MOBILISING THE STUDENT'S VOICE IN DATA SCIENCE EDUCATION: THE GREAT BARRIER REEF DATA PROJECT

Diana Warren
University of Sydney, Australia
diana.warren@sydney.edu.au

Like many World Heritage Areas, the Australian Great Barrier Reef (GBR), the world's largest coral reef system, is being threatened by climate change. Although much data is available for analysis, including complex spatial data, the domain knowledge required for investigation can be vast and the statistical tools complex. The purpose of our project was to investigate to what extent undergraduate students could engage with GBR data at the end of their first data science unit. Using projects from a large cohort with a detailed codebook, we explored the choices students made. Interesting findings emerged including the popularity of the GBR data, willingness to do independent research, and the strength of the student voice. This has implications for aligning data science curriculum with complex, global issues.

INTRODUCTION

Extending over 14 degrees latitude and 344,400 km^2, the Australian Great Barrier Reef (GBR) is the world's largest coral reef ecosystem and is internationally celebrated for its biodiversity. With an elaborate architecture of 3000 coral reefs, 600 continental islands, 300 coral cays, and around 150 inshore mangrove islands, the GBR is one of the most complex natural ecosystems in the world and home to a vast world of plants and animals, including more than 100 species of jellyfish (Great Barrier Reef Marine Park Authority, 2022).

Like many World Heritage Areas, the GBR is now threatened by climate change. At the recent United Nationals Educational, Scientific, and Cultural Organization (UNESCO) summit on the World Heritage Convention (June 2021), four key areas were tabled: GBR (Australia), Sagarmatha National Park (Nepal), Huascaran National Park (Peru), and Belize Barrier Reef Reserve System (Belize). Evidence including Figure 1, leads to the strong imperative: "Climate change is the greatest threat to the Great Barrier Reef," with effects from major coral bleaching to the rapid feminisation of green turtles. "If we are to secure a future for the Great Barrier Reef and coral reef ecosystems globally, there is an urgent and critical need to accelerate actions to reduce global greenhouse gas emissions. This must happen in parallel to taking actions to build the Reef's resilience" (Great Barrier Reef Marine Park Authority, 2019, summary).

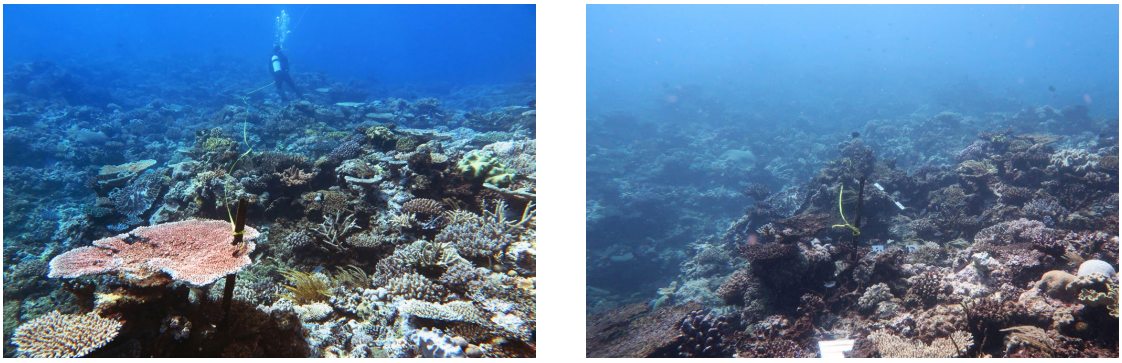

Figure 1. The Thetfold Reef near Cairns Australia, before (2016) and after (2017) coral bleaching

Given the global consequences of climate change and the wider need for knowledge and responsibility, the question arises how GBR data can be used in undergraduate data science education. Much GBR data is available for analysis, but is it accessible for undergraduate students, given the need for expert domain knowledge? Can students use foundational statistical tools to analyse GBR data? Does GBR data enable students to express their concern for the environment? The aim of our

In S. A. Peters, L. Zapata-Cardona, F. Bonafini, & A. Fan (Eds.), *Bridging the Gap: Empowering & Educating Today's Learners in Statistics. Proceedings of the 11th International Conference on Teaching Statistics (ICOTS11 2022), Rosario, Argentina.* International Association for Statistical Education. iase-web.org ©2022 ISI/IASE

"Great Barrier Reef Data Project" was to analyse data projects from a large undergraduate cohort to examine how students engaged with GBR data at the end of their first unit in data science.

