Car Price Prediction using Linear Regression

ARNAB ROY

Department - Statistics

Roll No.- 411

Session - 2019-2022

Supervisor- Dr. Ayan Chandra

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Arnab Roy

Signature of the Student

Contents

1 Introduction

The term *regression* was first introduced by Francis Galton.In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" towards the average height in the population as a whole.1 In other words, the height of the children of unusually tall or unusually short parents tends to move toward the average height of the population.

The modern interpretation of regression is, however, quite different.Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.

In regression analysis, there are few assumptions-

1. The regression model is assumed to be linear in the parameters.

2. the linear regression analysis requires all variables to be multivariate normal.

3. linear regression assumes that there is little or no multicollinearity in the data.

4. The regression model is assumed to be homoscedastic.

So, before fitting a regression model, we have to modify our dataset by cleaning or transforming the variables.

2 Statement of the problem

A Chinese automobile company *Geely Auto* aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends.Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market.Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

In this project, we will use this dataset and analyse the relationship between the price of the cars and its features and draw conclusion based on the results.

3 Objective

Given the data in hand, the primary objective is to find –

- Which variables play significant role in predicting the price of a car.
- How well those variables describe the price of a car.

In other words, here our main objective is to extract maximum knowledge from this data by finding the variables that are significant in determining car price and to fit a regression model using that variables to predict the price of the cars.

4 Description of the dataset

The primary description of the dataset is given below:

i. **Total no. of observations**: 205

ii. **Response Variable**: Price of the cars(in dollars)

iii. **Total no. of Predictor Variables**: 24

  (a) **No. of Categorical Variables**: 11

  (b) **No. of Continuous Variables**: 13

5 Visualising the Response Variable

At first, we visualize the response variable,i.e., price of the cars(in dollars) by **histogram** and **boxplot**.

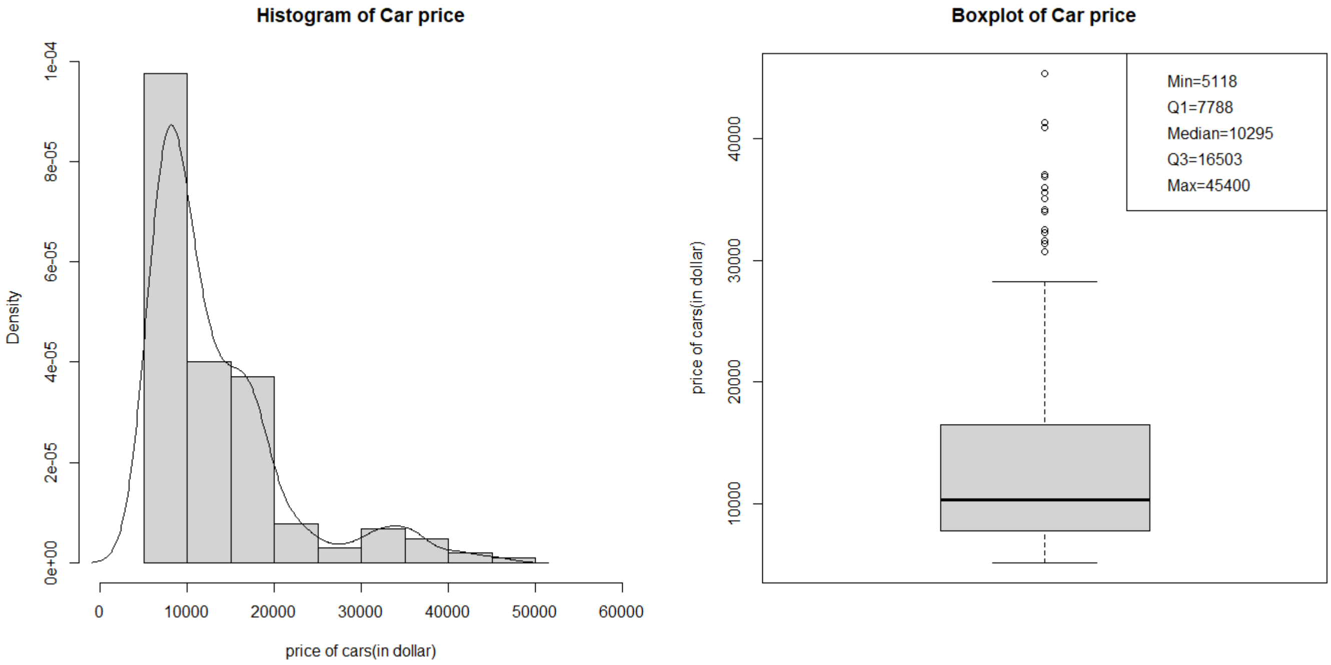

Figure 1: Histogram & Boxplot of price of cars

Now, the summary of the price dataset is given below:

Index	**Value**
count	205
mean	13276.71
std	7988.85
min	5118
P_{25}	7788
P_{50}	10295
P_{75}	16503
P_{85}	18500
P_{90}	22563
max	45400

Table 1: Summary of Price dataset

N.B. P_z denotes the z^{th} percentile

Inference:

- The plot seemed to be right-skewed, meaning that the most prices in the dataset are low(Below 15,000).
- There is a significant difference between the mean and the median of the price distribution.
- The data points are far spread out from the mean, which indicates a high variance in the car prices.(85% of the prices are below 18,500, whereas the remaining 15% are between 18,500 and 45,400).

6 Visualising the Categorical Data

6.1 Fuel Type

In America, mainly, 2 types of fuel used in the cars.

i. Diesel

ii. Gas

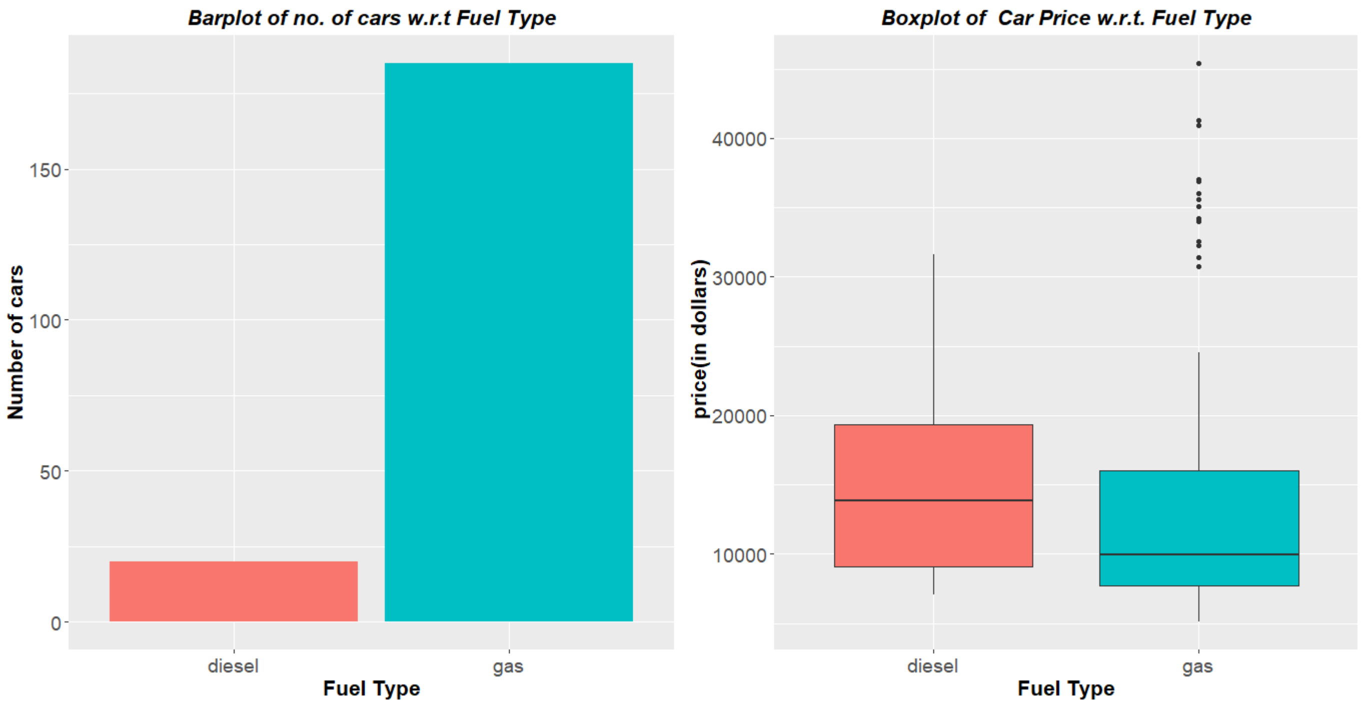

Figure 2: Suitable Diagrams w.r.t **Fuel Type**

Inference:

- *Gas* fueled cars are more preferable than *diesel*.
- The price range of *gas* fueled cars are lower than *diesel*.

6.2 Car Company

There are total 22 types of car companies in the dataset.

