RLabAssignment

title: "DS311 - R Lab Assignment" author: "Sandeep Kahlon" date: "3/12/2022" output: pdf_document: default html_document: theme: united highlight: tango df_print: paged —

R Assignment 1

- In this assignment, we are going to apply some of the build in data set in R for descriptive statistics analysis.
- To earn full grade in this assignment, students need to complete the coding tasks for each question to get the result.
- After finished all the questions, knit the document into HTML format for submission.

Question 1

Using **mtcars** data set in R, please answer the following questions.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Loading the data
data(mtcars)

# Head of the data set
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a. Report the number of variables and observations in the data set.

```
# Enter your code here!

dim(mtcars)
```

```
## [1] 32 11
```

```
# Answer:
print("There are total of 11 variables and 32 observations in this data set.")
```

```
## [1] "There are total of 11 variables and 32 observations in this data set."
```

b. Print the summary statistics of the data set and report how many discrete and continuous variables are in the data set.

```
# Enter your code here!
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

```
# Answer:
print("There are 5 discrete variables and 6 continuous variables in this data set.")
```

```
## [1] "There are 5 discrete variables and 6 continuous variables in this data set."
```

c. Calculate the mean, variance, and standard deviation for the variable **mpg** and assign them into variable names m, v, and s. Report the results in the print statement.

```r
# Enter your code here!
m <- mean(mtcars$mpg)
m
```

```
## [1] 20.09062
```

```r
v <- var(mtcars$mpg)
v
```

```
## [1] 36.3241
```

```r
s <- sd(mtcars$mpg)
s
```

```
## [1] 6.026948
```

```r
print(paste("The average of Mile Per Gallon from this data set is 20.01 with variance 36.32 and standard
```

```
## [1] "The average of Mile Per Gallon from this data set is 20.01 with variance 36.32 and standard devi
```

d. Create two tables to summarize 1) average mpg for each cylinder class and 2) the standard deviation of mpg for each gear class.

```r
# Enter your code here!
#Table 1 -- Cylinder Class
cyl <- mtcars %>%
  group_by(cyl) %>%
  summarize(AvgMPG = mean(mpg))
cyl
```

```
## # A tibble: 3 x 2
##     cyl AvgMPG
##   <dbl>  <dbl>
## 1     4   26.7
## 2     6   19.7
## 3     8   15.1
```

```r
#Table 2 -- Gear Class
gear <- mtcars %>%
  group_by(gear) %>%
  summarize(MPGstdev = sd(mpg))
gear
```

```
## # A tibble: 3 x 2
##    gear MPGstdev
##   <dbl>    <dbl>
## 1     3     3.37
## 2     4     5.28
## 3     5     6.66
```

e. Create a crosstab that shows the number of observations belong to each cylinder and gear class combinations. The table should show how many observations given the car has 4 cylinders with 3 gears, 4 cylinders with 4 gears, etc. Report which combination is recorded in this data set and how many observations for this type of car.

```
# Enter your code here!
combo <- mtcars %>%
  group_by(cyl, gear) %>%
  summarize(Instances = length(mpg))
```

```
## `summarise()` has grouped output by 'cyl'. You can override using the `.groups`
## argument.
```

```
combo
```

```
## # A tibble: 8 x 3
## # Groups:   cyl [3]
##     cyl  gear Instances
##   <dbl> <dbl>     <int>
## 1     4     3         1
## 2     4     4         8
## 3     4     5         2
## 4     6     3         2
## 5     6     4         4
## 6     6     5         1
## 7     8     3        12
## 8     8     5         2
```

```
print("The most common car type in this data set is a car with 8 cylinders and 3 gears. There are total
```

```
## [1] "The most common car type in this data set is a car with 8 cylinders and 3 gears. There are total
```

Question 2

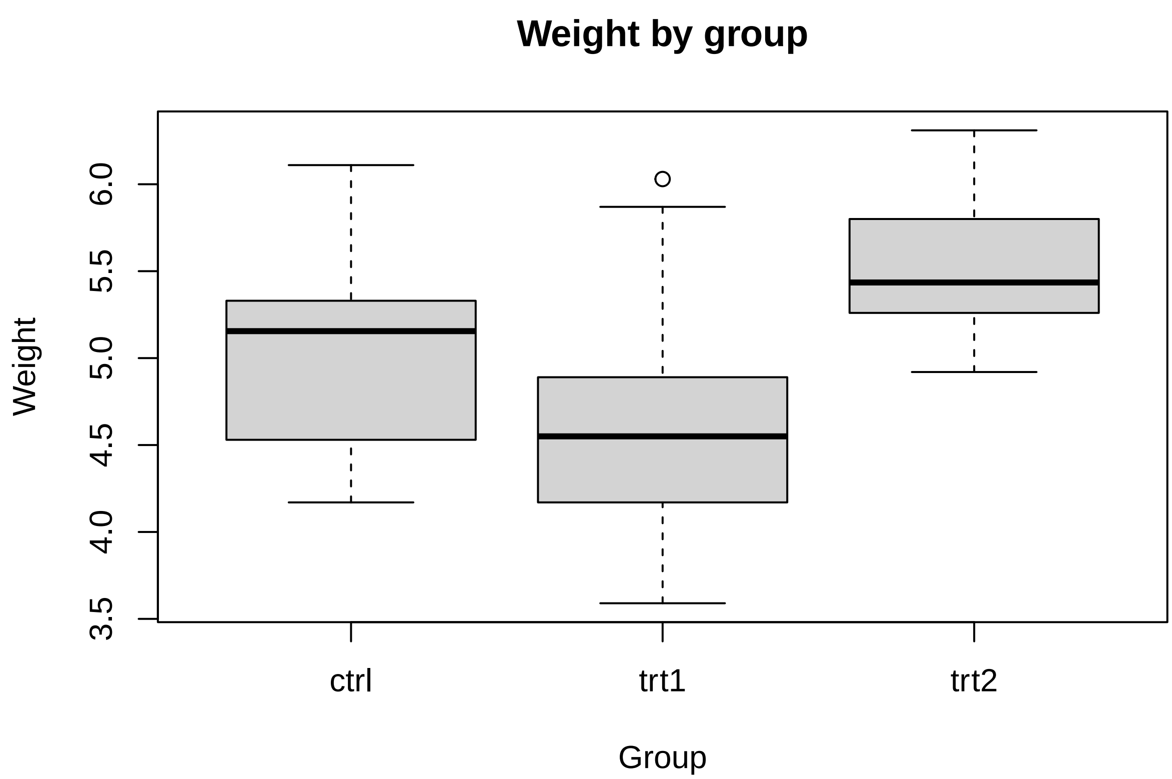
Use different visualization tools to summarize the data sets in this question.

a. Using the **PlantGrowth** data set, visualize and compare the weight of the plant in the three separated group. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your findings in this graph.

