found in Appendix B.

Answers on assignments were first coded with the intended SLO and whether they were correct, partially correct, or incorrect. These codes were applied for SLOs 1, 3, and 4. These SLOs pertain to stating whether the data is observational, explaining if we can make causal claims based on the data, and using the data to create multivariate graphs, respectively. These items were clearly separated into the correct, partially correct, or incorrect categories. In contrast, for SLO 2 (explaining relationships among variables) some answers were coded as plausible. This coding was used for Questions 8 and 13 of Assignment 3 because these questions were not clearly categorizable into "correct" or "incorrect" labels. Question 8 asked students to justify choosing a variable and explain the relationship they thought it might have with the other variables. This requires speculation which may be plausible, but without further research may not be deemed "correct" or "incorrect". Similarly, Question 13 had students choose between two graphs of multiple variables to explain which they thought best represented the data. For this question, most graphs were similar enough that there was not necessarily a correct answer to the responses, thus they were coded as “plausible” if they were reasonable answers.

In Assignment 2, all questions were mapped to a learning outcome, but in Assignments 1 and 3 some items were not mapped for various reasons. For example, Assignment 1, Question 2 had them create an unreadable plot before having them create the proper line plot, which was more about learning how to code than assessing the SLOs for this study. Questions 5 and 6 were an extension of the first part of the activity that had the students create more line graphs with more additional aesthetic attributes, but since some students did not get this far, questions 3 and 4 were used to assess their ability to create and reason with graphs in this first assignment.

In Assignment 3, Questions 2, 3, 4, 5, 6, 8, 12, 13, 15 were not coded for this part of the analysis. Questions 2, 3, 4, 5, and 6 do not pertain to any of the SLOs because they ask students to create and interpret scatterplots. Question 8 has them predict the relationship they think they will see before they create a DAG. Questions 12 and 13 have the students create a novel visualization and comment on its utility. Question 15 is a prediction question based on reasoning with their graph, but perhaps beyond what is described in SLO 2. Questions 12, 13, and 15 do not clearly map to the SLOs in a unique way compared to Questions 1, 7, 9, 11, 10, and 14 which were used for analysis. See Table 3 for a full mapping of the learning outcome and question analyzed on each assignment.

Table 3.2
Summary of Questions Mapped to SLOs

SLOs	HW1	HW2	HW3
1. Create graphs displaying the relationships among three or four variables in one plot	Q3 Use the WHO-TB.csv dataset to make a line plot visualization of tuberculosis deaths across time.] To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set group = Country. Paste your plot below.	Q3 Create a scatterplot to look at the relationships among the variables in Question 2. Paste your plot below.	Q9 Create a plot incorporating the four variables.
2. Explain the relationships among three to four variables using graphs	Q4 What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see in your plots. (Limit your response to 5 sentences or less)	Q6 Based on your DAG and the plot above, what variable(s) do you think are associated with the price of a house? Justify your answer using evidence from your plot.	Q11 Provide a justification for your DAG using your plot as evidence.

(Table continues on next page)

Table 3.2
Summary of Questions Mapped to SLOs (continued)

SLOs	HW1	HW2	HW3
3. Identify data as observational	Q1 Is this observational data? Explain your answer.	Q1 Is the data observational? Explain your answer.	Q1 What type of data is this (observational or experimental)? Can we use this data to make causal claims?
4. Explain the limitations in making causal claims with observational data	Not Assessed	Not Assessed	Q1 What type of data is this (observational or experimental)? Can we use this data to make causal claims?

(Table continues on next page)

Table 3.2

Summary of Questions Mapped to SLOs (continued)

SLOs	HW1	HW2	HW3
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	Q7 We see changes in tuberculosis death rates over time in some regions, but we might wonder what is causing these changes. Draw a DAG to incorporate two or three variables that you think might be associated with the tuberculosis death rate in a country. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.	Q2 Draw a DAG to propose variables you think have an effect on the price of a house (we will ignore baths for this activity).	Q7 Choose another variable from the list of variables in the dataset. Create a DAG to propose relationships among the four variables (3 possible causal variables and price). Paste it below.

(Table continues on next page)

Table 3.2
Summary of Questions Mapped to SLOs (continued)

SLOs	HW1	HW2	HW3
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	-	Q4 Draw a DAG to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related.	Q10 Draw your final DAG to represent how your three variables affect price and each other.
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	Q8 Provide a justification for your drawing in Question #7 that explains your proposed relationships among the variables you chose.	Q7 What other variables do you think are associated with the price of a house that are not considered in this dataset?	Q14 Are there any other variables (not in the current data set) that you think might affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars? Choose one or two other variables and explain your answer.

Other initial codes included ideas based on the literature reviewed in Chapter 2. For example, from studies by Kuhn et al. (2015) and Abdelhadi (2016) we know in multivariate thinking the context plays a big role in students' interpretations, especially if causality is questioned. For this reason, a code was created to mark student responses in which their explanations of plots or DAGs seemed to go beyond the context or variables given. When this occurred, it was coded "context-interference". This code was further refined in the second round of coding to describe the ways in which the context played a role in students' answers. This is discussed further in Chapter 5.

Another code added inspired from the multivariate thinking literature was a code for "variable-level-confusion". (Kuhn et al., 2008) described that sometimes students would confuse the variable level for a categorical variable with the variable itself. Though more quantitative data was used for these activities, this idea did emerge in the data in a few different ways, which is discussed further in the results chapter. Similarly, the literature describes how students often use a single variable to discuss associations, when it is possible more variables should be discussed. The codes "considering-all-vars" and "not considering-all-vars" were used to sort whether students considered all variables in the dataset or just some of the variables.

In light of research describing the difficulty students have interpreting scatterplots (e.g., Batanero, Estepa, & Godino, 1996; Batanero, Estepa, Godino, & Green, 1996; Cobb et al., 2003; Gil & Gibbs, 2017; Moritz, 2004; Zieffler & Garfield, 2009) a code was created to indicate if a student response describing their plot did not match what was depicted in the plot. In this case, responses were coded as "plot-description-mismatch".

Other initial codes were created from classroom observations of students working on the activities. From these observations, the students' interactions with DAGs and common mistakes were noted. All the codes generated for SLO 5 (wrong arrows, including directed arrows, forgot variable, updated DAG, and marking relationships between independent variables or not) were all codes created based on observation notes from watching the students work on in-class activities. Another code created based on in-class observations was "dag-description-mismatch". This code was used for SLO 2 questions in which students described their DAGs in a way that was not aligned with what was depicted in the DAG.

Table 3.3 provides the provisional list of the qualitative codes used in this study as well as their origin and a description of how each code was applied. Codes that emerged after this initial set of codes will be discussed further in the Chapter 5.

Table 3.3

Provisional Codes with Description

Code	Origin	Description
Learning Outcome 1		Create graphs displaying the relationships among three or four variables in one plot
LO1\correct		Graph was created with 3 or more variables in the way specified by the assignment or in a logical way given the variables.
LO1\incorrect		Graph was not created with 3 or more variables in the way specified by the assignment or in a logical way given the variables.
Learning Outcome 2		Explain the relationships among three to four variables using graphs
LO2\Correct		Provides a description of the variables in a way that is aligned with what is depicted in the graph

(Table continues on next page)

Table 3.3
Summary of Questions Mapped to SLOs (continued)

Code	Origin	Description
LO2\plot-description-mismatch	Class Observations	Provides a description of the variables in a way that is not aligned with what is depicted in the graph
LO2\aggregate-reasoning	Literature: (Konold et al., 2015)	Provides a description of the relationships among the variables at a high-level summarizing across all variables
LO2\case-reasoning	Literature: (Konold et al., 2015)	Provides a description for the variables on an individual level singling out certain cases in the graph
LO2\considering-all-variables	Literature: Abdelhadi, 2016; Casparo & Grulich, 2019; Kuhn, 2007; Kuhn et al., 2015; Ridgway et al., 2007; Gil and Gibbs, 2017	Provides a description of all the variables in the plot leaving none out
LO2\not-considering-all-variables	Literature: Abdelhadi, 2016; Casparo & Grulich, 2019; Kuhn, 2007; Kuhn et al., 2015; Ridgway et al., 2007; Gil and Gibbs, 2017	Provides a description of all the variables in the plot leaving one or more out of the description
LO2\dag-description-mismatch	Class Observations	Provides a description of a DAG which does not match what they have drawn in that DAG as the relationships among the variables

(Table continues on next page)

Table 3.3
Summary of Questions Mapped to SLOs (continued)

Code	Origin	Description
LO2\partially-correct		Provides an incomplete but correct description of the nature of the relationships seen in the plot or DAG
LO2\plausible		Provides a description of a DAG describing all variables in a way that is plausible
Learning Outcome 3		Identify data as observational
LO3\correct		Correctly identified data as observational
LO3\incorrect		Did not identify data as observational
LO3\partially correct		Identified data as observational, but gave wrong reasoning why it was observational
Learning Outcome 4		Explain the limitations in making causal claims with observational data
LO4\correct		Describes how causal claims cannot be made with observational data
LO4\incorrect		Describes how we can make causal claims with observational data

(Table continues on next page)

Table 3.3
Summary of Questions Mapped to SLOs (continued)

Code	Origin	Description
Learning Outcome 5		Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables
LO5\directed-arrows	Class Observations	DAG created contains directed arrows
LO5\forgot-variable	Class Observations	DAG created does not contain all needed variables
LO5\no-relationships-IVs	Class Observations	DAG created had no relational arrows between any variables except those with the outcome variable
LO5\not-correct-arrows	Class Observations	DAG created displays casual arrows that could not possibly exist
LO5\plausible		DAG created is plausible (opposite of not-correct-arrows)
LO5\relationships-between-IVs	Class Observations	DAG created shows relational arrows among any variables and with the outcome variable
Learning Outcome 6		Evaluate DAGs already created to assess if they accurately represent relationships among given variables
LO6\incorrect		DAG description does not match the plot/DAG provided

(Table continues on next page)

Table 3.3
Summary of Questions Mapped to SLOs (continued)

Code	Origin	Description
LO6\incorrect-dag-arrows	Class Observations	DAG created displays casual arrows that could not possibly exist given the information displayed in the plot created
LO6\updated-DAG	Class Observations	DAG was updated from a previous question after evaluating a graph of the variables
Learning Outcome 7		Develop hypothesis about variables not investigated and their relation to an outcome variable
LO7\plausible		Described possible/logical variables that could affect the system of variables in a meaningful way
context-interference	Literature: Abdelhadi, 2016; Kuhn et al., 2015	Description brings in outside context the conflicts with what is in the DAG/graph or in some way is adding to their response
variable-level-confusion	Literature: Kuhn 2008	Student expresses incorrect reasoning around the variable or the level of interest

3.5.5 Reliability.

The second coder and I met to review the initial proposed code set and proposed additional codes. After reviewing the codes and the initial assignments the second coder and I worked individually in our own copies of the NVivo project to code the remaining four submissions from each assignment. Once completed, the two individual projects were merged, and reliability analysis was conducted. Averaged across all assignments and codes, the average Kappa was 0.9 and the percent agreement was 98.46. This meets the typical recommended coding reliability analysis standard (e.g., O'Connor & Joffe, 2020). Though these are acceptable measures, to ensure that there were not any systematic differences in the coding scheme applied between the two researchers, individual codes across all assignments were investigated. The codes between the researchers were compared for any code in an assignment with percent agreement below 80% or Kappa below 0.8. Typically a lower percent agreement or Kappa value indicated one of the following scenarios: 1) the coders had coded same paragraph with the same code, but one researcher had only highlighted one sentence within that paragraph, 2) the same region was coded, but not entirely the same highlighted area, 3) one researcher forgot to code for a learning outcome, 4) one coder coded the wrong learning outcome, or 5) one researcher coded something as a learning outcome correctly that the other person did not code at all. These discrepancies were discussed between the coders and resolved as accidental for forgotten coding or prompted discussions about the code meanings which were then refined. I went on to code the remainder of the assignments.

3.5.6 Cognitive Interview for Final Assignment

Three volunteers (not those observed during the class activities) participated in a cognitive interview asking questions about their final multivariate thinking assignment. To recruit students the researcher made an announcement at the beginning class in each section exactly a week before the assignment was due. and sent out a Canvas announcement with details about the think-aloud. Students were offered a $10 Amazon gift card as an incentive for their time. The volunteers met individually with the researcher outside of class for a half hour long section conducted via Zoom. They talked through their answers on the assignment while being recorded. Their responses on the assignment, the researcher's observation notes, and excerpts from their recording were qualitatively analyzed (via Saldaña (2016) exploratory method) to gain insight into their final multivariate reasoning ability. Particularly, attention was paid to their reasoning as it pertains to the SLOs and any remaining misconceptions or challenges they faced while talking through this assignment. Quotes and themes are presented in Chapter 4.

3.6 Chapter Summary

A unit on multivariate thinking was created for a data visualization course. For this unit 10 activities and three assignments were created and implemented in two sections of a communication and visualization course in Fall 2021. To further gain insight into students' multivariate reasoning throughout the unit, one student from each section was observed. The results from qualitative analysis of the assignments were analyzed by the researcher and a co-coder for evidence of multivariate thinking pertaining to seven SLOs. Finally, three students were interviewed at the end of the unit to provide rationale for their answers on the last assignment. The results are presented in the next chapter.

Chapter 4

Results

This chapter presents results from the analysis of the qualitative data collected for this study. First, results from the classroom observations of two students are provided, presenting key features of their multivariate reasoning and challenges they faced as they worked through the class activities in the multivariate thinking unit. Then, results from analysis of the class assignments are presented, highlighting the changes in correct response rate across the three assignments and notable themes that emerged during the assignment coding. Finally, results from three cognitive interviews with students are detailed in relation to the SLOs, with emphasis on the reasoning underlying their responses on the last assignment.

4.1 Results from Class Observations

This section provides results from the two participants observed in class while completing the multivariate thinking unit in the communication and visualization course. Responses to class activities, excerpts from class discussion, and observations relevant to developing multivariate thinking across the unit are presented. Results are detailed with emphasis on creating and reasoning about multivariate visualizations and DAGs.

4.1.1 Section 1 Observation

Results from the student observed in Section 1 will be presented first, using the pseudonym Jordan. Notably, Jordan worked alone for the entirety of the multivariate thinking unit, as did most of this class section. Jordan produced little discussion about the answers provided in the activities, except when prompted by the researcher. Any quotes provided are from interacting with the researcher or from the class activities turned in through Canvas.

Jordan had sporadic attendance at the start of the semester, and as a result was behind on the in-class activities when observations started during the fifth week of the semester. However, when they did attend, they were very attentive and focused on the activities, and were able to get back on track by the second week of observation.

4.1.1.1 Creating visualizations.

Jordan had a slow start creating visualizations in RStudio due to their intermittent in-person attendance. While working on the first few activities in the multivariate thinking unit they created stacked bar charts and line plots but had a lot of basic R and debugging questions. They asked questions about how to know which arguments to update in the `ggplot` function, which layers to add to create the desired graph style, and when to add parentheses. However, once Jordan established a basic understanding of the code needed for graphing, they were able to become more fluent and even excited about writing the code to create visualizations after the first few classes of the observation.

Observer notes from Activity 5: Women in Stem, the first activity where they create line plots, indicated that Jordan had made significant progress in their ability to update the arguments of the functions without much help or looking back at previous work. In this activity, they were able to add many aesthetic features to the plot such as a title, annotations, arrows, and different shapes. By the time Jordan completed the High Peaks activity, the observer notes revealed that the coding was no longer an issue and their focus had shifted to the context of the data and the relationships among the variables in the graphs they created.

In one of the last activities in the unit, Activity 9: Evaluation, the students were asked to critique the graph they see at the start of the assignment. One learning goal in creating visualizations in this course, particularly multivariate ones, is to ensure variable relationships are clearly depicted and that students can communicate answers to a particular research question. Jordan had many thoughtful critiques about the clarity of the graph as indicated by their statement: *"X and y are not well defined. There is no title to tell me about the Age of what? . . . I wouldn't facet wrap it . . . I just would do it differently."*

Though the critiques based on the axis and heading titles would more clearly define what is depicted in the visualization, some of the other comments regarding the aesthetics would not have necessarily made the graph clearer. For example, they commented *"It doesn't seem like the different teaching, tenure track, [etc., are] comparable. It feels like they are their own little things and I can't extrapolate."* Jordan wanted to get rid of the facet wrap by the rank of the instructor because they felt the different positions were not comparable due to their different teaching goals. However, one of the goals of creating the graph was to see if there were differences within the ranks of the instructors, so this critique did not align with the research question being asked. Jordan was able to create graphs more easily throughout the multivariate thinking unit but creating graphs to help reason about a specific research question was still challenging at the end of the unit. These challenges are discussed in further detail in the next section about reasoning with visualizations.

4.1.1.2 Reasoning about visualizations

Initially, Jordan had little trouble reasoning about multivariate visualizations. In working on Activity 1: Hexadecimals, they were able to summarize the stacked bar graph and thoughtfully bring in some outside context to supplement their analysis. When asked *"Do iPhone (iOS) and Android users stream Music using the same services? Explain by referring to your* *bar chart,"* they responded, *"I would say, yes, because what they use is not mutually exclusive, for the most part, the most popular brands for apple and android users [are] the same streaming services. Except for Apple Music, which is a little unfair because Apple music exists [only] on the IOS device."* Though their answer is aligned with what is depicted in the graph, they did not reference explicit evidence from their graph itself to justify their answer.

Jordan occasionally had difficulty reasoning about the variables themselves when discussing relationships among them. For example, in the Activity 6: High Peaks, they noticed that Time and Distance had a strong positive relationship with each other, which is clearly depicted in the graph. They explained this happened because these variables (Time and Distance) are essentially *"measuring the same thing."* When asked to consider a case in the graph where the distance was short, but the time was long, the student was quick to determine that it must be from a steeper or more difficult hike, but remained adamant that the two variables measured the same thing and thus only one was needed to explain the Difficulty rating of the hike.

Next in Activity 6: High Peaks, Jordan described the relationship between ascent, length, and difficulty, stating that, *"based on the graph I see more the darker ones (higher ascent) up where the length is more. When you have more distance to cover (indicating greater length) you will have more to go up by default."* This is not necessarily true, because you could have a longer hike with a less steep slope, as indicated in some points on the plot. Their statement indicated that they were having a bit of trouble interpreting the variables and the relationships they were seeing in these visualizations, often trying to justify the relationships depicted with outside knowledge, regardless of the extent to which their justification fit with the graph.

Another example of bringing in outside knowledge to help reason about visualization occurred during Activity 7/8 World Data. When Jordan explored the relationships among income, region, life expectancy, and population they discussed all the relationships in detail, and though the descriptions aligned with the plot, they were not supported by evidence from the plot, but instead only with outside knowledge. For example: *"region impacts life expectancy because the resources you have in those regions are big determiners in what happens in your life, and thus how long you'll live."* Ideally, the student would have been able to give specific details about the plot that led them to this conclusions so that we would know that their answer was a combination of outside information and their graph, and not solely outside information.

This pattern of reasoning about the context more than the visualization continued into Activity 9: Evaluation. Jordan explained they did not think the average evaluation score variable was related to any of the other variables even though it had a clear relationship to the age variable and the beauty average rating. When investigating these variables further, they expressed some frustration about using some of the variables in the graph to investigate their impact on the average evaluation scores of the instructors. They explained,

> Pretty privilege exists, but if you have a question about that, make that your graph. If you are with a group [of students] for that long you can't get away with [relying on pretty privilege to help your evaluation score].

The activity asked the student to investigate the possible variables that affect the evaluation score, but Jordan strongly believed that the beauty average should not be considered and if we did want to see how that rating affected the scores, we should include only that variable and the score in our visualization. The student was concerned about reasoning with too many variables at once and struggled with the context of the research question and how to answer it using the variables and visualizations.

Jordan's concern about working with too many variables was addressed in Activity 10: SAT which covered Simpson's Paradox. In this activity the student was prompted to plot the SAT score and school expenditure to see if there was a relationship. Though Jordan predicted it would be a direct relationship and they were surprised when it appeared to be an indirect relationship, they quickly tried to justify the indirect relationship by saying that the relationship was negative *"because it isn't the school paying [for the SAT itself], it is the parents,"* implying that it does not matter what the school is spending on the students, because it is the parents spending that will impact the students' SAT score. Though this still did not explain the negative relationship seen, I encouraged the students to keep moving through the activity to see what else might be happening here. Once they colored the points of the plot with the fraction of the students taking the SAT, they saw that the trend in the data appeared to change to the direct relationship they initially predicted. Though we talked through this idea the student concluded that the data was from *"two different samples"* and that is why we saw this phenomenon in the data.

In exploring Simpson's Paradox, I used this SAT example to stress the importance of looking at more than one variable when we are trying to find potential causal variables for an outcome variable. Many of Jordan's concerns around reasoning with multiple variables throughout the unit were related to looking at too many variables at once. Discussing Simpson's Paradox allowed us to circle back to their thoughts on Activity 9: Evaluation and discuss why we could not only look at the instructors' evaluation score and one other variable because there could be other variables affecting the system.

4.1.1.3 Creating DAGs

In the introduction to the DAGs activity (Activity 4: Directed Acyclic Graphs), Jordan did a great job leading the small group they worked with to come up with the DAGs in the activity. They thought through each context and clearly communicated ideas to their group both verbally and through the DAG.

In the next activity with DAGs, Activity 5: Women in Stem, Jordan excelled at creating DAGs on their own. Figure 4.1 depicts the DAG they created. In this activity, the student was asked to come up with hypothetical causal variables that they would also want to explore related to a women's income level. They correctly used directed arrows and even depicted the potential relationships among all variables and not just between the outcome variable and the other variables.

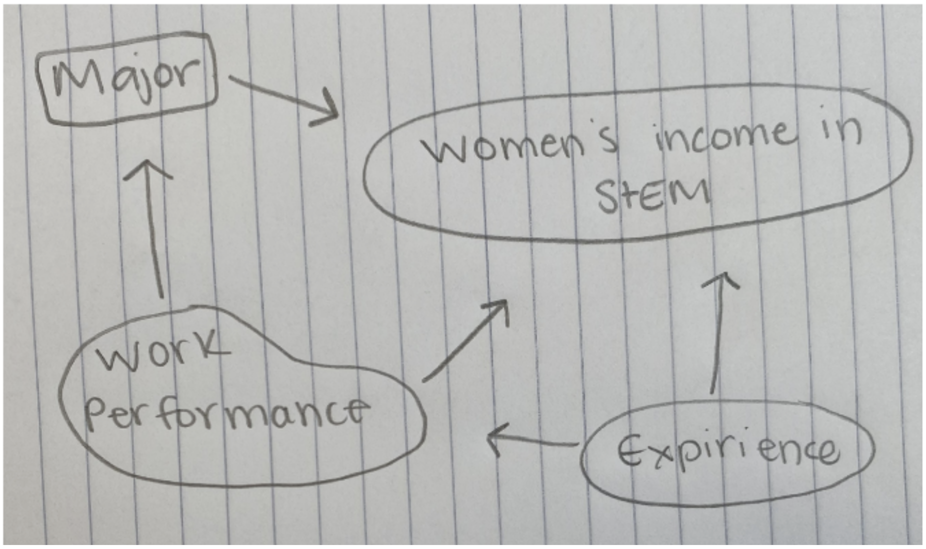


Figure 4.1. Jordan's DAG of potential variables influencing a Women's Income

However, when Jordan needed to draw DAGs based on a data visualization, they found this a bit more challenging. For example, near the end of the unit, in Activity 7/8 World Data, when asked *"Do you think the relationship between population size and life expectancy is as strong as the relationship you saw between region and life expectancy?"* They responded, *"No because I'm not seeing much of a difference between population size and life expectancy for Africa, or really very many of the other countries."* They determined that the population did not have a clear relationship with the other variables, but then proceeded to keep it in their final DAG as seen below in Figure 4.2.

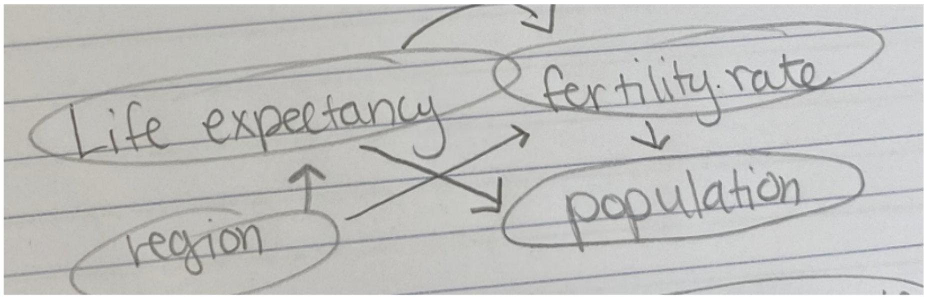


Figure 4.2. Jordan's DAG for Activity 7/8 World Data

4.1.1.4 Summary of Jordan's Observations.

Jordan, though initially behind the class, quickly learned the code needed for creating multivariate plots in R. They were able to create stacked bar charts, line plots, and scatterplots with customized aesthetics. However, their reasoning about their visualizations did not show as much growth as their ability to create them. They often had correct interpretations about the graph but used only outside knowledge to support their answers rather than evidence from the plot. They also struggled to understand what some variables meant in context and why they were needed to answer the research questions. Jordan was able to create DAGs for proposing relationships among hypothetical variables, but when updating or creating DAGs they occasionally did not incorporate what they learned from their visualization.

4.1.2 Section 2 Observations

Next, observations from Section 3 are presented using the pseudonym Kennedy. Notably, Kennedy worked with a group of one to three other students for the entirety of the unit, and thus had more opportunities for discussions of the context, coding, and visualizations in the activities.

4.1.2.1 Creating visualizations.

In the initial activities for this unit, though the students had been coding for approximately a month, the questions around creating the graphs still proved difficult for Kennedy and their group. Their main discussion and questions the first weeks of observation centered around downloading the data, importing it into R, and altering the script as needed to create the visualizations in the class activities. Observer notes from the hexadecimal activity suggest many questions of the form, *"how do I download this [data]?"* and *"how do I know where to put parenthesis/commas?"* At this point Kennedy was still developing the foundation of writing the code for creating visualizations.

After getting their basic coding question answered in the first activity, the group felt more confident in Activity 2: Bar Charts from Summary Data, noting *"hey we are getting good at this!"* in reference to writing the code to create a bar chart. Then, their questions shifted from syntax to focus on determining which variables should be mapped to the x and y axis. When they reached the Activity 6: High Peaks, the group quickly figured out how to create and modify the code to make a scatterplot: *"We need geom_point() not geom_hist(). But we need x, y - oh they are in the question."*

In Activities 7/8: World Data, and through the remaining activities of the multivariate thinking unit, Kennedy spent even less time coding the visualizations, having developed some fluency from the previous activities. Writing the code for scatterplots became easier and they also were able to do advanced plot work like add an annotation to the plot, change the size of the points to correspond to a variable, add a title, subtitle, caption, and adjust the x and y axis labels.

4.1.2.2 Reasoning about visualizations.

This student began the first activities of the unit demonstrating the ability to describe the relationship between two variables. In Activity 1: Hexadecimals, the student was asked about the music streaming preferences for IOS and Android users. To answer this question, they created a stacked bar chart with the streaming service mapped to the *x-axis* and colored by the phone operating system. The student gave a complete description of the graph with evidence from the plot to back up the answer:

> The IOS and the Android users [both use] Spotify, Pandora, or "Other", because on top of Pandora, Spotify, and Other (indicating the bars for each of those streaming services) they had pink and blue and the other [streaming services] just had blue.

When working on the line plots in Activity 2: Fan Cost Index, Kennedy experienced more difficulty reasoning about the plot, noting in their group discussion that the line plots looked *"less cluttered"* but were *"more difficult"* to read than the bar charts. Ultimately, as a group they pieced together how to read the change in variables over time in these plots.

In Activity 5: Women in Stem, Kennedy and their group mates discussed whether the graph was linear. Kennedy described to their groupmate that the *"scatter"* or variability in the points meant it could not be described as linear. They wrote in their final answer that, *"[t]his scatterplot is non-linear, the points are just going everywhere. This scatterplot has a zero slope. I would say that some parts of the scatterplot are strong. The higher portion of women the lower the income."* Their final answer claims there is no relationship between the variables for income and proportion of women in a certain STEM career, but then concludes by describing a relationship between the two. However, they correctly answered the research question at the end of the assignment by reasoning about the varying incomes among the different stem majors, which relate to different proportions of women.

In the next activity, Activity 6: High Peaks, where the questions repeatedly ask about the linearity, slope, and strength of a scatterplot, Kennedy got the description correct each time. Once they started describing the relationships among three variables in this activity, they began to make a claim without using evidence from the plot. However, when prompted to add in evidence from the plot to support their answer, they were able to elaborate more using the visualization.

Near the end of the unit, in Activities 7/8: World Data, the student was tasked with identifying relationships among four variables. In part one of the activities, Kennedy was asked about relationships among life expectancy, region, and fertility, after exploring each pair of the variables in turn. In describing the relationship among all three variables, the student could do this in detail using evidence from the plot. The quote below is their description of the relationships among life expectancy, region, and fertility:

> I would say that the region you live in does affect life expectancy. Some reasons have a high life expectancy and other such as Africa have a low life expectancy. For region and fertility rate it does not really have an effect because many of the countries are similar such as Asia, Europe, and the Americas. Africa is a bit different, so I think it does have more of [an] effect. This plot is weak linear with a strong group of points at the top and get less strong as it goes down. The Americas, Africa, and Europe are located in the same part of the graph, but Africa is more at the bottom. The main take away from this plot would be the three countries at the top are the same, but Africa is different.

However, when asked to describe relationships among three variables in the Activity 8: World Data, without first describing relationships between each pairing of the variables, their answer did not include specific evidence from their plot. The plot Kennedy created choosing their own variables is in Figure 4.3, and their description of the visualization follows.

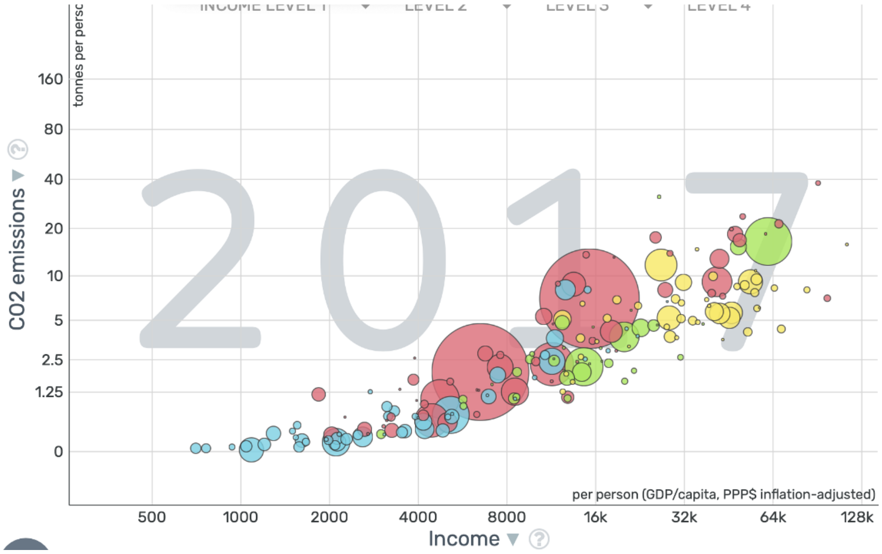


Figure 4.3. Kennedy's Graph Created in Gapminder with Income, CO2 Emissions, Sized by Population and Colored by Region

> I saw that as CO2 emission got - higher income did as well. It did not go up as fast for all countries. I saw that the population got higher for Asia, the Americas, and some parts of Africa. There are lots of differences between regions because some regions still have low Co2 emission.

In their response they indicated the *"population got higher for."* which implies a change over time. When asked to further explain the relationship between population and the other variables they replied: *"I would say that large population countries give off more CO2 and have a higher income. Especially when you look at Asia."* I clarified, *"so as we have more income (get higher on the y-axis) we see bigger countries and higher CO2 emissions (farther to the right of the x-axis)? Does that seem true for most of the graph? Or just Asia?"* And they replied, *"it seems true for most of the graph."* Kennedy's description of relationships depicted in this graph did not match what is pictured in the graph, even after encouraging them to check their work.

4.1.2.3 Creating DAGs.

In the initial activity introducing DAGs (Activity 4: DAGs), Kennedy worked in a group of four to create and discuss the DAGs. They all were engaged throughout the activity and came up with detailed thoughtful DAGs. When Kennedy created the DAGs in Activity 5: Women in Stem, they were able to think of variables that might affect Women's Income and considered relationships among all the variables in the model, as seen in Figure 4.4.

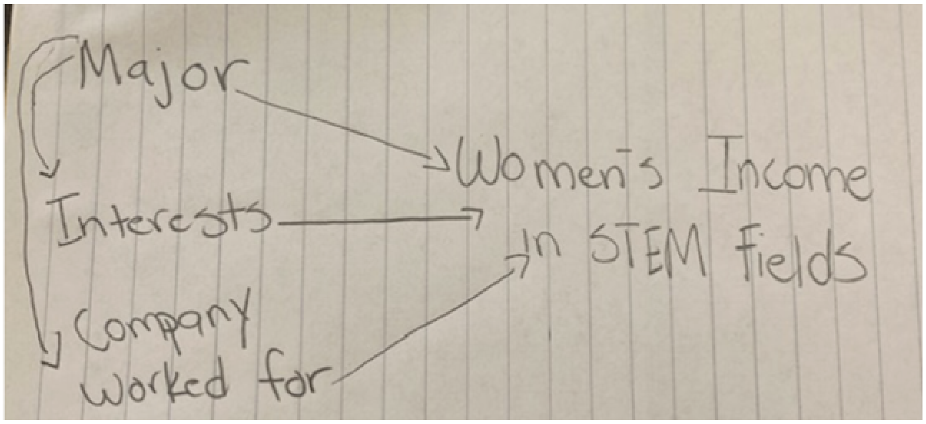


Figure 4.4. Kennedy's Depiction of Variables that Might Influence a Women's Income in a STEM Career

In Activity 9: Evaluation, there were more variables to consider, and it is at this point the student's work indicated there might be some confusion about the direction the arrows should point. As seen in Figure 4.5, the student's DAG suggested that the instructor's evaluation score potentially affects the age of the person. They also included a double ended arrow to indicate that the rank of the person potentially impacts their average beauty score and that the average beauty score potentially impacts their rank.

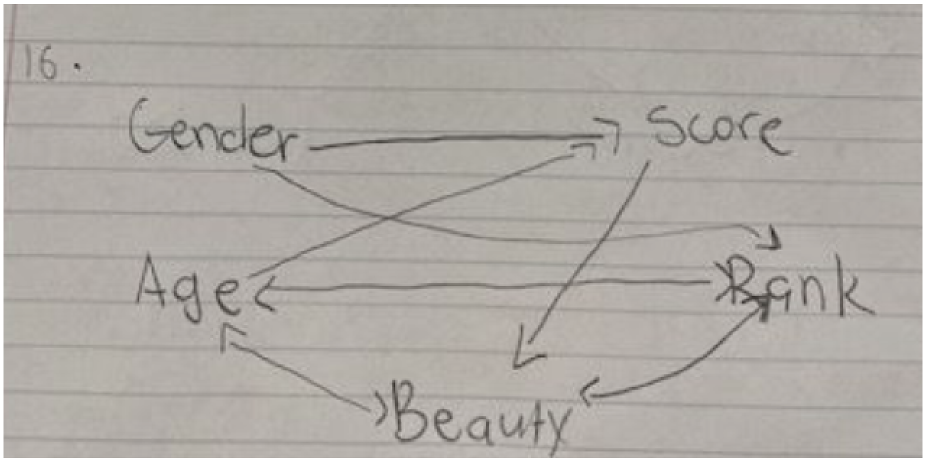


Figure 4.5. Kennedy's DAG for Activity 9: Evaluation

Even after completing the activity and exploring, in detail, the relationships among the variables described, Kennedy still concluded the activity with a final DAG depicting a few potential causal variables that may not be possible. For example, in Figure 4.5, they included a double ended arrow between age and average evaluation score for the instructor. Though it might be possible that age impacts score (perhaps as the instructors age they get better at teaching and receive higher scores), it does not seem possible that the average evaluation score could have an impact on the instructors' ages.

Kennedy eliminated the double ended arrow between age and rank in favor of a single headed arrow from age to rank as seen in Figure 4.6. After extensive discussion with their group about the context throughout the activity and exploration of the interactions among all these variables, drawing the DAGs became more difficult. They explained that all their DAGs were different *"based on what I was feeling"*, not based necessarily on the graph.

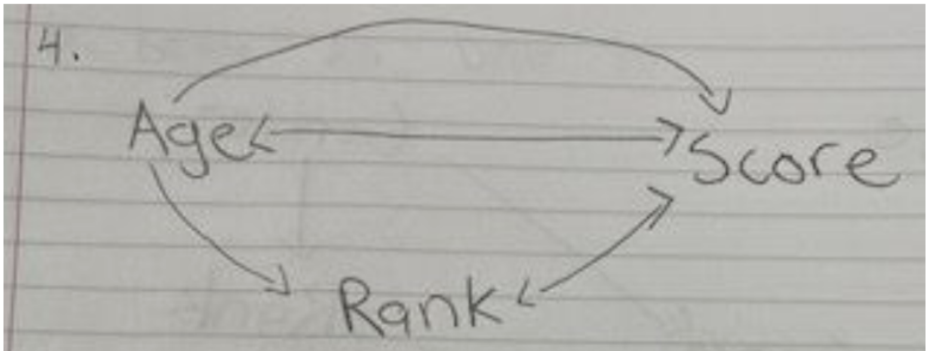


Figure 4.6. Kennedy's Final DAG in Activity 9: Evaluation

4.1.2.4 Summary of Kennedy's Observation.

Kennedy worked diligently throughout the unit having thoughtful discussions with their classmates in their group about creating DAGs, interpreting plots, and writing code in R to create plots. They came a long way in improving their graphing skills, creating various scatterplots with creative aesthetics. But their proposed DAGs still were not always aligned with what was depicted in the plots and occasionally depicted causal relationships that were not possible. The student correctly described the relationships among multiple variables in a plot when instructed to focus on individual variables first, but sometimes struggled to make those same detailed descriptions when asked more broad questions about all the relationships depicted in a visualization.

4.1.2.5 Summary of Results from Class Observations

Both observed students progressed in their ability to create multivariate visualizations using R. Though this initially took more of their attention away from reasoning about the graph and context, it became less of an issue throughout the five weeks of observation.

When reasoning about the graphs they were creating, both students began with strong interpretations of the two to three variables depicted in the stacked bar charts. They had more difficulty explaining relationships when three or more variables were used to create a scatterplot. They both struggled to use the graph itself to provide rationale for their answers unless given explicit instruction on what relationships to comment on. They both displayed some context bias or bringing in outside knowledge when elaborating on their reasoning when describing the graphs.

Jordan had more difficulty reasoning about the context of the problems when trying to describe the graphs and answer research questions, while Kennedy had more issues determining what was happening in the lines plots and being overly general in their interpretations of some relationships.

Similarly, both students were skilled at drawing DAGs in Activity 4: DAGs, easily able to determine variables that might be potentially affecting an outcome variable. Yet, they both struggled to create DAGs based on evidence from a plot. Though Jordan remained consistently good at creating the DAGs, Kennedy showed some continued difficulty in getting the arrows correct.

4.2 Results from Class Assignments

This section describes results from the three class assignments. First, an analysis of the correctness of items across the assignments is presented with respect to each of the SLOs. Then, a discussion of the results from the qualitative coding is detailed with a focus on the codes and themes that emerged from the analysis, supported by examples of student work. Finally, a summary of the class assignment results concludes the section.

4.2.1 Results Across Assignments

To gain insight into the student's proficiencies regarding the learning outcomes over the course of the unit, the assignments were coded for correctness. Select questions on each assignment were mapped to the SLO to get a sense of students' overall change in proficiency in these SLOs across the unit. The results as a percentage of correct responses on each SLO for all assignments are presented in Table 4.1. There is a complete table of those marked "correct", "partially correct" or "incorrect" in Appendix B. The items were not written the exact same way and the contexts were different across the assignments, making these difficult to compare quantitatively over time. Thus, they are qualitatively compared to give insight into how the students performed with student responses as evidence.

Table 4.1
Percentage of Correct on each SLO in Each Assignment

SLO	HW 1 (n=38)	HW 2 (n=37)	HW 3 (n=33)
1. Create graphs displaying the relationships among three or four variables in one plot	50%	91.90%	93.80%
2. Explain the relationships among three to four variables using graphs	50%	73%	21.20%
3. Identify data as observational	26.30%	29.70%	66.70%
4. Explain the limitations in making causal claims with observational data	Not Assessed	Not Assessed	18.20%
5. Create directed acyclic graphs (DAGs) to guide analysis of relationships among variables	71%	73%	66.70%
6. Evaluate DAGs already created to assess if they accurately represent relationships among given variables	NA	67.60%	69.70%
7. Develop hypothesis about variables not investigated and their relation to an outcome variable	94.70%	91.90%	97%

4.2.1.1 Student Learning Outcome 1.

As seen in Table 4.1, the percent correct increased across the assignments for SLO 1. This SLO focused on creating multivariate graphs in R Studio. Since this is the main SLO of the course as well as an important SLO for this unit, the students spent a significant amount of time learning to create plots in R Studio. They learned how to create both multivariate visualizations, adjust the aesthetics, and add annotations to plots in every in-class activity and assignment. By the end of the unit, they were nearly all able to create the scatterplot needed in the final assignment. The responses coded “incorrect” for this learning outcome on the final assignment were multivariate plots that had the wrong variables mapped to the x and y axes.

4.2.1.2 Student Learning Outcome 2.

For SLO 2 (reasoning with multiple variables), there was an initial improvement in overall class performance from Assignment 1 to Assignment 2, however there was a sharp decline in the percent correct on Assignment 3. Half of the students on Assignment 1 were able to summarize how relationships among regions and tuberculosis rates changed over time, often commenting in the aggregate and occasionally singling out countries as seen in the example below:

> Tuberculosis deaths have generally gone down over time. More developed regions had the lowest starting rates, and lowest ending rates of TB. There have been some marked increases in TB rates, outbreaks, in some regions in some years, notably 2006-2007 in South-East Asia.

However, other students, whose responses were marked “partially correct” (n=12), did not consider all the variables in their explanation of the plot. In the following examples, the responses comment on tuberculosis over time, but do not mention the regions, making it unclear if the student did not consider the region variable or thought the regions all followed the same pattern.

- *“Overtime, TB has leveled off, slightly decreased, or hardly increased as seen at the end of the plot in the year 2002.”*

- *"From what I can see on the graph, it appears that the overall tuberculosis deaths went down from 2000-2015. Most of the visible lines at the top of the graph show downward trends, particularly towards the left of the graph. However, most of the lines at the bottom of the graph are indistinguishable, so I cannot say this is true for all of them."*

For SLO 2 on Assignment 2, 73% of responses were coded as correct. Generally, students did well explaining the relationships among the house related variables, such as the following example:

> I think Bedrooms, sqft, and age are all variables associated with the price of a house. On the plot you can see the older houses are cheaper. Also, the houses with the largest [number] in sqft are the most expensive. Finally the less bedrooms the house has the lower it is in price, as you can see the orange dots indicating 2 bedrooms are all found at the lower price of $200,000.

Students' responses that were coded as "incorrect" or "partially correct" often did not include a description of all the variables or did not include any evidence from the plot to support their answer. One such answer marked "incorrect" gave no description of evidence from the plot or the nature of relationships among the variables: *"I think the three variables that play a role in purchasing a house is the age of the house, the square feet of the house, and the number of bedrooms in the house."*

On Assignment 3 it was less common that students' work was coded as "correct", with only 21.2% providing a full description of their plot. This question, like Assignment 2, required them to comment on the relationships among the four variables they used to create a visualization. However, on this assignment, not as many students referenced their plot to provide evidence for the reasoning behind their DAG, and they often left out some of the relationships that could be seen in the plot. For example, one student simply wrote: *"In the plot above, make affects the price of the car because all of the corresponding points for each* *car make follow the same linear trend lines,"* and did not comment on any other possible relationships with mileage, make, type, and price in this visualization.

4.2.1.3 Student Learning Outcome 3.

SLO 3, in which the students needed to determine whether the data they were using in the assignment was observational and explain why, proved challenging for students with only 26.3% and 29.7% of responses coded as correct on assignments one and two respectively. Correct responses mentioned that the data was observational and gave correct reasoning mentioning something about it not being an experiment for example *"observational, this data was not collected from an experiment in a lab. Instead it was most likely collected from some type of survey or database."*

There was much variability in the responses on this question, particularly in the responses that did not correctly identify the data as observational. The following quotes are a few of those responses from assignments one and two:

- *"No this is not observational data because it is actual statistics and not information like a survey in a way."* (Assignment 1)
- *"The data is not classified as observational because the information was gathered through research. Observational data is information gathered without the subject of the research."* (Assignment 1)
- *"No it's not observational because there is quantitative data."* (Assignment 2)

Some students identified the data as observational, but then incorrectly explained why it was considered observational. These responses were marked "partially correct" and made up 42.1% and 59.5% of the responses on Assignments 1 and 2 respectively. On Assignment 1, four of the responses tried to justify the data as observational by contrasting the data collection method with that of an experiment. One student claimed it was observational because, *"nothing in this set is being manipulated or changed. It is not survey data,"* potentially trying to explain the control of variables in an experimental setting but then conflating experimental with survey data. Another four students responded that the data was observational because it was collected without needing subjects to partake in the research. One such response follows:

> The data in this set is observational data. The data can be collected without the direct participation of subjects. This can be contrasted with experimental data, in which data is collected from subjects with direct participation.

Assignment 2 also included responses that contrasted the data to experimental data, with one response explaining, *"It is not data from experiments that we collected, which means it can change very easily."* The response correctly identifies the data as not experimental, but the reasoning about the data changing is not clear. Another response indicated that it was observational because it was *"taken from a sample"* and that the variables were *"not under control of the researcher,"* like responses seen in Assignment 1.

However, Assignment 2 had a couple new explanations about why the data was observational. These explanations commented on the data collection or measurement properties of the variables. One student commented that, *"this data is observational data because we are able to observe and change [the] measurements."* However, it is unclear what they mean by *"change the measurements."* Another response explained the data was observational *"because all the variables were measurable,"* which, while correct in reasoning that the variables were measurable, was not specific enough to completely explain why the data was observational.

On the final assignment, students improved identifying the data as observational, with 66.7% of responses marked "correct". These responses included some contrasts to experimental data that were more refined than previous assignments. For example, one response noted the data was observational because *"this data was not collected from an experiment in a lab. Instead, it was most likely collected from some type of survey or database."* Another explained, *"this data is observational because the researcher is observing the car data and* *not applying a treatment to it."* Others still claimed the data was observational without justification or unsuccessfully compared it to an experiment as seen in assignments one and two. Overall, there was an improvement in the percentage of correct responses for this SLO across the three assignments.

4.2.1.4 Student Learning Outcome 4.

SLO 4, pertains to making causal claims with data. This SLO was asked on Assignment 3, and only 18% of students answered this question correctly. Most students were able to identify the data as observational, but then thought this meant that they could use it to make causal claims. Ten of the responses noted this without any explanation, *"this data is observational. We can make causal claims based on this data."* But others justified their answer in various ways often alluding to creating plots and assessing relationships:

- *"We can use this data to make causal claims because we can examine how the variables affect each other."*
- *"We can also make causal claims with the data given since there are enough variables and information about the cars."*

Notably, one student claimed that we could make causal claims simply because this was real data:

> I would say that this data is observational data because the data was collected from the features and miles of the cars. It is not our own opinion. We can use this data to make causal claims because it is not people's opinion.

In these responses we can see how there was a lot of confusion among the students about when we can and cannot make causal claims with data and why.

4.2.1.5 Student Learning Outcomes 5/6.

Most students were able to create (SLO 5) and update (SLO 6) DAGs throughout all three assignments. In Assignment 1, they used DAGs to propose variables that they thought might cause differences between tuberculosis rates in the various countries and regions. Regardless of the student's background knowledge they were able to make reasonable inferences about what might affect these rates with respect to health care, politics, and economic factors. See Figure 4.7 for a detailed DAG created on the first assignment.

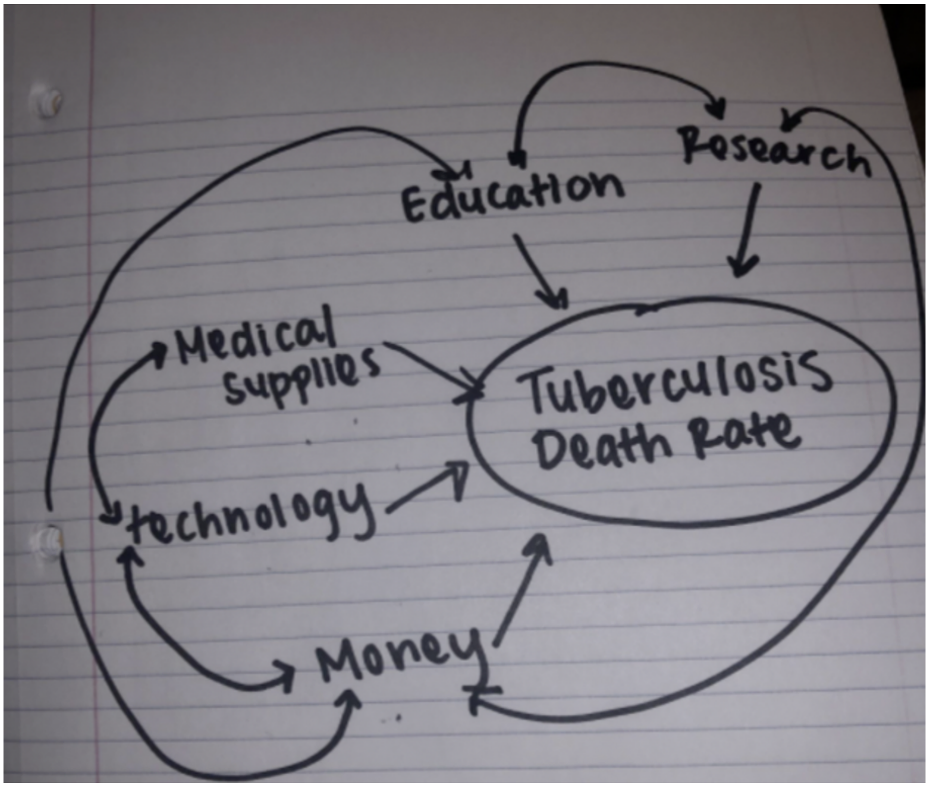


Figure 4.7. A DAG with Directed Arrows and Many Potential Causal Variables Created in Assignment 1

In Assignments 2 and 3 the students had to use the data they were given to make DAGs that predicted what relationships they thought they would see in the data prior to plotting it. Later in the assignment they refined the DAGs based on the visualizations they created.

Again, most students were able to complete this task and justify their answers.

Responses were marked "incorrect" or "partially correct" when coding the student's DAGs because they included lines instead of directed arrows or included some arrows that were not temporally possible (e.g., drawing an arrow from price to age, indicating that the price of the house affects the age of the house). Other "partially correct" responses had directed arrows only between the outcome variable and potential causal variables but not between the potential causal variables where a graph indicates there might be a relationship as seen in Figure 4.8 on Assignment 2.

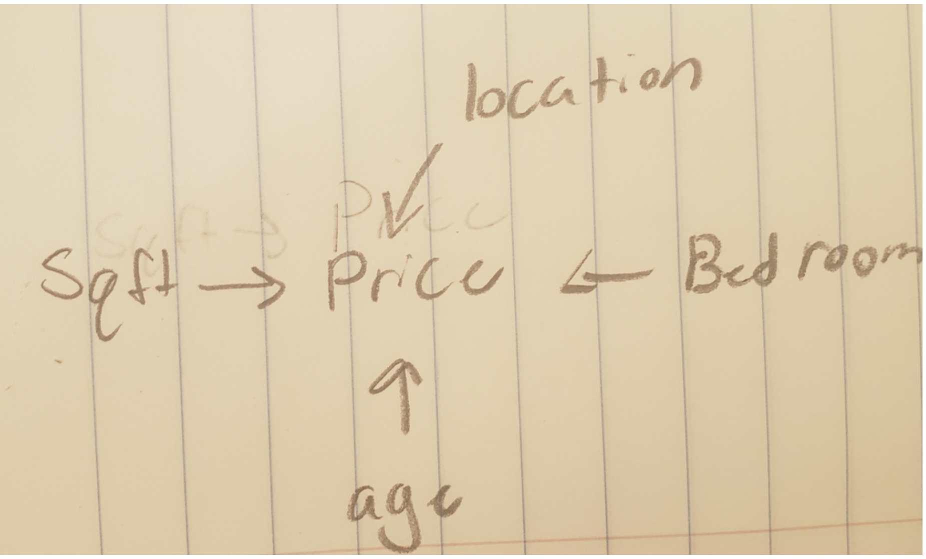
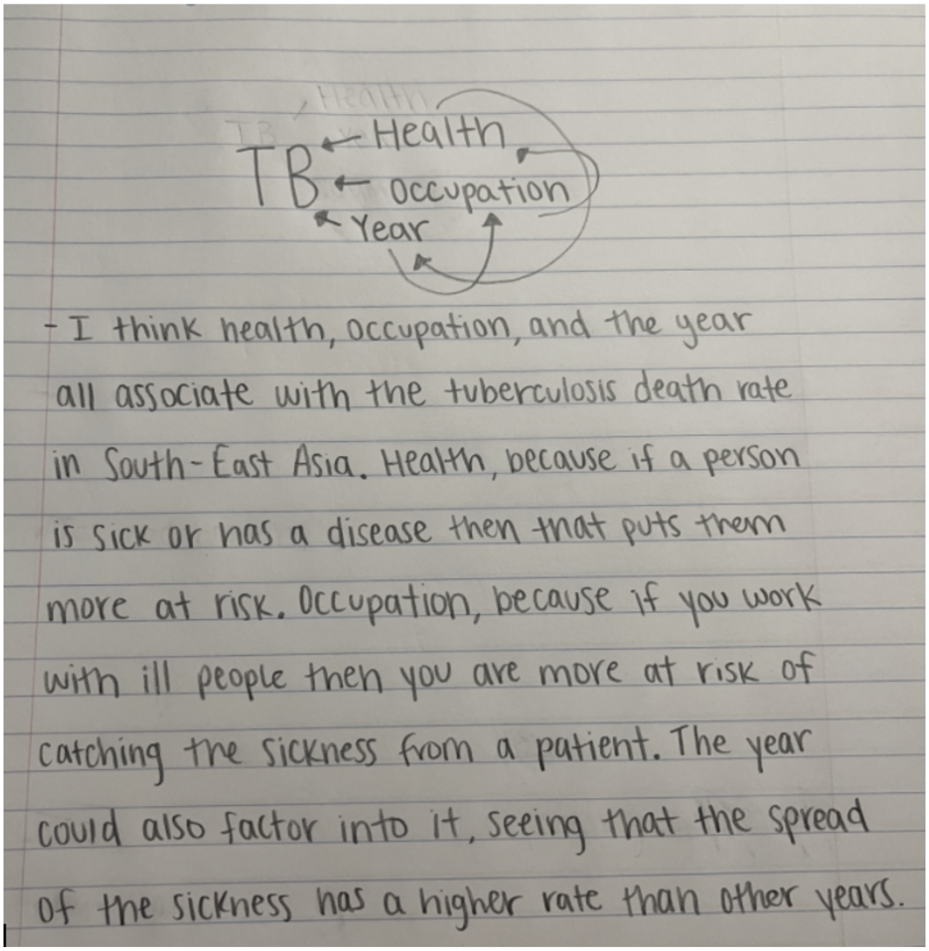


Figure 4.8. A DAG from Assignment 2 with no relationships proposed among the variables except with price.

Mostly, students did well creating DAGs to help them make predictions and communicate about the relationships they saw in the graphs.

4.2.1.6 Student Learning Outcome 7.

Nearly all students (91.9%-97%) were able to hypothesize about other variables that could affect the outcome variable and justify their answers reasonably. Incorrect student responses varied greatly, though they were few. For example, one student wrote on Assignment 2 that in determining what factors affect a house price,*"variables include, square feet and age. This is because these are the main factors that determine the price of a house, other variables are unnecessary."* This student only listed the variables investigated in the assignment and did not come up with any other variables. A few students (n=5) on Assignment 1 provided DAGs of potential variables affecting the tuberculosis rate. The proposed variables were not fully justified in their description and did not accurately describe factors that could affect the tuberculosis rate at the country level. In Figure 4.9, the arrow indicating that year affects the occupation of the person is of note.


TB
Health
Occupation
Year

- I think health, occupation, and the year all associate with the tuberculosis death rate in South-East Asia. Health, because if a person is sick or has a disease then that puts them more at risk. Occupation, because if you work with ill people then you are more at risk of catching the sickness from a patient. The year could also factor into it, seeing that the spread of the sickness has a higher rate than other years.

Figure 4.9. Example of a DAG with Questionable Direction of Arrows

Overall, this SLO was consistently achieved throughout all the assignments.

4.2.2 Codes and Themes Across Assignments

In addition to coding for correct and incorrect responses the data was also coded for notable multivariate thinking challenges identified in the literature review such as any context that interfered with their multivariate reasoning ("context-interference") or any confusion about the variables ("variable-level-confusion"). In addition to these codes, the "misc-interesting" code was created for anything that did not fall into those categories but was unusual or unique. These sub-codes are presented next along with exemplar student responses.

4.2.2.1 Context Interference.

The code "context-interference" was initially coded for any response on an assignment in which the context played a large role in the written response, in a way that went beyond what was asked in the assignment. This code was used 25 times. In a second iteration of coding of this data, four distinct sub-codes emerged within the responses coded for context interference.

The most prominent sub-code (coded nine times) pertaining to context interference was one indicating that the student was using the context to try to help support their conclusions. In these responses the students essentially used what they knew about the topic to make sense of what they were seeing in a graph, but they did not extrapolate beyond what was seen in the plot. For example, one student wrote on Assignment 1 about Tuberculosis, *"on the line plot Tuberculosis deaths looks like it's been going down over the years. I believe the main reason why it's been trending downward is because there seems to be more resources to combat Tuberculosis."*

The next most frequent sub-code within context interference (coded six times) was a code indicating that the student was making an unsupported claim about the data based on context alone, and not on evidence from the plot. This occurred most often on Assignment

3 (five out of the six times). Often, students responded only with outside knowledge, leading them to inadequately answer the question based on the plots. Occasionally, their response with outside knowledge contradicted what was depicted in the plots. On Assignment 3, for example, when asked what type of car they would recommend their friend get in Question #15, a student gave the response:

> Finding a GM vehicle for $40,000 I would recommend them buying it new. Seeing as a vehicle adds mileage it's price decreases, $40,000 would be near the more luxury offerings of a GM made vehicle. If they are forking out that much money for a car, I would recommend they [buy] it new as 2005 was almost 20 years ago. The 2005 car would not be worth $40,000 unless it was new, and even then, specs on cars change and parts become outdated.

In this response, this student is ignoring the points in the plot depicting cars at the $40,000 price point to solely give a recommendation based on their contextual knowledge.

Another common occurrence within responses with context interference was extrapolation of evidence from the plot. This occurred five times, most commonly on Assignment 1 (4 times). For example, when describing what conclusions they can draw from their plot, this student noted the decrease in Tuberculosis rates and then wrote, *"from this data, I can infer that the world has become more interconnected and medicine has advanced and become more accessible,"* which is not represented in the plot alone.

