FINDING PATTERNS IN DATA

AND EDA

WEEKLY ASSIGNMENT

NORTHEASTERN UNIVERSITY - SILICON VALLEY

SUBMITTED TO:

DR. VENKATA DIVVURI

(PROFESSOR)

SUBMITTED BY:

ANUBHAV RASTOGI

(STUDENT)

INTRODUCTION

Data mining is the process of finding patterns in the large data sets. It is the process of predicting the outcomes on the basis of discovered patterns and the correlations between various factors or variables. It can be performed through various techniques such as regression, clustering, principal component analysis, etc. In the present assignment we have performed data mining through regression and have also done the exploratory data analysis to understand and analyze the dataset. In the present assignment, we have used Car Sales dataset which contains 157 entries along with 15 column variables. The dataset contains the information about different car models manufactured by different car companies and includes variables such as pricing, sales, resale value, vehicle launch date and other car specifications. The dataset has been taken from Kaggle and can be downloaded from this link.

ANALYSIS

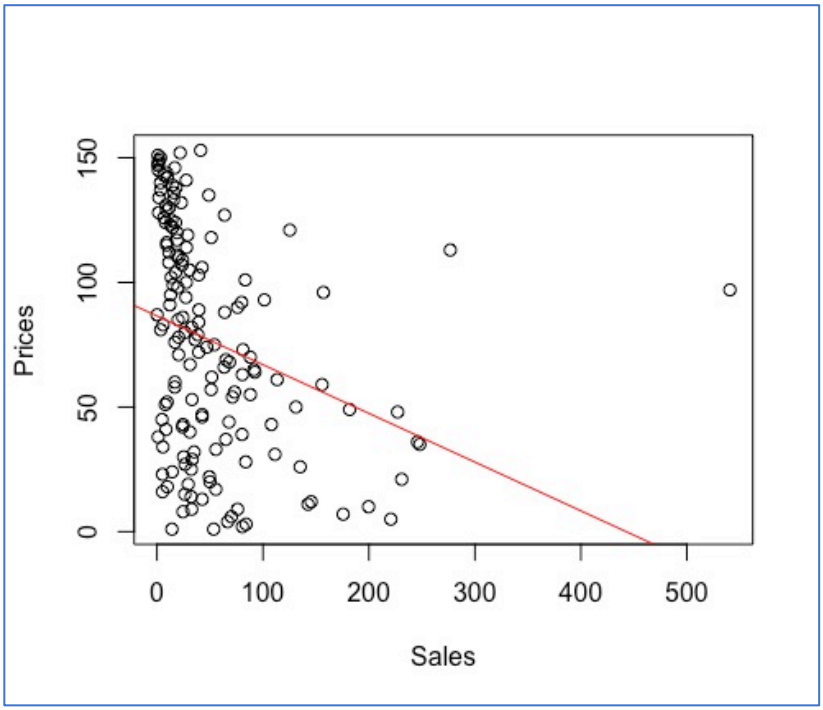
In the present assignment, we have used car datasets because the data set contains categorical variables as well as numeric variables. Further, the dataset highlights interesting specifications about different cars along with its pricing, resale value, etc. which helps a potential buyer in narrowing down his search with respect to purchase vehicle. Also, the dataset contains the sales record or sales count of each vehicle which helps in identifying the trending vehicle or the vehicle most liked by people.

FINDING PATTERNS IN DATA AND EDA

Firstly, we have imported the Car sales dataset. In order to read the csv file, we have used the command **read.csv( )**. After importing the dataset, we have printed the head so as to see first 5 rows of it. The dataset contains a few factor variables which we have converted into numeric variables. Following command has been used to do the same.

