Title
Data Visualization on Day One: Bringing Big Ideas into Intro Stats Early and Often

Journal
Technology Innovations in Statistics Education, 10(1)

Authors
Wang, Xiaofei
Rush, Cynthia
Horton, Nicholas Jon

Publication Date
2017

DOI
10.5070/T5101031737

1. INTRODUCTION

A number of calls to infuse our statistics curricula with statistical computing to expose students to authentic data experiences have been recently published (Nolan and Lang 2012; Hardin, Hoerl, Horton, Nolan, Baumer, Hall-Holt, Murrell, Peng, Roback, Temple Lang, and Ward 2015; Horton and Hardin 2015). In the revised 2016 Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report, instructors are recommended to "integrate real data with a context and a purpose", to "use technology to explore concepts and analyze data", and to "emphasize the multivariate nature of the discipline" (ASA GAISE College working group: R. Carver, Everson, Gabrosek, Rowell, Horton, Lock, Mocko, Rossman, Velleman, Witmer, and Wood 2016). With booming interest in data science, it is now more important than ever to bring statistical software into the classroom early, to enable analysis of real-world data, to expose students to the excitement and potential of statistics, and to provide examples where insights are extracted from data.

Chance, Ben-Zvi, Garfield, and Medina (2007) provides a broad overview of how technology benefits learning even at the introductory level, citing various tools ranging from graphing calculators to computing software. With the advent of R (R Core Team 2016) and RStudio (RStudio Team 2016), students now have access to a free, open-source software with an established prominence in industry. A powerful software such as R does come with a steep learning curve for some; we believe that using data visualization as the entry point to learning R and introductory statistics could lessen the anxiety associated with both. To facilitate this initial foray into the world of data visualization using R, we advocate the use of (1) R Markdown (Allaire, Cheng, Xie, McPherson, Chang, Allen, Wickham, Atkins, and Hyndman 2016), a system that is integrated into RStudio, because its workflow inherently encourages well-documented reproducible analysis (Baumer, Çetinkaya-Rundel, Bray, Loi, and Horton 2014) and (2) the `mosaic` package (Pruim, Kaplan, and Horton 2016b), whose modeling language provides a simplified interface to multivariate descriptive statistics, linear models, and graphical displays.

Much has been written about the nuances of data visualization techniques; see for example Tufte and Graves-Morris (1983); Cleveland (1994) and, more recently, Wickham (2009, 2010). Nolan and Perrett (2016) describes the potential for visualization to inform statistical thinking and suggests ways to incorporate this capacity into statistics courses, noting that computational advancements in recent decades have made it far easier for students to apply advanced graphical tools at the level of introductory statistics.

In this article, we present an in-class introductory multivariate data visualization activity designed for a single class period during the first week of class. The activity begins with a brief, instructor-led introduction to exploratory data analysis in R. Students are then placed in small groups tasked with exploring a new dataset to produce three visualizations that describe particular insights that are not immediately obvious from the data. Upon completion, students will have produced a series of univariate and multivariate visualizations on a real dataset and practiced describing them.

Table 1: Abridged codebook for the 'HELPrct' dataset

Variable	Description
sex	'male' or 'female'
age	subject age at baseline (in years)
racegrp	race/ethnicity: 'black', 'hispanic', 'white', or 'other'
anysub	use of any substance post-detox: 'no' or 'yes'
cesd	Center for Epidemiologic Studies Depression measure at baseline (high scores indicate more depressive symptoms)
substance	primary substance of abuse: 'alcohol', 'cocaine', or 'heroin'
mcs	SF-36 Mental Component Score (measured at baseline, lower scores indicate worse status)
pcs	SF-36 Physical Component Score (measured at baseline, lower scores indicate worse status)

2. THE ACTIVITY

2.1. Overview

The proposed activity begins with a 15-minute tutorial led by the instructor on how to generate basic numeric summaries and visualizations in R. The lecture introduces the basic functionality used by the `mosaic` R package to simplify the process of generating multivariate graphical displays and summary statistics. The focus of this tutorial is not on the mechanics of these exploratory tools (for example, how we compute the heights of a histogram) but rather on enhancing one's comprehension of the relationships between variables in a dataset. To that end, a complex, multivariate dataset should be selected to serve as the basis of this tutorial. In our case, we used the baseline data from the Health Evaluation and Linkage to Primary Care Clinical Trial ('HELPrct'), which enrolled subjects without primary medical care while they were attending a substance-use detoxification unit (Samet, Larson, Horton, Doyle, Winter, and Saitz 2003). Table 1 summarizes a portion of the variables collected at baseline that might be of interest to study.

We recommend starting out with some motivating questions for class discussion:

- What does one row of the dataset represent?
- Who is included in this dataset? Who is not?
- What kinds of variables are included?

After discussing these answers, the students have a better sense of the scope of the dataset, but questions remain regarding the relationships between the variables. The next step is then to show some simple summary statistics and visualizations that help shed light on more targeted questions:

Table 2: Abridged codebook for the 'CPS85' dataset

Variable	Description
wage	wage (US dollars per hour)
educ	number of years of education
race	'NW' (nonwhite) or 'W' (white)
age	age in years
sex	'F' (female) or 'M' (male)
married	'Married' or 'Single'
exper	number of years of work experience (inferred from age and education)
union	'Union' or 'Not Union'
sector	sector of employment: 'clerical', 'const', 'manag', 'manuf', 'other', 'prof', 'sales', or 'service'

- What are the proportions of men and women in this dataset?
- What are the proportions of different primary substances of abuse in this dataset?
- What does the relationship look like between depression score and overall mental health?

