# 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.

* # IST 687, Standard Homework Heading
* #
* # Student name: Thadhani Hitesh Chandrakumar
* # Homework number: PE06
* # Date due: Wed 2nd Oct 2019 11:59PM
* #
* # Attribution statement: (choose the statements that are true)
* # 1. I did this work by myself, with help from the book and the professor
* # 2. I did this homework with help from the book and the professor and these Internet sources: <provide the urls
* # 3. I did this homework with coaching from <Name of another student> but did not cut and paste any code

### Prep Exercise

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

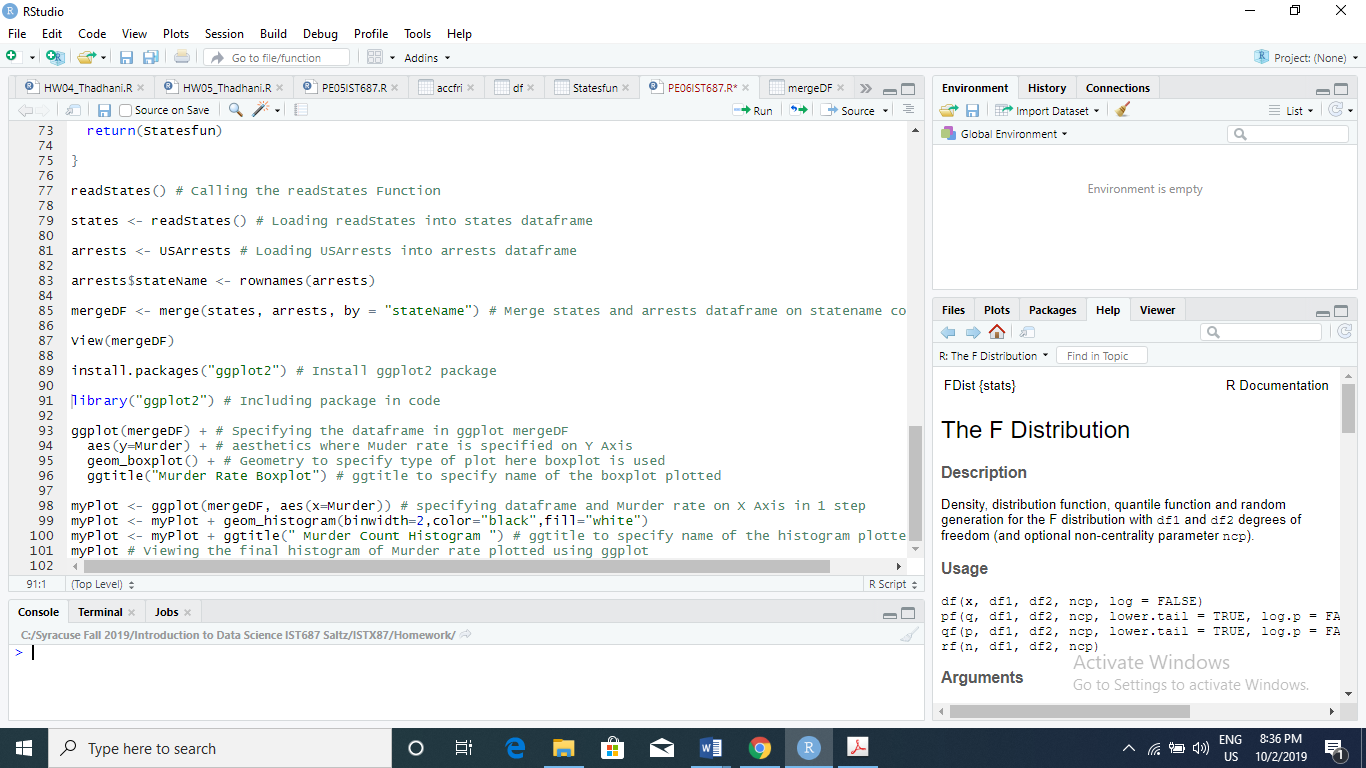
1. Code and Execute the following block of code. Be sure to understand each line. As a reminder, you created the ‘readStates’ function in a previous homework assignment, so you should be able to reuse that code. Add comments before each line to explain in detail what each line of code does.

states <- readStates() # Loading readStates into states dataframe

arrests <- USArrests # Loading USArrests into arrests dataframe

arrests$stateName <- rownames(arrests) # Rownames of arrests dataframe are assigned to statename column of arrests.

mergeDF <- merge(states, arrests, by = "stateName") # Merge states & arrests dataframes using stateName column to make it one dataframe including statenames and their corresponding incident rates



