MATH1312 Regression Analysis - Project 2021

Adity Bhargav(3814507) and Anushka Sharma(3817004)

Table of Contents

Introduction

With the rise of the vehicle business, car price prediction has become a popular study topic. The number of automobiles on the road has been rising, and this has increased the popularity of this subject. A huge variety of diverse attributes play a vital part in a dependable and accurate forecast of the price of a car. Fuel type, engine type, car dimensions, among other factors, will be investigated in this project in order to gain a better knowledge of this subject and identify the key features that majorly affect the price of a car.The key features identified will be used to build a model that will be help to predict the car price interval when the key features are provided.

Data Source and Description

The dataset that will be used for the analysis has been taken from- Car Price Prediction Multiple Linear Regression. (2021), from https://www.kaggle.com/hellbuoy/car-price-prediction. The dataset includes variables that influence car costs in the United States.

There are 205 observations and 26 variables in the presented dataset, which represents various car attributes. The variables along with their description is given in the below table:

Data Description of car dataset

S_No	Feature Name	Description
1	Car_ID	Unique id of each observation (Interger)

2	Symboling	Its assigned insurance risk rating, A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.(Categorical)
3	carCompany	Name of car company (Categorical)
4	fueltype	Car fuel type i.e gas or diesel (Categorical)
5	aspiration	Aspiration used in a car (Categorical)
6	doornumber	Number of doors in a car (Categorical)
7	carbody	body of car (Categorical)
8	drivewheel	type of drive wheel (Categorical)
9	enginelocation	Location of car engine (Categorical)
10	wheelbase	Weelbase of car (Numeric)
11	carlength	Length of car (Numeric)
12	carwidth	Width of car (Numeric)
13	carheight	height of car (Numeric)
14	curbweight	The weight of a car without occupants or baggage. (Numeric)
15	enginetype	Type of engine. (Categorical)
16	cylindernumber	cylinder placed in the car (Categorical)
17	enginesize	Size of car (Numeric)
18	fuelsystem	Fuel system of car (Categorical)
19	boreratio	Boreratio of car (Numeric)
20	stroke	Stroke or volume inside the engine (Numeric)
21	compressionratio	compression ratio of car (Numeric)
22	horsepower	Horsepower (Numeric)
23	peakrpm	car peak rpm (Numeric)
24	citympg	Mileage in city (Numeric)
25	highwaympg	Mileage on highway (Numeric)
26	price(Dependent variable)	Price of car (Numeric)

Response and Predictors

Out of the 26 variables listed above, price variable is our **response variable** and the other 25 variables are the predictors.

Problem Statement

Geely Auto, a Chinese car manufacturing company, wants to break into the US market by establishing a manufacturing facility there and making automobiles locally to compete with their American and European equivalents.

They have hired an automobile consultancy firm to help them understand the elements that influence car pricing. They are particularly interested in learning about the factors that influence car pricing in the American market, as they may differ significantly from those in China. The Chinese company wants to find answer of the following questions-

1.) Which variables are important in forecasting a car's price?

2.) How well those variables accurately describe a car's pricing?

Business Goal Objective

With the supplied independent variables,the pricing of car will be modeled. It will be utilized by management to determine how prices vary in relation to the independent factors. They can then adjust the car's design, commercial strategy, and other factors to fulfill specified price targets.

Methodology

The dataset car will be investigated by first performing data-prepossessing steps in order to check and deal with missing/special values, perform necessary data conversion (if required) and other basic data cleaning steps if needed.Car_ID variable will be removed from the model as it has contains unique values for each observation and does not contribute any information on the car attributes. Descriptive statistics will also be generated for the response variable price. Exploratory analysis will be carried out next for all the categorical variables by plotting them individually against the response variable price using barplots and boxplots. For the numerical variables in the dataset, correlation plot and matrix will be plotted to find correlation between each predictor with the response variable. The correlation matrix will also help in identifying whether there is any presence of any multicollinearity in the dataset. If there is presence of multicollinearity, it will be dealt by mean-centering approach.

Variable Selection

The exploratory analysis and correlation plots will help in identifying the variables that have significant effect on the response variable price. The categorical and numerical variables that show significance correlation and effect on the price variable will be kept in the original dataset and the rest of the variables will be removed from the dataset. This approach will be used in order to reduce the complexity of the dataset and make it more clean by only dealing with relevant variables in the dataset that significantly effect the price of a car.

Model Fitting

The multiple linear regression approach will be used to fit a multiple linear model to the car dataset will all the selected variables. The model fitting will be performed manually and by using other methods like backward elimination, forward selection and stepwiseregression. The adequacy of each model obtained from a different model will be

tested in order to identify the best model with significant predictors.The model adequacy will be tested by conducting anova tests and detailed residual analysis where all model assumptions will be tested. In case the model assumption of constant variance is not satisfied in the first model, transformation of the response variable will be done to satisfy the assumption.

The best models obtained from different methods of model fitting will be compared through partial F-tests to find the best model. The best model found from the Partial F-tests will then be used to predict the price of assumed future values of all the predictors in that model.

Prediction

After evaluating and identifying the best model, car price predictions interval will be calculated for random observations to check whether the best model is able to give accurate predictions or not.

Results

Importing packages

```
library(magrittr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(corrplot)
library(Hmisc)
library(ggcorrplot)
library(RColorBrewer)
library(QuantPsyc)
library(car)
library(stringr)
library(TSA)
library(DAAG)
library(qpcR)
library(olsrr)

options(warn=-1)
```

Importing the dataset

```
car <- read.csv("CarPrice_Assignment.csv")
head(car,5)
```

```
##   car_ID symboling                  CarName fueltype aspiration doornumber
## 1      1         3       alfa-romero giulia      gas        std        two
## 2      2         3      alfa-romero stelvio      gas        std        two
## 3      3         1 alfa-romero Quadrifoglio      gas        std        two
## 4      4         2              audi 100 ls      gas        std       four
## 5      5         2               audi 100ls      gas        std       four
```

```
##       carbody drivewheel enginelocation wheelbase carlength carwidth
carheight
## 1 convertible        rwd          front      88.6     168.8     64.1
48.8
## 2 convertible        rwd          front      88.6     168.8     64.1
48.8
## 3   hatchback        rwd          front      94.5     171.2     65.5
52.4
## 4       sedan        fwd          front      99.8     176.6     66.2
54.3
## 5       sedan        4wd          front      99.4     176.6     66.4
54.3
##   curbweight enginetype cylindernumber enginesize fuelsystem boreratio
stroke
## 1       2548       dohc           four        130       mpfi      3.47
2.68
## 2       2548       dohc           four        130       mpfi      3.47
2.68
## 3       2823       ohcv            six        152       mpfi      2.68
3.47
## 4       2337        ohc           four        109       mpfi      3.19
3.40
## 5       2824        ohc           five        136       mpfi      3.19
3.40
##   compressionratio horsepower peakrpm citympg highwaympg price
## 1                9        111    5000      21         27 13495
## 2                9        111    5000      21         27 16500
## 3                9        154    5000      19         26 16500
## 4               10        102    5500      24         30 13950
## 5                8        115    5500      18         22 17450
```

```r
#Checking the structure of all variables
str(car)
```

```
## 'data.frame':    205 obs. of  26 variables:
##  $ car_ID          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ symboling       : int  3 3 1 2 2 2 1 1 1 0 ...
##  $ CarName         : Factor w/ 147 levels "alfa-romero giulia",..: 1 3 2 4
5 9 5 7 6 8 ...
##  $ fueltype        : Factor w/ 2 levels "diesel","gas": 2 2 2 2 2 2 2 2 2
2 ...
##  $ aspiration      : Factor w/ 2 levels "std","turbo": 1 1 1 1 1 1 1 1 2 2
...
##  $ doornumber      : Factor w/ 2 levels "four","two": 2 2 2 1 1 2 1 1 1 2
...
##  $ carbody         : Factor w/ 5 levels "convertible",..: 1 1 3 4 4 4 4 5
4 3 ...
##  $ drivewheel      : Factor w/ 3 levels "4wd","fwd","rwd": 3 3 3 2 1 2 2 2
2 1 ...
##  $ enginelocation  : Factor w/ 2 levels "front","rear": 1 1 1 1 1 1 1 1 1
```

```
1 ...
##  $ wheelbase       : num  88.6 88.6 94.5 99.8 99.4 ...
##  $ carlength       : num  169 169 171 177 177 ...
##  $ carwidth        : num  64.1 64.1 65.5 66.2 66.4 66.3 71.4 71.4 71.4
67.9 ...
##  $ carheight       : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 52
...
##  $ curbweight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086
3053 ...
##  $ enginetype      : Factor w/ 7 levels "dohc","dohcv",..: 1 1 6 4 4 4 4 4
4 4 ...
##  $ cylindernumber  : Factor w/ 7 levels "eight","five",..: 3 3 4 3 2 2 2 2
2 2 ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuelsystem      : Factor w/ 8 levels "1bbl","2bbl",..: 6 6 6 6 6 6 6 6
6 6 ...
##  $ boreratio       : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13
3.13 ...
##  $ stroke          : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 3.4 ...
##  $ compressionratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 7 ...
##  $ horsepower      : int  111 111 154 102 115 110 110 110 140 160 ...
##  $ peakrpm         : int  5000 5000 5000 5500 5500 5500 5500 5500 5500
5500 ...
##  $ citympg         : int  21 21 19 24 18 19 19 19 17 16 ...
##  $ highwaympg      : int  27 27 26 30 22 25 25 25 20 22 ...
##  $ price           : num  13495 16500 16500 13950 17450 ...
```

Data-prepossesing

Data Cleaning & pre-processing is the most crucial part in data analysis because the accuracy of data reflects better on the outcome of the analysis. The main objective is to curate the data and prepare it for further exploration and modelling. To ensure that the heart failure clinical records dataset is clean and well-organized, the measures outlined below were followed:

1.) The dataset was checked for missing and special values as a real dataset may contain some missing and special values that need to be dealt with earlier to avoid any problems during analysis.

```
# Checking for missing values

colSums(is.na(car))

##           car_ID        symboling          CarName         fueltype
##                0                0                0                0
##       aspiration       doornumber          carbody       drivewheel
##                0                0                0                0
##   enginelocation        wheelbase        carlength         carwidth
##                0                0                0                0
##        carheight       curbweight       enginetype   cylindernumber
```

```
##                0                0                0                0
##       enginesize       fuelsystem        boreratio           stroke
##                0                0                0                0
## compressionratio       horsepower          peakrpm          citympg
##                0                0                0                0
##       highwaympg            price
##                0                0
```

```
#Checking for special values

is.specialorNA <- function(x){
  if (is.numeric(x)) (is.infinite(x) | is.nan(x) | is.na(x))
}
sapply(car, function(x) sum( is.specialorNA(x) ))
```

```
##           car_ID        symboling          CarName         fueltype
##                0                0                0                0
##       aspiration       doornumber          carbody       drivewheel
##                0                0                0                0
##   enginelocation        wheelbase        carlength         carwidth
##                0                0                0                0
##        carheight       curbweight       enginetype   cylindernumber
##                0                0                0                0
##       enginesize       fuelsystem        boreratio           stroke
##                0                0                0                0
## compressionratio       horsepower          peakrpm          citympg
##                0                0                0                0
##       highwaympg            price
##                0                0
```

From the above output it was observed that the dataset car do not contain any missing or special values.

2.) The column car_ID was removed from the dataset as it only represents the serial number of all the observations and does not contribute significant information in the dataset.

```
#Removing ID column

car <- car[-1]
```

3.) The variable CarName contains values with a lot of typos that generated 147 different levels for it. This needs to be dealt with in order to generate clean and better values for the variable carCompany. The carName variable was changed to carCompany and the values were split to include only the company name of the car for a better understanding of the dataset.

```
#Generating car_company column using car names

car$CarName <- sapply(car$CarName, function(x)  str_split(x,' ')[[1]][1])
```

```r
#Changing name of variable to carCompany

colnames(car)[2] <- "carCompany"

#Dealing with typos


#Correcting typos

car$carCompany[car$carCompany == "maxda"] <- "mazda"
car$carCompany[car$carCompany == "porschce" |car$carCompany == "porcshce"] <-
"porsche"
car$carCompany[car$carCompany == "toyouta"] <- "toyota"
car$carCompany[car$carCompany == "vokswagen"] <- "volkswagen"
car$carCompany[car$carCompany == "vw"] <- "volkswagen"
car$carCompany[car$carCompany == "Nissan"] <- "nissan"


#Changing first character of each value of carCompany to uppercase

car$carCompany <- str_to_title(car$carCompany)

#Applying type conversion to convert the variable to factor

car$carCompany <- as.factor(car$carCompany)

#Checking reduced levels of variable carCompany

levels(car$carCompany)
```

```
##  [1] "Alfa-Romero" "Audi"        "Bmw"         "Buick"       "Chevrolet"
##  [6] "Dodge"       "Honda"       "Isuzu"       "Jaguar"      "Mazda"
## [11] "Mercury"     "Mitsubishi"  "Nissan"      "Peugeot"     "Plymouth"
## [16] "Porsche"     "Renault"     "Saab"        "Subaru"      "Toyota"
## [21] "Volkswagen"  "Volvo"
```

The typos were successfully corrected and the initial 147 unique values of carCompany variable were reduced to only 22.

Data Exploration

Summary Statistics

The summary statistics for the response variable price were generated to get an idea about its range and values.

```r
#Calculating summary statistics for the response variable price
```

```
par(mfrow=c(1,1))
summaryPrice<- car %>% summarise(Min= min(price, na.rm = TRUE),
                                 Q1= quantile(price, probs = 0.25,na.rm
= TRUE),
                                 Median= median(price, na.rm = TRUE),
                                 Q3= quantile(price,probs = 0.75, na.rm
= TRUE),
                                 Max= max(price, na.rm = TRUE),
                                 Mean= mean(price, na.rm = TRUE),
                                 SD= sd(price, na.rm = TRUE),
                                 N= n(),
                                 Missing= sum(is.na(price)))
summaryPrice
```

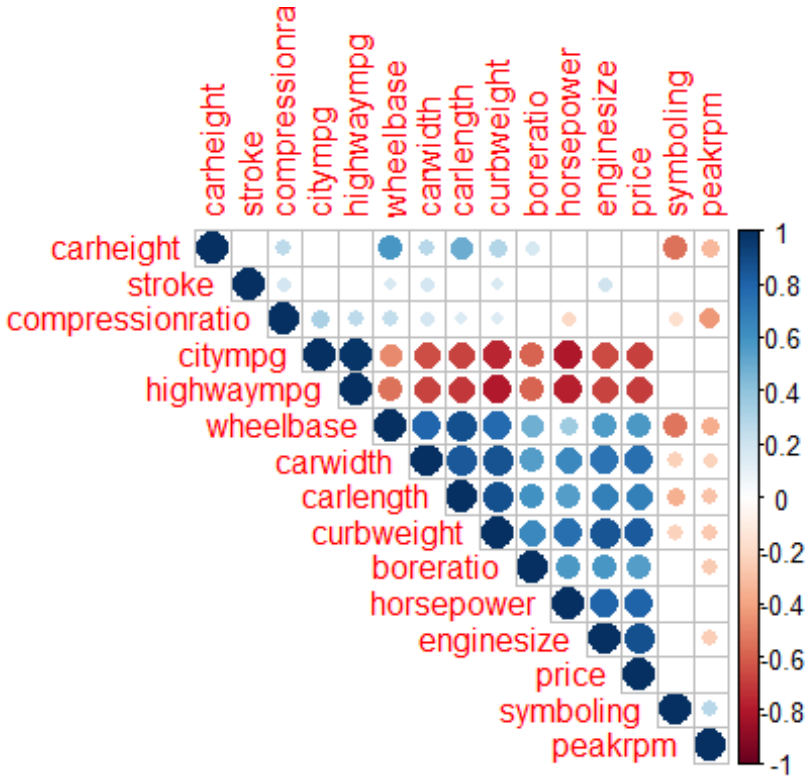
```
##    Min   Q1 Median    Q3   Max     Mean       SD   N Missing
## 1 5118 7788  10295 16503 45400 13276.71 7988.852 205       0
```

From the above summary statistics for price, it can be noted that the cars in the provided dataset fall between the price range of $5,118 to $45,400 with a mean price of $13,276.71.

