DS311 - R Lab Assignment

Thomas Cowart

2023-11-16

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.

```
# 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!
num_variables <- ncol(mtcars)
num_observations <- nrow(mtcars)
# 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
str(mtcars)
```

```
## ’data.frame’:    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```r
# Answer:
print("There are 7 discrete variables and 4 continuous variables in this data set.")
```

```
## [1] "There are 7 discrete variables and 4 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 Miles Per Gallon from this data set is '20.09', with variance '36.32', an
```

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

```
# Enter your code here!
mean_mpg_by_cyl <- aggregate(mpg ~ cyl, mtcars, mean)
sd_mpg_by_gear <- aggregate(mpg ~ gear, mtcars, sd)

print(mean_mpg_by_cyl)
```

```
##   cyl      mpg
## 1   4 26.66364
## 2   6 19.74286
## 3   8 15.10000
```

```
print(sd_mpg_by_gear)
```

```
##   gear      mpg
## 1    3 3.371618
## 2    4 5.276764
## 3    5 6.658979
```

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!
cyl_gear_crosstable <- table(mtcars$cyl, mtcars$gear)
print(cyl_gear_crosstable)
```

```
##    
##      3  4  5
##   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)
```

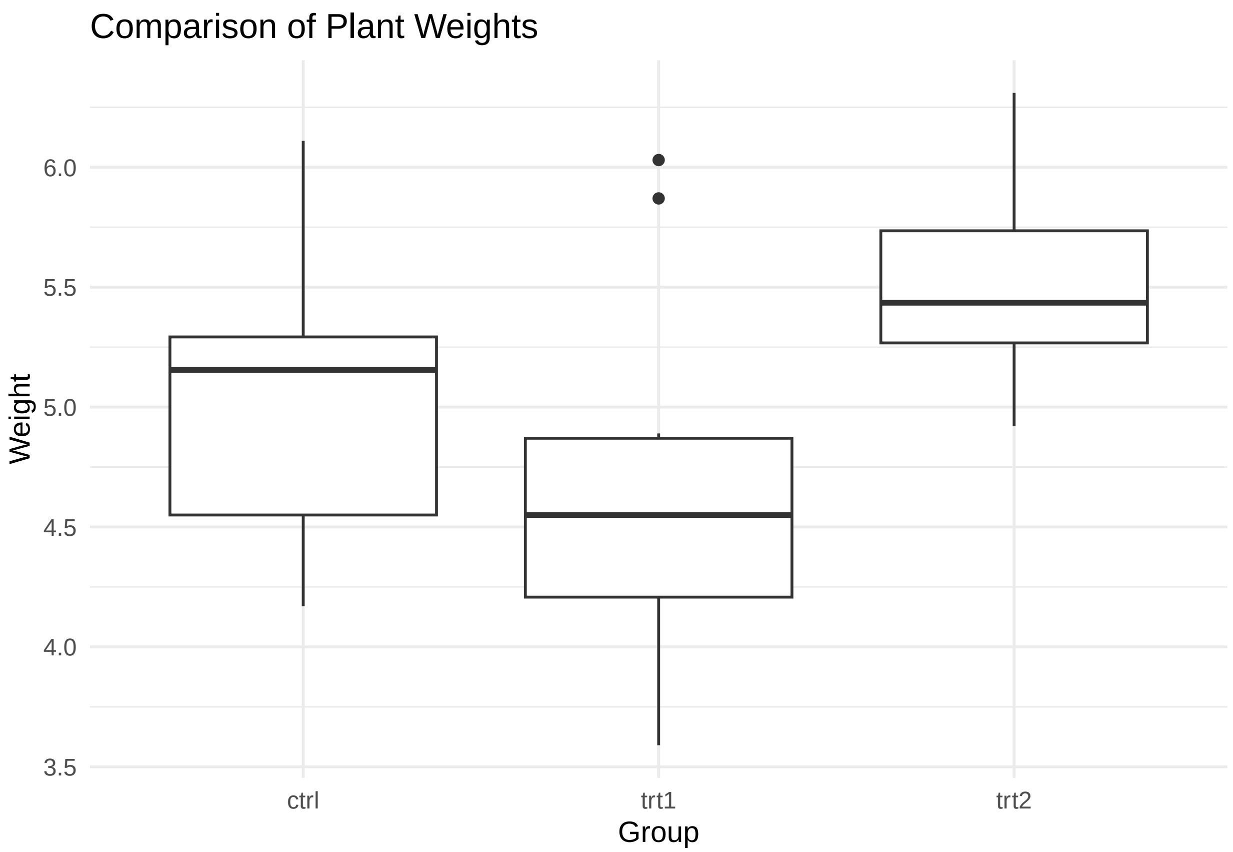
```
##   weight group
## 1   4.17  ctrl
## 2   5.58  ctrl
## 3   5.18  ctrl
## 4   6.11  ctrl
## 5   4.50  ctrl
## 6   4.61  ctrl
```

