# UNIVERSITY OF CALCUTTA

PROJECT WORK, B.Sc. 6th SEMESTER (H)

## CARS PRICE ANALYSIS : OPTIMIZING CAR PRICES FOR BUYERS AND SELLERS


UNIVERSITY ROLL NO. : 203012-21-0166

UNIVERSITY REGISTRATION NO. : 012-1111-0977-20

SUBJECT : STATISTICS (H)

PAPER : DSE-B2

SESSION : 2023

# CARS PRICE ANALYSIS

SANNIDHYA DAS

10.04.2023

# ACKNOWLEDGEMENT :

I am indebted to number of person for helping me in the preparation of this project. Firstly, Dr. Apurba Roy, Vice- Principal, Asutosh College, university of Calcutta. Without whose help I couldn't have been a part of this prestigious college. I owe a deep debt of gratitude to my supervisor Dr. Sankha Bhattacharya for necessary guidance, for this presentation of this dissertation, valuable comments and suggestions. I am extremely grateful to him for the necessary stimulus, support and valuable time. Special thanks to Dr. Dhiman Dutta , Head of the Department of Statistics, Asutosh college. I am greatly indebted to Dr. Shirsendu Mukhopadhay, Dr. Parthasarathi Bera and Prof. Ondrila Bose (Faculty members) often took pains and stood by me in adverse circumstances. Without their encouragement and inspiration it was not possible for me to complete this project. Finally my earnest thanks go to my friends who were always beside me when I needed them without any excuses and made these three years worthwhile. This project is not only a mere project. It is the memories spend with the whole department which has created a mutual understanding among us. There are many emotions related to this piece of work, especially respect and duty towards teachers and vice versa; educational attachment with my friends; social attachment with my college.

**Sannidhya Das**

**Student, Department of Statistics**

**Asutosh College**

# Contents

# 1 SELF DECLARATION :

I,Sannidhya Das,a student of B.Sc in Statistics Honours (SemesterVI), of University of Calcutta, Registration no. 012-1111- 0977-20, roll no.-203012-21-0166 hereby declare that I have done this project work titled as "Cars Price Analysis" under supervision of Dr. Sankha Bhattacharya (Assistant professor, Department of Statistics, Asutosh College) as a part of my B.Sc. Sem-VI examination according to the paper DSE-B2

I further declare that the piece of project work has NOT been published elsewhere for any degree or diploma or taken from any published project.

# 2 OBJECTIVE OF THE PROJECT :

The price of a car is determined by many factors, such as its brand, model, mileage, condition, features, and engine performance. Understanding how these factors influence the market value of a car can help consumers make informed decisions when buying or selling a car, and can also help car manufacturers design and market their products more effectively. In this project, I analyze the important factors that influence the price of a car using a data set of US market survey conducted by a consulting firm. I use descriptive statistics to explore the distribution and relationship of these variables, and inferential statistics to test hypotheses and compare groups. I also use regression analysis to build a predictive model that can estimate the price of a car based on its characteristics.

The model can be used to predict the price of a car given its features and condition, and can also provide insights into which factors have the most impact on the price. In this project, I try to give a comprehensive and general analysis of the car price in the US market, and also provide a practical tool for consumers and manufacturers to estimate and optimize the value of a car.

# 3 INTRODUCTION :

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts. They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. Essentially, the company wants to know:

- Which variables are significant in predicting the price of a car .
- How well those variables describe the price of a car Based on various market surveys, the consulting firm has gathered a large dataset of different types of cars across the American market.

I have to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

# 4 METHODOLOGY :

## 4.1 DATA HANDLING :

I have taken the data from a US market survey conducted by a Consulting firm from the website *kaggle.*

There are several variables ,Categorical and Numerical types including dimensions of a car , performance of a car , mileage and various other factors . I will analyse the data in 3 parts. First, in Part A, I'll do Exploratory data analysis (EDA) . Then I will make a Analysis of variance model and then in last part , I proceed with Regression Analysis .

In Exploratory Data Analysis part , I analyze data sets to summarize their main characteristics with visual methods(like barplots , scatterplots ,correlation plot , histogram etc). I used EDA for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. It helps me to uncover underlying structure, extract important variables, detect outliers and anomalies, and test underlying assumptions. In Anova portion I check wheather there is a statistically significant difference between

the means of three or more independent groups that have been split on two independent variables (eg : CompanyName and fueltype or CompanyName and carbody ) . Our analysis revealed that the standard ANOVA method was not applicable in this case, so we did not conduct it.

In Regression Analysis part , I choose the covariates in three different way and then compare between them which model explains the Response variable more appropriately . Then on that section I give some prediction of price when some choices of variables are given . Then i give a prediction model based on the estimated values of regression coefficients .

In the last section, I present a regression diagnostic plot to demonstrate the effectiveness of my analysis and provide some suggestions for improving the accuracy.

## 4.2 VARIABLES :

Here I defined all the variables which are used in my Data Analysis and the variables which are related with Cars Data .

**Car_ID :** Unique ID for each car

**CarName :** CompanyName + Model Name

**Fueltype :** Car Fuel Type (gas or diesel)

**Aspiration :** In a car engine, aspiration refers to the process of drawing air into the combustion chamber. There are two types of aspiration standard or turbo .

**Doornumber :** Number of car doors (two or four)

**Carbody :** Type of car body (convertible, sedan, hatchback, wagon, or hardtop)

**Drivewheel :** Car drive wheel (rear wheel drive, 4 wheel drive or front wheel drive)

**Enginelocation :** Location of car engine (front of rear)

**Wheelbase :** Car wheel base in inches . The wheelbase is the horizontal distance between the centers of the front and rear wheels. For road vehicles with more than two axles (e.g. some trucks), the wheelbase is the distance between the steering (front) axle and the centerpoint of the driving axle group. In the case of a tri-axle truck, the wheelbase would be the distance between the steering axle and a point midway between the two rear axles.

**Carlength :** Car length in inches

**Carwidth :** Car width in inches

**Carheight :** Car height in inches

**Curbweight :** Car weight in pounds

**Enginetype :** Car engine type (dohc, dohcv, I, ohc, ohcf, ohcv, or rotor)

**Cylindernumber :** Number of car cylinders (two, three, four, five, six, eight, or twelve)

**Enginesize :** Size of engine (numerical values of cubic inches)

**Fuelsystem :** Type of car fuel system (1bbl, 2 bbl, 4 bbl, idi, mfi, mpfi, spdi, or spfi)

**Boreratio :** Car Bore-Stroke Ratio is the ratio between the dimensions of the engine cylinder bore diameter to its piston stroke-length

**Stroke :** Car strokes (numerical value in strokes)

**Compressionratio :** Car compression ratio (ratio between the volume of the cylinder with the piston in the bottom position, Vbottom (largest volume), and in the top position, Vtop (smallest volume))

**Horsepower :** Car horsepower (numerical values of horsepower)

**Peakrpm :** Car peak RPM (revolutions per minute)

**Citympg :** Car city MPG (miles per gallon)

**Highwaympg :** Car highway MPG (miles per gallon)

**Symboling :** Acturian assessment of risk of the car (numerical values where -3 is safe, +3 is risky)

**Price :** total price of car in dollars

Later on I split the variable CarName into Company Name and Model Name . Since Car_ID and Model Names are all unique quantities and don't require to think much , I drop them when I proceed further .

## 4.3 SOFTWARE USED :

- RStudio version 4.1.1

- LyX
- MiKTeX Console
- TeXworks

### 4.4 DATA SOURCES :

I have downloaded the data from *kaggle* [Click kaggle to get directed to the website]
The raw datasets are attached below in APPENDIX- I.

### 4.5 Comparison by Diagrammatic Representation :

The use of diagrams to illustrate statistical data is very essential. The greatest way for representing any numerical data obtained in statistics is through diagrammatic representations. "A picture is worth a thousand words," according to one famous quote. In comparison to tabular or textual representations, the diagrammatic display of data provides an immediate understanding of the true scenario to be defined by the data.

It efficiently converts the exceedingly complex ideas contained in numbers into a more concrete and readily understandable form. Although diagrams are less certain, they are far more efficient in displaying data than tables. There are numerous types of diagrams in common use. Similarly, the diagrammatic representation of data gives a lot of information regarding the numerical data.

- Bar Diagram :
- Histogram:
- Scatter Plot:
- Boxplot:
- Correlation Plot:

## 5 Analysis and Discussions :

### 5.1 EXPLORATORY DATA ANALYSIS :

#### 5.1.1 THEORY :

Exploratory data analysis (EDA) is a process of examining and summarizing data sets using various techniques such as visualization, descriptive statistics, and clustering. EDA helps to identify patterns, outliers, anomalies, and relationships in the data, as well as to generate hypotheses and questions for further analysis. EDA is often the first step in data science projects, before applying more formal methods such as hypothesis testing and machine learning. And similarly we first perform EDA to understand the variables and their relations with other variables .In EDA i will show the relations of variables using various diagrams mentiond as above .

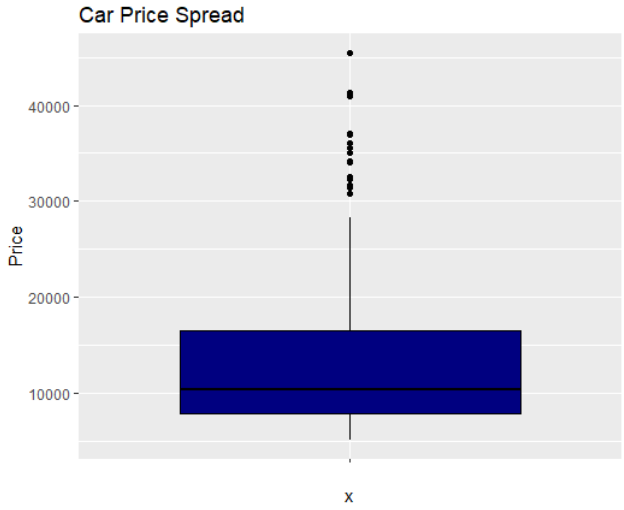
#### 5.1.2 ANALYTICAL RESULTS :

Let us now see the results or outputs given by the codes given in codes for exploratory data analysis in APPENDIX III. I've arranged them in a tabulated form to make them easily understandable.

