Problem Set 1

Author Name

2023-09-07

Note: *In all of these r codeblocks, I set eval=FALSE. You will need to undo that.*

I will be grading this problem set based on the following criteria: - Quality of code (33%): Is it well-commented? Is it easy to follow? Can I run it? - Quality of graphs (33%): Are they well-labeled? Do they have titles? Do they have legends? Are they formatted well? - Quality of answers (33%): Are they clear? Do they answer the question?

1. Open the .gitignore and add *.csv,* .dta, *.Rdata, etc.

This will prevent you from pushing any large data files to GitHub. GitHub cannot receive data files that are larger than 100MB. You can use Git Large File Storage to submit larger files, but even that has limits.

2. Create folders for data, code, documentation, output, and any necessary subfolders within this repository.

Reorganize the data files and code within these folders.

3. Skim the [paper]https://opportunityinsights.org/wp-content/uploads/2023/07/CollegeAdmissions_Paper.pdf and also reference the non-technical summary and a New York Times report. Briefly explain how the data were generated.

4. Correct the download_data.R to download the .csv to your preferred data folder.

```
school_data <- read.csv(paste0(data,'CollegeAdmissions.csv'))
```

5. First, look at the data. Any quirks about it?

```
school_data
```

6. Look at the variable `rel_apply` and `rel_attend`. What do these mean?

ANSWER HERE

7. Check the documentation for an explanation of what each row is – use that information to calculate the number of rows that there should be in the data. Does this match the number of rows in the dataset?

Answer how you calculated the number of rows and whether it matches the number of rows in the dataset.

```
some_function_to_count_rows(school_data)
```

8. What about the variable `rel_attend_cond_app`. What does this mean? Can you verify that it is calculated correctly?

Hint: You can use mutate() to create a new variable in a dataframe. filter() is useful for subsetting data. Also, the function signif() is very useful for matching up significant figures. Last, stopifnot() is a great way to catch code before there is a mistake.

```
school_data %>% mutate(new_var=someexpression)
```

9. The codebook mentions that most of these data are test-score-reweighted. What does that mean?

ANSWER HERE.

10. Replicate Figures 2a and 2b from the paper. I recommend using ggplot2. Interpret these graphs. Save them to your output folder.

11. Replicate Figures 3a and 3b from the paper. Save them to your output folder. Interpret these graphs.

11. Replicate Figures 4a and 4b from the paper. Save them to your output folder. Interpret these graphs.

12. Come up with your own cool visualization of the data. Save it to your output folder. Interpret this graph.

13. Based on what you have seen so far, do you think that the admissions process is meritocratic? Why or why not? What else would you want to know?