

Final Project, DATA 22700 Data Visualization

Due on Thursday of finals week

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

This project requires individual work, i.e., each student will *work alone*.

You should submit your project in two parts:

1. The analysis should be submitted as an iPython notebook in a literate programming style.
2. The write up should be submitted as either a PDF document.

Technical specification

First, **you must choose a dataset** to analyze. This should not be one of the datasets we provided to the class earlier in the quarter. Similarly, you should not find a dataset assembled for a previous analysis (e.g., by a data journalist) and merely reproduce the original analysis. In the past, high-quality projects involved either:

1. Finding a large 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 acquired 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.

Exercise 5, i.e., the **project check-in**, asks for a draft report motivating the problem or question that you plan to address in your project. It also asks to introduce the dataset the you plans to analyze. See the specification for Exercise 5 for details.

For the final project, you are asked to **produce an original data analysis** and a **report**. The former should be done in an iPython notebook in a literate programming style.

Data Analysis Specifications

- The analysis should be *compelling*:
 - Analysis choices should not seem arbitrary.
 - The analysis should identify patterns of interest.
 - The patterns identified should 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.
 - Course staff should be able to trace a student's reasoning about data analysis and visualization design choices, i.e., sufficient information about reasoning and choices should be provided.
- 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.
- Analysis visualizations do not have to be 100% polished, but should be easily readable.

Report Specifications

Students will **produce a written report** (write-up) 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. Optionally, it may be helpful to finalize images for the written report in graphics editing software like Figma, by adding annotations or labels.
- 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, including figures and citations.
- Students must *present a clear narrative in technical writing*:
 - Arguments should cohere (be logically consistent and form a whole).
 - Arguments should rely on valid logic.
 - Fallacies or baseless/unsubstantiated assertions should be avoided.
 - The style of writing should be formal and factual, while also presenting a story about the data

Note:

Storytelling can be difficult, so it may help to look at examples of academic papers and data journalism that make a compelling argument and reflect on what they do well. Good academic writing often starts by identifying a problem, summarizing a solution or findings about the nature of that problem (this should include some background), 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 particular formula.

We encourage you to be creative, and demonstrate what you've learned about how to perform rigorous analysis and visualization.

Evaluation (rubric):

The project is intentionally open-ended. Submissions will be evaluated on

- Choice of dataset.
- Quality of the analysis (see specifications above).
- Quality of the visualizations (see specifications above for Analysis and Report parts).
- Quality of write-up (see specifications above).

The following table provides a summary of the criteria that were explained above.

Criteria	Brief Description
Data set choice	Dataset is rich, complex, and clearly suited to analysis goals.
Analysis: compelling, dataset of clear value	Analysis is insightful, clear, and demonstrates strong value from the dataset.
Analysis: reproducibility and design choices	Workflow clearly reproducible; design choices well explained.
Analysis: appropriate methods, visualizations	Chosen methods and visualizations are entirely appropriate and effective.
Analysis: graphical polish	Plots and figures are visually clear, consistent, and professional.
Report: length	Appropriate length; concise but complete.
Report: consistency and truth	Claims well supported by evidence; coherent argument.
Report: narrative clarity	Writing is clear, logical, and engaging throughout.

Criteria	Brief Description (continued)
Report: captions	Every figure and table has an informative caption.
Report: writing/polish	No typos or grammatical errors; professional style.
Formatting: PDF + Notebook	Correct file formats submitted; all materials included.
Scholarly integrity: citations	Citations are complete, consistent, and properly formatted.

The project serves the purpose of a final in this class: you are required to work alone, put your best foot forward, and demonstrate what you've learned to do.