The final sub-code applied to five responses that suggested a misunderstanding about the variables or context. One student thought the "sound variable" in the cars data was in relation to the sound the car made and not the sound system used in the car. Other occurrences were when students were confused about what variables to include in their DAGs. One student included a variable for "deciding on a car" as a part of his model for variables that increase the price of a car. This student thought the assignment was about factors that affect a person's decision to buy a car and not about factors that influence the price of a car.

4.2.2.2 Variable-level Confusion.

The code “variable-level-confusion” was used to indicate that the students had trouble distinguishing a potential causal variable and the levels within that variable. This code was used only eight times, and all its uses were in the creation of DAGs. Half the time this code was used it was on the Tuberculosis assignment for choosing potential variables that might affect the death rates.

In the literature, participants in the studies often worked with categorical data, but in this study, students worked with both categorical and quantitative data, so this code was also used to indicate confusion between a quantitative variable and a particular amount of that variable. For example, out of the eight times this was coded, two of those times was to indicate the student labeled their DAG with an amount of a variable rather than the variable alone. In a DAG created to model potential causes of tuberculosis deaths, one student wrote their potential variable “Limited Health Supplies” instead of “Amount of Health Supplies”. Similarly, one student was commenting on tuberculosis rates in an entire country but included variables at the individual level. They had arrows from “how [hygienic] are you” to “tuberculosis deaths in the Americas”. Figure 4.10 depicts an example use of variable-level confusion in which the student has used both country level variables and individual level variables.

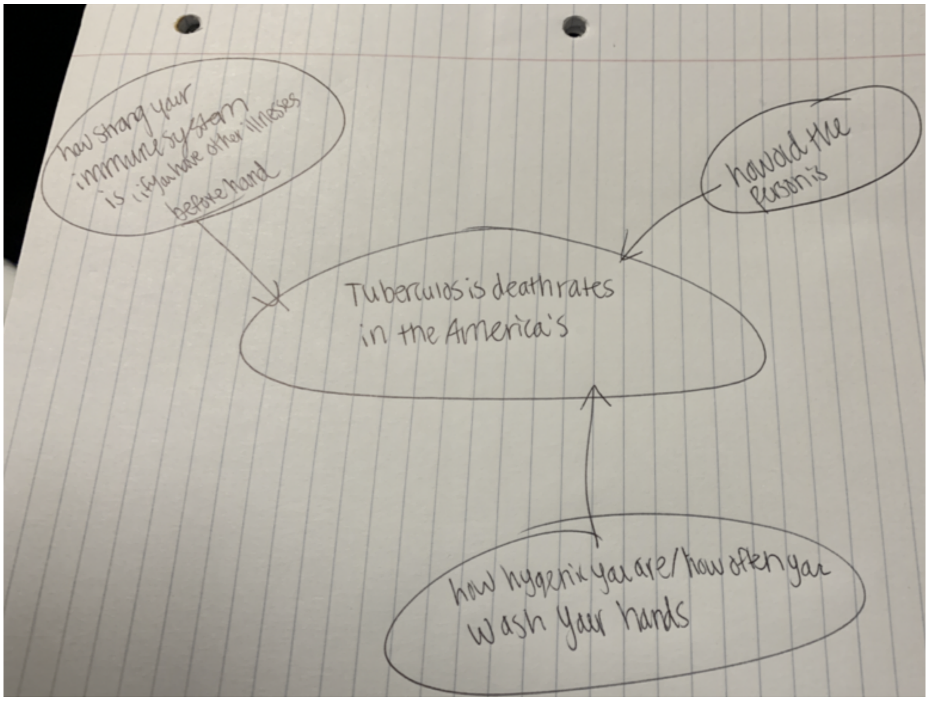

how strong your immune system is / if you have other illnesses before hand
how old the person is
Tuberculosis death rates in the America's
how hygenic you are / how often you wash your hands

Figure 4.10. Example DAG of Student Confusing Variables and Degree of that Variable

The most common use of this code was to denote misinterpretation around what variables should be included in the DAGs. Four responses used them as a framing for what the DAG was supposed to be denoting. Two of these responses added a marker for "relationships" and one response indicated the student thought Assignment 2 was about factors that affect one's decision to buy a house, instead of what variables affect the price of the house. Those examples are seen in Figure 4.11 below.

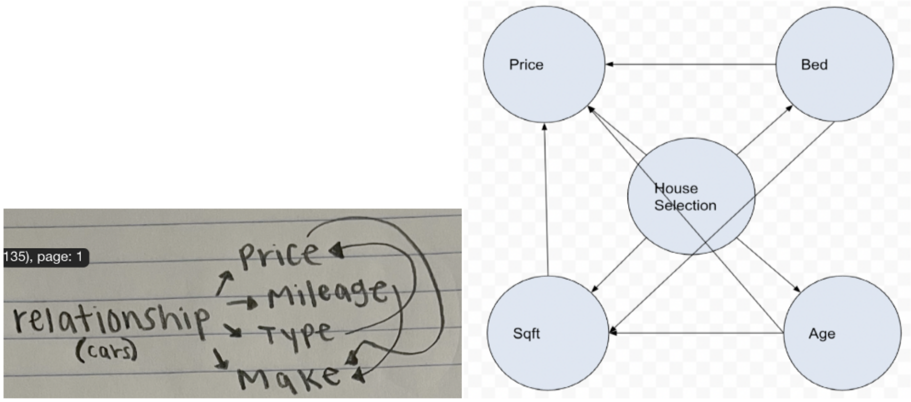


Figure 4.11. A Student's DAG Including an Extra "Variable" for Relationship, on the Left. On the Right, a DAG with "House Selection" in the Center.

The one other occurrence of this code was in reference to a student using multiple variables within the same node. This student wrote that one variable potentially affecting tuberculosis rates was "a country's scientific prowess and economic strength" not indicating that these might be separate variables.

4.2.2.3 Miscellaneous Code.

The code "misc-interesting" was used to classify unusual responses that did not fit within other codes throughout the first round of coding. This code was used 12 times. Further analysis of the responses assigned this code identified three subcategories: misunderstanding the question being asked, responding to the question in a way that contradicts the plot created in the assignment, and, in one case, just confusion about the entire assignment.

For students that misunderstood the question being asked, they often understood the context but did not respond directly to the question. This code appeared three times. One such example in the Cars assignment happened in the final question asking students to give car qualities they would want in a $40,000 car based on their plot. For this question, one student gave a reasonable answer, but it was framed in terms of reselling a car and not buying a car. The other two responses in this category were from students recommending that their friend buy a new car or a different type of car for less money:

> I would recommend that they buy this car as a Chevy coupe or a Cadillac sedan. Those both seem to have a pretty steady mileage rate at that price point, so you would be able to buy a nice car with less mileage for $40,000. If they were to get any other car, they risk either having higher mileage with that high of a price, such as a convertible. Or, they would be able to get a car with way less miles for cheaper if they were to get a chevy hatchback. Based on the data, that specific car tend[s] to be lower in mileage and a lower price.

Most commonly (six occurrences) the miscellaneous coded responses had to do with students providing a rationale that contradicts what is seen in the plot or potentially misreading their plots. For example, the response below describes how age does not affect price based on the plot, but then continues to list age as a variable affecting price in the next question, though hedging the answer by saying they all affect price to varying degrees.

> 5. The DAG above represents the variables that [affect] the price of a house. The two variables listed that have an impact on the price are square feet and bed. As shown on the plot, houses with lower square feet are cheaper, and houses with higher square feet are more expensive. The colors of the points are the number of beds in the house, and the plot shows a variety of colors plotted at different points. The age of the house however does not seem to have an effect on the price.
> 6. There are many variables that are associated with the price of a house. These include: age, bed, bath, and square feet. However, not all of these variables have the same level of effect when determining the price.

More often the students did not catch that their response did not match the plot while making conclusions. For example, in the last question of Assignment 3 one student wrote:

> I would recommend my friend buying this car if it [has] four doors. As we see from the plot majority of the cars do [have] four doors. Also, cars that are $40,000 seem to have good mileage to them as well, so I would suggest buying the car. Lastly, I think my friend should by this car if [it] is a Sedan, Hatchback, or Wagon because they seem to have the least miles.

This answer does not match what is depicted in the plot as the cars are more likely to have two doors if they were priced around $40,000, and the sedan, hatchback and wagon styles were least likely to have prices around $40,000.

4.2.3 Summary of Results from Class Assignments

Overall, students improved in their ability to create graphs and identify the data as observational over the course of the unit. Their ability to discuss relationships among two and three variables was strong to start the unit, but they found reasoning with more than three variables challenging later in the unit. They remained apt in their ability to create and update a DAG and provide logical potential causal variables. Most students still needed more practice explaining when we can make causal claims with data, despite the focus of this unit on multivariate thinking and causal claims.

A small number of responses indicated that students were making similar mistakes to those identified in the literature such as confusion around variables and their levels and confusion about contexts (both in terms of the dataset and what the activity was asking of them). Some students aptly used their outside knowledge to provide a rationale for what they had described seeing in their visualization while others answered questions about their graph using only outside knowledge. Occasionally, those that used only outside knowledge provided responses that directly contradicted evidence from their plots.

4.3 Results from Final Cognitive Interviews

This section discusses the results of three cognitive interviews with students after they turned in the final assignment in the multivariate thinking unit. One student was from Section 1 and two students were from Section 2. Each student took part in an interview in which they gave a rationale for their answers on the assignment. Notable responses to questions, reasoning, and discussions with the student are presented in this section.

4.3.1 Results from Cognitive Interview with Student 1

The first interview was conducted with a student via Zoom and lasted for approximately 20 minutes. The student started the interview by talking through their logic for the first question, explaining that the cars data was experimental because *"it was based on different variables and taken from research, and it wasn't human based."* Because the participant was unsure of this answer and it was not correct, we had a short discussion about the difference between experimental and observational data. They noted that the data was observational because there were no experimental or control groups randomly assigned to make it an experiment. Though the student indicated they understood that it was observational data, and we could not make causal claims with it, they then made a statement about the *"higher mileage causing a higher price in cars"* when describing their scatterplot in the next question.

Then the student described their visualization with four variables. The participant was able to detail the relationship between all the variables in the plot accurately and seemed to have a good understanding of what was happening in the plot, demonstrating their skill at multivariate thinking.

Next the participant explained their DAGs for this assignment. They updated their DAG, found in Figure 4.12. The top DAG was their initial prediction of the relationships among the variables and the bottom represents their updated theory after analyzing the graph with all the variables. They described how they initially thought that the make of the car would affect the mileage of the car, but that was not depicted in the graph. They updated their final DAG to reflect only the relationships they noticed in the graph and added additional comments to clarify the relationships.

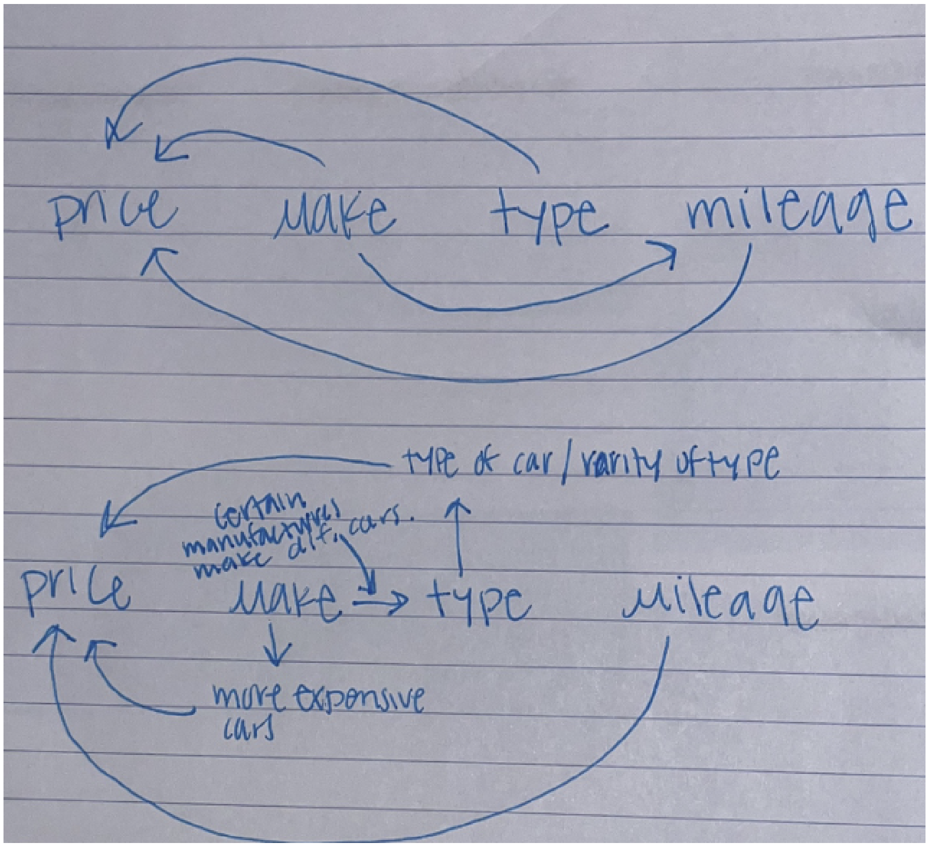
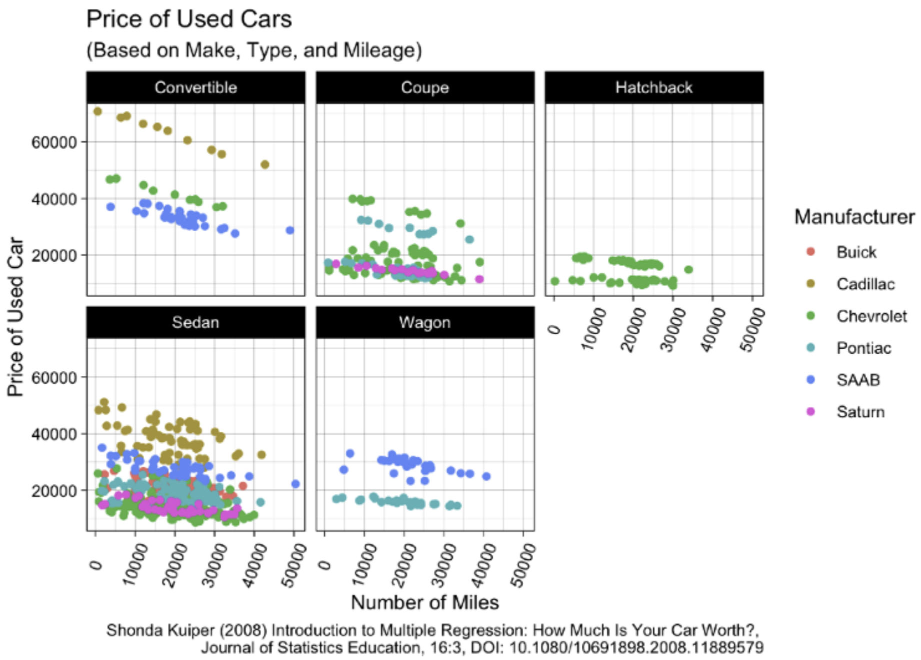

price
make
type
mileage
type of car / rarity of type
certain manufacturers make dif. cars.
price
make
type
mileage
more expensive cars

Figure 4.12. Student's DAGs for Assignment 3 with Initial Prediction (top DAG) and Final DAG (bottom DAG)

In the last question of the cars assignment the student was asked what features a used GM car would need to have to justify a purchase price of $40,000. The student returned to their plot in Figure 4.13 to answer this question.



Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

Figure 4.13. Multivariate Cars Visualization from the Student's Assignment used for Answering the Final Question

> I looked at this graph. . . .I said that . . . based on the plot, if a car is a Buick, to not buy the car because . . . it's . . . there (indicating with their mouse that the points were all under $40,000 for these cars). So I didn't think that would make sense. And then I said, if the car is a Chevrolet, only to buy it if it was . . . a certain type like the coupe or the convertible with 10,000 miles or less. And then I said just to not buy the car if it was a Pontiac. I would only buy the car if it was 20,000 miles or less [if it's a] SAAB. Not to buy the car if it was a Saturn [because] same thing with . . . the [B]uick it's like up here." (indicating with the mouse that the $40,000 price point is much higher than any of the point on the plot for Saturn cars.) But yeah so like for like [C]hevrolet and stuff it'd be like (hovering the mouse over only the cars that were priced around $40,000). . . yeah I said not to buy . . . it under 30,000 miles, (while gesturing at the plot in the $30,000 range). . . .I would say, only buy it if it was . . . a [C]adillac honestly.

The student's reasoning for selecting the cars they described was directly related to the plot, as evidenced by them highlighting certain points on the plot with their mouse, as well as their overall description of the graphs. However, they did confuse the price and mileage axis at the end of the description. It was unclear if this was an accidental switch or if the student thought the y-axis was mapped to mileage.

4.3.2 Results from Cognitive Interview with Student 2

Interview 2 took place over Zoom and lasted approximately 15 minutes. This student responded to the first question asking about the nature of the data by writing that it is observational, but we can make causal claims with it. When prompted about the reasoning behind that response, they described that they knew that the data was not from an experiment so it must be observational data but thought we could make causal claims with it regardless. A short discussion followed explaining the difference between experimental and observational data.

This student did not make any updates to their DAG that they initially created after looking at the visualization. Their DAG and visualization are pictured in Figure 4.14. Notably the DAG features all variables potentially affecting price, but with no relationships between them. They described in their final DAG that they did not think that there was any relationship between the make, mileage, and type of car variables so they did not include any arrows.

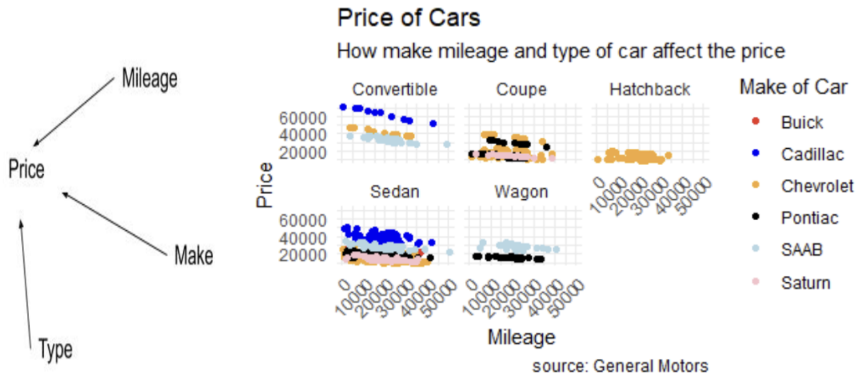


Figure 4.14. DAG (left) and Plot (right) from Assignment 3.

Student 2: I think that type, make, and mileage all affect the price, but between the other three variables type, make, and mileage I don't think that there really was a relationship between those. They all affect the price and not so much the specific mileage or type of car.

Me: If we ... think about them individually for a minute, like mileage and the make of the car. So ... the makes - the different ... Saturn Buick, Cadillac - I definitely don't see any of those lasting particularly longer than the others, so that definitely makes sense. Then the type of car and the make of the car. So the type of the car. The only one in this one that kind of stands out to me is, we only have ... chevy making hatchbacks. And then ... only three of the different brands make convertibles, so it's almost like the different makes [effect] type because you only have so many different types for each make.

Student 2: Yeah that's true.

Me: Yeah so you ... could have had a line there.

Student 2: Mm hmm.

With further discussion and examination of the plot, they noticed that there were different types of cars for each make of car, so there might be a relationship between type and make, but they did not follow up with any questions or further discussion for clarification.

When creating their own graph to display the relationships among the variables, the student created a unique plot that they had not made in the course before, in which a categorical variable was mapped to the *x-axis*. Figure 4.15 depicts this plot. The student acknowledged that it was not helpful for determining the relationships among the variables, stating, "I didn't really like this graph much. I thought it was hard to read the mileage numbers, they are all the same size." Because of this, they used their plot in question to answer the final questions about the graph.

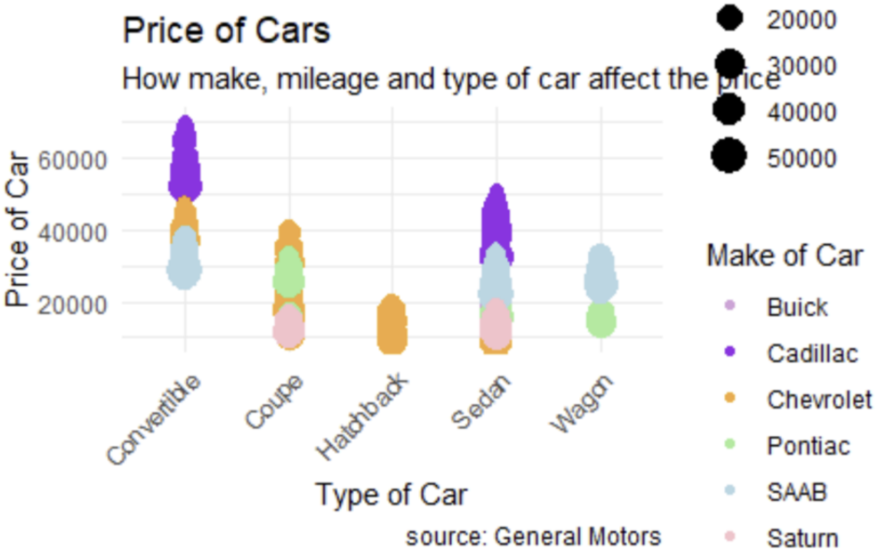


Figure 4.15. Student 2's Uique Plot for Assignment 3

In answering the final question about buying a car for $40,000, they used their plot from Figure 4.14 stating:

> I would recommend they buy the car, as like a Chevy, Coupe, or Cadillac sedan because they have a steady mileage ..So you can see, they have a pretty rea-

> sonable price. . . .I guess the mileage should go up, but not the price point. . . [I]f they were to get any other car they risk either having a higher mileage for that price. Like a convertible would be a lot higher but they could get a car with way less mileage for cheaper if they were to get a chevy hatchback. So I don't know if they have a few options. It depends on what [they] were looking for and their preference.

In this response the student mostly read their answer from the activity. They did not give any additional specific details, and only gestured to the plot vaguely a couple times when answering the questions. They framed their answer in terms of what used car the person should buy that would last a long time, rather than what used cars would be worth a purchase price of $40,000. Their last sentence makes this clear by indicating that they could get something for cheaper, but it depends on the buyer's preference. It is unclear whether they did not understand the prompt or how to use the graph to interpret the plot. Their response became even more unclear when I followed up to ask what features from the plot helped them answer the question:

> It was pretty easy to see that they shouldn't buy convertible just because the price was high. . . [A] wagon was way lower than their price point, so the other ones were a little bit confusing because, like a Sedan there you can get basically any makeup the car for that price. But and if it's from 2005 the mileage is probably higher. The response seems to further indicate that they misunderstood what was happening in the graph evidenced by their explanations of which cars were typically priced higher. Additionally, their understanding of the question or context itself might have been misguided because they are responding that a car from 2005 would have a higher mileage, but all the cars in the graph are from 2005.

4.3.3 Results from Cognitive Interview with Student 3

Interview 3 lasted approximately 12 minutes. Notably, the student was connected to Zoom on their phone (in their car), so they did not have access to the documents themselves, only the screen I was sharing via Zoom. This student had a strong grasp of the first question asking about the nature of the data and whether causal claims could be made. They explained they had taken another psychological research methods course and they drew on information from that to answer the opening question of this assignment. When asked about their reasoning for their response they replied:

> This data came from General Motors right, if I remember correctly. I knew . . . from past kind of research methods class that if . . . it's not coming from an experiment, where you know you have randomization of groups - different . . . trial groups like a control group, and . . . all of those factors you can't really make causal claims unless all of those . . . randomization and . . . confound and things are kind of accounted for.

This student, far more than the others talked about drawing on their own experiences for making their predictions about the relationships among the variables:

> I just kind of was thinking about based on my personal experience of like—my parents have always bought exclusively used cars and so . . . , deciding not only when to turn your own car in because of mileage . . . I know we bought one brand new car in our life, and that was by far the most expensive one so that was kind of what I drew on for that.

Next the student talked through updating their DAG after seeing evidence from their visualization. Their plot is pictured in Figure 4.16 and their initial DAG and updated DAG are pictured in Figure 4.17. They were able to clearly articulate what features in the graph made them get rid of lines in their DAG and keep others.

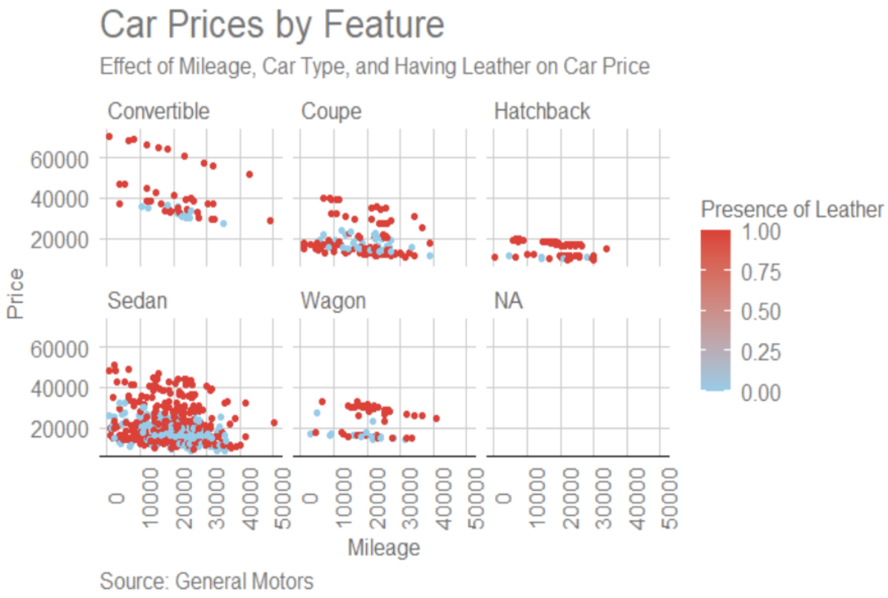


Figure 4.16. Graph of Mileage, Leather, Price, and Type of Car.

> It seemed like things were pretty evenly spread [in the plot], which I don't know why that is because, again I don't understand cars… I didn't add the arrow for leather again because I didn't think the relationship was super strong, so I don't know I didn't really want to. But definitely it seems like there was more leather for certain types of cars than others.

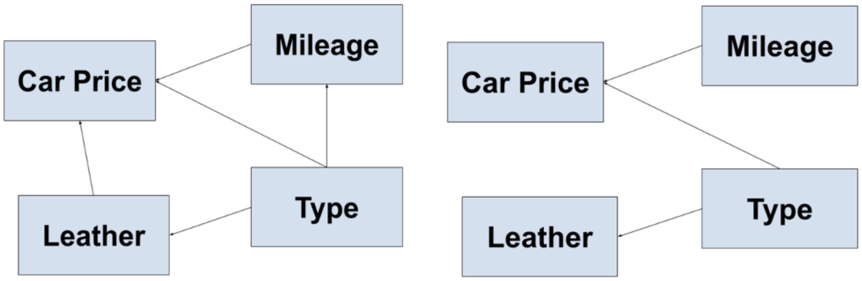


Figure 4.17. Predicted DAG (left) and Updated DAG (right) for Assignment 3.

To create their own graph (Figure 4.18), this student put a categorical variable on the *x-axis*. When asked to describe their plot, they made note of the categorical variable on the *x-axis*, describing, "… of course that's how you get these weird looking columns … because you put a discrete variable on a continuous scale versus having two … continuous variables like you did with mileage and price". The difficulty reading this plot led them to use their original plot of the four variables, created in the previous question, to answer the final question on the assignment.

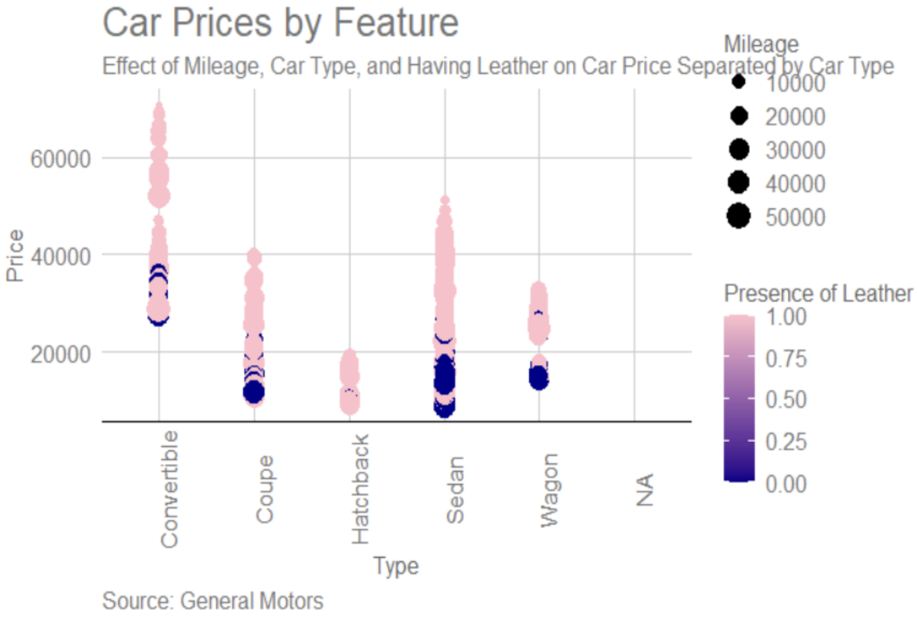


Figure 4.18. Graph of the Four Variables; Mileage, Leather, Type, and Price.

Using the plot in Figure 4.16, the student described their method for determining what features they would expect if they paid $40,000 for a used car. In their description we can see that they understood they could answer the question by looking for cars that were priced, on average, $40,000 or more. Once they found those points, they looked at the graph for the typical mileage, presence of leather, and type of car at those points.

> I was looking on the graphs and seeing … for what conditions that $40,000 would either be below the average or about the average price…Definitely should have

> leather and 20,000 miles or less because all of those kinds of conditions lined up in that point that would make $40,000 below or average. And yeah I just kind of looked at where points were at $40,000 and it was those where it was either convertible, certain types of sedans . . . and then all the ones that had leather were $40,000 ..That was definitely something that should have been expected at that price point and then 20,000 was [where] that price and the mileage intersected.

4.3.4 Summary of Results from Cognitive Interviews

The cognitive interviews with three students provided insight into some of the challenges the students faced completing the multivariate thinking assignment and gave a sense of their final multivariate skills at the end of this unit. The students demonstrated they could easily create multivariate graphs that helped them answer research questions when they were explicitly given instructions on how to do so, but they created less insightful plots when told to create something new. However, they were good at critiquing their own graphs and realizing that they were not as useful as the scatterplots they had initially created.

Two of the students in these interviews did not understand what observational data was or whether we could make causal claims with it, but after a brief one-on-one discussion they seemed to indicate they had a better idea that this data was observational. The third student was more informed on the difference between observational and experimental data and the ability to make causal inferences from previous coursework.

All three of the participants interviewed claimed they did not know much about the context of cars. However, they were easily able to draw on their own experiences with cars to create DAGs and discuss other hypothetical variables that might be affecting the price of a used car.

Two of the participants were able to correctly use their graphs to answer the final question on the assignment. One student gave a very clear description of their process of determining features of cars that were prices at or above $40,000 on average, while the other student went through each faceted graph to describe the cars in each type that would be reasonably priced $40,000 to give a complete answer - only confusing the mileage axis with the price axis at the end of the task. The third student struggled to make it clear what cars they thought fit this description and even suggested some that did not fit this description, potentially due to confusion about the question itself.

4.3.5 Chapter Summary

Over the course of the multivariate thinking unit, students improved in their ability to create multivariate graphs using R. The percentage of correct graphs increased over the three assignments. The two students observed in class showed progress in their coding skills for creating multivariate visualizations while working on the activities. Initially, the coding slowed them down as they worked on the assignment, but it became less of an issue throughout the five weeks of observation. The students in the cognitive interviews also demonstrated they could easily create multivariate graphs that helped them answer research questions when they were explicitly given instructions on how to do so and were good at critiquing the graphs they created for the assignment.

Overall students' reasoning with multiple variables improved throughout the unit, until the assignments and activities asked them to reason with more than three variables. The initial increase and subsequent decrease in percentage correct across the three assignments illustrates the difficulty the students had as more variables were introduced. The students observed in class demonstrated that they could readily interpret the two to three variables depicted in the stacked bar charts and scatterplots. However, they had a difficult time picking out each of the relationships depicted in a plot with three or more variables on various in-class activities.

At the end of the unit students still were not able to easily determine if it was appropriate to make causal claims with their data. Only 18% of students answered the question on the final assignment about this correctly. The cognitive interviews revealed the extent that two of the students were unsure of their answers, but a brief discussion of this topic helped them identify the data as observational in the moment.

Students remained consistently apt in their ability to create and update a DAG and provide logical potential causal variables. Although students observed in class and those in the cognitive interview expressed varying degrees of unfamiliarity with some of the contexts used, they were easily able to draw on their own experiences to create DAGs and discuss other hypothetical variables that might be affecting the systems in the activities and assignment.

Overall, observing the two students in class gave insight into students' reasoning as they learned to create and reason with many variables. Analysis of the SLOs across the three assignments helped identify trends in the students' performance on SLOs and identified similar challenges seen in the literature, such as confusion about variables and the context of the data. Finally, the cognitive interviews with three students elucidated some of the concepts they were still trying to understand and gave a sense of their final multivariate skills at the end of this unit.

Chapter 5

Discussion

This study investigated the development of student's multivariate thinking in an undergraduate, introductory data visualization course. As a part of this study, a collection of in-class activities and homework assignments were created and implemented in Fall 2021. The study also consisted of observing two students as they worked in class and conducting cognitive interviews at the end of the unit. Audio from these sessions, observer notes, submissions of student in-class work, and all student assignments were qualitatively analyzed to answer the research questions. To facilitate multivariate thinking, a five-week unit consisting of 10 activities and three assignments were created to introduce students to multivariate thinking, while they also learned to create visualizations in RStudio. The activities and assignments incorporated GAISE suggested SLOs to promote multivariate thinking and integrated DAGs to communicate and make predictions about relationships among variables. The materials for this unit went through multiple rounds of feedback and revisions, including four think-aloud interviews to improve the three assignments. The think-aloud interviews resulted in updates to the assignments that further aligned them with the course content, improved the clarity of the items, and helped eliminate redundant items.

A student from each section of the course was observed as they worked on the in-class activities, to gain a sense of the development of multivariate thinking throughout the unit. These students talked aloud to their group or to me as they worked on the activities, asking questions, having discussions about the context, or collaborating on coding, as needed. Additionally, assignments from all students in both sections were collected to provide additional information about the students' reasoning over the course of the unit. Responses on the assignments were qualitatively coded for correctness, any expected common challenges or misunderstandings based on the literature, and any new themes that emerged during coding. The challenges and new themes were then further investigated.

At the end of the unit, cognitive interviews were conducted with three students to obtain insight into their multivariate reasoning abilities as they explained their responses on the last assignment. These discussions provided further evidence of the students' rationale for their graphs and DAGs and illuminated several challenges they faced completing the assignment. This chapter provides a discussion of the results pertaining to each research question, followed by limitations and implications for teaching and research.

5.1 Research Question 1

In this section the first set of research questions are discussed, highlighting relevant qualitative results from the in-class observations and all students' assignments throughout the unit. The questions addressed are: *How does students' multivariate thinking develop as they take part in a series of activities designed to introduce and promote reasoning with multiple variables? How do student responses to questions requiring multivariate thinking change throughout the semester?*

As seen in Table 5 (in Chapter 4) there was fluctuation among the students' performance on the SLOs across the unit. The table gives a general impression of their competencies in these SLOs throughout the unit. Regarding SLO 1 (creating multivariate visualizations), the students' abilities to create multivariate visualizations, specifically scatterplots and line plots, improved over the course of the unit. From Assignment 1 to Assignment 3 we see an increase in the percentage correct from 50% to 93.8%. Every activity in the unit (except the activity introducing DAGs) required the students to create at least one visualization with two or more variables and learn new customizations to the plots (e.g., adding titles, customizing colors, changing themes). Since this was a main goal of the course and the multivariate thinking unit, building on this skill was the focus of many of the activities and assignments.

Additionally, use of R to create the graphs with the **ggplot2** package was likely new to the students in the course, so the coding took the majority of their attention, especially when new plots or customizations were introduced. Both students observed in the class made this apparent, because learning to code the graphs took up most of their class time in the first assignments, but became easier in later activities. The students observed in class spent time during the initial activities orienting themselves to using R Studio, uploading data, and debugging their code to create graphs, leaving them less time to reason through the relationships in the graphs. However, with the introduction of DAGs and the subsequent activities focusing mainly on scatterplots, they then focused more on the relationships among the variables.

SLO 2 (reasoning with three or more variables) was also a primary focus of the multivariate thinking unit. Results indicated that there were more correct responses reasoning about relationships among variables on Assignment 2 than Assignment 1 (50% to 73%), but on Assignment 3 there was a decrease in the percentage of correct responses (with only 21.2% correct responses). Most students on the last assignment were not able to support their reasoning using evidence from the plot or they did not include a description of all the relationships in the plot, and thus it was not coded as "correct". In both Assignments 2 and 3 the students created plots with four variables. However, the mappings of the plot and the contexts were different. For Assignment 2, the students worked with data in the context of house prices and created a graph with variables mapped to the x-axis, y-axis, color, and size of the points. In Assignment 3 however, they worked with data about cars and created a plot with variables mapped to the x-axis, y axis, color, and faceted on the fourth variable.

Though they had seen faceted graphs previously, the variable they faceted on in this case had more levels than previous graphs they created using faceting. On Assignment 3, responses indicated students potentially wrestled with the car context or the assignment itself being unclear. It remains uncertain how much, if at all, the context or different style of graph played a role in the student's reasoning about the relationships among these variables.

Students observed in class also demonstrated the ability to reason with their graphs in the initial activities but had more difficulty reasoning with the newer plots featuring more variables. Jordan found it difficult to make sense of so many variables at once and expressed concern over whether considering all these variables could really help answer the research questions posed in the activities. Teaching multivariate thinking is ideal for addressing these issues, however discussions on the importance of stratifying on a variable to avoid Simpson's Paradox was only included in the last activity, which is perhaps too late. Kennedy incorrectly overgeneralized some relationships in the graphs created in Activity 7/8: World Data. They made conclusions looking at only a couple countries (points in the graph) to generalize relationships across the entire graph (for all regions). Overall, students in the course improved in their reasoning, though there were continued challenges working with three or more variables on later activities.

Regarding SLO 3, identifying data as observational improvement was seen over the course of the unit. In Assignment 1, 26.3% of the students correctly identified the data as observational, but in Assignment 3, the percentage correct increased to 66.7%. This concept was explained to students in Activity 4: DAGs and was asked on subsequent activities and assignments. Most students were able to deduce that the data used in class and on assignments was always observational, potentially because of reviewing their feedback on their assignments and activities.

SLO 4, explaining if a causal claim could be made with the data, proved difficult. This SLO was also introduced in the Activity 4: DAGs but did not show up as frequently on other in-class activities or the first two assignments. As demonstrated in Fry's study (2017), this is a difficult topic to fully understand in a short period of time. Even trying to simplify this idea to not go into full details of study design still was not enough to give students a surface level understanding of when we can and cannot make causal claims. The activities and assignments did not cover these concepts enough and were not supported with fruitful classroom discussions. Thus, students did not make as much progress on SLOs 3 and 4 as the other SLOs.

Generally, most students were consistently able to create and update DAGs (SLOs 5 and 6) throughout the semester, with ease, after their initial introduction. Though there was some confusion initially during the in-class activity, providing feedback on their in-class assignments helped around two thirds of the students consistently get these questions correct on the assignments. The remaining one third of the class was consistently getting this wrong in some way throughout the term. This was mostly due to small errors in their lack of correct facing arrowheads or their confusion about the context. This could be because they did not attend and participate in Activity 4: DAGs in class, did not review their feedback on activities or assignments, or generally did not understand the proper form for creating the DAGs.

Jordan created DAGs that were consistently thoughtful, typically aligned with the relationships they described in their plots, and updated them when new information was discovered. However, Kennedy had a little more trouble determining the arrows for their DAGs throughout the unit. When probed for reasoning behind their DAGs, they occasionally did not cite evidence from the plot but rather their instinct or “feeling”. Overall, students’ use of DAGs was adequate, yet there was more work to be done to get students to consistently use the right form to create the DAGs (e.g., use arrowheads) and to create and update them based on the evidence from their graphs.

For SLO 7, students were able to hypothesize about other possible confounding variables relatively easily from the introduction to the DAGs activity onward. On each of the assignments at least 90% of the students were able to consider other variables that might affect the systems. Students observed in class were also able to brainstorm many variables alone or working with a group despite their level of outside knowledge about the contexts.

5.1.1 Summary of Research Question 1

In general, the students in this study showed some development of multivariate thinking skills over the course of the semester. They demonstrated improvement in their ability to create multivariate plots, identify data as observational, consider potential confounding variables, and create and update DAGs to model potential relationships among variables. Student assignments and class observations indicate that the students came a long way in their ability to create the graphs using R, but not as far in their ability to reason with many variables. This is a difficult skill, especially incorporating evidence from the plot in their responses and considering all possible relationships. To master this skill likely takes more discussion, examples, and work than can be accomplished in only 5 weeks.

The students' ability to determine if casual claims could be made with data did not develop much over the course of the unit. Developing this understanding needs extensive time and discussion that could not be given in this unit. Ultimately, multivariate thinking coupled with the skills of creating multivariate graphs and understanding the nuances of making causal claims are all complex skills that take time and much scaffolding to develop, especially when combined.

5.2 Research Question 2

The second set of research questions are discussed in this section, highlighting relevant qualitative results from the in-class observations, all students' assignments throughout the unit, and the final cognitive interviews. The questions addressed are: *What challenges surrounding multivariate thinking still persist after taking part in the intervention? Do any new challenges emerge after the completion of these activities?*

One prominent challenge remaining in students' multivariate thinking was reasoning about the relationships among three or more variables. Table 5 showed an initial increase in the percentage correct from 50% to 73% from Assignment 1 to Assignment 2, then a decrease on Assignment 3 to 21.2% correct. On the last assignment students struggled to describe relationships among all the variables in the graph and to support their answers with evidence from the plots they created.

The cognitive interviews with students gave further insight into the students' challenges with multivariate visualizations, beyond determining and describing relationships among the variables. In the interview with Kennedy, they gave a reasonable answer about which variables might affect the price of the car based on their graph, but then had difficulty reading the graph to determine the features of a car that would fit the $40,000 price point. In another student's interview, they confused the x- and y-axis labels either by accident or demonstrating a misunderstanding about which variable was plotted to which axis. These two instances call into question the students' ability to read their graphs and answer basic questions about them. Most questions on the activities and assignments focused on determining the relationships among variables, but seldom asked the students to directly read the graphs to answer questions. Perhaps, this too should have been a SLO for the unit.

Similarly, Kennedy, when observed in class, though in general displaying skillful reasoning when describing graphs, had some issues overgeneralizing interpretations in the Activity 8: World Data. They correctly described relationships among all variables in a visualization depicting life expectancy, region, and fertility. However, in interpreting their scatterplot displaying income, CO2, population, and region they claimed that in Asia larger populations with higher income have higher CO2 emissions. Even though there was no clear connection between population and income, or CO2 emissions displayed in their graph, they noticed there were two Asian countries with large populations, and then generalized this direct relationship to all the variables. It is unknown whether this is due to contextual knowledge that led them to believe that all these variables must have a direct relationship or if it came from a misreading of the graph. But in either case, the conclusions were not aligned with the visualization.

Another challenge, though less prominent, was the creation of DAGs. Consistently, around a third of the class got the questions with DAGs incorrect or partially correct on the assignments. Students still had issues with using arrows instead of lines, putting arrows in directions that are not temporally possible for causality, or generally not understanding the variables or context. While some students did not include any arrows between their potentially causal variables, it is unclear whether they did not identify any potential relationships among those variables as evidenced by their graphs, or if they simply did not consider that there might be relationships between them. Since the students were not individually tracked across the study it is unknown whether it was the same set of students continually making these mistakes across the assignments or if some learned from their feedback and others made these same mistakes on later assignments.

In particular, Kennedy, despite working with a group and having discussions about creating their DAGs, still seemed unclear about drawing the DAGs, often drawing incorrect arrows between variables. These issues, though minor, could indicate a conceptual misunderstanding or a simple oversight on the part of the students. If it was an issue of oversight, it might have been mitigated with additional class discussion of the DAGs or more feedback on their assignments. But if it is a conceptual misunderstanding, this warrants further investigation in future studies.

A third challenge was the context, as expected from previous literature studying multivariate thinking. The contexts did not appear to preclude any students from completing the assignments or even thinking of possible variables that would affect the outcome variable (they all consistently got this SLO correct across all three assignments at $> 90\%$). Often context was used to help support the students' answers, however it did occasionally play a role in students' final reasoning about the relationships among the variables. For example, some students (n=6) used only the context to answer questions about relationships in their graphs, occasionally contradicting the information depicted in the plot. Other students (n=4) made generalizations in their conclusion, extrapolating beyond the variables depicted in the plot using outside knowledge. Again, these mistakes were not common and seemed unique within each assignment. However, it does indicate that the context plays a role in how they are thinking about these problems and drawing conclusions.

The final challenges in the students' reasoning pertained to the study design within the activities and assignments. Foremost, most students struggled to identify observational data. More students were getting this correct by the end of the unit (with a correct response rate of 66.7%), but it remains unclear whether they understood what made the data observational in comparison to experimental data. Additionally, most students did not correctly answer the question on Assignment 3 about making causal claims with the cars data (only 18.2% correct). As discussed in response to Research Question 1, these challenges remained for students and needed more class time, discussions, and coverage on in-class assignments and exams to help students understand these concepts.

The observation with Jordan indicated some difficulties reasoning about how to answer the research questions posed in the activities using the data. Specifically, they expressed concern over considering multiple variables when we only want to know about one variable or the relationship between two variables. A main goal of the multivariate thinking unit was to expose students to the importance of assessing multivariate relationships and encouraging awareness of confounding variables. This student may have gained some insight into the importance of investigating multiple variables to answer simple seeming research questions through discussions about this and seeing Simpson's Paradox in Activity 10: SAT. Other students in class and in the cognitive interviews never questioned the need to look at the relationships of multiple variables to answer the research question. However, given the fragmented nature of the hybrid course, it is unknown if they also struggled with this concept. Jordan's concerns indicate that the importance of looking at graphs with multiple variables was not as clear as it needed to be in this collection of activities. It is possible a class discussion or resequencing of the activities could have helped make these points more salient.

5.2.1 Summary of Research Question 2

As discussed in answering Research Question 1, though the students demonstrated some growth in development of their multivariate thinking skills over the course of the unit, there were continued challenges when reasoning with multiple variables, creating DAGs, identifying data as observational and knowing that we cannot make causal claims with observational data. Kennedy was still a bit apprehensive about needing to investigate multiple variables at once if we were only interested in one variable or the relationship between two variables. Though other students did not outright ask this question it is possible they too had questions about the purpose of reasoning with multiple variables. In addition, some students had a difficult time using the context in conjunction with evidence from their graph to support their answers without extending their response to contexts beyond what was in the graph.

5.3 Limitations

Several limitations to this study are discussed next. In general, there are limitations to the generalizations that can be made from this study, but more specifically there are limitations pertaining to the course structure, use of R to create plots, difficulties coding, and the data sets.

A significant limitation to this study was the ongoing prevalence of the COVID-19 pandemic in which this data was collected. The pandemic forced the hybrid online/in person and synchronous/asynchronous nature of the course. Throughout the duration of the study both in-person attendance and attendance over Zoom decreased. Though some students came to the classroom to discuss their answers with a group, ask questions to the instructor, and for general accountability, about half the class or more typically did not come into the classroom. Some students asked questions or worked on the class materials during class time via Zoom, but a small number of students did their work without interacting synchronously and instead may or may not have asked questions over email. This made any planned "in-class" discussion or review of important concepts difficult. With only a few people in the classroom and a few people online it was challenging to have a meaningful discussion about the activities.

Additionally, the students were allowed to work at their own pace on the materials. Some students fell behind throughout the semester, while others worked ahead in the materials since all were openly accessible throughout the semester via Canvas (the university learning management system). Thus, any information sent to the class through email or in an attempted class discussion about an activity or concept did not always reach the students at the time it was most needed. Students that were ahead of schedule had forgotten about the nuances of the activity during in-class reviews and those that were behind schedule did not yet understand the premise of the questions or activities. This made pivoting to discuss some of the challenges the students had with the material difficult, such as creation of DAGs or reasoning about the graphs.

In addition to the unique hybrid structure of the course, it is also a unique course in that it focuses solely on visualizations and is geared toward undergraduates in their first or second year of college, often looking to fulfill a mathematical thinking credit. The population that chose this course over a traditional math or statistics course may be systematically different than those that chose more traditional math or statistics courses, and thus generalizations of the results are limited.

The coding requirement for this course presents certain limitations to the content that could be added for this study. In teaching coding to novice coders there are a lot of basics that need to be covered to ensure they have the software properly downloaded, can upload data, and can debug their code. These are skills the students worked on in the first few weeks of the semester, but they continued to build on these skills and discovered new challenges (such as embedding an image of a DAG into their R Markdown files) throughout the unit. The coding did get easier for the students, they learned to debug or ask for help when needed, and they seemed to understand the general structure of the code for creating graphs, but there were still other technology challenges they ran into throughout the course of the semester. Issues uploading the pictures of their DAGs, problems with their downloaded data files converting to .pages files instead of the .csv files they needed, and many other individual computer system issues naturally took up some of the students' time in addition to answering questions in activities and assignments. Each time these problems occurred they required time to troubleshoot and took time away from students working on thinking about the graphs they had created. Also not unexpectedly, some of the students dealt with the anxiety of working with technology and coding for the first time (e.g., Chang (2005)). The time it took to create the plots and debug code could have been spent working on reasoning about the relationships among the variables in graphs or discussing when making causal claims is appropriate.

The contexts used in the materials for this unit may have been another limiting factor for developing multivariate thinking or interpreting results. Given GAISE guidelines, real data was sought to use for these activities and assignments, however this was a challenging task. The datasets needed to be engaging, culturally inclusive, relevant and relatable to a diverse group of students, and contain more than two variables. The datasets also needed to lend themselves to multivariate thinking in various scenarios For example they needed to display Simpson's Paradox or contain a surprising or unexpected relationship (perhaps when faceted) for the students to explore. Searching for 10 data sets that fit these requirements was a challenge and the resulting datasets do not adequately hit all the marks. Notably, the women in STEM and evaluation scores activities focused on gender as a binary construct which was, rightfully, noted by Jordan as a limitation in the dataset. It is unknown in what ways the selected contexts may have limited the students' ability to process the information in the graphs beyond what has already been described, but it assumably played a role.

5.4 Implications for Teaching

When teaching multivariate thinking, it is important to consider the complexity of this topic. It is difficult to develop, even when given five weeks of focus. Though the GAISE guidelines emphasize that introductory statistics courses should include multivariate thinking, not all courses do at this time and additionally those that include it are likely teaching it in the form of multiple linear regression. However, it is multivariate visualizations that are ubiquitous in the media. Teaching about multivariate reasoning using visualizations provides an accessible entry point for students before diving into multiple linear regression. This section discusses some implications for teaching multivariate thinking.

Using DAGs to introduce and facilitate multivariate thinking provided a framework for discussing relationships among variables before creating the graphs. Using the DAGs allowed the students to make predictions and have conversations about relationships to pique their interest. However, given that some students did not put much thought into them or learn to use the arrows to imply causation there were difficulties conveying their intent. It is possible that with further instruction on creating and updating DAGs based on evidence from their graphs, they could have used them more effectively to help develop multivariate thinking. Additionally, if provided the opportunity to present their DAGs and graph to the class more often, the students would have had more experience using them for communication and explanation to answer a research question. This task could have made it more clear to them how the DAGs could be beneficial beyond stating their predictions before creating graphs.

The course this unit was created for focused on **ggplot2** to create static multivariate visualizations. In contrast, recommendations from the literature suggested employing an easy-to-use, point and click software to create interactive visualizations as a better place to start. However, using coding gave students experience with this skill that they may not have had otherwise, and their coding skills did improve over the course of the unit. Students faced general challenges with the technology itself, but they did end up able to code the graphs they needed and seemed to enjoy the customization of working with the `ggplot()` function to create the graphs. Given additional scaffolding and better attendance in the course (to troubleshoot issues earlier) the coding could have been less a burden on the students' time.

Notably, the students observed in class both improved in their ability to code the graphs, which could be due to their working with partners. Jordan worked with the researcher, who helped them more quickly debug in the code, which moved the activity along many times. By contrast the student that worked in a group often collaborated on writing the code and helped others debug their work. If instructors provide scaffolding for learning coding and allow students to work in groups to code, it could be a manageable way to create multivariate visualizations and explore multivariate thinking. However, this also requires the flexibility to potentially not include as much other content to accommodate for these scaffolds and discussions.

With proper scaffolding and use of groups helping the students code, ideally, they would be able to spend more time reasoning or interpreting the graphs they created. However, this must also be facilitated with detailed questions about the relationships among the variables and class time for small and large group discussions. Often when observing the students in class, if they were working on a section by themselves, they rushed through the interpretations of their graphs. It was only if the researcher slowed them down or if they answered extensive questions within the activity about the graphs they created that they focused on individual relationships (e.g., Activity 7/8: World Activity).

Finding relevant, engaging, ready to use, multivariate datasets is a challenge for teaching this material. The topics discussed in this study did occasionally excite or spur some meaningful discussions among the students about the context, measurement aspects of the variables, and the study design. Giving students the chance to explore multivariate thinking activities is ripe with the opportunity to get students thinking about the source of the data and their relationship to the datasets. (e.g., activities and framework by Lee, Wilkerson, Stokes, & McBride, 2022). However, having adequate time in class to discuss all the students' concerns and to reason through the graphs together would help meaningfully facilitate these discussions.

5.5 Implications for Research

There are many avenues for future research involving multivariate thinking in statistics, particularly when reasoning with multivariate visualizations. This study provides an in-depth summary and analysis of students' thinking as they begin to develop multivariate thinking skills and communicate about multivariate visualizations, which has, to date, not been explored to this extent in the statistics education research literature. Additionally, this study provides a first exploration of SLOs that might be needed for developing multivariate thinking. There are opportunities for further inquiry into teaching and learning about multivariate thinking.

This study provided evidence of students' reasoning as they worked with many variables through exploring traditional visualizations (mainly scatterplots and stacked bar charts). However, it only scratched the surface of reasoning with these graphs and did not extend into many other traditional, more modern, or novel visualizations with three or more variables (e.g., network graphs).

In considering how to incorporate DAGs into a course, one should consider also teaching about mediator and moderator variables and identifying the DAGs that accompany them. These were not covered in this unit, but their inclusion might have made the DAGs more meaningful and given the students more insight into the ways in which variables can interact and confound each other. However, it takes a specific dataset and knowledge of the context surrounding that dataset to truly help students understand these concepts and their implications. Finding such data sets might prove challenging but would be invaluable for helping students wrestle with these concepts. Simulated data or even shiny apps that could somehow introduce these concepts with interactivity could provide the additional support needed to help introduce these concepts.

Additionally, though students were asked to update their DAGs after creating a visualization, this study did not investigate the details of when they chose to update their DAGs. This is of interest for further research. For example, it is unknown what level of evidence was needed for the students to decide to update their DAG from their prediction after seeing their graph. Based on Kennedy's reasoning in class for Activity 9: Evaluation in which they said they "updated the DAGs based on feeling", it might be of note how much the graphs play a role in updating the DAGs, if at all. Similarly, results indicated occasions where students did not update their DAG with new information based on their plot or included arrows for relationships they did not think were strong. So, future research could investigate how much evidence a student needs to see in the graph to convince them to update their predicted graph.

For this study line plots were taught before scatterplots as suggested in the literature. This ensured time was the third variable to help prime students to think with multiple variables. Students observed in class, Kennedy and their group in particular, had difficulty reasoning with line plots. Perhaps the graphs themselves needed more description and introduction, but they initially caused as much, if not more confusion than the scatterplots. It is possible that they were confused about the context instead of the graph itself, or that it seemed much more complex than what they had initially seen in bar graphs which caused them difficulties. Further research on sequencing of graphs to introduce multivariate reasoning is needed to confirm that time as the third variable is the best option – especially while using **ggplot2** for creating graphs.

Another potential resequencing that could help students is to introduce Simpson's Paradox earlier in the unit. Simpson's Paradox provides a good motivation for reasoning with more than just the variables of explicit interest. Perhaps introducing this idea earlier would have mitigated some of Jordan's confusion about why we need to look at multiple variables to answer a research question. Discussing mediating and moderating variables also would provide more motivation for the need to reason with multiple variables.

Though multivariate thinking for this unit was broken into the SLOs suggested by GAISE with DAGs added, these SLOs do not necessarily make up a trajectory of developing multivariate learning. Though they serve an initial blueprint, they were not studied as a formal learning trajectory, thus leaving this for future work. In addition, it is worth critiquing and evaluating these SLOs as a whole. There may be more learning milestones in multivariate thinking that are not currently a part of the outcomes, or there may be redundancies in the outcomes. Additional work to define the needed skills of multivariate thinking and potential sequencing of those skills is an area of future work that can be developed from this research. Furthermore, once the learning goals are more well defined, there will need to be an assessment of the learning goals.

References

(2020, March). QSR International Pty Ltd. Retrieved from https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home

Abdelhadi, R. M. D. (2016). *What matters? Assessing and developing inquiry and multivariable reasoning skills in high school chemistry* (PhD thesis). Columbia Univeristy.

Adams, B., Baller, D., Jonas, B., Joseph, A.-C., & Cummiskey, K. (2021). Computational skills for multivariable thinking in introductory statistics. *Journal of Statistics and Data Science Education*, *29*, S123–S131. http://doi.org/10.1080/10691898.2020.1852139

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352. http://doi.org/10.1016/0010-0277(94)00640-7

Batanero, C., Estepa, A., & Godino, J. D. (1996). Evolution of students' understanding of statistical association in a computer-based teaching environment. In J. Garfield & G. Burhill (Eds.), (pp. 191–205). Presented at the Research on the role of technology in teaching and learning statistics: Proceedings of the 1996 IASE round table conference, Voorburg,The Netherlands: International Statistical Institute.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathemateics Education*, *27*(2), 151–169. Retrieved from https://www.jstor.org/

stable/749598

Brown, A. L., & Campione, J. C. (1994). *Guided discovery in a community of learners.* The MIT Press.