Generating correlation graph for all numeric variables of dataset car

```
#Selecting all numeric variables
numeric_col <- select_if(car, is.numeric)

##Finding correlations between numeric variables in the dataset
res2<-rcorr(as.matrix(numeric_col))
corrplot(res2$r, type="upper", order="hclust",
         p.mat = res2$P, sig.level = 0.05, insig = "blank")
```

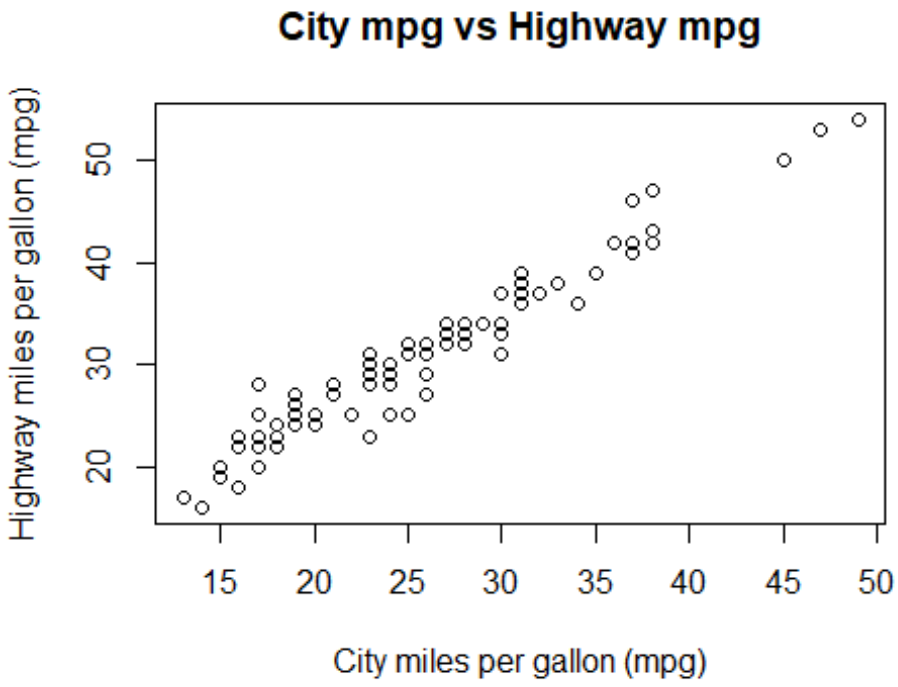


The above corrplot shows correlations between all numeric variables in the dataset. The blanks in the corrplot represent insignificant relationships that exists between numeric variables. From the output, it can be noted that car price is highly positively correlated with the dimensions of the car, i.e., the width, length and the curbweight. Moreover, cars with bigger engines and higher horsepower tend to have a higher price. However, the price of a car is negatively correlated with the miles per gallon for city and highway. Apart from this, significant correlations between the predictor variables can be observed. For instance, the city and highway mpg. This suggests the presence of multicollinearity among the predictors in the dataset.

The multicollinearity was checked visually between citympg and highwaympg by plotting a scatterplot-

```
# Plotting scatterplot

plot(car$citympg , car$highwaympg, main = "City mpg vs Highway mpg",
     xlab = "City miles per gallon (mpg)",
     ylab = "Highway miles per gallon (mpg)")
```



From the above scatterplot, a strong positive linear relationship between the two variable can be noticed. This indicates multicollinearity, which will be dealt with later.

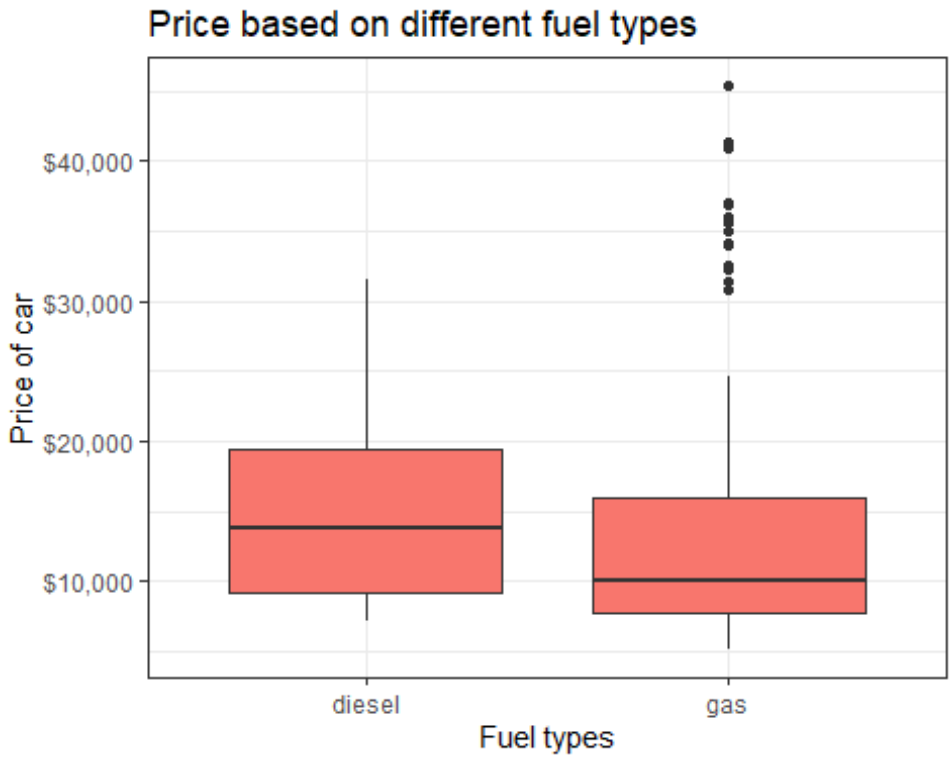
Exploratory Analysis of categorical variables in the dataset

The exploratory analysis for categorical variables in the dataset were carried out to identify the significant cateogrical variables in the dataset car. Bar plots and box-plots were plotted for each categorical variable with the response variable price to understand its effect on the latter.

Fuel type vs Price

```
#Analysing different fueltype with price

p1<- ggplot(data = car, aes(x= fueltype, y= price, fill= "#F46036"))
p1 + geom_boxplot() +
  ggtitle(label = "Price based on different fuel types") +
  labs(x= "Fuel types", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```



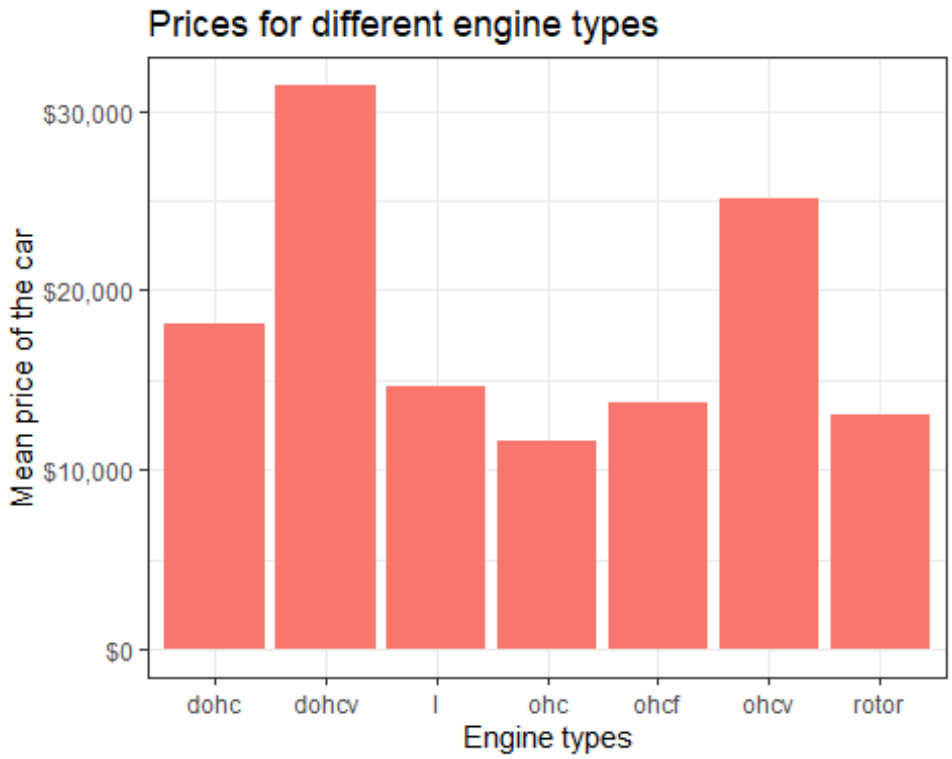
The above boxplot shows that prices of cars vary with the type of fuel the car runs on. It can be noted that price of diesel cars are higher than cars running on gas.

Engine type vs Price

```
#Analysing enginetype with price

p2 <- car %>%
  group_by(enginetype) %>%
  summarise(m = mean(price)) %>%
  ggplot(aes(x= enginetype, y = m, fill= "#F46036")) +
  geom_bar(stat = "identity") +
  labs(y = "Mean price of the car", x = "Engine types")+
  ggtitle(label = "Prices for different engine types") +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)

p2
```



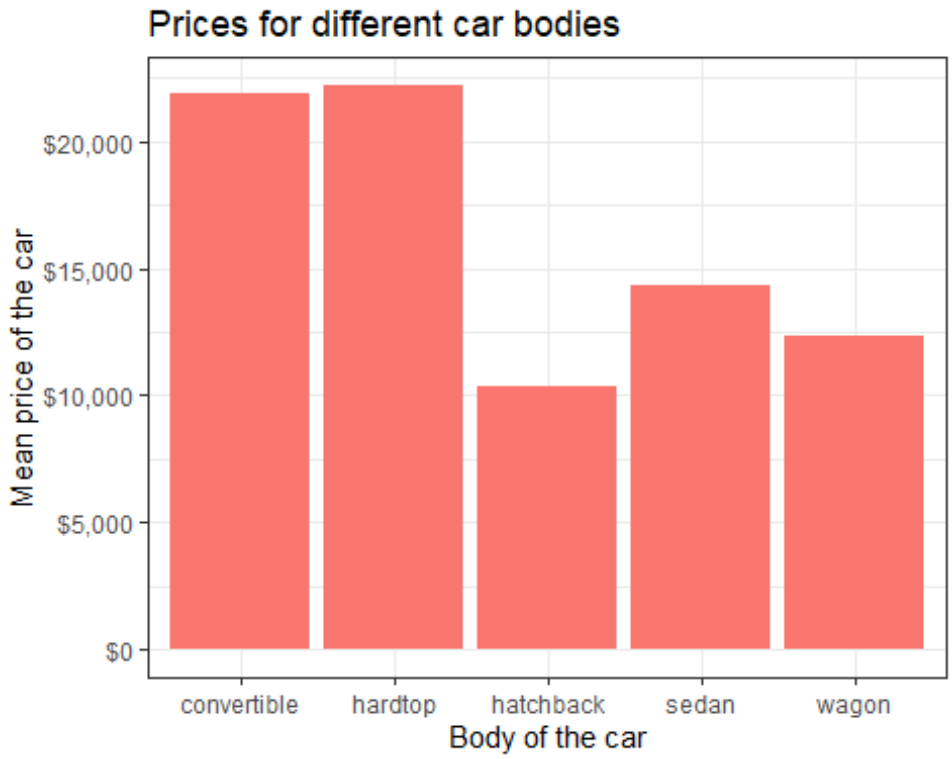
The above barplot shows that cars with dohcv type engine have the expensive cars as their mean price is significantly higher in comparison to cars with other engine types such as ohc, dohc etc. Also, cars with ohc engines tend to have lowest average price in comparison to others.

Car body vs Price

```
#Analysing carbody with price

p3 <- car %>%
  group_by(carbody) %>%
  summarise(m1 = mean(price)) %>%
  ggplot(aes(x= carbody, y = m1, fill= "#F46036")) +
  geom_bar(stat = "identity") +
  labs(y = "Mean price of the car", x = "Body of the car")+
  ggtitle(label = "Prices for different car bodies") +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)

p3
```

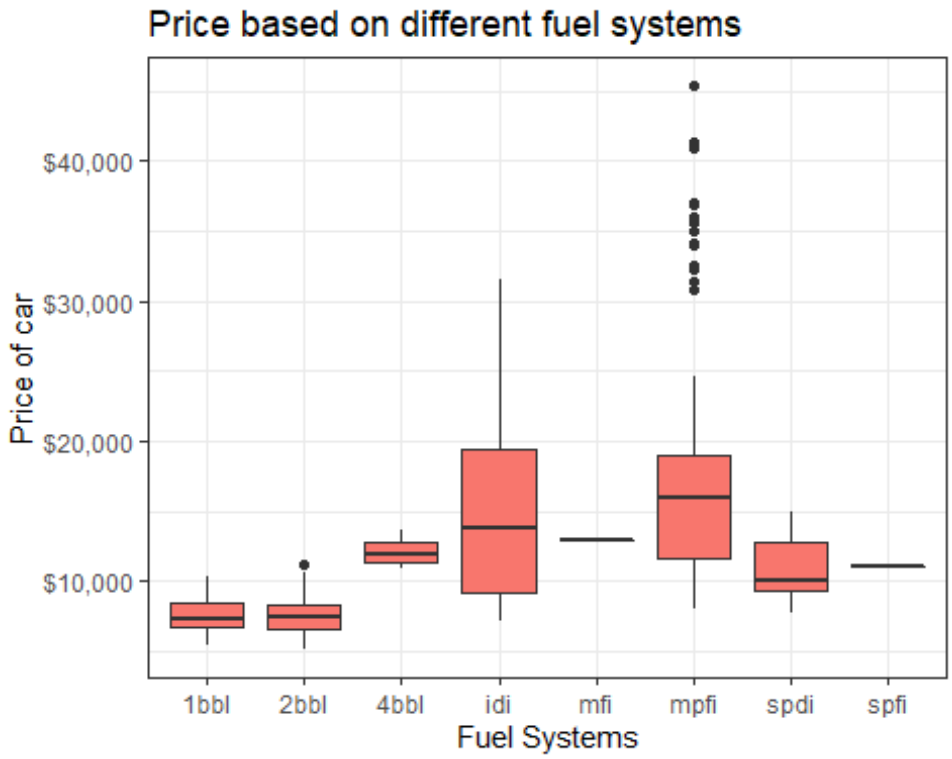


From the above barplot, it is observed that number of cylinders in a car significantly effect the price of a car. It can be seen that cars with eight number of cylinders have the highest mean price followed by cars with twelve number of cylinders. Moreover, cars with three number of cylinders have the lowest average price.

Fuel system vs Price

```
#Analysing fuel system with price

p5<- ggplot(data = car, aes(x= fuelsystem, y= price, fill= "#F46036"))
p5 + geom_boxplot() +
  ggtitle(label = "Price based on different fuel systems") +
  labs(x= "Fuel Systems", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```

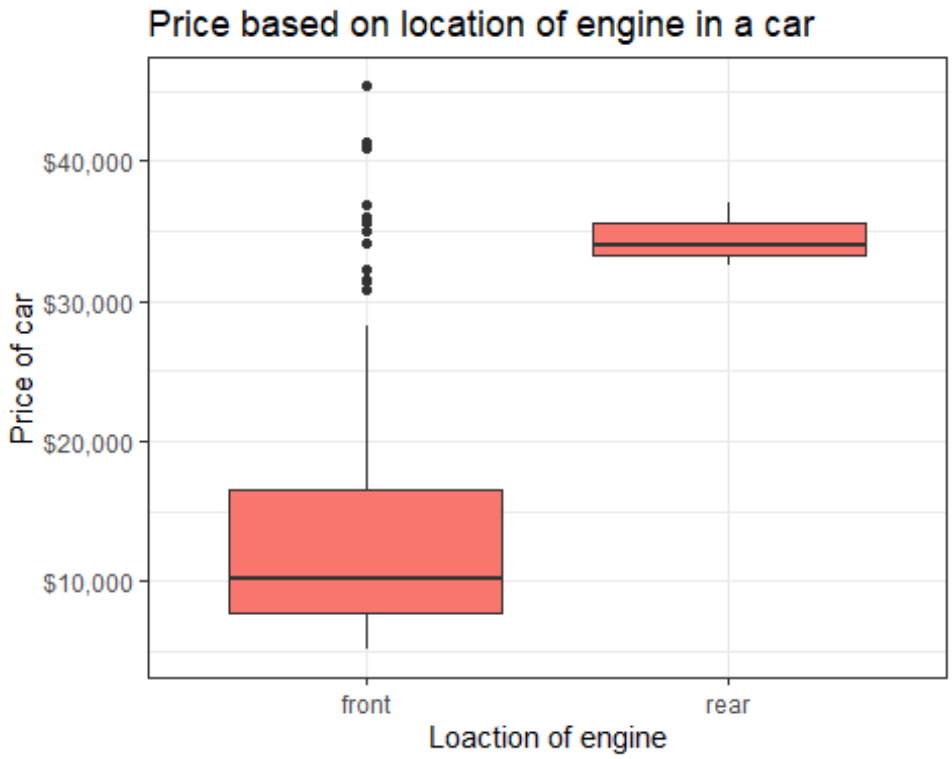


The above boxplot shows that prices of cars seem to slightly vary with the type of fuel systems the car uses but it does not seem to provide any significant information. The fuel system used in a car do not seem to effect its price.

