DS311 - R Lab Assignment

Your Name

2023-03-26

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!
ncol(mtcars)
```

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

```
nrow(mtcars)
```

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

```
# 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 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.

```
# Enter your code here!
m <- mean(mtcars$mpg)
v <- var(mtcars$mpg)
s <- sd(mtcars$mpg)

m
```

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

```
v
```

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

```
s
```

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

```
# print(paste("The average of Mile Per Gallon from this data set is "20.10" with variance "36.32" and s
```

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!
avg_mpg_per_cyl <- aggregate(mpg ~ cyl, data=mtcars, mean)
avg_mpg_per_cyl
```

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

```
sd_mpg_per_gear <- aggregate(mpg ~ gear, data=mtcars, sd)
sd_mpg_per_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!
obs_cyl_gear <- table(mtcars$cyl, mtcars$gear)
obs_cyl_gear
```

```
##
##       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 o
```

Question 2

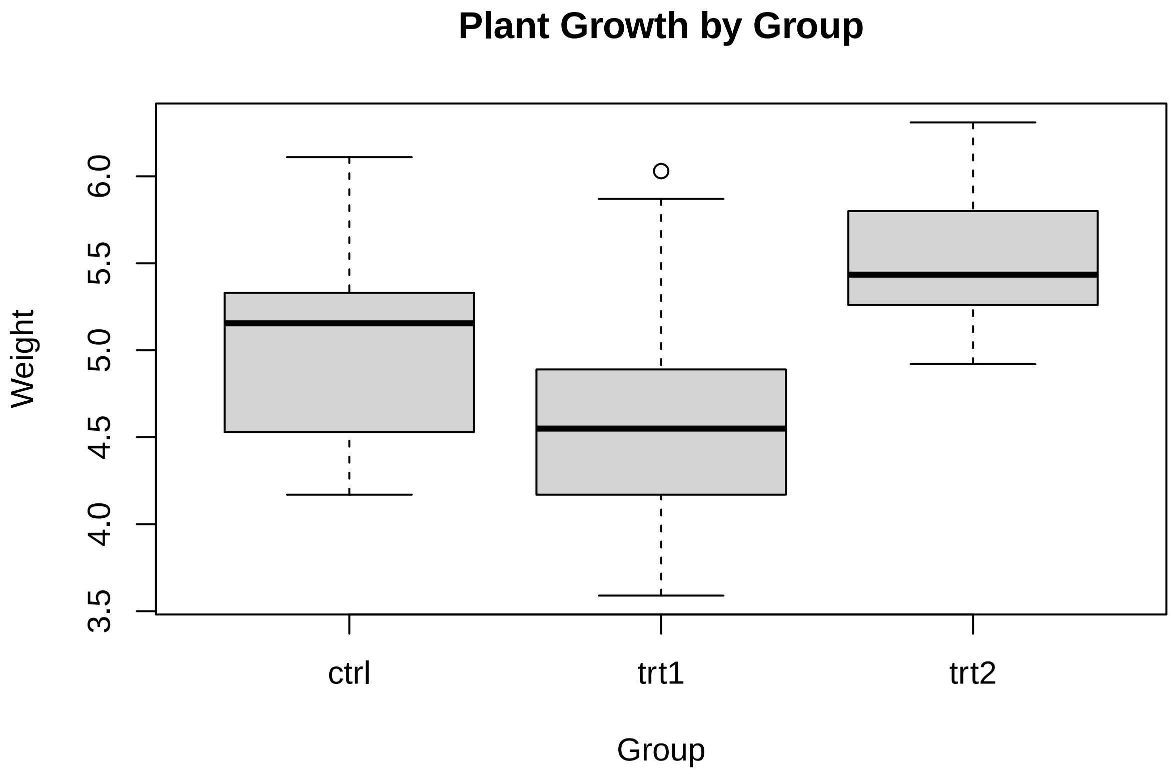
Use different visualization tools to summarize the data sets in this question.

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!
boxplot(weight ~ group, data = PlantGrowth, main = "Plant Growth by Group", xlab = "Group", ylab = "Wei
```



```
ctrl <- subset(PlantGrowth, group=='ctrl')
trt1 <- subset(PlantGrowth, group=='trt1')
trt2 <- subset(PlantGrowth, group=='trt2')
summary(ctrl)
```

```
##      weight        group   
##  Min.   :4.170   ctrl:10  
##  1st Qu.:4.550   trt1: 0  
##  Median :5.155   trt2: 0  
##  Mean   :5.032            
##  3rd Qu.:5.293            
##  Max.   :6.110            
```

```
summary(trt1)
```

```
##      weight        group   
##  Min.   :3.590   ctrl: 0  
##  1st Qu.:4.207   trt1:10  
##  Median :4.550   trt2: 0  
##  Mean   :4.661            
##  3rd Qu.:4.870            
##  Max.   :6.030            
```

```
summary(trt2)
```

```
##      weight        group   
##  Min.   :4.920   ctrl: 0  
##  1st Qu.:5.268   trt1: 0  
##  Median :5.435   trt2:10  
##  Mean   :5.526            
##  3rd Qu.:5.735            
##  Max.   :6.310            
```

