Consider only the below columns and prepare a prediction model for predicting Price.

Corolla<-Corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

Model -- model of the car

Price -- Offer Price in EUROs

Age\_08\_04 -- Age in months as in August 2004

Mfg\_Month -- Manufacturing month (1-12)

Mfg\_Year -- Manufacturing Year

KM -- Accumulated Kilometers on odometer

Fuel\_Type -- Fuel Type (Petrol, Diesel, CNG)

HP -- Horse Power

Met\_Color -- Metallic Color? (Yes=1, No=0)

Color -- Color (Blue, Red, Grey, Silver, Black, etc.)

Automatic -- Automatic ( (Yes=1, No=0)

cc -- Cylinder Volume in cubic centimeters

Doors -- Number of doors

Cylinders -- Number of cylinders

Gears -- Number of gear positions

Quarterly\_Tax -- Quarterly road tax in EUROs

Weight -- Weight in Kilograms

Mfr\_Guarantee -- Within Manufacturer's Guarantee period (Yes=1, No=0)

BOVAG\_Guarantee -- BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)

Guarantee\_Period -- Guarantee period in months

ABS -- Anti-Lock Brake System (Yes=1, No=0)

Airbag\_1 -- Driver\_Airbag (Yes=1, No=0)

Airbag\_2 -- Passenger Airbag (Yes=1, No=0)

Airco -- Airconditioning (Yes=1, No=0)

Automatic\_airco -- Automatic Airconditioning (Yes=1, No=0)

Boardcomputer -- Boardcomputer (Yes=1, No=0)

CD\_Player -- CD Player (Yes=1, No=0)

Central\_Lock -- Central Lock (Yes=1, No=0)

Powered\_Windows -- Powered Windows (Yes=1, No=0)

Power\_Steering -- Power Steering (Yes=1, No=0)

Radio -- Radio (Yes=1, No=0)

Mistlamps -- Mistlamps (Yes=1, No=0)

Sport\_Model -- Sport Model (Yes=1, No=0)

Backseat\_Divider -- Backseat Divider (Yes=1, No=0)

Metallic\_Rim --Metallic Rim (Yes=1, No=0)

Radio\_cassette -- Radio Cassette (Yes=1, No=0)

Tow\_Bar -- Tow Bar (Yes=1, No=0)

**#install psych packages for some data analysis**

install.packages("psych")

library(psych)

**#importing dataset**

Toyotta\_corolla <- read.csv(file.choose(),header = T)

View(Toyotta\_corolla)

nrow(Toyotta\_corolla)

>1439

str(Toyotta\_corolla)

|  |
| --- |
| str(Toyotta\_corolla)  'data.frame': 1436 obs. of 38 variables:  $ Id : int 1 2 3 4 5 6 7 8 9 10 ...  $ Model : chr "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" " TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" "TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors" ...  $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  $ Age\_08\_04 : int 23 23 24 26 30 32 27 30 27 23 ...  $ Mfg\_Month : int 10 10 9 7 3 1 6 3 6 10 ...  $ Mfg\_Year : int 2002 2002 2002 2002 2002 2002 2002 2002 2002 2002 ...  $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  $ Fuel\_Type : chr "Diesel" "Diesel" "Diesel" "Diesel" ...  $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  $ Met\_Color : int 1 1 1 0 0 0 1 1 0 0 ...  $ Color : chr "Blue" "Silver" "Blue" "Black" ...  $ Automatic : int 0 0 0 0 0 0 0 0 0 0 ...  $ cc : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  $ Doors : int 3 3 3 3 3 3 3 3 3 3 ...  $ Cylinders : int 4 4 4 4 4 4 4 4 4 4 ...  $ Gears : int 5 5 5 5 5 5 5 5 5 5 ...  $ Quarterly\_Tax : int 210 210 210 210 210 210 210 210 100 185 ...  $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ...  $ Mfr\_Guarantee : int 0 0 1 1 1 0 0 1 0 0 ...  $ BOVAG\_Guarantee : int 1 1 1 1 1 1 1 1 1 1 ...  $ Guarantee\_Period: int 3 3 3 3 3 3 3 3 3 3 ...  $ ABS : int 1 1 1 1 1 1 1 1 1 1 ...  $ Airbag\_1 : int 1 1 1 1 1 1 1 1 1 1 ...  $ Airbag\_2 : int 1 1 1 1 1 1 1 1 0 1 ...  $ Airco : int 0 1 0 0 1 1 1 1 1 1 ...  $ Automatic\_airco : int 0 0 0 0 0 0 0 0 0 0 ...  $ Boardcomputer : int 1 1 1 1 1 1 1 1 0 1 ...  $ CD\_Player : int 0 1 0 0 0 0 0 1 0 0 ...  $ Central\_Lock : int 1 1 0 0 1 1 1 1 1 0 ...  $ Powered\_Windows : int 1 0 0 0 1 1 1 1 1 0 ...  $ Power\_Steering : int 1 1 1 1 1 1 1 1 1 1 ...  $ Radio : int 0 0 0 0 0 0 0 0 1 0 ...  $ Mistlamps : int 0 0 0 0 1 1 0 0 0 0 ...  $ Sport\_Model : int 0 0 0 0 0 0 1 0 0 0 ...  $ Backseat\_Divider: int 1 1 1 1 1 1 1 1 0 1 ...  $ Metallic\_Rim : int 0 0 0 0 0 0 0 0 1 0 ...  $ Radio\_cassette : int 0 0 0 0 0 0 0 0 1 0 ...  $ Tow\_Bar : int 0 0 0 0 0 0 0 0 0 0 ... |
|  |
| |  | | --- | | > | |

# in our data set contain 38 bariables and 1436 records

**#here we are taking few variable for price prediction**

Corolla<-Toyotta\_corolla[c("Price","Age\_08\_04","KM","HP","cc","Doors","Gears","Quarterly\_Tax","Weight")]

head(Corolla)

|  |
| --- |
| Price Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight  1 13500 23 46986 90 2000 3 5 210 1165  2 13750 23 72937 90 2000 3 5 210 1165  3 13950 24 41711 90 2000 3 5 210 1165  4 14950 26 48000 90 2000 3 5 210 1165  5 13750 30 38500 90 2000 3 5 210 1170  6 12950 32 61000 90 2000 3 5 210 1170 |
|  |
| |  | | --- | | > | |

**# here price is our target or dependant variable, others are our independent variable**

**#seeing the summary and structure and checking the aspects of data set**

str(Corolla)

|  |
| --- |
| data.frame': 1436 obs. of 9 variables:  $ Price : int 13500 13750 13950 14950 13750 12950 16900 18600 21500 12950 ...  $ Age\_08\_04 : int 23 23 24 26 30 32 27 30 27 23 ...  $ KM : int 46986 72937 41711 48000 38500 61000 94612 75889 19700 71138 ...  $ HP : int 90 90 90 90 90 90 90 90 192 69 ...  $ cc : int 2000 2000 2000 2000 2000 2000 2000 2000 1800 1900 ...  $ Doors : int 3 3 3 3 3 3 3 3 3 3 ...  $ Gears : int 5 5 5 5 5 5 5 5 5 5 ...  $ Quarterly\_Tax: int 210 210 210 210 210 210 210 210 100 185 ...  $ Weight : int 1165 1165 1165 1165 1170 1170 1245 1245 1185 1105 ... |
|  |
| |  | | --- | | > | |

nrow(Corolla)

