**Module Assignment**

**Module 1**

**QMB-6304 Analytical Methods for Business**

R script:

**#Debjani Sarma**

**rm(list = ls())**

**install.packages("rio")**

**install.packages("moments")**

**library(rio)**

**library(moments)**

**Preprocessing**

1. Load the file “6304 Module 1 Assignment Data.xlsx” into R. This file contains information on crime in each of the 67 counties in Florida. This will be your master data frame.

**> crime.info=import("6304 Module 1 Assignment Data.xlsx")**

1. Using the numerical portion of your U number as a random number seed, take a random sample of 30 counties using the method presented in class. This will be your primary data frame.

**set.seed(24173877)**

**> my.crime= crime.info[sample(1:nrow(crime.info),30),]**

**> colnames(my.crime)=tolower(make.names(colnames(my.crime)))**

**Analysis**

Using R calculate and report the following using your primary data frame:

1. The structure of the data frame using the str() command.

str(my.crime)

**> str(my.crime)**

**'data.frame': 30 obs. of 12 variables:**

**$ county : chr "Citrus" "Lafayette" "Lake" "Hendry" ...**

**$ population : num 149383 8690 366742 40953 203951 ...**

**$ total.crimes : num 2333 47 6341 868 3723 ...**

**$ murder : num 7 0 18 5 9 1 26 3 32 8 ...**

**$ rape : num 25 0 138 6 94 33 160 8 239 29 ...**

**$ robbery : num 38 2 130 25 43 6 264 9 304 39 ...**

**$ aggravated.assault : num 335 24 687 115 514 ...**

**$ burglary : num 334 9 1035 276 427 ...**

**$ larceny : num 1403 10 3799 365 2331 ...**

**$ vehicle.theft : num 191 2 534 76 305 57 772 55 771 151 ...**

**$ crime.rate.per.100k.popln : num 1562 541 1729 2120 1825 ...**

**$ clearance.rate.per.100.offenses: num 34.9 87.2 26.3 27.3 31.5 47.3 36.6 44.3**

**> attach(my.crime)**

1. Mean, Median, Standard Deviation, Skewness, and Kurtosis of the Population variable. Based on these descriptive measurements how closely do you think this variable conforms to a theoretical normal distribution?

**> mean(population)**

**[1] 314987.7**

**> median(population)**

**[1] 161896.5**

**> sd(population)**

**[1] 448511.8**

**> skewness(population)**

**[1] 2.272512**

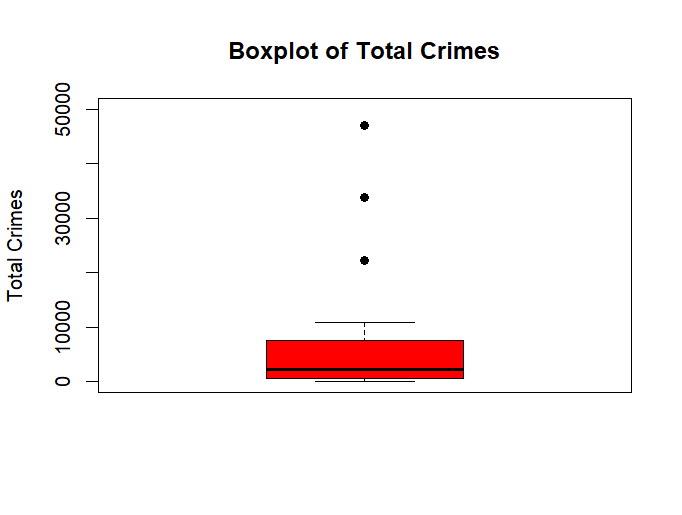
**> kurtosis(population)**

**[1] 7.868467**

The data is rightly skewed or positively skewed.

1. A boxplot of the Total Crimes variable. Based on this boxplot what can you say about the symmetry/skewness of this variable?

