Term Paper, Part II: Data Collection

PS 270: Understanding Political Numbers

Due Monday, April 1, 2019

For this assignment, you will collect and organize data for your term paper. Your goal is to **create two documents** and upload them to Canvas.

1. A data table that is ready to be imported into R and analyzed *as is*. Your data should include a variable that identifies each unit of observation. (If your data describe countries, there must be a `country` variable in addition to all of the other data). This should be uploaded as a `.csv` file.[1]
2. A codebook describing your unit of analysis and variables. For each variable, you should describe the data source and how it is measured/coded. Your codebook should include a brief update on what else needs to happen before your data are ready for regression analysis in R. See the discussion below for more detail on what exactly that means. Upload your codebook as a `.pdf` document.

As you prepare your data, you aim to complete the following three steps.

1. **Spreadsheets.** You must have data in a spreadsheet/table. Depending on your project, you may be downloading an existing dataset or building your own dataset in a spreadsheet.
2. **Merging.** If data are contained in *multiple* spreadsheets, these sheets must be "merged" (or "joined") along a common variable. For example, two datasets of state-level data will need to be joined along the "state" variable. This is also required if you have multi-level data, e.g. if you want to match individuals to data about the state in which they live.
3. **Recoding.** Data must be "recoded" to make it ready for analysis. Examples of recoding could be converting a categorical variable into a dummy variable(s), collapsing variable values on a scale,[2] converting any missing data to `NA`,[3] rescaling variables into a more intuitive scale,[4] or other tasks that require you to modify the variables in your dataset.

Everybody's data are a little different, so you may not need to engage with all three steps. Survey data don't typically need to be merged, but they usually need some attention with recoding. Global economic data don't typically need as much recoding (other than dealing with `NA`s), but they often require you to merge data tables from multiple sources. Consult your notes *and readings* for help with these steps! You can of course ask us for help, but we may try pointing you toward class resources as a first resort.

---

[1] Export a `.csv` file from R with `write_csv(dataset, here("data", "dataset-file-name.csv"))`. Remember: you can always use `select()` to keep only the variables you want.

[2] One example would be collapsing "strong Democrats," "weak Democrats", and "Independent leaning Democrats" into a category called "Democrats."

[3] Sometimes data come with missing data coded as `-99` or something similar, which we don't want.

[4] For example, it may be easier to analyze GDP per capita if it is measured in *thousands* of dollars rather than *single* dollars. Remember, regression coefficients are the effect of a "one-unit" increase in an independent variable.