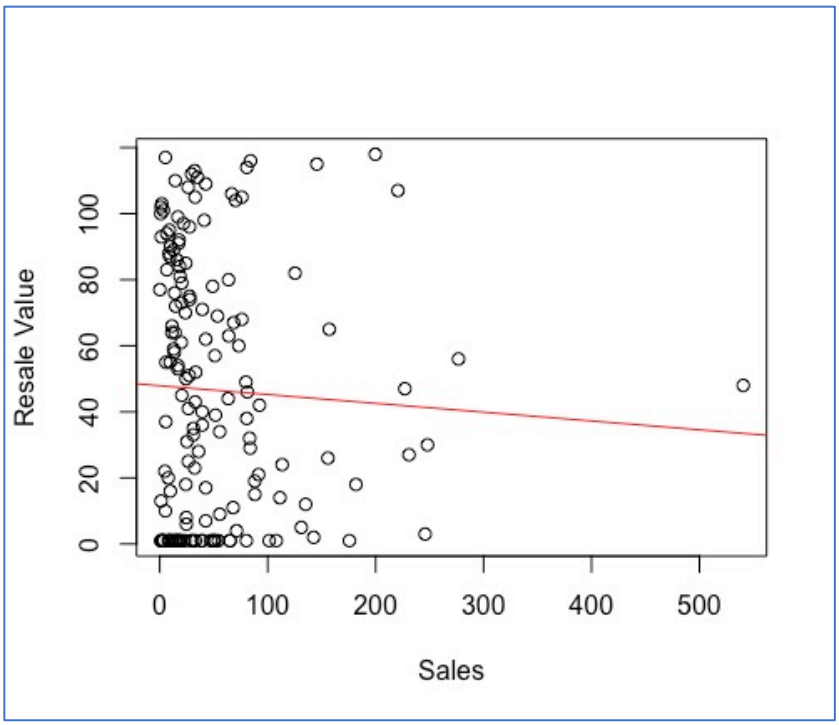
```
# Converting factor variable into numeric variable
mydata[, c(4,6:14)] <- sapply(mydata[, c(4,6:14)], as.numeric)
head(mydata)
```

Next, we have performed data visualization through scatterplots and box plots. With the help of scatter plots, we have figured out the correlation between Prices and the sales of cars and also between the sales and the resale value of vehicle after 4 years.



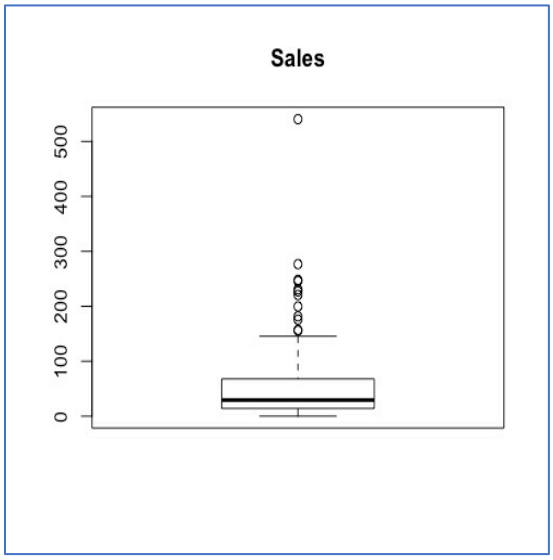
It can be observed from the graph that when prices increase, sales decrease and vice-versa. Thus, there is a negative correlation between the two.

