DS311 - R Lab Assignment

Nytina Cooks

1/26/2022

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.

```
# Loading the data
data(mtcars)

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

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

6 rows | 1-10 of 12 columns

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

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

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

```
dim(mtcars)
```

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

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

```
## [1] "There are total of 32  variables and 11 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 _____ discrete variables and _____ continuous variables in this data set.")
```

```
## [1] "There are _____ discrete variables and _____ 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!

mean(mtcars[["mpg"]]) -> m
var(mtcars[["mpg"]]) -> v
sd(mtcars[["mpg"]]) -> s


# print
(paste("The average of Mile Per Gallon from this data set is ", 20.09 , " with variance ", 36.32
         , " and standard deviation", 6.03 , "."))
```

```
## [1] "The average of Mile Per Gallon from this data set is  20.09  with variance  36.32  and s
tandard deviation 6.03 ."
```

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

```
# Enter your code here!


aggregate(mtcars$mpg, by=list(Category=mtcars$cyl), FUN=mean)
```

Category <dbl>	x <dbl>
4	26.66364
6	19.74286
8	15.10000

3 rows

```
aggregate(mtcars$mpg, by=list(Category=mtcars$gear), FUN=sd)
```

Category <dbl>	x <dbl>
3	3.371618
4	5.276764
5	6.658979

3 rows

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!
#install.packages("janitor")

library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
tabyl(mtcars, gear, cyl)
```

gear <dbl>	4 <dbl>	6 <dbl>	8 <dbl>
3	1	2	12
4	8	4	0
5	2	1	2

3 rows

```
print("The most common car type in this data set is a car with _8__ cylinders and _3___ gears. T
        here are total of __12___ cars belong to this specification in the data set.")
```

```
## [1] "The most common car type in this data set is a car with _8__ cylinders and _3___ gears.
There are total of __12___ cars belong to this specification in the data set."
```

Question 2

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

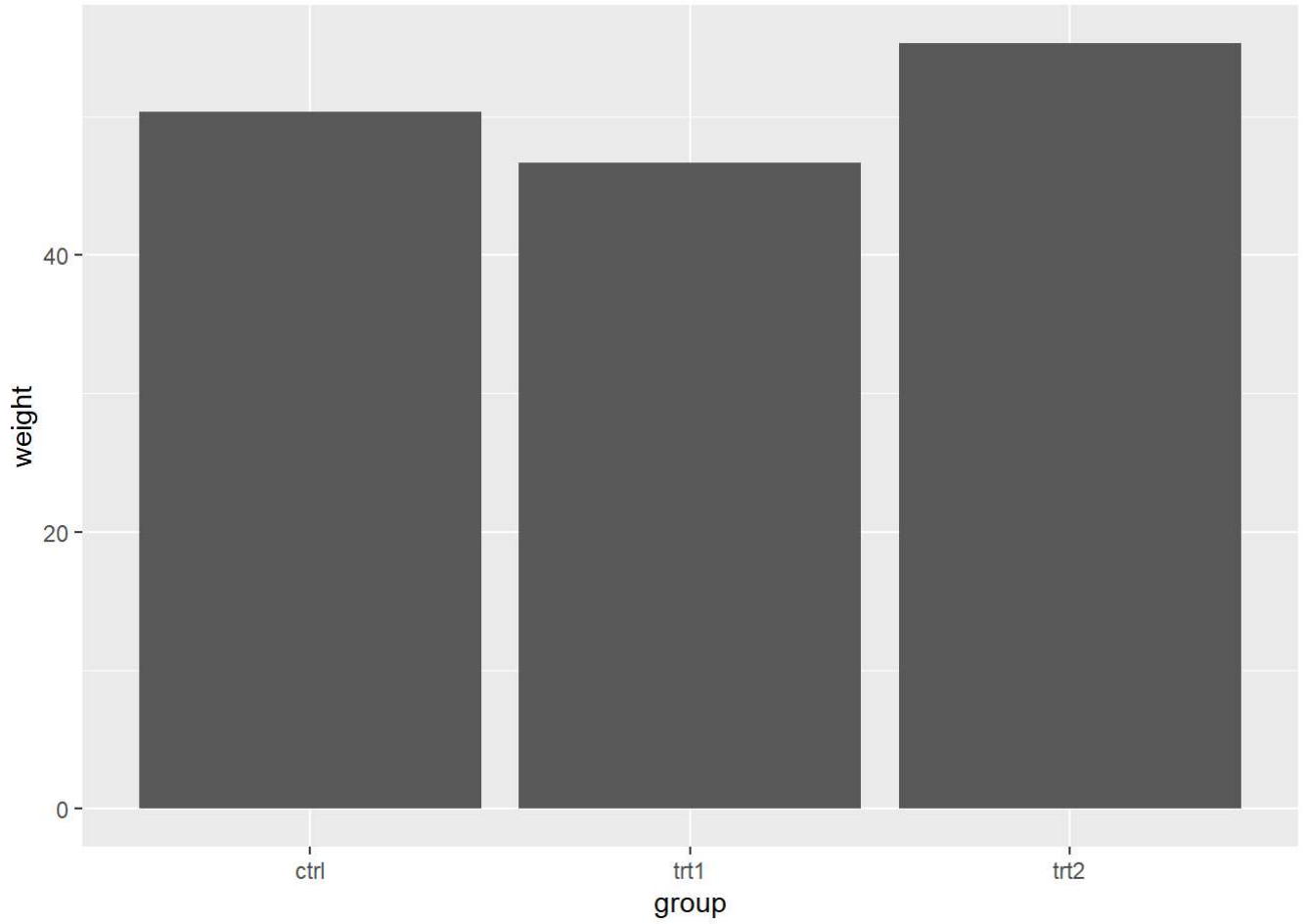
a. Using the **PlantGrowth** data set, visualize and compare the weight of the plant in the three separated group. Give labels to the title, x-axis, and y-axis on the graph. Write a paragraph to summarize your findings in this graph.