Caspari, I., & Graulich, N. (2019). Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning. *International Journal of Physics and Chemistry Education*, *11*(2), 31–43. http://doi.org/10.1039/c8rp00131f

Chang, S. E. (2005). Computer anxiety and perception of task complexity in learning programming-related skills. *Computers in Human Behavior*, *21*(5), 713–728. http://doi.org/https://doi.org/10.1016/j.chb.2004.02.021

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120. http://doi.org/10.1111/1467-8624.00081

Cobb, P., McClain, K., & Gravemeijer, K. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*(1), 1–78. http://doi.org/10.1207/S1532690XCI2101_1

Committee, G. C. R. A. R. (2016). *Guidelines for assessment and instruction in statistics education college report 2016.* Retrieved from http://www.amstat.org/education/gaise

Committee on How People Learn II: The Science and Practice of Learning, Board on Behavioral, Cognitive, and Sensory Sciences, Board on Science Education, Division of Behavioral and Social Sciences and Education, & National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures.* Washington, D.C.: National Academies Press. http://doi.org/10.17226/24783

Cummiskey, K., Adams, B., Pleuss, J., Turner, D., Clark, N., & Watts, K. (2020). Causal inference in introductory statistics courses. *Journal of Statistics Education*, *28*(1), 2–8. http://doi.org/10.1080/10691898.2020.1713936

delMas, R. (2004). A comparison of mathematical and statistical reasoning. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 77–95). Dordrecht: Springer Netherlands. http://doi.org/10.1007/1-4020-2278-6

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research*. Dordrecht: Springer Netherlands. http://doi.org/10.1007/978-94-007-6094-3

Fry, E. B. (2017). *Introductory statistics students' conceptual understanding of study design and conclusions* (PhD thesis). University of Minnesota.

Fry, E. B. (2017). *Introductory statistics students' conceptual understnadning of study design and conclusions* (PhD thesis). University of Minnesota, Minneapolis, MN. Retrieved from https://iase-web.org/documents/dissertations/17.ElizabethBrondosFry.Dissertation.pdf

Gapminder. (n.d.). Retrieved from https://www.gapminder.org/tools/#$chart-type=bubbles&url=v1

Gil, E., & Gibbs, A. L. (2017). Promoting modeling and covariational reasoning among secondary school students in the context of big data. *Statistics Education Research Journal*, *16*(2), 163–190. Retrieved from http://iase-web.org/Publications.php?p=SERJn

Gopnik, A., & Schulz, L. (Eds.). (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford ; New York: Oxford University Press.

Greenland, S., Pearl, J., & Robins, J. M. (1999). Causal diagrams for epidemiologic research: *Epidemiology*, *10*(1), 37–48. http://doi.org/10.1097/00001648-199901000-00008

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hernán, M. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*(2), 176–184. http://doi.org/10.1093/aje/155.2.176

Hernán, Miguel A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, *32*(1), 42–49. http://doi.org/https://doi.org/10.1080/09332480.2019.1579578

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 945–960. Retrieved from https://www.jstor.org/stable/2289064

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician*, *69*(2), 138–145. http://doi.org/10.1080/00031305.2015.1032435

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *arXiv:1907.07271 [Stat]*. Retrieved from http://arxiv.org/abs/1907.07271

Kahneman, D. (2013). *Thinking, fast and slow* (1st pbk. ed). New York: Farrar, Straus; Giroux.

Kaplan, D. T. (2017). *Statistical modeling: A fresh approach* (2nd ed.). Retrieved from https://dtkaplan.github.io/SM2-bookdown/preface-to-this-electronic-version.html

Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, *91*(5), 710–726. http://doi.org/10.1002/sce.20214

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. San Diego: Academic Press.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development*, *23*(4), 435–451. http://doi.org/10.1016/j.cogdev.2008.09.006

Kuhn, D., Ramsey, S., & Arvidsson, T. S. (2015). Developing multivariable thinkers. *Cognitive Development*, *35*, 92–110. http://doi.org/10.1016/j.cogdev.2014.11.003

Kuiper, S. (2008). Introduction to multiple regression: How much is your car worth? *Journal of Statistics Education*, *16*(3), 10. http://doi.org/10.1080/10691898.2008.11889579

Lee, H., Wilkerson, M. H., Stokes, D., & McBride, C. (2022, April 11). *Telling stories with data: Strategies and tools for building data fluency*. Retrieved from https://www.fi.ncsu.edu/event/telling-stories-with-data-strategies-and-tools-for-building-data-fluency/

Lorch, Robert F., Lorch, E. P., Freer, B., Calderhead, W. J., Dunlap, E., Reeder, E. C., … Chen, H.-T. (2017). Very long-term retention of the control of variables strategy following a brief intervention. *Contemporary Educational Psychology*, *51*, 391–403. http://doi.org/10.1016/j.cedpsych.2017.09.005

Lorch, Robert F., Lorch, E. P., Wheeler, S. L., Freer, B. D., Dunlap, E., Reeder, E. C., … Chen, H.-T. (2019). Oversimplifying teaching of the control of variables strategy. *Psicología Educativa*, *26*(1), 7–16. http://doi.org/10.5093/psed2019a13

Lübke, K., Gehrke, M., Horst, J., & Szepannek, G. (2020). Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, *28*(2), 133–139. http://doi.org/10.1080/10691898.2020.1752859

Mason, R. L., & Young, J. C. (2004). Multivariate thinking. *Quality Progress*, *37*(4), 89–91. Retrieved from http://207.67.83.164/data/subscriptions/qp/2004/0404/qp0404statistics.html

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). *Qualitative data analysis: A methods sourcebook* (Fourth edition). Los Angeles: SAGE.

Morgan, S. L., & Winship, C. (n.d.). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge University Press.

Moritz, J. (2004). Reasoning about covariation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–255). Dordrecht: Springer Netherlands. http://doi.org/10.1007/1-4020-2278-6_10

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 160940691989922. http://doi.org/10.1177/1609406919899220

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688. Retrieved from https://www.jstor.org/stable/2337329

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, U.K. ; New York: Cambridge University Press.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. http://doi.org/10.1214/09-SS057

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. New Jersey: Wiley.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (First edition). New York: Basic Books.

Peters, J., Janzing, D., & Scholkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. Massachusetts Institute of Technology: MIT Press. Retrieved from https://library.oapen.org/bitstream/handle/20.500.12657/26040/11283.pdf?sequ

Powell, S. (2018). The book of why: The new science of cause and effect. *Journal of Multidisciplinary Evaluation*, *14*(31), 47–54. Retrieved from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/507

Prodromou, T. (2014). Drawing inference from data visualisations. *International Journal of Secondary Education*, *2*(4), 66. http://doi.org/10.11648/j.ijsedu.20140204.12

Pruim, R., Kaplan, D. T., & Horton, N. J. (2017). The mosaic package: Helping students to 'think with data' using r. *The R Journal*, *9*(1), 77–102. Retrieved from https://journal.r-project.org/archive/2017/RJ-2017-024/index.html.

Ridgway, J., Nicholson, J., & McCusker, S. (2007). Reasoning with multivariate evidence. *International Electronic Journal of Mathematics Education*, *2*(3), 26.

Ridgway, J., Nicholson, J., & McCusker, S. (2009). The next great leap - from official data to public knowledge (p. 8). Presented at the IASE/ISDatellite. Retrieved from https://iase-web.org/documents/papers/sat2009/4_2.pdf?1402524995

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. http://doi.org/10.1037/h003735

Ryff, C., Almeida, D., Ayanian, J., Binkley, N., Carr, D. S., Coe, C., … Williams, D. (2019). Midlife in the united states (MIDUS 3), 2013-2014. Inter-university Consortium for Political; Social Research [distributor]. http://doi.org/10.3886/ICPSR36346.v7

Sabbag, A. G., & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal*, *14*(2), 93–116.

Saldaña, J. (2016). *The coding manual for qualitative researchers* (3E [Third edition]). Los Angeles ; London: SAGE.

Schield, M. (2004). Statistical literacy curriculum design (pp. 54–74). Presented at the Curricular development in statistics education, Sweden. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.144.8102&rep=rep1&type=pdf

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. http://doi.org/10.1016/j.dr.2015.12.001

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*(1), 70. http://doi.org/10.1186/1471-2288-8-70

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488–511. http://doi.org/10.1016/j.cogdev.2008.09.005

Sutherland, S., & Ridgway, J. (2017). Interactive visualizations and statistical literacy. *Statistics Education Research Journal*, *16*(1), 26–33. Retrieved from http://iase-web.org/Publications.php?p=SERJ

Suzuki, E., Shinozaki, T., & Yamamoto, E. (2020). Causal diagrams: Pitfalls and tips. *Journal of Epidemiology*, *30*(4), 153–162. http://doi.org/10.2188/jea.JE20190192

Tu, Y.-K., & Gilthorpe, M. S. (2012). *Statistical thinking in epidemiology*. Boca Raton, FL: CRC Press.

Valero-Mora, P. M., & Ledesma, R. D. (2011). Using interactive graphics to teach multivariate data analysis to psychology students. *Journal of Statistics Education*, *19*(1), 5. http://doi.org/10.1080/10691898.2011.11889600

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

Wang, X., Rush, C., & Horton, N. J. (2017). Data visualization on day one: Bringing big ideas into intro stats early and often. *Technology Innovations in Statistics Eduvation*, *10*(1). Retrieved from https://doi.org/10.5070/T5101031737

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed. 2016). Cham: Springer International Publishing : Imprint: Springer. http://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019a). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. http://doi.org/10.21105/joss.01686

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019b). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. http://doi.org/10.21105/joss.01686

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B., & Gagliardi, L. (2018). Directed acyclic graphs: A tool for causal studies in paediatrics. *Pediatric Research*, *84*(4), 487–493. http://doi.org/10.1038/s41390-018-0071-3

Wood, K. E. (2015). *Evolution of scientific reasoning in control of variables for undergraduate physics lab* (PhD thesis). University of Cincinnati.

Xie, Y. (2021). *formatR: Format R code automatically*. Retrieved from https://CRAN.R-project.org/package=formatR

Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal*, *8*(1), 7–31. Retrieved from https://www.stat.auckland.ac.nz/~iase/serj/SERJ8(1)_Zieffler_Garfield.pdf

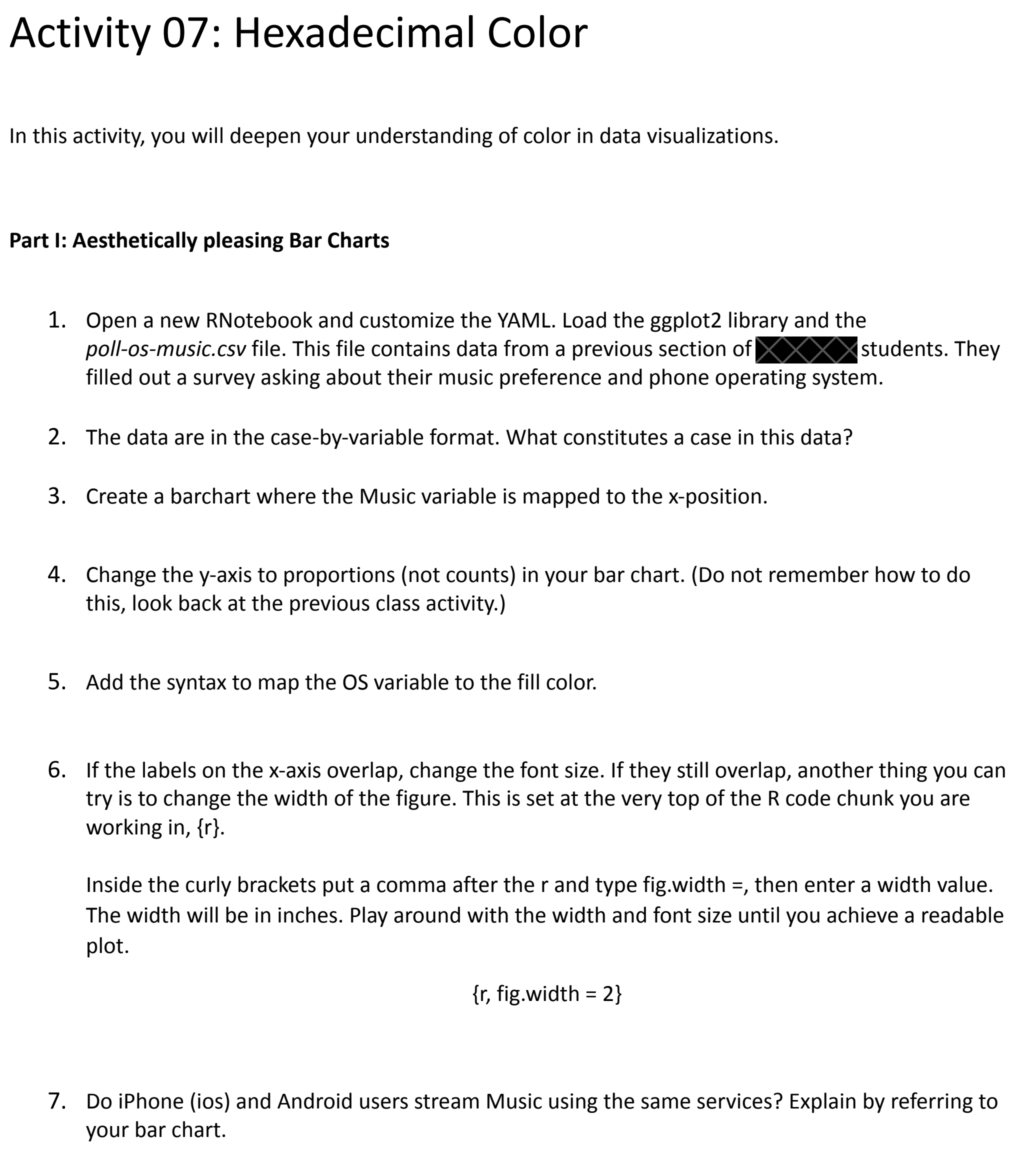
Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149. http://doi.org/10.1006/drev.1999.0497

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. http://doi.org/10.1016/j.dr.2006.12.001

Appendix A

Course Materials

A.1 Activity 1



Activity 07: Hexadecimal Color

In this activity, you will deepen your understanding of color in data visualizations.

Part I: Aesthetically pleasing Bar Charts

1. Open a new RNotebook and customize the YAML. Load the ggplot2 library and the *poll-os-music.csv* file. This file contains data from a previous section of [redacted] students. They filled out a survey asking about their music preference and phone operating system.

2. The data are in the case-by-variable format. What constitutes a case in this data?

3. Create a barchart where the Music variable is mapped to the x-position.

4. Change the y-axis to proportions (not counts) in your bar chart. (Do not remember how to do this, look back at the previous class activity.)

5. Add the syntax to map the OS variable to the fill color.

6. If the labels on the x-axis overlap, change the font size. If they still overlap, another thing you can try is to change the width of the figure. This is set at the very top of the R code chunk you are working in, {r}.

   Inside the curly brackets put a comma after the r and type fig.width =, then enter a width value. The width will be in inches. Play around with the width and font size until you achieve a readable plot.

   {r, fig.width = 2}

7. Do iPhone (ios) and Android users stream Music using the same services? Explain by referring to your bar chart.

Adding Labels

8. Change the x-axis label to something more informative.

9. Using the same idea as changing the label on the x-axis, we will change the label on the y-axis to Proportion.

```
labs(x = “your informative label”,
     y = “Proportion”)
```

Color

To change the fill color associated with the different OS types, we will add a scale_fill_manual() layer to our syntax.

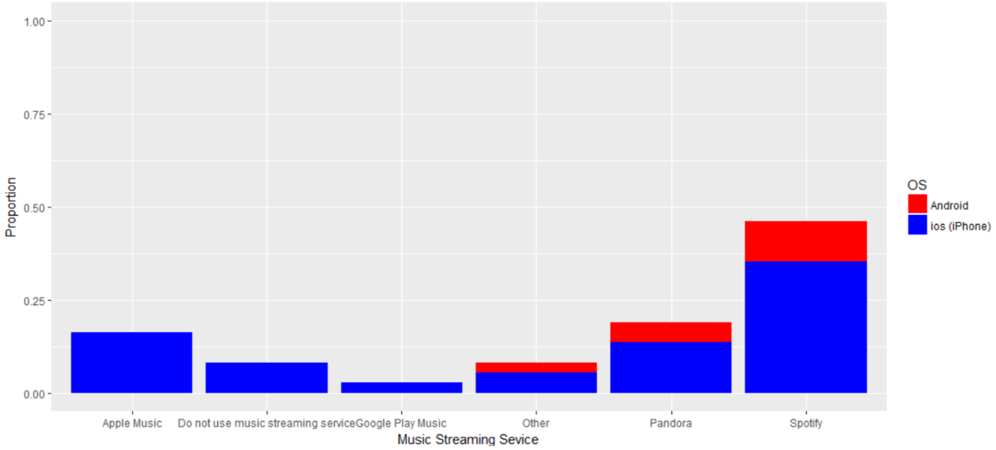
- Remember we are literally adding layers using the + sign.
- The syntax we add will look like this:

```
scale_fill_manual()
```

- Inside the parentheses there are two things we can change.
  - The first thing we can do is change the color values associated with the fill scale. To do this we include values = c( )
  - Inside the c( ) function we include color names inside of quotation marks and separated by commas.
  - We will make the Android “red” and ios (iPhone) “blue”.

```
scale_fill_manual(values = c ("red", "blue"))
```

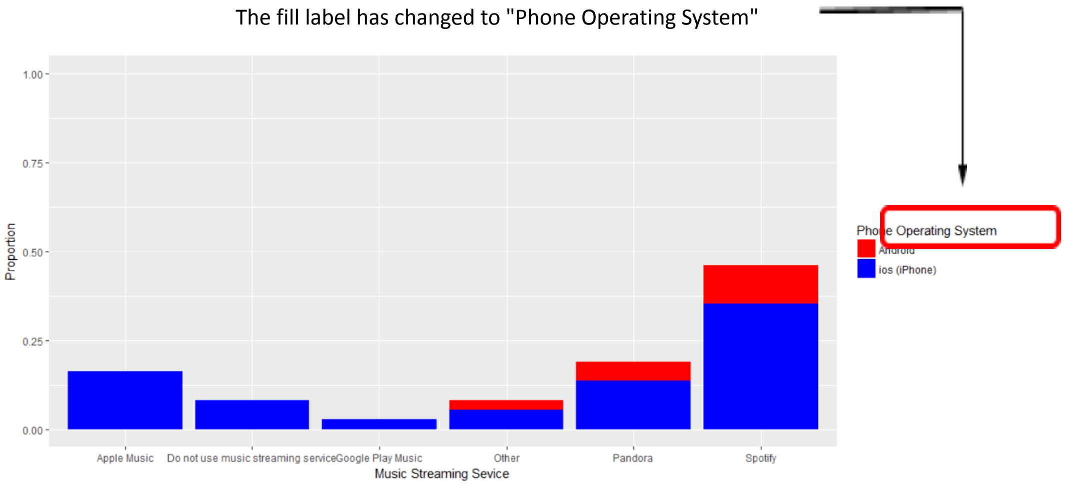
Note: The fill colors will be mapped to the categories of the fill variable alphabetically. Since we gave the color “red” first, it is mapped to “Android”. We gave the color “blue” second, so it was mapped to “ios (iPhone)”.



Customizing the title of the legend

Inside the scale_fill_manual() parentheses, we can also change the label associated with the fill scale. To do this we include name = "…", where the … is the label you want.

10. Change the OS label on the color legend to Phone Operating System.



The fill label has changed to "Phone Operating System"

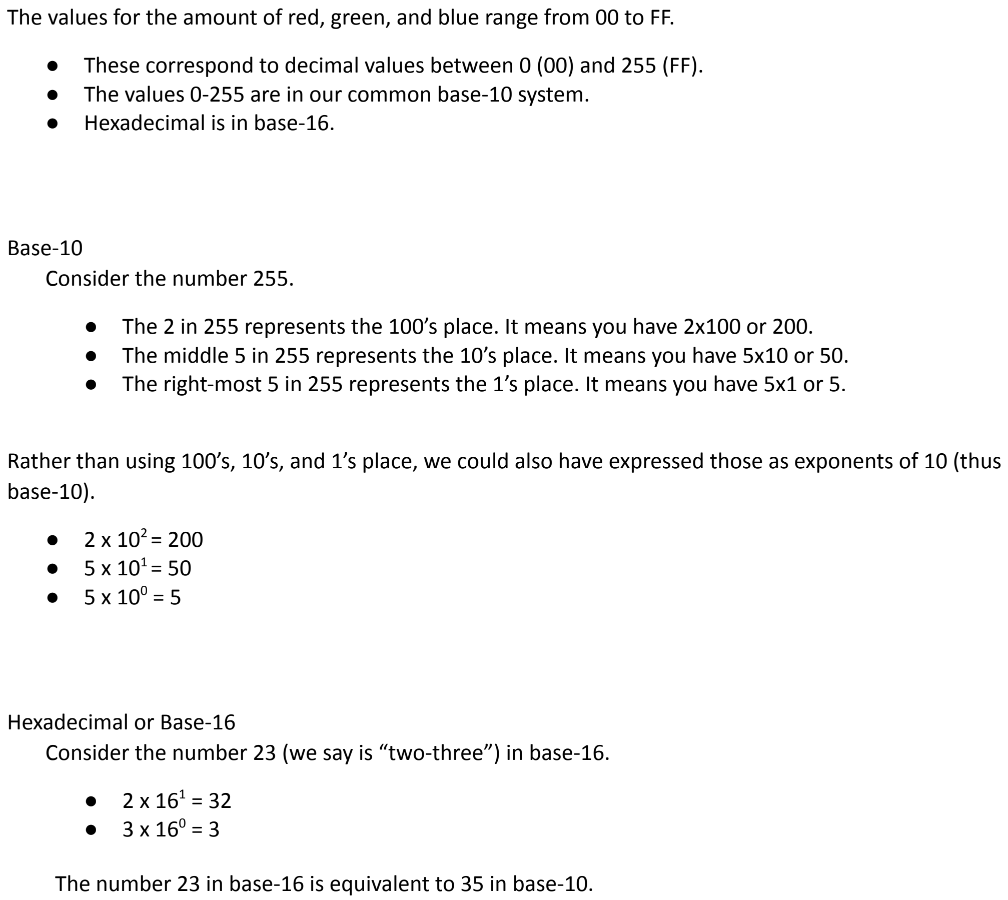
11. Change the “red” and “blue” fill colors in your bar chart to two different fill colors. Remember: to get the color names you can run colors() in an R code chunk. There are 657 color names you can choose from.

Part II Hexadecimal

A hex triplet is a six-digit, three-byte hexadecimal number used in HTML, CSS, SVG and other computing applications to represent colors. The bytes represent the red, green and blue components of the color. One byte represents a number in the range 00 to FF (in hexadecimal notation). – Wikipedia; Web colors

The color red is denoted by FF0000.

- The first two characters, FF, represent the amount of red.
- The second two characters, 00, represent the amount of green.
- The third two characters, 00, represent the amount of blue.



The values for the amount of red, green, and blue range from 00 to FF.

- These correspond to decimal values between 0 (00) and 255 (FF).
- The values 0-255 are in our common base-10 system.
- Hexadecimal is in base-16.

Base-10

Consider the number 255.

- The 2 in 255 represents the 100's place. It means you have 2x100 or 200.
- The middle 5 in 255 represents the 10's place. It means you have 5x10 or 50.
- The right-most 5 in 255 represents the 1's place. It means you have 5x1 or 5.

Rather than using 100's, 10's, and 1's place, we could also have expressed those as exponents of 10 (thus base-10).

- $2 \times 10^2 = 200$
- $5 \times 10^1 = 50$
- $5 \times 10^0 = 5$

Hexadecimal or Base-16

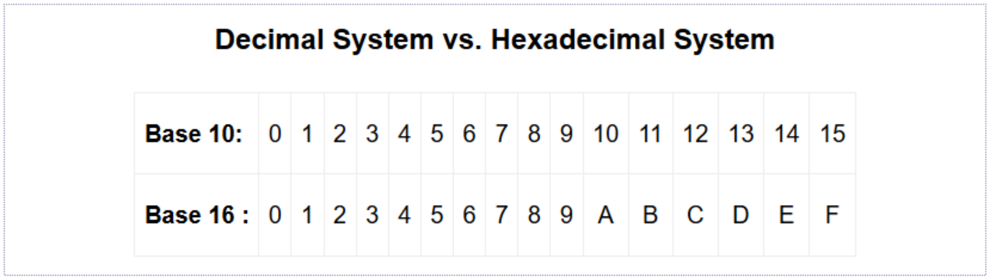
Consider the number 23 (we say is "two-three") in base-16.

- $2 \times 16^1 = 32$
- $3 \times 16^0 = 3$

The number 23 in base-16 is equivalent to 35 in base-10.

Why does base-16 have letters?
In base-10 we need unique digits to cover all of the values between 0 and 9. This is because once we hit the value of 10 we express it by putting a digit in the 10^1 place.

In base-16 we need a single character to cover each value between 0 and 15.



Decimal System vs. Hexadecimal System

Base 10:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Base 16 :	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F

How to convert from base-10 to base-16
To show the conversion from base-10 to hexadecimal we will convert the RGB decimal code for orchid, rgb(218, 112, 214), to base-16.

Starting with the first number, 218, figure out how many time 16 divides into 218.

218 / 16 = 13 with remainder of 10.

The first place in the hexadecimal color code is how many time 16 divides into the base-10 number and the remainder will be the second place.

13 = D and 10 = A so 218 is the same as DA in base-16.

We do the same process for 112. One-hundred twelve divided by 16 is 7, no remainder. So 112 in hexadecimal is 70. Remember each hexadecimal value will have two places ranging from 00 to FF.

For our last conversion, two hundred fourteen divided by 16 is 13 with a remainder of 6. So the hexadecimal number is D6.

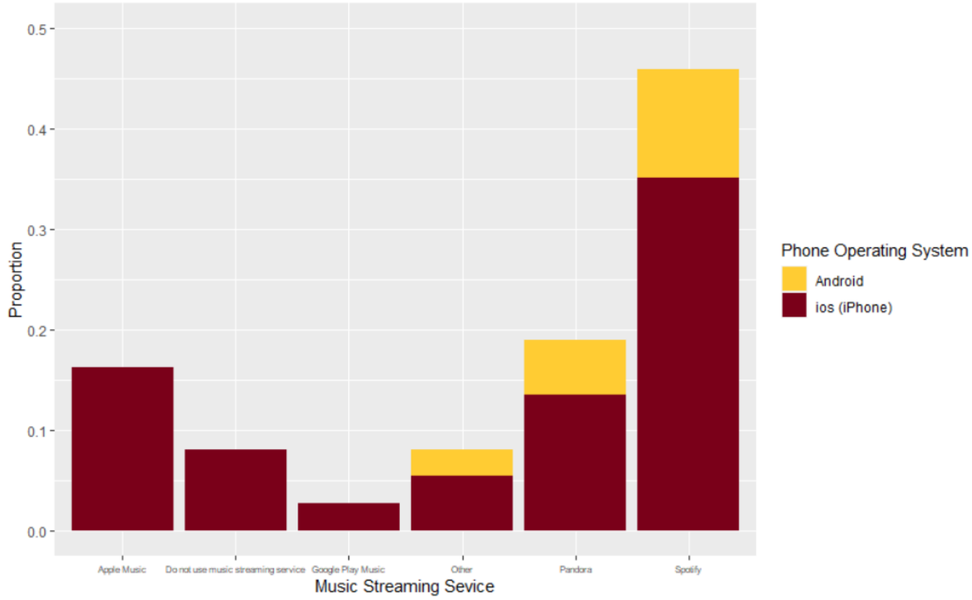
The hex code for orchid is #DA70D6

Your Turn

Color your bar chart with the University of Minnesota's official colors, gold color rgb (255, 204, 51) and maroon rgb (122, 0, 25).

12. Convert the rgb color code to hex code.

13. In your code, replace the color word with the hex codes contained in quotes and starting with a # symbol.



Part III: Putting It All Together

14. Open the *midus.csv* data file used in the last activity.

15. Review the midus codebook and choose one categorical and continuous variable that interests you. What variables did you pick?

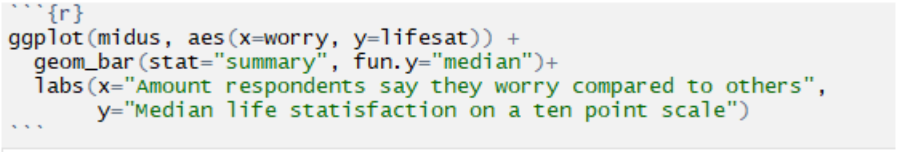
16. Load the *dplyr* package and use it to filter out the na responses. (If you do not remember how to do this, refer back to last class's activity.)

17. Make a bar chart of the categorical variable. Summarize what you see in your bar chart in context.

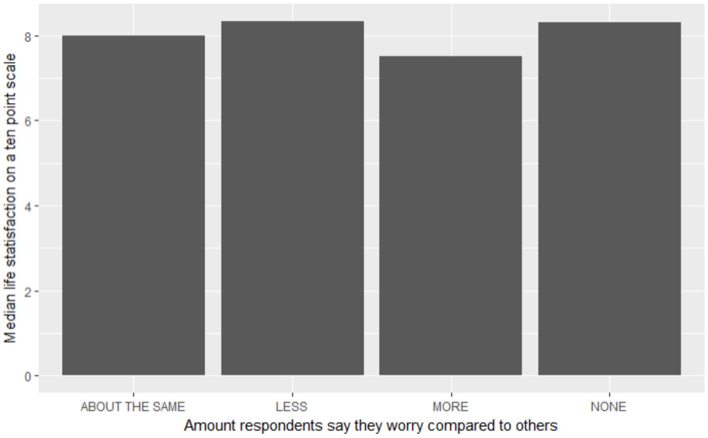
Next, we are going to create a bar chart with our categorical variable on the x-axis and the median of the continuous variable on the y-axis.

In the aes of the ggplot layer add y ='your continuous variable' after your x variable.

Now we need to tell R to find and graph the median of your continuous variable. Below is the code for my two variables. See if you can graph your chosen variables by modifying the code below.

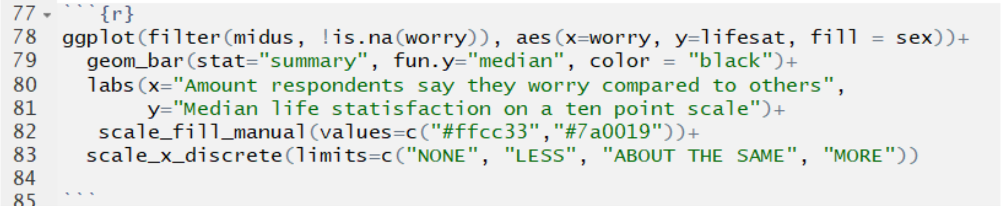


~~~
```{r}
ggplot(midus, aes(x=worry, y=lifesat)) +
  geom_bar(stat="summary", fun.y="median")+
  labs(x="Amount respondents say they worry compared to others",
       y="Median life statisfaction on a ten point scale")
```
~~~



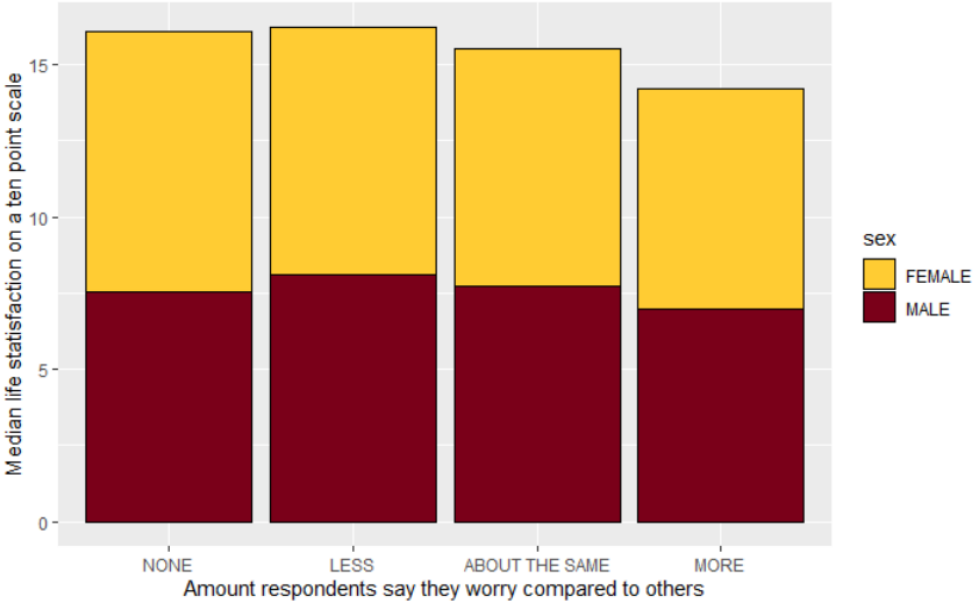
18. Paste your plot here.

Now let's fill on sex to see if there is a difference for men and women.

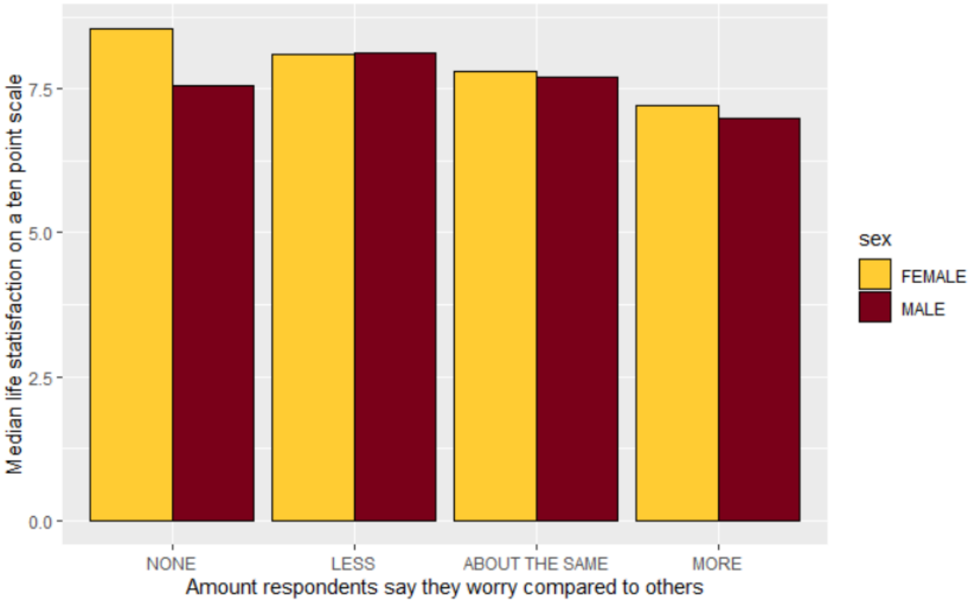


```{r}
ggplot(filter(midus, !is.na(worry)), aes(x=worry, y=lifesat, fill = sex))+
  geom_bar(stat="summary", fun.y="median", color = "black")+
  labs(x="Amount respondents say they worry compared to others",
       y="Median life statisfaction on a ten point scale")+
   scale_fill_manual(values=c("#ffcc33","#7a0019"))+
  scale_x_discrete(limits=c("NONE", "LESS", "ABOUT THE SAME", "MORE"))

```



This plot may be easier to read if we make a side-by-side bar chart. Add in the position = “dodge” to your code.

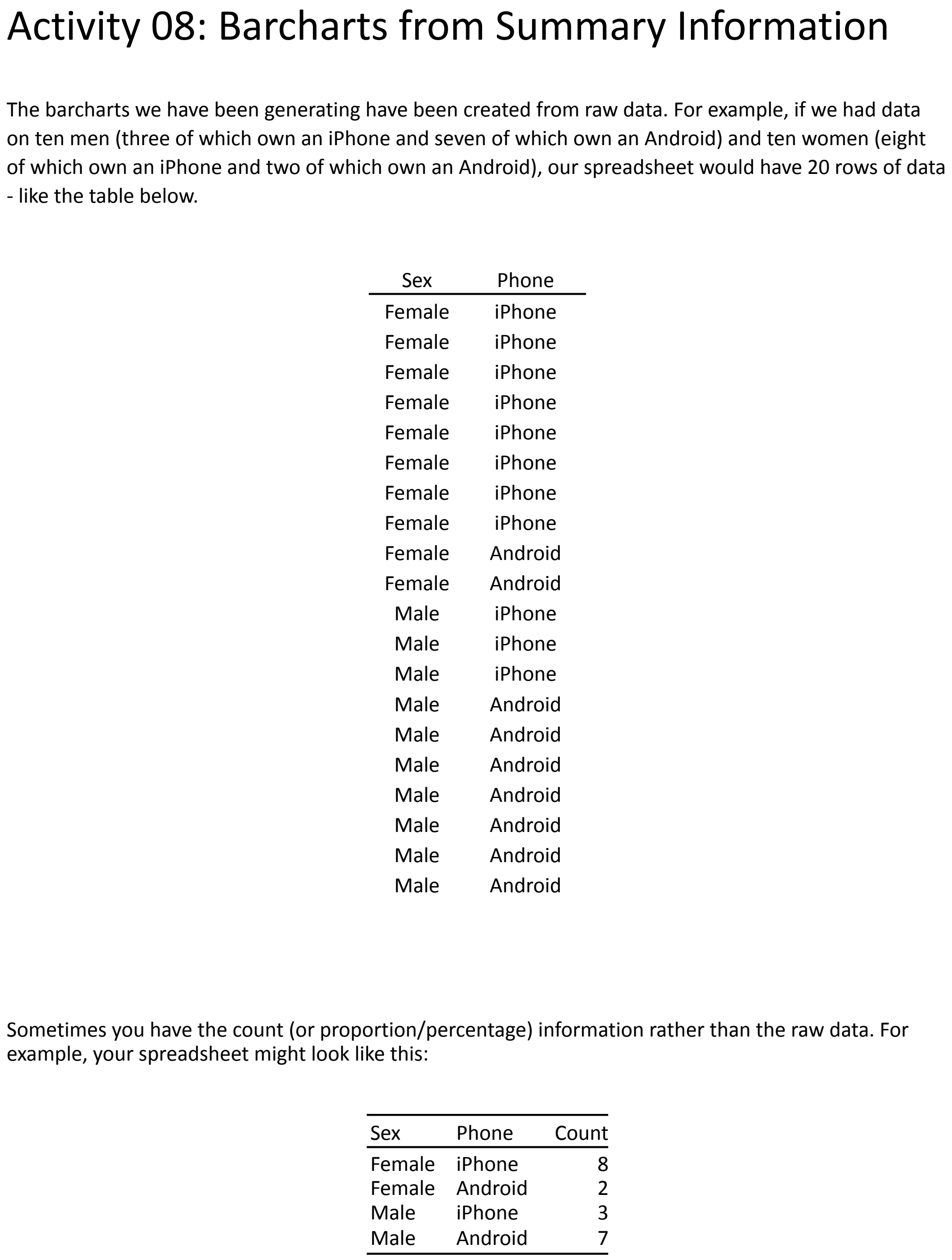


Both men and women report high median life satisfaction and these medians vary little by the amount they say they worry compared to others. The largest difference we see is in the group that professes to not worry at all. In this group women report a higher median life satisfaction than men.

19. Paste your final side-by-side bar chart below.

20. Give a description of your graph. Make sure to mention which variables you chose to look at and if you think there is a difference between how medians for males and females compare on your chosen variables.

A.2 Activity 2



Activity 08: Barcharts from Summary Information

The barcharts we have been generating have been created from raw data. For example, if we had data on ten men (three of which own an iPhone and seven of which own an Android) and ten women (eight of which own an iPhone and two of which own an Android), our spreadsheet would have 20 rows of data - like the table below.

Sex	Phone
Female	iPhone
Female	iPhone
Female	iPhone
Female	iPhone
Female	iPhone
Female	iPhone
Female	iPhone
Female	iPhone
Female	Android
Female	Android
Male	iPhone
Male	iPhone
Male	iPhone
Male	Android
Male	Android
Male	Android
Male	Android
Male	Android
Male	Android
Male	Android

Sometimes you have the count (or proportion/percentage) information rather than the raw data. For example, your spreadsheet might look like this:

Sex	Phone	Count
Female	iPhone	8
Female	Android	2
Male	iPhone	3
Male	Android	7

When you have summary information we can still use ggplot() to create a bar chart, but we need to add two things to our syntax.

- First, we need to map the *Count* variable to the *y*-position.
- Second, we include stat = "identity" in the geom_bar() layer.

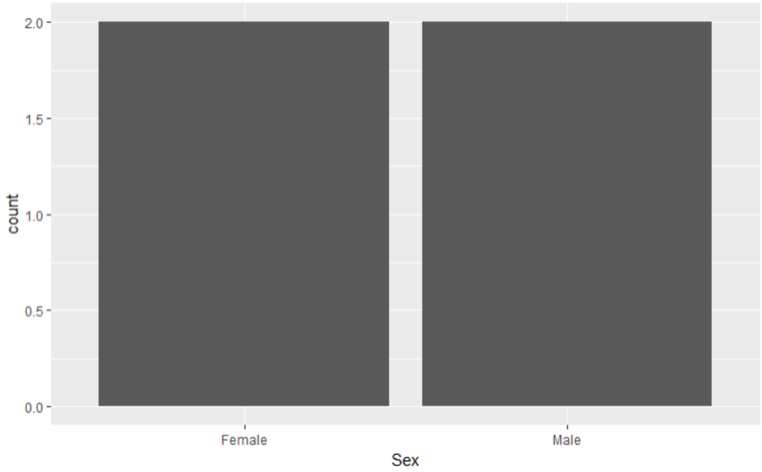
***Note**: Always make sure to check the layout (case by variable or summary) of the data before starting your coding*.

Part I: Bar chart from summary information

1. Open a new RNotebook, customize the YAML, load the ggplot2 library and the *phone-by-sex.csv* file.

2. Run head() to see the data.

3. Create a bar chart of the Sex variable.



4. Something is wrong here, why do the bars only go up to two? The data set contains responses from 10 men and 10 women. Look back at your data and see if you can figure out what happened.

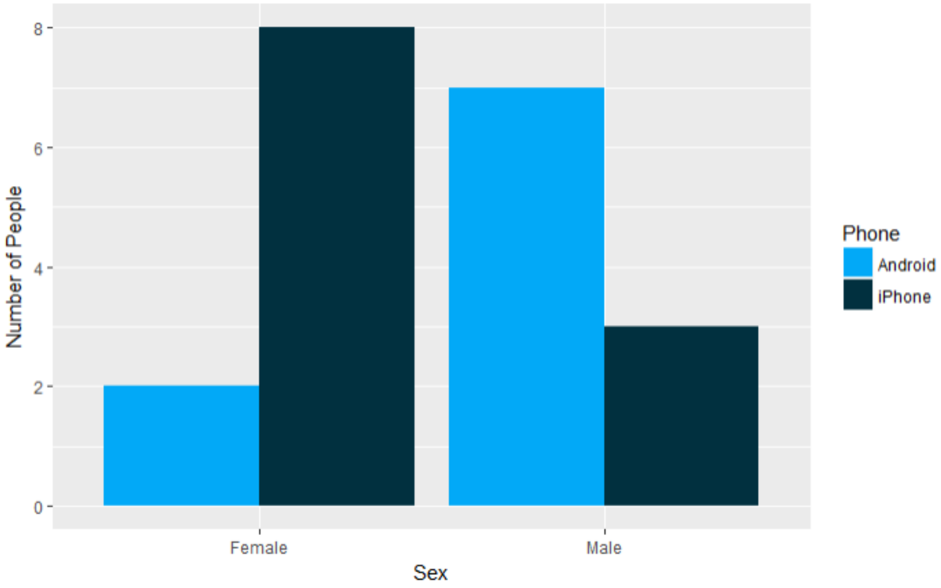
The above plot is created from the raw data. Every bar goes up to two because there are two rows in the spreadsheet for each sex.

5. Open another R code chunk and copy your bar chart code to the new chunk. Add the syntax y = Count inside the aes() function in the ggplot () layer. This will map the counts (in the Count column) to the y-position. Also include stat=”identity” in the geom_bar() layer. Run your code.

You now have a bar chart with two columns (male and female) each 10 units tall.

6. Copy your bar chart syntax to a new R code chunk and fill the bar chart based on Phone.

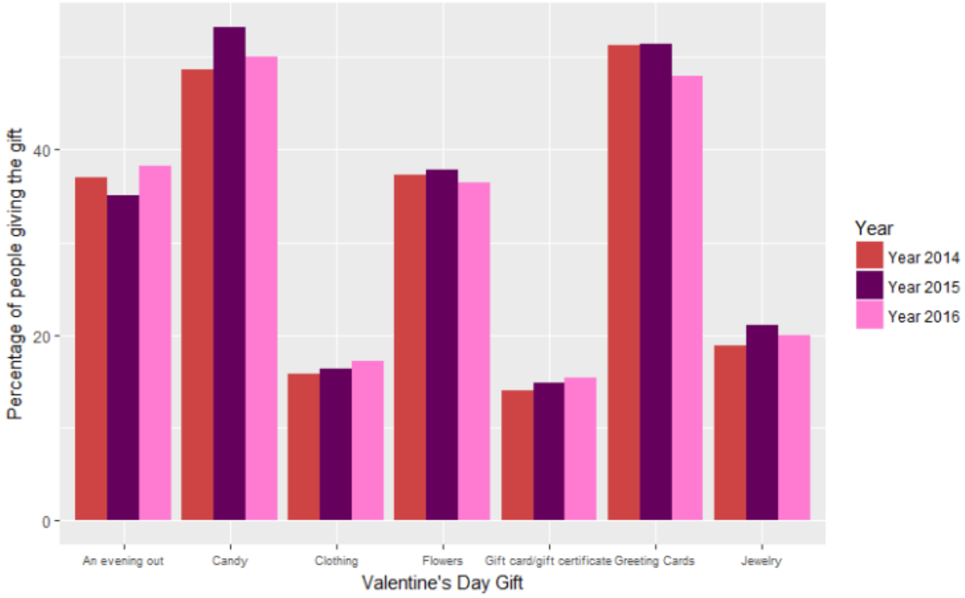
7. Again copy the above code and this time make a side-by-side bar chart. Change the y-axis label to something more informative and change the colors of the bars. Give the Android phones the rgb color (2, 169, 247) and the iPhones (1, 48, 63)



Part II: Your Turn – Valentine's Day Gifts

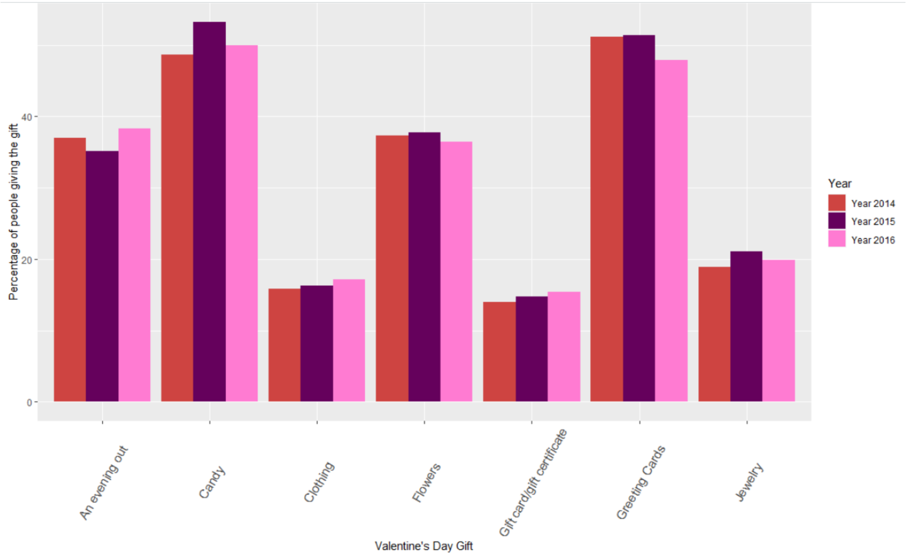
8. Load the *valentine-gifts.csv* data. These data give the percentage of each gift type given for Valentine's Day in 2014, 2015, and 2016.

9. Recreate the following plot. The color palette uses rgb (206, 68, 65) for the year 2014, (101, 1, 92) for 2015 and (255, 123, 210) for the year 2016. In addition to the colors, be sure to match the labels provided in the graph below.

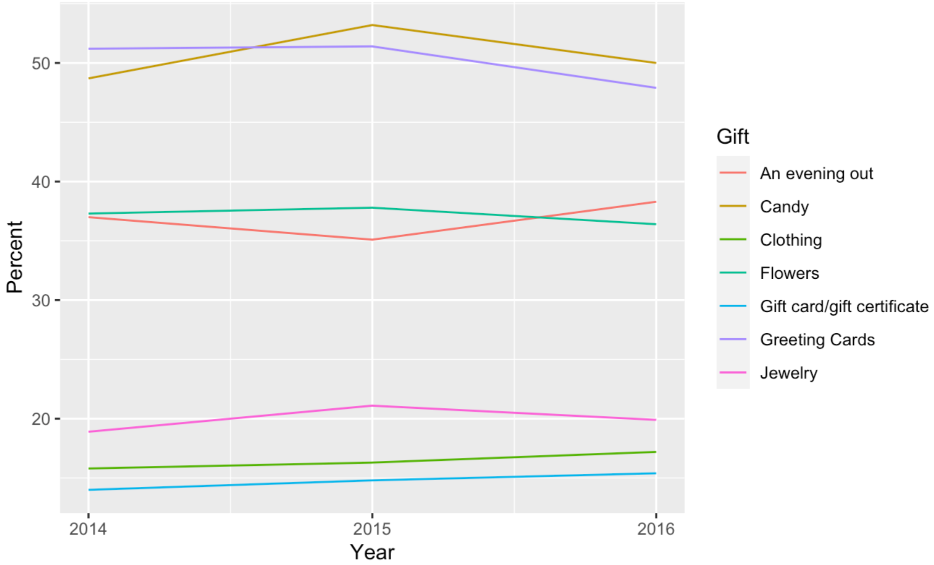


10. Let's make the Valentine's Day Gift words a bit easier to read. Try changing the font size of the gift words on the x-axis. Check out this webpage for how to change font size. You may need to play around with the font size value to make the bar labels more readable. Change the size value to see how the labels change.

11. Another way to make the bar labels easier to read without decreasing the font size is to put the words at an angle. Check out this webpage for how to rotate the tick mark labels.

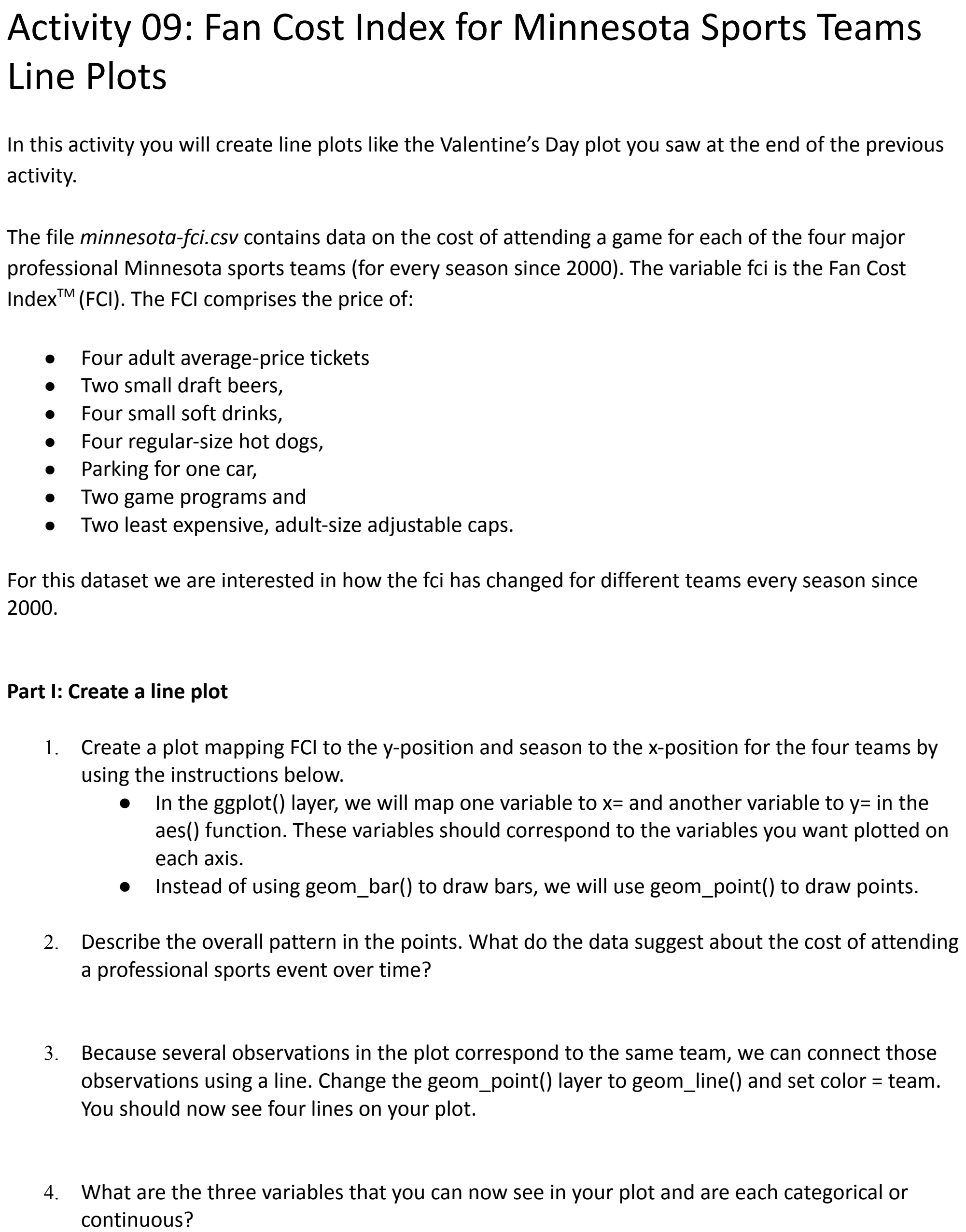


An alternate way to view these same three variables is with a line plot like the one below.



12. Consider your final side-by-side bar graph for Valentine's Day gifts and the line plot above.
    a. What features standout on the bar graph?
    b. What features standout in the line plot?
    c. Which graph would you use to convey the relationship among these variables? Explain your answer.

A.3 Activity 3



Activity 09: Fan Cost Index for Minnesota Sports Teams Line Plots

In this activity you will create line plots like the Valentine's Day plot you saw at the end of the previous activity.

The file *minnesota-fci.csv* contains data on the cost of attending a game for each of the four major professional Minnesota sports teams (for every season since 2000). The variable fci is the Fan Cost Index$^{\text{TM}}$ (FCI). The FCI comprises the price of:

- Four adult average-price tickets
- Two small draft beers,
- Four small soft drinks,
- Four regular-size hot dogs,
- Parking for one car,
- Two game programs and
- Two least expensive, adult-size adjustable caps.

For this dataset we are interested in how the fci has changed for different teams every season since 2000.

Part I: Create a line plot

1. Create a plot mapping FCI to the y-position and season to the x-position for the four teams by using the instructions below.
   - In the ggplot() layer, we will map one variable to x= and another variable to y= in the aes() function. These variables should correspond to the variables you want plotted on each axis.
   - Instead of using geom_bar() to draw bars, we will use geom_point() to draw points.

2. Describe the overall pattern in the points. What do the data suggest about the cost of attending a professional sports event over time?

3. Because several observations in the plot correspond to the same team, we can connect those observations using a line. Change the geom_point() layer to geom_line() and set color = team. You should now see four lines on your plot.

4. What are the three variables that you can now see in your plot and are each categorical or continuous?

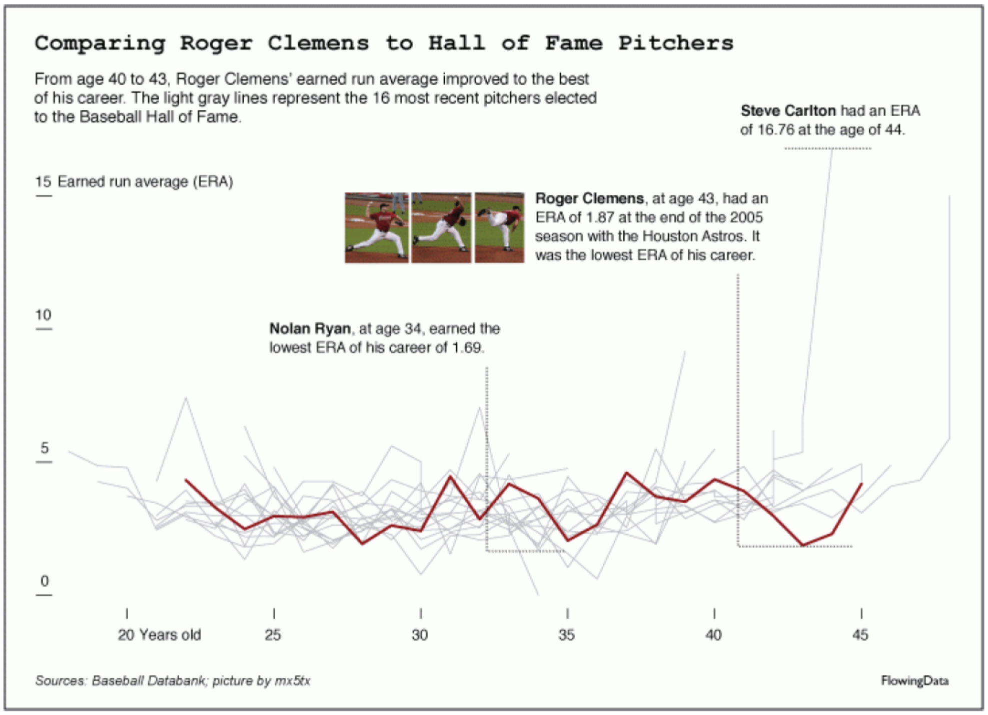
5. The Minnesota Wild line had a break in it during the 2004–2005 season. Do some research to find out why?

6. Describe the change in cost over time for each of the four teams.

7. How do the change in costs over time compare across the four teams? Explain.

Part II: Highlighting a Particular Team

Take a look at the line plot below. Nathan Yau (the person who writes the visualization blog Flowing Data) created a lineplot to show how Hall of Fame pitcher Roger Clemens' ERA (a baseball pitching statistic) compares to other pitchers in the Baseball Hall of Fame. In the lineplot displaying Roger Clemens' ERA over time, the designer of the plot used color to highlight a particular line (Roger's). All the other lines were colored gray to de-emphasize them. In this way, it is easy for a reader to see Roger's trend, while also being able to compare his trend to others.



Comparing Roger Clemens to Hall of Fame Pitchers

From age 40 to 43, Roger Clemens' earned run average improved to the best of his career. The light gray lines represent the 16 most recent pitchers elected to the Baseball Hall of Fame.

Steve Carlton had an ERA of 16.76 at the age of 44.

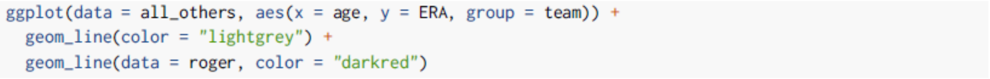
Roger Clemens, at age 43, had an ERA of 1.87 at the end of the 2005 season with the Houston Astros. It was the lowest ERA of his career.

Nolan Ryan, at age 34, earned the lowest ERA of his career of 1.69.

Sources: Baseball Databank; picture by mx5tx

FlowingData

If the designer were using ggplot to create this effect, they would need two separate data sets. One would be the data for Roger's line only. The other dataset would include the data for all the other pitchers. Let's say these two datasets were imported as roger and all_others, respectively. Further, imagine that in BOTH datasets there was a pitcher_name variable, an age variable, and an ERA variable. The syntax to create this plot would look like this:



```
ggplot(data = all_others, aes(x = age, y = ERA, group = team)) +
  geom_line(color = "lightgrey") +
  geom_line(data = roger, color = "darkred")
```

The first geom_line() layer will use the data from the all_others data set and maps age to the *x*-position and ERA to the *y*-position. It also groups together the data by team, so that a separate line is drawn for each team. Since we want to color ALL the pitcher's lines the same color, we set the color to a specific value (not a variable) and do NOT include it inside aes(). (This is referred to as a fixed-aesthetic.)