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

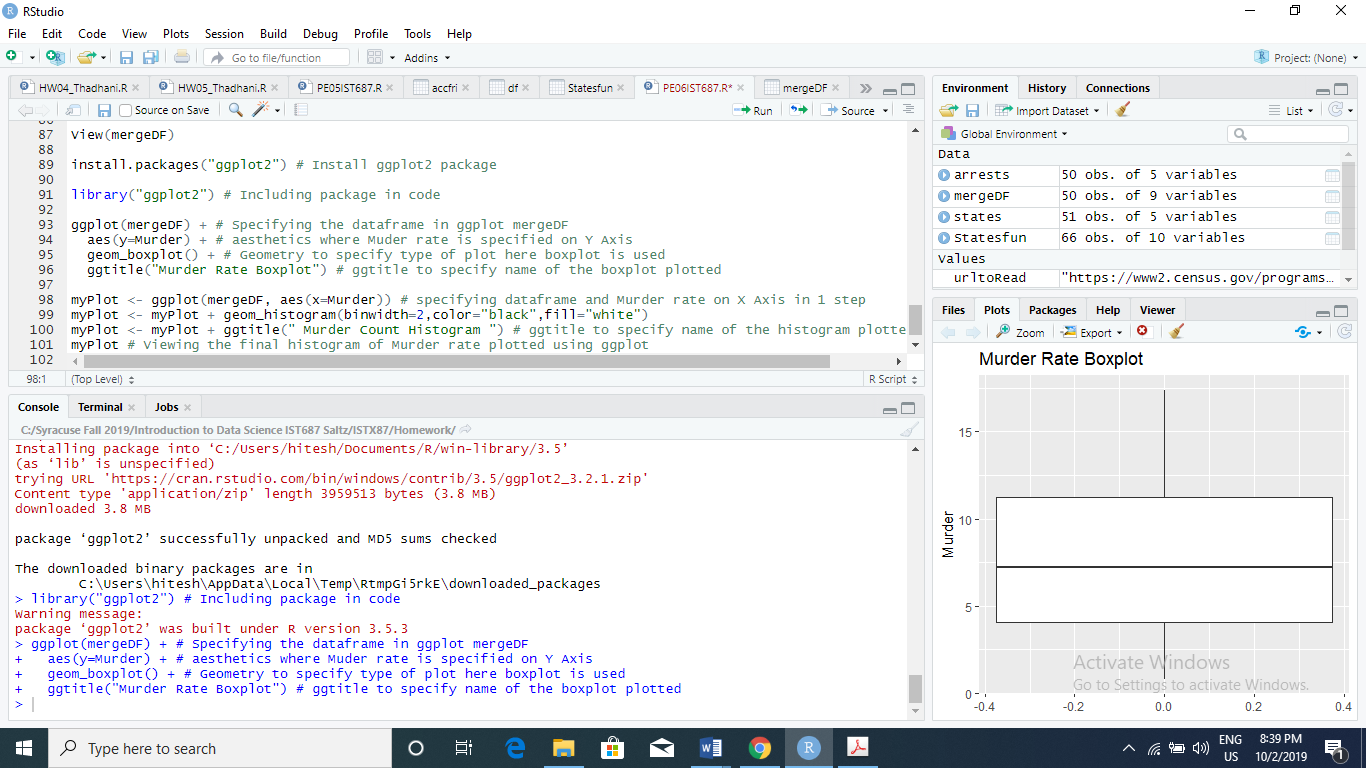
1. Install and library the ggplot2 package.
2. Code and Execute the following block of code (actually type, do not copy/paste). Add comments before each line to explain in detail what each line of code does. Add an appropriate title for the chart (using ‘ggtitle’)

ggplot(mergeDF) + # ggplot takes mergeDF dataframe as 1st input which we want to use to create visualization

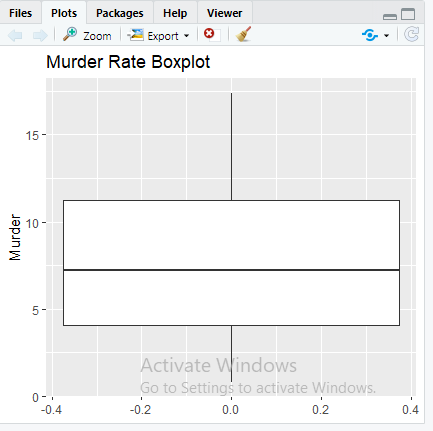
aes(y=Murder ) + # ggplot has Murder on Y-Axis as 2nd input

geom\_boxplot() + # Type of Plot is Boxplot using 3rd input geometry

ggtitle("Murder Rate Boxplot") # ggtitle gives the title of the boxplot created using ggplot



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



The Boxplot of Murder rate on Y Axis has median (Q2) roughly around 7.25 and Q1 (Quantile 1) is roughly around 4 and Q3 (Quantile 3) roughly around 11.25. The IQR (Inter Quantile Range) is 7.25. The Range is 17.5(Max value upper whisker) – 1(Min value lower whisker) = 16.5 roughly not exact to the scale.

