INFUSING DATA VISUALIZATION INTO INTRO STAT USING TABLEAU

Stacey Hancock
Montana State University
stacey.hancock@montana.edu

The revised Guidelines for Assessment and Instruction in Statistics Education (GAISE) include two new emphases for teaching statistical thinking in the introductory statistics course, one of which is to "give students experience with multivariable thinking" (ASA, 2016). One of the simplest and most efficient ways to introduce multivariable thinking is through data visualization. The traditional college introductory statistics course covers univariate and bivariate plots such as boxplots, histograms, and scatterplots. However, students are rarely exposed to data visualizations of three or more variables. We will discuss how to infuse multivariable data visualization into the introductory statistics course using the software Tableau by dedicating one class period to data visualization and including data visualization as an integral part of the statistical investigative process throughout the course.

INTRODUCTION

Two recent sets of guidelines in the statistics education community have underscored the importance of exposing students to multivariable data sets from real applications. The revised *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) emphasize integrating "real data with a context and a purpose," and giving students "experience with multivariable thinking" (ASA, 2016, p. 6). An emphasis on real applications and complex data is also found throughout the *Curriculum Guidelines for Undergraduate Programs in Statistical Science* (ASA, 2014). At the introductory level, a simple and efficient way to incorporate multivariable data at the start of the course is through data visualization with interactive software.

The term "multivariable thinking" refers to the ability to explore relationships among more than two variables and to investigate how these variables interact, a skill that emphasizes the multivariate nature of statistics. For example, a simple scatterplot of two quantitative variables can be transformed into a "multivariable" plot by using different plotting symbols or colors for different groups. This type of plot could allow students to visualize how the presence of a third confounding variable limits the scope of conclusions that can be drawn about the association between two variables.

Creating meaningful data visualizations to communicate information is an important skill in its own right. While statistical graphics emerged with the earliest attempts to analyze data (Beniger & Robyn, 1978), and much has been written on best practices for data visualization (e.g., Tufte, 1992; Cleveland, 1994), with the rise of data science and an increasingly data-infused society, the teaching and understanding of effective data visualizations has become even more crucial.

In this paper, we argue that instructors can effectively introduce multivariable thinking and emphasize data visualization and exploration by dedicating one 75-minute class period to data visualization using the software Tableau, then revisiting data visualization concepts throughout the course by incorporating additional problems into class activities and assignments already in use.

WHY TABLEAU?

Our decision to use Tableau in the intro stat classroom over other statistical software packages was based on several reasons. Tableau is a powerful, authentic, data visualization software that is used among data science professionals. Tableau is free for academic use and easy for students to download to their own computers. It also provides a venue for students to share visualizations and create a public online portfolio of their work that can be highlighted during a job interview.

Tableau has a drag-and-drop interface which facilitates data exploration and the creation of quality graphics without requiring programming skills and with minimal training in how to use the software. While others have advocated using R (e.g., Wang et al., 2017), and the `mosaic` R package aims to ease the cognitive load of learning programming in conjunction with statistics, we

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 [© 2018 ISI/IASE]

feel that even this slight increase in mental effort may detract from a focus on the underlying concepts of data visualization and statistical thinking.

INTRODUCTORY STATISTICS AT MONTANA STATE UNIVERSITY

Each semester, we offer Introductory Statistics (Stat 216) to 20-25 sections of 40 students. We teach a simulation-based curriculum with a flipped classroom format using the textbook *Introduction to Statistical Investigations* (ISI) by Tintle et al. (2016). We use the textbook's online applets in place of a statistical software package. Students read selected passages and watch videos, then take a short online quiz on the material prior to coming to class. Class time is dedicated to "explorations"–in-class activities where students work in teams of three to discover new concepts, with the instructor fostering discussion and asking targeted questions.

We cover descriptive statistics and plots (Sections P.2, 5.1, 6.1, and 10.1) during the first week of the 15-week semester. Starting in fall 2017, we added one 75-minute classroom activity on data visualization with Tableau to the first week of class. The addition of this Tableau class day is not dependent on the curriculum and could be integrated into a variety of courses.

TABLEAU CLASS DAY

Our Tableau class day occurs on the second class period of the semester, after one class on descriptive statistics and basic plots where students learn which plots are appropriate for which data types. Students are given instructions outside of class to download and install Tableau on their laptops. If a student does not own a laptop, he or she can share a laptop with another student during class, and can use Tableau in the student computer labs for assignments.

In the first 15 minutes, instructors work through examples of data visualizations that appear in the news, and critique them with input from students through class discussion. Throughout the discussion, instructors guide students to think about basic principles of data visualization, such as understanding how humans perceive differences (e.g., compare position rather than angle) and how to use that understanding to make important comparisons easy (Robbins, 2013; Cleveland, 1994). Given an example data visualization, class discussion focuses on the following questions:

1. What are the observational units in these data?
2. What are the variables or summary measures displayed in the plot and what "graphical perception tasks" (Cleveland, 1994) are used for each variable, e.g., position, area, angle? What are the variable types (quantitative or categorical)?
3. What would the data set look like in a spreadsheet?
4. What is the overall message the author is attempting to communicate through the visualization?
5. What comparisons are easy to make using this data visualization? Hard to make?
7. How could we improve this visualization?