Result:

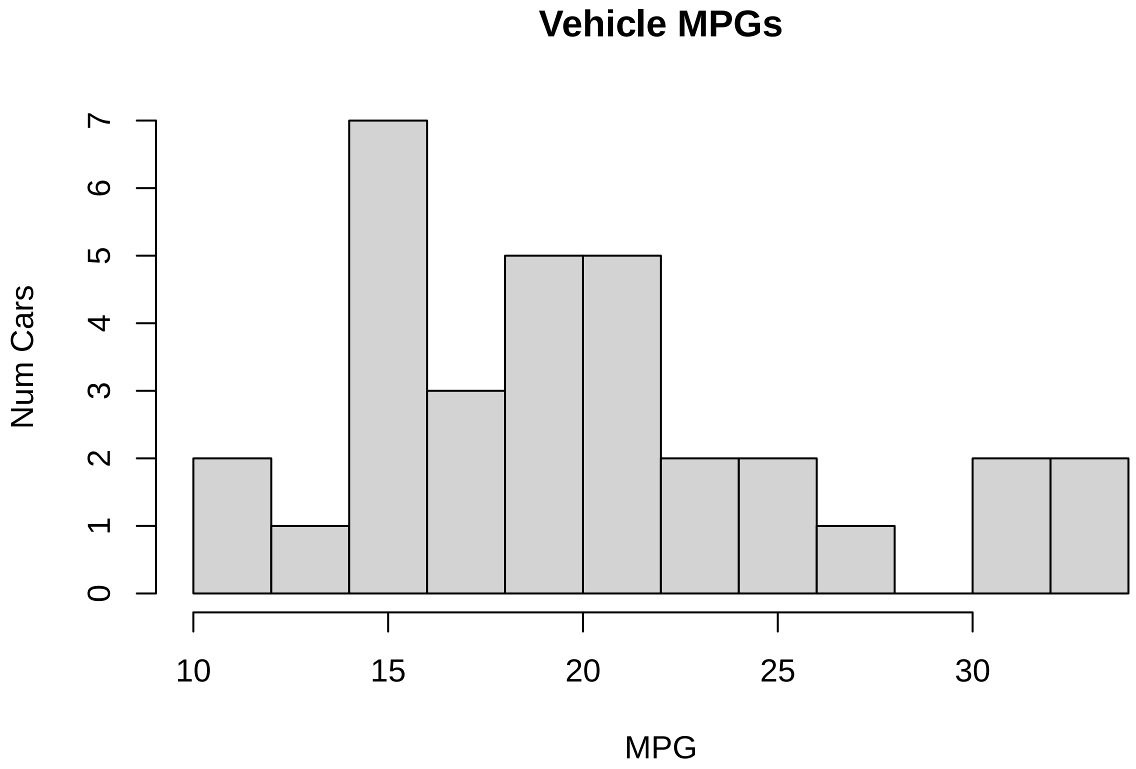
=> Report a paragraph to summarize your findings from the plot! From the results, we start to discover that the ctrl group has an IQR between 4.5 and 5.4, its median lying within the weight 5.2. The trt1 group has the most diversity, though the data has a smaller IQR, the data has an outlier in its size. Group trt2 has the heftier size with the lowest weight at 4.9.