CONTEXT

DATA1001 (Foundations of Data Science) is the flagship unit in a suite of first-year units in data science and statistics at the University of Sydney, undertaken by over 4,000 students each year. Each cohort is very diverse, with majors ranging from Ancient Greek to Wildlife Conservation, with high-performing students differentiated in the Advanced DATA1901 stream. Given such diversity, the whole DATA1001/1901 unit is taught through data stories. Careful attention is given to finding datasets that motivate, interest, and challenge all students (American Statistical Association, 2014; Fergusson & Bolton, 2018), including data that evokes student social conscience and agency (Stenalt & Lassesen, 2021).

Capstone Project 3

The DATA1001/1901 unit is assessed through three authentic, collaborative data projects (see a similar approach in Dierker et al., 2012, with a single research project). Worth 15% of students' grades, Project 3 is the capstone assessment in which students choose one of three data sets from different domains and write a client report in R Markdown within the RStudio Integrated Development Environment (IDE). The Briefing given to students appears in Table 1.

Table 1. Student Briefing for Project 3 in DATA1001/1901

Choose one of the three given datasets. Propose a client. Write a report for that client, with a concrete recommendation backed up by evidence from the data, and extra evidence (if appropriate). Prepare a one-minute video presentation (Client Briefing). Note: This project is demonstrating evidence-based decision making. Your report and video should be designed for your client, so they should use more or less technical language, depending on your client's level of statistical thinking.

Choice of Datasets

The students have a choice of three multivariable datasets—each with $p > 10$ variables—that are directly downloaded from government websites or research reports. The marking rubric is purposely general, allowing students complete freedom in terms of specifying a client, formulating a research question, and then choosing the appropriate statistical tools for their investigations. In Semester 2 for 2021, the three data sets focused on Economics, Social Science, and the Environment, as summarised in Table 2.

Table 2. Choice of datasets for Project 3 in DATA1001/1901

Subject	Source	Size
International airlines operating from Australia (Flights)	https://data.gov.au/dataset/ds-dga-e82787e4-a480-4189-b963-1d0b6088103e/details	n = 89312; p = 15; size = 9.5MB
Penalty notices in Australia (Penalties)	https://www.revenue.nsw.gov.au/help-centre/resources-library/statistics	n = 284404; p = 25; size = 59.2MB
Great Barrier Reef (GBR) chlorophyll monitoring	https://researchdata.edu.au/great-barrier-reef-1992-2009/677311	n = 19174; p = 14; size = 1.7MB

Domain Knowledge

The data sets are authentic—they are not cleaned and come with whatever data dictionary and documentation is provided by the source. Hence, each dataset requires students to investigate whatever domain knowledge is needed to understand the nature of the variables and how those variables might relate in context. In Semester 2 for 2021, the first two datasets were more accessible to students because most students have some past knowledge or personal experience of airlines and traffic penalties, allowing them to anticipate and more easily research what variables such as "Max_Seats" or "School_Zone_Ind" measure. In contrast, the third dataset (GBR) involves technical terms such as

"Secchi_Depth" and "Trichodesmium" that require careful research into specialised terms of reference. This led to four research questions (RQ) concerning the GBR data:

- Given a choice of three datasets, what proportion of students choose to analyse the GBR data, and are advanced students more likely to choose it?
- For their chosen investigation, do students investigate data for technical variables and consult resources to develop domain knowledge?
- For their chosen investigation, what level of statistical tools do students choose to analyse data?
- Does the project enable students to voice their social concern for the environment?

METHODOLOGY

In the second semester of 2021, 627 DATA1001 students and 33 DATA1901 students submitted Project 3, giving rich observational data. The submission files were produced in the RStudio IDE using RMarkdown, with output in the html form. Data analysis had four stages.

- First, the 660 projects were de-identified, resulting in an ID column in the data frame. Students' client briefing videos were not used to maintain the anonymity of each student.
- Second, the four research questions (thematic framework) were summarised into eight qualitative and two quantitative variables, forming a codebook, as summarised in Table 3.
- Third, the 660 projects were coded for research question 1 (the first two coding variables).
- Finally, the 322 projects concerning the GBR data (308 for DATA1001 and 14 for DATA1901) were coded for research questions 2–4 (the next eight variables).

