Homework 7

Introduction to Data Science (QR2)

due: **11/3/25**

Instructions: Please submit solutions on Populi. Only a knitted pdf of an `Rmarkdown` file will be accepted.

In this set of exercises, you will get your project data in shape and perform some exploratory data analysis. Please follow the steps below and turn in a professionally written document. You can think of this as a mini-paper that will feed into your final project presentation and write-up.

1. Write out a brief outline of possible analysis steps. You are not married to this outline, but I want you to think carefully about a potential plan-of-attack.

2. Is there missing data? If so, in what columns and how persistent is it? Decide what to do with these observations.

3. Some columns of data are categorical with many categories that may not be discernible.[1] Choose the set of levels you want for each categorical variable and clean the data appropriately. For example, the gender variable might have {male, female, unreported, and unknown} levels. Do you drop the "unknown" and "unreported" observations, keep them, or merge them? Make a decision and discuss. For those not working with the medical school data, clean your data set and report your choices.

4. Once you believe your data is sufficiently cleaned, provide a thorough summary of the data. This includes number of observations, number of variables, content of the variables, and summary tables of your variables, etc. For continuous variables, it is common to report minimum, maximum, mean, and standard deviation. For categorical variables, you can report the proportion of observations in each category. The summary tables should not be screenshots. Please print out your summary tables and display them professionally in your document. There are many ways to do this – ask `ChatGPT` or another AI and it will lead the way.

5. Investigate one bivariate relationship via a plot or table. For example, how is admission rate related to race or how is MCAT related to gender? No regression modeling allowed! Comment on your findings.

[1] I can provide direction for the medical school data if the variable levels are difficult to parse. However, I have given you all of the information provided by these medical schools.