Auto Miles-per-Gallon Predictive Modeling

Phase I: Data Summarization & Exploration

Akshay Sharma [redacted] *& Vishesh Jain* [redacted]

08 April 2018

Contents

1 INTRODUCTION

The objective of this project is to build a model to predict miles-per-gallon of a car. The data set was sourced from the UCI Machine Learning Repository. This project has two phases. Phase I presents the data summmarization and exploration which is covered in this report and Phase II will present the model building and predictions. This Report has been divided into sections which explains the Data Pre-processing methods, Data Exploration through visualisation and finally a Summary.

2 DATA SET

The UCI Machine Learning Repository provides two data sets, but only `auto-mog.data`, and `auto-mpg.names` were useful in this project. `auto-mpg.names` contains the details of attributes or variables. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition. This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file `auto-mpg.data-original`.The data set contains 398 instances, 8 desciptive features and 1 target feature.

2.1 Target Feature

The response feature in `mpg` *(Miles-per-Gallon)* which is a continuous numerical variable. The goal is to predict the numerical value of miles-per-gallon.

2.2 Descriptive Feature

The variable description is produced here from `auto-mpg.names` file:

- cylinders: multi-valued discrete
- displacement: continuous
- horsepower: continuous
- weight: continuous
- acceleration: continuous
- model year: multi-valued discrete
- origin: multi-valued discrete
- car name: string (unique for each instance)

Most of the descriptive features are self-explanatory.

3 DATA PREPROCESSING

3.1 Preliminaries

In this project we will use the following packages

```
library(knitr)
library(mlr)
library(GGally)
library(data.table)
library(dplyr)
library(ggplot2)
library(MASS)
library(forcats)
library(wesanderson)
```

3.2 Data Import

```
#Importing the dataset through URL
ampg.file <-
  fread("http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
        sep = "\t", header = FALSE)

#Adding the text column to a vector using textConnections()
con <- textConnection(ampg.file$V1)
ampg.2 <- read.table(con)

#Adding the remaining column from our data
ampg.2$newcol <- ampg.file$V2

#Giving column names to our data frame

colnames(ampg.2) <- c("mpg", "cylinders","displacement","horsepower","weight",
                      "acceleration","modelYear","origin","carName")
```

3.3 Data Cleaning & Transformation

With `str` and `summarizeColumns` (see Table 1), we noticed the following anomalies:

- There are 6 missing value in `Horsepower`
- The feature `cylinder` is given as a numeric feature
- `modelYear` and `Origin` are also taken as numeric while these can be taken as factor.
- `Horsepower` is taken as a factor as there are some non-numeric values present in the dataset.
- `modelYear` is given in numbers such as 70, 71 to 82.

```
str(ampg.2)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : Factor w/ 94 levels "?","100.0","102.0",..: 17 35 29 29 24 42 47 46 48 40 ...
```

```
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ modelYear   : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ carName     : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebe
```

```
summarizeColumns(ampg.2) %>% knitr::kable( caption =  'Feature Summary before Data Preprocessing')
```

Table 1: Feature Summary before Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
mpg	numeric	0	23.514573	7.8159843	23.0	8.89560	9	46.6	0
cylinders	integer	0	5.454774	1.7010042	4.0	0.00000	3	8.0	0
displacement	numeric	0	193.425879	104.2698382	148.5	86.73210	68	455.0	0
horsepower	factor	0	NA	0.9447236	NA	NA	1	22.0	94
weight	numeric	0	2970.424623	846.8417742	2803.5	945.15750	1613	5140.0	0
acceleration	numeric	0	15.568091	2.7576889	15.5	2.52042	8	24.8	0
modelYear	integer	0	76.010050	3.6976266	76.0	4.44780	70	82.0	0
origin	integer	0	1.572864	0.8020549	1.0	0.00000	1	3.0	0
carName	character	0	NA	0.9849246	NA	NA	1	6.0	305

3.3.1 Data Type Conversion

First, lets convert the `horsepower` column into factor and remove the non numeric value denoted by '?'.

```
ampg.2$horsepower <- as.numeric(as.character(ampg.2$horsepower))
```

Second, converting `Cylinders` & `Origin` to factor variables.

```
#as.factor() method is used for conversion, it takes unique values in the variable as levels

ampg.2$cylinders <- as.factor(ampg.2$cylinders)
ampg.2$origin <- as.factor(ampg.2$origin)
```

Third, adding 1900 to the values of `modelYear` for better presentation and converting the same to factor.

```
ampg.2$modelYear <- ampg.2$modelYear + 1900
ampg.2$`modelYear` <- as.factor(ampg.2$`modelYear`)
```

We computed the level table for each factor column. The tables showed:

- We have only 4 number of instances for cars with `3 cylinders`
- Only 3 number of instances for cars with `5 cylinders`
- We have approximately equal number of instances for each `modelYear`
- There are more number of instances from `origin 1` as compared to `origin 2 & 3` combined

```
sapply( ampg.2[ sapply(ampg.2, is.factor)], table)
```

```
## $cylinders
##
##   3   4   5   6   8
##   4 204   3  84 103
##
## $modelYear
##
## 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982
##   29   28   28   40   27   30   34   28   36   29   29   29   31
##
## $origin
##
##   1   2   3
## 249  70  79
```

Looking at the `car names`, a new column can be included in the data set that can give the *company name* information. This information can further be used to analyse if the `MPG` is effected by the manufacturing company or not.

```
ampg.2$company_name <- gsub("([A-Za-z]+).*", "\\1", ampg.2$`carName`)

#Checking the unique values in the new column
unique(ampg.2$company_name)
```

```
##  [1] "chevrolet"  "buick"      "plymouth"   "amc"        "ford"
##  [6] "pontiac"    "dodge"      "toyota"     "datsun"     "volkswagen"
## [11] "peugeot"    "audi"       "saab"       "bmw"        "chevy"
## [16] "hi"         "mercury"    "opel"       "fiat"       "oldsmobile"
## [21] "chrysler"   "mazda"      "volvo"      "renault"    "toyouta"
## [26] "maxda"      "honda"      "subaru"     "chevroelt"  "capri"
## [31] "vw"         "mercedes"   "cadillac"   "vokswagen"  "triumph"
## [36] "nissan"
```

Looking at the above results, following changes can be done:

- Replace maxda with mazda
- Replace chevroelt and chevy with chevrolet
- Replace vokswagen and vw with volkswagen
- Replace toyouta with toyota

```
ampg.2$company_name <- ifelse(ampg.2$company_name=='vw' |
                                ampg.2$company_nam=="vokswagen","volkswagen",
                              ampg.2$company_name)
ampg.2$company_name <- ifelse(ampg.2$company_name=="chevroelt" |
                                ampg.2$company_nam=="chevy","chevrolet",
                              ampg.2$company_name)
ampg.2$company_name <- ifelse(ampg.2$company_name=="maxda","mazda",
                              ampg.2$company_name)
ampg.2$company_name <- ifelse(ampg.2$company_name=="toyouta","toyota",
                              ampg.2$company_name)

unique(ampg.2$company_name)
```

```
##  [1] "chevrolet"  "buick"      "plymouth"   "amc"        "ford"
##  [6] "pontiac"    "dodge"      "toyota"     "datsun"     "volkswagen"
## [11] "peugeot"    "audi"       "saab"       "bmw"        "hi"
## [16] "mercury"    "opel"       "fiat"       "oldsmobile" "chrysler"
## [21] "mazda"      "volvo"      "renault"    "honda"      "subaru"
## [26] "capri"      "mercedes"   "cadillac"   "triumph"    "nissan"
```

Now converting the given column into a factor with 31 levels.

```
ampg.2$company_name <- as.factor(ampg.2$company_name)
```

3.4 Handling Missing Values

We observed above that there are 6 missing values in horsepower. As we have 95% of our dataset as complete, we will remove the rows with missing values.

Doing so, we will loose only 1.5% of the data which will be below our threshold of 5%.

```
ampg.2 <- ampg.2 %>% filter(!is.na(horsepower))
```

3.5 Summarization

Taking the summarization after data processing.

```
summarizeColumns(ampg.2) %>% knitr::kable( caption =  'Feature Summary After Data Preprocessing')
```

Table 2: Feature Summary After Data Preprocessing

name	type	na	mean	disp	median	mad	min	max	nlevs
mpg	numeric	0	23.44592	7.8050075	22.75	8.59908	9	46.6	0
cylinders	factor	0	NA	0.4923469	NA	NA	3	199.0	5
displacement	numeric	0	194.41199	104.6440039	151.00	90.43860	68	455.0	0
horsepower	numeric	0	104.46939	38.4911599	93.50	28.91070	46	230.0	0
weight	numeric	0	2977.58418	849.4025600	2803.50	948.12270	1613	5140.0	0
acceleration	numeric	0	15.54133	2.7588641	15.50	2.52042	8	24.8	0
modelYear	factor	0	NA	0.8979592	NA	NA	26	40.0	13
origin	factor	0	NA	0.3750000	NA	NA	68	245.0	3
carName	character	0	NA	0.9872449	NA	NA	1	5.0	301
company_name	factor	0	NA	0.8775510	NA	NA	1	48.0	30

4 DATA EXPLORATION

We first explore each feature individually and check their distribution with the target variable as well.

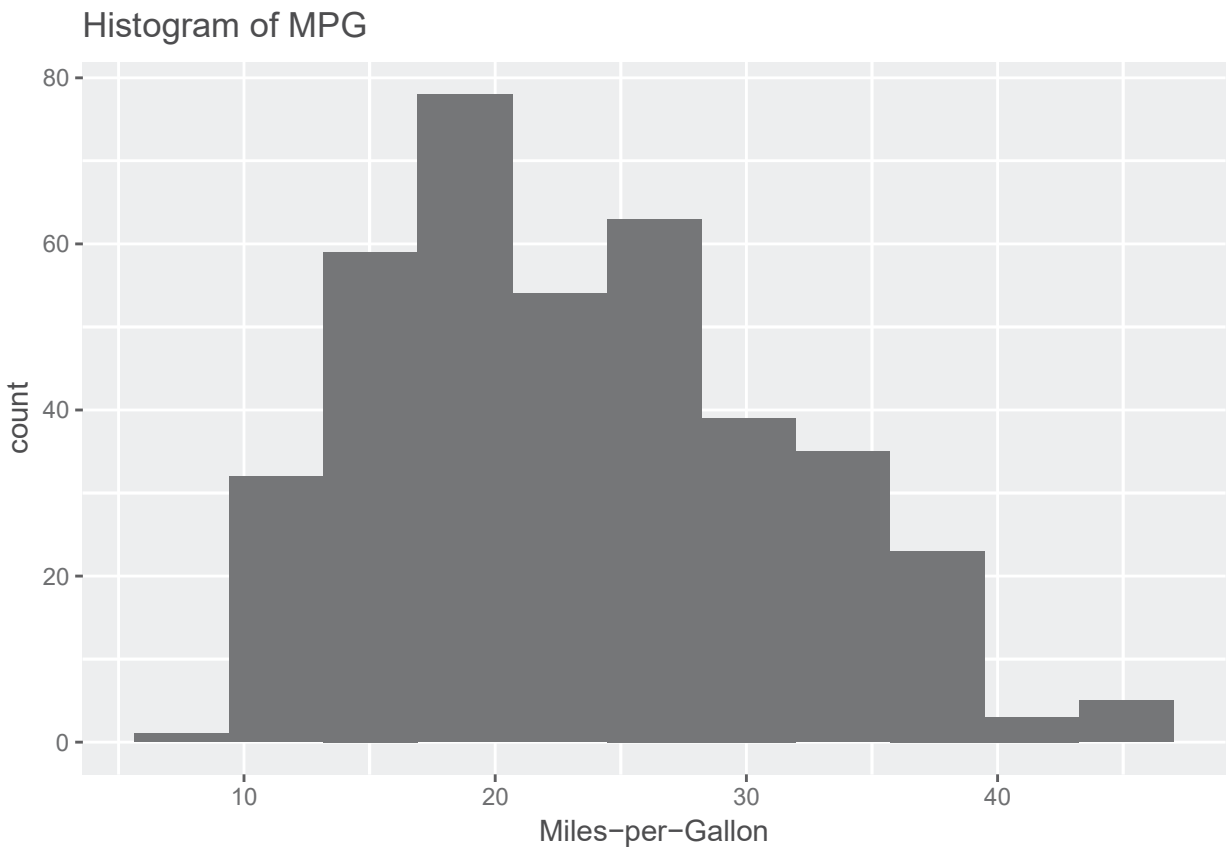
4.1 Univariate Visualisation

4.1.1 Numerical Feature

4.1.1.1 MPG - Target Feature

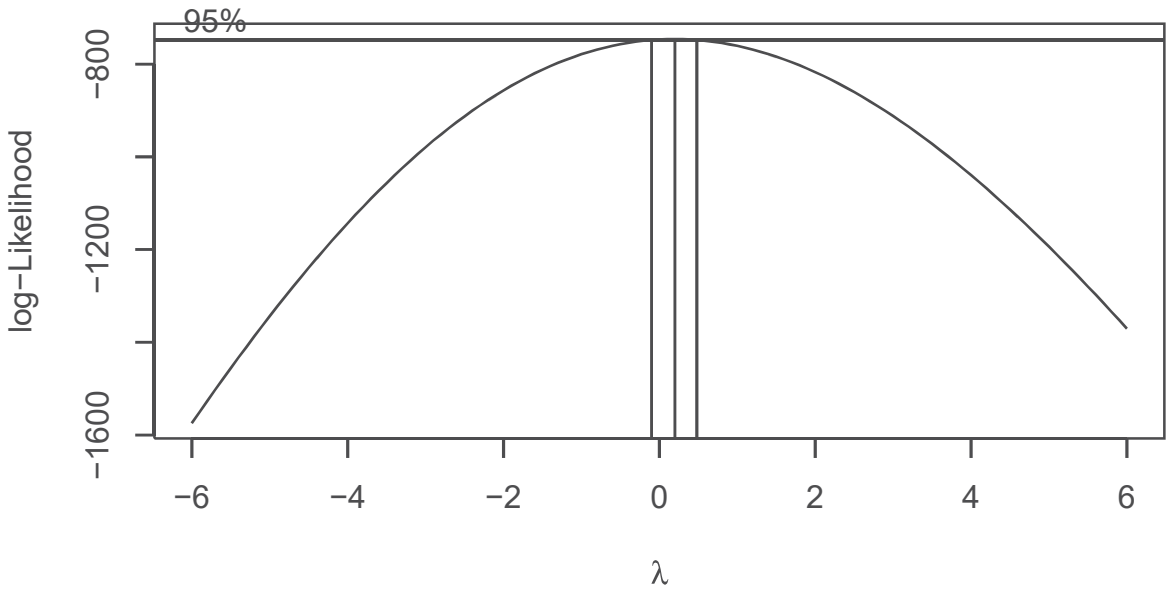
Looking at the histogram of `MPG` below, the distribution is found to be right-skewed. To make it closer to a normal distribution, we can use Box-Cox Transformation approach.

```
#Building function to calculate the binwidth

binwidth <- function(x){
  bw = diff(range(x, na.rm = TRUE)) / (2 * IQR(x, na.rm = TRUE) / length(x)^(1/3))
  return (bw)
  }

b = binwidth(ampg.2$mpg)
ggplot(ampg.2, aes(x = (mpg))) + geom_histogram(bins = b) +
  labs(title = "Histogram of MPG", x = "Miles-per-Gallon")
```



Box-Cox Method of Maximum Log-Likelihood

```
Box = boxcox(mpg ~ 1,
             data = ampg.2,
             lambda = seq(-6,6,0.1)      # Try values -6 to 6 by 0.1
             )
```