Table 3. Summary of codebook for analysis of Project 3 html submissions

RQ	Coding Variable	Question	Values
1	Unit	What unit was the student enrolled in?	Data1001; Data1901
1	DataChoice	What data did the student use?	Flights; Penalties; GBR
2	TechnicalLanguage	Did the student use technical language, such as Trichodesmium or Clorophyll-a, in their project?	Yes; No
2	DomainKnowledge	Did the project require background research?	Yes; No
2	Papers	How many research papers were cited?	Integer > =0
3	StatisticalTools	What level of statistical tools were used in the analysis?	Level1: Numerical and graphical summaries Level2: Hypothesis testing Level3: Advanced analysis (not part of Data1001/1901)
4	ClimateChange	Did the student use the word "climate change" (or equivalents like "climate warming," "climate conditions," or "global warming") in their project?	Yes; No
4	ClimateChange Strength	How many times did the student use the word "climate change" (or its equivalents)?	Integer >= 0
4	SocialVoice	Did the student display a social concern for the environment in their project?	StrongVoice: Communicated strong concern for environment, usually in emotive language SomeVoice: Implied some concern for the environment NoVoice: No evidence of concern for environment
4	EmotiveWord	For students who demonstrated a "Strong Voice," what was their most emotive word?	Single word: e.g., dire, drastic, concern

A solo researcher coded all projects to ensure consistent application of codes. To minimise measurement bias and subjectivity—especially with the final two variables—a detailed codebook (Table 3) was written before coding began, which was tested and refined on a small random sample.

RESULTS

The results offer insights into the popularity of the GBR data, the willingness of students to conduct their own independent research, and the strength of the student voice.

Research Question 1: What Proportion of Students Chose the GBR Data, and Were Advanced Students More Likely to Choose It?

As Figure 2 depicts, the GBR data was the most popular choice for students (308/627 = 49%), followed by Flights (256/627 = 41%) and Penalties (63/627 = 10%). The DATA1901 students predominately chose GBR and Flights (both 14/33 = 42%). Although the GBR data was smaller in size than the other two datasets, it had the added complexity of variables requiring expert domain knowledge. Hence, the popularity of the GBR data could suggest that students were motivated to do their own research if the subject matter interested them, aligning with Wild (2015)'s conjecture, "We should target 'What I can do with data and what data can do for me' to build a desire to learn more" (para. 10).

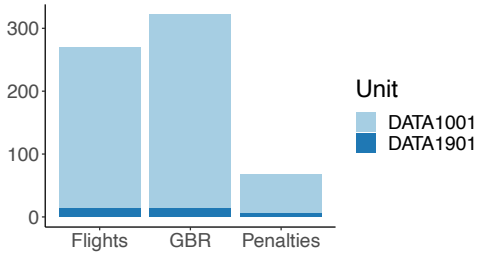

Figure 2. Datasets chosen by the DATA1001/1901 cohort for Project 3

Research Question 2: For Their Chosen Investigation, Did Students Investigate Data for Technical Variables and Consult Resources to Develop Domain Knowledge?

Although the GBR data had some variables that required only common knowledge (e.g., "Temperature" and "Salinity"), most DATA1001 students (259/308 = 84%) and all DATA1901 students chose to focus on variables that required expert domain knowledge (e.g., "Clorophyll-a" or "Trichodesmium"), as shown in Figure 3. As a result, most students cited at least two research papers, indicating their independent research (see Figure 4), with DATA1001 students citing more papers on average than the advanced students (DATA1001: median = 3, mean = 3.682; DATA1901: median = 2.5, mean = 2.786). Hence, it appears students were willing and able to acquire the domain knowledge needed to investigate the GBR data. Perhaps doing independent research increases student's interest in the data, concurring with the construct of "autonomy" in self-determination theory (Ryan & Deci, 2017).

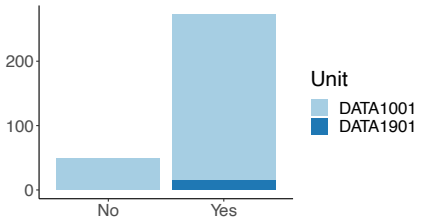

Figure 3. Technical language used for GBR data

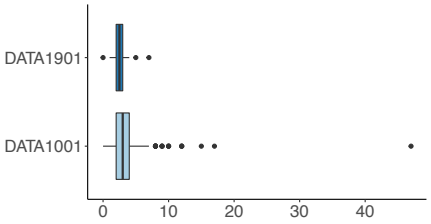

Figure 4. Number of papers cited for GBR data

Research Question 3: For Their Chosen Investigation, What Level of Statistical Tools Were Chosen by Students to Analyse the GBR Data?

Given the freedom in marking criteria, there was a wide range of analyses deployed, from simple bar charts (coded Level1 in "StatisticalTools"), to hypothesis testing (Level2), to interactive spatial maps (Level3), as summarised in Table 4. Hence, the GBR data allowed students to apply the full range of learning outcomes from DATA1001/1901 as well as to apply content beyond the course, supporting Ridgway et al.'s (2018) call to invigorate data science teaching by addressing "contexts that relate to social upheaval … such as global warming" (p. 1).