We distribute to students a handout (see Appendix A) that contains the code to generate a variety of univariate (histograms and barplots) and bivariate (boxplots and scatterplots) plots. We demonstrate how some of these plots can be drawn in R by typing the corresponding code from the handout into an R Markdown file and compile the results. We also show how easy it is to load the help files for the relevant functions to look at ways of adding third or fourth grouping variables to make multivariate visualizations. We then ask the students to practice statistical thinking by reflecting on what is learned from each plot and discuss as a class. Some plots are not particularly insightful (see Figure 1 for example) relative to others (Figure 2). And additional information might be gleaned by including another variable (Figure 3). Through these demonstrations, we encourage students to think about what interesting questions might be answered given the data and the tools at their disposal.

After the brief introduction to R, we then provide students with a different dataset of similar complexity. For this second dataset, we used data from the 1985 Current Population Survey ('CPS85') (Berndt 1991). Table 2 summarizes some of the variables contained in this dataset. Note that getting into RStudio and loading the R Markdown template may take some time when students first use R/RStudio and RMarkdown.

The students are provided a single R Markdown worksheet (Appendix B provides an example) in which they are asked to produce three meaningful plots of the new dataset to provide insight on some facet of interest, and write a couple of sentences about each plot to discuss what they learned from it. Students should complete this task in groups of about two or three over a period of 40 to 50 minutes. The deliverable is a compiled R Markdown file, in HTML form, that collates the code, plots, and descriptions.