The second geom_line() layer will use the data from the roger data set. It will also map age to the *x*-position and ERA to the *y*-position. (If there is no mapping in the specific layer, the mapping from the first ggplot() layer is used.) We again set a fixed-aesthetic of color (this time a dark red color to differentiate it from the grey lines) for the line.

8. The file *nhl-fci.csv* includes data on the FCI for all the teams except the Minnesota Wild. Create a data file for the Minnesota Wild (use the data values from the *minnesota-fci.csv* file). Be sure that this data has the EXACT SAME variable names as the *nhl-fci.csv* data.

9. Use the data from the file *nhl-fci.csv* to create a line plot showing the cost of attending an NHL game over time. Then add the line for the Minnesota Wild to this plot. Use color to highlight the Wild's line. (Note: Make sure that you change the line colors so that they are differentiable from the theme. Grey lines on a grey background can be problematic.)

10. One way to de-emphasize the other NHL team's lines is to make them semi-transparent. To do this include the syntax alpha=value (where *value* is a number between 0 and 1) in the geom_line() layer for the other NHL teams. Setting the alpha value to 0 will make the lines completely transparent, and setting it to 1 makes the lines completely opaque (1 is the default value if this option is not included). Try different alpha values until you get a plot that shows the lines, but de-emphasizes them and makes the Wild line stand out.

11. Add a label for every other year to the x-axis (starting with 2000).

12. Compare how the cost of attending a Minnesota Wild game has changed over time to how the cost of attending other NHL teams' games has changed over time.

A.4 Activity 4

Activity 10: Introduction to Directed Acyclic Graphs

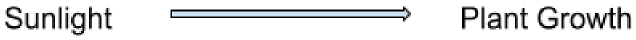
Part I: Acyclic Directed Graphs to Support Investigation

The data you have analyzed so far in this course is referred to as *observational data*. Data from an observational study are not collected as a part of an experiment where treatments are assigned. This type of data is commonly collected through our daily activities, on our fitness trackers, through our patterns of technology use, online shopping habits, etc. It is often multivariate, containing many different variables that might have a variety of complex relationships among and between them. An important part of learning to create visualizations to analyze data is being able to investigate multivariate relationships in order to learn more about the data. To do this it is important to have a theory about the nature of these relationships before we begin analysis.

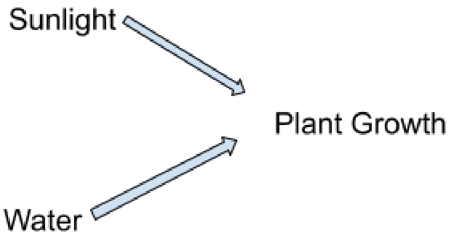
Introduction to Directed Acyclic Graphs

One tool we use to investigate relationships among variables in a meaningful way is a directed acyclic graph (DAG). DAGs are different from the graphs that you have seen in the course thus far. Instead of plotting the data, these graphs propose a relationship among the variables we are interested in.

For example, if we think that the variable sunlight affects plant growth, we create a DAG by placing an arrow pointing from sunlight to plant growth.



We can make this diagram more complex by adding other variables we think affect plant growth. For example, plant growth is likely dependent on other factors, such as water. The DAG below depicts how both sunlight and water affect plant growth.



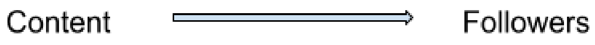
Take a few minutes to discuss with your partner whether you think the graph above needs any more relational arrows between the variables listed. Draw them and explain your reasoning below.

Class discussion and presentation of a few final plant diagrams.

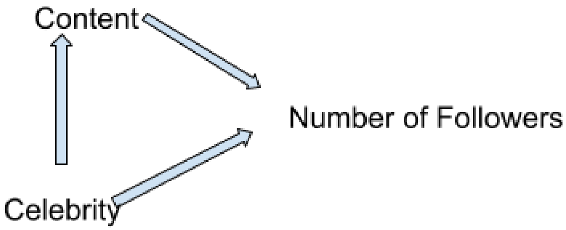
Instagram Followers

Now it is your turn to create DAGs to model the relationships among variables. Suppose we want to construct DAGs to consider what variables affect the number of followers an Instagram account has. One variable we might consider is the type of content (i.e. health, politics, hobbies, or general lifestyle tips) an account posts about.

Below draw the DAGs depicting the relationship between the variables *Content* and *Number of Followers.*



Another variable that might affect the number of followers an Instagram account has is whether or not the account is for a celebrity. We might incorporate this into your DAG like we see below. Celebrities might have more followers, and what they are famous for might affect their content. Hence, we need arrows between Celebrity and Content and between Celebrity and Number of Followers.



Take a minute with your partner to consider other variables that you think might affect the number of followers a person has on Instagram.

- Discuss what variables you think affect the number of followers an account has. Choose two to incorporate into your DAG.
- Discuss what type of relationship you think those variables have with the number of followers.
- Be sure to consider that the variables you chose might not only affect the number of followers but also the celebrity status or the content of the creator.
- Draw all arrows (and potentially double arrow heads) that you think should be in your DAG.
- Put your final DAG below with an explanation for each relationship in the graph.

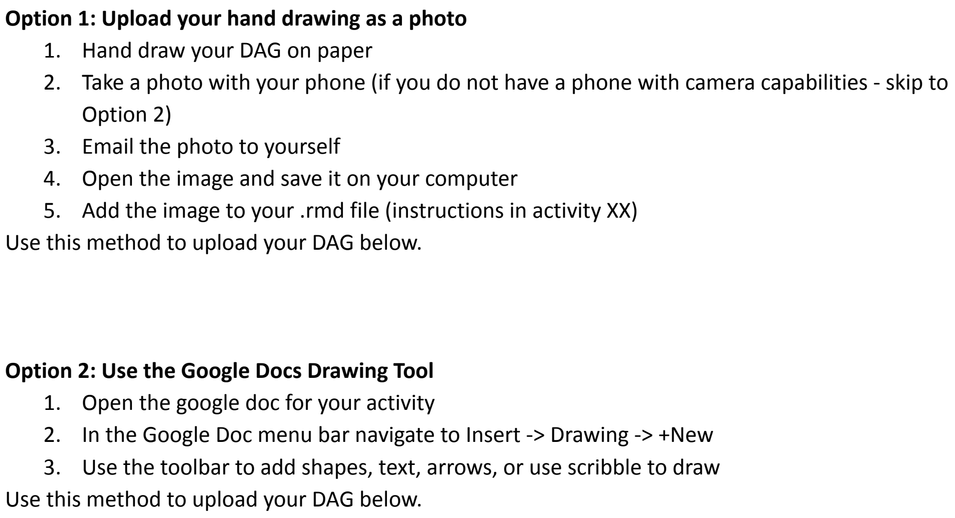
Next, you will join with another pair to compare DAGs and have a discussion about how you created yours. Have students add to my diagram on the board from their groups in pairs.

Is there anything that your partner group came up with that you did not? Incorporate any other variables and arrow heads that you think should be in the graph.

Class discussion and presentation of as many final graphs as time allows.

Part II: Upload DAGs to assignments

DAGs are a useful tool to propose and summarize relationships among variables that we are interested in. They can give us a starting point for investigating variables in a dataset. Then they can be modified based on exploratory analysis. Though there are more advanced techniques for using DAGs to determine causality, but we won't get to those in this class. However, we will continue to use them to propose and discuss relationships among variables. You will need to create DAGs and be able to upload them into your class activities and assignments. Below are a couple options to try.



Option 1: Upload your hand drawing as a photo

1. Hand draw your DAG on paper
2. Take a photo with your phone (if you do not have a phone with camera capabilities - skip to Option 2)
3. Email the photo to yourself
4. Open the image and save it on your computer
5. Add the image to your .rmd file (instructions in activity XX)

Use this method to upload your DAG below.

Option 2: Use the Google Docs Drawing Tool

1. Open the google doc for your activity
2. In the Google Doc menu bar navigate to Insert -> Drawing -> +New
3. Use the toolbar to add shapes, text, arrows, or use scribble to draw

Use this method to upload your DAG below.

A.5 Activity 5

Activity 11: Scatterplots Themes and Titles

Much like the line plots that you explored in previous activities, scatterplots are used to map the values on two quantitative variables to a two-dimensional space. This mapping allows you to understand the relationship between the two variables. We can also add features like colors and shapes to look at even more variables in the same plot. In this activity, we will use scatterplots to explore the relationships among the variables in this dataset.

We create scatterplots in ggplot2 much like we created line plots. The only difference is we will use the layer geom_point() instead of geom_line().

The data set *women-stem.csv* contains data on 76 majors in STEM fields. (Here STEM is defined as any major categorized as engineering, computing, science, math or health.) The data are from the American Community Survey 2010-2012 Public Use Microdata Series and were made available from fivethirtyeight. The variables in this data set are:

- **Women:** This variable indicates the proportion of female graduates with this major.
- **Income:** This variable indicates the median income (in thousands of dollars) for a person with this major working full-time, year-round.
- **Major:** STEM major
- **Category:** Type of STEM major (e.g. Engineering)

Open a new RNotebook and customize the YAML, load the libraries, and import the data file, *women stem.csv*.

Part I: Creating and Describing the Plot

We want you to create a scatterplot that allows you to examine the relationship between the proportion of female graduates (*x*-position) and median income (*y*-position) for the 76 STEM majors.

1. Create the scatterplot to examine this relationship. Give the axes appropriate labels, source, and include a title for the plot.

   Use the code below to create a scatterplot:

   ```
   ggplot(data =women_stem, aes(x = Women, y = Income)) +geom_point()
   ```

To describe scatterplots we can comment on linearity, slope, and the strength of their relationship:

a. Linearity: how closely the points appear to follow a straight line
b. Slope: the general trend of the points upward or downward

c. Strength: the points closely follow a linear path or are they more scattered around the plot

2. Use the plot you created to describe the linearity, slope and strength of the relationship.
3. Explain why the direction/trend of the relationship between these variables would be described as negative.
4. To interpret this trend, fill in the blanks:

The _____(*higher* or *lower*) the proportion of women the __________ (*highe*r or *lower*) the income.

Part II: Further Exploration of Relationships

One explanation for this relationship may be that the type of STEM major that attracts women are the same majors that pay less well. To explore this, re-create your scatterplot, but now color the points by the type of STEM major (the *Category* variable).

5. In a new R code chunk, re-create the scatterplot and color the points by the type of STEM major.
6. Based on the resulting plot, is the explanation offered reasonable (that the type of STEM major that attracts women are the same majors that pay less well)? Explain.
7. What other factors do you think might affect income?
8. Create a DAG with at least three variables that you think might affect women's income in STEM fields.

Part III: Upgrading Plots

Sprucing Up the Plot: Themes and Titles

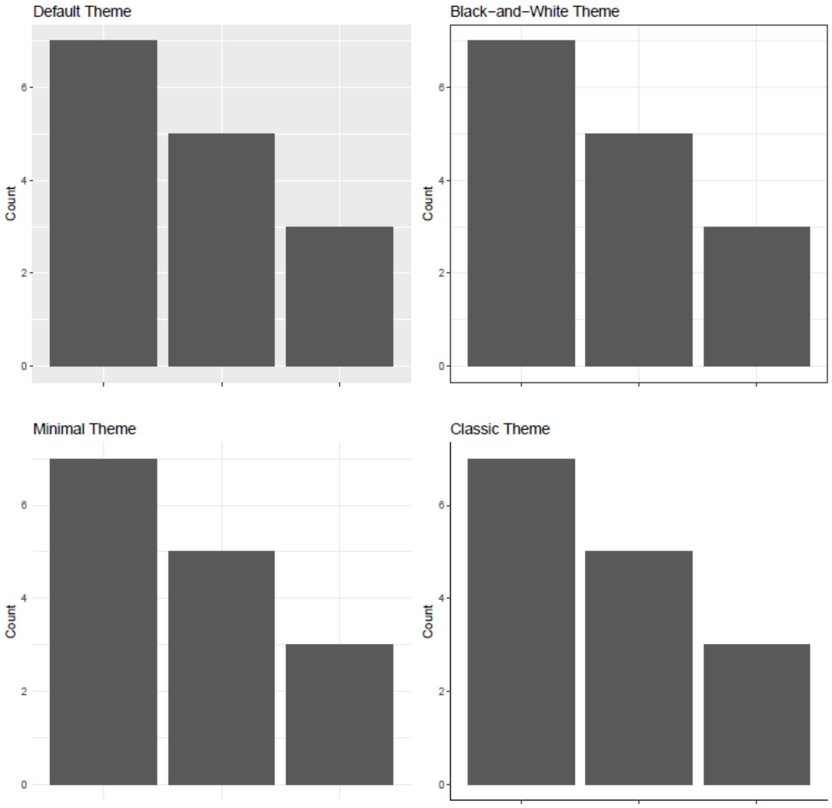
Once you have the bones of the plot you want, you can focus on the little things like changing the theme, adding a title, etc. These small things make the plot look great. However, you should not worry about the small things until you get the initial plot to look like you want it to.

Themes

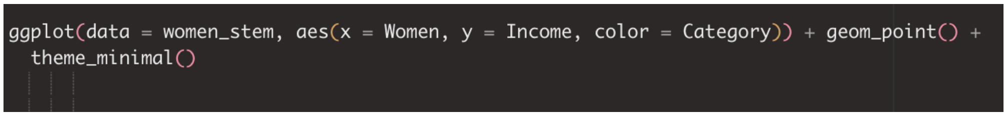
The first change you will make is to the plot's theme. By default, the background for ggplot plots is grey with a white grid system. There are several pre-programmed themes that come with ggplot, including:

- Black-and-White theme: theme_bw()
- Minimal theme: theme_minimal()
- Classic theme: theme_classic()

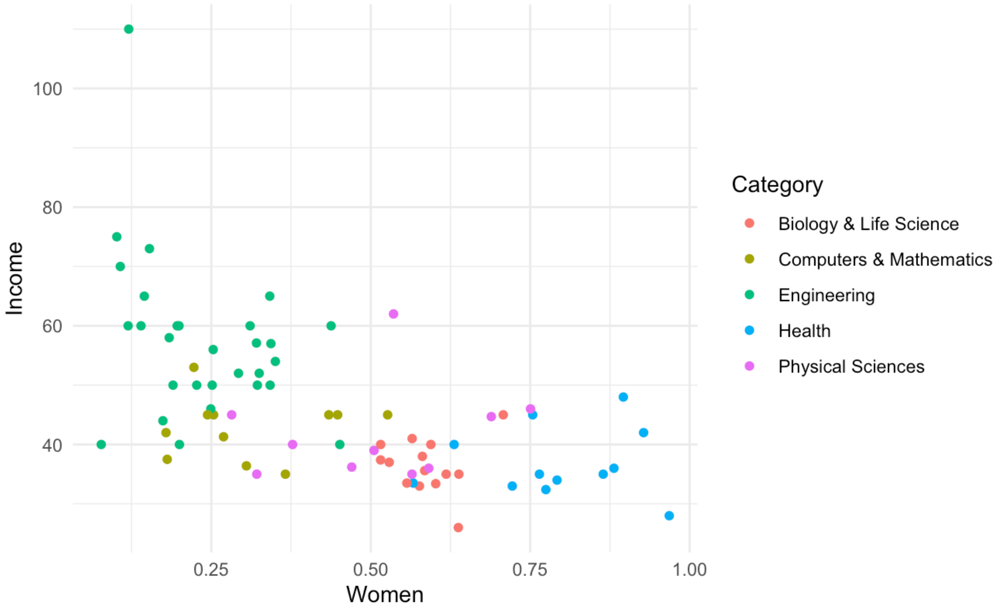
Below I show four plots of some generic data. In the top row, the left plot shows the default (grey) theme and the right plot shows the black-and-white theme. In the bottom row, the left plot shows the minimal theme and the right plot shows the classic theme.



Each of these themes makes different changes to the background color, grid, and axes (as well as to other theme elements). To use one of these themes, we just add the appropriate theme layer to the ggplot syntax. For example, to apply the minimal theme in our Women in Stem plot, we use the following syntax:



```
ggplot(data = women_stem, aes(x = Women, y = Income, color = Category)) + geom_point() +
  theme_minimal()
```

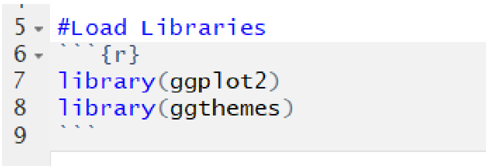


You can see a complete list of the ggplot themes and some examples at https://ggplot2.tidyverse.org/reference/ggtheme.html

1. Go to the website above and choose a theme. Change your code to reflect your choice and run your code.

More Themes from the ggthemes Package

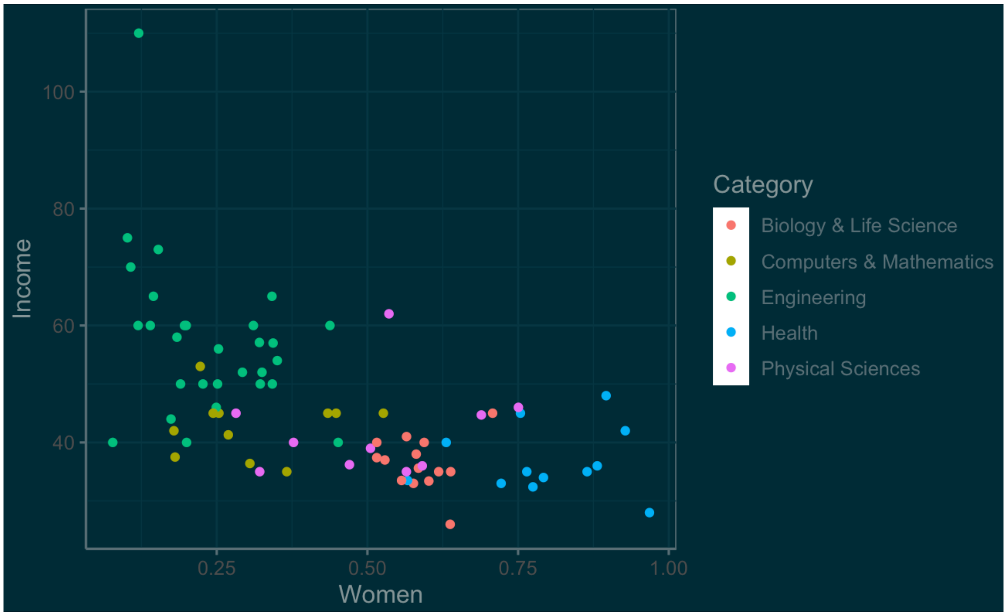
If those themes do not excite you, or you are looking for something different, you can install and load the ggthemes package. (You can install it by clicking Packages > Install and entering in ggthemes.) Load it by including library(ggthemes) in your R code chunk where to load the ggplot2() library.



~~~
#Load Libraries
```{r}
library(ggplot2)
library(ggthemes)
```
~~~

This package has many different themes and pre-programmed color palettes. You can see them and read about how to use them at https://jrnold.github.io/ggthemes/reference/index.html

Below we use the theme_solarized() layer to add a dark theme.



The package has themes that mimic plots from *The Economist*, *The Wall Street Journal, Excel*, and even fivethirtyeight.

2. Pick a theme from the ggthemes package and apply it to your plot.

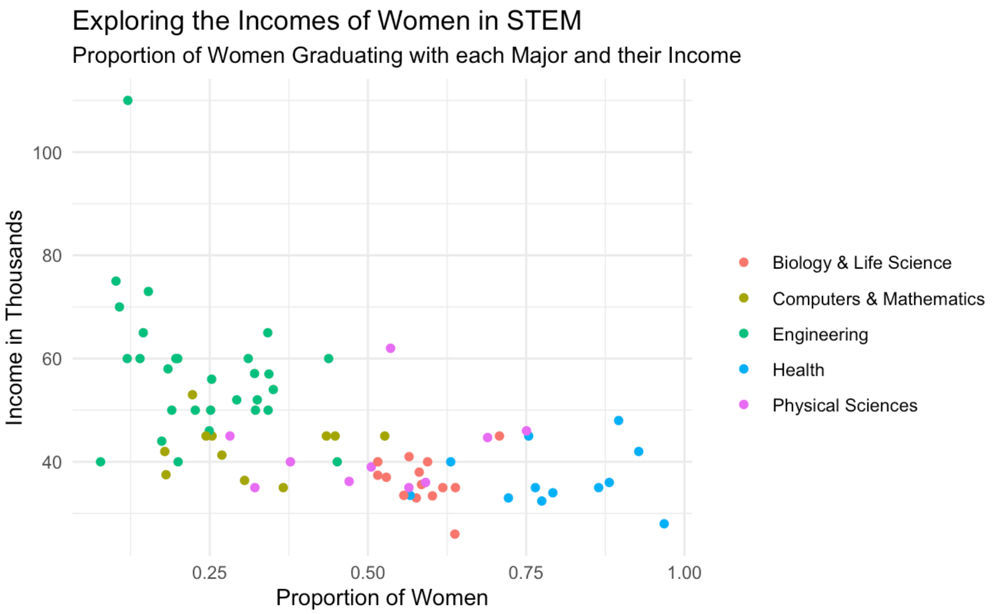
Part III: Titles and Labels All in One Layer

The labs() layer is a generic layer that allows us to add as many labels as we want (e.g. title, subtitle, x-axis).

In the labs() layer, we just specify which labels we want to use. We can even change the label above the fill colors in the legend by using color= "What you want the legend label to be."

Two other labels we can add/modify to the plot are subtitle= and caption=. These add a subtitle and caption to our plot respectively. Below we add these elements to our plot. In addition, we change the color label on the plot to be blank. (Do this by putting nothing, or a space, in the quotation marks.) The category labels for each major are descriptive enough that the legend does not need a label.

3. Modify your code to make your plot look like the one below.



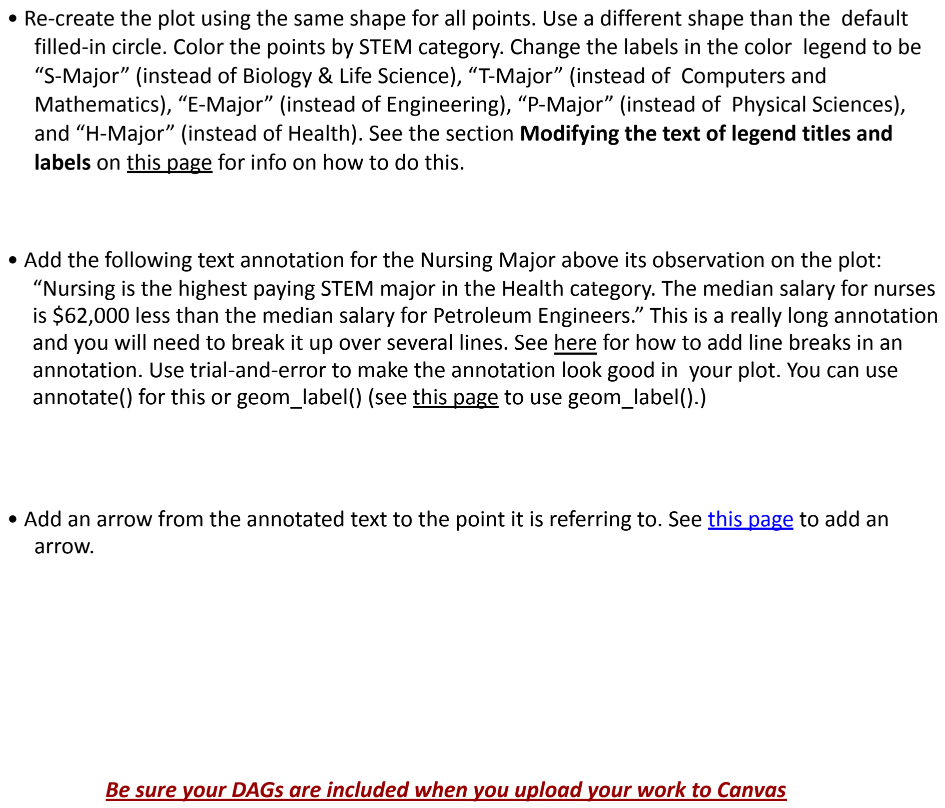
Part IV: More Plot Aesthetics

Using the base plot you made in Part I of this activity try making each of these changes. Make a new plot for each bullet listed below.

While these are all ways we can customize a scatterplot we do not necessarily want to add all these layers of code to our finalized plot.

- Change the points to labels that give the STEM major for all 76 points. See this webpage for examples of how to do that. You can also spruce up the labels. See this page to see examples of how to improve the labeling.

- Re-create the plot using points. Then, add an aesthetic so that each category of STEM major has a different shape. See this page for info.



- Re-create the plot using the same shape for all points. Use a different shape than the default filled-in circle. Color the points by STEM category. Change the labels in the color legend to be “S-Major” (instead of Biology & Life Science), “T-Major” (instead of Computers and Mathematics), “E-Major” (instead of Engineering), “P-Major” (instead of Physical Sciences), and “H-Major” (instead of Health). See the section **Modifying the text of legend titles and labels** on this page for info on how to do this.

- Add the following text annotation for the Nursing Major above its observation on the plot: “Nursing is the highest paying STEM major in the Health category. The median salary for nurses is $62,000 less than the median salary for Petroleum Engineers.” This is a really long annotation and you will need to break it up over several lines. See here for how to add line breaks in an annotation. Use trial-and-error to make the annotation look good in your plot. You can use annotate() for this or geom_label() (see this page to use geom_label().)

- Add an arrow from the annotated text to the point it is referring to. See this page to add an arrow.

Be sure your DAGs are included when you upload your work to Canvas

A.6 Activity 6

Activity 12: High Peaks

In this activity we will investigate the difficulty ratings for the 46 High Peaks in the Adirondack Mountains. These mountains are known as High Peaks because they have elevation around 4000ft. The peaks have varying difficulty, but what makes some more difficult than others? We will explore the difficulty rating throughout this activity.

You can find the dataset *high-peaks.csv*[1] in the course data folder. See the data information in the data set manual here: http://www.stat2.org/manuals/Stat2DataManual.pdf

1. Does a difficulty rating of 1 indicate that the hike is easy or difficult?

2. Create an aesthetically pleasing (think title, labels, theme, values on the x-axis, etc.) histogram of the difficulty variable. Paste your plot below.

3. Describe your plot.

4. What are some of variables in the dataset that you think will be related to the difficulty rating of the hike?

5. Are there any variables that you do not think will be related?

6. Draw a DAG to depict what variables you think will affect the difficulty rating.

We are interested in what variables affect the Difficulty rating. Let's start by using a scatterplot to look at the association between Difficulty and Time.

7. Paste your scatterplot below.

[1] Data originally from http://www.stat2.org

Use the plot to answer the following questions.

8. Describe the plot.

9. Explain why the direction/trend of the relationship between these variables would be described as positive.

10. To interpret this trend, fill in the blanks:

Mountains that take _____(*more* or *less*) time to climb tend to have a__________ (*highe*r or *lower*) difficulty rating.

11. Do you think TIme is an important variable to consider when determining the difficulty of the peak? Use your graph and description to support your answer,

Next we will investigate the relationship between *Elevation* and *Difficulty*.

12. Create the scatterplot for this relationship.
13. Describe the plot.

14. Describe why Elevation doesn't appear to have an association with Difficulty. Use your graph and description to support your answer.

Next we will investigate the relationship between *Length* and *Difficulty*.

15. Create the scatterplot for this relationship.
16. Describe the plot.
17. Do you think Length is an important variable to consider when determining the difficulty of the peak? Use your graph and description to support your answer.

Adding Color

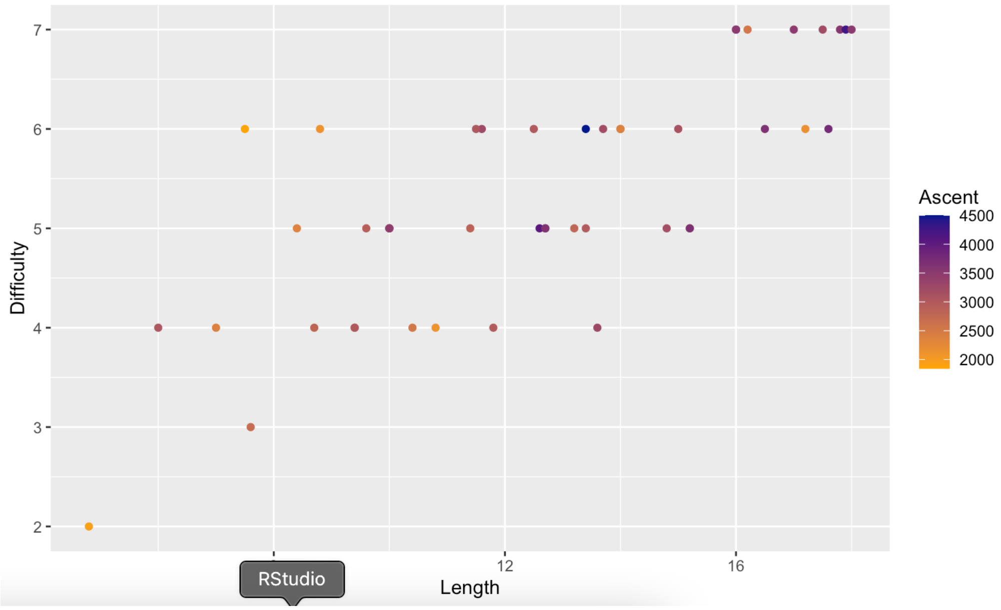
18. Create a plot with Length on the x-axis, Difficulty on the y-axis, and colored corresponding to the Time variable.

```
ggplot(data = HighPeaks, aes(y = Difficulty, x = Length, color = Time)) +geom_point() +
```

scale_color_gradient(low = "orange", high = "darkblue")

19. Experiment with the colors in scale_color_gradient() to see how it changes the color gradient. Paste a plot with a different color scheme below.
20. Based on your plot, describe the relationship among Time, Length, and Difficulty rating. Does it appear both affect the Difficulty rating?
21. Draw a DAG to display these relationships.

Below is a graph of another variable in the dataset that we have not yet considered, *Ascent*. This is a measure of the vertical distance in feet to the top of the mountain.

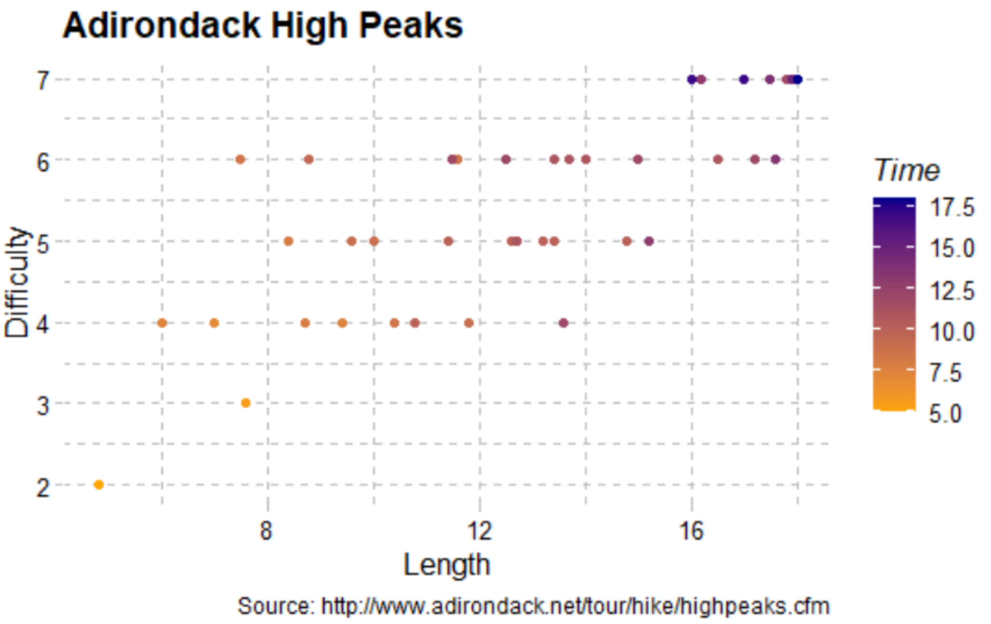


22. Do you think Ascent has an association with Difficulty? Explain why or why not based on this graph.

23. Draw a final DAG considering all variables in the dataset (elevation, time, length, and ascent), but only including those that you think affect the difficulty rating.

24. Explain why you included each variable in your DAG using evidence from your plots.

Part II: More titles and Themes:

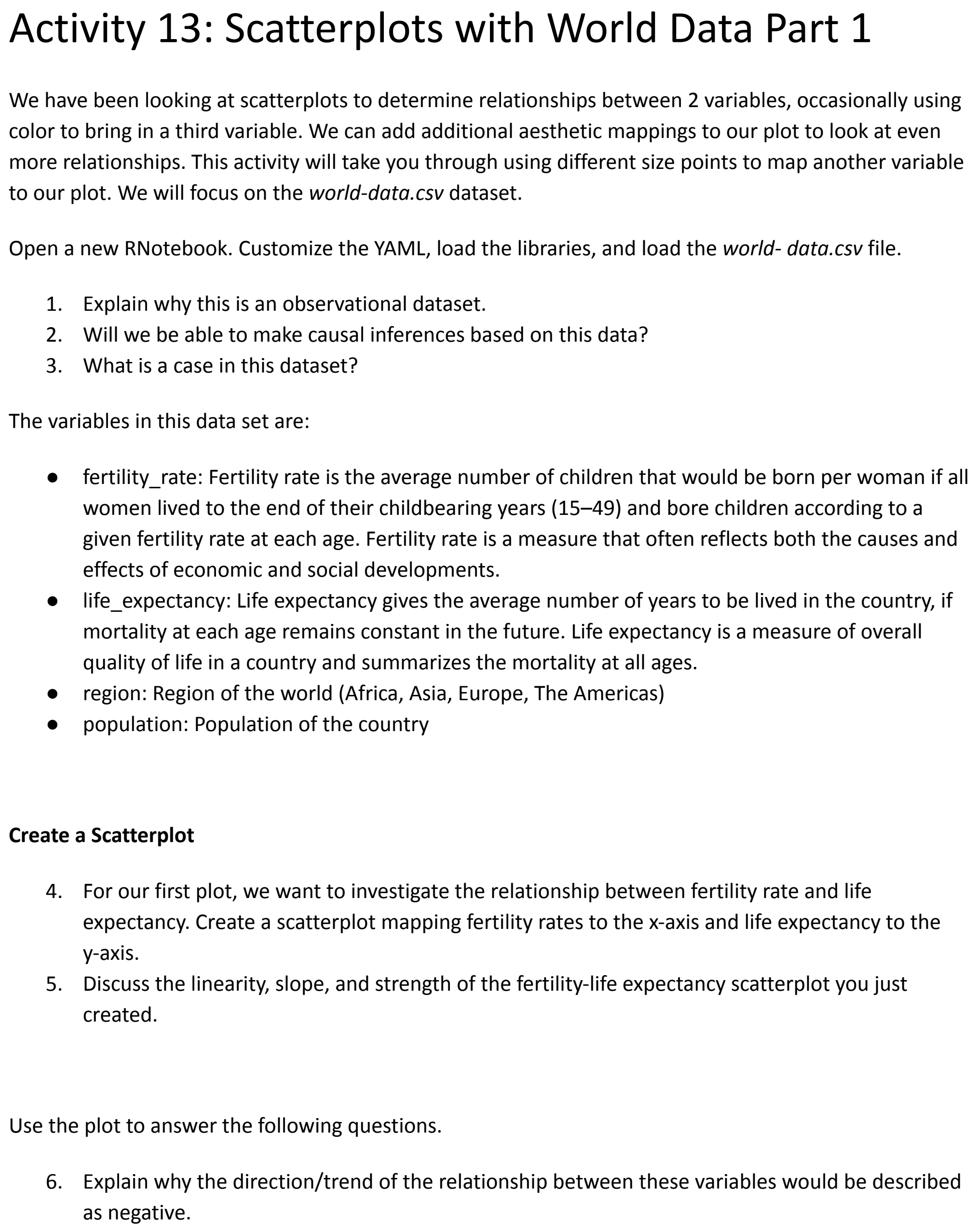


25. Re-create the plot above using theme_pander() from the *ggthemes* package. Making sure to add in the title and data source.

26. Paste your plot below.

Be sure your DAGs are included when you upload your work to Canvas

A.7 Activity 7



Activity 13: Scatterplots with World Data Part 1

We have been looking at scatterplots to determine relationships between 2 variables, occasionally using color to bring in a third variable. We can add additional aesthetic mappings to our plot to look at even more relationships. This activity will take you through using different size points to map another variable to our plot. We will focus on the *world-data.csv* dataset.

Open a new RNotebook. Customize the YAML, load the libraries, and load the *world- data.csv* file.

1. Explain why this is an observational dataset.
2. Will we be able to make causal inferences based on this data?
3. What is a case in this dataset?

The variables in this data set are:

- fertility_rate: Fertility rate is the average number of children that would be born per woman if all women lived to the end of their childbearing years (15–49) and bore children according to a given fertility rate at each age. Fertility rate is a measure that often reflects both the causes and effects of economic and social developments.
- life_expectancy: Life expectancy gives the average number of years to be lived in the country, if mortality at each age remains constant in the future. Life expectancy is a measure of overall quality of life in a country and summarizes the mortality at all ages.
- region: Region of the world (Africa, Asia, Europe, The Americas)
- population: Population of the country

Create a Scatterplot

4. For our first plot, we want to investigate the relationship between fertility rate and life expectancy. Create a scatterplot mapping fertility rates to the x-axis and life expectancy to the y-axis.
5. Discuss the linearity, slope, and strength of the fertility-life expectancy scatterplot you just created.

Use the plot to answer the following questions.

6. Explain why the direction/trend of the relationship between these variables would be described as negative.

7. To interpret this trend, fill in the blanks with the words *high* and *low*:

Countries that have ___________ fertility rates tend to also have ___________ life expectancies.

8. Do you think any of the other variables in the dataset affect life expectancy?

Sprucing Up the Plot: Colors, Labels, and another variable

9. Next, we will look at the relationship between life expectancy, fertility rate, and region of the world. Propose a DAG for the relationships among these variables and draw it below. Explain your reasoning for arrow directions on your plot.

10. To color the points by some variable we add color= inside the aes() function of the ggplot() layer (not fill=). To adjust the color values, since we used color=, we use the layer scale_color_manual() (not scale_fill_manual).

11. Color the points in your scatterplot by region of the world. Use the following HEX color values:
    a. Africa: 00D5E9 (blue)
    b. The Americas: 7FEB00 (green)
    c. Asia: FF5872 (red)
    d. Europe: FFE700 (yellow)

12. Add the following labels:
    a. Title: Population Aging around the World
    b. Subtitle: Relationship between fertility rate and life expectancy by region
    c. Color: (NO LABEL ON THE LEGEND)
    d. x: Fertility rate
    e. y: Life expectancy
    f. caption: Source: https://www.census.gov/

13. Based on your plot what do you think is the relationship between region and life expectancy?

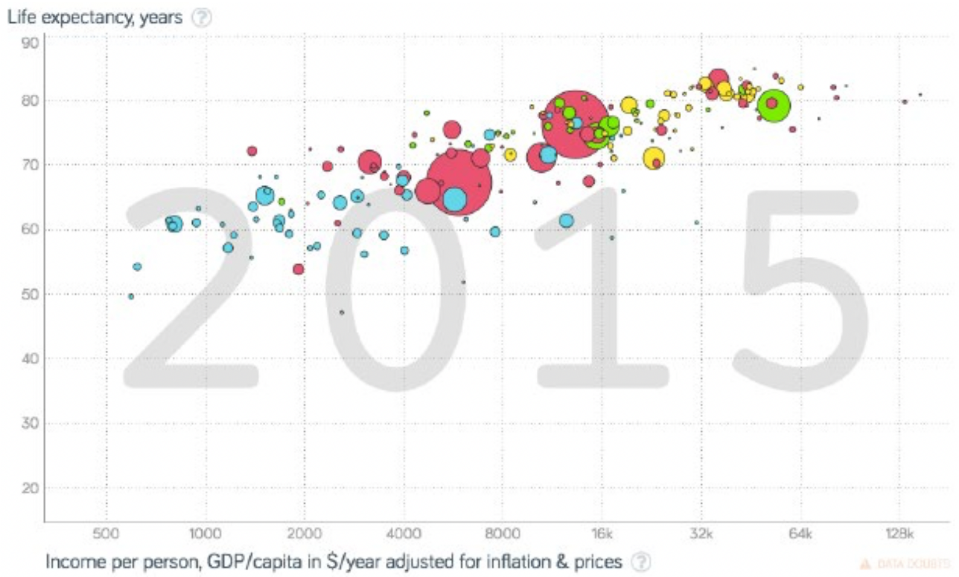
14. Based on your plot what do you think is the relationship between region and fertility rate?

15. Draw your final directed graph for the relationship among these variables and explain your drawing using evidence from your plot.

16. Write a short description of your scatterplot (what is the overall trend in the data?, are different regions of the world located in different areas of the graph or are the regions of the world scattered throughout the plot?). What is the main point or take-away from this visualization?

Bubble Chart: Scatterplot with a Size Mapping

So far we have used three variable/aesthetic combinations in the plot: fertility rate (mapped to the x-position), life expectancy (mapped to the y-position), and region (mapped to color). One thing that Hans Rosling did on Gapminder, was to also map each country's population to the size of the point. Below is a scatterplot from Gapminder.



Rosling referred to this plot as a bubble chart. Bubble charts are just scatterplots that include a size mapping. This allows us to consider a fourth variable (population).

17. Add population to your final directed graph from above to imply it will have an effect on life-expectancy. Put your drawing below.

18. Map each country's population to the size of the point. To do this, include the syntax size=population is the aes() function of the ggplot() layer.

19. You can change the label in the legend for the size mapping by adding to the labs() layer. Just include size= to this layer. Change the labels to say “Country population”.

Use the plot you created to answer the following questions.

20. Do countries with larger populations tend to have lower or higher fertility rates than countries with smaller populations?

21. Do countries with larger populations tend to have lower or higher life expectancies than countries with smaller populations?

22. Do countries with larger populations tend to come from particular world regions? Which regions?

23. Do you think the relationship between population size and life expectancy is as strong as the relationship you saw between region and life expectancy?

24. Draw your final DAG to summarize the relationships among the 4 variables. Keep in mind any relationships you do not think are very strong you may leave off your graph. You only want to include variables you can make a case impact life expectancy.

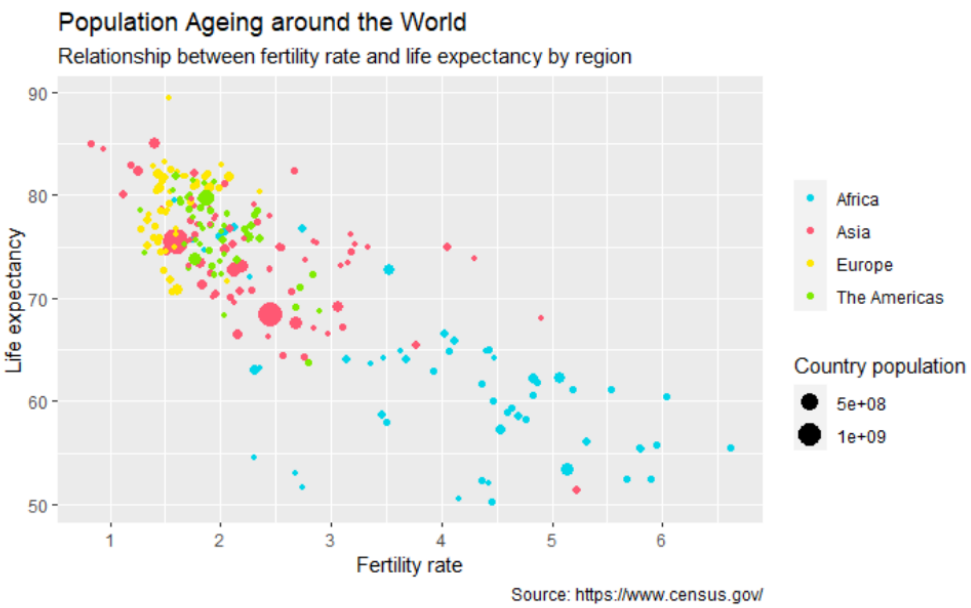
Be sure your DAGs are included when you upload your work to Canvas

A.8 Activity 8

Activity 14: Scatterplots with World Data Part 2

Recall: Gapminder Plot

Recall the plot that you made in the previous class activity. Create a new RNotebook and recreate the plot below.



Source: https://www.census.gov/

Part I: Adding Text Annotation to the Plot

There are times when it is useful to add annotations to the plot. In our plot, it might be useful to identify particular countries by adding text to the plot. To add an annotation, we add an annotate() layer to the plot. There are several types of annotations we can add to the plot, but in this class, you will primarily be adding text. We also need to give the *x* and *y coordinates* for the text, and the actual text we want to write.

For example, say we want to identify the European country that has an extremely high life expectancy (close to 90 years). Looking in the data we see that this country is Monaco.

country fertility_rate life_expectancy region population

Monaco 1.53 89.5 Europe 30581

The text we want to add is “Monaco”. The x-position (fertility rate) for Monaco is 1.53, and the y-position (life expectancy) is 89.5. These are good starting points for the text. The syntax to add this annotation would be:

```
annotate(geom = “text”, x = 1.53, y = 89.5, label = “Monaco”)
```

The geom=”text” argument indicates that the annotation is text. The label=argument specifies the actual text to be added to the plot. The text is centered at the x- and y-position, and you will need to adjust these values (via trial-and-error) to get the text exactly where you want it. Make sure the annotation is not covering the point you are annotating. Convention is to place the annotation to the right of the point. You can also include the argument size= in the annotate() layer to adjust the size of the text. The default size is size=4.

1. Add a text annotation to identify the United States.

Guides: Removing Parts (All) of the Legend

Every aesthetic that you map inside an aes() function (other than the position aesthetics) gets added to the legend. In our plot, aside from the *x*- and *y*-position aesthetics, we mapped population to size and region to color. Thus, the legend will include both size and color information. We can omit one (or both) of these from the legend by adding a guides() layer to the plot syntax and inside of that layer setting the appropriate aesthetic to FALSE. For example, to show the color part of the legend, but not show the size part of the legend we would use:

```
guides(size = FALSE)
```

If we wanted to omit both color and size from the legend, the syntax would be guides(color = FALSE, size = FALSE)

Your Turn: Omit the Size Guide from the Legend

Use the guides () layer to omit the size information from the legend.

If we remove the size information from the legend, we still should cue readers of our plot into how size is being used in the plot. This might mean changing the title or subtitle of the plot to include this information.

Make the change to your subtitle or title to include information about population size and add a theme to the plot.

Part II

Now you will explore additional country characteristics using the Gapminder website. Go to https://www.gapminder.org/tools/

This brings you to an interactive version of the plot you just created. Click the circle button with a triangle (play button) to see how these variables have changed over time.

1. Describe the relationship between income and life expectancy over the years. (You don't have to discuss population or region.)

Let's use this tool to explore a couple new variables:

- Click on the x-axis label to change it to CO2 emissions per person
- Click on the y-axis to change it to Income
- Drag through the years to get a sense of these relationships over time

2. Describe how the relationship between CO2 emissions per person and Income changes over the years.
3. Describe what you notice about the population over time. Is it the same for all regions?

4. Do CO2 emissions seem to be the same among regions or is there a lot of variability within regions?

5. Draw a DAG for the relationships that you think exist between the variables depicted in this plot for 2017.

Your Turn

Create a new plot in Gapminder using *at least two variables we have not explored* yet. Use the data for a year of your choosing.

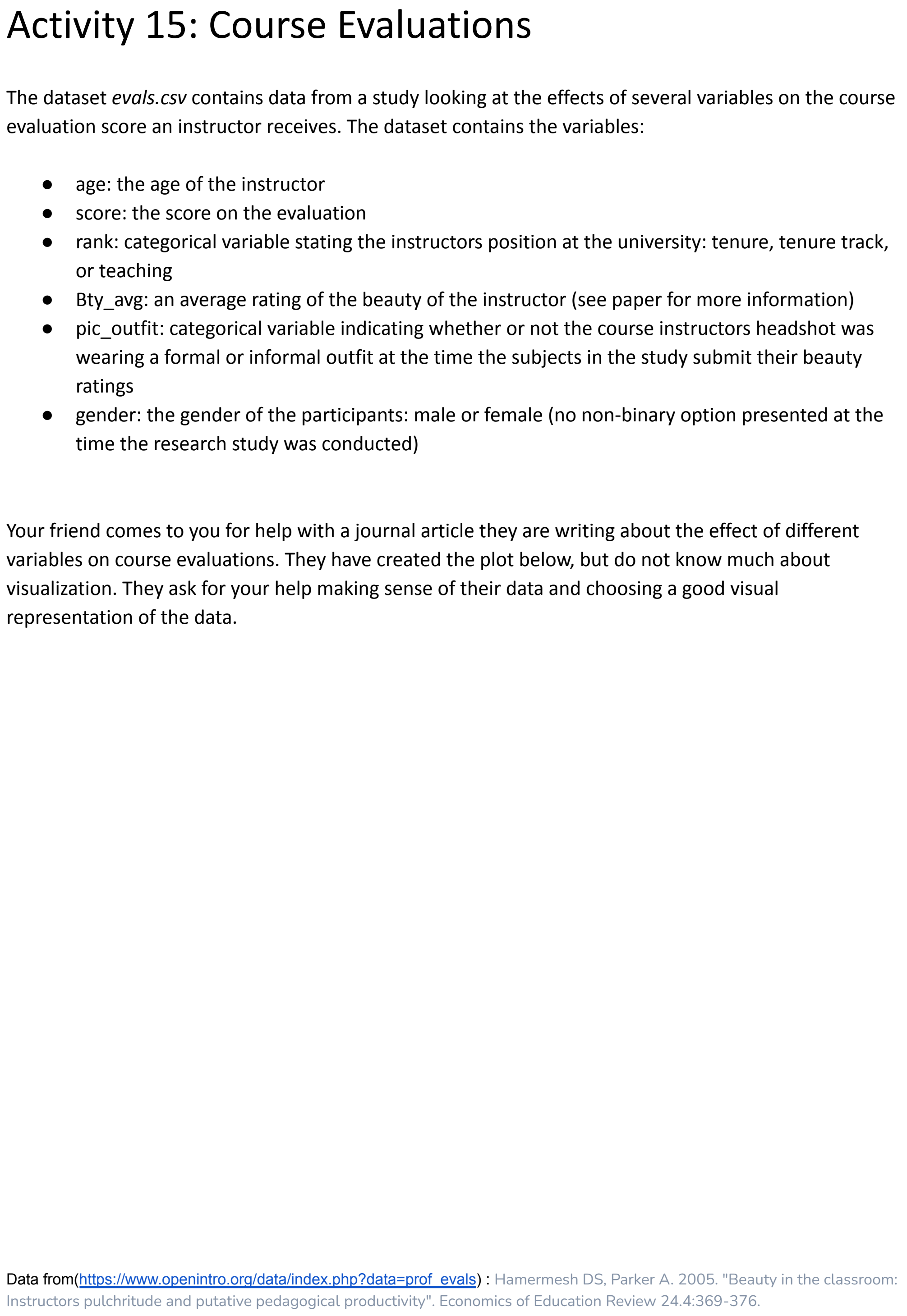
6. Paste a screenshot of your plot below.

7. Create a DAG to represent the relationships that you see among the variables in your plot.

8. Describe your DAG. Be sure to use evidence from your plot to support your reasoning for including each variable and the arrows between them.

Be sure your DAGs are included when you upload your work to Canvas

A.9 Activity 9



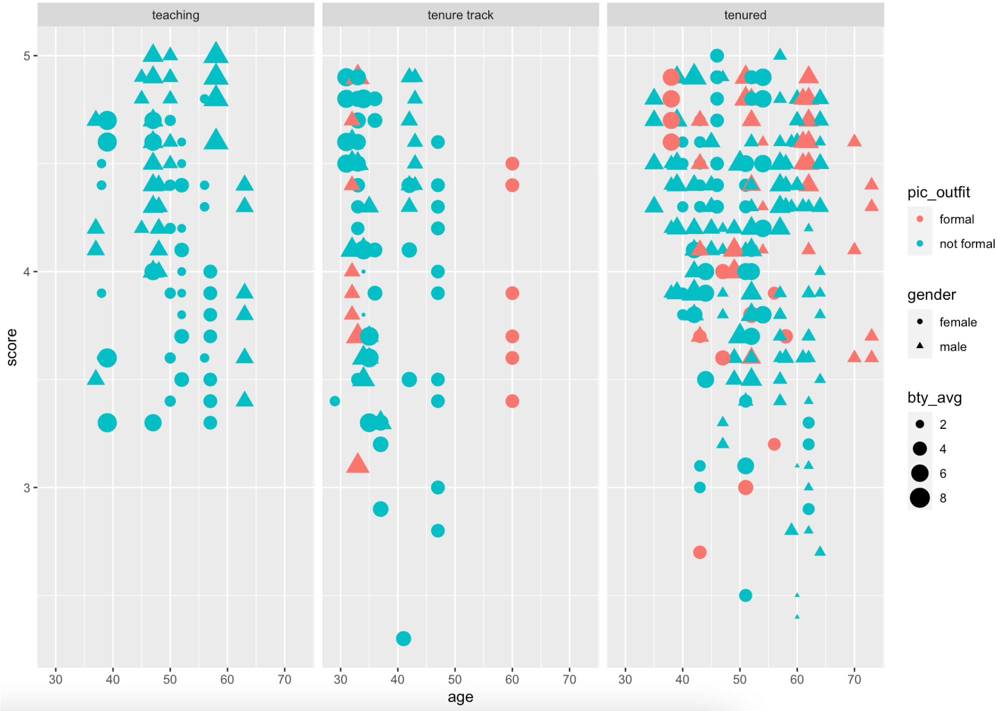
Activity 15: Course Evaluations

The dataset *evals.csv* contains data from a study looking at the effects of several variables on the course evaluation score an instructor receives. The dataset contains the variables:

- age: the age of the instructor
- score: the score on the evaluation
- rank: categorical variable stating the instructors position at the university: tenure, tenure track, or teaching
- Bty_avg: an average rating of the beauty of the instructor (see paper for more information)
- pic_outfit: categorical variable indicating whether or not the course instructors headshot was wearing a formal or informal outfit at the time the subjects in the study submit their beauty ratings
- gender: the gender of the participants: male or female (no non-binary option presented at the time the research study was conducted)

Your friend comes to you for help with a journal article they are writing about the effect of different variables on course evaluations. They have created the plot below, but do not know much about visualization. They ask for your help making sense of their data and choosing a good visual representation of the data.

Data from(https://www.openintro.org/data/index.php?data=prof_evals) : Hamermesh DS, Parker A. 2005. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". Economics of Education Review 24.4:369-376.



1. How many variables are depicted in this plot?

2. What aesthetics are each of the variables mapped to?

3. What are some critiques you have of this plot? Discuss with your partner and write your critiques below.

Data from(https://www.openintro.org/data/index.php?data=prof_evals) : Hamermesh DS, Parker A. 2005. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". Economics of Education Review 24.4:369-376.

Choosing Variables

When graphing variables, it can be helpful to view multiple variables at one time; however, as we saw in the graph above, this can also make the graphs difficult to discern relationships among the variables. Let’s consider which variables we want to feature in a plot.

Three Variables

4. First, consider the relationship among age, score, and rank. Draw a DAG that you think represents the relationships among these variables and paste it below.

5. Create a scatterplot with age on the x-axis, score on the y-axis, and colored by rank.

6. Describe the relationship among the variables as depicted in the plot.

7. Update your DAG if needed to match the relationships depicted in the plot. Paste it below.

Four Variables

8. Next, consider the *gender* variable. Do you think this variable will affect the score the instructor received? Do you think this variable is related to the age or the rank of the instructor? Explain your answer.

9. Draw a DAG to purpose a relationship among age, score, gender, and rank.

10. Add the gender variable to your plot by mapping age on the x-axis, score on the y-axis, coloring by gender, and using facet_wrap(~rank). Paste your plot below.

11. Explain the relationships among the variables that you see in your plot using evidence from the plot. Be sure to expand the size of your plot to make the relationships easier to discern.

12. Draw a DAG to represent the relationships among these four variables.

13. Now let’s, copy and paste the code for your previous plot to create a new plot. In the new plot you will color by bty_avg instead of gender. Paste this plot below.

14. Describe the relationships among these variables using evidence from your plot. Be sure to expand the size of your plot to make the relationships easier to discern.

Data from(https://www.openintro.org/data/index.php?data=prof_evals) : Hamermesh DS, Parker A. 2005. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". Economics of Education Review 24.4:369-376.

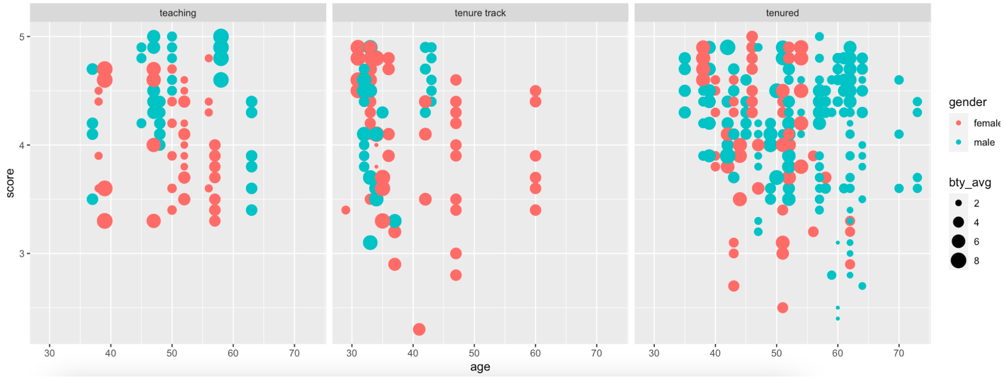
15. Draw a DAG to represent the relationships among these four variables.

Five Variables

16. Consider all five variables we have worked with thus far; score, age, rank, gender, beauty. Draw a DAG to represent the relationships that you think exist among these variables. We haven't explored the relationship between age and beauty average, but make a prediction here.

17. Create a scatterplot to depict all these relationships by using the code below.

```
ggplot(data = evals, aes(x = age, y = score, size = bty_avg, color = gender)) + geom_point() +
 facet_wrap(~rank)
```



18. This plot has become increasingly complicated. Do you think all these variables are needed to fully portray which variables have an effect on the score an instructor receives for a course? Explain which are needed and which are not using evidence from your plots.

19. Create a final DAG for the variables that you think affect the score an instructor receives. Paste it below.

20. Create a final plot to display the relationships among variable you think effect the score an instructor receives. Be sure to include
    - A descriptive title

Data from(https://www.openintro.org/data/index.php?data=prof_evals) : Hamermesh DS, Parker A. 2005. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". Economics of Education Review 24.4:369-376.

- Appropriate axis and legend labels
- A theme and colors that are not the default settings.