```
library(ggplot2)
# Load the data set
data("PlantGrowth")

# Head of the data set
head(PlantGrowth)
```

```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```r
# Enter your code here!
g_w <- plot(PlantGrowth$group, PlantGrowth$weight, main = "Weight by group",
     xlab = "Group",
     ylab = "Weight")
```

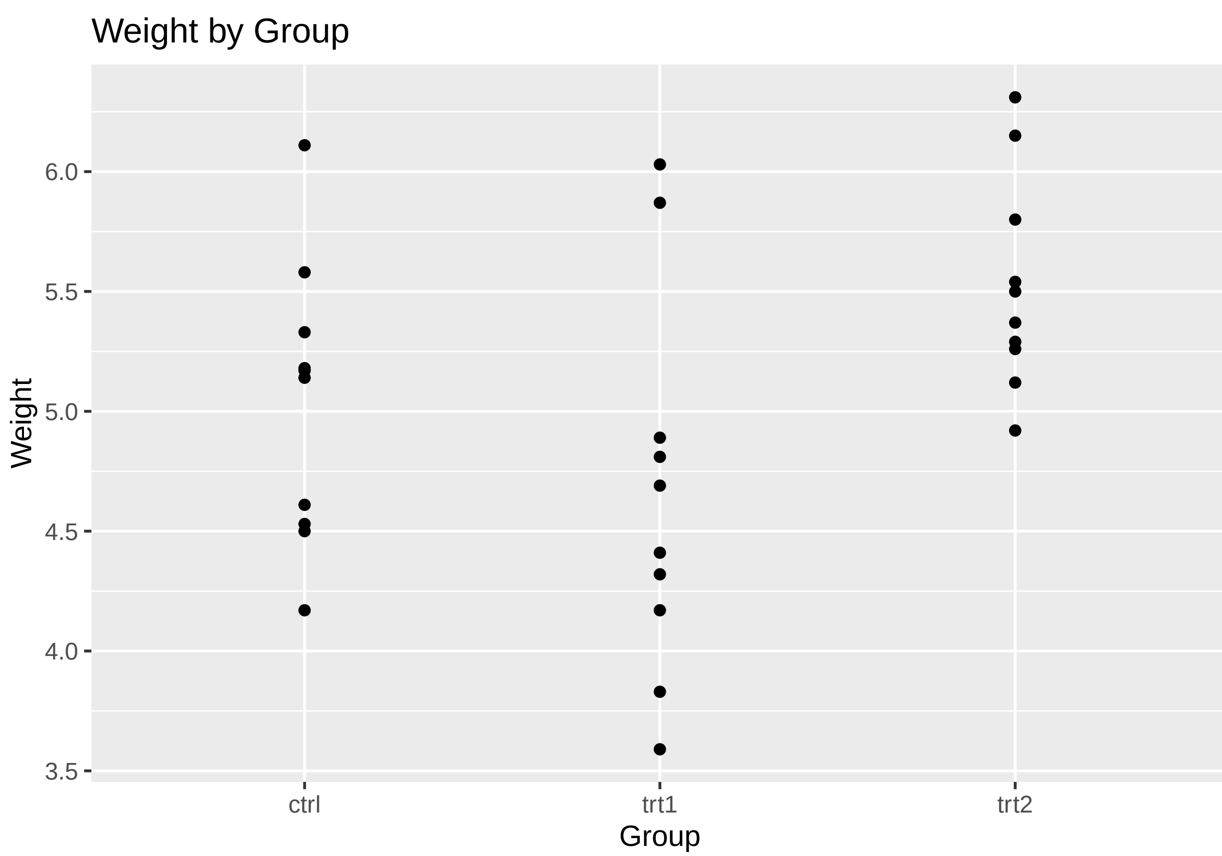



```r
g_w
```

```
## $stats
##       [,1] [,2]  [,3]
## [1,] 4.170 3.59 4.920
## [2,] 4.530 4.17 5.260
## [3,] 5.155 4.55 5.435
## [4,] 5.330 4.89 5.800
## [5,] 6.110 5.87 6.310
##
## $n
## [1] 10 10 10
```

```
##
## $conf
##          [,1]     [,2]     [,3]
## [1,] 4.755288 4.190259 5.165194
## [2,] 5.554712 4.909741 5.704806
##
## $out
## [1] 6.03
##
## $group
## [1] 2
##
## $names
## [1] "ctrl" "trt1" "trt2"
```