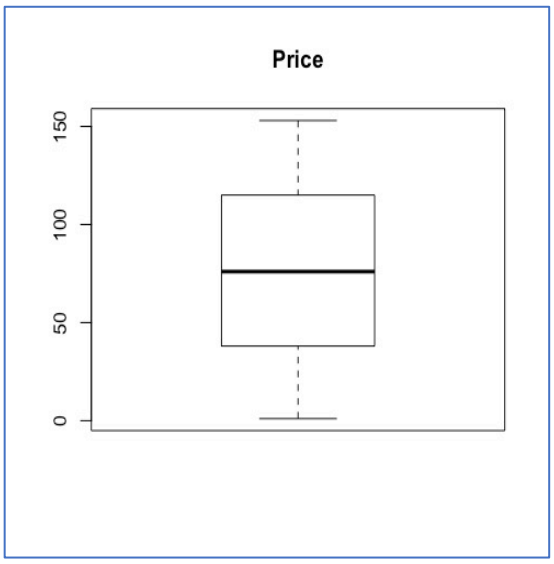
FINDING PATTERNS IN DATA AND EDA



The above plot depicts the correlation between the Resale value of car and its sales. It can be observed that there exists a weak negative correlation between the two. It implies that the resale value of vehicle is not much affected by its sales.

We have also created the box plots of sales and the prices which are as follows:



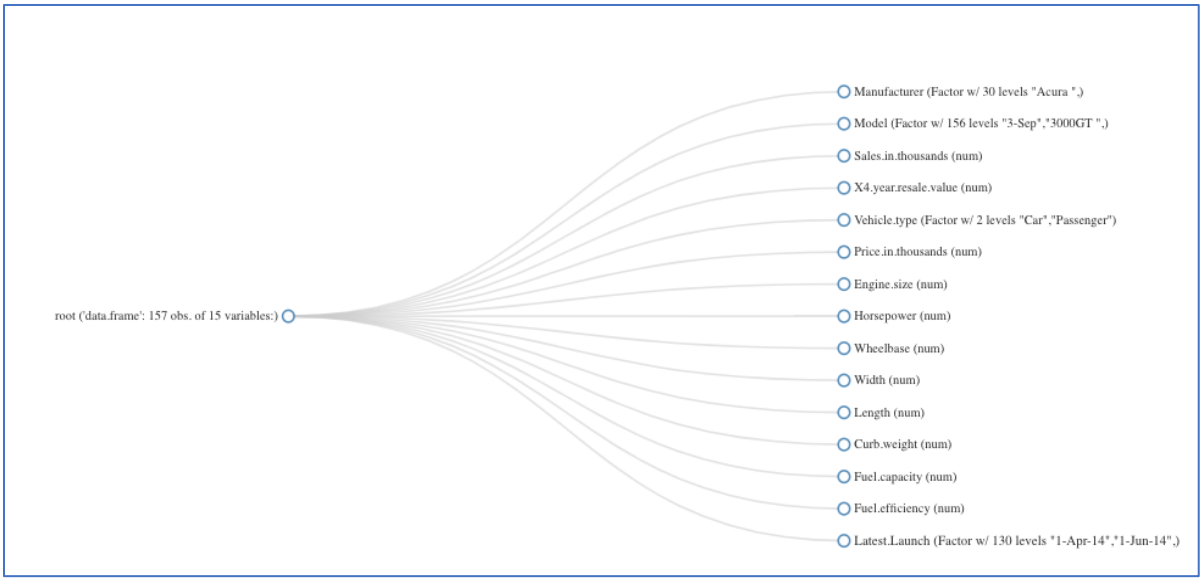


FINDING PATTERNS IN DATA AND EDA

It can be observed from the above graphs that there exist a lot of outliers with respect to the sales of the vehicles, but the Price box plot represents an even distribution. Outliers in sales box plot implies that there are a few vehicles whose sales has been so much as compared to other vehicles that it has resulted in the formation of an outlier in the box plot.

