EXPLORING MODERN DATA IN A LARGE INTRODUCTORY STATISTICS COURSE

Anna-Marie Fergusson and Liza Bolton
The University of Auckland, New Zealand
a.fergusson@auckland.ac.nz

The amount, availability, diversity and complexity of modern data means that students now require a wider range of data-related knowledge and skills. We present our initial attempts at increasing student awareness of what can be learned from modern data, within the context of working with large classes (n ≈ 450) made up of students from diverse majors. Through the telling of data exploration stories we cover: sourcing data through APIs, web scraping and open databases; working with different types of data/variables such as text, timestamps, location co-ordinates and images; and potential issues related to data quality, ownership, and privacy.

INTRODUCTION

Introductory statistics courses have tended to use data collected within formal studies to teach students about study design, statistical inference and statistical models. However, the amount, availability, diversity and complexity of data that is now available in our modern world requires educators to broaden their definitions of what data is and what it means to teach students how to learn from data (Finzer, 2013; Gould, 2010). The statistics curriculum needs to be rebalanced and reimagined, by using data sourced from the data revolution (Ridgway, 2015) and by designing learning experiences that reveal more of the world of data faster than is typically the case in standard introductory statistics courses (Wild, 2016). Coupled with recommendations to modernise the data used to teach statistics are continued calls to ground statistical investigations within meaningful contexts with authentic data. Exploiting these contexts can support development of important research questions, motivate students to learn, and assist development of technical understanding (e.g. American Statistical Association, 2014; Cobb, 2015). While the changes needed to modernise the introductory statistics course extend well beyond the use of modern data, this paper will limit discussion to changes made to our large introductory statistics course to place a greater emphasis on exploring modern data.

OUR COURSE CONTEXT

With an enrolment of over 4500 students per year, our introductory statistics course (STATS10x) is one of the largest courses offered at the University of Auckland. STATS10x is a required course for students majoring in Statistics or Data Science and is also a service course for number of client departments located within the Faculties of Business, Arts, and Science. There are no mathematics or statistics prerequisites for our course and very few demands for algebraic methods beyond substitution into formulae. Students use the software packages *iNZight* (data visualisation software driven by R, see https://www.stat.auckland.ac.nz/~wild/iNZight/) and VIT (Visual Inference Tools, see https://www.stat.auckland.ac.nz/~wild/VIT/) from the first lecture to explore data and to carry out simulation-based inference. The statistical programming language *R* is introduced in the stage two equivalent course (STATS20x). There are no scheduled labs for the course: the use of technology is demonstrated in lectures and students are not expected to bring laptops to lectures due to equity, engagement and manageability issues.

STATS10x is delivered through 50-minute lectures, held three times a week over a 12-week semester. Class sizes are large, on average 450 students per class. These lectures are recorded and made available for students who do not attend the lectures in person. The course material for STATS10x has been developed over many years by members of the teaching team and is made available to students through a workbook. Typically each lecture activity is based on a real story, media article or published study, with lecture slides used to guide students through the related statistical investigation. The workbook includes most of the information displayed on the lecture slides, usually with gaps left for students to complete during the lecture. Some students bring the physical workbook to each lecture and some use the digital version of the workbook with or without the gaps already filled out. Alongside the workbook, web clickers are used throughout all lectures to encourage engagement and gain feedback on student learning.

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.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

THE NEW EXPLORING DATA CHAPTER

As an outcome of the annual course review, the decision was made to rewrite three chapters of the workbook (Exploring univariate variables, Exploring relationships between variables and Probability) as one new chapter called *Exploring data*. It was also decided to move the new chapter to the beginning of the course. The learning outcomes for the chapter were:

- appreciate the amount, diversity and complexity of data in our modern world
- identify and use appropriate tools to explore data
- know which features to look for in data and interpret and communicate these features
- reason with proportions
- be critical of the nature of patterns and relationships discovered in data.