```r
gg <- ggplot(PlantGrowth, aes(x=group, y=weight)) + geom_point()
gg <- gg + labs(title = "Weight by Group", x="Group", y="Weight")
gg
```



Result: Group trt2 contains the highest average weight and the most coincise interquartaile range between the three respective groups. The maximum weight observation in the dataset is found in group trt2.

Group trt 1 contains the lowest average weight and consists of a wide upper quartile range with a respective outlier. The minimum weight observation in the dataset is found in group trt1.

Group ctrl contains the widest interquartile range with an average weight of ~5.5. This suggest observations are the most volatile with respect to weight in group ctrl.

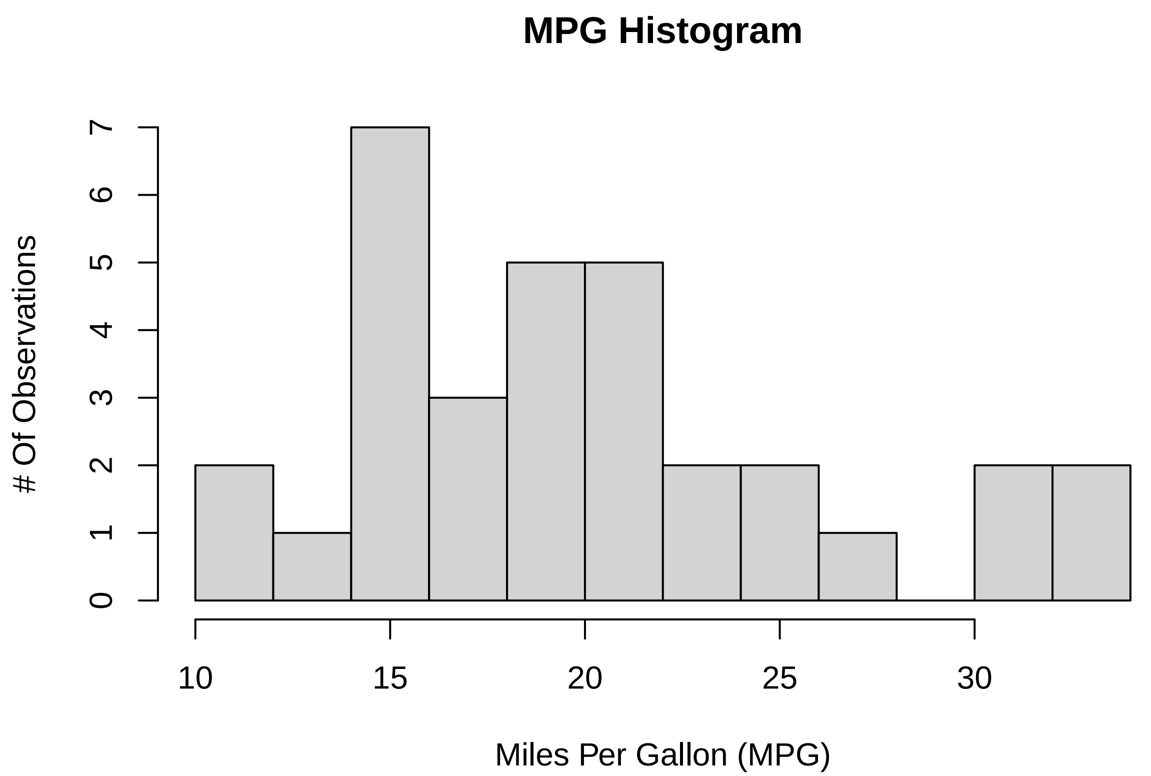
=> Enter your results here!

b. Using the **mtcars** data set, plot the histogram for the column **mpg** with 10 breaks. Give labels to the title, x-axis, and y-axis on the graph. Report the most observed mpg class from the data set.

```
attach(mtcars)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     cyl, gear

## The following object is masked from package:ggplot2:
##
##     mpg
```

```
hist(mpg,
     main = "MPG Histogram",
     xlab = "Miles Per Gallon (MPG)",
     ylab = "# Of Observations",
     breaks=10)
```




