1. Prepare a prediction model for profit of 50\_startups data.

Do transformations for getting better predictions of profit and

make a table containing R^2 value for each prepared model.

R&D Spend -- Research and devolop spend in the past few years

Administration -- spend on administration in the past few years

Marketing Spend -- spend on Marketing in the past few years

State -- states from which data is collected

Profit -- profit of each state in the past few years

Solution:

1.) **Business Problem:** To prepare a predication model for profit of start up

Or to predict profit for startup

2.) **Datasets:**

Independent Variables (X): R&DSpent (Research and devolop spend in the past few years), Continous

Administration (spend on administration in the past few years), Continous

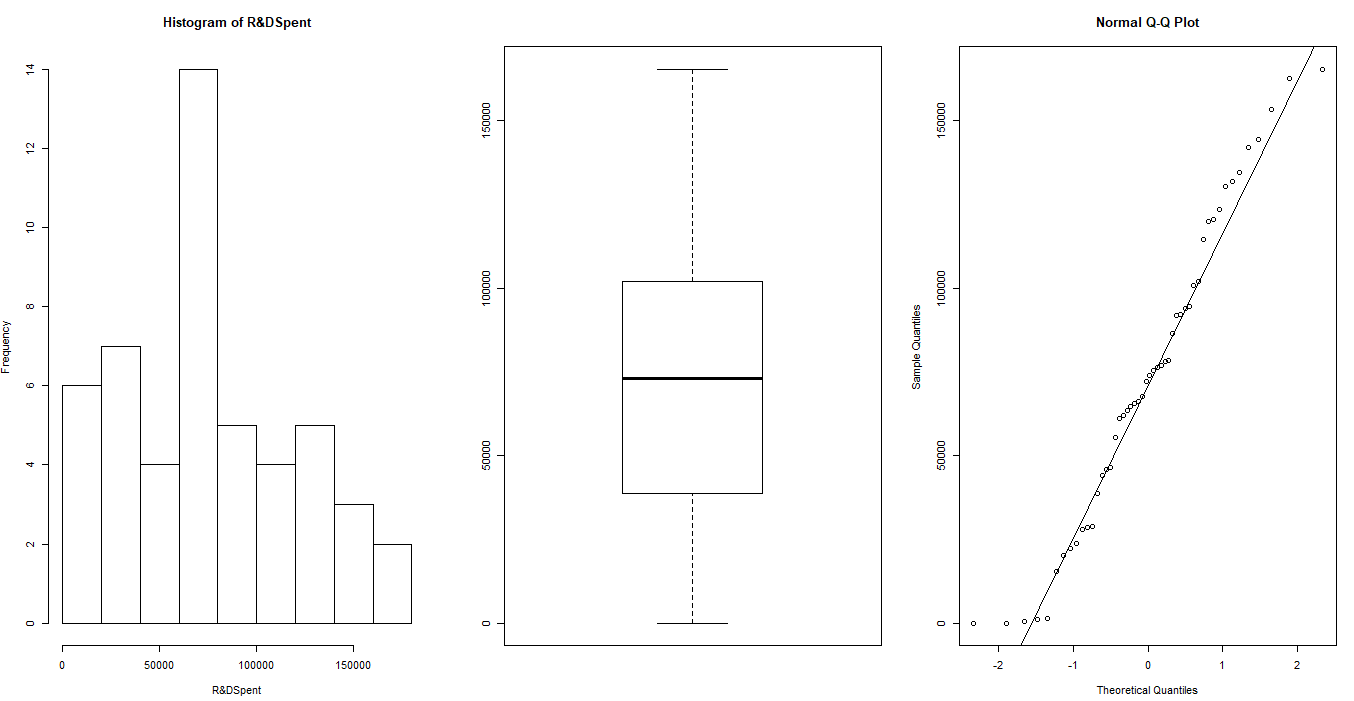
Marketing\_Spend (spend on Marketing in the past few years), Continous

State (states from which data is collected), Categorical

Dependent Variables (Y): Profit (profit of each state in the past few years) Continous

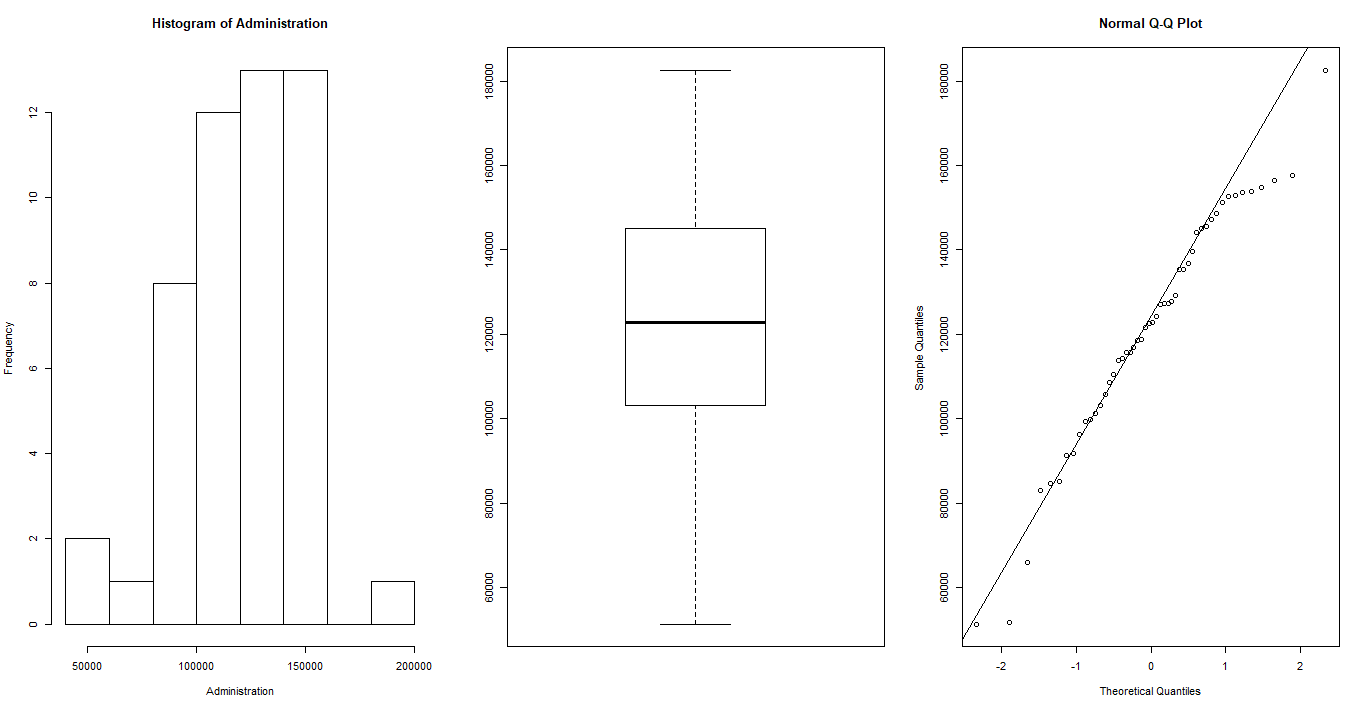
3.) **EDA:**

For R&D spent



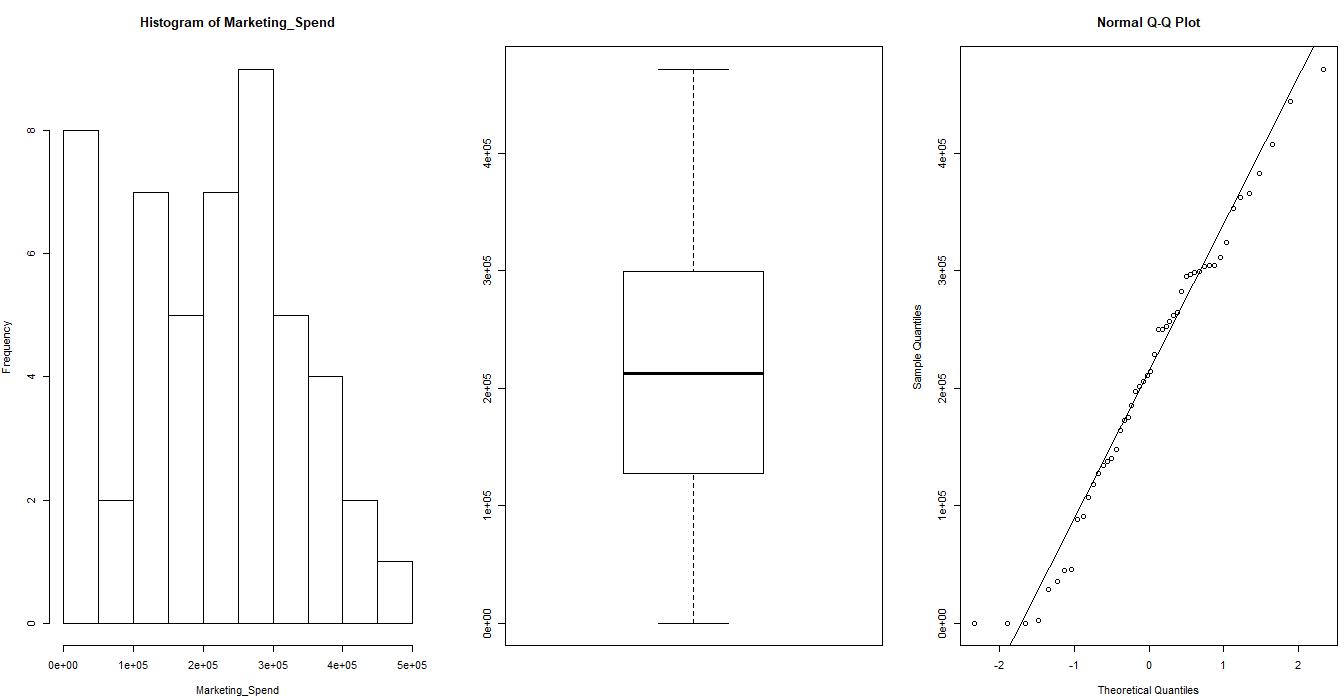
This is not normal distribution and have slight positive skewness. The box plot does not show outliers.