Engine location vs Price

```
##Analysing engine location with price

p6<- ggplot(data = car, aes(x= enginelocation, y= price, fill= "#F46036"))
p6 + geom_boxplot() +
  ggtitle(label = "Price based on location of engine in a car") +
  labs(x= "Loaction of engine", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```

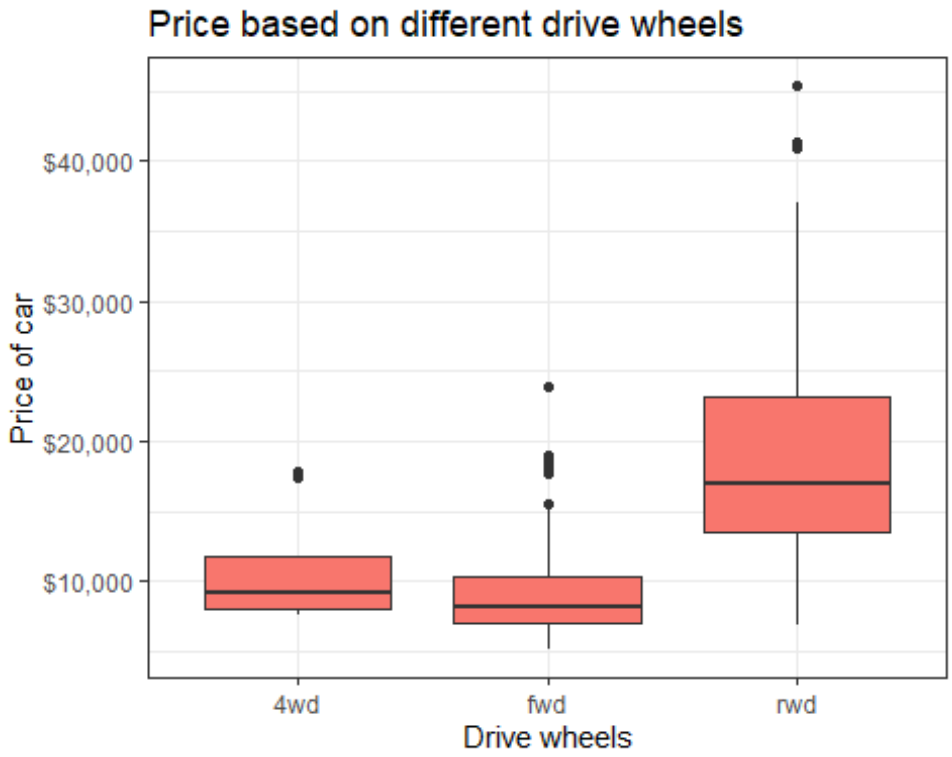


The above boxplot shows that location of engine in a car have significant effect on car price. It can be noted that the cars with engine located in the rear have significantly very high price in comparison to cars with engine located in the front.

Drive wheel vs Price

```
##Analysing drive wheel with price

p7<- ggplot(data = car, aes(x= drivewheel, y= price, fill= "#F46036"))
p7 + geom_boxplot() +
  ggtitle(label = "Price based on different drive wheels") +
  labs(x= "Drive wheels", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```

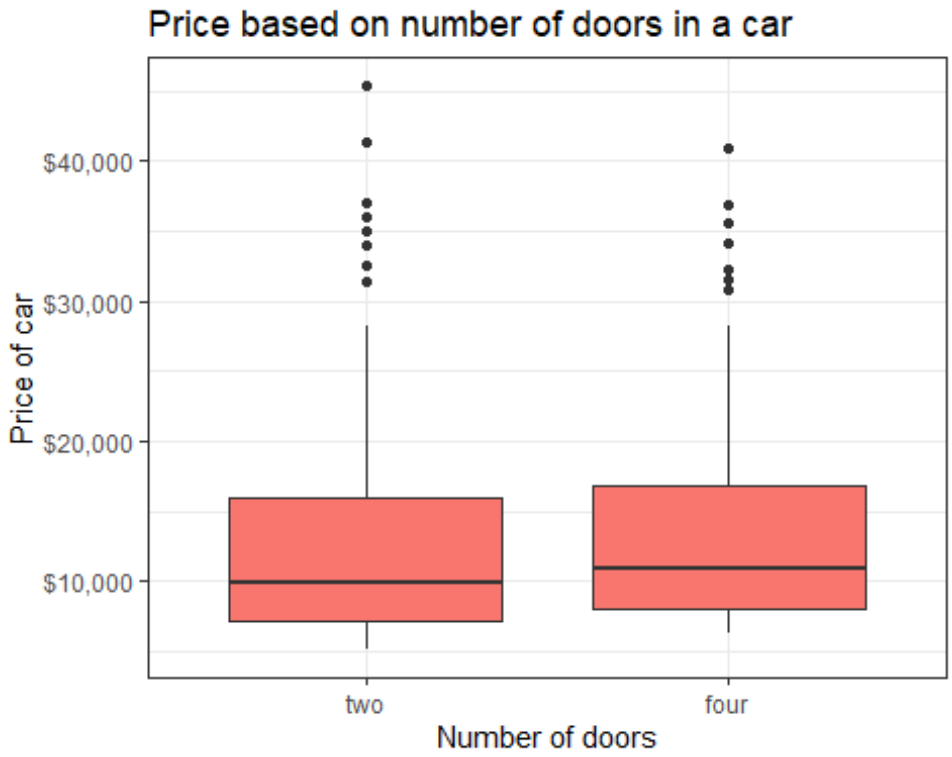


The above boxplot shows that type of drive wheels in a car have significant effect on car price. It can be seen that the cars with rear wheel drives have significantly higher price in comparison to cars with 4wd and front wheel drives.

Door number vs Price

```
##Analysing door number with price

p8<- ggplot(data = car, aes(x= factor(doornumber, levels = c("two", "four")),
y= price, fill= "#F46036"))
p8 + geom_boxplot() +
  ggtitle(label = "Price based on number of doors in a car") +
  labs(x= "Number of doors", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```

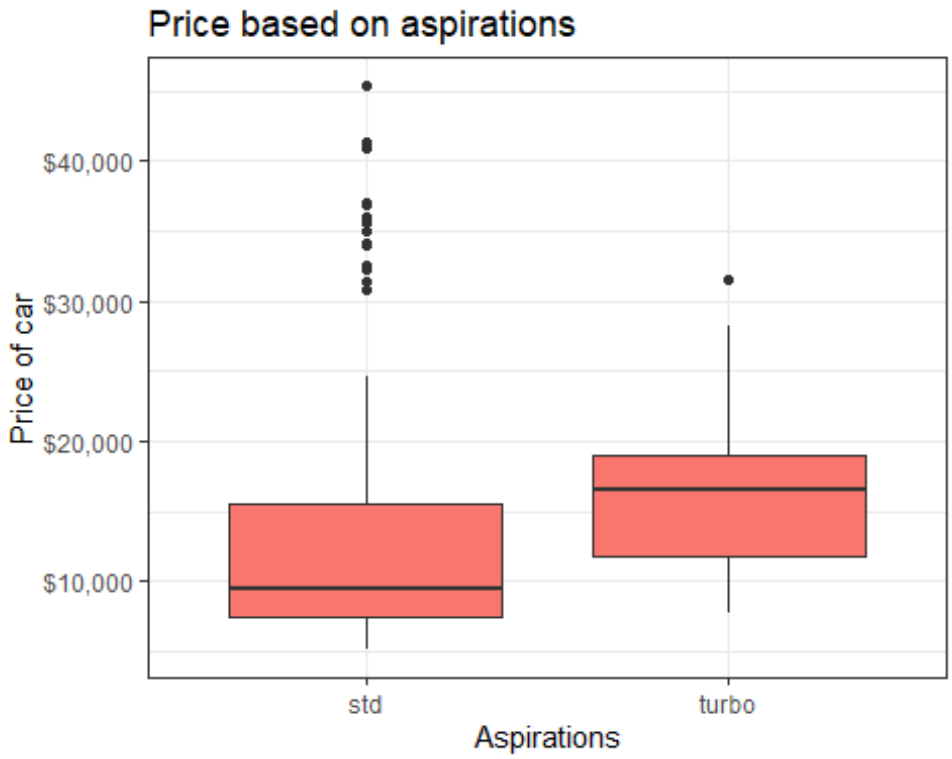


The boxplot shows that number of doors in a car do not contribute significant information in deciding the price of car.

Aspiration vs Price

```
#Analysing aspiration vs price

p9<- ggplot(data = car, aes(x= aspiration, y= price, fill= "#F46036"))
p9 + geom_boxplot() +
  ggtitle(label = "Price based on aspirations") +
  labs(x= "Aspirations", y= "Price of car")+
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)
```



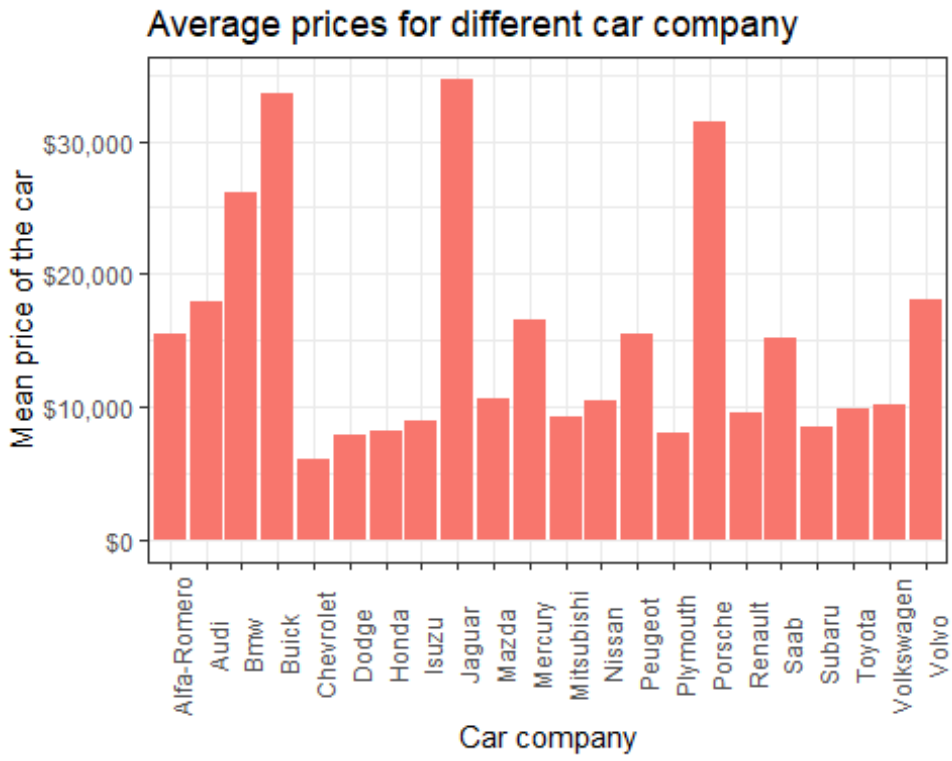
The above boxplot shows that aspiration variable significantly effect the price of a car.It can be seen that the cars with aspiration as turbo have significantly higher price in comparison to cars with aspiration as std.

Car company vs Price

```
#Analysing car company with price

p10 <- car %>%
  group_by(carCompany) %>%
  summarise(m3 = mean(price)) %>%
  ggplot(aes(x= carCompany, y = m3, fill= "#F46036")) +
  geom_bar(stat = "identity") +
  labs(y = "Mean price of the car", x = "Car company")+
  ggtitle(label = "Average prices for different car company") +
  theme(panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        plot.title = element_text(color = "black", size = 12, face = "bold"))
+ theme_bw() +
  theme(legend.position = "none") +
  scale_y_continuous(labels = scales::dollar)+
  theme(axis.text.x = element_text(angle = 90))

p10
```



The above barplot shows that the company which the car belongs to sigificantly effect its average price. It can be observed that Porsche, Jaguar and Buick cars have significantly very high average price while Chevrolet, Plymouth and Sabanu etc. cars have lower average price than cars that belong to other companies.

Regression Analysis

Before begining the regression analysis on the data, feature selection will be performed. From the above plot, it can be observed that some of the independent variables listed are not significantly influencing the price of car, therefore, these variables will be removed from the data set for further analysis.

```
#Removing unimportant variables

car <- car[ , -which(names(car) %in% c("fuelsystem","doornumber",
"symboling", "peakrpm", "compressionratio", "stroke", "carheight"))]

dim(car)

## [1] 205  18
```

Model 1 : Full Model

Fitting the multiple regression model with all predictors selected above.

```
model_1 <- lm(price ~ ., data = car)
summary(model_1)
```

```
##
## Call:
## lm(formula = price ~ ., data = car)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3457.1 -1041.9     0.0   790.5 10098.0
##
## Coefficients: (3 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -36755.861  12013.736  -3.059 0.002611 **
## carCompanyAudi          546.978   2135.409   0.256 0.798175
## carCompanyBmw          8112.457   2140.156   3.791 0.000214 ***
## carCompanyBuick        3534.661   2503.171   1.412 0.159920
## carCompanyChevrolet   -2052.483   2170.825  -0.945 0.345873
## carCompanyDodge       -2912.339   1701.297  -1.712 0.088914 .
## carCompanyHonda       -2065.971   1665.683  -1.240 0.216721
## carCompanyIsuzu        -738.927   1898.489  -0.389 0.697645
## carCompanyJaguar       2644.605   2582.548   1.024 0.307407
## carCompanyMazda       -1153.349   1668.702  -0.691 0.490489
## carCompanyMercury     -2614.127   2839.391  -0.921 0.358647
## carCompanyMitsubishi  -3581.056   1662.427  -2.154 0.032767 *
## carCompanyNissan      -1388.905   1598.059  -0.869 0.386117
## carCompanyPeugeot     -3529.292   4438.707  -0.795 0.427753
## carCompanyPlymouth    -2959.037   1692.914  -1.748 0.082450 .
## carCompanyPorsche      5068.362   2804.701   1.807 0.072674 .
## carCompanyRenault     -2881.195   2078.309  -1.386 0.167628
## carCompanySaab          693.207   1775.989   0.390 0.696831
## carCompanySubaru      -1144.126   1896.055  -0.603 0.547102
## carCompanyToyota      -2494.203   1580.028  -1.579 0.116458
## carCompanyVolkswagen  -2166.649   1616.693  -1.340 0.182138
## carCompanyVolvo         220.337   2015.975   0.109 0.913109
## fueltypegas              39.881   1050.317   0.038 0.969760
## aspirationturbo         790.235    695.491   1.136 0.257605
## carbodyhardtop        -2680.403   1219.891  -2.197 0.029477 *
## carbodyhatchback      -3195.892   1116.167  -2.863 0.004770 **
## carbodysedan          -2460.863   1169.154  -2.105 0.036909 *
## carbodywagon          -2980.192   1238.825  -2.406 0.017314 *
## drivewheelfwd          -297.604    896.154  -0.332 0.740267
## drivewheelrwd           140.976   1045.825   0.135 0.892944
## enginelocationrear    10023.322   2687.238   3.730 0.000268 ***
## wheelbase               157.967     77.281   2.044 0.042629 *
## carlength              -107.402     49.641  -2.164 0.032017 *
## carwidth                780.131    222.663   3.504 0.000599 ***
## curbweight                4.404      1.656   2.660 0.008634 **
## enginetypedohcv       -9263.847   4399.675  -2.106 0.036843 *
## enginetypel             -176.306   4212.110  -0.042 0.966666
## enginetypeohc           -429.182   1090.045  -0.394 0.694319
## enginetypeohcf                NA         NA      NA       NA
## enginetypeohcv        -2442.546   1236.159  -1.976 0.049928 *
```

```
## enginetyperotor          11.057   4206.842   0.003 0.997906
## cylindernumberfive    -5265.870   2892.943  -1.820 0.070638 .
## cylindernumberfour    -3994.422   3554.682  -1.124 0.262863
## cylindernumbersix     -5668.546   2639.413  -2.148 0.033285 *
## cylindernumberthree          NA         NA      NA       NA
## cylindernumbertwelve  -8067.533   3622.922  -2.227 0.027393 *
## cylindernumbertwo            NA         NA      NA       NA
## enginesize               62.112     24.006   2.587 0.010584 *
## boreratio             -4382.042   1784.363  -2.456 0.015154 *
## horsepower               35.064     18.611   1.884 0.061416 .
## citympg                 -92.272    134.299  -0.687 0.493062
## highwaympg              128.322    115.934   1.107 0.270063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1833 on 156 degrees of freedom
## Multiple R-squared:  0.9598, Adjusted R-squared:  0.9474
## F-statistic:  77.5 on 48 and 156 DF,  p-value: < 2.2e-16
```

The summary of the regression model suggests that the fitted model is significant because the overall p-value is less than the 5% significance level. Moreover, the adjusted r-squared value indicates that 94.74% of variation in the price of a car is explained by all the seventeen predictors. However, from the summary of the model, it can be noted that the certain attributes such as fuel type of the cars, drive wheels, citympg and highwaympg have p-values greater than 0.05 significance, therefore they can be considered insignificant in evaluating the response.

Checking for Multicollinearity

When an independent variable in a multiple regression equation is strongly correlated with one or more of the other independent variables, multicollinearity occurs. Multicollinearity is a concern since it reduces an independent variable's statistical significance.

Previously using correlation plot, multicollinearity between several predictor variables was noted.

Multicollinearity can also be detected using the vif() function. If the vif values for the respective predictor variables is greater than 10, then they are assumed to be multicollinear.

```
vif(model_1)
```

```
##       carCompanyAudi        carCompanyBmw      carCompanyBuick
##               9.1787              10.4830              14.3410
##  carCompanyChevrolet      carCompanyDodge      carCompanyHonda
##               4.1474               7.4151              10.0570
##      carCompanyIsuzu     carCompanyJaguar      carCompanyMazda
##               4.2085               5.8698              12.9250
##    carCompanyMercury carCompanyMitsubishi     carCompanyNissan
##               2.3886              10.0180              12.4840
```

```
##    carCompanyPeugeot   carCompanyPlymouth    carCompanyPorsche
##              61.0610               5.7689              11.4240
##    carCompanyRenault       carCompanySaab     carCompanySubaru
##               2.5468               5.4694              12.0920
##     carCompanyToyota carCompanyVolkswagen      carCompanyVolvo
##              20.0720               8.7912              12.5960
##          fueltypegas      aspirationturbo       carbodyhardtop
##               5.9279               4.3667               3.4061
##     carbodyhatchback         carbodysedan         carbodywagon
##              17.0980              20.7730              10.0300
##        drivewheelfwd        drivewheelrwd   enginelocationrear
##              11.8970              15.5730               6.3554
##            wheelbase            carlength             carwidth
##              13.1530              22.7800              13.8570
##           curbweight      enginetypedohcv          enginetypel
##              45.1390               5.7349              59.6750
##        enginetypeohc       enginetypeohcv      enginetyperotor
##              14.5570               5.5392              20.6650
##   cylindernumberfive   cylindernumberfour    cylindernumbersix
##              25.9380             134.2200              43.9500
## cylindernumbertwelve           enginesize            boreratio
##               3.8887              60.6970              14.1850
##           horsepower              citympg           highwaympg
##              32.8950              46.8840              38.7130
```

VIF is the Variance Inflation Factor, that indicates multicollinearity amongst the independent variables in the data. The results of VIF shows significant strong multicollinearity in variables citympg, highwaympg, horespower, enginsize, cylindernumber, curbweight and enginetype.

To deal with multicollinearity, mean centering approach will be used using the meanCenter() function.

