Group 12

Mullagan sean , shishir kumar, tejas kulkarni

Final project

Final report: This should be a professionally prepared report that addresses the following parts: cover page, executive summary, project motivation/background, data description, your BI model, diagrams, your findings and managerial implications/conclusions, references

**Executive Summary:**

Predicting manufacturer’s suggested retail price has always been a challenge for rise or decline on sales of cars in automobile sector. Retail customers mostly assumes that luxury cars with in-build high tech features called fully loaded would have high retail price than as compared to base model with 4 cylinders model engines. There are many more differences which brings plethora of variations in retail price. So to analyze perfect price and segmentation of cars our business intelligence model is certainly going to help our stakeholders to mark best prices in this competitive world and definitely going to increase sales.

To analyze and get best price prediction we are using past dataset which is easily available on (<https://www.kaggle.com/CooperUnion/cardataset>). This dataset has sixteen variables, car brand(make), model, engine type, year of manufacturing, number of doors, fuel type, transmission(manual or automatic), market category, driven wheels, vehicle size, vehicle style, highway mileage, city mileage, MSRP(Manufacturer’s suggested retail price). It has been notice that whenever a customer think about purchasing, he/she thinks most of these feature, but when it comes to price, they have to think twice. So this model is going to explore best price which will not only raise profit percentage but customer satisfaction related to price and features is also one the major aspect.

Dataset collected from the reference site is raw data needed to be mined before making analysis and prediction. Certain have been run to cleanup data then converted data as factor to work on R. After running certain codes available in R file some of the awesome conclusions have been drawn at last. Segmentation of data using clustering analysis has worked out to make decisions, as segmentation helps manufacturers to decide production of which grade of cars has to be increased and as more production can lead to drop of prices.

**Project motivation/background:**

We wanted a wide spread challenging dataset to work with. So when we found the “mtcars” data set it had a variety of categorical and numerical variables which makes it possible to have so many meaning full inferences and is also a challenging dataset to work on because of its range of values. Hence, we thought we could apply our selves on this data and that was our motivation.

**Data description:**

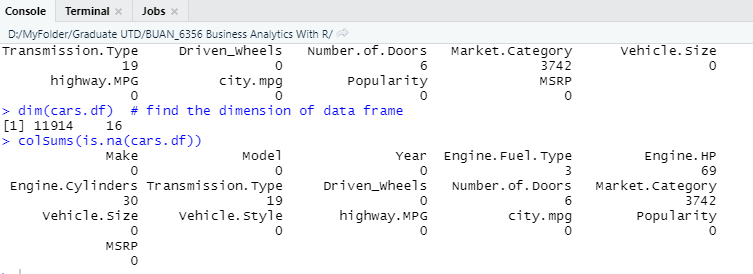
The Exploratory data analysis of our data set:

We used R-Studio to perform the EDA on our data set (Refer to the R-codes )

Dimensions: Number of Columns: 16

Number or Rows: 11914

The missing values in the data set:



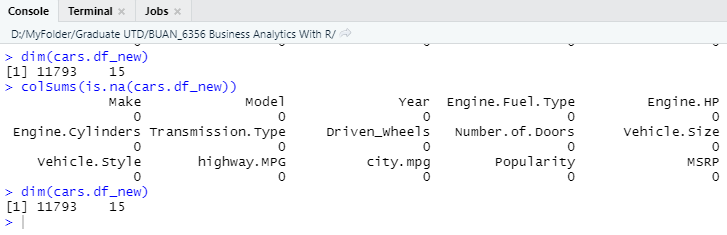
So, the attributes Engine.Fuel.Type , Engine.HP, Engine.Cylinders , Transmission.Type, Number of doors, Market.category, have missing values but compared to the 11914 rows it is very small. Hence, we will drop these records.

But on the other hand, attribute Market.Category has 3742 missing values which is a considerable amount of data to lose.

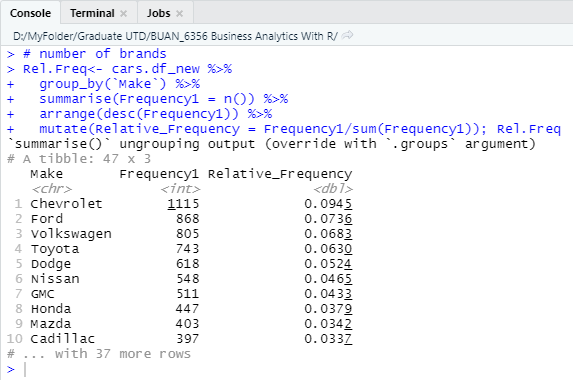
This attribute is a Market segmentation attribute derived from as per the consumer needs and we don’t want a pre-defined attribute in our exploration, hence we will drop the entire attribute.

