CMSC 320, Homework 4

Due: October 27th

October 1, 2023

1 Project Description

The goal of this project is to get you comfortable with the entirety of the data preparation and exploration process. You will be working off of two similar datasets. **This is a substantial project, and will be worth two homeworks. Start early.**

The two files can be found here:

- Dr. Alam dataset
- Morawski dataset

These datasets are a result of a set of survey questions sent to the two different 320 classes, including a set of demographic questions as well as morality ones. Each row represents a single student entry; each column represents the response.

2 The Dataset

The datasets do include a number of errors. These were not introduced on purpose, as a test, but instead were the result of my general incompetence. In order to proceed, you will need to find and fix them.

There is also one difference between the two datasets: Dr. Alam's class was asked a single, extra question: "Would you describe yourself as compassionate?" before the morality section of the survey. This is in canse you which to investigate the effects of priming.

3 The Assignment

For this assignment, you must come up with three interesting questions about the data, and answer them. You must submit:

- A PDF report of your questions and answers, as well as an explanation of why you think the answer is the way it is. The first section should consist of a list of data errors, how you fixed them, and why you fixed them that way. Below should be a section for each question. Each question must be stated clearly at the top of its section.
- The notebook you used to compute your answers

4 Grading Criteria

You will be graded on:

- How interesting your questions are. I will personally be going through every single assignment, and if I think your questions are boring, I will dock you points. Boring questions include things like "What is the gender makeup of the class?" "How many students answered the survey?" "Did Dr. Alam or Morawski's class answer faster?" Interesting questions may be things like "Did the priming question cause a statistically significant difference in answers between the two sections.

- Whether or not you got the question you asked correct. If you get the wrong answer, you will lose points. However, leniency will be given to particularly challenging questions.
- You will NOT be penalized if the answer to your interesting question is boring. Sometimes interesting questions have boring answers.

You will also be required to:

- Use a hypothesis test for at least one of your questions
- Provide at least one **gorgeous** graph or graphic. Keep in mind you can use pip to install extra packages in your colab notebook. Lackluster graphs will be penalized.
- Unless you are investigating differences between the two datasets, you must merge them.
- Write clearly and concisely. If a word or sentence is not required to communicate your point, omit it.

Extra points will be given for:

- Particularly creative or unique questions
- Any use of external data

5 The Grading Criteria Seems Really Subjective

That is correct! We are simulating you writing a report to your boss. Part of that job is coming up with things that will interest her, be eye catching, and useful. Thinking about subjective things like this will be a large portion of your job. I will likely be reading 500 of these assignments, and I will have no compunctions for taking off points for things that are lackluster, poorly written, or outright boring. The best way to avoid this is to have fun with it and be interested in your own questions. Curiosity is at the heart of data science; indulge it in this project and you'll be fine.