```
print("Most of the cars in this data set are in the class of 15 mile per gallon.")
```

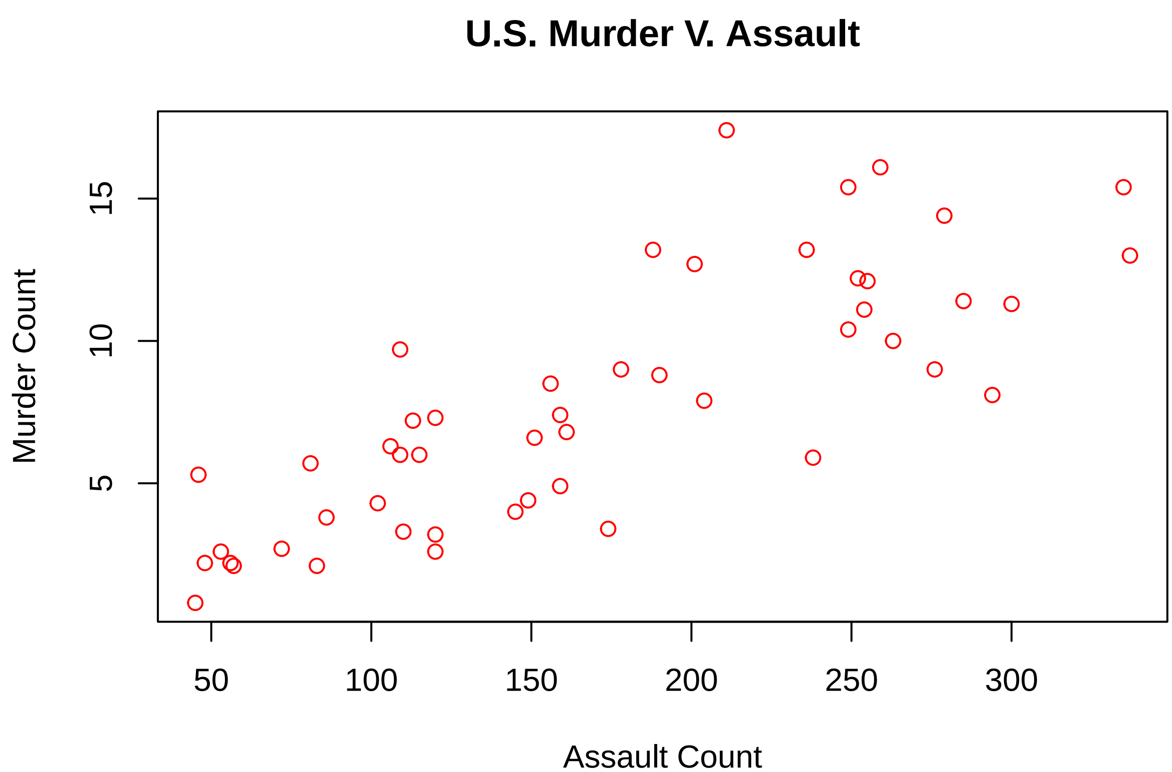
```
## [1] "Most of the cars in this data set are in the class of 15 mile per gallon."
```

c. Using the **USArrests** data set, create a pairs plot to display the correlations between the variables in the data set. Plot the scatter plot graph of **Murder** and **Assault**. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your results from both plots.

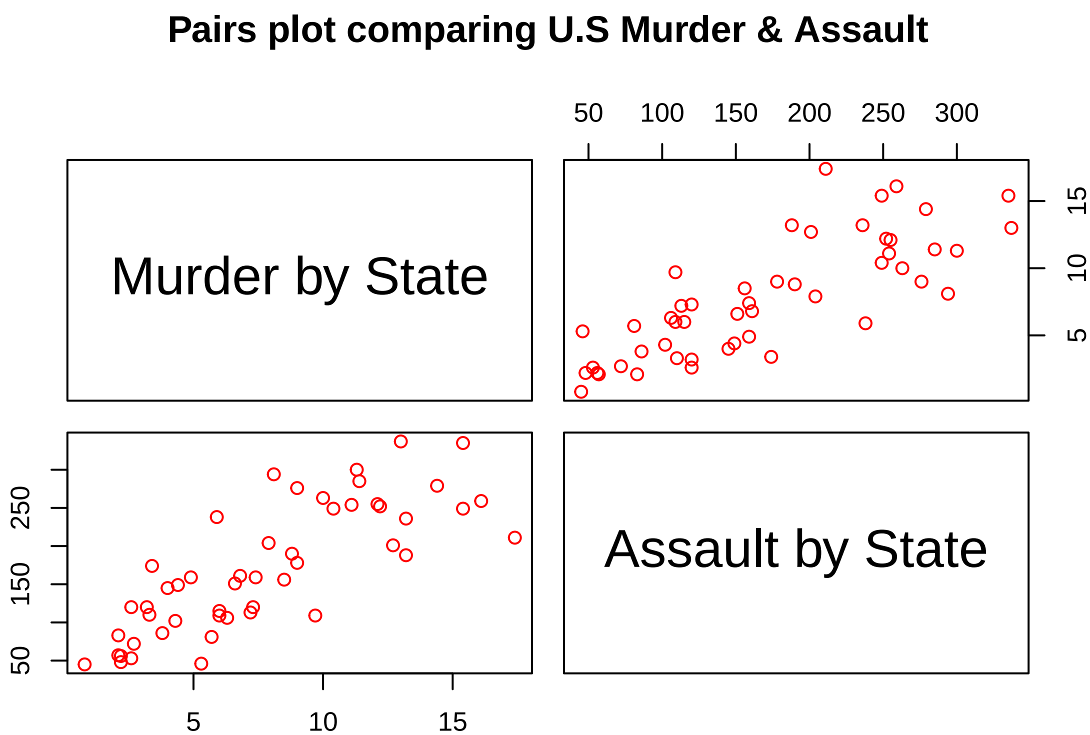
```
# Load the data set
data("USArrests")

# Head of the data set
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
# Enter your code here!
#Scatter Plot Murder V. Assault
plot(USArrests$Assault, USArrests$Murder,
     col = "Red",
     main = "U.S. Murder V. Assault",
     xlab = "Assault Count",
     ylab = "Murder Count")
```



```
#Pairs Plot Murder V. Assault
pairs(USArrests[,1:2],
      col="Red",
      labels = c("Murder by State", "Assault by State"),
      main = "Pairs plot comparing U.S Murder & Assault")
```



Result:

=> Enter your result here!

Assessing the Pairs plot. The majority of states contain low rates of murder and assault. However, there remains a significant number of states with high rates of both.A positive linear relationship is present between assault and murder in the United States.

Question 3

Download the housing data set from www.jaredlander.com and find out what explains the housing prices in New York City.

```
## Warning in download.file(url = "https://www.jaredlander.com/data/housing.csv", :
## URL https://www.jaredlander.com/data/housing.csv: cannot open destfile 'data/
## housing.csv', reason 'No such file or directory'
```

```
## Warning in download.file(url = "https://www.jaredlander.com/data/housing.csv", :
## download had nonzero exit status
```

a. Create your own descriptive statistics and aggregation tables to summarize the data set and find any meaningful results between different variables in the data set.

```
library(dplyr)
# Head of the cleaned data set
head(housingData)
```

```
##   Neighborhood Market.Value.per.SqFt      Boro Year.Built
## 1    FINANCIAL                200.00 Manhattan       1920
## 2    FINANCIAL                242.76 Manhattan       1985
## 4    FINANCIAL                271.23 Manhattan       1930
## 5      TRIBECA                247.48 Manhattan       1985
## 6      TRIBECA                191.37 Manhattan       1986
## 7      TRIBECA                211.53 Manhattan       1985
```

```
# Enter your code here!
#Avg Market Value per Sqft by neighborhood
housingData %>%
   group_by(Neighborhood) %>%
   summarize(Avg_MktVal_Sqft = round(mean(Market.Value.per.SqFt), digits=2),
             Stdev_MktVal_Sqft = round(sd(Market.Value.per.SqFt), digits=2),
             Var_MktVal_Sqft = round(var(Market.Value.per.SqFt), digits=2))
```

```
## # A tibble: 148 x 4
##    Neighborhood         Avg_MktVal_Sqft Stdev_MktVal_Sqft Var_MktVal_Sqft
##    <chr>                          <dbl>             <dbl>           <dbl>
##  1 ALPHABET CITY                  148.              37.8          1433.
##  2 ARROCHAR-SHORE ACRES            57.8             NA              NA
##  3 ASTORIA                         91.5             21.8           477.
##  4 BATH BEACH                      70.3             21.7           473.
##  5 BAY RIDGE                       68.0             16.6           275.
##  6 BAYSIDE                         71.4             22.3           498.
##  7 BEDFORD PARK/NORWOOD            38.2              1.34            1.79
##  8 BEDFORD STUYVESANT              83.2             13.0           169.
##  9 BELMONT                         56.4             NA              NA
## 10 BENSONHURST                     71.7             22.8           518.
## # ... with 138 more rows
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
#Avg + standard deviation of house year built by neighborhood
hd1 <- housingData %>% drop_na(Year.Built)

hd1 %>%
  group_by(Neighborhood) %>%
  summarize(Avg_House_Age= round(mean(Year.Built)),
            Stdev_House_age = round(sd(Year.Built), digits=2),
            Var_House_age = round(var(Year.Built), digits=2))
```

