WHERE IS WALDO IN STATISTIC CLASS? USING MAPS TO EXPORE MODERN DATA TYPES.

Megan Mocko
University of Florida, USA
Megan.Mocko@warrington.ufl.edu

To be prepared for the modern world, students need to learn how to work with multivariate relationships as well as geographical and text data. In this paper, three class activities to investigate geographic data in conjunction with other standard categorical and quantitative data are described. The activities are described using the statistical software JMP, but modifications are given for using R. Modifications for undergraduate and graduate level work are also given. The activities have students explore data at an international level as well as local level. This paper illustrates classroom activities that demonstrate necessary scaffolding to move students beyond univariate and bivariate understandings as the curriculum shifts to keep up with modern data. The results of a ten-question survey to graduate and undergraduate students are given.

INTRODUCTION

"Where's Waldo?" is an illustrated book series written by Martin Handford in which the viewer looks for Waldo in his red strip shirt over a crowded landscape – a geographical representation. Although many students see data in statistics classes, many do not see data in a geographic landscape or in a more modern interpretation. Gould (2010) stated that our current students expect to explore and understand data that they see every day. Ridgeway (2015) says, "Educators should place less emphasis on small samples and linear models and more emphasis on large samples, multivariate description and data visualization." Students in business and other disciplines will need to know how not only to deal with quantitative and categorical variables but also other forms of data, such as spatial or temporal. For example, a business in South Africa was able to improve non-regulated transportation by recording geographic locations, distance to stops and number of passengers picked up and dropped off in a non-standardized taxi system. After reviewing the data, they were able to reduce gas used and smooth out the routes (Scott-Clarke & Lewis, 2019*)*.

In this paper, three activities are described that have students combine traditional data such as quantitative and discrete variables with geographic locations. The learning objectives for these three activities were students performing tidying techniques, textual analysis, as well as creating visualizations or maps to explore multivariate relationships.

These activities were completed with undergraduate and graduate students during Spring of 2020 at a large research institution in the southeastern region of the United States. The undergraduate students were completing a second semester statistics course for Business and the graduate students were completing a seven-week Business statistics course. The graduate class was small (43 students); whereas the undergraduate class had 660 students enrolled.

METHOD

The three activities included data from three different sources available on the web. Each of these activities were conducted with students who used JMP, but adaptions for R are also given.

Activity One

The first data set comprised loans that were given by Kiva, which is a crowdsource lending organization for people around the world to receive small loans to help start or grow their business. According to their website in February 2020, there were 3.5 million borrowers and 1.8 million lenders providing 1.41 billion dollars in loans. The data was initially released in 2018 as a data set for a competition on Kaggle (https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding). The entire population distribution of all loans (671,206 loans) from 1/2004 to 6/2017 are included. Due to the size of the data set, Excel and JMP can become overloaded. Therefore, the data set was limited to only thirteen countries spanning from Turkey to India.

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

For the undergraduate students', step-by-step instructions are given to complete the assignment. Additionally, the instructor worked through the activity with the students in a live lecture that was also recorded, and a 30-minute recorded help session was available. The assignment was posted as a quiz in the course management system, Canvas by Instructure. This also allowed for the scoring to be done automatically and instantaneously. The students were given three attempts and the highest attempt counted. In the activity, the students first explored the loan amounts for the entire thirteen countries combined using summary statistics and a histogram and then they compared the loan amounts by repayment method.

The next step was to examine how the loan amounts differed by gender which takes some data cleaning. In the original data set, gender is listed for everyone in the group. So, if there was one lender, it may be "female" or "male". However, if there are two lenders, you might get "female, female" or "male, male" or "male, female". This continued for each additional set of possible combinations. The initial data set had 67 combinations making it very difficult to compare. The students were led through the process to clean up the data, so that gender was listed in five groups: "one female", "one male", "group of men", "group of women", and "mixed group".

They then graphed the average loan amounts onto a map and used JMP tools to view the histograms, mean loan amounts, as well as number of loans per country. The next part of the assignment had the students look at the hashtags given for each loan. They first used the text explorer tool to determine what types of hashtags were used in the data set. They then downloaded from JMP's website a script that counted the number of times the word "Single" was listed.

For the graduate level class, they worked through the above exercise in class with the instructor. At each stage, the students were asked to state, "What do you notice?" and "What are you wondering?". These questions come from the National Council of Teachers of Mathematics, Math Forum. After completing the in-class exercise, the students worked in small groups (2 to 4 students) to complete the assignment outside of class. The students were given a choice between using a pre-selected group of countries or selecting their own from Kaggle.com.