An early decision was to remove probability as a standalone topic. Instead, emphasis was shifted to the use of data to make statements about absolute or relative likelihoods, in particular risk. The focus on proportional reasoning was seen as a way to increase students experience and confidence working with categorical data, and to support development of modern graphicacy skills important to making meaning of the ever-expanding types of charts used for data visualisations and infographics. Ideas for new content related to exploring data were based on recommendations from statistics and data science educators (e.g. American Statistical Association, 2014; Cobb, 2015; Finzer, 2013; Gould, 2010; Horton, Baumer, & Wickham, 2014; Ridgway, 2015; Wild, 2015). These ideas included: using open data; working with databases; using social media data; using images, text and sounds as data; integrating computation; working with data structures, in particular hierarchal and rectangular; familiarity with data file types, such as CSV, XML, JSON; working with a variety of data visualisations, including geomapping; increasing use of categorical data; using simple predictive models, such as CART; increasing explorations with three or more variables (multivariate); using messy and complex data sets; appreciating issues related to data privacy, ownership and "big data"; and increasing responsibility for communication, in particular reproducibility of computations and analysis. Armed with an inspiring and comprehensive list of modern data-related knowledge and skills, the first author of this paper began the process of developing lecture activities.

DESIGNING THE LECTURE ACTIVITIES

The natural mode of instruction for teaching students about data exploration involves students actually exploring the data themselves using technology. We considered students conducting data explorations during lectures using laptops but due to the practical constraints of lecture delivery discussed previously, "laptop lectures" were not considered viable. As the new *Exploring data* chapter was students' first exposure to statistics at the tertiary level, we wanted activities that moved quickly, required active participation, were engaging, interesting and relevant, and inspired students about what was to come in the remainder of the course. Although students were not going to be conducting explorations during lectures, they were required to explore data as part of their course assessed assignments and so therefore needed to be shown good examples of the types of question and decisions made during a data exploration. In seeking alternative lecture designs, we modelled our teaching delivery on data scientists who often share data explorations through the form of a blog post (e.g. https://juliasilge.com/blog/women-in-film/). This mode of communication appealed for many reasons: data explorations are often of a personal nature, inspired by something that sparks the blogger's interest and curiosity; data are often *modern* in nature, providing relevant and interesting contexts; bloggers tend to be open and honest about the questions they asked of the data and the methods they used to explore the data, including the code used for any computations or analysis; a greater breadth of data-related knowledge and skills can be covered due to the story telling delivery; and readers of the blog post can gain confidence about what can be done with data even if they have not done it themselves. Because of the increasing popularity and promotion of data science there are a large number of data exploration blog posts available for students to learn from. These narrative style blog posts are good examples of how to change the culture of teaching statistics in order to enculturate students into data explorations where curiosity about data drives learning (Gould, 2010).

Several different data explorations were then undertaken to provide material for lecture activities, supplementary workbook examples, assignment tasks and assessment questions. Data contexts and exploratory methods included: data related to the London 2012 Olympic Games, obtained through web scraping; data on crimes reported to the police department from the San Francisco open data project (https://datasf.org/opendata/); data on Games of Thrones episodes, obtained through an API; data on recipes from an online database; data from Harry Potter fan fiction; and data on a lecturer's personal emails. The underlying principle for the design of our lectures were for lecturers to model data exploration through the telling of stories. To transition the written blog post form to the "live presentation" form, suitable for delivery in large lectures, care was taken to punctuate the story telling with questions posed to the students during the lecture in a variety of ways, for example, verbally to the lecture class, for small group discussions within the lecture, through web clickers, or interactive applications. A greater emphasis on visuals was also needed, as well as live modelling by the lecturer/instructor using software. We now present an example data exploration from the new *Exploring data* chapter which was developed from the first author's story of how she collected, explored and interrogated the data.

EXAMPLE DATA EXPLORATION: JE T'AIME PARIS

The story begins with the lecturer asking students what monument they think is the most photographed in the world. After getting responses from the students they are told about a study undertaken by Crandall, Backstrom, Huttenlocher, and Kleinberg (2009), which analysed nearly 35 million images posted by over 300 000 users on Flicker during a six-month period. The study found that the most photographed monument in the world is the Eiffel Tower. Students are asked to discuss in small groups how they think the researchers were able to determine that the Eiffel Tower was the most photographed monument. The lecturer then explains the personal motivation for the data exploration which was triggered by the first author's curiosity:

> *At the end of 2014, Anna travelled to Paris with her future husband. Like many other tourists, they took many photos of the Eiffel Tower and like some other tourists got engaged there. The experience got Anna thinking about the potential similarity of photos of the Eiffel Tower, in particular engagement photos.*

Students are shown a selection of public Instagram posts of the Eiffel Tower and asked to compare features of the different posts in terms of similarities and differences. For example, some photos are taken during the day and others at night; some photos include all the Eiffel Tower in the shot while others are close ups of the structure; some photos contain people in the photo and others do not. The lecturer then explains that Instagram makes all public posts available through an API and explains how data were collected for this exploration. To collect the data, the Instagram API was used to record the details of any post made within 500 metres of the Eiffel Tower (the criterion) over a period of two months in 2015. The API ran in real time using streaming data and updated a Google sheet each time a public Instagram post met the criterion (n = 9398). Students are shown a snippet of the raw data set (see Figure 1). Throughout the data exploration, the raw data set is re-displayed to strengthen the connection between the visualisations, manipulations and data.

