Unit 1 Homework

W203 Statistics for Data Science

Student Name: Youzhi Wu

Section Number: 05

Exercises

1.0: Load the dataset found in the file, cars.csv.

In [2]:
```
cars = read.csv("cars.csv")
```

2.0: What are the variables in the file?

In [3]:
```
head(cars)
```

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

The variables in the file includes: mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, and carb

3.0 Find the mean, median, minimum, maximum, 1st quartile and 3rd quartile for the mpg variable.

In [5]:
```
summary(cars)
```

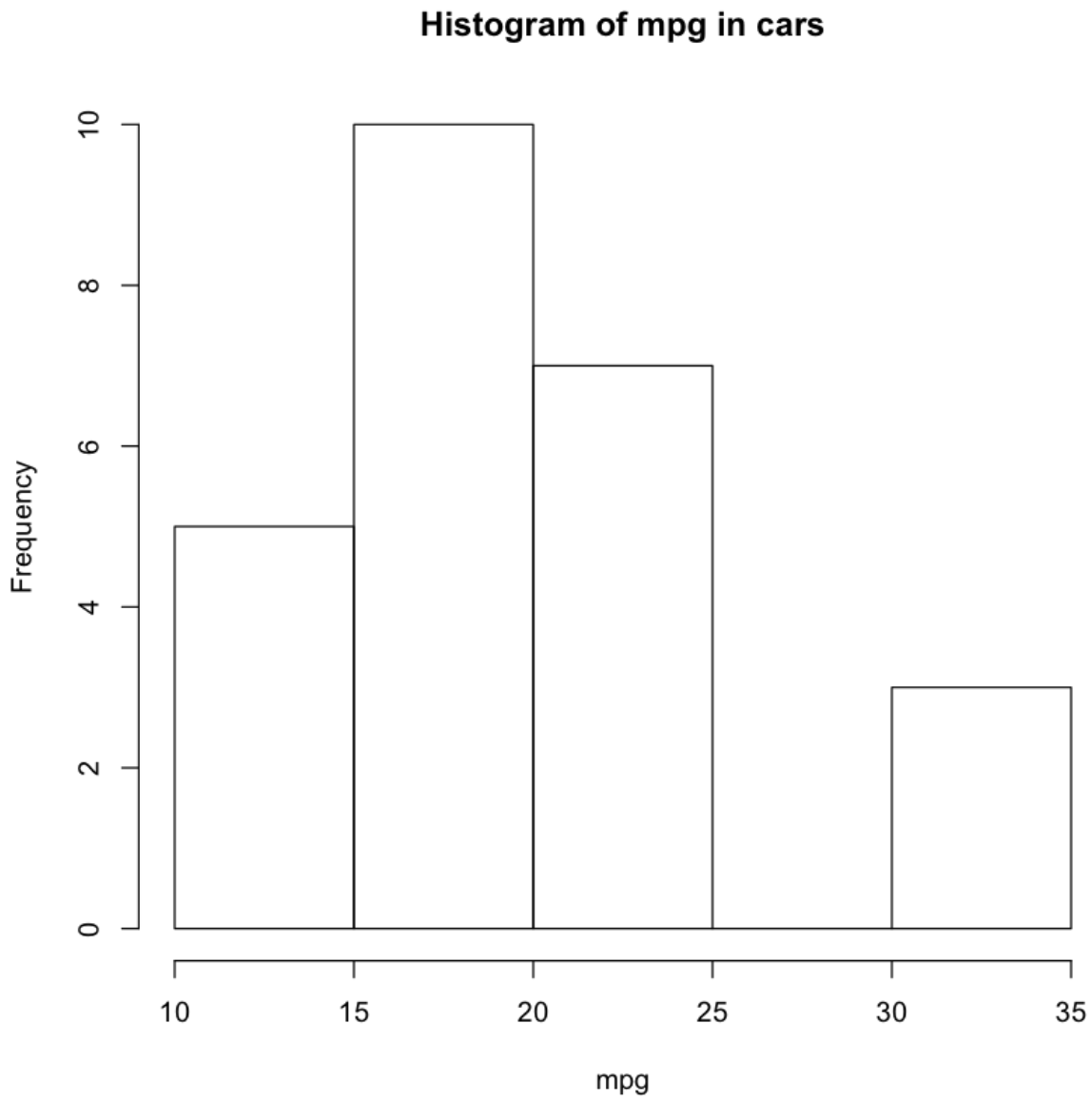
```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.20   1st Qu.:4.000   1st Qu.:146.7   1st Qu.: 97.0  
 Median :18.70   Median :6.000   Median :258.0   Median :123.0  
 Mean   :19.49   Mean   :6.261   Mean   :246.8   Mean   :141.6  
 3rd Qu.:21.50   3rd Qu.:8.000   3rd Qu.:350.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :245.0  
                 NA's   :2                                      
      drat             wt             qsec             vs              
am
 Min.   :2.760   Min.   :1.615   Min.   :15.41   Min.   :0.00   Min.   
:0.00
 1st Qu.:3.070   1st Qu.:2.875   1st Qu.:17.05   1st Qu.:0.00   1st 
Qu.:0.00
 Median :3.230   Median :3.440   Median :17.98   Median :0.00   Medi
an :0.00
 Mean   :3.493   Mean   :3.403   Mean   :18.22   Mean   :0.44   Mean   
:0.24
 3rd Qu.:3.900   3rd Qu.:3.780   3rd Qu.:19.44   3rd Qu.:1.00   3rd 
Qu.:0.00
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.00   Max.   
:1.00

      gear            carb      
 Min.   :3.0   Min.   :1.0  
 1st Qu.:3.0   1st Qu.:2.0  
 Median :3.0   Median :2.0  
 Mean   :3.4   Mean   :2.6  
 3rd Qu.:4.0   3rd Qu.:4.0  
 Max.   :4.0   Max.   :4.0  
```

Mean :19.49
Median :18.70
Min. :10.40
Max. :33.90
1st Qu.:15.20
3rd Qu.:21.50

4.0: Create a histogram of the mpg variable.

In [7]:
```
hist(cars$mpg, main = "Histogram of mpg in cars", xlab = "mpg")
```



5.0: What is the standard deviation of mpg variable?

In [8]:
```
sd(cars$mpg)
```

6.04744574179876

> The standard deviation of mpg variable is 6.04744574179876.

6.0: What is the variance of mpg variable?

In [9]:
```
var(cars$mpg)
```

36.5716

> The variance of mpg variable is 36.5716.

7.0: What is the relationship of the standard deviation to the variance? Why does the standard deviation and variance of the mpg variable differ?

> The standard deviation is the square root of the variance.
>
> - The standard deviation is expressed in the same units as the mean is;
> - The variance is expressed in squared units of the mean.

8.0 How many data points are there for the cyl variable?

In [12]:
```
length(cars$cyl)
```

25

> The total data points of cyl variable is 25. There are 2 data points having NA values in cyl variable. Excluding the two NA points, there are 23 data points with valid values.

9.0 What is the mean of the cyl variable?

In [16]:
```
# Calculate the mean of cyl variable for all 25 data points
cyl_mean = mean(cars$cyl)

# Calculate the mean of cyl variable for the 23 data points, excluding
two data points with NA values
cyl_grand_mean = mean(cars$cyl, na.rm = TRUE)
```

6.26086956521739

The mean of cyl variable (excluding the two data points with NA value) is 6.26086956521739.