1436

summary(Corolla**)**

|  |
| --- |
| Price Age\_08\_04 KM HP cc  Min. : 4350 Min. : 1.00 Min. : 1 Min. : 69.0 Min. : 1300  1st Qu.: 8450 1st Qu.:44.00 1st Qu.: 43000 1st Qu.: 90.0 1st Qu.: 1400  Median : 9900 Median :61.00 Median : 63390 Median :110.0 Median : 1600  Mean :10731 Mean :55.95 Mean : 68533 Mean :101.5 Mean : 1577  3rd Qu.:11950 3rd Qu.:70.00 3rd Qu.: 87021 3rd Qu.:110.0 3rd Qu.: 1600  Max. :32500 Max. :80.00 Max. :243000 Max. :192.0 Max. :16000  Doors Gears Quarterly\_Tax Weight  Min. :2.000 Min. :3.000 Min. : 19.00 Min. :1000  1st Qu.:3.000 1st Qu.:5.000 1st Qu.: 69.00 1st Qu.:1040  Median :4.000 Median :5.000 Median : 85.00 Median :1070  Mean :4.033 Mean :5.026 Mean : 87.12 Mean :1072  3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.: 85.00 3rd Qu.:1085  Max. :5.000 Max. :6.000 Max. :283.00 Max. :1615 |
|  |
| |  | | --- | | > | |

attach(Corolla)

**########VISULIZATION########**

**#Here we are using two types of visualization methods**

**#UNIVARIATE AND BIVARIATE ANALYSIS**

**#UNIVARIATE ANALYSIS**

**#Univariate analysis means one variable analysis.'Uni' means 'one' and 'variate'**

**#means'variable'. Univariate analysis is to analyse one variable or**

**#one features Univariate basically tells us how data in**

**#each feature is distributed and also tells us about central tendencies**

**#like mean, median, and mode.**

**#To do univariate data analysis we use following ploting mechanisms:**

**#Histograms**

**#Boxplot**

install.packages("ggplot2")

library(ggplot2)

**#histogram of price**

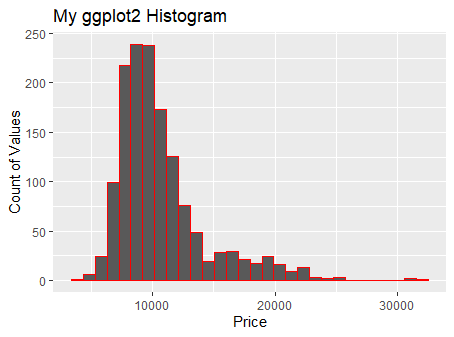
ggplot(Corolla, aes(x = Price)) +

geom\_histogram(col = "red") +

labs(title = "My ggplot2 Histogram",

x = "Price",

y = "Count of Values")



**# here we can see that most of the price is in the range of 10000 to 20000 euros**

**#histgram of Age\_08\_04**

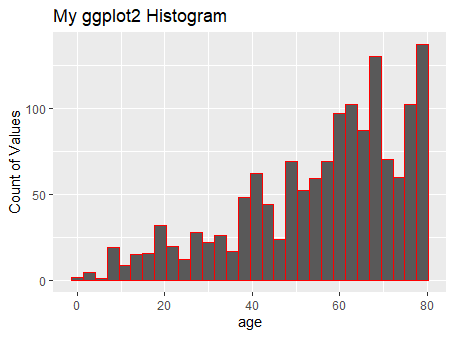
ggplot(Corolla, aes(x = Age\_08\_04)) +

geom\_histogram(col = "red") +

labs(title = "My ggplot2 Histogram",

x = "age",

y = "Count of Values")



**#histogram of KM**

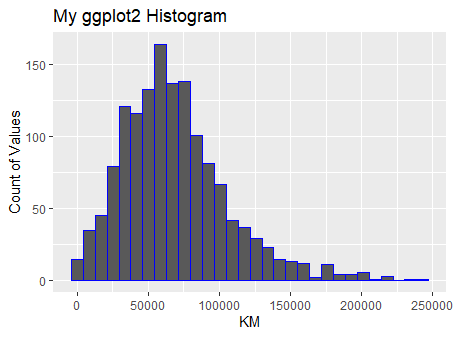
ggplot(Corolla, aes(x = KM)) +

geom\_histogram(col = "blue") +

labs(title = "My ggplot2 Histogram",

x = "KM",

y = "Count of Values") **##Accumulated Kilometers on odometer(KM) all in the range of 50000 to 150000**



**#histogram of HP**

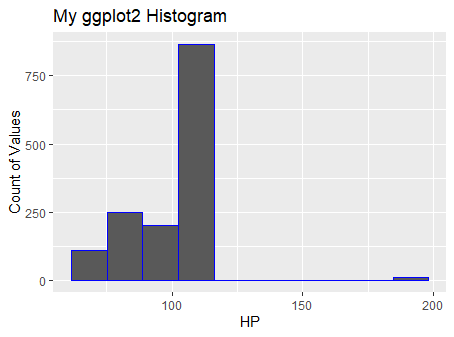
ggplot(Corolla, aes(x = HP)) +

geom\_histogram(bins=10,col = "blue") +

labs(title = "My ggplot2 Histogram",

x = "HP",

y = "Count of Values") **##we can se e that most of the carhorse power is 100 to 110 range**



**#histogram of cc**

ggplot(Corolla, aes(x = cc)) +

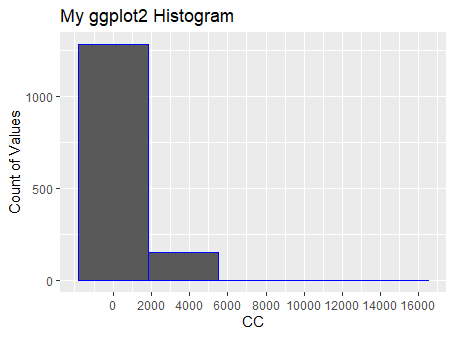
geom\_histogram(bins=5,col = "blue") +

scale\_x\_continuous(breaks = seq(0,max(Corolla),2000))+

labs(title = "My ggplot2 Histogram",

x = "CC",

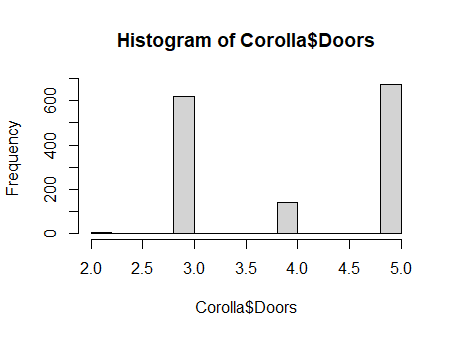
y = "Count of Values") ## **Cylinder Volume in cubic centimeters (cc)**



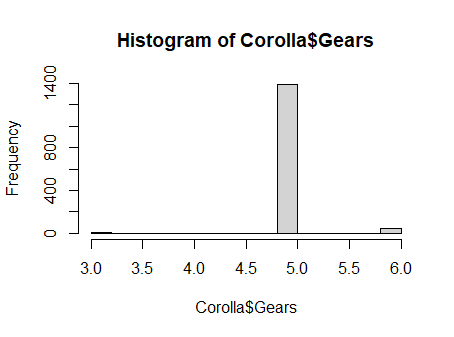
**#here most of the car cc is in the range of 1000 to 2000**

**#histogram of Doors**

hist(Corolla$Doors) **# here most of the cars doors 3 or 5**



hist(Corolla$Gears) # **Number of gear positions is 5**



**#histogram of Quarterly\_Tax**

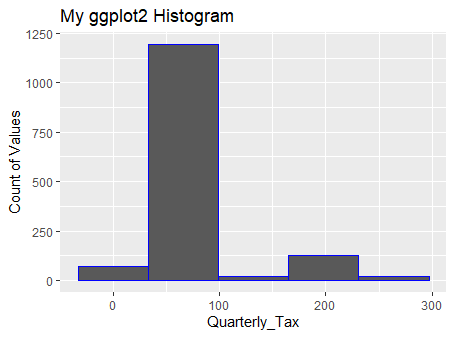
ggplot(Corolla, aes(x = Quarterly\_Tax)) +

geom\_histogram(bins=5,col = "blue") +

labs(title = "My ggplot2 Histogram",

x = "Quarterly\_Tax",

y = "Count of Values") **# Quarterly\_Tax is in the range of 150 to 250 euros**



**#histogram of Weight**

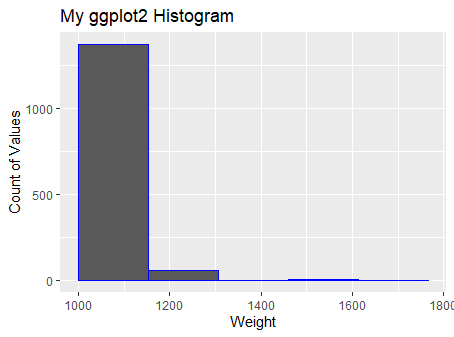
ggplot(Corolla, aes(x = Weight)) +

geom\_histogram(bins=5,col = "blue") +

labs(title = "My ggplot2 Histogram",

x = "Weight",

y = "Count of Values") ##**most of the car weight is in the range of 1000 to 1200**



