Project 1: Data Visualization and Exploratory Data Analysis

Jose Lazarte

September 14, 2019

Computing Arithmetic

To begin this project, basic calculations and functions will follow:

12 plus 55 is equal to 67

56 divided by 12 is equal to 4.6666667

13 to the power of 4 is equal to 2.8561×10^4

By using the c() vector function and the seq() function we can aquire:

```
ariArray <- c(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)


ariSeq <- seq(from=1, to=20, by=1)
```

The values of the previous vector and seq are as follows:

```
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
```

Descriptive Statistics

Utilizing the *"Auto MPG Data Set"* available on the UCI Machine Learning Repository, the following calculations were completed:

Mean and Standard Deviation

	Category	Mean Value	Standard Deviation
1	MPG	23.515	7.816
2	Acceleration	15.568	2.758
3	Horsepower	104.469	38.491
4	Displacement	193.426	104.270
5	Weight	2970.425	846.842

Cylinders Analysis

The mean of Cylinders can be found via $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, which results in an average of:

```
[1] 5.454774
```

However, this value does not provide a meaningful result as the average number of cylinders in the data set does not represent a connectino with other values. Rather than utilizing the mean of the cylinders I beleive that the frequency or mode of the cylinders will provide a deeper understanding of the data set. The amount of cylinders varies from car to car, therefore, cylinders should be used as a way to further categorize the other recorded values. This will be experienced later in the project when comparing a scatter plot of MPG to cylinders.

5-Number Summary

The following tables refer to the 5-Number Summary of the *"Auto MPG Data Set"*:

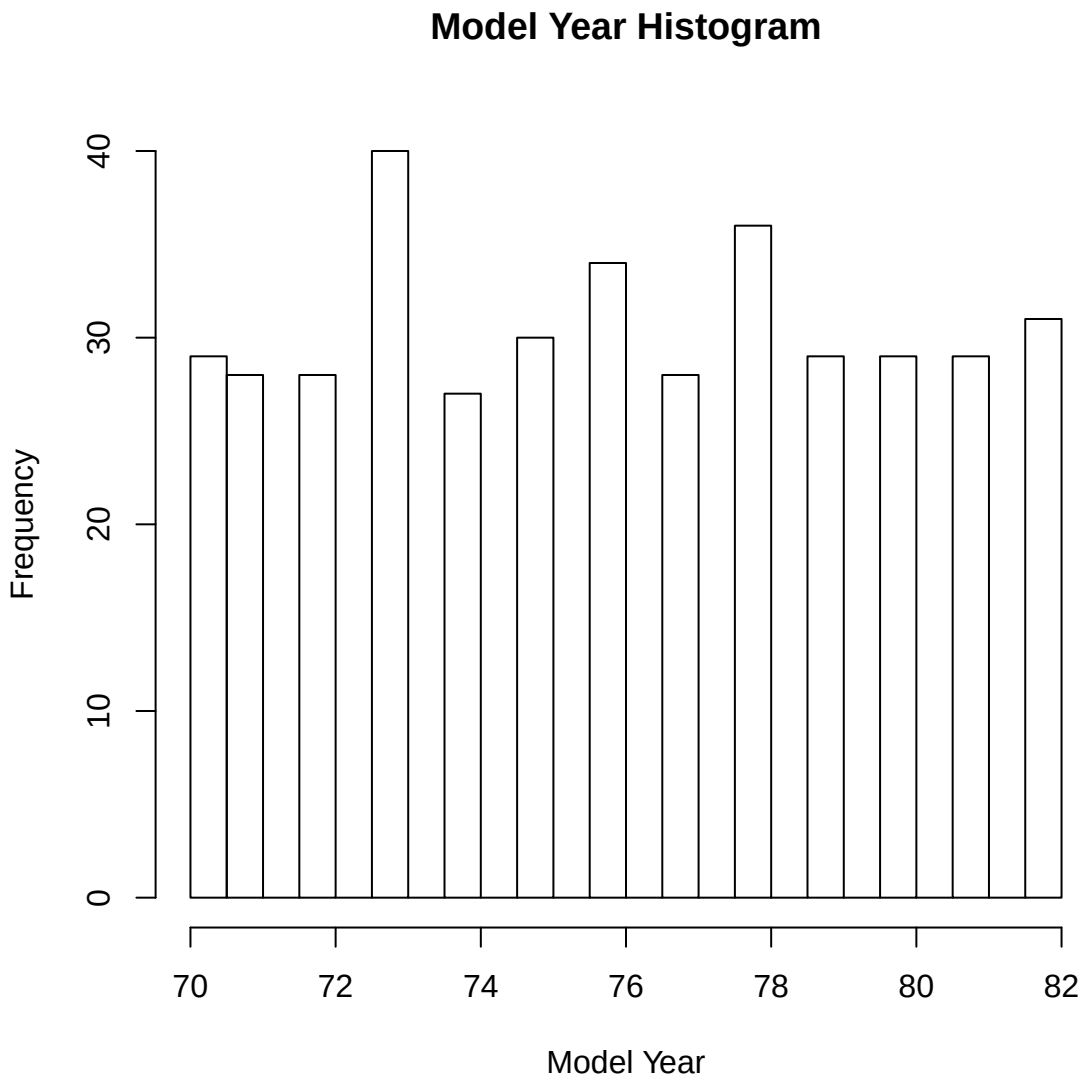
	N	Missing	Mean	SD	Min	Q1	Median	Q3	Max
MPG	398	0	23.51	7.82	9.00	17.50	23.00	29.00	46.60
Cylinders	398	0	5.45	1.70	3.00	4.00	4.00	8.00	8.00
Displacement	398	0	193.43	104.27	68.00	104.00	148.50	262.00	455.00
Horsepower	392	6	104.47	38.49	46.00	75.00	93.50	127.00	230.00
Weight	398	0	2970.42	846.84	1613.00	2223.00	2803.50	3609.00	5140.00
Acceleration	398	0	15.57	2.76	8.00	13.80	15.50	17.20	24.80
ModelYear	398	0	76.01	3.70	70.00	73.00	76.00	79.00	82.00
Origin	398	0	1.57	0.80	1.00	1.00	1.00	2.00	3.00