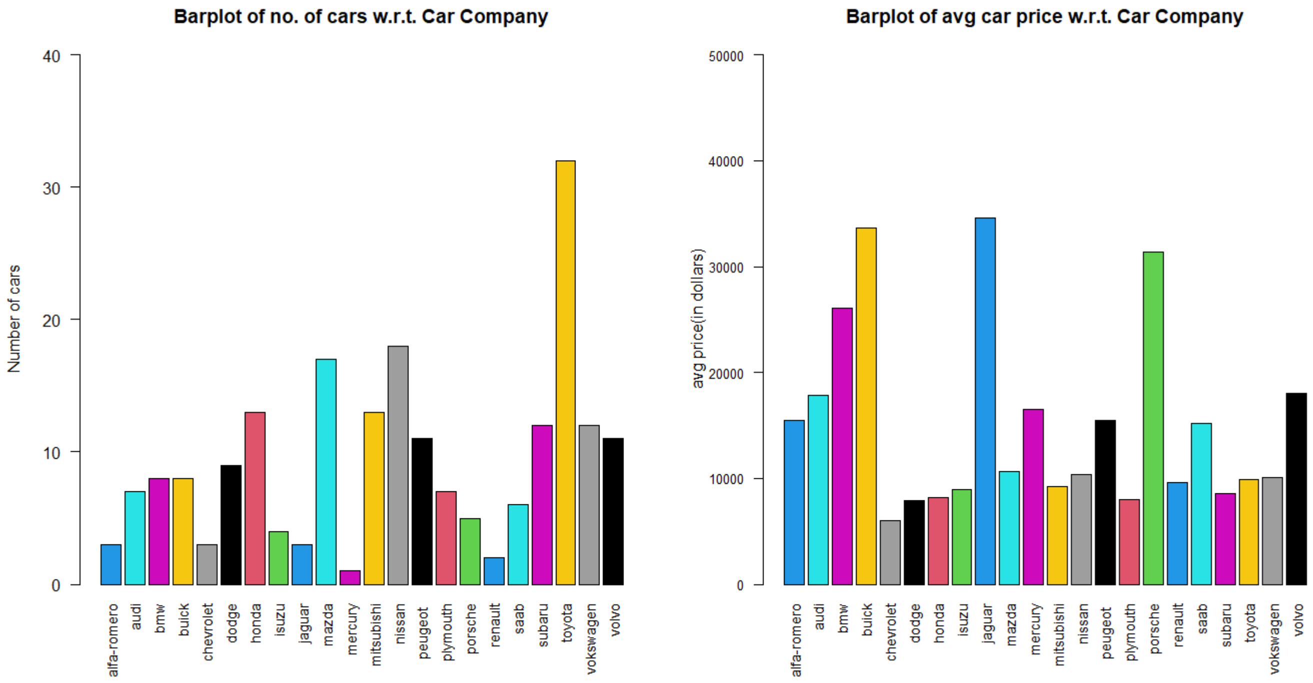

Figure 3: Suitable Diagrams w.r.t *Car Company*

Inference:

- *Toyota* seems to be most preferred car company in America and its average price is low.(below $10000)
- The car companies that sale high budget cars in American market are *Buick,Jaguar,Porsche*

6.3 Car Type

There are 6 types of cars in the dataset.

i. **Convertible:**A convertible or cabriolet is a passenger car that can be driven with or without a roof in place.

ii. **Hardtop:**This type of car has a rigid metal top and no center posts between windows.

iii. **Hatchback:**A hatchback is a car body configuration with a rear door that swings upward to provide access to a cargo area.

iv. **Sedan:**A sedan is a passenger car in a three-box configuration with separate compartments for engine, passenger, and cargo.

v. **Wagon:**a car with a longer body than usual, incorporating a large carrying area behind the seats and having an extra door at the rear for easy loading.

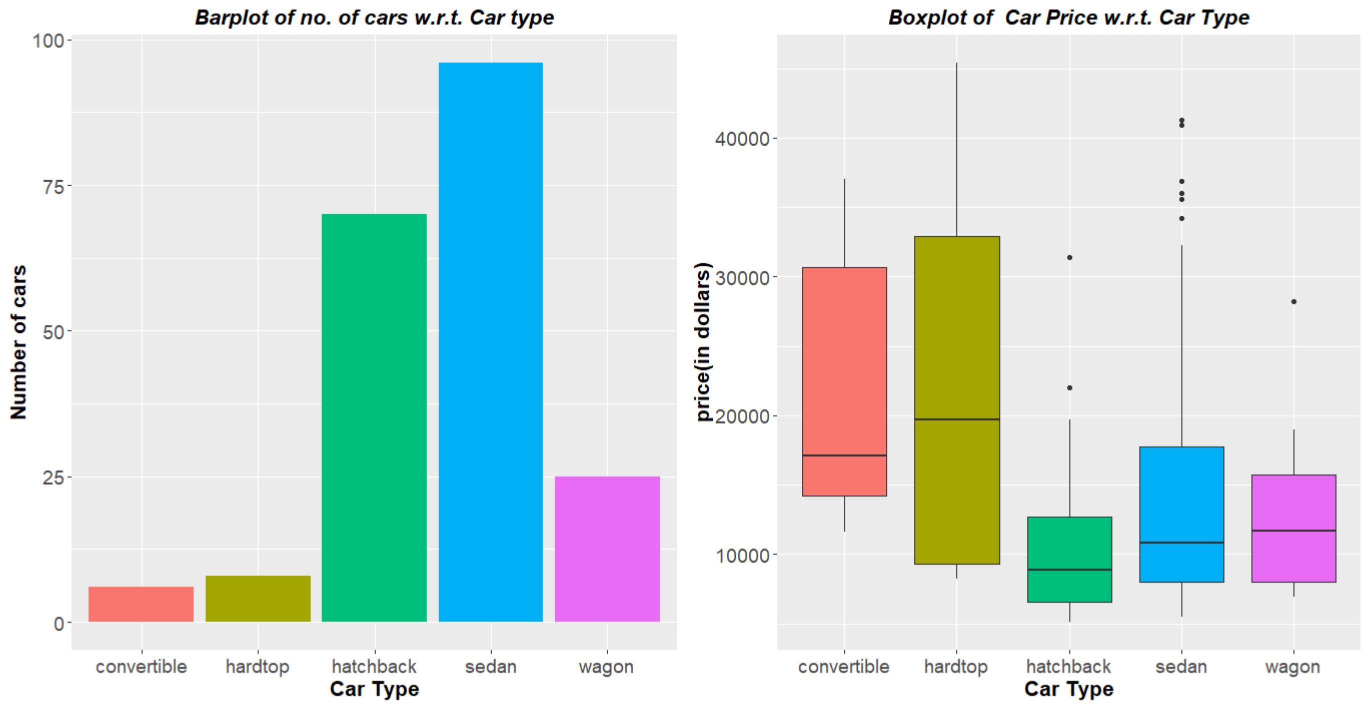

Figure 4: Suitable Diagrams w.r.t *Car Type*

Inference:

- *Sedan* seems to be most preferred car type.
- The price of *Convertible* and *Hardtop* cars seems to be high.

6.4 Engine Type

There are 7 types of engines used in the cars.

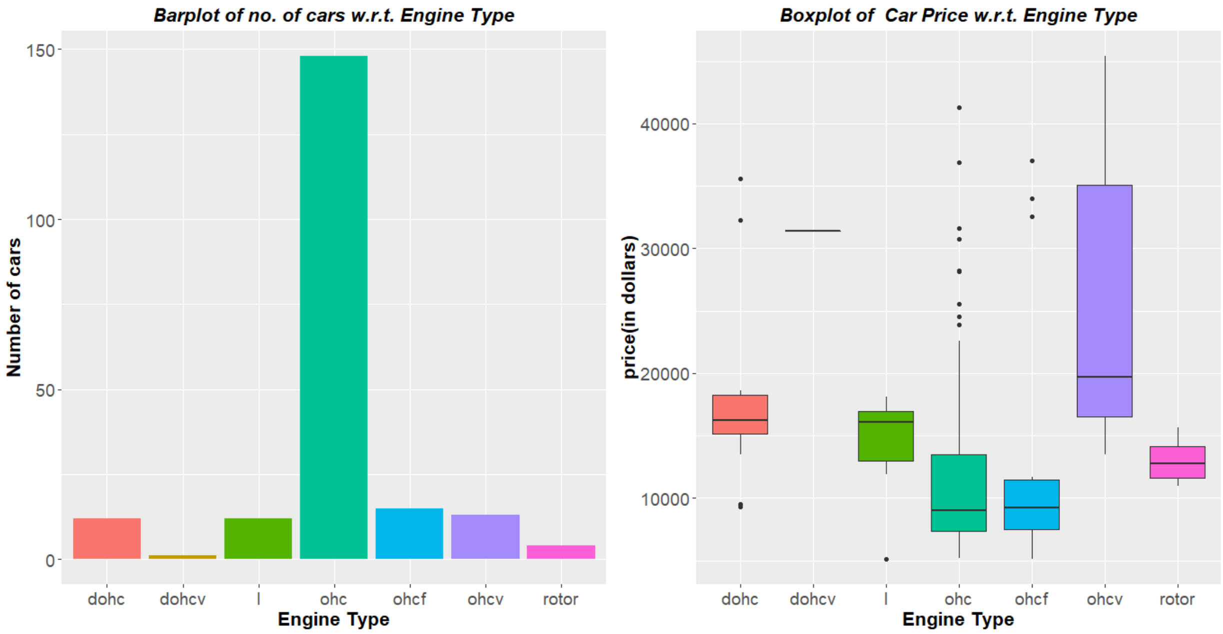

Figure 5: Suitable Diagrams w.r.t *Engine Type*

Inference:

- *ohc* Engine type seems to be most favored type and its price is also low(below $10000)
- *ohcv* has the highest price range (While dohcv has only one row in the dataset).