```
## # A tibble: 148 x 4
##    Neighborhood         Avg_House_Age Stdev_House_age Var_House_age
##    <chr>                        <dbl>           <dbl>         <dbl>
##  1 ALPHABET CITY                 1968            42.0         1760.
##  2 ARROCHAR-SHORE ACRES          1987            NA             NA
##  3 ASTORIA                       1990            29.3          857.
##  4 BATH BEACH                    1988            33.3         1109.
##  5 BAY RIDGE                     1995            10.4          108.
##  6 BAYSIDE                       1979            18.4          338.
##  7 BEDFORD PARK/NORWOOD          1980            17.7          312.
##  8 BEDFORD STUYVESANT            1998            24.1          580.
##  9 BELMONT                       2007            NA             NA
## 10 BENSONHURST                   1982            36.2         1311.
## # ... with 138 more rows
```

```r
#Prominent neighborhood and boro by listings
Hd2 <- housingData %>%
  group_by(Boro, Neighborhood) %>%
  summarize(Listings = length(Year.Built))
```

```
## `summarise()` has grouped output by 'Boro'. You can override using the
## `.groups` argument.
```

```r
Hd2 <- Hd2[order(Hd2$Listings,decreasing=TRUE),]
Hd2
```

```
## # A tibble: 149 x 3
## # Groups:   Boro [5]
##    Boro      Neighborhood              Listings
##    <chr>     <chr>                        <int>
##  1 Queens    FLUSHING-NORTH                 133
##  2 Manhattan UPPER EAST SIDE (59-79)        123
##  3 Manhattan HARLEM-CENTRAL                  94
##  4 Manhattan CHELSEA                         88
##  5 Manhattan UPPER WEST SIDE (59-79)         87
##  6 Manhattan UPPER EAST SIDE (79-96)         78
##  7 Manhattan TRIBECA                         74
```

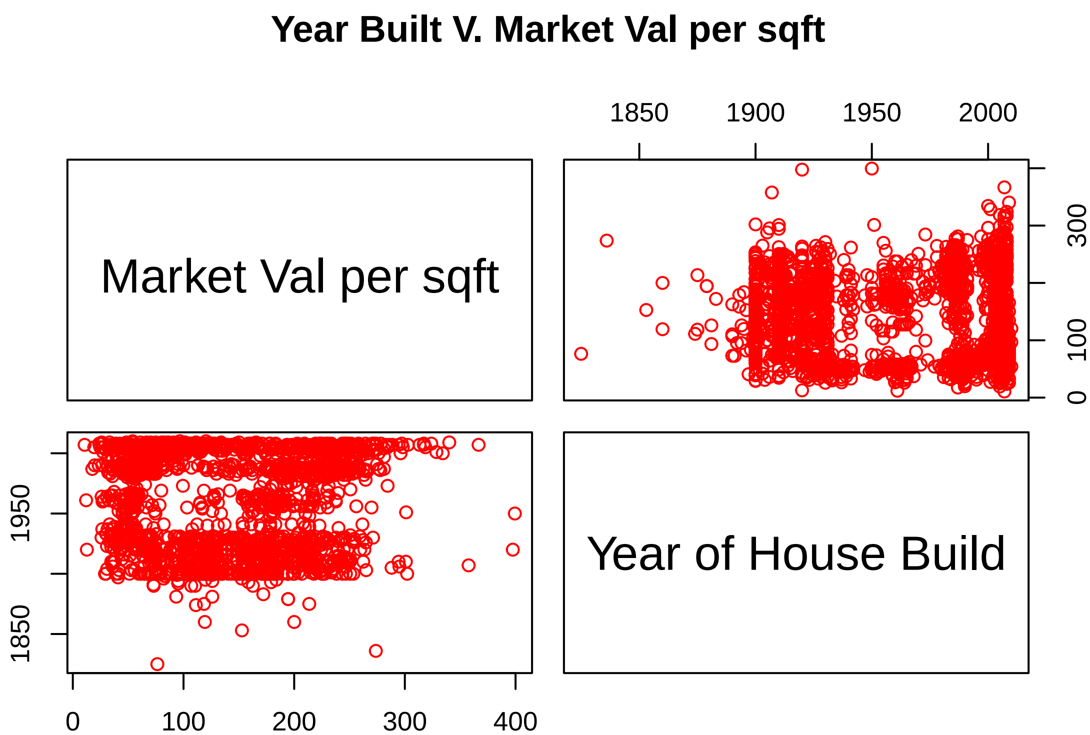
```
##  8 Manhattan UPPER WEST SIDE (79-96)         66
##  9 Brooklyn  WILLIAMSBURG-CENTRAL            60
## 10 Manhattan GREENWICH VILLAGE-CENTRAL       60
## # ... with 139 more rows
```