#### 5.1.3 GRAPHS :

**PLOTTING THE BOX PLOT OF PRICE VARIABLE :**

Here i have plotted a boxplot to understand the data .




| | |
|---|---|
| count | 205.000000 |
| mean | 13276.710571 |
| standard deviation | 7988.852332 |
| minimum | 5118.000000 |
| 25% | 7788.000000 |
| 50% | 10295.000000 |
| 75% | 16503.000000 |
| maximum | 45400.000000 |

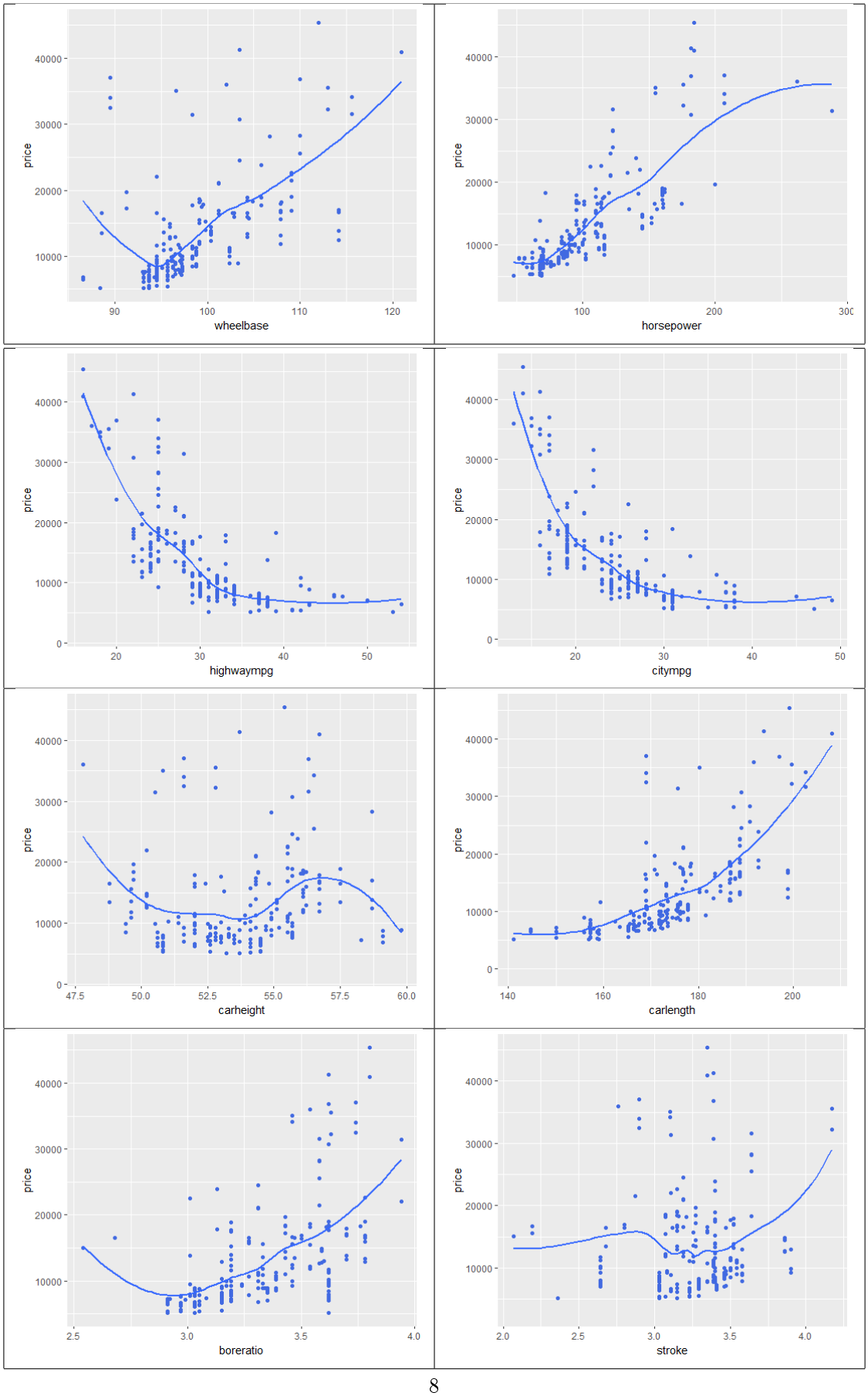
DISCUSSION:

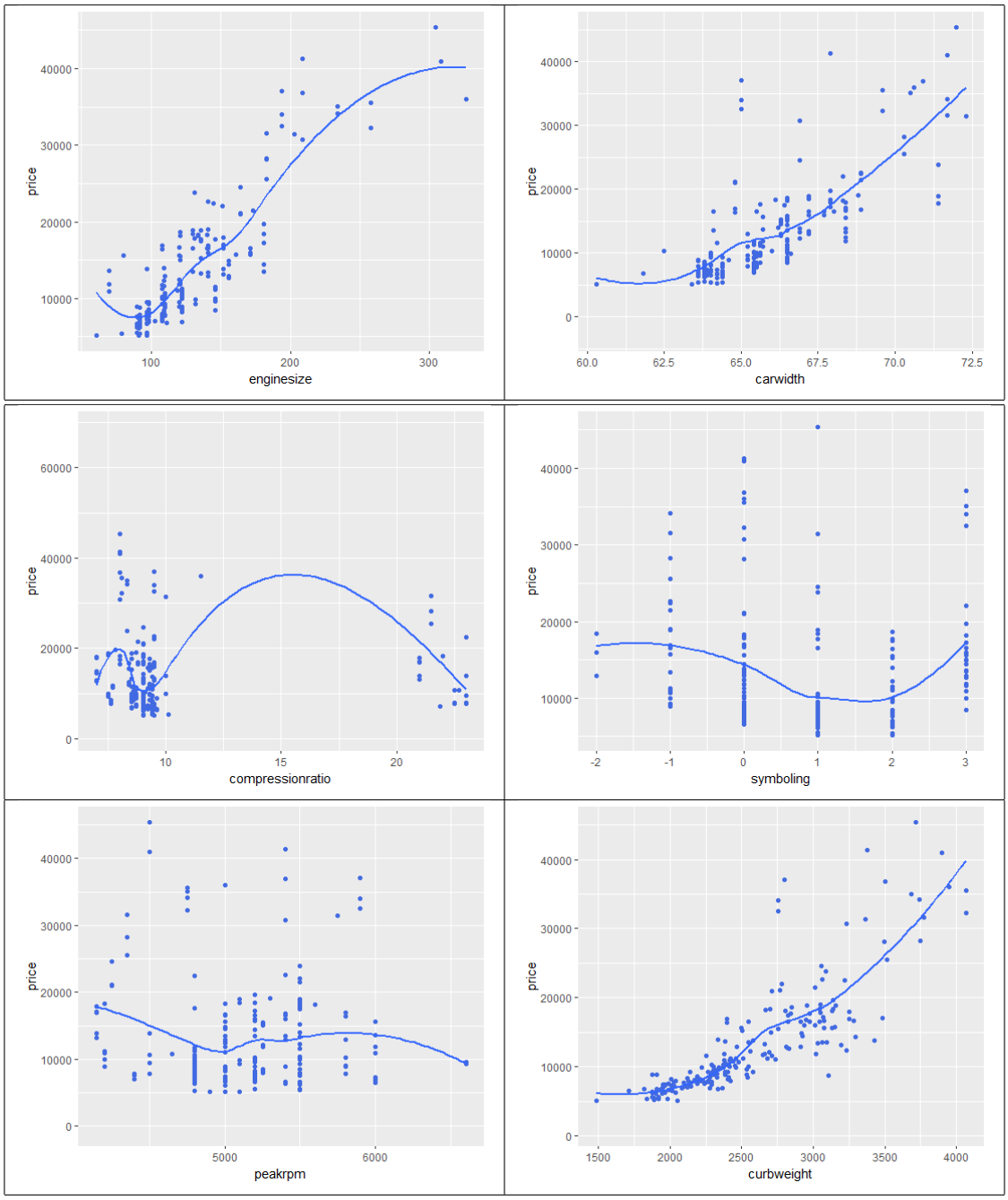
According to the boxplot, the price field has an average around 13K and a median around 10k with the most expensive car values at 45k and the cheapest cars at 5k.

**Visualization of independent variables:**

**Numerical Type Variables :**

**I. Check the linear relationship between the dependent variable "Price" and the numerical independent variables :**






DISCUSSION:

1. At first glance, the 3 variables are positively correlated but spread at higher values . We can make sure of this by looking at the Correlation Coefficient . Correlation Coefficient between Price and wheelbase: 57.781559829215 % , Correlation coefficient between Price and curbweight: 83.53048793372966 % , Correlation coefficient between Price and boreratio: 55.31732367984436 %

2. Carlength and Carwidth are more correlated than carheight which is more spread out but positive .We can make sure of this by looking at the Correlation Coefficient .Correlation coefficient between Price and carlength: 68.2920015677962 % , Correlation coefficient between Price and carwidth: 75.93252997415114 % , Correlation coefficient between Price and carheight: 11.933622657049444 %

3. Enginesize and Horsepower are positively correlated, but Stroke is more spread out (may not be related). We can make sure of this by looking at the Correlation Coefficient . Correlation coefficient between Price and enginesize: 87.41448025245117 % , Correlation coefficient between

Price and horsepower: 80.81388225362217 % , Correlation coefficient between Price and stroke: 7.944308388193101 %

4. Compressionratio, Peakrpm and symboling are not correlated .We can make sure of this by looking at the Correlation Coefficient . Correlation coefficient between Price and compressionratio: 6.798350579944266 % , Correlation coefficient between Price and peakrpm: -8.526715027785684 % , Correlation coefficient between Price and symboling: -7.997822464270351 %

5. Citympg & Highwaympg are negatively correlated. The more prices get lower, the higher the distances get, which means that the cheapest cars have better mileage than expensive cars . We can make sure of this by looking at the Correlation Coefficient . Correlation coefficient between Price and citympg: -68.57513360270397 % , Correlation coefficient between Price and highwaympg: -69.75990916465565 %

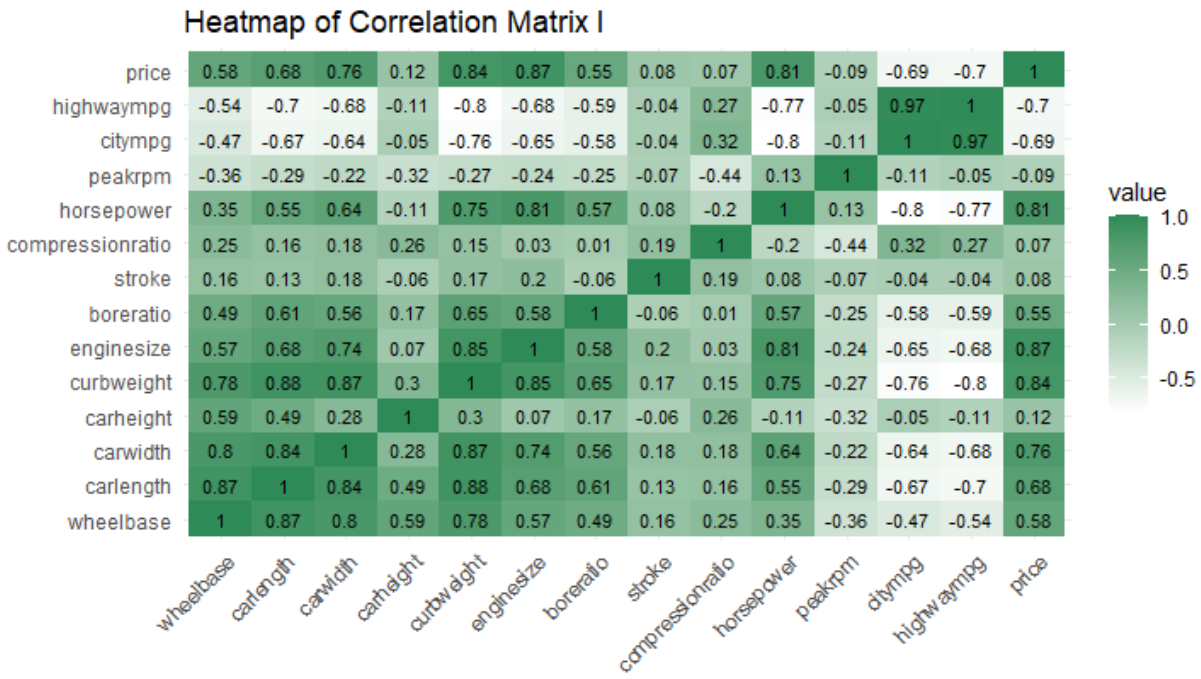
CONCLUSION :

