STAT 231: Problem Set 8A

Evan Daisy

due by 10 PM on Monday, May 3

In order to most effectively digest the textbook chapter readings – and the new R commands each presents – series A homework assignments are designed to encourage you to read the textbook chapters actively and in line with the textbook's Prop Tip of page 33:

"**Pro Tip**: If you want to learn how to use a particular command, we highly recommend running the example code on your own"

A more thorough reading and light practice of the textbook chapter prior to class allows us to dive quicker and deeper into the topics and commands during class. Furthermore, learning a programming language is like learning any other language – practice, practice, practice is the key to fluency. By having two assignments each week, I hope to encourage practice throughout the week. A little coding each day will take you a long way!

Series A assignments are intended to be completed individually. While most of our work in this class will be collaborative, it is important each individual completes the active readings. The problems should be straightforward based on the textbook readings, but if you have any questions, feel free to ask me!

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps8A.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps8A.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

1. "Tell the truth. Don't steal. Don't harm innocent people."

In the textbook, the authors state, "Common sense is a good starting point for evaluating the ethics of a situation. Tell the truth. Don't steal. Don't harm innocent people. But, professional ethics also require a neutral, unemotional, and informed assessment."

(1a) Assuming the numbers reported in Figure 6.1 are correct (truthful), do you think Figure 6.1 is an *unethical* representation of the data presented? Why or why not?

> ANSWER: I think Figure 6.1 is still an unethical representation of the data, because it violates an incredibly standard graphical convention which one would not violate unless they had the intention to mislead.

(1b) Pulling from the examples in the textbook, provide one example of a more nuanced ethical situation (one that you perhaps found surprising or hadn't considered before).

> ANSWER: Example 6.3.2 raises an interesting question in pointing out that data analysis produces claims that almost always bear the risk of being false while carrying the perception of being proven fact. It is important to evaluate institutions for possible discrimination, but a false positive in this case would prove disastrous to the institution's reputation. How does one balance the importance of conducting an analysis against the significant impacts of a false conclusion?

2. Does publishing a flawed analysis raise ethical questions?

In the course so far, we've touched upon some of the ethical considerations discussed in this chapter, including ethical acquisition of data (e.g., abiding by the scraping rules of a given website) and reproducibility. At the end of Section 6.3.4 (the "Reproducible spreadsheet analysis" example), the authors ask: Does publishing a flawed analysis raise ethical questions?

After reading Section 6.4.1 ("Applying the precepts") for the "Reproducible spreadsheet analysis" example, re-consider that question: Does publishing a flawed analysis raise ethical questions? And, a follow-up question for consideration: Does it depend on who published the flawed analysis (e.g., a trained data scientist? an economist who conducts data science work? a psychologist who works with data? a clinician who dabbles in data science?)

In 4-6 sentences, respond to those questions and explain your response.

> ANSWER: Publishing a flawed analysis does raise ethical questions, but only if it is feasible that this was done with the intent to deceive. If the individual publishing the flawed analysis is an economist or a dabbling clinician and they make no attempts to hide their workflow, then they may simply have accidentally carried out a flawed analysis, which can later be corrected by experts without ethical concerns being raised. If they are a trained data scientist who witholds information about their analysis process, or publicizes it in Excel or something to make the process less clear, then there may be reason to think that the analysis is deceptive. I would say the answer to this question depends entirely on the researcher's intent, which can be deduced from their background and their willingness to share their reproducible workflow.