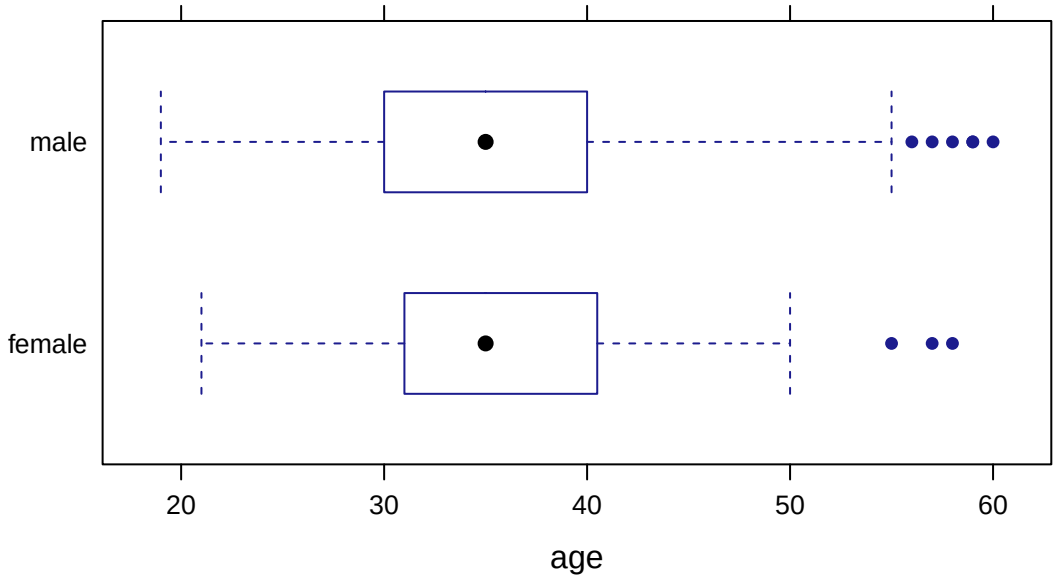


Figure 1: Boxplot of age by sex of subjects

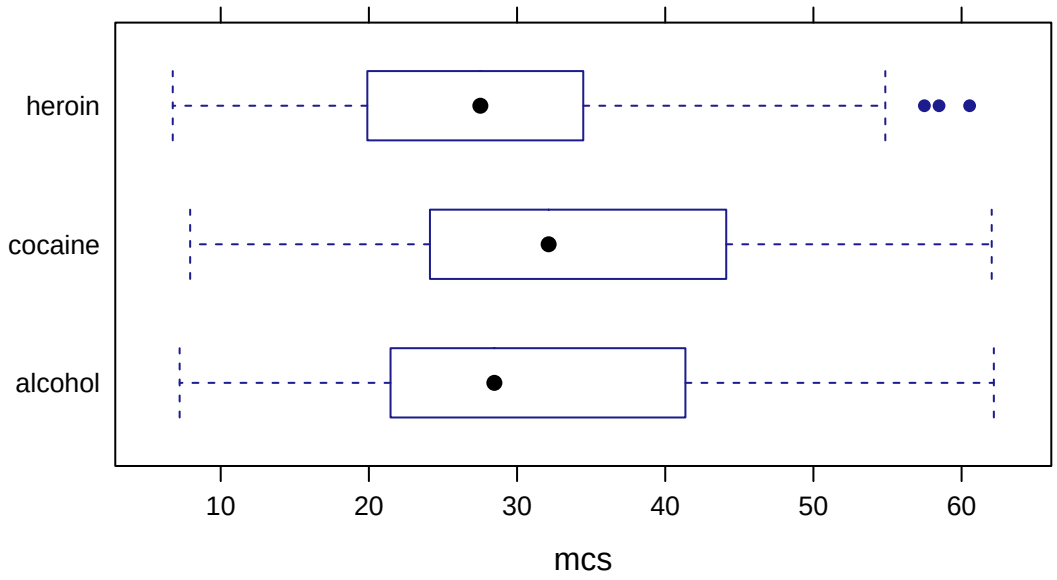


Figure 2: Boxplot of Mental Component Score by primary substance of abuse

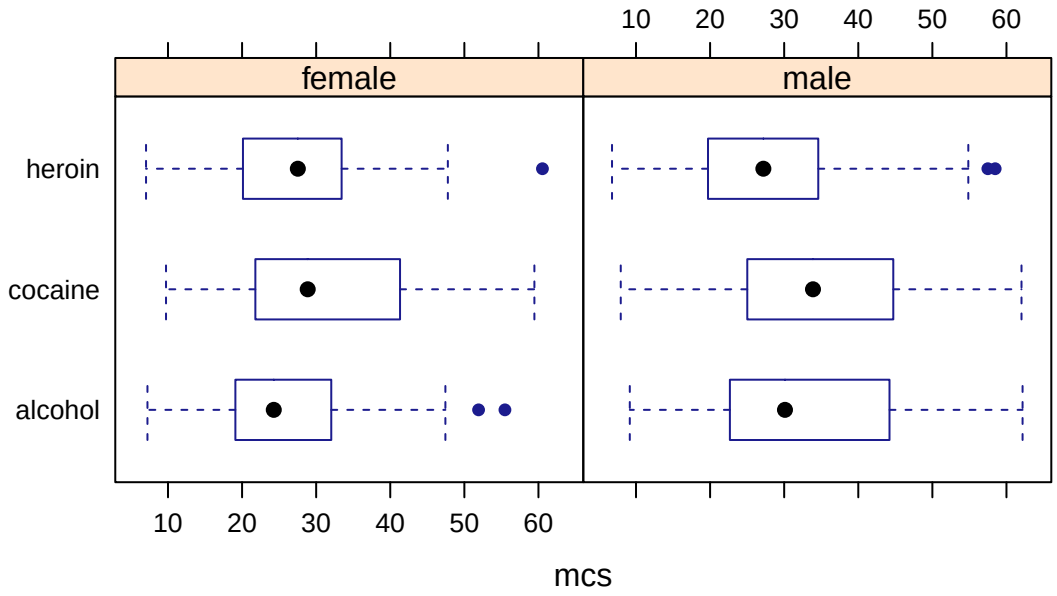


Figure 3: Boxplot of Mental Component Score by primary substance of abuse, grouped by sex

2.2. Requirements

The simplest environment to support this activity is an institutional RStudio server since in this setup, students need only a web browser to run their analyses. RStudio servers are licensed on a per-user basis but are provided free for academic institutions for teaching purposes. Having an RStudio server that all can access on day one helps to overcome several difficulties: students do not have to download RStudio and R on their own computers, which may present its own sort of problem when installation fails; also, the server can be preloaded with some packages that all can use (Çetinkaya-Rundel and Horton 2016); finally, working from a server eliminates the worry of having insufficient computing power on the students' own computers. Alternatively, the activity could be done within a computer lab where RStudio and necessary packages have been preloaded.

The R Markdown system enables students (and instructors) to break up R code into short, digestable chunks that can then be annotated. The beauty of this system is that allows for the easy creation of presentation-worthy reports showcasing the results of analysis; it seamlessly weaves together R code with plain text to produce a single file containing analysis and explanation. In this way, R Markdown provides a compelling argument for getting students used to the process of continuously documenting their work and encourages clear presentation of findings (Baumer et al. 2014). With native support within RStudio, and recently bolstered by a new R Notebook feature that provides automatic previews, R Markdown is straightforward for students to work with even on day one of class. Additionally, compiled R Markdown files

can be shared publicly via RPubs, a free platform for showcasing R Markdown output[1].

For graphing, we utilize the `mosaic` package (Pruim, Kaplan, and Horton 2016a; Pruim et al. 2016b) and take advantage of its simple-to-learn R syntax that helps unite the various different R functions used for data explorations. Conveniently, the syntax for modeling can be described by:

```
GOAL(Y ~ X, data= DATASET)
```

with variations depending on whether there are more (multivariate) or fewer (univariate) variables than the typical outcome and single predictor. Students need to pick a GOAL (e.g. create a scatterplot), specify the variables (X and Y) to study, and the DATASET containing these variables.