21. If we were to redo this study, we would include a more inclusive gender variable scale for instructors to select from. What *other variables* do you think might affect the course evaluation score that an instructor receives other than those in this dataset?

Be sure your DAGs are included when you upload your work to Canvas

Data from(https://www.openintro.org/data/index.php?data=prof_evals) : Hamermesh DS, Parker A. 2005. "Beauty in the classroom: Instructors pulchritude and putative pedagogical productivity". Economics of Education Review 24.4:369-376.

A.10 Activity 10

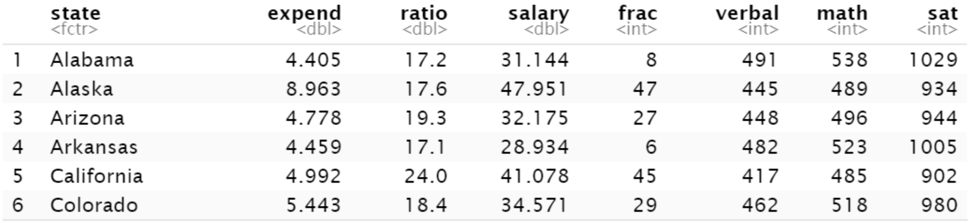
Activity 16: SAT Scores

In this activity, you will explore the *SAT* dataset in the *mosaic* package in R. This dataset contains the average SAT score for each of the 50 states in 1994-95 and other information about education in those states.

You will need to install the *mosaic* package and load the mosaic package with the library function in your notebook

Run the head function on the SAT dataset.

head(SAT)



	state	expend	ratio	salary	frac	verbal	math	sat
	<fctr>	<dbl>	<dbl>	<dbl>	<int>	<int>	<int>	<int>
1	Alabama	4.405	17.2	31.144	8	491	538	1029
2	Alaska	8.963	17.6	47.951	47	445	489	934
3	Arizona	4.778	19.3	32.175	27	448	496	944
4	Arkansas	4.459	17.1	28.934	6	482	523	1005
5	California	4.992	24.0	41.078	45	417	485	902
6	Colorado	5.443	18.4	34.571	29	462	518	980

The variables in this data that we will focus on in this activity are:

- *sat:* the average SAT score
- *expend*: the average expenditure per student in thousands of dollars
- *frac*: the percentage of students taking the SAT

For this activity, we will explore the question:

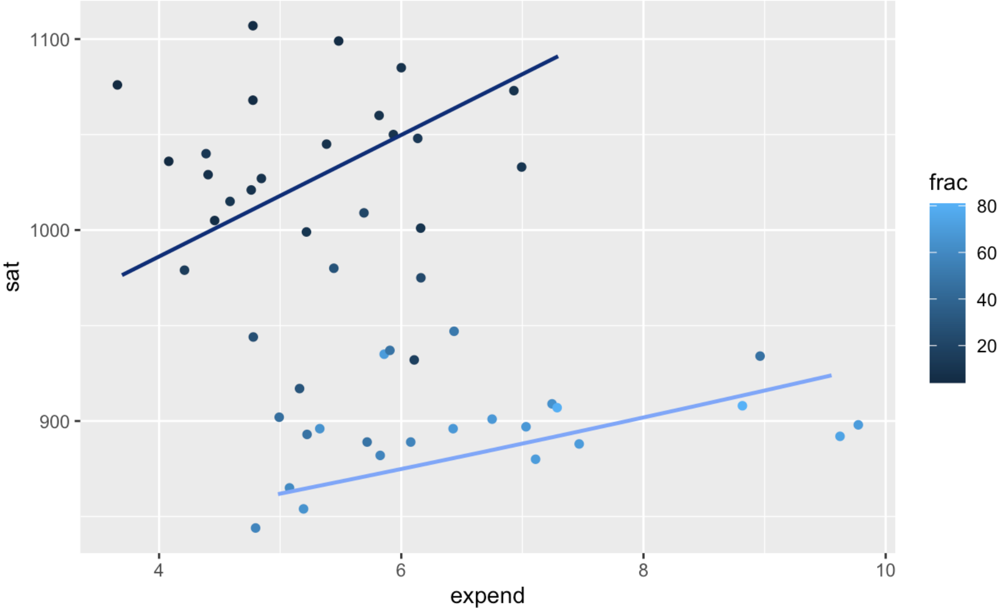
Is increased expenditure associated with higher SAT scores?

1. What do you predict is the answer to the research question?
2. Create a scatterplot with to display the relationship between *expend* and *sat* in the dataset. Be sure that you put *expend* on the x-axis and *sat* on the y-axis. Paste your plot below.
3. Describe the linearity, slope, and strength you see in the plot.
4. Based on your plot is increased expenditure associated with higher SAT scores? Does this match what you expected in Question #1?
5. Why do you suspect you see this relationship between these two variables?

To investigate further, let's look at the percentage of eligible students that took the SAT from each state. This is the number in the variable, *frac.*

6. Add *frac* to your existing plot by coloring the points based on the frac values. Paste your plot below.
7. Based on your plot, does the percentage of students taking the SAT seem to have a relationship with the SAT scores?

If you take a close look at your plot you will notice that it appears there are two groups. Most of the higher percentages of students taking the SAT (lightest colored points) are in the bottom half of the graph while at the top of the graph we see the lower percentages of students taking the SAT (the darker points).



Now it appears that for both lower and higher percentages of students taking the SAT, we see the larger the expenditure the higher the scores. This is called *Simpson's Paradox*. First, we saw a negative association between sat and expenditure, but once we control for another variable (percent of students taking the sat) we see the reverse of our initial trend.

8. Discuss how you think this occurred with your partner. Write a summary of your discussion below.
9. Create a final DAG that models the relationships among the three variables.
10. Provide a final answer to the research question using evidence from your graph.

Part II: Reverse Color Gradient

You must now create a graph that highlights the true relationship between expenditure and SAT score. To start we will reverse the color gradient for the *frac* variable. Typically, lighter colors represent lower numbers and darker colors represent higher numbers. Use *scale_color_gradient2(high = "your color choice")* from the High Peaks activity to put in a new color scale.

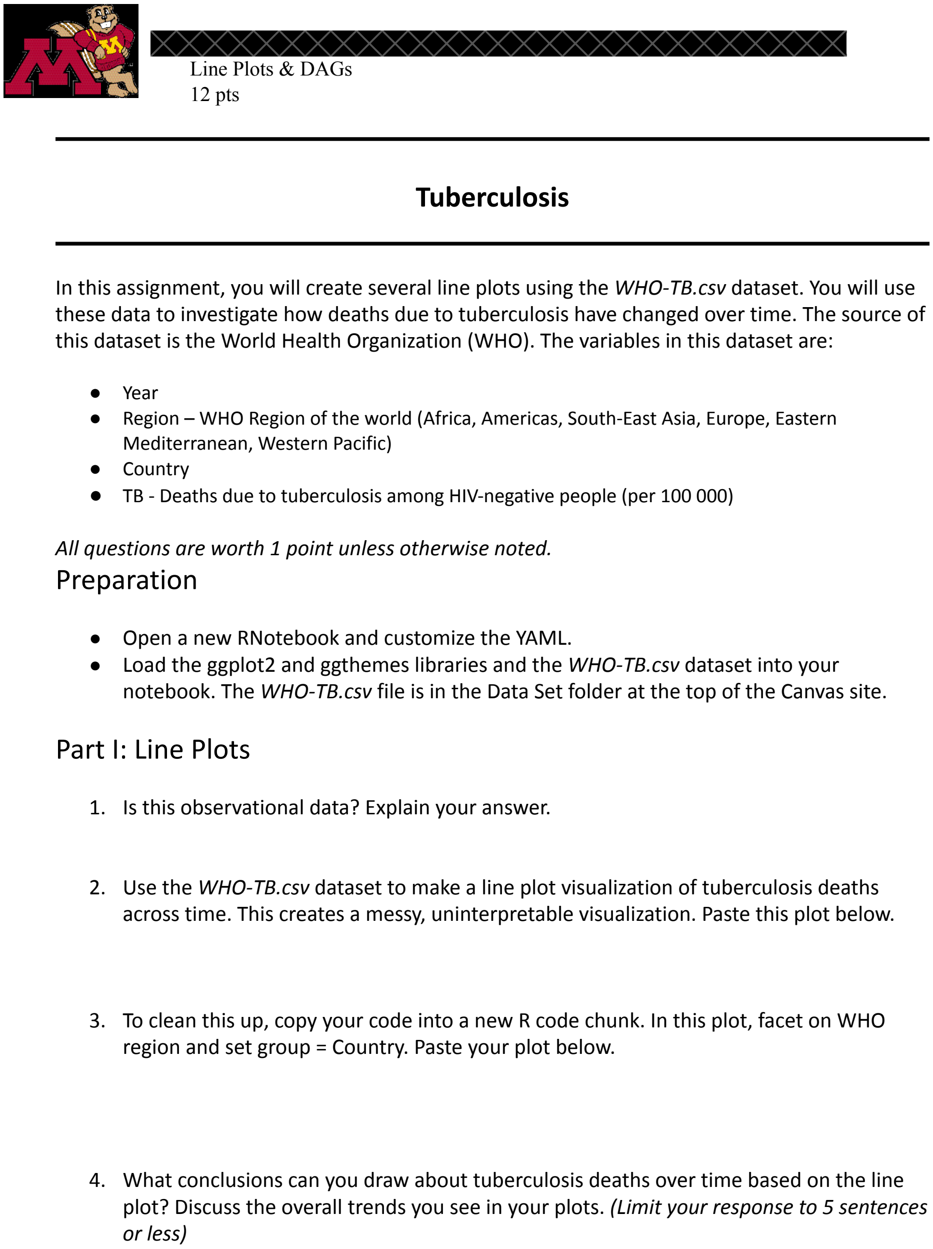
11. Update your plot from Question 6 to use a different color scale that will highlight the difference between the higher and lower values of *frac.*

Another important consideration when creating color scales is to make sure they are accessible for a color blind audience. Save our graph as an image (or take a screenshot) and upload it on this website https://www.color-blindness.com/coblis-color-blindness-simulator/

Then click through the different types of color blindness to determine if your color palette is still discernible for the majority of types of color blindness.

12. Once you have settled on an appropriate palette create your final plot. Be sure to
    - Have a color blind friendly palette
    - Use a theme other than the default
    - Put a thoughtful title on your graph
    - Change the x and y axis label to be more informative
    - Change the legend title to be more descriptive

A.11 Assignment 1



[redacted]

Line Plots & DAGs
12 pts

Tuberculosis

In this assignment, you will create several line plots using the *WHO-TB.csv* dataset. You will use these data to investigate how deaths due to tuberculosis have changed over time. The source of this dataset is the World Health Organization (WHO). The variables in this dataset are:

- Year
- Region – WHO Region of the world (Africa, Americas, South-East Asia, Europe, Eastern Mediterranean, Western Pacific)
- Country
- TB - Deaths due to tuberculosis among HIV-negative people (per 100 000)

All questions are worth 1 point unless otherwise noted.

Preparation

- Open a new RNotebook and customize the YAML.
- Load the ggplot2 and ggthemes libraries and the *WHO-TB.csv* dataset into your notebook. The *WHO-TB.csv* file is in the Data Set folder at the top of the Canvas site.

Part I: Line Plots

1. Is this observational data? Explain your answer.

2. Use the *WHO-TB.csv* dataset to make a line plot visualization of tuberculosis deaths across time. This creates a messy, uninterpretable visualization. Paste this plot below.

3. To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set group = Country. Paste your plot below.

4. What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see in your plots. *(Limit your response to 5 sentences or less)*

[redacted]

Line Plots & DAGs
12 pts

Part II: Closer Inspection

5. Choose a WHO region that has at least one country that has an increasing trend. Create a new dataset for your chosen region based on the *WHO-TB.csv* dataset. Create a line plot for your new dataset colored by country. Be sure to do each of the following to make an aesthetically pleasing plot:
    a. Use a theme other than the default theme
    b. Change the color palette
    c. Add an informative title to your plot
    d. Add a caption telling the source of the data
    e. Remove the label on the x-axis and change the y-axis label to something more informative.

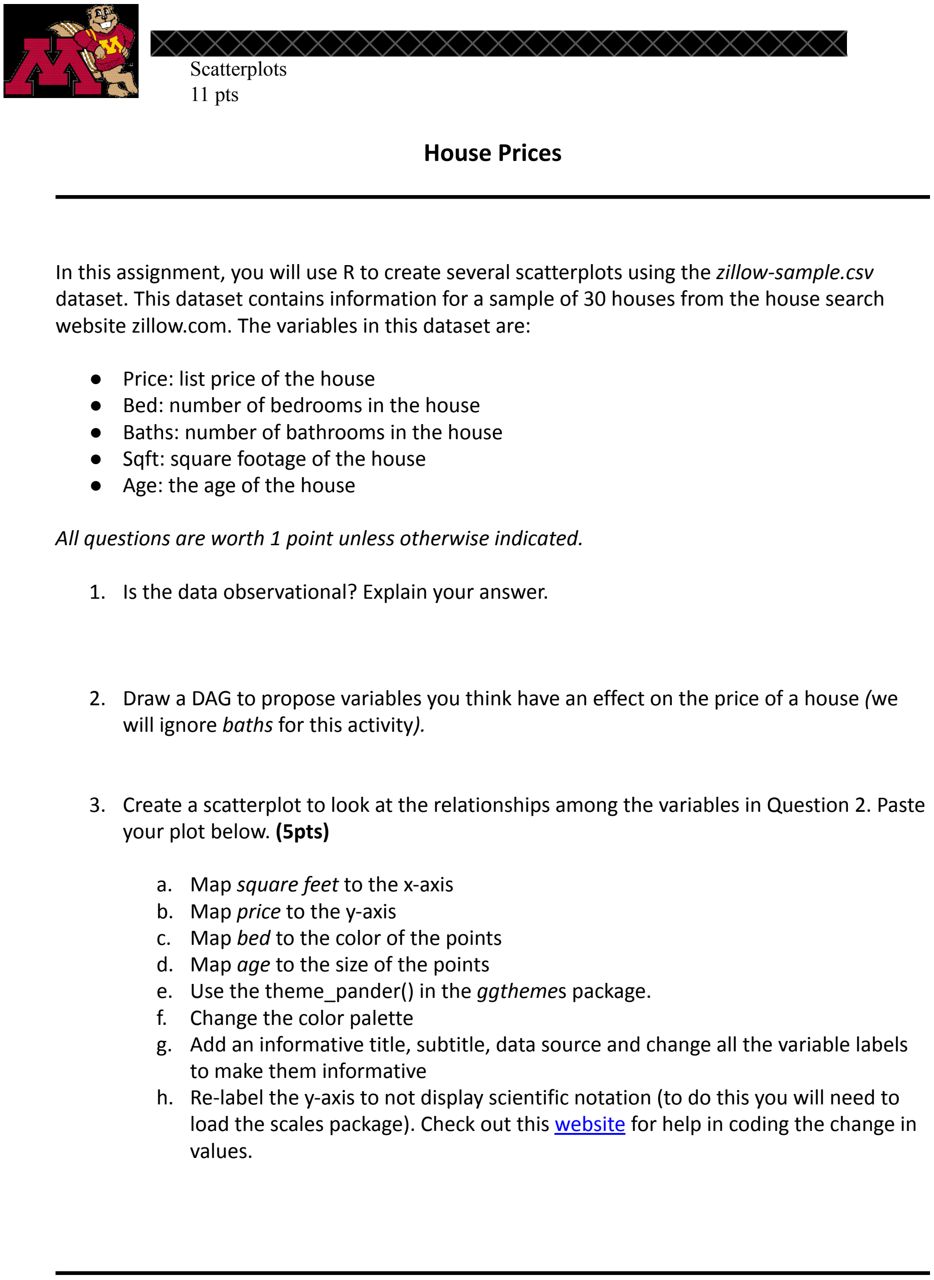
    Paste your final plot below. **(4 pts)**

6. What conclusions can you make about tuberculosis deaths in your chosen region over time using evidence from your plot?

7. We see changes in tuberculosis death rates over time in some regions, but we might wonder what is causing these changes. Draw a DAG to incorporate two or three *variables* that you think might be associated with the tuberculosis death rate in a country. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.

8. Provide a justification for your drawing in Question #7 that explains your proposed relationships among the variables you chose.

A.12 Assignment 2



[redacted]

Scatterplots
11 pts

House Prices

In this assignment, you will use R to create several scatterplots using the *zillow-sample.csv* dataset. This dataset contains information for a sample of 30 houses from the house search website zillow.com. The variables in this dataset are:

- Price: list price of the house
- Bed: number of bedrooms in the house
- Baths: number of bathrooms in the house
- Sqft: square footage of the house
- Age: the age of the house

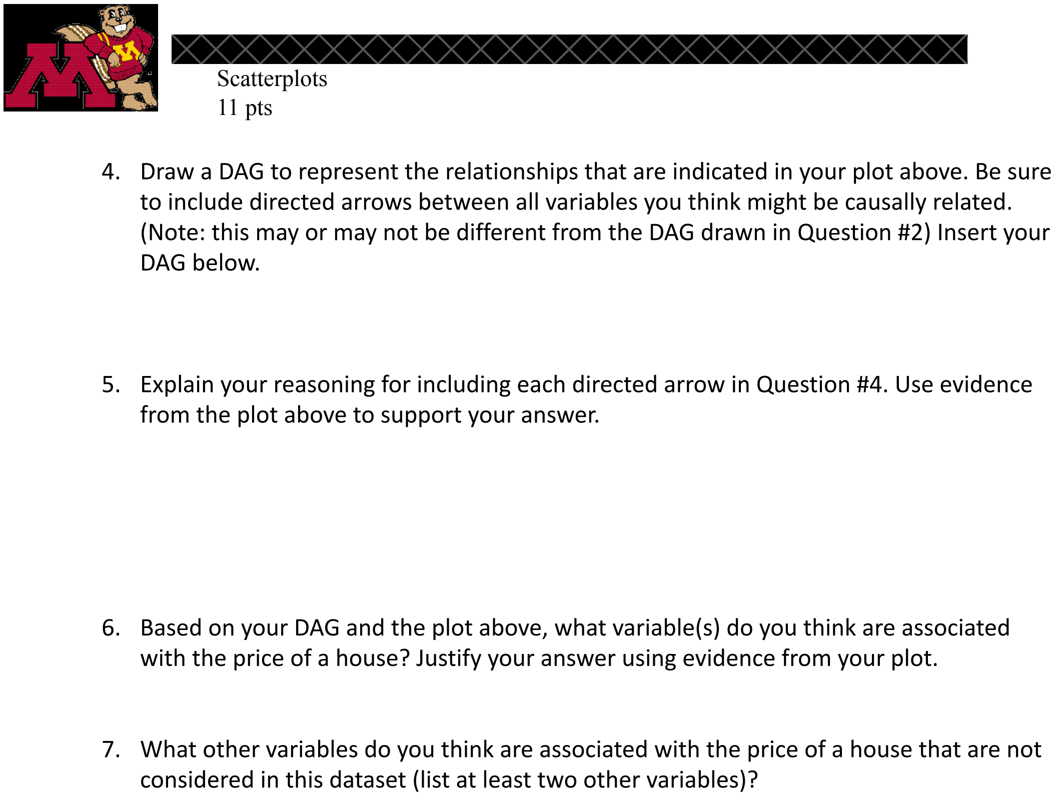
All questions are worth 1 point unless otherwise indicated.

1. Is the data observational? Explain your answer.

2. Draw a DAG to propose variables you think have an effect on the price of a house *(*we will ignore *baths* for this activity*).*

3. Create a scatterplot to look at the relationships among the variables in Question 2. Paste your plot below. **(5pts)**

   a. Map *square feet* to the x-axis
   b. Map *price* to the y-axis
   c. Map *bed* to the color of the points
   d. Map *age* to the size of the points
   e. Use the theme_pander() in the *ggtheme*s package.
   f. Change the color palette
   g. Add an informative title, subtitle, data source and change all the variable labels to make them informative
   h. Re-label the y-axis to not display scientific notation (to do this you will need to load the scales package). Check out this website for help in coding the change in values.



[redacted]

Scatterplots
11 pts

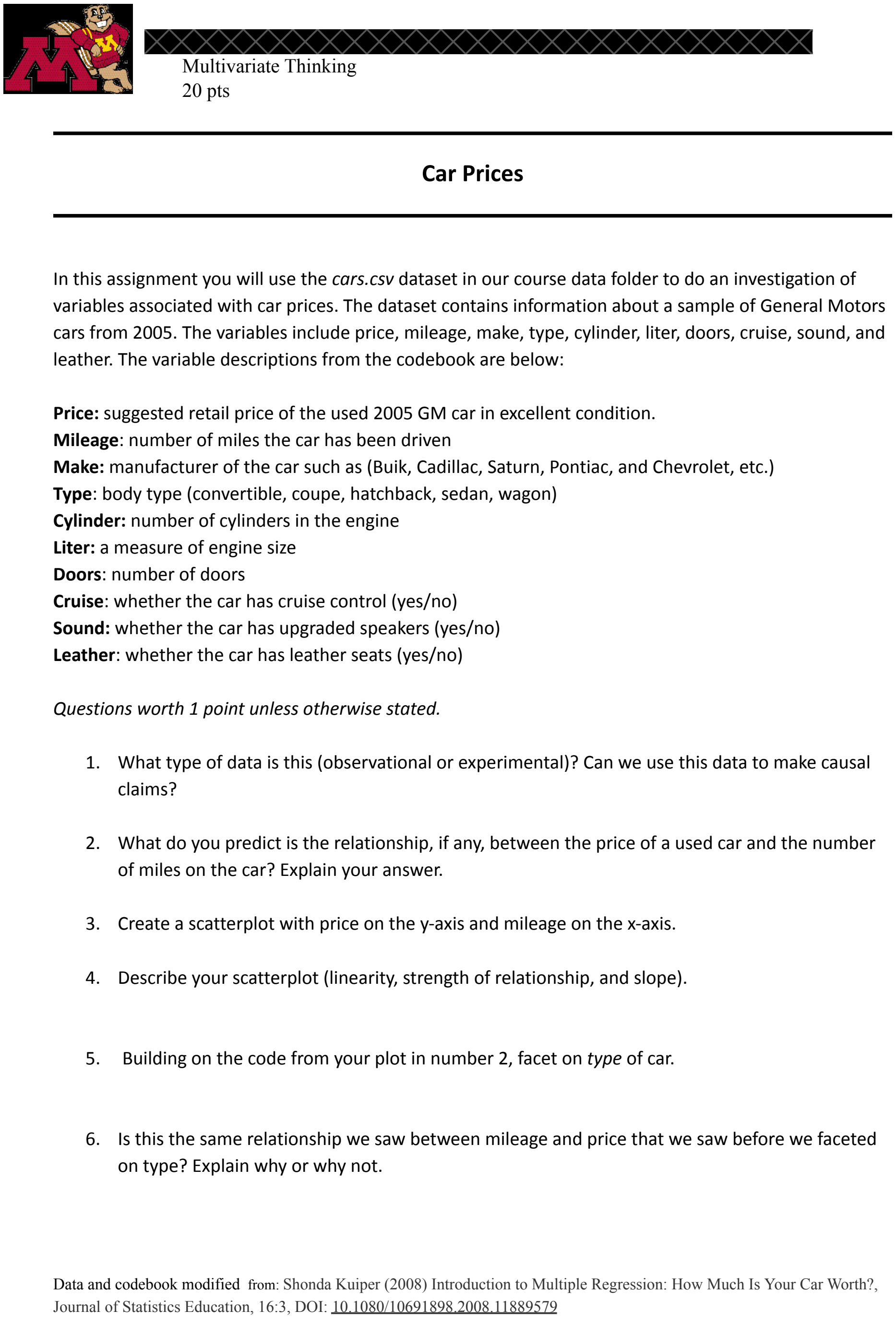
4. Draw a DAG to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related. (Note: this may or may not be different from the DAG drawn in Question #2) Insert your DAG below.

5. Explain your reasoning for including each directed arrow in Question #4. Use evidence from the plot above to support your answer.

6. Based on your DAG and the plot above, what variable(s) do you think are associated with the price of a house? Justify your answer using evidence from your plot.

7. What other variables do you think are associated with the price of a house that are not considered in this dataset (list at least two other variables)?

A.13 Assignment 3



[redacted]

Multivariate Thinking
20 pts

Car Prices

In this assignment you will use the *cars.csv* dataset in our course data folder to do an investigation of variables associated with car prices. The dataset contains information about a sample of General Motors cars from 2005. The variables include price, mileage, make, type, cylinder, liter, doors, cruise, sound, and leather. The variable descriptions from the codebook are below:

Price: suggested retail price of the used 2005 GM car in excellent condition.
Mileage: number of miles the car has been driven
Make: manufacturer of the car such as (Buik, Cadillac, Saturn, Pontiac, and Chevrolet, etc.)
Type: body type (convertible, coupe, hatchback, sedan, wagon)
Cylinder: number of cylinders in the engine
Liter: a measure of engine size
Doors: number of doors
Cruise: whether the car has cruise control (yes/no)
Sound: whether the car has upgraded speakers (yes/no)
Leather: whether the car has leather seats (yes/no)

Questions worth 1 point unless otherwise stated.

1. What type of data is this (observational or experimental)? Can we use this data to make causal claims?
2. What do you predict is the relationship, if any, between the price of a used car and the number of miles on the car? Explain your answer.
3. Create a scatterplot with price on the y-axis and mileage on the x-axis.
4. Describe your scatterplot (linearity, strength of relationship, and slope).
5. Building on the code from your plot in number 2, facet on *type* of car.
6. Is this the same relationship we saw between mileage and price that we saw before we faceted on type? Explain why or why not.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
20 pts

7. Choose another variable from the list of variables in the dataset. Create a DAG to propose relationships among the four variables (3 possible causal variables and price). Paste it below.

8. Explain how you think the variable chosen in Question #6 might (or might not) affect the price of the car, type of car, and mileage.

9. Create a plot incorporating the four variables. **(6pts)**
   a. Put a title and subtitle on the plot
   b. Use a theme other than the default theme
   c. Use a colorblind friendly color palette
   d. Add a caption with the source of the data to the plot
   e. Make the values on the x-axis readable by rotating the numbers
   f. If you have a legend, give the legend a meaningful title
   g. Remove the na's

10. Draw your final DAG to represent how your three variables affect price and each other. Paste it below.

11. Provide a justification for your DAG using your plot as evidence.

12. Create a new plot using your same 4 variables (map different variables to the x axes, colors, size, or facet on a different variable), but keep price mapped to the y-axis. Be sure to update the title, colors, and theme as well.

13. Do you think your plot in Question #9 or Question #12 depicts the relationships among the variables better? Explain your answer.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
20 pts

14. Are there any other variables (not in the current data set) that you think might affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars? Choose one or two other variables and explain your answer. *(Limit your response to 5 sentences or less)*

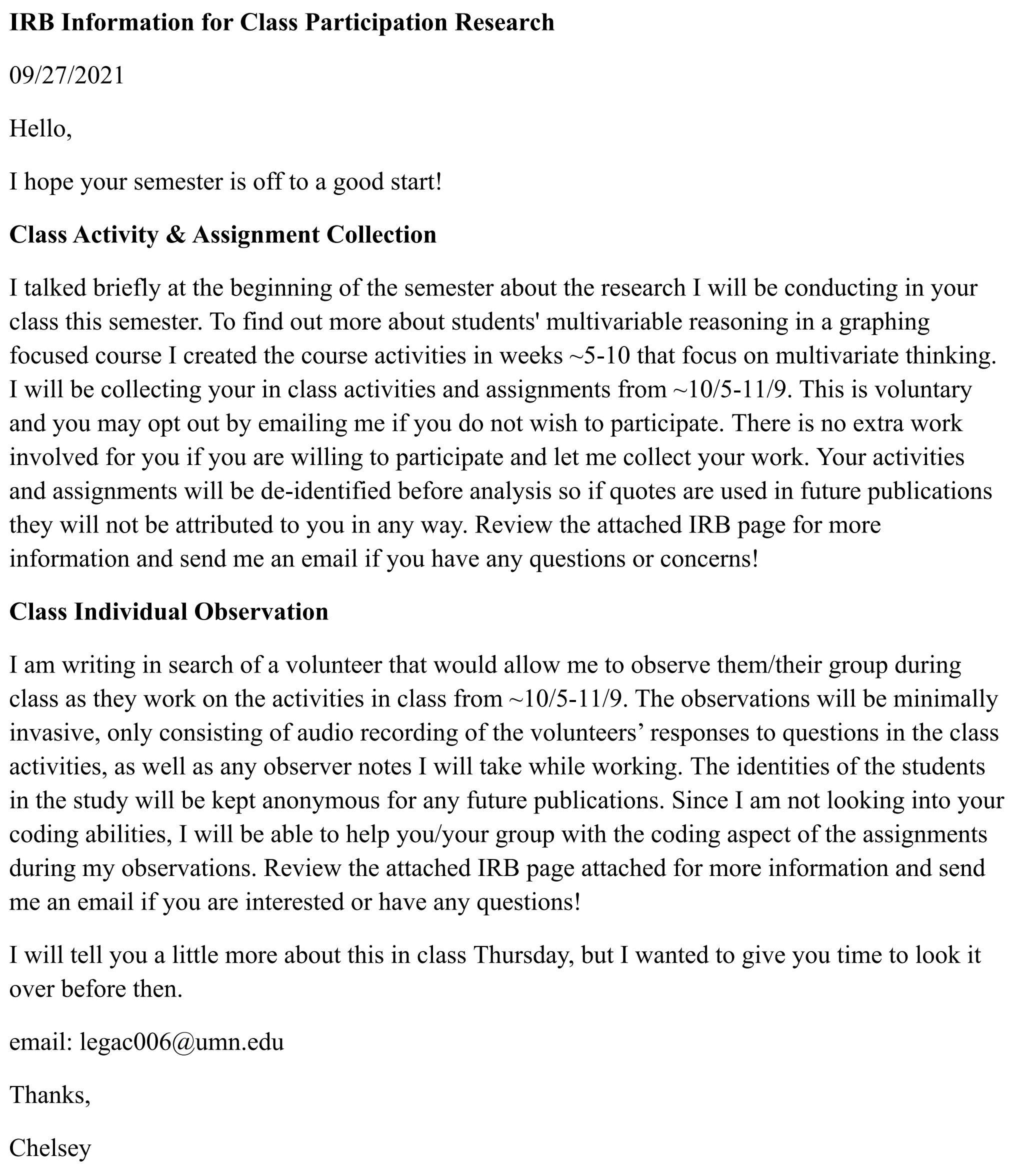
15. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions (based on your plot) would you recommend they buy it? Explain your answer below using evidence from your final plot.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

Appendix B

Appendix B

B.1 Class Email Announcements



IRB Information for Class Participation Research

09/27/2021

Hello,

I hope your semester is off to a good start!

Class Activity & Assignment Collection

I talked briefly at the beginning of the semester about the research I will be conducting in your class this semester. To find out more about students' multivariable reasoning in a graphing focused course I created the course activities in weeks ~5-10 that focus on multivariate thinking. I will be collecting your in class activities and assignments from ~10/5-11/9. This is voluntary and you may opt out by emailing me if you do not wish to participate. There is no extra work involved for you if you are willing to participate and let me collect your work. Your activities and assignments will be de-identified before analysis so if quotes are used in future publications they will not be attributed to you in any way. Review the attached IRB page for more information and send me an email if you have any questions or concerns!

Class Individual Observation

I am writing in search of a volunteer that would allow me to observe them/their group during class as they work on the activities in class from ~10/5-11/9. The observations will be minimally invasive, only consisting of audio recording of the volunteers' responses to questions in the class activities, as well as any observer notes I will take while working. The identities of the students in the study will be kept anonymous for any future publications. Since I am not looking into your coding abilities, I will be able to help you/your group with the coding aspect of the assignments during my observations. Review the attached IRB page attached for more information and send me an email if you are interested or have any questions!

I will tell you a little more about this in class Thursday, but I wanted to give you time to look it over before then.

email: legac006@umn.edu

Thanks,

Chelsey

Activity 10 Submission
10/19/2021

Hello,

Thank you to those that could make it to class in person or via zoom today! You did a great job embracing the DAGs!

For those that couldn't make it be sure to look through the activity, as there will be questions about DAGs in future class work.

Also, if you haven't done so already you will need to create a DAG (in a context of your choosing - with at least 3 variables), take a picture of it, and put it into an R markdown to practice doing this for future assignments. Then you can submit your .pdf file from the markdown you created to get credit for this activity.

Let me know if you have any questions about this!

Chelsey

Comments about using evidence from your plot to create your DAG
11/03/2021

Hello,

I wanted to send a quick email with some more information about drawing the DAGs. I thought it might be helpful to provide an example of the types of answers I'm looking for when asking you to justify your answers with evidence from the plot.

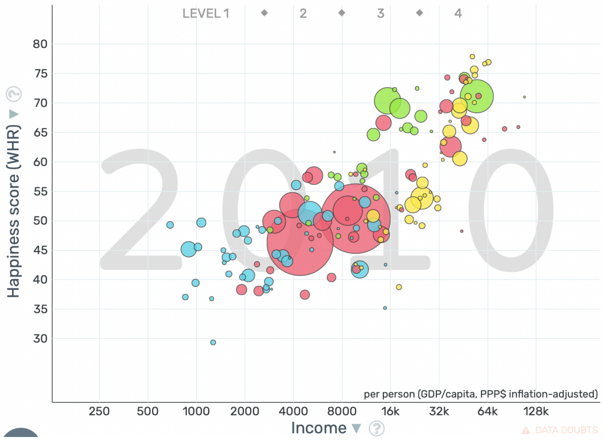
Here is my attempt at the last part of the World Activity Part 2. Answers will vary based on variables you chose, and even if you chose the same variables I would expect us to possibly interpret the plot differently and write different DAGs!

Variables: income, happiness score, region and population

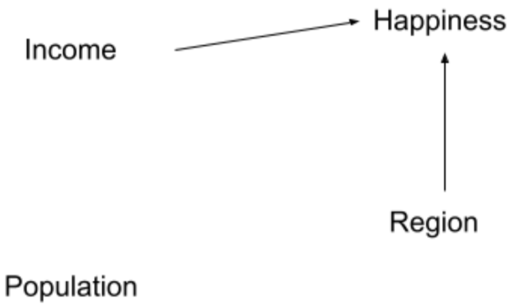
Case: Country

Observational data - so we can't make a causal claim about any of the variables affecting change in others because we don't have an experiment.

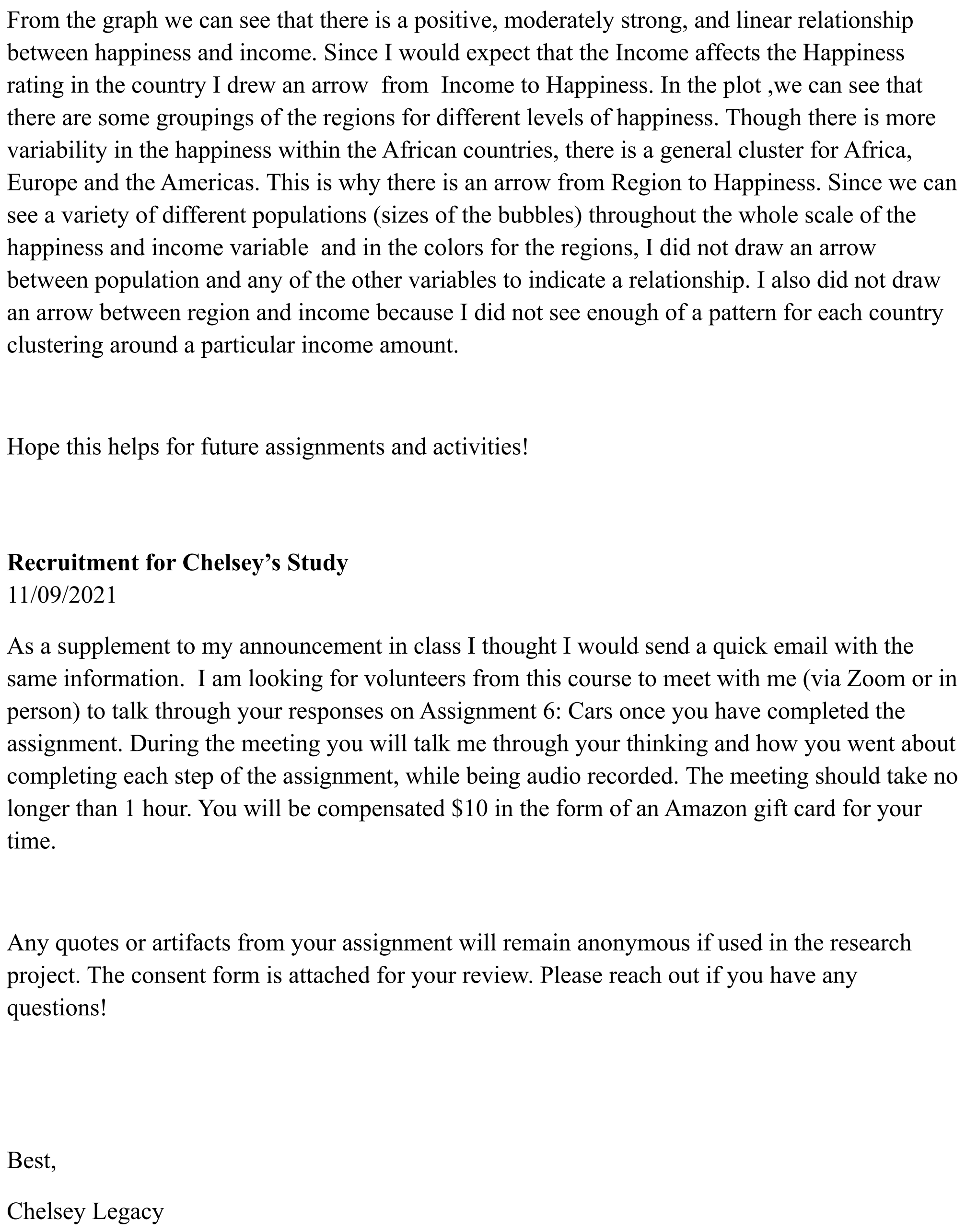
- **Paste a screenshot of your plot below.**



- **Create a DAG to represent the relationships that you see among the variables in your plot.**



- **Describe your DAG. Be sure to use evidence from your plot to support your reasoning for including each variable and the arrows between them.**



From the graph we can see that there is a positive, moderately strong, and linear relationship between happiness and income. Since I would expect that the Income affects the Happiness rating in the country I drew an arrow from Income to Happiness. In the plot ,we can see that there are some groupings of the regions for different levels of happiness. Though there is more variability in the happiness within the African countries, there is a general cluster for Africa, Europe and the Americas. This is why there is an arrow from Region to Happiness. Since we can see a variety of different populations (sizes of the bubbles) throughout the whole scale of the happiness and income variable and in the colors for the regions, I did not draw an arrow between population and any of the other variables to indicate a relationship. I also did not draw an arrow between region and income because I did not see enough of a pattern for each country clustering around a particular income amount.

Hope this helps for future assignments and activities!

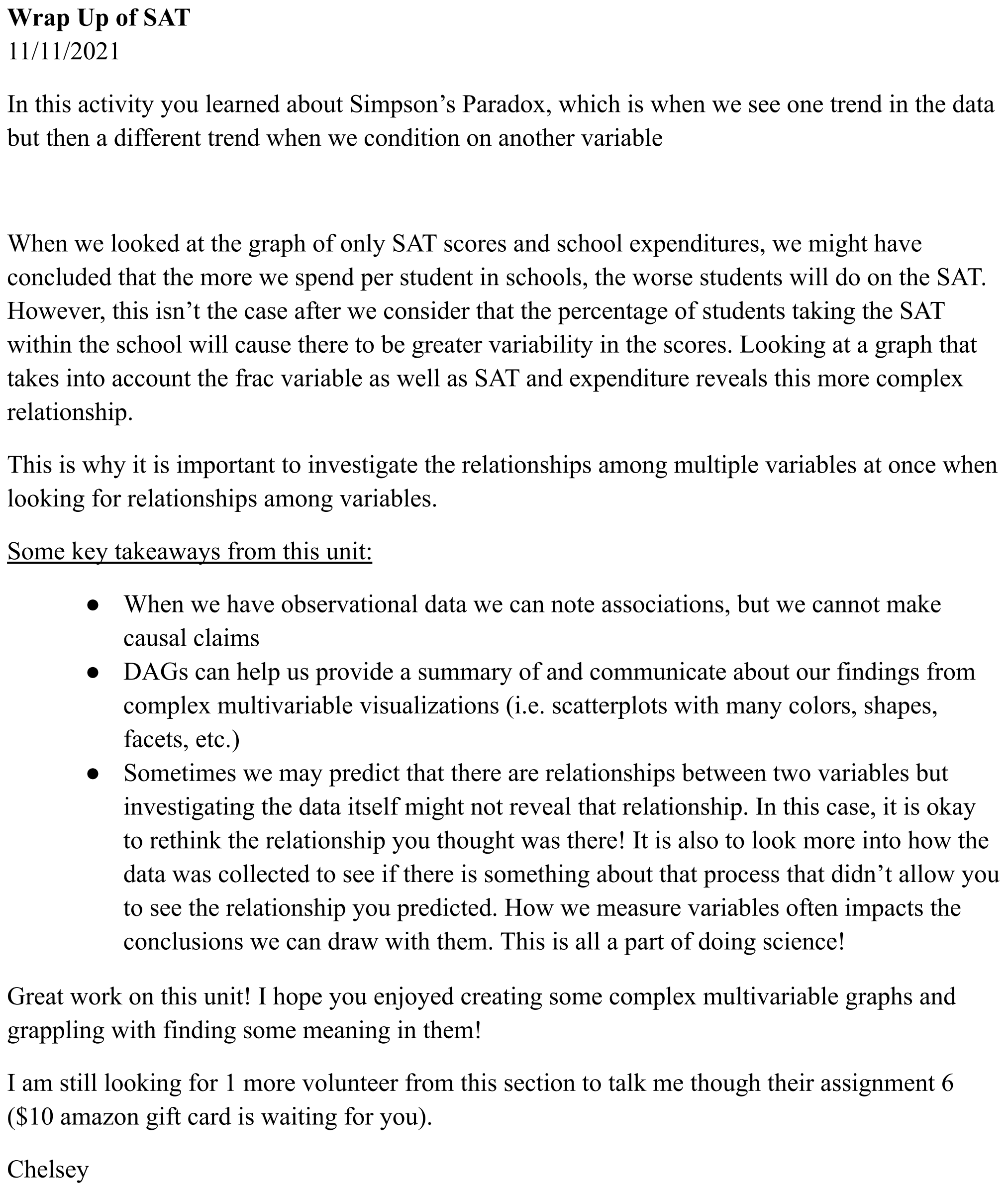
Recruitment for Chelsey's Study
11/09/2021

As a supplement to my announcement in class I thought I would send a quick email with the same information. I am looking for volunteers from this course to meet with me (via Zoom or in person) to talk through your responses on Assignment 6: Cars once you have completed the assignment. During the meeting you will talk me through your thinking and how you went about completing each step of the assignment, while being audio recorded. The meeting should take no longer than 1 hour. You will be compensated $10 in the form of an Amazon gift card for your time.

Any quotes or artifacts from your assignment will remain anonymous if used in the research project. The consent form is attached for your review. Please reach out if you have any questions!

Best,

Chelsey Legacy



Wrap Up of SAT
11/11/2021

In this activity you learned about Simpson's Paradox, which is when we see one trend in the data but then a different trend when we condition on another variable

When we looked at the graph of only SAT scores and school expenditures, we might have concluded that the more we spend per student in schools, the worse students will do on the SAT. However, this isn't the case after we consider that the percentage of students taking the SAT within the school will cause there to be greater variability in the scores. Looking at a graph that takes into account the frac variable as well as SAT and expenditure reveals this more complex relationship.

This is why it is important to investigate the relationships among multiple variables at once when looking for relationships among variables.

Some key takeaways from this unit:

- When we have observational data we can note associations, but we cannot make causal claims
- DAGs can help us provide a summary of and communicate about our findings from complex multivariable visualizations (i.e. scatterplots with many colors, shapes, facets, etc.)
- Sometimes we may predict that there are relationships between two variables but investigating the data itself might not reveal that relationship. In this case, it is okay to rethink the relationship you thought was there! It is also to look more into how the data was collected to see if there is something about that process that didn't allow you to see the relationship you predicted. How we measure variables often impacts the conclusions we can draw with them. This is all a part of doing science!

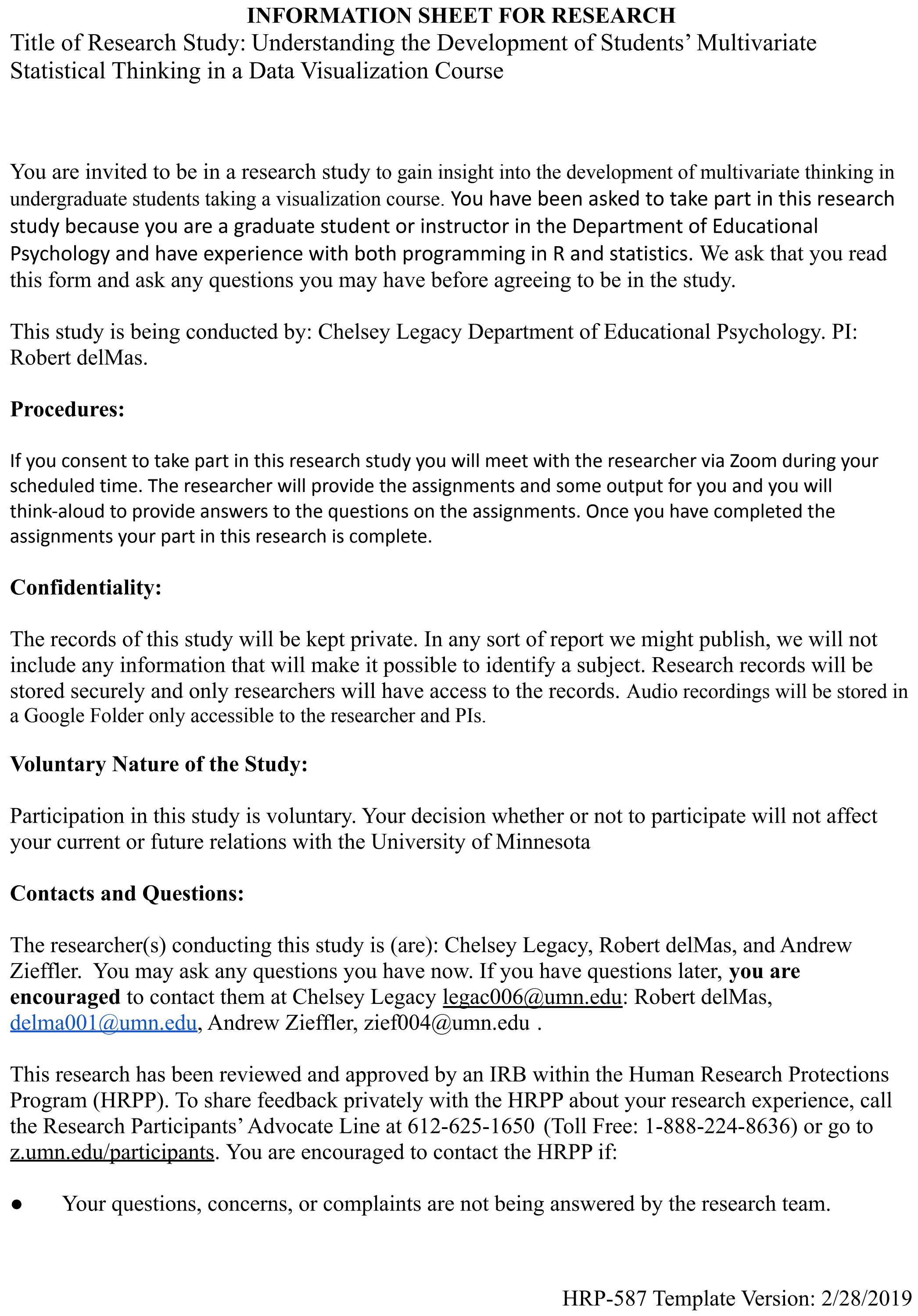
Great work on this unit! I hope you enjoyed creating some complex multivariable graphs and grappling with finding some meaning in them!

I am still looking for 1 more volunteer from this section to talk me though their assignment 6 ($10 amazon gift card is waiting for you).

Chelsey

B.2 IRB Consent

B.2.1 IRB for Initial Think-Aloud Consent



INFORMATION SHEET FOR RESEARCH

Title of Research Study: Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

You are invited to be in a research study to gain insight into the development of multivariate thinking in undergraduate students taking a visualization course. You have been asked to take part in this research study because you are a graduate student or instructor in the Department of Educational Psychology and have experience with both programming in R and statistics. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Chelsey Legacy Department of Educational Psychology. PI: Robert delMas.

Procedures:

If you consent to take part in this research study you will meet with the researcher via Zoom during your scheduled time. The researcher will provide the assignments and some output for you and you will think-aloud to provide answers to the questions on the assignments. Once you have completed the assignments your part in this research is complete.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Audio recordings will be stored in a Google Folder only accessible to the researcher and PIs.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota

Contacts and Questions:

The researcher(s) conducting this study is (are): Chelsey Legacy, Robert delMas, and Andrew Zieffler. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at Chelsey Legacy legac006@umn.edu: Robert delMas, delma001@umn.edu, Andrew Zieffler, zief004@umn.edu .

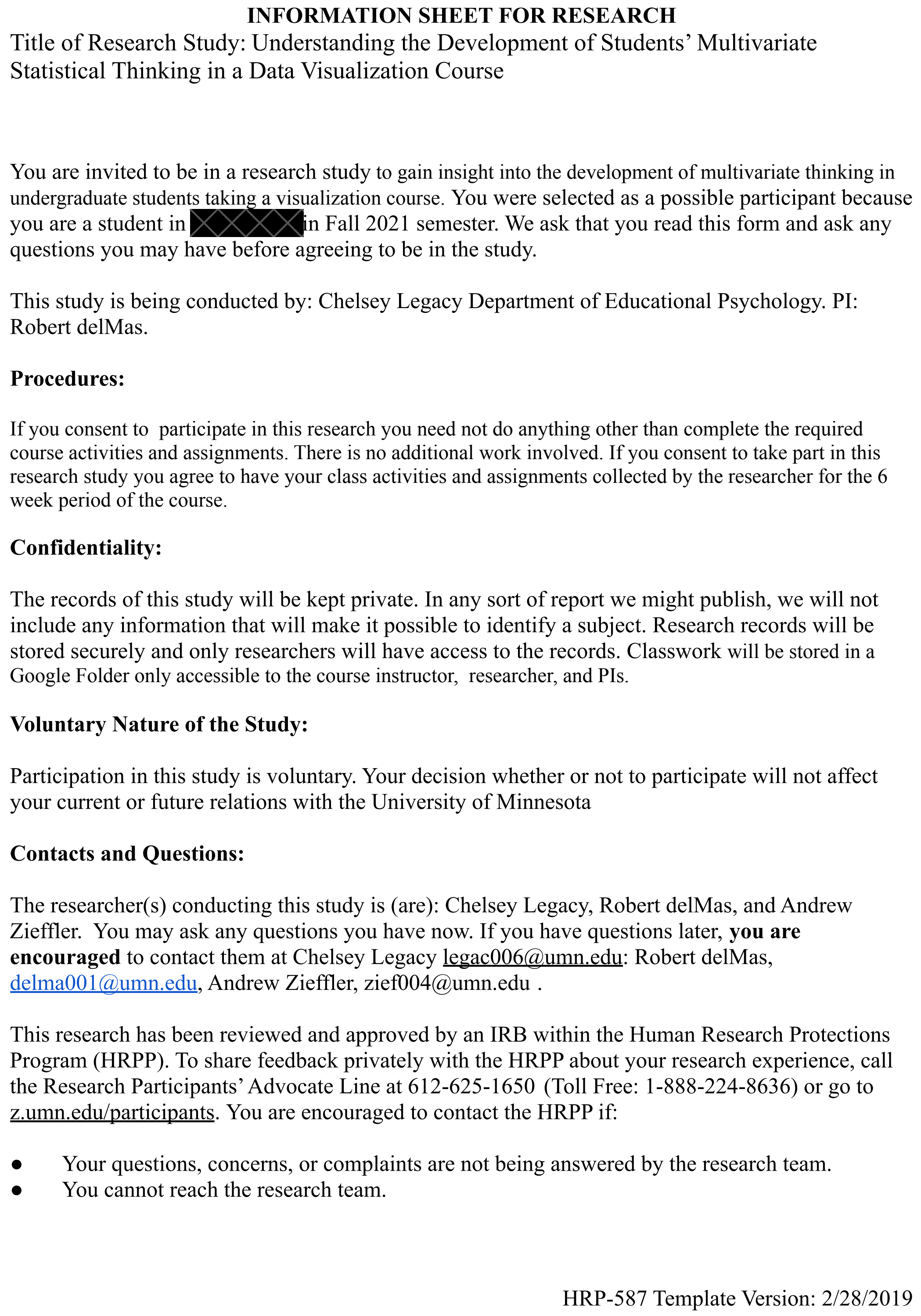
This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.

- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be given a copy of this information to keep for your records.

B.2.1.1 IRB for Class Data Collection Consent



INFORMATION SHEET FOR RESEARCH

Title of Research Study: Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

You are invited to be in a research study to gain insight into the development of multivariate thinking in undergraduate students taking a visualization course. You were selected as a possible participant because you are a student in [redacted] in Fall 2021 semester. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Chelsey Legacy Department of Educational Psychology. PI: Robert delMas.

Procedures:

If you consent to participate in this research you need not do anything other than complete the required course activities and assignments. There is no additional work involved. If you consent to take part in this research study you agree to have your class activities and assignments collected by the researcher for the 6 week period of the course.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Classwork will be stored in a Google Folder only accessible to the course instructor, researcher, and PIs.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota

Contacts and Questions:

The researcher(s) conducting this study is (are): Chelsey Legacy, Robert delMas, and Andrew Zieffler. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at Chelsey Legacy legac006@umn.edu: Robert delMas, delma001@umn.edu, Andrew Zieffler, zief004@umn.edu .

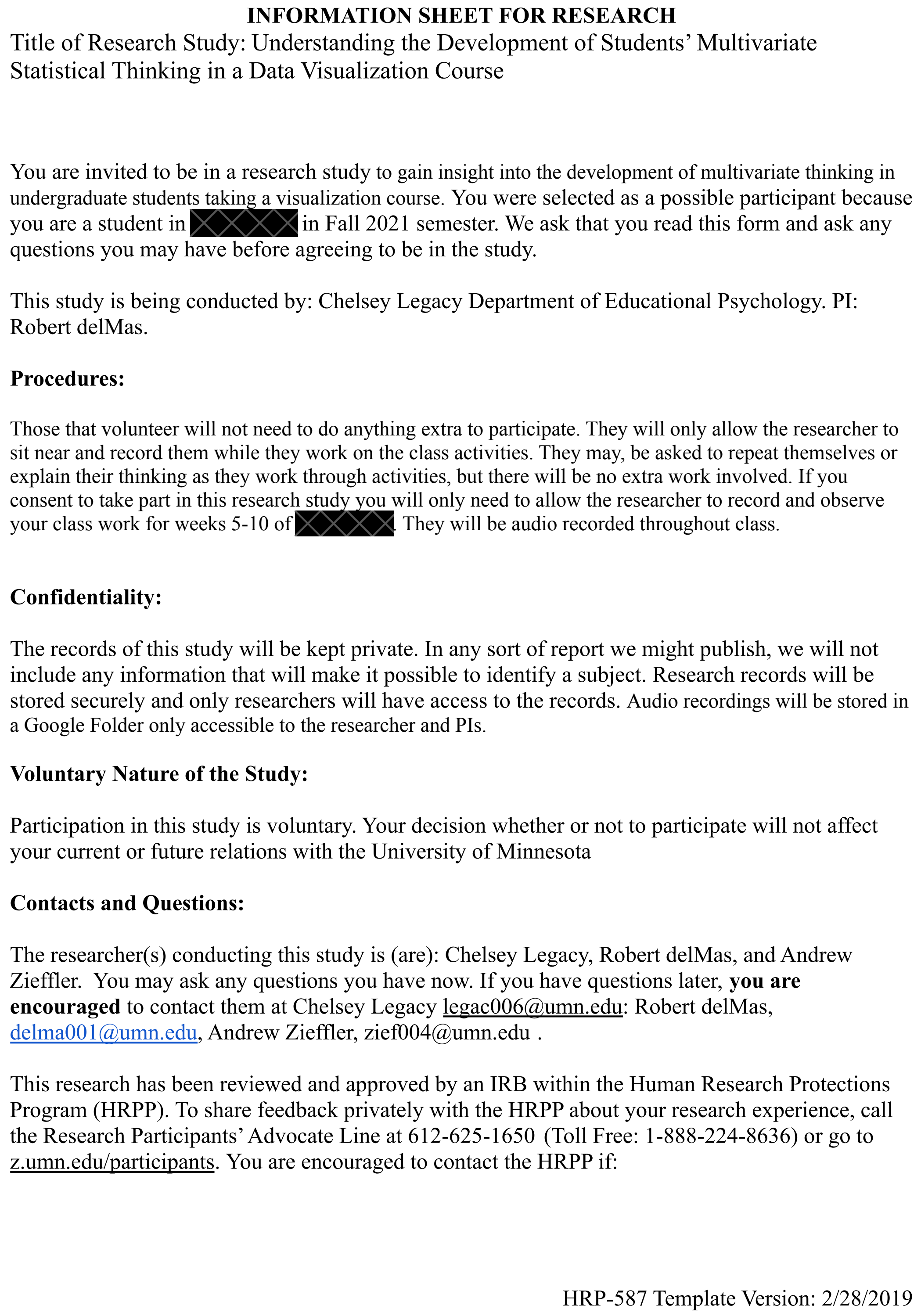
This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.

- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be given a copy of this information to keep for your records.

B.2.2 IRB for Class Observation Consent



INFORMATION SHEET FOR RESEARCH

Title of Research Study: Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

You are invited to be in a research study to gain insight into the development of multivariate thinking in undergraduate students taking a visualization course. You were selected as a possible participant because you are a student in [redacted] in Fall 2021 semester. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Chelsey Legacy Department of Educational Psychology. PI: Robert delMas.

Procedures:

Those that volunteer will not need to do anything extra to participate. They will only allow the researcher to sit near and record them while they work on the class activities. They may, be asked to repeat themselves or explain their thinking as they work through activities, but there will be no extra work involved. If you consent to take part in this research study you will only need to allow the researcher to record and observe your class work for weeks 5-10 of [redacted]. They will be audio recorded throughout class.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Audio recordings will be stored in a Google Folder only accessible to the researcher and PIs.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota

Contacts and Questions:

The researcher(s) conducting this study is (are): Chelsey Legacy, Robert delMas, and Andrew Zieffler. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at Chelsey Legacy legac006@umn.edu: Robert delMas, delma001@umn.edu, Andrew Zieffler, zief004@umn.edu .