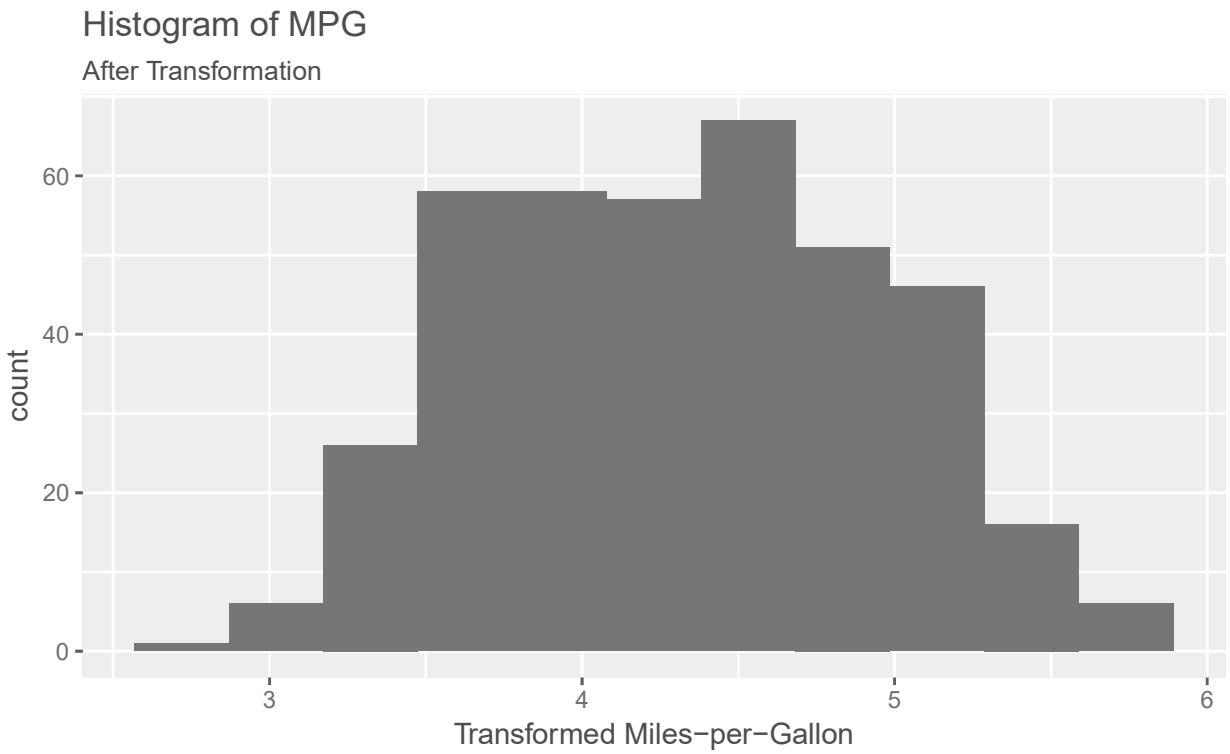



```
Cox = data.frame(Box$x, Box$y)              # Create a data frame with the results

Cox2 = Cox[with(Cox, order(-Cox$Box.y)),] # Order the new data frame by decreasing y

lambda = Cox2[1, "Box.x"]                   # Extract that lambda

lambda %>% kable()
```

0.2

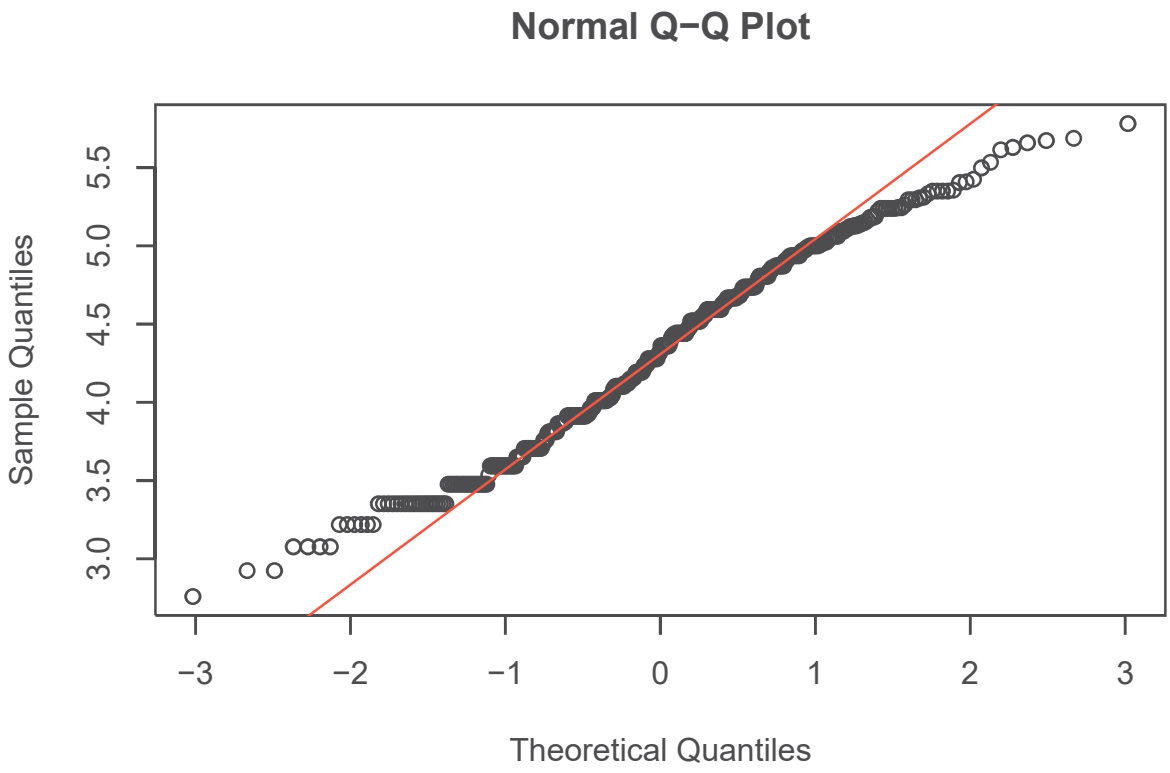
Using the above method, we found that *lambda* is *0.2*. We use this value to transform our target variable and observe the histogram again.

```r
ampg.2$tmpg = (ampg.2$mpg ^ lambda - 1)/lambda

ggplot(ampg.2, aes(x = (tmpg))) + geom_histogram(bins = binwidth(ampg.2$tmpg)) +
  labs(title = "Histogram of MPG",subtitle = "After Transformation",
       x = "Transformed Miles-per-Gallon")
```



Looking at the transformed histogram, we can say that the distribution is approximately normal. Lets confirm this by using Q-Q Plot as well.

```r
y = ampg.2$tmpg
qqnorm(y)
qqline(y, col = 2, lwd = 1, lty = 2)
```

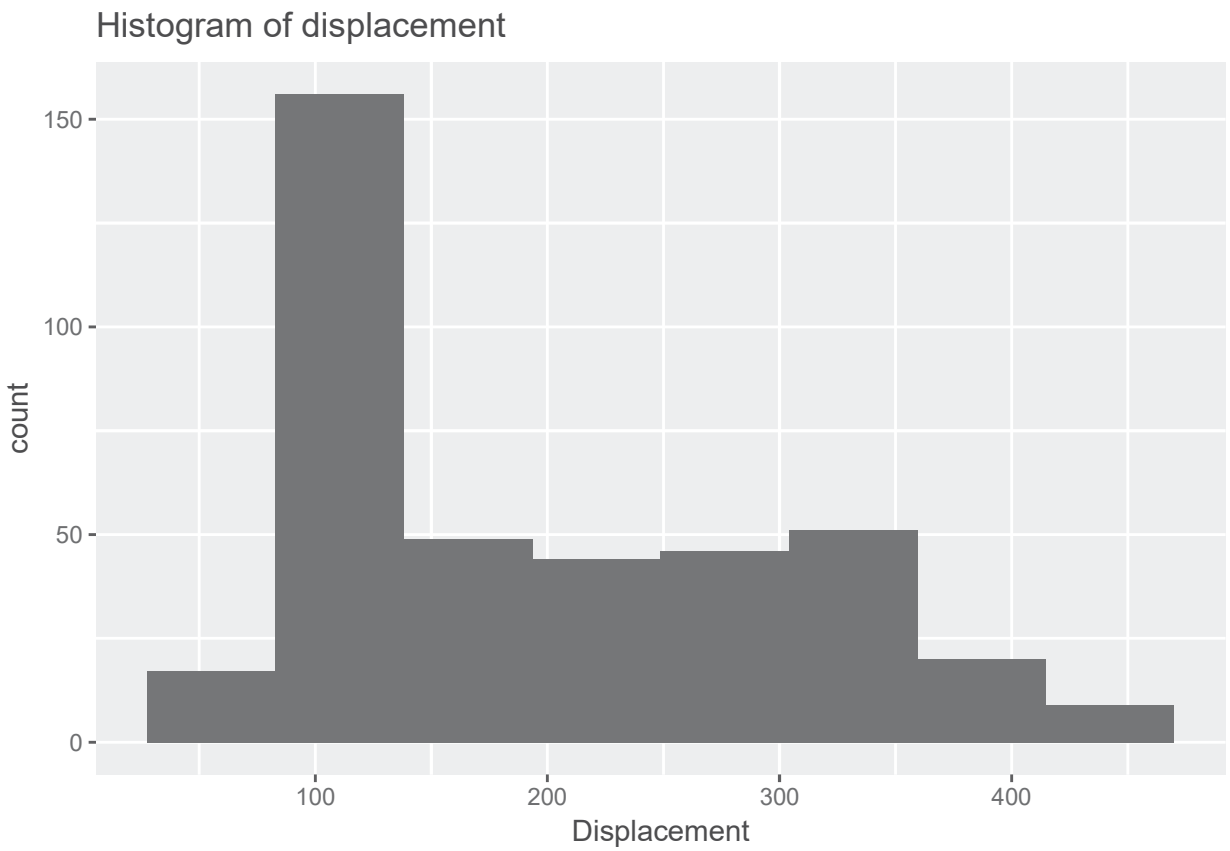


We see a small departure from the normal trend line, therefore it can be said the the `Tranformed MPG` is approximately normal. Lets explore the other descriptive features now and compare them with `tmpg`.

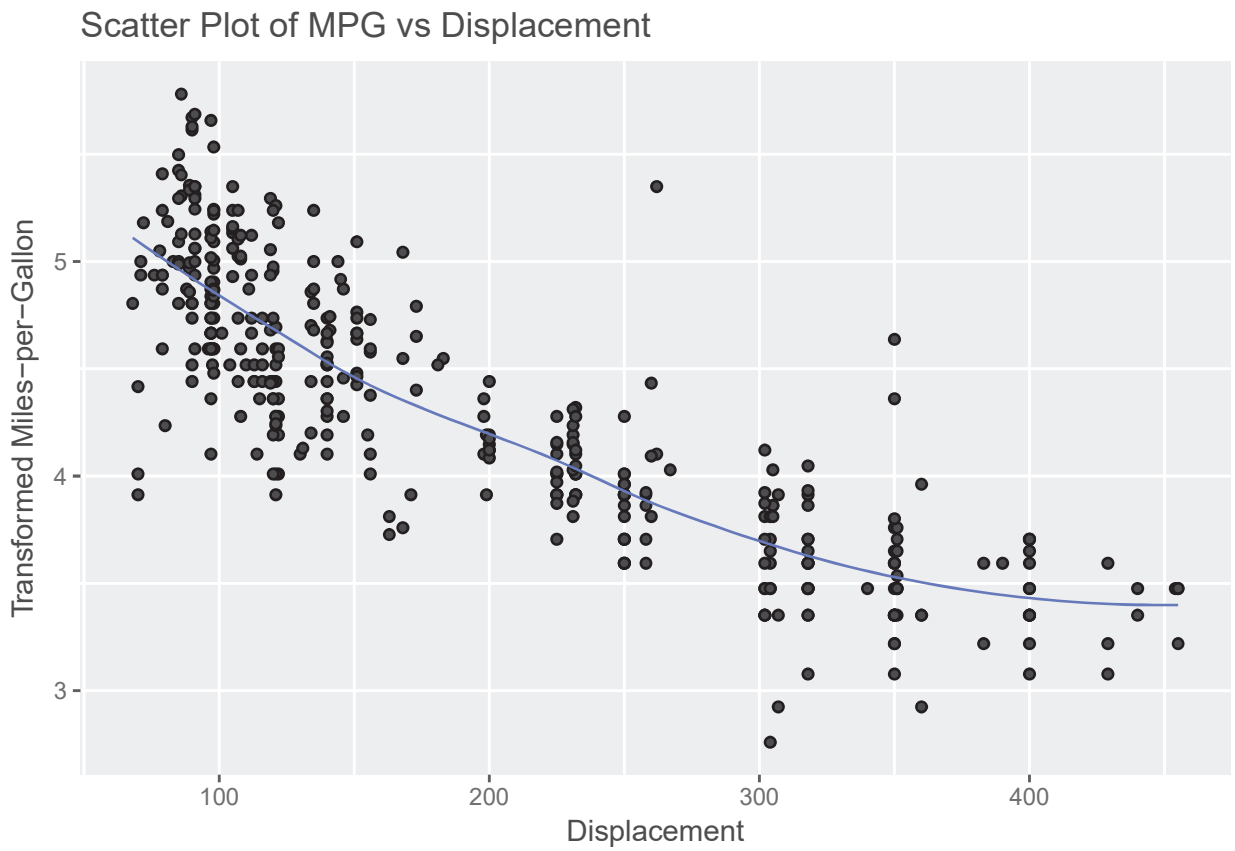
4.1.1.2 Displacement

Looking at the histogram below, we observe that our dataset is concentrated on low values of displacement. What could be the reason for this? Lets visualise `displacement` with `tmpg`.

```
ggplot(ampg.2, aes(x = displacement)) +
  geom_histogram(bins = binwidth(ampg.2$displacement), na.rm = TRUE) +
  labs(title = "Histogram of displacement", x = "Displacement")
```




```
ggplot(ampg.2, aes(x = (displacement), y = tmpg)) + geom_point() +  geom_smooth(se = F) +
  labs(title = "Scatter Plot of MPG vs Displacement",
       x = "Displacement", y="Transformed Miles-per-Gallon")
```



It can be seen in the scatter plot that displacement has a negative correlation with MPG. MPG decreases as displacement increases. There are more instances present for low displacement. This could be because MPG is an important factor for car demand?

