Unit 1 Homework 1

Statistics for Data Science

Unit 1 Homework 1

Exercise

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

```
cars = read.csv("~/Google Drive/_UCB_W203/_W203_2017Fall/Homework/Unit01_Homework1/cars.csv")
```

1. What are the variables in the file?

```
colnames(cars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

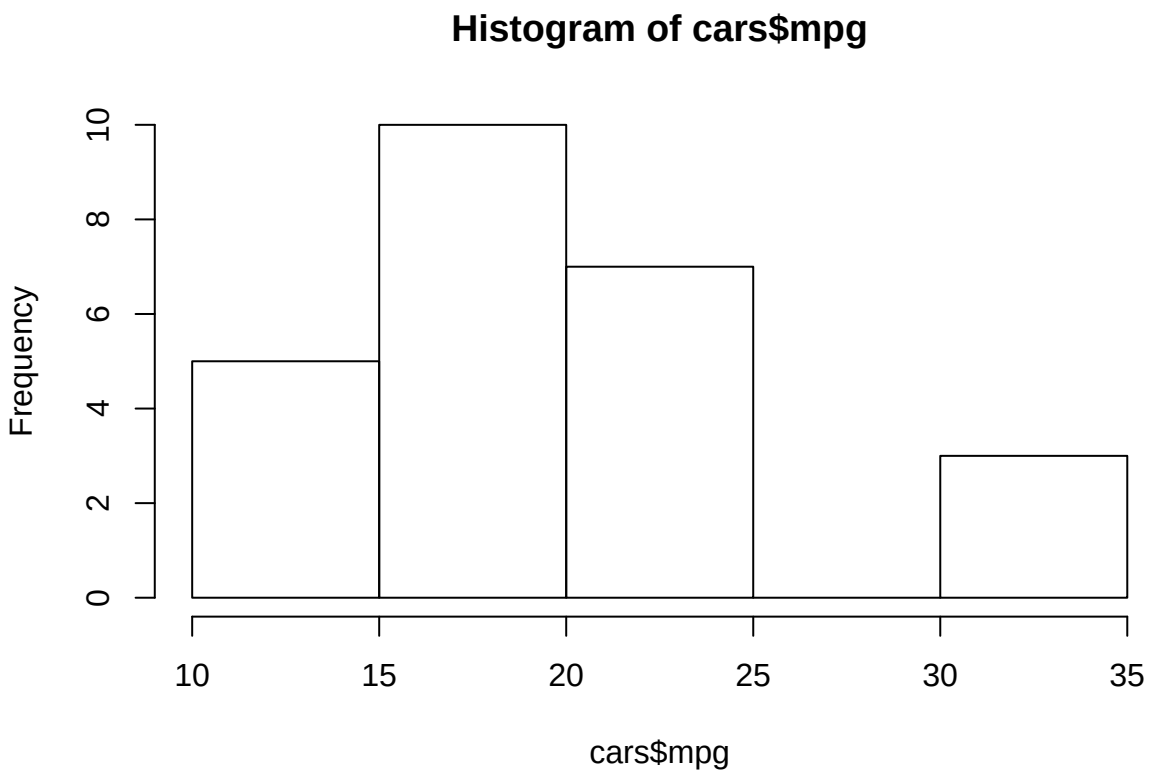
2. Find the mean, median, minimum, maximum, 1st quartile and 3rd quartile for the mpg variable.

```
summary(cars$mpg)
```

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

3. Create a histogram of the mpg variable.

```
hist(cars$mpg)
```



4. What is the standard deviation of mpg variable?

```
sd(cars$mpg)
```

```
## [1] 6.047446
```

5. What is the variance of mpg variable?

```
var(cars$mpg)
```

```
## [1] 36.5716
```

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

```
sd(cars$mpg)==var(cars$mpg)** 0.5
```

```
## [1] TRUE
```

By definition, standard deviation of a variable is the square root is the variance.

7. How many data points are there for the cyl variable?

```
length(cars$cyl[!is.na(cars$cyl)])
```

```
## [1] 23
```

Using '!' in front of the is.na() function negates the function (in other words, it gives us the variables that are not NA) and the length() function counts the number of data points in a given vector. We can combine these two functions to first find the values that are not NA and then counting the number of (non NA) values.

8. What is the mean of the cyl variable?

```
mean(cars$cyl, na.rm = TRUE)
```

```
## [1] 6.26087
```