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.

```
# Load the data set
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
```

```
# Enter your code here!
hist(mtcars$mpg,breaks = 10, main = "Vehicle MPGs", xlab = "MPG", ylab = "Num Cars")
```



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

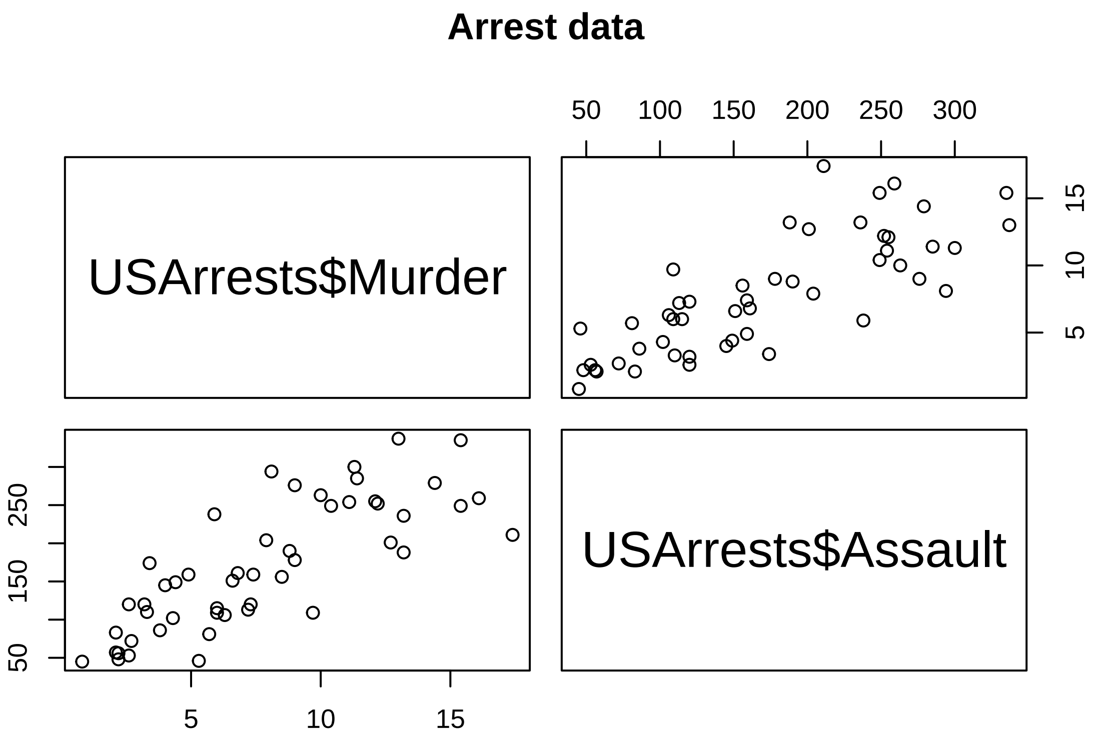
```
## [1] "Most of the cars in this data set are in the class of 15 miles per gallon."
```

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.

