DATA 23700 Winter 2025

Project

Due March 10, 2025

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 and answer questions about the relationships in that dataset. The analysis should be compelling, reproducible, and appropriate for the chosen dataset. Students produce 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: (1) the analysis should be submitted as an iPython or RMarkdown notebook in a literate programming style; and (2) the write up should be submitted as either a PDF document or an interactive web page (see below).

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. Similarly, students should not find a dataset assembled for a previous analysis (e.g., by a data journalist) and merely reproduce a previous analysis. Additionally, open-ended exploratory analysis of a clean dataset is insufficient for the project because such project tend to involve too little original work with the data source and not enough substantive conceptual interpretation of the data. Instead students should approach the project with a problem or question in mind. In the past, high-quality student projects involved either: (1) finding a large, messy dataset with many variables and a complex data generating process to analyze; or (2) sleuthing and fusion of multiple data sources to construct a dataset that can answer a targeted question. *It is important to choose a dataset that can support an 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.*

We will have a **project check-in (Exercise 5) due February 3**, where students are expected to submit a draft report motivating the problem or question they plan to address in their project, and presenting the dataset the student plans to analyze. See the specification for Exercise 5 for details.

Students will then **produce an original data analysis**. This should be done in an iPhython or RMarkdown 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 derived from the visualizations in the analysis, but they should be redesigned or polished as needed in order to facilitate clear communication. Students may find it helpful, for example, to finalized images for the written report in graphics editing software like Figma, by adding annotations or labels 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. Although students do not need to follow this formula exactly, *we expect formal, well-organized writing with an introduction that motivates the problem and questions and a concluding discussion of implications.* We encourage students to be creative, and demonstrate what they've learned about how to do rigorous analysis and visualization.

Although we suggest submitting the write-up as a static PDF document, students who are feeling ambitious may alternatively prepare and submit an interactive web page in the style of distill.pub. We only recommend this alternative to students who are comfortable with web development and want an opportunity to hone their skills. Please contact your instructor if you plan to take this option.

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 (i.e., not procrastinate) and show us what they've learned to do.