(+) positively correlated variables with Price: wheelbase, carlenght, carwidth, curbweight, enginesize, boreratio, horesepower

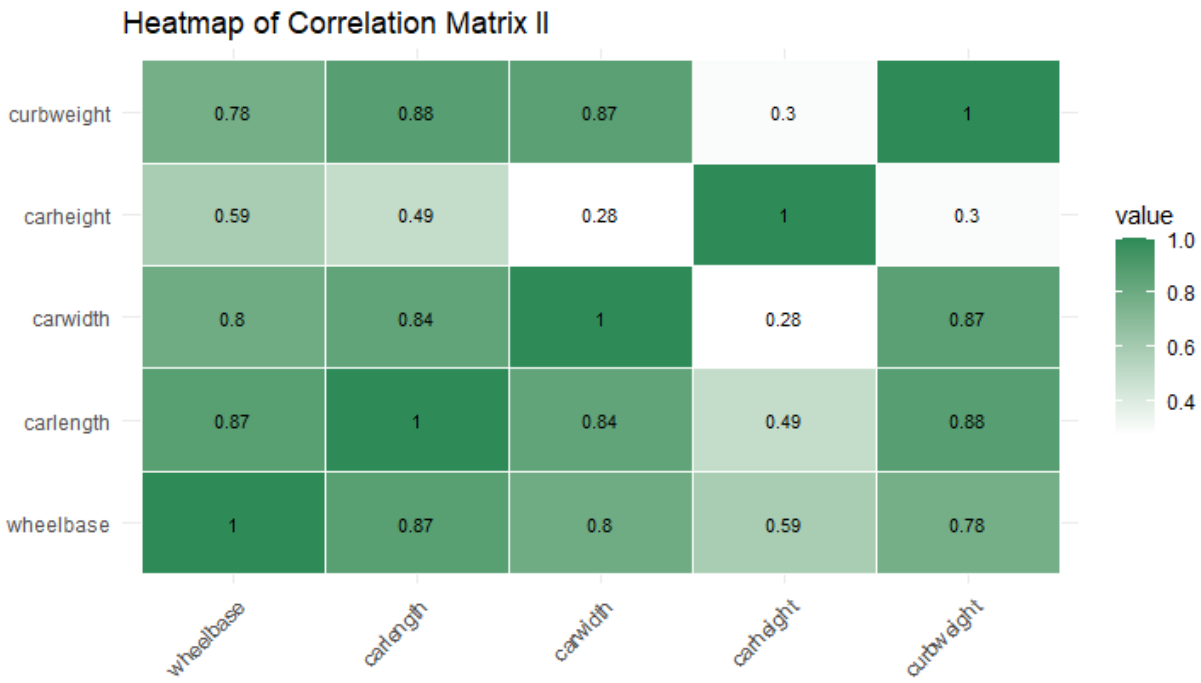
(-) negatively correlated variables with Price: citympg, highwaympg.

These variables should be kept for a better model, and the other variables should be ignored as they are not correlated with Price.

**II . Checking the multicollinearity between the correlated independent variables above and Price :**

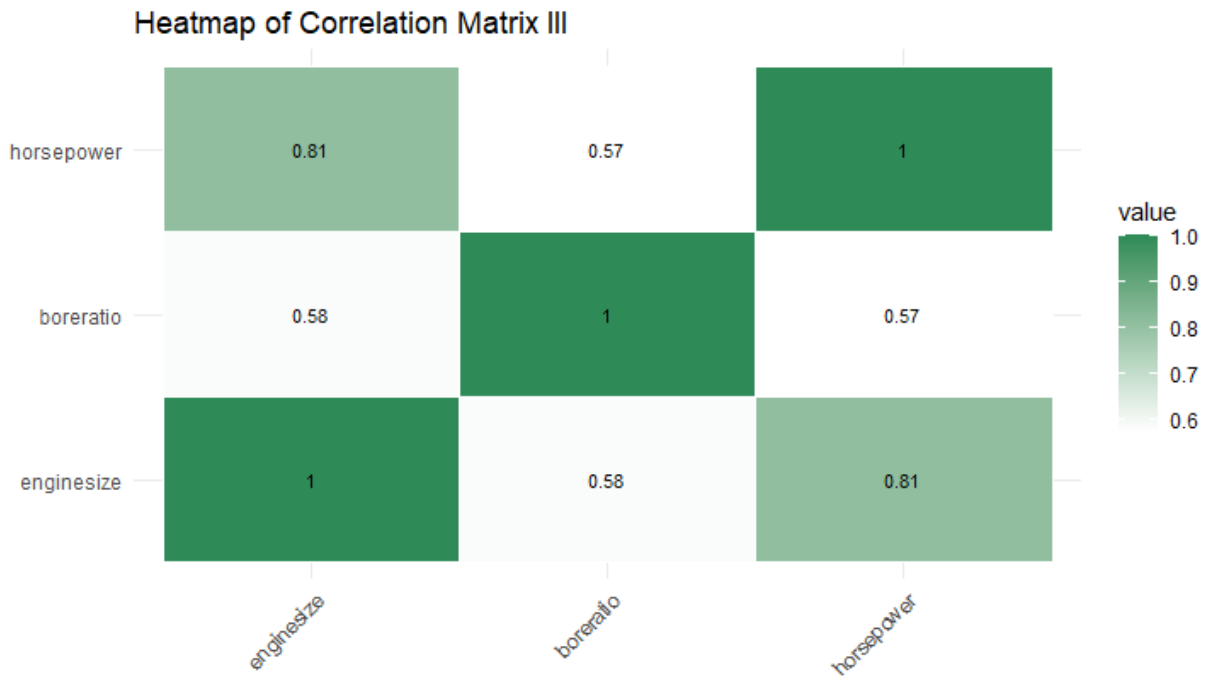


**a. Examination of the correlation between the variables specific to the dimensions of a car :**



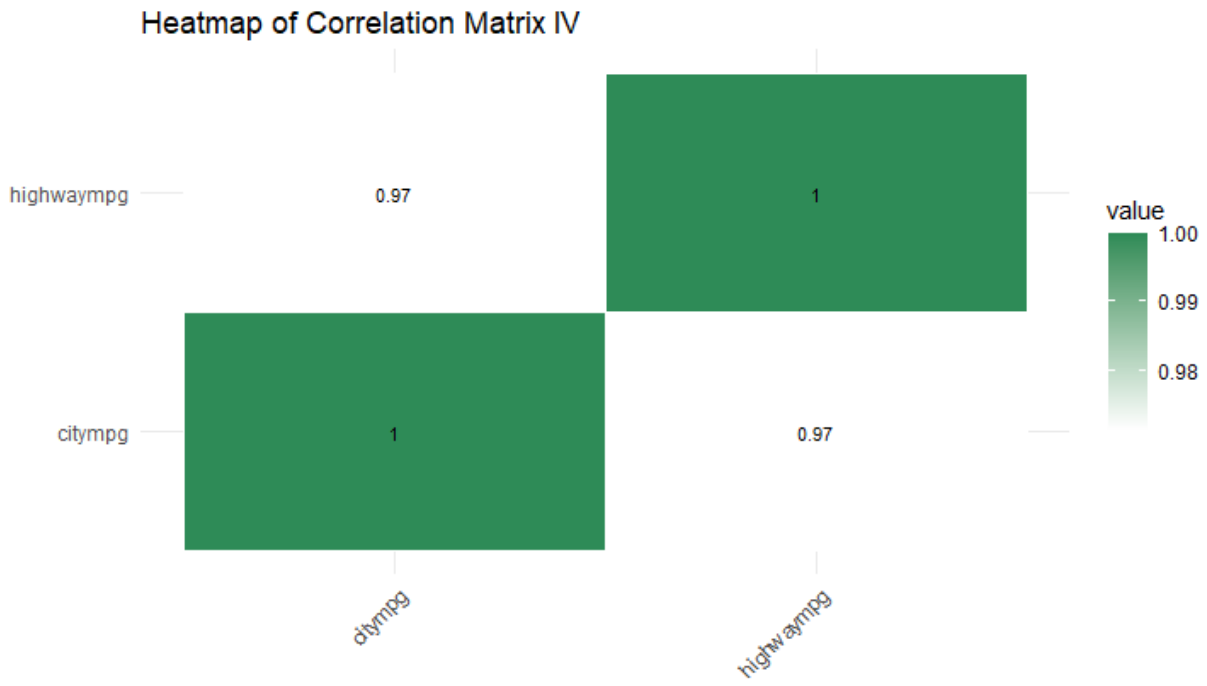

Note : Wheelbase , carlength, carwidth et curbweight [ 0.80 - 0.88 ] are very correlated and we have to keep only one between them .

**b. Examination of the correlation between the variables specific to the performance of a car :**




Note : Horsepower and enginesize are highly correlated and we need to keep only one .

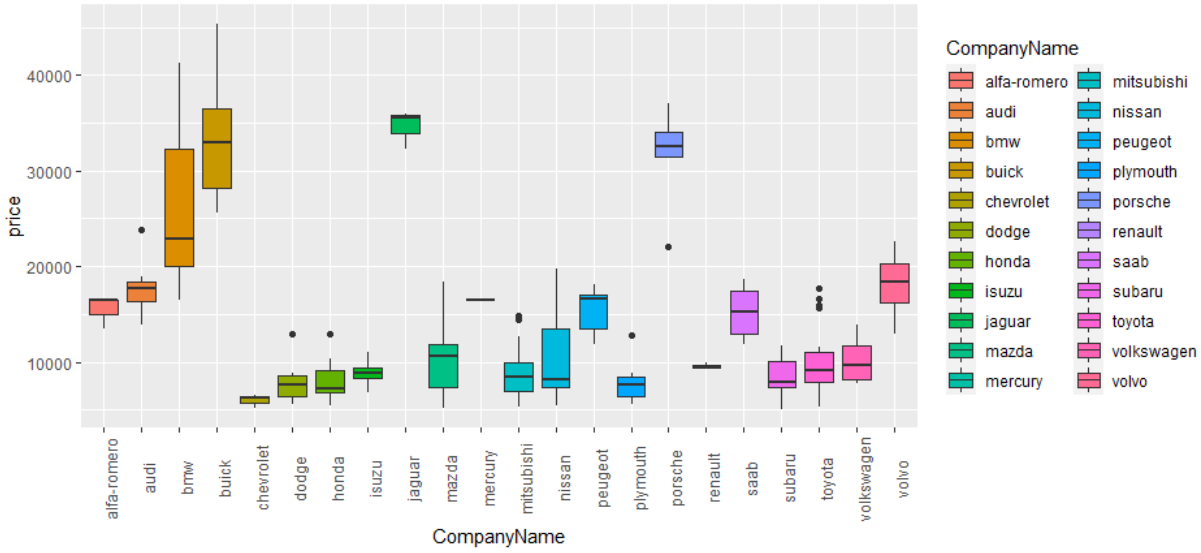
**c. Examining the correlation between citympg and highwaympg :**



