DS311 - R Lab Assignment

Chris Albert

8/22/2022

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 the **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!
ncol(mtcars)
```

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

```
nrow(mtcars)
```

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

```
# Answer:
print("There are total of 32 variables and 352 observations in this data set.")
```

```
## [1] "There are total of 32 variables and 352 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 6 discrete variables and 5 continuous variables in this data set.")
```

```
## [1] "There are 6 discrete variables and 5 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 <- sqrt(v)

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

```
## [1] "The average of Mile Per Gallon from this data set is  20.090625  with variance  36.3241028225806  and sta
ndard deviation 6.0269480520891 ."
```

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!
library(dplyr)
```

```
## 
## 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
```

```
avgclass = mtcars %>% group_by(mtcars$cyl) %>%
  summarise(mean_mpg = mean(mtcars$mpg))
print(avgclass)
```

```
## # A tibble: 3 × 2
##   `mtcars$cyl` mean_mpg
##          <dbl>    <dbl>
## 1            4     20.1
## 2            6     20.1
## 3            8     20.1
```

```
stndevgear = mtcars %>% group_by(mtcars$gear) %>%
  summarise(standard_deviation = sd(mtcars$mpg))
print (avgclass)
```

```
## # A tibble: 3 × 2
##   `mtcars$cyl` mean_mpg
##          <dbl>    <dbl>
## 1            4     20.1
## 2            6     20.1
## 3            8     20.1
```

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!
library(gmodels)
CrossTable(mtcars$cyl, mtcars$gear)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  32 
## 
## 
##              | mtcars$gear 
##   mtcars$cyl |         3 |         4 |         5 | Row Total | 
## -------------|-----------|-----------|-----------|-----------|
##            4 |         1 |         8 |         2 |        11 | 
##              |     3.350 |     3.640 |     0.046 |           | 
##              |     0.091 |     0.727 |     0.182 |     0.344 | 
##              |     0.067 |     0.667 |     0.400 |           | 
##              |     0.031 |     0.250 |     0.062 |           | 
## -------------|-----------|-----------|-----------|-----------|
##            6 |         2 |         4 |         1 |         7 | 
##              |     0.500 |     0.720 |     0.008 |           | 
##              |     0.286 |     0.571 |     0.143 |     0.219 | 
##              |     0.133 |     0.333 |     0.200 |           | 
##              |     0.062 |     0.125 |     0.031 |           | 
## -------------|-----------|-----------|-----------|-----------|
##            8 |        12 |         0 |         2 |        14 | 
##              |     4.505 |     5.250 |     0.016 |           | 
##              |     0.857 |     0.000 |     0.143 |     0.438 | 
##              |     0.800 |     0.000 |     0.400 |           | 
##              |     0.375 |     0.000 |     0.062 |           | 
## -------------|-----------|-----------|-----------|-----------|
## Column Total |        15 |        12 |         5 |        32 | 
##              |     0.469 |     0.375 |     0.156 |           | 
## -------------|-----------|-----------|-----------|-----------|
## 
## 
```

```
print("The most common car type in this data set is car with 4 cylinders and 4 gears. There are total of _8 cars be
```

```
## [1] "The most common car type in this data set is car with 4 cylinders and 4 gears. There are total of _8 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.

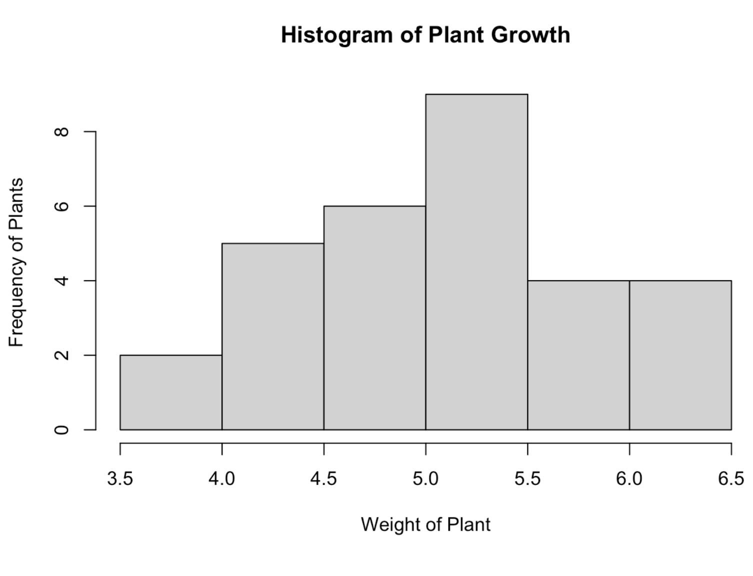
```
# Load the data set
data("PlantGrowth")

