DS311 - R Lab Assignment

William Lin

2023-10-27

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

```r
# 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.

```r
# Enter your code here!
row_count <- nrow(mtcars)
col_count <- ncol(mtcars)
dim(mtcars)
```

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

```r
# 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.

```r
# 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  
```

```r
# 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$mpg)
v <- var(mtcars$mpg)
s <- sd(mtcars$mpg)

print(paste("The average of Mile Per Gallon from this data set is", m, "with variance", v, "and standar
```

```
## [1] "The average of Mile Per Gallon from this data set is 20.090625 with variance 36.3241028225806 a
```

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!
m_cyl <- aggregate(mtcars$mpg, list(mtcars$cyl), FUN=mean)
```

```
s_gear <- aggregate(mtcars$mpg, list(mtcars$gear), FUN=mean)

m_cyl; s_gear
```

```
##   Group.1        x
## 1       4 26.66364
## 2       6 19.74286
## 3       8 15.10000
```

```
##   Group.1        x
## 1       3 16.10667
## 2       4 24.53333
## 3       5 21.38000
```

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!
mytable <- xtabs(~cyl+gear, data = mtcars)
ftable(mytable)
```

```
##     gear  3  4  5
## cyl
## 4         1  8  2
## 6         2  4  1
## 8        12  0  2
```

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

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

Question 2

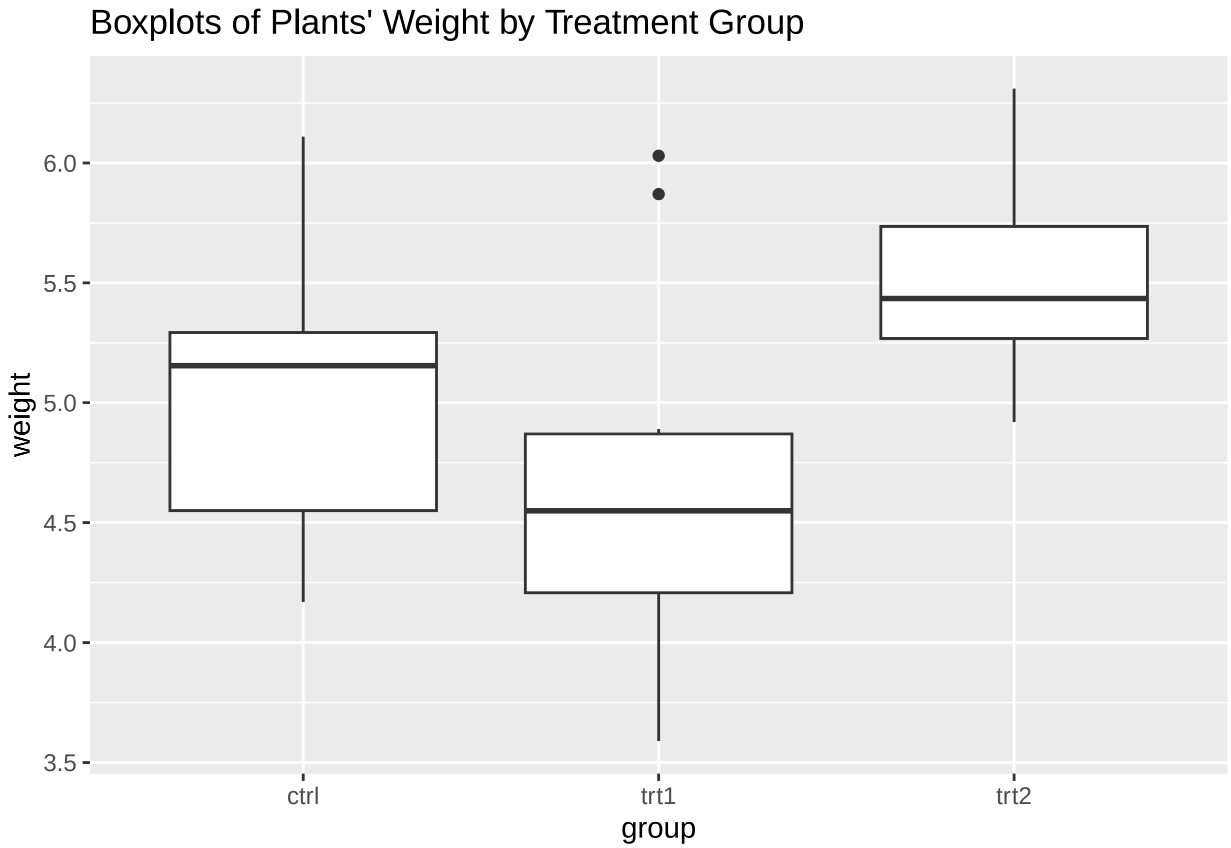
Use different visualization tools to summarize the data sets in this question.