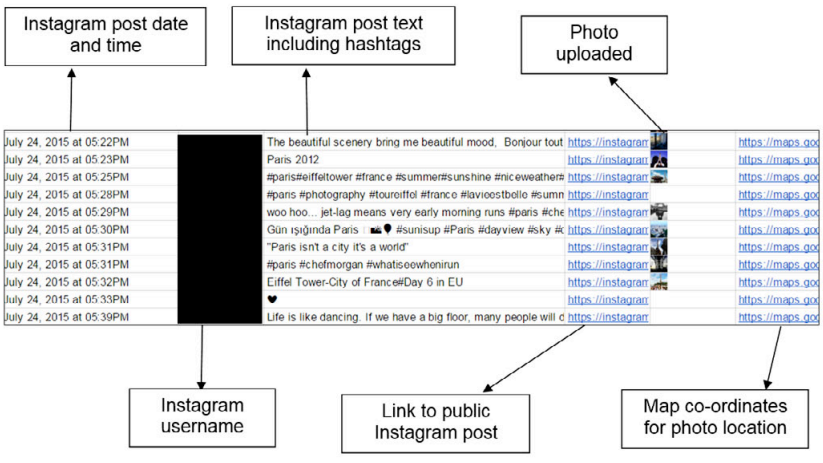

Figure 1. A snippet of the raw data set

Students are then asked how likely they think that any one of these posts features a photo of the Eiffel Tower, before being given a link to a web page that displays a random selection of 10 of the photos from the data set that are still public (see https://goo.gl/B9EwdL). Students are asked to examine each photo and record how many of the 10 photos are "of the Eiffel Tower". These counts are shared with the class through web clickers and used to estimate the proportion of photos in the data set that are "of the Eiffel Tower". The ownership of photos once they are shared publicly through platforms such as Instagram is discussed, and students are encouraged to read the terms and conditions they agree to when installing applications. After being given the conditional statement *It's likely that a photo taken within 500m of the Eiffel Tower features the Eiffel Tower,* students are asked if they trust the data enough to make this statement. After getting responses from the students, further examples of questions that could be asked about the data's relevance and credibility are provided in the workbook which the students discuss in groups: Were these photos actually taken at the location or uploaded to Instagram at the location? Has the automated process recorded the information about the Instagram posts accurately? What about Instagram users that do not publicly share their posts? Are photos uploaded by Instagram users similar to photos taken by everyone else? How many people turn off location sharing on Instagram and does this matter for our exploration? At this point, students do not have complete answers to these questions.

Students are then asked to look again at the raw data set snippet shown in Figure 1 and asked to reflect on the nature of the information that has been shared by these Instagram users. To highlight potential privacy issues, online applications such as https://iknowwhereyourcatlives.com/ are used to show how others have used similar public data for various purposes. The lecturer further explains how one of the people in the data set was able to be tracked as they walked around the Eiffel Tower, and another person was found to be making a large number of posts during work hours with their location clearly in an office block. After alarming students to the potential privacy and ethical issues associated with sharing location data through social media, students are shown how the data on location has been recorded in the data set using a query string as part of the URL, for example https://maps.google.com/?q=48.855530063,2.300052058&z=18. Students are then shown a visualisation of the data using Google maps to plot the location associated with each post (see Figure 2). Students are asked to discuss the visualisation in small groups and to share what they think the visualisation tells them. Common points of discussion are: the dots in the water (people taking photos from boats), the darkness of the points (transparency used to communicate frequency and to combat overprinting), the amount of dots in the top left and bottom right quadrants (common vantage points for photos).