As an example, a comparison of mean wages by sex could be generated through the command:

```
mean(wage ~ sex, data=CPS85) # wage "by" sex
```

```
##    F    M
## 7.88 9.99
```

while side by side boxplots could be generated by running (see Figure 4):

```
bwplot(sex ~ wage, data=CPS85)
```

Scatterplots of two quantitative variables can be generated using a similar command (see Figure 5):

```
xyplot(wage ~ age, data=CPS85)
```

Multivariate displays are straightforward to generate. Figure 6 modifies Figure 5 by adding sex as a grouping variable via the '`group=`' argument. The '`auto.key=TRUE`' argument asks for a figure legend matching colors to levels of the grouping variable. The '`type=`' argument optionally allows multiple geometric layers to be displayed on the same plot; popular options are "p" for points and "r" for least squares lines.

```
xyplot(wage ~ age, group=sex, type=c("p", "r"), auto.key=TRUE, data=CPS85)
```

[1] see https://rpubs.com/about/getting-started

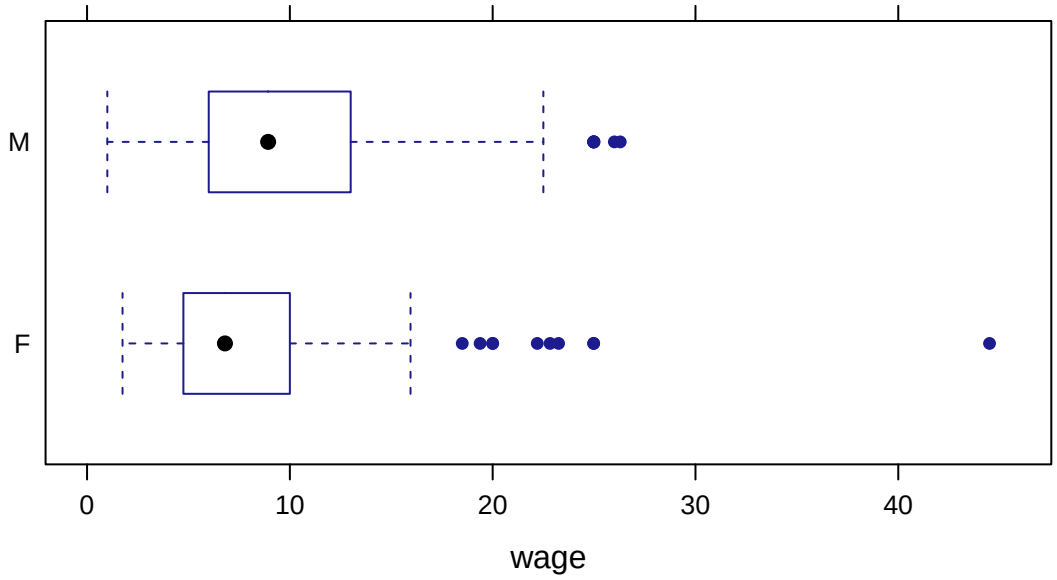


Figure 4: Boxplot of hourly wage by sex

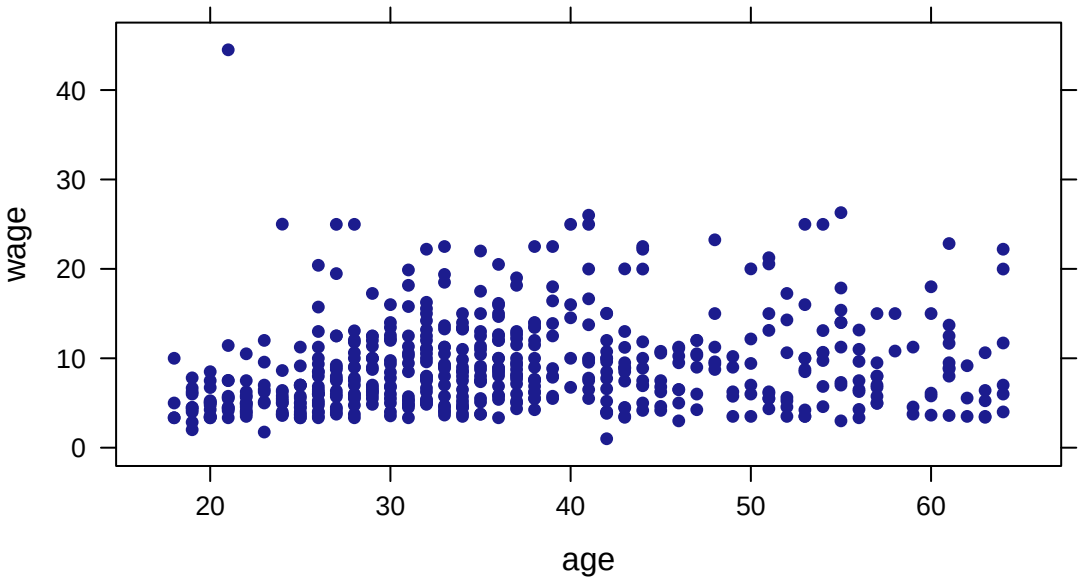


Figure 5: Scatterplot of hourly wage by age

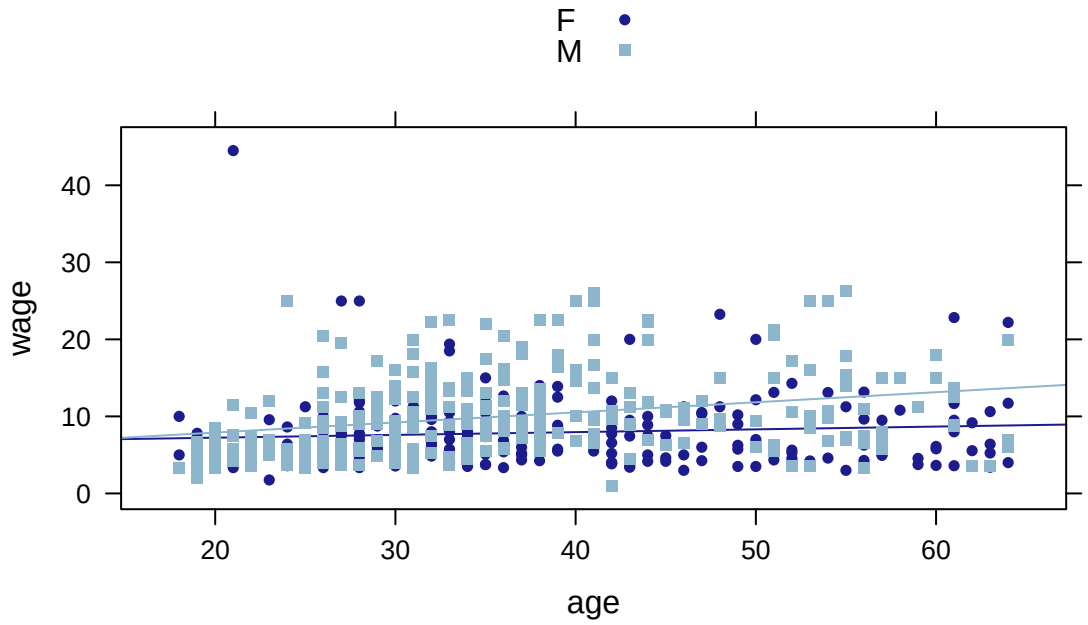


Figure 6: Scatterplot of hourly wage by age, grouped by sex

If we want to incorporate a grouping variable using facetting as opposed to using multiple colors (such as in Figure 3), we could use a vertical pipe to specify our grouping variable. Figure 3 was produced using the command:

```
bwplot(substance ~ mcs | sex, data=HELPrct)
```

We also recommend making use of the `mosaicData` R package (Pruim, Kaplan, and Horton 2015), which contains a number of ready-to-use datasets that would serve well for this activity. Both of our selected datasets are contained in this package.

We have used this lab activity in our classes with an average of 25 students. However, we believe that if course assistants are available to maintain the ratio of one assistant per 15 to 20 students, this activity can scale easily for classes with more students. The purpose of course assistants is to be able to help with hurdles as they arise – missing parentheses, misspelled variables/functions, incorrect capitalization – all of which are commonly experienced by a student working in R for the first time. We detail a number of other common pitfalls in the Discussion.

2.3. Selecting Datasets

It is important to use datasets that are of interest to students (Gould 2010). The datasets we have selected contain a good mix of demographic data (like age, race, and years of education), which students easily comprehend, along with some other aspect that is relevant to them (depression scores and alcohol usage in the 'HELPrct' dataset and wages, work experience, and