Dimensions after dropping: Number of Columns: 15

Number or Rows: 11793

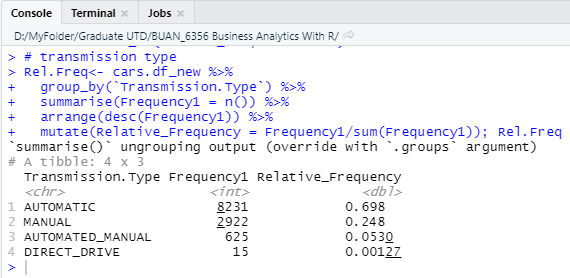


Looking at the brands here we have got 47 rows as grouped by make. Since this has 47 rows therefore cannot be categorized. Hence we be dropped.

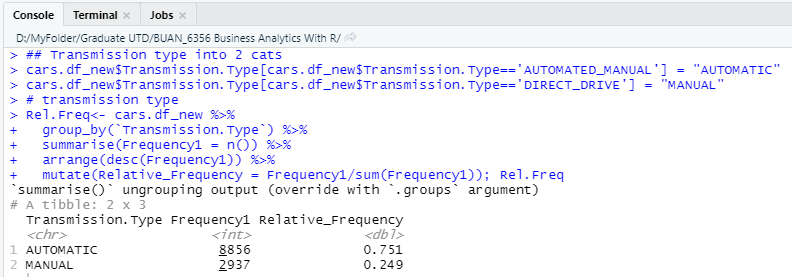


Some of the data has very few entries there they have been upgraded or merged with rest of the entries. Reducing as such would reduce categorical data but it’s not going to affect model as number of rows remain same.

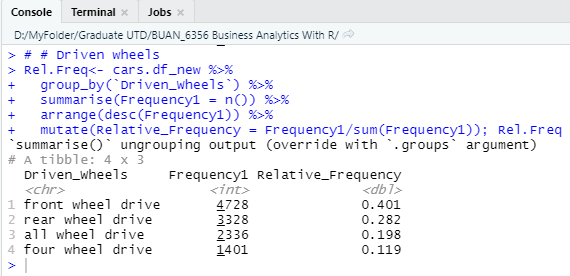
Considering Transmision type data



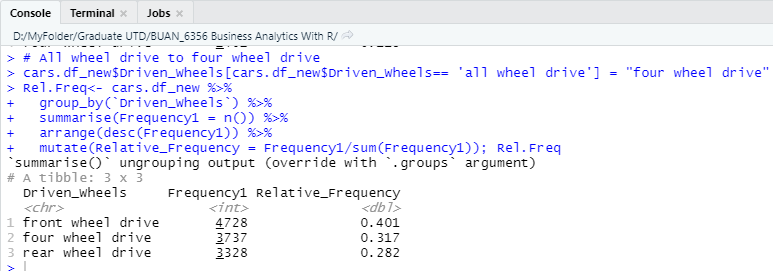
Here we can see Automated\_Manual and Direct\_Drive has less number of rows hence updated with Automatic and Manual.



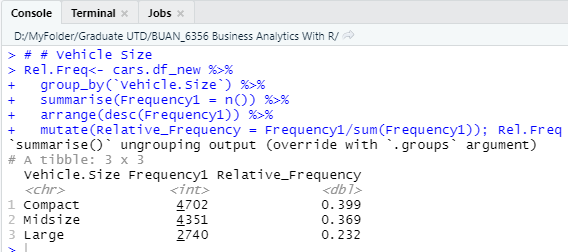
Driven wheels



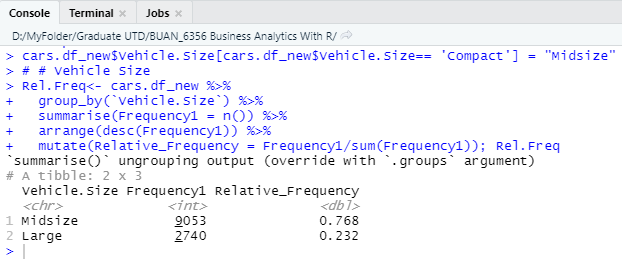
All wheel drive has been converted into four wheel drive