```
mean_center_func <- function(x) {
  if (is.numeric(x)) (sapply(x, function(y) meanCenter(y)))
}

mean_center_func(car)

#Checking for multicollinearity after mean centering
vif(model_1)
```

```
##       carCompanyAudi        carCompanyBmw      carCompanyBuick
##               9.1787              10.4830              14.3410
##  carCompanyChevrolet      carCompanyDodge      carCompanyHonda
##               4.1474               7.4151              10.0570
##      carCompanyIsuzu     carCompanyJaguar      carCompanyMazda
##               4.2085               5.8698              12.9250
##    carCompanyMercury carCompanyMitsubishi     carCompanyNissan
##               2.3886              10.0180              12.4840
```

```
##    carCompanyPeugeot   carCompanyPlymouth    carCompanyPorsche
##              61.0610               5.7689              11.4240
##    carCompanyRenault       carCompanySaab     carCompanySubaru
##               2.5468               5.4694              12.0920
##     carCompanyToyota carCompanyVolkswagen      carCompanyVolvo
##              20.0720               8.7912              12.5960
##          fueltypegas      aspirationturbo       carbodyhardtop
##               5.9279               4.3667               3.4061
##     carbodyhatchback         carbodysedan         carbodywagon
##              17.0980              20.7730              10.0300
##        drivewheelfwd        drivewheelrwd   enginelocationrear
##              11.8970              15.5730               6.3554
##            wheelbase            carlength             carwidth
##              13.1530              22.7800              13.8570
##           curbweight      enginetypedohcv         enginetypel
##              45.1390               5.7349              59.6750
##        enginetypeohc       enginetypeohcv      enginetyperotor
##              14.5570               5.5392              20.6650
##   cylindernumberfive   cylindernumberfour    cylindernumbersix
##              25.9380             134.2200              43.9500
## cylindernumbertwelve           enginesize           boreratio
##               3.8887              60.6970              14.1850
##           horsepower              citympg          highwaympg
##              32.8950              46.8840              38.7130
```

After applying mean centering on all the numeric variables, vif() returned values greater than 10 for several variables, indicating presence of multicollinearity.

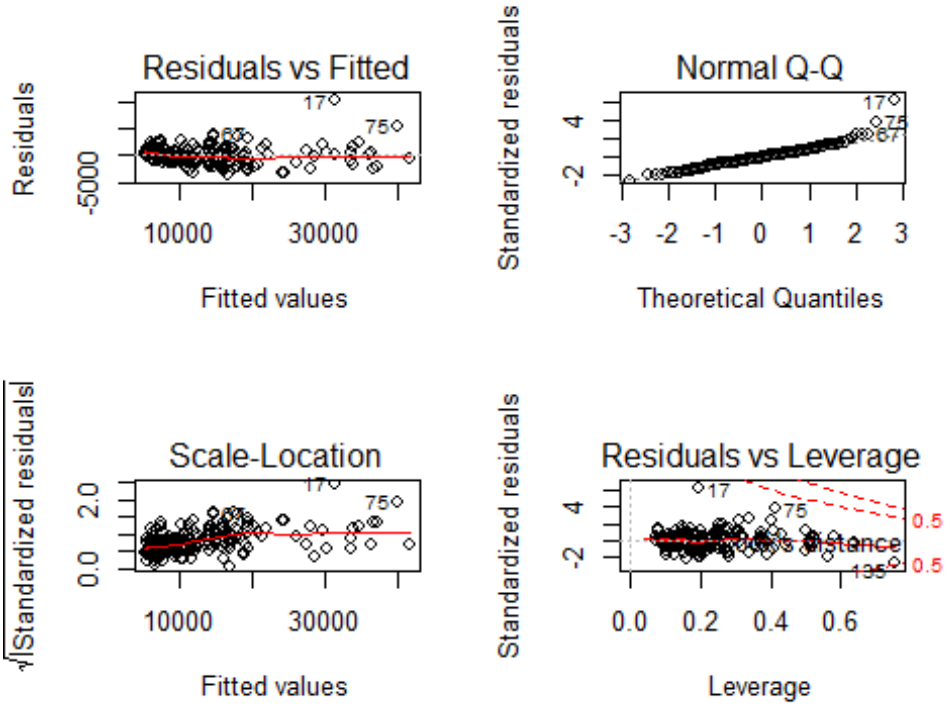
Residual Analysis

The analysis of residuals plays a crucial role in validating the regression model. Checking the underlying assumptions for residuals is important to evaluate the goodness of fit for the model. The residuals are assumed to have the following:

- Normal distribution
- Linearity
- Constant variance
- Independence of errors

Using the residual plots as well as some supporting statistical test, residual analysis will be performed.

```
par(mfrow=c(2,2))
plot(model_1)
```




```
par(mfrow=c(1,1))
```

Residuals vs Fitted values

The red line in the plot appear to be horizontal along 0 but the residuals are not evenly distributed. Moreover, they form a cluster around the lower fitted values on x-axis.Also, few outliers can be seen in the graph. Therefore, considering these observations, it can be stated that the residuals are non-linear in nature and the linearity assumption seems to be violated.

Scale-Location Plot

The scale location plot indicates some non-linearity, but it can also be observed that the dispersion of magnitudes appears to be lowest in the fitted values close to 40000, largest in the fitted values around 10000. This suggests the assumptiion of homoscedasticity is violated. To support the interpretation of non-constant variance, ncv test will be performed.

The statistical hypotheses for the ncv test is :

```
H0: Errors have constant variance
Ha: Errors have non-constant variance

ncvTest(model_1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 74.69626, Df = 1, p = < 2.22e-16
```

From the output of ncvTest, a p-value less than 0.05 can be seen. This implies there is enough statistical evidence to reject the null hypothesis. Therefore, the constant variance assumption for the residuals is violated.

Normal Q-Q Plot

From the normal Q-Q plot, it can be seen that the residuals fall over the dotted line but deviate from the tails. But it can be stated from the plot that the residuals seem to follow a normal distribution. But in addition to the plot, a Shapiro-Wilk test will be performed to support and verify the assumption of normality.

The statistical hypothese for the normality test is as follows:

```
H0: Errors are normally distributed
Ha: Errors are not normally distributed

shapiro.test(model_1$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model_1$residuals
## W = 0.92548, p-value = 1.083e-08
```

The Shapiro-Wilk test resulted in a p-value less than 0.05, which indicates that there is enough evidence to reject the null hypothesis hence stating that errors are not normally distributed. Therefore, the assumption of normal distribution of residuals is violated.

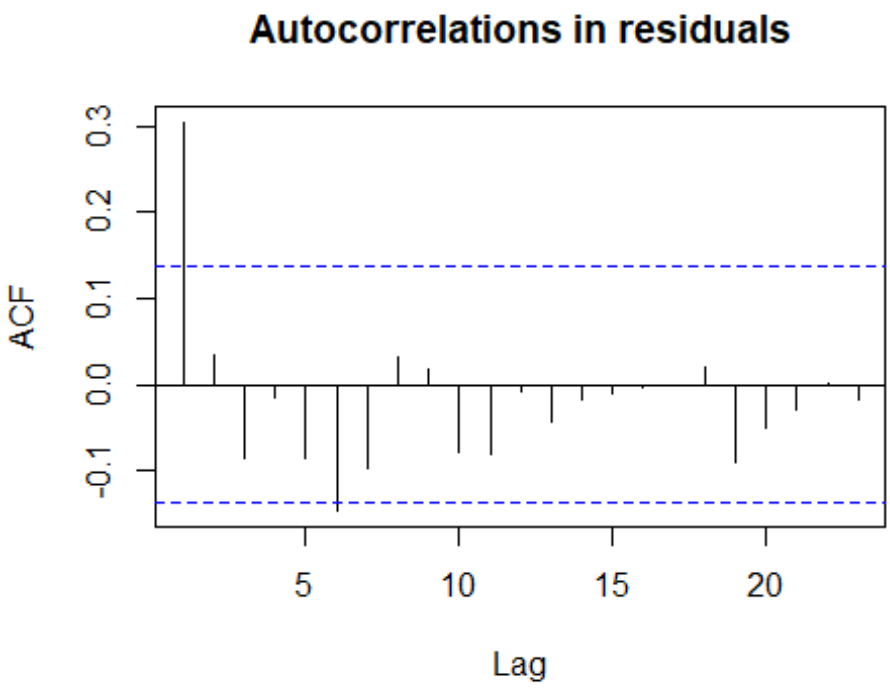
Residuals vs Leverage Plot

The Residuals vs. Leverage plots aid in identifying the model's most influential data points. From the plot, one point can be observed that falls below the Cook's distance and has value of y-axis higher than other values and have low leverage. This point can be considered influential and is worth investigating further.

Autocorrelation of errors

Autocorrelation in errors can be verified using the ACF plot as well as the Durbin Watson statistical test.

```
acf(model_1$residuals, main= "Autocorrelations in residuals")
```



From the ACF plot, two significant autocorrelations can be noted that falls beyond the confidence interval.Therefore, the errors can not be considered independent.

Durbin Watson test can be performed to support the independence check for residuals. The statistical hypothesis for the test of autocorrelation is given as:

```
H0: Errors are uncorrelated
Ha: Errors are correlated

durbinWatsonTest(model_1)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.3048747      1.366959       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson test results in a p-value less than the 5% significance level. This provides enough statistical evidence to reject the null hypothesis. Therefore, it can be stated that the residuals are correlated. And hence the assumption of independent residuals is violated.

Therefore, to conclude the residual analysis, it can be stated that the residuals violates all assumptions. Therefore, the model with all 17 predictors can not be considered a good model.

In order to deal with heteroscedasticity, log transformation will be performed on the response variable and a new model for the transformed data will be fitted.

Model 2

```
model_2 <- lm(log(price) ~ . , data = car)
summary(model_2)
```

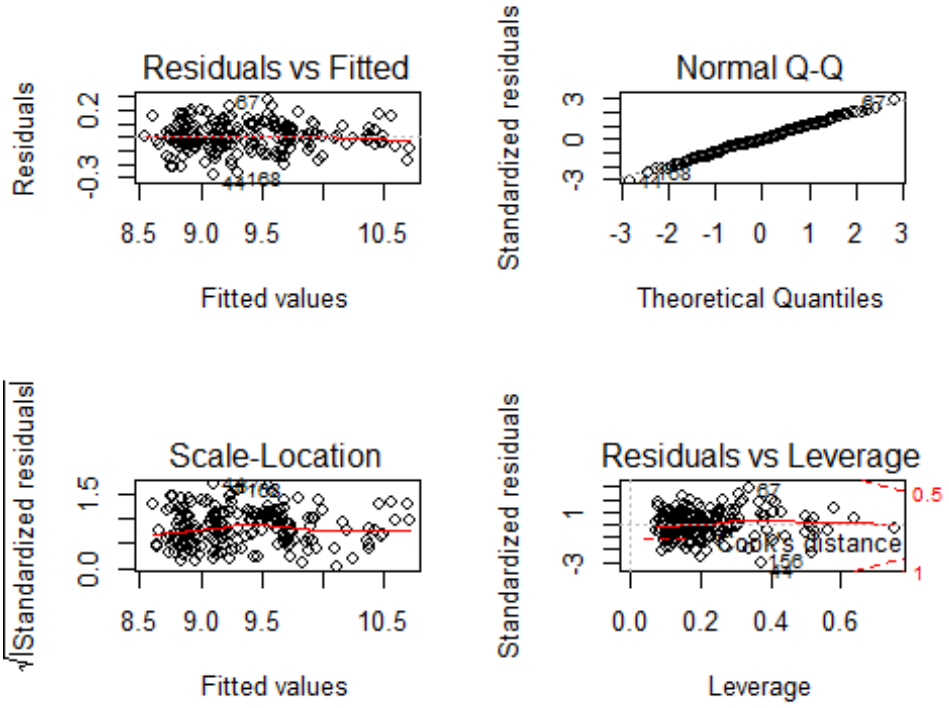
```
##
## Call:
## lm(formula = log(price) ~ ., data = car)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.277114 -0.055529 -0.002329  0.070253  0.269553
##
## Coefficients: (3 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.7725865  0.7631413   7.564 3.15e-12 ***
## carCompanyAudi         0.0800492  0.1356463   0.590 0.555956
## carCompanyBmw          0.3868354  0.1359479   2.845 0.005031 **
## carCompanyBuick       -0.0726194  0.1590074  -0.457 0.648519
## carCompanyChevrolet   -0.1256779  0.1378960  -0.911 0.363493
## carCompanyDodge       -0.2138309  0.1080705  -1.979 0.049619 *
## carCompanyHonda       -0.0985464  0.1058082  -0.931 0.353102
## carCompanyIsuzu       -0.0180413  0.1205965  -0.150 0.881273
## carCompanyJaguar      -0.2174248  0.1640497  -1.325 0.186990
## carCompanyMazda       -0.0828377  0.1060000  -0.781 0.435700
## carCompanyMercury     -0.1891567  0.1803649  -1.049 0.295918
## carCompanyMitsubishi  -0.2533691  0.1056014  -2.399 0.017606 *
## carCompanyNissan      -0.0964898  0.1015125  -0.951 0.343318
## carCompanyPeugeot     -0.5810111  0.2819573  -2.061 0.040998 *
## carCompanyPlymouth    -0.2284332  0.1075380  -2.124 0.035229 *
## carCompanyPorsche      0.2871310  0.1781613   1.612 0.109063
## carCompanyRenault     -0.1822446  0.1320192  -1.380 0.169426
## carCompanySaab         0.0165098  0.1128151   0.146 0.883839
## carCompanySubaru      -0.1811571  0.1204419  -1.504 0.134577
## carCompanyToyota      -0.1675529  0.1003672  -1.669 0.097044 .
## carCompanyVolkswagen  -0.1161742  0.1026962  -1.131 0.259690
## carCompanyVolvo       -0.0536385  0.1280595  -0.419 0.675897
## fueltypegas           -0.0667936  0.0667187  -1.001 0.318318
## aspirationturbo        0.0683697  0.0441793   1.548 0.123757
## carbodyhardtop        -0.1969239  0.0774904  -2.541 0.012021 *
## carbodyhatchback      -0.2414129  0.0709016  -3.405 0.000842 ***
## carbodysedan          -0.1827853  0.0742674  -2.461 0.014938 *
## carbodywagon          -0.2429835  0.0786931  -3.088 0.002388 **
## drivewheelfwd         -0.0048874  0.0569258  -0.086 0.931692
## drivewheelrwd          0.0499580  0.0664333   0.752 0.453182
## enginelocationrear     0.3652373  0.1706998   2.140 0.033939 *
## wheelbase              0.0102157  0.0049091   2.081 0.039069 *
## carlength             -0.0049885  0.0031533  -1.582 0.115676
```

```
## carwidth              0.0442493  0.0141441   3.128 0.002097 **
## curbweight            0.0004597  0.0001052   4.371 2.25e-05 ***
## enginetypedohcv      -0.5687629  0.2794779  -2.035 0.043535 *
## enginetypel           0.1874528  0.2675633   0.701 0.484601
## enginetypeohc        -0.0765134  0.0692423  -1.105 0.270856
## enginetypeohcf               NA         NA      NA       NA
## enginetypeohcv       -0.0831025  0.0785238  -1.058 0.291550
## enginetyperotor       0.1636147  0.2672287   0.612 0.541254
## cylindernumberfive   -0.0602233  0.1837667  -0.328 0.743566
## cylindernumberfour    0.0351754  0.2258019   0.156 0.876408
## cylindernumbersix    -0.1140532  0.1676619  -0.680 0.497350
## cylindernumberthree          NA         NA      NA       NA
## cylindernumbertwelve -0.1986660  0.2301367  -0.863 0.389324
## cylindernumbertwo            NA         NA      NA       NA
## enginesize            0.0015980  0.0015249   1.048 0.296297
## boreratio            -0.1799628  0.1133470  -1.588 0.114375
## horsepower            0.0020276  0.0011822   1.715 0.088306 .
## citympg              -0.0193289  0.0085310  -2.266 0.024843 *
## highwaympg            0.0134266  0.0073644   1.823 0.070192 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1164 on 156 degrees of freedom
## Multiple R-squared:  0.9592, Adjusted R-squared:  0.9466
## F-statistic: 76.34 on 48 and 156 DF,  p-value: < 2.2e-16
```

From the summary above, it can be noted that the transformed model is significant with a p-value less than 0.05, with an adjusted r-squared value of 0.9466. Residual analysis of the log transformed model will be performed to check the assumptions of residuals.

Residual Analysis

```
par(mfrow=c(2,2))
plot(model_2)
```




```
par(mfrow=c(1,1))
```

Residuals vs Fitted values

In the residual vs fitted values plot linearity seems to hold reasonably well, as the red line is close to the dashed line. Furthermore, the residuals appear to be equally variable across the entire range of fitted values. There is no indication of non-constant variance.

Scale-Location Plot

The standard linear regression assumption is that the variance is constant across the entire range. From the scale - location plot, residuals are observed to be randomly spread and the red line seems approximately horizontal, indicating presence of homoscedasticity. Therefore, it can be stated that log transformation on the response variable was useful in dealing with non-constant variance. However, to support the observation of the plot, p-value from ncv test will be used.

```
ncvTest(model_2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6323118, Df = 1, p = 0.42651
```

The ncv test (i.e., non-constant variance test) yielded a p-value = 0.42651. It is greater than the 5% significance level ,therefore, the null hypothesis stating constant variance for

residuals can not be rejected. To sum up, the assumption of constant variance is not violated for this model.