For example, a data visualization from the *New York Times* in the article "More Older People Are Finding Work, but What Kind?" (Bui, Aug 18, 2016) displays a scatterplot of "likelihood of being hired" for workers age 55-64 compared to workers age 30-49 on the *y*-axis versus salary on the *x*-axis (see http://nyti.ms/2pC2HZv; used in a workshop at USCOTS 2017 by Bergen & Iverson). Since position is the graphical perception task we are able to decode most accurately (Cleveland, 1994), comparisons of different types of jobs in terms of salary and likelihood of being hired come across well in this visualization. However, comparing number of jobs across job types (displayed as size of the dot, or area) is difficult.

After class discussion of one or two data visualization examples, instructors give a short 10-minute demonstration on how to import data into Tableau, how to open a new Tableau worksheet and dashboard, and the distinction between "dimensions" (variables that are not aggregated) and "measures" (aggregated variables, e.g., sum, average, count). Next, students are provided with a large, multivariable data set from the 1985 Current Population Survey (CPS) (Berndt, 1991), and create bar graphs, segmented bar graphs, histograms, boxplots, and scatterplots using these data in Tableau. This data set has a mix of categorical and quantitative variables, including variables such as hourly wage (US dollars per hour), number of years of education, sex, age (years), and sector of the economy (sales, clerical, etc.). An example of a scatterplot displaying age and wage by sex in Tableau, plus adding years of education as the area of the dot, is shown in Figure 1. (See Wang et al. (2017) for a data visualization activity in R using these data.)

Students then use Tableau to explore another multivariable data set on roller coasters in the United States, guided by a set of questions about the data. (The Tableau class activity used in spring 2018 can be found at http://www.math.montana.edu/courses/s216/TableauActivityS18.pdf.) The last question of the activity asks students to create a dashboard (a collection of several plots and supporting information) to summarize the relationship between at least three variables in the data set, including a written summary of what they found. This portion of the class period is purposely left somewhat unstructured, allowing students ample time to play with the software and explore the data in new ways.

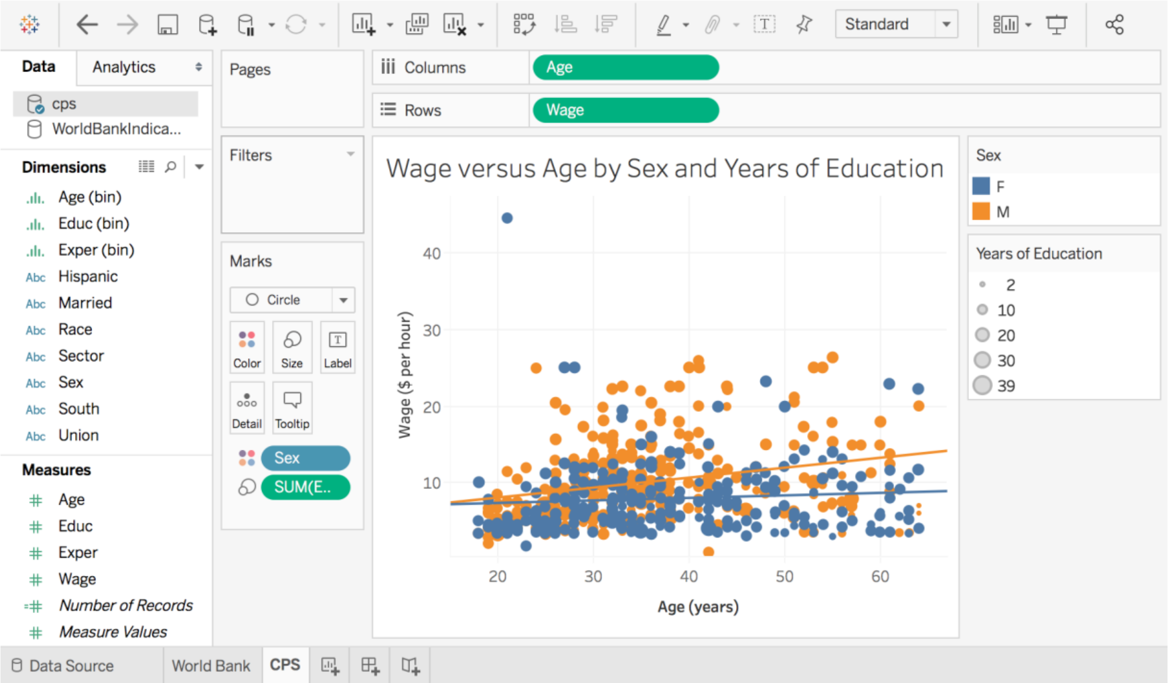

Figure 1. Tableau worksheet environment displaying CPS variables Age (position in *x*-direction), Wage (position in *y*-direction), Sex (color), and Years of Education (size).