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.

```
# 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
```

```
# Enter your code here!
library(ggplot2)
ggplot(PlantGrowth, aes(y=weight, x=group)) + geom_boxplot() + ggtitle("Boxplots of Plants' Weight by T
```

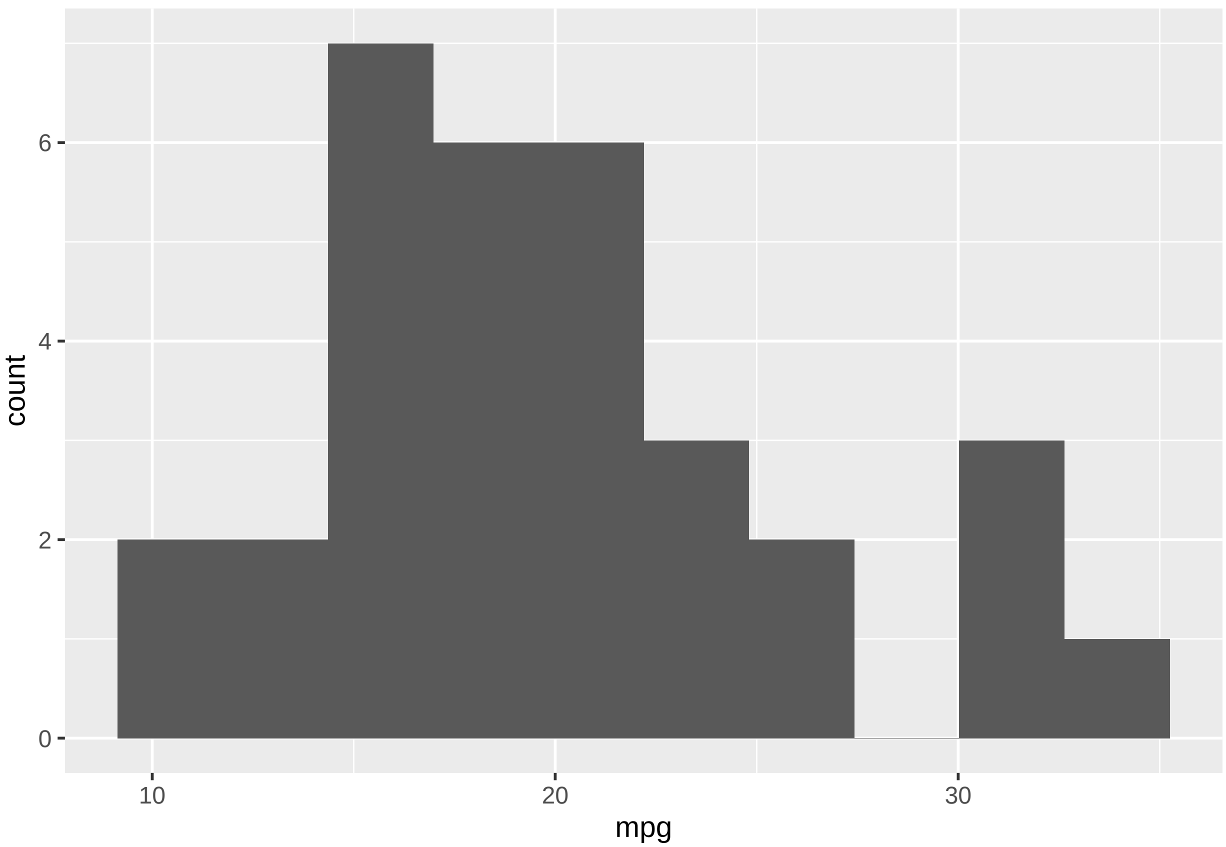


Result:

The median weight of plants in the control group is about 5.1 grams, and the interquartile range of weights in the control group are between roughly 4.5 and 5.25 grams. The median in the treatment 1 group is about 4.5 grams, and the median in the treatment 2 group is about 5.4 grams. Treatment 2 appears to be more effective in growing heavier plants than the control, while treatment 1 is less effective than the control. Furthermore, the treatment 2 group has the least variation in weights.

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.

```
ggplot(data=mtcars) + geom_histogram(aes(x=mpg), bins = 10)
```



```
print("Most of the cars in this data set are in the class of 15.0-17.5 miles per gallon.")
```

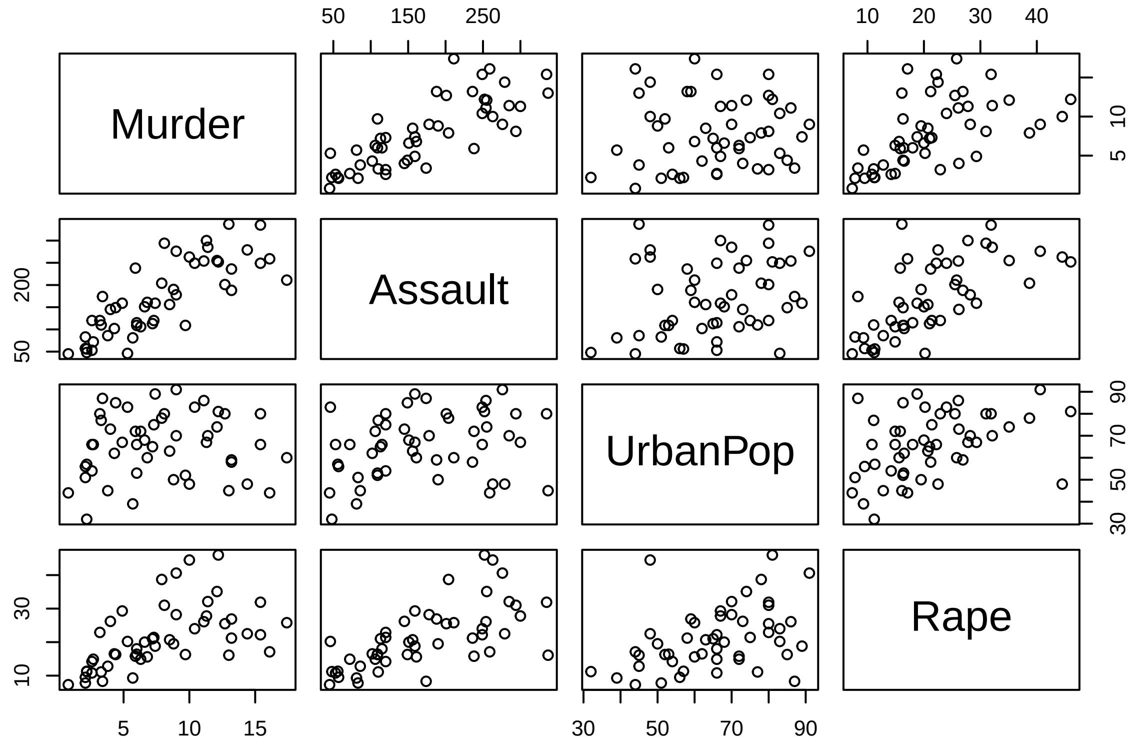
```
## [1] "Most of the cars in this data set are in the class of 15.0-17.5 miles per gallon."
```