Normal Q-Q Plot

The normal Q-Q plot suggests that the residuals are normally distributed because they fall well on the diagonal line. To aid the results, Shapiro-Wilk test will be used.

```
shapiro.test(model_2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_2$residuals
## W = 0.99555, p-value = 0.8138
```

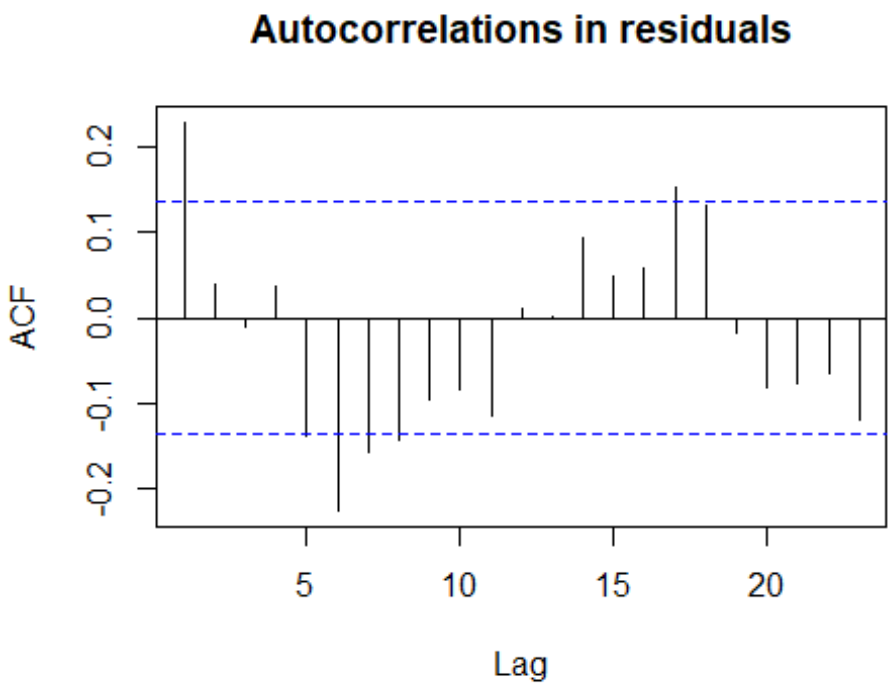
A p-value greater than the 5% significance level was obtained from the Shapiro-Wilk test, thus, the null hypothesis stating normal distribution of errors can not be rejected. Hence, the normality assumption for residuals is not violated here.

Residuals vs Leverage Plot

From the residual vs leverage plot, one point close to the Cook's distance can be observed that may have substantial influence on the regression model.

Autocorrelation of errors

```
acf(model_2$residuals, main= "Autocorrelations in residuals")
```



The ACF plot suggests that few significant autocorrelations between residuals exists. However, Durbin Watson test will be performed to test autocorrelation.

```
durbinWatsonTest(model_1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3048747      1.366959       0
##  Alternative hypothesis: rho != 0
```

The above test yields a p-value less than 0.05 indicating that there is enough evidence to reject the null hypothesis (H0: Residuals are uncorrelated). Therefore, it can be stated that the independence assumption for residuals is violated.

To conclude the residual analysis, it can be noted that applying log transformation on the response resulted in a better model with constant variance. The residuals were also observed to be linear in nature with a normal distribution. However, the assumption of independent residuals was violated. In corrospondence to the improved results, log transformed response variable will be used for further analysis during the course of this project.

Anova

Anova test will be performed to check the significance of the model and the predictors.

```
anova(model_2)
```

```
## Analysis of Variance Table
##
## Response: log(price)
##                  Df Sum Sq Mean Sq  F value    Pr(>F)
## carCompany       21 37.857 1.80271 133.0094 < 2.2e-16 ***
## fueltype          1  0.006 0.00633   0.4669   0.49543
## aspiration        1  1.685 1.68484 124.3124 < 2.2e-16 ***
## carbody           4  0.343 0.08564   6.3186 9.730e-05 ***
## drivewheel        2  3.140 1.56999 115.8388 < 2.2e-16 ***
## enginelocation    1  0.052 0.05168   3.8134   0.05263 .
## wheelbase         1  2.867 2.86673 211.5166 < 2.2e-16 ***
## carlength         1  1.166 1.16562  86.0032 < 2.2e-16 ***
## carwidth          1  0.906 0.90586  66.8374 9.564e-14 ***
## curbweight        1  1.199 1.19881  88.4516 < 2.2e-16 ***
## enginetype        5  0.172 0.03445   2.5419   0.03048 *
## cylindernumber    4  0.076 0.01906   1.4066   0.23430
## enginesize        1  0.021 0.02086   1.5388   0.21666
## boreratio         1  0.035 0.03537   2.6099   0.10822
## horsepower        1  0.070 0.07023   5.1817   0.02419 *
## citympg           1  0.026 0.02561   1.8899   0.17119
## highwaympg        1  0.045 0.04505   3.3239   0.07019 .
## Residuals       156  2.114 0.01355
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows the sum of squares for the regressors as well as the residuals. The sum of squares of residuals for the model with all the predictors is equal to 156. However, to analyse the significance of each predictor in the model, the p-values are examined. It can be interpreted from the anova output that the variables fueltype, cylinder number, engine size, boreratio, citympg and highway mpg are insignificant in analysisng the price of a car beacuse of the very high p-values. Using the results of anova, another model will be fitted with the variables that are significant as per the anova test.

Model 3

Model of significant variables from the above anova test is as follows:

```
model_3 <- lm(log(price) ~ carCompany + aspiration + carbody + drivewheel +
enginelocation + wheelbase + carlength + enginetype + carwidth  + horsepower
+ curbweight, data = car)

summary(model_3)
```

```
##
## Call:
## lm(formula = log(price) ~ carCompany + aspiration + carbody +
##     drivewheel + enginelocation + wheelbase + carlength + enginetype +
##     carwidth + horsepower + curbweight, data = car)
##
## Residuals:
```

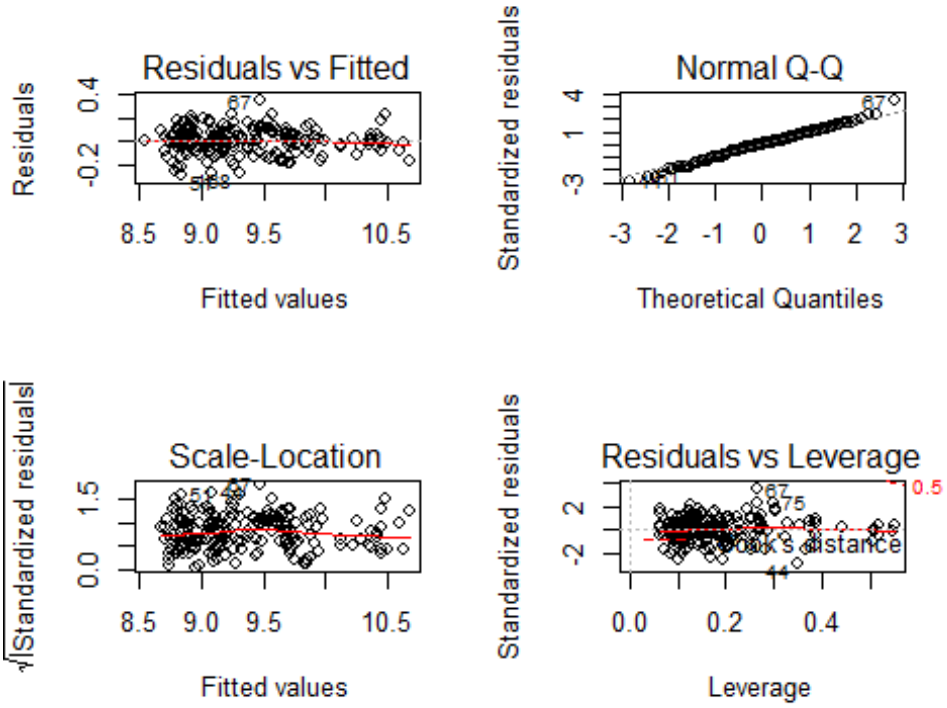
```
##      Min       1Q   Median       3Q      Max
## -0.28447 -0.05462  0.00174  0.06445  0.34380
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.396e+00  6.881e-01   7.841 5.25e-13 ***
## carCompanyAudi        2.008e-02  1.114e-01   0.180 0.857190
## carCompanyBmw         2.831e-01  1.067e-01   2.654 0.008743 **
## carCompanyBuick      -6.413e-02  1.135e-01  -0.565 0.572974
## carCompanyChevrolet  -2.347e-01  1.239e-01  -1.895 0.059891 .
## carCompanyDodge      -2.460e-01  1.000e-01  -2.460 0.014928 *
## carCompanyHonda      -1.464e-01  9.739e-02  -1.503 0.134661
## carCompanyIsuzu      -1.066e-01  1.063e-01  -1.003 0.317411
## carCompanyJaguar     -2.111e-01  1.272e-01  -1.660 0.098856 .
## carCompanyMazda      -1.309e-01  9.481e-02  -1.381 0.169110
## carCompanyMercury    -2.849e-01  1.579e-01  -1.804 0.073057 .
## carCompanyMitsubishi -2.804e-01  9.711e-02  -2.887 0.004409 **
## carCompanyNissan     -1.777e-01  8.982e-02  -1.978 0.049577 *
## carCompanyPeugeot    -5.468e-01  1.890e-01  -2.894 0.004319 **
## carCompanyPlymouth   -2.672e-01  9.951e-02  -2.685 0.008000 **
## carCompanyPorsche     2.599e-01  1.579e-01   1.646 0.101710
## carCompanyRenault    -1.849e-01  1.239e-01  -1.492 0.137522
## carCompanySaab       -4.340e-03  1.032e-01  -0.042 0.966504
## carCompanySubaru     -2.636e-01  9.798e-02  -2.691 0.007867 **
## carCompanyToyota     -2.315e-01  8.742e-02  -2.648 0.008890 **
## carCompanyVolkswagen -1.420e-01  9.462e-02  -1.501 0.135249
## carCompanyVolvo      -1.596e-01  1.039e-01  -1.536 0.126408
## aspirationturbo       6.840e-02  3.106e-02   2.202 0.029031 *
## carbodyhardtop       -2.094e-01  7.382e-02  -2.837 0.005118 **
## carbodyhatchback     -2.635e-01  6.541e-02  -4.028 8.58e-05 ***
## carbodysedan         -2.078e-01  6.766e-02  -3.071 0.002498 **
## carbodywagon         -2.822e-01  7.188e-02  -3.926 0.000126 ***
## drivewheelfwd        -9.978e-03  5.369e-02  -0.186 0.852803
## drivewheelrwd         4.187e-02  6.166e-02   0.679 0.498059
## enginelocationrear    3.262e-01  1.587e-01   2.056 0.041346 *
## wheelbase             1.123e-02  4.697e-03   2.391 0.017947 *
## carlength            -3.261e-03  2.929e-03  -1.113 0.267282
## enginetypedohcv      -4.523e-01  1.868e-01  -2.421 0.016564 *
## enginetypel           1.557e-01  1.558e-01   0.999 0.319267
## enginetypeohc        -2.104e-02  5.102e-02  -0.412 0.680524
## enginetypeohcf               NA         NA      NA       NA
## enginetypeohcv       -1.543e-02  6.360e-02  -0.243 0.808660
## enginetyperotor       1.681e-01  8.361e-02   2.011 0.045962 *
## carwidth              3.616e-02  1.337e-02   2.704 0.007566 **
## horsepower            2.440e-03  6.607e-04   3.693 0.000301 ***
## curbweight            4.527e-04  7.609e-05   5.949 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.117 on 165 degrees of freedom
```

```
## Multiple R-squared:  0.9564, Adjusted R-squared:  0.946
## F-statistic:  92.7 on 39 and 165 DF,  p-value: < 2.2e-16
```

The summary of the model shows a significant result with an overall p-value less than the 5% significance level. The adjusted R- squared value of 0.946 indicates that 94.6% variation in the price is explained by the predictors carCompany, aspiration, carbody, drivewheel, enginelocation, wheelbase, carlength, enginetype, carwidth, horsepower and curbweight. Comparing from the model with all 17 predictors, this model has an equivalent adjusted r- squared value. Therefore, it can be stated that model performs equally well with just 11 variables.

Residual Analysis

```
par(mfrow=c(2,2))
plot(model_3)
```




```
par(mfrow=c(1,1))
```

Residuals vs Fitted values

The residuals appear to be linear from the residual vs fitted values plot because the red line is close to the dashed line and the residulas are spread across the the fitted values. This indicates no signs of non-constant variance, moreover, the linearity assumption stands true for the residuals.

Scale-Location Plot

The residuals appear to be randomly distributed on the scale - location plot, and the red line appears to be about horizontal, suggesting the existence of homoscedasticity. These results can be tested using the ncv test.

```
ncvTest(model_3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.354339, Df = 1, p = 0.55167
```

The p-value for non-constant variance test was 0.55167. The null hypothesis of constant variance for residuals cannot be rejected since it is greater than the 5% significance level. To summarize, this model does not violate the assumption of constant variance.

Normal Q-Q Plot

Because the residuals lie nicely on the diagonal line in the normal Q-Q plot, the residuals appear to be normally distributed. The Shapiro-Wilk test will be employed to help the results.

```
shapiro.test(model_3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_3$residuals
## W = 0.99365, p-value = 0.5296
```

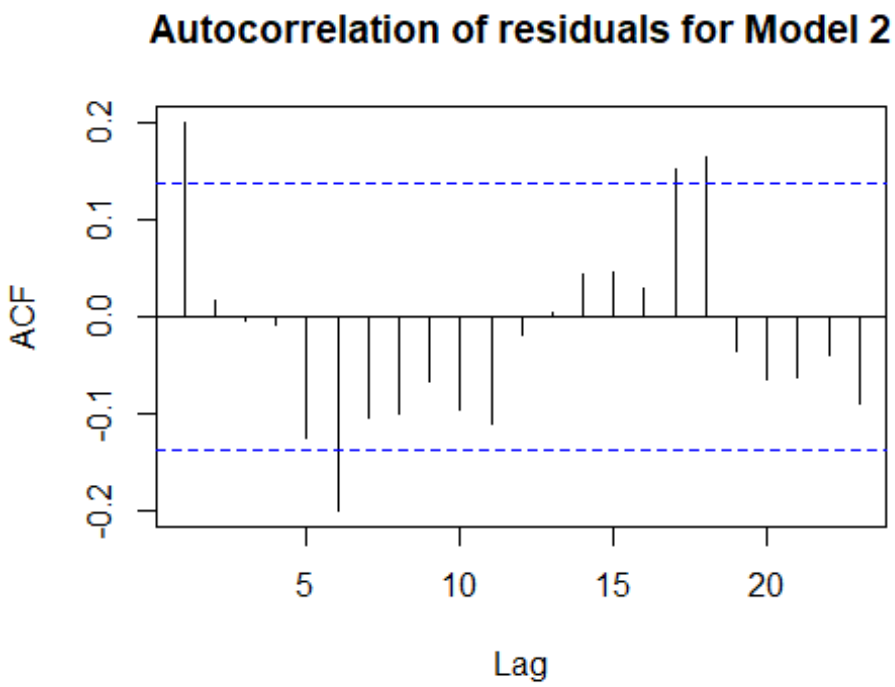
Here, the p-value is higher than the 0.05, therefore the residuals can be considered to be normally distributed. Hence the assumption of normal distribution of residuals is not violated.

Residuals vs Leverage Plot

From the residual vs leverage plot, one point close to the Cook's distance can be observed that may have substantial influence on the regression model.

Autocorrelation of errors

```
acf(model_3$residuals, main ="Autocorrelation of residuals for Model 2")
```




```
durbinWatsonTest(model_3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1        0.199958       1.58514       0
##  Alternative hypothesis: rho != 0
```

According to the ACF plot, there are substantial autocorrelations between residuals. In addition to that, the p-value obtained from the durbin watson test is less than 0.05, suggesting that there is sufficient evidence to reject the null hypothesis. As a result, the independence assumption for residuals might be violated.

To sum up the residual analysis, the model performs well since, aside from the independence of residuals assumptions, it meets all of the other criteria. As a result, the model that was developed can be considered significant.

Partial F-test

In order to compare the above two models, a partial F-test will be performed to check whether the variables fueltype, cylinder number, engine size, boreratio, citympg and highway mpg that were considered unimportant in evaluating the response from the anova test, are significant after taking the other 11 variables into consideration.

The hypothesis test for partial F-test is as follows:

```
H0: beta(fueltype) = beta(cylindernumber) = beta(enginesize) =
beta(boreratio) =     beta(citympg) = beta(highwaympg) = 0
Ha: Alteast one beta != 0

anova(model_3, model_2)

## Analysis of Variance Table
##
## Model 1: log(price) ~ carCompany + aspiration + carbody + drivewheel +
##     enginelocation + wheelbase + carlength + enginetype + carwidth +
##     horsepower + curbweight
## Model 2: log(price) ~ carCompany + fueltype + aspiration + carbody +
drivewheel +
##     enginelocation + wheelbase + carlength + carwidth + curbweight +
##     enginetype + cylindernumber + enginesize + boreratio + horsepower +
##     citympg + highwaympg
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    165 2.2601
## 2    156 2.1143  9   0.14575 1.1949 0.3019
```

The output shows the results of the partial F-test. Since F=1.1949 (p-value = 0.3019) we cannot reject the null hypothesis at the 5% level of significance. It appears that the variables fueltype, cylinder number, engine size, boreratio, citympg and highway mpg do not contribute significant information to the car price once the other eleven variables have been taken into consideration. As a result, model 3 might be considered noteworthy.

The following section consists of model building using:

- Forward selection technique
- Backward elimination technique
- Stepwise regression technique

All these methods build a model for the provided data by comparing the AIC values. AIC values are used to compare the quality of one model compared to others. It is a measure for model selection. The goal is to find a model with lowest AIC value.

Initially, a full and null model will be created. The full model contains all predictors in it, whereas, a null model is a model with no regression parameters.

