Reg No - 20414110001

Prediction of Price of Used Car

Presidency University, Kolkata

Suvendu Samanta

11/1/2021

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude and deep regards to my teachers for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. His blessings, help, time to time guidance will carry us a long way in the journey of life which we are about to embark.

I would also like to thank the institute authorities for giving me the chance to do the project and for providing the environment and necessary facilities required for the completion of my project.

I would like to thank our department professors for teaching all the necessary topics with immense care which was needed to make the project fruitful.

Finally, I would like to extend our gratitude and thanks to my friends (Specially Suvendu Samanta; Presidency University) and parents. Without their constant support and encouragement, it would not have been possible to proceed for me.

2021-11-16 Suvendu Samanta

CONTENTS

Introduction

In this modern society of economic war; We live in a society where more than 50% people belong to middle class category. And for a middle-class man to afford a car is a dream, and to chase their dreams often they give priority to buy a 2nd hand car or in other word a used car.

In this project looking at some factors I have fitted a model which can estimate an ideal price for a used car. But here I consider the Chinese 2nd hand car market as it is more evolved which can help the Indian companies to set up their environment fast.

Cars price Estimation

Study Objective:

Problem Statement:

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.

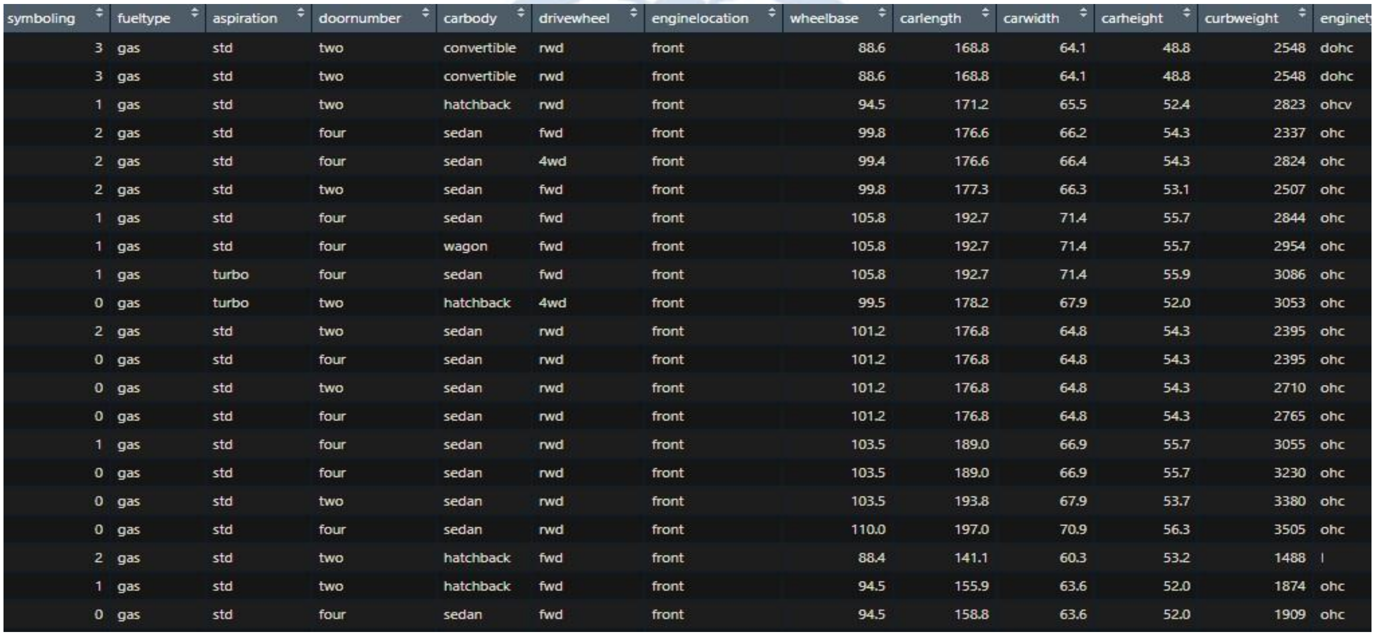
They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. The company wants to 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 data set of different types of cars across the American market.

Data Description:

The following data has been collected from online cars website; carsdekho.com. The whole data can be used from here; https://drive.google.com/file/d/1csctWdjNJPRqnb5VtMsar436sxkYaVLp/view?usp=sharing

Here is some glimpse of the data set;



symboling	fueltype	aspiration	doornumber	carbody	drivewheel	enginelocation	wheelbase	carlength	carwidth	carheight	curbweight	enginet
3	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc
3	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548	dohc
1	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv
2	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc
2	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc
2	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc
1	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc
1	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc
1	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc
0	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52.0	3053	ohc
2	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc
0	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc
0	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc
0	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc
1	gas	std	four	sedan	rwd	front	103.5	189.0	66.9	55.7	3055	ohc
0	gas	std	four	sedan	rwd	front	103.5	189.0	66.9	55.7	3230	ohc
0	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc
0	gas	std	four	sedan	rwd	front	110.0	197.0	70.9	56.3	3505	ohc
2	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l
1	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52.0	1874	ohc
0	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52.0	1909	ohc

Business Goal:

We are required to model the price of cars with the available independent variables. It 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 management to understand the pricing dynamics of a new market.

The dataset aims to determine what factors influence the price of a car and proposes to answer a couple of questions of which, only the first has been chosen and is re-formulated to fulfill our objective of the exercise:

From the features present in the dataset, which of them linearly affect the price of a car?

```
# Library Load
library(tidyverse)
library(gridExtra)
library(DataExplorer)
library(car)
library(nortest)
library(lmtest)
```

Let load the data set here:

```
data<- read.csv(file="C://Users//DELL//Desktop//presidency university//proj_junior/
/CarPrice_sorted.csv", head=TRUE)
```

Checking for the data type….

```
str(data)

## '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         : chr  "alfa-romero giulia" "alfa-romero stelvio" "alfa-romer
o Quadrifoglio" "audi 100 ls" ...
##  $ fueltype        : chr  "gas" "gas" "gas" "gas" ...
##  $ aspiration      : chr  "std" "std" "std" "std" ...
##  $ doornumber      : chr  "two" "two" "two" "four" ...
##  $ carbody         : chr  "convertible" "convertible" "hatchback" "sedan" ...
##  $ drivewheel      : chr  "rwd" "rwd" "rwd" "fwd" ...
##  $ enginelocation  : chr  "front" "front" "front" "front" ...
##  $ 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      : chr  "dohc" "dohc" "ohcv" "ohc" ...
##  $ cylindernumber  : chr  "four" "four" "six" "four" ...
##  $ enginesize      : int  130 130 152 109 136 136 136 136 131 131 ...
##  $ fuelsystem      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
##  $ 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 ...
```

let’s make some factors with this vaariables; and make eliminating the car_ID and Names;

```
# Changing discretes column as a factor.

data$fueltype <- as.factor(data$fueltype)
data$aspiration <- as.factor(data$aspiration)
data$doornumber <- as.factor(data$doornumber)
data$carbody <- as.factor(data$carbody)
data$drivewheel <- as.factor(data$drivewheel)
data$enginelocation <- as.factor(data$enginelocation)
data$enginetype <- as.factor(data$enginetype)
data$fuelsystem <- as.factor(data$fuelsystem)
data$cylindernumber <- as.factor(data$cylindernumber)

# Dropping columns with no predictive value

data$car_ID <- NULL
data$CarName <- NULL

# Transforming feature cylindernumber from text to its numeric equivalent

levels(data$cylindernumber) <- c("8","5","4","6","3","12","2")
data$cylindernumber <-   as.numeric(as.character (data$cylindernumber))
```

Here is the summary of the dataset:

```
summary(data)

##    symboling         fueltype   aspiration  doornumber       carbody
##  Min.   :-2.0000   diesel: 20   std  :168   four:115   convertible: 6
##  1st Qu.: 0.0000   gas   :185   turbo: 37   two : 90   hardtop    : 8
##  Median : 1.0000                                       hatchback  :70
##  Mean   : 0.8341                                       sedan      :96
##  3rd Qu.: 2.0000                                       wagon      :25
##  Max.   : 3.0000
##
##  drivewheel enginelocation   wheelbase        carlength        carwidth
##  4wd:  9    front:202      Min.   : 86.60   Min.   :141.1   Min.   :60.30
##  fwd:120    rear :  3      1st Qu.: 94.50   1st Qu.:166.3   1st Qu.:64.10
##  rwd: 76                   Median : 97.00   Median :173.2   Median :65.50
##                            Mean   : 98.76   Mean   :174.0   Mean   :65.91
##                              3rd Qu.:102.40   3rd Qu.:183.1   3rd Qu.:66.90  
##                              Max.   :120.90   Max.   :208.1   Max.   :72.30  
##                                                                              
##    carheight       curbweight   enginetype  cylindernumber    enginesize   
##  Min.   :47.80   Min.   :1488   dohc : 12   Min.   : 2.00   Min.   : 61.0  
##  1st Qu.:52.00   1st Qu.:2145   dohcv:  1   1st Qu.: 4.00   1st Qu.: 97.0  
##  Median :54.10   Median :2414   l    : 12   Median : 4.00   Median :120.0  
##  Mean   :53.72   Mean   :2556   ohc  :148   Mean   : 4.38   Mean   :126.9  
##  3rd Qu.:55.50   3rd Qu.:2935   ohcf : 15   3rd Qu.: 4.00   3rd Qu.:141.0  
##  Max.   :59.80   Max.   :4066   ohcv : 13   Max.   :12.00   Max.   :326.0  
##                                 rotor:  4                                  
##    fuelsystem   boreratio        stroke      compressionratio   horsepower  
##  mpfi   :94   Min.   :2.54   Min.   :2.070   Min.   : 7.00    Min.   : 48.0  
##  2bbl   :66   1st Qu.:3.15   1st Qu.:3.110   1st Qu.: 8.60    1st Qu.: 70.0  
##  idi    :20   Median :3.31   Median :3.290   Median : 9.00    Median : 95.0  
##  1bbl   :11   Mean   :3.33   Mean   :3.255   Mean   :10.14    Mean   :104.1  
##  spdi   : 9   3rd Qu.:3.58   3rd Qu.:3.410   3rd Qu.: 9.40    3rd Qu.:116.0  
##  4bbl   : 3   Max.   :3.94   Max.   :4.170   Max.   :23.00    Max.   :288.0  
##  (Other): 2                                                                  
##     peakrpm        citympg        highwaympg        price      
##  Min.   :4150   Min.   :13.00   Min.   :16.00   Min.   : 5118  
##  1st Qu.:4800   1st Qu.:19.00   1st Qu.:25.00   1st Qu.: 7788  
##  Median :5200   Median :24.00   Median :30.00   Median :10295  
##  Mean   :5125   Mean   :25.22   Mean   :30.75   Mean   :13277  
##  3rd Qu.:5500   3rd Qu.:30.00   3rd Qu.:34.00   3rd Qu.:16503  
##  Max.   :6600   Max.   :49.00   Max.   :54.00   Max.   :45400  
## 
```

Let check if there exist any missing values in the data set:

```
# Missing Value

lapply(data,function(x) { sum(is.na(x))})
```

```
## $symboling
## [1] 0
## 
## $fueltype
## [1] 0
## 
## $aspiration
## [1] 0
## 
## $doornumber
## [1] 0
## 
## $carbody
## [1] 0
## 
## $drivewheel
## [1] 0
##
## $enginelocation
## [1] 0
##
## $wheelbase
## [1] 0
##
## $carlength
## [1] 0
##
## $carwidth
## [1] 0
##
## $carheight
## [1] 0
##
## $curbweight
## [1] 0
##
## $enginetype
## [1] 0
##
## $cylindernumber
## [1] 0
##
## $enginesize
## [1] 0
##
## $fuelsystem
## [1] 0
##
## $boreratio
## [1] 0
##
## $stroke
## [1] 0
##
## $compressionratio
## [1] 0
##
## $horsepower
## [1] 0
##
## $peakrpm
## [1] 0
##
## $citympg
## [1] 0
##
```

```
## $highwaympg
## [1] 0
## 
## $price
## [1] 0
```

There are 205 observations and 24 variables in the dataset, 8 categorical and 16 continuous and no missing values. The Variable “price” is our dependent numerical variable and objective of study. = Main Objective =

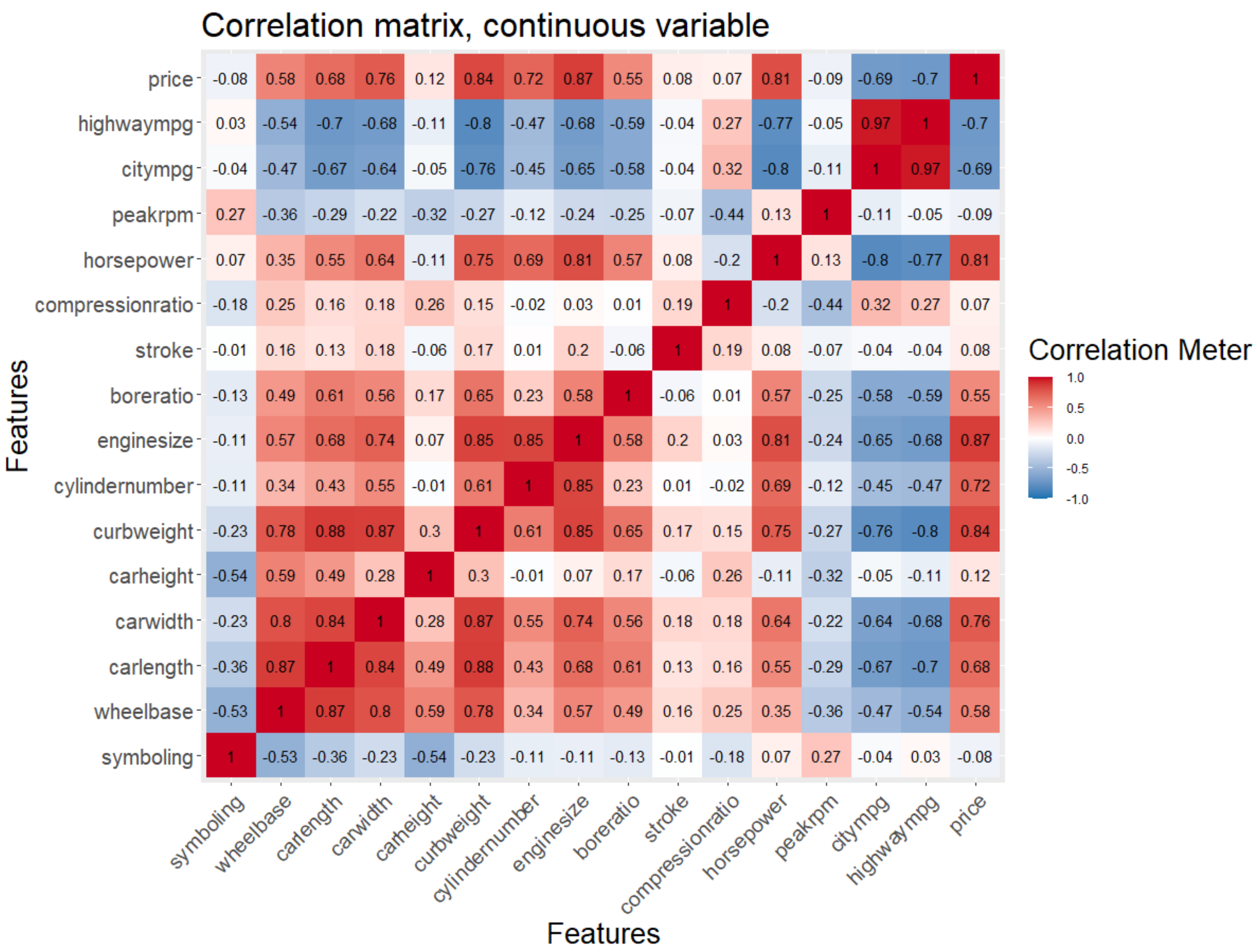
Perform a multiple regression model in order to know which explanatory variables of the dataset have a linear influence in our target variable “price”, a predictive multivariate model will be built with a backward approach. (backwards)