```r
#Top Neighborhood in each Boro by listings
Hd2 %>% slice(1)
```

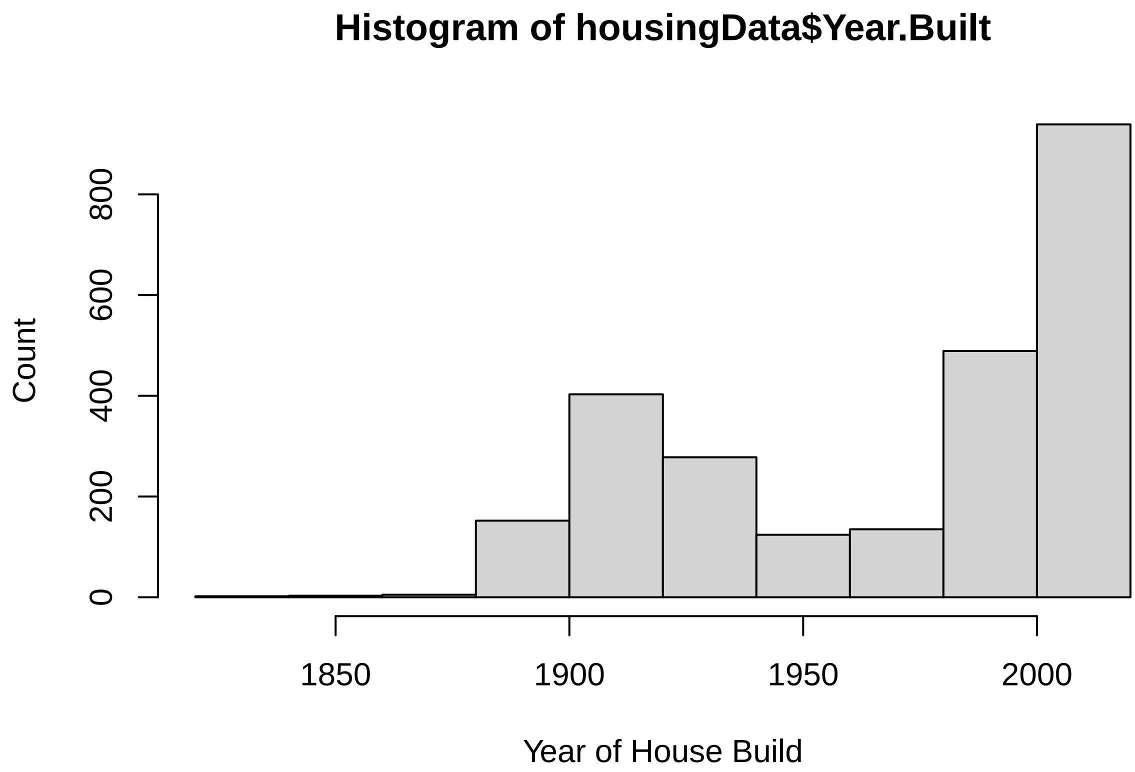
```
## # A tibble: 5 x 3
## # Groups:   Boro [5]
##   Boro          Neighborhood            Listings
##   <chr>         <chr>                      <int>
## 1 Bronx         RIVERDALE                     17
## 2 Brooklyn      WILLIAMSBURG-CENTRAL          60
## 3 Manhattan     UPPER EAST SIDE (59-79)      123
## 4 Queens        FLUSHING-NORTH               133
## 5 Staten Island NEW SPRINGVILLE                9
```

b. Create multiple plots to demonstrates the correlations between different variables. Remember to label all axes and give title to each graph.

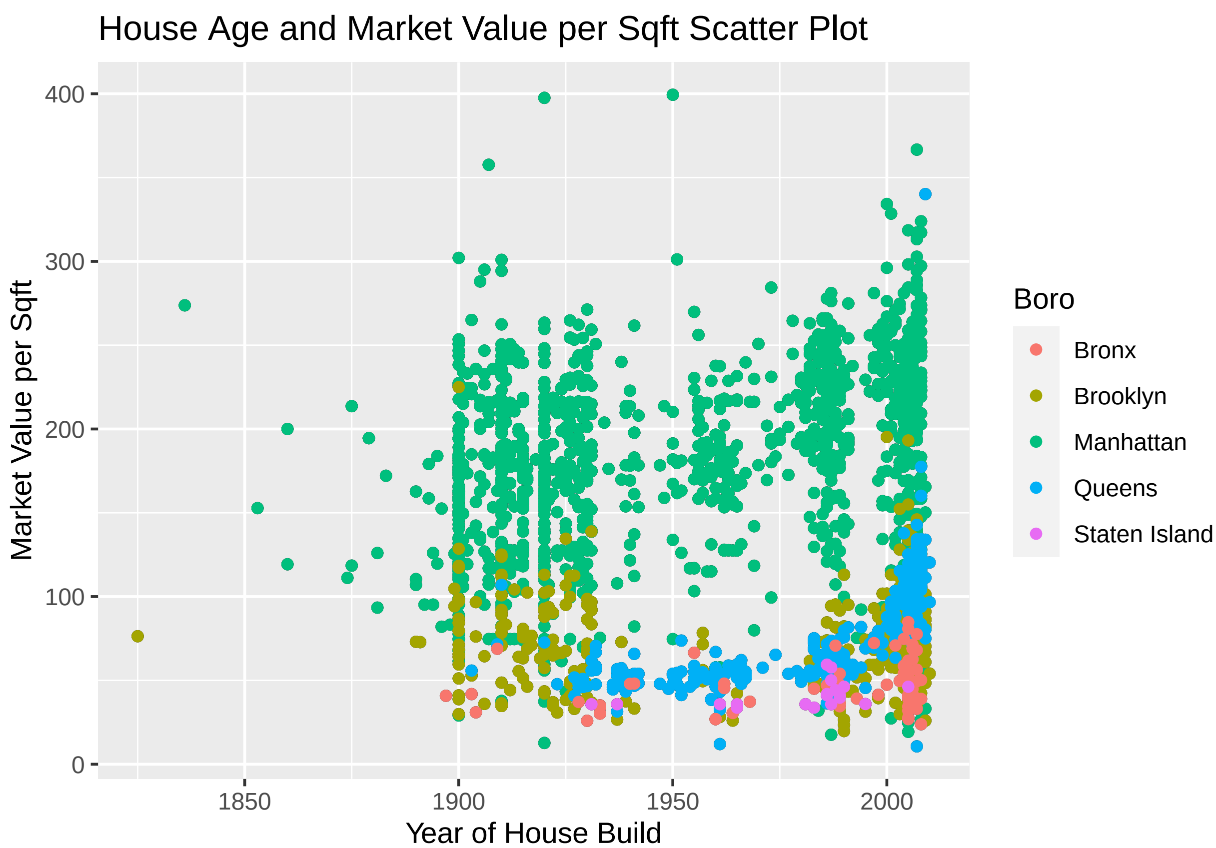
```r
# Enter your code here!
#Pair Plot to determine relationship
pairs(housingData[, c(2,4)],
      col = "Red",
      labels = c("Market Val per sqft", "Year of House Build"),
      main="Year Built V. Market Val per sqft")  #Stationary relationship
```



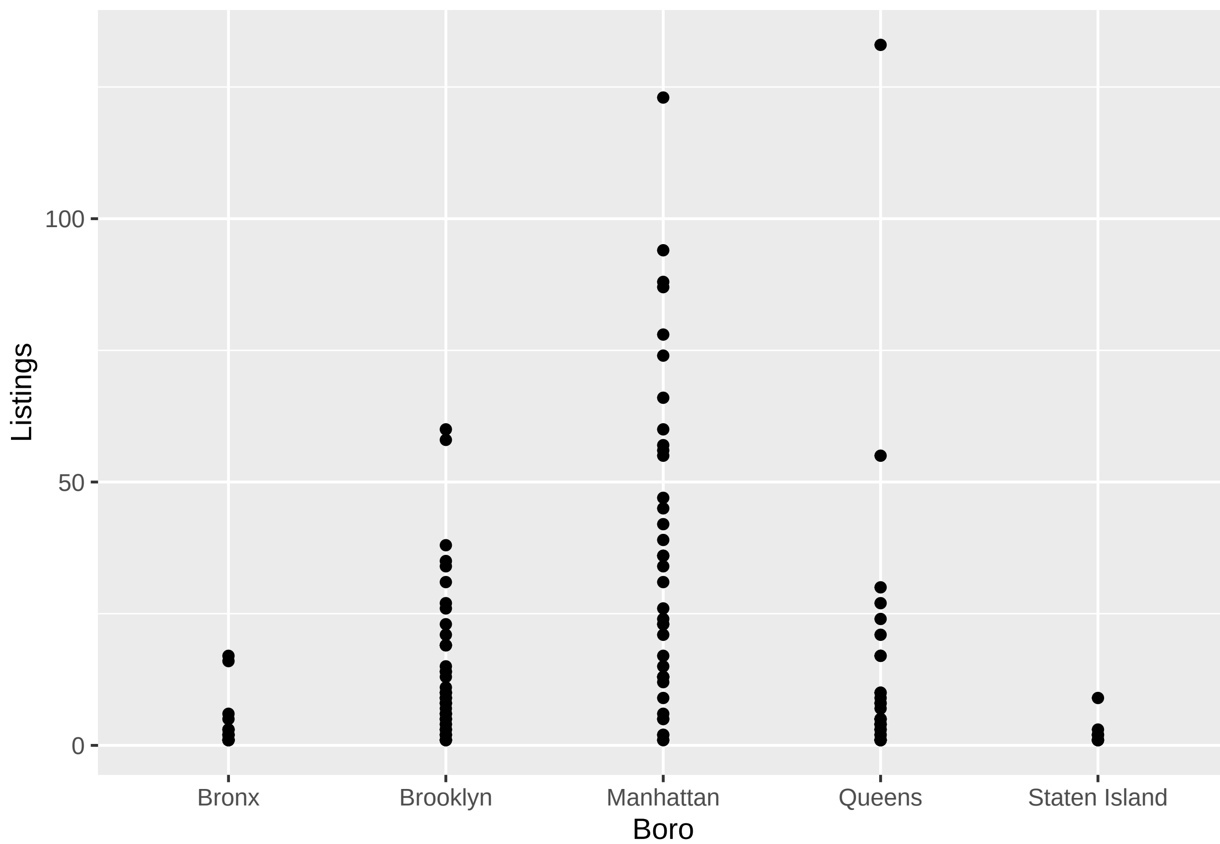