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 with **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
```

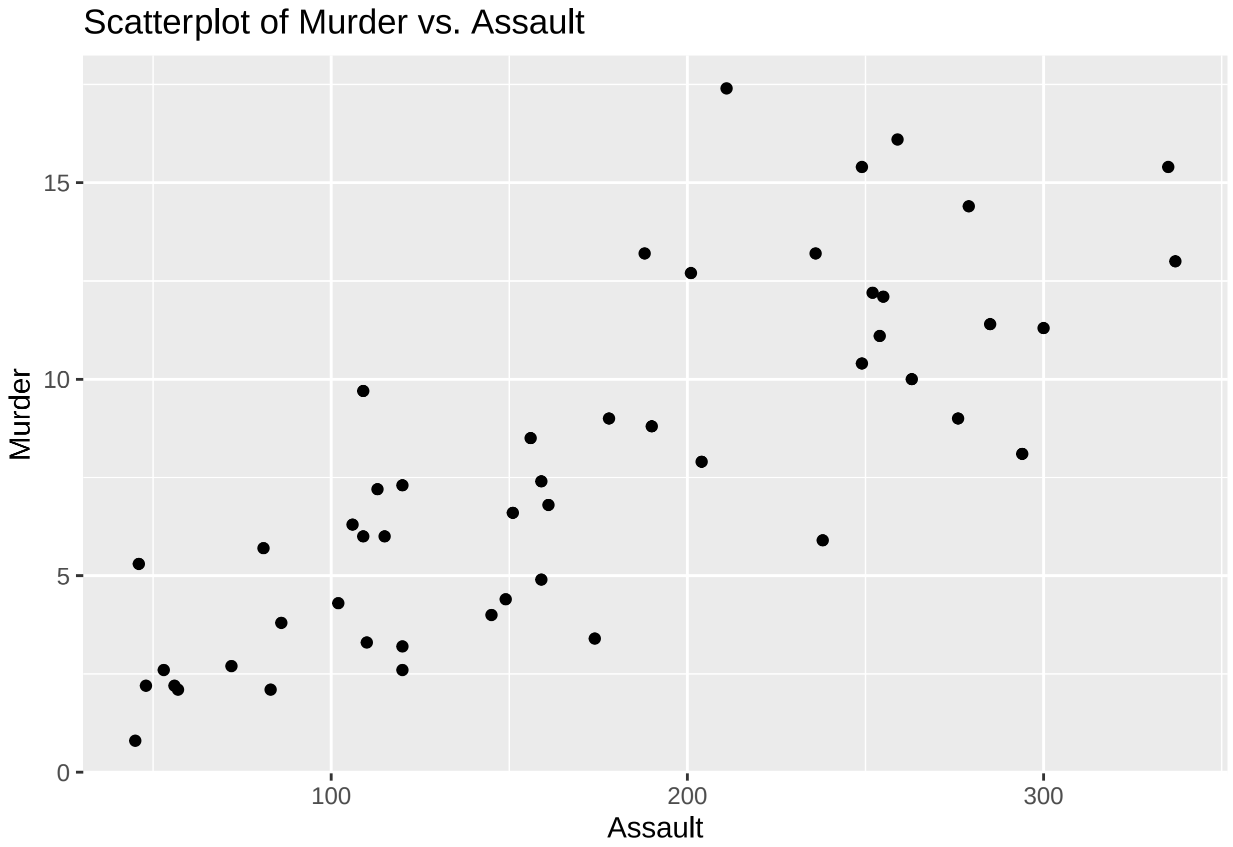
```
# Enter your code here!
pairs(USArrests)
```



```
cor(USArrests)
```

```
##              Murder  Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

```
ggplot(USArrests, aes(x=Assault, y=Murder)) + geom_point() + ggtitle("Scatterplot of Murder vs. Assault
```



Result:

Murder and assault appear to have a strong positive linear relationship. Murder and rape appear to have a moderate positive linear relationship, as do assault and rape. The percentage of urban population does not appear to be correlated with the other variables.