#Exploratory Data Analysis: ***This EDA will focus on finding a linear relationship between each feature and the dependent variable price of the vehicle.*** Quick look at what the dataset looks like:

```
#EDA
head(data,8)

##   symboling fueltype aspiration doornumber     carbody drivewheel
## 1         3      gas        std        two convertible        rwd
## 2         3      gas        std        two convertible        rwd
## 3         1      gas        std        two   hatchback        rwd
## 4         2      gas        std       four       sedan        fwd
## 5         2      gas        std       four       sedan        4wd
## 6         2      gas        std        two       sedan        fwd
## 7         1      gas        std       four       sedan        fwd
## 8         1      gas        std       four       wagon        fwd
##   enginelocation wheelbase carlength carwidth carheight curbweight enginetype
## 1          front      88.6     168.8     64.1      48.8       2548       dohc
## 2          front      88.6     168.8     64.1      48.8       2548       dohc
## 3          front      94.5     171.2     65.5      52.4       2823       ohcv
## 4          front      99.8     176.6     66.2      54.3       2337        ohc
## 5          front      99.4     176.6     66.4      54.3       2824        ohc
## 6          front      99.8     177.3     66.3      53.1       2507        ohc
## 7          front     105.8     192.7     71.4      55.7       2844        ohc
## 8          front     105.8     192.7     71.4      55.7       2954        ohc
##   cylindernumber enginesize fuelsystem boreratio stroke compressionratio
## 1              4        130       mpfi      3.47   2.68              9.0
## 2              4        130       mpfi      3.47   2.68              9.0
## 3              6        152       mpfi      2.68   3.47              9.0
## 4              4        109       mpfi      3.19   3.40             10.0
## 5              5        136       mpfi      3.19   3.40              8.0
## 6              5        136       mpfi      3.19   3.40              8.5
## 7              5        136       mpfi      3.19   3.40              8.5
## 8              5        136       mpfi      3.19   3.40              8.5
##   horsepower peakrpm citympg highwaympg price
## 1         111     5000       21          27 13495
## 2         111     5000       21          27 16500
## 3         154     5000       19          26 16500
## 4         102     5500       24          30 13950
## 5         115     5500       18          22 17450
## 6         110     5500       19          25 15250
## 7         110     5500       19          25 17710
## 8         110     5500       19          25 18920
```

Continuous variables correlation Next step is to study the relationship between the continuous variables in order to know, which of them will be the best predictors or even to detect signs of collinearity.

```
# Fig Size funtion

fig <- function(width, heigth){
  options(repr.plot.width = width, repr.plot.height = heigth)
}

# Correlation matrix, continuous variable

fig(14, 12)

plot_correlation(
  data,
  type = "c",
  cor_args = list("use" = "pairwise.complete.obs"),
  title = "Correlation matrix, continuous variable",
  theme_config = list(title = element_text(size=20),
                      axis.text.y = element_text(size = 15),
                      axis.text.x = element_text(hjust = 1, angle = 45,size = 15)))
```



A high correlation (collinearity) is observed in 1 pair of variables:

```
# Collinearity check
cor(data$citympg, data$highwaympg, use = "complete.obs")

## [1] 0.971337
```

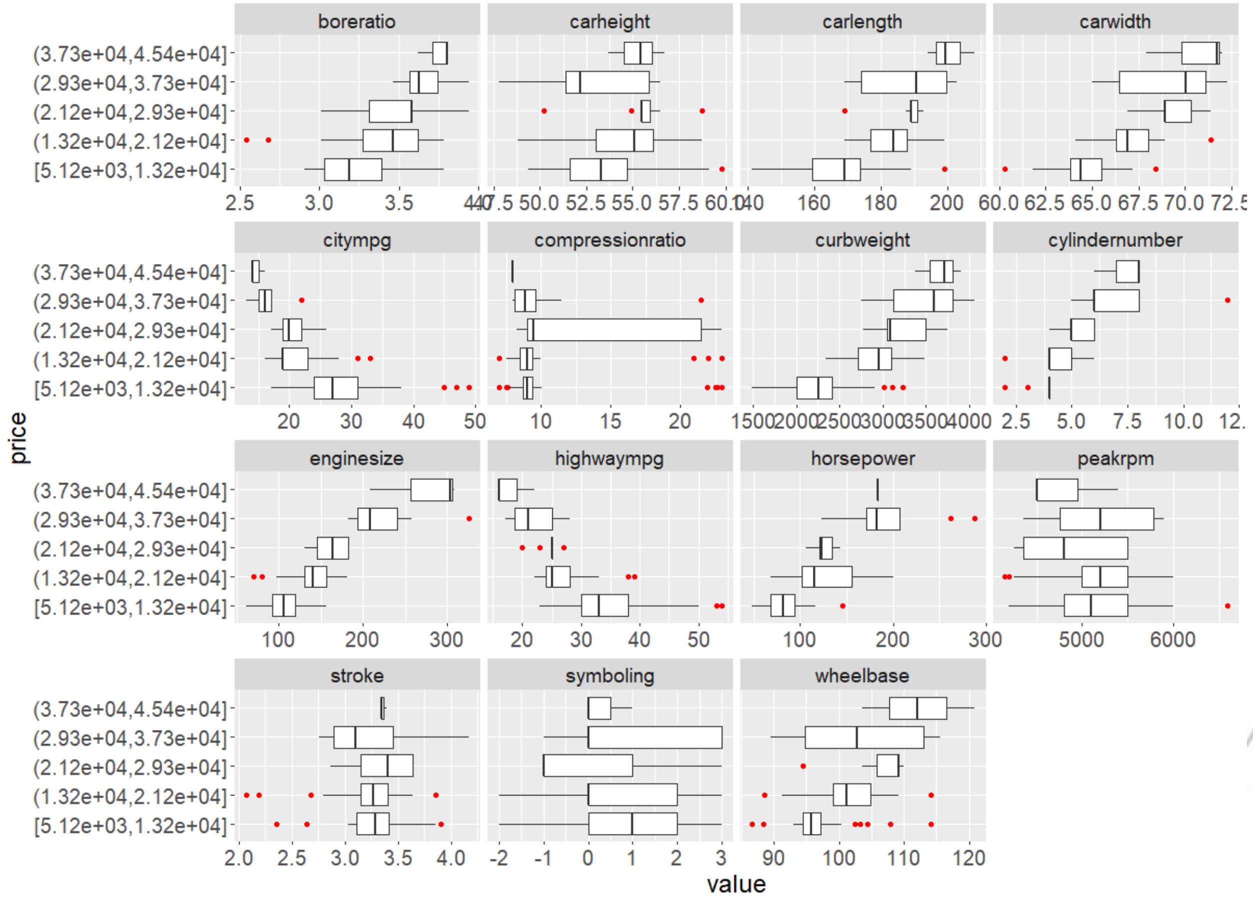
There is a high correlation between “citympg” - “highwaympg”, indicating a possible collinearity problem, it will probably not be useful to introduce both pairs in the model.

```
# Boxplot funtion by "price"
fig(18,16)
plot_boxplot(data, by = "price", nrow = 5L,
             geom_boxplot_args = list("outlier.color" = "red"),
             theme_config = list(text = element_text(size=18),
                                 axis.text.y = element_text(size = 15),
                                 axis.text.x = element_text(size = 15))
)
```

The features: boreratio, carlength, carwidth, curbweight, cylindernumber, enginesize, horsepower and wheelbase, have a significant and positive relationship with the price of the vehicle, a clear indication that a larger vehicle with more horsepower obviously it has a higher cost.

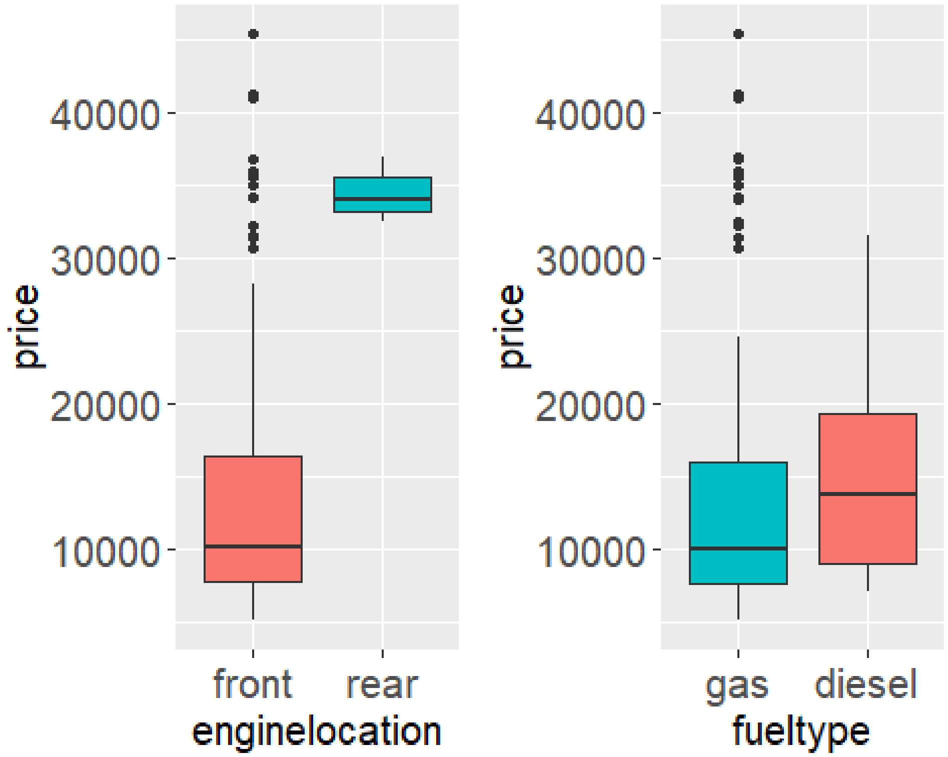
There are other features that have no or little influence on the price of the vehicle such as carheight, compressionration, peakrpm, stroke and symboling.

The citympg and highwaympg variables are negatively related to the Vehicle value. The most expensive vehicles have lower performance in fuel consumption.



Distribution of discretes variables in relation to the dependent variable

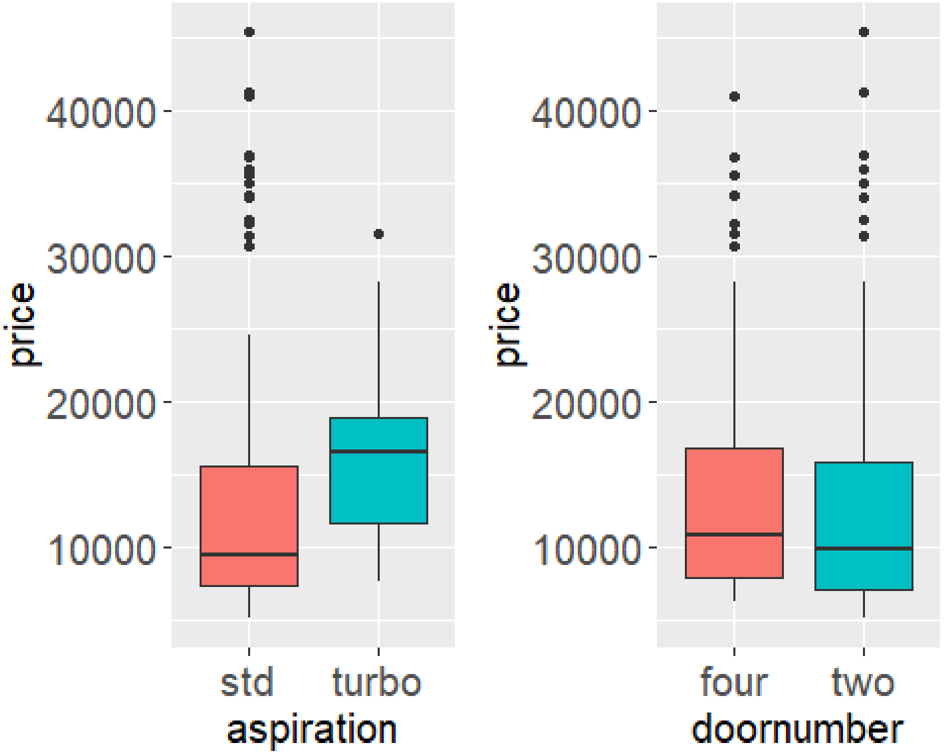
```
fig(17,8)

pl1 <- ggplot(data) +
  aes(x = reorder(enginelocation,price), y = price, fill = enginelocation) +
  geom_boxplot() +
  scale_fill_hue() +
  labs(x = "enginelocation") +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))


pl2 <- ggplot(data) +
  aes(x = reorder(fueltype, price), y = price, fill = fueltype) +
  geom_boxplot() +
  scale_fill_hue() +
  theme_gray() +
  labs(x = "fueltype") +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



There is a significant relationship between the price and the location of the engine, there seems to be no relationship with the type of fuel.

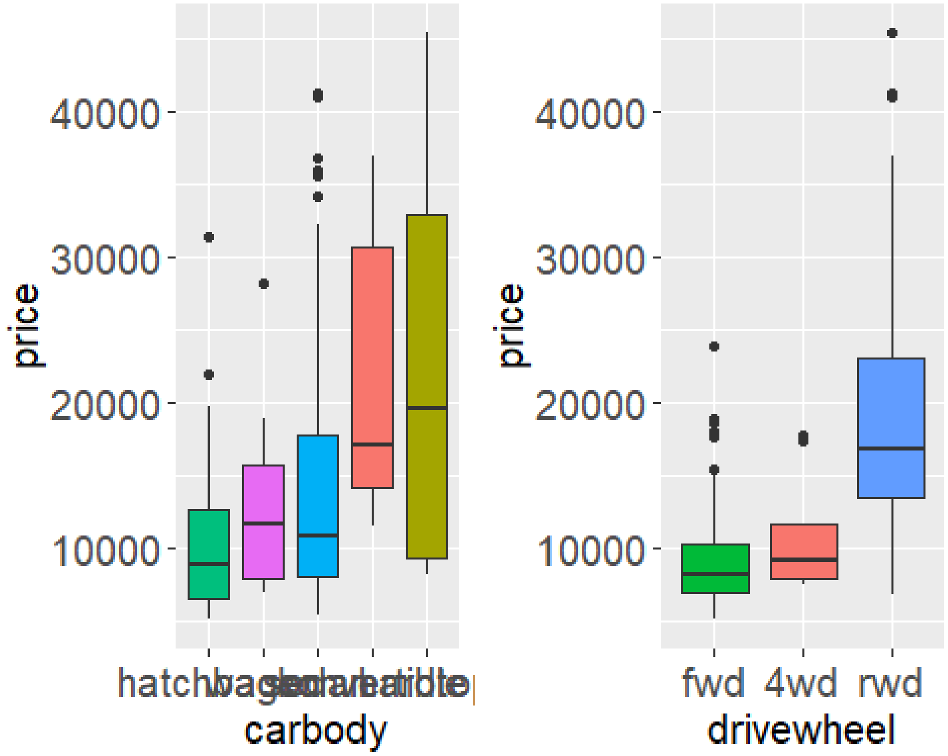
```
pl1 <- ggplot(data) +
  aes(x = aspiration, y = price, fill = aspiration) +
  geom_boxplot() +
  scale_fill_hue() +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))
```