```
#Full model
full=lm(log(price)~., data=car)

#Null model
null=lm(log(price)~1, data=car)
```

Model 4 - Backward Elimination

Backward elimination begins with a complete model that includes all possible predictor variables, and then eliminates the variable that affects the model's AIC value in each phase.

```
model_4 <- step(full, data=car, direction="backward")
```

```
## Start:  AIC=-839.73
## log(price) ~ carCompany + fueltype + aspiration + carbody + drivewheel +
##     enginelocation + wheelbase + carlength + carwidth + curbweight +
##     enginetype + cylindernumber + enginesize + boreratio + horsepower +
##     citympg + highwaympg
##
##
## Step:  AIC=-839.73
## log(price) ~ carCompany + fueltype + aspiration + carbody + drivewheel +
##     wheelbase + carlength + carwidth + curbweight + enginetype +
##     cylindernumber + enginesize + boreratio + horsepower + citympg +
##     highwaympg
##
##                   Df Sum of Sq    RSS     AIC
## - drivewheel       2   0.02268 2.1370 -841.54
## - fueltype         1   0.01358 2.1279 -840.42
## - enginesize       1   0.01488 2.1292 -840.29
## - cylindernumber   4   0.08054 2.1948 -840.06
## <none>                         2.1143 -839.73
## - aspiration       1   0.03246 2.1468 -838.60
## - carlength        1   0.03392 2.1482 -838.47
## - boreratio        1   0.03417 2.1485 -838.44
## - horsepower       1   0.03987 2.1542 -837.90
## - highwaympg       1   0.04505 2.1594 -837.41
## - wheelbase        1   0.05869 2.1730 -836.11
## - citympg          1   0.06958 2.1839 -835.09
## - carwidth         1   0.13265 2.2470 -829.25
## - enginetype       4   0.24310 2.3574 -825.42
## - carbody          4   0.26575 2.3801 -823.46
## - curbweight       1   0.25889 2.3732 -818.05
## - carCompany      20   1.65355 3.7679 -761.28
##
## Step:  AIC=-841.54
## log(price) ~ carCompany + fueltype + aspiration + carbody + wheelbase +
##     carlength + carwidth + curbweight + enginetype + cylindernumber +
##     enginesize + boreratio + horsepower + citympg + highwaympg
##
##                   Df Sum of Sq    RSS     AIC
## - enginesize       1   0.01140 2.1484 -842.45
## - cylindernumber   4   0.08080 2.2178 -841.93
## <none>                         2.1370 -841.54
## - fueltype         1   0.02172 2.1587 -841.47
## - aspiration       1   0.02178 2.1588 -841.46
## - boreratio        1   0.02912 2.1661 -840.77
## - carlength        1   0.03702 2.1740 -840.02
## - highwaympg       1   0.04544 2.1824 -839.23
## - wheelbase        1   0.05398 2.1910 -838.43
## - horsepower       1   0.05521 2.1922 -838.31
```

```
## - citympg         1   0.07315 2.2101 -836.64
## - carwidth        1   0.13032 2.2673 -831.41
## - enginetype      4   0.25180 2.3888 -826.71
## - carbody         4   0.26212 2.3991 -825.82
## - curbweight      1   0.35346 2.4905 -812.16
## - carCompany     20   1.78810 3.9251 -756.90
##
## Step:  AIC=-842.45
## log(price) ~ carCompany + fueltype + aspiration + carbody + wheelbase +
##     carlength + carwidth + curbweight + enginetype + cylindernumber +
##     boreratio + horsepower + citympg + highwaympg
##
##                   Df Sum of Sq    RSS     AIC
## - aspiration       1   0.01430 2.1627 -843.09
## - cylindernumber   4   0.08196 2.2304 -842.77
## - boreratio        1   0.01793 2.1663 -842.75
## <none>                         2.1484 -842.45
## - fueltype         1   0.03385 2.1822 -841.24
## - carlength        1   0.03413 2.1825 -841.22
## - highwaympg       1   0.04173 2.1901 -840.51
## - wheelbase        1   0.06166 2.2101 -838.65
## - citympg          1   0.06618 2.2146 -838.23
## - horsepower       1   0.10902 2.2574 -834.30
## - carwidth         1   0.12666 2.2751 -832.71
## - carbody          4   0.25568 2.4041 -827.40
## - enginetype       4   0.31653 2.4649 -822.27
## - curbweight       1   0.38096 2.5294 -810.98
## - carCompany      20   1.88757 4.0360 -753.19
##
## Step:  AIC=-843.09
## log(price) ~ carCompany + fueltype + carbody + wheelbase + carlength +
##     carwidth + curbweight + enginetype + cylindernumber + boreratio +
##     horsepower + citympg + highwaympg
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         2.1627 -843.09
## - boreratio        1   0.02946 2.1922 -842.32
## - highwaympg       1   0.03399 2.1967 -841.89
## - cylindernumber   4   0.09998 2.2627 -841.82
## - carlength        1   0.04327 2.2060 -841.03
## - wheelbase        1   0.05864 2.2213 -839.60
## - citympg          1   0.06088 2.2236 -839.40
## - fueltype         1   0.08497 2.2477 -837.19
## - carwidth         1   0.13251 2.2952 -832.90
## - carbody          4   0.25029 2.4130 -828.64
## - horsepower       1   0.21656 2.3793 -825.53
## - enginetype       4   0.33934 2.5020 -821.21
## - curbweight       1   0.38088 2.5436 -811.84
## - carCompany      20   1.92162 4.0843 -752.75
```

Backward elimination process performs model selection based on the AIC value. The model with lowest AIC value is chosen. From the output, it can be noted that the AIC value of the full model is -839.73. In the first step, eliminating drivewheel variable from the model resulted in a lower AIC value of -841.54. Similarly, comparisons based on AIC values were done in the following phases in order to keep only the variables in the model that showed any potential for improvement. With the following variables, the procedure completed with a final AIC value of -843.09: carCompany, fueltype, carbody, wheelbase, carlength, carwidth, curbweight, enginetype, cylindernumber, boreratio, horsepower, citympg, and highwaympg.

```
summary(model_4)
```

```
##
## Call:
## lm(formula = log(price) ~ carCompany + fueltype + carbody + wheelbase +
##     carlength + carwidth + curbweight + enginetype + cylindernumber +
##     boreratio + horsepower + citympg + highwaympg, data = car)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250465 -0.057611 -0.000552  0.070538  0.292043
##
## Coefficients: (2 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.925e+00  7.260e-01   8.161 9.28e-14 ***
## carCompanyAudi        2.331e-02  1.282e-01   0.182  0.85595
## carCompanyBmw         3.545e-01  1.244e-01   2.850  0.00495 **
## carCompanyBuick      -6.399e-02  1.547e-01  -0.414  0.67970
## carCompanyChevrolet  -1.652e-01  1.324e-01  -1.248  0.21388
## carCompanyDodge      -2.445e-01  1.033e-01  -2.367  0.01915 *
## carCompanyHonda      -1.438e-01  1.013e-01  -1.420  0.15745
## carCompanyIsuzu      -3.929e-02  1.150e-01  -0.342  0.73310
## carCompanyJaguar     -1.797e-01  1.560e-01  -1.151  0.25129
## carCompanyMazda      -1.205e-01  1.003e-01  -1.201  0.23144
## carCompanyMercury    -2.453e-01  1.638e-01  -1.498  0.13615
## carCompanyMitsubishi -2.902e-01  9.981e-02  -2.907  0.00416 **
## carCompanyNissan     -1.351e-01  9.503e-02  -1.422  0.15692
## carCompanyPeugeot    -4.682e-01  2.483e-01  -1.886  0.06115 .
## carCompanyPlymouth   -2.513e-01  1.036e-01  -2.427  0.01634 *
## carCompanyPorsche     2.187e-01  1.686e-01   1.297  0.19637
## carCompanyRenault    -2.136e-01  1.285e-01  -1.661  0.09860 .
## carCompanySaab       -5.571e-02  9.996e-02  -0.557  0.57806
## carCompanySubaru     -5.616e-01  2.145e-01  -2.619  0.00968 **
## carCompanyToyota     -2.010e-01  9.405e-02  -2.137  0.03409 *
## carCompanyVolkswagen -1.639e-01  9.609e-02  -1.706  0.09003 .
## carCompanyVolvo      -7.168e-02  1.123e-01  -0.638  0.52412
## fueltypegas          -1.326e-01  5.290e-02  -2.507  0.01317 *
## carbodyhardtop       -1.770e-01  7.530e-02  -2.350  0.01997 *
## carbodyhatchback     -2.319e-01  6.965e-02  -3.330  0.00108 **
## carbodysedan         -1.804e-01  7.250e-02  -2.488  0.01388 *
```

```
## carbodywagon         -2.473e-01  7.697e-02  -3.213  0.00159 **
## wheelbase             1.008e-02  4.840e-03   2.083  0.03885 *
## carlength            -5.417e-03  3.027e-03  -1.789  0.07546 .
## carwidth              4.321e-02  1.380e-02   3.131  0.00207 **
## curbweight            4.901e-04  9.232e-05   5.308 3.64e-07 ***
## enginetypedohcv      -7.732e-01  2.399e-01  -3.222  0.00154 **
## enginetypel           9.679e-02  2.198e-01   0.440  0.66022
## enginetypeohc        -5.202e-02  5.980e-02  -0.870  0.38568
## enginetypeohcf        3.369e-01  1.675e-01   2.011  0.04601 *
## enginetypeohcv       -9.528e-02  7.694e-02  -1.238  0.21735
## enginetyperotor       9.947e-03  1.767e-01   0.056  0.95519
## cylindernumberfive   -1.595e-01  1.262e-01  -1.264  0.20822
## cylindernumberfour   -7.528e-02  1.582e-01  -0.476  0.63488
## cylindernumbersix    -1.855e-01  1.413e-01  -1.313  0.19105
## cylindernumberthree          NA         NA      NA       NA
## cylindernumbertwelve -2.614e-01  2.232e-01  -1.171  0.24329
## cylindernumbertwo            NA         NA      NA       NA
## boreratio            -1.157e-01  7.836e-02  -1.476  0.14184
## horsepower            3.383e-03  8.453e-04   4.003 9.56e-05 ***
## citympg              -1.771e-02  8.345e-03  -2.122  0.03535 *
## highwaympg            1.131e-02  7.135e-03   1.586  0.11478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1163 on 160 degrees of freedom
## Multiple R-squared:  0.9582, Adjusted R-squared:  0.9467
## F-statistic: 83.43 on 44 and 160 DF,  p-value: < 2.2e-16
```

The summary of the model obtained from backward elimination techniques shows significant results. The overall p-value indicates that the model is significant. Moreover, 94.67% variation in the price can be explained by the thirteen predictors selected from backward elimination.

Model 5 - Forward Selection

Forward selection method begins with an empty model that gradually adds variables one by one. In each step,the variable which improves the model and reduces the AIC value is added. The process stops when the model no longer improves with adding variables

```
model_5 <- step(null, scope=list(lower=null, upper=full),
direction="forward")
```

```
## Start:  AIC=-280.08
## log(price) ~ 1
##
##                 Df Sum of Sq    RSS     AIC
## + curbweight     1    41.128 10.651 -602.26
## + enginesize     1    35.841 15.938 -519.64
## + horsepower     1    35.314 16.466 -512.96
## + carCompany    21    37.857 13.922 -507.35
```

```
## + carwidth        1    33.350 18.429 -489.86
## + highwaympg      1    31.116 20.664 -466.40
## + citympg         1    30.829 20.950 -463.57
## + carlength       1    30.530 21.249 -460.67
## + cylindernumber  6    28.064 23.716 -428.16
## + drivewheel      2    24.512 27.267 -407.55
## + wheelbase       1    20.512 31.267 -381.49
## + boreratio       1    19.303 32.476 -373.71
## + enginetype      6    11.080 40.700 -317.44
## + carbody         4     6.632 45.148 -300.18
## + enginelocation  1     3.640 48.139 -293.03
## + aspiration      1     3.432 48.348 -292.14
## + fueltype        1     0.911 50.868 -281.72
## <none>                        51.779 -280.08
##
## Step:  AIC=-602.26
## log(price) ~ curbweight
##
##                  Df Sum of Sq     RSS     AIC
## + carCompany     21    6.7778  3.8730 -767.64
## + horsepower      1    2.9153  7.7355 -665.82
## + enginelocation  1    2.5160  8.1348 -655.51
## + carbody         4    2.1405  8.5103 -640.25
## + cylindernumber  6    2.0671  8.5837 -634.49
## + citympg         1    1.1329  9.5179 -623.32
## + enginesize      1    1.0229  9.6280 -620.96
## + enginetype      6    1.3019  9.3489 -616.99
## + drivewheel      2    0.7614  9.8895 -613.47
## + highwaympg      1    0.5911 10.0597 -611.97
## + wheelbase       1    0.5099 10.1409 -610.32
## + fueltype        1    0.2023 10.4486 -604.19
## + carwidth        1    0.1854 10.4654 -603.86
## <none>                        10.6508 -602.26
## + boreratio       1    0.0951 10.5558 -602.10
## + aspiration      1    0.0597 10.5911 -601.41
## + carlength       1    0.0468 10.6041 -601.16
##
## Step:  AIC=-767.64
## log(price) ~ curbweight + carCompany
##
##                  Df Sum of Sq    RSS     AIC
## + horsepower      1   0.54162 3.3314 -796.52
## + carbody         4   0.56418 3.3088 -791.91
## + enginetype      6   0.56181 3.3112 -787.77
## + enginelocation  1   0.30086 3.5722 -782.22
## + citympg         1   0.24644 3.6266 -779.12
## + drivewheel      2   0.25251 3.6205 -777.46
## + highwaympg      1   0.20178 3.6713 -776.61
## + enginesize      1   0.12000 3.7530 -772.09
## + cylindernumber  6   0.29543 3.5776 -771.91
```

```
## + aspiration      1   0.10862 3.7644 -771.47
## + carwidth        1   0.04224 3.8308 -767.89
## <none>                        3.8730 -767.64
## + boreratio       1   0.02576 3.8473 -767.01
## + fueltype        1   0.02321 3.8498 -766.87
## + carlength       1   0.00279 3.8702 -765.79
## + wheelbase       1   0.00251 3.8705 -765.77
##
## Step:  AIC=-796.52
## log(price) ~ curbweight + carCompany + horsepower
##
##                   Df Sum of Sq    RSS     AIC
## + carbody          4   0.40246 2.9290 -814.91
## + enginetype       6   0.44263 2.8888 -813.75
## + enginelocation   1   0.25767 3.0737 -811.02
## + aspiration       1   0.05331 3.2781 -797.83
## + drivewheel       2   0.08316 3.2483 -797.70
## + cylindernumber   6   0.20404 3.1274 -797.48
## + fueltype         1   0.04456 3.2868 -797.28
## <none>                         3.3314 -796.52
## + citympg          1   0.02106 3.3104 -795.82
## + wheelbase        1   0.01582 3.3156 -795.50
## + highwaympg       1   0.01446 3.3169 -795.41
## + carwidth         1   0.00905 3.3224 -795.08
## + carlength        1   0.00626 3.3252 -794.91
## + boreratio        1   0.00585 3.3256 -794.88
## + enginesize       1   0.00205 3.3294 -794.65
##
## Step:  AIC=-814.91
## log(price) ~ curbweight + carCompany + horsepower + carbody
##
##                   Df Sum of Sq    RSS     AIC
## + enginetype       6   0.36218 2.5668 -829.97
## + enginelocation   1   0.15566 2.7733 -824.11
## + cylindernumber   6   0.21101 2.7179 -818.24
## + aspiration       1   0.06405 2.8649 -817.45
## + wheelbase        1   0.06062 2.8683 -817.20
## + fueltype         1   0.03452 2.8944 -815.35
## <none>                         2.9290 -814.91
## + drivewheel       2   0.04916 2.8798 -814.38
## + carwidth         1   0.01957 2.9094 -814.29
## + citympg          1   0.01833 2.9106 -814.20
## + highwaympg       1   0.01446 2.9145 -813.93
## + carlength        1   0.01124 2.9177 -813.70
## + enginesize       1   0.01025 2.9187 -813.63
## + boreratio        1   0.00610 2.9229 -813.34
##
## Step:  AIC=-829.97
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype
```

