**IST 687 PREP EXERCISE 06**

**Student Name: Bhavish Kumar**

**Prep Exercise No: 06**

**Date Due: 2nd October 2019**

**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(dfStates) *### Obtaing the cleaned datframe returned from*

*readStates function*

arrests <- USArrests *### Creating a new dataframe that contains USArrests data*

arrests$stateName <- rownames(arrests) *### Creating a new 'stateName'*

*column in the arrests dataframe that*

*contains the rownames of the dataframe*

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

*### Merging the 2 dataframes on stateName column to create a new dataframe that has columns from both the tables*

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

1. Install and library the ggplot2 package.

install.packages(“ggplot2”)  
library(“ggplot2”)

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’)

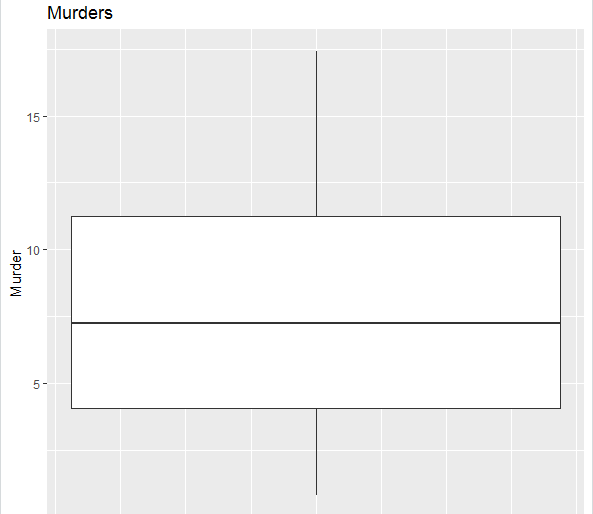
ggplot(mergeDF) + *### This is to create a ggplot with the mergeDF dataframe*

aes(y=Murder ) + *### Using the murder column of the dataframe as Y axis*

geom\_boxplot() + *### Specifying that we want to produce a boxplot*

ggtitle("Murders") *### Providing a title for the Boxplot*

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



From the above boxplot we can observe that the median (50th percentile) value for murders is around 7.25 and the 25th and 75th percentile values are around 4 and 11 respectively. Using this we can measure the interquartile range. We can also observe the smallest and largest values too.

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)) *### Creating a ggplot with*

*dataframe and using Murder column in X axis.*

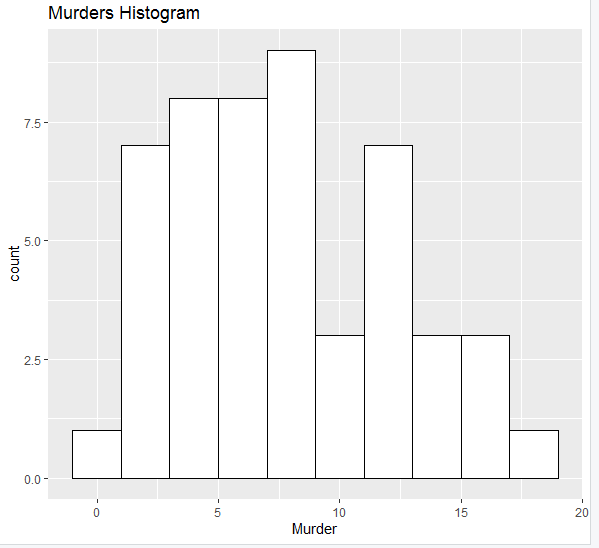
myPlot <- myPlot + geom\_histogram(binwidth=2, color="black", fill="white")

*### Producing a black bordered histogram filled with white color with binwidth 2 (i.e. shows the frequency of murder rate between 1 to 3, 3 to 5 and so on)*

myPlot <- myPlot + ggtitle("Murders Histogram") *### Adding a title to the histogram*

myPlot *### Displaying the histogram*

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



We can observe the histogram of murder rate column. It follows a normal distribution having 2 peaks (bi modal). We can observe that murder rate between 7 to 9 has maximum number of states and there is another peak between 11 to 13.