sector in the 'CPS85' dataset). In addition to the two datasets presented here, we also recommend the 2013 New York City flight delays dataset (Wickham 2016), the IMDB movie ratings dataset (Wickham 2015), and the 2006 New Haven residential property dataset (Emerson, Green, and Hartigan 2012), accessible in R packages `nycflights13`, `ggplot2movies`, and `barcode`, respectively. We should note that larger datasets, say, with sample size larger than 1000, could be tricky for students to work with due to the time it takes to plot a large number of points and excessive overplotting.

Having about 8 to 15 variables provides a good variety of different possibilities for exploration. Too few variables may yield a smaller chance to identify questions of interest and too many variables would overwhelm. A good mix of categorical and quantitative variables would provide more opportunities to practice with different kinds of data summaries and visualizations. Needless to say, having a large number of rows in the dataset is also important so that, for example, if a student wanted to include explorations of two variables by subgroups of a third, the sample sizes would still be substantial enough to render interesting visualizations.

We also recommend selecting datasets that are already built into R or an R package (or include code to download the data directly into the R Markdown file using the `read.csv()` function in conjunction with a weblink). The benefit of using a dataset that can be loaded with minimal effort is that we can head straight into exploratory data analysis. An added bonus is that most built-in datasets are described using a help page that provides a codebook and description of the variables. The ability to access and interpret help files is an important skill that will inevitably become useful as students delve into the deeper levels of R computing as time progresses.

3. RESULTS

The authors collected student feedback concerning this activity from introductory statistics students at Amherst College. We received permission to report anonymous student feedback and to show examples of student work from those groups of students who have provided consent (Amherst College Institutional Review Board approval #15-028). Out of 50 total students, 39 completed an electronic consent form affirming that their work may be shared, while the other 11 did not complete the form. Rather than pick a few of the examples to share in this article, we have posted all of them online[2].

The students noted strengths and weaknesses of the activity when queried at the one month mark of the course. The following positive aspects of the activity were excerpted from the survey:

- The activity got us to immediately start working in R in a hands-on way. We started with a brief overview and then got to experience the eccentricities of coding in R. Eventually, I became more comfortable with working with R.
- I got to work with data.

[2] http://xiaofei-wang.com/research/vislab/

- I got to meet and work with my classmates. I appreciate that the activity was self-guided and that it gave opportunity for teamwork.
- The activity was challenging and gave us an accurate depiction of how the class is run. We were able to ask for help from the instructor and the lab assistant when we got stuck.