For Administration



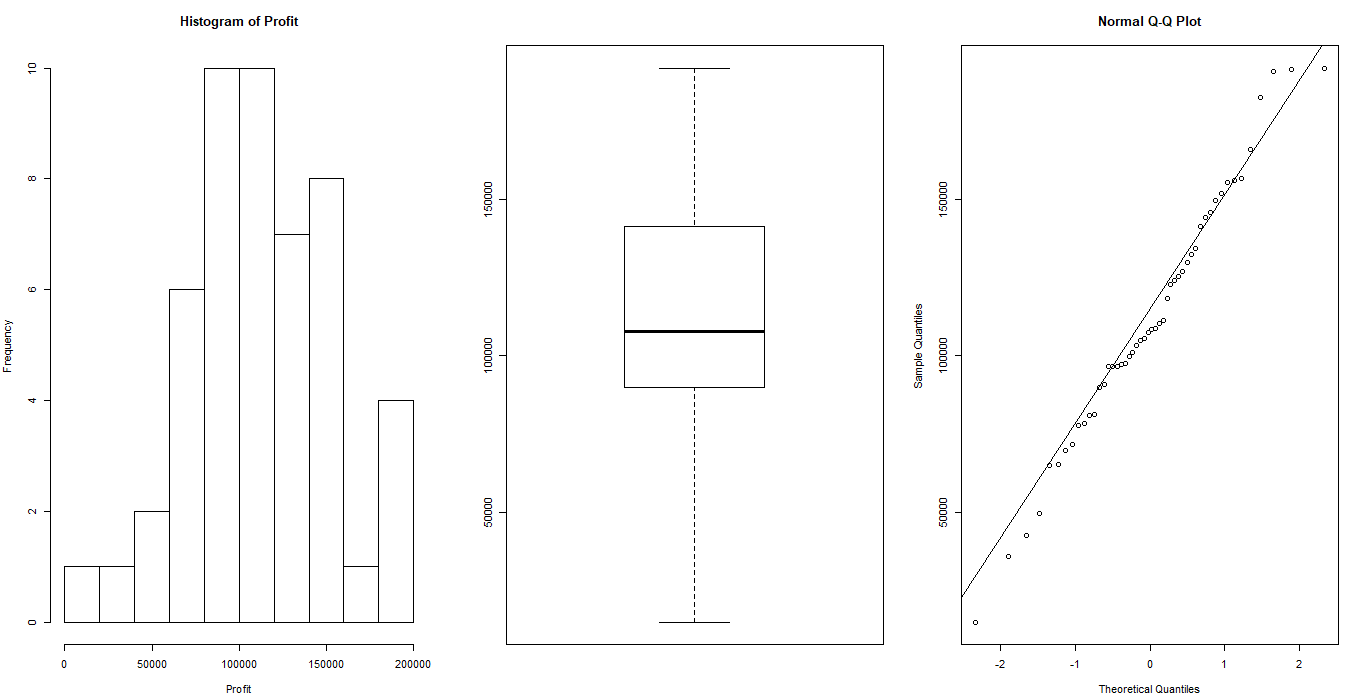
This is not normal distribution with negative skewness. The boxplot doesnot show any outlier.

For Marketing Spend



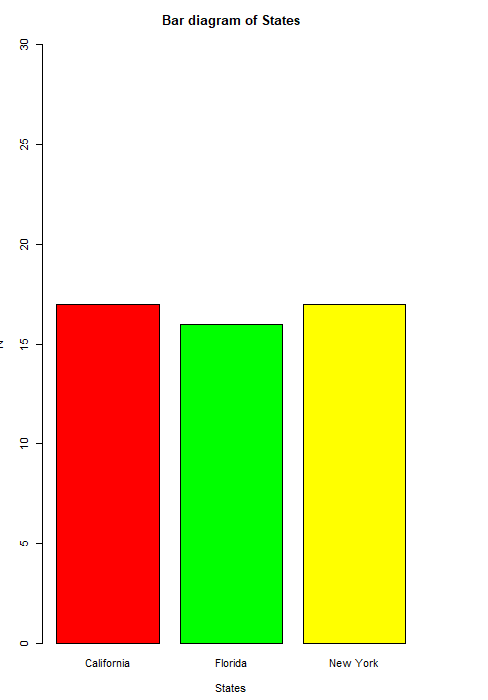
This is not normal distribution and with positive skewness. It doesnot have any outliers.

For Profit



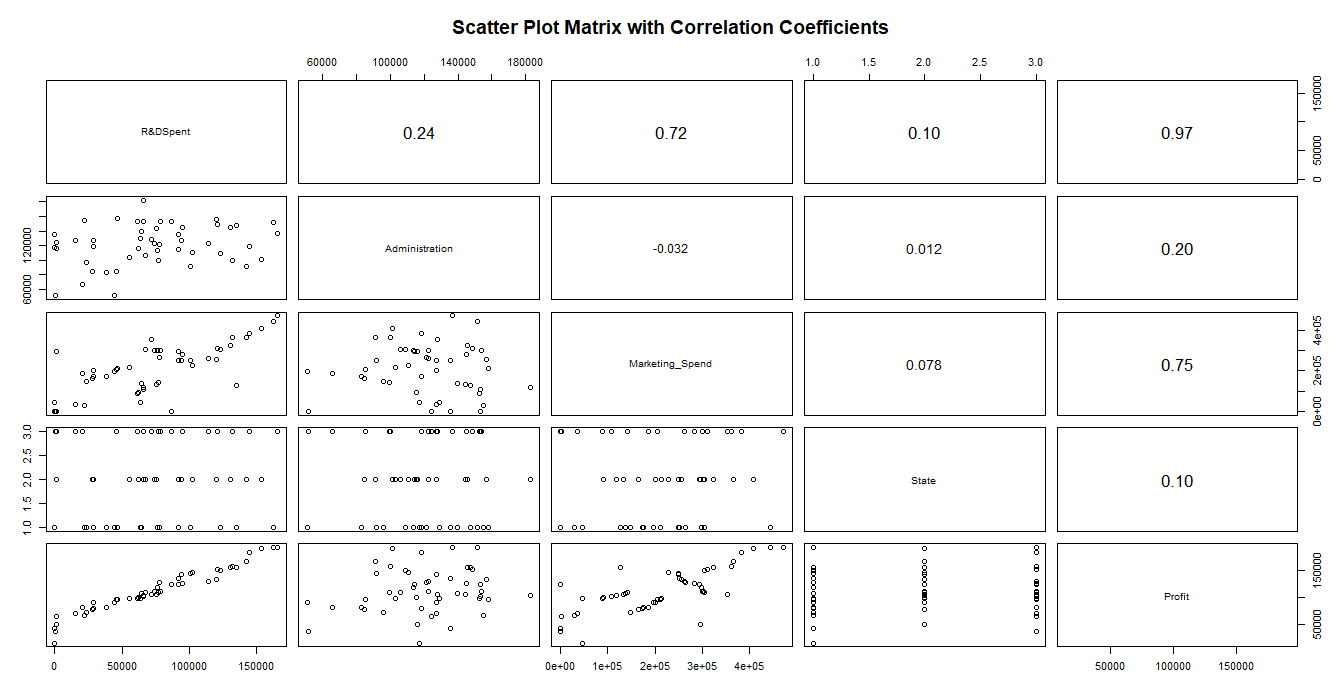
The data is negatively skewed and does not contain any outliers

For State



The above diagram shows that that the data is equally distributed to among all the three categories of the data.