```
# Enter your code here!
install.packages("ggplot2")
```

```
## Installing package into '/Users/thomascowart/Library/R/arm64/4.3/library'
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/yc/43xbnrgd1fb0pgy34rx7x_580000gn/T//RtmpCyLpqy/downloaded_packages
```

```
library(ggplot2)
ggplot(PlantGrowth, aes(x = group, y = weight)) +
    geom_boxplot() +
    labs(title = "Comparison of Plant Weights",
         x = "Group",
         y = "Weight") +
    theme_minimal()
```



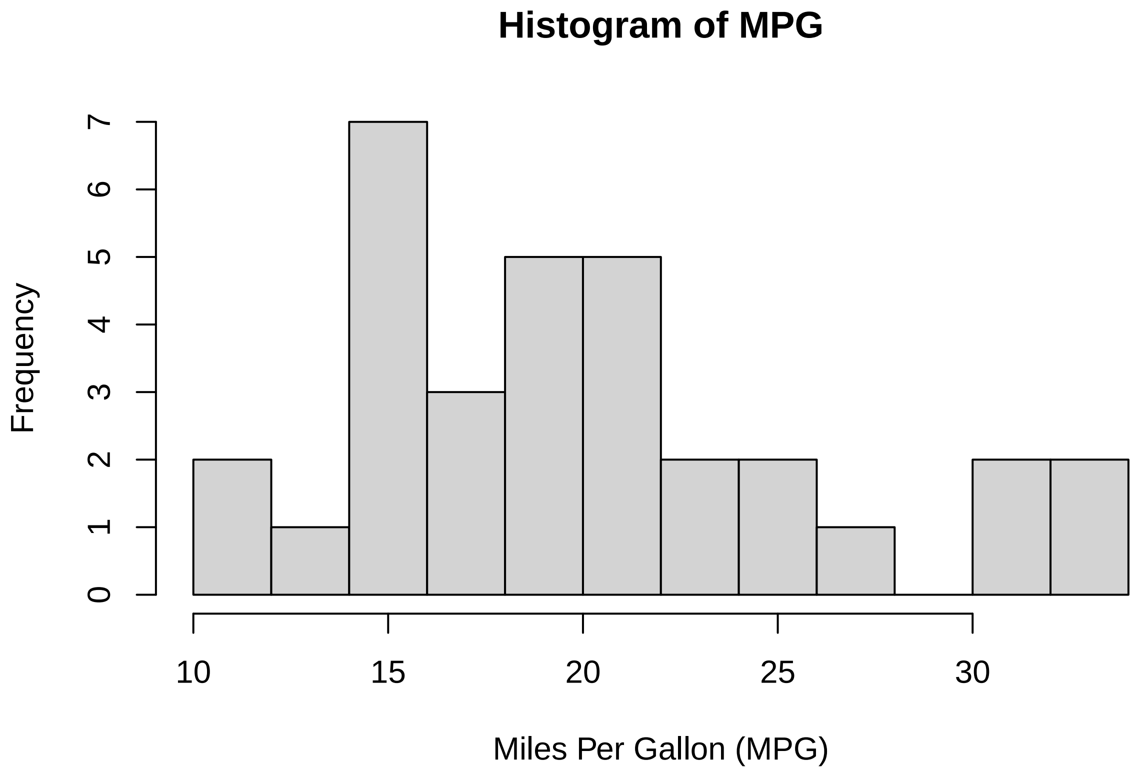
Result:

=> Report a paragraph to summarize your findings from the plot!

The control group's median weight is just above 5.0, without outliers. Treatment 1 has a similar median weight but includes two significant outliers, indicating some plants had much higher weights. Treatment 2 shows a higher median weight near 5.3 and a wider range, suggesting this treatment may increase plant weight more consistently. Overall, Treatment 2 appears to have the most positive impact on plant weight.

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.