6.5 Door Number

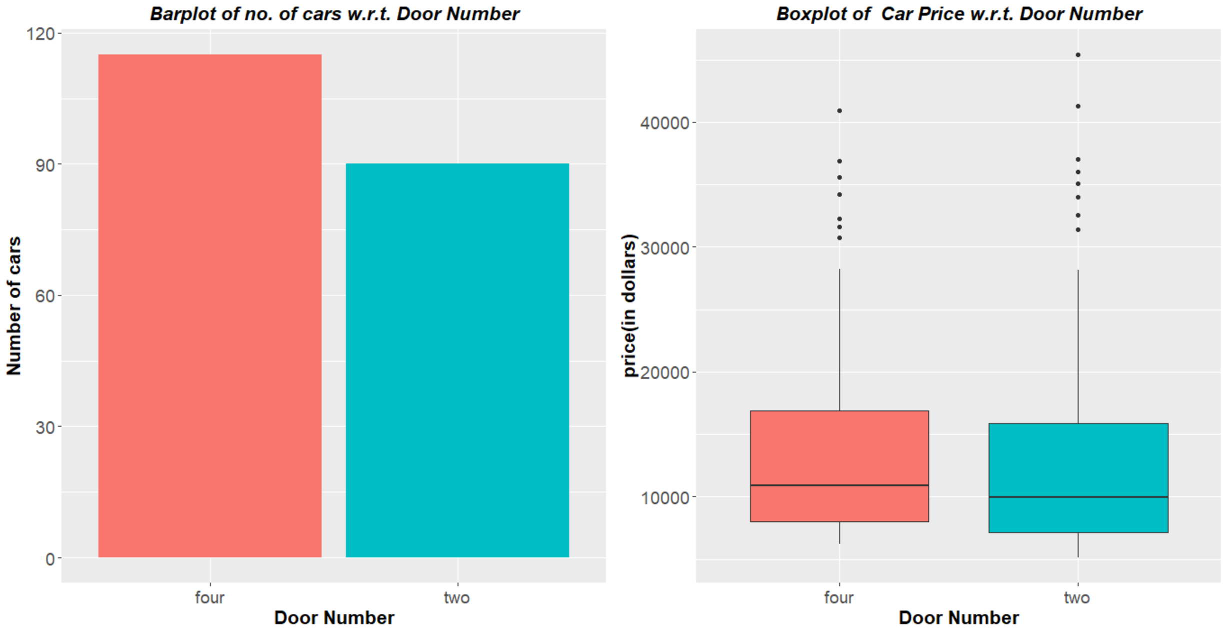

Figure 6: Suitable Diagrams w.r.t *Door Number*

Inference:

doornumber variable is not affecting the price much. There is no significant difference between the categories in it.

6.6 Aspiration

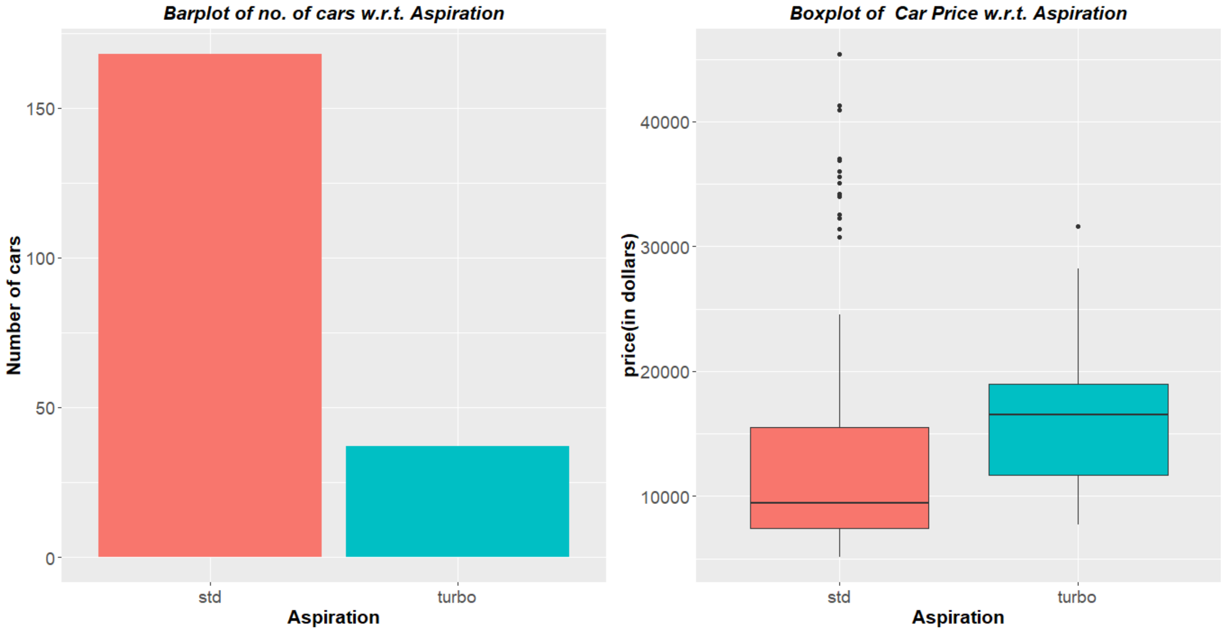

Figure 7: Suitable Diagrams w.r.t *Aspiration*

Inference:

- cars with *std* aspirated engines are more preferred than *turbo*.
- It seems aspiration with *turbo* have higher price range than the *std*.

6.7 Engine Location

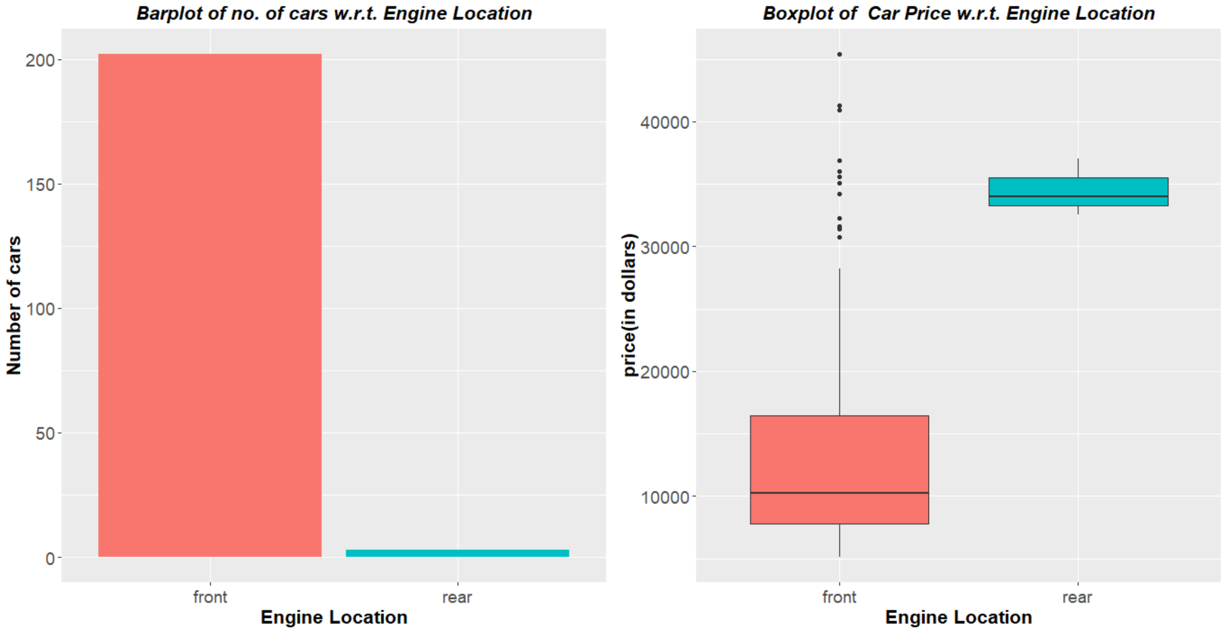

Figure 8: Suitable Diagrams w.r.t *Engine Location*

Inference:

- Cars with *front* engine are most favourable.
- The price of the cars with *rear* engine seems to be higher than the cars with *front* engine.Although,there is very few datapoints for the category *rear* to make an inference.

6.8 Cylinder Number

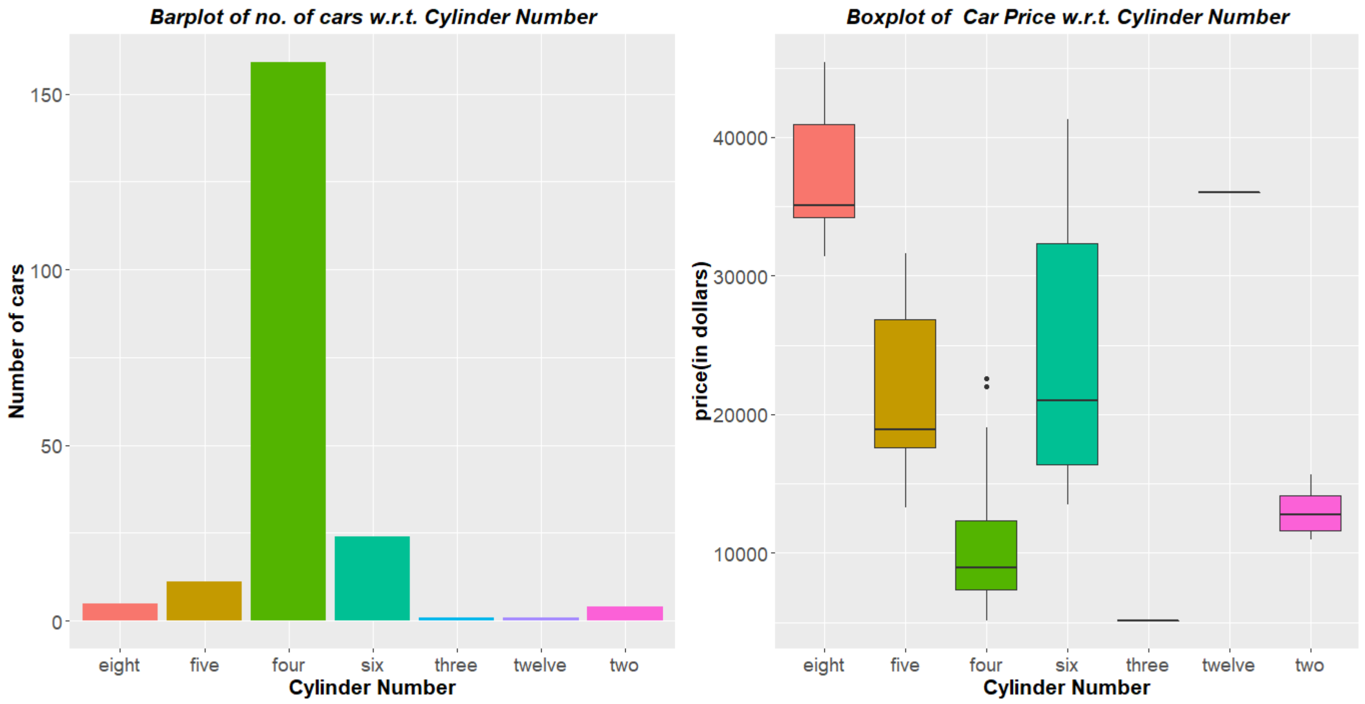

Figure 9: Suitable Diagrams w.r.t *Cylinder Number*

Inference:

- From the dataset, it seems that cars which have *four* no. of cylinders are most preferable to customers.
- Cars with *eight* cylinders have highest price range.