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

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

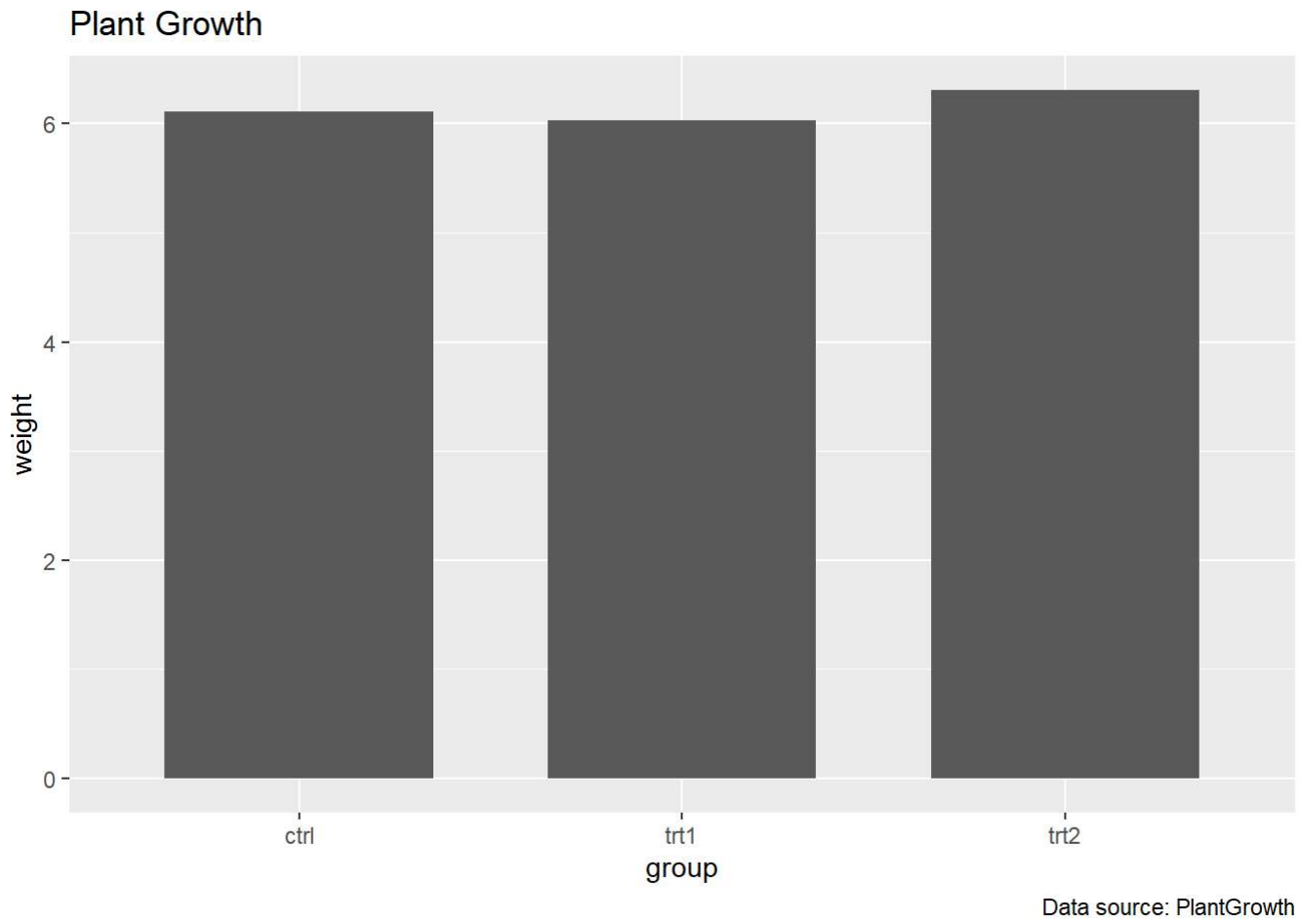
	weight <dbl>	group <fct>
1	4.17	ctrl
2	5.58	ctrl
3	5.18	ctrl
4	6.11	ctrl
5	4.50	ctrl
6	4.61	ctrl

6 rows

```
# Enter your code here!
library(ggplot2)

ggplot(PlantGrowth, aes(x = group, y = weight)) +
  geom_bar(
    aes(),
    stat = "identity", position = position_stack()
    ) +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF"))+
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF"))
```




```
p <- ggplot(PlantGrowth, aes(x = group, y = weight)) +
  geom_bar(
    aes(),
    stat = "identity", position = position_dodge(0.8),
    width = 0.7
    ) +
  scale_color_manual(values = c("#0073C2FF", "#EFC000FF"))+
  scale_fill_manual(values = c("#0073C2FF", "#EFC000FF"))
p <- p + labs(title = "Plant Growth",
              caption = "Data source: PlantGrowth")
p
```