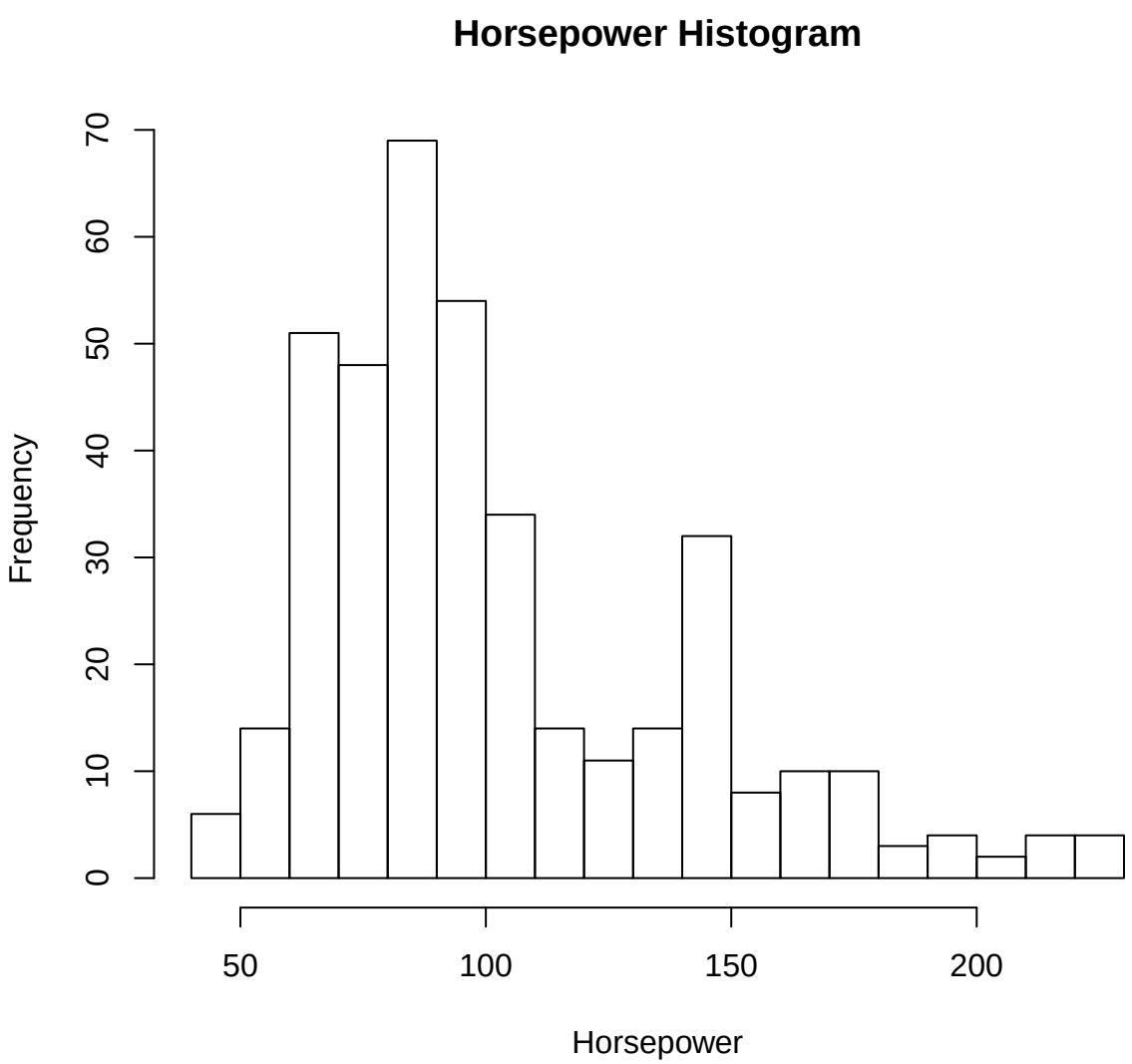
Utilizing papeR library we are able to take a 5-number summary of the numerical data

Data Visualization

Utilizing the *"Auto MPG Data Set"* we will now visualize and analyze the data with the use of graphs and other tools.