6.9 Fuel System

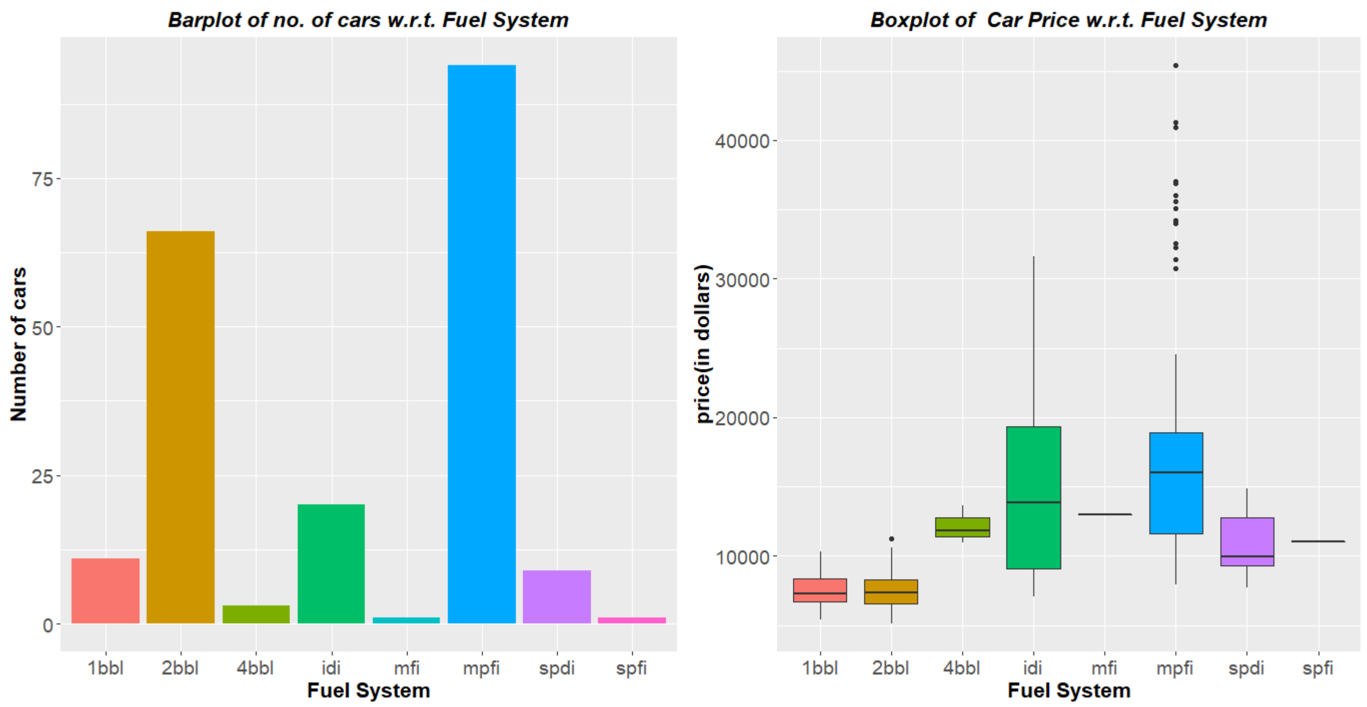

Figure 10: Suitable Diagrams w.r.t *Fuel System*

Inference:

- *mpfi* and *2bbl* are most common type of fuel systems.
- *mpfi* and *idi* have the highest price range. But there are few data for other categories to derive any meaningful inference.

6.10 Drive Wheel

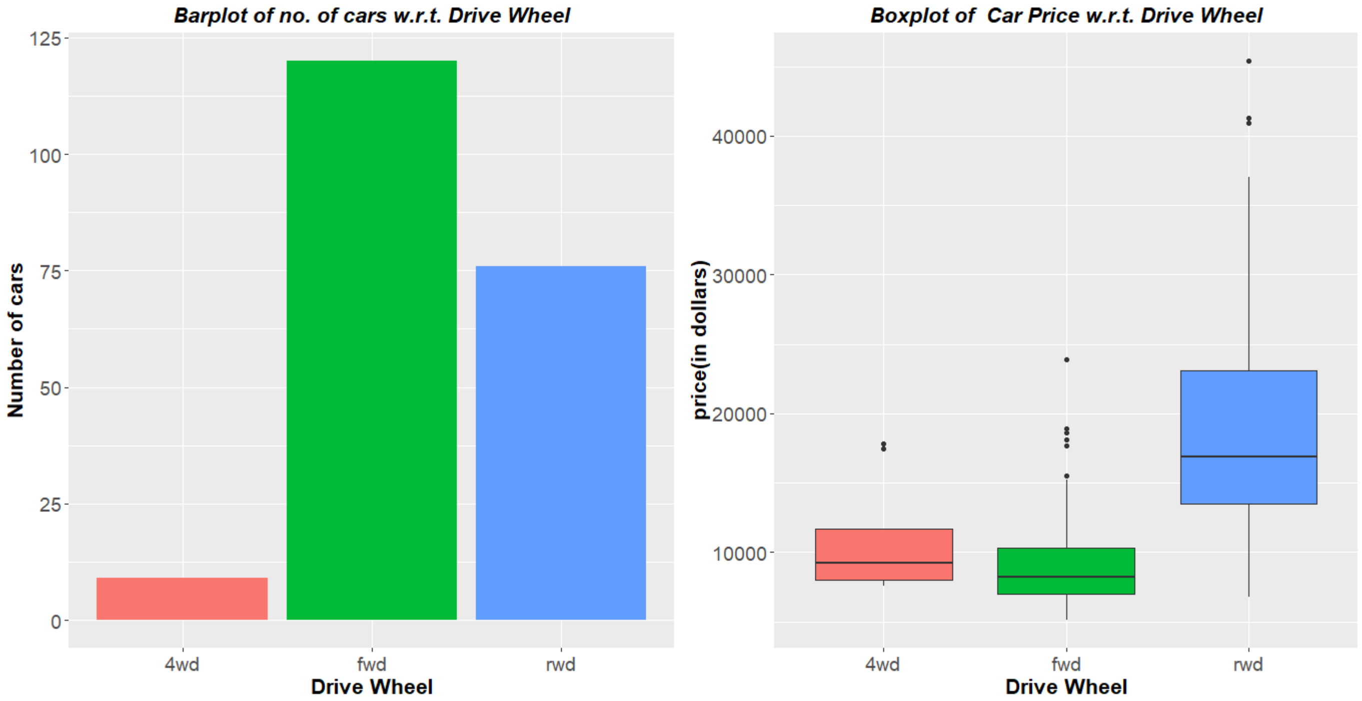

Figure 11: Suitable Diagrams w.r.t *Drive Wheel*

Inference:

- *fwd* is most preferable drivewheel
- Most high ranged cars seem to prefer *rwd* drivewheel.

7 Dummy Variable

In statistics and econometrics, particularly in regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.

Reference Category: One can include k–1 dummy variables, where k stands for the total number of categories in the ordinal/nominal variable. The category that is left out of the equation is called 'the reference category'.

Now, based on the above graphical analysis of the categorical data,the significant categorical variables(the categorical variables ,which have significant differences in prices among the categories) are-

i. Engine Type

ii. Fuel Type

iii. Car Body

iv. Aspiration

v. Cylinder Number

vi. Drivewheel

To use these variables in regression analysis,we are converting them into dummy variables. A table of significant categorical variables and the reference category of each variable is given below:

Significant Categorical Variable	**Reference Category**
Engine Type	dohc
Fuel Type	Diesel
Car Body	Convertible
Aspiration	Std
Cylinder Number	Eight
Drivewheel	4wd

8 Visualising the Numerical Data

In the dataset, there are 13 predictors which are continuous.

8.1 Bivariate Analysis

To find the plausible inter relationship between the response variable(i.e. price) and each of the predictors, we'll analyze them by using the following tools –

i. Graphical analysys

ii. Correlation Heatmaps

8.1.1 Graphical Analysis

Here, we draw scatterplots of the responce variable(price) vs each of the predictors and observe which predictors have no positive/negative correlation with the price.