**#Boxplot**

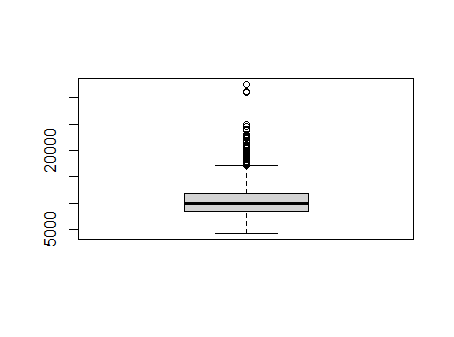
**#Boxplot is used for numerical data representations.in Boxplot data representation is in quartile basis. a boxplot contains five parts:Minimum, First quartile(25% of data),**

**#second quartile(50% of data, also called median),Third quartile(75% of data)**

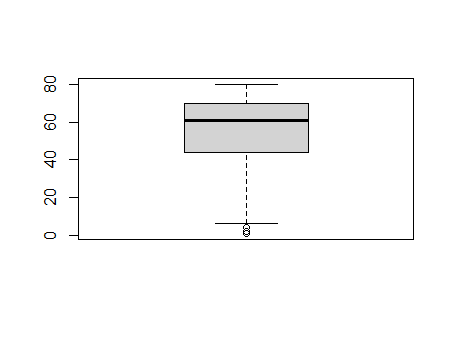
**#and maximum.If data below the minimum boundaries and above the maximum boundaries**

**#considered as Outliers. Main task of boxplot is to detect outliers.**

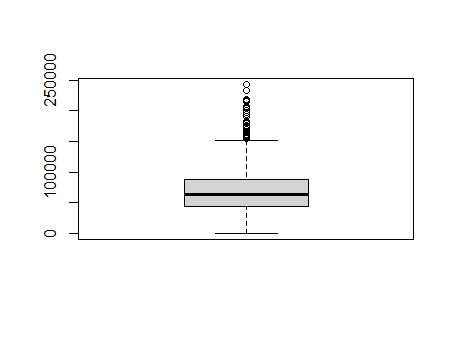
boxplot(Corolla$Price) **# we can see that price variable has many outliers**



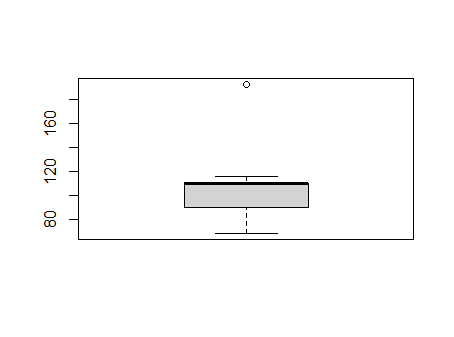
boxplot(Corolla$Age\_08\_04) **# few outliers detected**



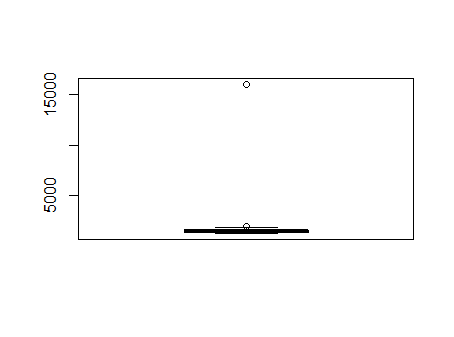
boxplot(Corolla$KM) **#many outliers**



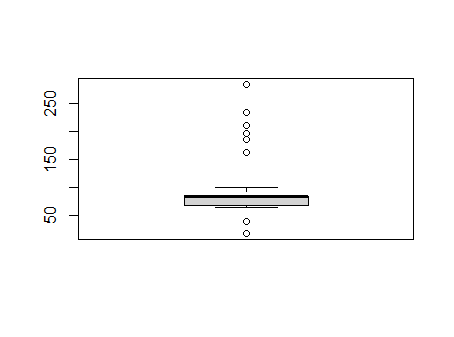
boxplot(Corolla$HP) **##one outliers**



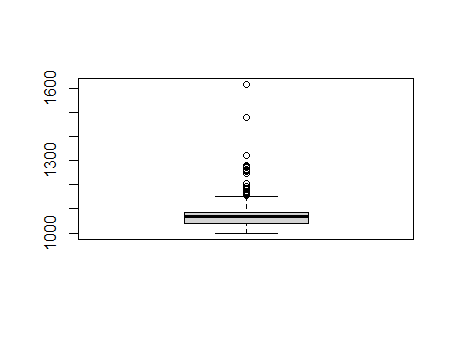
boxplot(Corolla$cc) **#one ouliers**



boxplot(Corolla$Quarterly\_Tax) **# many outliers are detected**



boxplot(Corolla$Weight) **# many outliers are detected**



**# here we see that boxplot representation of varible, and also detects many outliers**

**# for better prediction we have to remove those outliers from our data set**

**# here we see that boxplot representation of varible, and also detects many outliers**

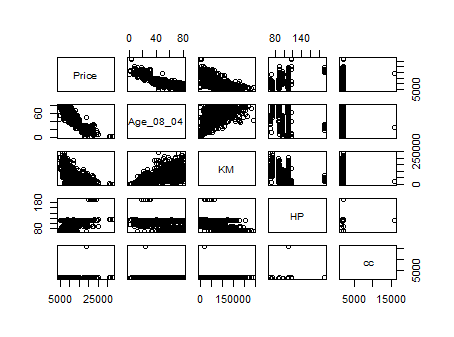
**# for better prediction we have to remove those outliers from our data set**

**##BIVARIATE ANALYSIS###**

**#Pair plot**

**#here i am dividing data sets in to two parts for more convenience to visualize pair plot**

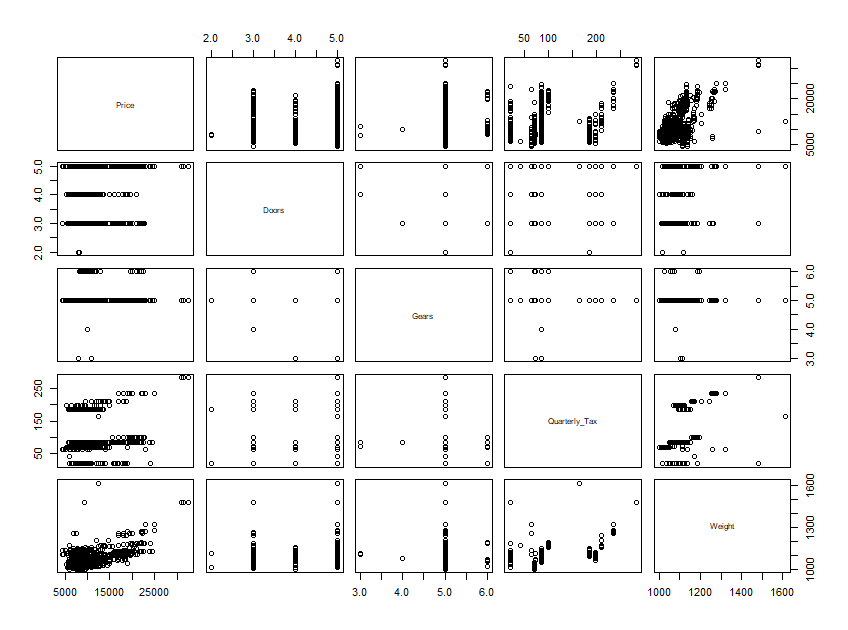
pairs(Corolla[c("Price","Age\_08\_04","KM","HP","cc")]) # in this plot we can see that