```
pl2 <- ggplot(data) +
  aes(x = doornumber, y = price, fill = doornumber) +
  geom_boxplot() +
  scale_fill_hue() +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



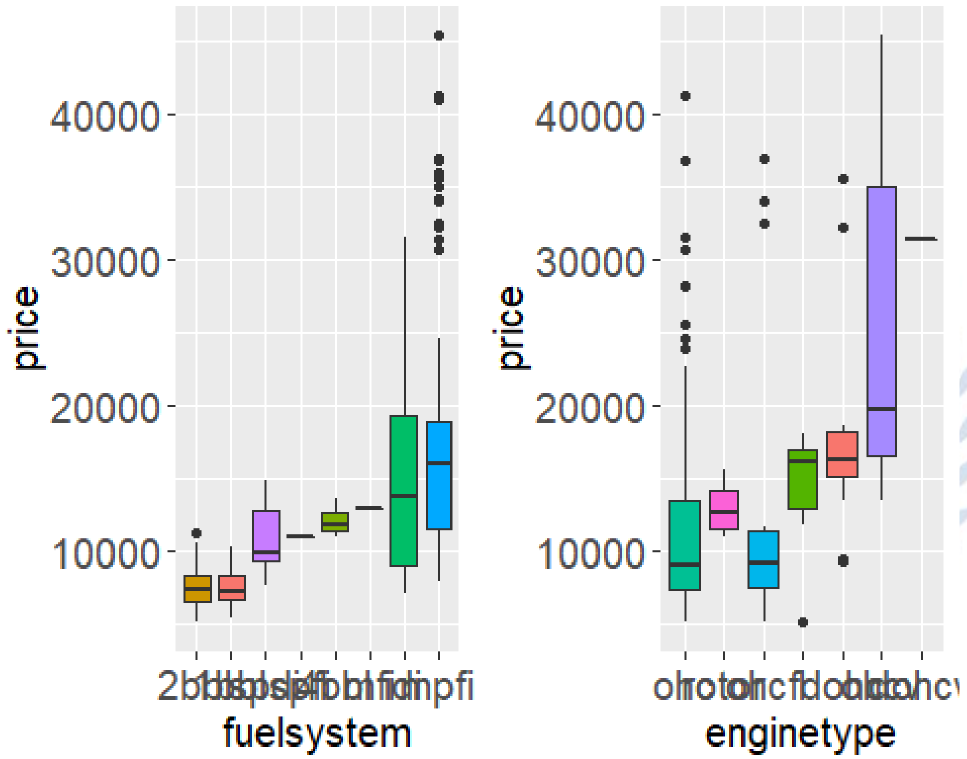
There does not seem to be a significant relationship between the price and the type of engine intake, or with the number of doors.

```
  pl1 <- ggplot(data) +
  aes(x = reorder(carbody,price), y = price, fill = carbody) +
  geom_boxplot() +
  scale_fill_hue() +
  labs(x = "carbody") +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))


pl2 <- ggplot(data) +
  aes(x = reorder(drivewheel,price), y = price, fill = drivewheel) +
  geom_boxplot() +
  scale_fill_hue() +
  labs(x = "drivewheel") +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



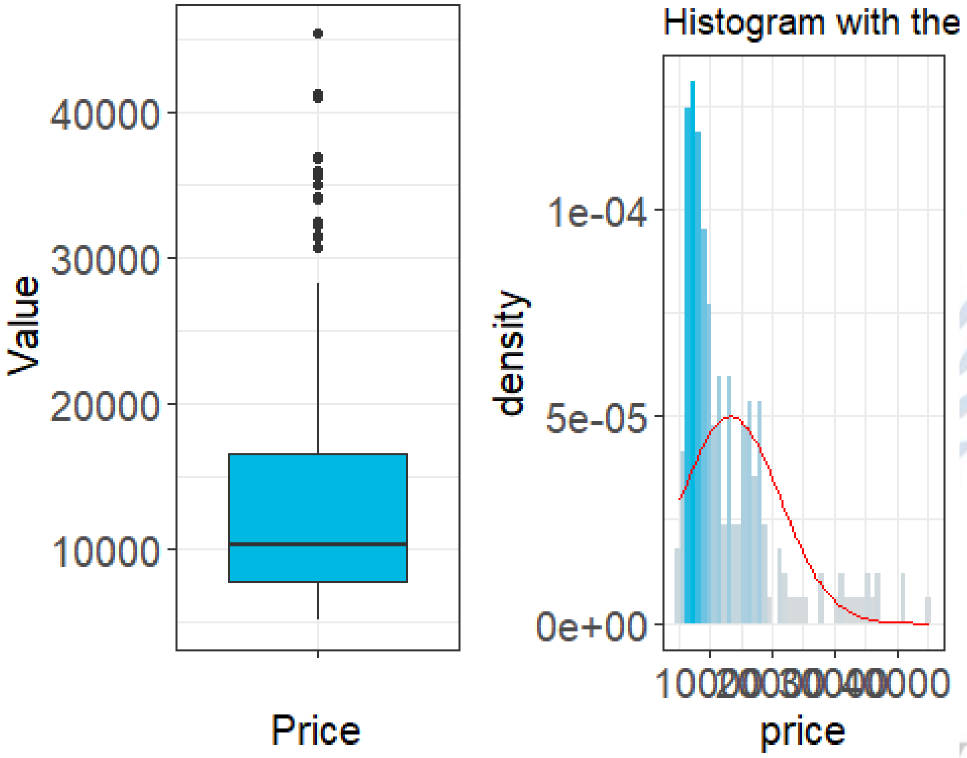
There is a significant relationship between the price and the shape of the vehicle, as well as the type of wheel drive.

```
pl1 <- ggplot(data) +
  aes(x = reorder(fuelsystem,price), y = price, fill = fuelsystem) +
  geom_boxplot() +
  scale_fill_hue() +
  labs(x = "fuelsystem")  +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

pl2 <- ggplot(data) +
  aes(x = reorder(enginetype,price), y = price, fill = enginetype) +
  geom_boxplot() +
  scale_fill_hue() +
  labs(x = "enginetype") +
  theme_gray() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))


grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



There is a significant relationship between the price and the form or type of fuel injection to the engine, as well as the type of engine. #Analyzing Independent variable Vehicle Price

```
#Boxplot of price

pl1 <- ggplot(data) +
  aes(x = "", y = price) +
  geom_boxplot(fill = "#00B9E3") +
  labs(x = "Price", y ="Value") +
  theme_bw() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))
#Histogram of price

pl2 <- ggplot(data = data, aes(x = price)) +
  geom_histogram(bins = 50,aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#00B9E3") +
  stat_function(fun = dnorm, colour = "red",
                args = list(mean = mean(data$price),
                            sd = sd(data$price))) +
  ggtitle("Histogram with theorical normal dist. curve") +
  theme_bw() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



```
summary(data$price)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5118    7788   10295   13277   16503   45400
```

The histogram shows a large skew of the data to the right, which means that most of the prices in the dataset are low. There is a significant difference between the mean and the median. The price distribution is widely separated from the average, indicating a high variation (75% of prices are below 16,500, while the remaining 15% are between 16,500 and 45,400).

#Insights found so Far:

- A high correlation is observed in the pair of features: “citympg” - “highwaympg”, possibly indicating a collinearity problem.

- The features: boreratio, carlength, carwidth, curbweight, cylindernumber, enginesize, horsepower and wheelbase, have a significant and positive relationship with the price of the vehicle.

- There are features with little or no influence on the price of the vehicle such as carheight, compressionration, peakrpm, stroke and symboling.

- Although the city-mpg and highwaympg variables are negatively related to the value of the Vehicle, they are not independent variables that influence its price, instead, they are dependent variables of the same fatures of the dataset that increase the price of the vehicle. That is why larger, heavier and more powerfull vehicles, in addition to having a higher cost, have lower fuel efficiency, therefore these variables are not considered independent themselves and will be discarded.

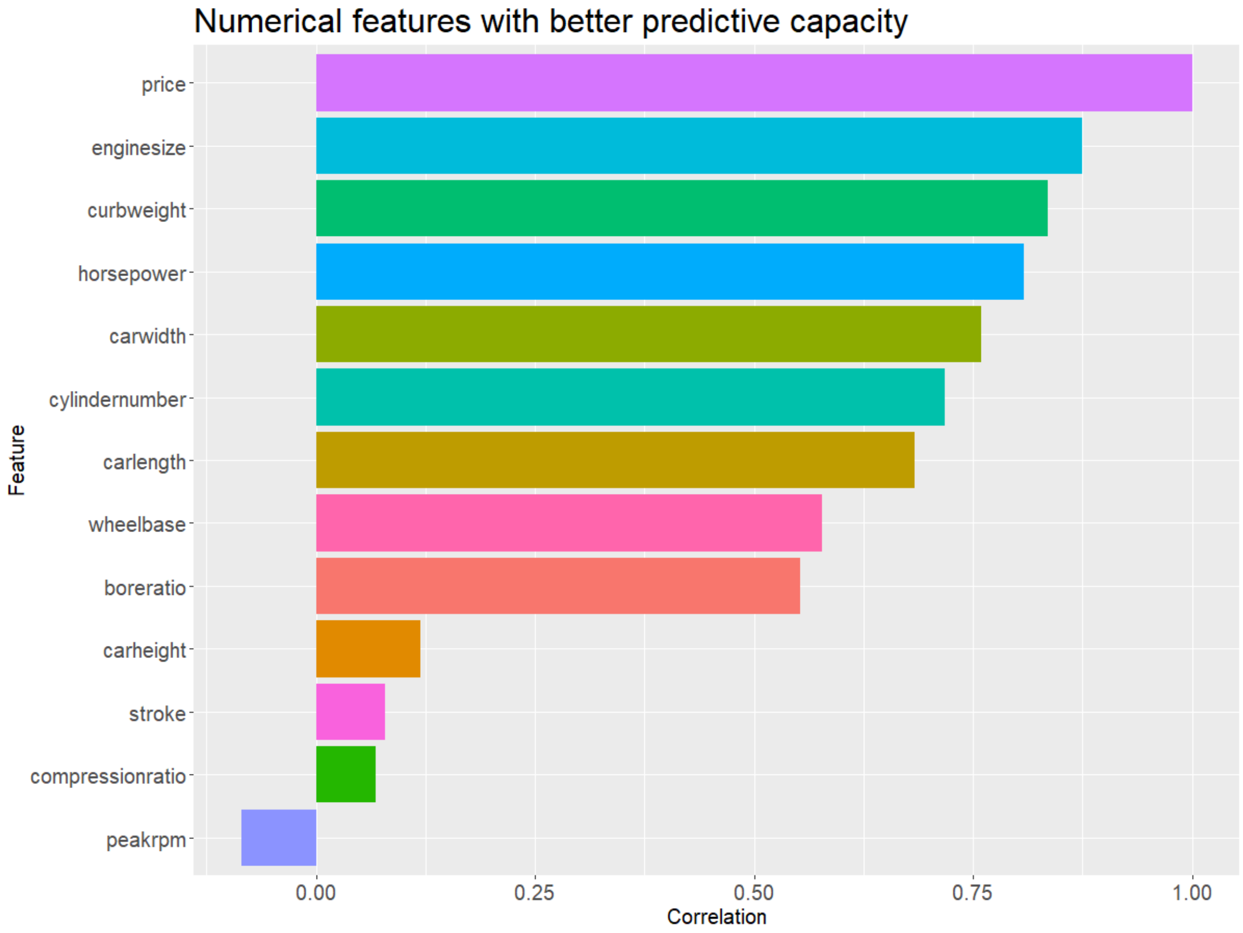
- The high skew to the right and outliers in the price variable, indicates some type of transformation is necessary to adjust it to a normal distribution.

- There is a significant relationship between the features “enginelocation”, “carbody”, “drivewheel”, “fuelsystem”, “enginetype” and price of vehicle.

- There is not any relationship between the features “fueltype”, “aspiration”, “doornumber” and price of vehicle.

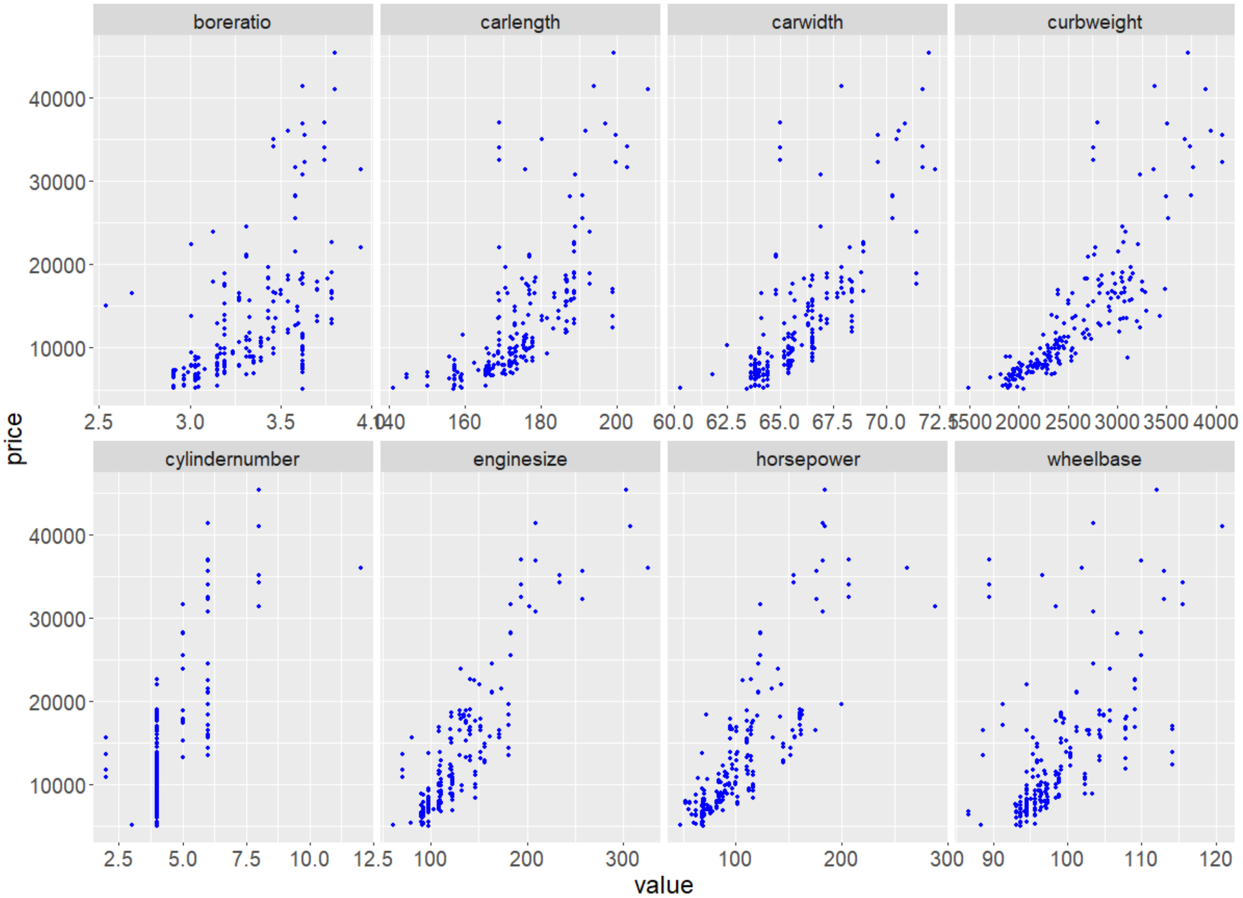
- Bar plot of numerical variables with the best predictive capacity based on their correlation, they will to be used in the multivariate linear model.