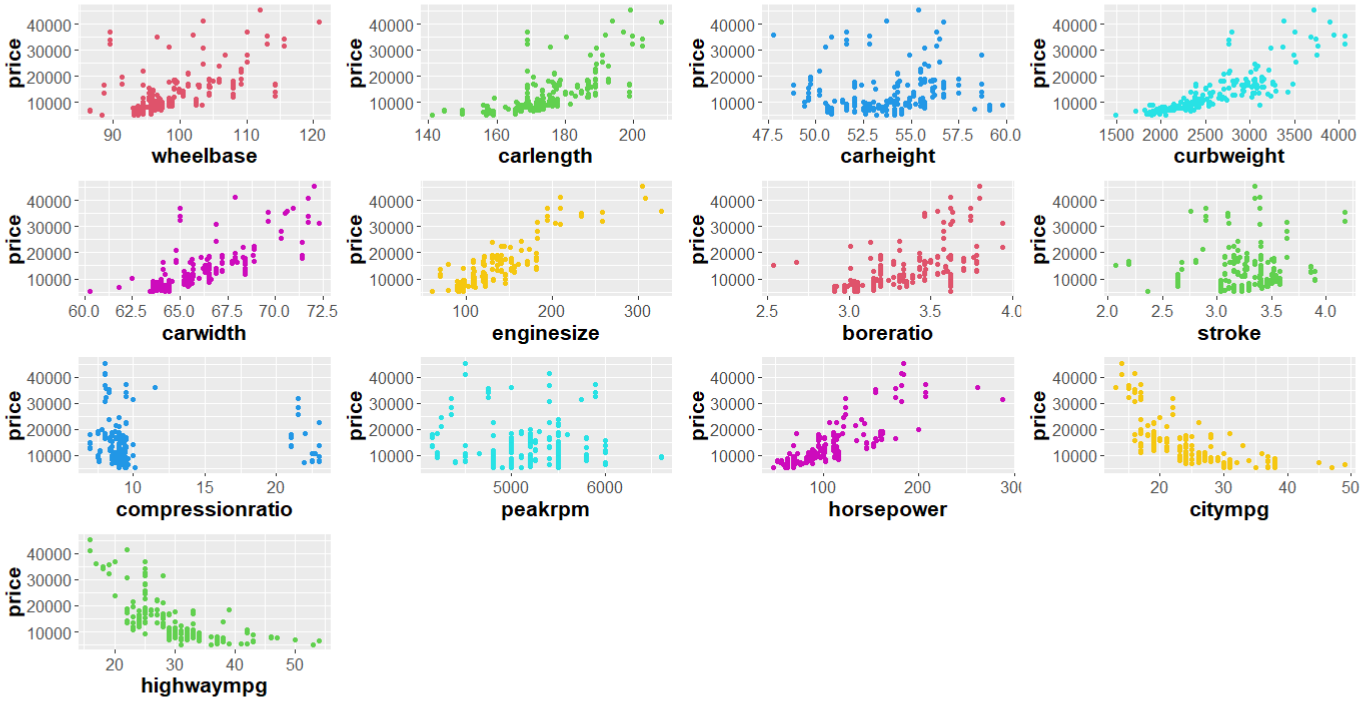

Figure 12: Scatterplot of Price vs Predictors

Analysing the graphs,we can observe that there's no inter relationship between price and the predictors Car Height,Stroke,Compression Ratio,Peakrpm.

8.1.2 Correlation Heatmap

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using colored cells to represent data. The correlation heatmap of the response variable(price) and the 13 predictors is shown below:

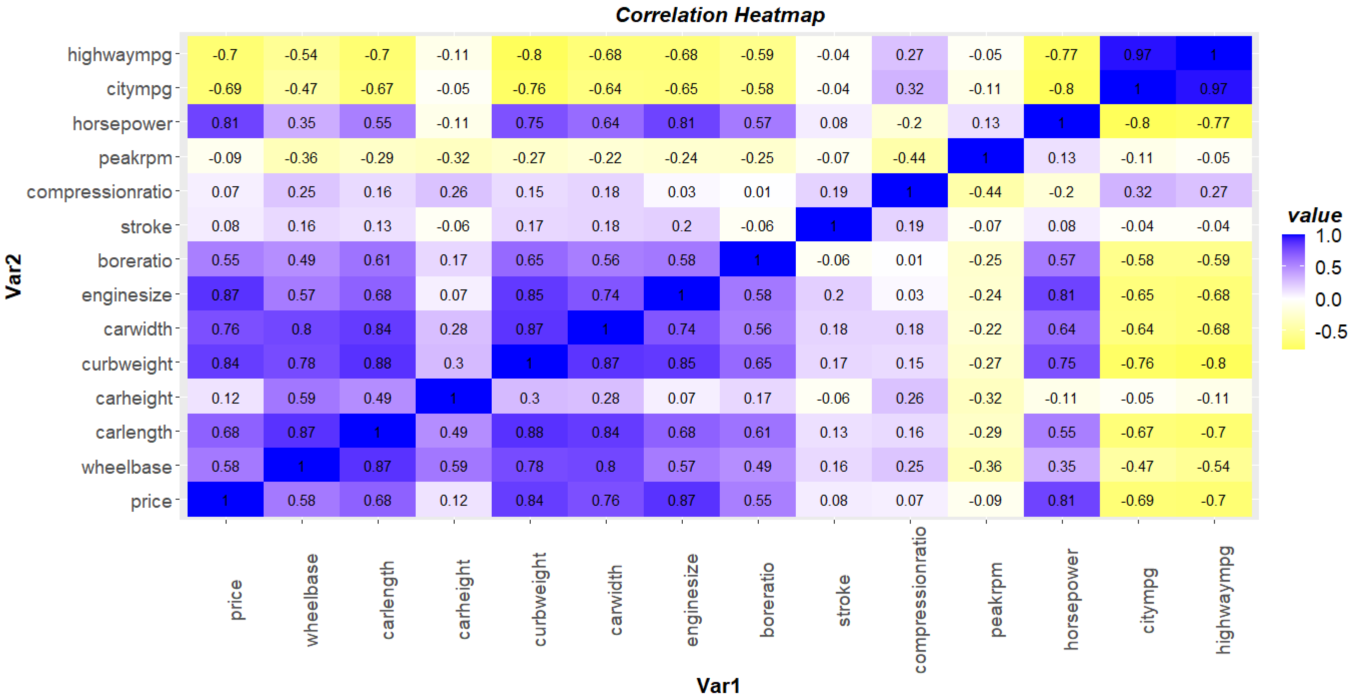

Figure 13: Correlation Heatmap

From the first column(or last row), the values of the correlation coefficients r between price and Carheight, Compression Ratio,Peakrpm,Stroke are very low.

Var1	**Var2**	r
Price	Carheight	0.12
Price	Compressionratio	0.07
Price	Sroke	0.08
Price	Peakrpm	-0.09

From graphical analysis and correlation heatmap, it seems that the predictors *Carheight, Compression Ratio,Peakrpm,Stroke* are not affecting the price significantly.So we will discard these four predictors from our regression model.

8.2 Multicolinearity

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a statistical model.

To detect multicolinearity,we will find partial correlations between 2 predictors ignoring the other independent variables.

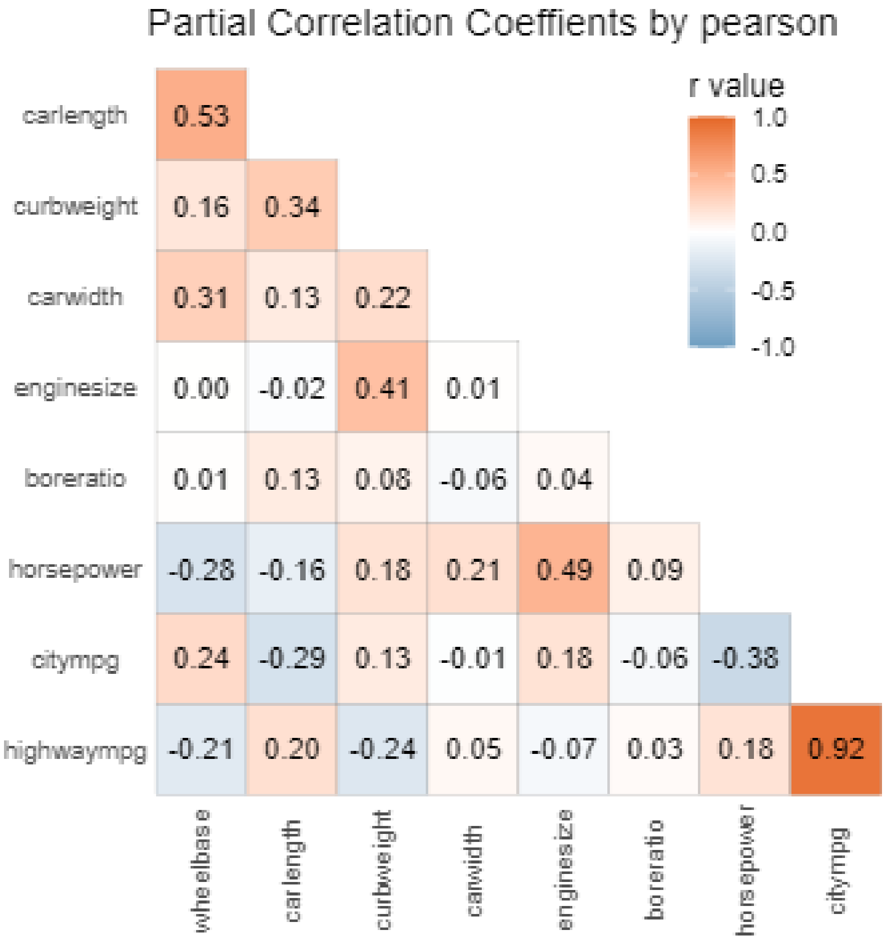

Figure 14: Partial Correlation Coefficients

From the above diagram, we can see the partial correlation coefficient between the predictors citympg and highwaympg is high(= 0.92). So it's clear that there's a presence of multicolinearity between *citympg* and *highwaympg*.

To remove the effect of multicolinearity we'll discard one of the two predictors, citympg and highwaympg. Here,we'll discard the predictor *highwaympg* from the model.

8.3 Transformation of Skewed Variables

In regression analysis, we assume that variables have normal distributions. Non-normally distributed variables (highly skewed or kurtotic variables, or variable with substantial outliers) can distort relationships and significance of tests. That's why, it is very much important to detect the presence of skewness in the response and the numerical variables.

After analysing the scatterplots between two variables,Correlation Heatmaps and multicolinearity, we have discarded total 5 numerical variables from our regression analysis.The following table shows the skewness of the response variable and the remaining 8 numerical variables.

Variable	**Skewness**
Price	1.76464423
Wheelbase	1.04251361
Carlength	0.15481032
Curbweight	0.67640218
Carwidth	0.89737535
Engine Size	1.93337485
Horsepower	1.39500643
Boreratio	0.02000863
Citympg	0.65883775

Table 2: Skewness of Variables

From the table, it's clear that the response variable, *Price* and the 3 numerical predictors *Wheelbase,Engine Size,Horsepower* are highly skewed.