4.) Scatter Plot and Co-relations



The above plots describes the plot among different variables and the correlation matrix among different variables.

There is positive high correlation(r=0.97) and linear relationship between R&D spent and profit.

There is positive moderate correlation(r=0.75) and linear relationship between Marketing\_Spend and profit.

The co-relation coefficient(0.20) between administration and profit is very low and from the scatterplot it is very difficult to conclude anything about the relationship.

**5.) Model Builiding**

The model is built using lm function In R

**Model 1**

X = R&DSpent, Marketing\_Spend, Administration, States

Y = Profit

Summary

Call:

lm(formula = Profit ~ ., data = startup)

Residuals:

Min 1Q Median 3Q Max

-33504 -4736 90 6672 17338

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.013e+04 6.885e+03 7.281 4.44e-09 \*\*\*

`R&DSpent` 8.060e-01 4.641e-02 17.369 < 2e-16 \*\*\*

Administration -2.700e-02 5.223e-02 -0.517 0.608

Marketing\_Spend 2.698e-02 1.714e-02 1.574 0.123

StateFlorida 1.988e+02 3.371e+03 0.059 0.953

StateNew York -4.189e+01 3.256e+03 -0.013 0.990

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9439 on 44 degrees of freedom

Multiple R-squared: 0.9508, Adjusted R-squared: 0.9452

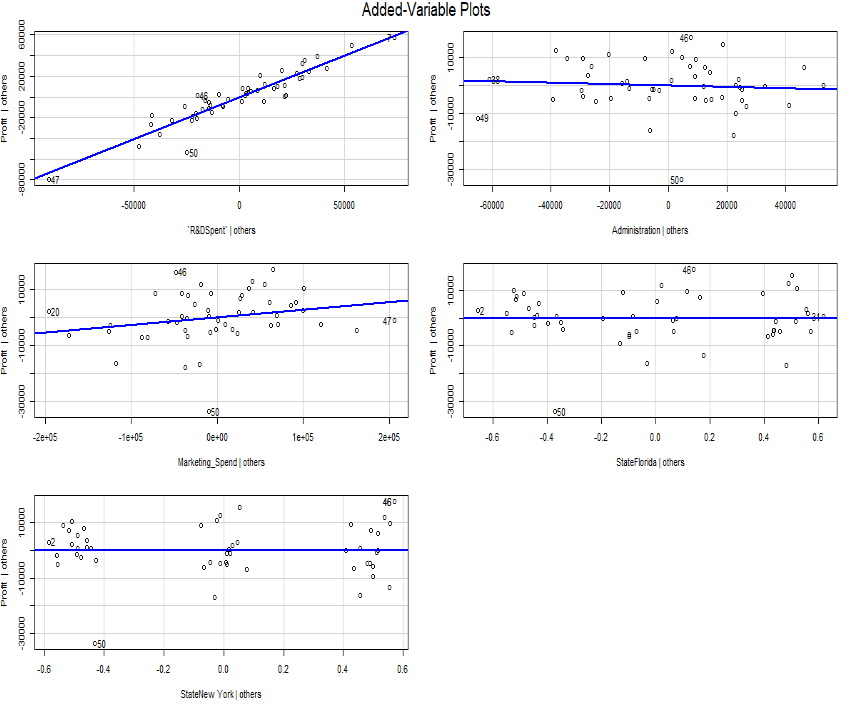
F-statistic: 169.9 on 5 and 44 DF, p-value: < 2.2e-16

From the summary, only R&DSpent is significant and other variables are not significant

The R-squared value is very good.

6.) Regression line

**Profit = 50125.34 + 0.806 R&DSpent - 0.027Administration + 0.0269Marketing\_Spend + 198.8StateFlorida - 41.88702 StateNew York**



7.) Evaluation

As administration is not significant and also there is no good co-relation between the profit and adminstration , the following model is built after ignoring the administration variable.

**Model 2**

X = R&DSpent, Marketing\_Spend, States

Y = Profit

Model Summary:

Call:

lm(formula = startup$Profit ~ startup$`R&DSpent` + startup$Marketing\_Spend +

startup$State, data = startup)

Residuals:

Min 1Q Median 3Q Max

-33621 -4721 -363 6526 17133

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.696e+04 3.119e+03 15.053 <2e-16 \*\*\*

startup$`R&DSpent` 7.967e-01 4.245e-02 18.771 <2e-16 \*\*\*

startup$Marketing\_Spend 2.975e-02 1.615e-02 1.842 0.072 .

startup$StateFlorida 1.408e+02 3.342e+03 0.042 0.967

startup$StateNew York -1.952e+01 3.229e+03 -0.006 0.995

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9362 on 45 degrees of freedom

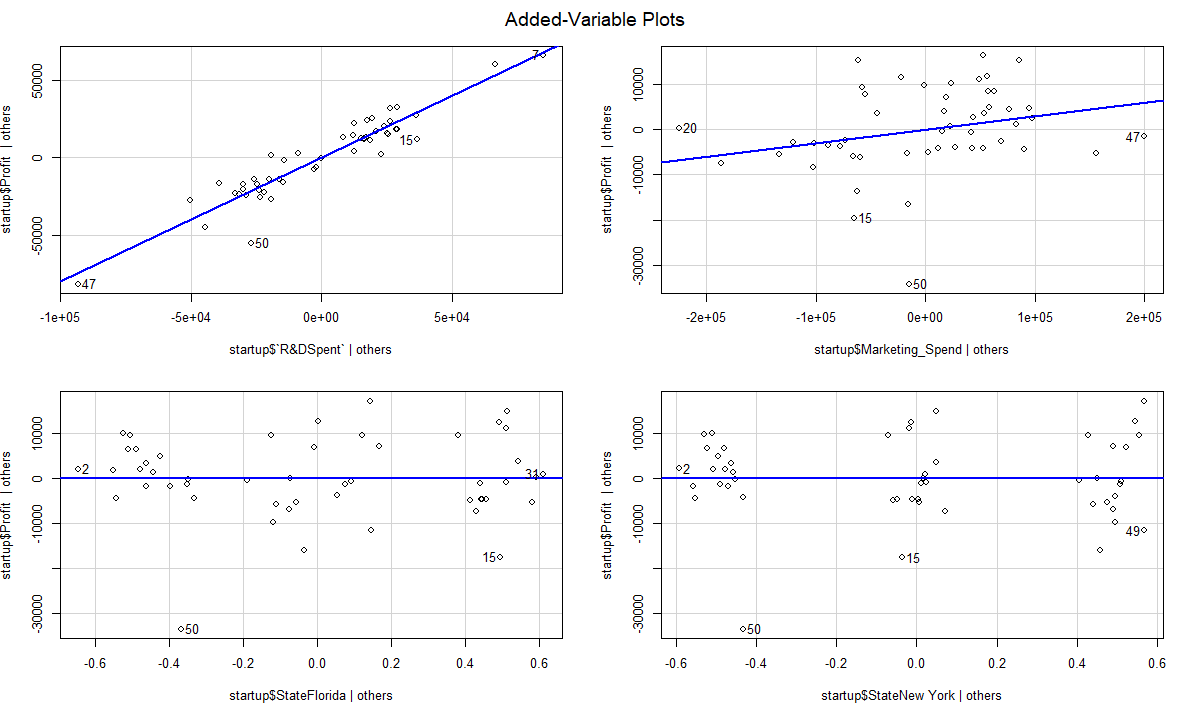
Multiple R-squared: 0.9505, Adjusted R-squared: 0.946

F-statistic: 215.8 on 4 and 45 DF, p-value: < 2.2e-16

From the model summary, it can be observed that p-value for marketing spend is getting

reduced from the previous model p-value. There is decimal reduction in p-value but that

does not matter.



Regression line:

**Profit = 46960+ 0.796 R&DSpent + 0.0297Marketing\_Spend + 140.8StateFlorida – 19.52StateNew York**

**Model 3**

X = R&DSpent, Marketing\_Spend

Y = Profit

Model Summary:

Call:

lm(formula = startup$Profit ~ startup$`R&DSpent` + startup$Marketing\_Spend)

Residuals:

Min 1Q Median 3Q Max

-33645 -4632 -414 6484 17097

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.698e+04 2.690e+03 17.464 <2e-16 \*\*\*

startup$`R&DSpent` 7.966e-01 4.135e-02 19.266 <2e-16 \*\*\*

startup$Marketing\_Spend 2.991e-02 1.552e-02 1.927 0.06 .