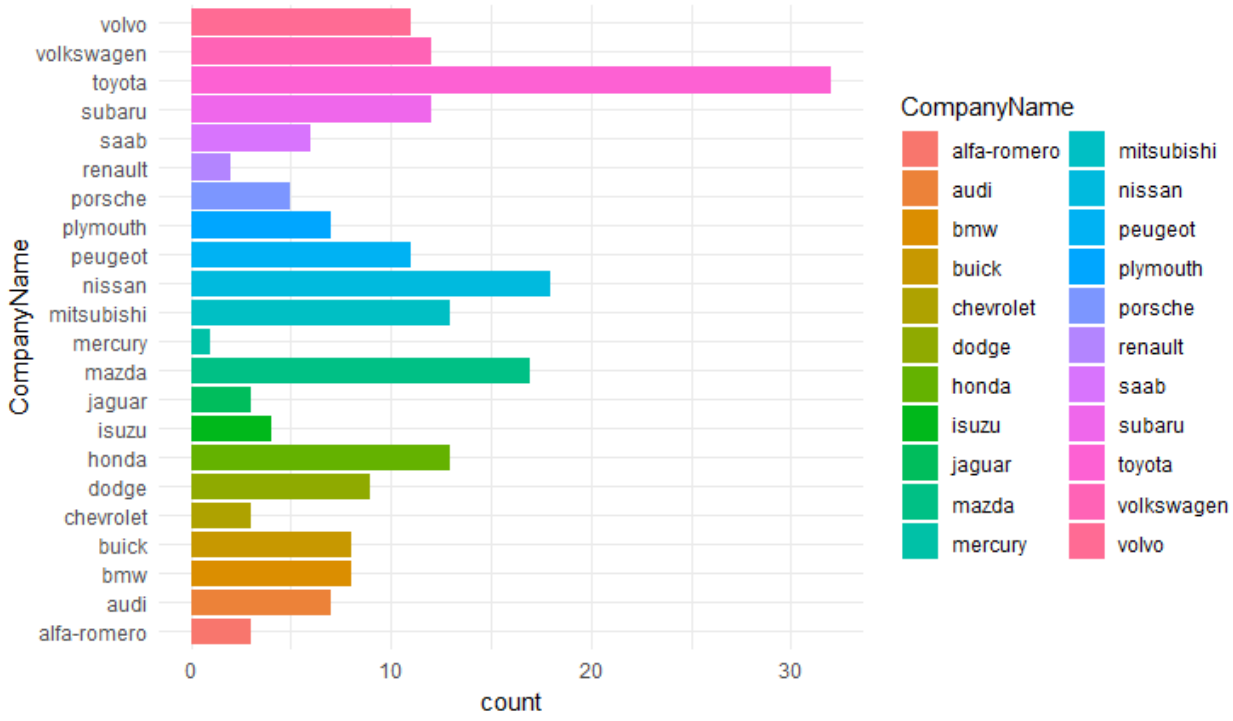

Note : citympg and highwaympg are highly correlated and we need to keep one of them .

**Categorical Type Variables :**

**Price VS CompanyName :**

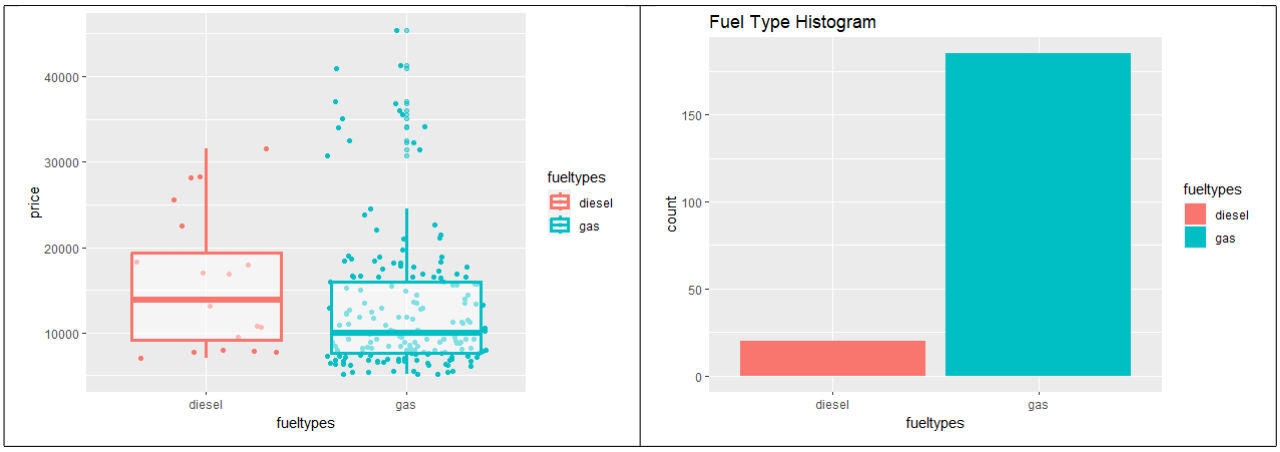



From the above Boxplot it is clear that Jaguar & buick seems to have the highest Average price range cars. Car companies like Renault & Mercury are having only one to two Cars Models . Note that Toyota , Nisaan and Mazada 's average prices clusters around 10k .




Looking at the above histogram, Toyota seems to be very popular, followed by Nissan and Mazda.

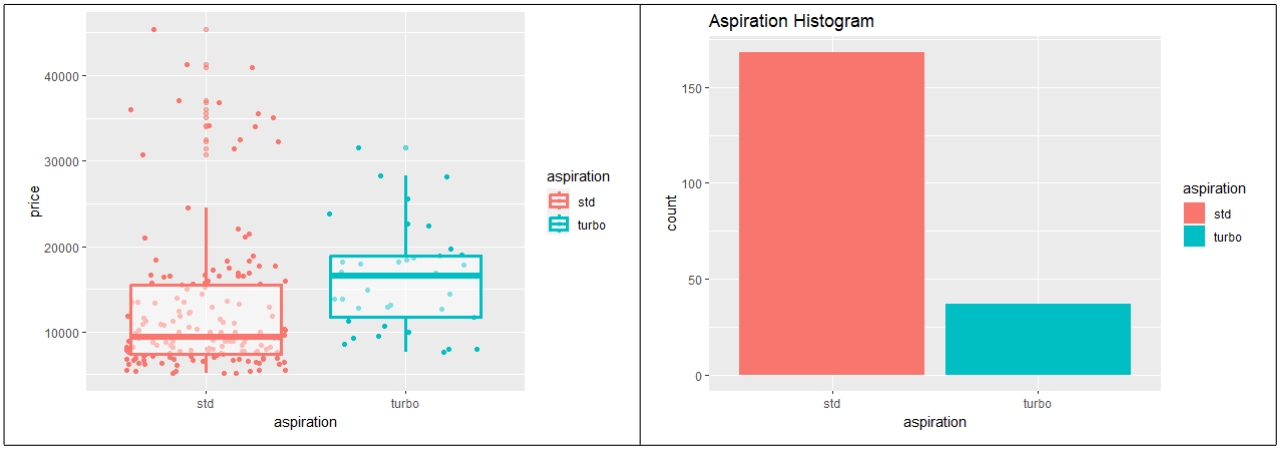
**Price VS fueltype :**




The average price of a diesel car is higher than that of gas cars, which explains, according to the histogram, why the company sold more gas cars than diesel cars.

Note: Existence of Outliers for Gas

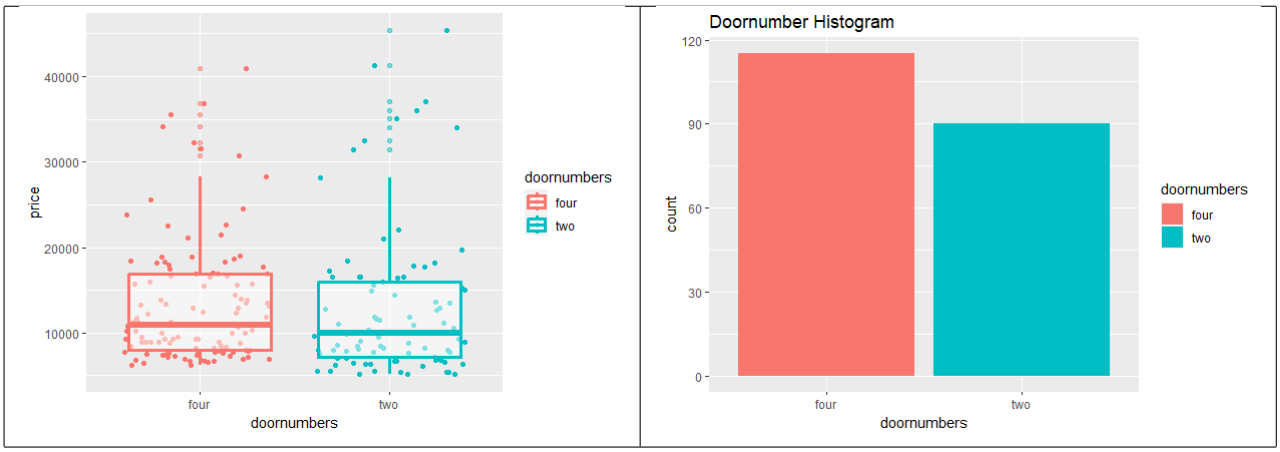
**Price VS aspiration :**




The average price of cars with turbo aspiration is higher than that of standard aspiration, which explains, according to the histogram, why the company sells cars with standard aspiration more than of cars with turbo aspiration.

Note: Existence of Outliers for Turbo and std .

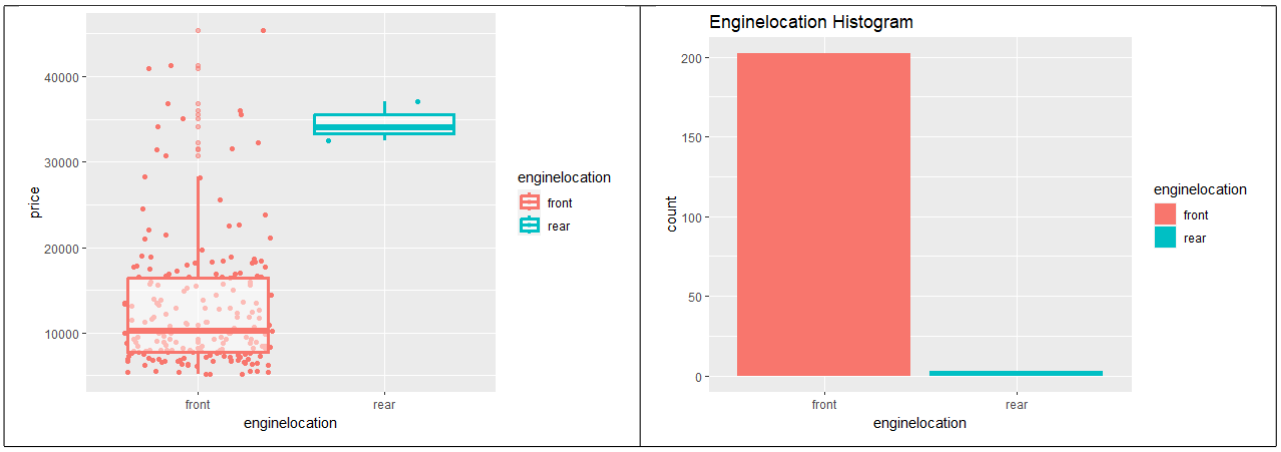
**Price VS number of Doors :**



doornumber values are pretty close, which means the price is not affected by number of doors in a car
.