```
# Positive and negatively influencing features in the price variable

df <-
  subset(
    data,
    select = c(
      "wheelbase","carlength","carwidth", "carheight", "curbweight", "cylindernumbe
r", "enginesize",
      "boreratio", "stroke", "compressionratio", "horsepower", "peakrpm", "price"
    )
  )

corr_car_data <-
  as.data.frame(t(cor(data$price, df, method = "pearson")))

corr_car_data$key <- rownames(corr_car_data)

ggplot(data = corr_car_data) +
  aes(x = reorder(key, V1),
      y = V1,
      fill = key) +
  geom_bar(stat = "identity") +
  ylab("Correlation") +
  xlab("Feature") +
  ggtitle("Numerical features with better predictive capacity") +
  theme(legend.position = "none",
        title = element_text(size = 20),
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))  +
  coord_flip()
```



Scatter plot of continuous features with best prediction the price of the vehicle.

```
# Scatter plot with better prediction of price value
fig(19,10)
df <-
  subset(
    data,
    select = c(
      "wheelbase","carlength","carwidth", "curbweight", "cylindernumber", "enginesi
ze",
      "boreratio", "horsepower", "price"
    )
  )

plot_scatterplot(df, by ="price", ncol = 4L, nrow = 2L,
                 geom_point_args = list("stroke" = 0.1, "colour" = "blue"),
                 theme_config = list(text = element_text(size=18),
                                     axis.text.y = element_text(size = 15),
                                     axis.text.x = element_text(size = 15)))
```



Experimental Design We will intend to verify if a multivariate linear regression can predict the price of vehicles, based on the selected or most favorable features of the dataset studied according to the model:

Where
x_{p}are the explanatory variables that best predict the price

Transformation of the dependent variable "price".

Determine favorable categorical variables through a single-factor design with fixed effects.

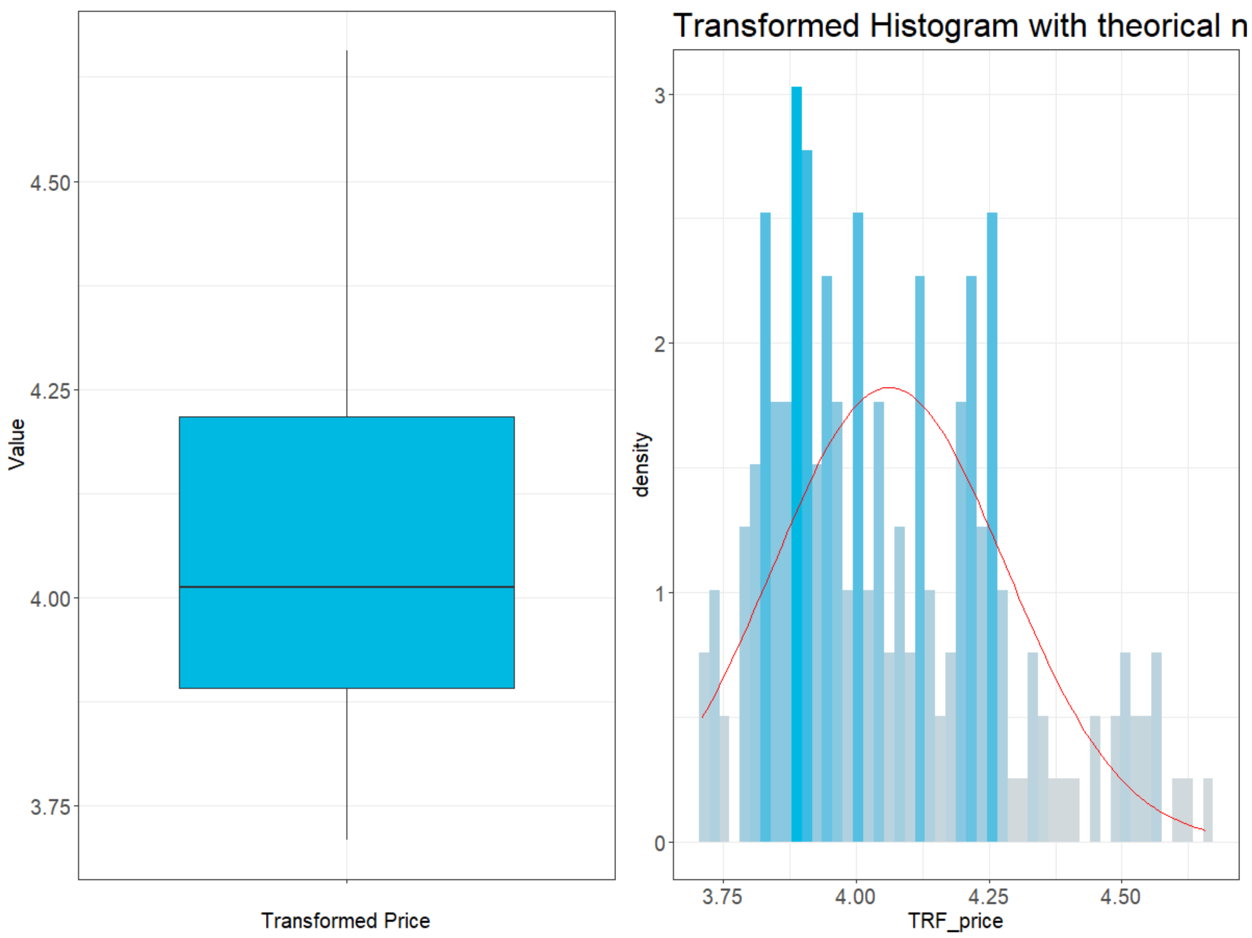
Using simple regression, determine which continuous variables have predictive capacity on the price of the vehicle.

Using multiple regression, determine the best linear equation that predicts the price of vehicles, using the selected features.

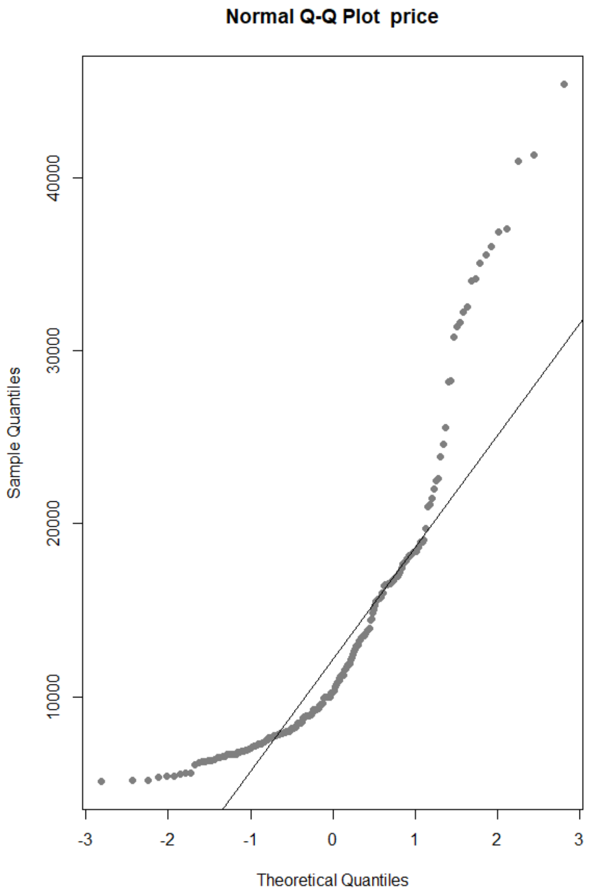
Diagnosis and validation of the linear model obtained.

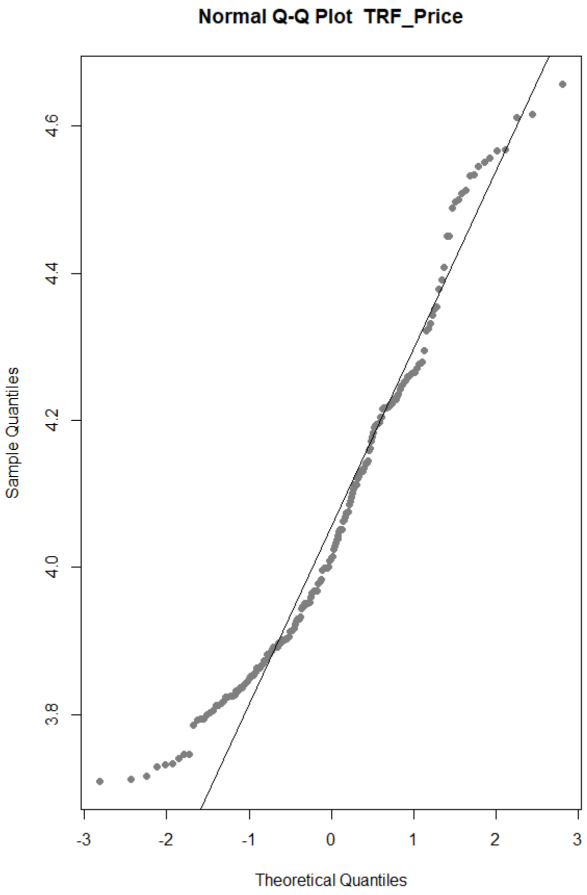
Transformation of the dependent variable “price”

```
# Logaritmic transformation of the variable price

data$TRF_price <- log10(data$price)

#Boxplot

pl1 <- ggplot(data) +
  aes(x = "", y = TRF_price) +
  geom_boxplot(fill = "#00B9E3") +
  labs(x = "Transformed Price", y ="Value") +
  theme_bw() +
  theme(legend.position = "none",
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

#Histogram

pl2 <- ggplot(data = data, aes(x = TRF_price)) +
  geom_histogram(bins = 50,aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#00B9E3") +
  stat_function(fun = dnorm, colour = "red",
                args = list(mean = mean(data$TRF_price),
                            sd = sd(data$TRF_price))) +
  ggtitle("Transformed Histogram with theorical normal dist. curve") +
  theme_bw() +
  theme(legend.position = "none",
        title = element_text(size = 20),
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

grid.arrange(ncol = 2, nrow = 1,pl1,pl2)
```



```
# Q-Q plot comparison

par(mfrow=c(1,2))

qqnorm(data$price, pch = 19, col = "gray50",main = "Normal Q-Q Plot  price")
qqline(data$price)

qqnorm(data$TRF_price, pch = 19, col = "gray50", main = "Normal Q-Q Plot  TRF_Price
")
qqline(data$TRF_price)
```





Although the vehicle price variable after the transformation does not follow a normal distribution, it is much closer and the outliers have been eliminated, with this simple logarithmic transformation.

Single-factor design with fixed effects Analysis of variance (ANOVA) will be applied to all the categorical variables of the dataset with levels greater than 2 (K > 2), to determine their predictive capacity.

We are interested to know if the means of the vehicle´s price are the same in each levels or treatments of each discrete variable (or factor), according to the following hypothesis test:

That is, H0: contrasts that the variable cannot predict a clear trend in the price of the vehicle, because there are no differences in the means of the levels, compared to the alternative H1: that at least one mean differs from the other.

The variable of interest or *dependent variable*, is the price of the vehicle. The factor: The *categorical variables* of the dataset with more than 2 levels. Factor levels: all treatments or levels within each *discrete variable*. Model *unbalanced*: the levels do not have the same number of elements.

T-Student test. Categorical variables with only 2 levels: enginelocation, fueltype, aspiration, doornumber.