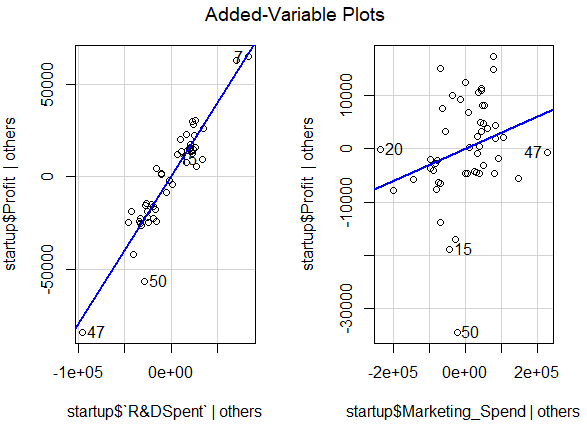
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

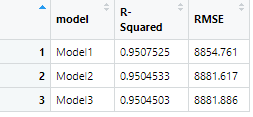
Residual standard error: 9161 on 47 degrees of freedom

Multiple R-squared: 0.9505, Adjusted R-squared: 0.9483

F-statistic: 450.8 on 2 and 47 DF, p-value: < 2.2e-16



The below table summarizes the R-squared value and RMSE value of the models generated above by various combinations.



If we compare the three models then we can say that the first model is the best model with good R-squared value and less RMSE value.

2.) Predict Price of the computer

A dataframe containing :

price : price in US dollars of 486 PCs

speed : clock speed in MHz

hd : size of hard drive in MB

ram : size of Ram in in MB

screen : size of screen in inches

cd : is a CD-ROM present ?

multi : is a multimedia kit (speakers, sound card) included ?

premium : is the manufacturer was a "premium" firm (IBM, COMPAQ) ?

ads : number of 486 price listings for each month

trend : time trend indicating month starting from January of 1993 to November of 1995.

Solution:-

**1.) Business Problem:** To predict price of the computer

**2.) Datasets:**

Independent Variable (x) : Speed (clock speed in MHz), continuous

hd (size of hard drive in MB), continuous

ram (size of Ram in in MB), continuous

Screen (size of screen in inches), continuous

Cd (is a CD-ROM present), categorical

Multi (is a multimedia kit (speakers, sound card) included), categorical

premium (is the manufacturer was a "premium" firm (IBM, COMPAQ)), categorical

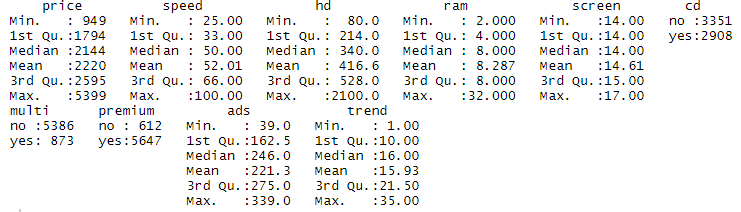
ads (number of 486 price listings for each month), continuous

trend (time trend indicating month starting from January of 1993 to November of 1995), continuous.

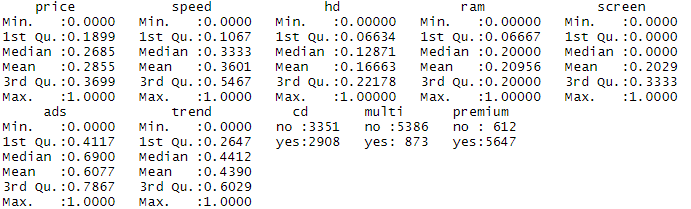
Dependent Variable (y): price (price in US dollars of 486 PCs), continous

**3.) EDA**

**Summary**



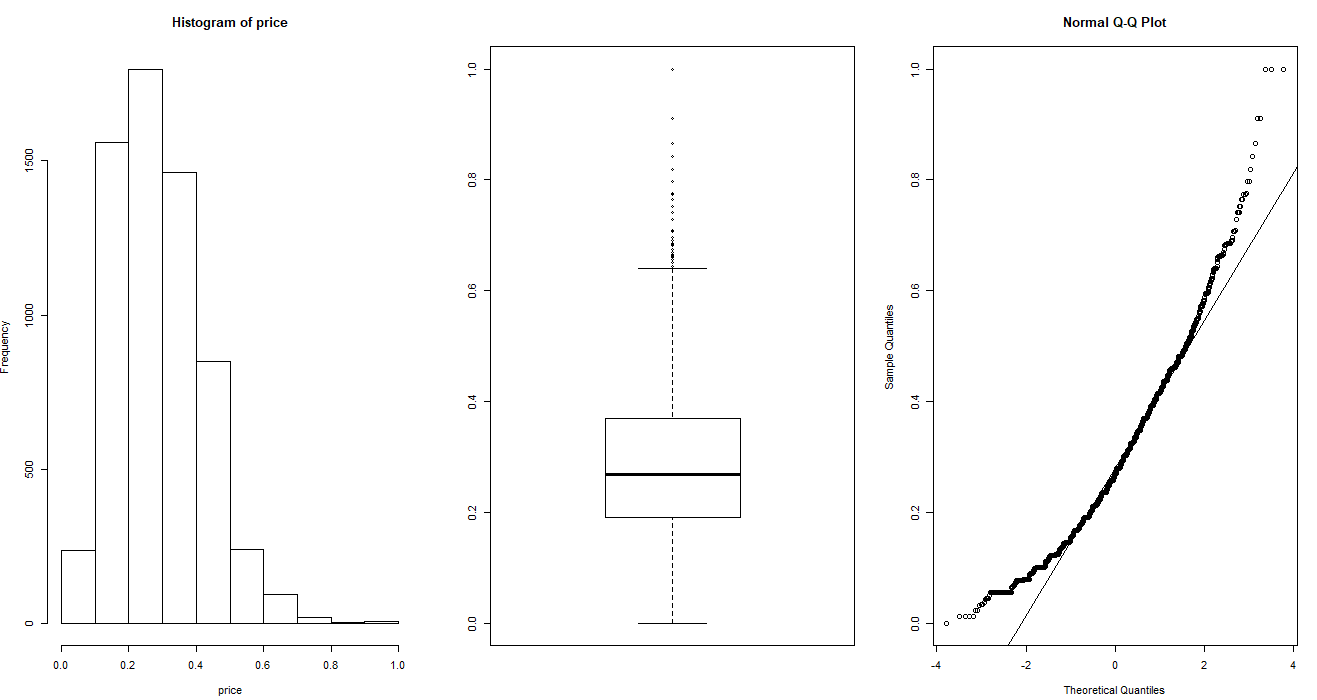
**Applying Normalization to make data scale free and unit less**

