DATA 240 **Data Science Applications**

Project 2

Assigned: February 02 *Due: February 23/26*

Abstract

For this project we'll be working with data from the SATS survey, which measures students' attitudes towards statistics. You'll be required to load data, wrangle it, explore it enough to understand it, and derive some insights from it. You will need to both write a report and give a final presentation. We will simulate having a boss in the workplace with the help of Dr. Marjorie Bond, who will describe the project, give you the requirements, and listen to your final presentations.

Below you will find more details about the project. However, keep in mind that Dr. Bond is the boss – anything she says supersedes this document!

The final presentation is due in-class on Tuesday, February 23. The final report is due by midnight on Friday, February 26. **Note**: waiting until after the presentation to start on the report will probably result in a very bad grade. Having at least a draft of your report will make writing the presentation easier.

The Files

You can find the data files and some extra files in this shared folder. You will need to login with your Monmouth College account to access it. The `.csv` files contain data about instructors, students, and courses. The "Codebook" file contains descriptions of the various columns. The file `Report content.pdf` describes what a good report should look like. The remaining files present an analysis of a different, but similar, survey. The `SOMAS.Rmd` file is the source file for the report in the `SOMAS.html`, and the file `USRESP_Final_Summer2020.pdf` contains a more thorough description of the SOMAS survey used in the report, and an analysis of it.

Resources

It may be helpful to review R. Two online textbooks might be particularly useful: Modern Dive and R for Data Science. You can find a more complete set of R documentation sources listed here.

Throughout the presentation and the report, your goal should be to use the data to tell a story. You might find this blog post to be a useful resource and jumping-off point for the concept of data storytelling.

First Steps: Loading and Wrangling

You will first need to load the data from the csv files into data frames. This data set has already been cleaned up significantly, but there may still be some invalid student data. In particular, you should look for students that repeated the same pattern for all responses, such as all the same value, or 1 then 2 then 3 then 4, etc. for consecutive items.

Explore and Understand Your Data

You should mostly follow Dr. Bond's lead here, but you will definitely want to review the codebook to make sure you understand the data. You might ask what is the distribution of how the course survey information was obtained? What were the CIDs in which the researchers had to find the info through the web? Who were the instructors of those courses? You might also print out the instructor IDs, CIDs, and year for the course codes having more than 1 CID.

The **checkpoint** presentation should come around this point. You should present for 10-12 minutes about what you have done, showing me how you have explored the data and some visualizations, and your plan for the rest of the project.

You should be trying to get a handle on how researchers and teachers might use this data. That includes weaknesses of it! For example, you might study the attitude construct "effort" globally and with several subsets, explaining why researchers might be hesistant to work with this construct.

Final Presentation and Report

Finally, you need to give a final 20 minute presentation and create a report using R Markdown. These are best worked on in tandem. You could delay polishing the report until after the presentation, but if you wait to start the report until after the presentation you are likely to fail.

For the presentation, imagine that you are presenting to the CEO of a large company. The CEO does not have time to dig into the data or learn about statistical jargon. You need to quickly but clearly present the key findings and insights in the data, including how it might be used or how a future survey might collect data differently. **Dr. Bond is your guide here – she is your CEO!**

Your final report should be an R markdown report. When exported as a PDF or Microsoft Word document, it should be between 3 and 10 pages long. While the presentation contains only the most important insights, this should contain much more detail. You can assume that the reader is a coworker who has taken STAT101, COMP151, and DATA151. You should describe not only the main insights, but also create more visualizations and describe how you approached the project – what went well, what went horribly wrong, etc.