```
# Enginelocation

t1 <- data %>% filter(enginelocation == "front")
t2 <- data %>% filter(enginelocation == "rear")

t.test(t1$TRF_price,t2$TRF_price)

##
##  Welch Two Sample t-test
##
## data:  t1$TRF_price and t2$TRF_price
## t = -21.647, df = 6.5946, p-value = 2.254e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5352133 -0.4286030
## sample estimates:
## mean of x mean of y
##  4.055633  4.537541
```

The p-value is <0.05 and the 95% confidence interval does not include zero, we can affirm that the samples differ in their means, the variable “enginelocation” has predictive capacity.

```
# fueltype

t1 <- data %>% filter(fueltype == "gas")
t2 <- data %>% filter(fueltype == "diesel")

t.test(t1$TRF_price,t2$TRF_price)

##
##  Welch Two Sample t-test
##
## data:  t1$TRF_price and t2$TRF_price
## t = -1.9503, df = 23.561, p-value = 0.06314
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.200921413  0.005784993
## sample estimates:
## mean of x mean of y
##  4.053167  4.150735
```

The p-value is >0.05 and the 95% confidence interval does include zero, we can not confirm the samples differ in their means, the variable “fueltype” has not predictive capacity and it is discarded.

```
# aspiration

t1 <- data %>% filter(aspiration == "std")
t2 <- data %>% filter(aspiration == "turbo")

t.test(t1$TRF_price,t2$TRF_price)
```

```
##
##  Welch Two Sample t-test
##
## data:  t1$TRF_price and t2$TRF_price
## t = -4.5705, df = 67.933, p-value = 2.115e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.20989172 -0.08231285
## sample estimates:
## mean of x mean of y
##  4.036316  4.182418
```

The p-value is <0.05 and the 95% confidence interval does not include zero, we can affirm that the samples differ in their means, the variable “aspiration” has predictive capacity.

```
# doornumber

t1 <- data %>% filter(doornumber == "four")
t2 <- data %>% filter(doornumber == "two")

t.test(t1$TRF_price,t2$TRF_price)

##
##  Welch Two Sample t-test
##
## data:  t1$TRF_price and t2$TRF_price
## t = 1.215, df = 174.92, p-value = 0.226
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02378436  0.09997579
## sample estimates:
## mean of x mean of y
##  4.079410  4.041315
```

The p-value is >0.05 and the 95% confidence interval does include zero, we can not confirm the samples differ in their means, the variable “doornumber” has not predictive capacity and it is discarded.

ANOVA test Categorical features with 2 or more levels: carbody, drivewheel, enginetype, fuelsystem.

```
# fuelsystem

anova <- aov(data$TRF_price ~ data$fuelsystem)
summary(anova)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## data$fuelsystem   7  5.048  0.7212   30.12 <2e-16 ***
## Residuals       197  4.718  0.0239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verified the p-value is very small, in addition, the contrast statistic F is greater than 1, therefore, there is sufficient evidence to conclude that the means in the price of the vehicle are not equal (or at least in one of the them), for the different levels of the variable "fuelsystem".

And, consequently, the type of injection of a vehicle can influence the price of the car.

```
# enginetype

anova <- aov(data$TRF_price ~ data$enginetype)
summary(anova)

##                    Df Sum Sq Mean Sq F value   Pr(>F)    
## data$enginetype   6  2.090  0.3483   8.983 1.11e-08 ***
## Residuals       198  7.676  0.0388                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verified the p-value is very small, in addition, the contrast statistic F is greater than 1, therefore, there is sufficient evidence to conclude that the means in the price of the vehicle are not equal (or at least in one of the them), for the different levels of the variable "enginetype".

And, consequently, the engine type of a vehicle can influence the price of the car.

```
# drivewheel

anova <- aov(data$TRF_price ~ data$drivewheel)
summary(anova)

##                    Df Sum Sq Mean Sq F value Pr(>F)    
## data$drivewheel   2  4.623  2.3116    90.8 <2e-16 ***
## Residuals       202  5.143  0.0255                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verified the p-value is very small, in addition, the contrast statistic F is greater than 1, therefore, there is sufficient evidence to conclude that the means in the price of the vehicle are not equal (or at least in one of the them), for the different levels of the variable "drivewheel".

And, consequently, the traction type of a vehicle can influence the price of the car.

```
# carbody

anova <- aov(TRF_price ~ carbody, data = data)
summary(anova)

##                Df Sum Sq Mean Sq F value   Pr(>F)    
## carbody       4  1.251 0.31270   7.344 1.54e-05 ***
## Residuals   200  8.515 0.04258                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verified the p-value is very small, in addition, the contrast statistic F is greater than 1, therefore, there is sufficient evidence to conclude that the means in the price of the vehicle are not equal (or at least in one of the them), for the different levels of the variable “carbody”.

And, consequently, the shape of a vehicle can influence the price of the car.

Statistically significant categorical variables in predicting the price of the vehicle: “enginelocation”,“aspiration”, “fuelsystem”, “enginetype”, “carbody”, “drivewheel”

Simple Linear Regression:

Simple linear regression will be applied to all continuous variables in the dataset in order to determine their predictive capability.

A regression model is generated to explain the linear relationship between the value of the vehicle and each continuous feature present in the dataset.

The main objective is to perform a hypothesis test on the slope of the regression line beta to know if there is a linear relationship between the feature under test and the price of the vehicle. Null Hypothesis: In the population from sample, there is no linear relationship between the price of the vehicle and the study feature. Altenative Hypothesis: In the population from sample, there is a linear relationship between the price of the vehicle and the study feature.

If the null hypothesis is rejected, the feature under test is linearly related to the dependent variable price of the vehicle.

To quantify the magnitude of the association between price and the test variable, the 95% confidence interval for the slope of the regression line is revised.

Continuous features: “symboling”, “wheelbase”, “carlength”, “carwidth”, “carheight”, “curbweight”, “cylindernumber”, “enginesize”, “boreratio”, “stroke”, “compressionratio”, “horsepower”, “peakrpm”.

```
# Continuous variables under test

nombre_col <- c("symboling",  "wheelbase",
                "carlength",  "carwidth",
                "carheight",  "curbweight",
                "cylindernumber", "enginesize",
                "boreratio",  "stroke",
                "compressionratio", "horsepower", "peakrpm" )

# empty dataframe

df <-
  structure(
    list(
      Variable = character(),
      Intercept = numeric(),
      beta = numeric(),
      t_value = numeric(),
      p_value = numeric(),
      IC_2.5 = numeric(),
      IC_95 = numeric(),
      R_squared = numeric()
    ),
    class = "data.frame"
  )
# Extraction and tabulation of the most important values of the simple regression o
f each continuous variable.

for (i in nombre_col) {

  regresion <- lm(data$TRF_price ~ data[[i]])
  sumario <- summary(regresion)
  confianza <-  confint(regresion, level = 0.95)

  vector <-
    data.frame(
      Variable = i,
      Intercept = round(sumario$coefficients[1], 2), beta = round(sumario$coefficie
nts[2], 4),
      t_value = round(sumario$coefficients[6], 2), p_value = sumario$coefficients[8
],
      IC_2.5 = round(confianza[2, 1], 4), IC_95 = round(confianza[2, 2], 4),
      R_squared = sumario$r.squared
    )
  df <- rbind(df,vector)
}

# Features with no lineal relationship

df %>% filter(p_value > 0.05)

##            Variable Intercept    beta t_value   p_value  IC_2.5  IC_95
## 1         symboling      4.08 -0.0156   -1.27 0.2058845 -0.0398 0.0086
## 2            stroke      3.84  0.0684    1.40 0.1621660 -0.0277 0.1645
## 3 compressionratio      4.02  0.0047    1.22 0.2248159 -0.0029 0.0123
## 4           peakrpm      4.27  0.0000   -1.27 0.2072374 -0.0001 0.0000
##     R_squared
## 1 0.007870711
## 2 0.009602493
## 3 0.007249427
## 4 0.007824113
```

With a contrast p-value greater than 0.05, and a confidence interval that includes zero, we can conclude that we do not have sufficient statistical evidence to reject the null hypothesis, that is:

It can be stated that there is no statistically significant linear relationship between the price of the vehicle and the variables:

“symboling”, “stroke”, “compressionratio”, “peakrpm”, so they will be discarded from the model.

```
# Features with lineal relationship

df %>% filter(p_value < 0.05)

##          Variable Intercept   beta t_value      p_value IC_2.5  IC_95  R_squared
## 1       wheelbase      1.80 0.0229   11.54 5.130555e-24 0.0190 0.0268 0.39614581
## 2       carlength      1.69 0.0136   17.08 3.979675e-41 0.0120 0.0152 0.58961498
## 3        carwidth     -1.33 0.0819   19.17 2.016454e-47 0.0734 0.0903 0.64407762
## 4       carheight      3.28 0.0146    2.35 1.968835e-02 0.0024 0.0268 0.02650307
## 5      curbweight      3.11 0.0004   28.00 1.229604e-71 0.0003 0.0004 0.79430295
## 6 cylindernumber      3.50 0.1279   11.61 3.128245e-24 0.1062 0.1496 0.39906083
## 7      enginesize      3.51 0.0044   21.37 7.692717e-54 0.0040 0.0048 0.69219853
## 8       boreratio      2.42 0.4932   10.98 2.487891e-22 0.4047 0.5818 0.37279022
## 9      horsepower      3.59 0.0046   20.87 2.116506e-52 0.0041 0.0050 0.68200371
```

With a contrast p-value less than 0.05, and a confidence interval that not includes zero, we can conclude that we do have sufficient statistical evidence to reject the null hypothesis, that is:

It can be stated that there is a statistically significant linear relationship between the price of the vehicle and the features:

"wheelbase", "carlength", "carwidth", "curbweight", "cylindernumber", "enginesize", "boreratio" and "horsepower".

The determination coefficient "R_squared" values of these variables are the highest, indicating these features of the dataset have the best explicability of the price of the vehicle.

```
# Variable Carheight

regresion <- lm(data$TRF_price ~ data$carheight)
summary(regresion)

## 
## Call:
## lm(formula = data$TRF_price ~ data$carheight)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.35322 -0.17921 -0.04039  0.14457  0.57999 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    3.279519   0.333482   9.834   <2e-16 ***
## data$carheight 0.014577   0.006201   2.351   0.0197 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2164 on 203 degrees of freedom
## Multiple R-squared:  0.0265, Adjusted R-squared:  0.02171 
## F-statistic: 5.527 on 1 and 203 DF,  p-value: 0.01969
```

```
confint(regresion, level = 0.95)

##                       2.5 %     97.5 %
## (Intercept)    2.621986703 3.93705092
## data$carheight 0.002351058 0.02680365
```

The variable “Carheight” is not statistically significant, and it will not be included in the final model, since it is at the limit of statistical significance (P-value = 0.02), in addition, the coefficient of determination R_square just explains 2.6% of the price.

#Multivariable Linear Regression We are going to identify which features of those already selected, are linearly associated with the variable price of the vehicle, now taking into account the simultaneous effect they have on each other.

We will use step method and direction backward.

```
# Remove discarded variable from model

desechadas <- c("symboling", "stroke", "carheight", "compressionratio", "peakrpm",
                "fueltype", "doornumber", "citympg", "highwaympg", "price")

# Data frame with significant variables for the model

CarPrice_predictoras <- data[ , !(names(data) %in% desechadas)]

# Linear model with accepted predictor variables

modelo <- lm(TRF_price ~., data = CarPrice_predictoras)
summary(modelo)

##
## Call:
## lm(formula = TRF_price ~ ., data = CarPrice_predictoras)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17421 -0.03736 -0.00143  0.03389  0.16219
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.341e+00  3.061e-01   7.648  1.3e-12 ***
## aspirationturbo      1.296e-02  2.298e-02   0.564 0.573508
## carbodyhardtop      -1.316e-01  3.981e-02  -3.306 0.001146 **
## carbodyhatchback    -1.262e-01  3.449e-02  -3.660 0.000334 ***
## carbodysedan        -9.642e-02  3.561e-02  -2.708 0.007446 **
## carbodywagon        -1.410e-01  3.849e-02  -3.664 0.000329 ***
## drivewheelfwd       -2.668e-02  2.961e-02  -0.901 0.368798
## drivewheelrwd        2.088e-02  3.183e-02   0.656 0.512716
## enginelocationrear   2.444e-01  6.474e-02   3.774 0.000219 ***
## wheelbase            1.715e-03  2.313e-03   0.742 0.459317
## carlength           7.037e-04  1.360e-03   0.517 0.605551
## carwidth            1.535e-02  6.037e-03   2.542 0.011886 *
## curbweight          1.466e-04  4.672e-05   3.137 0.002003 **
## enginetypedohcv    -6.376e-02  1.026e-01  -0.622 0.534975
## enginetypel        -1.567e-02  3.558e-02  -0.440 0.660122
## enginetypeohc       5.716e-02  2.494e-02   2.292 0.023086 *
## enginetypeohcf      4.008e-02  4.108e-02   0.976 0.330549
## enginetypeohcv     -2.563e-02  3.336e-02  -0.768 0.443446
## enginetyperotor     1.695e-01  7.806e-02   2.171 0.031299 *
## cylindernumber      9.440e-03  1.628e-02   0.580 0.562817
## enginesize          4.416e-04  5.797e-04   0.762 0.447232
## fuelsystem2bbl     -1.818e-02  2.311e-02  -0.787 0.432455
## fuelsystem4bbl     -1.380e-02  8.176e-02  -0.169 0.866155
## fuelsystemidi       2.493e-02  3.458e-02   0.721 0.471823
## fuelsystemmfi      -1.944e-02  7.384e-02  -0.263 0.792644
## fuelsystemmpfi      2.914e-02  2.627e-02   1.109 0.268897
## fuelsystemspdi     -3.128e-02  3.500e-02  -0.894 0.372787
## fuelsystemspfi     -9.613e-03  7.315e-02  -0.131 0.895600
## boreratio          -4.162e-02  4.370e-02  -0.952 0.342193
## horsepower          1.385e-03  5.393e-04   2.569 0.011037 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06618 on 175 degrees of freedom
## Multiple R-squared:  0.9215, Adjusted R-squared:  0.9085
## F-statistic: 70.86 on 29 and 175 DF,  p-value: < 2.2e-16
```

The model with all the variables selected as best predictors, has a high adjusted R_square (0.9085). It is able to explain 90.85% of the variability observed in the price of vehicles.

The p-value of the model is significant (<0.05), as well as, the contrast statistic F, therefore it can be accepted that the relationship is not random; at least one of the regression coefficients is nonzero.

The next step is to build a “backwards” model, to select the best set of predictors.