**> boxplot(total.crimes,col="red",pch=19,ylim=c(0,50000),main="Boxplot of Total Crimes",ylab="Total Crimes")**

****Based on the boxplot, the total crimes variable is rightly skewed or positively skewed.

1. Quartiles of the Aggravated Assault variable. Show your quartiles running from the minimum to maximum values for the variable, incrementing by .20.

quantile(aggravated.assault,probs = seq(0,1,.2))

**> quantile(aggravated.assault,probs = seq(0,1,.2))**

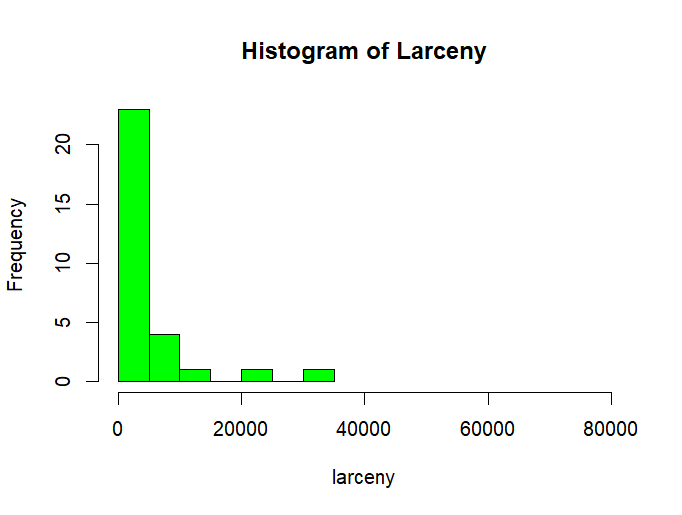
**0% 20% 40% 60% 80% 100%**

**6.0 78.6 236.4 540.8 1230.2 4578.0**

1. A simple histogram of the Larceny variable. Color your histogram green and give it an appropriate main title. Make sure the bottom axis of your histogram covers a range from 0 to 80,000. Based on this graphical tool would you say from this histogram the distribution of Larceny follows a symmetric distribution, or a skewed distribution?

hist(larceny,col="green",main="Histogram of Larceny",xlim = c(0,80000))

**> hist(larceny,col="green",main="Histogram of Larceny",xlim = c(0,80000))**

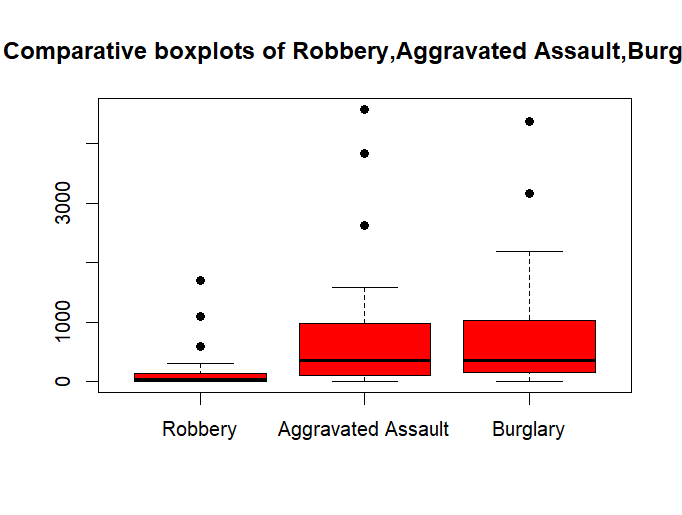


Histogram of Larceny follows rightly skewed (or positively skewed) distribution

1. Three comparative boxplots for the Robbery, Aggravated Assault, and Burglary variables. Your boxplots should be colored red and shown side by side in a single graphic with an appropriate main title and labels for the crime categories on the bottom axis. Based on these boxplots what can you say about the similarity in the number of crimes in these categories?

boxplot(robbery,aggravated.assault,burglary,col="red",main="Comparative boxplots of Robbery,Aggravated Assault,Burglary",names=c("Robbery","Aggravated Assault","Burglary"),pch=19)

**> boxplot(robbery,aggravated.assault,burglary,col="red",main="Comparative boxplots of Robbery,Aggravated Assault,Burglary",names=c("Robbery","Aggravated Assault","Burglary"),pch=19)**



Based on the boxplot, it can be said, Robbery is compared to Burglary and Aggravated Assault. Aggravated Assault and Burglary has similar IQR, median and lower whisker.There are 3 outliers of Robbery and Aggravated Assault and 2 outliers for Burglary.

1. Use R to determine and report the name of the county in primary data frame with the maximum number of Total Crimes.

max(total.crimes)

county.max.totalcrime = subset(my.crime,total.crimes=="47045")

county.max.totalcrime$county

**> county.max.totalcrime$county**

**[1] "Broward"**