Summary 

The data sets does not contain any NA values

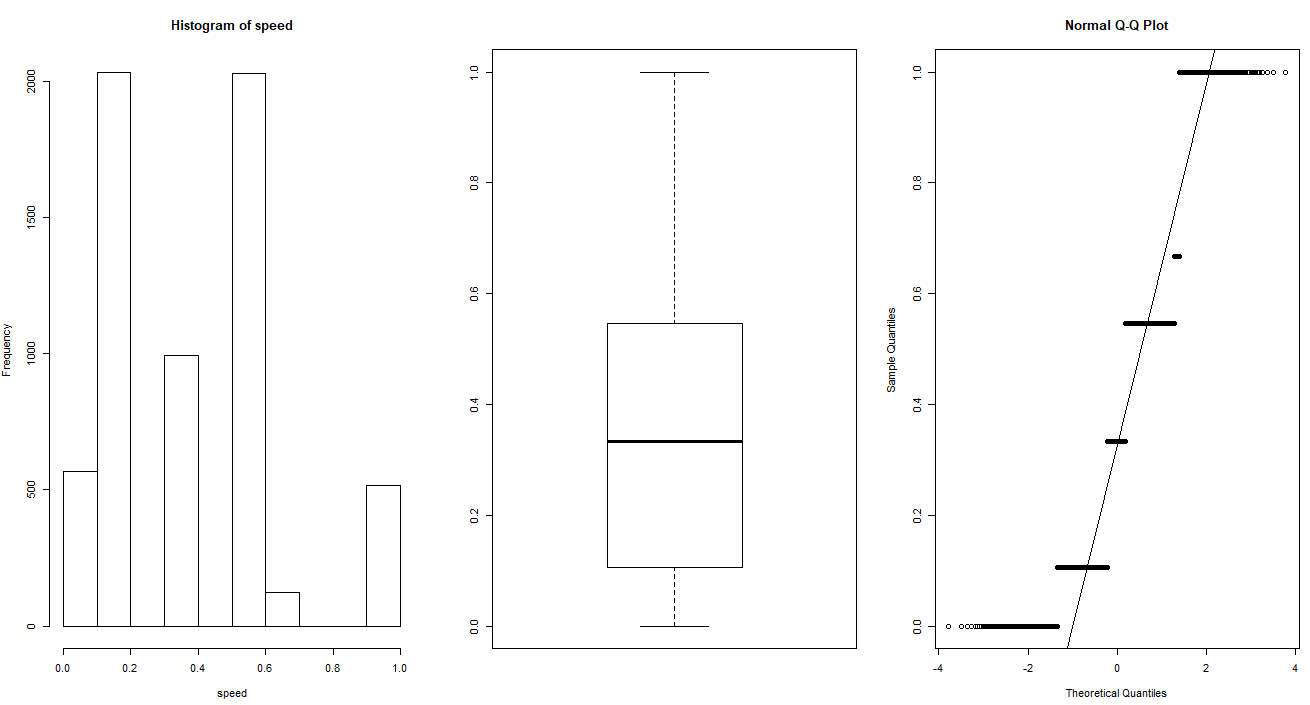
**Graphical Representations**

For Price



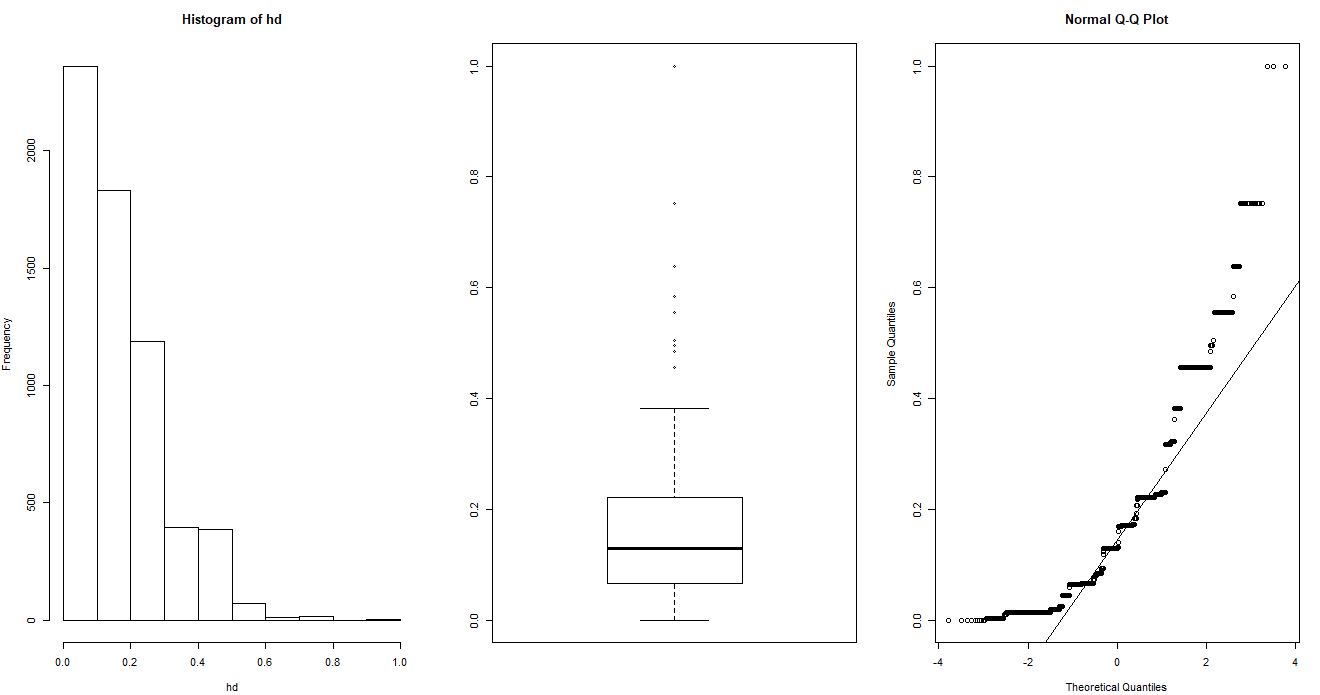
The above graphs describe that the data is not normally distributed with positive skewness and there are lot of outliers.

For Speed



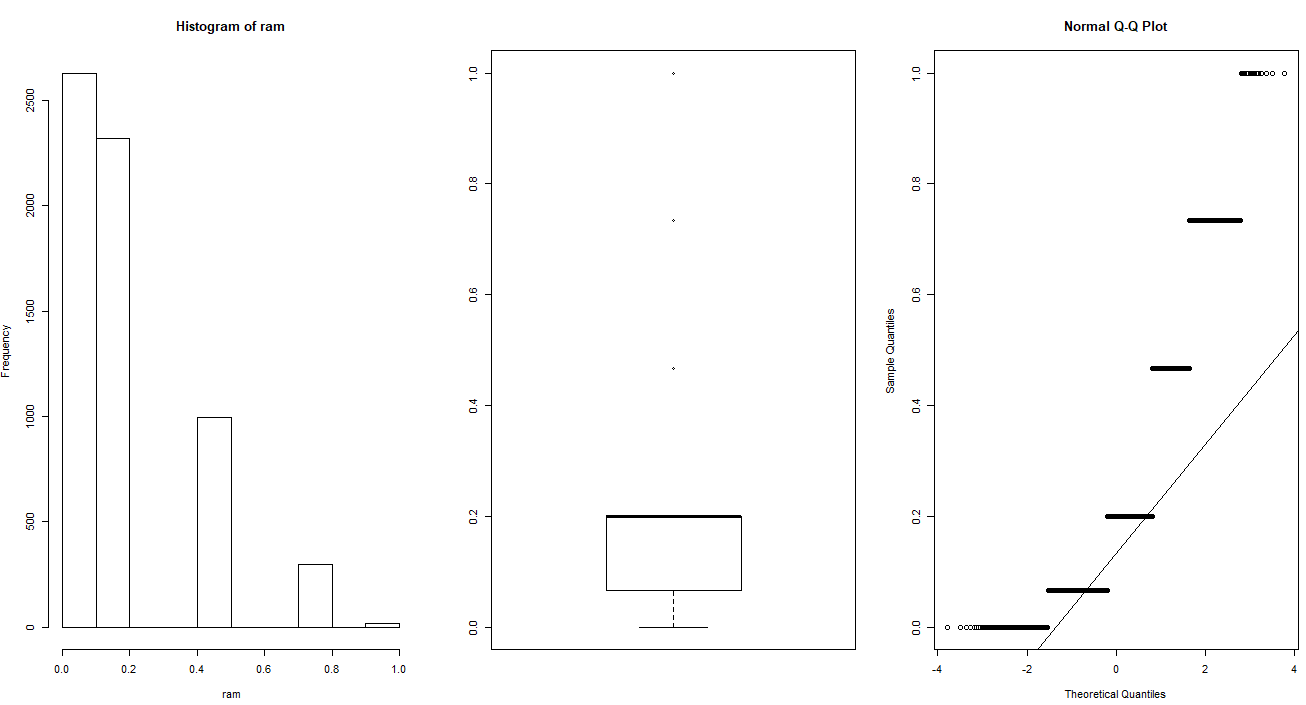
The above graphs describe that the data is positively skewed without any outliers.

For hd



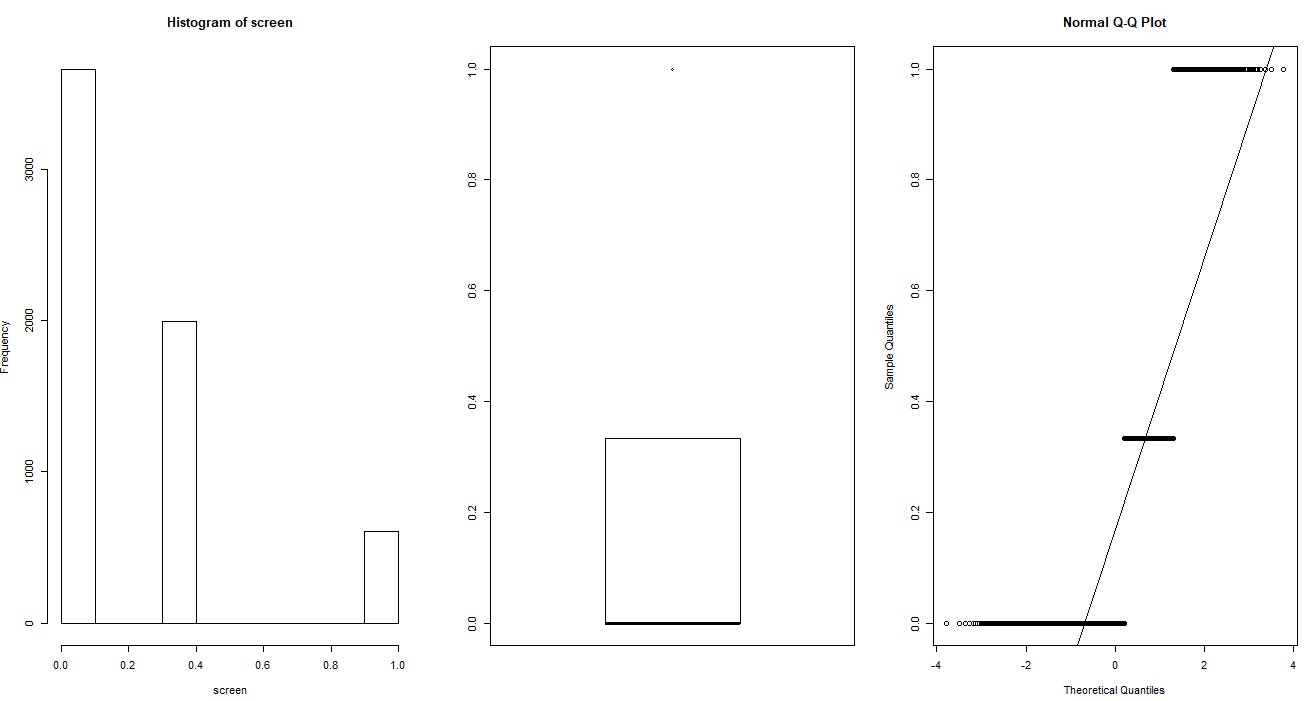
The above graphs describe that the data is not normally distribute and with positive skewness, with peakedness and there are lot of outliers.

