DATA 31500 Autumn 2024

Assignment 1

Due October 10, 2024

The objective of this assignment is to do exploratory data analysis. Students will use an IDE, programming language, and grammar of graphics API of their choice to produce a computational notebook exploring a chosen dataset. The notebook will be written in a literate programming style, such that the analysis is well-documented, with textual interpretation and rationale interleaved with programmatically generated visualizations.

Students will *work alone*.

Students should submit their computational notebook as a PDF on Gradescope.

Technical Specification

First, **students must choose a dataset** to analyze. This should not be a dataset we discuss in class. Similarly, students should not merely reproduce an existing analysis. Rather students should seek a dataset through a source such as an open-source repository, government website, academic project, or similar. *A high-quality dataset will enable an analysis of sufficient depth to demonstrate analytical skills and of sufficient interest/meaning to support questions and interpretation.*

Students will then **do exploratory data analysis**. This should be done in an `iPhython` or `RMarkdown` notebook in a `literate programming` style. The purpose of this format is well-documented code, which is achieved by interleaving textual interpretation and rationale with code blocks implementing the analysis.

The suggested approach is roughly:

1. Load the dataset and wrangle it as needed for initial visualization
2. Take an initial tour of the data using univariate and bivariate summaries
3. Note any anomalies, patterns, or variables of interest as you discover them
4. Pursue things that caught your attention during the initial tour by creating additional visualizations
5. Offer interpretation and ask questions in response to any informative views of data you create
6. (You might optionally decide to incorporate statistical modeling here, but you are not asked to)
7. Conclude with main take-aways and anything you want to follow up on

As you work, be sure to document your thinking about why you are visualizing the data in a particular way, what stands out to you in the visualizations you create, and what you think

it could mean. Be systematic as you tour the dataset. Let your questions guide you toward more complex (i.e., higher-dimensional) queries as the session progresses.

The analysis should be *complelling*: analysis choices should not seem arbitrary and should identify patterns of interest or questions about the data. The analysis should be *reproducible*: course staff should be able to re-run the analysis to produce the same results and should also be able to trace a student's reasoning about data analysis and visualization design choices. The analysis should be *appropriate for the data*: students should apply techniques that are suitable based on what we've learned about things like data types and encodings.

Last, students will **run and export their notebook as a PDF**. Here's how you do this for `iPython` and `RMarkdown` notebooks.