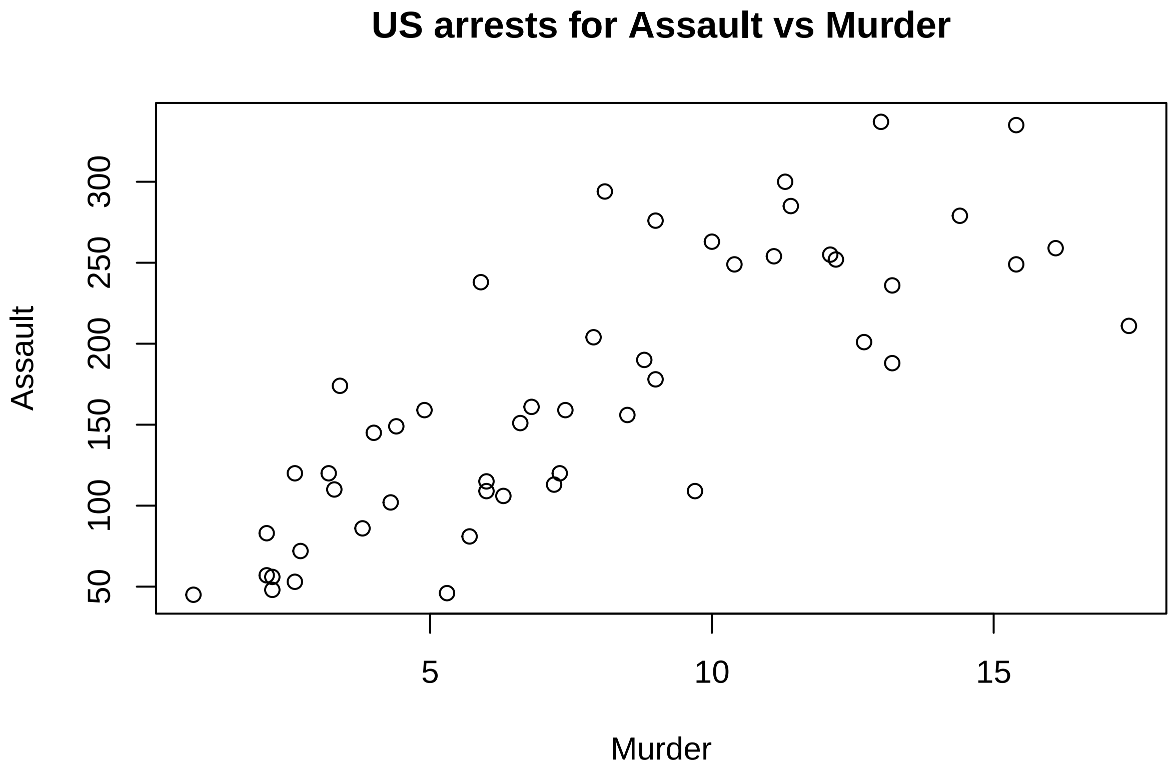
```
# Load the data set
data("USArrests")

# 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$Murder ~ USArrests$Assault, main = "Arrest data")
```



```
plot(USArrests$Murder, USArrests$Assault, main = "US arrests for Assault vs Murder", xlab = "Murder", y
```



Result:

=> Report a paragraph to summarize your findings from the plot! There are similar patterns between arrests for assault and arrests for murder. Low instances of assault/murder tend to be where the population lies.

