STAT 231: Problem Set 8A

Kevin Ma

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 that how it is a little unethical because what the graph is presenting is the complete opposite of what a person would reasonably interpret when looking at a graph of this type. At first glance it seems that the number of murders decreased after the law was passed but instead it had increased.

(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: I thought that algorithmic bias is something that I found surprising considering that some of the issues appear to be unsolvable. For example, the textbook describes how an algorithm the returns predictions about how likely a criminal is to commit another crime was biased against black defendants. I think it is almost impossible to reduce bias in this respect because almost any factor that is correlated with how likely a person is going to commit another crime is likely tied to the defendant's race. As such, this scenario is particularly tricky as there is no easy solution on how to solve this issue.

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: I think that publishing a flawed analysis does raise ethical questions because people can read flawed analysis and believe what they are reading even though there are serious mistakes. As most readers are not going to be experts on the subject, it is unreasonable to put the responsibility on the reader to refute the researcher's results. After reading the section I agree with the textbook stating that data science professionals have an ethical obligation to use tools that are reliable, verifiable, and conducive to reproducible data analysis. I think it also does matter on who published the flawed analysis because research from a supposedly reputable source is more likely to be believed than analysis from an amateaur researcher.