```
# Model Backward

modelo_backward <- step(object =modelo, direction = "backward", trace = 1)
```

```
## Start:  AIC=-1085.76
## TRF_price ~ aspiration + carbody + drivewheel + enginelocation +
##     wheelbase + carlength + carwidth + curbweight + enginetype +
##     cylindernumber + enginesize + fuelsystem + boreratio + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - carlength       1  0.001172 0.76757 -1087.5
## - aspiration      1  0.001393 0.76779 -1087.4
## - cylindernumber  1  0.001472 0.76787 -1087.4
## - wheelbase       1  0.002409 0.76881 -1087.1
## - enginesize      1  0.002541 0.76894 -1087.1
## - boreratio       1  0.003973 0.77037 -1086.7
## - fuelsystem      7  0.053391 0.81979 -1086.0
## <none>                        0.76640 -1085.8
## - drivewheel      2  0.030783 0.79718 -1081.7
## - carwidth        1  0.028301 0.79470 -1080.3
## - horsepower      1  0.028901 0.79530 -1080.2
## - curbweight      1  0.043096 0.80949 -1076.5
## - enginetype      6  0.100300 0.86670 -1072.5
## - enginelocation  1  0.062393 0.82879 -1071.7
## - carbody         4  0.113619 0.88002 -1065.4
##
## Step:  AIC=-1087.45
## TRF_price ~ aspiration + carbody + drivewheel + enginelocation +
##     wheelbase + carwidth + curbweight + enginetype + cylindernumber +
##     enginesize + fuelsystem + boreratio + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - aspiration      1  0.000914 0.76848 -1089.2
## - cylindernumber  1  0.001406 0.76898 -1089.1
## - enginesize      1  0.002064 0.76963 -1088.9
## - boreratio       1  0.003361 0.77093 -1088.5
## - wheelbase       1  0.004505 0.77207 -1088.2
## - fuelsystem      7  0.054038 0.82161 -1087.5
## <none>                        0.76757 -1087.5
## - drivewheel      2  0.029685 0.79725 -1083.7
## - horsepower      1  0.030112 0.79768 -1081.6
## - carwidth        1  0.032929 0.80050 -1080.8
## - curbweight      1  0.059435 0.82700 -1074.2
## - enginetype      6  0.103967 0.87154 -1073.4
## - enginelocation  1  0.063114 0.83068 -1073.2
## - carbody         4  0.114848 0.88242 -1066.9
##
## Step:  AIC=-1089.2
## TRF_price ~ carbody + drivewheel + enginelocation + wheelbase +
##     carwidth + curbweight + enginetype + cylindernumber + enginesize +
##     fuelsystem + boreratio + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - cylindernumber  1  0.001181 0.76966 -1090.9
## - enginesize      1  0.001411 0.76989 -1090.8
## - boreratio       1  0.003489 0.77197 -1090.3
## - wheelbase       1  0.005248 0.77373 -1089.8
## - fuelsystem      7  0.053129 0.82161 -1089.5
## <none>                        0.76848 -1089.2
## - drivewheel      2  0.029062 0.79754 -1085.6
## - carwidth        1  0.033380 0.80186 -1082.5
## - horsepower      1  0.057923 0.82641 -1076.3
## - curbweight      1  0.059964 0.82845 -1075.8
## - enginetype      6  0.103735 0.87222 -1075.2
## - enginelocation  1  0.064129 0.83261 -1074.8
## - carbody         4  0.116240 0.88472 -1068.3
```

```
##
## Step:  AIC=-1090.89
## TRF_price ~ carbody + drivewheel + enginelocation + wheelbase +
##     carwidth + curbweight + enginetype + enginesize + fuelsystem +
##     boreratio + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - wheelbase       1  0.005080 0.77474 -1091.5
## - fuelsystem      7  0.053080 0.82274 -1091.2
## <none>                        0.76966 -1090.9
## - enginesize      1  0.008816 0.77848 -1090.5
## - boreratio       1  0.013118 0.78278 -1089.4
## - drivewheel      2  0.034096 0.80376 -1086.0
## - carwidth        1  0.035895 0.80556 -1083.5
## - curbweight      1  0.059086 0.82875 -1077.7
## - enginelocation  1  0.063274 0.83294 -1076.7
## - horsepower      1  0.064714 0.83438 -1076.3
## - enginetype      6  0.110433 0.88010 -1075.4
## - carbody         4  0.120383 0.89005 -1069.1
##
## Step:  AIC=-1091.54
## TRF_price ~ carbody + drivewheel + enginelocation + carwidth +
##     curbweight + enginetype + enginesize + fuelsystem + boreratio +
##     horsepower
##
##                  Df Sum of Sq     RSS     AIC
## <none>                        0.77474 -1091.5
## - fuelsystem      7  0.056045 0.83079 -1091.2
## - enginesize      1  0.011227 0.78597 -1090.6
## - boreratio       1  0.012504 0.78725 -1090.3
## - drivewheel      2  0.037274 0.81202 -1085.9
## - carwidth        1  0.057326 0.83207 -1078.9
## - horsepower      1  0.059988 0.83473 -1078.2
## - enginelocation  1  0.060774 0.83552 -1078.1
## - enginetype      6  0.111421 0.88617 -1076.0
## - curbweight      1  0.079305 0.85405 -1073.6
## - carbody         4  0.115741 0.89049 -1071.0
```

```
summary(modelo_backward)
```

```
##
## Call:
## lm(formula = TRF_price ~ carbody + drivewheel + enginelocation +
##     carwidth + curbweight + enginetype + enginesize + fuelsystem +
##     boreratio + horsepower, data = CarPrice_predictoras)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.174774 -0.040944 -0.000444  0.029996  0.161193
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         2.361e+00  2.904e-01   8.131 6.84e-14 ***
## carbodyhardtop     -1.204e-01  3.748e-02  -3.213 0.001557 ** 
## carbodyhatchback   -1.144e-01  3.169e-02  -3.611 0.000396 ***
## carbodysedan       -7.786e-02  3.075e-02  -2.532 0.012199 *  
## carbodywagon       -1.213e-01  3.350e-02  -3.622 0.000380 ***
## drivewheelfwd      -2.305e-02  2.842e-02  -0.811 0.418302    
## drivewheelrwd       2.612e-02  3.045e-02   0.858 0.392251    
## enginelocationrear  2.296e-01  6.126e-02   3.747 0.000241 ***
## carwidth            1.917e-02  5.266e-03   3.639 0.000357 ***
## curbweight          1.671e-04  3.905e-05   4.281 3.03e-05 ***
## enginetypedohcv    -8.440e-02  8.583e-02  -0.983 0.326775    
## enginetypel        -1.070e-02  3.387e-02  -0.316 0.752469    
## enginetypeohc       5.783e-02  2.427e-02   2.382 0.018248 *  
## enginetypeohcf      4.396e-02  3.832e-02   1.147 0.252848    
## enginetypeohcv     -2.814e-02  2.985e-02  -0.943 0.347111    
## enginetyperotor     1.445e-01  7.187e-02   2.011 0.045873 *  
## enginesize          5.738e-04  3.563e-04   1.611 0.109035    
## fuelsystem2bbl     -1.298e-02  2.213e-02  -0.587 0.558050    
## fuelsystem4bbl     -3.624e-03  8.019e-02  -0.045 0.964006    
## fuelsystemidi       3.341e-02  2.974e-02   1.123 0.262766    
## fuelsystemmfi      -1.980e-02  7.229e-02  -0.274 0.784508    
## fuelsystemmpfi      3.340e-02  2.491e-02   1.341 0.181745    
## fuelsystemspdi     -2.752e-02  3.302e-02  -0.834 0.405610    
## fuelsystemspfi     -1.248e-02  7.181e-02  -0.174 0.862236    
## boreratio          -5.251e-02  3.089e-02  -1.700 0.090927 .  
## horsepower          1.449e-03  3.892e-04   3.723 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06579 on 179 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9096 
## F-statistic:  83.1 on 25 and 179 DF,  p-value: < 2.2e-16
```

The backward model has marginally improved the coefficient of determination R_square to 0.9096, has increased the test statistic F to 83.1, in addition to reducing the number of predictor variables to 10:

```
# Best lineal model until now.

modelo_backward$call

## lm(formula = TRF_price ~ carbody + drivewheel + enginelocation + 
##     carwidth + curbweight + enginetype + enginesize + fuelsystem + 
##     boreratio + horsepower, data = CarPrice_predictoras)
```

However, there are two not significant predictors, “enginesize” and “boreratio”,their p-value > 0.05, an indication they may not contribute to the model.

Regression will be performed again eliminating these 2 features.

```
# Remove discarded variable from model
# New removed features: enginesize y boreratio

desechadas <- c("symboling", "stroke", "carheight", "compressionratio", "peakrpm",
"fueltype", "doornumber","citympg", "highwaympg", "price", "enginesize", "boreratio
")

# Data frame with significant variables for the model

CarPrice_predictoras <- data[ , !(names(data) %in% desechadas)]

# New backward model with less variables

modelo <- lm(TRF_price ~., data = CarPrice_predictoras)

modelo_backward2 <- step(object =modelo, direction = "backward", trace = 1)

## Start:  AIC=-1088.6
## TRF_price ~ aspiration + carbody + drivewheel + enginelocation +
##     wheelbase + carlength + carwidth + curbweight + enginetype +
##     cylindernumber + fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - carlength       1  0.000488 0.77124 -1090.5
## - aspiration      1  0.001000 0.77176 -1090.3
## - wheelbase       1  0.003427 0.77418 -1089.7
## <none>                        0.77076 -1088.6
## - fuelsystem      7  0.055752 0.82651 -1088.3
## - cylindernumber  1  0.017747 0.78850 -1085.9
## - drivewheel      2  0.026432 0.79719 -1085.7
## - carwidth        1  0.026983 0.79774 -1083.5
## - horsepower      1  0.033430 0.80419 -1081.9
## - curbweight      1  0.063469 0.83422 -1074.4
## - enginelocation  1  0.072287 0.84304 -1072.2
## - enginetype      6  0.122877 0.89363 -1070.3
## - carbody         4  0.113955 0.88471 -1068.3
##
## Step:  AIC=-1090.47
## TRF_price ~ aspiration + carbody + drivewheel + enginelocation +
##     wheelbase + carwidth + curbweight + enginetype + cylindernumber +
##     fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - aspiration      1  0.000754 0.77200 -1092.3
## - wheelbase       1  0.005141 0.77638 -1091.1
## <none>                        0.77124 -1090.5
## - fuelsystem      7  0.056215 0.82746 -1090.0
## - cylindernumber  1  0.017446 0.78869 -1087.9
## - drivewheel      2  0.026024 0.79727 -1087.7
## - carwidth        1  0.030911 0.80215 -1084.4
## - horsepower      1  0.034449 0.80569 -1083.5
## - enginelocation  1  0.072424 0.84367 -1074.1
## - curbweight      1  0.081609 0.85285 -1071.8
## - enginetype      6  0.126080 0.89732 -1071.4
## - carbody         4  0.114568 0.88581 -1070.1
##
## Step:  AIC=-1092.27
## TRF_price ~ carbody + drivewheel + enginelocation + wheelbase +
##     carwidth + curbweight + enginetype + cylindernumber + fuelsystem +
##     horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - wheelbase       1  0.005465 0.77746 -1092.8
## <none>                        0.77200 -1092.3
## - fuelsystem      7  0.055653 0.82765 -1092.0
## - cylindernumber  1  0.017423 0.78942 -1089.7
## - drivewheel      2  0.025594 0.79759 -1089.6
## - carwidth        1  0.031866 0.80386 -1086.0
## - horsepower      1  0.056570 0.82857 -1079.8
## - enginelocation  1  0.073637 0.84563 -1075.6
## - curbweight      1  0.081439 0.85344 -1073.7
## - enginetype      6  0.125448 0.89745 -1073.4
## - carbody         4  0.114842 0.88684 -1071.8
##
## Step:  AIC=-1092.82
## TRF_price ~ carbody + drivewheel + enginelocation + carwidth +
##     curbweight + enginetype + cylindernumber + fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## <none>                        0.77746 -1092.8
## - fuelsystem      7  0.057177 0.83464 -1092.3
## - cylindernumber  1  0.018466 0.79593 -1090.0
## - drivewheel      2  0.028775 0.80624 -1089.4
## - horsepower      1  0.051111 0.82857 -1081.8
## - carwidth        1  0.051595 0.82906 -1081.7
## - enginelocation  1  0.070666 0.84813 -1077.0
## - enginetype      6  0.125490 0.90295 -1074.1
## - carbody         4  0.109727 0.88719 -1073.8
## - curbweight      1  0.125826 0.90329 -1064.1
```

