**IST687 – HW 3 - Cleaning/Munging Dataframes**

**Reminders of things to practice from last week:**

Make a data frame data.frame( )

Index of max/min which.max( ) which.min( )

Sort value or order rows sort( ) order( )

Descriptive statistics mean( ) sum( ) max( )

**This Week:** Often, in data science, when you get a dataset, it is not in the format you want. So, you have to use R code to refine the dataset into something more useful and sometimes merging multiple datasets. As Chapter 6 of Introduction to Data Science mentions, this is called “data munging.” In this homework, you will read in a dataset from the web and work on that dataset (in a dataframe) so that it can be useful. Then, you will explore the distribution if a variable within the dataset and finally, merge two datasets (dataframes).

**Step A: Use read.csv( ) and url( ) to read a CSV file form the web into a data frame**

1. Use R code to read directly from a URL on the web. Store the dataset into a new dataframe, called dfStates. Use stringsAsFactors=FALSE. The URL is:

https://www2.census.gov/programs-surveys/popest/datasets/2010-2017/state/asrh/scprc-est2017-18+pop-res.csv

**Step B: Clean the dataframe**

1. Use View( ), head( ), and tail( ) to examine the data frame.
2. Remove unneeded columns and rows by using the minus sign in the rows or columns of the [ , ] accessor.
3. Remove the last Row (for Puerto Rico)
4. Make sure there are exactly 51 rows (one per state + the district of Columbia).   
   Hint: remove Puerto Rico and the summary for the united states
5. Make sure there are precisely 4 columns, with the following names:  
   stateName, population, popOver18, percentOver18.   
   Hint: use colnames( ) and you will need to remove some columns

**Step C: Create a Function**

1. Create a function that takes no parameters and returns the clean dataframe created in step 6 above.

**Step D: Explore the dataframe**

1. Calculate the average population of the states
2. Find the state with the highest population (use which.max)
3. Create a histogram of the state populations, what do you observe?
4. Sort the data frame by population (hint the use ‘order’ function)
5. Show the 10 states with the lowest populations
6. Use barplot( ) to create a plot of each of the population from the sorted dataframe. What do you observe?