DATA VISUALIZATION THROUGHOUT THE COURSE

Data visualization using Tableau was integrated into the course through additional questions on existing in-class activities and homework assignments following the statistical investigative process, and its required use on the course project. Assignments and activities were revised to use data sets of more than two variables, and students were asked to produce appropriate visualizations of these data to address a specific research question. Students presented Tableau visualizations of their project data as part of their presentation and written report.

ASSESSMENT

Assessment of the Tableau classroom activity included a pre- and post- online survey (spring 2018 only), a short survey administered in week 14 (fall 2017 only), and the Tableau visualizations and summaries students created in class (graded based only on completion). The post-survey and week 14 survey asked students to reflect on whether Tableau had enhanced their understanding of statistics and the strengths and weaknesses of the activity.

Assessment items on multivariate thinking through data visualization were included on several homework assignments and on each exam. For example, on the first midterm exam in spring 2018, using multivariable plots of a movie data set, students were asked to identify observational units, list the variables and their types displayed in the plots, and answer questions about the data set. On the second midterm exam, students were asked to explain why stress level was a confounding variable in the relationship between happiness level and income based off of information provided in three segmented bar graphs of the data.

RESULTS AND DISCUSSION

During the Tableau activity, students appeared to enjoy using the program, but expressed frustration if they were unable to get Tableau to create a desired plot. In week 14 of fall 2017, only 4 of 34 students agreed that Tableau had enhanced their understanding of statistics. Though several students commented that Tableau helped make data easier to interpret and gave them a better understanding of how to visualize data, many stated that they did not receive adequate instruction in using the program and did not revisit it enough throughout the semester. Indeed, since Tableau is a software targeted towards "doing" statistics rather than "teaching" statistics, students require a suitable amount of instruction in the use of the software in order to relieve their frustration. Consequently, in spring 2018, we moved the instructor demonstration of Tableau to the previous class, and dedicated the entire Tableau class day to discussion of data visualization principles followed by hands-on practice with Tableau. Additionally, we required the use of Tableau homework assignments rather than giving students the option of using either Tableau or the ISI online applets (as in fall 2017).

In the fall 2017 week 14 survey and the spring 2018 post-survey, the majority of student responses to the activity were positive; overall, students enjoyed working in a team on this activity (though some expressed it was more helpful when each student had their own laptop), valued instructor guidance throughout the activity, and viewed the activity as a good introduction to the variety of ways one can display data.

In summary, the Tableau class day gave students exposure to multivariable thinking through data visualization in the first weeks of the semester. Students explored real-world data sets, created quality visualizations in a drag-and-drop software environment, and communicated a data story through visual and written summaries. Students recognized the value of using a data visualization tool beyond the ISI online applets and gained a better understanding of what type of variables produce which plots. If a goal of the course is to develop student proficiency in creating quality data visualizations, instructors need to provide sufficient instruction in the use of the software and should revisit the software regularly throughout the semester in a variety of contexts.

REFERENCES

American Statistical Association Undergraduate Guidelines Workgroup (2014). *2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science*. Alexandria, VA: American Statistical Association. http://www.amstat.org/asa/education/Curriculum-Guidelines-for-Undergraduate-Programs-in-Statistical-Science.aspx

Beniger, J. R. & Robyn, D. L. (1978). Quantitative Graphics in Statistics: A Brief History. *The American Statistician, 32*(1), 1-11.

Bergen, S., & Iverson, T. (2017, May). Web Scraping and Data Visualization with Python and Tableau. Workshop presented at the *United States Conference on Teaching Statistics, State College, PA.*

Berndt, E. R. (1991). *The Practice of Econometrics: Classic and Contemporary*. Reading, MA: Addison-Wesley.

Bui, Q. (2016, August 18). More Older People Are Finding Work, but What Kind? Retrieved Nov 9, 2017, from: http://nyti.ms/2pC2HZv.

Cleveland, W. (1994). *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

GAISE College Report ASA Revision Committee (2016). *2016 Guidelines for Assessment and Instruction in Statistics Education College Report*. Alexandria, VA: American Statistical Association. http://www.amstat.org/education/gaise

Robbins, N. B. (2013). *Creating More Effective Graphs.* Chart House.

Tintle, N. L., Chance, B. L., Cobb, G. W., Rossman, A. J., Roy, S., Swanson, T. M., & VanderStoep, J. L. (2016). *Introduction to Statistical Investigations*. Hoboken: Wiley.

Tufte, E. R. (1992). *The Visual Display of Quantitative Information*. Graphics Pr.

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 Education, 10*(1).