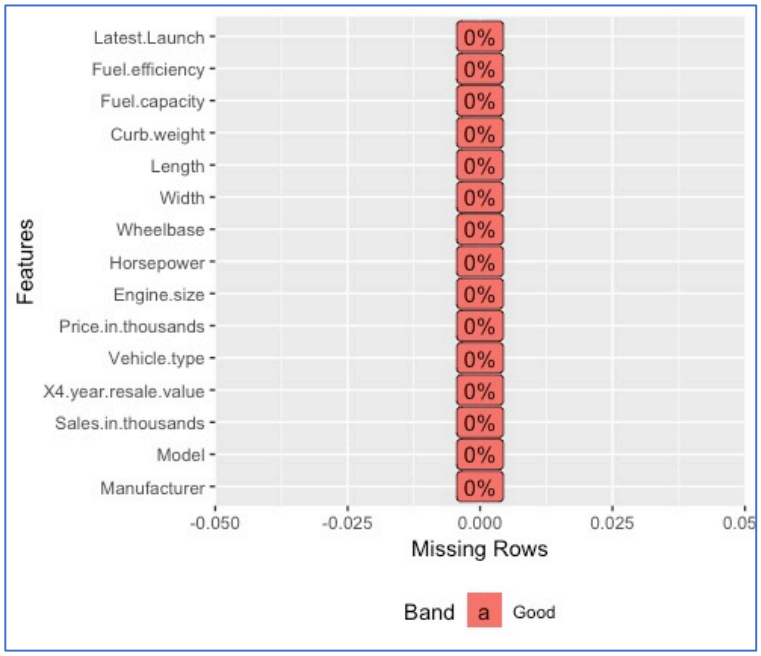
We have also performed data normalization. It is one of the important process which aims at removal of the redundant data. The process of Data Normalization improves the data integrity and is done by subtracting the mean and dividing it by the standard deviation.

In order to further explore the data, we have used a library known as Data Explorer library. It helps in easy EDA (Exploratory Data Analysis). EDA refers to the process of analyzing and summarizing the data sets as per their characteristics and often involves visual methods. It is a process of carefully observing the data and what the data tell us beyond modeling or hypothesis testing. With the help of Data Explorer library, we have created various charts which are as follows:

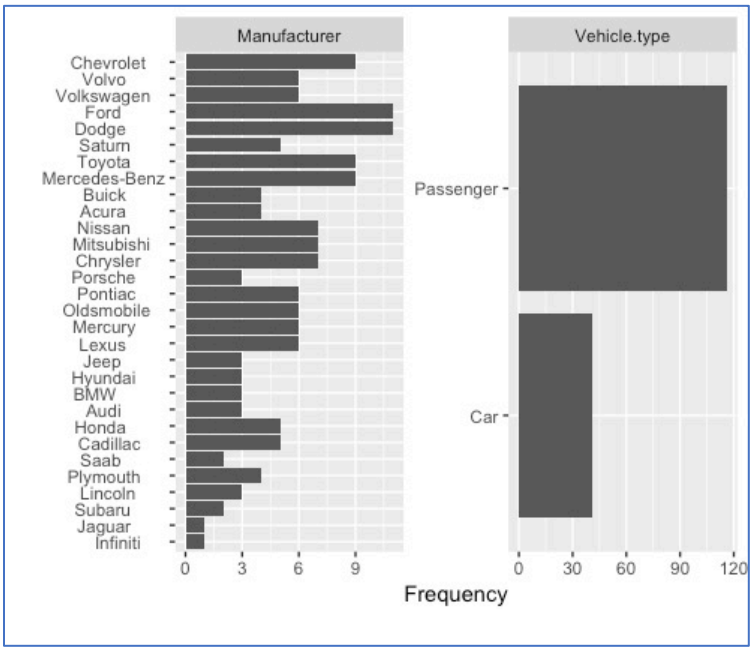


FINDING PATTERNS IN DATA AND EDA

The above chart tells us the data dimension and shows the continuous and categorical variables present in the dataset. The following chart explains us whether there are any missing values in the input dataset. It can be seen that there are no missing values and the dataset is good.