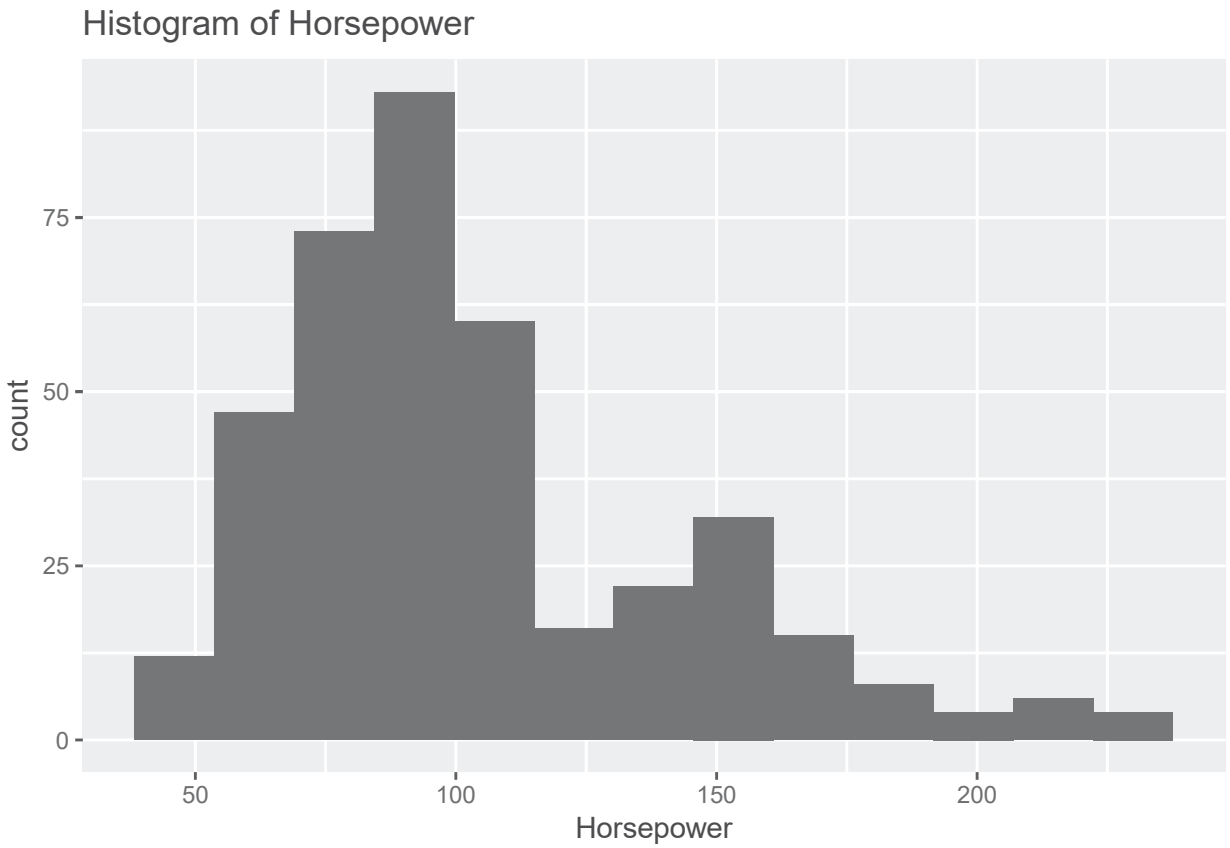
Also, we see that there are multiple values of MPG at same displacement, in multi-variate analysis we will attempt to explore this behavior.

4.1.1.3 Horsepower

We see approximately the same behavior in horsepower as displacement, the distribution is right skewed.

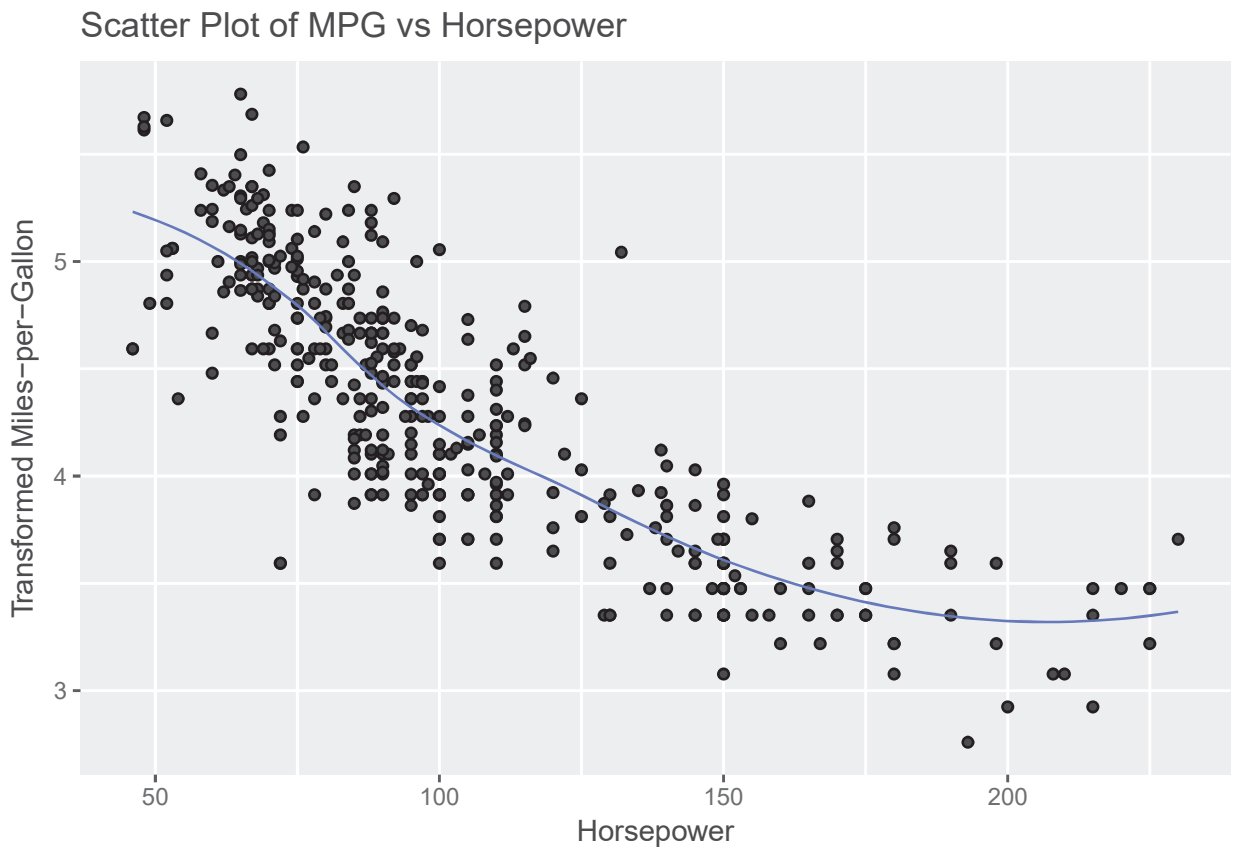
Also, it can be said that the distribution is approximately bi-modal, which can be an effect of other variable as well.

```
ggplot(ampg.2, aes(x = horsepower)) +
  geom_histogram(bins = binwidth(ampg.2$horsepower), na.rm = TRUE) +
  labs(title = "Histogram of Horsepower", x = "Horsepower")
```



Lets visualise the correlation as well.

```
ggplot(ampg.2, aes(x = (horsepower), y = tmpg)) + geom_point() +
  geom_smooth(se = F) +
  labs(title = "Scatter Plot of MPG vs Horsepower",
       x = "Horsepower", y="Transformed Miles-per-Gallon")
```

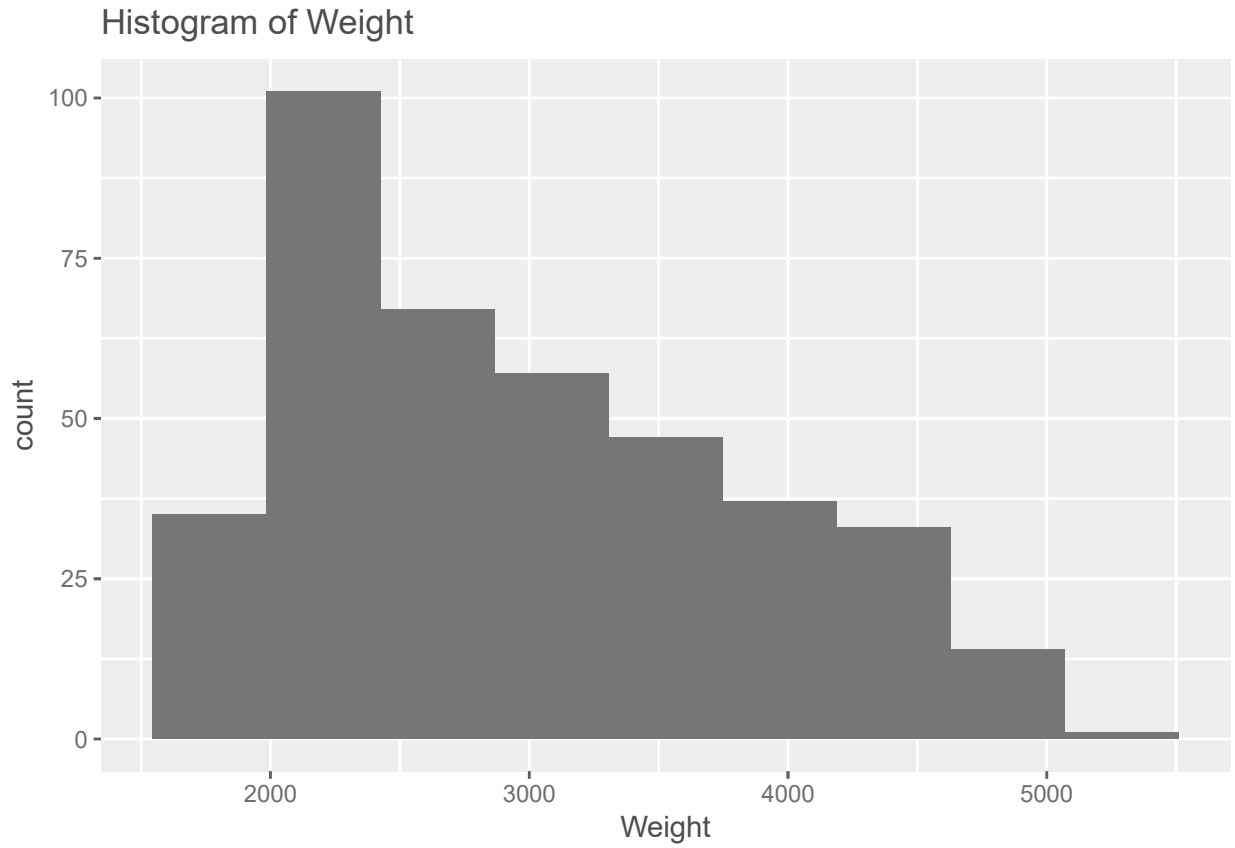


We can see below, that the `horsepower` too has a negative correlation with `mpg`. With increase in `horsepower`, `mpg` decreases.

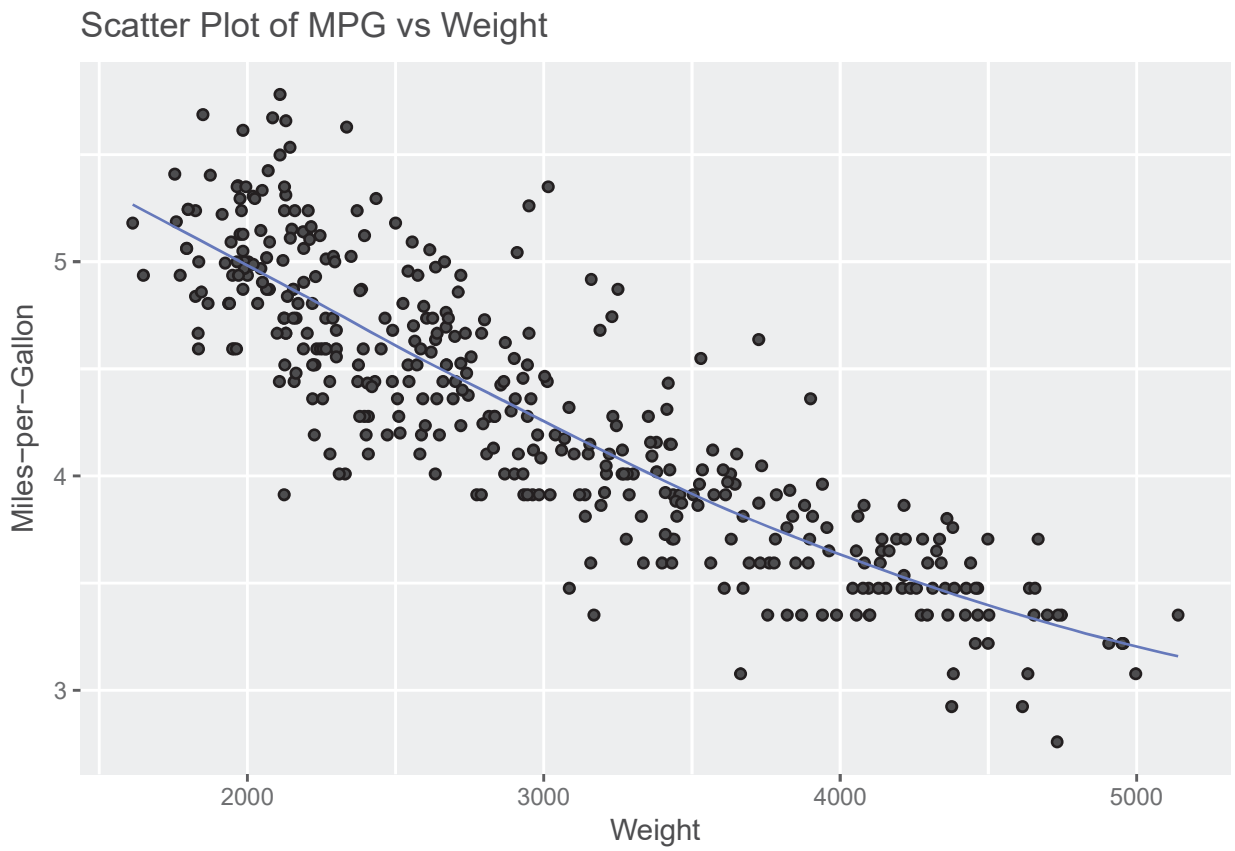
4.1.1.4 Weight

Like the distribution of previous two numerical feature, the distribution of `weight` is alos right skewed.

```
ggplot(ampg.2, aes(x = weight)) + geom_histogram(bins = binwidth(ampg.2$weight), na.rm = TRUE) +
  labs(title = "Histogram of Weight", x = "Weight")
```




```
ggplot(ampg.2, aes(x = weight, y = tmpg)) + geom_point() + geom_smooth(se = F) +
  labs(title = "Scatter Plot of MPG vs Weight", x = "Weight", y="Miles-per-Gallon")
```



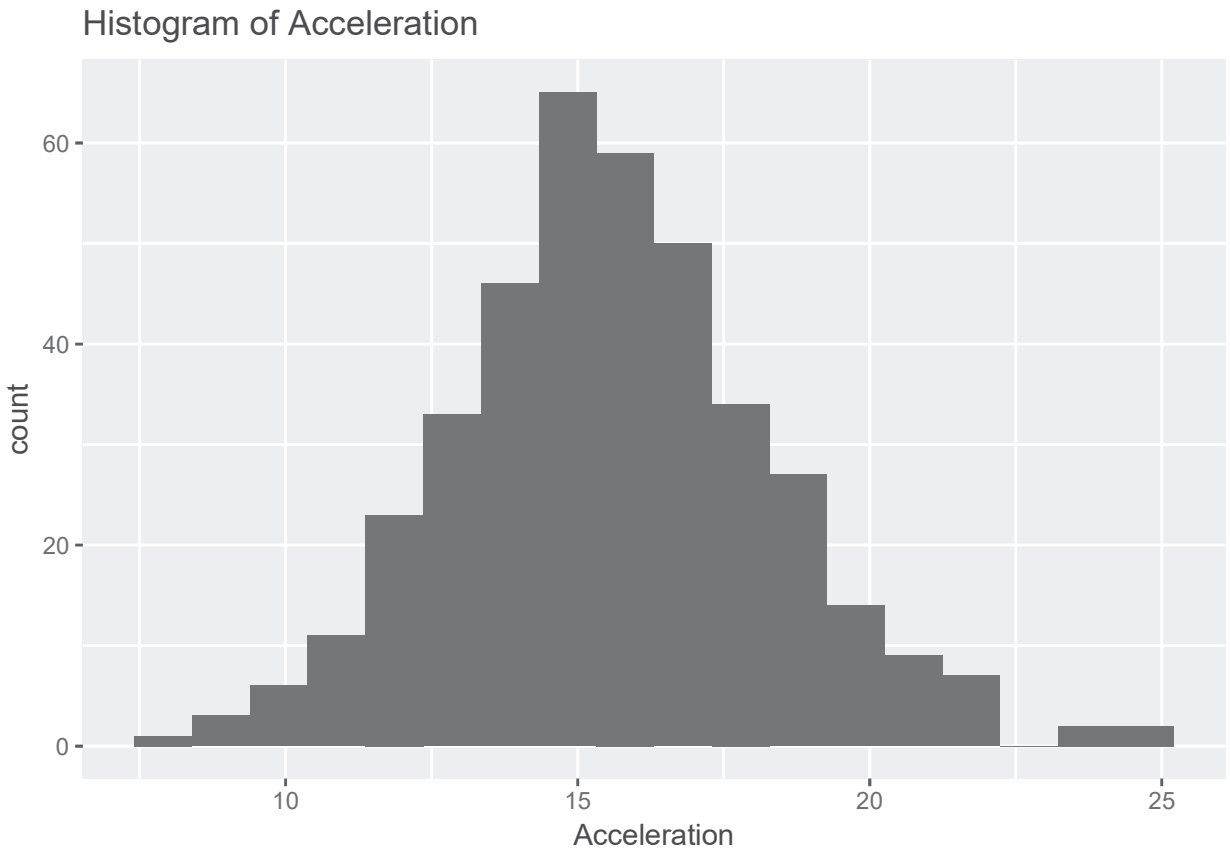
Also, through scatter plot, we can see that there is negative correlation between `weight` and `MPG`, with increase in `weight`, `MPG` decreases.