Below are areas where students noted room for improvement:

- The pace was a bit fast. A bit more direction would improve the experience.
- It would be beneficial to discuss the plots (histograms, bargraphs, boxplots, etc.) and compare and contrast them. They utilize different types of variables; I realized this after forming an incorrect plot.
- You asked us to write descriptions before we really knew how to describe.

In summary, most students acknowledged the importance of R and appreciated the chance to experience it early on with support from the instructor and lab assistant. Moreover, the activity served as an ice breaker between classmates, facilitating subsequent group-based activities.

At the same time, not all students appreciated the whirlwind tour of R offered by this activity. Some students craved more guidance and a slower pace. Some students expressed that they felt lost through the exercise. Indeed, the group work began after only fifteen minutes of instruction. However, we believe it is acceptable for a first-week activity to leave students with questions unanswered and a desire to learn more. We recommend setting the right expectations by conveying to students that the activity is intended to provide a first exposure to R, with mastery to be achieved later in the semester.

At the end of the class period, students were able to code and generate a presentable HTML file containing three plots and summaries using a reproducible analysis framework. Incorporating this experience into the first week of class dispelled the sense that the early classes would simply be a rehash of mean, median, and mode, and provided additional confidence for subsequent interactions with R.

Given the critical feedback, a follow-up class discussion might be inserted after the students have learned more about data visualization through readings and in-class examples. This discussion might begin with a critique of a few lab submissions from the activity. For example, we might find an instance where a student plotted a histogram but called it a bargraph in the accompanying description or vice versa. Upon revisiting the activity, some students will have learned that bargraphs are meant for categorical variables and histograms are meant for quantitative variables; reviewing the mistakes from the lab activity helps reinforce these new concepts. We might also take the opportunity to critique the phrasing of some of the descriptions that were written. A student's description of a graph might say "this is a histogram that depicts the distribution of x", to which we can now agree as a class that a better insight would discuss "surprisingly, we see that x is actually bimodal, with peaks at 3 and 5." This

exercise gives students a second look to reflect upon their previous work and see how far they have come since their initial foray.

4. DISCUSSION

An important learning outcome in any statistics class is for students to begin to think like a statistician. Specifically, we believe this consists of repeated practice with posing statistical questions and answering them with evidence backed by data. Software makes this an achievable practice, even in the first week of class, if we exploit the students' curiosity about the world around them. If we provide a dataset about which students have some contextual understanding (even better, misunderstanding), they will naturally pose interesting questions.

In our experience, there are some common pitfalls that students encounter during this activity. When students first download the template R Markdown file, some browsers will force a .txt extension on the file. Students have to change the extension back to .Rmd in order to proceed. One way around this issue is to simply copy and paste the contents that appear as raw text in their browser window into a brand new R Markdown file. Sometimes this act of copying and pasting introduces leading whitespace before some code chunks, which have to be manually deleted in order for the file to compile. Some students may try a lot of different plots in the RStudio console before picking their favorites to submit. In the process of copying their work from the console to their R Markdown script, they may include the "`+`" and "`>`" signs that then break the compilation process. Some students have difficulty distinguishing between code chunks from regular text. In several instances, students learned that the "`#`" symbol creates a comment in R, but placed these comments outside of a code chunk, in which case the comment gets printed as top-level header font in Markdown. Furthermore, some students do not realize that a button or keyboard shortcut allows them to create code chunks, so instead they manually type in the code chunk headers and footers. With incorrect syntax, they then run into compiling issues. All of these pitfalls are part of the learning curve; we expect that students will run into a number of these issues sooner or later when learning R. Experiencing these issues in class gives students immediate assistance when they do arise and helps minimize the friction of learning new software. To make the activity run as smoothly as possible, we highly recommend having one course assistant per 15 to 20 students during this activity.

To reiterate, our proposed activity does not aim to produce experts at data visualization or R coding; rather, it is intended to serve as a pedagogical tool to inspire multivariate thinking. Our goal is to motivate students to ask good, statistical questions and then attempt to answer them with data and a minimal amount of computing. Being able to shed light on those questions, albeit without the rigor of considering significance, within a class period in the first week of class is extremely empowering and helps to whet their appetite for more to come.

The activity is extensible depending on how much time can be allotted for it. On some occasions, we have asked students to share their compiled HTML files on RPubs. This approach is attractive since it allows student findings to be shared with the class as a whole by sharing the appropriate RPubs link. In the instances where we added this step, we found

that students took more time to polish their work, taking greater ownership in the final published product. If additional time is available (perhaps in a second class period), some groups of students can present their plots to the rest of their class. This helps to develop communication skills, overcome the fear of speaking to ones' classmates, and share insights, all early in the course.

5. REFERENCES

Allaire, J. J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., Atkins, A., and Hyndman, R. (2016), *rmarkdown: Dynamic Documents for R*, R package version 0.9.5.

ASA GAISE College working group: R. Carver, Everson, M., Gabrosek, J., Rowell, G. H., Horton, N. J., Lock, R., Mocko, M., Rossman, A., Velleman, P., Witmer, J., and Wood, B. (2016), "Guidelines for Assessment and Instruction in Statistics Education: College Report," http://www.amstat.org/education/gaise.

Baumer, B., Çetinkaya-Rundel, M., Bray, A., Loi, L., and Horton, N. J. (2014), "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics," *Technology Innovations in Statistics Education*, 8, http://escholarship.org/uc/item/90b2f5xh.