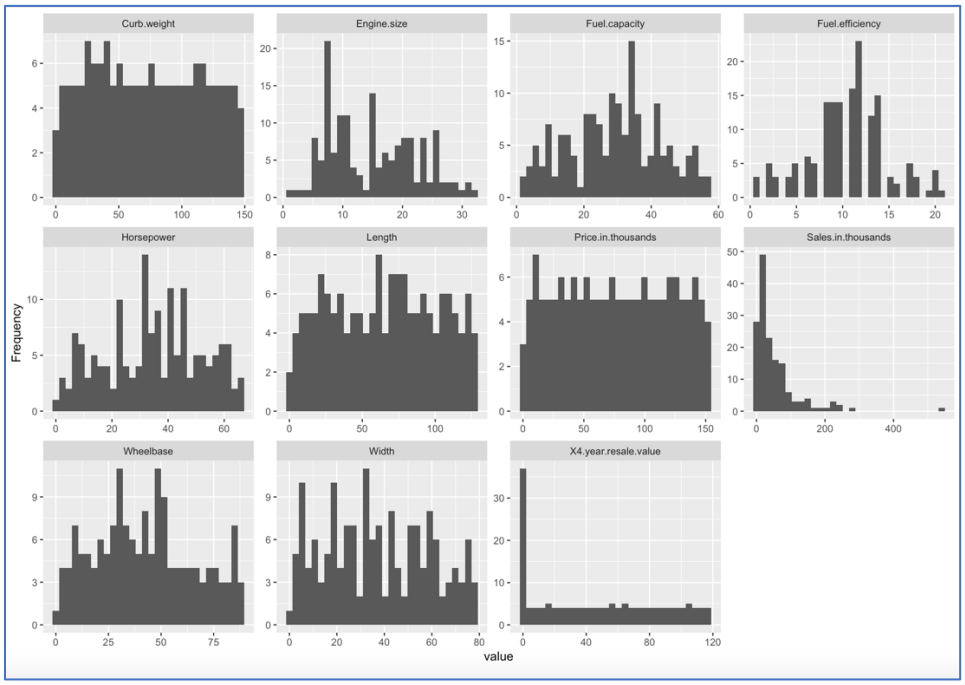


Following chart represents the bar plots for categorical variables.

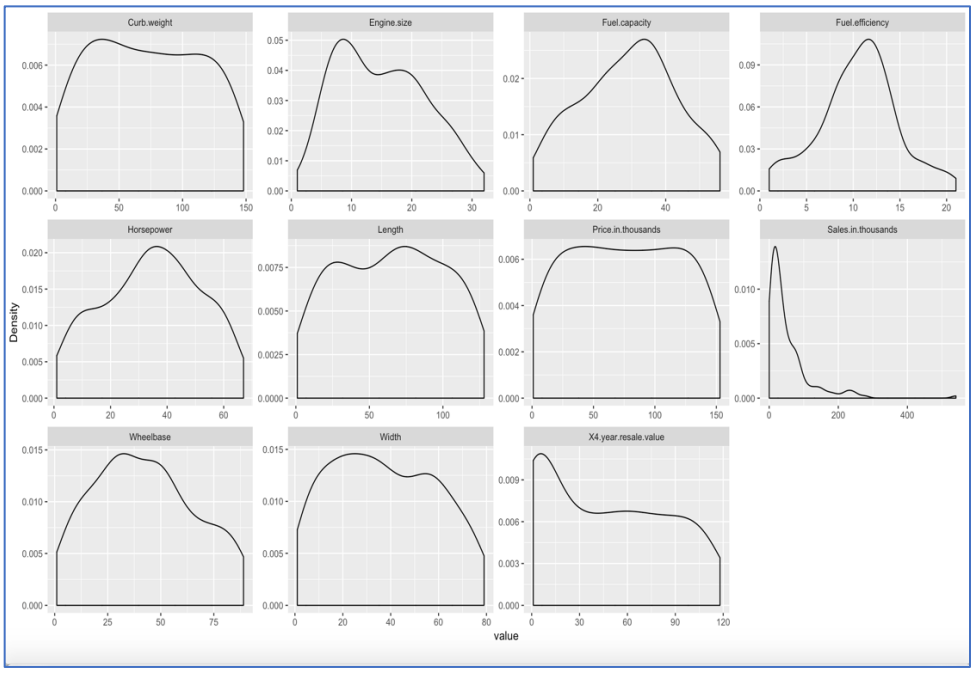


FINDING PATTERNS IN DATA AND EDA

We have also visualized histograms and density plots for continuous variables. Following two charts represents the same.