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
```

```r
# Enter your code here!
avg_val_by_neigborhood <- aggregate(Market.Value.per.SqFt ~ Neighborhood, data = housingData, mean)
avg_val_by_neigborhood
```

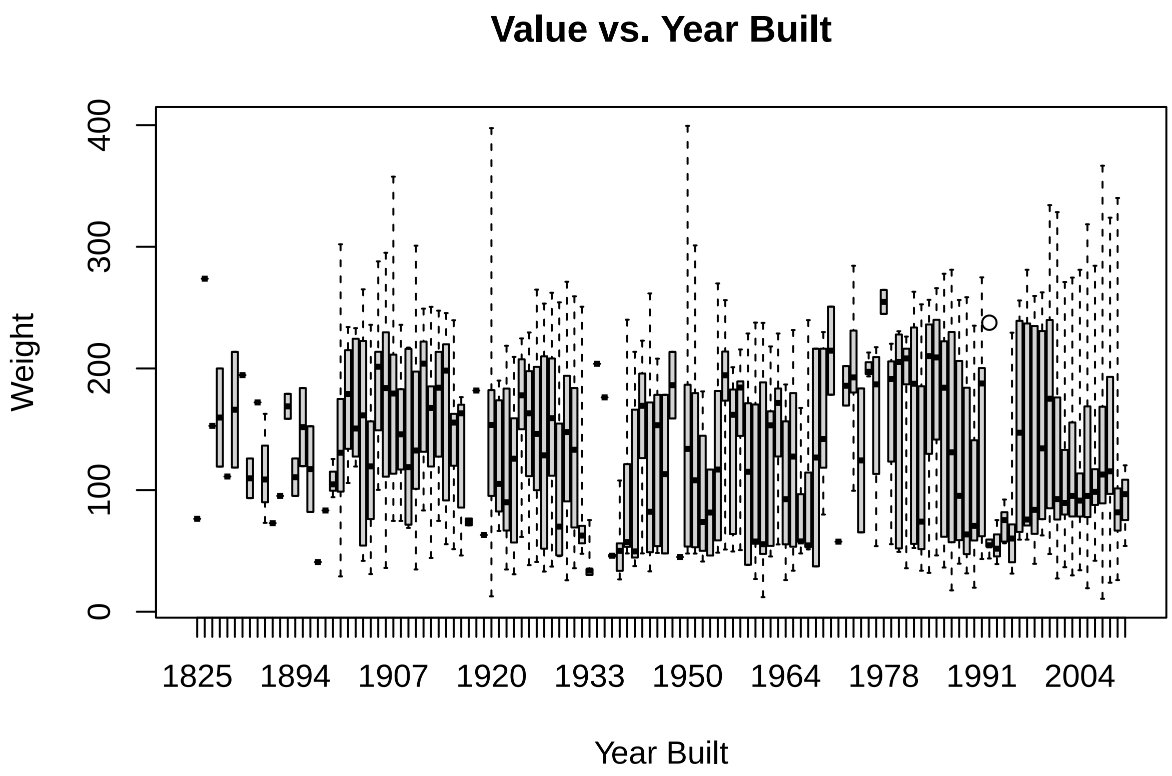
```
##                   Neighborhood Market.Value.per.SqFt
## 1                ALPHABET CITY             148.35500
## 2         ARROCHAR-SHORE ACRES              57.75000
## 3                      ASTORIA              91.48167
## 4                   BATH BEACH              70.34000
## 5                    BAY RIDGE              68.03500
## 6                      BAYSIDE              71.42111
## 7         BEDFORD PARK/NORWOOD              38.24500
## 8           BEDFORD STUYVESANT              83.24172
## 9                      BELMONT              56.45000
## 10                 BENSONHURST              71.70429
## 11                BERGEN BEACH              73.27000
## 12                 BOERUM HILL              96.57600
## 13                BOROUGH PARK              64.10857
## 14                   BRIARWOOD              75.36250
## 15              BRIGHTON BEACH              81.91429
## 16               BRONX-UNKNOWN              32.06500
## 17                   BRONXDALE              28.94333
## 18            BROOKLYN HEIGHTS             114.11778
## 19               BUSH TERMINAL              60.95000
## 20                    BUSHWICK              76.13500
## 21                    CANARSIE              46.58000
## 22             CARROLL GARDENS              93.40556
## 23                     CHELSEA             215.94932
## 24                   CHINATOWN             154.17952
## 25                 CITY ISLAND              40.83000
## 26                CIVIC CENTER             174.06696
## 27                     CLINTON             176.70032
## 28                CLINTON HILL              88.97385
## 29                 COBBLE HILL             120.69800
## 30            COBBLE HILL-WEST              85.71125
## 31               COLLEGE POINT              65.05000
## 32                CONEY ISLAND              55.05750
## 33                      CORONA              94.20706
## 34               CROWN HEIGHTS              64.26286
## 35       DOWNTOWN-FULTON FERRY             103.26857
## 36        DOWNTOWN-FULTON MALL             132.42500
## 37          DOWNTOWN-METROTECH             122.48000
## 38               DYKER HEIGHTS              68.36000
## 39               EAST NEW YORK              36.99167
## 40                EAST TREMONT              72.33333
## 41                EAST VILLAGE             207.46115
## 42                    ELMHURST              69.80564
## 43                FAR ROCKAWAY              74.88500
## 44                     FASHION             194.81067
## 45                   FINANCIAL             199.30917
## 46           FLATBUSH-CENTRAL              65.71167
## 47   FLATBUSH-LEFFERTS GARDEN              46.27000
## 48             FLATBUSH-NORTH              54.00000
## 49                   FLATIRON             223.30311
## 50             FLUSHING-NORTH              80.16992
## 51             FLUSHING-SOUTH              89.62750
## 52       FLUSHING MEADOW PARK              58.59000
## 53               FOREST HILLS              70.20706
## 54                FORT GREENE              81.76900
## 55                   GLENDALE              57.39667
## 56                    GOWANUS              82.45333
## 57                   GRAMERCY             188.68471