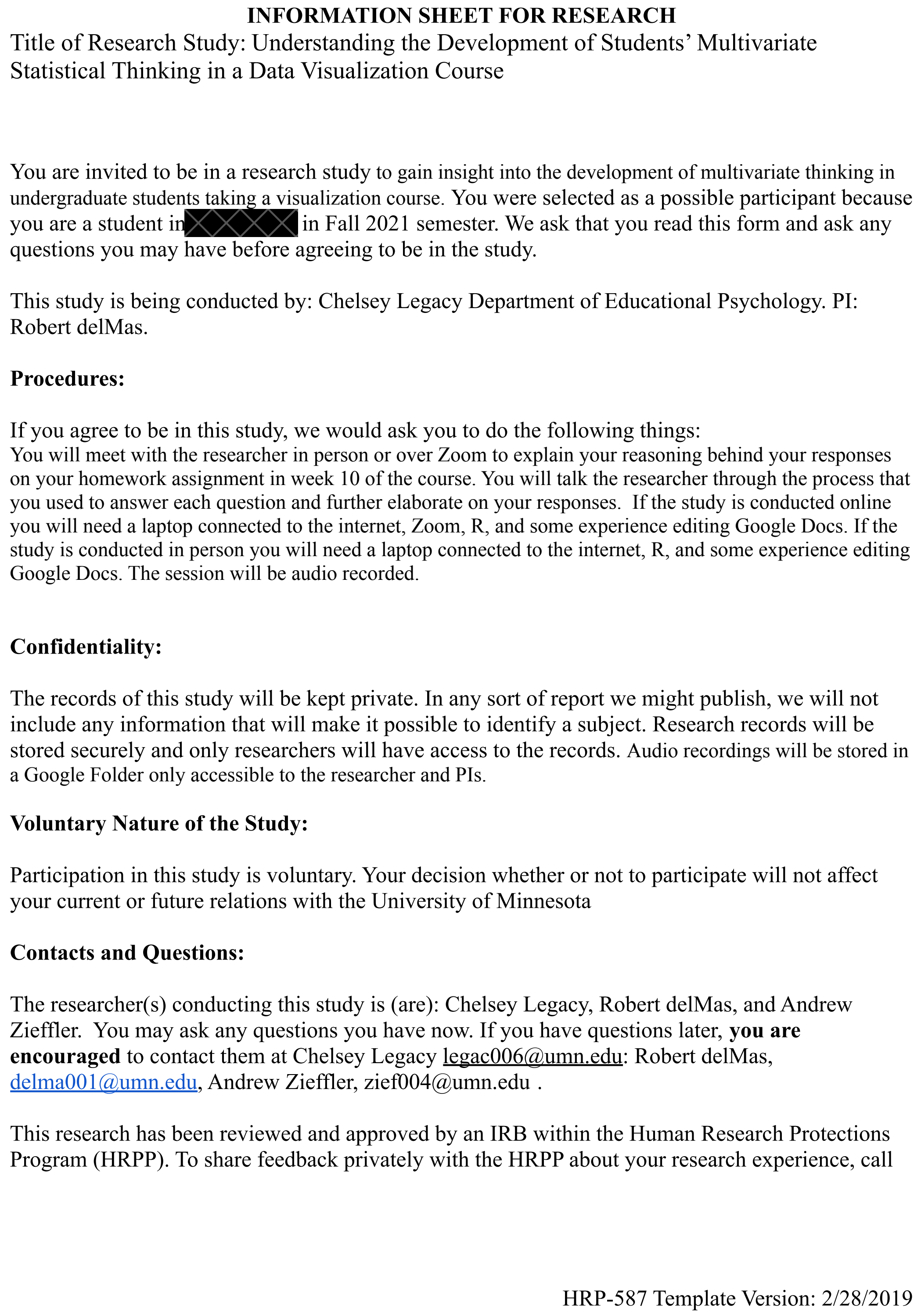
This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

HRP-587 Template Version: 2/28/2019

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be given a copy of this information to keep for your records.

B.3 IRB for Final Think-Aloud Consent



INFORMATION SHEET FOR RESEARCH

Title of Research Study: Understanding the Development of Students' Multivariate Statistical Thinking in a Data Visualization Course

You are invited to be in a research study to gain insight into the development of multivariate thinking in undergraduate students taking a visualization course. You were selected as a possible participant because you are a student in[redacted] in Fall 2021 semester. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by: Chelsey Legacy Department of Educational Psychology. PI: Robert delMas.

Procedures:

If you agree to be in this study, we would ask you to do the following things:
You will meet with the researcher in person or over Zoom to explain your reasoning behind your responses on your homework assignment in week 10 of the course. You will talk the researcher through the process that you used to answer each question and further elaborate on your responses. If the study is conducted online you will need a laptop connected to the internet, Zoom, R, and some experience editing Google Docs. If the study is conducted in person you will need a laptop connected to the internet, R, and some experience editing Google Docs. The session will be audio recorded.

Confidentiality:

The records of this study will be kept private. In any sort of report we might publish, we will not include any information that will make it possible to identify a subject. Research records will be stored securely and only researchers will have access to the records. Audio recordings will be stored in a Google Folder only accessible to the researcher and PIs.

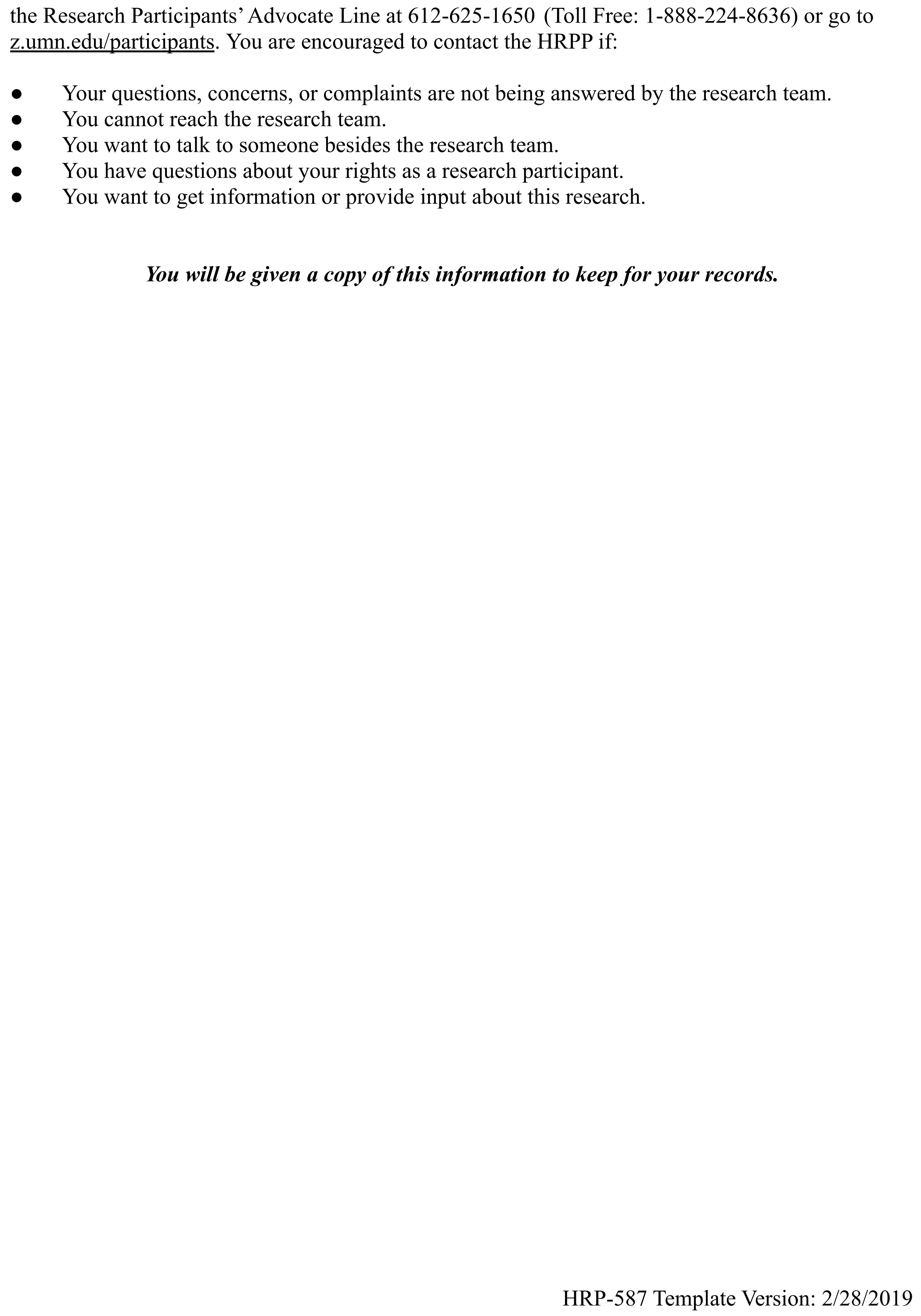
Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota

Contacts and Questions:

The researcher(s) conducting this study is (are): Chelsey Legacy, Robert delMas, and Andrew Zieffler. You may ask any questions you have now. If you have questions later, **you are encouraged** to contact them at Chelsey Legacy legac006@umn.edu: Robert delMas, delma001@umn.edu, Andrew Zieffler, zief004@umn.edu .

This research has been reviewed and approved by an IRB within the Human Research Protections Program (HRPP). To share feedback privately with the HRPP about your research experience, call



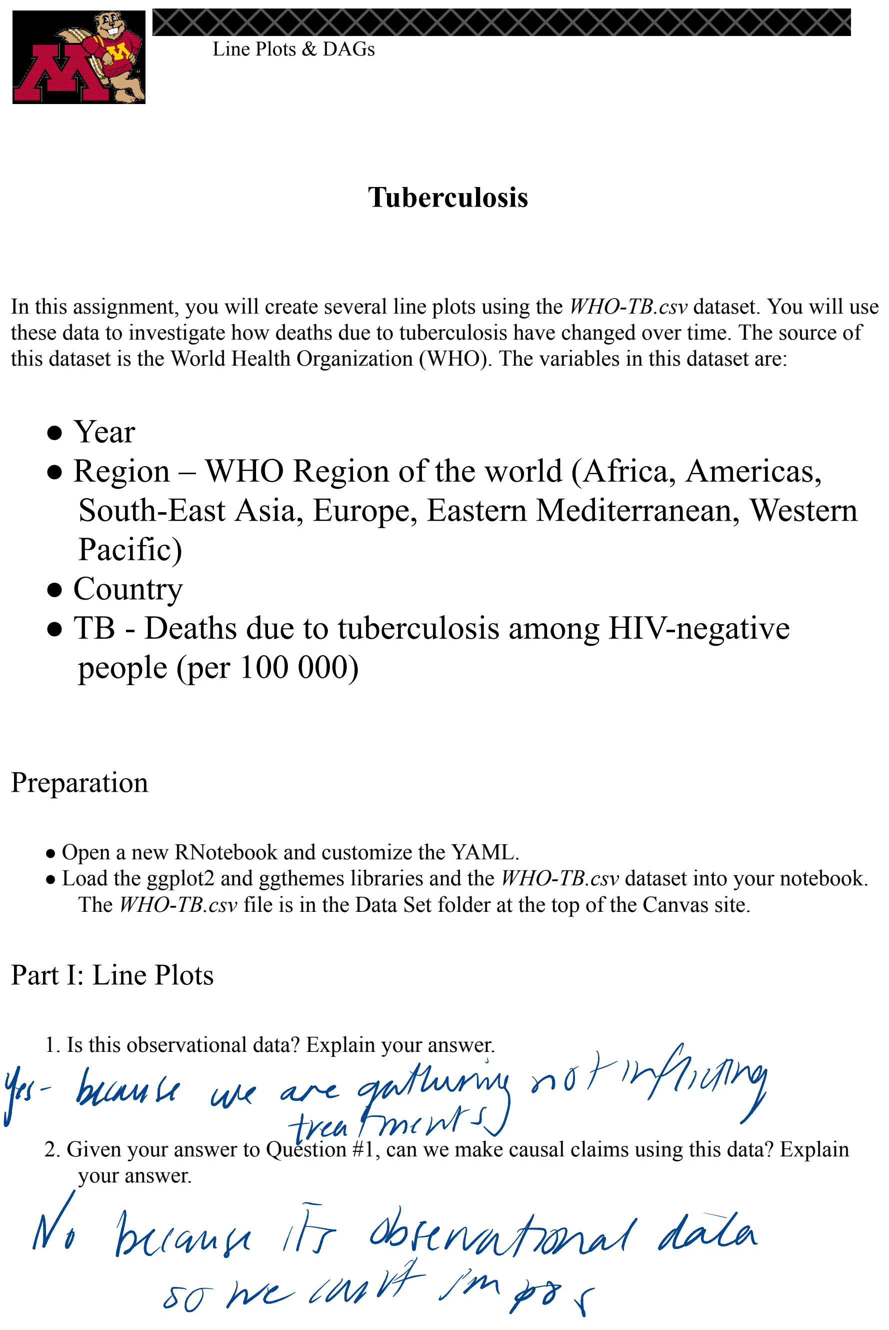
the Research Participants' Advocate Line at 612-625-1650 (Toll Free: 1-888-224-8636) or go to z.umn.edu/participants. You are encouraged to contact the HRPP if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research participant.
- You want to get information or provide input about this research.

You will be given a copy of this information to keep for your records.

B.4 Think-Aloud Feedback Notes

B.4.1 Round 1: Assignment 1



Line Plots & DAGs

Tuberculosis

In this assignment, you will create several line plots using the *WHO-TB.csv* dataset. You will use these data to investigate how deaths due to tuberculosis have changed over time. The source of this dataset is the World Health Organization (WHO). The variables in this dataset are:

- Year
- Region – WHO Region of the world (Africa, Americas, South-East Asia, Europe, Eastern Mediterranean, Western Pacific)
- Country
- TB - Deaths due to tuberculosis among HIV-negative people (per 100 000)

Preparation

- Open a new RNotebook and customize the YAML.
- Load the ggplot2 and ggthemes libraries and the *WHO-TB.csv* dataset into your notebook. The *WHO-TB.csv* file is in the Data Set folder at the top of the Canvas site.

Part I: Line Plots

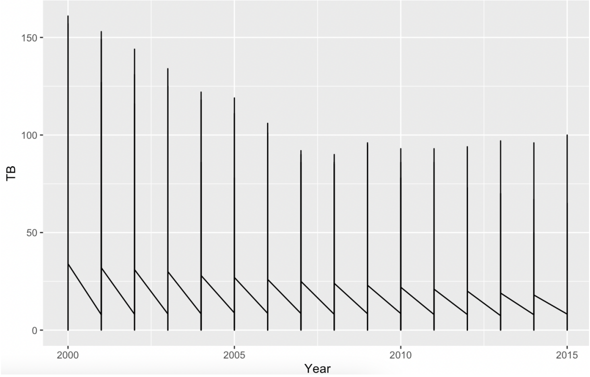
1. Is this observational data? Explain your answer.

Yes- because we are gathering not inflicting treatments)

2. Given your answer to Question #1, can we make causal claims using this data? Explain your answer.

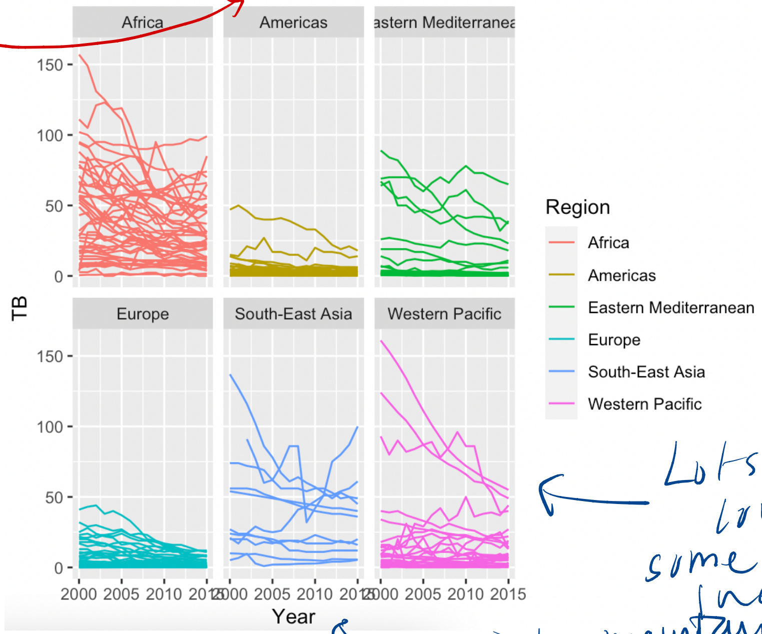
No because its observational data so we can't impos

3. Use the *WHO-TB.csv* dataset to make a line plot visualization of tuberculosis deaths across time [redacted] and grouped by country. This creates a messy, uninterpretable visualization. Paste this plot below.



4. To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set [redacted] = County. Paste your plot below.

group = Country



Lots low & remain low

some ↑ some dramatic increase & erratic (at least 1 country erratic)

Remain low, maintain

has steady increase

5. What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see and trends within each region.

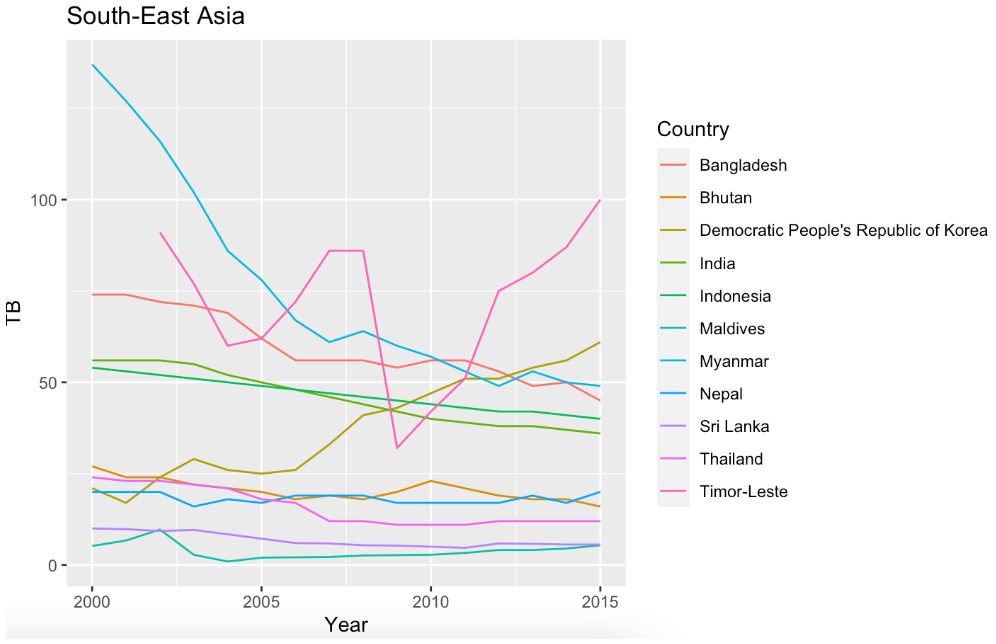
- Overall decreasing/maintaing low over time, decreased over time range of 15 yrs.
- Africa, spaghetti
- Some countries increasing
- some go down & up

Americas remained low mostly

Europe is ~~great~~ low & remain low.

Part II: Closer Inspection

6. Choose a WHO region that has at least one country that has an increasing trend. Create a new dataset for your chosen region based on the *WHO-TB.csv* dataset. Create a line plot for your new dataset colored by country. Paste your line plot below.



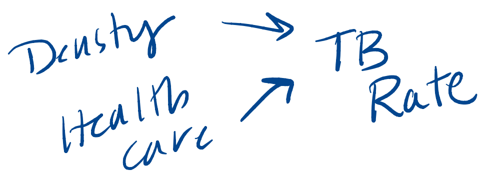
7. What conclusions can you make from your visualization about tuberculosis deaths in your chosen region over time?

Alot maintaining low rate   1 country ↑
4 countries decreasing   1 erratic

8. What variables (other than country/region) do you think might affect the number of deaths due to Tuberculosis? Draw a DAG to incorporate *at least two variables* that you think affect the number of Tuberculosis deaths. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.

- Spread through particles maybe density
- General healthcare / availibility of Health care

9. Provide a justification for your drawing in Question #8 that explains your proposed relationships among the variables you chose.



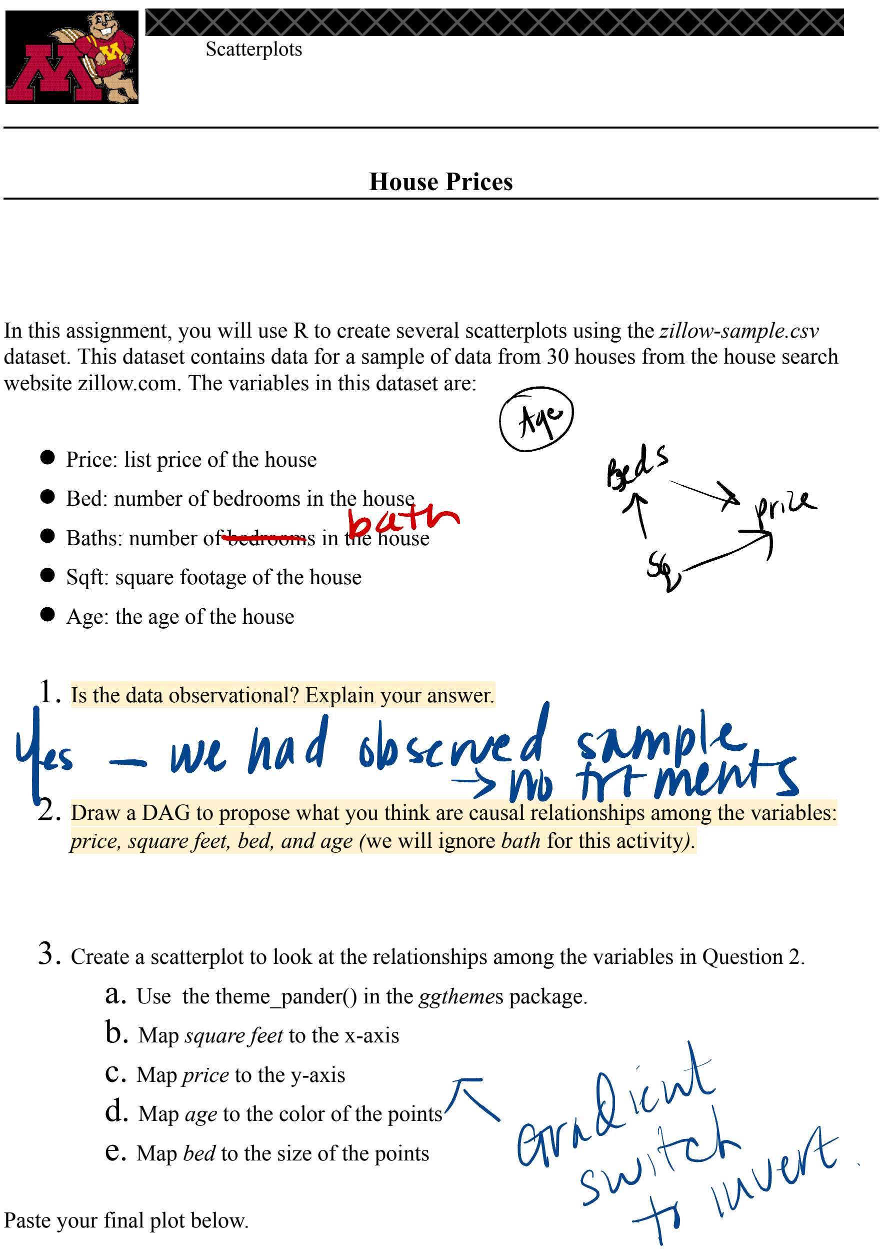
Reflections:

Make Plot for $\boxed{6}$ pretty

- change labels + title = (1pt)
- add theme (1pt)
- ~~add title~~
- Caption: add source of data as caption or subtitle. (1pt)
- Rescale x-axis by years (1pt)

4pts

B.4.2 Round 1: Assignment 2



[redacted]

Scatterplots

House Prices

In this assignment, you will use R to create several scatterplots using the *zillow-sample.csv* dataset. This dataset contains data for a sample of data from 30 houses from the house search website zillow.com. The variables in this dataset are:

- Price: list price of the house
- Bed: number of bedrooms in the house
- Baths: number of ~~bedrooms~~ bath s in the house
- Sqft: square footage of the house
- Age: the age of the house

1. Is the data observational? Explain your answer.

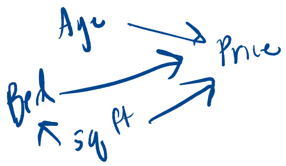
Yes – we had observed sample
→ no trtments

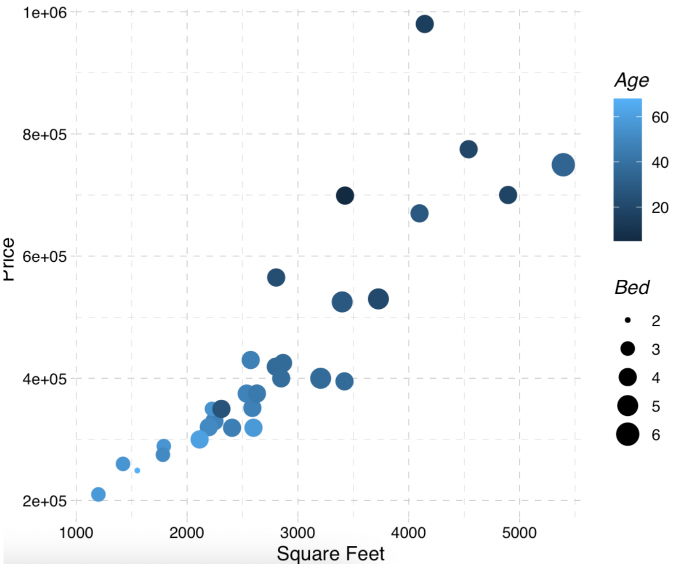
2. Draw a DAG to propose what you think are causal relationships among the variables: *price, square feet, bed, and age (*we will ignore *bath* for this activity*)*.

3. Create a scatterplot to look at the relationships among the variables in Question 2.
   a. Use the theme_pander() in the *ggthemes* package.
   b. Map *square feet* to the x-axis
   c. Map *price* to the y-axis
   d. Map *age* to the color of the points
   e. Map *bed* to the size of the points

Gradient switch to invert.

Paste your final plot below.





4. Update your DAG from Question 2 to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related. Insert your DAG below.

5. Explain your reasoning for including each directed arrow in Question #4. Use evidence from the plot above to support your answer.

Age affects house
sqft affects price
bed affects price
# bed affects price & sqft

As sqft ↑ price ↑
younger hs expensives
that the
sqft → Age as well.

6. Based on your DAG and the plot above, what variable(s) do you think affect the price of a house? Justify your answer using evidence from your plot.

sqft affects because linear relationship. → ↑
Age affects the price as well
maybe w/ bed → (apper have less but not as strong.
because we have 2 bed at low & high a thing & a relationship
but range of prices for 4 beds

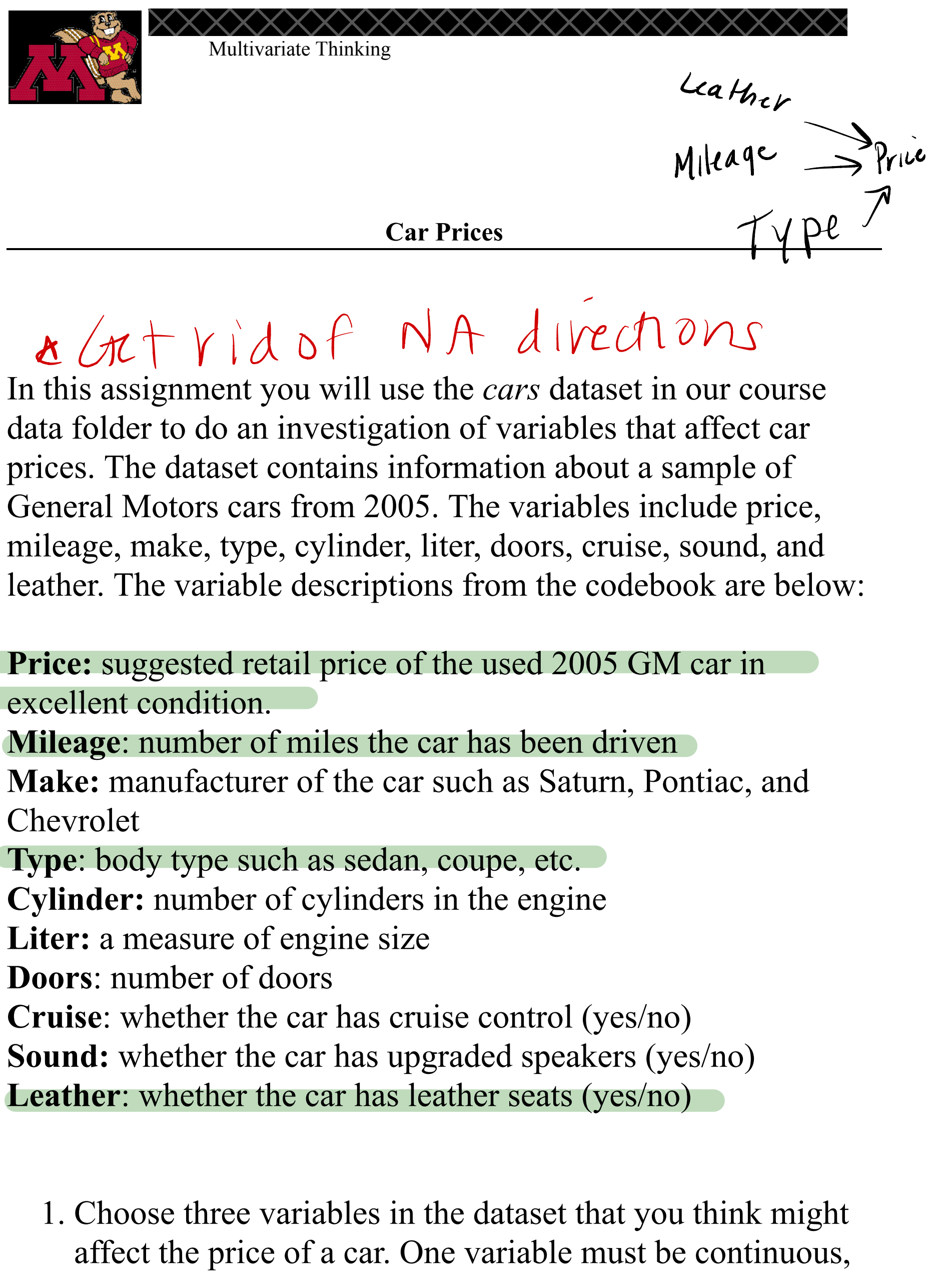
7. What other variables do you think affect the price of a house that are not considered in this dataset?

-Updated:
• Location
school district
• Amenities
- size of lot
garage stalls
or outbuildings

#3 Add title
Add relabel y axis & instructions
Reverse color gradient ? color blind.

(4pts)

B.4.3 Round 1: Assignment 3



[redacted]

Multivariate Thinking

Leather

Mileage → Price

Type

Car Prices

Get rid of NA directions

In this assignment you will use the *cars* dataset in our course data folder to do an investigation of variables that affect car prices. The dataset contains information about a sample of General Motors cars from 2005. The variables include price, mileage, make, type, cylinder, liter, doors, cruise, sound, and leather. The variable descriptions from the codebook are below:

Price: suggested retail price of the used 2005 GM car in excellent condition.
Mileage: number of miles the car has been driven
Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
Type: body type such as sedan, coupe, etc.
Cylinder: number of cylinders in the engine
Liter: a measure of engine size
Doors: number of doors
Cruise: whether the car has cruise control (yes/no)
Sound: whether the car has upgraded speakers (yes/no)
Leather: whether the car has leather seats (yes/no)

1. Choose three variables in the dataset that you think might affect the price of a car. One variable must be continuous,

but the other two variables may be whatever you chose.

2. Draw a DAG to propose a relationship among the variables you chose. Paste it below.

3. Provide a plot of three of the variables that you are investigating.

4. Update your DAG based on that plot.

5. Provide a plot of all your variables and price. Be sure to
   a. Use a theme and colors other than the default settings
   b. Add a thoughtful title and labels to your axes

6. Draw your final DAG to represent how your 4 variables affect price and each other. Paste it below.

7. Provide a justification for your DAG using your plot as evidence.

8. Are there any other variables (not in the current data table) that you think affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars?

9. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions (refer to the variables above) would you recommend they buy it? Explain your answer below using evidence from your final plot and DAG.

Based on your plot

convertible & leather seats
not if hatchback wagon or coupe

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

& ~~and if~~ not for a sedan

① Start w/ Price & Mileage

3rd) facet by type

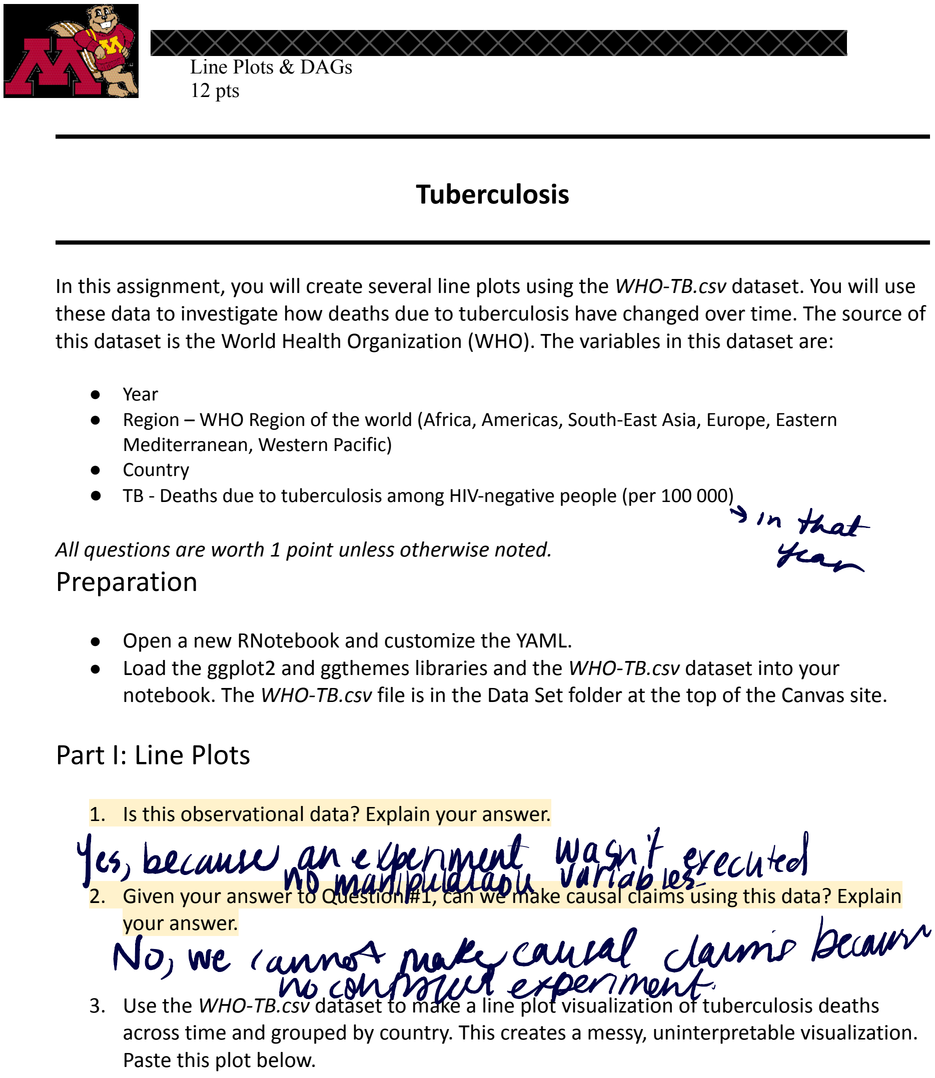
4th var → color or whatever you want

↳ Let them pick as they go

↳ Make them redo w/ 4 variables & compare.

- Fix up final plot!

B.4.4 Round 2: Assignment 1



[redacted]

Line Plots & DAGs
12 pts

Tuberculosis

In this assignment, you will create several line plots using the *WHO-TB.csv* dataset. You will use these data to investigate how deaths due to tuberculosis have changed over time. The source of this dataset is the World Health Organization (WHO). The variables in this dataset are:

- Year
- Region – WHO Region of the world (Africa, Americas, South-East Asia, Europe, Eastern Mediterranean, Western Pacific)
- Country
- TB - Deaths due to tuberculosis among HIV-negative people (per 100 000) → in that year

All questions are worth 1 point unless otherwise noted.

Preparation

- Open a new RNotebook and customize the YAML.
- Load the ggplot2 and ggthemes libraries and the *WHO-TB.csv* dataset into your notebook. The *WHO-TB.csv* file is in the Data Set folder at the top of the Canvas site.

Part I: Line Plots

1. Is this observational data? Explain your answer.

Yes, because an experiment wasn't executed
no manipulation variables.

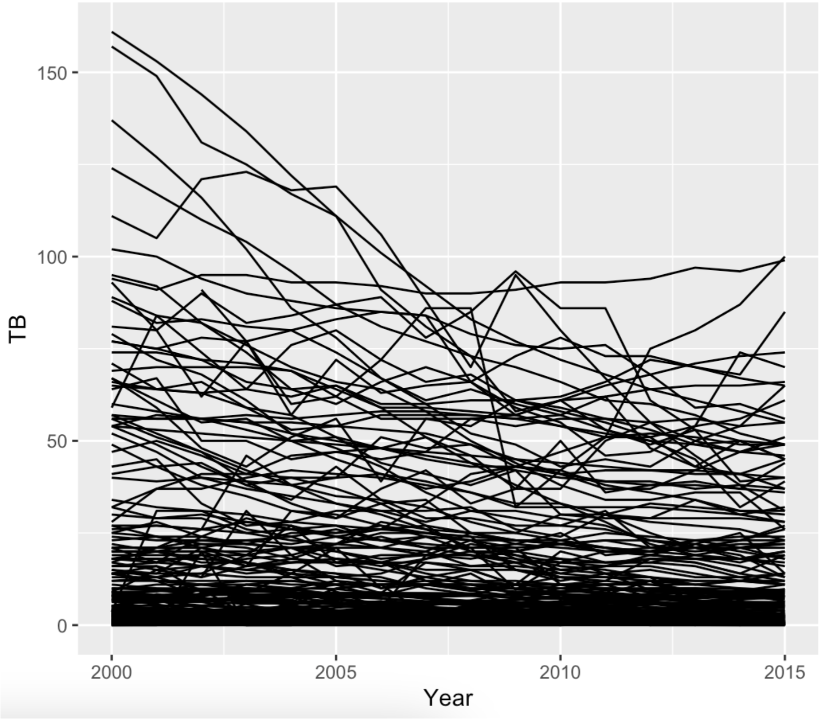
2. Given your answer to Question #1, can we make causal claims using this data? Explain your answer.

No, we cannot make causal claims because
no controlled experiment.

3. Use the *WHO-TB.csv* dataset to make a line plot visualization of tuberculosis deaths across time and grouped by country. This creates a messy, uninterpretable visualization. Paste this plot below.

[redacted]

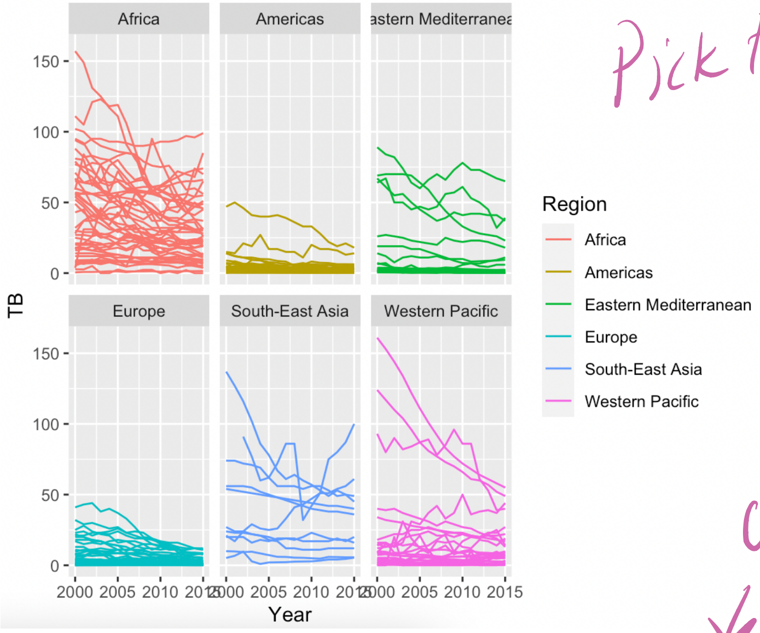
Line Plots & DAGs
12 pts



4. To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set group = County. Paste your plot below.

[redacted]

Line Plots & DAGs
12 pts



Pick two regions

could go on here . . . .

Most variation in Trends appears to be in Africa.

5. What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see and trends within each region.

Difficult to draw many firm conclusions when considering all 6 regions - seems to be mild decrease in deaths over time but potentially vizually biased by handful of countries. in each region that show more Δ.

Part II: Closer Inspection

6. Choose a WHO region that has at least one country that has an increasing trend. Create a new dataset for your chosen region based on the *WHO-TB.csv* dataset. Create a line plot for your new dataset colored by country. Be sure to do each of the following to make an aesthetically pleasing plot:
   a. Use a theme other than the default theme
   b. Change the color palette
   c. Add an informative title to your plot
   d. Add a caption telling the source of the data

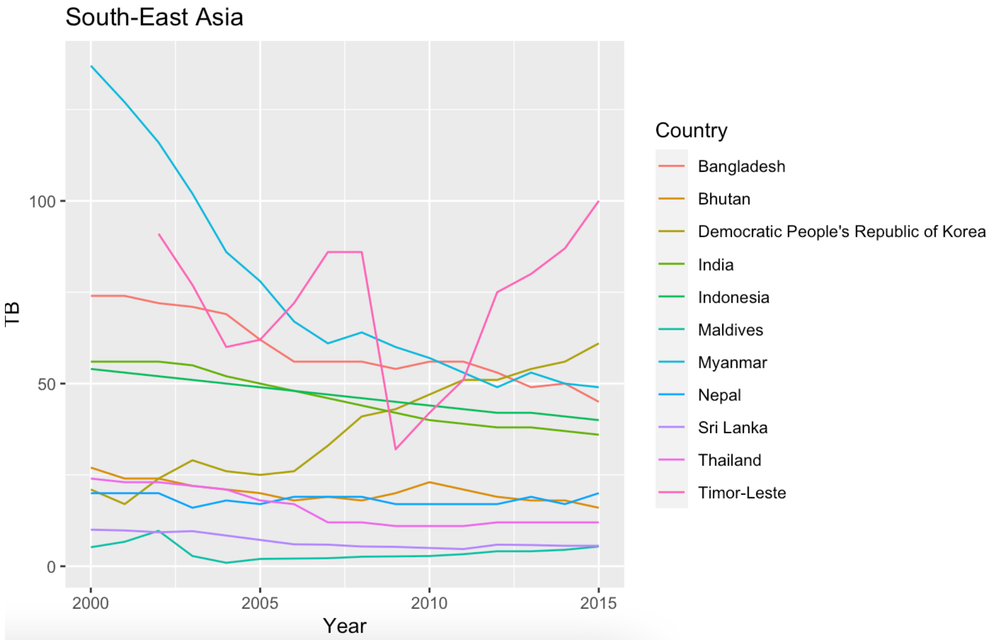
   Paste your final plot below. **(4 pts)**

The grouping at bottom of plots. Hard to visually dis entangle.

All regions seem to have experienced some countries decline / increase. exceptions.

[redacted]

Line Plots & DAGs
12 pts



7. What conclusions can you make from your visualization about tuberculosis deaths in your chosen region over time?

Overall the region appears to show some decline in TB deaths.
But Thialand Green one are exceptions

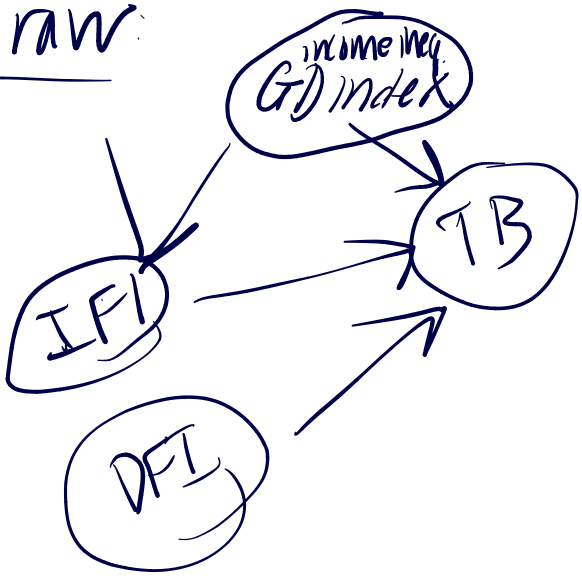
8. What variables (other than country/region) do you think might affect the number of deaths due to Tuberculosis? Draw a DAG to incorporate *at least two variables* that you think affect the number of Tuberculosis deaths. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.

9. Provide a justification for your drawing in Question #8 that explains your proposed relationships among the variables you chose.

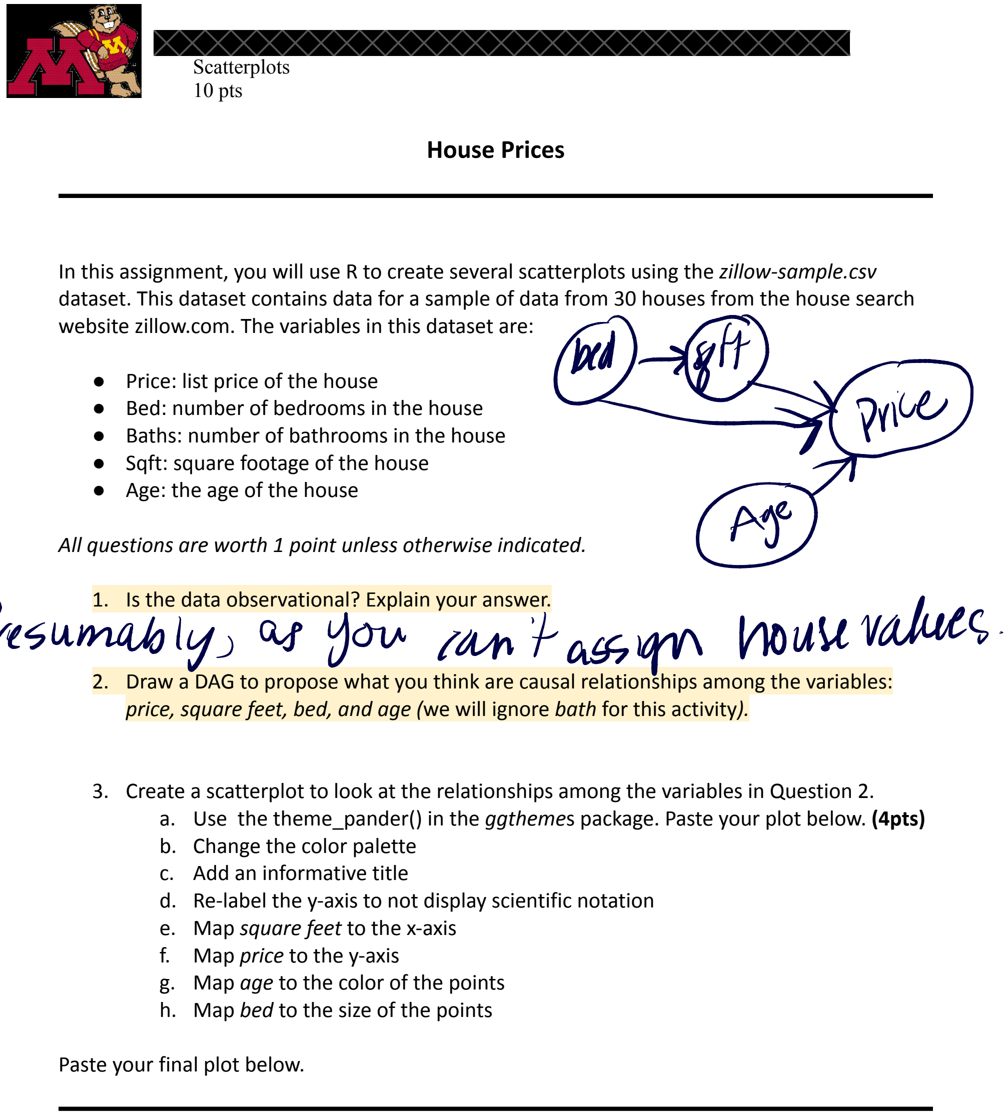
#5 put in sentence limit.

#8

Draw:



B.4.5 Round 2: Assignment 2



[redacted]

Scatterplots
10 pts

House Prices

In this assignment, you will use R to create several scatterplots using the *zillow-sample.csv* dataset. This dataset contains data for a sample of data from 30 houses from the house search website zillow.com. The variables in this dataset are:

- Price: list price of the house
- Bed: number of bedrooms in the house
- Baths: number of bathrooms in the house
- Sqft: square footage of the house
- Age: the age of the house

All questions are worth 1 point unless otherwise indicated.

1. Is the data observational? Explain your answer.

Presumably, as you can't assign house values.

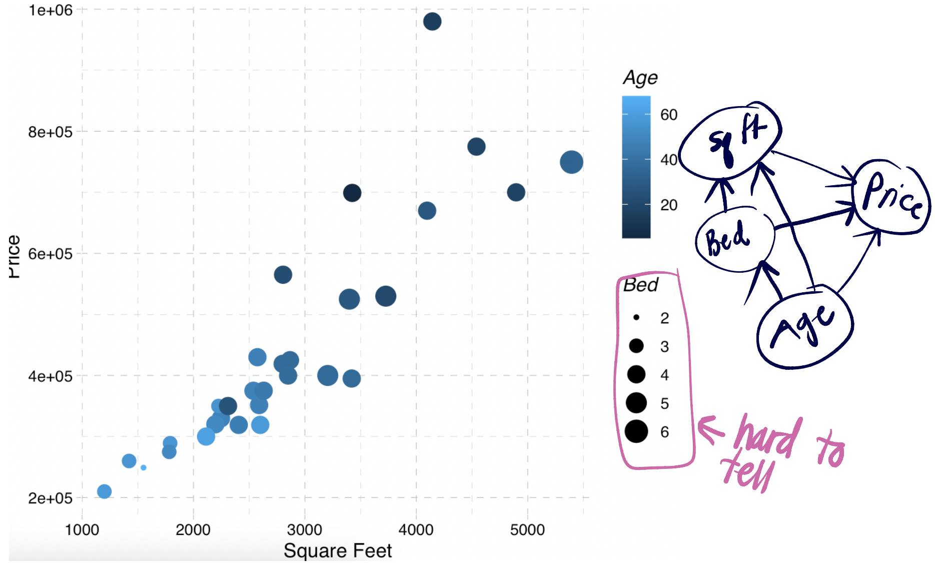
2. Draw a DAG to propose what you think are causal relationships among the variables: *price, square feet, bed, and age (*we will ignore *bath* for this activity*).*

3. Create a scatterplot to look at the relationships among the variables in Question 2.
   a. Use the theme_pander() in the *ggthemes* package. Paste your plot below. **(4pts)**
   b. Change the color palette
   c. Add an informative title
   d. Re-label the y-axis to not display scientific notation
   e. Map *square feet* to the x-axis
   f. Map *price* to the y-axis
   g. Map *age* to the color of the points
   h. Map *bed* to the size of the points

Paste your final plot below.

[redacted]

Scatterplots
10 pts



indicate to make new.

4. Update your DAG from Question 2 to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related. Insert your DAG below.

5. Explain your reasoning for including each directed arrow in Question #4. Use evidence from the plot above to support your answer.

implies causality

6. Based on your DAG and the plot above, what variable(s) do you think affect the price of a house? Justify your answer using evidence from your plot.

All of them appear to be associated but potentially to varying degrees AS sq ft seems to be most closely tied to trend.

7. What other variables do you think affect the price of a house that are not considered in this dataset?

Location; property tax rate) & school buget/capita

Age also can see a trend as lighter colors - Clu

Put experiment in for
final assignment (?)

B.4.6 Round 2: Assignment 3

[redacted]

Multivariate Thinking
17 pts

Car Prices

use of affect again.

In this assignment you will use the *cars* dataset in our course data folder to do an investigation of variables that affect car prices. The dataset contains information about a sample of General Motors cars from 2005. The variables include price, mileage, make, type, cylinder, liter, doors, cruise, sound, and leather. The variable descriptions from the codebook are below:

Price: suggested retail price of the used 2005 GM car in excellent condition.
Mileage: number of miles the car has been driven
Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
Type: body type such as sedan, coupe, etc.
Cylinder: number of cylinders in the engine
Liter: a measure of engine size
Doors: number of doors
Cruise: whether the car has cruise control (yes/no)
Sound: whether the car has upgraded speakers (yes/no)
Leather: whether the car has leather seats (yes/no)

Questions worth 1 point unless otherwise stated.

1. What do you predict is the relationship, if any, between the price of a used car and the number of miles on the car? Explain your answer.

↑ miles Price↓ car usage decreases in value.

2. Create a scatterplot with price on the y-axis and mileage on the x-axis.
3. Describe your scatterplot (linearity, strength of relationship, and slope).

Potentially linear (-) slight relationship.

4. Building on the code from your plot in number 2, facet on type of car.
5. Describe the scatterplot for each of the different types of car.

Conv. strong-linear | weak linear neg neg sedan wagon SW moderate neg

6. Is this the same relationship we saw between mileage and price we saw before we faceted on type? Explain why or why not.

In general yes, but varying strength across cars. No curvature anywhere, no positive, varying strength

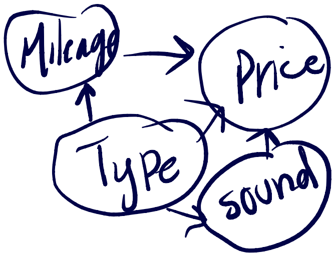
Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts

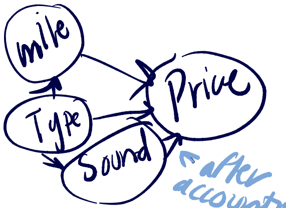
7. Choose a fourth variable from the list of variables above. Create a DAG to propose relationships among the four variables. Paste it below.

8. Explain your proposed relationships.



9. Create a plot incorporating the four variables. **(4pts)**
   a. Put a title on the plot
   b. Use a theme other than the default theme
   c. Use a colorblind friendly color palette
   d. Add a caption with the source of the data to the plot

10. Draw your final DAG to represent how your 4 variables affect price and each other. Paste it below.



← after accounting for type no.

11. Provide a justification for your DAG using your plot as evidence.

12. Create a new plot using your same 4 variables (i.e. map different variables to the a and y axes, colors, size, etc). Be sure to update the title, colors, and theme as well.

maybe notice interaction effects.

13. Do you think your plot in Question #9 or Question #12 depicts the relationships among the variables better? Explain your answer.

swap colors - hard to see relationships here

obscurs all relations

14. Are there any other variables (not in the current data set) that you think affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars? Explain your answer.

Eh rewrite - Limit one other variable

15. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions (based on your plot) would you recommend they buy it? Explain your answer below using evidence from your final plot and DAG.

Don't buy if its wagon, sedan, coupe only convertible & 20,000 miles. Sound perfect. All under 70,000

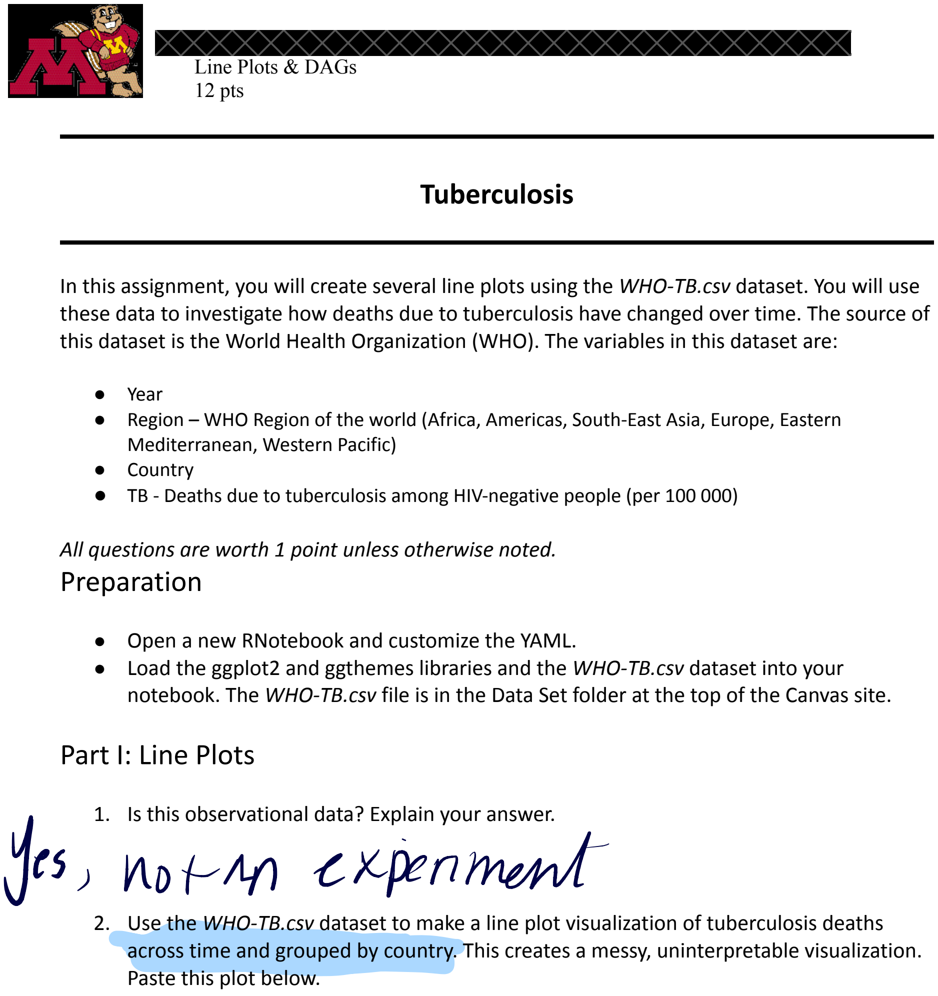
Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

Te cat w/ no colors was easier
as facet.

Explaining
"outward facing DAG"
given forced uncertainty
design elements
- could be useful

Could be useful.

B.4.7 Round 3: Assignment 1



[redacted]

Line Plots & DAGs
12 pts

Tuberculosis

In this assignment, you will create several line plots using the *WHO-TB.csv* dataset. You will use these data to investigate how deaths due to tuberculosis have changed over time. The source of this dataset is the World Health Organization (WHO). The variables in this dataset are:

- Year
- Region – WHO Region of the world (Africa, Americas, South-East Asia, Europe, Eastern Mediterranean, Western Pacific)
- Country
- TB - Deaths due to tuberculosis among HIV-negative people (per 100 000)

All questions are worth 1 point unless otherwise noted.