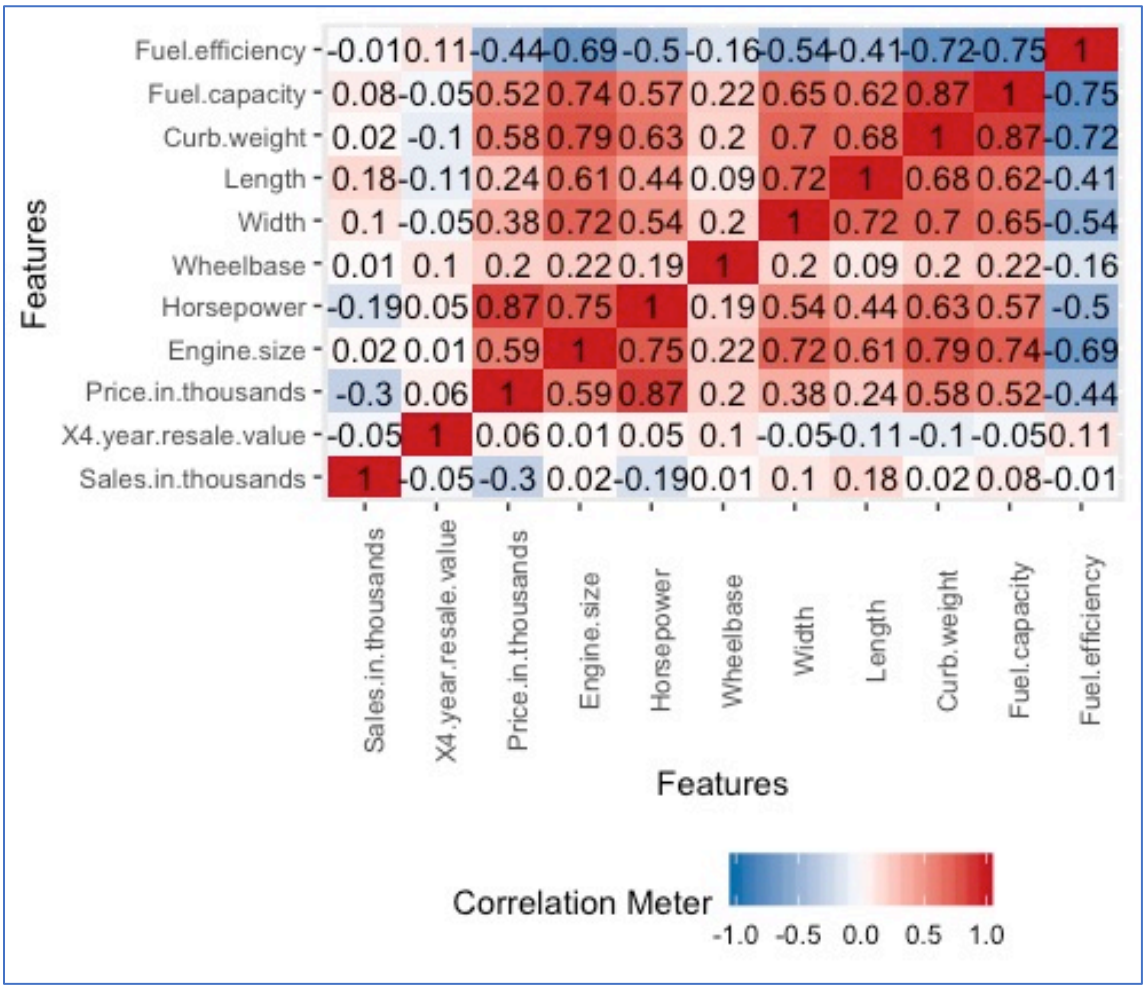
Histograms



Density Plots

FINDING PATTERNS IN DATA AND EDA

Following is a multivariate chart representing the correlation analysis of different variables.