Histograms for Car Year and Car Horsepower

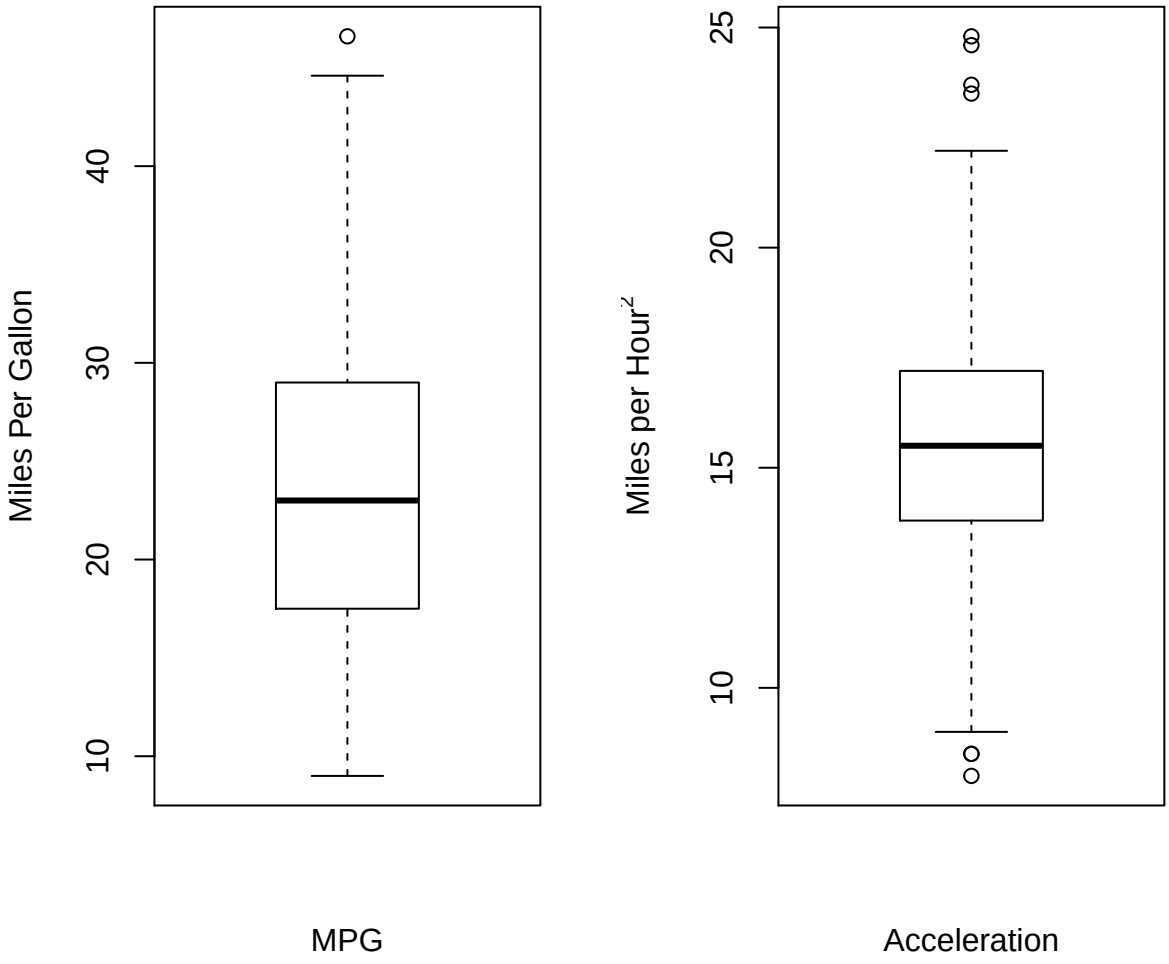




The first histogram represents the year the car was made and how frequent they occur in the data set. With a bin number of 20, the data takes a uniform shape with increases in frequency near the 78th and 73rd division. This tells us that for the most part, the year of the cars is very symmetrical and uniform.

