Final Project

Data Wrangling and Visualization

Prof. Jack Reilly

F2025

Overview

Pose and answer a descriptive research question on an aspect of policy or social behavior (very broadly construed - policy, international affairs, economics, politics, sociology, geography, etc - all are fair game) using a large data set. Justify and use all appropriate methods, and present your findings clearly in writing, tables, and graphics. If you are uncertain where to begin, replicating a prior paper with a new dataset is an excellent starting point.

Requirements

In considering your question, there are only three specific requirements:

1. You may use any large dataset you wish, but this dataset should be reasonably sized (N>1000) and involve pre-collected data.[1] Second, this data should involve a reasonable degree of management: I'm looking for more than few perfunctory "here's a graph of a single variable".

2. You should create a reasonable number of graphics, visualizations, or data analyses. I expect at least three major graphic or analysis elements, and you should consider three a necessary, but not necessarily sufficient, condition. We did not cover statisticasl inference in this class, but you are welcome (and encouraged) to use reasonable statistical techniques learned elsewhere - difference in means, regression, etc - to analyze your data.

3. You are strongly encouraged to include a social network or geographical map in your analysis. If it doesn't make sense, though, don't jam it in just for the sake of meeting a requirement.

[1] In other words, I'm *not* expecting you to collect your own data.

Submission

You will turn in three things to me, each of which you will be evaluated on: a report (written in quarto), a presentation, and a series of .R script files.

Report Guidelines

The report should follow standard research paper conventions, but be heavily slanted towards the data half of the equation. Think of it as a research paper without the front half: I do not expect a broad literature review or well-developed empirical theory. Instead, you just pose a data question, motivate it briefly, and answer it. You may find the following suggested structure to be useful:

1. Introduction: What is your research question, and why you find it interesting? What are your independent variables and dependent variables? A short literature review may or may not be useful.
2. Hypothesis: What do you think the answer to your question will be? Why do you expect this? What kind of relationship do you expect between your variables? What confounding variables do you need to control for?
3. Research Design: What data are you using, and what visualization or analysis technique(s) are you using on that data? Why?
4. Analysis and Results: Was your hypothesis correct? What can the data tell you? Make sure to present your findings clearly, concisely, and in an aesthetically appropriate fashion.
5. Conclusion: What can we take from your analysis? If your hypothesis was correct, what does this mean? If your hypothesis was incorrect, what does this mean? You paper should also include an abstract (~250 words) and title page.

On the other hand, if you project is more exploratory in nature (or a replication) you may need to follow a different outline.

Presentation Guidelines

You should target a 5 minute presentation (with slides). Remember: less is more when it comes to text on slides (images and graphics are excellent, though) and practice, practice, practice. It's easy to go on a lot longer (or shorter) than you plan.

Odds and Ends

- .R File Guidelines: As always, any analyses you present in your paper or presentation should be replicated, without change, upon your .do file. Standard rules from class apply - be sure to appropriately comment, annotate, and place headers in your .do file.

- Paper formatting: I don't have particular preferences for specific paper formats, so long as you choose reasonable defaults. Standard quarto is fine.
- Graphics, Tables, etc: All graphics and tables should be professionally and tastefully laid out in order to make digestion of your information as easy as possible for the reader. We talked frequently about this in class, now it's time to put your knowledge to work!
- Replication projects are highly encouraged! In fact, I think it's one of the better ways to go about this project.

Where Should I Find my Data?

Any number of places! Some suggestions:

- Sources from class handouts, readings, and lectures in this class
- "A Dataset of Political Datasets"
- "Datasets for Economists"
- Bauder, Julia. A Reference Guide to Data Sources
- Data sources of the articles and books you read in your other classes

Deadlines and Due Dates

- **Before Thanksgiving Break** (Nov 20): You should identify your project by the class before Thanksgiving break. This means talking to Professor Reilly in or after class and getting the go-ahead. More specificity is always better than less, but don't view this as a huge hurdle: I just want to know - roughly - what direction you're going in. (The journey itself might change a bit in between proposal and final project - that's fine.)
- **Last Day of Class** (Dec 9): Presentations will take place during the last class of the semester.
- **End of Finals Week** (Dec 16): Reports are due at the end of finals week (both a .qmd and a .pdf) along with .R files. They can be submitted directly to Blackboard as a .zip or as a link to a github repository.