```
##
##                   Df Sum of Sq    RSS     AIC
## + carwidth         1  0.154282 2.4125 -840.68
## + wheelbase        1  0.107359 2.4594 -836.73
## + aspiration       1  0.040765 2.5260 -831.26
## + fueltype         1  0.035319 2.5314 -830.81
## <none>                         2.5668 -829.97
## + highwaympg       1  0.008812 2.5580 -828.68
## + carlength        1  0.008234 2.5585 -828.63
## + enginesize       1  0.004582 2.5622 -828.34
## + boreratio        1  0.004230 2.5625 -828.31
## + citympg          1  0.003221 2.5636 -828.23
## + drivewheel       2  0.015610 2.5512 -827.22
## + cylindernumber   4  0.027525 2.5393 -824.18
##
## Step:  AIC=-840.68
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth
##
##                   Df Sum of Sq    RSS     AIC
## + aspiration       1  0.058865 2.3536 -843.75
## + fueltype         1  0.033493 2.3790 -841.55
## + wheelbase        1  0.032308 2.3802 -841.45
## <none>                         2.4125 -840.68
## + carlength        1  0.007465 2.4050 -839.32
## + boreratio        1  0.004976 2.4075 -839.11
## + highwaympg       1  0.001740 2.4108 -838.83
## + citympg          1  0.000006 2.4125 -838.68
## + enginesize       1  0.000000 2.4125 -838.68
## + drivewheel       2  0.007357 2.4051 -837.31
## + cylindernumber   4  0.033319 2.3792 -835.53
##
## Step:  AIC=-843.75
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth + aspiration
##
##                   Df Sum of Sq    RSS     AIC
## + wheelbase        1  0.053217 2.3004 -846.43
## <none>                         2.3536 -843.75
## + enginesize       1  0.009217 2.3444 -842.55
## + citympg          1  0.005367 2.3483 -842.21
## + fueltype         1  0.003147 2.3505 -842.02
## + boreratio        1  0.001411 2.3522 -841.87
## + carlength        1  0.000079 2.3535 -841.75
## + highwaympg       1  0.000001 2.3536 -841.75
## + drivewheel       2  0.015291 2.3383 -841.08
## + cylindernumber   4  0.045070 2.3085 -839.71
##
## Step:  AIC=-846.43
## log(price) ~ curbweight + carCompany + horsepower + carbody +
```

```
##     enginetype + carwidth + aspiration + wheelbase
##
##                   Df Sum of Sq    RSS     AIC
## <none>                          2.3004 -846.43
## + carlength        1 0.0189016 2.2815 -846.13
## + fueltype         1 0.0049680 2.2954 -844.88
## + boreratio        1 0.0037871 2.2966 -844.77
## + enginesize       1 0.0028093 2.2976 -844.68
## + highwaympg       1 0.0021907 2.2982 -844.63
## + drivewheel       2 0.0233843 2.2770 -844.53
## + citympg          1 0.0010250 2.2994 -844.53
## + cylindernumber   4 0.0291542 2.2713 -841.05
```

In this situation, the null model's AIC score was -280.08. In the first stage, it was discovered that include the curbweight variable reduced the AIC value to -602.26, hence it was included in the model. The addition of the carcompany variable improved the model's performance in correlation, resulting in a lower AIC value of -767.64 Other variables were also incorporated, which aided in the improvement of the model. Eight predictors out of 17 variables were added to the model by the end of the last stage, resulting in the lowest AIC value of -846.43.

```
summary(model_5)
```

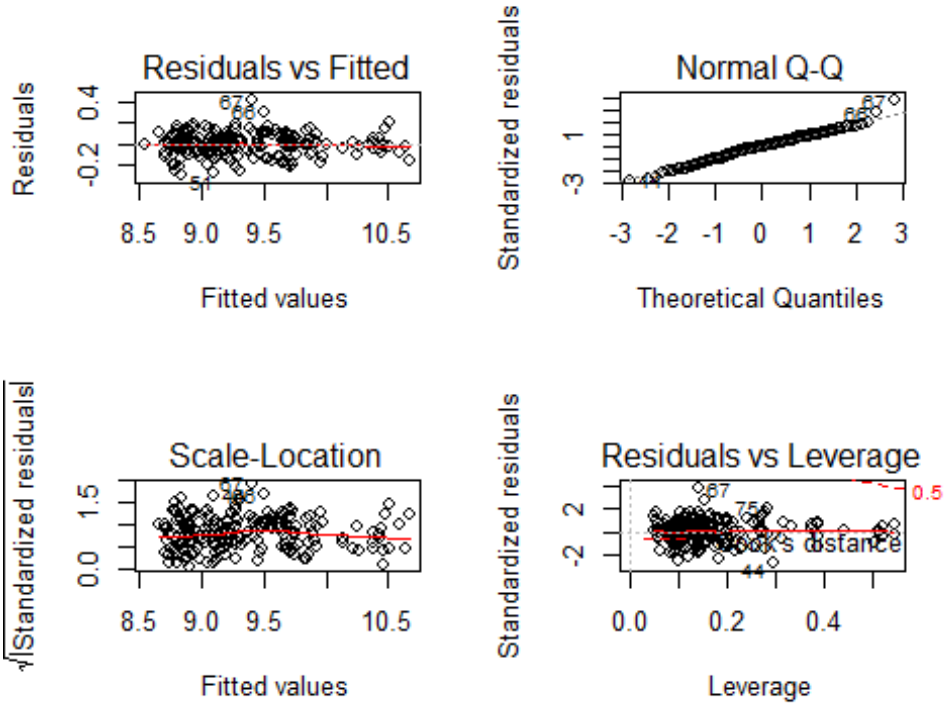
```
##
## Call:
## lm(formula = log(price) ~ curbweight + carCompany + horsepower +
##     carbody + enginetype + carwidth + aspiration + wheelbase,
##     data = car)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28448 -0.06323  0.00048  0.06321  0.41261
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.553e+00  6.580e-01   8.440 1.41e-14 ***
## curbweight            4.438e-04  6.008e-05   7.387 6.70e-12 ***
## carCompanyAudi        1.325e-02  1.058e-01   0.125 0.900414
## carCompanyBmw         3.155e-01  1.047e-01   3.015 0.002971 **
## carCompanyBuick      -2.220e-02  1.092e-01  -0.203 0.839073
## carCompanyChevrolet  -2.228e-01  1.203e-01  -1.852 0.065776 .
## carCompanyDodge      -2.402e-01  9.511e-02  -2.526 0.012463 *
## carCompanyHonda      -1.402e-01  9.226e-02  -1.520 0.130396
## carCompanyIsuzu      -8.291e-02  1.042e-01  -0.796 0.427206
## carCompanyJaguar     -1.866e-01  1.182e-01  -1.578 0.116401
## carCompanyMazda      -1.326e-01  9.336e-02  -1.420 0.157391
## carCompanyMercury    -2.435e-01  1.555e-01  -1.566 0.119329
## carCompanyMitsubishi -2.921e-01  9.317e-02  -3.135 0.002031 **
## carCompanyNissan     -1.863e-01  8.795e-02  -2.118 0.035611 *
## carCompanyPeugeot    -5.032e-01  1.871e-01  -2.690 0.007863 **
```

```
## carCompanyPlymouth   -2.642e-01  9.707e-02  -2.721 0.007185 **
## carCompanyPorsche     3.107e-01  1.535e-01   2.024 0.044534 *
## carCompanyRenault    -2.137e-01  1.220e-01  -1.751 0.081723 .
## carCompanySaab       -5.737e-02  9.737e-02  -0.589 0.556513
## carCompanySubaru     -5.491e-01  1.836e-01  -2.990 0.003207 **
## carCompanyToyota     -2.233e-01  8.571e-02  -2.605 0.010016 *
## carCompanyVolkswagen -1.502e-01  9.207e-02  -1.631 0.104693
## carCompanyVolvo      -1.210e-01  1.014e-01  -1.194 0.234136
## horsepower            2.476e-03  6.431e-04   3.850 0.000168 ***
## carbodyhardtop       -2.061e-01  7.235e-02  -2.848 0.004949 **
## carbodyhatchback     -2.650e-01  6.499e-02  -4.078 7.00e-05 ***
## carbodysedan         -2.231e-01  6.594e-02  -3.383 0.000892 ***
## carbodywagon         -3.054e-01  6.946e-02  -4.397 1.94e-05 ***
## enginetypedohcv      -4.549e-01  1.811e-01  -2.513 0.012930 *
## enginetypel           1.461e-01  1.554e-01   0.940 0.348544
## enginetypeohc        -3.616e-02  4.973e-02  -0.727 0.468170
## enginetypeohcf        2.706e-01  1.550e-01   1.746 0.082659 .
## enginetypeohcv       -2.012e-02  5.991e-02  -0.336 0.737422
## enginetyperotor       1.911e-01  8.208e-02   2.328 0.021106 *
## carwidth              3.057e-02  1.257e-02   2.433 0.016019 *
## aspirationturbo       7.113e-02  2.947e-02   2.414 0.016866 *
## wheelbase             8.060e-03  4.089e-03   1.971 0.050320 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.117 on 168 degrees of freedom
## Multiple R-squared:  0.9556, Adjusted R-squared:  0.9461
## F-statistic: 100.4 on 36 and 168 DF,  p-value: < 2.2e-16
```

The model constructed using the forward selection strategy has a significant overall p-value. Furthermore, the selected eight variables are able to explain 94.61 percent of the variation in determining the price of a car. This model's adjusted R-squared value is quite close to the above models with additional predictors. As a result, the variables curbweight, carCompany, horsepower, carbody, enginetype, carwidth, aspiration, and wheelbase are adequate in determining a car's pricing.

Residual Analysis

```
par(mfrow=c(2,2))
plot(model_5)
```




```
par(mfrow=c(1,1))
```

Residuals vs Fitted values

Because the red line is close to the dashed line and the residuals are dispersed throughout the fitted values, the residuals seem to be linear in the residual versus fitted values figure. This shows that there are no evidence of non-constant variance, and the linearity assumption holds for the residuals.

Scale-Location Plot

On the scale - location plot, the residuals appear to be randomly distributed, and the red line looks to be horizontal, indicating the presence of homoscedasticity. The ncv test can be used to verify these results.

```
ncvTest(model_5)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.2219168, Df = 1, p = 0.63758
```

The ncv test yielded a p-value of 0.569 > 0.05, showing that the assumption of constant variance was not violated.

Normal Q-Q Plot

The residuals appear to be normally distributed because they lie well on the diagonal line in the normal Q-Q plot. To aid the results, the Shapiro-Wilk test will be used.

```
shapiro.test(model_5$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model_5$residuals
## W = 0.98711, p-value = 0.0599
```

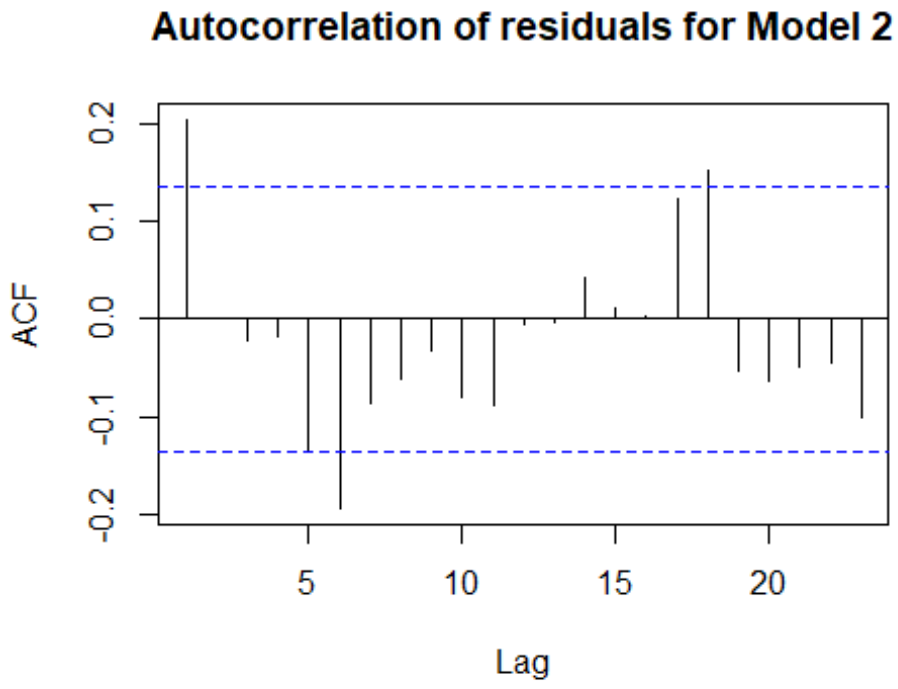
Here, the p-value is slightly higher than the 0.05, therefore the residuals can be considered to be normally distributed. Hence the assumption of normal distribution of residuals is not violated.

Residuals vs Leverage Plot

From the residual vs leverage plot, one point close to the Cook's distance can be observed that may have substantial influence on the regression model.

Autocorrelation of errors

```
acf(model_5$residuals, main ="Autocorrelation of residuals for Model 2")
```



```
durbinWatsonTest(model_5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2046924      1.573541       0
##  Alternative hypothesis: rho != 0
```

There are significant autocorrelations between residuals, according to the ACF plot. Furthermore, the durbin watson test yielded a p-value of less than 0.05, indicating that there is sufficient evidence to reject the null hypothesis. As a result, the residuals' independence assumption may be broken.

The residual analysis of the model constructed using the forward selection method produced significant findings. The residuals have a normal distribution and are linear. They also have a constant variance. However, the residuals for this model were shown to violate the independency assumption as well. However, when compared to other models, with only eight predictors, this model performs equally better.

Model 6 - Stepwise Regression

Stepwise regression is a technique for fitting regression models in which the selection of predictor variables is automated. Each step considers whether a variable should be added to or subtracted from the collection of independent variables based on the AIC values.

```
model_6 <- step(null, scope = list(upper=full), data=car, direction="both")
```

```
## Start:  AIC=-280.08
## log(price) ~ 1
##
##                   Df Sum of Sq    RSS     AIC
## + curbweight       1    41.128 10.651 -602.26
## + enginesize       1    35.841 15.938 -519.64
## + horsepower       1    35.314 16.466 -512.96
## + carCompany      21    37.857 13.922 -507.35
## + carwidth         1    33.350 18.429 -489.86
## + highwaympg       1    31.116 20.664 -466.40
## + citympg          1    30.829 20.950 -463.57
## + carlength        1    30.530 21.249 -460.67
## + cylindernumber   6    28.064 23.716 -428.16
## + drivewheel       2    24.512 27.267 -407.55
## + wheelbase        1    20.512 31.267 -381.49
## + boreratio        1    19.303 32.476 -373.71
## + enginetype       6    11.080 40.700 -317.44
## + carbody          4     6.632 45.148 -300.18
## + enginelocation   1     3.640 48.139 -293.03
## + aspiration       1     3.432 48.348 -292.14
## + fueltype         1     0.911 50.868 -281.72
## <none>                         51.779 -280.08
##
```

```
## Step:  AIC=-602.26
## log(price) ~ curbweight
##
##                   Df Sum of Sq    RSS     AIC
## + carCompany     21     6.778  3.873 -767.64
## + horsepower      1     2.915  7.735 -665.82
## + enginelocation  1     2.516  8.135 -655.51
## + carbody         4     2.141  8.510 -640.25
## + cylindernumber  6     2.067  8.584 -634.49
## + citympg         1     1.133  9.518 -623.32
## + enginesize      1     1.023  9.628 -620.96
## + enginetype      6     1.302  9.349 -616.99
## + drivewheel      2     0.761  9.889 -613.47
## + highwaympg      1     0.591 10.060 -611.97
## + wheelbase       1     0.510 10.141 -610.32
## + fueltype        1     0.202 10.449 -604.19
## + carwidth        1     0.185 10.465 -603.86
## <none>                        10.651 -602.26
## + boreratio       1     0.095 10.556 -602.10
## + aspiration      1     0.060 10.591 -601.41
## + carlength       1     0.047 10.604 -601.16
## - curbweight      1    41.128 51.779 -280.08
##
## Step:  AIC=-767.64
## log(price) ~ curbweight + carCompany
##
##                   Df Sum of Sq     RSS     AIC
## + horsepower      1    0.5416  3.3314 -796.52
## + carbody         4    0.5642  3.3088 -791.91
## + enginetype      6    0.5618  3.3112 -787.77
## + enginelocation  1    0.3009  3.5722 -782.22
## + citympg         1    0.2464  3.6266 -779.12
## + drivewheel      2    0.2525  3.6205 -777.46
## + highwaympg      1    0.2018  3.6713 -776.61
## + enginesize      1    0.1200  3.7530 -772.09
## + cylindernumber  6    0.2954  3.5776 -771.91
## + aspiration      1    0.1086  3.7644 -771.47
## + carwidth        1    0.0422  3.8308 -767.89
## <none>                         3.8730 -767.64
## + boreratio       1    0.0258  3.8473 -767.01
## + fueltype        1    0.0232  3.8498 -766.87
## + carlength       1    0.0028  3.8702 -765.79
## + wheelbase       1    0.0025  3.8705 -765.77
## - carCompany     21    6.7778 10.6508 -602.26
## - curbweight      1   10.0493 13.9223 -507.35
##
## Step:  AIC=-796.52
## log(price) ~ curbweight + carCompany + horsepower
##
##                   Df Sum of Sq    RSS     AIC
```