**# KM and Age\_08\_04 has some linear relation ship between price, but other variables such as**

**# HP,cc has not make any impact on price**

pairs(Corolla[c("Price","Doors","Gears","Quarterly\_Tax","Weight")])

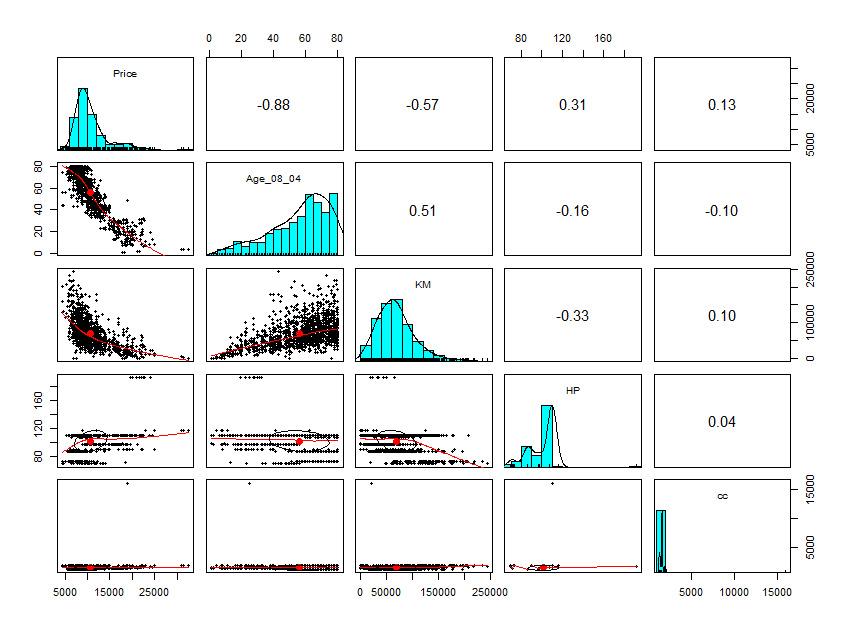


**#here only Weight has form a linear relationship**

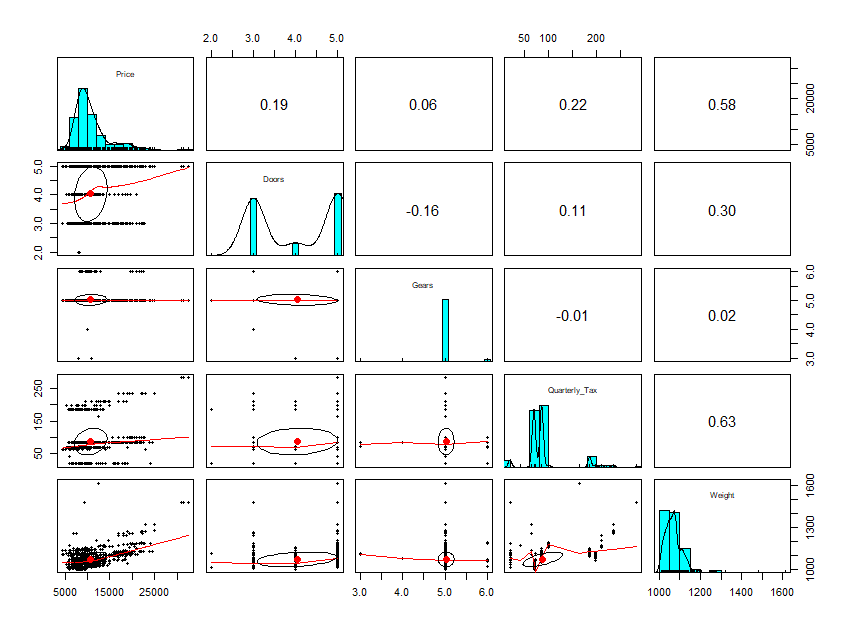
**#with respect to price**

**#using pairs.panels to show the correlation between variables**

pairs.panels(Corolla[c("Price","Age\_08\_04","KM","HP","cc")])



pairs.panels(Corolla[c("Price","Doors","Gears","Quarterly\_Tax","Weight")])



**# here price is our target variable, lets check the correlation between price and other variable**

**cor(Corolla)**

|  |
| --- |
| rice 1.00000000 -0.876590497 -0.56996016 0.31498983 0.12638920 0.18532555  Age\_08\_04 -0.87659050 1.000000000 0.50567218 -0.15662202 -0.09808374 -0.14835921  KM -0.56996016 0.505672180 1.00000000 -0.33353795 0.10268289 -0.03619661  HP 0.31498983 -0.156622020 -0.33353795 1.00000000 0.03585580 0.09242450  cc 0.12638920 -0.098083739 0.10268289 0.03585580 1.00000000 0.07990330  Doors 0.18532555 -0.148359215 -0.03619661 0.09242450 0.07990330 1.00000000  Gears 0.06310386 -0.005363947 0.01502333 0.20947715 0.01462935 -0.16014143  Quarterly\_Tax 0.21919691 -0.198430508 0.27816470 -0.29843172 0.30699580 0.10936323  Weight 0.58119759 -0.470253184 -0.02859846 0.08961406 0.33563740 0.30261764  Gears Quarterly\_Tax Weight  Price 0.063103857 0.219196911 0.58119759  Age\_08\_04 -0.005363947 -0.198430508 -0.47025318  KM 0.015023328 0.278164697 -0.02859846  HP 0.209477146 -0.298431717 0.08961406  cc 0.014629352 0.306995798 0.33563740  Doors -0.160141430 0.109363225 0.30261764  Gears 1.000000000 -0.005451955 0.02061328  Quarterly\_Tax -0.005451955 1.000000000 0.62613373  Weight 0.020613284 0.626133733 1.00000000 |
|  |
|  |

**# check partial correlation**

**###Partial Correlation matrix**

install.packages("corpcor")

library(corpcor)

|  |
| --- |
| [,1] [,2] [,3] [,4] [,5] [,6] [,7]  [1,] 1.000000000 -0.776238352 -0.402745405 0.28521314 -0.03556185 -0.001069746 0.07958671  [2,] -0.776238352 1.000000000 0.002383081 0.24531845 -0.02014628 -0.002800916 0.05107486  [3,] -0.402745405 0.002383081 1.000000000 -0.06039653 0.05108725 0.026724172 0.10050633  [4,] 0.285213137 0.245318454 -0.060396533 1.00000000 0.09871851 0.068175272 0.20769268  [5,] -0.035561846 -0.020146283 0.051087249 0.09871851 1.00000000 -0.016060377 -0.01198838  [6,] -0.001069746 -0.002800916 0.026724172 0.06817527 -0.01606038 1.000000000 -0.18924933  [7,] 0.079586710 0.051074865 0.100506331 0.20769268 -0.01198838 -0.189249333 1.00000000  [8,] 0.079548117 0.015830863 0.261673195 -0.38254954 0.12380803 -0.074825415 0.03732241  [9,] 0.387523482 0.094746528 0.187502181 0.12427899 0.16043171 0.231960007 -0.02325832  [,8] [,9]  [1,] 0.07954812 0.38752348  [2,] 0.01583086 0.09474653  [3,] 0.26167319 0.18750218  [4,] -0.38254954 0.12427899  [5,] 0.12380803 0.16043171  [6,] -0.07482541 0.23196001  [7,] 0.03732241 -0.02325832  [8,] 1.00000000 0.51026027  [9,] 0.51026027 1.00000000 |
| |  | | --- | | > | |