For ram,

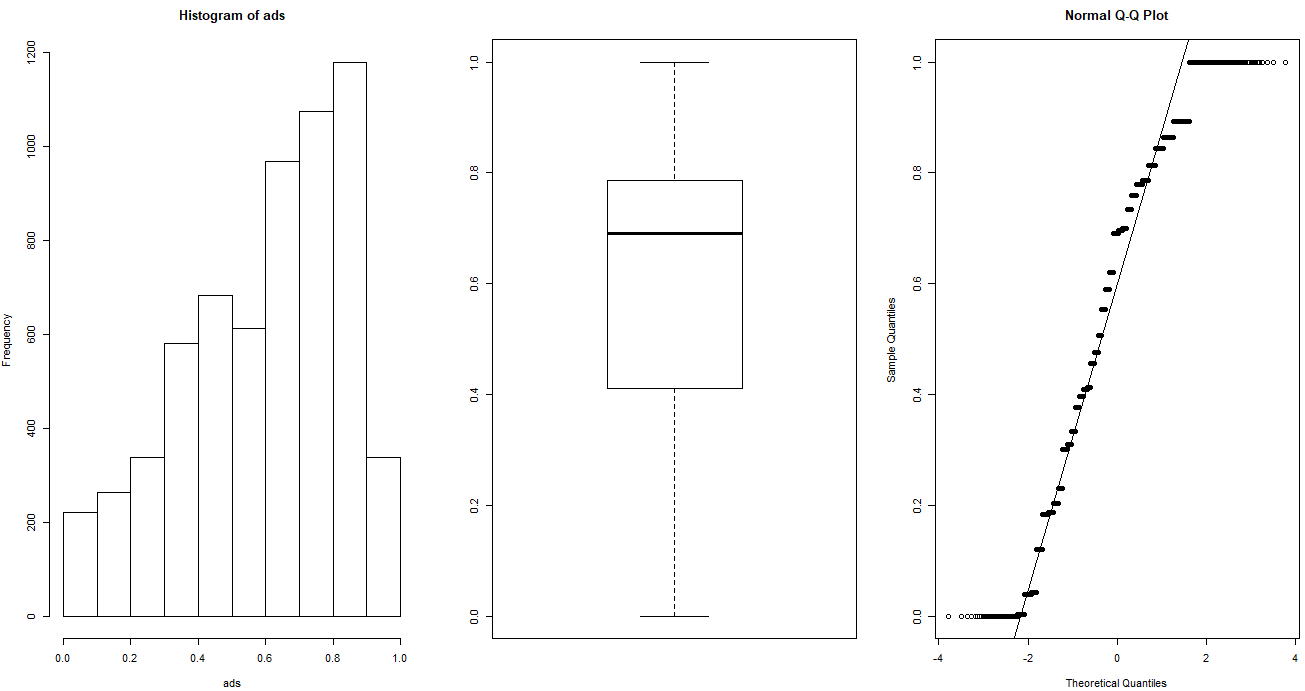


The above graphs describe that the data is not normally distribute and with positive skewness, with peakedness and there are lot of outliers.