Question 3

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

Note: Check your working directory to make sure that you can download the data into the data folder.

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.

```r
# 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!
summary(housingData)
```

```
##  Neighborhood       Market.Value.per.SqFt     Boro             Year.Built
##  Length:2530        Min.   : 10.66        Length:2530        Min.   :1825
##  Class :character   1st Qu.: 75.10        Class :character   1st Qu.:1926
##  Mode  :character   Median :114.89        Mode  :character   Median :1986
##                     Mean   :133.17                           Mean   :1967
##                     3rd Qu.:189.91                           3rd Qu.:2005
##                     Max.   :399.38                           Max.   :2010
```

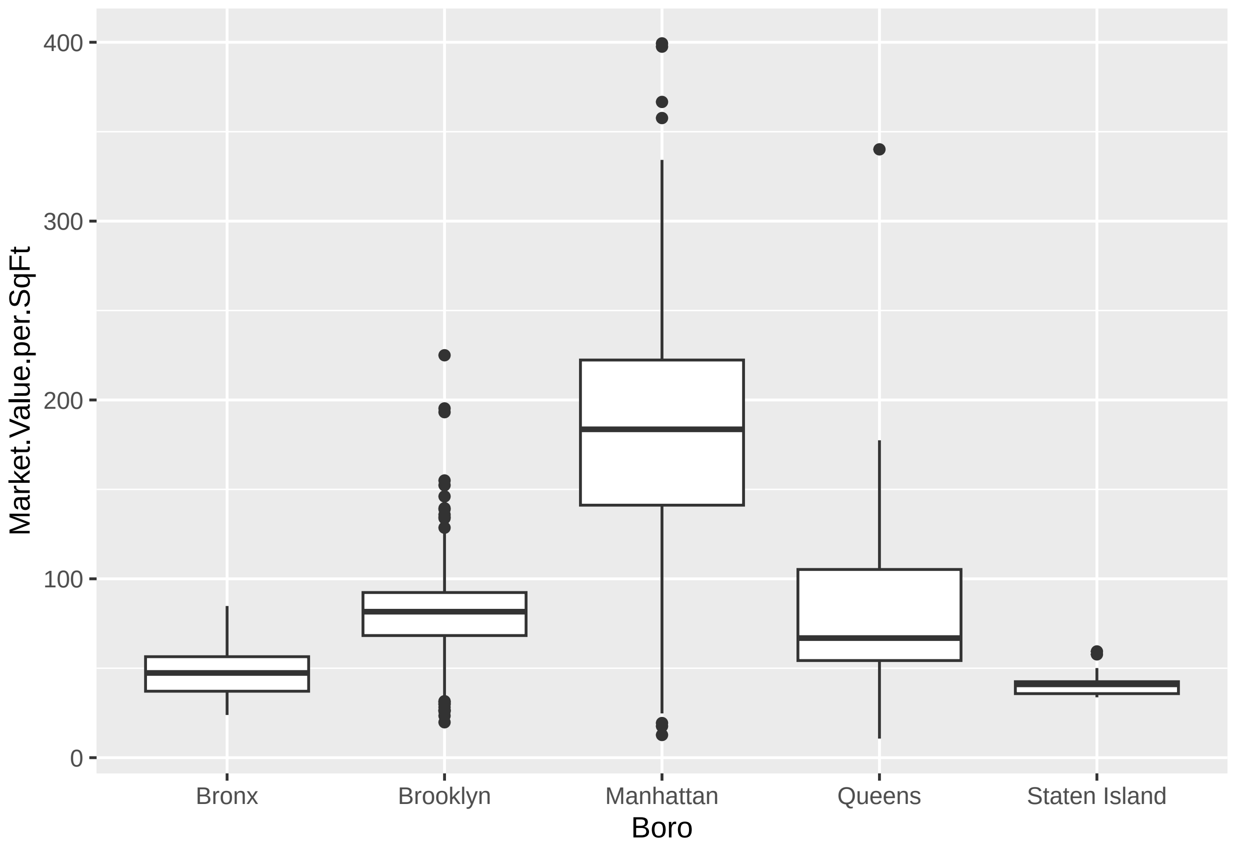
```
m_h <- mean(housingData$Market.Value.per.SqFt)
v_h <- var(housingData$Market.Value.per.SqFt)
s_h <- sd(housingData$Market.Value.per.SqFt)

print(paste("The average of market value per square footage from this data set is", m_h, "with variance
```

```
## [1] "The average of market value per square footage from this data set is 133.173098814229 with varia
```

```
m_Boro <- aggregate(housingData$Market.Value.per.SqFt, list(housingData$Boro), FUN=mean)
m_Boro
```