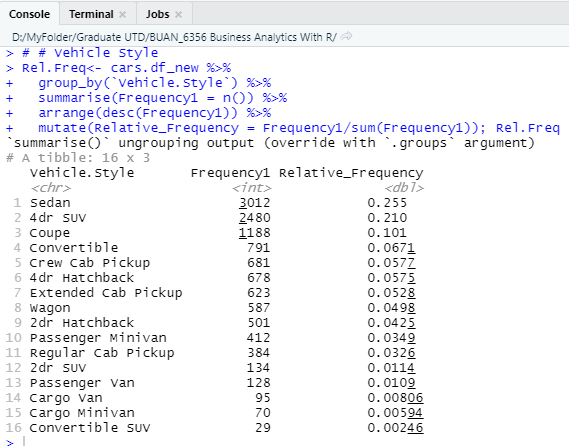
Size



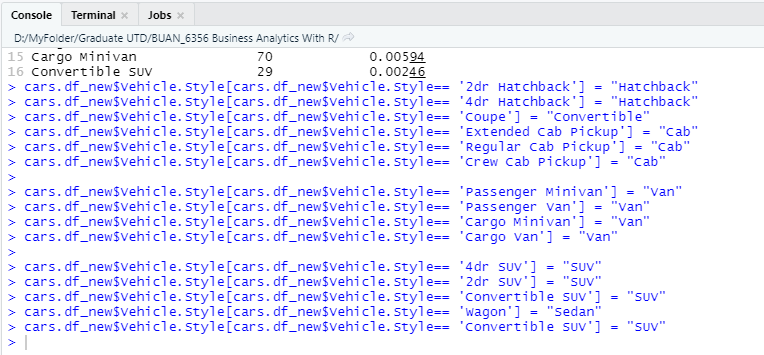
As compact and Mid-size is approximately same so renamed compact-size as Midsize