# 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!
hist <-PlantGrowth$weight
hist(hist,
     main = 'Histogram of Plant Growth',
     xlab = 'Weight of Plant',
     ylab = 'Frequency of Plants'
     )
```

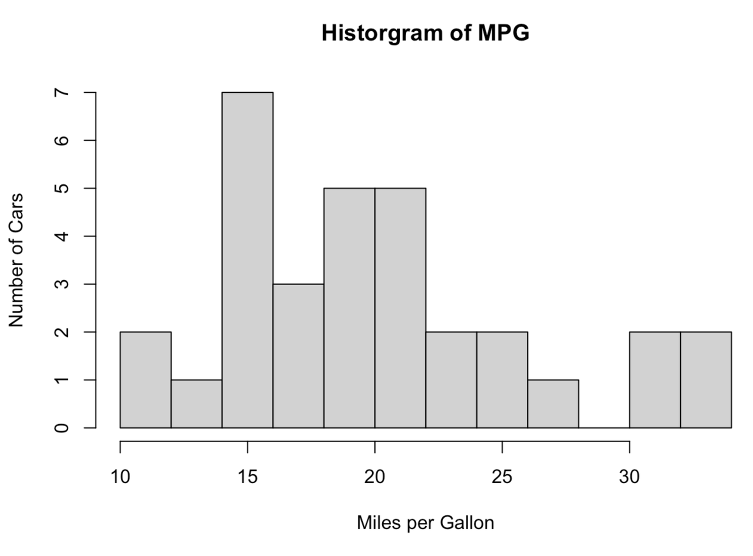


Result:

=> Report a paragraph to summarize your findings from the plot! There are more than 8 plants in between 5.0 and 5.5, with the least amount being between 3.5 and 4.0. There are no frequenceis that exceed 10, and there are none that are lower than 10.

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.

```
#library(ggplot2)
miles <- mtcars$mpg
hist(miles,
     breaks = 10,
     main = 'Historgram of MPG',
     xlab = 'Miles per Gallon',
     ylab = 'Number of Cars'
)
```




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

```
## [1] "Most of the cars in this data set are in the class of ______ 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 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 <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
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