Since, all of these variables take positive values, we will use log transformation to these variables to reduce the skewness.

Log Transformation: Transformed Variable=log(Skewed Variable)

After applying the log transformation to those 4 variables,we get the following skewness:

Skewed Variable	**Skewness**(before the log transformation)	**Skewness**(After the log transformation)
Price	1.76464423	0.66795492
Wheelbase	1.04251361	0.87691004
Engine Size	1.93337485	0.85153808
Horsepower	1.39500643	0.47929772

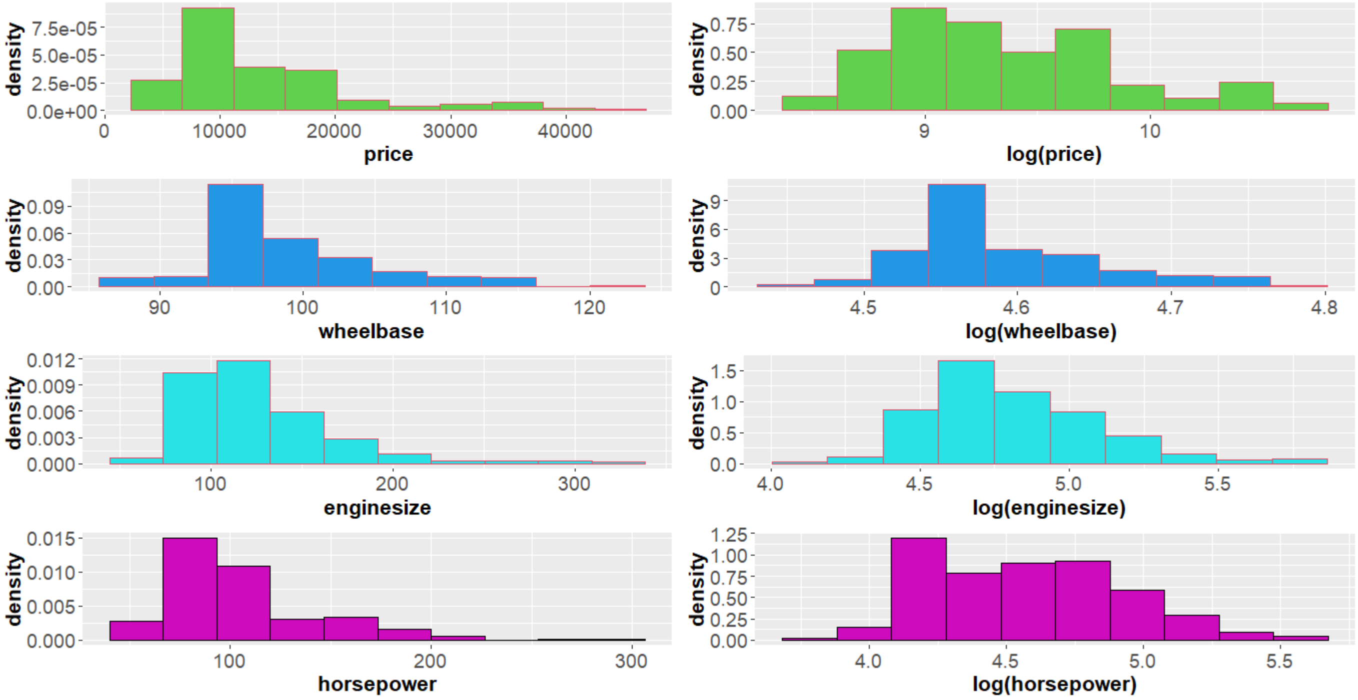

Figure 15: ***Histogram*** of skewed variables *before applying the log transformation* and *after applying the log transformation*

So,it is clear that by log transformation, the skewness has been reduced.

9 Model Building

After eliminating and transforming some of the numerical variables, we will regress the dummy variables and numerical variables on the response variable, price. Before building the model,we will define some terms related to the model building.

1. **Multiple Linear Regression:** Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of the response variable.Let there be k explanatory variables or regressors x_1,x_2,.........,x_k. In this type of regression, the response variable y may be linearly related to x_1,x_2,.........,x_k so that

   $$Y = \beta_0 + \beta_1 x_1 + ...\beta_k x_k + \epsilon$$

   Where,
   Y= predicted values of the response
   β_0, β_1,β_2,.........,β_k are unknown constants which are to be estimated by least square method.
   ϵ is the random error component.

2. **Least Square Estimation:**The method of least squares is used to estimate the unknown constants in regression model,i.e.,β_0, β_1,β_2,.........,β_k such that the sum of the squares of the errors(sum of the squares of the differences between the observations y_i and predicted values of y_i) is a minimum.

3. **R-squared:**R-squared is the proportion of the variation in the dependent variable that is predictable from the independent variables in a linear model. It is also known as coefficient of determination and is denoted by R^2.

   $$R - squared = \frac{Explained\ Variation}{Total\ Variation}$$

4. **Testing of Hypothesis on Individual Regression Coefficients (t-test):**

Statistical hypothesis are statements about relationships. The statistical hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The null hypothesis is denoted by H_0. The alternative hypothesis is the negation of the null hypothesis, denoted by H_1. The t-test is used to check the significance of individual regression coefficients in the multiple linear regression model. Adding a significant variable to a regression model makes the model more effective, while adding an unimportant variable may make the model worse. The hypothesis statements to test the significance of a particular regression coefficient, β_j , are:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The test statistic for this test, under null hypothesis, has the t-distribution with degrees of freedom $(n-2)$:

$$T = \frac{\widehat{\beta_j}}{se(\widehat{\beta_j})}$$

where the estimate of the standard error of β_j is $se(\widehat{\beta_j})$. We reject the null hypothesis if the test statistic lies in the critical region:

$$W : \left\{ |T|_{obs} > t_{\frac{\alpha}{2}, n-2} \right\}$$

Where α is the desired level of significance. The rejection of this Null Hypothesis implies that the corresponding predicting variable has significant effect on predicting response variable y.

5. **P-value:** A p-value is a statistical measurement used to validate a hypothesis against observed data. A P-value lesser than α implies that the test rejects the null hypothesis. For two-sided test,

$$P - value = 2min\{P(T > T_{obs}), P(T < T_{obs})\}$$

Here, we will fit a log-linear model,using OLS.

MODEL 1:

$$\ln Y = \alpha + \sum_{i=1}^{n} \beta_i x_i + u$$

where, Y =Predicted value of the response variable.

α =Intercept term of the model

$x_i = i^t h$ predictor.

β_i =Slope coefficient corresponding to $i^t h$ predictor.

u=Error term of the model.

Results of Testing of Hypothesis of the model coefficients are given below:

term	estimate	std error	T_{obs}	p value
(Intercept)	7.323590	2.235429	3.276145	0.001266
drivewheelfwd	-0.014104	0.066247	-0.212908	0.831643
drivewheelrwd	0.061868	0.073380	0.843113	0.400303
cylindernumberfive	-0.664874	0.147487	-4.508009	0.000012
cylindernumberfour	-0.817573	0.166699	-4.904481	0.000002
cylindernumbersix	-0.432503	0.117635	-3.676652	0.000313
cylindernumberthree	-0.489736	0.285608	-1.714712	0.088148
cylindernumbertwelve	-0.289500	0.183775	-1.575296	0.116974
cylindernumbertwo	-0.799389	0.252841	-3.161625	0.001846
enginetypedohcv	-0.744446	0.238596	-3.120111	0.002111
enginetypel	0.112509	0.080367	1.399943	0.163281
enginetypeohc	0.235676	0.054880	4.294353	0.000029
enginetypeohcf	0.162720	0.081956	1.985449	0.048639
enginetypeohcv	-0.185182	0.078523	-2.358312	0.019450
carbodyhardtop	-0.256748	0.088273	-2.908579	0.004096
carbodyhatchback	-0.319108	0.076767	-4.156843	0.000050
carbodysedan	-0.248483	0.081154	-3.061876	0.002543
carbodywagon	-0.335254	0.088578	-3.784850	0.000210
aspirationturbo	-0.113541	0.049037	-2.315436	0.021736
fueltypegas	-0.294663	0.071358	-4.129332	0.000056
log(wheelbase)	-0.348671	0.505426	-0.689856	0.491188
carlength	0.001431	0.003049	0.469300	0.639433
curbweight	0.000361	0.000105	3.425328	0.000763
carwidth	0.037052	0.015108	2.452474	0.015158
log(enginesize)	-0.547004	0.216293	-2.528997	0.012312
boreratio	0.124443	0.109440	1.137089	0.257038
log(horsepower)	0.755642	0.128939	5.860448	0.000000
citympg	-0.004685	0.004887	-0.958501	0.339117

Using the 205 observations,we get,

Multiple R-square: 0.9247

Inference:

The R-squared value implies that approximately 92.47% of the total variation in the response is explained by this new model.

To obtain the final model,we will list down the p-values of the numerical predictors from the above model and see which variables are significant at 0.05 level of significance.

Numerical Predictors	**p-value**
log(Wheelbase)	0.491188
Carlength	0.639433
Curbweight	0.000763
Carwidth	0.015158
log(Enginesize)	0.012312
Boreratio	0.257038
log(Horsepower)	2.21×10^{-8}
Citympg	0.339117

Table 3: Table of p-values of numerical predictors

From the tables, it's clear that the p-values of *Curbweight,Carwidth,log(Enginesize),log(Horsepowert)* less than 0.05. So, these predictors are significant in the model at 0.05 level of significance.

Now, taking the dummy variables and the significant numerical predictors,we will build a new model.

MODEL 2:

$$\ln Y = \alpha + \sum_{i=1}^{n} \beta_i x_i + u$$

where, Y =Predicted value of the response variable.

α =Intercept term of the model

$x_i = i^t h$ significant predictor.

