DATA 22700 Spring 2023

Project

Due May 19, 2023

In this project, students produce an original data analysis and a short written report. Students first find and choose a dataset to analyze. Then, they apply analysis and visualization techniques learned throughout the quarter to uncover a story (or stories) about the relationships in that dataset. The analysis should be compelling, reproducible, and appropriate for the chosen dataset. Students write a short written report to accompany the analysis. The report should interleave text and visualizations to present a narrative account of what students found in the dataset.

Students will *work alone*.

Students should submit their project in two parts on Gradescope: an iPython notebook and a PDF document.

Technical specification

First, **students must choose a dataset** to analyze. This should not be one of the datasets we provided to students earlier in the quarter. We will have a *project check-in on April 27*, where students are expected to present their chosen dataset to course staff for approval. It is important to choose a dataset that can support analysis of sufficient depth to demonstrate skills aquired in the course and of sufficient interest to support a narrative about the analysis akin to a technical report or a piece of data journalism.

Students will then **produce an original data analysis**. This should be done in an iPhython notebook in a literate programming style. The analysis should be *complelling*: analysis choices should not seem arbitrary and should identify patterns of interest that can be woven into a narrative account of 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, encodings, and models.

Last, students will **produce a written report** about the analysis. The write-up should clearly follow from the analysis. All claims in the write up should be consistent with something shown about the data. Visualizations in the write-up should be a subset of the visualizations in the analysis, although we encourage students to polish images for the written report in graphics editing software like Figma (e.g., adding annotations if needed). Figures in the report should have captions. Sources should be cited; we are not strict about citation format as long as the provenance of information is clear. The write-up should be concise (no more than 4 pages, single spaced) and well-written.

Students must *present a clear narrative in technical writing*. This means that arguments should cohere, rely on valid logic, and avoid fallacies or baseless/unsubstantiated assertions. The style of writing should be formal and factual, while also presenting a story about the

data. Storytelling can be difficult, so it may help to look at examples of academic papers and data journalism that make a complelling arguement and reflect on what they do well. Good academic writing in computer science often starts by identifying a problem, summarizing a solution or findings about the nature of that problem, then presenting the approach to the problem in depth, and concluding with a discussion of what was found. However, students do not need to follow this formula. We encourage students to be creative, and demonstrate what they've learned about how to do rigorous analysis and visualization.

The project is intentionally open-ended. Submissions will be evaluated on choice of dataset, quality of analysis, quality of visualizations, and quality of write-up according to the criteria outlined above. The project serves the purpose of a final in this class, so students should put their best foot forward (not procrastinate) and show us what they've learned to do.