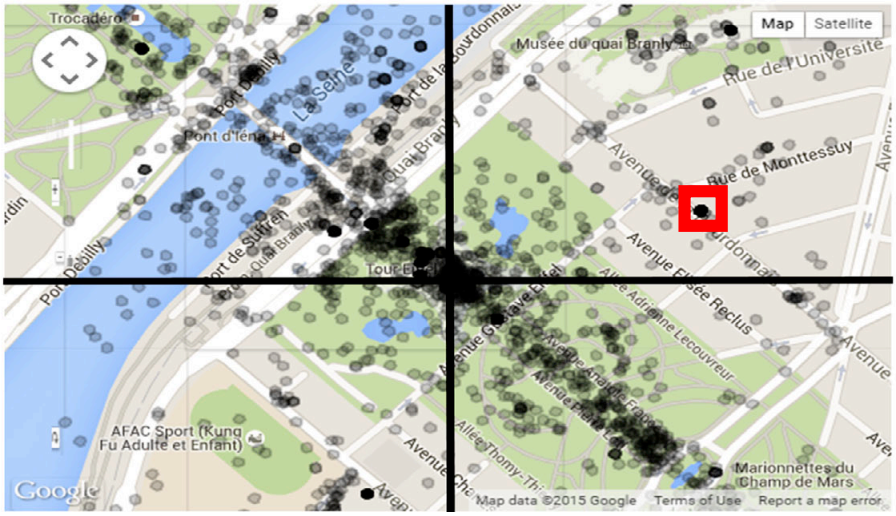

Figure 2. Locations of Instagram posts with the default co-ordinate shown by the red box

The lecturer then explains how an issue came up during the analysis of the data. Over half of the posts had the same location recorded (the default co-ordinate shown in Figure 2). The lecture then tells the story of how the data was checked for accuracy. A check of a random selection of posts showed that while some of the photos matched the location recorded, others did not. Checking the *default* co-ordinate revealed photos that did not match what was visible from this location which was confirmed by using Google street view. Students are then shown how Stack Overflow was used to try to find a solution to the problem (see https://stackoverflow.com/questions/31447121/why-am-i-getting-duplicate-location-co-ordinates-from-the-instagram-api/).

The story then moves to focus on what people write in their posts and how we could analyse the words used. Students are shown a large number of posts from the data set and asked to discuss the ways that the text used is similar or different, for example, number of words used, number of emojis used, language used, or particular hashtags used. Sentiment analysis is discussed briefly before moving to using hashtags as a proxy for the key words or general feeling of the post. The lecturer then explains the computational process of identifying which parts of the text of the post represented the hashtags and associated issues with this process, for example, users not leaving a space between multiple hashtags, the encoding of emojis if used as part of hashtags, and the use of different languages. Decisions made during the hashtag identification process are also explained, for example, the decision to convert all hashtags to lower case. Students are shown a JSON string which lists hashtags with the counts of their use across the data set and challenged to find an "inappropriate" hashtag as a fun way of introducing students to this format. The hashtag counts data is then used to identify the top 10 hashtags used and the mean number of hashtags used per post. This analysis also reveals that a high proportion of the posts do not have hashtags.

While time conversions between 24- and 12-hour time and different time zones are part of the high school curriculum, it is unlikely that many students have performed large scale time conversions as part of manipulating data. Timestamps are notoriously difficult to work with and an added complication in this story is that the timestamps recorded in the Google sheet reflect the time in New Zealand rather than Paris. Students are shown how this discrepancy was discovered by comparing the timestamp recorded in the Google sheet and the timestamp shown in the meta tags within the HTML for the post for one of the posts in the data set. To explore whether Paris is indeed *the city that never sleeps*, students are shown how the timestamp was used to create new variables in the data set, including the hour of the day (see Figure 3).