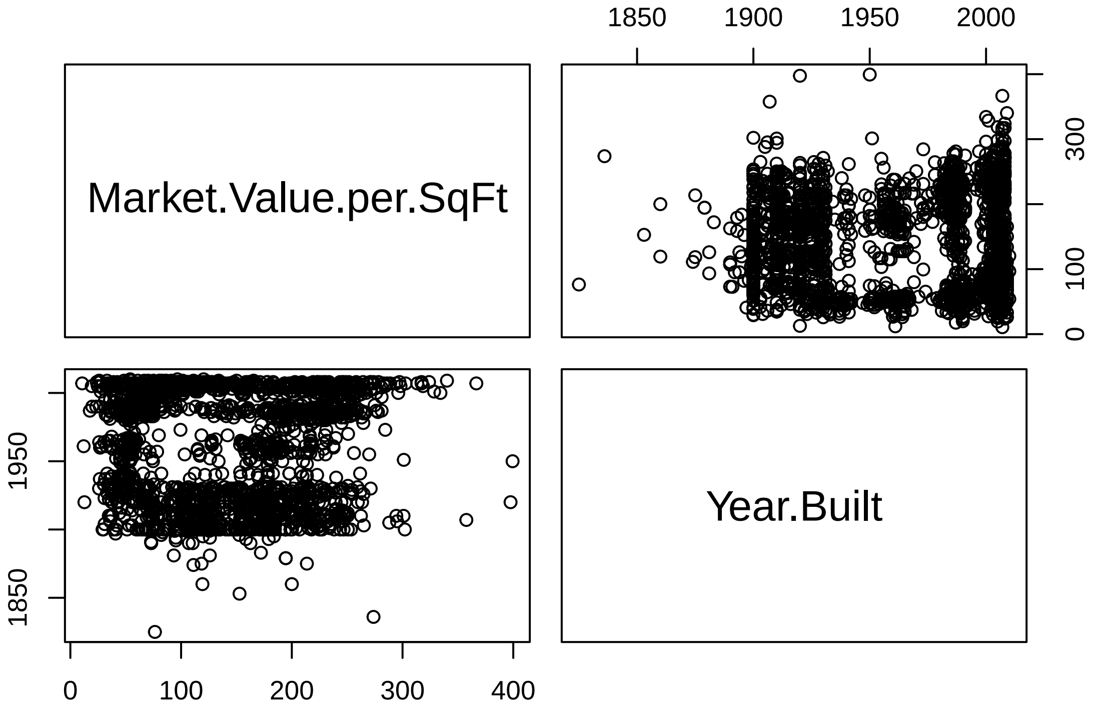
```
##         Group.1         x
## 1         Bronx  47.93232
## 2      Brooklyn  80.13439
## 3     Manhattan 180.59265
## 4        Queens  77.38137
## 5 Staten Island  41.26958
```

```
ggplot(housingData, aes(y=Market.Value.per.SqFt, x=Boro)) + geom_boxplot()
```