## create training and test data

install.packages("caTools")

library(caTools)

**##use caTools function to split, SplitRatio for 70%:30% splitting**

data= sample.split(Corolla,SplitRatio = 0.3)

**## here I am using 70% of data for training and 30% data for testing**

**#subsetting into Test data**

test\_data1 =subset(Corolla,data==TRUE)

**#subsetting into Train data**

train\_datas =subset(Corolla,data==FALSE)

**## check number of records present in the data set**

nrow(test\_data1)

319

nrow(train\_datas)

1117

head(train\_datas)

|  |
| --- |
| Price Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight  2 13750 23 72937 90 2000 3 5 210 1165  3 13950 24 41711 90 2000 3 5 210 1165  4 14950 26 48000 90 2000 3 5 210 1165  5 13750 30 38500 90 2000 3 5 210 1170  6 12950 32 61000 90 2000 3 5 210 1170  8 18600 30 75889 90 2000 3 5 210 1245 |
|  |
| |  | | --- | | > | |

head(test\_data1)

|  |
| --- |
| Price Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight  1 13500 23 46986 90 2000 3 5 210 1165  7 16900 27 94612 90 2000 3 5 210 1245  10 12950 23 71138 69 1900 3 5 185 1105  16 22000 28 18739 192 1800 3 6 100 1185  19 16750 24 25563 110 1600 3 5 19 1065  25 16250 29 25813 110 1600 3 5 85 1120 |
|  |
| |  | | --- | | > | |

**# create first model**

Corolla\_model1<-lm(Price~Age\_08\_04+KM+HP+cc+Doors+Gears+Quarterly\_Tax+Weight,data=train\_datas)

summary(Corolla\_model1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | Call:  lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears +  Quarterly\_Tax + Weight, data = train\_datas)  Residuals:  Min 1Q Median 3Q Max  -10701.3 -792.6 -34.2 790.9 6248.5  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -8.628e+03 1.671e+03 -5.162 2.89e-07 \*\*\*  Age\_08\_04 -1.208e+02 3.024e+00 -39.931 < 2e-16 \*\*\*  KM -2.056e-02 1.449e-03 -14.194 < 2e-16 \*\*\*  HP 2.738e+01 3.211e+00 8.528 < 2e-16 \*\*\*  cc -1.124e-01 9.191e-02 -1.223 0.22164  Doors -3.668e+01 4.599e+01 -0.798 0.42520  Gears 7.294e+02 2.266e+02 3.218 0.00133 \*\*  Quarterly\_Tax 1.758e+00 1.560e+00 1.127 0.25991  Weight 1.984e+01 1.290e+00 15.372 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1356 on 1108 degrees of freedom  Multiple R-squared: 0.8664, Adjusted R-squared: 0.8655  F-statistic: 898.5 on 8 and 1108 DF, p-value: < 2.2e-16 | |  | | |  | | --- | | > | | |
|  |
|  |

**#here we can see that gear,cc,Doors** Quarterly\_Tax  **variables are not significant**

**#lets check individually those variables significant or not**

Corolla\_model1\_gear<-lm(Price~Gears)

summary(Corolla\_model1\_gear)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | Call:  lm(formula = Price ~ Gears, data = train\_datas)  Residuals:  Min 1Q Median 3Q Max  -6385.2 -2335.2 -935.2 1214.8 21764.8  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 3725.1 2954.0 1.261 0.2076  Gears 1402.0 587.7 2.386 0.0172 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3688 on 1115 degrees of freedom  Multiple R-squared: 0.005079, Adjusted R-squared: 0.004187  F-statistic: 5.692 on 1 and 1115 DF, p-value: 0.01721 | |  | | |  | | --- | | > | |   Corolla\_model1\_cc<-lm(Price~cc,data=train\_datas)  summary(Corolla\_model1\_cc) |
| |  | | --- | | Call:  lm(formula = Price ~ cc, data = train\_datas)  Residuals:  Min 1Q Median 3Q Max  -6967.3 -2340.1 -890.1 1159.9 21289.7  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9109.3331 383.6556 23.744 < 2e-16 \*\*\*  cc 1.0505 0.2329 4.511 7.14e-06 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3665 on 1115 degrees of freedom  Multiple R-squared: 0.01792, Adjusted R-squared: 0.01704  F-statistic: 20.35 on 1 and 1115 DF, p-value: 7.138e-06 | |  | | |  | | --- | | > | |   Corolla\_model1\_Doors<-lm(Price~Doors,data=train\_datas)  summary(Corolla\_model1\_Doors)   |  | | --- | | Call:  lm(formula = Price ~ Doors, data = train\_datas)  Residuals:  Min 1Q Median 3Q Max  -7128.7 -2238.8 -1028.7 971.3 21021.3  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 7754 474 16.359 < 2e-16 \*\*\*  Doors 745 114 6.532 9.82e-11 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3629 on 1115 degrees of freedom  Multiple R-squared: 0.03686, Adjusted R-squared: 0.036  F-statistic: 42.67 on 1 and 1115 DF, p-value: 9.82e-11 | |  | | |  | | --- | | > | |   Corolla\_model1\_Tax<-lm(Price~Quarterly\_Tax,data=train\_datas)  summary(Corolla\_model1\_Tax) |
|  |
| Call:  lm(formula = Price ~ Quarterly\_Tax, data = train\_datas)  Residuals:  Min 1Q Median 3Q Max  -7937.2 -2266.4 -816.4 1233.6 17415.6  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 8841.28 248.74 35.544 <2e-16 \*\*\*  Quarterly\_Tax 22.06 2.57 8.582 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 3581 on 1115 degrees of freedom  Multiple R-squared: 0.06197, Adjusted R-squared: 0.06112  F-statistic: 73.66 on 1 and 1115 DF, p-value: < 2.2e-16 |
|  |
| |  | | --- | | > | |

**# check individually we can see that all variables are significant variable**

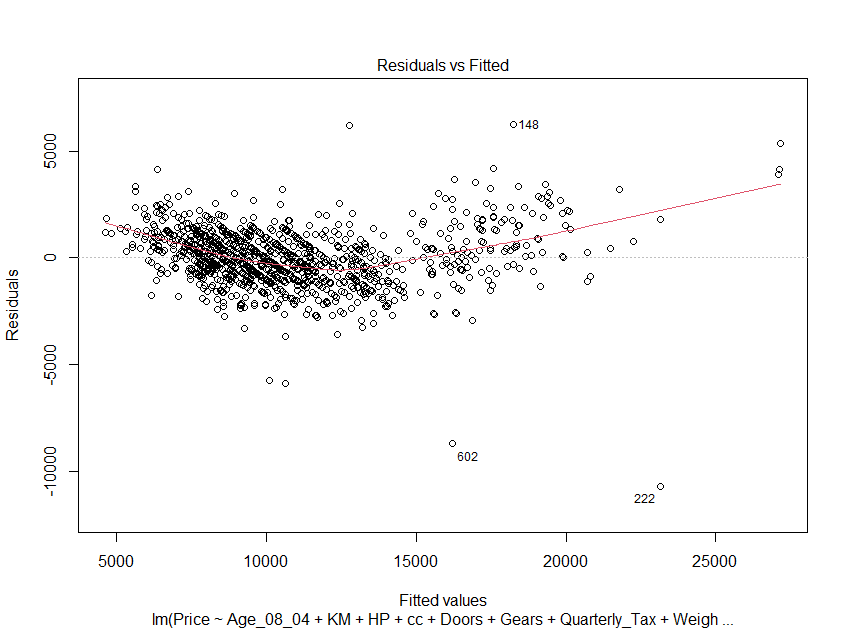
**# lets perform another analysis to improve the model performance**

