Homework 2

Week 3

Question 1

In this question, we will explore data about the jobs that college graduates received, depending on their major while in college. The data come from the US Census Bureau's 2010-2012 American Community Survey, and were aggregated and posted publicly at this website.

a. Begin by reading the file "recent-grads.csv" into R. How many rows and columns are there in the data? What is the unit of analysis of the data?

b. How many different "major categories" is the data divided into? Which category has the most majors in it? How many majors are in this category?

c. In total, how many women are included in the dataset? What percentage of people in the dataset are women?

d. Which major had the highest percentage of women graduates? Which one had the lowest percentage? What where those two percentages?

e. How many people in the data majored in a field in the "Health" category? What percentage of those people have a full-time, year-round job?

f. Create a variable that tells you the spread in (i.e. the difference between) the 25th and 75th percentiles of income earned by people who received each major. Among majors with an unemployment rate of less than 6%, which one had the largest spread in salaries?

Question 2

For this question, we will clean a messy dataset so that it can be used for analysis. The data we will use ("exit-poll-2016.RData") is a version of the Pennsylvania exit poll survey from the November 2016 election. The data include responses from 2957 Election Day voters. Before digging into the questions below, take a couple minutes to load the data and familiarize yourself with what is in it.

a. What is our unit of observation in these survey data? Does the dataset contain exactly one row for every observation? If not, write code to reorganize the data so that each observation shows up as one, and only one, row. After you've done that, write a line of code to check whether there is one row per respondent (who are each assigned a unique number in the id column). Lastly, recode the numeric values of any new variables into meaningful labels (such as 'favorable' and 'unfavorable').

b. The information about each respondent's education has been split across four columns. Reformat the data so that there is a single variable called educ that contains four possible values ('hs', 'some college','bachelors','postgrad'). Set any missing values to NA.

c. Now that the dataset is in good order, we can move on to cleaning individual variables/columns. Start by splitting the sex.age.race variable into three separate columns. Clean those new variables so that missing or unknown values are coded as NA.

d. Create a new variable called third.party . This variable should equal 0 if the respondent voted for Clinton or Trump and 1 if the respondent voted for another candidate. If the respondent did not vote, or we do not know whether or not they voted, this variable should be set to NA. After you're done, remove PRSPA16 from the dataset.

e. Often data scientists or social science researchers will refer to a variable that is coded as True/False or 0/1 as a 'dummy' or 'indicator' variable. Convert the married variable into a dummy variable, where 1 (or TRUE) indicates that somebody is married and 0 (or FALSE) indicates that they are not.

f. Recode the PHIL3 and partyid variables so that their values have meaningful labels, rather than just arbitrary numbers.

g. Make sure that every column has a name that is adequately descriptive and uniformly formatted.