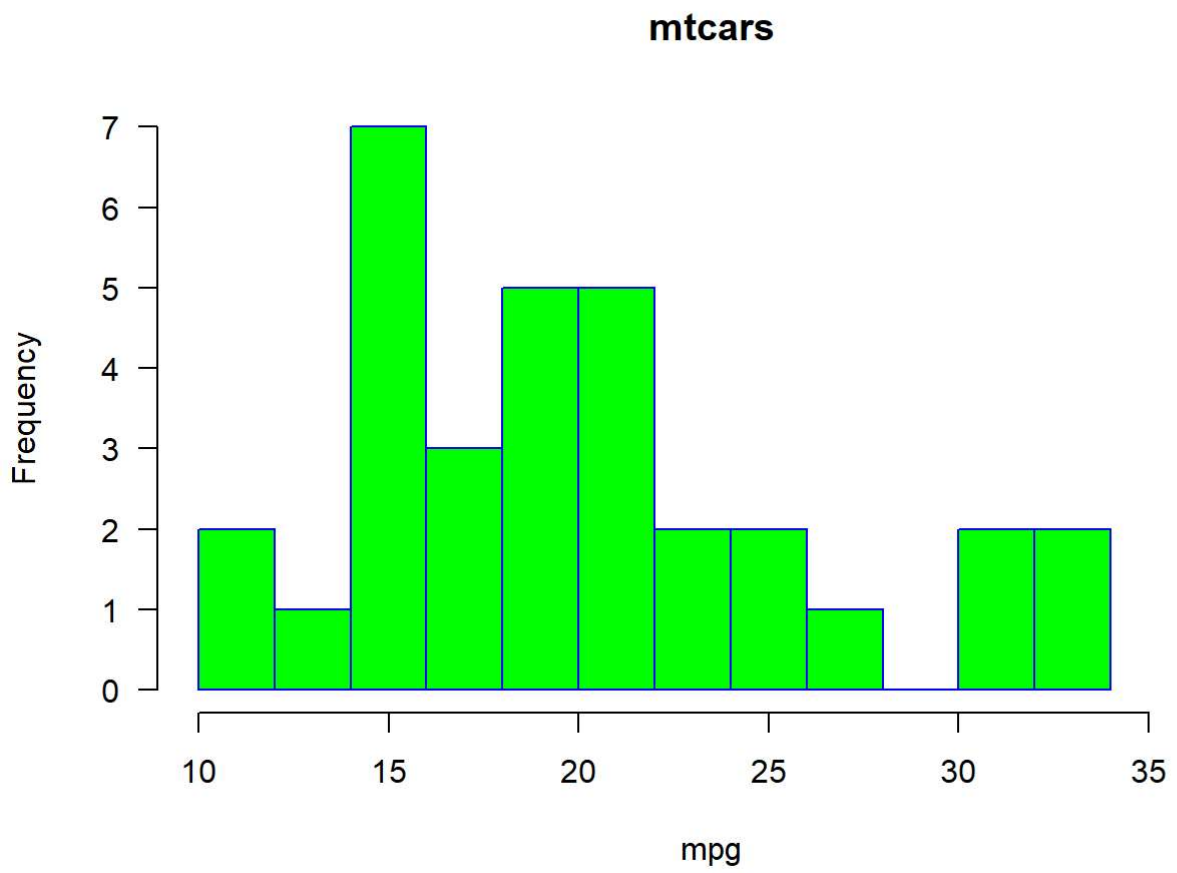
Result:

=> Enter your results here!

For this data set the data is split into weights for 3 different groups of plants ctrl, trt1, and trt2. The ctrl group reaches a max growth of 6.11 units while the trt1 group has a max of 6.03, and trt2 group has a max of 6.31.