```
hist(mtcars$mpg, breaks = 10, main = "Histogram of MPG", xlab = "Miles Per Gallon (MPG)", ylab = "Frequ
```



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

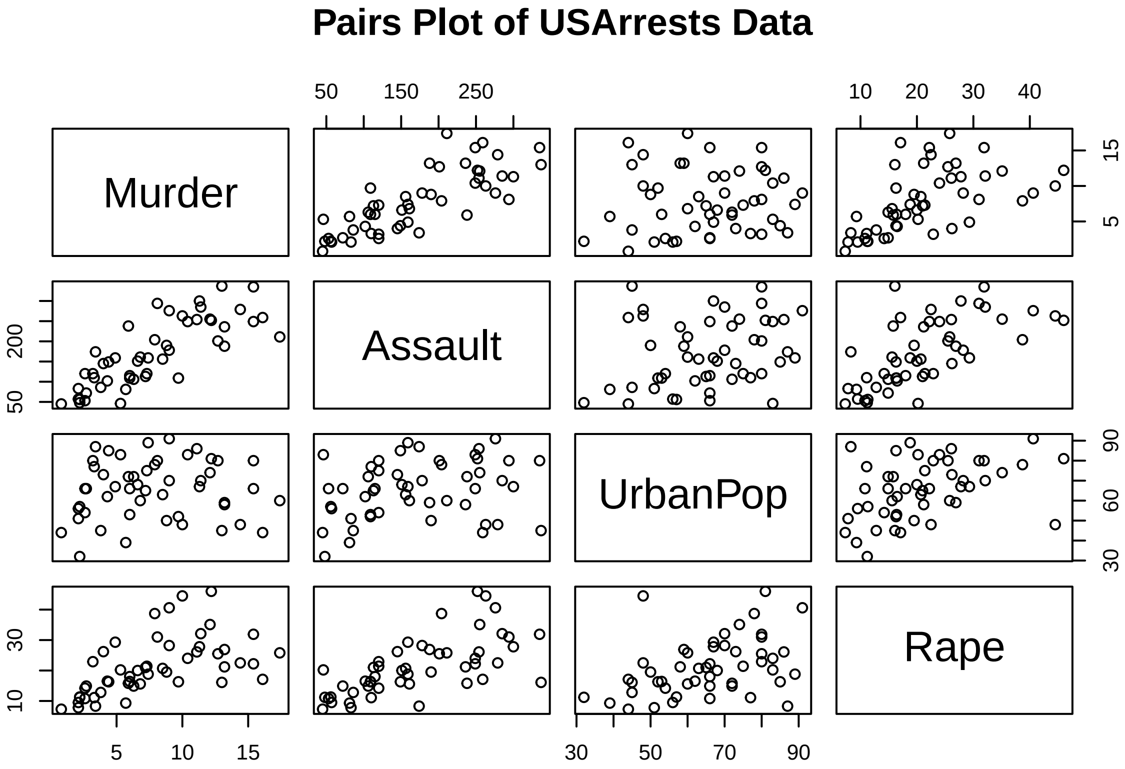
```
## [1] "Most of the cars in this data set are in the class of 15 miles per gallon."
```

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.

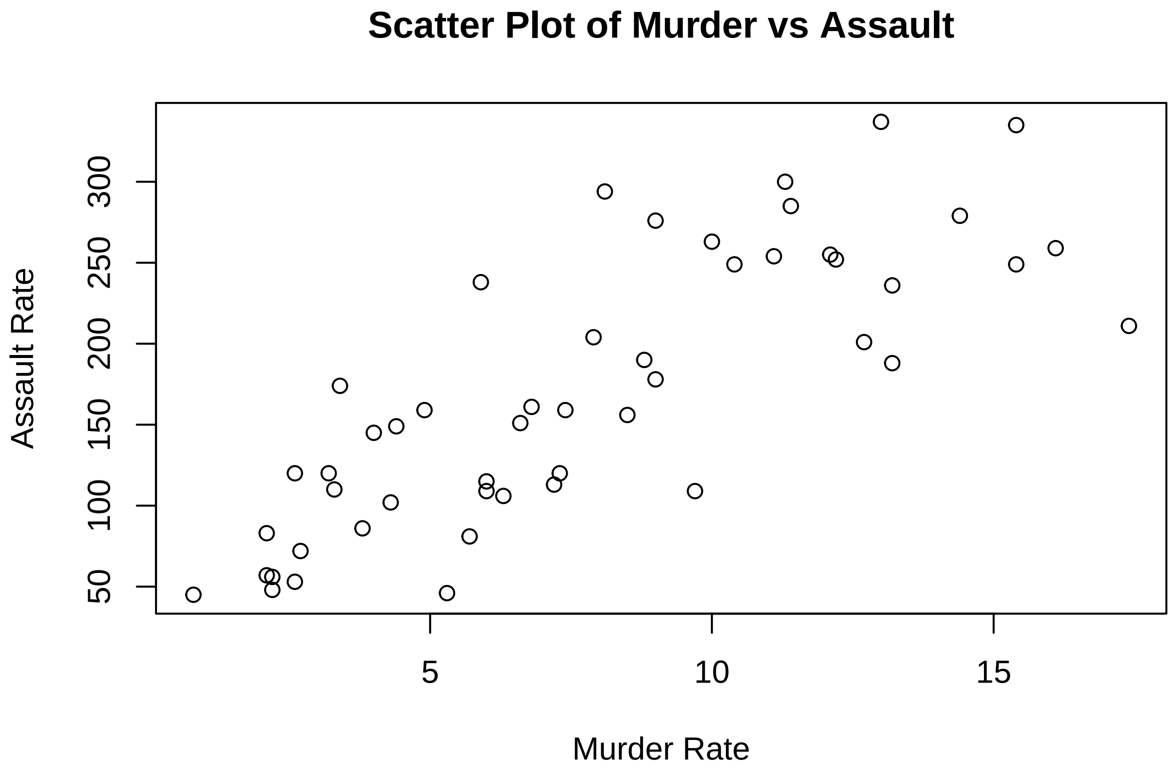
```r
# Load the data set
data("USArrests")

