JSC370 2025: Midterm Project

Due Friday, March 7th, 2025 by 11:59pm Eastern Time

Learning Objective To apply the skills learned in the first part of JSC 370 by analyzing and interpreting a dataset of your choice. The dataset should be complex enough that test your skills in data wrangling and cleaning.

Narrative Through this project you will launch your portfolio of data science projects. This midterm is a stepping stone for the final project, which will be published on your github website. The first step in any data analysis is to have a dataset for which you have formulated an interesting question. If you do not have a dataset to work with, you may find inspiration one from our list of suggestions. With your dataset, formulate a clear and concise question to answer and conduct exploratory data analysis, data visualization, and some statistical analysis to explore/answer this question.

Deliverable: A knitted R markdown (or Jupyter Notebook if using Python) written report (.html or .pdf) with embedded tables and figures with a link to the project-specific github repository that you create. The report should have the following sections:

- Introduction: provide background on your dataset(s) and a clear formulated question(s) or hypothesis.
- Methods: include how and where the data were acquired, how you cleaned and wrangled the data, what tools you used for data exploration. **Please note: you cannot use a Kaggle dataset on its own. You must show your ability to use one or more of APIs, data scraping, merging (wrangling).**
- Preliminary Results: provide summary statistics in tabular form and publication-quality figures (e.g. `ggplot2`), use `kable` to write nice tables in Rmarkdown.
- Summary: write about what you found so far from your data in terms of the formulated question. Include a plan (e.g. list of steps and modeling) of what you will do for the final project.

In your report, please *do not* include unformatted output or dataset summaries (e.g. output from head(), str(), etc.). You should summarize these aspects of your data within the text.