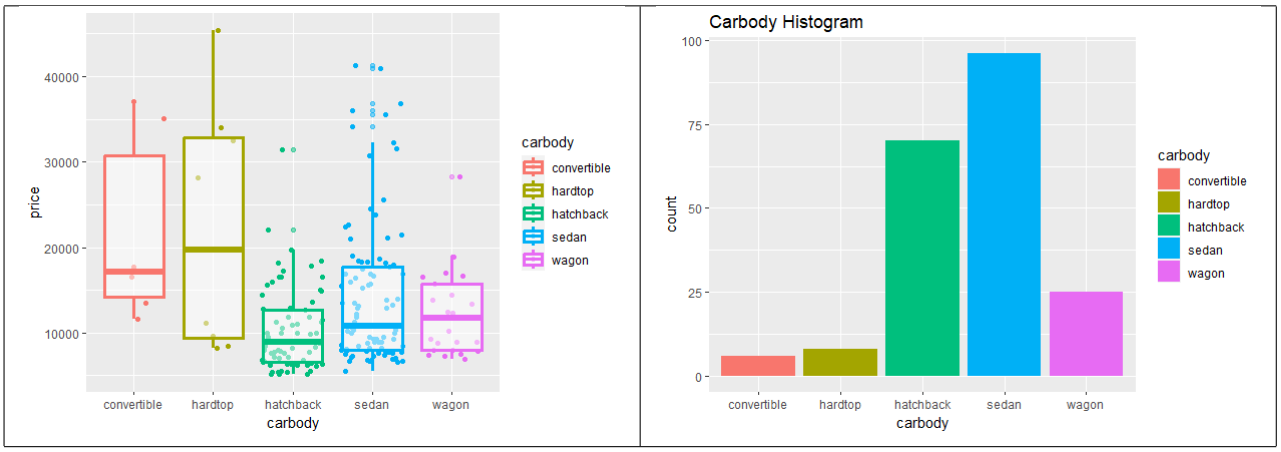
Note: Existence of Outliers in four and two .

**Price VS enginelocation :**



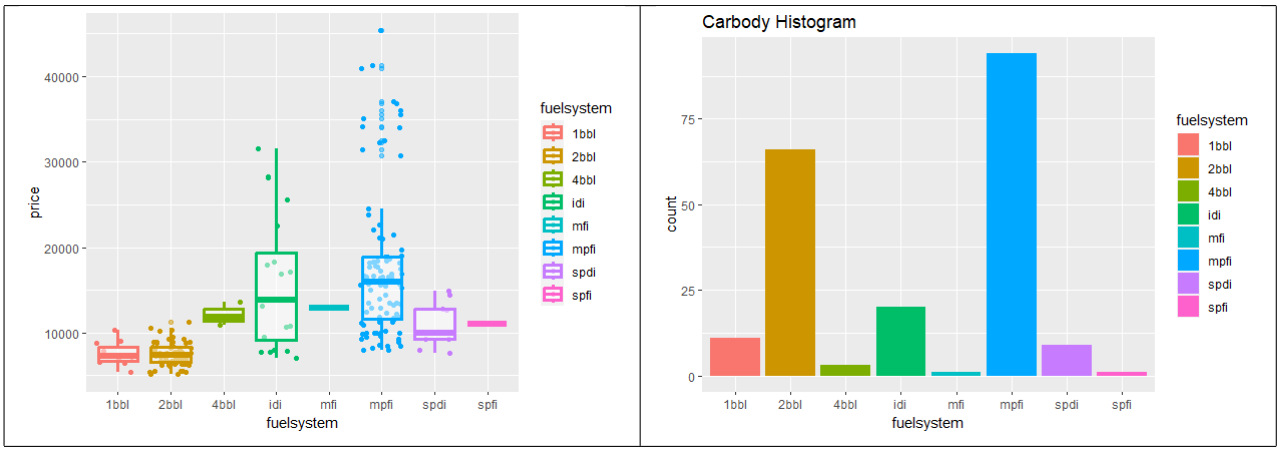

It is clear that rear cars are very expensive, which is why the company sold more cars with front rear.
Note: Existence of Outliers in front .

**Price VS carbody :**



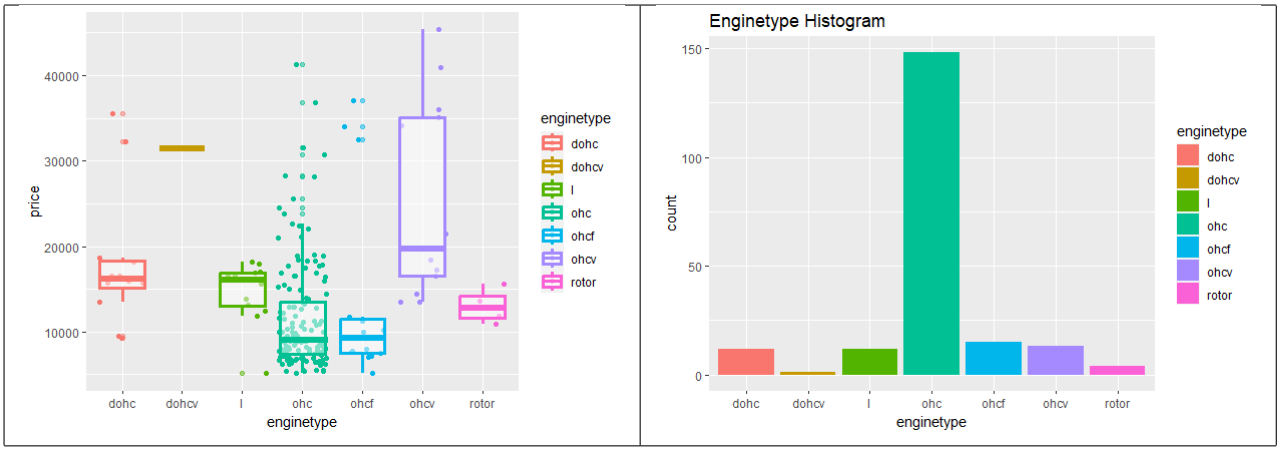

It seems that sedan is the most favored . hardtop has the highest average price.
Note: Existence of Outliers for several values.

**Price VS fuelsystem :**



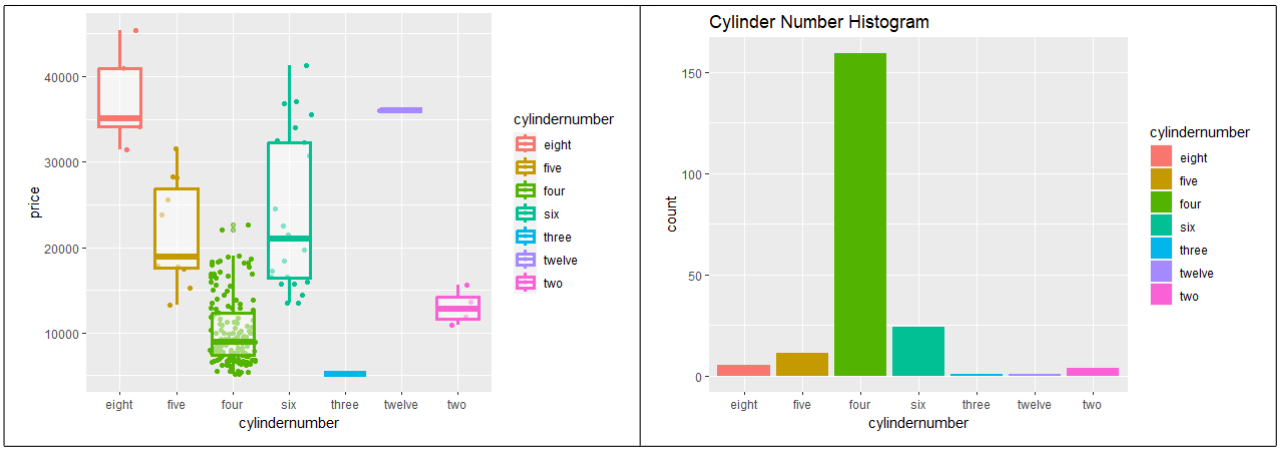

mpfi is the most favored type of fuelsystem , even though it has the highest average price.
Note: Existence of Outliers for mpfi and 2bbl .

**Price VS enginetype :**




ohc is the most favored engine type.
Note: Existence of Outliers for several values.

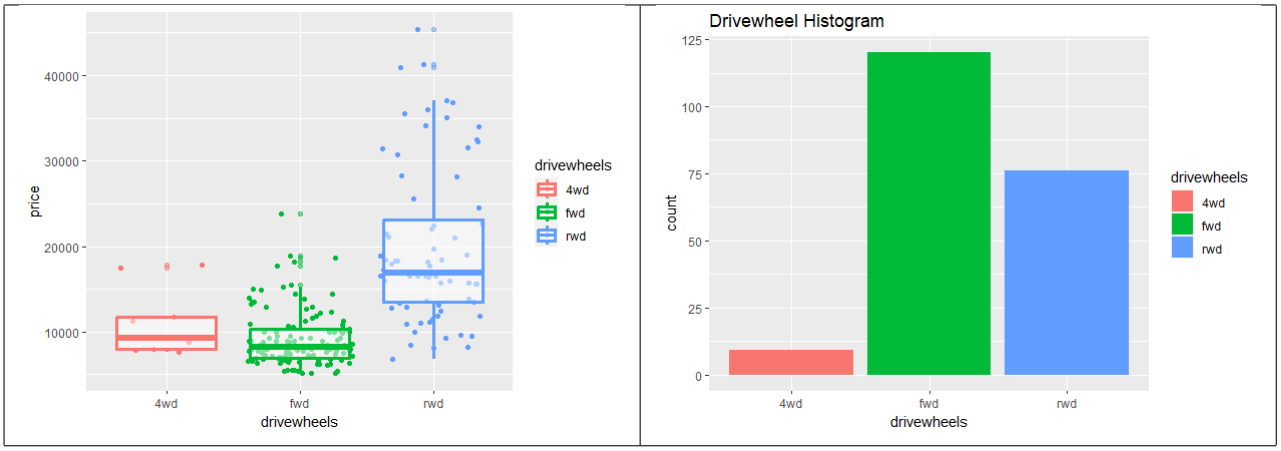
**Price VS cylindernumber :**




The four-cylinder seems to be the most favored . We can see that expensive cars have eight-cylinder , and four-cylinder are the cheapest.

Note: Existence of Outliers for four .