β_i =Slope coefficient corresponding to $i^t h$ significant predictor.

u=Error term of the model

Results of Testing of Hypothesis of the model coefficients are given below:

term	estimate	std.error	statistic	p.value
(Intercept)	5.059895	0.928358	5.450373	0.000000
cylindernumberfive	-0.556410	0.129092	-4.310160	0.000027
cylindernumberfour	-0.670070	0.128799	-5.202460	0.000001
cylindernumbersix	-0.374180	0.109646	-3.412600	0.000794
cylindernumberthree	-0.329620	0.232366	-1.418550	0.157751
cylindernumbertwelve	-0.332660	0.174706	-1.904120	0.058480
cylindernumbertwo	-0.549060	0.190554	-2.881380	0.004438
enginetypedohcv	-0.659290	0.206848	-3.187340	0.001692
enginetypel	0.106633	0.076805	1.388356	0.166735
enginetypeohc	0.236434	0.053836	4.391711	0.000019
enginetypeohcf	0.216782	0.068370	3.170690	0.001786
enginetypeohcv	-0.163420	0.073964	-2.209510	0.028393
drivewheelfwd	-0.022890	0.065068	-0.351760	0.725428
drivewheelrwd	0.063122	0.070089	0.900604	0.368996
aspirationturbo	-0.118870	0.047003	-2.529050	0.012291
carbodyhardtop	-0.270420	0.083653	-3.232620	0.001458
carbodyhatchback	-0.336490	0.070639	-4.763550	0.000004
carbodysedan	-0.257720	0.069594	-3.703150	0.000283
carbodywagon	-0.343670	0.076822	-4.473610	0.000014
fueltypegas	-0.261060	0.064619	-4.039980	0.000079
curbweight	0.000385	0.000092	4.191515	0.000043
carwidth	0.036813	0.013084	2.813559	0.005441
log(enginesize)	-0.401010	0.169638	-2.363910	0.019143
log(horsepower)	0.822897	0.112243	7.331353	0.000000

Using the 205 observations,we get,

Multiple R-square: 0.9234

Inference:

The R-squared value implies that approximately 92.34% of the total variation in the response is explained by this new model.

10 Residual Plot

A residual plot is a graph that shows the residuals on the vertical axis and the predicted values on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data.

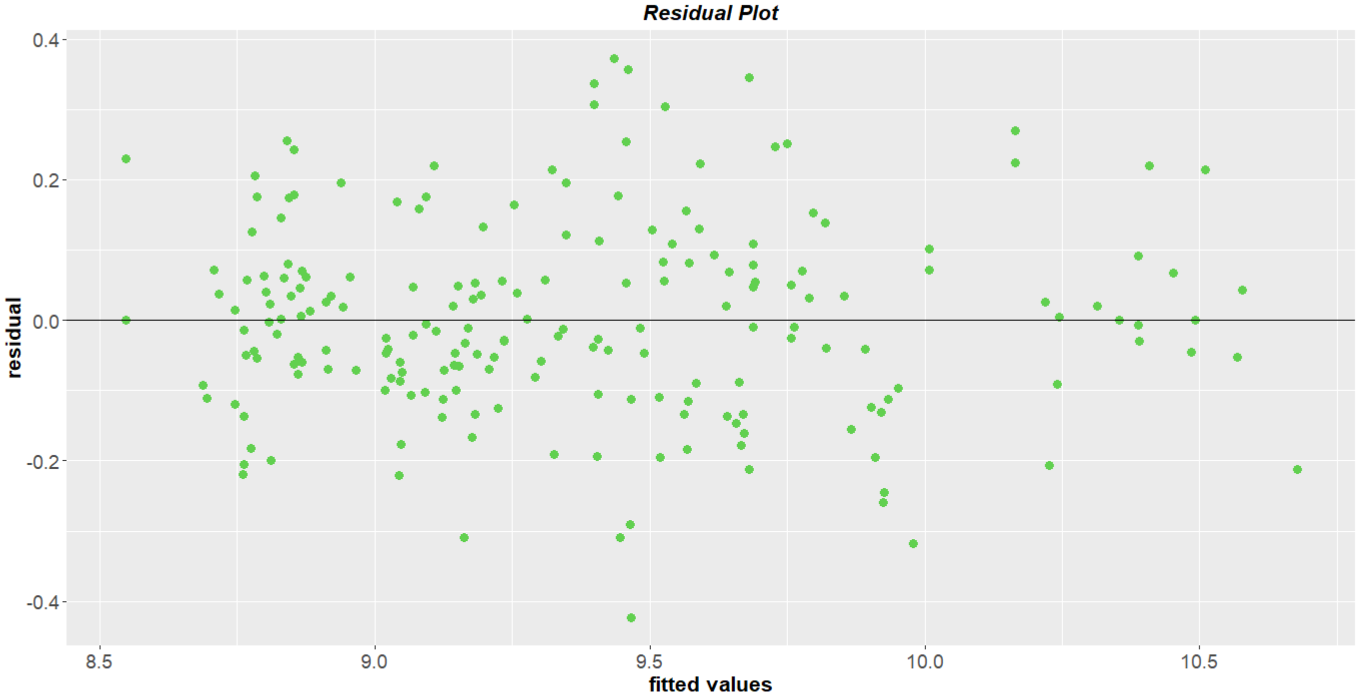

Figure 16: Residual Plot

From the above residual plot,it's clear that the residuals are randomly scattered over the predicted values. It indicates that our fitted model is good.

11 Reference

- Basic Econometrics by Damodar N. Gujarati
- Fundamentals of Statistics(Volume one) by Gun,Gupta,Dasgupta
- https://www.kaggle.com/code/goyalshalini93/car-price-prediction-linear-regressio
  notebook

12 Acknowledgement

I would like to express my gratitude to the Rector and Principal of my college, Rev. Dr Dominic Savio, S.J. for giving me the opportunity to work on this project. I would also like to thank the Vice Principal of the Arts and Science department, Prof Betram Da'Silva, and The Dean of Science Dr. Tapati Dutta. I am also very grateful to my project supervisor Dr. Ayan Chandra for guiding me with the project's guidelines. He constantly supported me and guided me with his valuable advice whenever I was stuck with my project. He took notice of the problems I faced, all the concepts I did not have. He also provided me with necessary materials and references that helped me complete this project and also helped me to gain knowledge about different aspects of theoretical statistics and its application.

I would like to thank my family and my friends as well. They kept me motivated to work constantly on my project. They also helped me to understand some of the concepts and helped me to write some codes. My work was made easier by my family and my friends.

13 Appendix

Code 1: R Code for the Paper

```
rm(list=ls())

##Calling the Libraries

library(ppcor)
library(reshape2)
library(ggplot2)
library(moments)
library(MASS)
library(ggiraphExtra)
library(gridExtra)
library(snpar)
library(car)
library(writexl)
library(broom)
##Calling the dataset

setwd("C:/Users/USER/Desktop/DISSERTATION")
data=read.csv(’Assignment.csv’)
attach(data)

##Visualising the Response Variable

#i. Summary of the price dataset

d=summary(price)
d