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.

```
# histogram with added parameters

hist(mtcars$mpg,
     main="mtcars",
     xlab="mpg",
     border="blue",
     col="green",
     xlim=c(10,37),
     las=1,
     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.

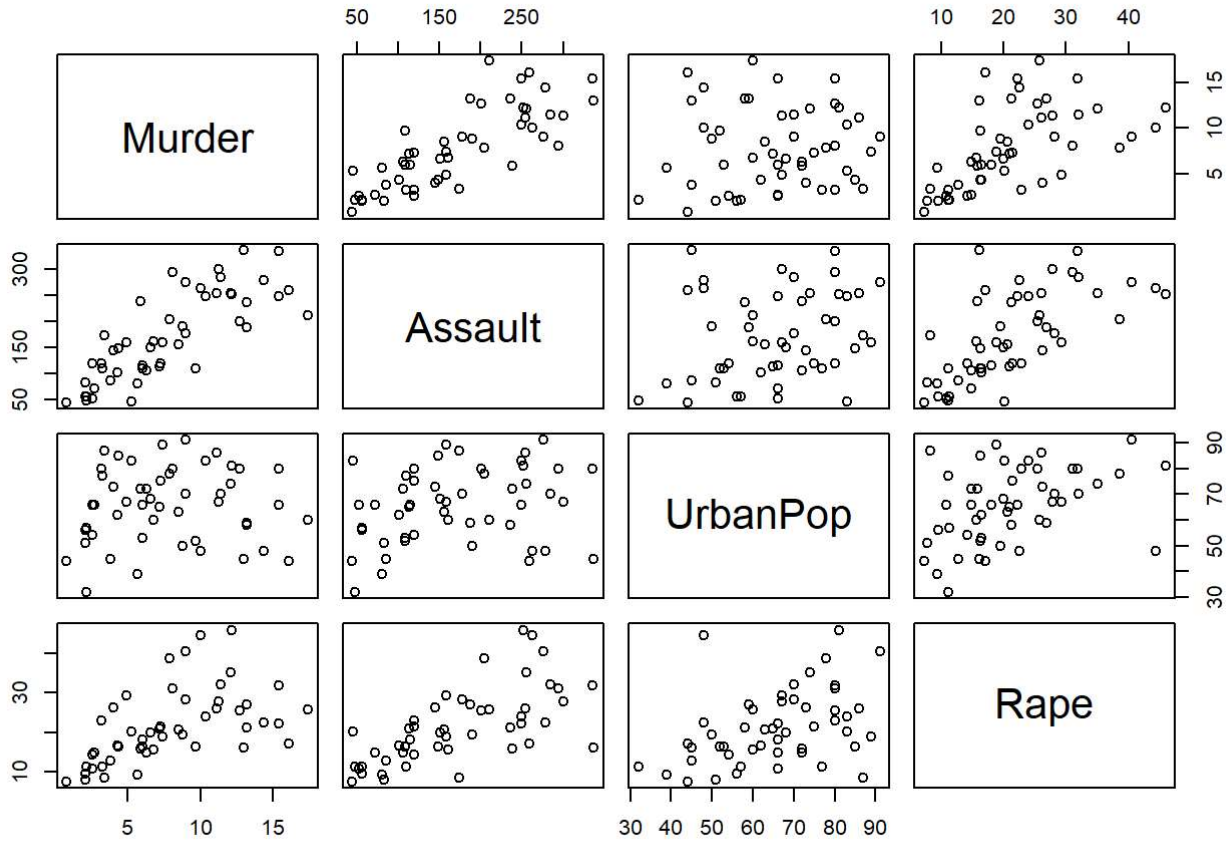
```
# Load the data set
data("USArrests")

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

	Murder <dbl>	Assault <int>	UrbanPop <int>	Rape <dbl>
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