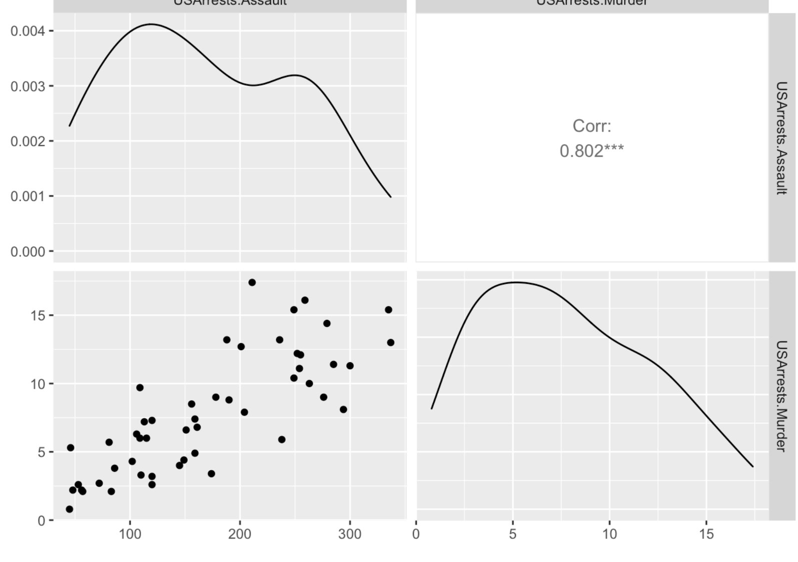
6 rows

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

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
Arrests_Frame <- data.frame(USArrests$Assault,USArrests$Murder)
ggpairs(Arrests_Frame)
```



Result:

=> Report a paragraph to summarize your findings from the plot! Murder and Assaults have a high correlation of .83. In the bottom left, more assaults mean more arrests. This support the high correlation between high murders and high assalts. If needed, one could plot linear regression on the bottom left to try to predict the number of assaults for murders and versa.

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.ggs

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

	Neighborhood <chr>	Market.Value.per.SqFt <dbl>	Boro <chr>	Year.Built <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

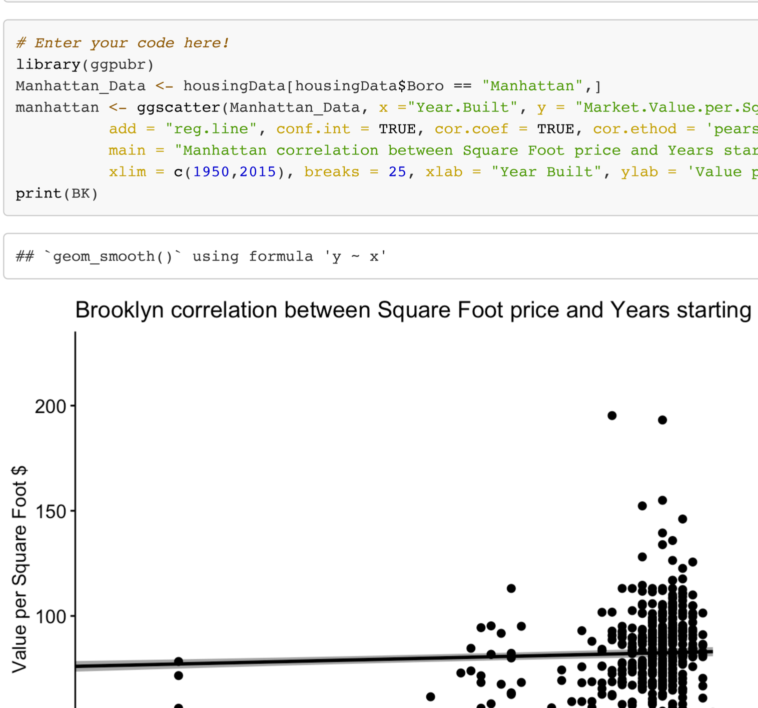
6 rows

```
# Enter your code here!
library(ggpubr)
Brooklyn_Data <- housingData[housingData$Boro == "Brooklyn",]
BK <- ggscatter(Brooklyn_Data, x ="Year.Built", y = "Market.Value.per.SqFt",
          add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.ethod = 'pearson',
          main = "Brooklyn correlation between Square Foot price and Years starting in 1980",
          xlim = c(1950,2015), breaks = 25, xlab = "Year Built", ylab = 'Value per Square Foot $')


#b. Create multiple plots to demonstrates the correlations between different variables. Remember to label all axes 
```

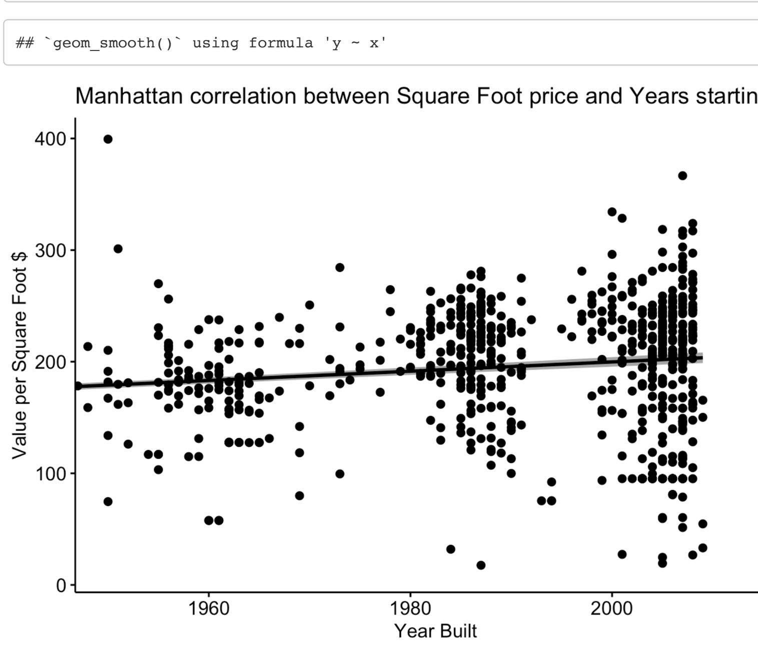
```
# Enter your code here!
library(ggpubr)
Manhattan_Data <- housingData[housingData$Boro == "Manhattan",]
manhattan <- ggscatter(Manhattan_Data, x ="Year.Built", y = "Market.Value.per.SqFt",
          add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.ethod = 'pearson',
          main = "Manhattan correlation between Square Foot price and Years starting in 1980",
          xlim = c(1950,2015), breaks = 25, xlab = "Year Built", ylab = 'Value per Square Foot $')
print(BK)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
print(manhattan)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



c. Write a summary about your findings from this exercise. Prices have strong correlation with the time built. Prices seem to be more expensive when the building is brand new. This could be due to better construction guidelines and safety standards. And the general improvement of building in infrastructure over the last decades. => Enter your answer here! Prices have strong correlation with the time built. Prices seem to be more expensive when the building is brand new. This could be due to better construction guidelines and safety standards. And the general improvement of building in infrastructure over the last decades. The outliers with old buildings having high prices could be in historical distiricts or are preserved buildings with historic value.