For screen,

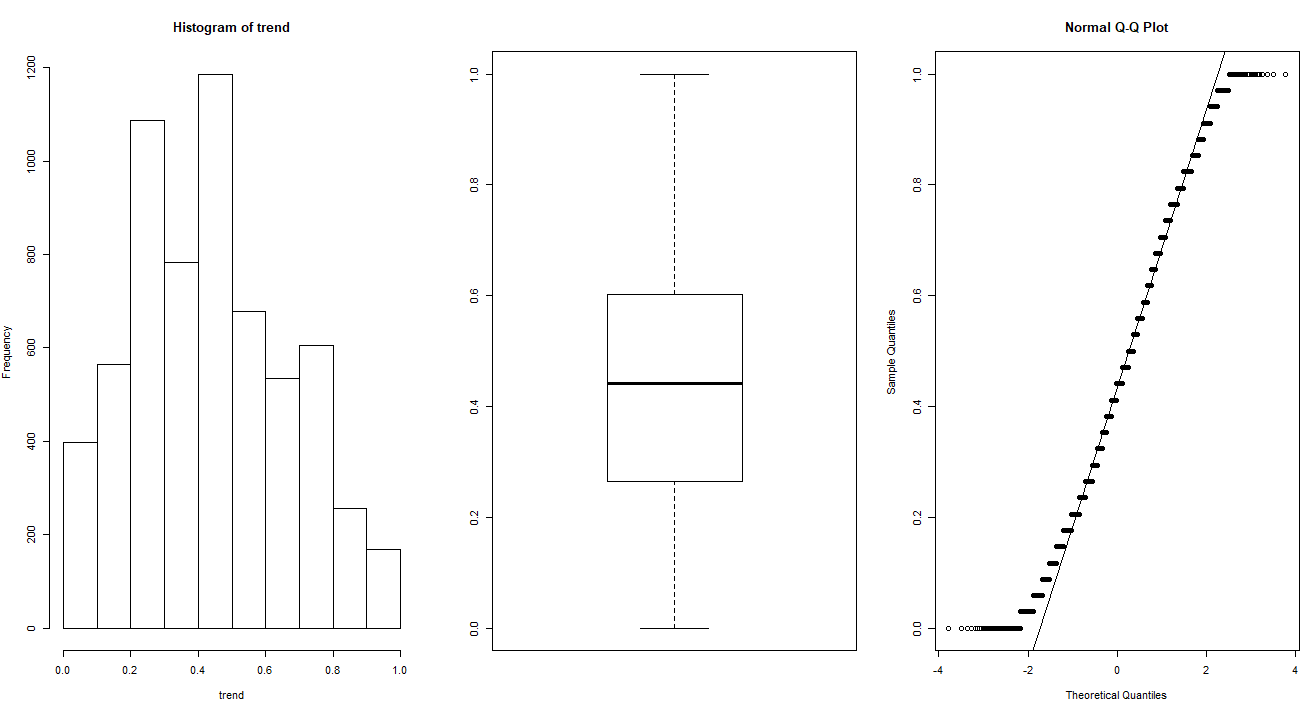


For ads



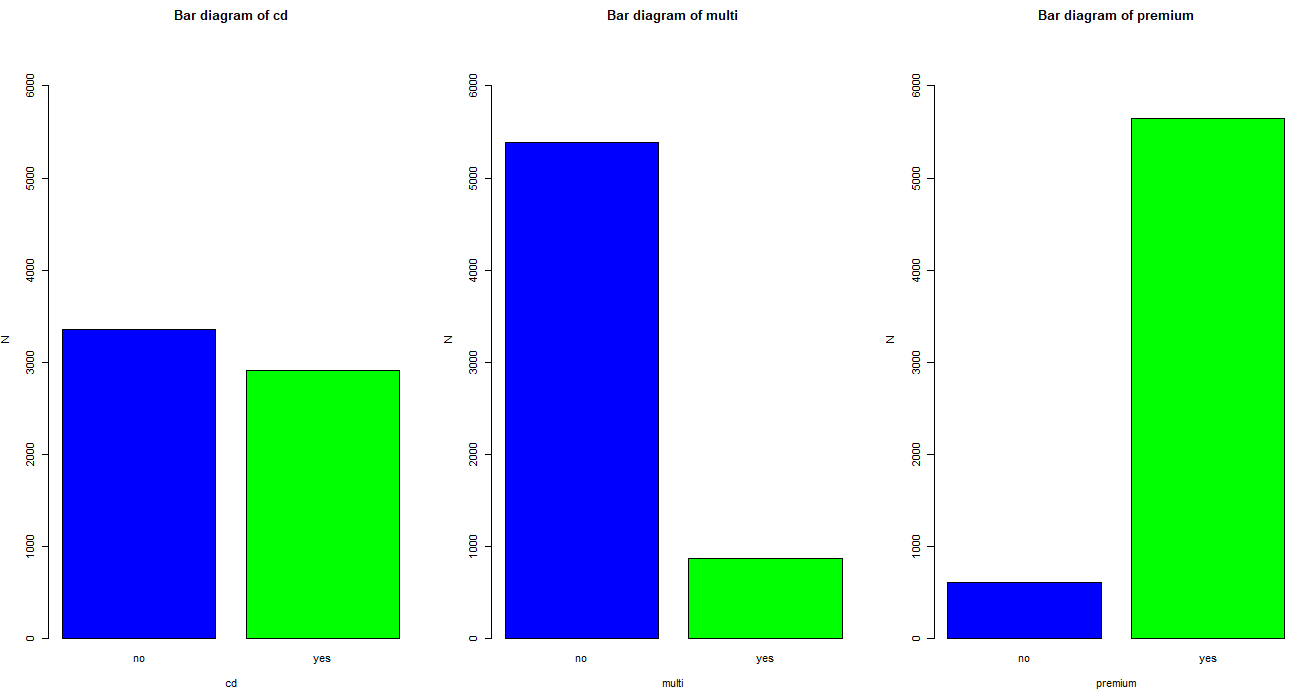
The data is negatively skewed with no outliers.

For trends



The data is not normally distributed but positive skewed with to very little extent but no outliers.

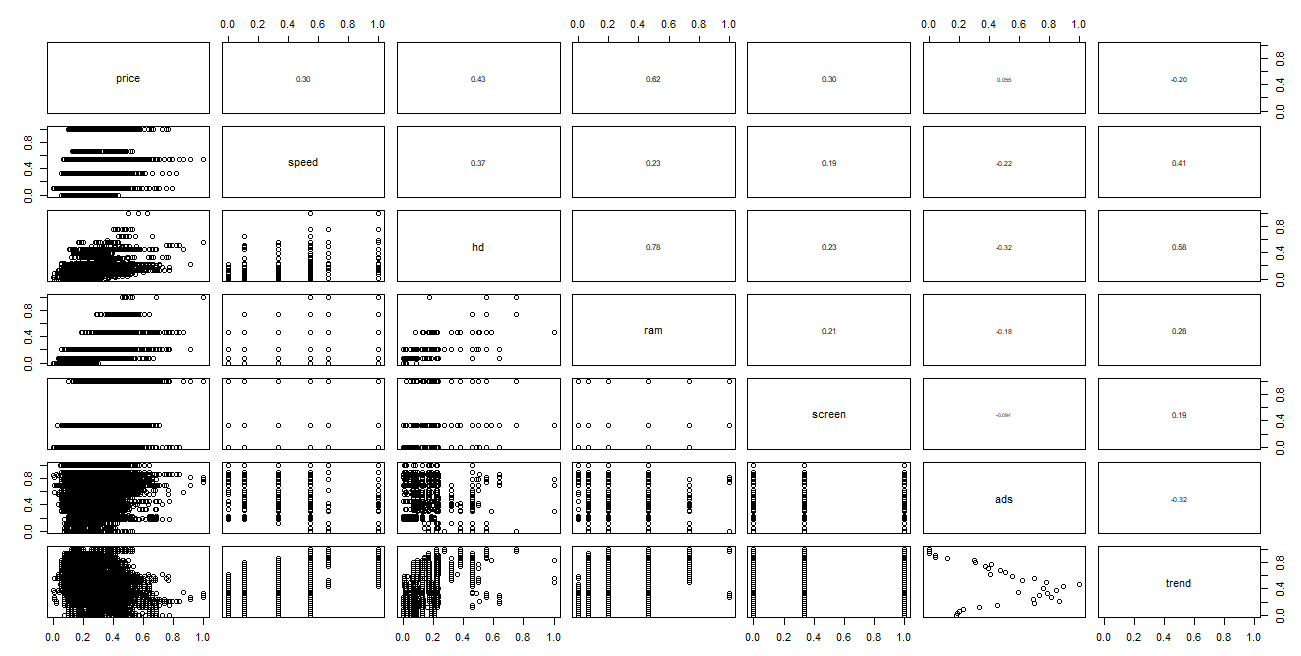
For cd, multi, premium



Outliers

In the above plot, some outliers are found in some variables. But this outliers are not the anomaly inside the datasets. This outliers are have much influence in our response and should be considered for building models

**4.) Scatter plot and co-relation matrix**



From co-relation matrix and scatter plots, it can be conclude that there is no strong co-relation among the variables and even very difficult to interpret about the relationship of predictors with the response variable.

**5.) Model Building**

Model Summary

Call:

lm(formula = price ~ ., data = computer\_data)

Residuals:

Min 1Q Median 3Q Max

-0.24579 -0.03915 -0.00258 0.03292 0.44968

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.325411 0.004215 77.202 < 2e-16 \*\*\*

speed 0.157083 0.003119 50.364 < 2e-16 \*\*\*

hd 0.354875 0.012535 28.311 < 2e-16 \*\*\*

ram 0.325321 0.007187 45.265 < 2e-16 \*\*\*

screen 0.082981 0.002696 30.776 < 2e-16 \*\*\*

ads 0.044312 0.003459 12.809 < 2e-16 \*\*\*

trend -0.396154 0.004804 -82.470 < 2e-16 \*\*\*

cdyes 0.013689 0.002138 6.402 1.65e-10 \*\*\*

multiyes 0.023444 0.002565 9.141 < 2e-16 \*\*\*

premiumyes -0.114433 0.002774 -41.259 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06188 on 6249 degrees of freedom

Multiple R-squared: 0.7756, Adjusted R-squared: 0.7752

F-statistic: 2399 on 9 and 6249 DF, p-value: < 2.2e-16

From the above summary, R-squared value is 0.7756 and all the variables are significant.

RMSE =0.0618

**6.) Regression Line**

**Price = 0.32541139 + 0.15708338speed + 0.35487532hd + 0.32532107ram +0.08298138screen + 0.04431177ads -0.39615410trend + 0.01368915cdyes + 0.02344355 multiyes - 0.11443252premiumyes**