Comparing the behavior with previous two features, the correlation of `weight` with `MPG` is more linear.

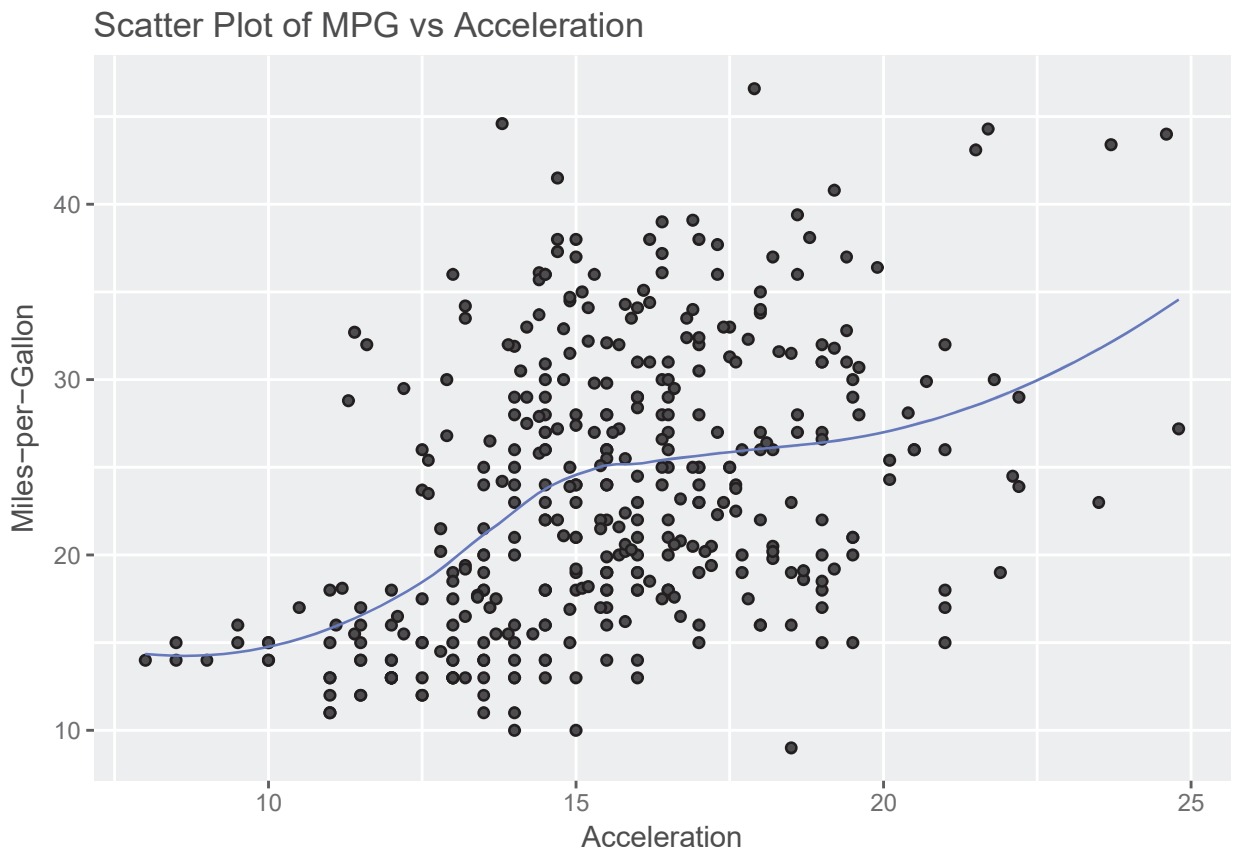
4.1.1.5 Acceleration

We see an approximaterly normal distribution of acceleration in the dataset with an outlier on the right tail.

```
ggplot(ampg.2, aes(x = acceleration)) +
  geom_histogram(bins = binwidth(ampg.2$acceleration), na.rm = TRUE) +
  labs(title = "Histogram of Acceleration", x = "Acceleration")
```




```
ggplot(ampg.2, aes(x = acceleration, y = mpg)) + geom_point() + geom_smooth(se=F) +
  labs(title = "Scatter Plot of MPG vs Acceleration", x = "Acceleration", y="Miles-per-Gallon")
```



Looking at the below scatter plot, we can say that `acceleration` has a very weak positive correlation with `MPG`, so, it is possible that this is not a predictive feature.

NOTE: Exploration of numerical features revealed that all numerical features except `acceleration` has a negative correlation with our target feature `mpg`.

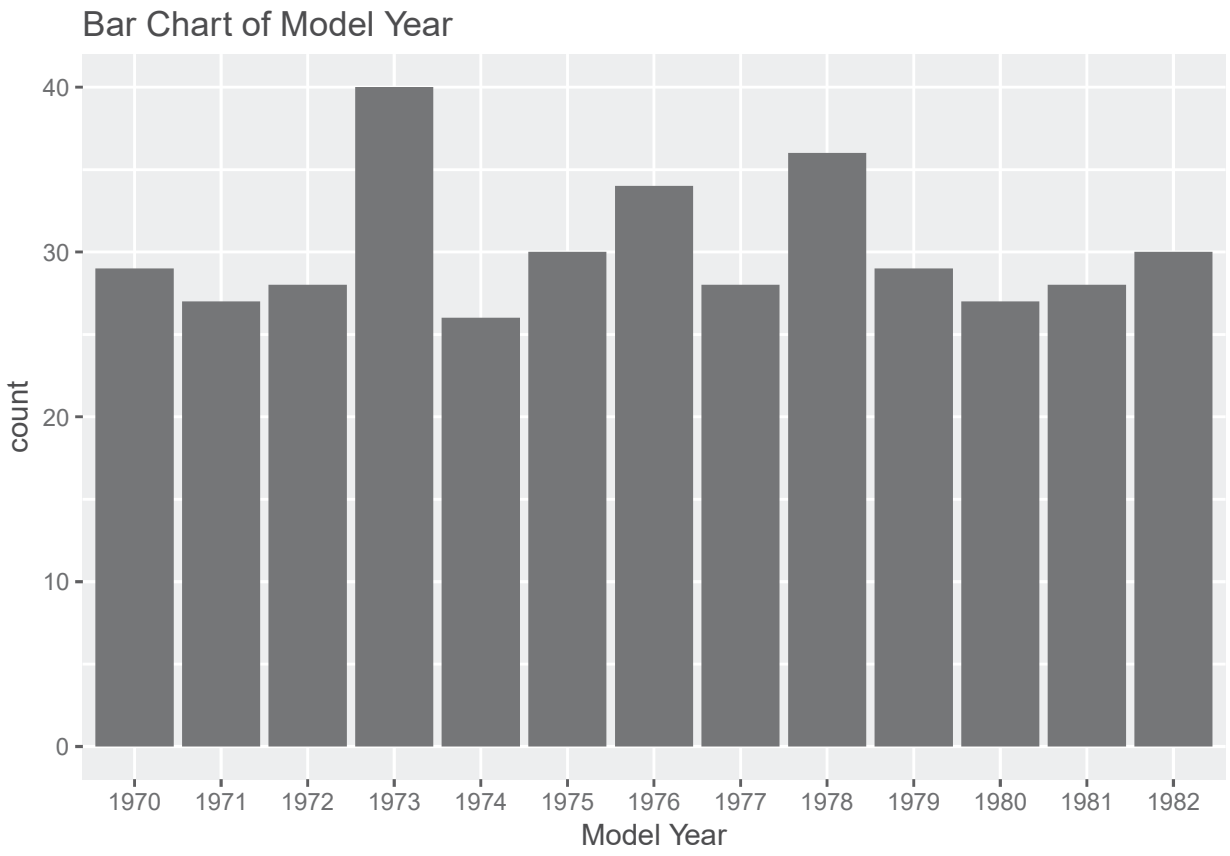
4.1.2 Categorical Features

4.1.2.1 Model Year

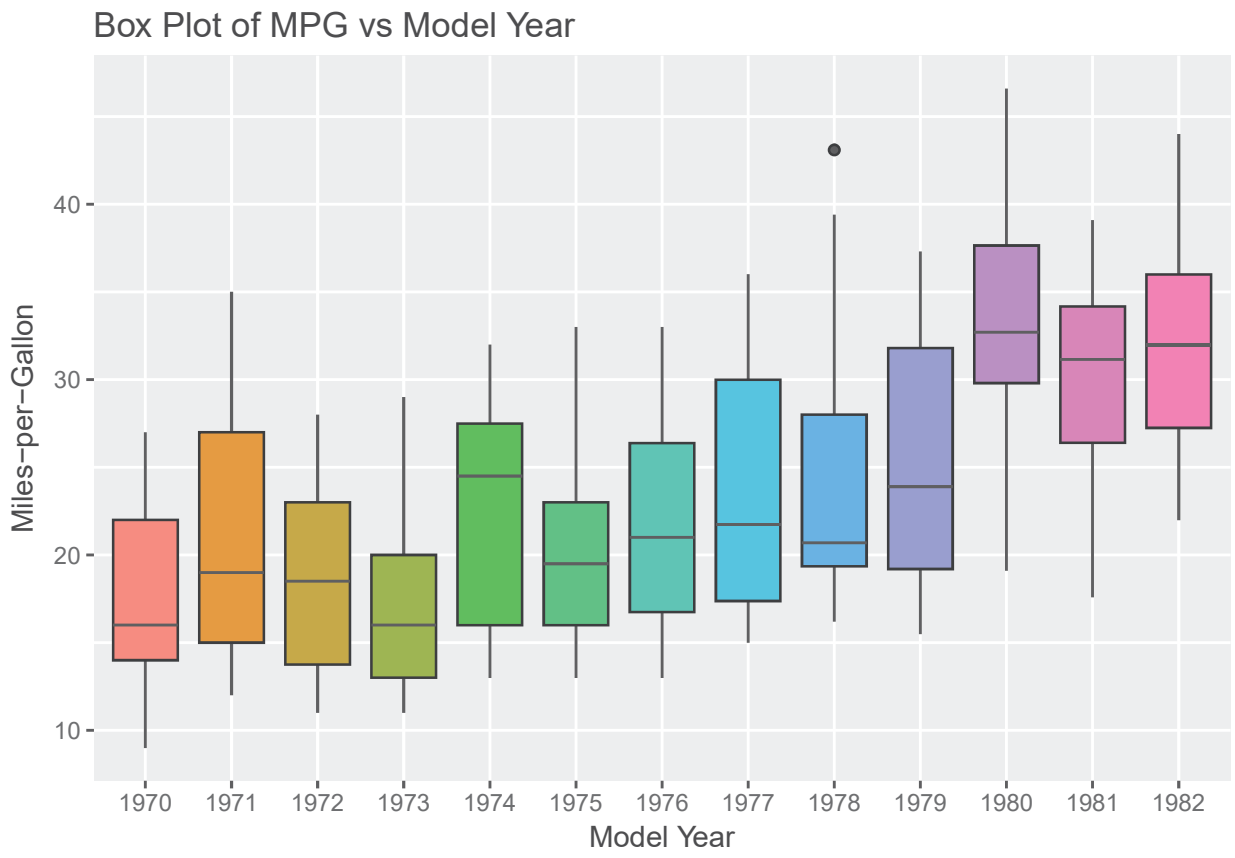
As seen in Data Summarization above, there are approximately even number of instances in each `model year`.

Lets compare this with the `Transformed MPG`.

```
ggplot(ampg.2, aes(x = `modelYear`)) +
    geom_bar() + labs(title = "Bar Chart of Model Year", x = "Model Year")
```




```
ggplot(ampg.2, aes(x = `modelYear`, y=mpg, fill = `modelYear`)) + geom_boxplot() +
  labs(title = "Box Plot of MPG vs Model Year", x = "Model Year", y = "Miles-per-Gallon") +
  guides(fill = FALSE)
```



We observe that the median MPG is increasing as we go from year 1970 to 1982. Also, there is an outlier in Model Year 1978, where the MPG > 40.

