**IST687 – Getting Started with ggplot2**

**Reminders of things to practice from previous weeks:**

Descriptive statistics mean( )

Install a package install.packages( )

?command Ask R for help with a command

**This week:** Data visualization is important because many people can make sense of data more easily when it is presented in graphic form. As a data scientist, you will have to present complex data to decision makers in a form that makes the data interpretable for them. From your experience with Excel and other tools, you know that there are a variety of common data visualizations (e.g., pie charts). How many of them can you name?

The most powerful tool for data visualization in R is called ggplot. Written by computer/data scientist Hadley Wickham, this “graphics grammar” tool builds visualizations in layers. This method provides immense flexibility, but takes a bit of practice to master.

**Step A: Load and Merge datasets**

1. Read in the census dataset (using the function created in HW 3)
2. Copy the USArrests dataset into a local variable (similar to HW 2)
3. Create a merged dataframe -- with the attributes from both dataframes  
   *Hint: get the state names from the USArests dataframe with the rownames   
   Hint: use the merge command*

**Step B: Explore the Data – Understanding distributions**

1. Create a histogram using GGPLOT for the population and a different histogram for the murder rate

*Hint: Don’t forget to install and library the ggplot2 package.*

Ensure each line of code is explained (comments) in terms of what it is doing. Then build similar code to create histograms of each of the other three variables in the merged data frame. What parameter will you have to adjust to make the other histograms look right?

1. Create a boxplot for the population, and a different boxplot for the murder rate.
2. Create a block comment explaining which visualization (boxplot or histogram) you thought was more helpful (explain why)

**Step C: Which State had the Most Murders – bar charts**

1. Calculate the number of murders per state
2. Generate a bar chart, with the number of murders per state  
   *Hint: use the geom\_col function*
3. Generate a bar chart, with the number of murders per state. Rotate text (on the X axis), so we can see x labels, also add a title named “Total Murders”.
4. Generate a new bar chart, the same as in the previous step, but also sort the x-axis by the murder rate
5. Generate a third bar chart, the same as the previous step, but also showing percentOver18 as the color of the bar

**Step D: Explore Murders – scatter chart**

1. Generate a scatter plot – have population on the X axis, the percent over 18 on the y axis, and the size & color represent the murder rate