```
## + carbody         4    0.4025 2.9290 -814.91
## + enginetype      6    0.4426 2.8888 -813.75
## + enginelocation  1    0.2577 3.0737 -811.02
## + aspiration      1    0.0533 3.2781 -797.83
## + drivewheel      2    0.0832 3.2483 -797.70
## + cylindernumber  6    0.2040 3.1274 -797.48
## + fueltype        1    0.0446 3.2868 -797.28
## <none>                       3.3314 -796.52
## + citympg         1    0.0211 3.3104 -795.82
## + wheelbase       1    0.0158 3.3156 -795.50
## + highwaympg      1    0.0145 3.3169 -795.41
## + carwidth        1    0.0091 3.3224 -795.08
## + carlength       1    0.0063 3.3252 -794.91
## + boreratio       1    0.0059 3.3256 -794.88
## + enginesize      1    0.0020 3.3294 -794.65
## - horsepower      1    0.5416 3.8730 -767.64
## - curbweight      1    2.1709 5.5024 -695.66
## - carCompany     21    4.4041 7.7355 -665.82
##
## Step:  AIC=-814.91
## log(price) ~ curbweight + carCompany + horsepower + carbody
##
##                  Df Sum of Sq    RSS     AIC
## + enginetype      6    0.3622 2.5668 -829.97
## + enginelocation  1    0.1557 2.7733 -824.11
## + cylindernumber  6    0.2110 2.7179 -818.24
## + aspiration      1    0.0641 2.8649 -817.45
## + wheelbase       1    0.0606 2.8683 -817.20
## + fueltype        1    0.0345 2.8944 -815.35
## <none>                       2.9290 -814.91
## + drivewheel      2    0.0492 2.8798 -814.38
## + carwidth        1    0.0196 2.9094 -814.29
## + citympg         1    0.0183 2.9106 -814.20
## + highwaympg      1    0.0145 2.9145 -813.93
## + carlength       1    0.0112 2.9177 -813.70
## + enginesize      1    0.0103 2.9187 -813.63
## + boreratio       1    0.0061 2.9229 -813.34
## - carbody         4    0.4025 3.3314 -796.52
## - horsepower      1    0.3799 3.3088 -791.91
## - curbweight      1    2.1468 5.0757 -704.20
## - carCompany     21    3.4607 6.3897 -697.01
##
## Step:  AIC=-829.97
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype
##
##                  Df Sum of Sq    RSS     AIC
## + carwidth        1   0.15428 2.4125 -840.68
## + wheelbase       1   0.10736 2.4594 -836.73
## + aspiration      1   0.04077 2.5260 -831.26
```

```
## + fueltype        1   0.03532 2.5315 -830.81
## <none>                        2.5668 -829.97
## + highwaympg      1   0.00881 2.5580 -828.68
## + carlength       1   0.00823 2.5585 -828.63
## + enginesize      1   0.00458 2.5622 -828.34
## + boreratio       1   0.00423 2.5625 -828.31
## + citympg         1   0.00322 2.5636 -828.23
## + drivewheel      2   0.01561 2.5512 -827.22
## + cylindernumber  4   0.02753 2.5392 -824.18
## - enginetype      6   0.36218 2.9290 -814.91
## - carbody         4   0.32201 2.8888 -813.75
## - horsepower      1   0.31138 2.8782 -808.50
## - carCompany     21   3.02100 5.5878 -712.50
## - curbweight      1   2.07189 4.6387 -710.66
##
## Step:  AIC=-840.68
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth
##
##                  Df Sum of Sq    RSS     AIC
## + aspiration      1   0.05887 2.3536 -843.75
## + fueltype        1   0.03349 2.3790 -841.55
## + wheelbase       1   0.03231 2.3802 -841.45
## <none>                        2.4125 -840.68
## + carlength       1   0.00746 2.4050 -839.32
## + boreratio       1   0.00498 2.4075 -839.11
## + highwaympg      1   0.00174 2.4107 -838.83
## + citympg         1   0.00001 2.4125 -838.68
## + enginesize      1   0.00000 2.4125 -838.68
## + drivewheel      2   0.00736 2.4051 -837.31
## + cylindernumber  4   0.03332 2.3792 -835.53
## - carwidth        1   0.15428 2.5668 -829.97
## - carbody         4   0.30071 2.7132 -824.60
## - horsepower      1   0.26564 2.6781 -821.27
## - enginetype      6   0.49690 2.9094 -814.29
## - curbweight      1   1.05575 3.4682 -768.27
## - carCompany     21   2.91133 5.3238 -720.42
##
## Step:  AIC=-843.75
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth + aspiration
##
##                  Df Sum of Sq    RSS     AIC
## + wheelbase       1   0.05322 2.3004 -846.43
## <none>                        2.3536 -843.75
## + enginesize      1   0.00922 2.3444 -842.55
## + citympg         1   0.00537 2.3483 -842.21
## + fueltype        1   0.00315 2.3505 -842.02
## + boreratio       1   0.00141 2.3522 -841.87
## + carlength       1   0.00008 2.3535 -841.75
```

```
## + highwaympg      1   0.00000 2.3536 -841.75
## + drivewheel      2   0.01529 2.3383 -841.08
## - aspiration      1   0.05887 2.4125 -840.68
## + cylindernumber  4   0.04507 2.3086 -839.71
## - carwidth        1   0.17238 2.5260 -831.26
## - horsepower      1   0.17761 2.5312 -830.83
## - carbody         4   0.31477 2.6684 -826.01
## - enginetype      6   0.48212 2.8357 -817.54
## - curbweight      1   0.97486 3.3285 -774.70
## - carCompany     21   2.89207 5.2457 -721.45
##
## Step:  AIC=-846.43
## log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth + aspiration + wheelbase
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         2.3004 -846.43
## + carlength       1   0.01890 2.2815 -846.13
## + fueltype        1   0.00497 2.2954 -844.88
## + boreratio       1   0.00379 2.2966 -844.77
## + enginesize      1   0.00281 2.2976 -844.68
## + highwaympg      1   0.00219 2.2982 -844.63
## + drivewheel      2   0.02338 2.2770 -844.53
## + citympg         1   0.00102 2.2994 -844.53
## - wheelbase       1   0.05322 2.3536 -843.75
## - aspiration      1   0.07977 2.3802 -841.45
## - carwidth        1   0.08106 2.3815 -841.33
## + cylindernumber  4   0.02915 2.2713 -841.05
## - horsepower      1   0.20297 2.5034 -831.10
## - carbody         4   0.35510 2.6555 -825.01
## - enginetype      6   0.47659 2.7770 -819.84
## - curbweight      1   0.74714 3.0475 -790.78
## - carCompany     21   2.91341 5.2138 -720.70
```

```
summary(model_6)
```

```
##
## Call:
## lm(formula = log(price) ~ curbweight + carCompany + horsepower +
##     carbody + enginetype + carwidth + aspiration + wheelbase,
##     data = car)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28448 -0.06323  0.00048  0.06321  0.41261
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.553e+00  6.580e-01   8.440 1.41e-14 ***
## curbweight            4.438e-04  6.008e-05   7.387 6.70e-12 ***
```

```
## carCompanyAudi         1.325e-02  1.058e-01   0.125 0.900414
## carCompanyBmw          3.155e-01  1.047e-01   3.015 0.002971 **
## carCompanyBuick       -2.220e-02  1.092e-01  -0.203 0.839073
## carCompanyChevrolet   -2.228e-01  1.203e-01  -1.852 0.065776 .
## carCompanyDodge       -2.402e-01  9.511e-02  -2.526 0.012463 *
## carCompanyHonda       -1.402e-01  9.226e-02  -1.520 0.130396
## carCompanyIsuzu       -8.291e-02  1.042e-01  -0.796 0.427206
## carCompanyJaguar      -1.866e-01  1.182e-01  -1.578 0.116401
## carCompanyMazda       -1.326e-01  9.336e-02  -1.420 0.157391
## carCompanyMercury     -2.435e-01  1.555e-01  -1.566 0.119329
## carCompanyMitsubishi  -2.921e-01  9.317e-02  -3.135 0.002031 **
## carCompanyNissan      -1.863e-01  8.795e-02  -2.118 0.035611 *
## carCompanyPeugeot     -5.032e-01  1.871e-01  -2.690 0.007863 **
## carCompanyPlymouth    -2.642e-01  9.707e-02  -2.721 0.007185 **
## carCompanyPorsche      3.107e-01  1.535e-01   2.024 0.044534 *
## carCompanyRenault     -2.137e-01  1.220e-01  -1.751 0.081723 .
## carCompanySaab        -5.737e-02  9.737e-02  -0.589 0.556513
## carCompanySubaru      -5.491e-01  1.836e-01  -2.990 0.003207 **
## carCompanyToyota      -2.233e-01  8.571e-02  -2.605 0.010016 *
## carCompanyVolkswagen  -1.502e-01  9.207e-02  -1.631 0.104693
## carCompanyVolvo       -1.210e-01  1.014e-01  -1.194 0.234136
## horsepower             2.476e-03  6.431e-04   3.850 0.000168 ***
## carbodyhardtop        -2.061e-01  7.235e-02  -2.848 0.004949 **
## carbodyhatchback      -2.650e-01  6.499e-02  -4.078 7.00e-05 ***
## carbodysedan          -2.231e-01  6.594e-02  -3.383 0.000892 ***
## carbodywagon          -3.054e-01  6.946e-02  -4.397 1.94e-05 ***
## enginetypedohcv       -4.549e-01  1.811e-01  -2.513 0.012930 *
## enginetypel            1.461e-01  1.554e-01   0.940 0.348544
## enginetypeohc         -3.616e-02  4.973e-02  -0.727 0.468170
## enginetypeohcf         2.706e-01  1.550e-01   1.746 0.082659 .
## enginetypeohcv        -2.012e-02  5.991e-02  -0.336 0.737422
## enginetyperotor        1.911e-01  8.208e-02   2.328 0.021106 *
## carwidth               3.057e-02  1.257e-02   2.433 0.016019 *
## aspirationturbo        7.113e-02  2.947e-02   2.414 0.016866 *
## wheelbase              8.060e-03  4.089e-03   1.971 0.050320 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.117 on 168 degrees of freedom
## Multiple R-squared:  0.9556, Adjusted R-squared:  0.9461
## F-statistic: 100.4 on 36 and 168 DF,  p-value: < 2.2e-16
```

The model created using the stepwise regression approach is identical to the model created using the forward selection. With an adjusted R-squared value of 0.9461 and eight predictors, the entire model is significant. The variables present in the model are : curbweight, carCompany, horsepower, carbody, enginetype, carwidth, aspiration and wheelbase. But according to the model_3 created above, drivewheel, enginelocation and carlength were also significant along with these eight variables in evaluating the response.

And the model_4 created using backward elimination consisted of highwaympg, citympg, boreratio, cylindernumber and fueltype, along with the eight significant variables.

To test the statistical significance of extra variables on the explanatory power by the inclusion of these variables in the equation, a partial F-test will be performed.

Comparing model 3 and model 5

The hypothesis test for partial F-test is as follows:

```
H0: beta(drivewheel) = beta(enginelocation) = beta(carlength) = 0
Ha: Alteast one beta != 0

anova(model_5, model_3)

## Analysis of Variance Table
##
## Model 1: log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth + aspiration + wheelbase
## Model 2: log(price) ~ carCompany + aspiration + carbody + drivewheel +
##     enginelocation + wheelbase + carlength + enginetype + carwidth +
##     horsepower + curbweight
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    168 2.3004
## 2    165 2.2601  3  0.040355 0.9821 0.4027
```

The p-value from the partial F-test is equal to 0.4027, which is greater than the 5% significance level, therefore the null hypothesis can not be rejected, which implies drive wheel, engine location and car length do not contribute significantly in determining the price of a car, once other eight variables are taken into consideration.

Comparing model 4 and model 5

The hypothesis test for partial F-test is as follows:

```
H0: beta(highwaympg) = beta(citympg) = beta(boreratio) = beta(cylindernumber)
= beta(fueltype) = 0
Ha: Alteast one beta != 0

anova(model_5, model_4)

## Analysis of Variance Table
##
## Model 1: log(price) ~ curbweight + carCompany + horsepower + carbody +
##     enginetype + carwidth + aspiration + wheelbase
## Model 2: log(price) ~ carCompany + fueltype + carbody + wheelbase +
carlength +
##     carwidth + curbweight + enginetype + cylindernumber + boreratio +
##     horsepower + citympg + highwaympg
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    168 2.3004
## 2    160 2.1627  8   0.13771 1.2735 0.2608
```

The p-value from the partial F-test is equal to 0.2608, which is greater than the 5% significance level, therefore the null hypothesis can not be rejected, which implies highwaympg, citympg, boreratio, cylindernumber and fueltype do not contribute significantly in determining the price of a car, once the other eight variables are taken into consideration.

As a result, the model obtained from forward selection is regarded as the most appropriate model.

The model equation is as follows:

```
Y.hat = 5.5533579 +  0.0004438*curbweight + 0.0132543*carCompanyAudi +
0.3154945*carCompanyBmw -0.0222016*carCompanyBuick -
0.2227836*carCompanyChevrolet -0.2402383*carCompanyDodge -
0.1402331*carCompanyHonda -0.0829145*carCompanyIsuzu -
0.1865527*carCompanyJaguar  -0.1325969*carCompanyMazda -
0.2435115*carCompanyMercury -0.2920559*carCompanyMitsubishi -
0.1863161*carCompanyNissan -0.5032313*carCompanyPeugeot -
0.2641818*carCompanyPlymouth + 0.3106696*carCompanyPorsche -
0.2136935*carCompanyRenault -0.0573704*carCompanySaab -
0.5491463*carCompanySubaru -0.2232509*carCompanyToyota  -
0.1501937*carCompanyVolkswagen -0.1210310*carCompanyVolvo +
0.0024760*horsepower  -0.2060506*carbodyhardtop -0.2650181*carbodyhatchback -
0.2230771*carbodysedan -0.2230771*carbodywagon -0.4548953*enginetypedohcv +
0.1460861*enginetypel -0.0361609*enginetypeohc + 0.2706304*enginetypeohcf -
0.0201179*enginetypeohcv + 0.1910720*enginetyperotor + 0.0305733*carwidth +
0.0711283*aspirationturbo + 0.0080603*wheelbase + 0.117
```

Prediction

The linear regression models can be used for predictions about the response variable using the predictor variables. The above model (model_5) will be used to predict the value of price for two new observations.

The observations used for prediction are as follows :

1. carCompany = Honda , aspiration = turbo, carbody = hatchback, wheelbase = 96, carwidth = 64, curbweight = 2005, enginetype = ohc, horsepower = 100

```
new_obs1 = data.frame(carCompany = "Honda",
                      aspiration = "turbo",
                      carbody = "hatchback" ,
                      wheelbase = 96 ,
                      carwidth = 64,
                      curbweight = 2005,
                      enginetype = "ohc",
                      horsepower = 100 )

predict_1 <- round(exp(predict(model_5,new_obs1,interval="prediction", level
= 0.95)),1)
predict_1
```

```
##       fit    lwr     upr
## 1 8526.6 6639.1 10950.8
```

The prediction of price for a Honda car with the above mentioned attributes is $15,258, with a 95% prediction interval between 11625.1 and 20026.2 .

2. carCompany = Mazda , aspiration = std, carbody = hatchback, wheelbase = 95, carwidth = 66.5, curbweight = 2385, enginetype = rotor, horsepower = 100

```
new_obs2 = data.frame(carCompany = "Mazda",
                      aspiration = "std",
                      carbody = "hatchback" ,
                      wheelbase = 95 ,
                      carwidth = 66.5,
                      curbweight = 2385,
                      enginetype = "rotor",
                      horsepower = 100 )

predict_2 <- round(exp(predict(model_5,new_obs2,interval="prediction", level
= 0.95)),1)
predict_2
```

```
##       fit    lwr     upr
## 1 12729.8 9819.3 16503.1
```

The prediction of price for a Mazda car with the above mentioned attributes is $12729.8, with a 95% predicition interval between 9819.3 and 16503.1.

Discussion

The results generated from the detailed multiple linear regression helped in identifying the significant predictors that affect the price of the car. The results obtained in this project helped in answering the two main questions that were listed in the problem statement. The analysis generated from this project can help the Chinese automobile organization to focus more on particular attributes of a car which will help them deciding the price for it. The management team of the Chinese organization can utilize the results to understand how prices change as a result of the independent variables They can then tweak the car's design, marketing strategy, and other aspects to meet price targets.

The business goal that was stated initially seem to be achieved by the extensive regression analysis conducted in this project. Other than the Chinese organization, other manufacturing companies that wish to enter the American automobile sector, can get a meaningful insight from the results of this analysis. There is also practical application of the best model evaluated from the analysis as the Chinese organization can input the significant car attributes that were included in the model and can get a price prediction interval to decide the prices for a particular car.

Conclusion

The final regression model that was chosen as the best model is a good and significant model as it explains a good variation in the response variable with a minimum number of variables and a good AIC value among other models that were generated during the course of this project. Out of the original 26 variables, the best model obtained from forward selection/stepwise-regression only had 8 variables that significantly affect the price of a car. The variables that significantly affect the price of car are curbweight, carcompany, horsepower, carbody, enginetype,carwidth,aspiration and wheelbase.These variables help in explaining/describing about 94.61% of variation in the price of car around its mean. The predictions obtained for random observations that were not included in the dataset from the best selected model seem to be accurate as the model used seems to explain maximum variation in the response variable price.

However, the analysis conducted has certain limitations that may have caused in not generating the most appropriate regression analysis. The dataset provided was not sufficient as certain other factors like location where the car was on sale, and other important attributes like mileage, whether the car has automatic or manual gears, sunroof, stereo system in the car etc. were not considered in predicting the car price due to their absence from the dataset provided. A detailed analysis can be performed in future with these factors included to give a more prominent and better model that will help in identifying more factors that influence the car price than the once identified in this analysis.

References

- Car Price Prediction Multiple Linear Regression. (2021). Retrieved 28 May 2021, from https://www.kaggle.com/hellbuoy/car-price-prediction
- Correlation matrix : A quick start guide to analyze, format and visualize a correlation matrix using R software - Easy Guides - Wiki - STHDA. (2021). Retrieved 28 May 2021, from http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software