Preparation

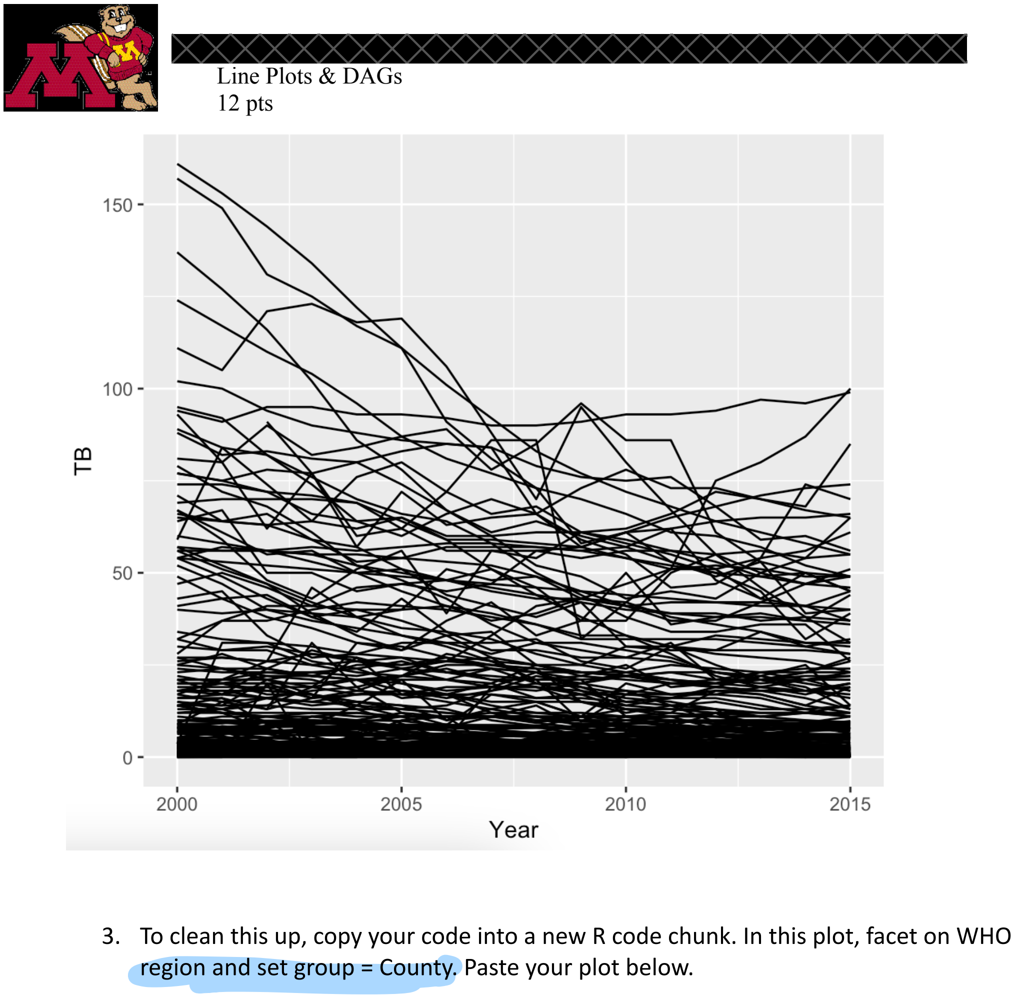
- Open a new RNotebook and customize the YAML.
- Load the ggplot2 and ggthemes libraries and the *WHO-TB.csv* dataset into your notebook. The *WHO-TB.csv* file is in the Data Set folder at the top of the Canvas site.

Part I: Line Plots

1. Is this observational data? Explain your answer.

Yes, not an experiment

2. Use the *WHO-TB.csv* dataset to make a line plot visualization of tuberculosis deaths across time and grouped by country. This creates a messy, uninterpretable visualization. Paste this plot below.

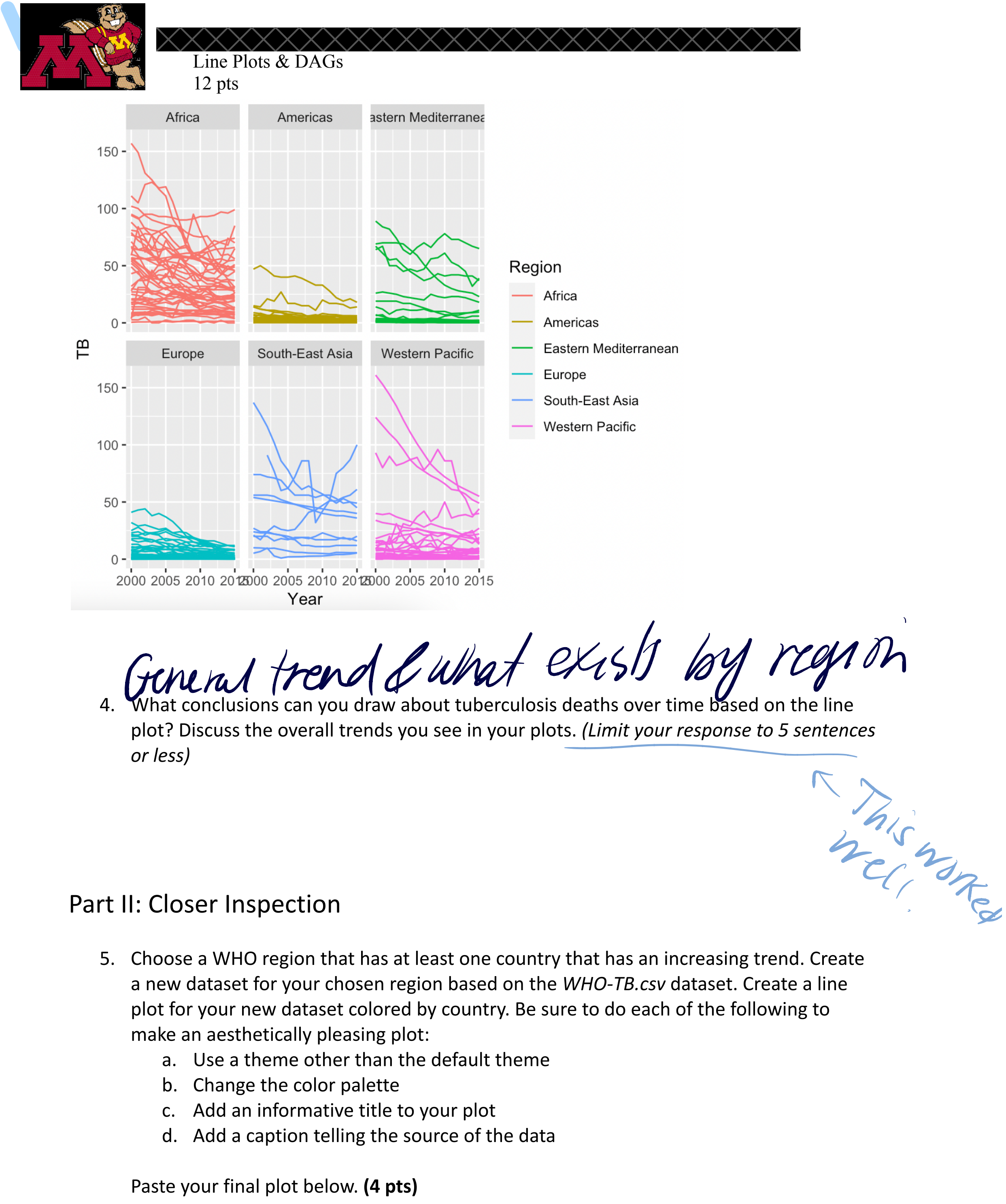


[redacted]

Line Plots & DAGs
12 pts

TB
150
100
50
0
2000
2005
2010
2015
Year

3. To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set group = County. Paste your plot below.



[redacted]

Line Plots & DAGs
12 pts

Africa
Americas
astern Mediterranea
Europe
South-East Asia
Western Pacific
150
100
50
0
TB
2000 2005 2010 2015
Year
Region
Africa
Americas
Eastern Mediterranean
Europe
South-East Asia
Western Pacific

General trend & what exists by region

4. What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see in your plots. *(Limit your response to 5 sentences or less)*

This worked well.

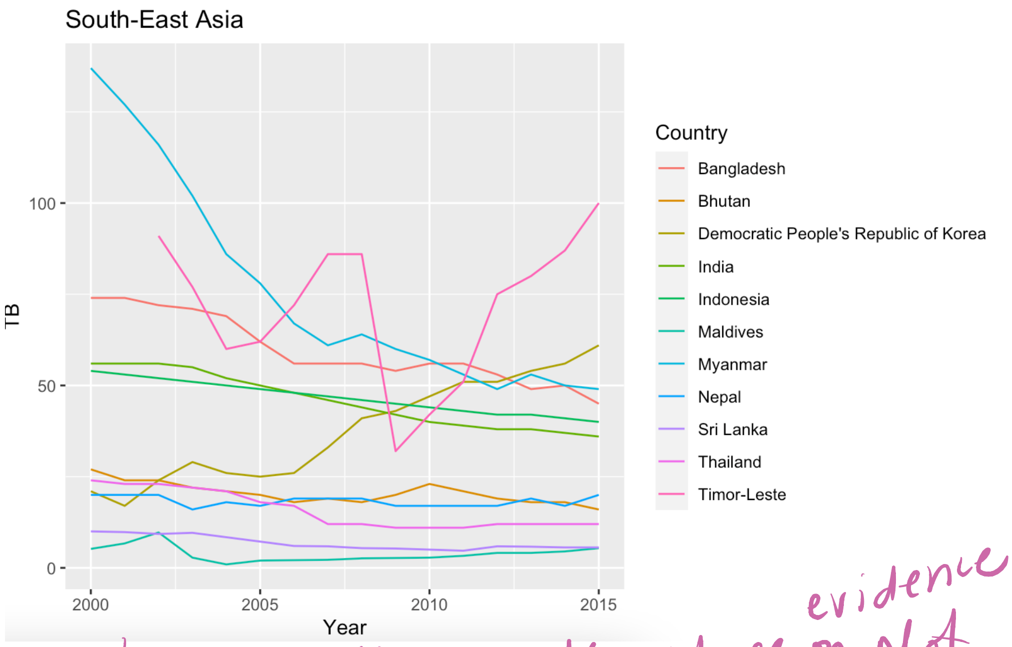
Part II: Closer Inspection

5. Choose a WHO region that has at least one country that has an increasing trend. Create a new dataset for your chosen region based on the *WHO-TB.csv* dataset. Create a line plot for your new dataset colored by country. Be sure to do each of the following to make an aesthetically pleasing plot:
    a. Use a theme other than the default theme
    b. Change the color palette
    c. Add an informative title to your plot
    d. Add a caption telling the source of the data

   Paste your final plot below. **(4 pts)**

[redacted]

Line Plots & DAGs
12 pts



Rewrite to make sure they base on plot evidence

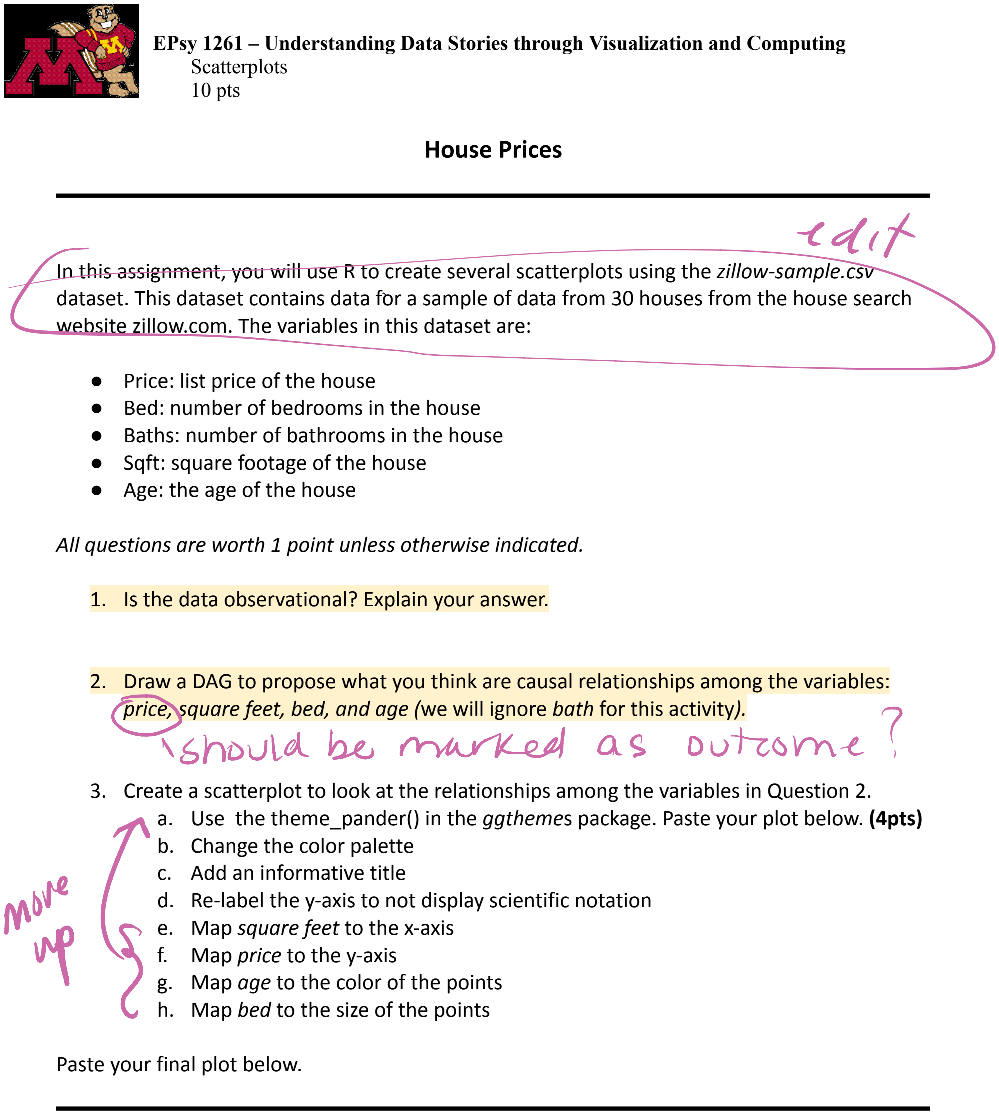
6. What conclusions can you make from your visualization about tuberculosis deaths in your chosen region over time?

7. What variables (other than country/region) do you think might affect the number of deaths due to Tuberculosis? Draw a DAG to incorporate *at least two variables* that you think affect the number of Tuberculosis deaths. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.

rewrite to explain + add to lineplot activity

8. Provide a justification for your drawing in Question #8 that explains your proposed relationships among the variables you chose.

B.4.8 Round 3: Assignment 2



EPsy 1261 – Understanding Data Stories through Visualization and Computing
Scatterplots
10 pts

House Prices

edit

In this assignment, you will use R to create several scatterplots using the *zillow-sample.csv* dataset. This dataset contains data for a sample of data from 30 houses from the house search website zillow.com. The variables in this dataset are:

- Price: list price of the house
- Bed: number of bedrooms in the house
- Baths: number of bathrooms in the house
- Sqft: square footage of the house
- Age: the age of the house

All questions are worth 1 point unless otherwise indicated.

1. Is the data observational? Explain your answer.

2. Draw a DAG to propose what you think are causal relationships among the variables: *price, square feet, bed, and age* (we will ignore *bath* for this activity*).*

   should be marked as outcome ?

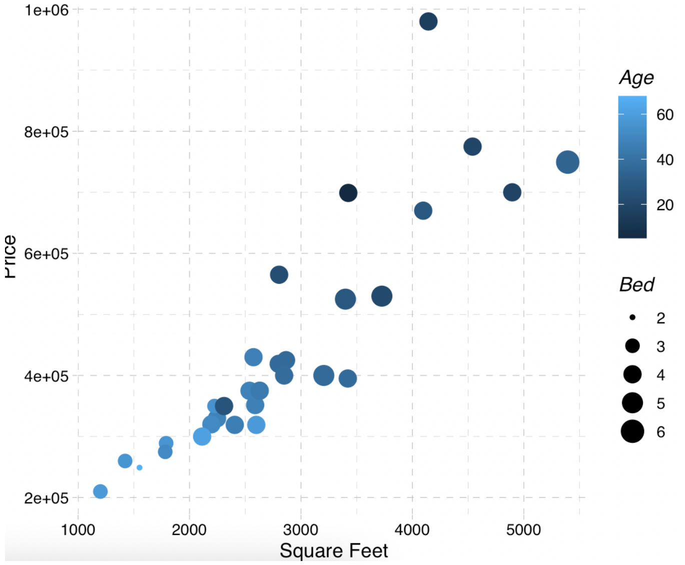
3. Create a scatterplot to look at the relationships among the variables in Question 2.
   a. Use the theme_pander() in the *ggthemes* package. Paste your plot below. **(4pts)**
   b. Change the color palette
   c. Add an informative title
   d. Re-label the y-axis to not display scientific notation
   e. Map *square feet* to the x-axis
   f. Map *price* to the y-axis
   g. Map *age* to the color of the points
   h. Map *bed* to the size of the points

move up

Paste your final plot below.

[redacted]

Scatterplots
10 pts



4. Draw a DAG to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related. (Note: this may or may not be different from the DAG drawn in Question #2) Insert your DAG below.

Include double sided arrows

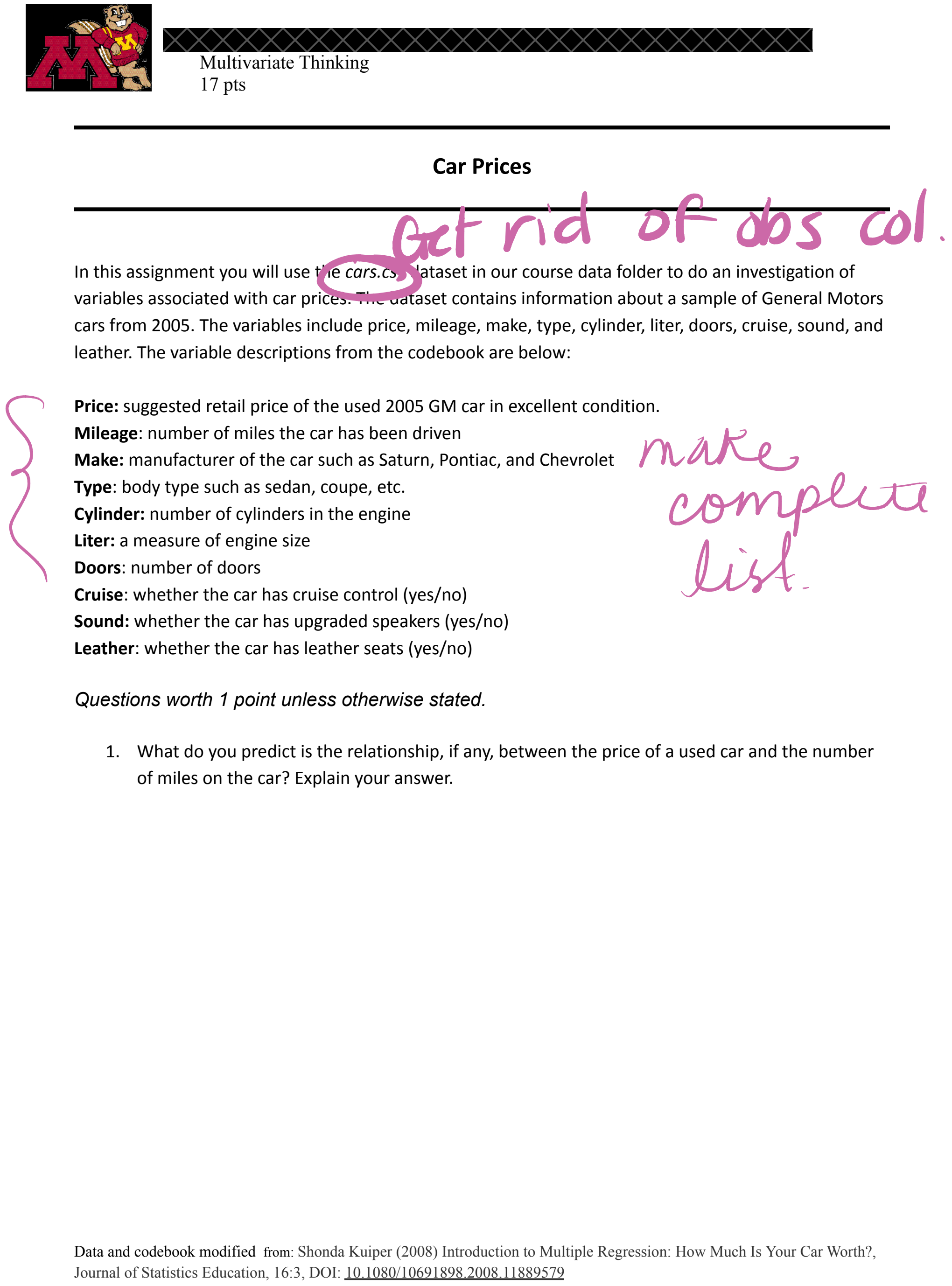
5. Explain your reasoning for including each directed arrow in Question #4. Use evidence from the plot above to support your answer.

consider a swap here

6. Based on your DAG and the plot above, what variable(s) do you think are associated with the price of a house? Justify your answer using evidence from your plot.

7. What other variables do you think are associated with the price of a house that are not considered in this dataset?

B.4.9 Round 3: Assignment 3



[redacted]

Multivariate Thinking
17 pts

Car Prices

Get rid of obs col.

In this assignment you will use the *cars.csv* dataset in our course data folder to do an investigation of variables associated with car prices. The dataset contains information about a sample of General Motors cars from 2005. The variables include price, mileage, make, type, cylinder, liter, doors, cruise, sound, and leather. The variable descriptions from the codebook are below:

Price: suggested retail price of the used 2005 GM car in excellent condition.
Mileage: number of miles the car has been driven
Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
Type: body type such as sedan, coupe, etc.
Cylinder: number of cylinders in the engine
Liter: a measure of engine size
Doors: number of doors
Cruise: whether the car has cruise control (yes/no)
Sound: whether the car has upgraded speakers (yes/no)
Leather: whether the car has leather seats (yes/no)

make complete list.

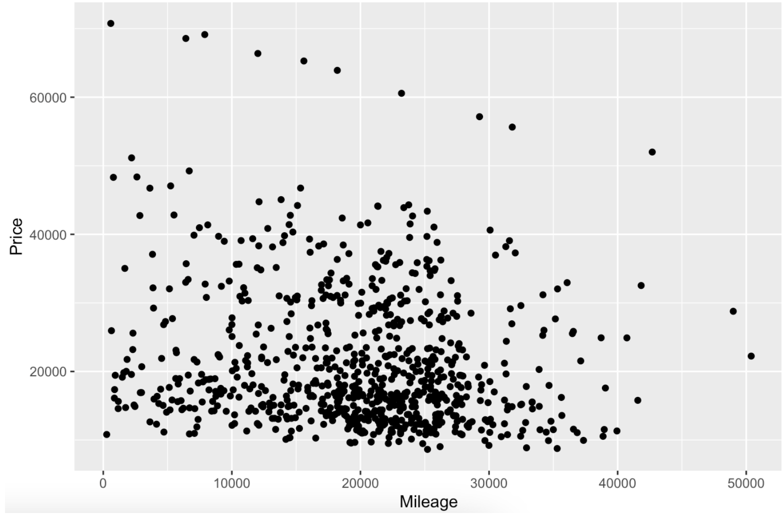
Questions worth 1 point unless otherwise stated.

1. What do you predict is the relationship, if any, between the price of a used car and the number of miles on the car? Explain your answer.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

Multivariate Thinking
17 pts

2. Create a scatterplot with price on the y-axis and mileage on the x-axis.



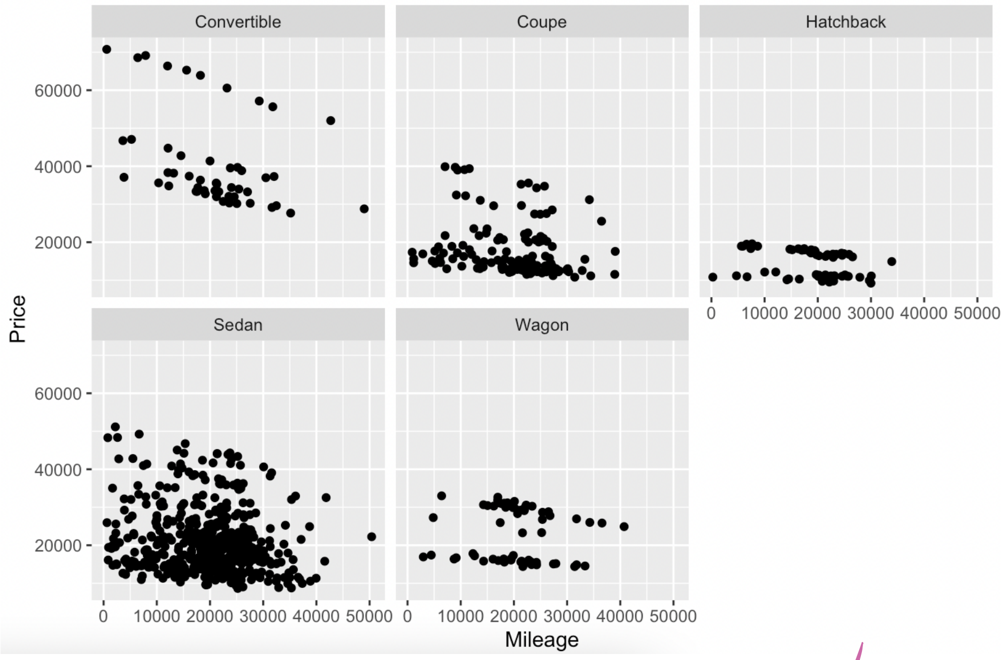
3. Describe your scatterplot (linearity, strength of relationship, and slope).

4. Building on the code from your plot in number 2, facet on *type* of car.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts



~~5. Describe the scatterplot for each of the different types of car.~~

that

6. Is this the same relationship we saw between mileage and price we saw before we faceted on type? Explain why or why not.

7. Choose a fourth variable from the list of variables above. Create a DAG to propose relationships among the four variables. Paste it below.

8. Explain your proposed relationships.

9. Create a plot incorporating the four variables. **(4pts)**
   a. Put a title on the plot
   b. Use a theme other than the default theme
   c. Use a colorblind friendly color palette
   d. Add a caption with the source of the data to the plot

10. Draw your final DAG to represent how your 4 variables affect price and each other. Paste it below.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts

11. Provide a justification for your DAG using your plot as evidence.

leave price on y

12. Create a new plot using your same 4 variables (i.e. map different variables to the x and y axes, colors, size, etc). Be sure to update the title, colors, and theme as well.

will they detect interactions(?)

Ask Suzanne

13. Do you think your plot in Question #9 or Question #12 depicts the relationships among the variables better? Explain your answer.

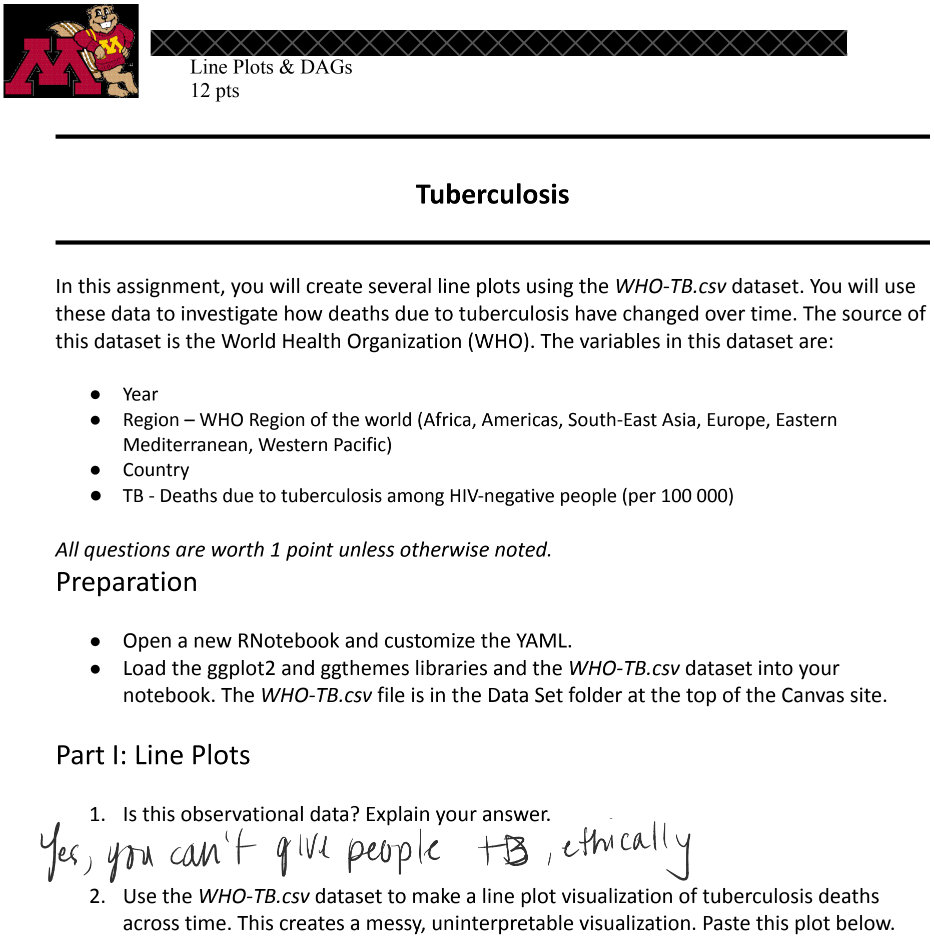
14. Are there any other variables (not in the current data set) that you think affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars? Explain your answer. *(Limit your response to 5 sentences or less)*

15. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions (based on your plot) would you recommend they buy it? Explain your answer below using evidence from your final plot and DAG.

takeaway - they are time consuming

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

B.4.10 Round 4: Assignment 1



[redacted]

Line Plots & DAGs
12 pts

Tuberculosis

In this assignment, you will create several line plots using the *WHO-TB.csv* dataset. You will use these data to investigate how deaths due to tuberculosis have changed over time. The source of this dataset is the World Health Organization (WHO). The variables in this dataset are:

- Year
- Region – WHO Region of the world (Africa, Americas, South-East Asia, Europe, Eastern Mediterranean, Western Pacific)
- Country
- TB - Deaths due to tuberculosis among HIV-negative people (per 100 000)

All questions are worth 1 point unless otherwise noted.

Preparation

- Open a new RNotebook and customize the YAML.
- Load the ggplot2 and ggthemes libraries and the *WHO-TB.csv* dataset into your notebook. The *WHO-TB.csv* file is in the Data Set folder at the top of the Canvas site.

Part I: Line Plots

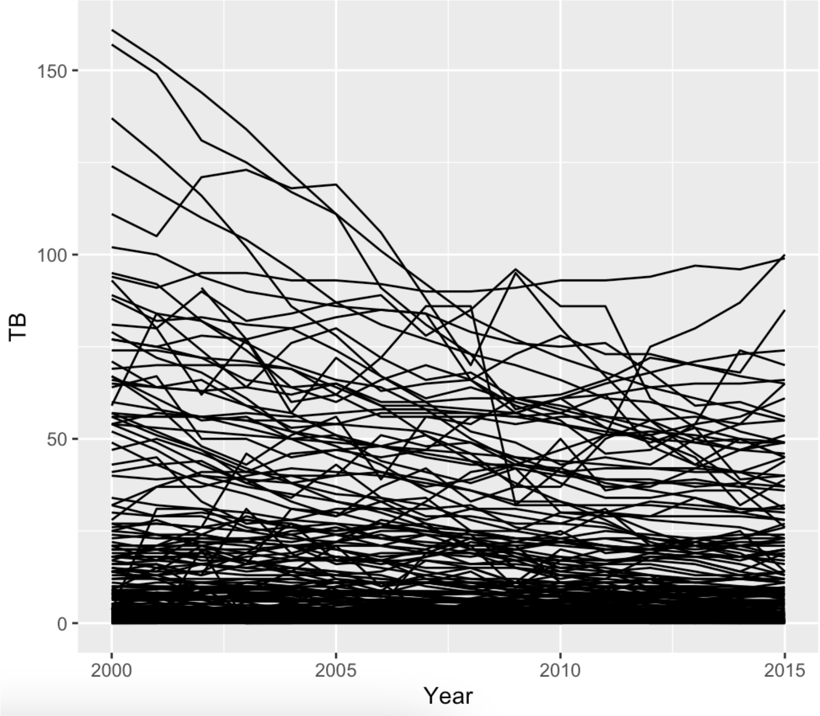
1. Is this observational data? Explain your answer.

Yes, you can't give people TB, ethically

2. Use the *WHO-TB.csv* dataset to make a line plot visualization of tuberculosis deaths across time. This creates a messy, uninterpretable visualization. Paste this plot below.

[redacted]

Line Plots & DAGs
12 pts

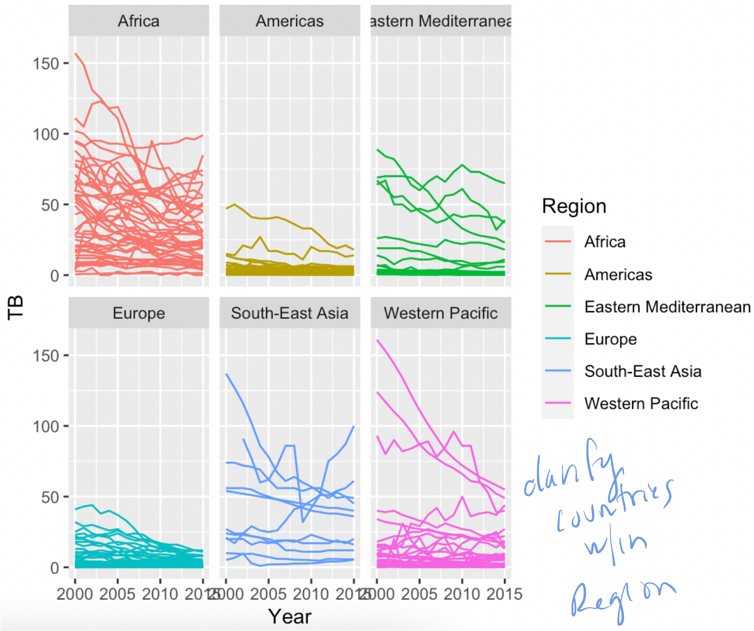


Maybe confusing but they might know it.

3. To clean this up, copy your code into a new R code chunk. In this plot, facet on WHO region and set group = Country. Paste your plot below.

[redacted]

Line Plots & DAGs
12 pts



clarify countries w/in Region

4. What conclusions can you draw about tuberculosis deaths over time based on the line plot? Discuss the overall trends you see in your plots. *(Limit your response to 5 sentences or less)*

Decreasing voer time in most places.
One line in Africa is concurning
Overall, Americas not much Δ in general
but certain countries had much more,
Noting SE Asia has
one line going up.

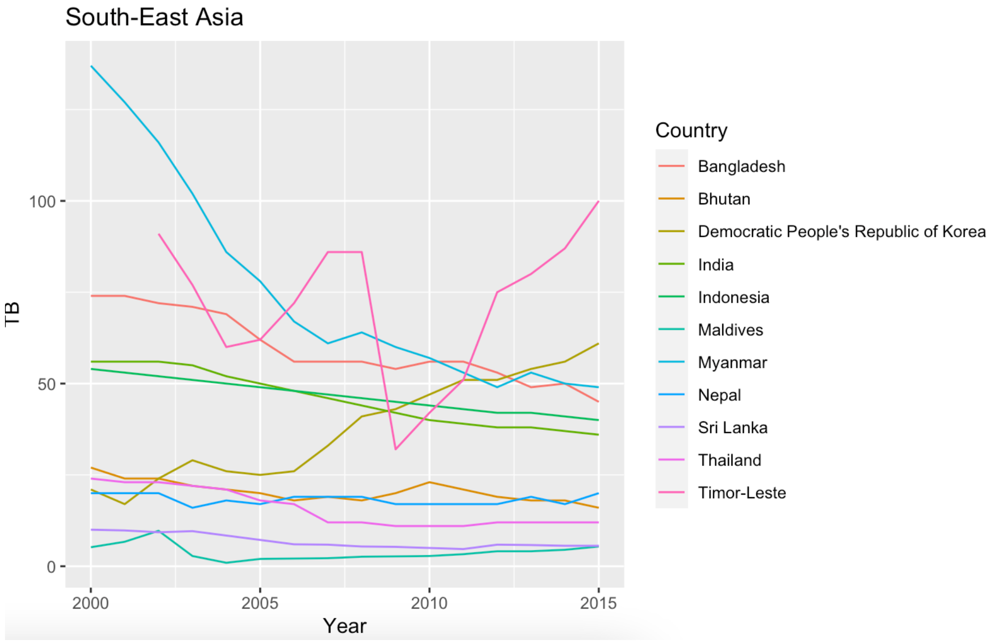
Part II: Closer Inspection

5. Choose a WHO region that has at least one country that has an increasing trend. Create a new dataset for your chosen region based on the *WHO-TB.csv* dataset. Create a line plot for your new dataset colored by country. Be sure to do each of the following to make an aesthetically pleasing plot:
   a. Use a theme other than the default theme
   b. Change the color palette
   c. Add an informative title to your plot
   d. Add a caption telling the source of the data

   Paste your final plot below. **(4 pts)**

[redacted]

Line Plots & DAGs
12 pts

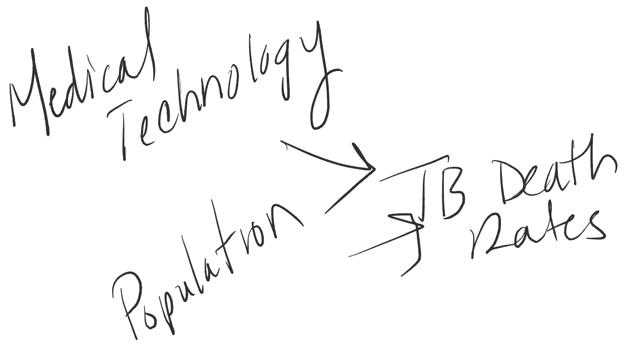


6. What conclusions can you make about tuberculosis deaths in your chosen region over time using evidence from your plot?

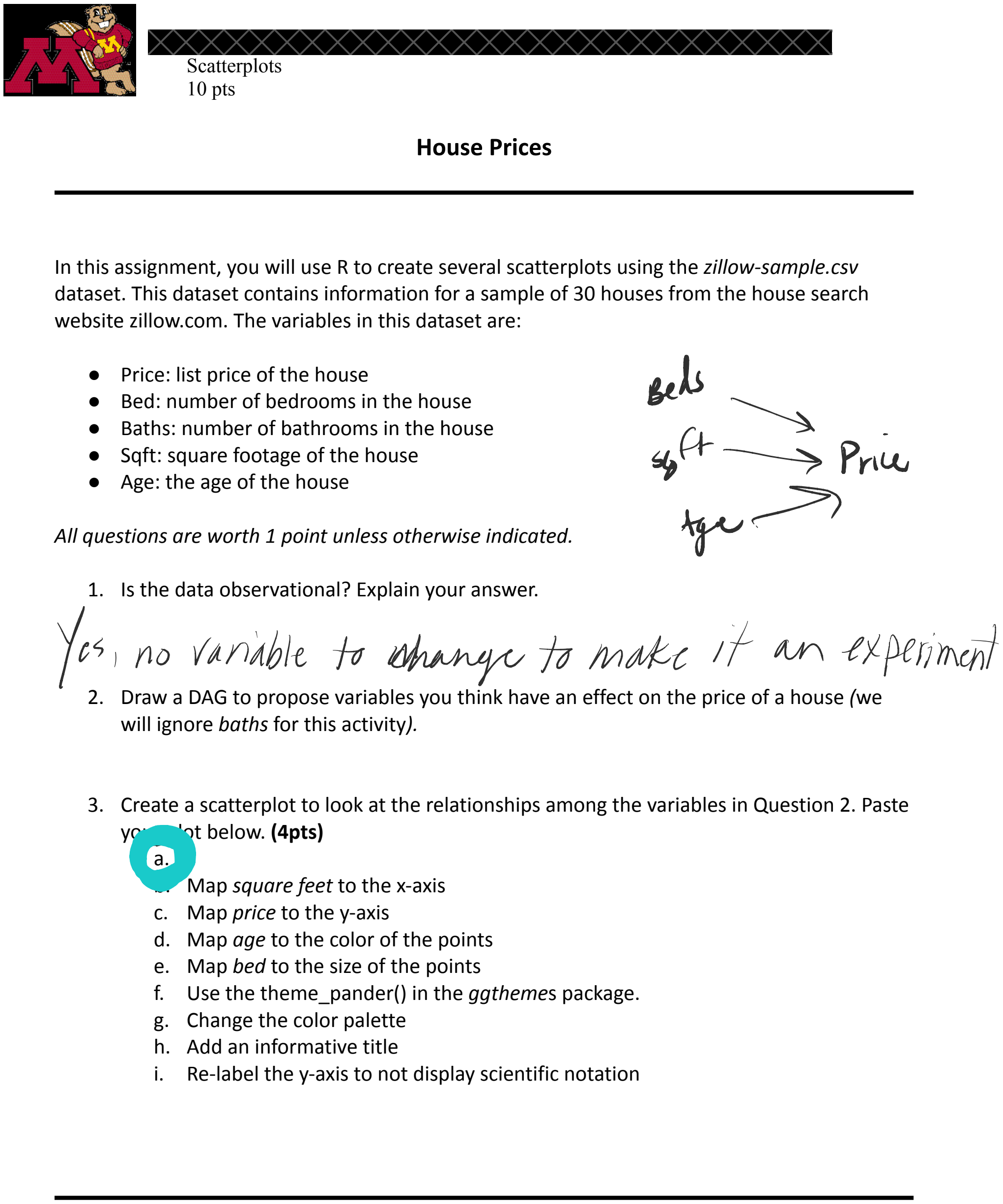
slight decrease in countries that started w/ more. Less decrease for countries that started w/ less. They are in ideal areas already. Timore-leste concerning

7. We see changes in tuberculosis death rates over time in some regions, but we might wonder what is causing these changes. Draw a DAG to incorporate two or three *variables* that you think might be associated with the tuberculosis death rate in a country. Be sure to consider the relationships among all the variables you propose. Insert your final drawing below.

8. Provide a justification for your drawing in Question #8 that explains your proposed relationships among the variables you chose.



B.4.11 Round 4: Assignment 2



[redacted]

Scatterplots
10 pts

House Prices

In this assignment, you will use R to create several scatterplots using the *zillow-sample.csv* dataset. This dataset contains information for a sample of 30 houses from the house search website zillow.com. The variables in this dataset are:

- Price: list price of the house
- Bed: number of bedrooms in the house
- Baths: number of bathrooms in the house
- Sqft: square footage of the house
- Age: the age of the house

All questions are worth 1 point unless otherwise indicated.

1. Is the data observational? Explain your answer.

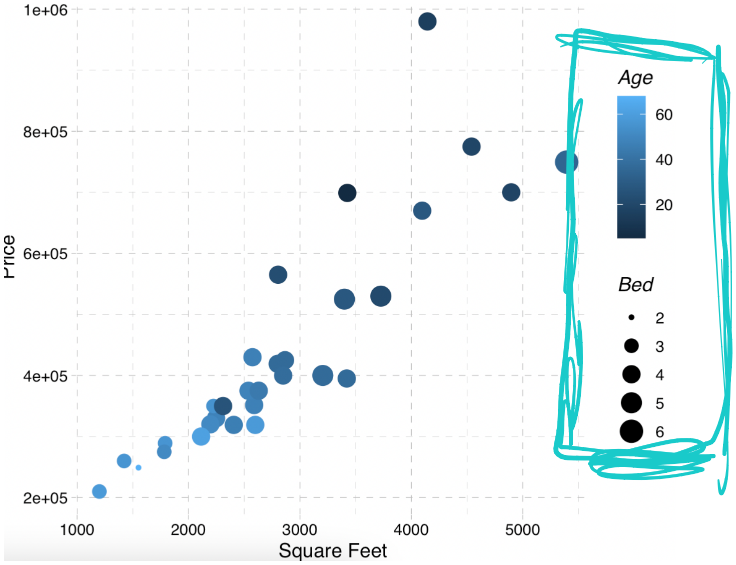
Yes, no variable to change to make it an experiment

2. Draw a DAG to propose variables you think have an effect on the price of a house *(*we will ignore *baths* for this activity*).*

3. Create a scatterplot to look at the relationships among the variables in Question 2. Paste your plot below. **(4pts)**
   a.
   b. Map *square feet* to the x-axis
   c. Map *price* to the y-axis
   d. Map *age* to the color of the points
   e. Map *bed* to the size of the points
   f. Use the theme_pander() in the *ggthemes* package.
   g. Change the color palette
   h. Add an informative title
   i. Re-label the y-axis to not display scientific notation

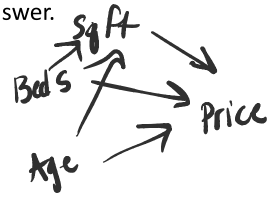
[redacted]

Scatterplots
10 pts



4. Draw a DAG to represent the relationships that are indicated in your plot above. Be sure to include directed arrows between all variables you think might be causally related. (Note: this may or may not be different from the DAG drawn in Question #2) Insert your DAG below.

5. Explain your reasoning for including each directed arrow in Question #4. Use evidence from the plot above to support your answer.



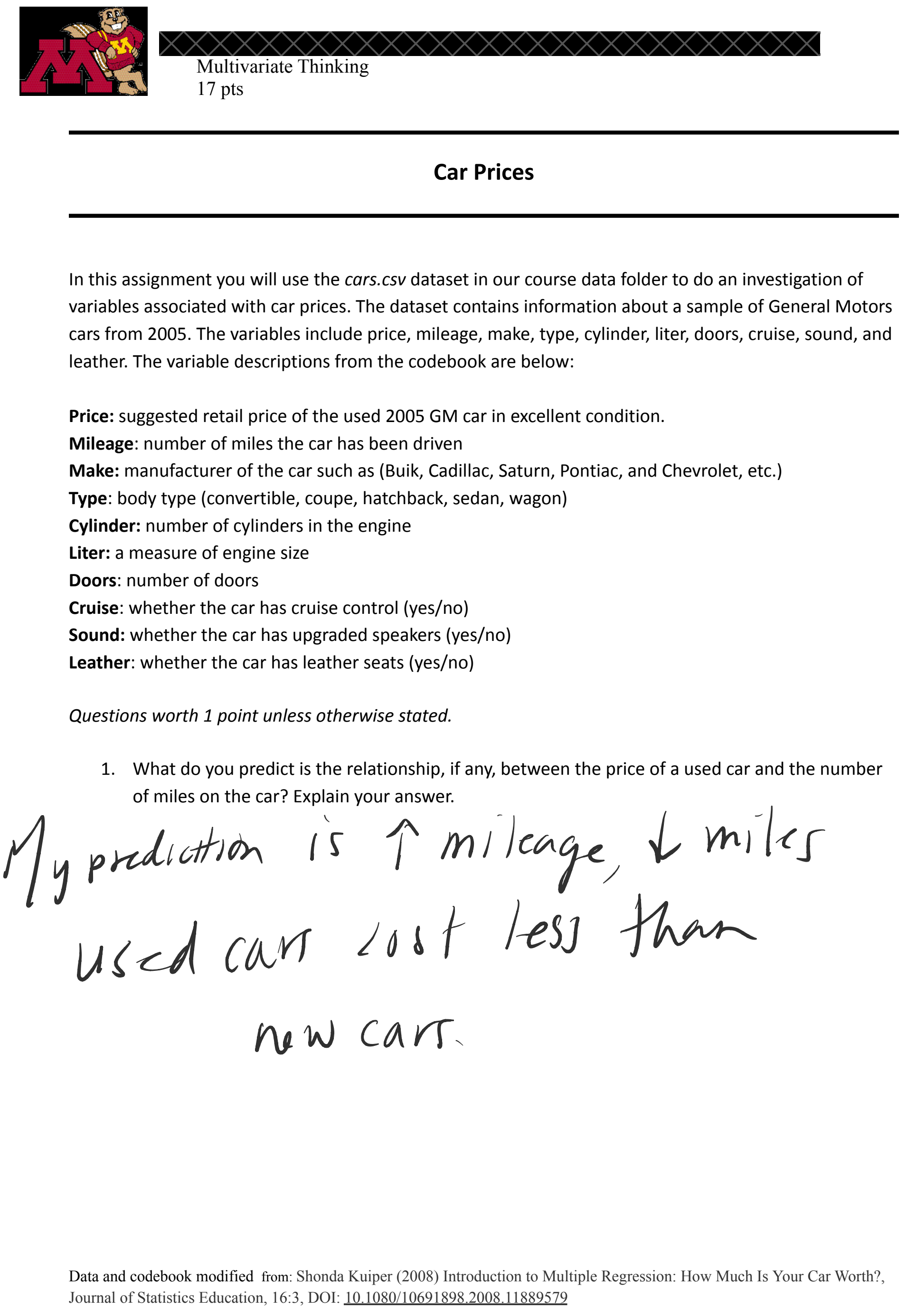
6. Based on your DAG and the plot above, what variable(s) do you think are associated with the price of a house? Justify your answer using evidence from your plot.

Age, bed, sq ft,

7. What other variables do you think are associated with the price of a house that are not considered in this dataset?

Views, sq ft land, quality of neighborhood, schools

B.4.12 Round 4: Assignment 3



[redacted]

Multivariate Thinking
17 pts

Car Prices

In this assignment you will use the *cars.csv* dataset in our course data folder to do an investigation of variables associated with car prices. The dataset contains information about a sample of General Motors cars from 2005. The variables include price, mileage, make, type, cylinder, liter, doors, cruise, sound, and leather. The variable descriptions from the codebook are below:

Price: suggested retail price of the used 2005 GM car in excellent condition.
Mileage: number of miles the car has been driven
Make: manufacturer of the car such as (Buik, Cadillac, Saturn, Pontiac, and Chevrolet, etc.)
Type: body type (convertible, coupe, hatchback, sedan, wagon)
Cylinder: number of cylinders in the engine
Liter: a measure of engine size
Doors: number of doors
Cruise: whether the car has cruise control (yes/no)
Sound: whether the car has upgraded speakers (yes/no)
Leather: whether the car has leather seats (yes/no)

Questions worth 1 point unless otherwise stated.

1. What do you predict is the relationship, if any, between the price of a used car and the number of miles on the car? Explain your answer.

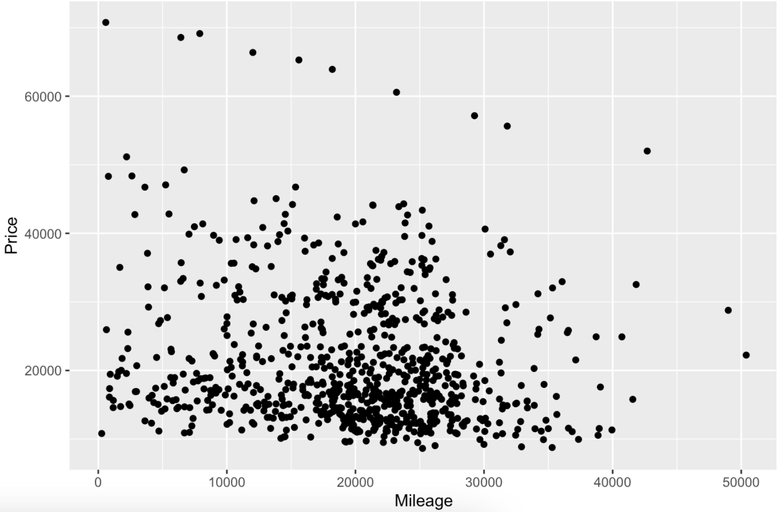
My prediction is ↑ mileage, ↓ miles used cars cost less than new cars.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts

2. Create a scatterplot with price on the y-axis and mileage on the x-axis.



3. Describe your scatterplot (linearity, strength of relationship, and slope).

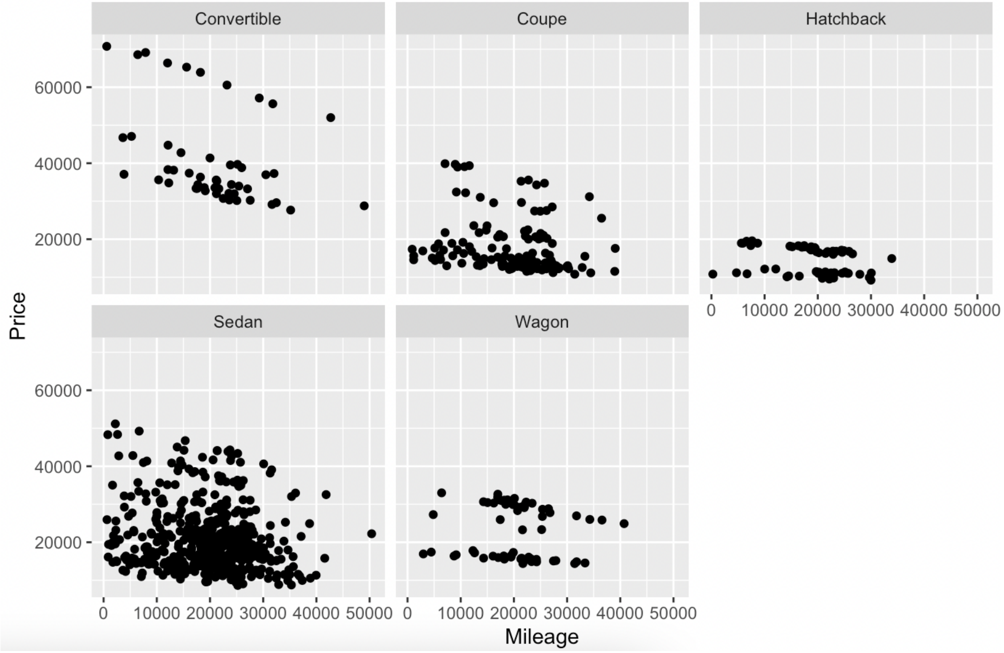
weak/mod neg slope, not too linear

4. Building on the code from your plot in number 2, facet on *type* of car.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts



5. Is this the same relationship we saw between mileage and price that we saw before we faceted on type? Explain why or why not.

Still negative relationship, Hatchback less (-) Sedan lots of variability

6. Choose a fourth variable from the list of variables above. Create a DAG to propose relationships among the four variables. Paste it below.

total explanatory

Type
Mileage → Price
Leather

7. Explain your proposed relationships.

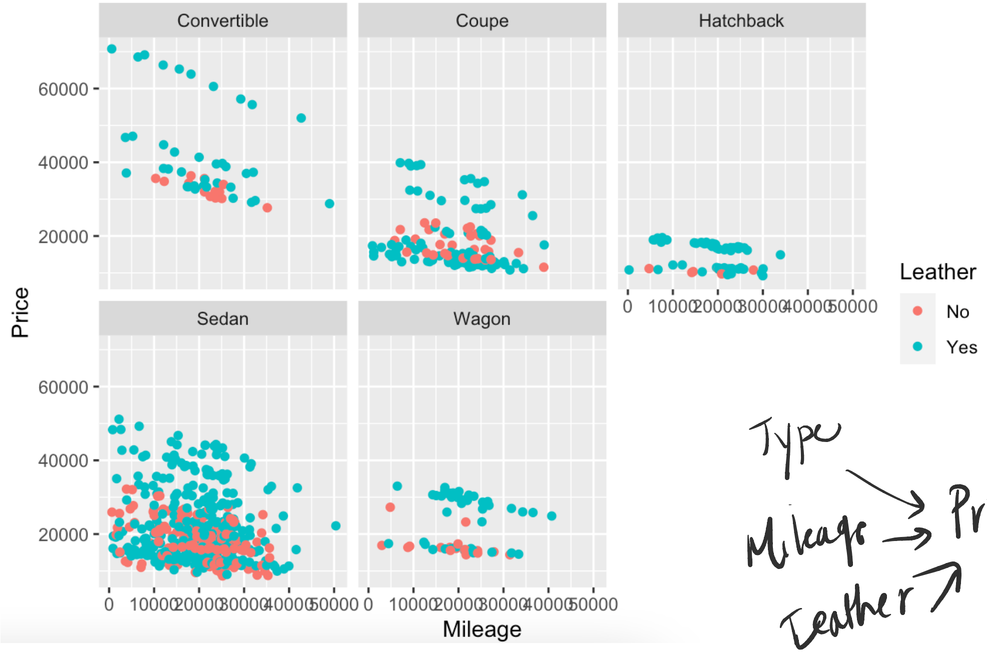
→ make only about new variable

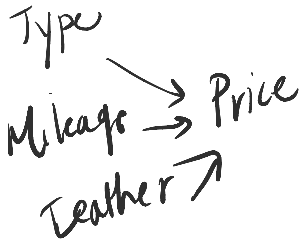
8. Create a plot incorporating the four variables. **(4pts)**
   a. Put a title on the plot
   b. Use a theme other than the default theme
   c. Use a colorblind friendly color palette
   d. Add a caption with the source of the data to the plot

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts





9. Draw your final DAG to represent how your 4 variables affect price and each other. Paste it below.

10. Provide a justification for your DAG using your plot as evidence.

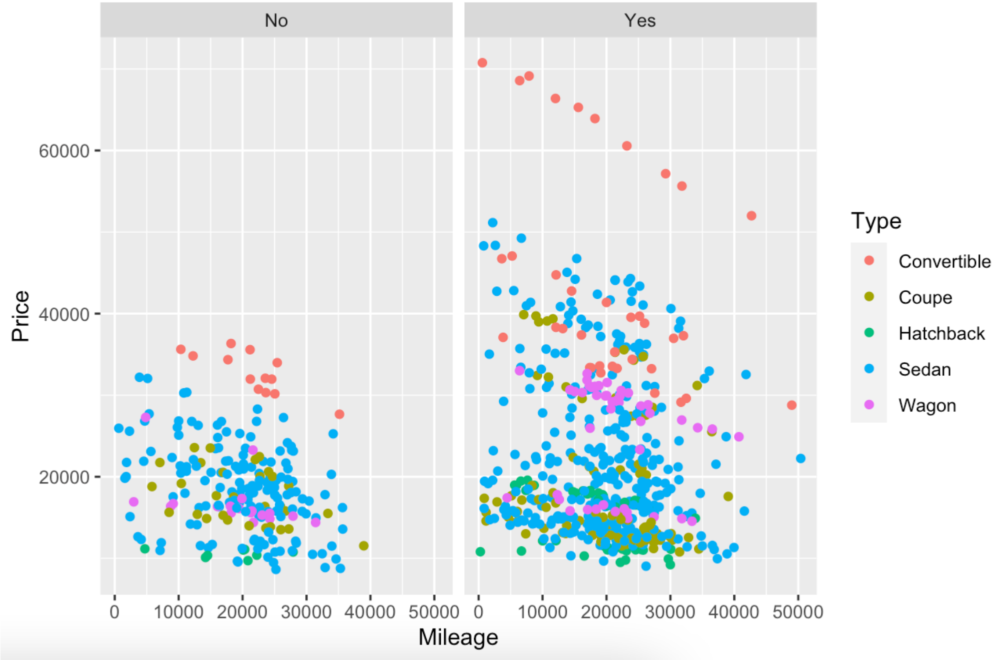
Type seems to not affect — Maybe hatch backs don't go as high, not enough, type doesn't affect leather logically not here no

11. Create a new plot using your same 4 variables (map different variables to the x axes, colors, size, or facet on a different variable), but keep price mapped to the y-axis. Be sure to update the title, colors, and theme as well.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[redacted]

Multivariate Thinking
17 pts



12. Do you think your plot in Question #9 or Question #12 depicts the relationships among the variables better? Explain your answer.