## 58                 GRANT CITY              47.60000
## 59                  GRAVESEND              75.63526
## 60                GREAT KILLS              33.74000
## 61                 GREENPOINT              86.18053
## 62  GREENWICH VILLAGE-CENTRAL             142.57767
## 63     GREENWICH VILLAGE-WEST             202.13667
## 64                GRYMES HILL              50.09000
## 65                    HAMMELS             139.07200
## 66             HARLEM-CENTRAL             102.79106
## 67                HARLEM-EAST             139.93972
## 68               HARLEM-UPPER              79.25667
## 69                HARLEM-WEST              95.20500
## 70  HIGHBRIDGE/MORRIS HEIGHTS              61.82000
## 71                  HILLCREST              53.95000
## 72                     HOLLIS             109.56000
## 73               HOWARD BEACH              55.06000
## 74                     INWOOD              62.05500
## 75            JACKSON HEIGHTS              47.79238
## 76                    JAMAICA             104.76600
## 77            JAMAICA ESTATES              79.69500
## 78              JAVITS CENTER             125.09000
## 79                 KENSINGTON              56.87500
## 80                KEW GARDENS              69.64300
## 81   KINGSBRIDGE HTS/UNIV HTS              23.86000
## 82    KINGSBRIDGE/JEROME PARK              58.37800
## 83                   KIPS BAY             191.31769
## 84               LITTLE ITALY             142.52308
## 85                LITTLE NECK              65.85000
## 86           LONG ISLAND CITY             108.16667
## 87            LOWER EAST SIDE             173.56262
## 88                    MADISON              71.26000
## 89           MANHATTAN VALLEY             111.30043
## 90                    MASPETH              53.32750
## 91             MIDDLE VILLAGE              78.35857
## 92                MIDTOWN CBD             234.36154
## 93               MIDTOWN EAST             211.04750
## 94               MIDTOWN WEST             222.06489
## 95                    MIDWOOD              79.50273
## 96        MORNINGSIDE HEIGHTS              74.63000
## 97       MORRIS PARK/VAN NEST              26.90000
## 98        MORRISANIA/LONGWOOD              44.21250
## 99     MOTT HAVEN/PORT MORRIS              30.96000
## 100               MURRAY HILL             206.26795
## 101              NEW BRIGHTON              41.47667
## 102   NEW BRIGHTON-ST. GEORGE              41.06000
## 103           NEW SPRINGVILLE              40.47000
## 104           OAKLAND GARDENS              66.94000
## 105                OCEAN HILL              37.92900
## 106       OCEAN PARKWAY-NORTH              76.51111
## 107       OCEAN PARKWAY-SOUTH              75.08000
## 108                OZONE PARK              54.10000
## 109                PARK SLOPE              88.01774
## 110          PARK SLOPE SOUTH              95.84200
## 111               PARKCHESTER              32.67500
## 112      PELHAM PARKWAY SOUTH              30.55000
## 113          PROSPECT HEIGHTS              79.16200
## 114                 REGO PARK              62.13630
## 115                 RIDGEWOOD              64.28667
## 116                 RIVERDALE              57.10176
## 117             ROCKAWAY PARK              88.13600
## 118  SCHUYLERVILLE/PELHAM BAY              49.68000
## 119            SHEEPSHEAD BAY              79.79704
## 120               SILVER LAKE              35.80500
## 121                      