The year 1973 is seen with lowest median MPG while the year 1980 saw the highest median MPG.

```
##Running anova test

summary(aov(mpg~`modelYear`, data = ampg.2))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## modelYear    12  10236   853.0    23.8 <2e-16 ***
## Residuals   379  13583    35.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

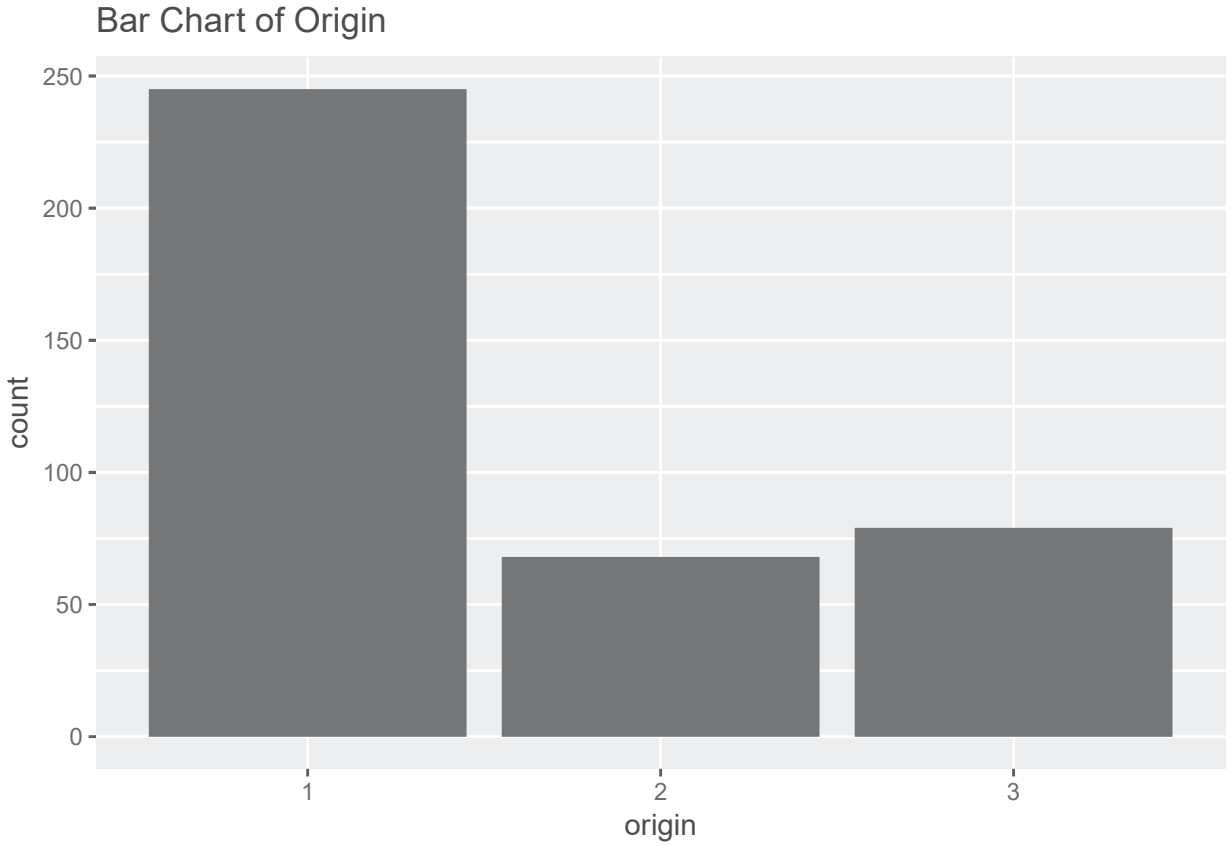
We also ran an anova test and found that the model year is statistically significant with MPG.

4.1.2.2 Origin

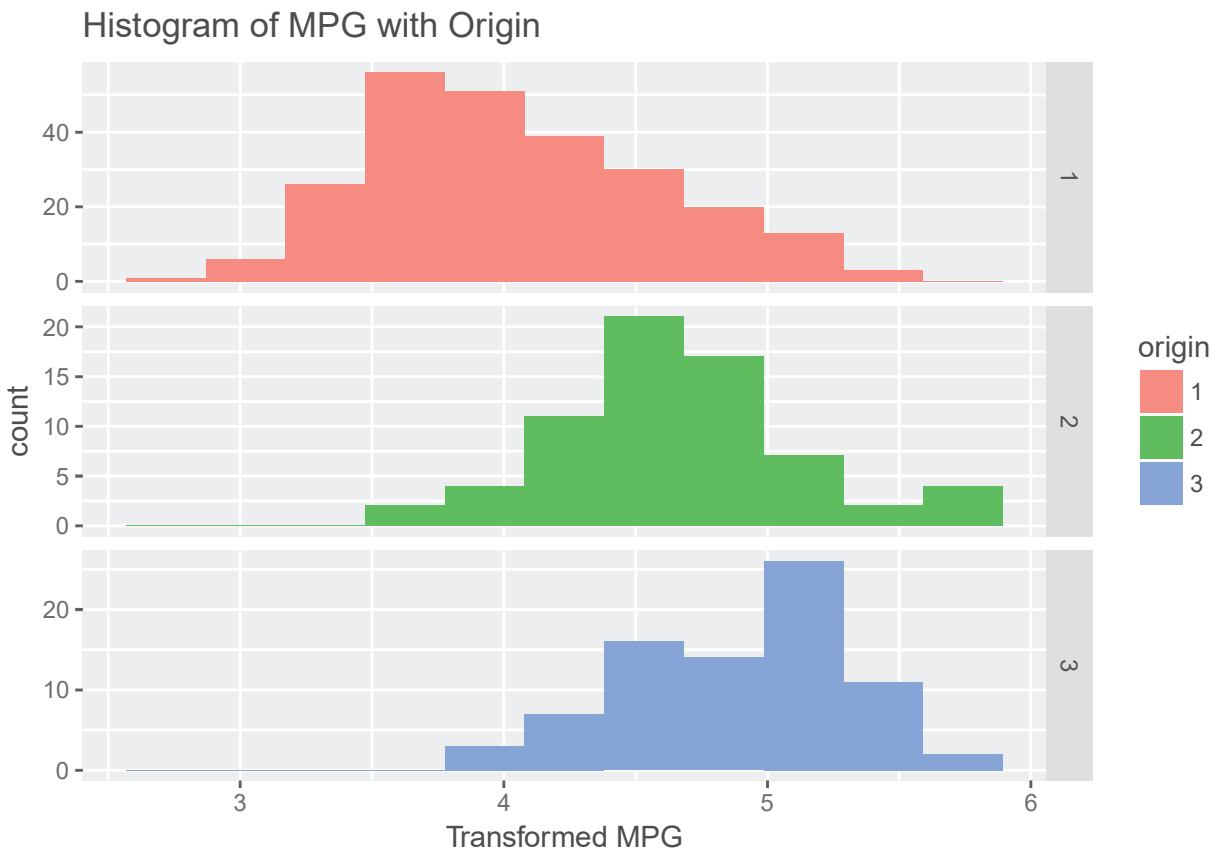
We can see that the number of instances from origin 1 are greater than the number of instances from origin 2 and 3.

Lets have a look at the distribution of MPG in each origin.

```
ggplot(ampg.2, aes(x = origin)) +
    geom_bar() + labs(title = "Bar Chart of Origin")
```




```
ggplot(ampg.2, aes(x = tmpg, fill = origin)) + geom_histogram(bins = binwidth(ampg.2$tmpg)) +
  facet_grid(origin~., scales = "free_y") +
  labs(title = "Histogram of MPG with Origin", x="Transformed MPG")
```

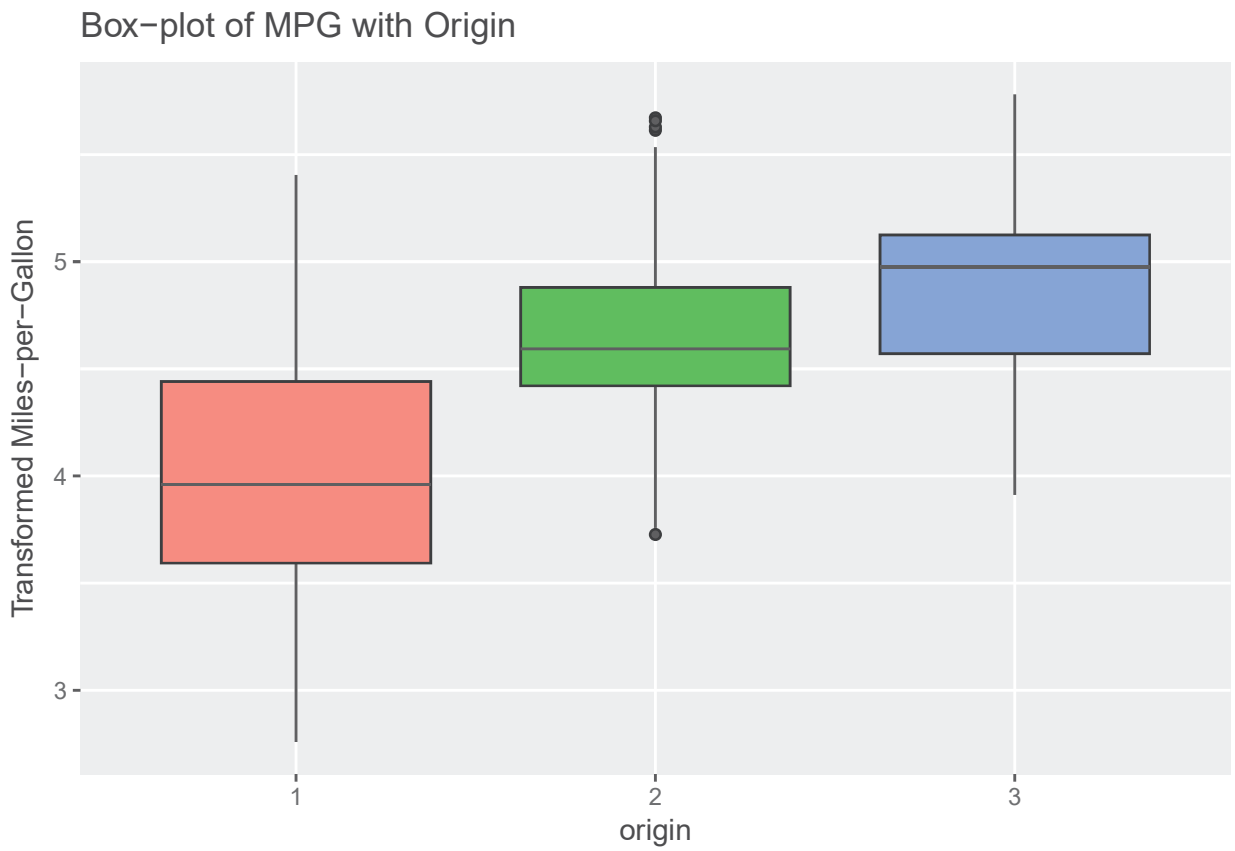


Looking at the histogram above, we have following observations:

- The distribution in `origin 1` is right skewed
- The distribution in `origin 2` is approximately normal
- The distribution in `origin 3` is left skewed

We will explore the distribution further through a box plot.

```
ggplot(ampg.2, aes(x = origin, y=tmpg, fill = origin)) + geom_boxplot() +
  labs(title = "Box-plot of MPG with Origin", y="Transformed Miles-per-Gallon") +
  guides(fill = FALSE)
```



The Boxplot confirms the existence of outliers in `origin 2`. Also, the origin 3 has the highest mean MPG. This change in `MPG` with origin shows the effect of `origin` on our target feature.

```
summary(aov(mpg~origin, data = ampg.2))
```

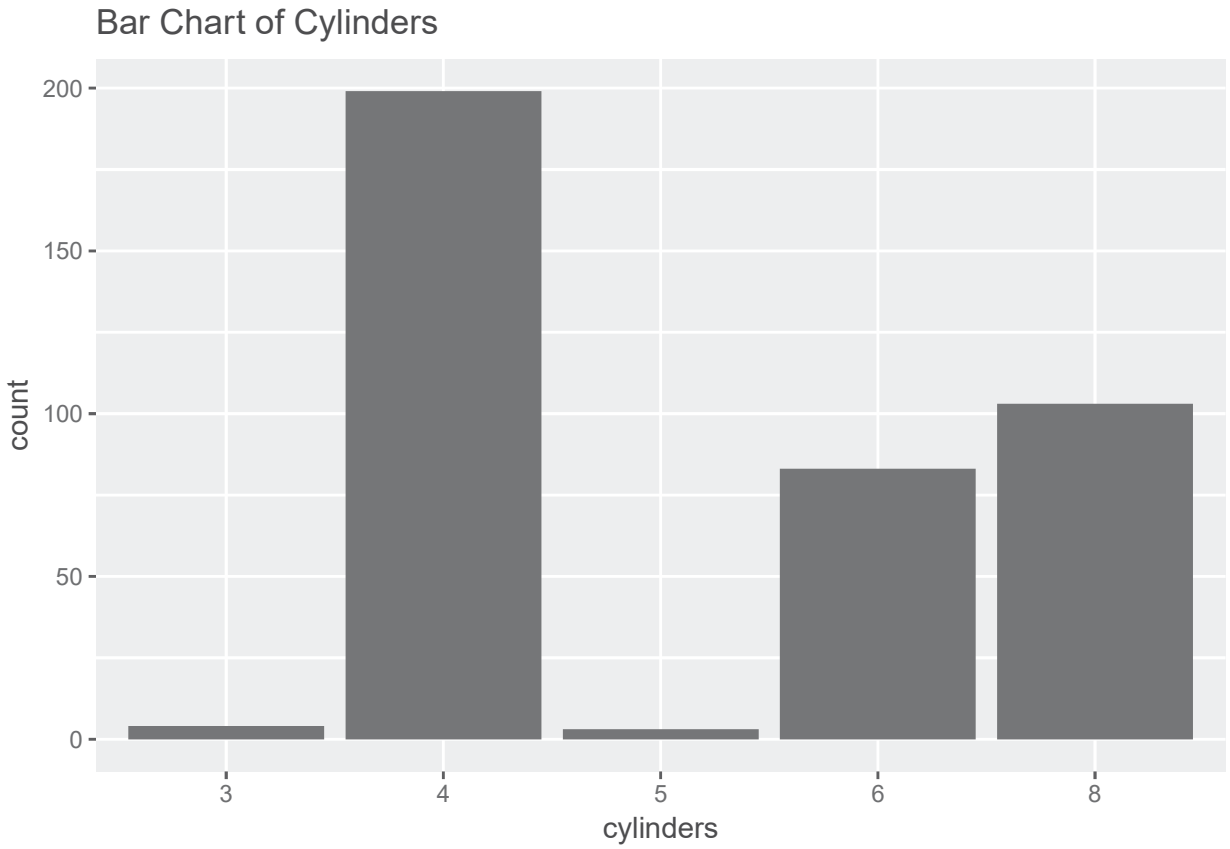
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## origin        2   7904    3952    96.6 <2e-16 ***
## Residuals   389  15915      41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The same is confirmed using ANOVA test, thus, this could be a predictive feature.

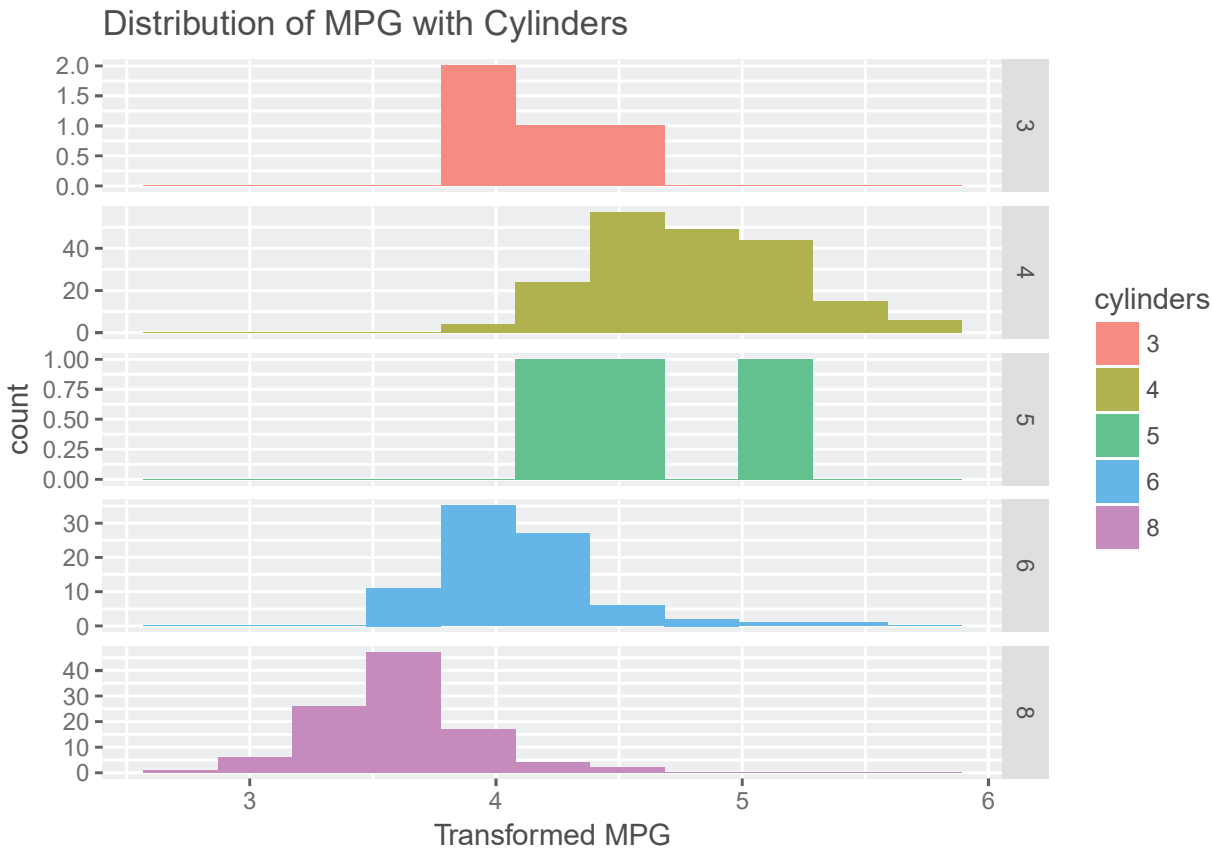
4.1.2.3 Cylinders

The number of instances of cars with 3 and 5 cylinders is very less as compared to cylinders 4, 6 & 8.