**# use some diagnostic plot to detect outliers**

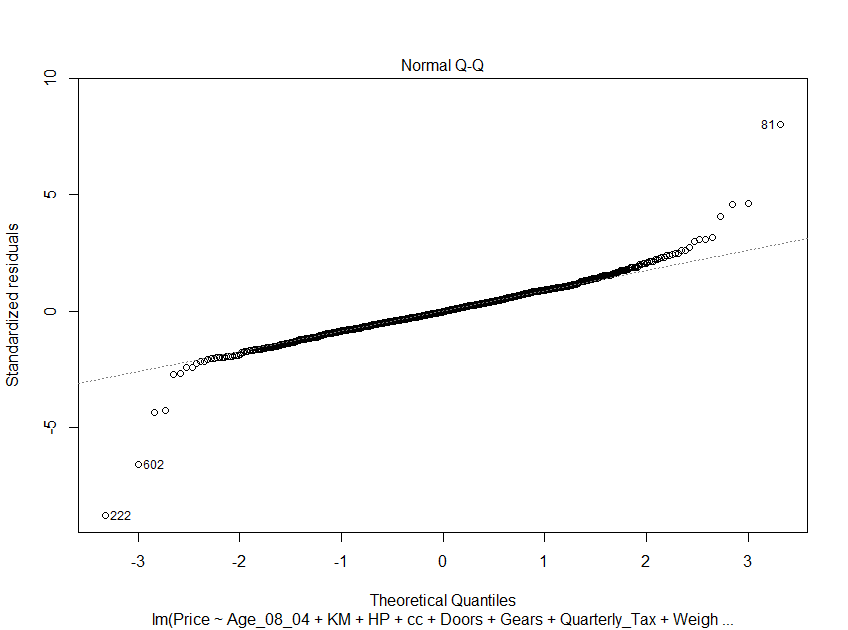
install.packages("car")

library(car)

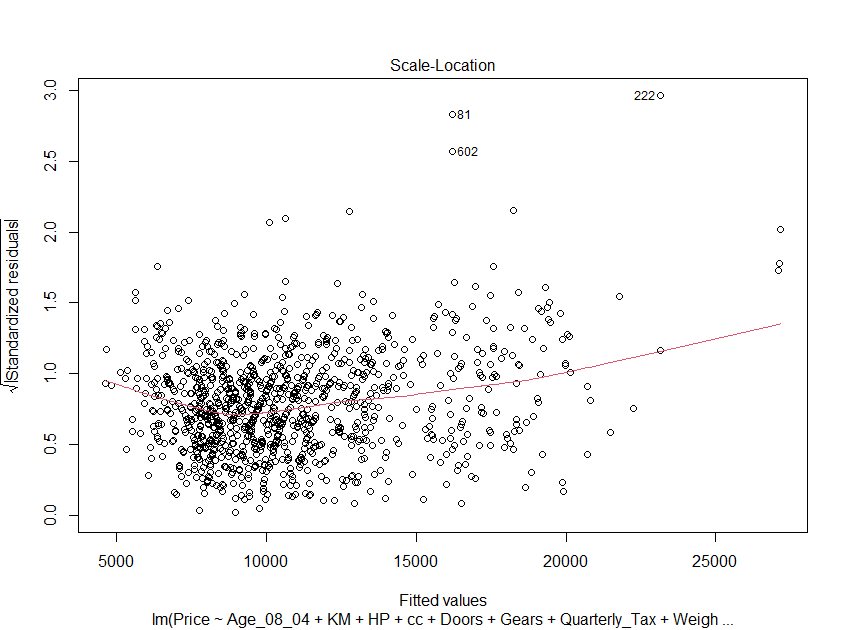
plot(Corolla\_model1) **# Residual Plots, QQ-Plos, Std. Residuals vs Fitted, Cook's distance**

****

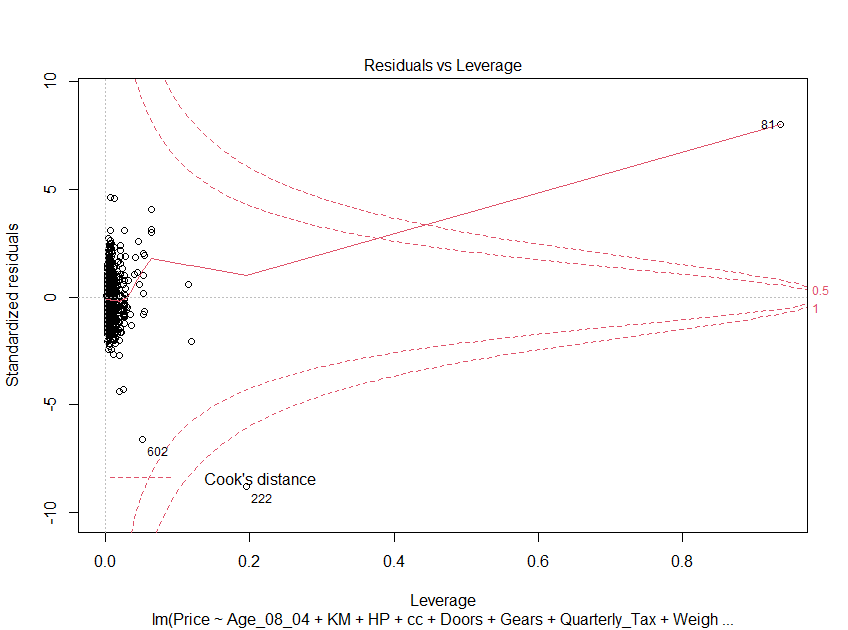
**# residual vs fitted plot showing records 602,222,148 are outliers**

****

**# normal Q-Q plot showing records 602,222,81 are outliers**

****

**#Std. Residuals vs Fitted showing records 602,81,222 are outliers**

****

**#Cook's distance showing records 602,222,81 are outliers**

**# in all plots 602,81,222 are considered outliers, so remove all those records**

**#from our training data set to improve model performance**

**#create second model,remove the otliers**

Corolla\_model2<-lm(Price~Age\_08\_04+KM+HP+cc+Doors+Gears+Quarterly\_Tax+Weight,data=train\_datas[-c(222,81,602),])

summary(Corolla\_model2)

