STAT 231: Problem Set 7A

Jack Dove

due by 5 PM on Monday, October 26

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 lanugage 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 ps7A.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 ps7A.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: My opinion of this figure is two-sided. When I look closely, I see red (blood) pouring down and filling up the plot right after 2005, which corresponds to the trend of murders increasing after the stand your ground law. However, my main thought is that, for someone who looks at that plot quickly, they will see a line chart with murders decreasing (they don’t see the reversed scale) after 2005. Overall, as Reuters publishes research and news to the public, their stakeholders are the general public: since the general public is not majoring in data science and likely doesn’t have the time to look at this plot twice, the y axis scale is an unethical representation of the data in question.

(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: The CEO scenario stuck with me after reading the chapter. I thought of myself as a statistical consultant for a hedgefund, with my CEO asking me for a model of corn prices based on temperature. Even though my linear regression model had a coefficient of 0.1 (10 cents increase per degree of temperature), the CEO responds to me that our hedgefund is ESG-friendly, so our models should account for the negative affects on supply of a rising climate (raising the price); he asks me to raise the coefficient to 1.2. My main takeaway from this question was that, in the business world, statistics aren’t just used to say “that’s cool” but for a profit-focused angle. Therefore, I must remember my stakeholders before I deem my analysis as “correct.”

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: A flawed analysis absolutely raises ethical questions. While these textbook examples do provide evidence of this idea, the most memorable example I've seen was when an Amherst Alumnus working at the IMF found flawed economic statistics and analysis published by the Greek Government. He proposed (and found) ethical issues with the figures: the Greeks had dramatically understated their outstanding debts, potentially to instill confidence in foreign investors. And yet, when he brought this dilemma to light, critics asked similar questions of him: did he inflate debt figures in the following IMF report due to anti-Greek sentiments? After reading this chapter and considering my own example, I draw the line between "mistake" and "ethical deviation" at the magnitude of the stakeholder, not the statistician. If your stakeholders are the two people that read your blog, you should follow ethical guidelines, but it is not critical. If your stakeholder is the stability of the global economy, no matter your allegiances, is is critical that your work is reproducible, clearly reported, and especially legal. Fortunately, stakeholder and staticician levels typically sort themselves out, as an amateur statistician would typically not work for the Greek Economic Agency, and a data scientist would not take a job with a blog with two readers. Overall, I believe flawed analysis raises ethical questions, and it is the magnitude of the recipients, not the makers, of the analysis that affect the weights of those questions. (Apologies for the length: I felt that the example provided beneficial context to my idea.)