```
ggplot(ampg.2, aes(x = cylinders)) +
    geom_bar() + labs(title = "Bar Chart of Cylinders")
```




```
ggplot(ampg.2, aes(x = tmpg, fill = cylinders)) + geom_histogram(bins = binwidth(ampg.2$tmpg)) +
  facet_grid(cylinders ~ ., scales = "free_y") +
  labs(title = "Distribution of MPG with Cylinders", x = "Transformed MPG")
```

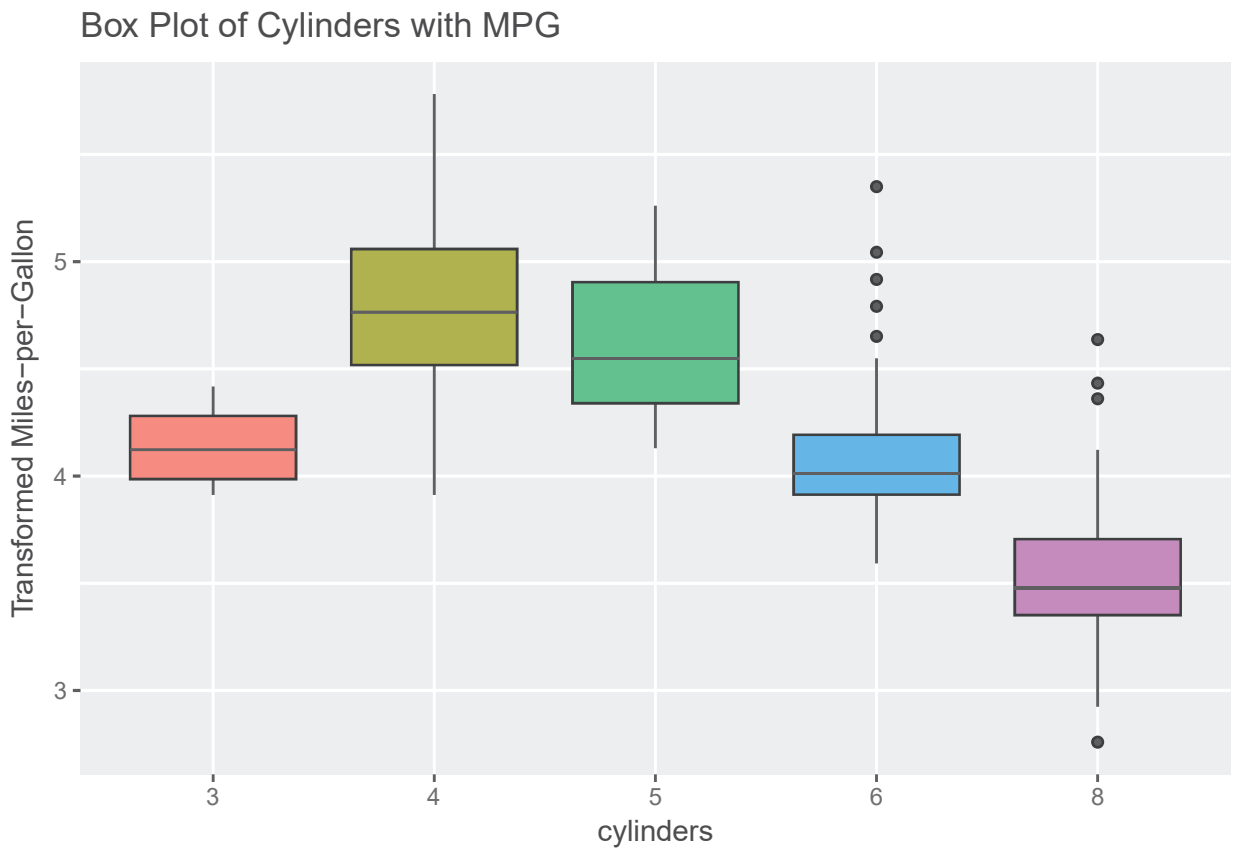


Looking at the distribution of MPG with cylinders:

- Due to few instances, nothing can be said about the 3 & 5 cylinder cars
- the cars with 4,6 & 8 cylinders shows an approximate normal distribution with few outliers.

We will explore this behavior further using a box plot.

```
ggplot(ampg.2, aes(x = cylinders, y=tmpg, fill = cylinders)) +
  geom_boxplot() +
  labs(title = "Box Plot of Cylinders with MPG", y = "Transformed Miles-per-Gallon") +
  guides(fill = FALSE)
```



The Box-plot confirms the existence of outliers for instances of 6 & 8 cylinder cars. Also, it shows that 4 cylinder cars have highest mean MPG while 8 cylinders have the lowest.

```
summary(aov(tmpg~cylinders, data = ampg.2))
```

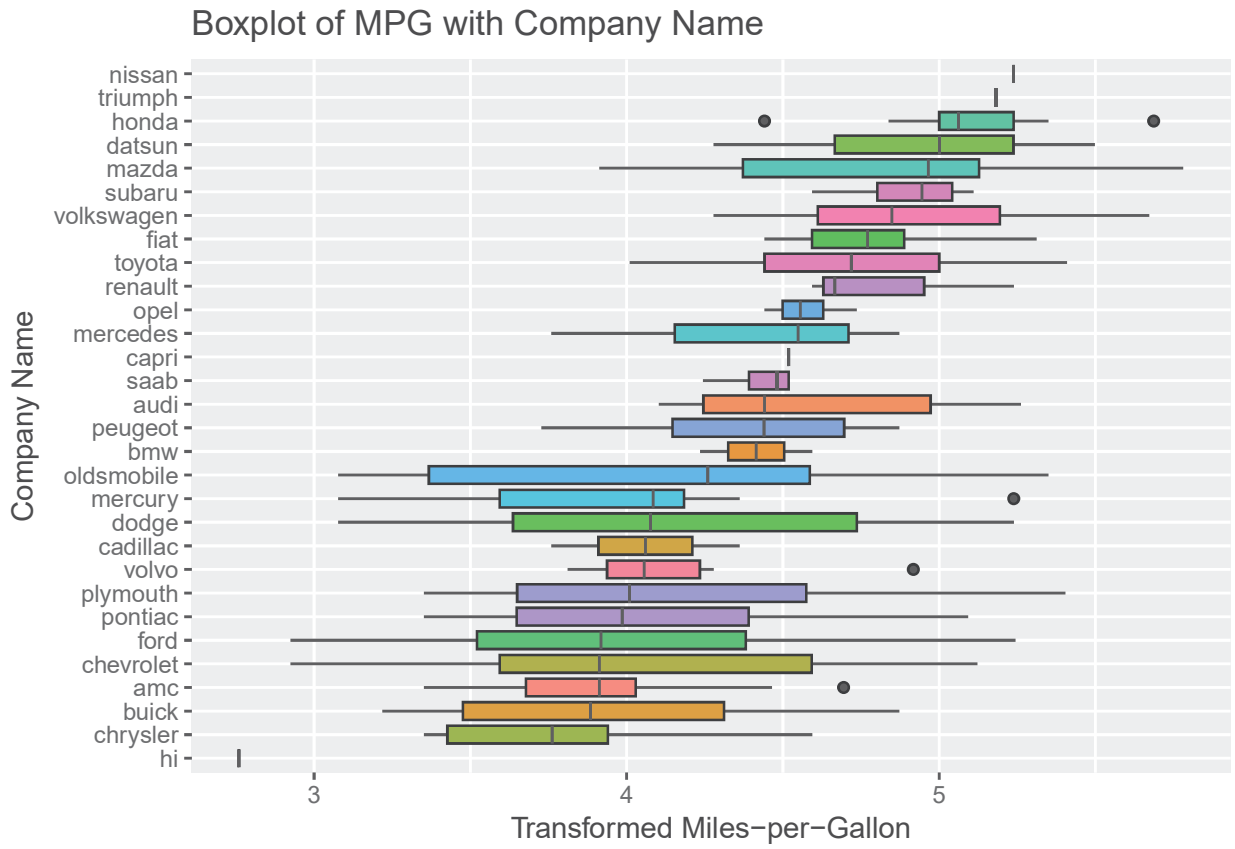
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## cylinders     4 108.69  27.173   224.8 <2e-16 ***
## Residuals   387  46.77   0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test shows that number of cylinders is a significant feature for MPG.

4.1.2.4 Company Name

We observe that the cars from *Triumph* has the *highest* mean `MPG` and *lowest* from *hi*.

```
ggplot(ampg.2, aes(x = fct_reorder(company_name, tmpg, data = ampg.2, desc = TRUE),
                   y=tmpg, fill = company_name)) +
  geom_boxplot() + guides(fill = FALSE) +coord_flip() +
  labs(title = "Boxplot of MPG with Company Name", y = "Transformed Miles-per-Gallon",
       x = "Company Name")
```



4.2 Multivariate Analysis

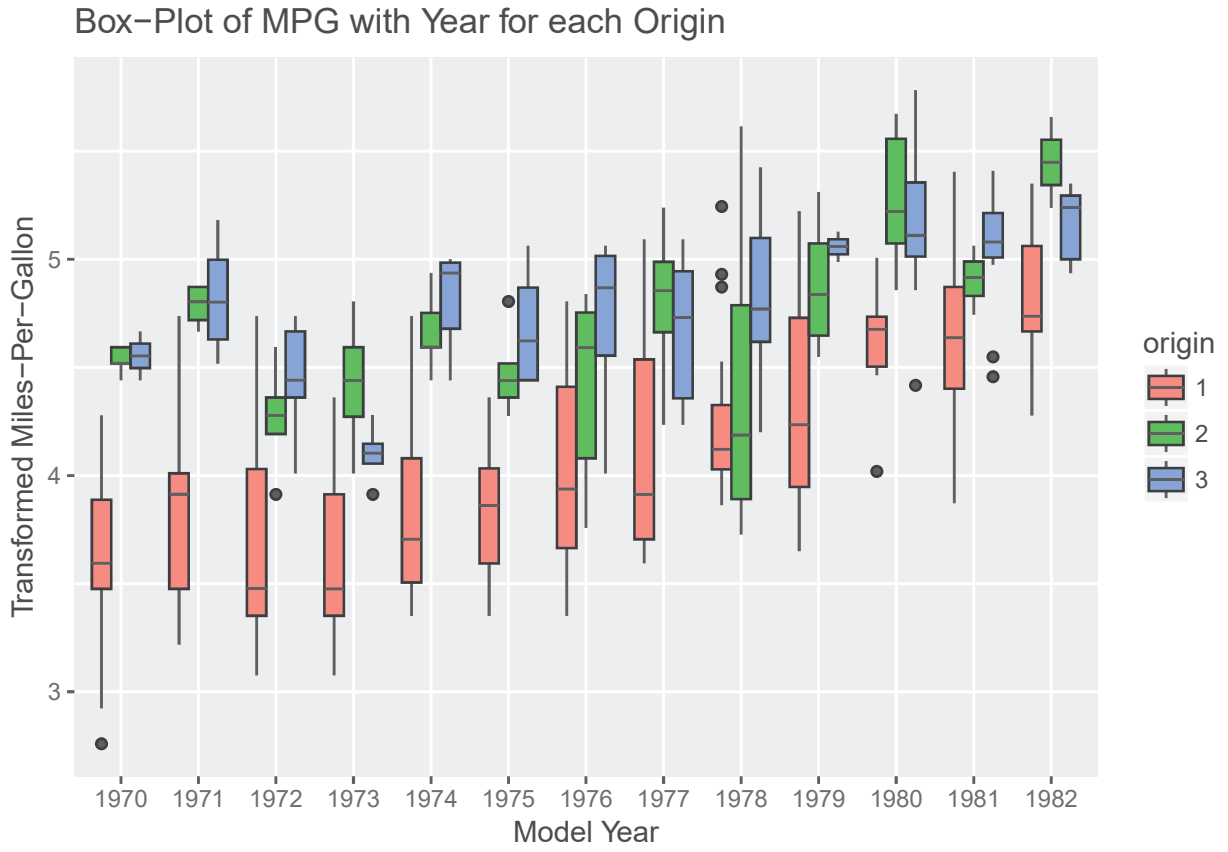
As we saw in univariate analysis, numerical features have effect of other variables as well besides MPG. In this section we will explore the impact of other variables on our numerical descriptive features.

4.2.1 MPG, Origin and Model Year

While MPG for origin 1 is clearly increasing with year, we saw a lot of variation for instances with `origin` 2 and 3, but, there is an overall increase in `MPG` for all three origins.

This could be an effect of technological development in a decade.