SOHO             162.72473
## 122                 SOUNDVIEW              43.40333
## 123          SOUTH OZONE PARK              40.78000
## 124               SOUTHBRIDGE             159.53333
## 125                 SUNNYSIDE              61.61818
## 126               SUNSET PARK              80.58348
## 127               THROGS NECK              53.70667
## 128             TOMPKINSVILLE              35.81000
## 129                   TRIBECA             180.18473
## 130   UPPER EAST SIDE (59-79)             216.83715
## 131   UPPER EAST SIDE (79-96)             202.45179
## 132  UPPER EAST SIDE (96-110)             167.41600
## 133   UPPER WEST SIDE (59-79)             200.24391
## 134   UPPER WEST SIDE (79-96)             171.84515
## 135  UPPER WEST SIDE (96-116)             134.09353
## 136  WASHINGTON HEIGHTS LOWER              65.29600
## 137  WASHINGTON HEIGHTS UPPER              93.50833
## 138         WEST NEW BRIGHTON              39.69000
## 139                WHITESTONE              72.90000
## 140            WILLIAMSBRIDGE              42.46000
## 141      WILLIAMSBURG-CENTRAL              79.97017
## 142         WILLIAMSBURG-EAST              84.32605
## 143        WILLIAMSBURG-NORTH              84.10577
## 144        WILLIAMSBURG-SOUTH              82.27618
## 145           WINDSOR TERRACE              70.21200
## 146                 WOODHAVEN              38.61000
## 147                  WOODSIDE              80.52625
## 148           WYCKOFF HEIGHTS              84.93000
```

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

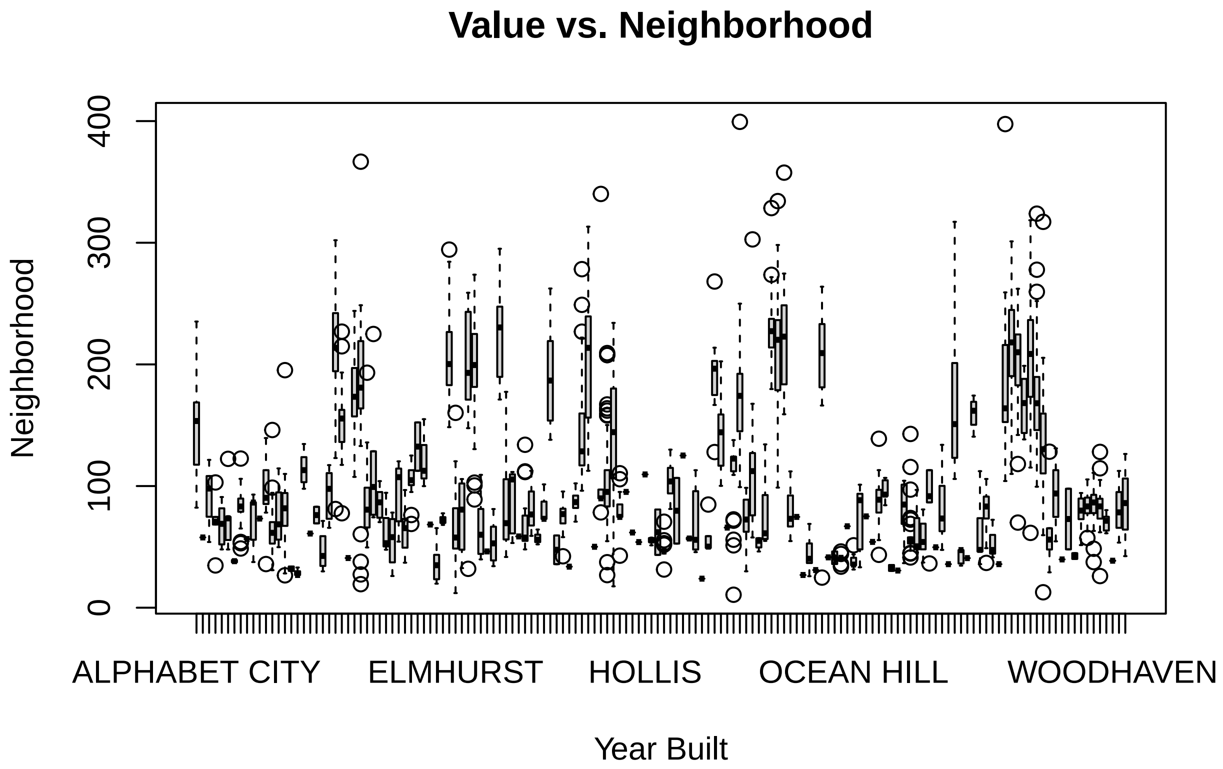
```
##     Year.Built Market.Value.per.SqFt
## 1         1825              76.36000
## 2         1836             273.77000
## 3         1853             152.79000
## 4         1860             159.64500
## 5         1874             111.17000
## 6         1875             166.05000
## 7         1879             194.52000
## 8         1881             109.70500
## 9         1883             172.10000
## 10        1890             113.28750
## 11        1891              72.83000
## 12        1892              95.21000
## 13        1893             168.85000
## 14        1894             110.62000
## 15        1895             151.77500
## 16        1896             117.26500
## 17        1897              40.83000
## 18        1898              83.25000
## 19        1899             108.16000
## 20        1900             137.55908
## 21        1901             172.36778
## 22        1902             167.62167
## 23        1903             147.97000
## 24        1904             123.09333
## 25        1905             187.76583
## 26        1906             169.03364
## 27        1907             173.80000
## 28        1908             150.35000
## 29        1909             135.23667
## 30        1910             147.36257
## 31        1911             179.76067
## 32        1912             159.51636
## 33        1913             175.93500
## 34        1914             160.29286
## 35        1915             147.08673
## 36        1916             128.20714
## 37        1917              73.87000
## 38        1918             181.84000
## 39        1919              63.11000
## 40        1920             145.30862
## 41        1921             122.39125
## 42        1922             118.33250
## 43        1923             115.47625
## 44        1924             165.94091
## 45        1925             147.51316
## 46        1926             148.36423
## 