```r
#Distribution of house ages
hist(x= housingData$Year.Built,
     xlab = "Year of House Build",
     ylab = "Count")
```



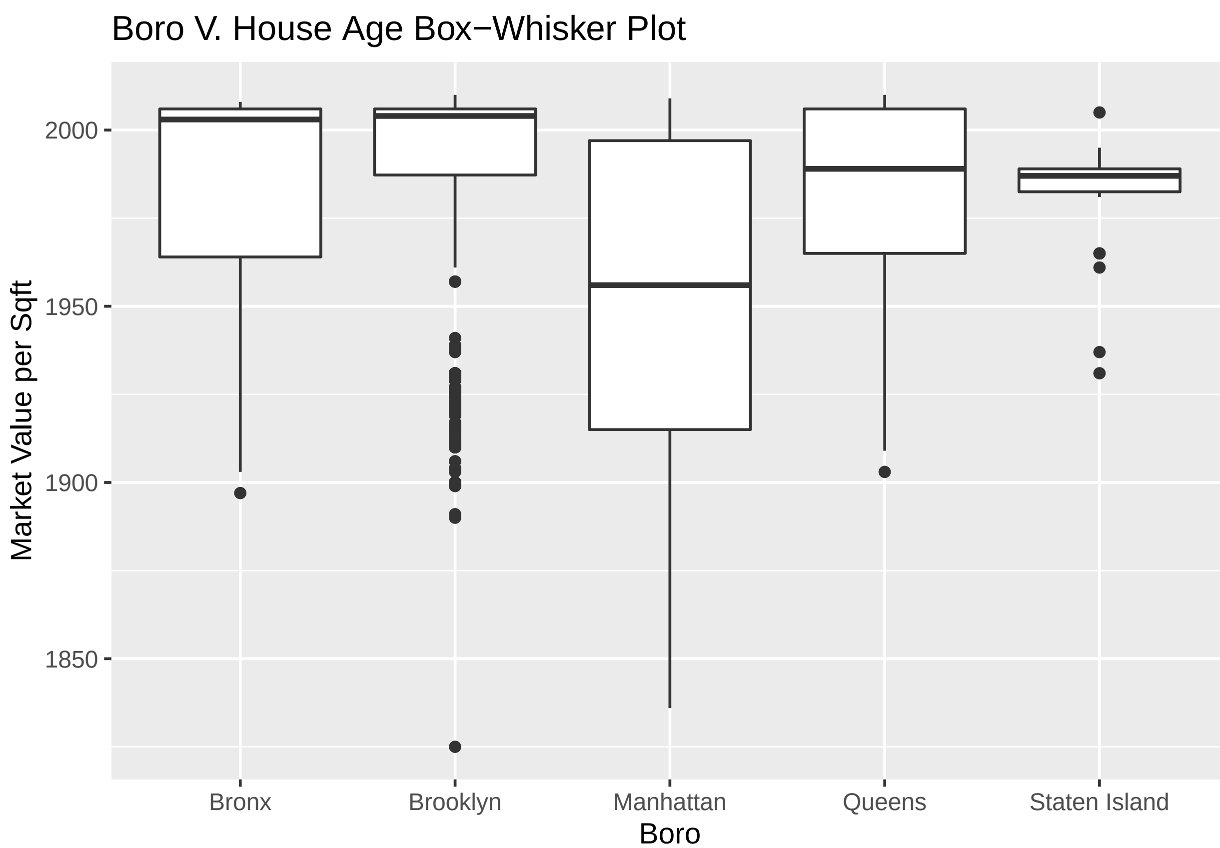