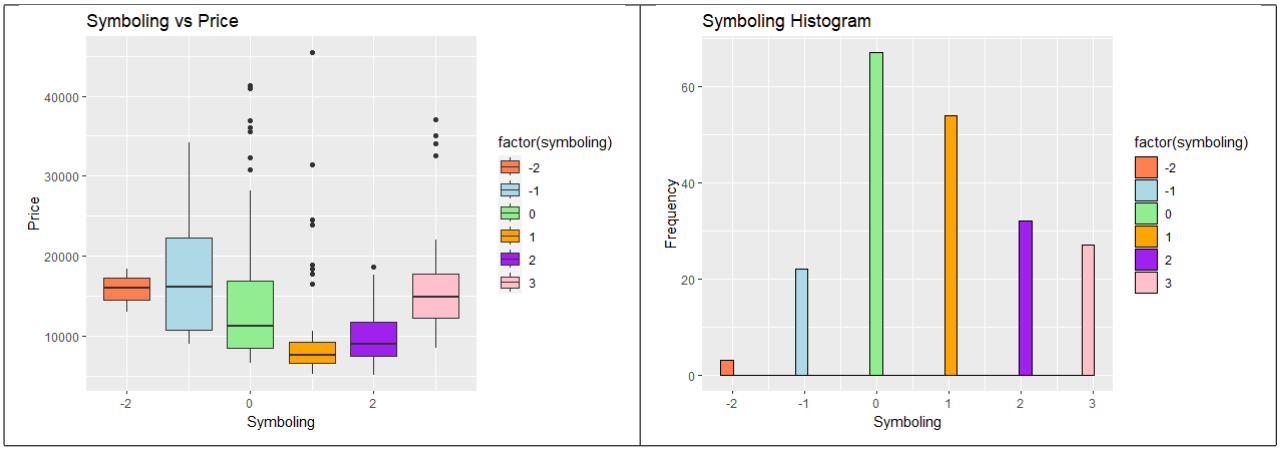
**Price VS drivewheel :**




FWD is the most favored, followed by RWD , and 4WD is the least favored even though it is cheaper than RWD .

Note: Existence of Outliers for several values.

**Price VS symboling :**




It seems that symboling 0 and 1 are the most favored . Cars with symboling -1 and -2 are the most expensive in the sense of average price , which is logical because it means that the car is more secure.

Note: Existence of Outliers for several values.

**CONCLUSIONS FROM EXPLORATORY DATA ANALYSIS :**

Toyota, Nissan and Mazda cars are more budget friendly and reliable than other brands. They offer a variety of options for different needs and preferences. Gas cars are cheaper than diesel cars in terms of initial cost and maintenance, but diesel cars have better fuel efficiency and lower emissions . Aspiration type affects the performance and price of the car. Std aspiration type is cheaper but less powerful than turbo or supercharged types. The number of doors does not affect the price significantly, so one can choose according to their convenience and style. Front engine location is more common and cheaper than rear or mid engine location, but it also has less traction and balance. Sedan is the most popular body style due to its stable average pricing and spacious interior. For fuel system, customer can choose which system will be more suitable for them depending on their driving habits and environment. Although MPFI is the most advanced and efficient system, it is not the cheapest option. OHC is the most common

and affordable engine type, which has lower noise and vibration than other types. In case of choosing drivewheel, FWD is the most favored and cheap option, but it has less stability and handling than RWD or 4WD. For symboling, most of the customers buy 0 and 1 type cars, which have lower insurance risk and cost than higher numbers.

A buyer can use the above analysis to make a smart purchase of a car at a lower price. And for a car company, I would suggest dividing their customers into two segments: common and premium. Then, they can use the above analysis to maximize their profit as according their customers .

## 5.2 REGRESSION ANALYSIS :

### 5.2.1 THEORY :

Here we are going to implement Multiple Linear Regression model to the different variables (e.g : carheight , stroke, horsepower boreratio etc.)and price would be our response or dependent variable . It is also noting that here i fit three Multiple Linear Regression model to show which fits more to our dataset . The model is given as :

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ..... + \beta_p x_{ip} + e_i$ , where i = 1(1)n . and

$y_i$ = Response Variable or Dependent Variable

$x_{i1}, x_{i2}, ....., x_{ip}$ denotes the np tupples of observations on $x_1, x_2, ....., x_p$ where $x_1, x_2, ....., x_p$ are Covariates or better to say independent variables . Here $\beta_0, \beta_1, \beta_2, ....., \beta_p$ are the unknown constants which are determinded by using the least square theory . But here in project i use lm( ) function to fit the model and then apply summary( ) function on that fitted model . By using the summary( ) function i get all the values of $\beta's$ .

Let $\rho_{y.123....p}$ denote the population multiple correlation coefficient of y on $x_1, x_2, .....x_p$.

Here we are interested in testing $H_0 : \rho_{y.123....p} = 0$ against $H_1 : H_0$ is not true. $\equiv H_0 : \beta_1 = \beta_2 = ..... = \beta_p = 0$ against $H_1 : H_0$ is not true.

**Multiple Linear Regression Model 1:**

Let us assume Price as our response variable and Carheight , Stroke , Compression Ratio , Peakrpm , Carlenght , Carwidth , Curbweight , Engine size , Highway mpg , City mpg , Symboling , Horse Power , Boreratio and Wheelbase as our independent covariates , and we have 205 total observations . Then our MLR model will be as follows :

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ..... + \beta_{13} x_{i14} + e_i$ , where i = 1(1)205 .

**Hypothesis :**

$H_0 : \beta_1 = \beta_2 = ..... = \beta_{14} = 0$ against $H_1 : H_0$ is not true.

**Multiple Linear Regression Model 2 :**

Let us assume Price as response variable and Carheight , Stroke , Compressionratio , Peakrpm , Symboling ,Horsepower , Boreratio and Wheelbase as independent covariates . We have 205 total sample observations based on this we propose a MLR model as follows :

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ..... + \beta_{13} x_{i8} + e_i$ , where i = 1(1)205 .

**Hypothesis :**

$H_0 : \beta_1 = \beta_2 = ..... = \beta_8 = 0$ against $H_1 : H_0$ is not true.

**Multiple Linear Regression Model 3 :**

Let us assume Price as response variable and Whelbase , Boreratio , Horsepower , City mpg and Symboling as independent covariates . We have 205 observations based on those observations we prepare a MLR model such that :

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + e_i$ , where i = 1(1)205 .

**Hypothesis :**

$H_0 : \beta_1 = \beta_2 = ..... = \beta_5 = 0$ against $H_1 : H_0$ is not true.

### 5.2.2 METHODOLOGY :

**Estimates of Model parameters for MLR Model 1 :**

$\beta_0 = -5.165e+04$ , $\beta_1 = 1.948e+02$ , $\beta_2 = -3.056e+03$ , $\beta_3 = 2.865e+02$ , $\beta_4 = 2.358e+00$ , $\beta_5 = -9.482e+01$ , $\beta_6 = 4.666e+02$ , $\beta_7 = 1.878e+00$ , $\beta_8 = 1.168e+02$ , $\beta_9 = 1.913e+02$ , $\beta_{10}$ =-2.869e+02 , $\beta_{11} = 2.859e+02$ , $\beta_{12} = 3.250e+01$ , $\beta_{13} = -9.844e+02$ , $\beta_{14}$ =1.677e+02

Table 5.3.2(a) : Table for MLR Model 1

| No. of Covariates | Degrees of Freedom | F - value | p-value of F - statistic | Adjusted $R^2$value |
|---|---|---|---|---|
| 14 | 190 | 78.05 | < 2.2e-16 | 0.841 |

**Estimates of Model parameters for MLR Model 2 :**

$\beta_0 = -3.649e+04$ , $\beta_1 = -1.540e+01$ , $\beta_2 = -1.728e+03$ , $\beta_3 = 2.991e+02$ , $\beta_4 = -5.747e-01$ , $\beta_5 = 4.757e+02$ , $\beta_6 = 1.513e+02$ , $\beta_7 = -1.323e+03$ , $\beta_8 = 4.495e+02$

Table 5.3.2(b) : Table for MLR Model 2

| No. of Covariates | Degrees of Freedom | F - value | p-value of F - statistic | Adjusted $R^2$value |
|---|---|---|---|---|
| 8 | 196 | 85.48 | < 2.2e-16 | 0.7681 |

**Estimates of Model parameters for MLR Model 3 :**

$\beta_0 = -60411.24$ , $\beta_1 = 551.27$ , $\beta_2 = -16.70$ , $\beta_3 = 149.43$ , $\beta_4 = 128.86$ , $\beta_5 = 592.42$

Table 5.3.2(c) : Table for MLR Model 3

| No. of Covariates | Degrees of Freedom | F - value | p-value of F - statistic | Adjusted $R^2$value |
|---|---|---|---|---|
| 5 | 199 | 124.5 | < 2.2e-16 | 0.7517 |

**DISCUSSION 1 :**

In MLR , we have several explanatory variables to predict the outcome of a response variable . There are more than one independent variables . The best model in MLR is the one that has highest adjusted $R^2$ value . The adjusted $R^2$ value is a modified version of $R^2$that adjusts for the number of predictors in the model . It penalizes models that have too many predictors and rewards models that have just enough predictors to explain the variation in the response variable . In the above three models with different number of predictors viz, 14, 8, 5 predictors . The adjusted $R^2$ values for these models are 0.841 , 0.7681 , 0.7517 respectively .

Based on these values we can conclude that the Model 1 with 14 predictors is the best because it has the highest adjusted $R^2$ value of 0.841 . This means that MLR Model 1 explains more of the variation in the response variable than other two models . But from the above heatmaps ( heatmap II , heatmap III and heatmap IV ) it is clear that in our Cars data there are some variables having high correlation between themselves . And we know in regression analysis if there are high correlations between two or more predictor variables , this can lead to unstable estimates of regression coefficients and reduces predictive power of the model .

To address this issue , there are several methods that can be used such as PCR , PLSR , Ridge Regression , Lasso Regression but in this project we are not going to tackle the multicolinearity problem with these methods rather we just eliminate those multicolinear variables and checking their adjusted $R^2$ values .

**DISCUSSION 2 :**

To build MLR Model 2, I performed variable selection on the main dataset with 26 variables. I removed the variables that had high multicollinearity with each other and created a new dataframe called Cars3 with 8 predictor variables (see APPENDIX III for details). Then I fitted a MLR Model on Cars3 and computed the adjusted R-squared value, which was 0.7681. This indicates that the model explains about 77% of the variation in the response variable.

**DISCUSSION 3 :**

Now I try to improve adjusted $R^2$ value by implementing new technique to avoide multicolinrarity . Here I check first which variables have approximate no correlation with price ,

| Corr(stroke , price) | Corr(compressionratio , price) | Corr(peakrpm , price) | Corr(symboling , price) |
|---|---|---|---|
| 0.0794 | 0.0679 | -0.085 | -0.0799 |

and drop them . Then from heatmap II It is clear that wheelbase , carlength , carwidth , curbweight are highly correlated so I choose any one among them . In heatmap III , enginesize and horsepower are highly correlated so I choose any one variable among them and similarly from heatmap IV , highway mpg and city mpg are found to be highly correlated so I choose any one from them . Rest of the variables are dropped .

Note: In this analysis, I keep the symboling variable as a predictor of car price. Symboling is a measure of how risky a car is according to its insurance classification . I believe that symboling is an important factor that affects the car price, as it reflects the safety and reliability of the car .

