# Prep Exercise (PE06) Data Prep for Visualizations using GGPLOT

### General Instructions

1. For this exercise you will answer all of the questions in this document and turn it in to Blackboard.
2. Before you get started make sure to **read Chapters 12 of *An Introduction to Data Science* and execute the code throughout the chapter to gain familiarity.**
3. Getting Started: 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?

As usual we’ll use the Prep Ex to clean up our data and this time, we will also merge two datasets, using the built-in merge( ) function, which provides a similar capability to a JOIN in SQL. Many analytical strategies require joining data from different sources based on a “key” – a field that two datasets have in common. Specifically, we’ll revisit the USArrests dataframe that is built-in to R as well as the census dataset.

We’ll also start to use one of the most powerful tools for data visualization in R - 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. But first we’ll be cleaning up our data.

### Prep Exercise

**Step 1: Use the merge command to create a new dataframe**

1. Code and Execute the following block of code. Be sure to understand each line. As a reminder, you created the ‘readStates’ function in a previous homework assignment, so you should be able to reuse that code. Add comments before each line to explain in detail what each line of code does.

**#obtaining the cleaned census dataframe from readStates() function**

states <- readStates()

**#copying the USArrests dataframe in R into arrests dataframe**

arrests <- USArrests

**#creating a new column stateName in arrests dataframe that takes row names**

arrests$stateName <- rownames(arrests)

**#merging states and arrests dataframes by stateName**

mergeDF <- merge(states, arrests, by = "stateName")

**Step 2: Use ggplot to start to explore our merged dataframe**

1. Install and library the ggplot2 package.
2. Code and Execute the following block of code (actually type, do not copy/paste). Add comments before each line to explain in detail what each line of code does. Add an appropriate title for the chart (using ‘ggtitle’)

**#creating a ggplot object for mergeDF dataframe**

ggplot(mergeDF) +

**#setting y-axis to Murder for the ggplot object created**

aes(y=Murder ) +

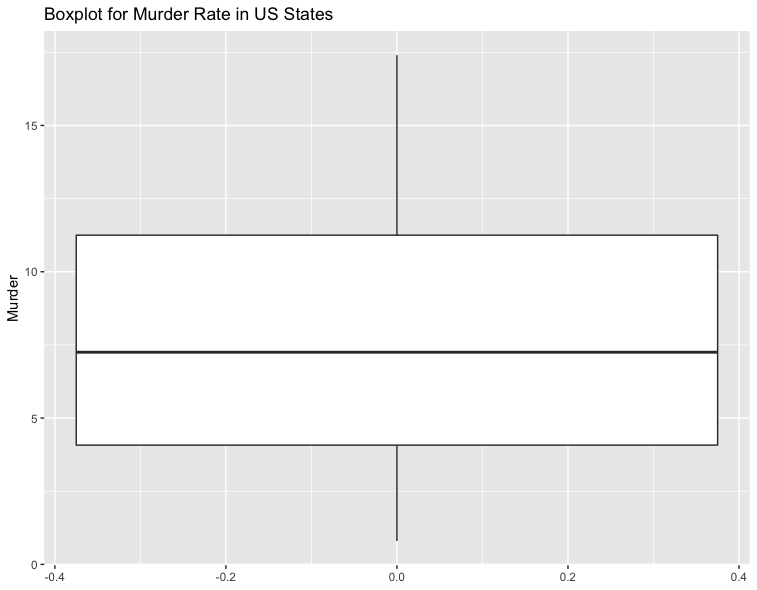
**#creating a boxplot**

geom\_boxplot() +

**#adding title to the plot**

ggtitle("**Boxplot for Murder Rate in US States**")

1. Cut and paste an image of the visualization created by the ggplot and explain what you see



**The box plot visualizes the following five summary statistics for Murder data.**

* **Median is approximately 7.5**
* **Lower hinge is less than 5**
* **Upper hinge is more than 10**
* **Lower whisker is closer to 0**
* **Upper whisker is around 17.5**

**From the plot we can infer that few states have murder rate below the first quartile and significantly large number of states have murder rate above the third quartile. Murder rate for most of the states is above the median (second quartile).**

1. Code and Execute the following block of code (actually type, do not copy/paste). Add comments before each line to explain in detail what each line of code does. Add an appropriate title for the chart (using ‘ggtitle’)

**#initializing a ggplot object for mergeDF dataframe with Murder as x-axis**

myPlot <- ggplot(mergeDF, aes(x=Murder))

**#creating a histogram with binwidth as 2, color of bins as white and bin border as black**

myPlot <- myPlot + geom\_histogram(binwidth=2, color="black", fill="white")

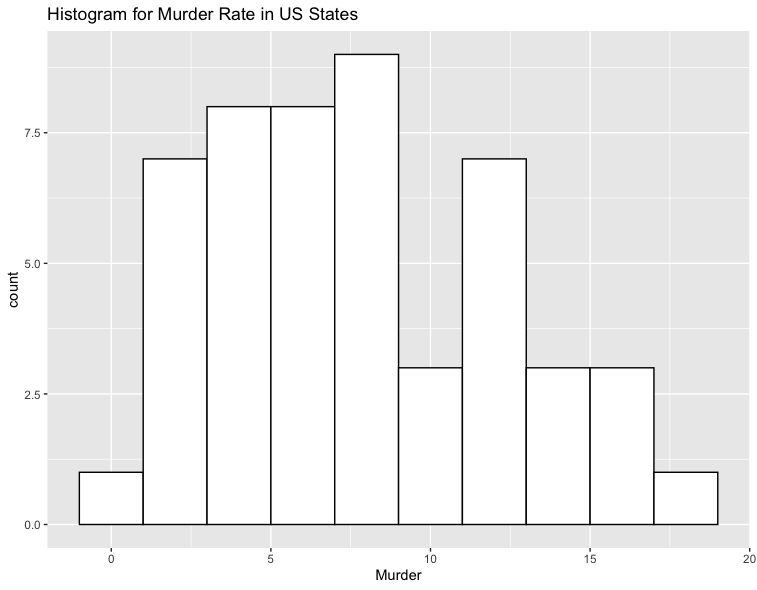
**#adding title to the plot**

myPlot <- myPlot + ggtitle("**Histogram for Murder Rate in US States**")

**#displaying the plot**

myPlot

1. Cut and paste an image of the visualization created by the ggplot and explain what you see

****

**Large number of states have murder rate less than 10. Few states have murder rate between 10 and 15. Very few states have murder rate greater than 15.**

1. **List any additional resources you used here.**
2. **Be sure to save your R file as this will become the starting code for your homework.**

***You must submit all Prep Exercises to blackboard prior to the deadline specified for each assignment.*** PE assignments are due on the evening prior to the lecture class. Late PE assignments will not be accepted for credit.