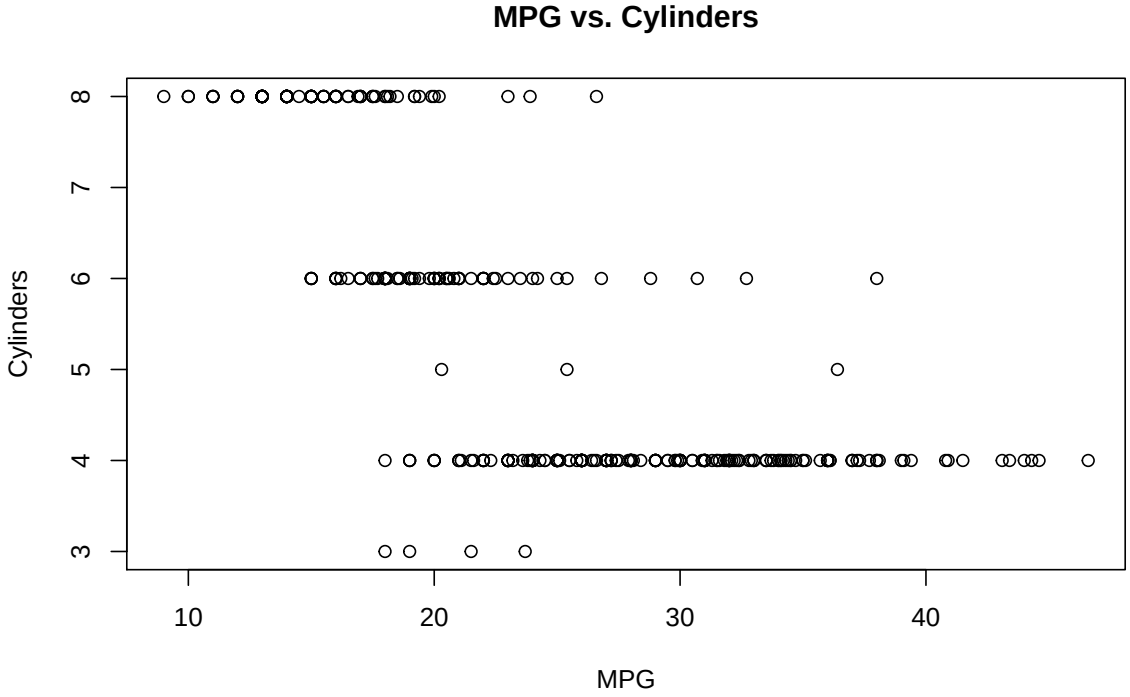
Meanwhile, the second histogram represents horsepower the cars achieved and how frequently certain numbers appeared. Same as the last histogram, a bin number of 20 was used, however, the histogram shows that the data is right-skewed. This can mean that the horsepower across the data set frequently occurs at 80th to 90th division. There may be outliers as the histogram shows that the data has reached around the 220th division, therefore, it may skew the data but further analysis is required.

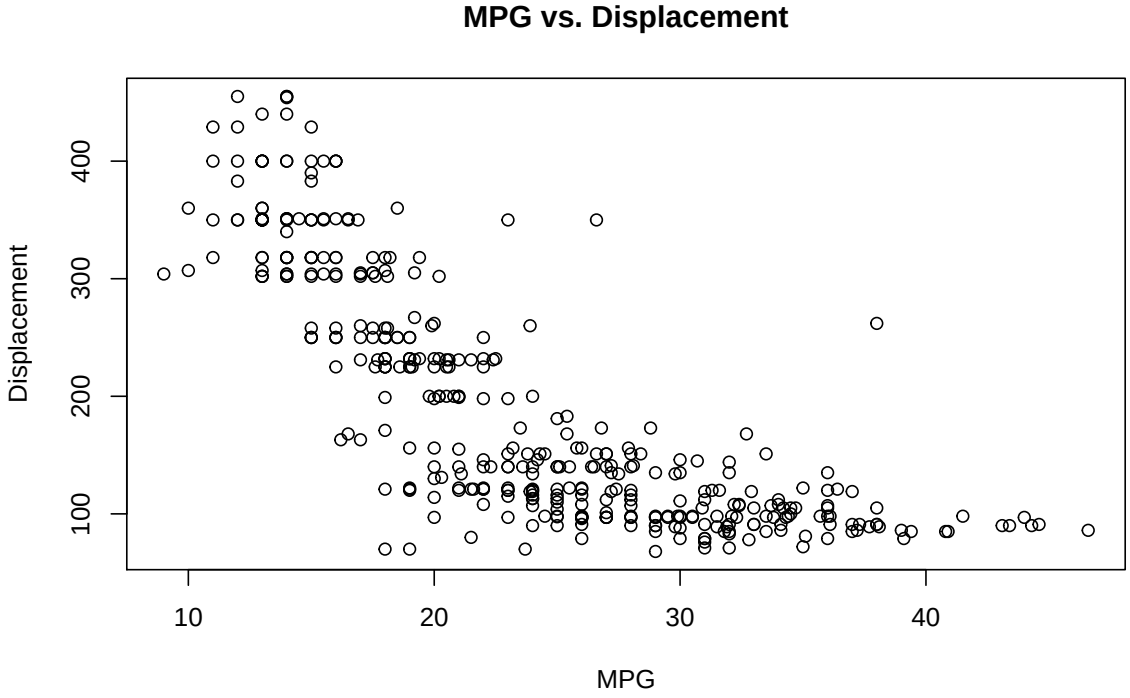
Box-and-Whisker Plot of MPG and Acceleration