```
(sumario2 <- summary(modelo_backward2))
```

```
##
## Call:
## lm(formula = TRF_price ~ carbody + drivewheel + enginelocation +
##     carwidth + curbweight + enginetype + cylindernumber + fuelsystem +
##     horsepower, data = CarPrice_predictoras)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -0.18246 -0.04030  0.00000  0.03382  0.16494 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         2.224e+00  2.822e-01   7.882 2.97e-13 ***
## carbodyhardtop     -1.142e-01  3.716e-02  -3.073 0.002451 ** 
## carbodyhatchback   -1.121e-01  3.149e-02  -3.558 0.000478 ***
## carbodysedan       -7.865e-02  3.059e-02  -2.571 0.010953 *  
## carbodywagon       -1.247e-01  3.286e-02  -3.794 0.000203 ***
## drivewheelfwd      -2.017e-02  2.732e-02  -0.738 0.461290    
## drivewheelrwd       2.162e-02  2.840e-02   0.761 0.447541    
## enginelocationrear  2.425e-01  5.995e-02   4.045 7.76e-05 ***
## carwidth            1.823e-02  5.274e-03   3.456 0.000683 ***
## curbweight          1.798e-04  3.331e-05   5.397 2.11e-07 ***
## enginetypedohcv    -1.295e-01  8.569e-02  -1.511 0.132442    
## enginetypel        -1.565e-02  3.339e-02  -0.469 0.639869    
## enginetypeohc       5.416e-02  2.382e-02   2.274 0.024156 *  
## enginetypeohcf      2.013e-02  3.283e-02   0.613 0.540567    
## enginetypeohcv     -3.795e-02  3.133e-02  -1.211 0.227469    
## enginetyperotor     1.655e-01  7.390e-02   2.239 0.026378 *  
## cylindernumber      1.942e-02  9.394e-03   2.068 0.040099 *  
## fuelsystem2bbl     -1.870e-02  2.179e-02  -0.858 0.391810    
## fuelsystem4bbl     -1.669e-02  7.983e-02  -0.209 0.834668    
## fuelsystemidi       2.898e-02  2.953e-02   0.981 0.327705    
## fuelsystemmfi      -1.934e-02  7.212e-02  -0.268 0.788863    
## fuelsystemmpfi      2.829e-02  2.443e-02   1.158 0.248454    
## fuelsystemspdi     -2.687e-02  3.315e-02  -0.811 0.418688    
## fuelsystemspfi     -2.300e-02  7.128e-02  -0.323 0.747298    
## horsepower          1.340e-03  3.895e-04   3.440 0.000723 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06572 on 180 degrees of freedom
## Multiple R-squared:  0.9204, Adjusted R-squared:  0.9098 
## F-statistic: 86.71 on 24 and 180 DF,  p-value: < 2.2e-16
```

This model with the best predictor features, has a high adjusted R_square (0.9098). It is able to explain 90.98% of the variability observed in the price of vehicles.

The p-value of the model is significant (<0.05), as well as the test statistic F, therefore it can be accepted that the relationship is not random.

We can say, this model is good at predicting the price of cars, using the following predictor variables:

carbody drivewheel *enginelocation* carwidth *curbweight* enginetype *cylindernumber* fuelsystem *horsepower

Diagnosis and validation of the linear model obtained. Multicolinearity Test:

```
#  Multicollinearity test

vif(modelo_backward2)

##                       GVIF Df GVIF^(1/(2*Df))
## carbody         2.741826  4        1.134371
## drivewheel      3.993699  2        1.413656
## enginelocation  2.460122  1        1.568478
## carwidth        6.046077  1        2.458877
## curbweight     14.207537  1        3.769289
## enginetype     53.051216  6        1.392274
## cylindernumber  4.869416  1        2.206675
## fuelsystem     26.195165  7        1.262704
## horsepower     11.203835  1        3.347213
```

Thera are variables with values greater than 10, a clear indication of multicollinearity. Regression will be carried out again, eliminating the variable with the highest index and verifying the behavior of the new model.

```
# Remove discarded variable from model
# New removed feature: enginetype

desechadas <- c("symboling", "stroke", "carheight", "compressionratio", "peakrpm",
"fueltype",
                "doornumber", "citympg", "highwaympg", "price", "enginesize", "bore
ratio", "enginetype" )

# Data frame with significant variables for the model

CarPrice_predictoras <- data[ , !(names(data) %in% desechadas)]


# New backward model with less variables

modelo <- lm(TRF_price ~., data = CarPrice_predictoras)

modelo_backward3 <- step(object =modelo, direction = "backward", trace = 1)

## Start:  AIC=-1070.27
## TRF_price ~ aspiration + carbody + drivewheel + enginelocation +
##     wheelbase + carlength + carwidth + curbweight + cylindernumber +
##     fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - aspiration      1  0.000481 0.89411 -1072.2
## - wheelbase       1  0.002032 0.89566 -1071.8
## - carlength       1  0.003691 0.89732 -1071.4
## - cylindernumber  1  0.005716 0.89935 -1071.0
## - fuelsystem      7  0.063001 0.95663 -1070.3
## <none>                        0.89363 -1070.3
## - drivewheel      2  0.021901 0.91553 -1069.3
## - carwidth        1  0.036517 0.93015 -1064.1
## - curbweight      1  0.052733 0.94636 -1060.5
## - horsepower      1  0.055343 0.94898 -1060.0
## - carbody         4  0.123121 1.01675 -1051.8
## - enginelocation  1  0.117596 1.01123 -1046.9
##
## Step:  AIC=-1072.16
## TRF_price ~ carbody + drivewheel + enginelocation + wheelbase +
##     carlength + carwidth + curbweight + cylindernumber + fuelsystem +
##     horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - wheelbase       1  0.002445 0.89656 -1073.6
## - carlength       1  0.003333 0.89745 -1073.4
## - cylindernumber  1  0.005438 0.89955 -1072.9
## - fuelsystem      7  0.062521 0.95663 -1072.3
## <none>                        0.89411 -1072.2
## - drivewheel      2  0.021493 0.91561 -1071.3
## - carwidth        1  0.036959 0.93107 -1065.9
## - curbweight      1  0.053295 0.94741 -1062.3
## - horsepower      1  0.071298 0.96541 -1058.4
## - carbody         4  0.123522 1.01763 -1053.6
## - enginelocation  1  0.117803 1.01192 -1048.8
##
## Step:  AIC=-1073.6
## TRF_price ~ carbody + drivewheel + enginelocation + carlength +
##     carwidth + curbweight + cylindernumber + fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - cylindernumber  1  0.005355 0.90191 -1074.4
## - carlength       1  0.006395 0.90295 -1074.1
## - fuelsystem      7  0.061140 0.95770 -1074.1
## <none>                        0.89656 -1073.6
## - drivewheel      2  0.025799 0.92236 -1071.8
## - carwidth        1  0.048436 0.94499 -1064.8
## - curbweight      1  0.063480 0.96004 -1061.6
## - horsepower      1  0.072571 0.96913 -1059.7
## - carbody         4  0.121176 1.01773 -1055.6
## - enginelocation  1  0.116834 1.01339 -1050.5
##
## Step:  AIC=-1074.38
## TRF_price ~ carbody + drivewheel + enginelocation + carlength +
##     carwidth + curbweight + fuelsystem + horsepower
##
##                  Df Sum of Sq     RSS     AIC
## - fuelsystem      7  0.055851 0.95776 -1076.1
## - carlength       1  0.004025 0.90594 -1075.5
## <none>                        0.90191 -1074.4
## - drivewheel      2  0.022229 0.92414 -1073.4
## - carwidth        1  0.051872 0.95379 -1064.9
## - curbweight      1  0.080178 0.98209 -1058.9
## - carbody         4  0.132809 1.03472 -1054.2
## - horsepower      1  0.106651 1.00856 -1053.5
## - enginelocation  1  0.114685 1.01660 -1051.8
## 
## Step:  AIC=-1076.06
## TRF_price ~ carbody + drivewheel + enginelocation + carlength + 
##     carwidth + curbweight + horsepower
## 
##                  Df Sum of Sq     RSS     AIC
## - carlength       1  0.007368 0.96513 -1076.5
## <none>                        0.95776 -1076.1
## - drivewheel      2  0.053159 1.01092 -1069.0
## - carwidth        1  0.075620 1.03338 -1062.5
## - curbweight      1  0.100559 1.05832 -1057.6
## - carbody         4  0.159764 1.11753 -1052.4
## - enginelocation  1  0.130590 1.08835 -1051.9
## - horsepower      1  0.197949 1.15571 -1039.5
## 
## Step:  AIC=-1076.49
## TRF_price ~ carbody + drivewheel + enginelocation + carwidth + 
##     curbweight + horsepower
## 
##                  Df Sum of Sq     RSS     AIC
## <none>                        0.96513 -1076.5
## - drivewheel      2  0.053026 1.01816 -1069.5
## - carwidth        1  0.102130 1.06726 -1057.9
## - enginelocation  1  0.129720 1.09485 -1052.6
## - carbody         4  0.164246 1.12938 -1052.3
## - curbweight      1  0.185106 1.15024 -1042.5
## - horsepower      1  0.190586 1.15572 -1041.5
```

```
(sumario3 <- summary(modelo_backward3))
```

```
## 
## Call:
## lm(formula = TRF_price ~ carbody + drivewheel + enginelocation + 
##     carwidth + curbweight + horsepower, data = CarPrice_predictoras)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.161752 -0.043683 -0.005067  0.037541  0.208891 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          2.057e+00  2.774e-01   7.415 3.68e-12 ***
## carbodyhardtop      -1.057e-01  3.872e-02  -2.731 0.006888 ** 
## carbodyhatchback   -1.202e-01  3.201e-02  -3.757 0.000228 ***
## carbodysedan       -7.632e-02  3.142e-02  -2.429 0.016052 *
## carbodywagon       -1.347e-01  3.393e-02  -3.969 0.000101 ***
## drivewheelfwd      -1.656e-02  2.578e-02  -0.643 0.521261
## drivewheelrwd       3.274e-02  2.644e-02   1.238 0.217048
## enginelocationrear  2.608e-01  5.107e-02   5.106 7.82e-07 ***
## carwidth            2.291e-02  5.057e-03   4.531 1.03e-05 ***
## curbweight          1.723e-04  2.824e-05   6.100 5.65e-09 ***
## horsepower          1.404e-03  2.269e-04   6.189 3.52e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07053 on 194 degrees of freedom
## Multiple R-squared:  0.9012, Adjusted R-squared:  0.8961
## F-statistic: 176.9 on 10 and 194 DF,  p-value: < 2.2e-16
```

This new model has a high adjusted R_squared= 0.8960. It is able to explain 89.6% of the variability observed in the price of vehicles. With just a decrease of 1.369 % of the coefficient of determination and a very significant improvement in the test statistic F = 176.90.

The p-value of the model is significant (<0.05), as well as the test statistic F, therefore it can be accepted that the relationship is not random.

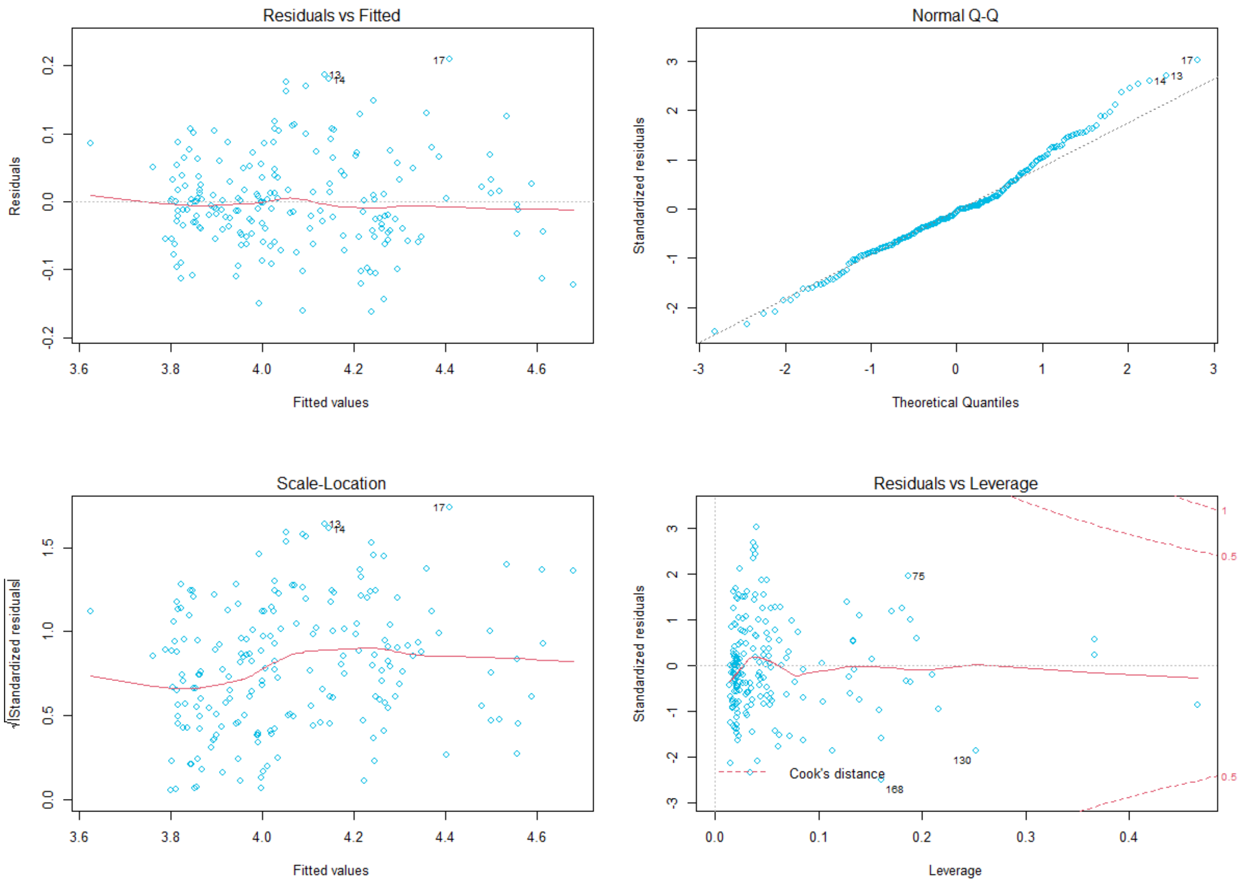
```
#  Multicollinearity test

vif(modelo_backward3)