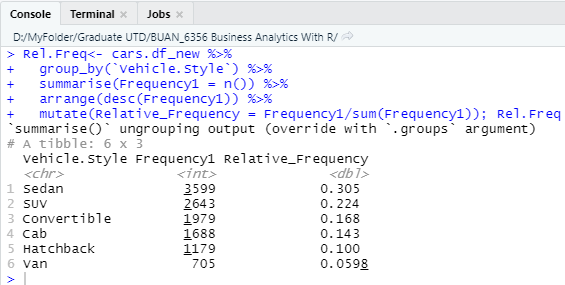


Vehicle Style

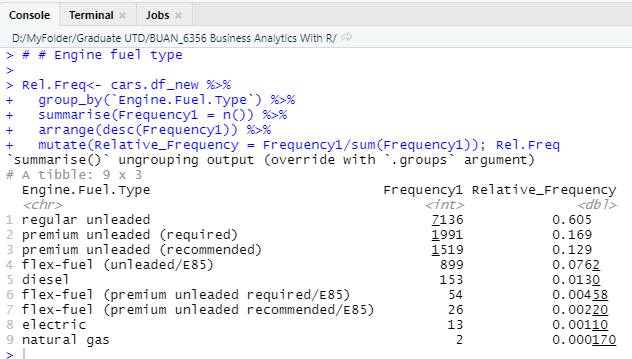


Updated Vehicle Style

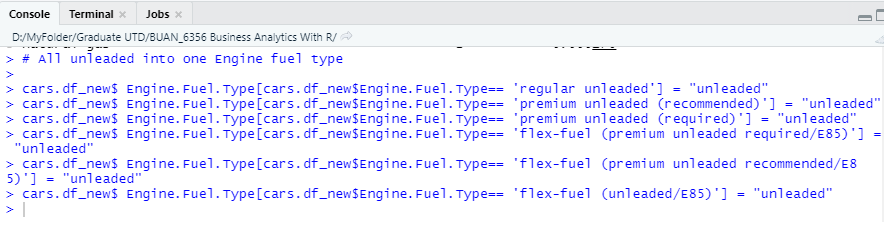


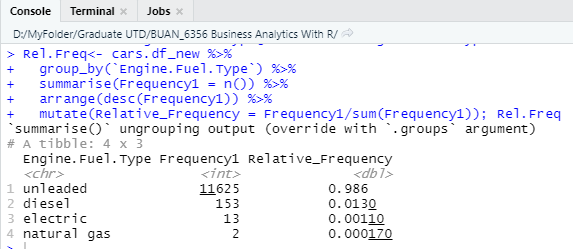


Engine

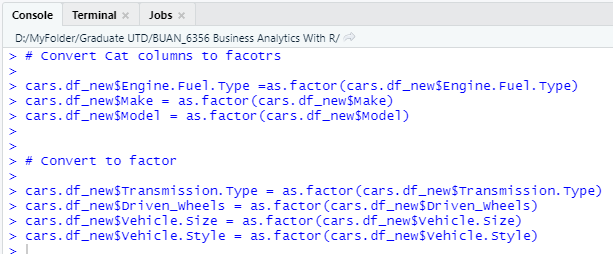


Updated Engine

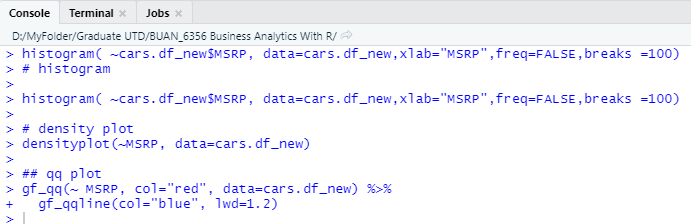




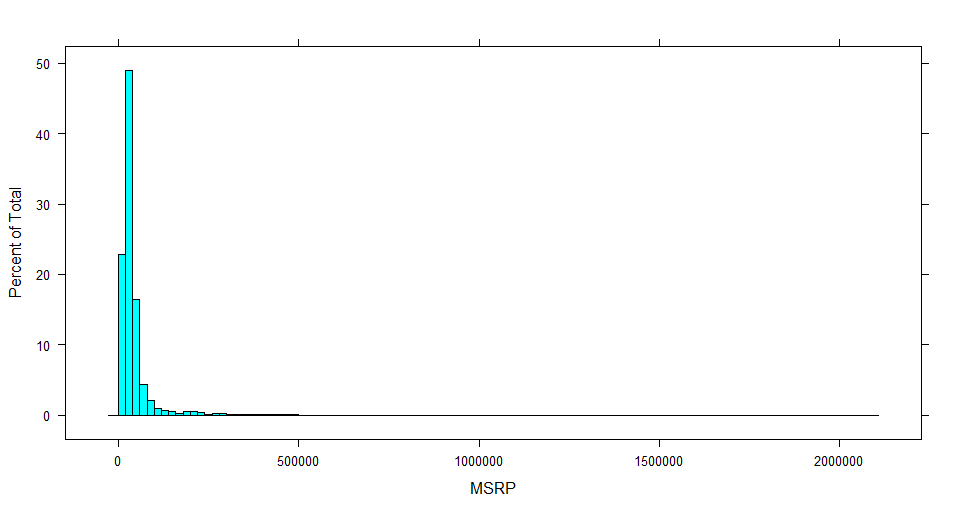
Converting all categorical data into factor for analysis.



Now that the data is clean and mutated as per requirement, we can plot the target variable (MSRP) to analyze its distribution

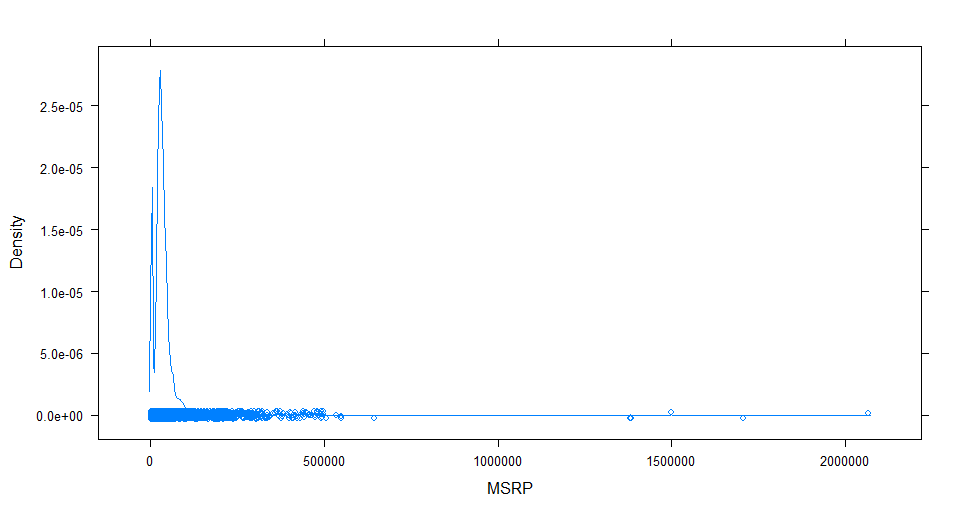


Histogram of MSRP



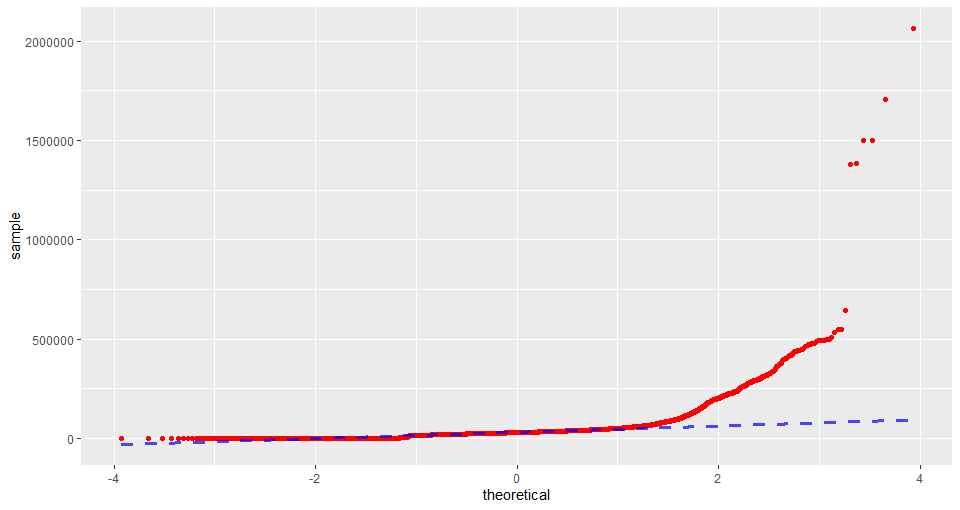
We can see from above graph that the Target Variable MSRP is heavily positively skewed.