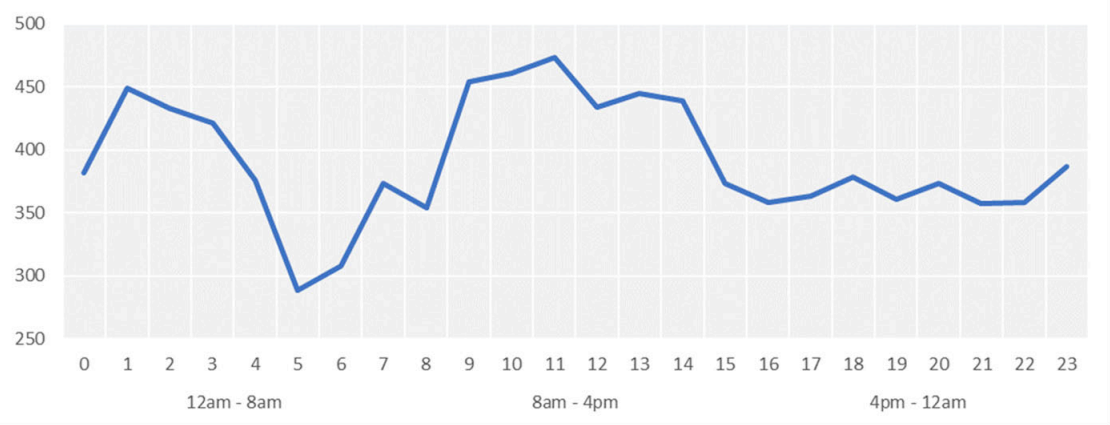

Figure 3. Total number of posts by hour of day

Students are asked to discuss what different plots of the data reveal about when posts are being made and to consider the earlier questions concerning what the timestamp and locations are capturing. Due to the nature of the data set, triangulation is possible by comparing the photo, photos, the timestamp and the location co-ordinates. The story ends by the lecturer returning to the initial motivation for the exploration: to determine the similarity of Eiffel Tower photos. The use of humans to classify photos as part of a modelling process is discussed using the "Selfcity" project (see http://selfiecity.net/selfiexploratory/). The use of computational techniques to "cluster" the

photos by similarity based on graphical features are also discussed. Students are then given access to the tidied data set and challenged to explore the data further to find new insights.

DISCUSSION AND CONCLUSION

The new *Exploring data* chapter has been taught for six semesters with minimal changes to the lecture activities during this time. As is often the case with storytelling, with each retelling additional details are added and key events in the story are elaborated further when the listener displays interest. Four data explorations were intended to be covered but after three semesters of teaching, the fourth data exploration was moved to a reading example. There are plans to update the data used for these explorations, to develop new data explorations and to supplement *iNZight*-based data exploration with parallel examples of *R* code. Additionally, as our introductory statistics course has a very large weighting towards assessment conducted using multiple choice questions (MCQs), further work is needed to create MCQs that assess ideas related to exploring modern data.

Although the nature of large lecture classes imposes some practical restrictions on the design of lecture activities, we are confident the new chapter serves as an exciting introduction to statistics at the tertiary level. The telling of data exploration stories, supported by the demonstration of skills such as obtaining data through computation and producing data visualisations, appears to engage students and quickly broaden their awareness of modern data sourced from the data revolution (Ridgway, 2015). Students who do not have strong computing skills are able to participate in learning from the first lecture. The contexts explored use data that students have personal experience with and the use of explorations conducted by the teaching team further emphasises the human perspective. By narrating modern data explorations we are able to model the inquisitive stance needed to learn from data (Gould, 2010) and the habits of mind (Finzer, 2013) that are essential for statistics and data science. Students participate in the exploration by being challenged to ask questions about the data, the methods used to manipulate the data, the graphics used to visualise the data, and the conclusions made about what the data tells us. We believe our approach helps to elevate the position of modern data within the statistics curriculum and supports a culture of teaching statistics and data science where students' curiosity drives learning.

REFERENCES

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

Cobb, G.W. (2015). Mere renovation is too little too late: We need to rethink our undergraduate curriculum from the ground up. *The American Statistician 69*(4), 266–282.

Crandall, D. J., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. In Proceedings of the 18th international conference on World Wide Web (pp. 761-770). Association for Computing Machinery. *http://www.cs.cornell.edu/~dph/papers/photomap-www09.pdf*.

Finzer, W. (2013). The data science education dilemma. *Technology Innovations in Statistics Education, 7*(2). http://escholarship.org/uc/item/7gv0q9dc.

Gould, R. (2010), Statistics and the modern student. *International Statistical Review*, *78*(2), 297–315.

Horton, N. J., Baumer, B. S., & Wickham, H. (2014). Teaching precursors to data science in introductory and second courses in statistics. In K. Makar, B. de Sousa & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014),* Flagstaff, Arizona, USA. Voorburg, The Netherlands: International Statistical Institute.

Ridgway, J. (2015) Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528–549.

Wild, C. J. (2015) Discussion: Locating statistics in the world of finding out. *International Statistical Review, 84*, 194–202.