47        1927             131.63357
## 48        1928             153.68375
## 49        1929             106.32121
## 50        1930             142.28936
## 51        1931             129.51731
## 52        1932              91.74333
## 53        1933              40.97000
## 54        1934             203.80000
## 55        1935             176.23000
## 56        1936              46.04333
## 57        1937              51.77250
## 58        1938              99.23857
## 59        1939              93.65083
## 60        1940             154.89857
## 61        1941             111.83733
## 62        1942             128.38600
## 63        1947             113.13500
## 64        1948             186.25000
## 65        1949              44.98000
## 66        1950             141.96182
## 67        1951             132.98833
## 68        1952              97.95143
## 69        1954              81.56500
## 70        1955             130.17538
## 71        1956             178.42786
## 72        1957             127.24091
## 73        1958             159.77000
## 74        1959             108.62692
## 75        1960             104.91200
## 76        1961             106.63000
## 77        1962             129.26294
## 78        1963             152.82937
## 79        1964             103.15000
## 80        1965             121.01313
## 81        1966              79.94375
## 82        1967              91.94000
## 83        1968             126.76000
## 84        1969             157.28000
## 85        1970             214.59000
## 86        1971              57.60000
## 87        1972             185.72000
## 88        1973             196.75500
## 89        1974             124.42500
## 90        1975             201.26667
## 91        1977             161.32250
## 92        1978             254.69000
## 93        1979             155.71333
## 94        1980             161.74500
## 95        1981             175.96800
## 96        1982             151.30364
## 97        1983             114.79917
## 98        1984             179.48333
## 99        1985             182.66868
## 100       1986             157.62328
## 101       1987             142.14055
## 102       1988             126.43686
## 103       1989             109.25390
## 104       1990              99.31500
## 105       1991             145.76105
## 106       1992              83.92333
## 107       1993              55.45000
## 108       1994              73.13500
## 109       1995          75.77375
## 110       1996         152.36750
## 111       1997         137.41364
## 112       1998         138.25125
## 113       1999         145.93217
## 114       2000         165.47296
## 115       2001         124.74295
## 116       2002         117.92442
## 117       2003         121.56193
## 118       2004         113.79702
## 119       2005         122.70817
## 120       2006         119.73598
## 121       2007         134.12665
## 122       2008         144.34935
## 123       2009          96.52619
## 124       2010          90.36667
```

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!
boxplot(Market.Value.per.SqFt ~ Year.Built, data = housingData, main = "Value vs. Year Built", xlab = "
```



```
boxplot(Market.Value.per.SqFt ~ Neighborhood, data = housingData, main = "Value vs. Neighborhood", xlab
```



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

=> Within the market value by year built, the area with the largest range in pricing is between 1918 - 1959.