```
#Scatter plot comparing age of home and  market value per square feet for the listed Boros
g <- ggplot(housingData, aes(x=housingData$Year.Built, y=housingData$Market.Value.per.SqFt)) + geom_poi
g <- g + geom_point(aes(color=Boro))
g <- g + labs(title="House Age and Market Value per Sqft Scatter Plot",
              x = "Year of House Build",
              y = "Market Value per Sqft")
g
```



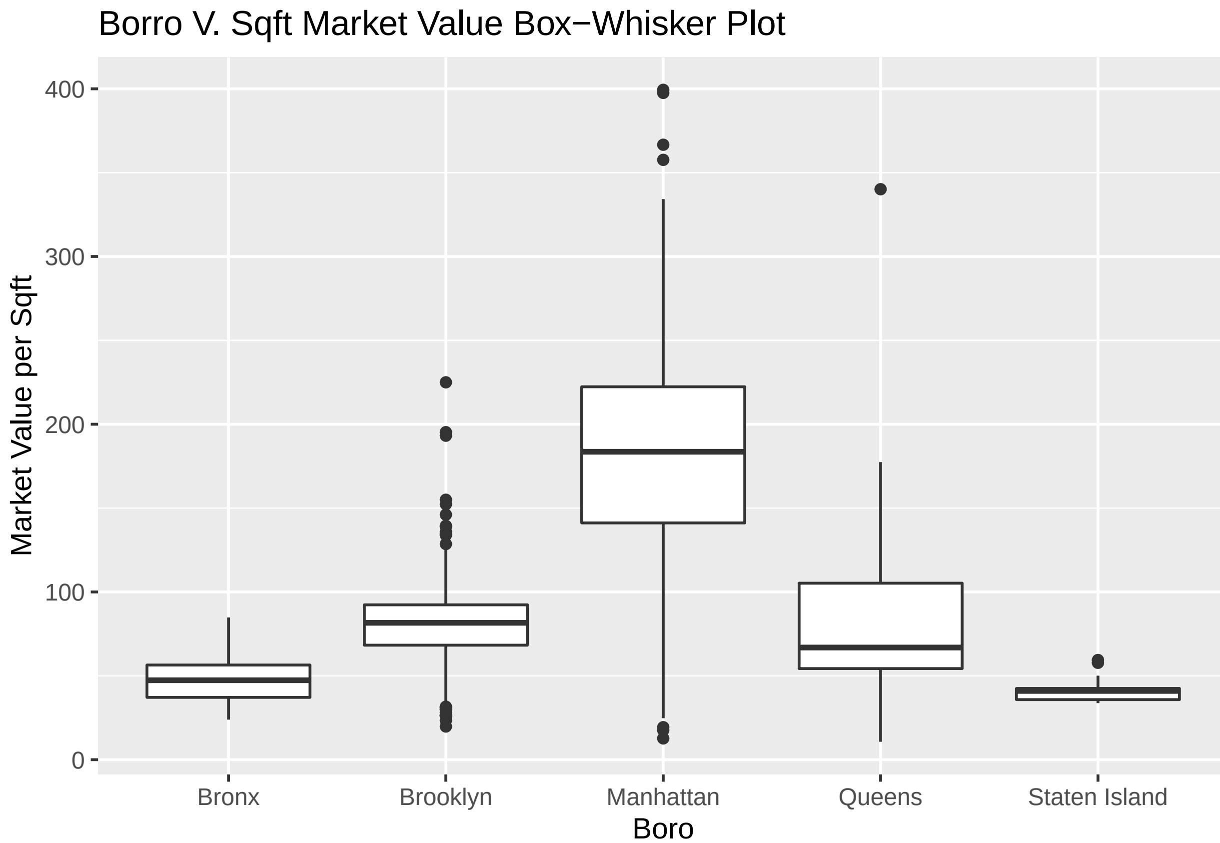

```
#Listings per Boro
ggplot(Hd2, aes(y=Listings, x=Boro)) + geom_point()
```




```r
#Box + Whisker plot
Bw <- ggplot(housingData, aes(y=Year.Built, x=Boro)) + geom_boxplot()
Bw <- Bw + labs(title="Boro V. House Age Box-Whisker Plot",
                x="Boro",
                y="Market Value per Sqft")
Bw
```




```
Bw1 <- ggplot(housingData, aes(y=Market.Value.per.SqFt, x=Boro)) + geom_boxplot()
Bw1 <- Bw1 + labs(title = "Borro V. Sqft Market Value Box-Whisker Plot",
                  x = "Boro",
                  y="Market Value per Sqft")
Bw1
```



c. Write a summary about your findings from this exercise.

Enter your answer here!

In this exercise I utilized R programming language to perform data manipulation, statistics, and visualization. I leveraged a variety of statistical and visual packages such as dplyr and ggplot to create insightful interpretation of data and detailed graphics. The exercise reinforced, strengthened, and added to my prior R programming knowledge.