Programming Assignment 1

SOC Methods Camp

September 3rd, 2019

Step zero: Loading and examining the data

- Load the anespilot16.csv data, view its dimensions, check what type (class) of object it is, and view the first few rows and first 10 variables
- Install (if not already) and load the dplyr and ggplot2 packages

Step 1: Creating variables of interest to review different data types

Task 1: Create a vector, *varsinclude*, that contains the following variables of interest for our analysis:

Feelings about free trade:

- freetrade: views about free trade

Rankings of which issues are most important:

- ISSUES_OC14_10: where the respondent ranks unemployment as an important issue (out of 21 issues)

Feelings about candidates:

- ftsanders: feeling thermometer for Sanders
- fttrump: feeling thermometer for Trump
- fthrc: feeling thermometer for Hillary Clinton

Demographic variables:

- gender
- race
- educ
- birthyr

Question 1: What type (i.e. numeric, logical, etc.) of vector should *varsinclude* be? Confirm your answer by using the "class" command

Task 2: Use either the appropriate dplyr command to create a new data.frame, anes2, that only includes those relevant variables. *Hint: if using dplyr, the command one_of may be useful*

Then, check the dimensions (should be 1200 x 9)

Now, we're going to work on getting the variables into more usable form.

Task 3: Using the codebook for the study and the appropriate R command, create a new factor variable- racenew- that collapses the racial groups into the following categories and labels them appropriately:

- 1 = white

- 2 = black
- 3 = hispanic
- 4 = other (includes all non-hispanic categories)

Save the results into a new data.frame, *anes3*. Now we're going to clean some of the other variables.

Task 4: Use the codebook and the appropriate R command in dplyr, create labels for the levels of the following variables. When creating these labels, it's up to you whether you want to create new variables or just save over the existing names

- gender
- educ
- freetrade

Save the results in *anes3*

Step 2: logical statements and filtering

Task 1: To make sure the continuous variables are usable, use the "summary" command to view the range of the following heat thermometer variables we will be examining (make sure to use indexing so it only summarizes those three variables):

- fttrump
- ftsanders
- fthrc

The feeling thermometers should go from 0-100, but what is the range of these three variables?

Task 2: Use the codebook to:

1. For each of the variables, recode values above 100 to NA
2. Exclude from the data observations that have NA for at least one of the heat thermometer variables - there should be 1188 observations remaining

You want to see if there's a meaningful enough subset of respondents who oppose free trade to look at whether they're more likely to support Trump or Sanders than Clinton.

Task 3: Using either dplyr, find and print the number and percentage of respondents who *either* oppose free trade a little, a moderate amount, or a lot

Question 1: How many and what % of respondents oppose free trade at least a little bit?

Task 4: Now you want to plot the distribution of feelings about Trump versus Sanders versus Clinton among those who oppose making free trade deals with other countries. To make this plotting easier by avoiding logical statements inside the plotting function, create a dataframe, *opposers*, composed of just those who oppose free trade at least a little (i.e. oppose a little, moderate, or greatly)

Task 5: Now, using *ggplot2* and *grid.arrange* from the package *gridExtra* (use the help file to find out the syntax), create the following 3 plots side by side (so a 3 column grid). No need to add titles and axis labels yet, but you can for extra credit:

- Density of feelings about Trump among opposers
- Density of feelings about Sanders among opposers
- Density of feelings about Hillary among opposers

Keeping in mind that 0 = dislike and 100 = like, what does the graph show?

Step 3: look at education as a confounder

You notice a trend but are wondering if it stems from an omitted variable- a person's education level, where persons with lower educational attainment might both be more likely to support trump or sanders and also more likely to oppose free trade because of the sectors in which they work

Task 1: Going back to the *anes3* data, do the following:

1. Create a numeric version of the freetrade variable called freetradenum- this will allow you to take the mean
2. use dplyr to find the mean views about trade (higher = more opposition) for each educational category, using *na.rm* in the *mean* command to exclude NA's from this calculation (the *mean* command won't run if you have missing values), and then to rank the education groups from most opposed to least opposed.

Question 1: Which educational groups are most opposed to free trade?

Task 2: Now, use dplyr to find the mean thermometer rating for trump, sanders, and clinton by education group, creating three summary measures using the *summarise* command: *trumpheat*, *sandersheat*, *hilheat*. Display the data so that rows are education categories and the columns are the candidates.

Task 3: You're also interested in descriptively exploring what role gender might play in these views– find the mean degree of opposition to free trade for each education and gender category (e.g., males with 2 year, females with 2 year, males with some college, etc.) using dplyr

Then, order from most opposed (highest score) to least opposed (lowest score)

What patterns, if any, do you notice?

Task 4: You're confused about why postgrad females seem one of the most opposed to free trade– use subsetting in dplyr to find out how many males versus females with a postgrad education were surveyed

Step 4: for loops for sampling

You're still perplexed by the somewhat unexpected finding that postgrad females seem to oppose free trade so much. Given that the sample is somewhat small, you want to make sure that there is not one observation that is unduly influencing the results. One way to do this is to use a loop to iteratively remove each observation, take the mean of the freetrade variable in the rest of the sample with that observation removed, and then plot those means.

Note: Of course, a simpler way would just be to plot the values of that variable and look for outliers, but the structure of this loop format can be adapted to do more advanced analysis other than taking the mean, like running a regression model and extracting the coefficient of interest. For simple functions like means, it is pretty clear how an outlier input can contribute to a skewed output, but for more complicated functions like regressions (which you'll be doing a lot of), it is harder to see how one outlier input is affecting the entire model.

Task 1: Subset the data to include only females with a postgrad education and save it as a dataframe called *pgfem*

Task 2: to make sure the "meat" part of the for loop is correct, practice running with one iteration

Remove the first row of the *pgfem* dataframe and save it as a new object called *pgfem.minus1*. Then take the mean of the opposition to free trade variable in *pgfem.minus1* and save it as an object called *ftmean*.

Question 1: What type of object is ftmean?

Task 3: then, use that general code to generate a for loop

Task 4: Plot

Use ggplot2 to do a density plot of the means. *Hint: before inputting, you'll have to turn the vector into something that adheres to ggplot's plotting requirements such as a data.frame*

Task 5: Compare to mean of all females with post-grad education

What was the mean value of the *freetradenum* variable in the entire *pgfem* data? To your plot, add a red vertical dashed line that intersects the x-axis at this value.