Density plot of MSRP

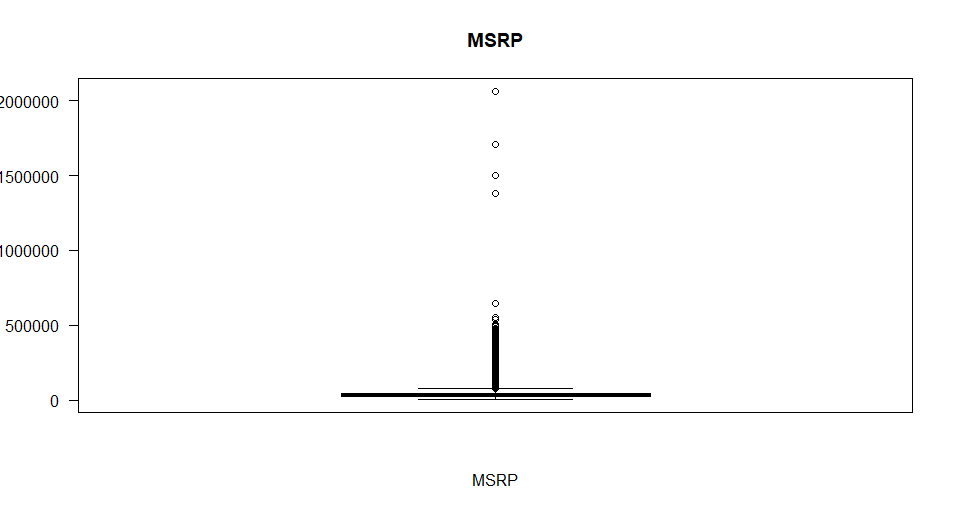


The same is confirmed by the density plot showing that the Variable is positively skewed and must be because of the presence of outliers.

QQ Plot of MSRP

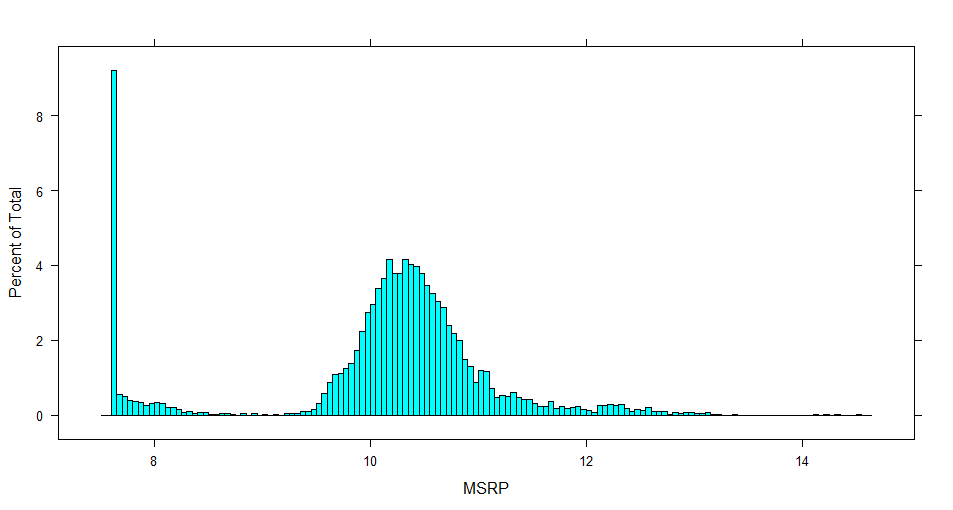


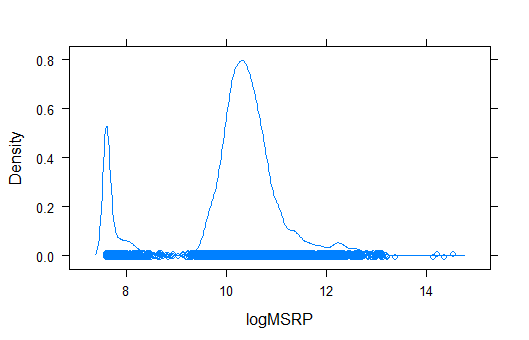
It clear now the data is normally distributed but is skewed because of the presence heavy outliers.