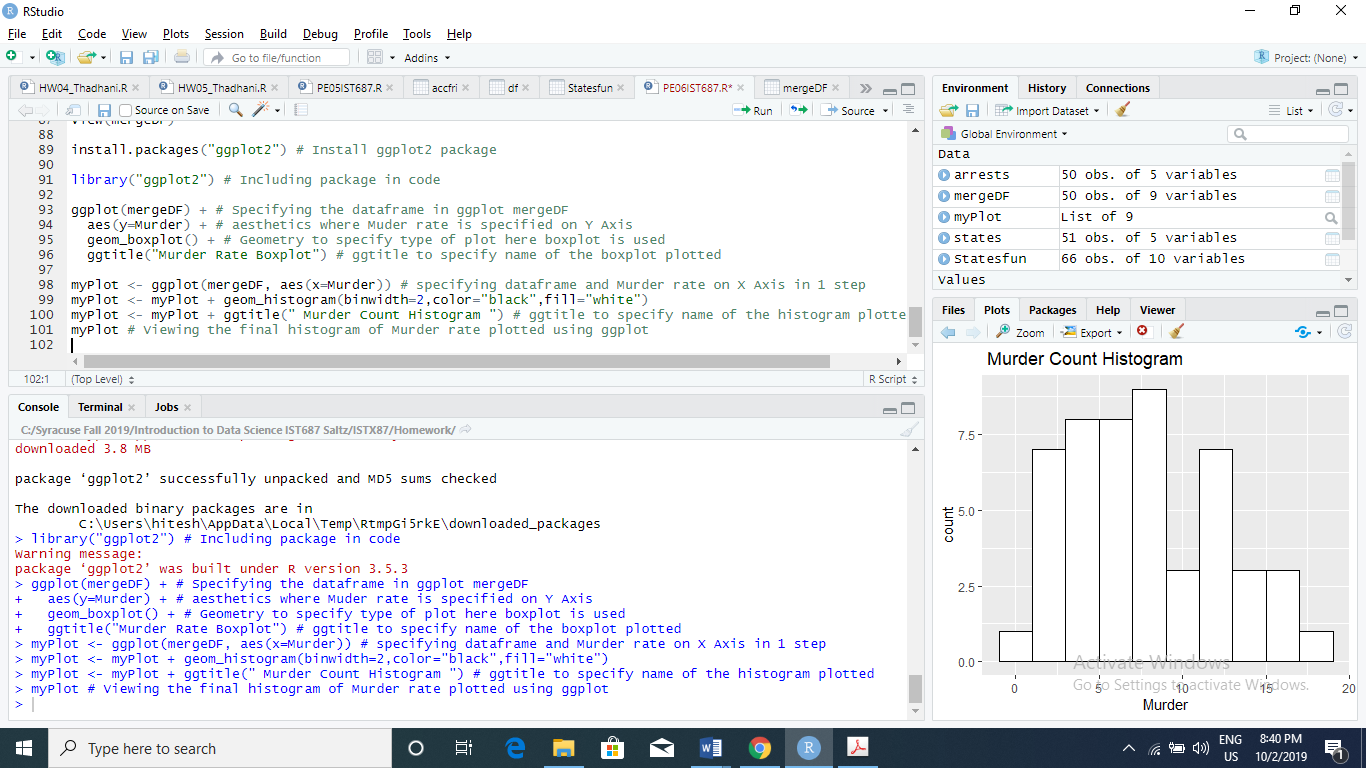
1. Code and Execute the following block of code (actually type, do not copy/paste). Add comments before each line to explain in detail what each line of code does. Add an appropriate title for the chart (using ‘ggtitle’)

myPlot <- ggplot(mergeDF, aes(x=Murder)) # Input mergeDF dataframe with X Axis taking Murder Rate(aes= aesthetics taking input X/Y Axis parameter in the plot)

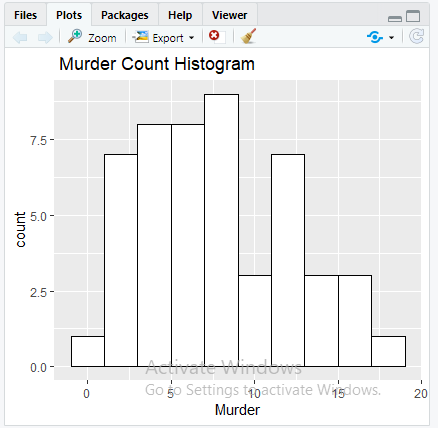
myPlot <- myPlot + geom\_histogram(binwidth=2, color="black", fill="white") # Geometry which can be used to specify type of plot here histogram is used(boxplot, histogram etc.) with binwidth means width of histogram plot 2 bins for a single bar, color is black which is the outlines of the bars, fill is white which means bars in histogram are filled with white color and outer line of black color specified.

myPlot <- myPlot + ggtitle("Murder Count Histogram") # ggtitle adds the title to the histogram we created using ggplot

myPlot # Display the histogram



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



The histogram is about the Murder rates plotted using ggplot. The histogram has only 1 variable Murder rate which is on X-Axis.

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.