Table 4. Level of statistical tools deployed for analysing the GBR data

Unit / Level	Level1	Level2	Level3	Totals
DATA1001	107	185	16	308
DATA1901	4	6	4	14

Research Question 4: Did Students Voice Their Social Concern for The Environment? If So, How Strong Was Their Voice?

As displayed in Figure 5, the specific words "climate change" (or equivalent words such as "global warming") were only mentioned in about 30% of GBR projects (93/308 for DATA1001 and 4/14 for DATA1901). However, Figure 6 shows that only 7% of DATA1001 students (22/308) and 0% of DATA1901 students voiced no concern for the environment. ("NoVoice"). Sixty-seven percent of the DATA1001 students (205/308) and 93% (13/14) of the DATA1901 students communicated with a "StrongVoice," with the word cloud in Figure 7 revealing that the most common words were "impact" and "damage." A possible confounder is the disjunction between a student's personal voice, and the voice they used in the project, given that the marking criteria rewarded a "persuasive argument." However, surveying the projects based on the penalties data, as a type of control, reveals that students do not tend to use emotive words when the focus is economic (e.g., government revenue) but only when discussing issues such as safety (e.g., potential road deaths).

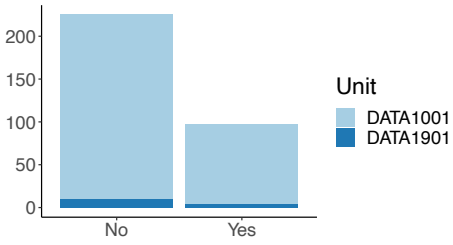

Figure 5. Climate change mentioned with GBR data

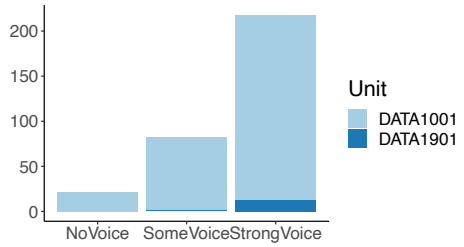

Figure 6. Strength of the student voice concerning the environment with GBR data

Figure 7. Word Cloud of the most emotive word used by those deploying a "StrongVoice" with regards the environment with GBR data

CONCLUSION

The GBR project gives evidence that even first-year undergraduate data science curriculum can align with complex, global, social issues. Students appear willing and able to conduct independent research for data requiring expert domain knowledge. This has implications for further aligning of curriculum with current social concerns, including Indigenous issues in Australia. Further work could compare the three different datasets for grade distribution and complexity of analysis.

ACKNOWLEDGMENTS

The two images in Figure 1 were provided, with permission, by the Commonwealth of Australia, Great Barrier Reef Marine Park Authority, and derive from the Australian Institute of Marine Science, Long-term Monitoring Program. Student data was analysed in accordance with ethics approval from the University of Sydney (Protocol no: 2022/243).

REFERENCES

American Statistical Association. (2014). *Curriculum guidelines for undergraduate programs in statistical science*. https://www.amstat.org/asa/files/pdfs/EDU-guidelines2014-11-15.pdf

Dierker, L., Kaparakis, E., Rose, J., Selya, A., & Beveridge, D. (2012). Strength in numbers: A multidisciplinary, project-based course in introductory statistics. *Journal of Effective Teaching*, *12*(2), 4–14.

Fergusson, A. M., & Bolton, L. (2018). Exploring modern data in a large introductory statistics course. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan*. ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3C1.pdf

Great Barrier Reef Marine Park Authority. (2019). *Position statement: Climate change.* Australian Government. https://elibrary.gbrmpa.gov.au/jspui/bitstream/11017/3460/5/v1-Climate-Change-Position-Statement-for-eLibrary.pdf

Great Barrier Reef Marine Park Authority. (2022). *Reef facts*. Australian Government. https://www.gbrmpa.gov.au/the-reef/reef-facts

Ridgway, J., Ridgway, R., & Nicholson, J. (2018). Data science for all: A stroll in the foothills. In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* ISI/IASE. https://iase-web.org/icots/10/proceedings/pdfs/ICOTS10_3A1.pdf

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness.* Guilford Press. https://doi.org/10.1521/978.14625/28806

Stenalt, M. H., & Lassesen, B. (2021). Does student agency benefit student learning? A systematic review of higher education research. *Assessment & Evaluation in Higher Education*, *47*(5), 653–669. https://doi.org/10.1080/02602938.2021.1967874

Wild, C. (2015). Further, faster, wider [Online discussion]. *The American Statistician Special Issue on Statistics and the Undergraduate Curriculum*, *69*. https://www.stat.auckland.ac.nz/~wild/preprints/15.Further-Faster-Wider.pdf