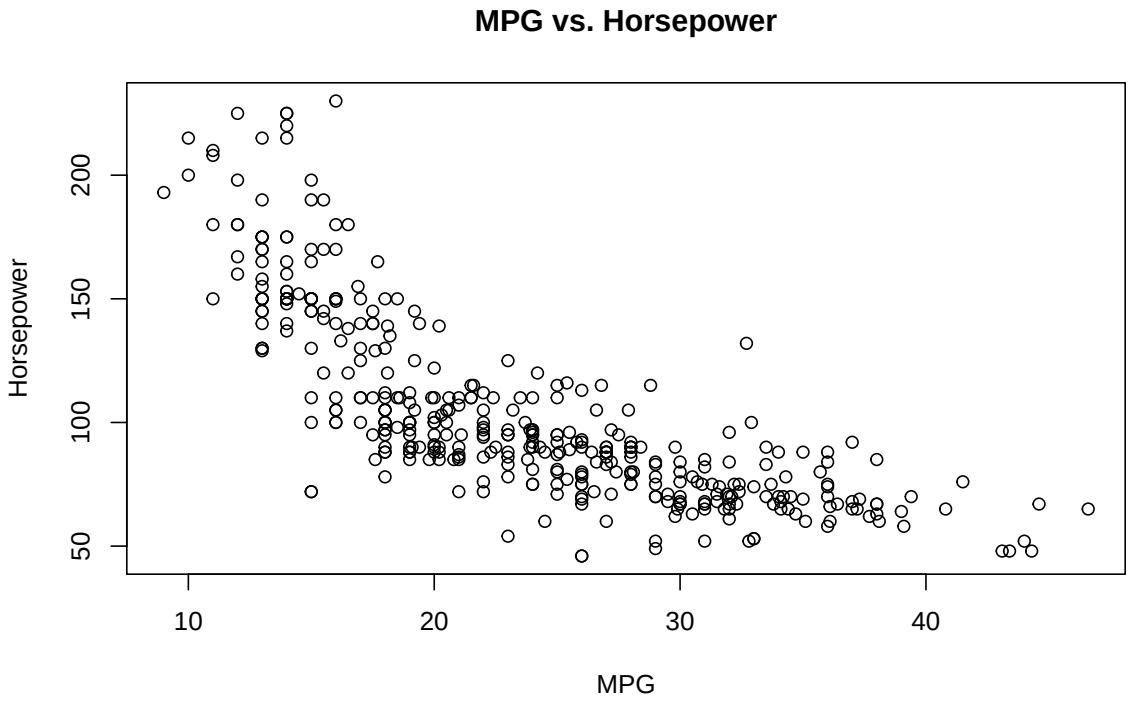


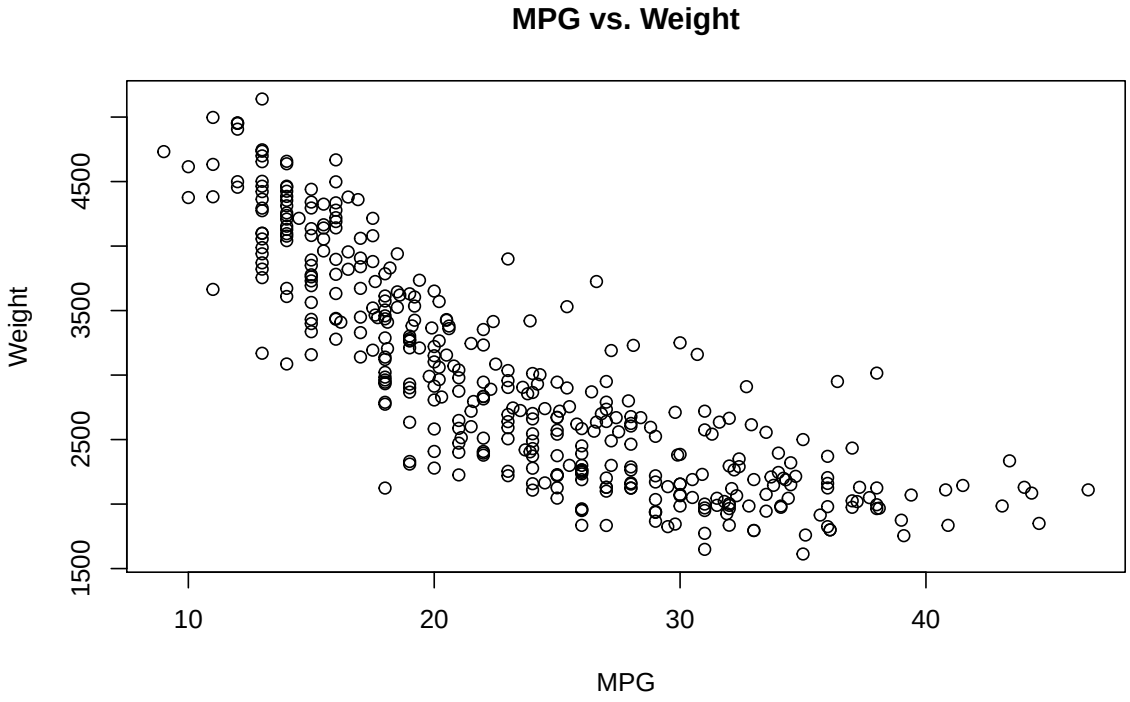
These box-and-whisker plots represent Miles Per Gallon and Acceleration, each is unique in their own way. A key difference is that the acceleration box-and-whisker has many outliers at both ends of the whiskers. This could mean that the sample size was not sufficient or too large for acceleration. It could also mean that many other categories in the data set can have outliers that cannot be seen through the current methods.