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

Çetinkaya-Rundel, M. and Horton, N. J. (2016), "Technology Lowering Barriers: Get Started with R at the Snap of a Finger," in *Electronic Conference on Teaching Statistics*, https://www.causeweb.org/cause/ecots/ecots16/breakouts/7.

Chance, B., Ben-Zvi, D., Garfield, J., and Medina, E. (2007), "The Role of Technology in Improving Student Learning of Statistics," *Technology Innovations in Statistics Education*, 1, http://escholarship.org/uc/item/8sd2t4rr.

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

Emerson, J. W., Green, W. A., and Hartigan, J. A. (2012), *barcode: Barcode Distribution Plots*, R package version 1.1.

Gould, R. (2010), "Statistics and the Modern Student," *International Statistical Review*, 78, 297–315.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, P., Murrell, P., Peng, R., Roback, P., Temple Lang, D., and Ward, M. D. (2015), "Data Science in Statistics Curricula: Preparing Students to 'Think with Data'," *The American Statistician*, 69, 343–353.

Horton, N. J. and Hardin, J. S. (2015), "Teaching the Next Generation of Statistics Students to 'Think With Data': Special Issue on Statistics and the Undergraduate Curriculum," *The American Statistician*, 69, 259–265, http://amstat.tandfonline.com/doi/full/10.1080/00031305.2015.1094283.

Nolan, D. and Lang, D. T. (2012), "Computing in the Statistics Curricula," *The American Statistician*, 64, 97–107.

Nolan, D. and Perrett, J. (2016), "Teaching and Learning Data Visualization: Ideas and Assignments," *The American Statistician*, 70, 260–269.

Pruim, R., Kaplan, D., and Horton, N. J. (2015), *mosaicData: Project MOSAIC (mosaic-web.org) Data Sets*, R package version 0.13.0.

— (2016a), "The mosaic Package: Helping Students to 'Think with Data' Using R," *R Journal*, in press. https://journal.r-project.org/archive/2017/RJ-2017-024/RJ-2017-024.pdf.

— (2016b), "Mosaic: Project MOSAIC Statistics and Mathematics Teaching Utilities," *R Journal*, R package version 0.14.4. https://cran.r-project.org/web/packages/mosaic/index.html.

R Core Team (2016), "R: A Language and Environment for Statistical Computing," Vienna, Austria.

RStudio Team (2016), "RStudio: Integrated Development Environment for R," Boston, MA.

Samet, J. H., Larson, M. J., Horton, N. J., Doyle, K., Winter, M., and Saitz, R. (2003), "Linking Alcohol-and Drug-Dependent Adults to Primary Medical Care: A Randomized Controlled Trial of a Multi-Disciplinary Health Intervention in a Detoxification Unit," *Addiction*, 98, 509–516.

Tufte, E. R. and Graves-Morris, P. (1983), *The Visual Display of Quantitative Information*, vol. 2, Graphics press, Cheshire, CT.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer Science & Business Media, https://github.com/hadley/ggplot2-book.

— (2010), "A Layered Grammar of Graphics," *Journal of Computational and Graphical Statistics*, 19, 3–28.

— (2015), *ggplot2movies: Movies Data*, R package version 0.0.1.

— (2016), *nycflights13: Flights that Departed NYC in 2013*, R package version 0.2.0.

Appendix A: Lab Handout

The Template

The template for most functions (from the `mosaic` package in R) is:

```
goal(     ~     , data =     )
```

Getting R to Work

Each command you type should be guided by the following 2 questions:

1. What do you want R to do?
2. What must R know to do that?

Exploring the Data

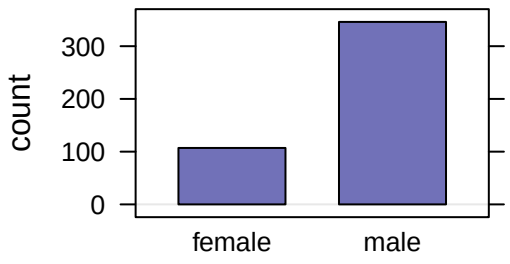
In this course, we'll work with datasets that have a combination of quantitative and categorical variables. Oftentimes, an important first step (before doing any analysis) is to explore the data. Here are some plots that are frequently used to visually display the data.

Univariate Summaries

```
  tally(~ sex, data=HELPrct)
```

```
##
## female   male
##    107    346
```

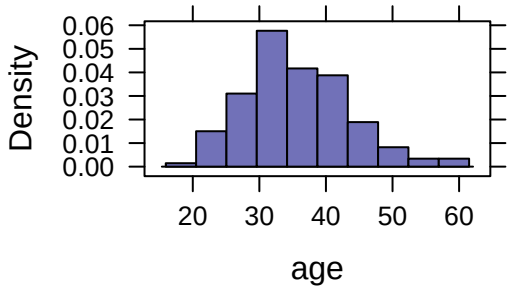
```
  bargraph(~ sex, data=HELPrct)
```




```
  favstats(~ age, data=HELPrct)
```

```
##  min Q1 median Q3 max mean   sd   n missing
##   19 30     35 40  60 35.7 7.71 453       0
```