|  |
| --- |
| Call:  lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears +  Quarterly\_Tax + Weight, data = train\_datas[-c(222, 81, 602),  ])  Residuals:  Min 1Q Median 3Q Max  -10689.1 -793.0 -33.8 789.6 6246.8  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -8.603e+03 1.673e+03 -5.144 3.19e-07 \*\*\*  Age\_08\_04 -1.208e+02 3.032e+00 -39.854 < 2e-16 \*\*\*  KM -2.057e-02 1.449e-03 -14.193 < 2e-16 \*\*\*  HP 2.736e+01 3.213e+00 8.515 < 2e-16 \*\*\*  cc -1.121e-01 9.195e-02 -1.219 0.22292  Doors -3.632e+01 4.609e+01 -0.788 0.43088  Gears 7.304e+02 2.267e+02 3.222 0.00131 \*\*  Quarterly\_Tax 1.770e+00 1.560e+00 1.135 0.25683  Weight 1.981e+01 1.291e+00 15.343 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1356 on 1105 degrees of freedom  Multiple R-squared: 0.8662, Adjusted R-squared: 0.8652  F-statistic: 894 on 8 and 1105 DF, p-value: < 2.2e-16 |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | >  **##check multicollinearity present in a data set**  **# here using Variance Inflation Factors(VIF) technique helps to identify the multicollinearity**  vif(Corolla\_model2) **# VIF is > 10 => collinearity**   |  | | --- | | Age\_08\_04 KM HP cc Doors Gears  1.932408 1.744250 1.416542 1.138232 1.166225 1.101029  Quarterly\_Tax Weight  2.569808 2.917583 | |  | | |  | | --- | | > | |   **# here we get VIF values of all variable less than 10, so we can say that there is no multicollinearity present in our data set**  **#The Akaike information criterion (AIC) is a mathematical method for evaluating**  **#how well a model fits the data it was generated from.**  **#In statistics, AIC is used to compare different possible models and determine**  **#which one is the best fit for the data**.  install.packages("MASS")  library("MASS")  stepAIC(Corolla\_model2)  AIC=16061.89  Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears + Quarterly\_Tax +  Weight  Df Sum of Sq RSS AIC  - Doors 1 5032 2002722479 16060  <none> 2002717447 16062  - cc 1 4308214 2007025661 16062  - Gears 1 10931696 2013649143 16066  - Quarterly\_Tax 1 22205219 2024922666 16072  - HP 1 210842431 2213559878 16171  - Weight 1 389544267 2392261714 16258  - KM 1 411744945 2414462392 16268  - Age\_08\_04 1 2937323614 4940041061 17066  Step: AIC=16059.9  Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax + Weight  Df Sum of Sq RSS AIC  <none> 2002722479 16060  - cc 1 4314615 2007037094 16060  - Gears 1 11215320 2013937799 16064  - Quarterly\_Tax 1 22252379 2024974858 16070  - HP 1 211467901 2214190380 16170  - KM 1 411838608 2414561086 16266  - Weight 1 413356615 2416079094 16267  - Age\_08\_04 1 2937674539 4940397018 17064  Call:  lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +  Weight, data = train\_datas[-c(222, 81, 602), ])  Coefficients:  (Intercept) Age\_08\_04 KM HP cc Gears  -6.802e+03 -1.197e+02 -2.145e-02 3.353e+01 -1.410e-01 5.246e+02  Quarterly\_Tax Weight  5.176e+00 1.814e+01 | |
|  |
|  |
| **## Lower the AIC (Akaike Information Criterion) value better is the model.**  **#AIC is used only if you build multiple models**  **#here we get two AIC model, if choose lowesr AIC model value for prediction**    **#following are the AIC model, we can choose lowest AIC model**    **#AIC=16061.89**  **#Price ~ Age\_08\_04 + KM + HP + cc + Doors + Gears + Quarterly\_Tax + Weight**  **#AIC=16059.9**  **#Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax + Weight**    **# here AIC=16059.9 is the lowest value, so choose this model for prediction**  Corolla\_model3<-lm(Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax + Weight,data=train\_datas[-c(222,81,602),] )  summary(Corolla\_model3) #Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +  Weight, data = train\_datas[-c(222, 81, 602), ])  Residuals:  Min 1Q Median 3Q Max  -8253.4 -801.6 -33.6 807.6 6457.1  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -6.802e+03 1.602e+03 -4.246 2.36e-05 \*\*\*  Age\_08\_04 -1.197e+02 2.973e+00 -40.278 < 2e-16 \*\*\*  KM -2.145e-02 1.422e-03 -15.081 < 2e-16 \*\*\*  HP 3.353e+01 3.103e+00 10.807 < 2e-16 \*\*\*  cc -1.410e-01 9.133e-02 -1.544 0.122968  Gears 5.246e+02 2.108e+02 2.489 0.012967 \*  Quarterly\_Tax 5.176e+00 1.477e+00 3.506 0.000474 \*\*\*  Weight 1.814e+01 1.200e+00 15.109 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1346 on 1106 degrees of freedom  Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  F-statistic: 1026 on 7 and 1106 DF, p-value: < 2.2e-16  # perform other transformmation methods to improve r-squared values  #logerithamic transformation  Corolla\_model4<-lm(Price ~ log(Age\_08\_04) +log( KM) + log(HP) + log(cc)+ log(Gears)+log(Quarterly\_Tax) + log(Weight),data=train\_datas[-c(222,81,602),])  summary(Corolla\_model4) ##Multiple R-squared: 0.8424, Adjusted R-squared: 0.8414  lm(formula = Price ~ log(Age\_08\_04) + log(KM) + log(HP) + log(cc) +  log(Gears) + log(Quarterly\_Tax) + log(Weight), data = train\_datas[-c(222,  81, 602), ])  Residuals:  Min 1Q Median 3Q Max  -13162.4 -826.9 -10.2 844.4 8289.6  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -105555.39 9486.45 -11.127 < 2e-16 \*\*\*  log(Age\_08\_04) -5426.26 129.71 -41.835 < 2e-16 \*\*\*  log(KM) 89.26 54.28 1.644 0.1004  log(HP) 5019.41 318.32 15.769 < 2e-16 \*\*\*  log(cc) -2661.01 436.30 -6.099 1.47e-09 \*\*\*  log(Gears) 2676.14 1133.64 2.361 0.0184 \*  log(Quarterly\_Tax) 551.85 120.36 4.585 5.05e-06 \*\*\*  log(Weight) 18112.64 1513.16 11.970 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1462 on 1106 degrees of freedom  Multiple R-squared: 0.8424, Adjusted R-squared: 0.8414  F-statistic: 844.8 on 7 and 1106 DF, p-value: < 2.2e-16  #exponential transformation  Corolla\_model5<-lm(log(Price) ~ Age\_08\_04 + KM + HP + cc+ Gears +Quarterly\_Tax+ Weight,data=train\_datas[-c(222,81,602),])  summary(Corolla\_model5) ##Multiple R-squared: 0.8484, Adjusted R-squared: 0.8475  Call:  lm(formula = log(Price) ~ Age\_08\_04 + KM + HP + cc + Gears +  Quarterly\_Tax + Weight, data = train\_datas[-c(222, 81, 602),  ])  Residuals:  Min 1Q Median 3Q Max  -0.70595 -0.06543 0.00199 0.07381 0.51118  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 8.584e+00 1.390e-01 61.739 < 2e-16 \*\*\*  Age\_08\_04 -1.027e-02 2.580e-04 -39.815 < 2e-16 \*\*\*  KM -2.072e-06 1.234e-07 -16.790 < 2e-16 \*\*\*  HP 2.876e-03 2.693e-04 10.679 < 2e-16 \*\*\*  cc -5.035e-06 7.927e-06 -0.635 0.52540  Gears 5.138e-02 1.829e-02 2.808 0.00507 \*\*  Quarterly\_Tax 7.756e-04 1.281e-04 6.052 1.95e-09 \*\*\*  Weight 7.065e-04 1.042e-04 6.781 1.94e-11 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.1168 on 1106 degrees of freedom  Multiple R-squared: 0.8484, Adjusted R-squared: 0.8475  F-statistic: 884.4 on 7 and 1106 DF, p-value: < 2.2e-16  #square root transformation  Corolla\_model6<-lm(sqrt(Price) ~ Age\_08\_04 + KM + HP +cc+ Gears +Quarterly\_Tax+ Weight,data=train\_datas[-c(222,81,602),])  summary(Corolla\_model6) ##Multiple R-squared: 0.8667, Adjusted R-squared: 0.8659  Call:  lm(formula = sqrt(Price) ~ Age\_08\_04 + KM + HP + cc + Gears +  Quarterly\_Tax + Weight, data = train\_datas[-c(222, 81, 602),  ])  Residuals:  Min 1Q Median 3Q Max  -34.511 -3.549 0.110 3.740 28.490  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 4.649e+01 7.089e+00 6.557 8.41e-11 \*\*\*  Age\_08\_04 -5.524e-01 1.316e-02 -41.979 < 2e-16 \*\*\*  KM -1.030e-04 6.294e-06 -16.367 < 2e-16 \*\*\*  HP 1.519e-01 1.373e-02 11.062 < 2e-16 \*\*\*  cc -4.251e-04 4.042e-04 -1.052 0.29312  Gears 2.564e+00 9.328e-01 2.749 0.00607 \*\*  Quarterly\_Tax 3.167e-02 6.534e-03 4.847 1.43e-06 \*\*\*  Weight 5.919e-02 5.313e-03 11.141 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 5.955 on 1106 degrees of freedom  Multiple R-squared: 0.8667, Adjusted R-squared: 0.8659  F-statistic: 1028 on 7 and 1106 DF, p-value: < 2.2e-16  #quadratic model  Corolla\_model7<-lm(Price ~ Age\_08\_04+I(Age\_08\_04)^2 + KM+I(KM)^2 + HP+I(HP)^2 +cc+I(cc)^2 + Gears+I(Gears)^2 +Quarterly\_Tax+I(Quarterly\_Tax)^2+ Weight+I(Weight)^2,data=train\_datas[-c(222,81,602),])  summary(Corolla\_model7) ##Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  Call:  lm(formula = Price ~ Age\_08\_04 + I(Age\_08\_04)^2 + KM + I(KM)^2 +  HP + I(HP)^2 + cc + I(cc)^2 + Gears + I(Gears)^2 + Quarterly\_Tax +  I(Quarterly\_Tax)^2 + Weight + I(Weight)^2, data = train\_datas[-c(222,  81, 602), ])  Residuals:  Min 1Q Median 3Q Max  -8253.4 -801.6 -33.6 807.6 6457.1  Coefficients: (7 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -6.802e+03 1.602e+03 -4.246 2.36e-05 \*\*\*  Age\_08\_04 -1.197e+02 2.973e+00 -40.278 < 2e-16 \*\*\*  I(Age\_08\_04) NA NA NA NA  KM -2.145e-02 1.422e-03 -15.081 < 2e-16 \*\*\*  I(KM) NA NA NA NA  HP 3.353e+01 3.103e+00 10.807 < 2e-16 \*\*\*  I(HP) NA NA NA NA  cc -1.410e-01 9.133e-02 -1.544 0.122968  I(cc) NA NA NA NA  Gears 5.246e+02 2.108e+02 2.489 0.012967 \*  I(Gears) NA NA NA NA  Quarterly\_Tax 5.176e+00 1.477e+00 3.506 0.000474 \*\*\*  I(Quarterly\_Tax) NA NA NA NA  Weight 1.814e+01 1.200e+00 15.109 < 2e-16 \*\*\*  I(Weight) NA NA NA NA  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1346 on 1106 degrees of freedom  Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  F-statistic: 1026 on 7 and 1106 DF, p-value: < 2.2e-16  #polynomial model  Corolla\_model8<-lm(Price ~ Age\_08\_04+I(Age\_08\_04)^2+I(Age\_08\_04)^3 + KM+I(KM)^2+I(KM)^3 + HP+I(HP)^2+I(HP)^3 +cc+I(cc)^2+I(cc)^3 + Gears+I(Gears)^2+I(Gears)^3 +Quarterly\_Tax+I(Quarterly\_Tax)^2+ I(Quarterly\_Tax)^3+Weight+I(Weight)^2+I(Weight)^3,data=train\_datas[-c(222,81,602),])  summary(Corolla\_model8) ##Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  lm(formula = Price ~ Age\_08\_04 + I(Age\_08\_04)^2 + I(Age\_08\_04)^3 +  KM + I(KM)^2 + I(KM)^3 + HP + I(HP)^2 + I(HP)^3 + cc + I(cc)^2 +  I(cc)^3 + Gears + I(Gears)^2 + I(Gears)^3 + Quarterly\_Tax +  I(Quarterly\_Tax)^2 + I(Quarterly\_Tax)^3 + Weight + I(Weight)^2 +  I(Weight)^3, data = train\_datas[-c(222, 81, 602), ])  Residuals:  Min 1Q Median 3Q Max  -8253.4 -801.6 -33.6 807.6 6457.1  Coefficients: (7 not defined because of singularities)  Estimate Std. Error t value Pr(>|t|)  (Intercept) -6.802e+03 1.602e+03 -4.246 2.36e-05 \*\*\*  Age\_08\_04 -1.197e+02 2.973e+00 -40.278 < 2e-16 \*\*\*  I(Age\_08\_04) NA NA NA NA  KM -2.145e-02 1.422e-03 -15.081 < 2e-16 \*\*\*  I(KM) NA NA NA NA  HP 3.353e+01 3.103e+00 10.807 < 2e-16 \*\*\*  I(HP) NA NA NA NA  cc -1.410e-01 9.133e-02 -1.544 0.122968  I(cc) NA NA NA NA  Gears 5.246e+02 2.108e+02 2.489 0.012967 \*  I(Gears) NA NA NA NA  Quarterly\_Tax 5.176e+00 1.477e+00 3.506 0.000474 \*\*\*  I(Quarterly\_Tax) NA NA NA NA  Weight 1.814e+01 1.200e+00 15.109 < 2e-16 \*\*\*  I(Weight) NA NA NA NA  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1346 on 1106 degrees of freedom  Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657  F-statistic: 1026 on 7 and 1106 DF, p-value: < 2.2e-16   |  | | --- | |  | |  | | |  | | --- | |  | | |  | |  | |  | |  | |  |   **Following are our r squared values, choose better value for prediction**   |  | | --- | |  | |  | |  |  |  | | --- | |  | | |  | | --- | |  | |   **#Corolla\_model1:Multiple R-squared: 0.867, Adjusted R-squared: 0.866**  **#Corolla\_model2:Multiple R-squared: 0.866, Adjusted R-squared: 0.8656**  **#Corolla\_model3:Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657**  **#Corolla\_model4:Multiple R-squared: 0.8424, Adjusted R-squared: 0.8414**  **#Corolla\_model5:Multiple R-squared: 0.8484, Adjusted R-squared: 0.8475**  **#Corolla\_model6:Multiple R-squared: 0.8667, Adjusted R-squared: 0.8659**  **#Corolla\_model7:Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657**  **#Corolla\_model8:Multiple R-squared: 0.8666, Adjusted R-squared: 0.8657**  **# here we can see that model1 has highest r squared value**  **# lets take model1**  **#here i am taking Corolla\_model1 for prediction**  **#prediction**  price\_predn<-predict(Corolla\_model1,test)  head(price\_predn)  1 6 10 15 19 24  17008.00 15720.76 14581.75 19976.96 15273.18 15990.67  **# compare the predicted value in test\_data**  head(test\_data1)  Price Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight  1 13500 23 46986 90 2000 3 5 210 1165  7 16900 27 94612 90 2000 3 5 210 1245  10 12950 23 71138 69 1900 3 5 185 1105  16 22000 28 18739 192 1800 3 6 100 1185  19 16750 24 25563 110 1600 3 5 19 1065  25 16250 29 25813 110 1600 3 5 85 1120  test\_data1$predict\_price<-price\_predn  Price Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight predict\_price  1 13500 23 46986 90 2000 3 5 210 1165 17008.00  7 16900 27 94612 90 2000 3 5 210 1245 15720.76  10 12950 23 71138 69 1900 3 5 185 1105 14581.75  16 22000 28 18739 192 1800 3 6 100 1185 19976.96  19 16750 24 25563 110 1600 3 5 19 1065 15273.18  25 16250 29 25813 110 1600 3 5 85 1120 15990.67  head(test\_data1$predict\_price)  17008.00 15720.76 14581.75 19976.96 15273.18 15990.67  **# check the correlation of actual price and predicted price**  r<-cor(test\_data1$Price,test\_data1$predict\_price)  0.8744  r\_squared<-cor(test\_data1$Price,test\_data1$predict\_price)^2  0.7646  **# using ggplot, grapgically represent the actual price and predicted price**  ggplot(data=test\_data1,aes(x=Price,color="actual price"))+  geom\_density(aes(x=Price,color="actual price"))+  geom\_density(aes(x=predict\_price,color="predicted price"))+  scale\_color\_manual(values=c("predicted price"="blue","actual price"="red"))+  labs(title="density plt between the actual price and predicted price",  x="price",y= "")  C:\Users\sathi\OneDrive\Desktop\DATASCIENCE\11  **# in this plot showing there are lot of variation between actual price and predicted price**   |  | | --- | |  | |  | |  | |  | |