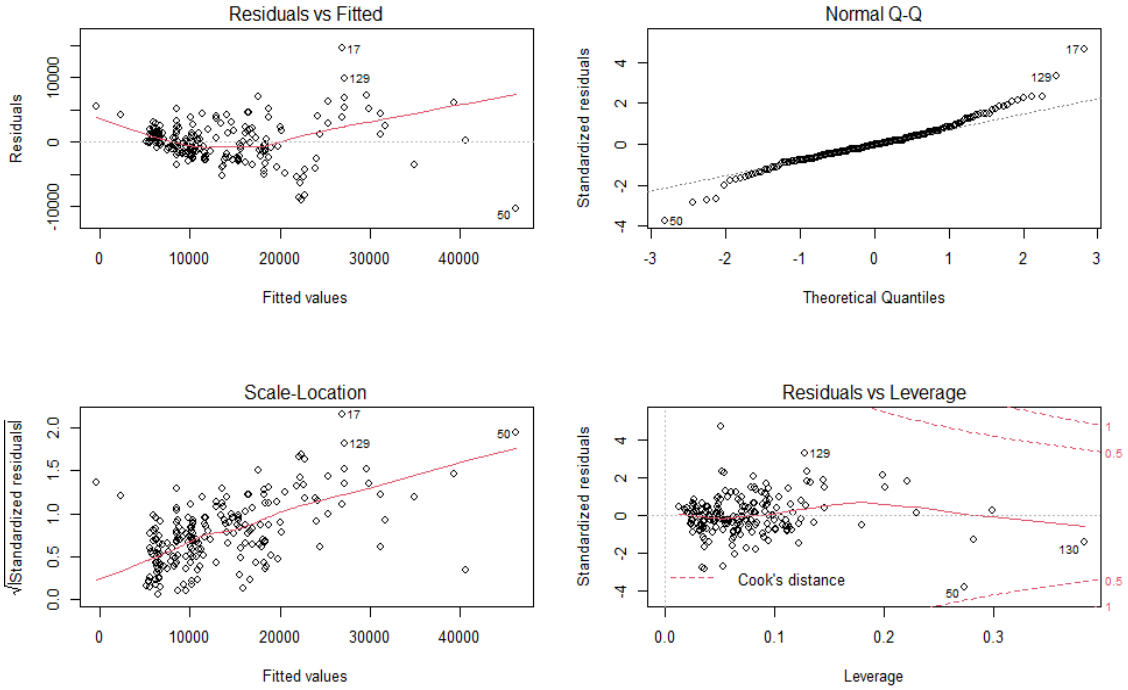
So ultimately my predictor variables for MLR Model 3 are wheelbase , boreratio , horsepower and citympg and after fitting my MLR Model I observed that my adjusted $R^2$ value is 0.7517 . This adjusted $R^2$ value is close to the previous adjusted $R^2$ value of MLR Model 2 .

**DISCUSSION 4 :**

In this case, Model 2 has a higher adjusted R-squared than Model 3, which means that it explains more variation in the response variable with fewer predictors. However, this does not necessarily mean that Model 2 is more appropriate than Model 3. The choice of the model also depends on the significance and relevance of the predictors. If all the predictors in Model 3 are important and significant for explaining the response variable, then Model 3 may be more realistic and reliable than Model 2. Therefore, one should not rely solely on the adjusted R-squared to choose the appropriate model, but also consider other factors such as the significance tests, the domain knowledge, and the purpose of the analysis.

**REGRESSION DIAGNOSTIC PLOT :**

To demonstrate the validity of my model, I performed a regression diagnostic plot on Model 3. This plot shows how well the model ts the data and whether there are any outliers or inuential points that might aect the results. The plot also helps to check the assumptions of linearity, homoscedasticity, and normality of residuals.




In the above four plots ,

**Residuals vs Fitted :**

Used to check the linear relationship assumptions. If we find equally spread residuals around a horizontal line without distinct patterns, that is a good indication we don't have non-linear relationships. In our first plot , there is no pattern in the residual plot. This suggests that we can assume linear relationship between the predictors and the outcome variables.

**Normal Q-Q :**

Used to examine whether the residuals are normally distributed. It's good if residuals points follow the straight dashed line. In our second plot, all the points fall approximately along this reference line, so we can assume normality.

**Scale-Location (or Spread-Location) :**

Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity. This is not the case in our third plot, where we have a heteroscedasticity problem. This is the reason why Our Model 3 has an adjusted $R^2$ value of 75.17 % , which means it explains most of the variation in the dependent variable. However, we could improve the adjusted $R^2$ value by addressing the issue of homoscedasticity, which is the assumption that the error terms have constant variance. Due to the syllabus constraints, we will not go into the details of how to test and correct for homoscedasticity, but we will assume that our Model 3 meets this assumption.

**Residuals vs Leverage :**

Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. In our fourth plot , the data don't present any influential points. Cook's distance lines (a red dashed line) are not shown on the Residuals vs Leverage plot because all points are well inside of the Cook's distance lines (except 1 observation ,but this can be ignored ) .

Remark : When data points have high Cook's distance scores and are to the upper or lower right of the leverage plot, they have leverage meaning they are influential to the regression results. The regression results will be altered if we exclude those cases.

### 5.2.3 RESULTS OF REGRESSION :

So according to our regression model 3 it is clear to us or better to say based on the market data of our project to predict a cars price the car parameters wheelbase , boreratio , horsepower , city mpg and symboling are the most important factors and a car price mainly depend on them .

Here's our regression equation :

$y = -60411.24 + 551.27x_1 - 16.70x_2 + 149.43x_3 + 128.86x_4 + 592.42x_5$ - - - - - - - - - - - - - - - - - - (i)

where y denotes the Car price ( in dollers ) , $x_1$ denotes wheelbase ( in inches) , $x_2$ denotes boreratio (numeric value ) , $x_3$ denotes horsepower ( hp ) , $x_4$ denotes citympg ( in MPG) , $x_5$ denotes the symboling ( numeric value ) .

# 6 COMMENT :

Suppose someone wants to buy a car then using our regression model he / she can get a estimated car price . Consider a person wants to buy a car having the specifications as follows wheelbase - 96.8 inch , boreratio - 3.19 , horsepower - 135 hp , citympg - 24 MPG , symboling - +3

then the estimated car price using (i) will be ,

y = - 60411.24 + (551.27 *96.8) - (16.70 *3.19 ) + (149.43 *135) +(128.86*24) + (592.42 * 3)

= 17941.373 $

So to purchase a car according to his / her specification he / she have to pay approox 17941.373 $ (U.S) .

# 7 REFERENCES :


- Fundamental of Statistics (Volume One & Two) – A.M. Goon, M.K. Gupta, B. Dasgupta
- An Introduction to R (R documentation ) —https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
- R for Data Science: Import, Tidy, Transform, Visualize, and Model Data – Hadley Wickham, Garrett Grolemund —https://pdfroom.com/books/r-for-data-science-import-tidy-transform-visualize-an 1vxdz4oZ2RV
- Chatterjee S. , Hadi A.S. , Price B. : Regression Analysis by Example , $3^{rd}$ Edn , John Wiley & Sons.


# 8 APPENDIX-I (Raw datasets)

Raw Dataset:

I have attached the google drive links of raw datasets below. Click the link below to get directed to https://drive.google.com/file/d/1S2ZFz2_hc1sm2beEcW_gxjyzHQBOkA9o/view?usp=share_link

# 9 APPENDIX-II

Here are some websites which helps me to understand my data :

- https://driving-test-success.com/gears/gearinfo.htm
- https://www.carbuyer.co.uk/tips-and-advice/146778/engine-size-explained

# 10 APPENDIX-III (R-Codes)

```
getwd()

setwd("C:/Users/sanni/OneDrive/Documents/My_R_files/My_College_Projects/Cars Price Prediction")

Cars = read.csv("car price.csv")

head(Cars)

dim(Cars)

str(Cars)

summary(Cars)

#Checking if i have missing observations or not

sum(is.na(Cars))

##Dataset is clean and no substitution of Null values is required

#CLEANING DATA #1. Separate the CarName variable to two columns : CompanyName and CarModel

install.packages("tidyr")

library(tidyr)

install.packages("rlang")

Cars <- separate(Cars, name, into = c("CompanyName", "CarModel"), sep = " ")

## in our analysis the two columns : ID and CarModel is unnecessary so we drop them now .

Cars$ID = NULL

Cars$CarModel = NULL

##Let's see in CompanyNames column, are there any repetitive values? unique(Cars$CompanyName)

#In reviewing the above data, we found that few company names were identical but misspelled, such as:

##'maxda' Et 'mazda' ================> mazda

##'porsche' Et 'porcshce' ============> porsche

##'toyota' Et 'toyouta' ==============> toyota

##'vokswagen' Et 'volkswagen','vw' ==> volkswagen
```

```
##'Nissan' Et 'nissan' ==============> nissan

#So we have to adjust things by replacing the values with one identical variable:

Cars$CompanyName <- gsub("maxda", "mazda", Cars$CompanyName)

Cars$CompanyName <- gsub("porcshce", "porsche", Cars$CompanyName)

Cars$CompanyName <- gsub("toyouta", "toyota", Cars$CompanyName)

Cars$CompanyName <- gsub("vokswagen", "volkswagen", Cars$CompanyName)

Cars$CompanyName <- gsub("vw", "volkswagen", Cars$CompanyName)

Cars$CompanyName <- gsub("Nissan", "nissan", Cars$CompanyName)

#2. Exploratory Data Analysis

##Since the independent variable (i.e Price) is continuous numerical variable, and there is many dependat variables, we we will use Multiple linear regression.

##Dependent variable visualization: Price

library(ggplot2)

ggplot(data=Cars, aes(x="", y=price)) + geom_boxplot(fill="navyblue", color="black")
+ labs(title="Car Price Spread", y="Price")

summary(Cars$price)

##According to the boxplot, the price field has an average around 13K and a median around 10k with the most expensive car values at 45k and the cheapest cars at 5k.

##Since we have mean > median, then our distribution is positively asymmetric, as we can see in the following histogram:

ggplot(data=Cars,aes(x=price))+geom_histogram(fill="lightblue",color="white")+ labs(title="Histogr
of Price", x="Price")

###Distribution plot

ggplot(data=Cars, aes(x=price)) + geom_density(fill="lightgreen", color="lightgreen")
+labs(title="Car Price Distribution Plot", x="Price")

##Conclusion

##Which means that most of the prices offered by this company are low.

##As seen below, we have 75% prices are around 16k, or 25% between 17k and 45k.

#Visualization of independent variables :
```

# ***Numerical***

##A. Checking the linear relationship between the dependent variable "Price" and the numerical independent variables

### we check this by drawing scatterplot

###A1. Price vs wheelbase

ggplot(data=Cars,aes(x=wheelbase,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$wheelbase,Cars$price)

# positively correlated as shown in the figure

###A2. Price vs Curbweight

ggplot(data=Cars,aes(x=curbweight,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$curbweight,Cars$price)

###A3. Price Vs Boreratio

ggplot(data=Cars,aes(x=boreratio,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$boreratio,Cars$price)

##At first glance, the 3 variables are positively correlated but spread at higher values.

##We can make sure of this by looking at the Coefficient of Correlation. calculated above.