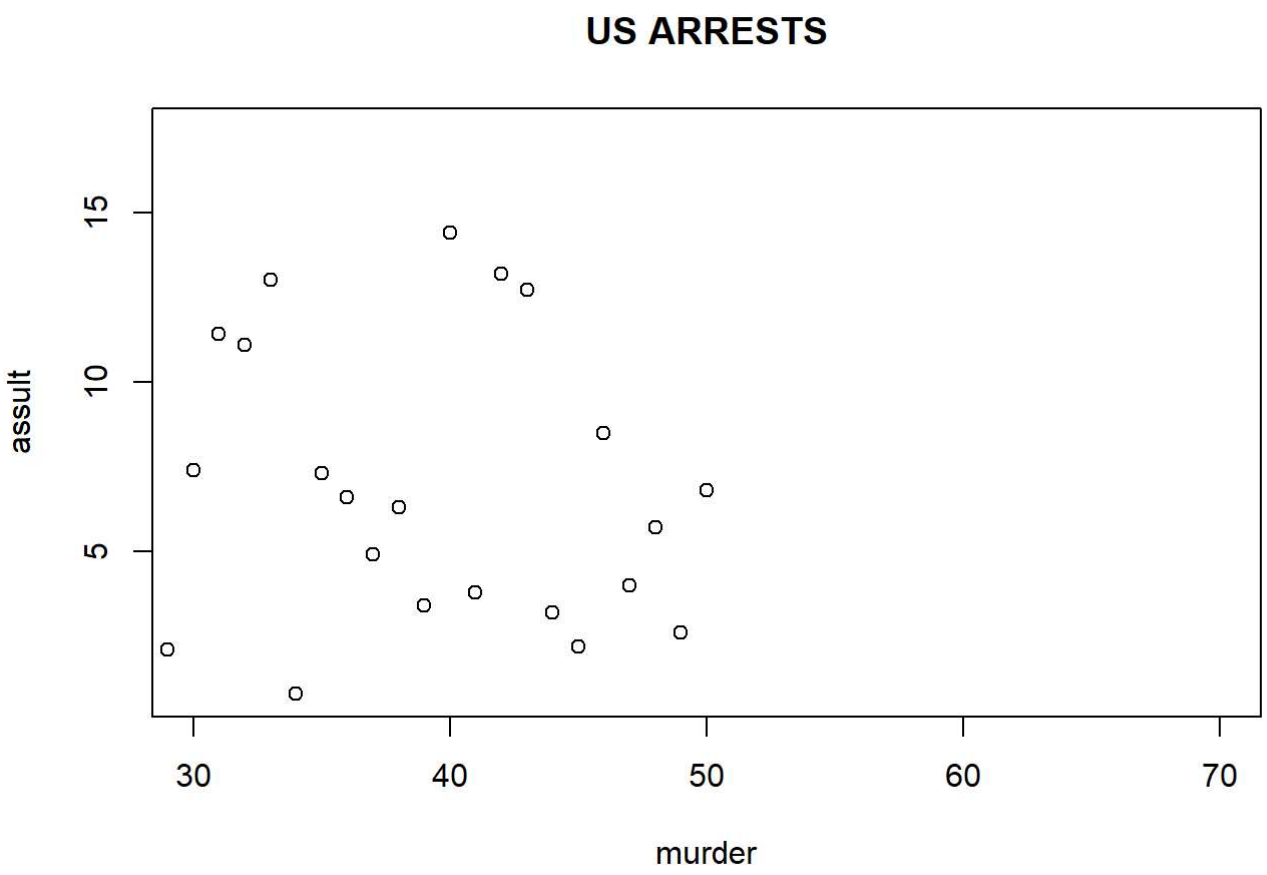
	Murder <dbl>	Assault <int>	UrbanPop <int>	Rape <dbl>
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

6 rows

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




```
plot(x= USArrests$Murder,y=USArrests$assult,xlab = "murder",ylab = "assult",xlim = c(50,50),main
        = "US ARRESTS")
```




```
print("the graphs using the data from the arrests dataset indicates that there is no apparent co
        rrelation between the number of murder arrests in a state and the number of assult arre
        sts. THe number of arrests overall appear to be more reliant on other variables such as
        population of the state which remains unseen in the dataset." )
```

```
## [1] "the graphs using the data from the arrests dataset indicates that there is no apparent c
orrelation between the number of murder arrests in a state and the number of assult arrests. THe
number of arrests overall appear to be more reliant on other variables such as population of the
state which remains unseen in the dataset."
```

Result:

=> Enter your result here!

Question 3

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

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
	<chr>	<dbl>	<chr>	<int>
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

6 rows

```
# Enter your code here!
summary(housingData$Market.Value.per.SqFt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.66   75.10  114.89  133.17  189.91  399.38
```

