R Homework #2

A Test of Two Means & More Data Cleaning

Due: November 20th, 2018 via email before lab; you are also expected to hand in a printed version of script and plots in lab. No late work will be accepted.

Important Details:

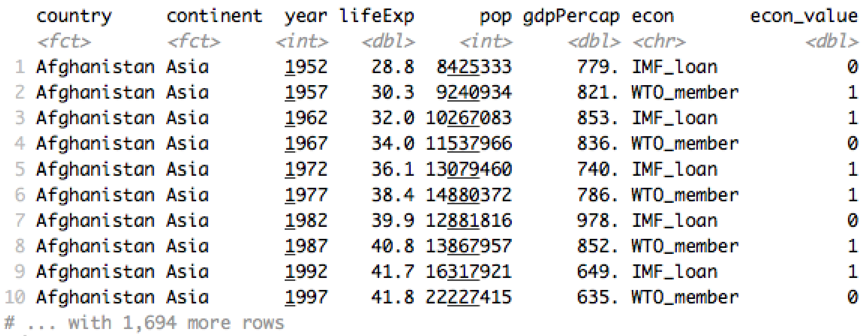
- Once again: three files. A script file, .png files of your plots, and a text/word document with responses to questions four and five.
- You may (indeed, I would encourage you to) work with other fellows to solve these problems, but you should list everyone you worked with as an annotation at the top of your R script file (e.g. #Collaborated with Dan and Kim).
- You may also visit the consultants at the Center for Social Science Computation and Research during their copious and friendly drop in hours (http://csscr.washington.edu/)
- You may use google to discover answers to challenges or questions as you are coding.
- A reminder on replication: I should be able to run your script from beginning to end on my own machine and get all the same results that you did. This is what it means for code to be "replicable" – the script file needs to contain everything for me to replicate your results and plots, without me doing any typing.
- You will be graded on – from most to least important – the replicability of your code, the clarity of your code and annotations, the elegance of your plots, and the thoughtfulness and thoroughness of your written responses.

Start a new script file and name it "lastname_HW2.R".

Question 1

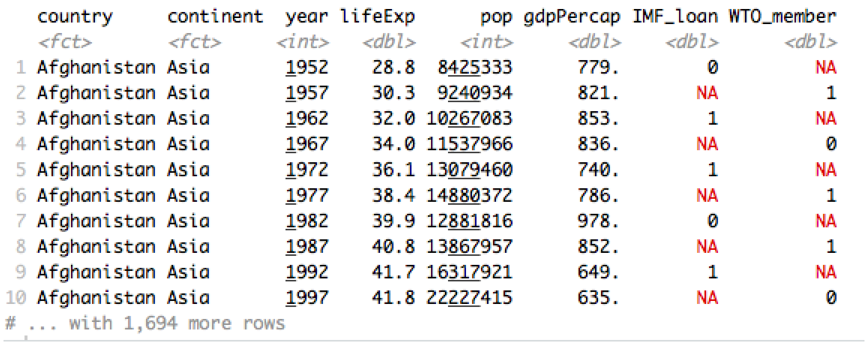
Load the gapminder data and the packages you will use (dplyr, tidyr, and ggplot2).

Save the gapminder data as an object called "data." Create two new variables (two new columns) in "data." The first should be called "econ" and the second should be called "econ_value." The variables will tell you (with made-up values) whether or not a country was in the World Trade Organization in a given year, and whether or not they had a loan with the International Monetary Fund. Fill in all rows of "econ" with alternating values of "IMF_loan" and "WTO_member," as shown in the snippet below. Fill in all rows of "econ_value" with a repeating pattern "0, 1, 1". (Hint: the function "rep_len" will be useful to you here.)



```
   country     continent  year lifeExp      pop gdpPercap econ       econ_value
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl> <chr>           <dbl>
 1 Afghanistan Asia       1952    28.8  8425333      779. IMF_loan            0
 2 Afghanistan Asia       1957    30.3  9240934      821. WTO_member          1
 3 Afghanistan Asia       1962    32.0 10267083      853. IMF_loan            1
 4 Afghanistan Asia       1967    34.0 11537966      836. WTO_member          0
 5 Afghanistan Asia       1972    36.1 13079460      740. IMF_loan            1
 6 Afghanistan Asia       1977    38.4 14880372      786. WTO_member          1
 7 Afghanistan Asia       1982    39.9 12881816      978. IMF_loan            0
 8 Afghanistan Asia       1987    40.8 13867957      852. WTO_member          1
 9 Afghanistan Asia       1992    41.7 16317921      649. IMF_loan            1
10 Afghanistan Asia       1997    41.8 22227415      635. WTO_member          0
# ... with 1,694 more rows
```

Next, clean this data using the tidyr function "spread" so that your data looks like this snippet:



```
   country     continent  year lifeExp      pop gdpPercap IMF_loan WTO_member
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>    <dbl>      <dbl>
 1 Afghanistan Asia       1952    28.8  8425333      779.        0         NA
 2 Afghanistan Asia       1957    30.3  9240934      821.       NA          1
 3 Afghanistan Asia       1962    32.0 10267083      853.        1         NA
 4 Afghanistan Asia       1967    34.0 11537966      836.       NA          0
 5 Afghanistan Asia       1972    36.1 13079460      740.        1         NA
 6 Afghanistan Asia       1977    38.4 14880372      786.       NA          1
 7 Afghanistan Asia       1982    39.9 12881816      978.        0         NA
 8 Afghanistan Asia       1987    40.8 13867957      852.       NA          1
 9 Afghanistan Asia       1992    41.7 16317921      649.        1         NA
10 Afghanistan Asia       1997    41.8 22227415      635.       NA          0
# ... with 1,694 more rows
```

Next, remove both of these new columns that you've created from "data" object.

Question 2

Research Question: did age impact survival among the passengers on the ill-fated Titanic?

First, load the .csv of Titanic data, setting header to true and strings as factors to false.

Next, calculate the mean age and standard deviation of those who survived and the mean age and standard deviation of those who did not. Print your results in a single, clean table.

Next, plot two histograms or density plots on top of each other – i.e. two histograms in one plot – where one histogram shows the distribution of ages of those who lived and one shows the distribution of ages of those who died.

Next, create two new dataframes – one called "lived" and one called "died" – separating the data into those who lived and those who did not. What is the difference in the mean ages of each group?

Finally, calculate a two-sample t-test for the mean ages of each group. How confident are we that we can reject the null hypothesis that the two population means are, in fact, equal? What is our margin of error, with 95% confidence, around the difference in means? Explain in a paragraph – in substantive terms that a person who doesn't know anything about statistics would understand – the results you have found. What is the answer to our research question about whether age and survival are related?

Question 3

Using your own data set for one of your variables, plot a histogram of its distribution. Divide it into two groups of interest – this is just an exercise, so they don't need to be groups necessarily relevant to your eventual analysis – and plot two overlaid histograms.

Question 4

Write out a table with all of your dependent and independent variables. Describe where you will get each dataset, report whether you have downloaded or collected the data set, describe what steps you will need to complete in cleaning, and report when you expect (approximately) to have completed those tasks – i.e. when you expect to have all of your cleaned data for your dependent and independent variables.

Question 5

Begin to think about "control variables" – the kinds of differences across your cases that could influence the relationship you are hypothesizing between your independent and dependent variables. Make a list of five things you would like to control for across your dependent variable cases and give a 1-2 sentence justification of each.