The deliverable was a slide presentation or a word document. In this assignment, they were asked to show at least four different graphs and one map that showed something interesting. Additionally, they were to compare the countries using visualizations that showed only one variable, a bivariate and multivariate relationship, as well as the hashtag text analysis. From their analysis, they were to suggest where there might be a business opportunity.

To complete this activity in R, an R markdown document that has been partially completed is the starting point. The R Markdown document includes the example of the thirteen countries first, which is then altered for a different group of countries. The packages needed would include the *mosaic package*, *str_count*, *ggplot2* and *tidyverse*.

Activity Two

For the second activity, the students explored data from the World Bank. The World Bank has a website (https://databank.worldbank.org/source/world-development-indicators#) that allows viewers to download data from a large assortment of variables for many countries over time.

In the undergraduate class, the students were once again led through an exploration of the data through a quizzing assignment mechanism; however, this time they did not have the option to attend a live walk through of the assignment. This time the data set they were given to explore included the labor force, percent women and percent men for countries in South America. They were asked to first explore the trend over time for both men and women in an overlay format and then using a trellis (grid) display. They were then led through a process to create pivot tables to compute max values, join two files, and to incorporate the linear trends onto a map. The students were asked questions as they completed these tasks to have them investigate the trend and compare trends across countries.

For the graduate class, they did a similar exercise in class with the House Price Index in the United States as described in the JMP mastering video titled "Unlocking the Power of Graph Builder" (https://www.jmp.com/en_gb/events/ondemand/mastering-jmp/unlocking-the-power-of-graph-builder-2019.html). Next, they were asked to complete the assignment on their own using data of their choice from the World Bank. In their assignment, they needed to make a plot of time versus the variable they chose by country. The students then either wrote a 200 to 500-word description or recorded a video presentation. One of the nice benefits of allowing the students to choose their own countries to explore is that some of them picked their country of origin or ancestry.

For this activity in R, students would use some of the same packages and codes learned in activity one, but would add on the use of additional features of the *ggplot2* package to arrange maps and plots together into one image, and facet wrap to create a grid display to contain each of the plots in one image.

Activity Three

For the third activity, the students were asked to explore data closer to home. In the large enrollment undergraduate class, the students were asked to collect data about houses that were for sale on Realtor.com. The class was divided up so that each student would collect data from one of eight cities in the state of the Florida. For each city, the students would find the numbers of bed + bathrooms, square footage, lot size, style, number of cars that would fit into a garage, year built, days on Realtor.com, distance from elementary school, distance from middle school, distance from high school, latitude and longitude, description and price. The latitude and longitude is not listed on Realtor.com, but students were given instructions on how to enter the street address at https://www.latlong.net/ in order to find these coordinates.

Later in the semester, the students returned to this data and were asked to create a map, conduct a text analysis, and create a model to predict price. Each student then presented a 200 to 500-word description or a 3 to 5-minute video to explain their model. After the due date for the assignment, the students were asked to complete a peer review of another student in the class

For the smaller enrollment graduate classes, the classes worked in teams. The teams were given cities located in the southeastern United States and they were assigned to collect data for these cities using Realtor.com. There were three data collection requirements. They had to collect the latitude and longitude data as well as the text from the realtor’s description. The teams were told that they had two objectives. The first objective was to find a regression model to predict house price as well as to create a map and look for geographical patterns. The students were then asked to create a 5 to 10-minute video presentation that explained why their model should be used for predicting house prices.

The second objective of their assignment was to act as a realtor company to determine which of the other team’s models they should buy. Each team wrote a description of the best model for that city and individually each student wrote a review of one team’s presentation.

For this activity in R, the students continued to use codes and packages learned in the previous assignments. In addition, they added using latitude and longitude as part of *ggplot2*. To perform the regression analysis, they also used the *lm* regression function from base R.

RESULTS

Thirty-nine graduate and 562 undergraduate students took the survey and gave consent for their data to be used. For the first five statements, the students were asked how much they strongly agreed, agreed, somewhat agreed, somewhat disagreed, disagreed, or strongly disagreed. The percentage that strongly agreed and agreed were combined. The percentages for graduate students are listed first and undergraduate students are listed second (Graduate %, Undergraduate %).