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

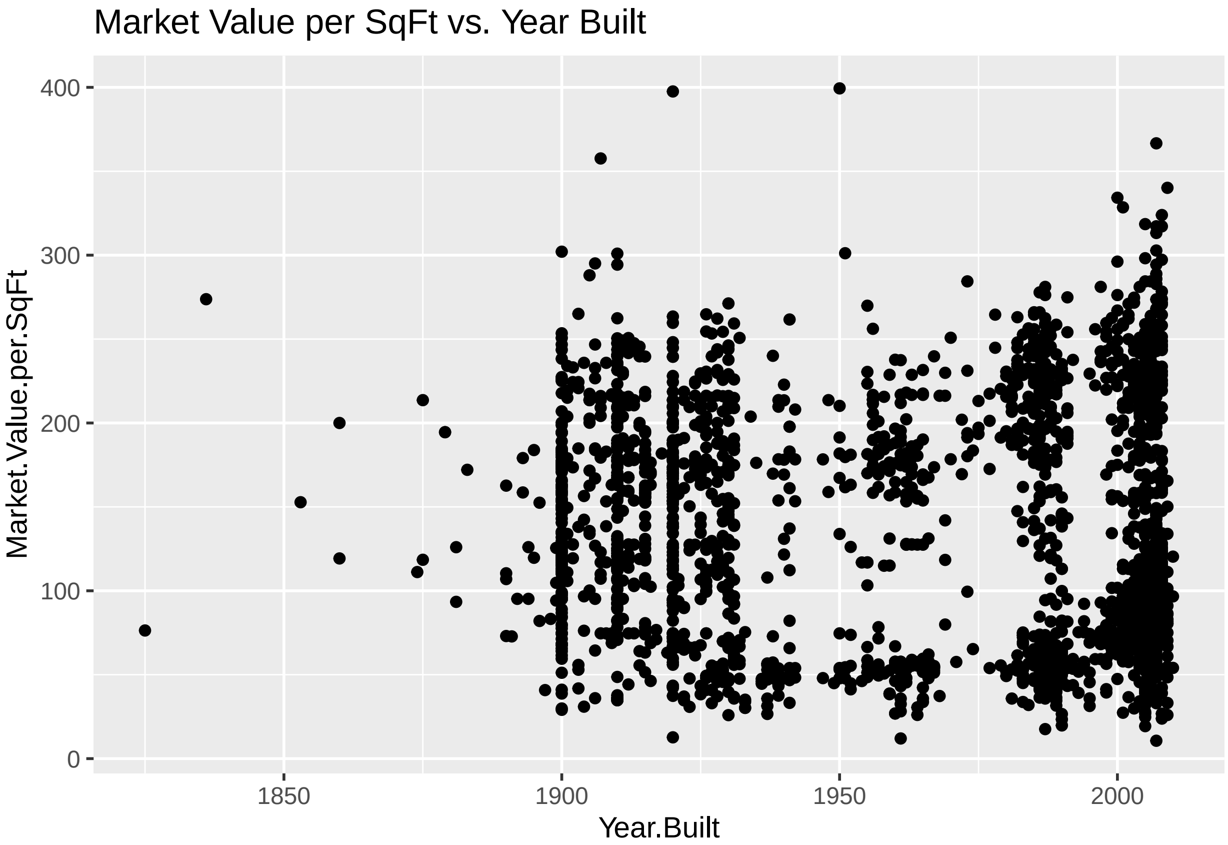
```
# Enter your code here!
pairs(housingData[, c(2,4)])
```



```
cor(housingData[, c(2,4)])
```

```
##                       Market.Value.per.SqFt  Year.Built
## Market.Value.per.SqFt            1.00000000 -0.09559073
## Year.Built                      -0.09559073  1.00000000
```

```
ggplot(housingData, aes(x=Year.Built, y=Market.Value.per.SqFt)) + geom_point() + ggtitle("Market Value p
```



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

There are 2530 houses in the dataset. The quantitative variables in the NYC housing data are market value per square footage and year built. Market value per square footage appears to have nearly 0 correlation with year built. The median and mean market value per square footage are 114.89 and 133.17. Grouping by borough, the following are listed in descending order of median market value: Manhattan, Brooklyn, Queens, Bronx, and Staten Island. Manhattan is the only borough with a median higher than the overall median. Therefore, Manhattan housing market values are skewing the distribution upward.