Problem Set 1

POLS 8810

Due 6 February 2025

This problem set and most of the future assignments will require the use of the data in the ANES data file (and codebook) available here. (Note that the first time you try to access ANES data, you will need to register for a free account.)

Your goal for this assignment is to prepare some data for future use in an analysis. This means that you will want to thoroughly examine the data and the codebook to ensure that the variables are coded in a way that makes sense for use in a regression analysis and, if not, to recode them as needed.

For this study, you are interested in finding variables to represent the following concepts: Party ID, gender, race, income, and education. The dataset may have more than one possible variable to capture each concept. In this situation, you will need to decide which one is best suited for your use. Since this is often a theoretical question—and you have no theory to guide you for this exercise—you should make data-driven decisions.

Your write-up should contain any information the reader would need to replicate your data manipulations, recoding of variables, variable transformations, etc. It should also provide a descriptive summary of the variables using figures and/or tables as appropriate. However, you should write this up as if this was the Data Section of a research paper, thus you would not talk about the specific things you do with software, but rather the substantive alterations to—and descriptions of—the variables you use. The handout "How to Write the Data/Methods/Results Sections of a Research Paper" should help you learn how to properly write this section of a paper. Additionally, it is very valuable to start reading published work both for the substance, but also as models for how to write like a political scientist. The Problem Set write up is to be submitted *both* via iCollege under /Assessments/Assignments/Problem Set 1 (in order to allow for automated plagiarism detection) and in physical hard copy following the instructions on the syllabus (for grading).

In addition to the write up, you should prepare a replication file with the code you use that should be submitted via email at the same time the problem set is submitted. Your replication file should be formatted as a R script (`.R`) not as an R Markdown file (`.Rmd`)

or any other file type. Good replication code should run on any machine. This means that I should be able to open your replication file on my local machine and get the outputs in your write up without any modifications to the code. *If your code does NOT replicate everything included in your write up, I will NOT grade your problem set submission.* Here is a Handout with some general guidelines for making replication code files.

The primary grade will be based on an evaluation of the student's understanding of how to appropriately prepare and describe the data being used. However, quality/clarity of writing, quality and aesthetics of tables and figures, and other considerations matter as well. The write up should be approximately 3 pages, with 5 pages as an absolute maximum. **Assignments that fail to follow these direction or that cut and paste raw output from the software will be returned ungraded.**

Finally, while I *strongly* encourage you to work together to develop an understanding of data, coding, and the collaborative nature of research, the final write up must be your own. Moreover, it must be typed and written in the manner of the data section of an article manuscript complete with any and all professional looking figures and/or tables.[1]

[1] In a perfect world, no one would ever make another table again (see, e.g., Gelman 2011). However, I accept that I live in the real world, so here is "A Short Guide to Making Good Tables" that I have made to provide some general principle. If you are working in LaTeX, here is a link for a guide to making tables in LaTeX.