Homework Exercise 1

Adam Yang

W203 Statistics for Data Science

Unit 1 Homework

Exercise

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

```
Cars <- read.csv("cars.csv")
```

1. What are the variables in the file?

```
names(Cars)
```

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

The variables/columns in the file are mpg, cyl, disp, hp, drat, wt, qsec, vs, am, gear, and carb.

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

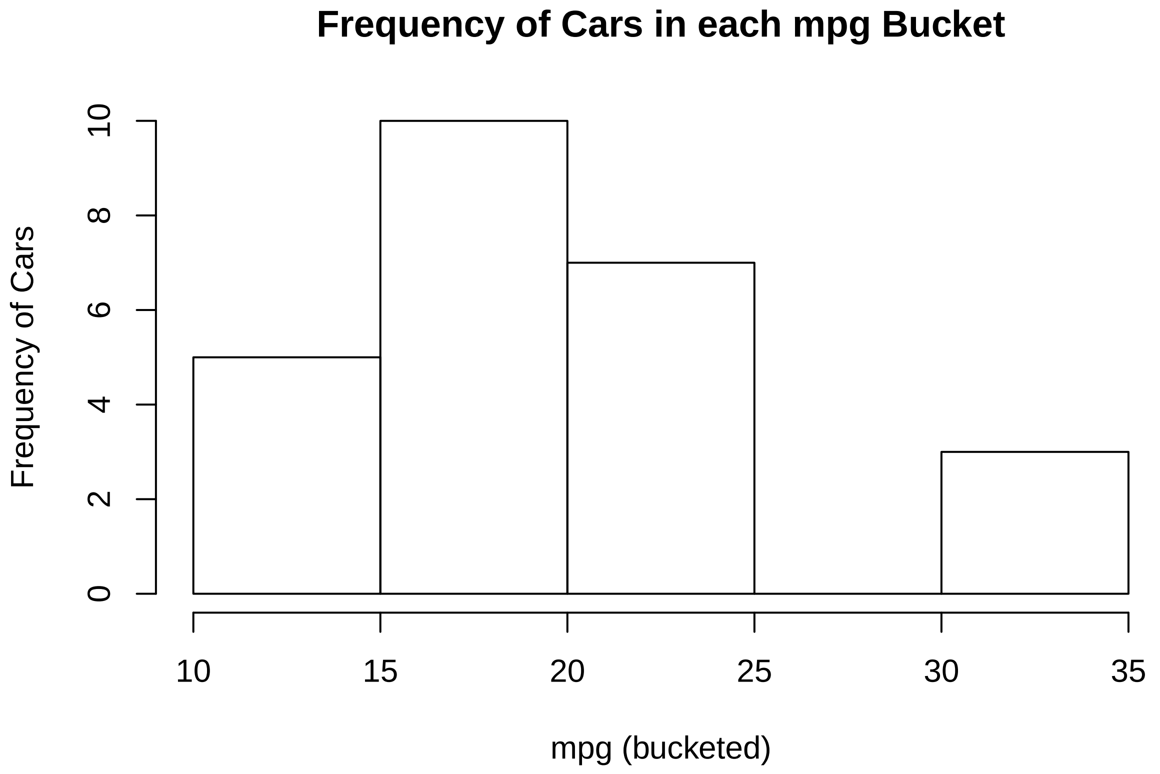
```
summary(Cars$mpg)
```

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

For the mpg variable, the mean = 19.49, the median = 18.70, the min = 10.40, the max = 33.90, the 1st quartile = 15.20, and the 3rd quartile = 21.50.

3. Create a histogram of the mpg variable.

```
hist(Cars$mpg,
     main = "Frequency of Cars in each mpg Bucket",
     xlab = "mpg (bucketed)",
     ylab = "Frequency of Cars"
)
```



4. What is the standard deviation of mpg variable?

```
sd(Cars$mpg)
```

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

The standard deviation of the mpg variable is 6.047446

5. What is the variance of mpg variable?

```
var(Cars$mpg)
```

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

The variance of the mpg variable is 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$mpg)^2 == var(Cars$mpg)
```

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

As shown above, the variance is the square of the standard deviation. The variance gives us an idea of how much the data points, on average, deviate from the mean. It emphasizes the outliers of the sample a bit more than the data near the mean which can be useful to understanding the data sample. The standard deviation gives us an idea of how much each data point deviates from the mean. An important difference is that the standard deviation has the same unit as the data points while the variance does not.

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

```
summary(Cars$cyl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   4.000   4.000   6.000   6.261   8.000   8.000       2
```

```
NROW(na.omit(Cars$cyl))
```

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

Looking at the summary of the cyl variable, we see that ther are 2 NA's. Therefore, if we want to count the number of data points in the cyl variable, we need to omit the NA's. We end up with 23 data points in cyl.

8. What is the mean of the cyl variable?

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

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

The mean of the cyl variable is 6.26087.