POLS 7012: Problem Set 1

Joe Ornstein

Due September 2, 2020

In this problem set, you will create an `R` script that performs some basic analysis of the 2019 ANES (American National Election Studies) Pilot Study. More information about that study is here. (Basically, they test a bunch of questions on a non-random opt-in Internet panel, so don't draw far-reaching conclusions from this dataset.)

Make sure to comment your code so that a reader will know what each line is doing, like this:

```
# Compute the median age
median(data$age)
```

When you're done, upload the `.R` file to eLC. Feel free to work with others in the class, but you must submit your own work.

Create an R Project

R Projects are a great way to organize your workflow. In a nutshell, they keep all your files in one place so that `R` knows where to look. See R4DS Chapter 8 for more detail. I recommend that whenever you start a new data analysis project, your first step should be to create an R Project.

In RStudio, click the "Create a project" button. Put it in a New Directory (where you can easily find it), and title it whatever you want.

Create a subfolder in your project folder called `data/` and put the the data file `anes_pilot_2019.RData` into that subfolder.

Now you're all set up!

Load The Data

The `load()` function loads an `.RData` file. To load the ANES data, run `load("data/anes_pilot_2019.RData")`. (Don't forget the quotation marks around the path.)

You should now have an object called `data` in your environment.

Summarize The Data

- The `nrow()` function counts the number of rows (i.e. observations) in a dataframe. How many observations are in this dataset?
- The `ncol()` function counts the number of columns (i.e. variables) in a dataframe. How many variables are in this dataset?
- What are the `names()` of the variables?

Clean Up The Data

- There is a variable for birth year, called `birthyr`, but no variable for age. Let's fix that. Create a variable called `age`, and set it equal to the current year minus birth year.

- What is the `median()` age of our survey respondents?
- Create a histogram of age with the `hist()` function. (We'll learn how to make prettier ones later.)

A lot of the variables have **missing values**, and it will trip up your data analysis if you don't know where those missing values are.

- Create a `table()` of the variable `vote16`. How many respondents skipped this question (code = -1)?
- In `R`, we typically represent missing values with `NA`. We can recode those values with the power of **indexing**. Try this: `data$vote16[data$vote16 == -1] <- NA`. (Read that line of code as "get the vote16 variable, but only the entries where it equals -1, and assign those entries the value `NA`).
- Create the table again. What happened?

Explore The Data

- Create a `table()` of the variable `liveurban`. Where are our respondents most likely to live? See the ANES codebook to learn what the labels mean.
- Create a two-way table (just the `table` function, but with two inputs) with `liveurban` and `vote20jb`. Who are the rural respondents in our sample most likely to vote for? The urban respondents?
- Skim the codebook and find three variables that you think are interesting. Summarize each one, either with a `table()` for categorical variables or the `mean()` for continuous variables.