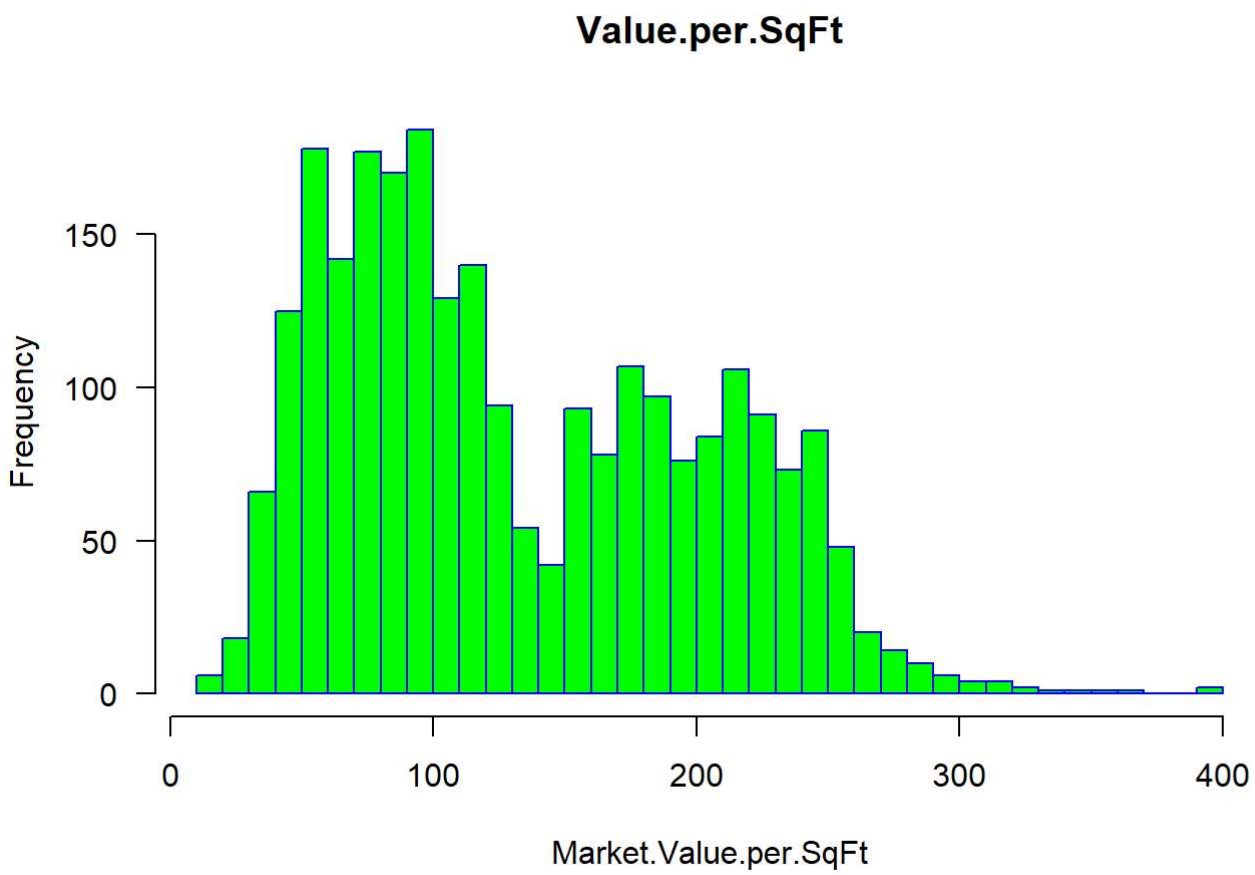
```
aggregate(housingData$Market.Value.per.SqFt, by=list(Category=housingData$Year.Built), FUN=mean)
```

Category	x
<int>	<dbl>
1825	76.36000
1836	273.77000
1853	152.79000
1860	159.64500
1874	111.17000
1875	166.05000
1879	194.52000
1881	109.70500
1883	172.10000
1890	113.28750

1-10 of 124 rows Previous **1** 2 3 4 5 6 … 13 Next

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!
hist(housingData$Market.Value.per.SqFt,
     main="Value.per.SqFt",
     xlab="Market.Value.per.SqFt",
     border="blue",
     col="green",
     xlim=c(10,400),
     las=1,
     breaks=55)
```




```
plot(x= housingData$Market.Value.per.SqFt,y=housingData$Year.Built,xlab = "value per sqf",ylab =
         "year built",xlim = c(10,400),main = "housing data")
```

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

Enter your answer here! through looking at the frequency table of the dataset we are able to determine that there are a largeer number of instances of market value of up to 100 per square feet. there is also more value per square foot for houses built closer to the years 1900 and 2000. I chose these graphs to show the significance of the two categories of year and value.