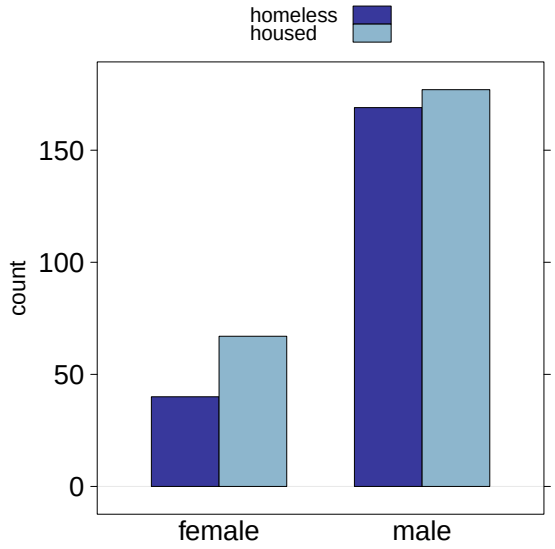
```
  histogram(~ age, data=HELPrct)
```



Categorical var. vs. categorical var.

```
tally(homeless ~ sex, data=HELPrct)
bargraph(~ sex, group = homeless,
                data=HELPrct,
                auto.key=TRUE)
```

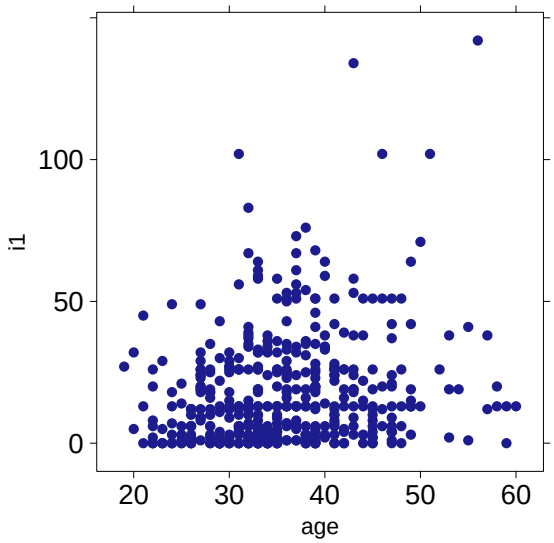
```
##           sex
## homeless   female male
##   homeless     40  169
##   housed       67  177
```



Quantitative var. vs. quantitative var.

```
cor(i1 ~ age, data=HELPrct)
xyplot(i1 ~ age, data=HELPrct)
```

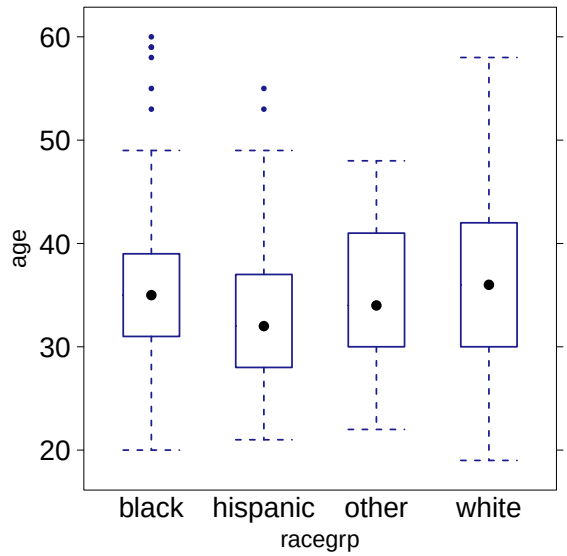
```
## [1] 0.207
```



Categorical var. vs. quantitative var.

```
favstats(age ~ racegrp, data=HELPrct)
bwplot(age ~ racegrp, data=HELPrct)
```

```
##    racegrp min   Q1 median   Q3 max mean   sd   n missing
## 1    black  20 31.0     35 39.0  60 35.7 7.08 211       0
## 2 hispanic  21 28.2     32 36.2  55 33.2 7.99  50       0
## 3    other  22 30.0     34 40.5  48 35.0 7.66  26       0
## 4    white  19 30.0     36 42.0  58 36.5 8.28 166       0
```



Helpful Tips

- R is case sensitive: `x` is not the same thing as `X`.
- In the console, `>` means R is ready for a new command, whereas `+` means R is *waiting for you to finish* an existing command. Hitting ESC gets you out of the latter scenario if you're there by accident.
- Not sure what a function like `msummary()` does? Type the function name preceded by a question mark, like this: `?msummary` to get help. Scroll down to Examples – replicate some of these on your own.
- If R throws you an error, read it before you panic. Usually, the error is more interpretable than you think!

Appendix B: Lab Activity

Instructions

Please delete this entire section before you submit your file to RPubs!

In your groups, explore the `CPS85` dataset within the `mosaicData` package to try to find some interesting insights. You may want to type `?CPS85` and `head(CPS85)` to get a glimpse at what this dataset contains. Next, start exploring the dataset using plots, tables, and other numeric summaries. Select 3 favorite plots and tell a story (in writing) about each of them. Extra brownie points if you can weave the 3 plots together into one cohesive story.

PLOT 1

```
# put the code for your plot here
```

(Include the description for your plot here.)

PLOT 2

```
# put the code for your plot here
```

(Include the description for your plot here.)

PLOT 3

```
# put the code for your plot here
```

(Include the description for your plot here.)