Scatterplot Analysis of MPG to the data

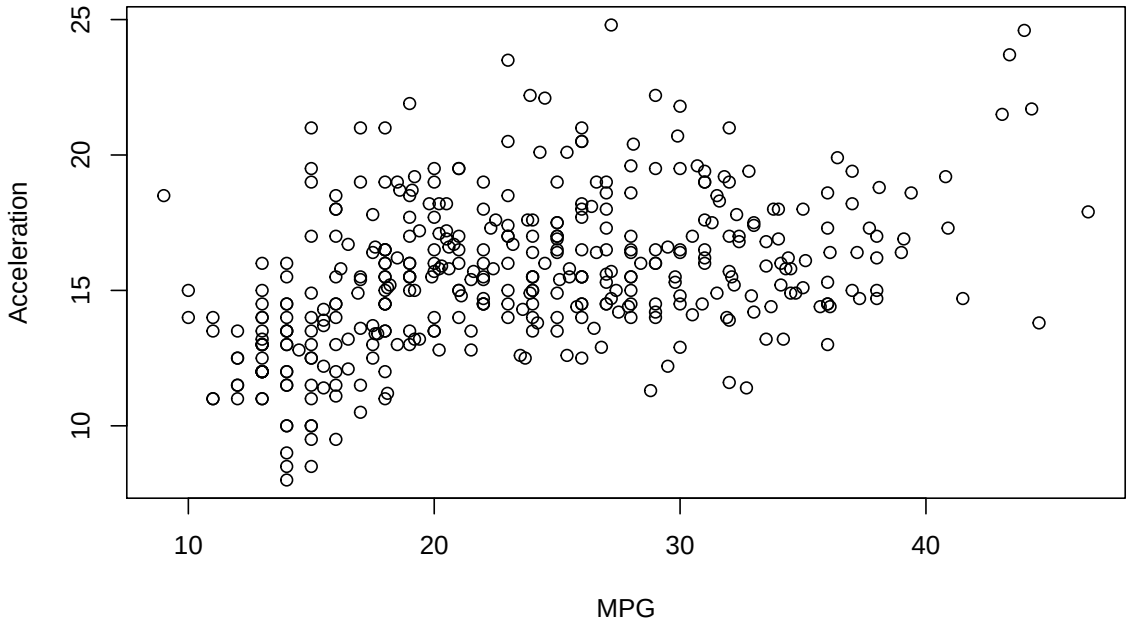




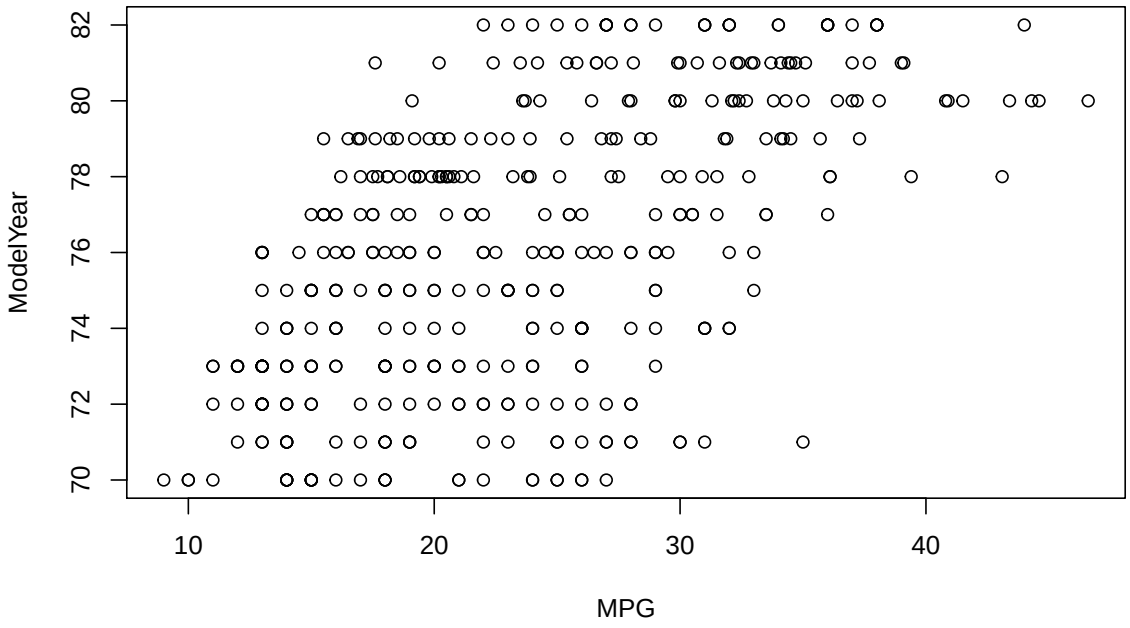

MPG vs. Horsepower
Horsepower
50
100
150
200
10
20
30
40
MPG

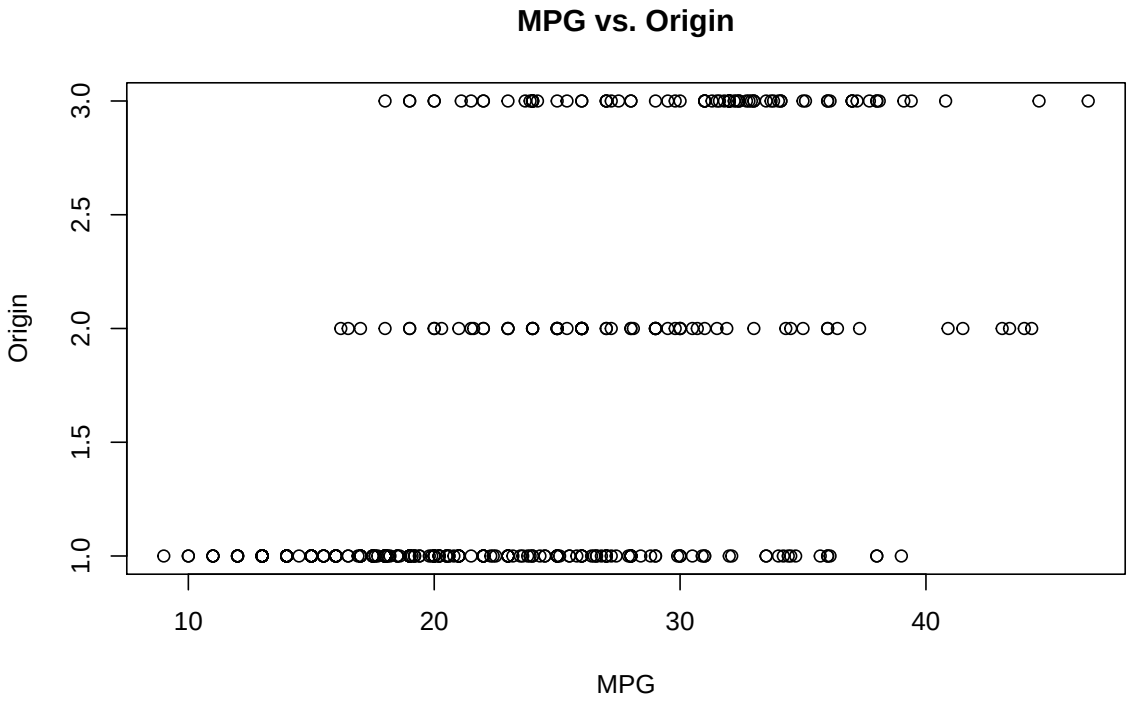

MPG vs. Weight
Weight
1500
2500
3500
4500
10
20
30
40
MPG

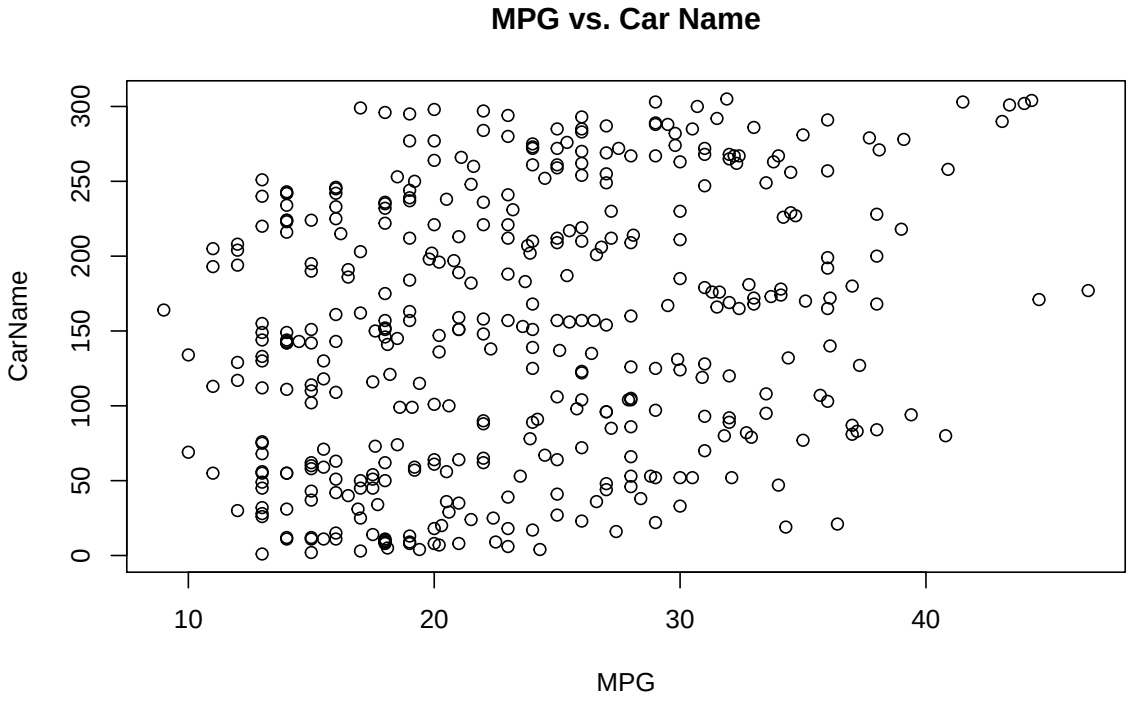
MPG vs. Acceleration



MPG vs. Model Year




MPG vs. Origin
Origin
MPG
1.0
1.5
2.0
2.5
3.0
10
20
30
40


MPG vs. Car Name
CarName
MPG
0
50
100
150
200
250
300
10
20
30
40

The scatter plots that were just presented all draw the relation between Miles per Gallon and other categories of the data.

The first scatter plot represents Miles Per Gallon vs Cylinders showing a negative correlation between the number of cylinders and mpg. In other words the scatter plot is telling us that miles per gallon improve as the number of cylinders go down. Although the data follows a linear trend, the visual can be improved by separating the number of cylinders into groups and further analysing the relation.