After doing the Exploratory analysis, we have performed regression. In the present assignment we have performed firstly the linear regression and have tried to predict whether the change in price brings about a change in the sales. The following chart represents the console output while performing the linear regression:

```
Call:
lm(formula = Price.in.thousands ~ Sales.in.thousands, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-82.770 -36.818  -1.367  37.180 116.181

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        86.52975    4.33955  19.940  < 2e-16 ***
Sales.in.thousands -0.19556    0.05042  -3.878 0.000155 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.84 on 155 degrees of freedom
Multiple R-squared:  0.08846,   Adjusted R-squared:  0.08258
F-statistic: 15.04 on 1 and 155 DF,  p-value: 0.000155
```

We have also performed, logistic regression and have predicted two things: firstly, we tried to predict the change in resale value with respect to the change in sales and secondly, we have tried to predict the effect of Engine size, fuel capacity and various other specifications on the Prices. Following the screenshots of the console after running the logistic regression.

```
Call:
glm(formula = X4.year.resale.value ~ Sales.in.thousands, data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-46.879  -39.764   -3.429   33.599   75.446

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        47.90412    3.93062  12.187   <2e-16 ***
Sales.in.thousands -0.02679    0.04567  -0.587    0.558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1505.838)

    Null deviance: 233923  on 156  degrees of freedom
Residual deviance: 233405  on 155  degrees of freedom
AIC: 1598.3

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = Price.in.thousands ~ Engine.size + Horsepower +
    Wheelbase + Width + Length + Curb.weight + Fuel.capacity +
    Fuel.efficiency, data = mydata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-111.713   -10.619     1.203    11.339    50.517

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.13342   11.37495  -0.363 0.716840
Engine.size     -1.18200    0.45356  -2.606 0.010095 *
Horsepower       2.41131    0.13821  17.447  < 2e-16 ***
Wheelbase        0.06372    0.06814   0.935 0.351249
Width           -0.10692    0.12015  -0.890 0.374960
Length          -0.31884    0.06884  -4.631 7.89e-06 ***
Curb.weight      0.29642    0.08594   3.449 0.000733 ***
Fuel.capacity    0.34820    0.24627   1.414 0.159485
Fuel.efficiency  0.44526    0.61457   0.725 0.469896
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 376.1433)

    Null deviance: 312108  on 156  degrees of freedom
Residual deviance:  55669  on 148  degrees of freedom
AIC: 1387.3

Number of Fisher Scoring iterations: 2
```

CONCLUSION

Thus, on the basis of the analysis, it can be concluded that data mining is the process of finding patterns in the data. It is very essential to perform data mining otherwise it will lead to inaccurate results thereby making the false predictions. Therefore, proper attention should be given to the input data while building the model and to do this, EDA should be done prior to it. With the help of DataExplorer library, it has become easy to perform Exploratory Analysis of the input dataset. In the present assignment also, we have performed EDA and created visualizations which has helped in understanding the dataset and then performed the regression to make the predictions.

REFERENCES

- Amr. (February 16, 2018). Simple Fast Exploratory Data Analysis in R with DataExplorer Package. Retrieved from https://towardsdatascience.com/simple-fast-exploratory-data-analysis-in-r-with-dataexplorer-package-e055348d9619

- Harshit Sinha. Car_sales. Retrieved from https://www.kaggle.com/hsinha53/car-sales/version/1#Car_sales.csv%20(https://www.kaggle.com/hsinha53/car-sales/version/1