JSC370 2026: Midterm Project

Due Wednesday, March 11th, 2026 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 EDA including data wrangling, cleaning, visualizations, and summary statistics.

Narrative 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. You may find inspiration from our list of suggestions. With your dataset, formulate clear and concise question(s) to answer and conduct exploratory data analysis, data visualization, and some statistical analysis to explore/answer this question.

Deliverable: A written report (rendered .html or .pdf, but .html is preferred) that is uploaded to Quercus. In the report there should be a link to a GitHub repo where we can find the .qmd or .ipynb code and dataset. The report should have the following sections (total 85 points):

- Introduction (10 points): provide background on your dataset(s) and clear formulated question(s) or hypotheses.
- Methods (20 points): include how and where the data were acquired, how you cleaned and wrangled the data, what tools (including statistical methods) you used for data exploration. You must use an API and/or scraping to acquire your data. You may supplement with additional data sources (e.g. merging with other datasets that are directly downloaded or that you have from another source), but API usage is mandatory. Please note: you cannot use a Kaggle dataset on its own.
- Preliminary Results (30 points): provide summary statistics in well formatted tables and include publication-quality figures.
- Summary (15 points): 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. The emphasis of the final project will be ML modeling, website development, and interactive visualizations.

In your report, please *do not* include unformatted output or dataset summaries. You should summarize these aspects of your data within the text. Additional things we will look for in your report are code and reproducibility GitHub (5 points), readability/style (5 points).