##                    GVIF Df GVIF^(1/(2*Df))
## carbody        1.908427  4        1.084138
## drivewheel     2.305861  2        1.232277
## enginelocation 1.549992  1        1.244987
## carwidth       4.825405  1        2.196681
## curbweight     8.865120  1        2.977435
## horsepower     3.300690  1        1.816780
```

Multicollinearity is not observed in the new set of predictive variables.

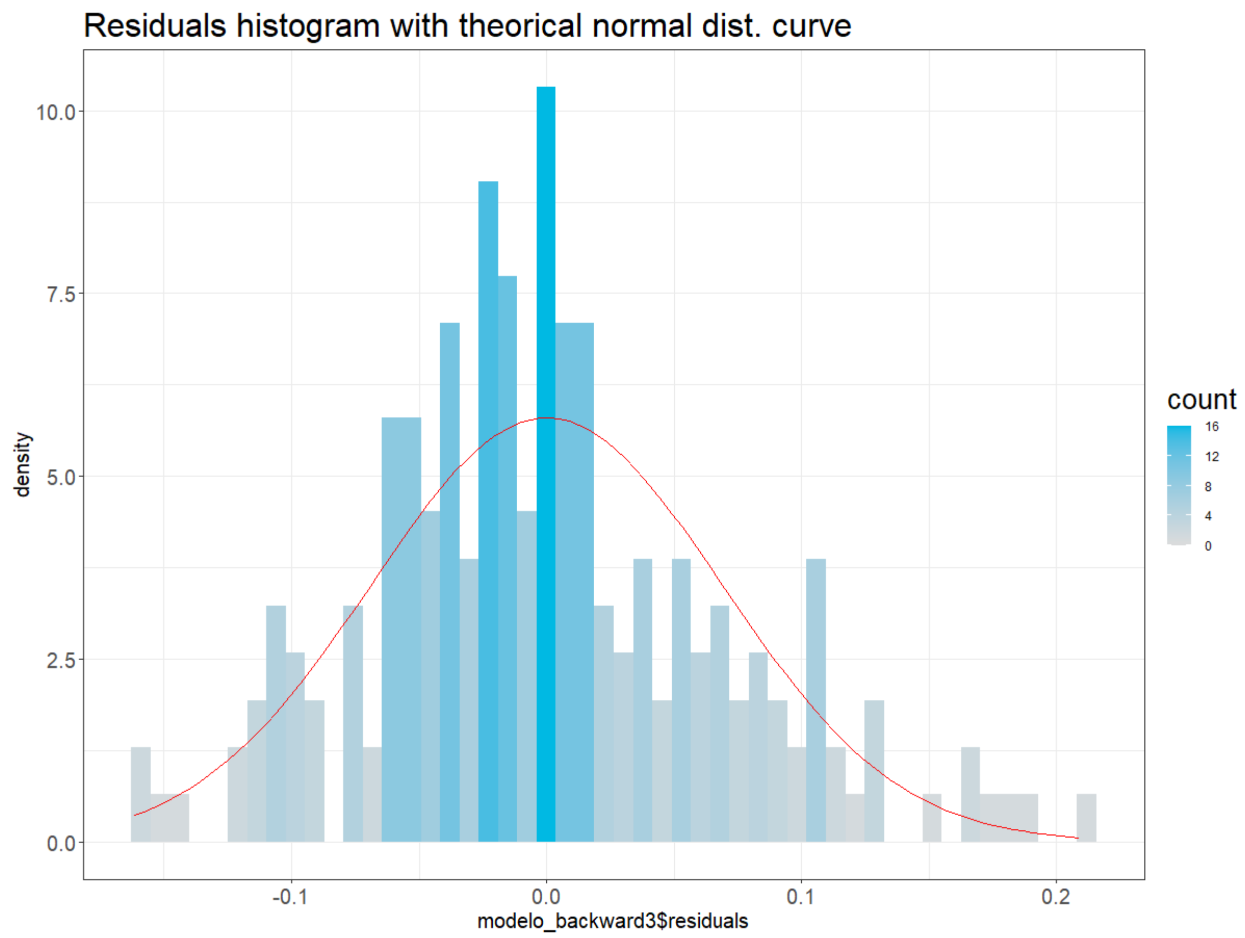
Therefore, we can say, this model is good enough to predict the price of cars, using the following predictive variables:

carbody drivewheel *enginelocation* carwidth *curbweight* horsepower #Tools to check model assumptions: +Q-Q graph of standardized residuals +Heteroscedasticity graph +Graph of Cook’s distances against fitted values +Graph of residuals against adjusted values.

```
par(mfrow=c(2,2))
plot(modelo_backward3, col =c("#00B9E3"))
```



```
ggplot(data = modelo_backward3, aes(x = modelo_backward3$residuals)) +
  geom_histogram(bins = 50,aes(y = ..density.., fill = ..count..)) +
  scale_fill_gradient(low = "#DCDCDC", high = "#00B9E3") +
  stat_function(fun = dnorm, colour = "red",
                args = list(mean = mean(modelo_backward3$residuals),
                            sd = sd(modelo_backward3$residuals))) +
  ggtitle("Residuals histogram with theorical normal dist. curve") +
  theme_bw() +
  theme(title = element_text(size = 20),
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))
```



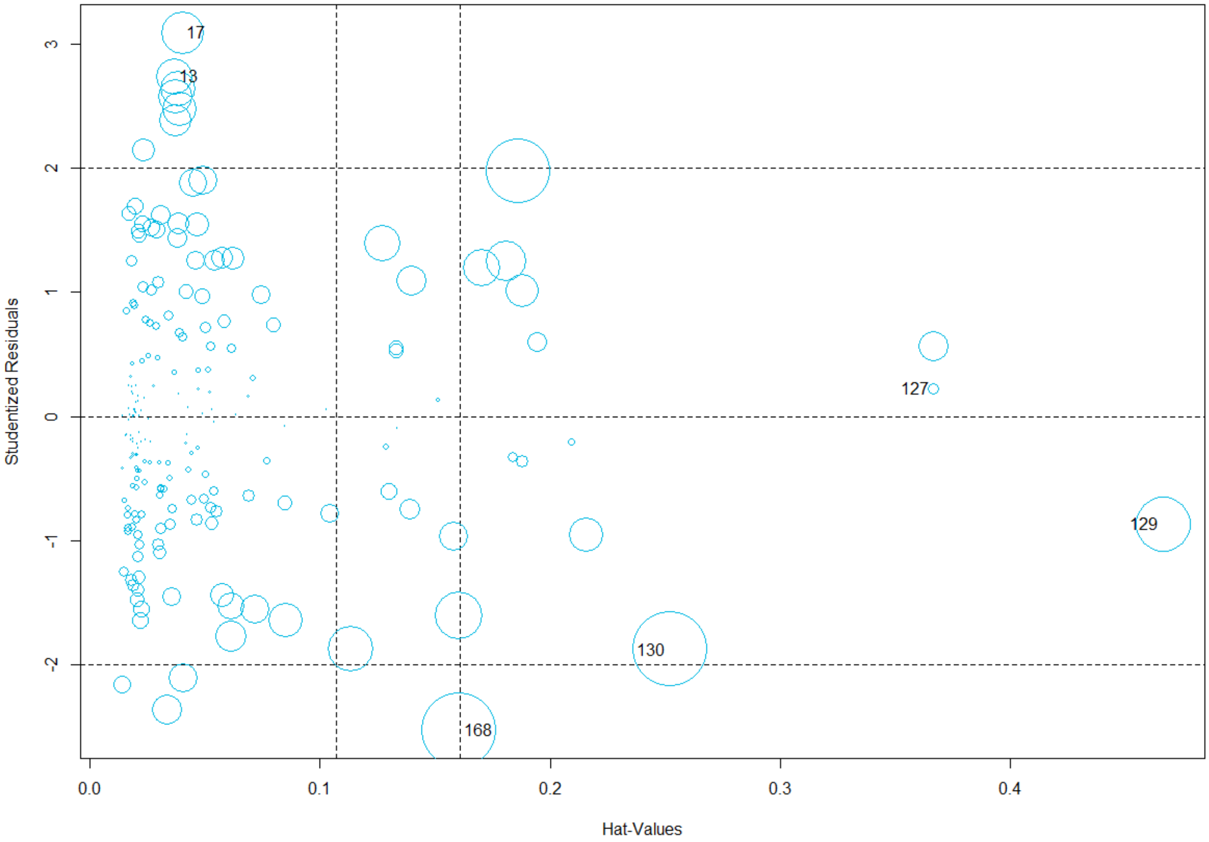
```
# Normality test

lillie.test(x = modelo_backward3$residuals)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  modelo_backward3$residuals
## D = 0.08785, p-value = 0.0005783
```

The null hypothesis of normality of the residuals is rejected.

```
# Homoscedasticity test

bptest(modelo_backward3)

##
##  studentized Breusch-Pagan test
##
## data:  modelo_backward3
## BP = 41.656, df = 10, p-value = 8.634e-06
```

The null hypothesis of homocedasticity of the residuals is rejected.

```
# Detection and visualization of outliers

outlierTest (modelo_backward3)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 17 3.088535          0.0023081      0.47316

par(mfrow=c(1,1))
influencePlot(modelo_backward3,col =c("#00B9E3") )
```



```
##         StudRes        Hat        CookD
## 13   2.7383333 0.03659512 0.025054453
## 17   3.0885347 0.04002864 0.034635212
## 127  0.2219798 0.36659486 0.002605389
## 129 -0.8643161 0.46637945 0.059432784
## 130 -1.8747015 0.25205224 0.106291366
## 168 -2.5239959 0.16039373 0.107655701
```

●Conclusion: With the results obtained, the following model goodness validations can be concluded:

1.An even distribution around 0 of the residuals is observed compared to the values adjusted by the model, the QQ plot reflects, there are indications of a lack of normality in the residuals, only of higher value, corroborated by the Lilliefors hypothesis test (Kolmogorov-Smirnov) who rejects the normality of the residuals.

2.In the same way, we verified the residuals are distributed uniformly at random around the x = 0 axis and do not form specific groups or clusters, this is an indication of the independence of the predictive variables against the dependent variable.

3.On the contrary, the Breusch-Pagan test provides evidence of lack of homoscedasticity.

4.Observations 13,17 and 168 seem to have a high level of influence, they can be considered as influential outliers.

5.The set of chosen predictors did not present variance inflation after their last adjustment.

6.Due to big disparity between the largest and smallest values of the independent variable, there is also a greater risk of presenting heteroscedasticity. The vehicle price variable was highly skewed to the left, although a transformation was carried out to normalize its distribution.

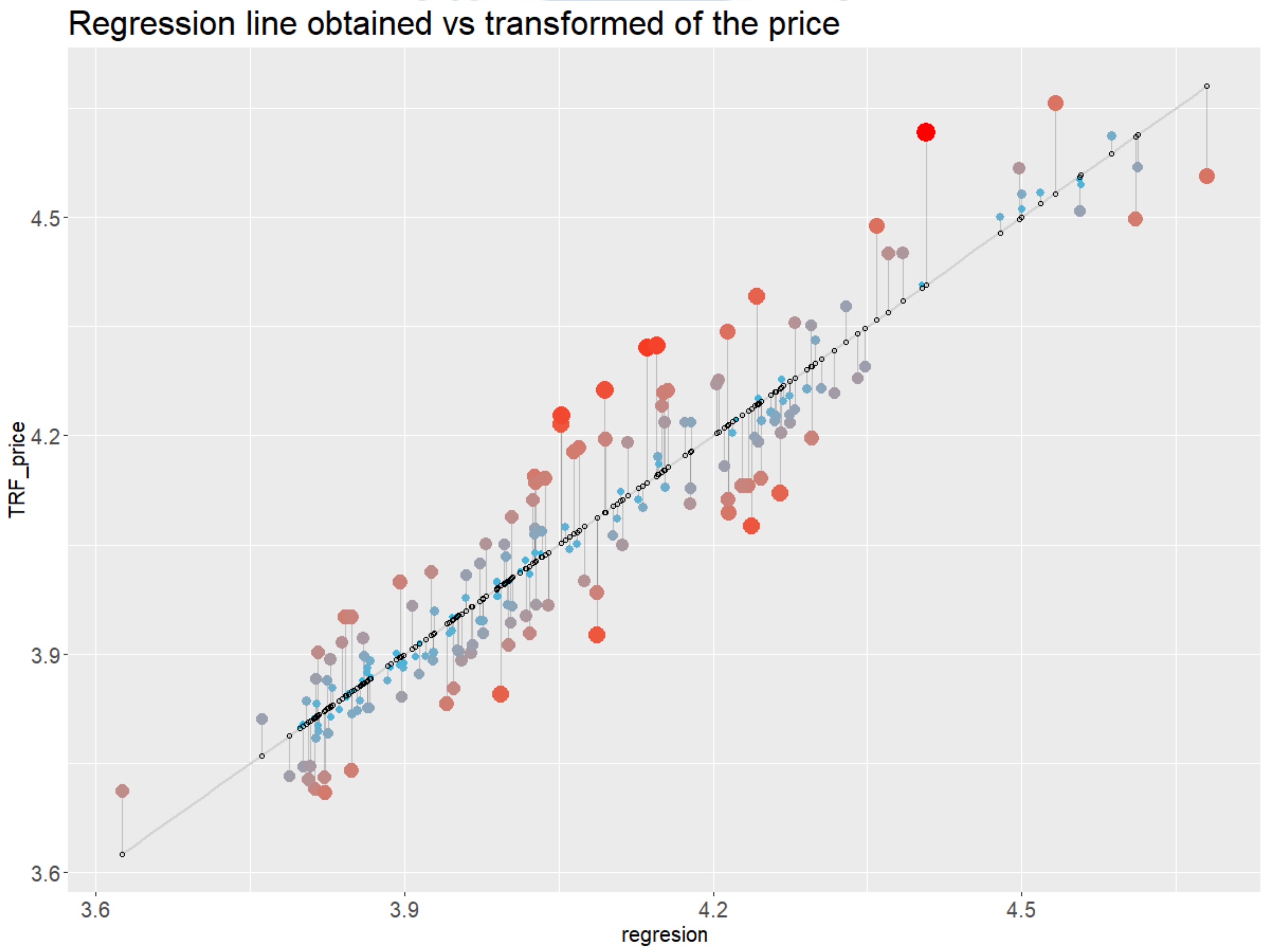
7.The main objective of the study is fulfilled but, with conditions, it has been determined that several features in the dataset can linearly predict the price of a vehicle, however, due to the nature of the distribution of the price values in the dataset, it is likely that a quadratic expression will define better its behavior.

Graph that shows the regression line obtained vs. the transform of the price, the size and color of the points, indicate the magnitude of the residuals; blue and small are values with better prediction and closer to the regression line, red and larger are values with less prediction. The size of the residual is the length of the small vertical lines, from each point to where it meets the regression line.

● geom_smooth() using formula 'y ~ x'

```
CarPrice_predictoras$regresion <- predict(modelo_backward3)           # Save vector
of regression values
CarPrice_predictoras$residuos <- residuals(modelo_backward3)          # Save residu
als
ggplot(CarPrice_predictoras, aes(x = regresion, y = TRF_price)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +       # regression
line
  geom_segment(aes(xend = regresion, yend = regresion), alpha = .2) + # vertical li
nes of residual
  geom_point(aes(color = abs(residuos), size = abs(residuos))) +      # size of poi
```

```
nts
  scale_color_continuous(low = "#00B9E3", high = "red") +            # color of th
e points mapped by residue size
  guides(color = FALSE, size = FALSE) +
  geom_point(aes(y = regresion), shape = 1) +
  ggtitle("Regression line obtained vs transformed of the price") +
  theme_gray() +
  theme(title = element_text(size = 20),
        axis.text = element_text(size = 15),
        axis.title = element_text(size=15))

## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



References:

- ✓ Element of Statistical Learning; Vol 2
- ✓ ISLR v2
- ✓ https://www.geeksforgeeks.org/
- ✓ Regression with R by Rob Tibshirani.