RSA Hiring Exercise

Instructions

The following are some basic exercises testing your basic data manipulation and analysis skills. These are not intended to be laborious coding tasks or stumpers but rather a way for you to demonstrate you can think critically about a task and use the necessary tools to complete it. You are allotted 2.5 hours to complete the assignment but it is designed to take no more than 2 hours to complete. If you get stuck on one of the parts I encourage you to move onto the other parts. Note that Part 1 will likely take the most time to complete.

You may use `R` or `Python` to complete these exercises but your final submission must include a PDF with your answers. Working in an Rmarkdown or Quarto file is recommended but not required. You are free to use AI coding assistants so long as you can explain each and every line of code you submit.

Part 0: Git Ready

Most of the lab's datasets are developed and maintained on Github so knowing how to work with Git commands is essential. Create a Git repository (it can be public or private and shared with zayne@mit.edu) and include all your work (including the data, your code, and the final PDF) in it. Your results must be reproducible, meaning the repository can be cloned to someone else's machine and the code will run without issue.

Part 1: Wrangling

`part1_data.csv` contains raw precinct-level election returns for Indiana we want to clean as part of the Precinct Project.

1. Produce a cleaned dataset saved to `part1_clean.csv` that adheres to `part1_codebook.md` reasonably well. You will need to merge in county FIPS codes using `part1_fips.csv`.
2. Are there any issues with the quality of the rawmdata? If so, explain how you found them and how you could go about correcting them (you do not need to implement any fixes).
3. Plot the distribution of votes across all candidates.
4. Is the cleaned data in a long or wide format? What about the raw data?

Part 2: Thinking

Suppose you're interested in the information contained in voter guides about ballot measures. Based on the content of the example guide from Arizona, `part2_voterguide.pdf`, answer the following:

1. What variables would you extract for each ballot measure in order to study the information voters are exposed to about ballot measures? What might you be able to measure using them?
2. How would you structure the dataset you create from the voter guide and why?

Part 3: Reviewing

`part3.ado` is a `STATA` script working with Election Administration and Voting Survey (EAVS) data. You needn't run it yourself or open it in `STATA`, you can open with a text editor if you don't have STATA installed. This exercise is designed to be doable even for someone without STATA experience.

1. What does the script do? You don't need to understand what every variable is or explain every single line but should explain what each chunk of code seems to do.
2. Are there any ways to improve the code?