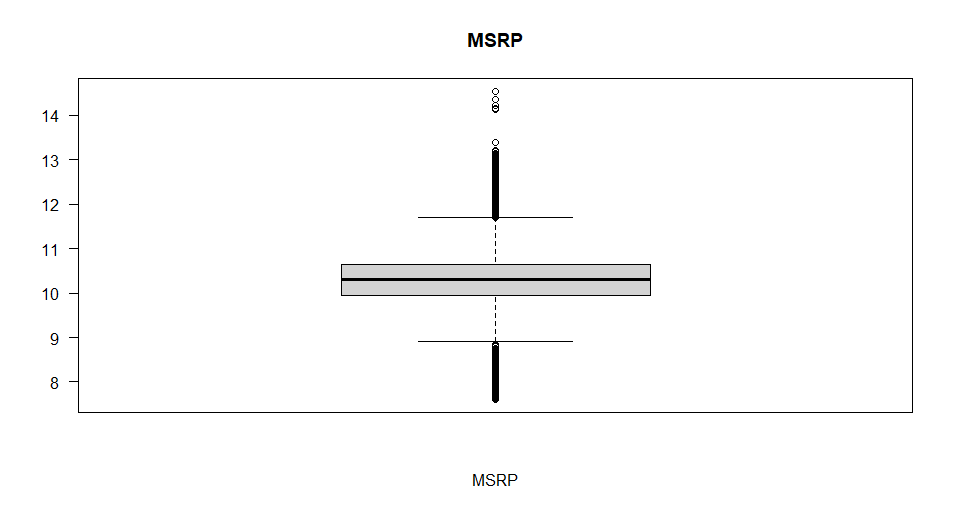


The above plots show the behavior of the target variable MSRP. The distribution of the variable is heavily positively skewed due to the presence of large number outliers which will need to be treated. To treat this outliers we can transform our data to log which will make it easier to analyze.

Data after log:







As seen from the above graphs the target variable has a bell shaped more normal curve. But it has two bell shapes.

**BI model:**

BI model is based in linear regression analysis to predict future values. In-build library function is applied here are:.

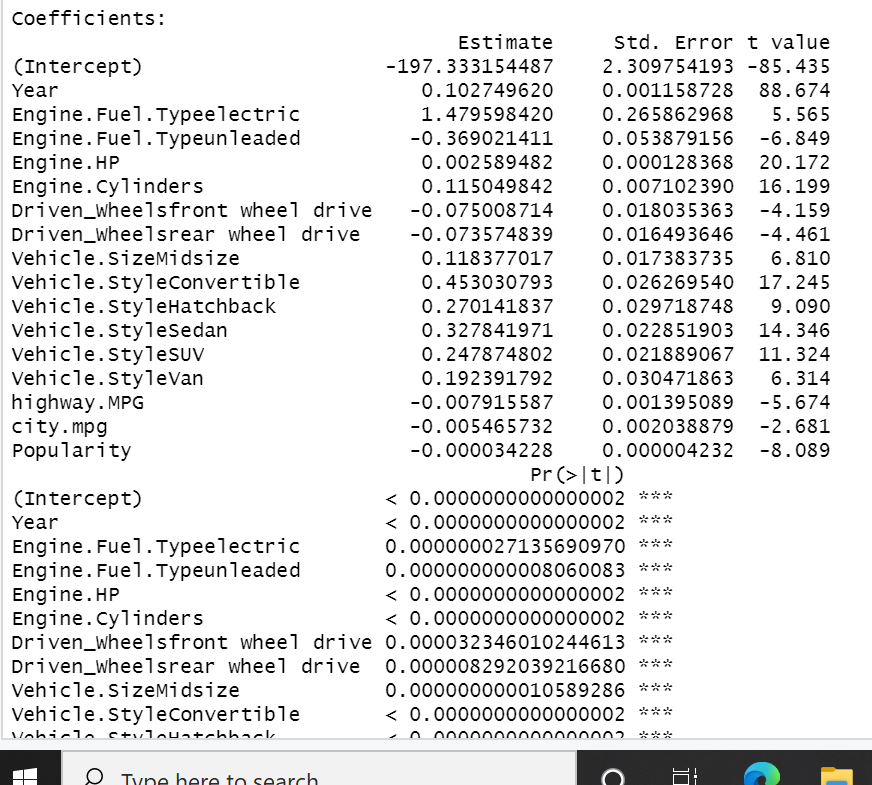
library(rpart)