```

```r

#ii. Histogram of the Price of the Car

hist(price,freq = F,xlim=c(0,60000),xlab='price of cars(in dollar)',main='Histogram of Car price')
lines(density(price))

#ii. Boxplot of the Price of the Car

boxplot(price,ylab='price of cars(in dollar)',main='Boxplot of Car price')
legend('topright',legend =c('Min=5118','Q1=7788','Median=10295','Q3=16503','Max=45400'))

##Visualising the Categorical Data

##a> FUEL TYPE

theme_new=theme(plot.title=element_text(size=16,hjust = 0.5,face = 'bold.italic'),
                plot.subtitle=element_text(size=10,hjust = 0.5,face = 'bold.italic'),
                legend.title=element_text(hjust = 0.5,face = 'bold.italic',size=16),
                legend.text=element_text(size=14),axis.title=element_text(face='bold',size=16),
                axis.text=element_text(hjust=0.5,size=14))


p1=ggplot(NULL, aes(fueltype,fill=fueltype))+geom_bar(show.legend = F)+theme_new+labs(x='Fuel Type',y = 'Number of cars',title=
  'Barplot of no. of cars w.r.t Fuel Type',fill='Index')

p2=ggplot(NULL,aes(fueltype,price,fill=fueltype))+geom_boxplot(show.legend = F)+theme_new+
  labs(title='Boxplot of  Car Price w.r.t. Fuel Type',x='Fuel Type',
       y=' price(in dollars)',col='Index')

grid.arrange(p1,p2,nrow=1)


##b> CAR COMPANY
par(mfrow=c(1,2))
y=table(Company)
y
barplot(y,cex.names =0.9,las=2,ylim=c(0,40),col=4:27,ylab='Number of cars',main='Barplot of no. of cars w.r.t. Car Company')
m2=aggregate(price,by=as.data.frame(Company),mean)
m2
barplot(m2[,2],names.arg=m2[,1],ylab='avg price(in dollars)',main='Barplot of avg car price w.r.t. Car Company',las=2,cex.names=0.9,cex.axis=0.8,ylim=c(0,50000),col=4:27)


###c> CAR TYPE

p1=ggplot(NULL,aes(carbody,fill=carbody))+geom_bar(show.legend = F)+
  labs( x='Car Type',y='Number of cars',
        title='Barplot of no. of cars w.r.t. Car type')+theme_new
```

```
p2=ggplot(NULL, aes(carbody,price,fill=carbody))+geom_boxplot(show.legend = F)+
  theme_new+labs(x='Car Type',y='price(in dollars)',
                 title='Boxplot of  Car Price w.r.t. Car Type')
grid.arrange(p1,p2,nrow=1)

###d> ENGINE TYPE

p1=ggplot(NULL,aes(enginetype,fill=enginetype))+geom_bar(show.legend = F)+
  labs( x='Engine Type',y='Number of cars',
        title='Barplot of no. of cars w.r.t. Engine Type')+theme_new
p2=ggplot(NULL, aes(enginetype,price,fill=enginetype))+geom_boxplot(show.legend = F)+
  theme_new+labs(x='Engine Type',y='price(in dollars)',
                 title='Boxplot of  Car Price w.r.t. Engine Type')
grid.arrange(p1,p2,nrow=1)


###e> DOOR NUMBER

p1=ggplot(NULL,aes(doornumber,fill=doornumber))+geom_bar(show.legend = F)+
  labs( x='Door Number',y='Number of cars',
        title='Barplot of no. of cars w.r.t. Door Number')+theme_new
p2=ggplot(NULL, aes(doornumber,price,fill=doornumber))+geom_boxplot(show.legend = F)+
  theme_new+labs(x='Door Number',y='price(in dollars)',
                 title=’Boxplot of  Car Price w.r.t. Door Number’)
grid.arrange(p1,p2,nrow=1)

###f> ASPIRATION


p1=ggplot(NULL,aes(aspiration,fill=aspiration))+geom_bar(show.legend = F)+
  labs( x=’Aspiration’,y=’Number of cars’,
        title=’Barplot of no. of cars w.r.t. Aspiration’)+theme_new
p2=ggplot(NULL, aes(aspiration,price,fill=aspiration))+geom_boxplot(show.legend = F)+
  theme_new+labs(x=’Aspiration’,y=’price(in dollars)’,
                 title=’Boxplot of  Car Price w.r.t. Aspiration’)
grid.arrange(p1,p2,nrow=1)

##g> ENGINE LOCATION

p1=ggplot(NULL,aes(enginelocation,fill=enginelocation))+geom_bar(show.legend = F)+
  labs( x=’Engine Location’,y=’Number of cars’,
        title=’Barplot of no. of cars w.r.t. Engine Location’)+theme_new
p2=ggplot(NULL, aes(enginelocation,price,fill=enginelocation))+geom_boxplot(show.legend = F)+
  theme_new+labs(x=’Engine Location’,y=’price(in dollars)’,
                 title=’Boxplot of  Car Price w.r.t. Engine Location’)
grid.arrange(p1,p2,nrow=1)
```

```


##h>CYLINDER NUMBER

p1=ggplot(NULL,aes(cylindernumber,fill=cylindernumber))+geom_bar(show.legend = F)+
  labs( x='Cylinder Number',y='Number of cars',
        title='Barplot of no. of cars w.r.t. Cylinder Number')+theme_new
p2=ggplot(NULL, aes(cylindernumber,price,fill=cylindernumber))+geom_boxplot(show.legend = F)+
  theme_new+labs(x='Cylinder Number',y='price(in dollars)',
                 title='Boxplot of  Car Price w.r.t. Cylinder Number')
grid.arrange(p1,p2,nrow=1)

##i>FUEL SYSTEM

p1=ggplot(NULL,aes(fuelsystem,fill=fuelsystem))+geom_bar(show.legend = F)+
  labs( x='Fuel System',y='Number of cars',
        title='Barplot of no. of cars w.r.t. Fuel System')+theme_new
p2=ggplot(NULL, aes(fuelsystem,price,fill=fuelsystem))+geom_boxplot(show.legend = F)+
  theme_new+labs(x='Fuel System',y='price(in dollars)',
                 title='Boxplot of  Car Price w.r.t. Fuel System')
grid.arrange(p1,p2,nrow=1)

##j>DRIVE WHEEL
```

```


p1=ggplot(NULL,aes(drivewheel,fill=drivewheel))+geom_bar(show.legend = F)+
  labs( x=’Drive Wheel’,y=’Number of cars’,
        title=’Barplot of no. of cars w.r.t. Drive Wheel’)+theme_new
p2=ggplot(NULL, aes(drivewheel,price,fill=drivewheel))+geom_boxplot(show.legend = F)+
  theme_new+labs(x=’Drive Wheel’,y=’price(in dollars)’,
                 title=’Boxplot of  Car Price w.r.t. Drive Wheel’)
grid.arrange(p1,p2,nrow=1)


###VISUALISING THE NUMERICAL DATA

theme_new1=theme(plot.title=element_text(size=16,hjust = 0.5,face = ’bold.italic’),
                 plot.subtitle=element_text(size=10,hjust = 0.5,face = ’bold.italic’),
                 legend.title=element_text(hjust = 0.5,face = ’bold.italic’,size=16),
                 legend.text=element_text(size=14),axis.title=element_text(face=’bold’,size=16),
                 axis.text=element_text(hjust=0.5,size=12))
##a. Scatterplots
p1=ggplot(NULL,aes(wheelbase,price))+geom_point(col=2)+theme_new1
p2=ggplot(NULL,aes(carlength,price))+geom_point(col=3)+theme_new1
p3=ggplot(NULL,aes(carheight,price))+geom_point(col=4)+theme_new1
```

```r
p4=ggplot(NULL,aes(curbweight,price))+geom_point(col=5)+theme_new1
p5=ggplot(NULL,aes(carwidth,price))+geom_point(col=6)+theme_new1
p6=ggplot(NULL,aes(enginesize,price))+geom_point(col=7)+theme_new1
p7=ggplot(NULL,aes(boreratio,price))+geom_point(col=10)+theme_new1
p8=ggplot(NULL,aes(stroke,price))+geom_point(col=11)+theme_new1
p9=ggplot(NULL,aes(compressionratio,price))+geom_point(col=12)+theme_new1
p10=ggplot(NULL,aes(peakrpm,price))+geom_point(col=13)+theme_new1
p11=ggplot(NULL,aes(horsepower,price))+geom_point(col=14)+theme_new1
p12=ggplot(NULL,aes(citympg,price))+geom_point(col=15)+theme_new1
p13=ggplot(NULL,aes(highwaympg,price))+geom_point(col=19)+theme_new1
grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11,p12,p13,ncol=4)

###b.CORRELATION HEATMAP
B1=data.frame(price,wheelbase,carlength,carheight,curbweight,carwidth,
              enginesize,boreratio,stroke,compressionratio,peakrpm,horsepower,citympg,highwaympg )

corr = data.matrix(cor(B1[sapply(B1,is.numeric)]))
mel = melt(corr)
ggplot(mel, aes(Var1,Var2))+geom_tile(aes(fill=value)) +
  geom_text(aes(label = round(value,2)))+
  scale_fill_gradient2(low='yellow',mid='white' ,high='blue') + labs(title = 'Correlation Heatmap')+theme_new+
  theme(axis.text.x = element_text(angle=90))
```

```



B=data.frame(price,wheelbase,carlength,curbweight,carwidth,
             enginesize,boreratio,horsepower,citympg,highwaympg )
B2=B[,-1]
B2

##Partial Correlation Coefficients between 2 predictors ignoring the
##effect of other predictors
pcor(B2)
ggCor(B2,what=2,label=1,interactive = T,xangle=90)

##obtaining the skewness of response and explanatory variables

skewness(B[,-10])
#Reducing  the  skewness of highly skewed variables by taking log transformastion
A1=data.frame(log(price),log(wheelbase),(carlength),(curbweight),(carwidth),log(enginesize),(boreratio),log(horsepower),(citympg))
pcor(A1)
A1
A1


##Histograms showing the reduction of skewness
q1=ggplot(NULL,aes(price))+
  geom_histogram(fill=3,col=2,bins = 10,
                 aes(y=..density..))+theme_new
```

```r
q2=ggplot(NULL,aes(log(price)))+
  geom_histogram(fill=3,col=2,bins = 10,
                 aes(y=..density..))+theme_new

q3=ggplot(NULL,aes(wheelbase))+
  geom_histogram(fill=4,col=2,bins = 10,
                 aes(y=..density..))+theme_new

q4=ggplot(NULL,aes(log(wheelbase)))+
  geom_histogram(fill=4,col=2,bins = 10,
                 aes(y=..density..))+theme_new

q5=ggplot(NULL,aes(enginesize))+
  geom_histogram(fill=5,col=2,bins = 10,
                 aes(y=..density..))+theme_new

q6=ggplot(NULL,aes(log(enginesize)))+
  geom_histogram(fill=5,col=2,bins = 10,
                 aes(y=..density..))+theme_new
q7=ggplot(NULL,aes(horsepower))+
  geom_histogram(fill=6,col=1,bins = 10,
                 aes(y=..density..))+theme_new
q8=ggplot(NULL,aes(log(horsepower)))+
  geom_histogram(fill=6,col=1,bins = 10,
                 aes(y=..density..))+theme_new

grid.arrange(q1,q2,q3,q4,q5,q6,q7,q8,ncol=2)

##Building the model
model1=lm(A1[,1]~drivewheel+cylindernumber+enginetype+carbody+
   aspiration+fueltype+A1[,2]+A1[,3]+A1[,4]+A1[,5]+A1[,6]+A1[,7]+
A1[,8]+A1[,9])
summary(lm(A1[,1]~drivewheel+cylindernumber+enginetype+carbody+aspiration+fueltype+A1[,2]+A1[,3]+A1[,4]+A1[,5]+A1[,6]+A1[,7]+A1[,8]+A1[,9]))


s1=tidy(model1)


write_xlsx(s1,"C:/Users/USER/Desktop/DISSERTATION/s1.xlsx")
##Building a new model removing the non-significant explanatory variables
model=lm(A1[,1]~cylindernumber+enginetype+drivewheel+aspiration+carbody+fueltype+A1[,4]+A1[,5]+A1[,6]+A1[,8])

summary(model)
s2 =tidy(model)
write_xlsx(s2,"C:/Users/USER/Desktop/DISSERTATION/s2.xlsx")

##Obtaining the residuals
res=resid(model)
res

##Obtaining the predicted values of price
Y_hat=(fitted(model))
##Residual Plot
ggplot(NULL,aes(Y_hat,res))+geom_point(col=3,size=3)+labs(x='fitted values',y='residual',title = 'Residual Plot')+
  theme_new+geom_hline(yintercept = 0)
abline(h=0)
```