###A4. Price vs Carlength

ggplot(data=Cars,aes(x=carlength,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$carlength,Cars$price)

###A5. price vs Carwidth

ggplot(data=Cars,aes(x=carwidth,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$carwidth,Cars$price)

###A6. price vs Carheight

ggplot(data=Cars,aes(x=carheight,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$carheight,Cars$price)

##Carlength and Carwidth are more correlated than carheight which is more spread out but positive.

##We can make sure of this by looking at the Coefficient of Correlation

###A7. price vs Enginesize

ggplot(data=Cars,aes(x=enginesize,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$enginesize,Cars$price)

###A8. price vs Horsepower

ggplot(data=Cars,aes(x=horsepower,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$horsepower,Cars$price)

###A9. price vs Stroke

ggplot(data=Cars,aes(x=stroke,y=price))+geom_point(colour="royalblue") + geom_smooth(fill=NA)

cor(Cars$stroke,Cars$price)

##Enginesize and Horsepower are positively correlated, but Stroke is more spread out (may not be related).

##We can make sure of this by looking at the Coefficient of Correlation

###A10. price vs compressionratio

ggplot(data=Cars,aes(x=compressionratio,y=price))+geom_point(colour="royalblue") + geom_smooth(fill=NA)

cor(Cars$compressionratio,Cars$price)

###A11. price vs peakrpm

ggplot(data=Cars,aes(x=peakrpm,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

```
cor(Cars$peakrpm,Cars$price)

###A12. price vs symboling

ggplot(data=Cars,aes(x=symboling,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$symboling,Cars$price)

##Compressionratio, Peakrpm and symboling are not correlated.

## We can make sure of this by looking at the Coefficient of Correlation

###A13. price vs citympg

ggplot(data=Cars,aes(x=citympg,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$citympg,Cars$price)

###A14. price vs highwaympg

ggplot(data=Cars,aes(x=highwaympg,y=price))+geom_point(colour="royalblue")

+ geom_smooth(fill=NA)

cor(Cars$highwaympg,Cars$price)

##Citympg & Highwaympg are negatively correlated.

##The more prices get lower, the higher the distances get, which means that the cheapest cars have better mileage than expensive cars.

##We can make sure of this by looking at the Coefficient of Correlation

## Conclusion

#(+) positively correlated variables with Price:  wheelbase, carlenght, carwidth, curb-weight, enginesize, boreratio, horesepower

#(-) negatively correlated variables with Price: citympg, highwaympg

#These variables should be kept for a better model, and the other variables should be ignored as they are not correlated with Price

# Checking the multicollinearity between the correlated independent variables above and Price

install.packages("reshape2")
```

```
library(reshape2)

# Calculate the correlation matrix

corr_matrix <- cor(Cars[,c("wheelbase", "carlength", "carwidth", "carheight","curbweight","enginesize"
"horsepower","peakrpm","citympg","highwaympg","price" )])

# Melt the correlation matrix into a long format

corr_melted <- melt(corr_matrix)

# Create the heatmap

ggplot(corr_melted, aes(x=Var1, y=Var2, fill=value)) + geom_tile() + scale_fill_gradient(low="white",
high="seagreen") + labs(title="Heatmap of Correlation Matrix I", x="", y="") + theme_minimal()
+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + geom_text(aes(label=round(value,2)),
color="black", size=3)

#a. Examination of the correlation between the variables specific to the dimensions of a
car

#i.e. weight, height etc

dimension_col_list = c('wheelbase', 'carlength', 'carwidth','carheight','curbweight')

Cars_matrix = cor(Cars[, dimension_col_list])

Cars_melted = melt(Cars_matrix)

ggplot(Cars_melted, aes(x=Var1, y=Var2)) + geom_tile(aes(fill=value), colour="white")
+ scale_fill_gradient(low="white", high="seagreen") + labs(title="Heatmap of Correla-
tion Matrix II", x="", y="") + theme_minimal() + theme(axis.text.x = element_text(angle
= 45, hjust = 1)) + geom_text(aes(label=round(value,2)), color="black", size=3)

##Wheelbase , carlength, carwidth et curbweight [ 0.80 - 0.88 ] are very correlated and
we have to keep only one between them.

#b. Examination of the correlation between the variables specific to the performance of a
car

dimension_col_list1 = c('enginesize','boreratio','horsepower')

data_matrix = cor(Cars[, dimension_col_list1])

data_melted = melt(data_matrix)

ggplot(data_melted, aes(x=Var1, y=Var2)) + geom_tile(aes(fill=value), colour="white")
+ scale_fill_gradient(low="white", high="seagreen") + labs(title="Heatmap of Correla-
tion Matrix III", x="", y="") + theme_minimal() + theme(axis.text.x = element_text(angle
= 45, hjust = 1)) + geom_text(aes(label=round(value,2)), color="black", size=3)

#Horsepower and enginesize are highly correlated and we need to keep only one.
```

```
#c. Examining the correlation between citympg and highwaympg

dimension_col_list2 = c('citympg','highwaympg')

data_matrix1 = cor(Cars[, dimension_col_list2])

data_melted1 = melt(data_matrix1)

ggplot(data_melted1, aes(x=Var1, y=Var2)) + geom_tile(aes(fill=value), colour="white")
+ scale_fill_gradient(low="white", high="seagreen") + labs(title="Heatmap of Correla-
tion Matrix IV", x="", y="") + theme_minimal() + theme(axis.text.x = element_text(angle
= 45, hjust = 1)) + geom_text(aes(label=round(value,2)), color="black", size=3)

#citympg and highwaympg are highly correlated and we need to keep one of them.

# *** Categorical ***

## Price VS CompanyName

ggplot(Cars, aes(x=CompanyName, y=price , fill=CompanyName)) + geom_boxplot()

+ theme(axis.text.x = element_text(angle = 90))

ggplot(Cars, aes(x=CompanyName,fill= CompanyName)) + geom_bar() + coord_flip()
+ theme_minimal()

#Looking at the above histogram, Toyota seems to be very popular, followed by Nissan
and Mazda.

## Price VS fueltype

ggplot(data=Cars,aes(x=fueltypes,y=price,color=fueltypes)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=fueltypes,fill=fueltypes)) + geom_bar() + ggtitle("Fuel Type His-
togram")

##The average price of a diesel car is higher than that of gas cars, which explains, according
to the histogram, why the company sold more gas cars than diesel cars.

#Price VS aspiration

ggplot(data=Cars,aes(x=aspiration,y=price,color=aspiration)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=aspiration,fill=aspiration)) + geom_bar() + ggtitle("Aspiration
Histogram")

##The average price of cars with turbo aspiration is higher than that of standard aspira-
tion, which explains, according to the histogram, why the company sells cars with standard
aspiration more than of cars with turbo aspiration.
```

```
#Price VS doornumber

ggplot(data=Cars,aes(x=doornumbers,y=price,color=doornumbers)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=doornumbers,fill=doornumbers)) + geom_bar() + ggtitle("Doornumber
Histogram")

#doornumber values are pretty close, which means the price is not affected by doornumber

#Price VS enginelocation

ggplot(data=Cars,aes(x=enginelocation,y=price,color=enginelocation)) + geom_jitter() +
geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=enginelocation,fill=enginelocation)) + geom_bar() + ggtitle("Enginelocation
Histogram")

#It is clear that rear cars are very expensive, which is why the company sold more cars
with front rear.

#Price VS carbody

ggplot(data=Cars,aes(x=carbody,y=price,color=carbody)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=carbody,fill=carbody)) + geom_bar() + ggtitle("Carbody His-
togram")

#It seems that sedan is the most favored.hardtop has the highest average price.

#Price VS fuelsystem

ggplot(data=Cars,aes(x=fuelsystem,y=price,color=fuelsystem)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=fuelsystem,fill=fuelsystem)) + geom_bar() + ggtitle("Carbody
Histogram")

#mpfi is the most favored type of fuelsystem , even though it has the highest average price.

#Price VS enginetype

ggplot(data=Cars,aes(x=enginetype,y=price,color=enginetype)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=enginetype,fill=enginetype)) + geom_bar() + ggtitle("Enginetype
Histogram")
```

```
#ohc is the most favored engine type.

#Price VS cylindernumber

ggplot(data=Cars,aes(x=cylindernumber,y=price,color=cylindernumber)) + geom_jitter()
+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=cylindernumber,fill=cylindernumber)) + geom_bar() + ggtitle("Cylinder
Number Histogram")

#The four-cylinder seems to be the most favored. We can see that expensive cars have
eight-cylinder , and four-cylinder are the cheapest.

#Price VS drivewheel

ggplot(data=Cars,aes(x=drivewheels,y=price,color=drivewheels)) + geom_jitter()

+ geom_boxplot(size=1.2,alpha=0.5)

ggplot(data=Cars,aes(x=drivewheels,fill=drivewheels)) + geom_bar() + ggtitle("Drivewheel
Histogram")

#FWD is the most favored, followed by RWD , and 4WD is the least favored even though
it is cheaper than RWD .

#Price VS symboling

ggplot(data = Cars) + geom_boxplot(aes(x = symboling, y = price, fill = factor(symboling)))
+ labs(title = "Symboling vs Price", x = "Symboling", y = "Price") + scale_fill_manual(values
= c("coral", "lightblue", "lightgreen", "orange", "purple" , "pink"))

ggplot(data = Cars) + geom_histogram(aes(x = symboling, fill = factor(symboling)), color
= "black") + labs(title = "Symboling Histogram", x = "Symboling", y = "Frequency")
+ scale_fill_manual(values = c("coral", "lightblue", "lightgreen", "orange", "purple" ,
"pink"))

#It seems that symboling 0 and 1 are the most favored.

#Cars with symboling -1 and -2 are the most expensive, which is logical because it means
that the car is more secure.

#Now we are going to subset our variables according to our use

Cars1 = subset(Cars, select = -c(carheight, stroke, compressionratio, peakrpm, carlength,
carwidth, curbweight, enginesize, highwaympg, citympg,symboling,horsepower,boreratio,wheelbase))

head(Cars1)

Cars2 = subset(Cars, select = c(carheight, stroke, compressionratio, peakrpm, carlength,
carwidth, curbweight, enginesize, highwaympg, citympg,symboling,horsepower,boreratio,wheelbase
, price))

head(Cars2)
```

```
Cars4 = Cars[, !(names(Cars) %in% c('carheight', 'stroke', 'compressionratio', 'peakrpm',
'carlength', 'carwidth', 'curbweight', 'enginesize', 'highwaympg'))]

head(Cars4)

#Multiple Regression

mat_a = subset(Cars2,select = -c(price))

numeric = mat_a[sapply(mat_a,is.numeric)]

descrcor=cor(numeric)

descrcor

library(caret)

highlyCorrelated=findCorrelation(descrcor,cutoff = 0.7)

highlycorcol = colnames(numeric)[highlyCorrelated]

highlycorcol

#we can find which columns are highly correlated and removing them from the model
Cars3 = Cars2[ , -which(colnames(Cars2) %in% highlycorcol)] dim(Cars3)

#Model3

library(car)

Model3= lm(price~. , data=Cars3)

summary(Model3)

#Model4

Model4 = lm(price~. , data=Cars2)

summary(Model4)

#Model5

Model5 = lm(price~wheelbase + boreratio + horsepower + citympg , data=Cars4)

summary(Model5)
```