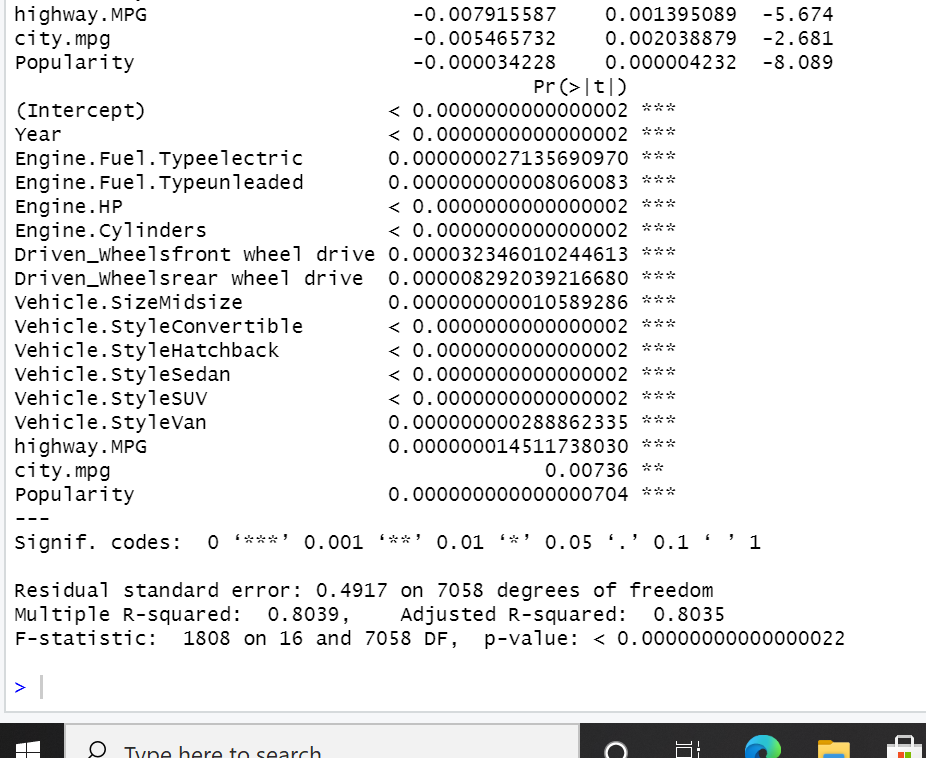
library(rpart.plot)

library(caret)

library(forecast)

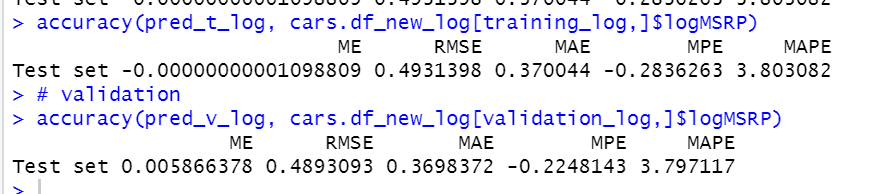
To make analysis of cleaned dataset, it has been divided randomly into training dataset, considering 60% of data, on which model will be build and trained and validation dataset to test and can be implemented to predict future values.





After preforming the regression on the logged data we have 16 independent variables with their respective coefficients. The coefficients for Engine.Fuel.Typeunleaded, Driven\_Wheelsfront wheel drive, Driven\_Wheelsrear wheel drive, highway.MPG, city.mpg, Popularity have negative coefficients which means that these factors when increases by 1 unit reduce the MSRP as per they coefficient value.

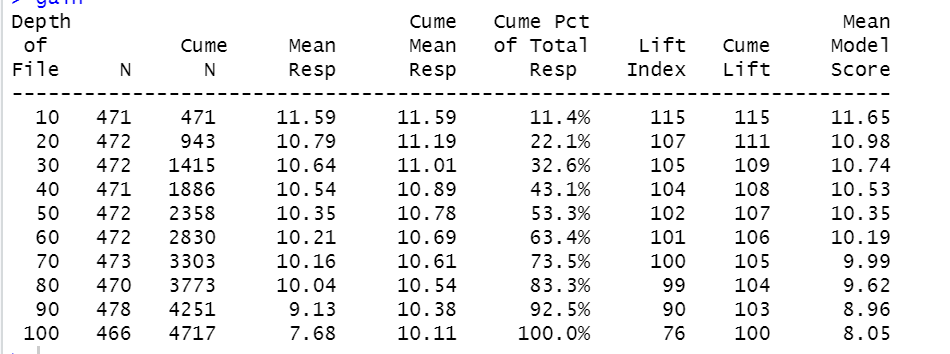
The R^2 for the model is 0.8039 and Adjusted R-squared is 0.8035 shows that the model is a good fit.



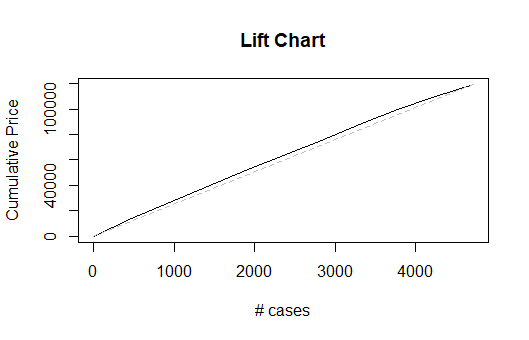
We can see the difference between RMSE score of predicted values and actual values on the train data set and the validation data set which is, **0.0038**.