#9 because visually you have more info in a convenient way two colors is easier to see them type as color.

13. Are there any other variables (not in the current data set) that you think affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars? Explain your answer. *(Limit your response to 5 sentences or less)*

used v. new, → mileage → price

14. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions (based on your plot) would you recommend they buy it? Explain your answer below using evidence from your final plot.

sound/older, type of tires

Hatchback, coupe / wagon, no

but great deal on convertible

sedan is high but not unusual

seats should be leather

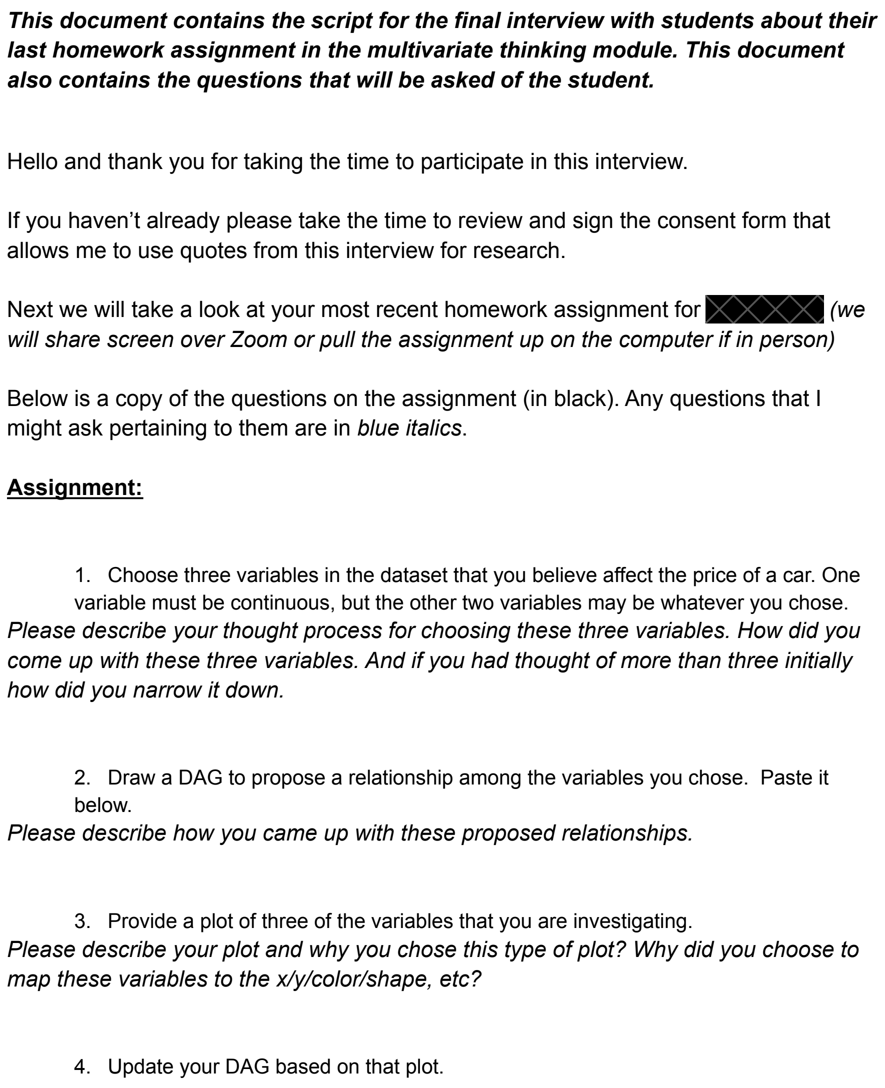
convertible nope for leather

Keep mileage below 30,000 for both

Any relation between type & other car would be difficult.

Data and codebook modified from: Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth?, Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

B.5 Final Interview Protcol



This document contains the script for the final interview with students about their last homework assignment in the multivariate thinking module. This document also contains the questions that will be asked of the student.

Hello and thank you for taking the time to participate in this interview.

If you haven't already please take the time to review and sign the consent form that allows me to use quotes from this interview for research.

Next we will take a look at your most recent homework assignment for [redacted] *(we will share screen over Zoom or pull the assignment up on the computer if in person)*

Below is a copy of the questions on the assignment (in black). Any questions that I might ask pertaining to them are in *blue italics*.

<u>**Assignment:**</u>

1. Choose three variables in the dataset that you believe affect the price of a car. One variable must be continuous, but the other two variables may be whatever you chose.

Please describe your thought process for choosing these three variables. How did you come up with these three variables. And if you had thought of more than three initially how did you narrow it down.

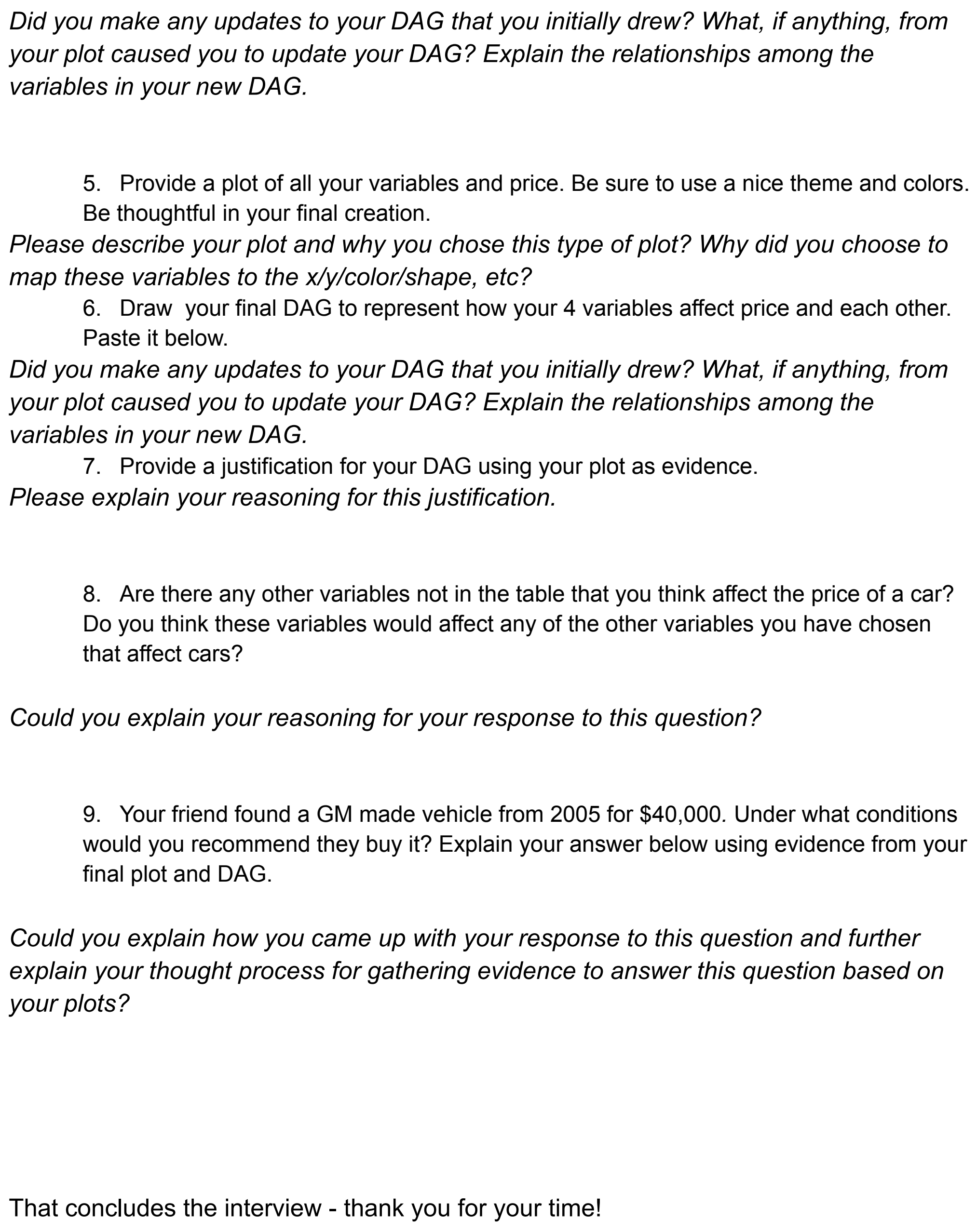
2. Draw a DAG to propose a relationship among the variables you chose. Paste it below.

Please describe how you came up with these proposed relationships.

3. Provide a plot of three of the variables that you are investigating.

Please describe your plot and why you chose this type of plot? Why did you choose to map these variables to the x/y/color/shape, etc?

4. Update your DAG based on that plot.



Did you make any updates to your DAG that you initially drew? What, if anything, from your plot caused you to update your DAG? Explain the relationships among the variables in your new DAG.

5. Provide a plot of all your variables and price. Be sure to use a nice theme and colors. Be thoughtful in your final creation.

Please describe your plot and why you chose this type of plot? Why did you choose to map these variables to the x/y/color/shape, etc?

6. Draw your final DAG to represent how your 4 variables affect price and each other. Paste it below.

Did you make any updates to your DAG that you initially drew? What, if anything, from your plot caused you to update your DAG? Explain the relationships among the variables in your new DAG.

7. Provide a justification for your DAG using your plot as evidence.

Please explain your reasoning for this justification.

8. Are there any other variables not in the table that you think affect the price of a car? Do you think these variables would affect any of the other variables you have chosen that affect cars?

Could you explain your reasoning for your response to this question?

9. Your friend found a GM made vehicle from 2005 for $40,000. Under what conditions would you recommend they buy it? Explain your answer below using evidence from your final plot and DAG.

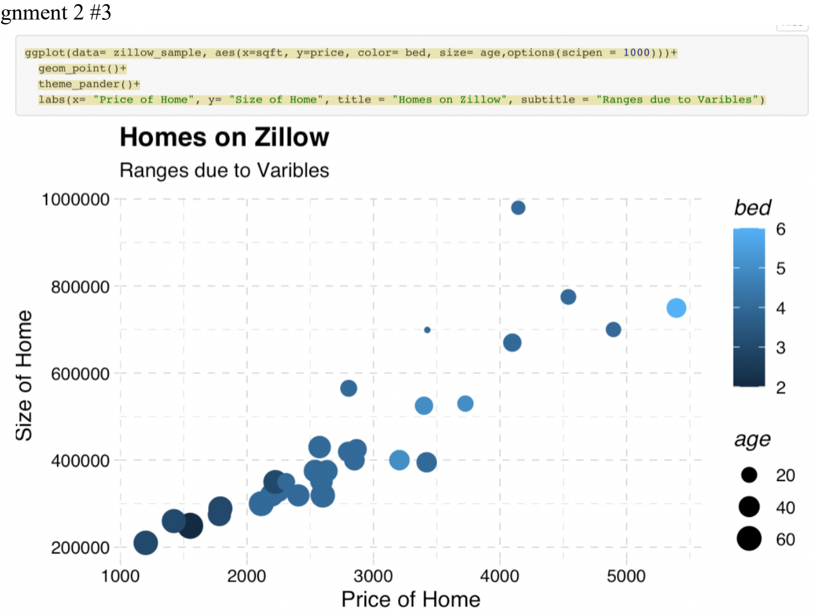
Could you explain how you came up with your response to this question and further explain your thought process for gathering evidence to answer this question based on your plots?

That concludes the interview - thank you for your time!

B.6 Codebook with Examples

LO1/correct: Graph was created with 3 or more variables in the way specified by the assignment or in a logical way given the variables.

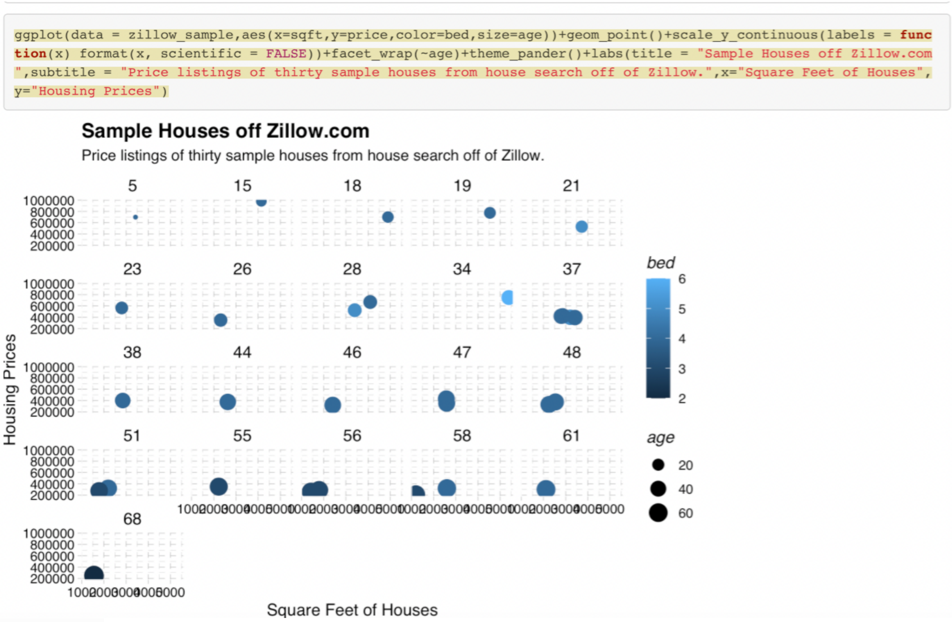
Assignment 2 #3



```
ggplot(data= zillow_sample, aes(x=sqft, y=price, color= bed, size= age,options(scipen = 1000)))+
  geom_point()+
  theme_pander()+
  labs(x= "Price of Home", y= "Size of Home", title = "Homes on Zillow", subtitle = "Ranges due to Varibles")
```

LO1/incorrect: Graph was not created with 3 or more variables in the way specified by the assignment or in a logical way given the variables.

Assignment 2 #3



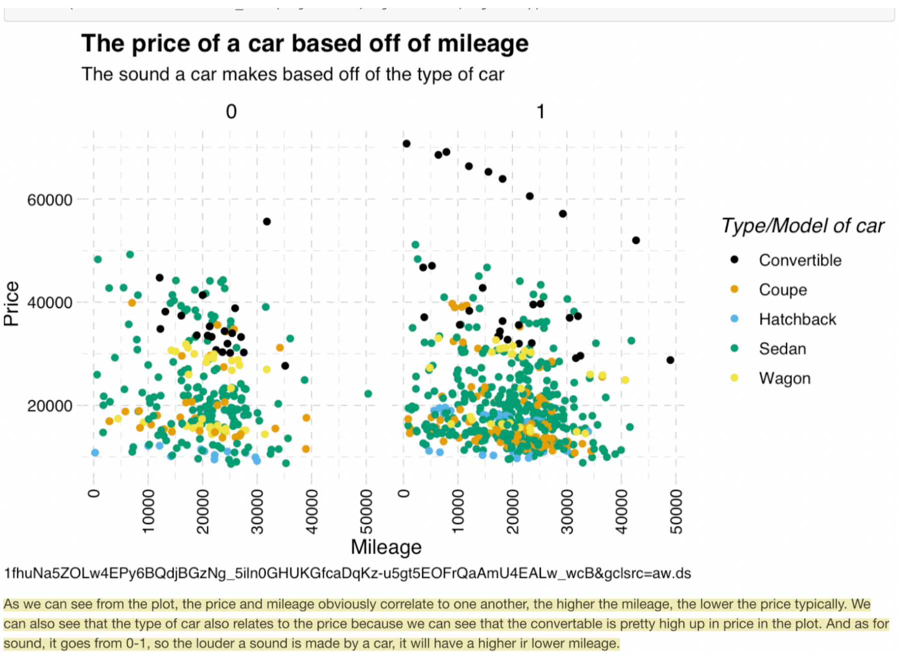
```
ggplot(data = zillow_sample,aes(x=sqft,y=price,color=bed,size=age))+geom_point()+scale_y_continuous(labels = func
tion(x) format(x, scientific = FALSE))+facet_wrap(~age)+theme_pander()+labs(title = "Sample Houses off Zillow.com
",subtitle = "Price listings of thirty sample houses from house search off of Zillow.",x="Square Feet of Houses",
y="Housing Prices")
```

LO2\Correct: Provides a description of the variables in a way that is aligned with that is depicted in the graph

Assignment 3 #15
“Based on the plot, spending $40,000 on a 2005 GM would be reasonable only if the car is a convertible or some sedans. The car should have leather and have around 20,000 miles or less.”

LO2\plot-description-mismatch: Provides a description of the variables in a way that is not aligned with that is depicted in the graph

Assignment 3 #11



1fhuNa5ZOLw4EPy6BQdjBGzNg_5iln0GHUKGfcaDqKz-u5gt5EOFrQaAmU4EALw_wcB&gclsrc=aw.ds

As we can see from the plot, the price and mileage obviously correlate to one another, the higher the mileage, the lower the price typically. We can also see that the type of car also relates to the price because we can see that the convertable is pretty high up in price in the plot. And as for sound, it goes from 0-1, so the louder a sound is made by a car, it will have a higher ir lower mileage.

LO2\aggregate-reasoning: Provides a description of the relationships among the variables at a high

Assignment 2 # 5
“According to the visualization, the bigger the house, the more expensive it becomes. This relationship between price and square footage was clearly shown through a strong linearity on the graph. Also, the larger points (the older houses) were placed lower on the graph, proving that older houses are cheaper and newer homes are more expensive. The colors on the graph are also quite interesting. The lighter shades were higher up, indicating that lots of bedrooms are costly.”

LO2\case-reasoning: Provides a description for the variables on an individual level singling out certain cases in the graph

Assignment 2 # 5
“The age of the house greatly affects the price making. I say this because the cheapest house are over 40 years old with bed size of around 2. The most expensive houses has a large square feet, 4+ beds, and the houses age is under 20 years. However, there is an outlier with a house around $700,000 with the house age of around 34. However, I think it still costs that much because the bed size is 6 and it has the largest square feet.”

LO2\considering-all-variables: Provides a description of all the variables in the plot leaving none out

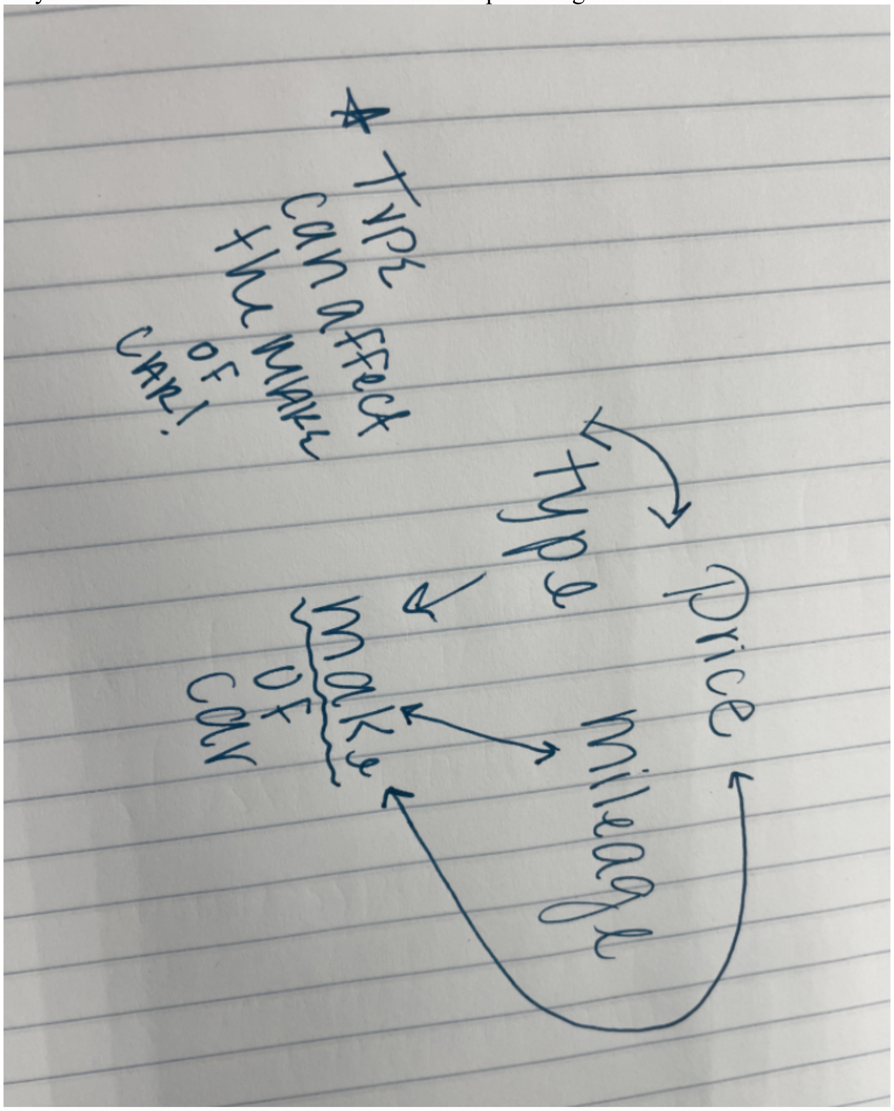
Assignment 2 #6

“The number of beds and the square feet of a listing all have a positive correlation with each other. As price goes up the square feet go up the number of beds go up. Age however has a negative correlation, as the price increases, the age tends to be lower. The number of beds is also affected by square feet most likely for the presented need of accomodating the space required to fit 5+ beds. Age also affects the square feet, looking at the graph, older homes tend to be smaller than those built 20ish years ago.”

LO2\not-considering-all-variables: Provides a description of all the variables in the plot leaving one or more out of the description

Assignment 3 #15
“I would say that if the car is less than 20K in the mileage for 40,000 dollars”

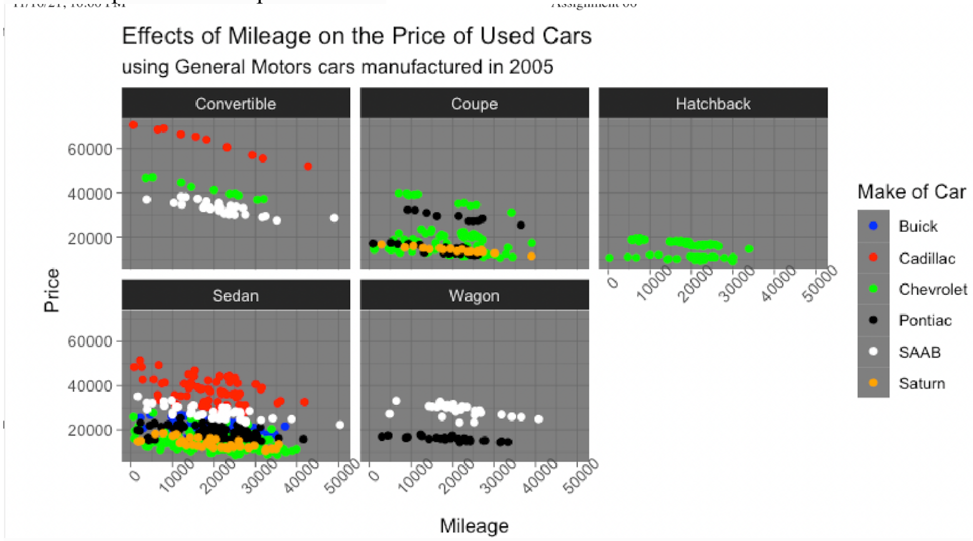
LO2\dag-description-mismatch: Provides a description of a DAG which does not match that they have drawn in that DAG as the relationships among the variables



"In my opinion, through my DAG, I believe that the make of a car can have an affect on price, mileage, and type. Price and make of a car can have a huge affect on one another. For example, a Buick is much less expensive than a Cadillac which shows that make can affect the price. The make can also affect mileage as if it is a more expensive car, there is an expectation that nicer cars will probably have a better mileage. Lastly, I claimed that the type of car can affect the make of it and vice versa (I meant to draw another arrow towards "type" from "make of car" in my

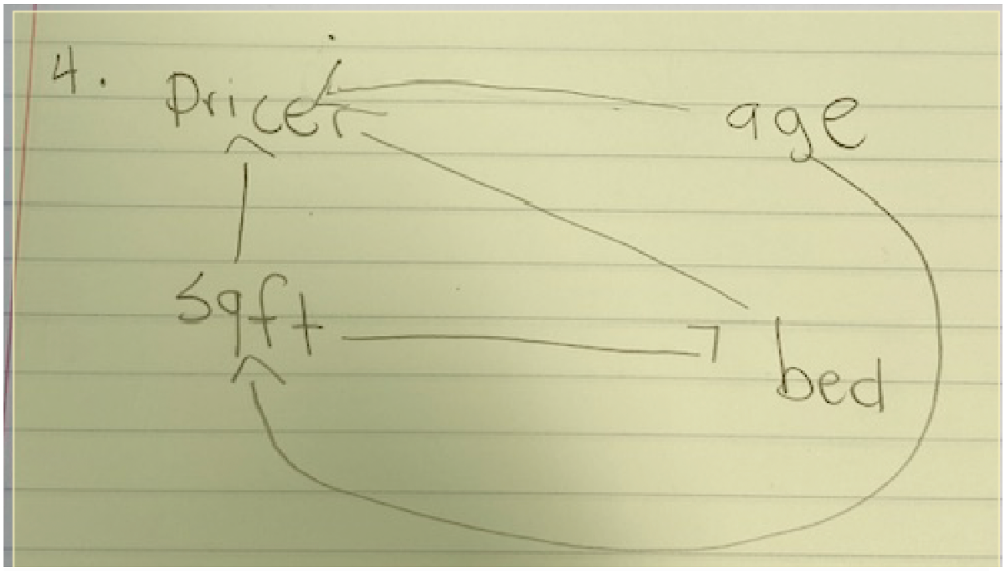
DAG). For example, in this data, it looks as though Buicks are more of the Sedan types and Cadillacs are more of the convertible types, and Chevrolets are the hatchback types!”

LO2\partially-correct: Provides an incomplete but correct description of the nature of the relationships seen in the plot or DAG



“In the plot above, make affects the price of the car because all of the corresponding points for each car make follow the same linear trend lines.”

LO2\plausible: Provides a description of a DAG describing all variables in a way that is plausible



I first drew an arrow to connect square feet and price. If you see on the plot the lower the square footage is the price stays downs. Next, square footage affects the number of beds. In between 2000 and 3000 square feet there is 4 to 5 beds. Once you get to over 5000 square feet you see a 6 bed. The higher the square feet the more beds you can have. Beds also have an affect on price. The more bedrooms a house has the higher the price is going to be. Lastly, age affects square feet because the younger house is the more square footage it has. Age and price effect each other because newer houses tend to cost more.

LO3\correct: Correctly identified data as observational

Assignment 3 #1
“The data we are using in the cars.csv dataset seem to be more observational than experimental. The data is observational because the researcher is not manipulating any of the variables.”

LO3\incorrect: Did not identify data as observational

Assignment 3 #1
“It is experimental data because variables are under control of tester. There is likely a relationship between these variables, so yes to casual claims I think.”

LO3\partially correct: Identified data as observational, but gave wrong reasoning why it was observational

Assignment 3 #1
“This is observational data because we are looking at a real life situation. It has to be observational, becuase the data would lose it’s significance if we made these data points up, we are trying to learn something based on what we see around us.”

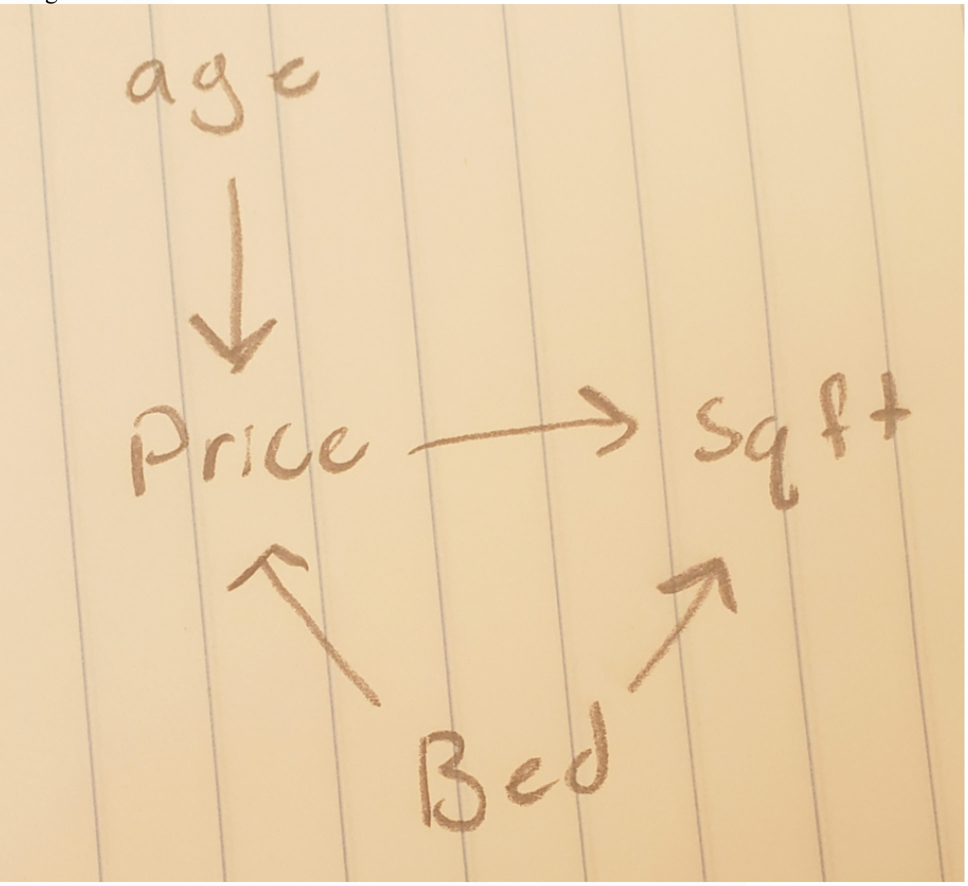
LO4\correct: Describes how we cannot make causal claims with observational data

Assignment 3 #1
"We can only make cause and effect statements when working with an experiment. Thus, we cannot make causal inferences with this data."

LO4\incorrect: Describes how we can make causal claims with observational data

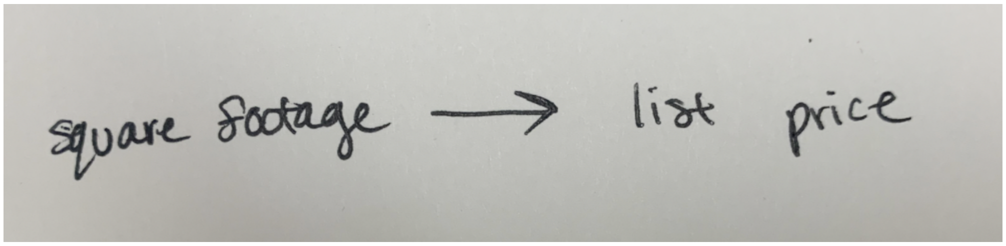
Assignment 3 #1
"We can use this data to make causal claims because we can examine how the variables affect each other."

LO5\directed-arrows: DAG created contains directed arrows
Assignment 2 #2

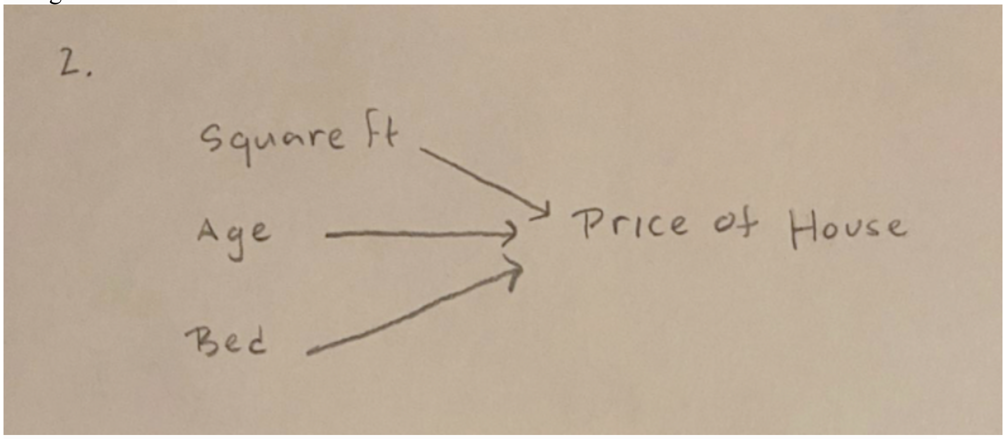


LO5\forgot-variable: DAG created does not contain all needed variables

Assignment 2 #2

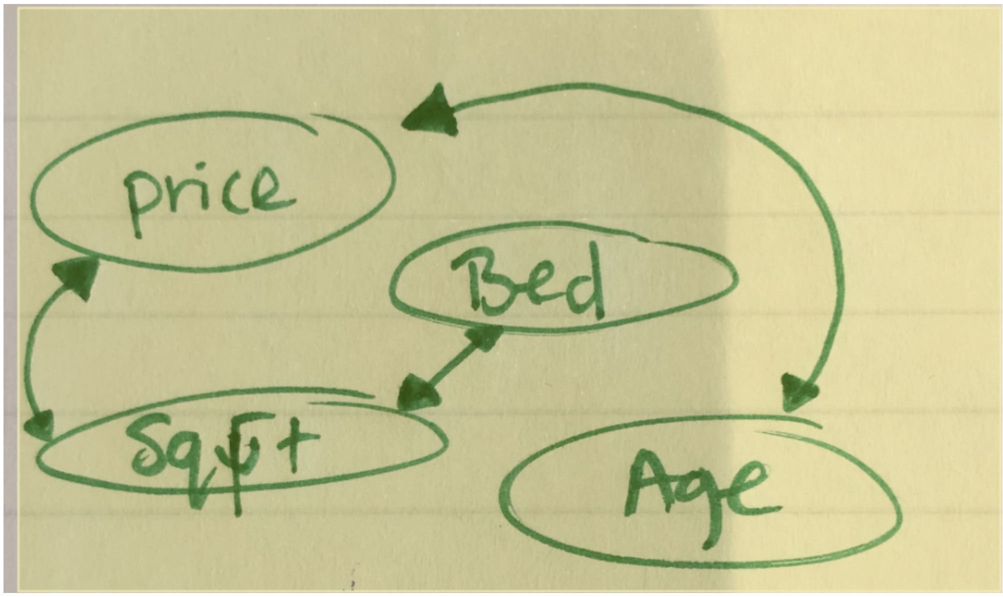


LO5\no-relationships-IVs: DAG created had no relational arrows between any variables except those with the outcome variable
Assignment 2 #2

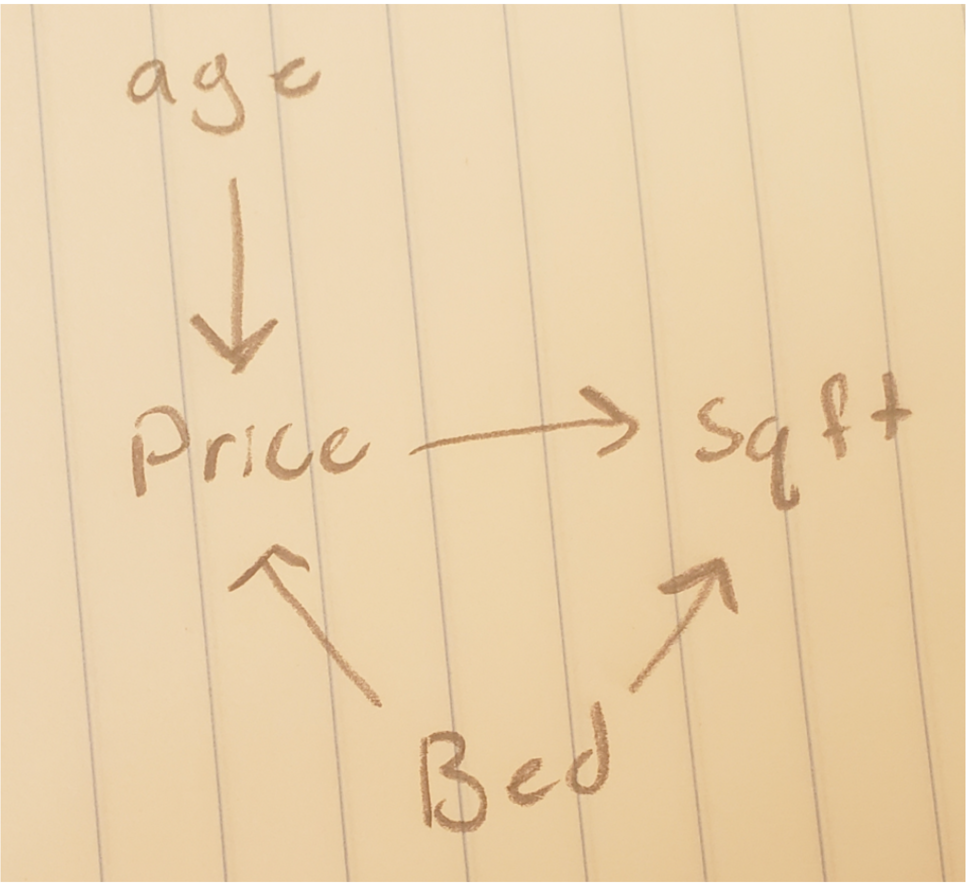


LO5\not-correct-arrows: DAG created displays casual arrows that could not possibly exist

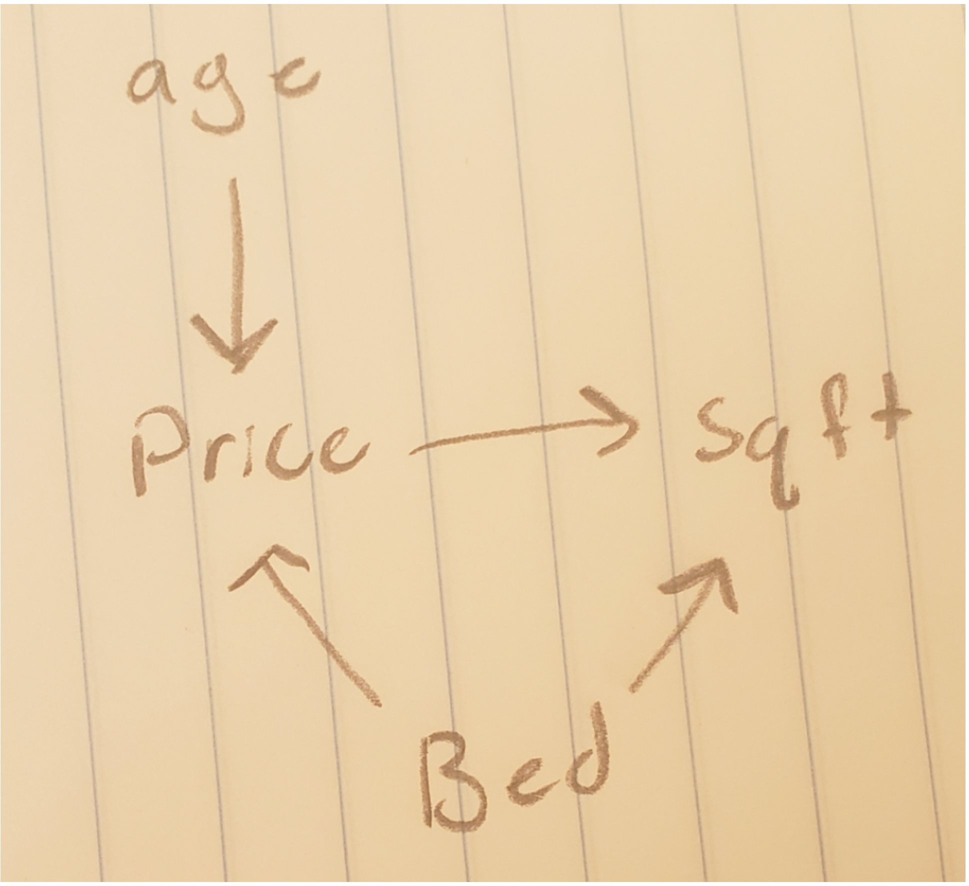
Assignment 2 #2



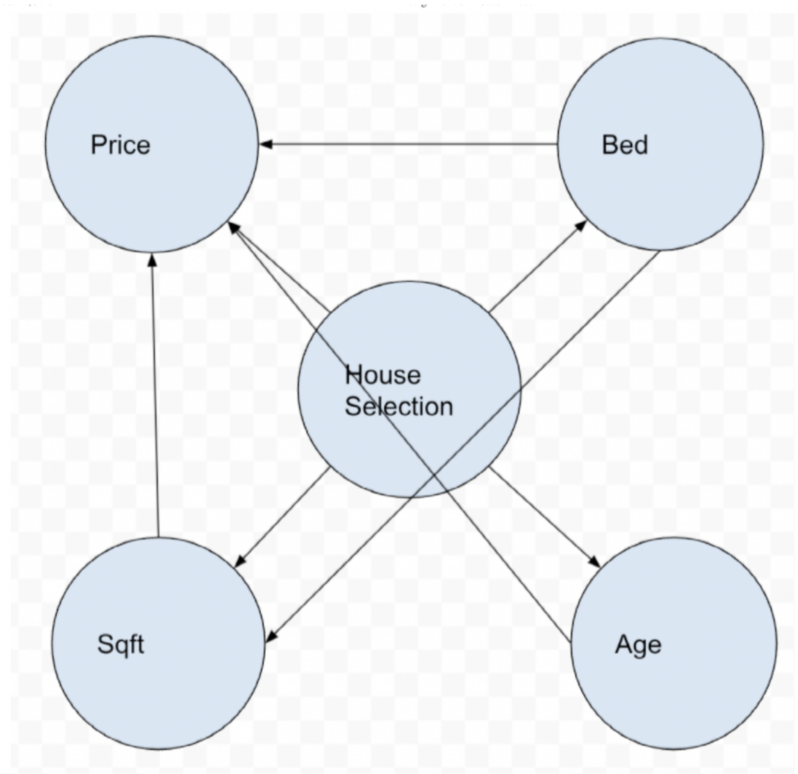
LO5\plausible: DAG created is plausible (opposite of not-correct-arrows)
Assignment 2 #2



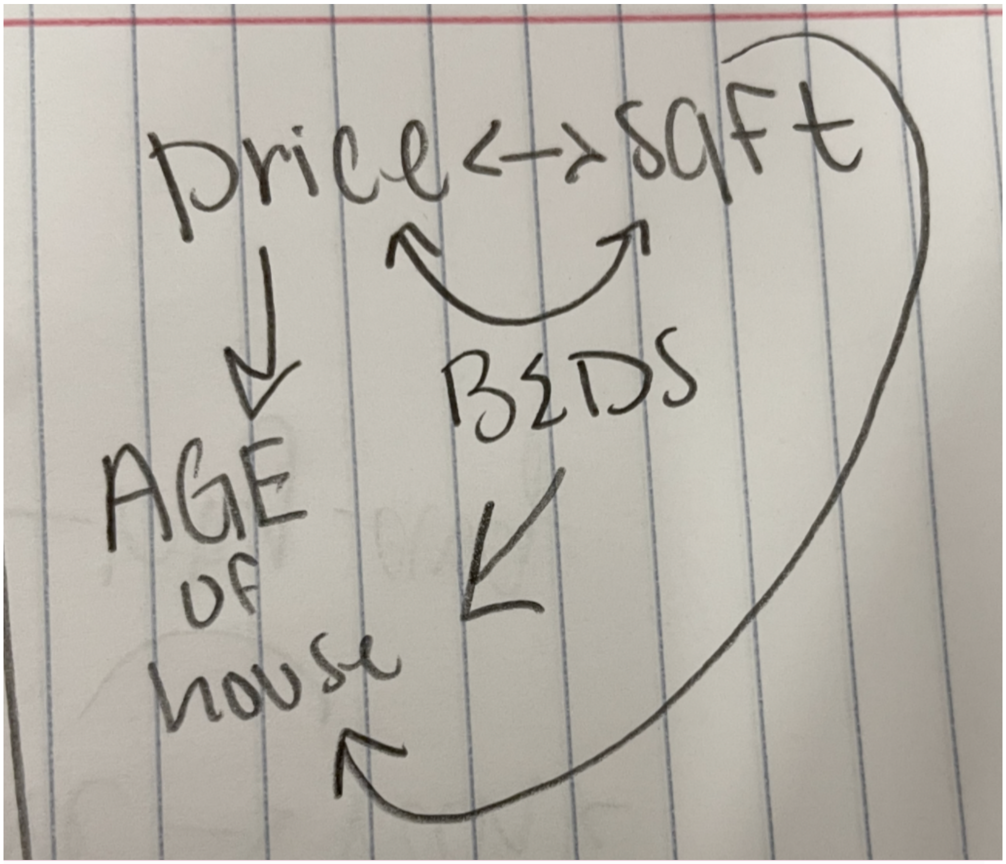
LO5\relationships-between-IVs: DAG created shows relational arrows among any variables and with the outcome variable
Assignment 2 #2



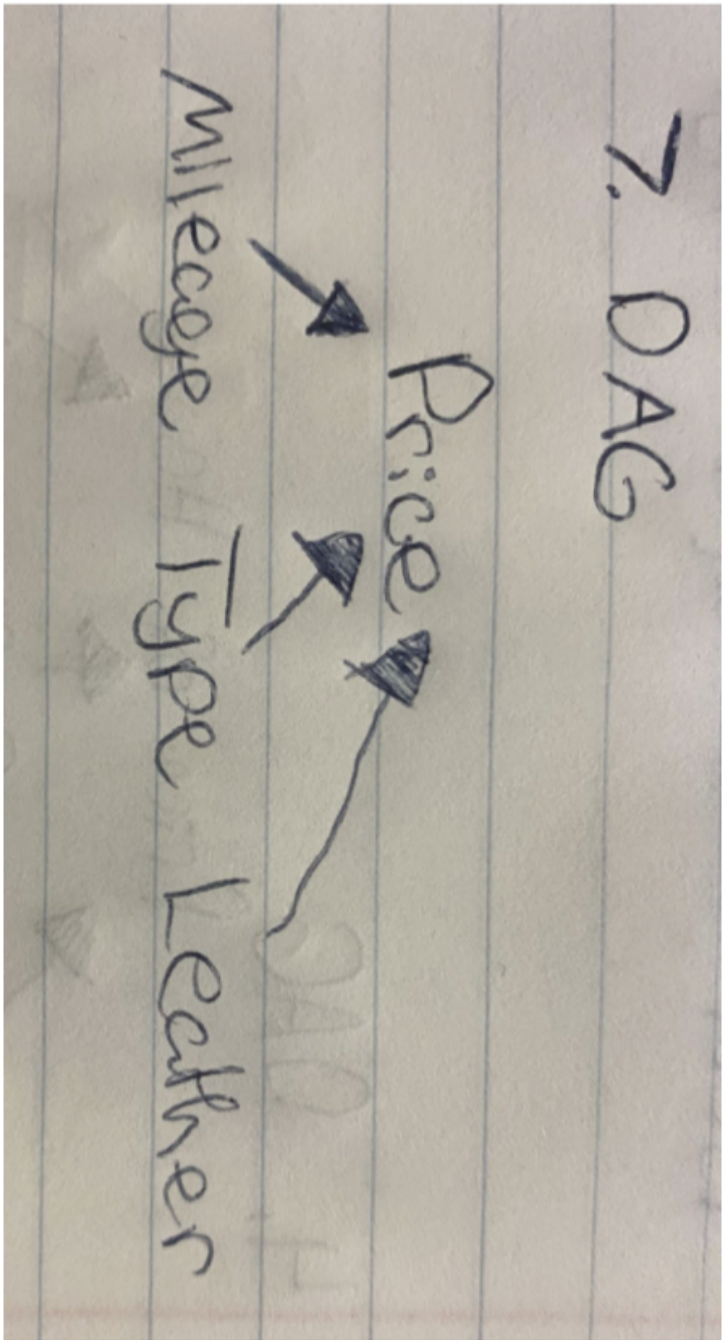
LO6\incorrect: DAG description does not match the plot/DAG provided
Assignment 2 # 6

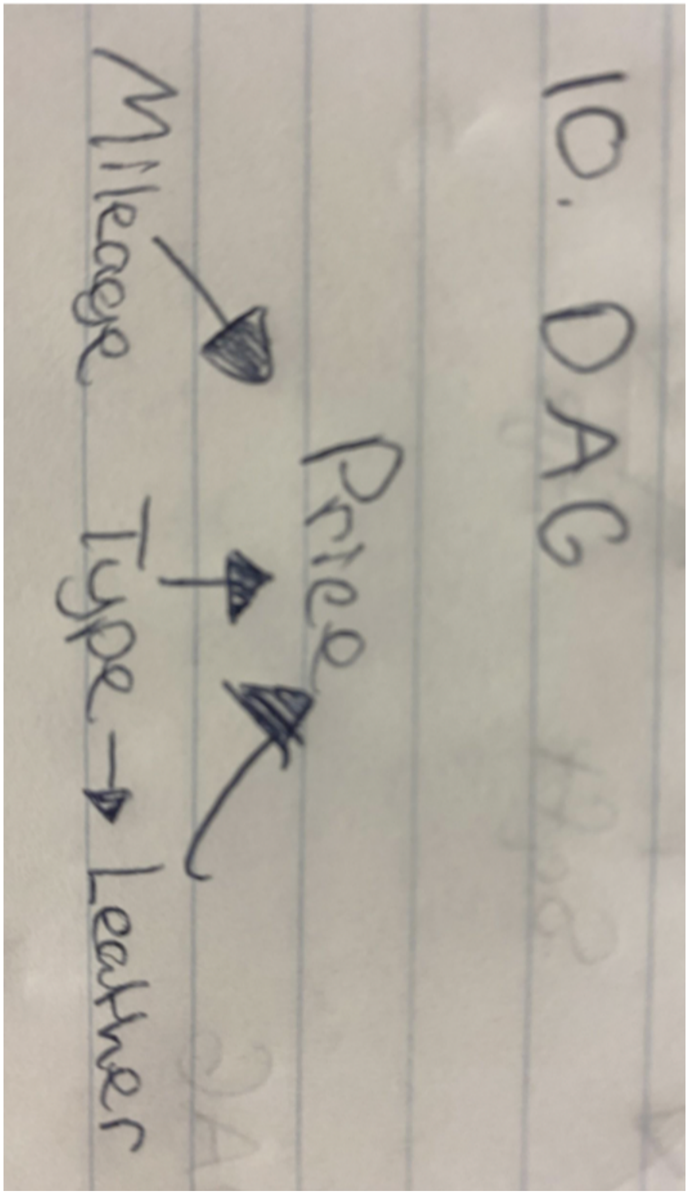


LO6\incorrect-dag-arrows: DAG created displays casual arrows that could not possibly exist given the information displayed in the plot created
Assignment 2 # 6



LO6\updated-DAG: DAG was updated from a previous question after evaluating a graph of the variables
Assignment 3 #7 and updated in #10


7. DAG
Price
Mileage
Type
Leather



LO7\plausible: Described possible/logical variables that could affect the system of variables in a meaningful way

Assignment 2 #7

"I feel like Location would factor in, specifically like a popular City, and then scenery, like if it's by a lake, forest, ocean, etc. If those two seem too similar, I'd say yard size would maybe work too."

LO7\not-plausible: Described variables that could not affect the system of variables in a meaningful way

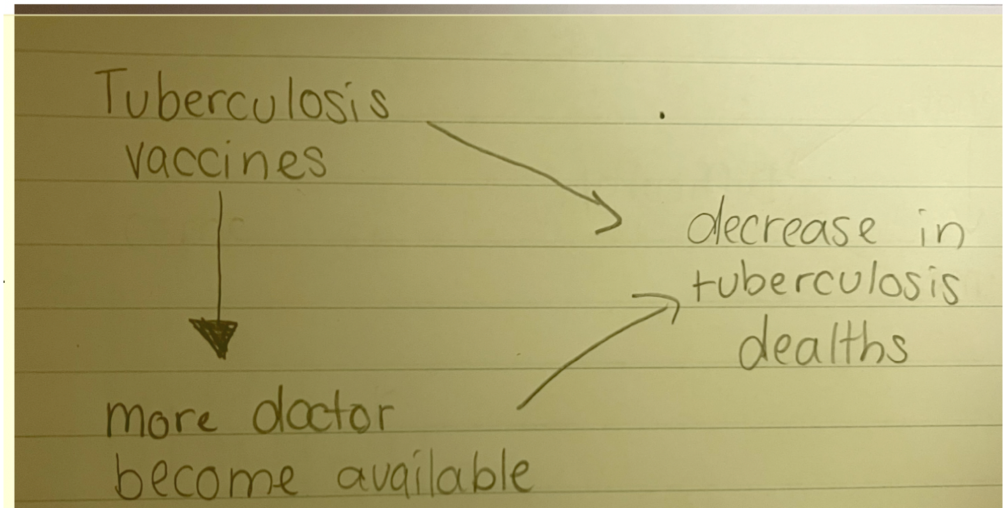
Assignment 2 #7
"Variables include, square feet and age. This is because these are the main factors that determine the price of a house, other variables are unnecessary."

Context-interference: Description brings in outside context the conflicts with what's in the DAG/graph or in some way is adding to their response

Assignment 3 #8:
"I think the manufacturer of the car itself is a brand, if the manufacturer has a good business reputation, the price of the car may be higher; and the car it produced should be more durable which means it takes longer duration to be maintained so the mileage should be left out more."

Variable-level-confusion: Student expresses incorrect reasoning around the variable or the level of interest

Assignment 1 #7


Tuberculosis vaccines
decrease in tuberculosis dealths
more doctor become available

B.7 Table of Coding Results Assignments

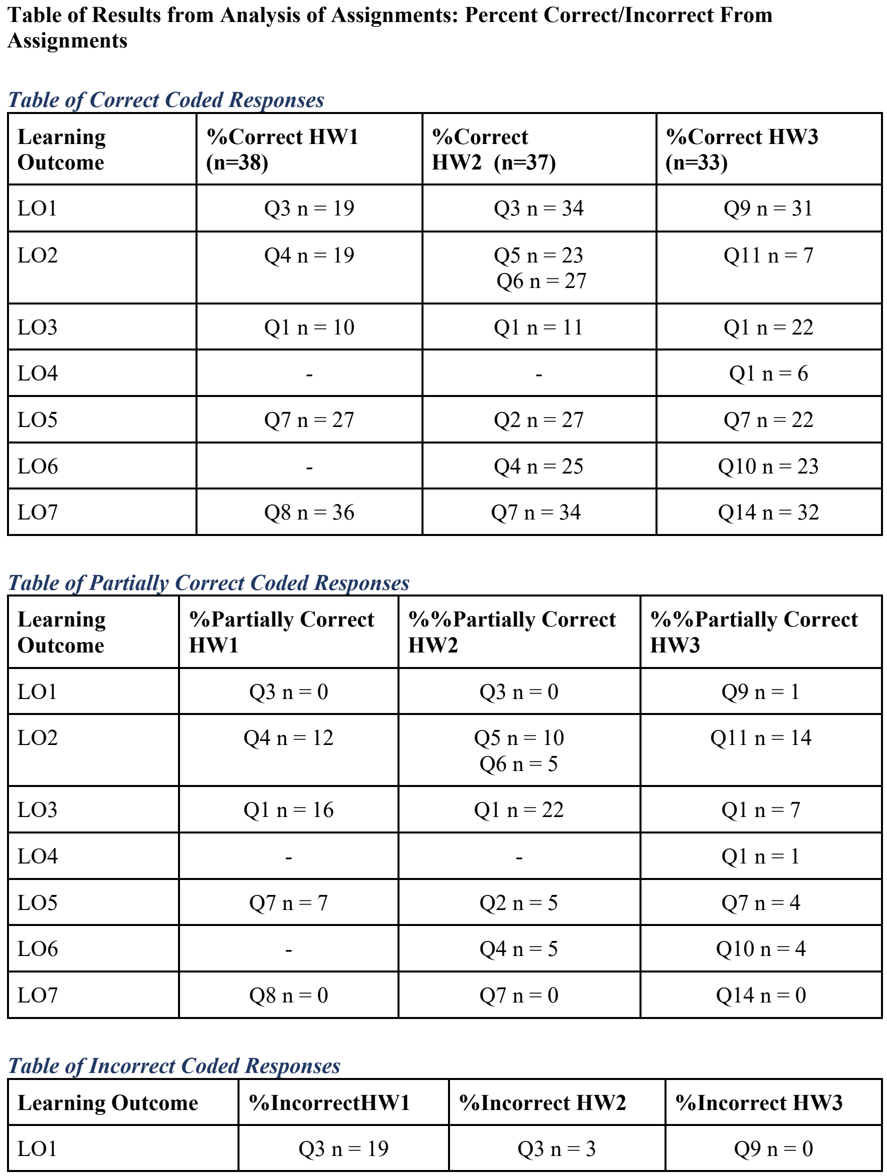


Table of Results from Analysis of Assignments: Percent Correct/Incorrect From Assignments

Table of Correct Coded Responses

Learning Outcome	%Correct HW1 (n=38)	%Correct HW2 (n=37)	%Correct HW3 (n=33)
LO1	Q3 n = 19	Q3 n = 34	Q9 n = 31
LO2	Q4 n = 19	Q5 n = 23 Q6 n = 27	Q11 n = 7
LO3	Q1 n = 10	Q1 n = 11	Q1 n = 22
LO4	-	-	Q1 n = 6
LO5	Q7 n = 27	Q2 n = 27	Q7 n = 22
LO6	-	Q4 n = 25	Q10 n = 23
LO7	Q8 n = 36	Q7 n = 34	Q14 n = 32

Table of Partially Correct Coded Responses

Learning Outcome	%Partially Correct HW1	%%Partially Correct HW2	%%Partially Correct HW3
LO1	Q3 n = 0	Q3 n = 0	Q9 n = 1
LO2	Q4 n = 12	Q5 n = 10 Q6 n = 5	Q11 n = 14
LO3	Q1 n = 16	Q1 n = 22	Q1 n = 7
LO4	-	-	Q1 n = 1
LO5	Q7 n = 7	Q2 n = 5	Q7 n = 4
LO6	-	Q4 n = 5	Q10 n = 4
LO7	Q8 n = 0	Q7 n = 0	Q14 n = 0

Table of Incorrect Coded Responses

Learning Outcome	%IncorrectHW1	%Incorrect HW2	%Incorrect HW3
LO1	Q3 n = 19	Q3 n = 3	Q9 n = 0

LO2	Q4 n = 5	Q5 n = 3 Q6 n = 5	Q11 n = 9
LO3	Q1 n = 8	Q1 n = 4	Q1 n = 4
LO4	-	-	Q1 n = 21
LO5	Q7 n = 0	Q2 n = 2	Q7 n = 4
LO6	-	Q4 n=5	Q10 n = 4
LO7	Q8 n = 1	Q7 n = 2	Q14 n = 0

Table of Non-response Coded Responses

Learning Outcome	**%**NA's**HW1**	**%**NA's **HW2**	**%**NA's**HW3**
LO1	Q3 n = 0	Q3 n = 0	Q9 n = 1
LO2	Q4 n = 2	Q5 n = 1 Q6 n = 0	Q11n = 3
LO3	Q1 n = 4	Q1 n = 0	Q1 n = 0
LO4	-	-	Q1 n = 5
LO5	Q7 n = 4	Q2 n = 3	Q7 n = 2
LO6	-	Q4 n = 2	Q10 n = 2
LO7	Q8 n = 1	Q7 n = 1	Q14 n = 1

Note: in HW2 10 students combined numbers 5 and 6