# 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, main = "Pairs Plot of USArrests Data")
```



```
plot(USArrests$Murder, USArrests$Assault, main = "Scatter Plot of Murder vs Assault",
     xlab = "Murder Rate", ylab = "Assault Rate")
```



Result:

=> Report a paragraph to summarize your findings from the plot!

The scatter plot shows a positive correlation between the two variables; as the Murder Rate increases, the Assault Rate tends to increase as well. The distribution of points suggests a linear relationship, with most data points clustered in the lower left of the plot, indicating that most observations have lower rates of both murders and assaults. There are a few states with higher rates of murder and assault, but these are less common. The plot does not indicate any outliers with extremely high or low rates when compared to the overall trend.

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.

```
# 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
```

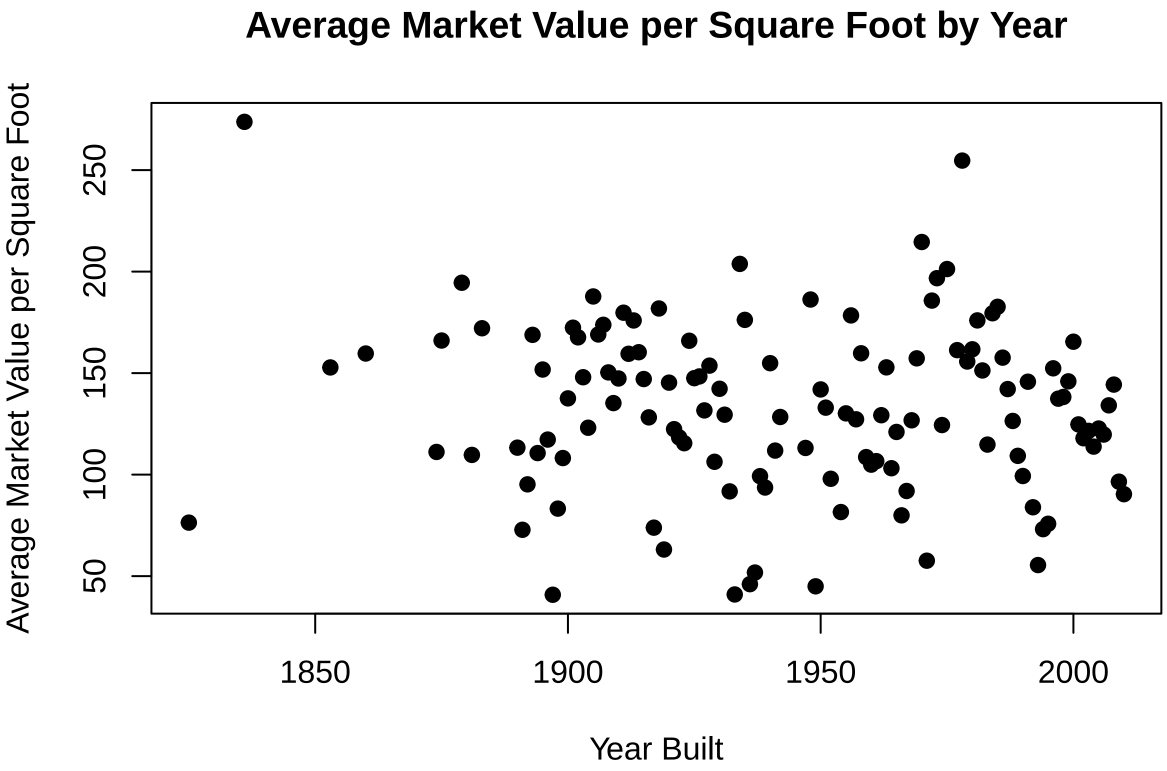
```
# Enter your code here!
summary(housingData)
```