Difference in Mean Absolute error : **0.00024.**

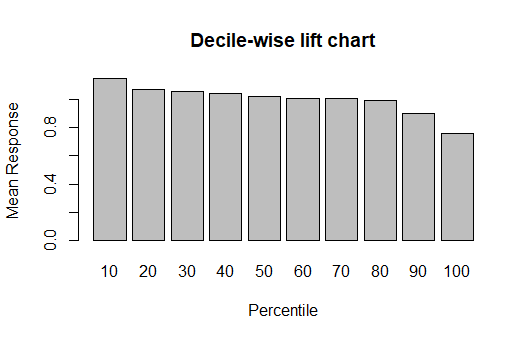
If we analyze gains in the model,



we have following results which can be interpreted through lift chart:



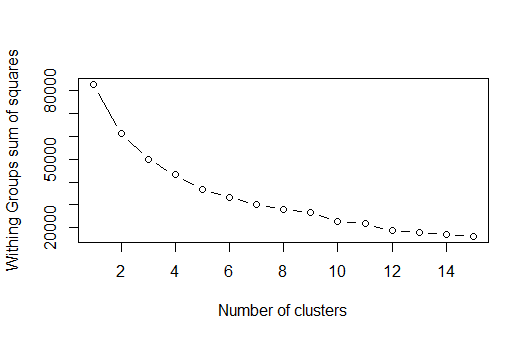
From the lift chart it can be observed that there is some Area between the Lift line and base line hence our model is better than having no model.



From the decile chart we conclude that the first decile has the highest cars with the highest avg MSRP which keeps on reducing as deciles progress. Which shows a good fit model.

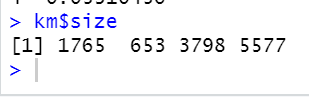
**Let’s see how clustering is going to enhance our model and help us to get some valuable business intelligence.**

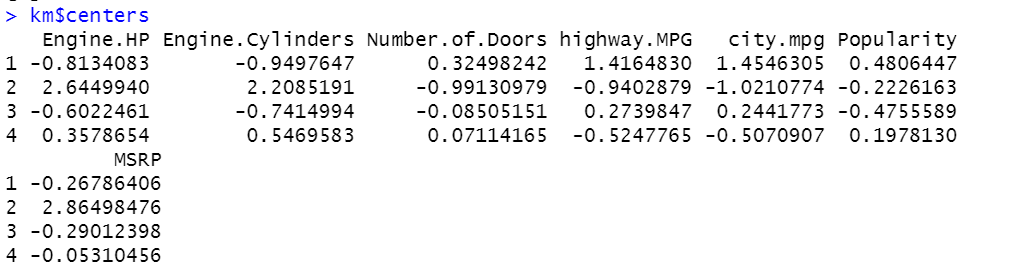
**We run a function to plot an WSS (elbow) graph to decide the optimum number of clusters.**

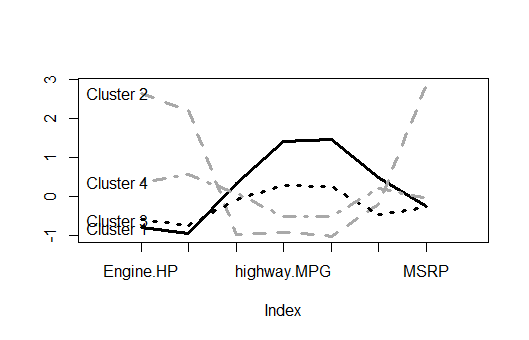
****

The graph bends at 4 (forming an elbow hence elbow graph) and that will be our optimum number of clusters.

Now we perform the K-means clustering and the achieve the following results,







From the above graph of we can see that the cluster 4 has higher specifications but not as much higher in price than cluster 3 and 4.

Cluster 1 has the lowest specification in terms of engineHP and Engine cylinders but the highest Highway and city mileage and the lowest price.

Cluster 3 is between 1 and 4 in which it has lower specs than 4 but higher specs than 1 but is priced the lowest.

The cluster 2 cars are the outliers where they have the highest specs lowest Mileages (highway and city) and really high MSRP.

This data could provide business intelligence for car rental companies to decide how many and what kind of cars to buy as per their requirements and will be able to get a much better overview to manage their business.

References:

Dataset : (<https://www.kaggle.com/CooperUnion/cardataset>).

Wss plot Function : <https://www.youtube.com/channel/UC519jKQcifYxG16822J3e_w> ( Easy ML )