The next scatter plot is Miles Per Gallon vs Displacement, similiar to MPG vs Cylinders, the graph follows a negative trend meaning that the miles per gallon improve as the displacement goes down. One of the key differences is that the grouping of data points is much more closer and therefore some outliers can be seen. The graph also follows a primarily nonlinear trend meaning that further analysis can provide different results.

The following two graphs have very similiar shape and spread, that is, a negative nonlinear trend with data points closer then the past two graphs. These graphs represent MPG vs Horsepower and MPG vs Weight, in other words, in order to improve miles per gallon, one must have a lower horsepower or a lighter vehicle. As these data points are a lot more concentrated than the last two graphs the correlation between MPG, Horsepower, and Weight must be strong.

The following scatter plot is titled MPG vs Acceleration and it has a positive linear trend. Although many of the data points in this graph are spread apart the relation between miles per gallon and acceleration must be recognized, that is to say, that in order to acquire high miles per gallon one must have higher acceleration. This correlation is not necceseraly true as the data is weak since many of the points are spread far apart.

Similiar to the last graph, MPG vs Model Year shares a low positive linear correlation. In other words, the higher the miles per gallon the newer the car should be. However, this is not necessarily true as the data points have a weak relationship.

MPG vs Origin can be assimilated to MPG vs Cylinders when it comes to the weak relationship. However, this graph follows a low positive linear trend. Similiar to the first scatter plot, the Origin category is better utilized as a method of filtering and separating data rather than as a way to directly compare data. For instance, separating the data into three seperate box-and-whisker plots will show a clearer visual of the data and the relation of the origin and MPG. Although weak, this scatter plot is tellings us that a higher MPG can be found in cars of higher origin, however, the correlation is weak and further analysis is needed.

The final scatter plot represents MPG vs Car Names, unlike any of the previous visuals this one does not have a relationship with MPG, therefore, the scatter plot is neither linear or nonlinear and cannot be positive

or negative.