```
ggplot(ampg.2, aes(x = `modelYear`, y = tmpg, fill = origin)) + geom_boxplot() +
  labs(title = "Box-Plot of MPG with Year for each Origin",
       x = "Model Year", y = "Transformed Miles-Per-Gallon")
```

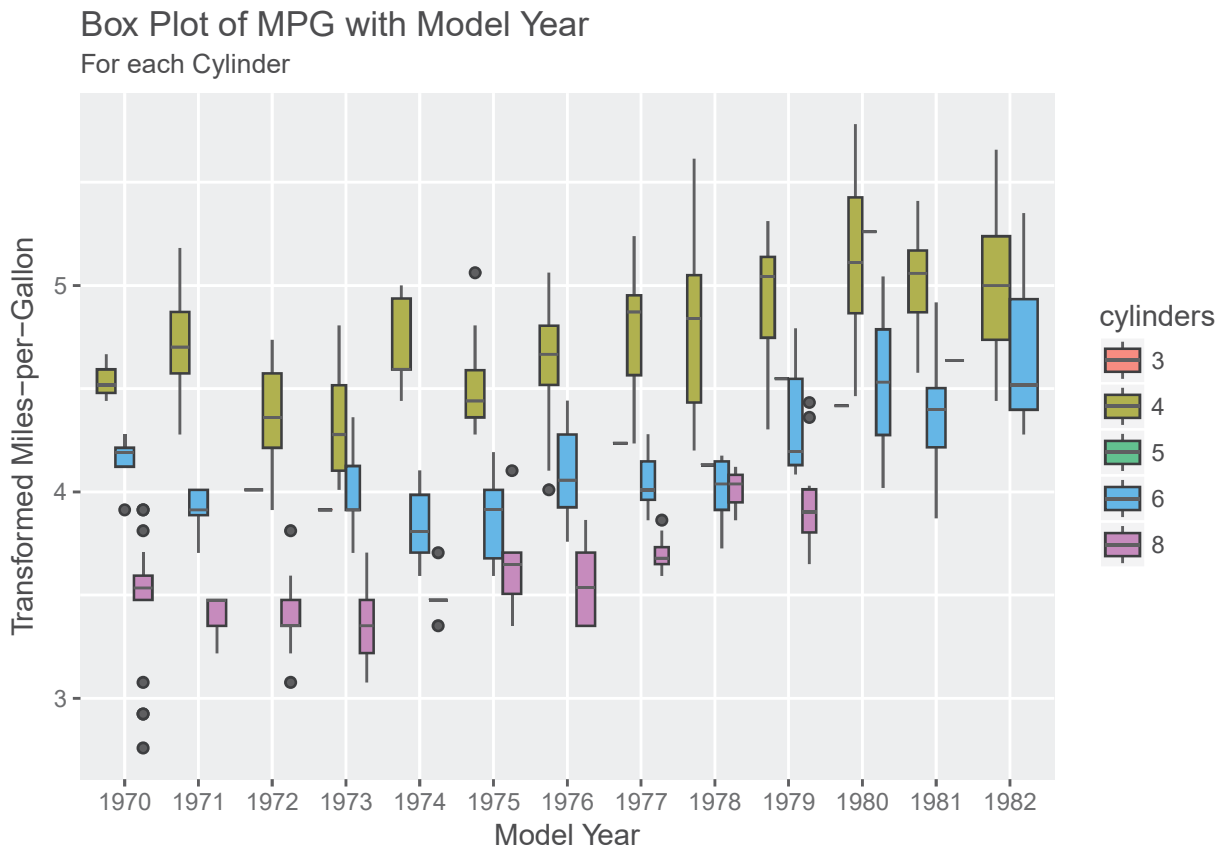


4.2.2 MPG, Model Year & Cylinders

This plot shows an interaction between model year and number of cylinders. Although, we saw in univariate analysis that the performance of cars has improved over the years, but, this shows that number of cylinders of a car plays a vital role in this change.

- As seen earlier also, there are very few data points for cylinders 3 & 5.
- The performance of each engine has improved over the years
- There is a very small change in performance of engines from 1970 to 1982

```
ggplot(ampg.2, aes(x = `modelYear`, y = tmpg, fill = cylinders)) + geom_boxplot() +
  labs(title = "Box Plot of MPG with Model Year",
       subtitle = "For each Cylinder", x = "Model Year",
       y = "Transformed Miles-per-Gallon")
```



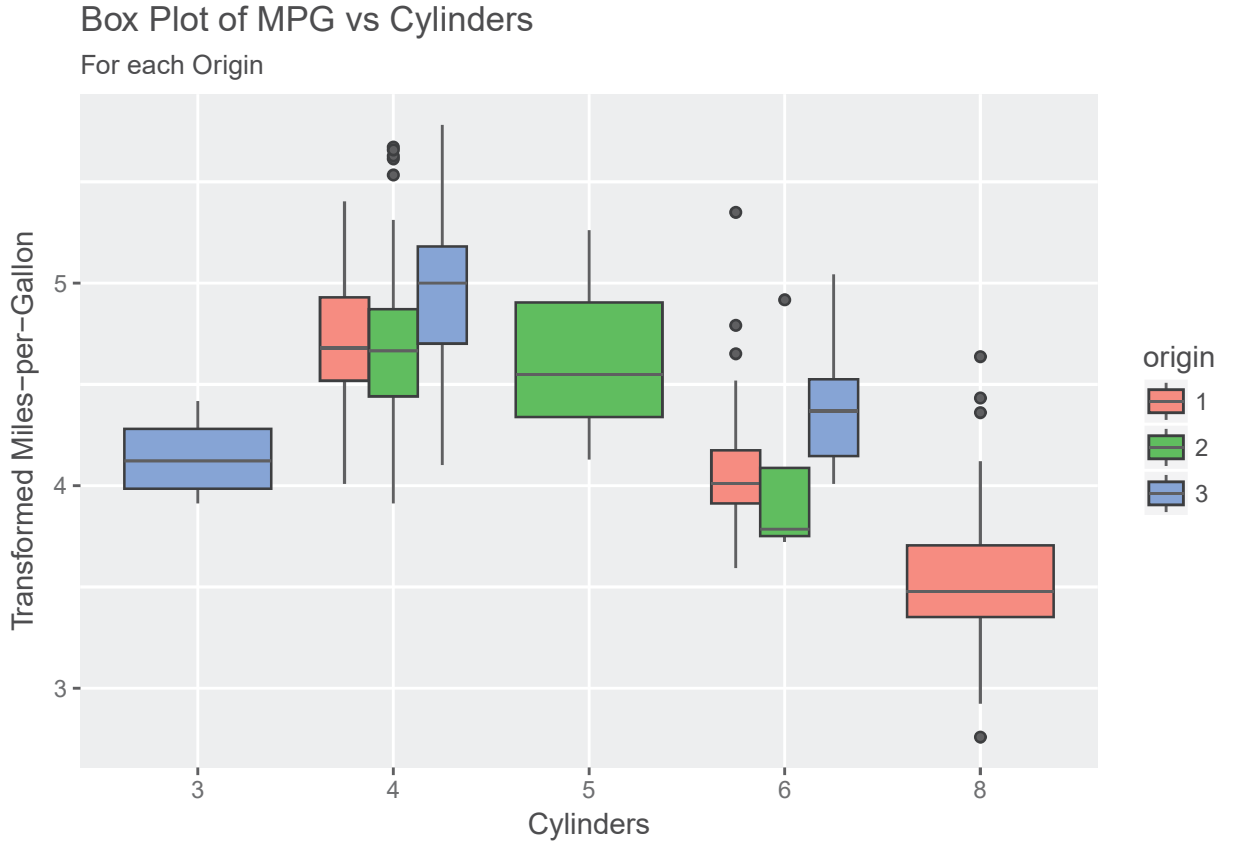
4.2.3 MPG, Origin and Cylinders

We had earlier seen that cars with 4 cylinders had best MPG, but this plot helps to explore this behavior further with `origin`. Cars with `4 cylinders` and from `origin 3` have the highest mean `MPG`.

There is a clear indication on following things: + Cars with cylinder 3 came from only origin 3 and cylinders 5 from origin 2 + Cars with 8 cylinders came only from origin 1 + Cars with 4 & 6 cylinders were from all three origins in the dataset

We can see that there is no change in behavior of MPG with cylinders in each origin, thus, there is no interaction between cylinders and origin.

```
ggplot(ampg.2, aes(x = cylinders, y = tmpg, fill = origin)) + geom_boxplot() +
  labs(title = "Box Plot of MPG vs Cylinders",
       subtitle = "For each Origin", x = "Cylinders",
       y = "Transformed Miles-per-Gallon")
```



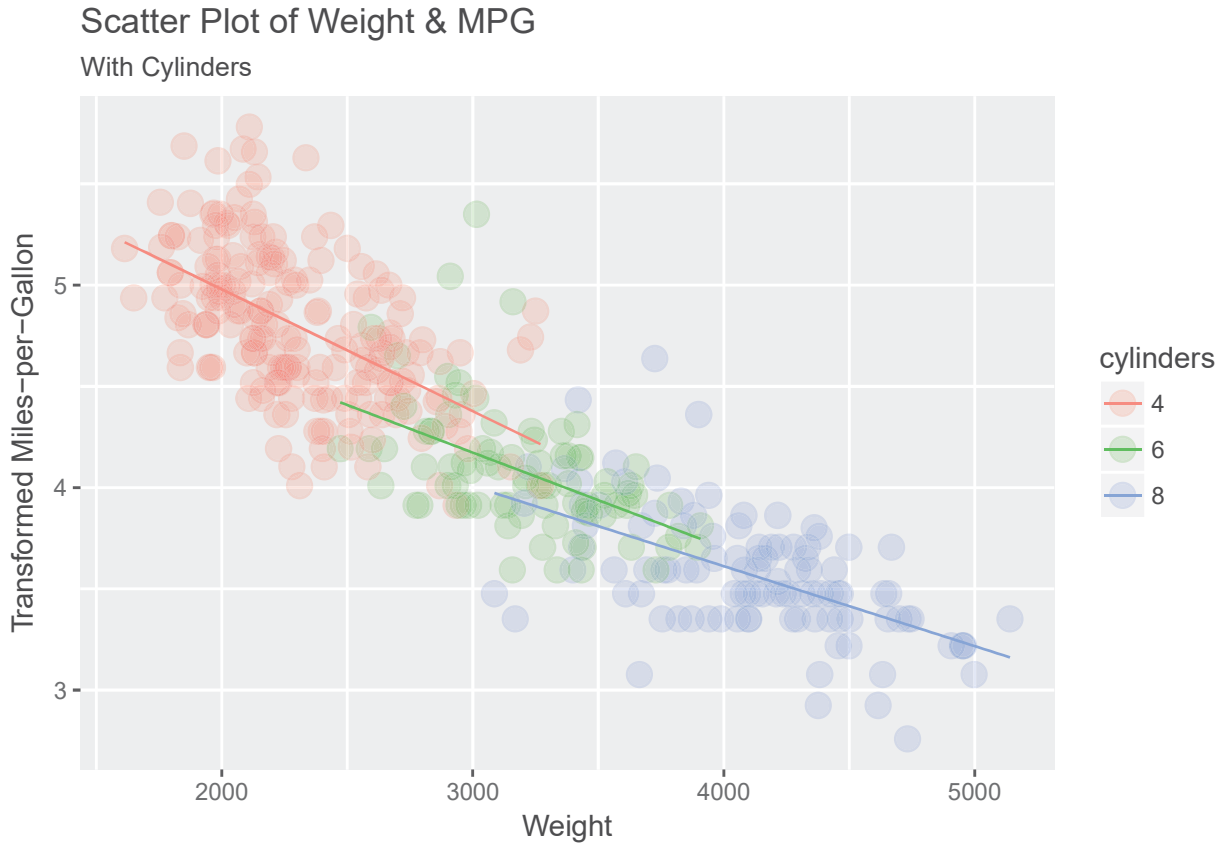
4.2.4 MPG,Weight and Cylinders

Keeping our focus on 4, 6 & 8 cylinders, as 3 & 5 had very few data points in the data set

With an increase in weight:

- Cars with 4 cylinders saw a steep decrease in MPG
- Cars with 6 cylinders had a less steep decrease in MPG
- Cars with 8 cylinders although had the same downward trend but less steeper as compared to above two

```
ggplot(ampg.2[ampg.2$cylinders!="3" & ampg.2$cylinders!="5",],
       aes(x = weight, y = tmpg, color = cylinders)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  labs(title = "Scatter Plot of Weight & MPG", subtitle = "With Cylinders",
       x = "Weight", y = "Transformed Miles-per-Gallon")
```



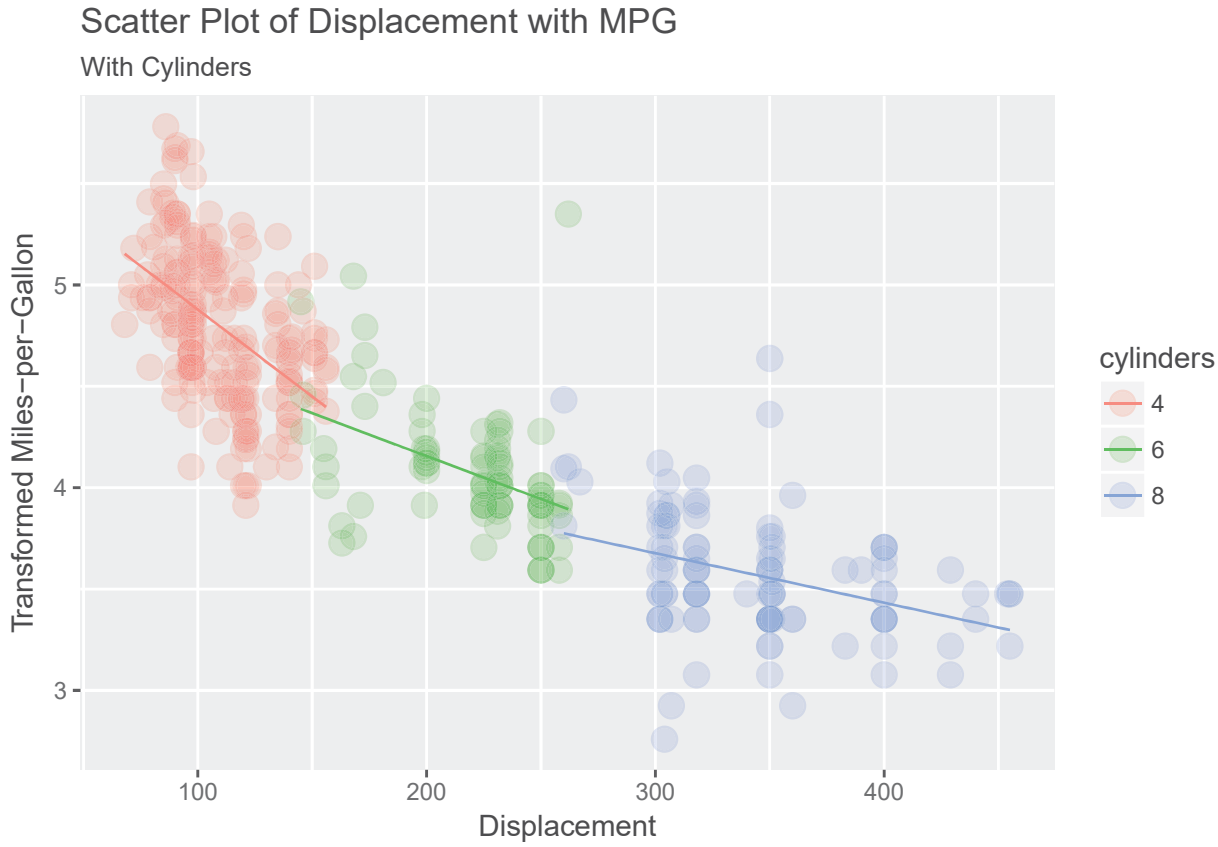
4.2.5 MPG, Displacement and Cylinders

Keeping our focus on 4, 6 & 8 cylinders, as 3 & 5 had very few data points in the data set

We see the same behavior with displacement as with weight we saw above.

We also see a clear distinction that cars with bigger number of cylinders have higher displacement which is expected as this displacement is *EngineDisplacement*.