1. I enjoyed working with the statistical software to examine data in maps. (58.9%, 67.0%)
2. The mapping assignments were valuable activities. (56.4%, 54.4%)
3. The mapping assignments allowed me to access modern data types. (74.4%, 79.3%)
4. I am an experienced computer software user. (25.7%, 26.7%)
5. It is important to learn how to format data. (82.1%, 86.8%)

In a second part of the survey, the students were asked, “In the mapping assignments, the data was not always in a format that was immediately able to be analyzed. Was this the first time that you had encountered this?”. Many graduate students (43.6%) and a little more than a third (35.3%) of undergraduate students had encountered the necessity of having to format data before analysis.

In terms of their favorite activity, 69.1% of the undergraduates and 46.2% of the graduate students picked the housing data as their favorite, the next favorite was the international data set.

Next, they were asked why they choose that as their favorite assignment. One student stated their favorite activity was the international data because; “I had an absolute blast on the International Data Set assignment because the data selection process itself was fascinating. I spent so much time just trying to decide which fascinating piece of data for my selected countries to use in the assignment because there was so much to choose from. In fact, I can definitely see myself going back to that website and using JMP to play with/visualize more data. To be honest, I do not remember the last time I had an assignment that enjoyable to do. I was genuinely interested in the relationships I was finding."

DISCUSSION

The activities described in this paper give students experience with important parts of the data scientific cycle ("import -> tidy -> transform -> visualize -> model -> communicate) as described by Grolemund and Wickham (2017). The students experience a small amount of tidying the data, as well as simplifying the data and creating graphics. Like the activities described by Loy, Kuiper and Chihara (2019), students gained valuable experience while working with real data and the resulting challenges. After they accomplished these processes, they examined trends for more than two variables. The activities described in this paper had the students expand beyond the standard discussion around bivariate relationships between two quantitative variables. The students were able to compare trends over time for various countries. Students learned how to create maps so that they could investigate possible geographical patterns and to use a very simple text analysis to explore command themes. As stated in Adrian et al (2020), maps are a common component of our everyday lives and should be a part of the introductory statistics curriculum. Additionally, the exercise gave the students experience with formatting data before it is analyzed. For each of the three graduate activities and one of the undergraduate activities, the students had to write about their findings. This process of communicating gave students practice explaining to others the trends that they had discovered.

Not only did these activities give the students experience working with multivariate data in the data scientific cycle, but also the students mostly valued and enjoyed the experience. The survey showed that over half of graduate and undergraduate students agreed at some level that these experiences were valuable, and a slightly higher percentage of the students found that these activities were enjoyable. This phase of the study focused on students' engagement with these activities, the next phase will focus on the statistical understanding gained by using a pre and posttest.

CONCLUSION

With appropriate scaffolding, students can dive further into relationships than just those of a bivariate nature. The use of maps in both graduate and undergraduate classes allowed students to look for general geographical trends as well as to see graphs that they commonly see in day to day life. Overall, the students found that the mapping assignments were considered valuable and a useful way to explore data.

REFERENCES

Adrian, D., Reischman, D., Anderson, K., Richardson, M. & Stephenson, P. (2020) Helping Introductory Statistics Students Find Their Way Using Maps, *Journal of Statistics Education*, (28)1, 56-74. DOI: 10.1080/10691898.2020.1721035

Gould, R. (2010). Statistics and the Modern Student. *International Statistical Review*, *78*(2), 297-315. https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1751-5823.2010.00117.x

Grolemund, G. & Wickham, H. (2017), R for Data Science. Newton, MA: O'Reilly.

Kaggle. Data Science for Good: Kiva Crowdfunding: Use kernels to assess the welfare of Kiva borrowers for $30K in prizes. Retrieved December 18, 2019 from https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding .

Kiva organization. About us. Retrieved February 9, 2020 from https://www.kiva.org/about .

Loy, A., Kuiper, S. Chihara, L. (2019) Supporting Data Science in the Statistics Curriculum. Journal of Statistics Education, (27)1, 2-11, DOI: https://doi.org/10.1080/10691898.2018.1564638.

National Council of Teachers of Mathematics. *The Math Forum*. Retrieved February 9, 2020 at https://www.nctm.org/mathforum/ .

Scott-Clarke, Ed & Lewis, Nell. How data is taming South Africa's infamous taxi buses. *CNN Business*. Retrieved October 31, 2019 at https://www.cnn.com/2019/10/31/business/data-south-africa-taxibuses/index.html

Ridgway, J. *(*2015*).* "Implications of the Data Revolution for Statistics Education*," International Statistical Review. Available at* http://onlinelibrary.wiley.com/doi/10.1111/insr.12110/abstract