```
##  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
```

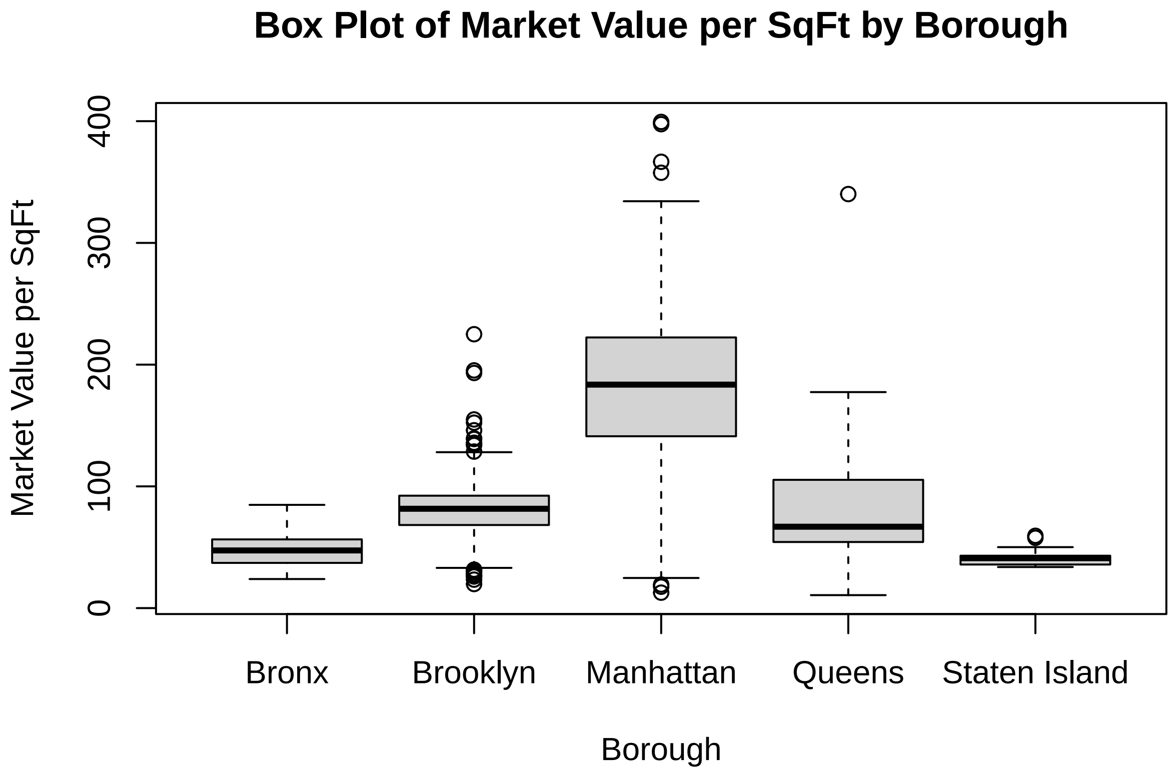
```
agg1 <- aggregate(Market.Value.per.SqFt ~ Year.Built, data = housingData, mean)
```

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!
plot(agg1$Year.Built, agg1$Market.Value.per.SqFt,
     main = "Average Market Value per Square Foot by Year",
     xlab = "Year Built", ylab = "Average Market Value per Square Foot", pch = 19)
```



```r
boxplot(housingData$Market.Value.per.SqFt ~ housingData$Boro,
        main = "Box Plot of Market Value per SqFt by Borough",
        xlab = "Borough",
        ylab = "Market Value per SqFt")
```



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

=> Enter your answer here!

1) The Scatter Plot

The scatter plot shows the average market value per square foot for properties built between 1850 and 2000. There is a wide variation in values, with a cluster of higher values around 1900 and a plateau from 1925 to 1975. Post-1975, values slightly decrease or stabilize. The data suggests that construction year is not a strong predictor of market value per square foot, given the high variability across the years.

2) The Box Plots

The box plots show that Manhattan has the highest median market value per square foot, followed by Brooklyn, Queens, and the Bronx. Manhattan also has the widest spread of values, indicating significant variability in property values. The Bronx has the lowest median value and less variability. There are outliers present in all boroughs, suggesting the presence of properties valued significantly different from the median.