```
ggplot(ampg.2[ampg.2$cylinders!="3" & ampg.2$cylinders!="5",],
       aes(x = displacement, y = tmpg, color = cylinders)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  labs(title = "Scatter Plot of Displacement with MPG",
       subtitle = "With Cylinders", x="Displacement",
       y = "Transformed Miles-per-Gallon")
```



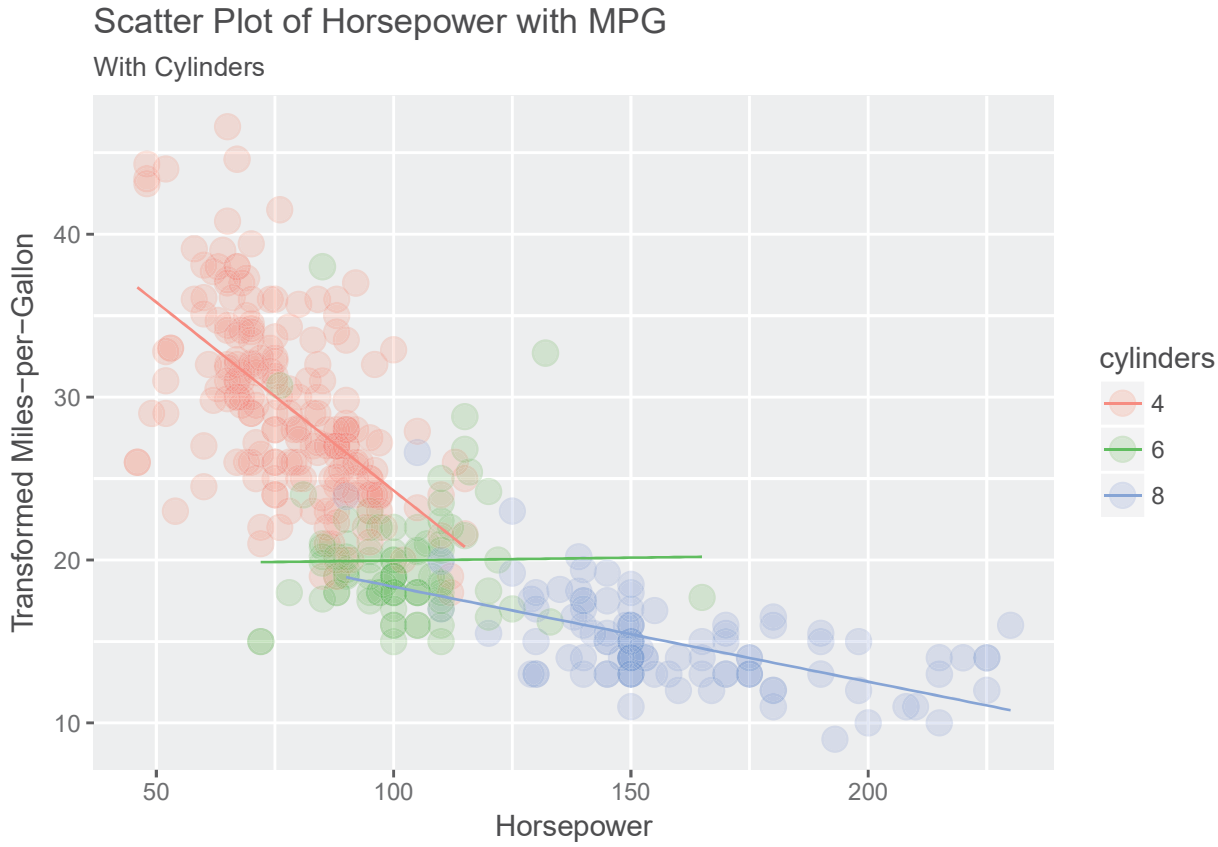
4.2.6 MPG, Horsepower and Cylinders

Keeping our focus on 4, 6 & 8 cylinders, as 3 & 5 had very few data points in the data set

Although we see the same behavior for cars with 4 and 8 cylinders.

But there is a drastic change in behavior of 6 cylinder cars. It show very small change in MPG with increase in horsepower.

```
ggplot(ampg.2[ampg.2$cylinders!="3" & ampg.2$cylinders!="5",],
       aes(x = horsepower, y = mpg, color = cylinders)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  labs(title = "Scatter Plot of Horsepower with MPG",
       subtitle = "With Cylinders", x = "Horsepower",
       y = "Transformed Miles-per-Gallon")
```

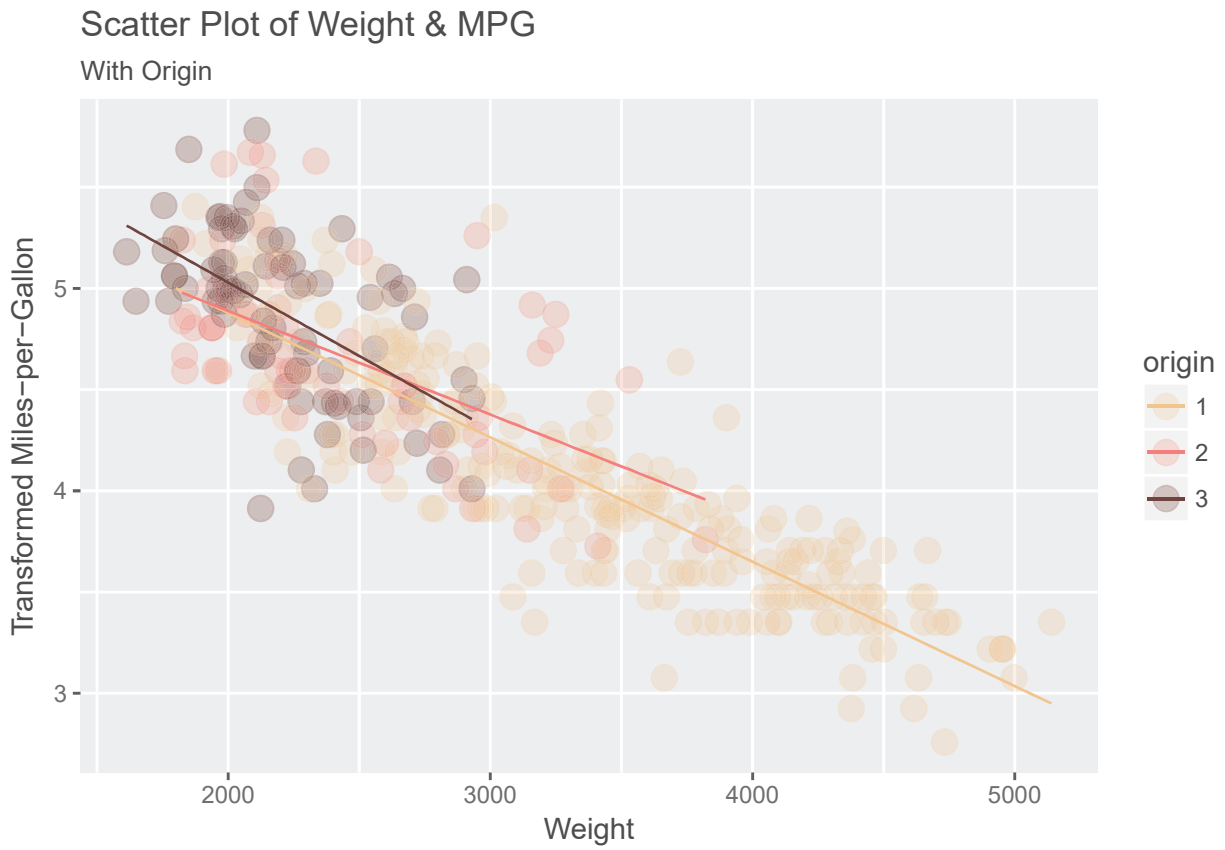


4.2.7 MPG,Weight and Origin

Looking at the below given plot,

- Cars with `origin 2` and `origin 3` are light-weight vehicles
- Cars from `origin 1` range from light to heavy
- With increase in `weight`, `MPG` decreases regardless of the origin of the car

```
ggplot(ampg.2, aes(x = weight, y = tmpg, color = origin)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  scale_color_manual(values=wes_palette(n=3, name="GrandBudapest")) +
  labs(title = "Scatter Plot of Weight & MPG", subtitle = "With Origin",
       x = "Weight", y = "Transformed Miles-per-Gallon")
```



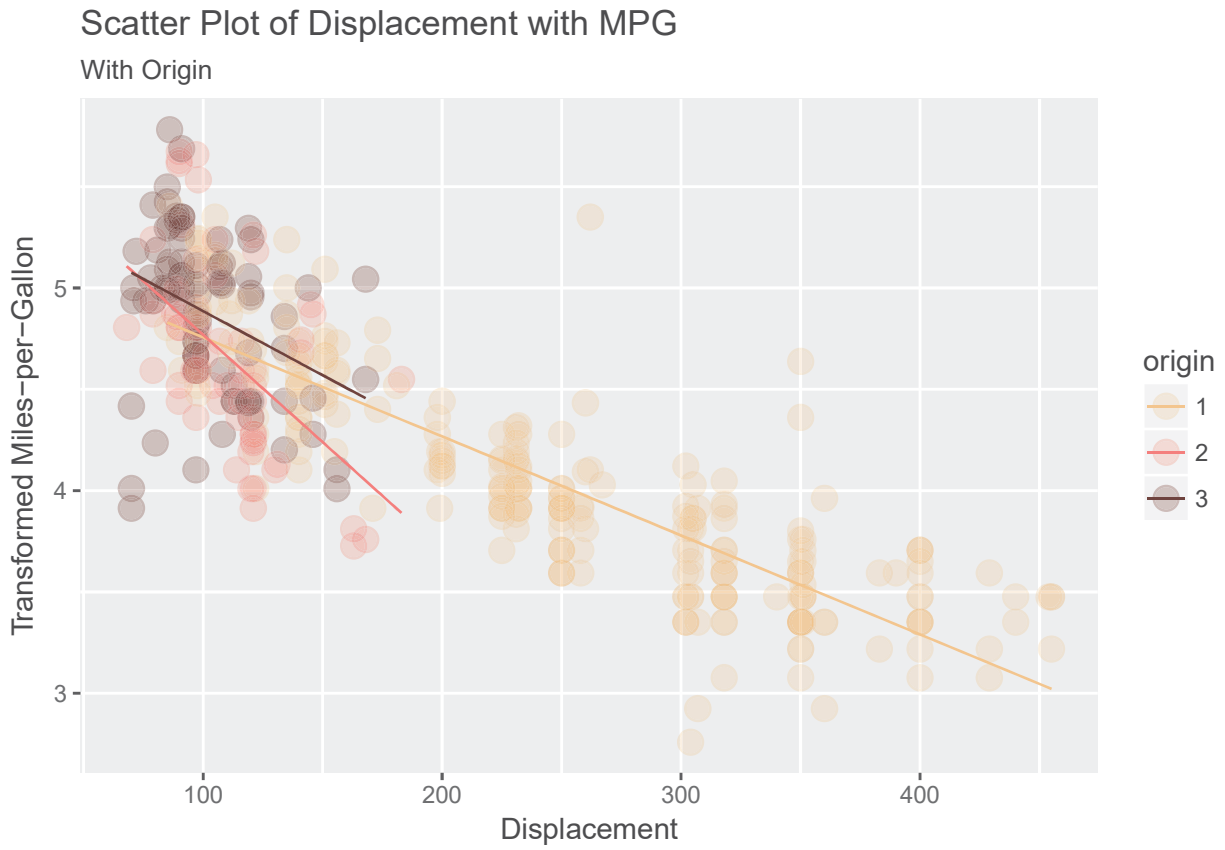
4.2.8 MPG, Displacement and Cylinders

Downward trend of `displacement` with `MPG` is as expected.

But, here we observe:

- The rate of decrement in `MPG` with increase in `displacement` is highest for cars from `origin 2`
- The rate of decrement being lowest for cars from `origin 1`, we have instances only from `origin 1` where the range of `displacement` is from less than 100 to more than 400.

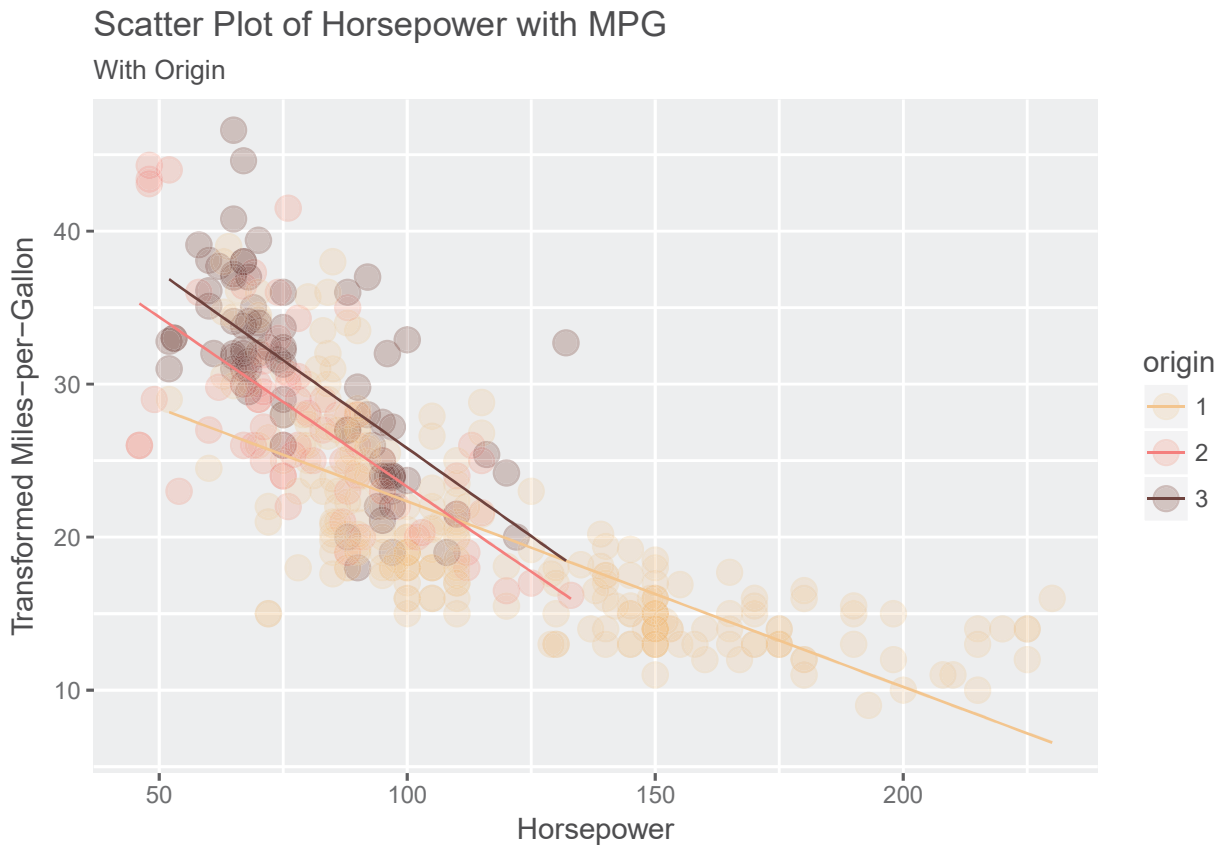
```
ggplot(ampg.2, aes(x = displacement, y = tmpg, color = origin)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  scale_color_manual(values=wes_palette(n=3, name="GrandBudapest")) +
  labs(title = "Scatter Plot of Displacement with MPG",
       subtitle = "With Origin", x="Displacement",
       y = "Transformed Miles-per-Gallon")
```



4.2.9 MPG, Horsepower and Origin

We observe an interesting behavior here, the rate of decrement of `MPG` with horsepower is almost identical for cars from `origin 2 & 3`, being higher than the cars from `origin 1`.

```
ggplot(ampg.2, aes(x = horsepower, y = mpg, color = origin)) +
  geom_point(size = 4, alpha = 0.2) +
  geom_smooth(se=F, method = "lm") +
  scale_color_manual(values=wes_palette(n=3, name="GrandBudapest")) +
  labs(title = "Scatter Plot of Horsepower with MPG",
       subtitle = "With Origin", x = "Horsepower",
       y = "Transformed Miles-per-Gallon")
```



5 SUMMARY

For the target feature, we transformed the `MPG` using Box-Cox method of maximum likelihood to make the distribution approximately normal. We used the transformed feature to further analyse our descriptive features. For categorical feature, we not only explored the distribution through bar chart, we also used histogram of `mpg` and box plot in each category to gain better understanding of the distribution.

Through exploratory and visual analysis, we found that `weight`, `horsepower` and `displacement` are negatively correlated with `mpg` which essentially means that `Miles-per-Gallon` decreases with increase in these parameters. However, we also saw that as we move forward from 1970 to 1982, `mpg` of these cars increased indicating towards technological advancement. We also observed that the `mpg` of 4 cylinder cars is generally higher other type of cars. Other thing to point out is that number of instances for `origin 1` cars is more than 2 & 3 indicating that `origin 1` produces more cars.

Hence, we can say that `weight`, `horsepower`, `displacement`, `origin`, `model year` and `cylinders` could be our predictive feature. We would like to explore this further in Phase II, where we will also work on feature selection and predictive modeling.