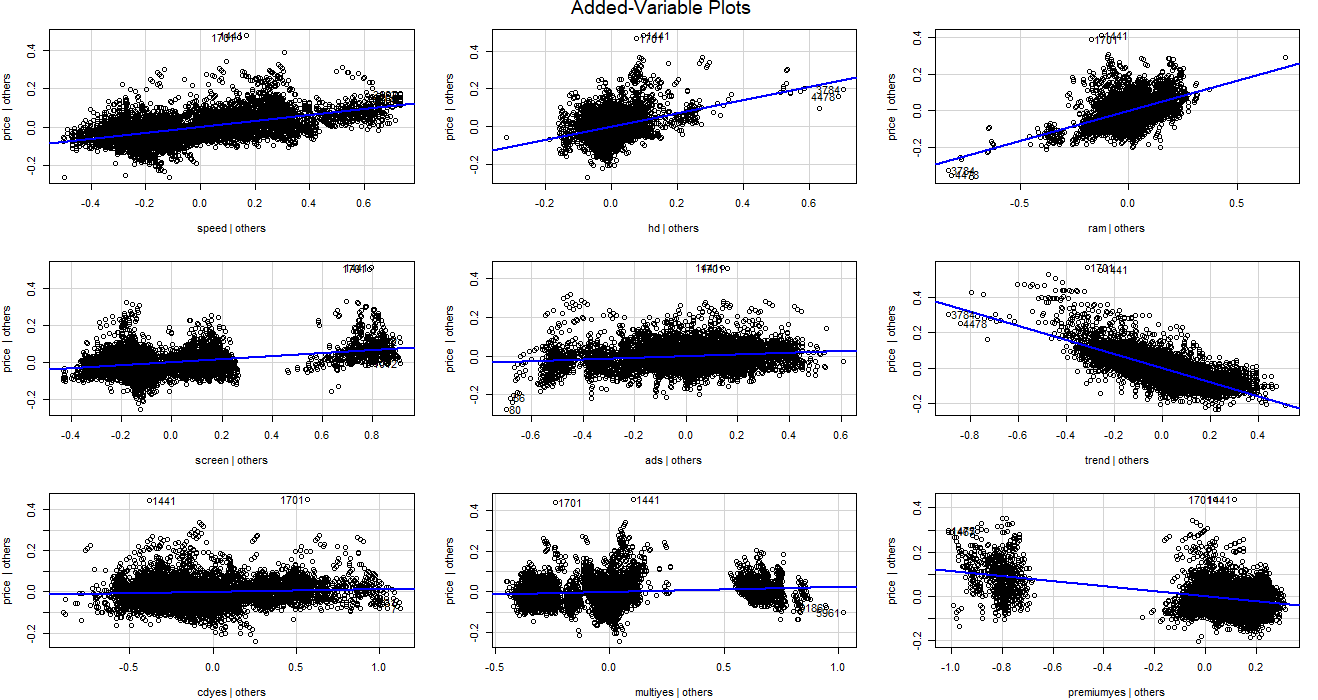
**7.) VIF**

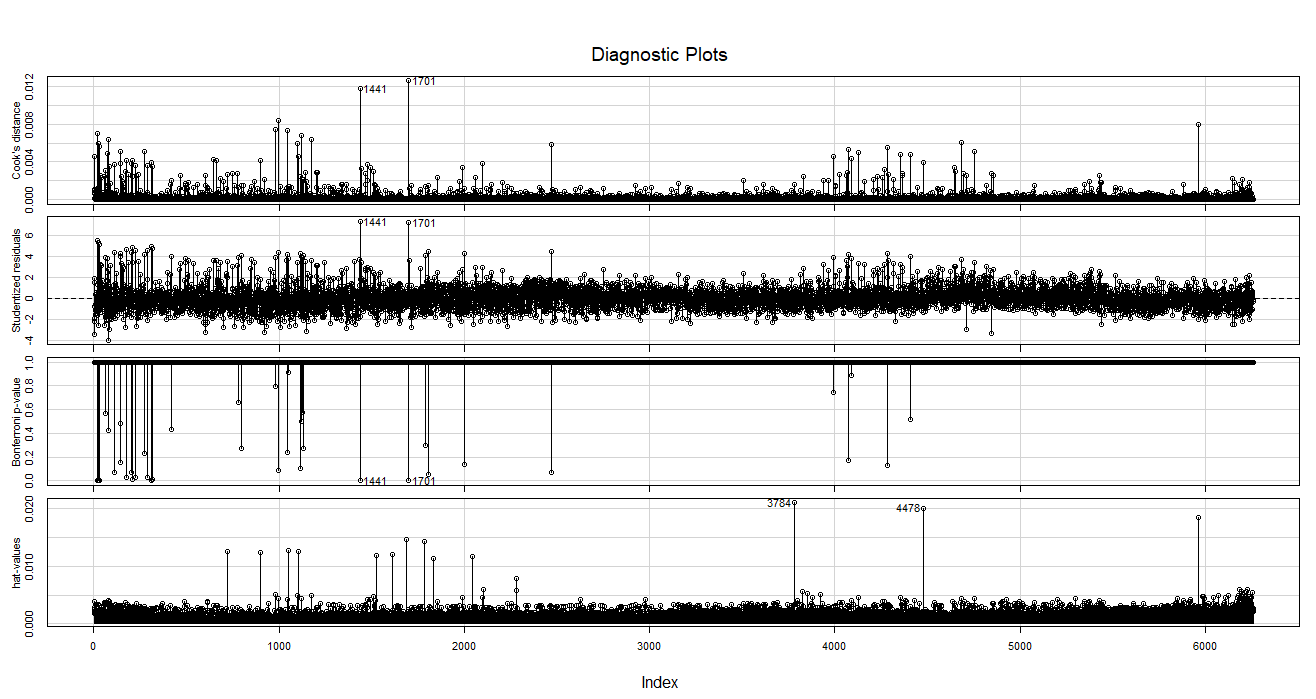
speed hd ram screen ads trend cd multi premium

1.265364 4.207395 2.974628 1.081644 1.217218 2.022790 1.859370 1.290568 1.109388

VIF < 10 for all the variables so there is no co-linearity problems.

**8.) AVPlots**



**9.) Identifying influential values**

The above influence index plot identifies two data points namely, 1441 and 1701 as the

most influencing value. There for we create another model after removing the two values.

**Model 2**

Model Summary

Call:

lm(formula = price ~ ., data = computer\_data[-c(1701), ])

Residuals:

Min 1Q Median 3Q Max

-0.24600 -0.03907 -0.00252 0.03292 0.45046

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.325459 0.004198 77.526 < 2e-16 \*\*\*

speed 0.156924 0.003106 50.515 < 2e-16 \*\*\*

hd 0.353518 0.012486 28.313 < 2e-16 \*\*\*

ram 0.326347 0.007160 45.582 < 2e-16 \*\*\*

screen 0.082318 0.002687 30.635 < 2e-16 \*\*\*

ads 0.044089 0.003446 12.796 < 2e-16 \*\*\*

trend -0.395322 0.004786 -82.605 < 2e-16 \*\*\*

cdyes 0.013389 0.002130 6.286 3.49e-10 \*\*\*

multiyes 0.023625 0.002554 9.249 < 2e-16 \*\*\*

premiumyes -0.114467 0.002762 -41.438 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06163 on 6248 degrees of freedom

Multiple R-squared: 0.7766, Adjusted R-squared: 0.7763

F-statistic: 2413 on 9 and 6248 DF, p-value: < 2.2e-16

R2 = 0.7766

RMSE = 0.06157

R2 value increased and RMSE value decreased by decimal values

Model 3 - sqaureroot Model

x: speed, hd, ram,screen,ads,trend,cd,multi,premium

y : √price

Model summary

Call:

lm(formula = sqrt(price) ~ ., data = computer\_data[-c(1701, 1401),

])

Residuals:

Min 1Q Median 3Q Max

-0.36223 -0.03505 0.00105 0.03486 0.27249

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.560943 0.003921 143.075 <2e-16 \*\*\*

speed 0.151883 0.002901 52.349 <2e-16 \*\*\*

hd 0.330229 0.011660 28.320 <2e-16 \*\*\*

ram 0.296460 0.006686 44.337 <2e-16 \*\*\*

screen 0.077463 0.002509 30.869 <2e-16 \*\*\*

ads 0.038495 0.003218 11.963 <2e-16 \*\*\*

trend -0.382198 0.004470 -85.508 <2e-16 \*\*\*

cdyes 0.022862 0.001990 11.488 <2e-16 \*\*\*

multiyes 0.022873 0.002386 9.587 <2e-16 \*\*\*

premiumyes -0.108599 0.002580 -42.097 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05755 on 6247 degrees of freedom

Multiple R-squared: 0.7811, Adjusted R-squared: 0.7808

F-statistic: 2477 on 9 and 6247 DF, p-value: < 2.2e-16

R2 = 0.7811, RMSE = 0.0577

Among the three models squareroot model is the best model because R-squared model is more and RMSE is less

The regression line can be :

**√price = 0.56094310 + 0.15188277speed + 0.33022899 hd + 0.29645958ram + 0.07746256screen + 0.03849523 ads -0.38219843 trend + 0.02286196cdyes + 0.02287263multiyes -0.10859870 premiumyes**

**3.)**

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

Solution: -

**1.) Business problem:** To prepare a model for predicting car price

**2.) Datasets:**

Input variable (x):

x1 : Age\_08\_04

X2 : KM

X3 : HP

X4 : cc

X5 : Doors

X6 : Gears

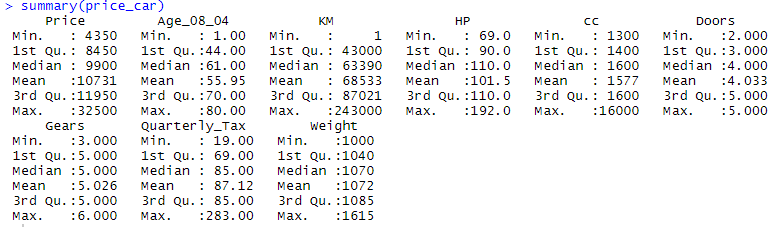
X7 : Quarterly\_Tax

X8 : Weight

Output variable(y): Price

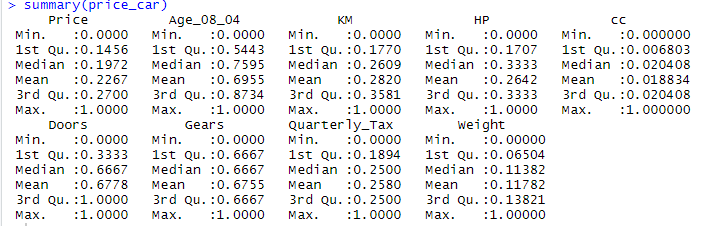
**3.) EDA**

Data summary



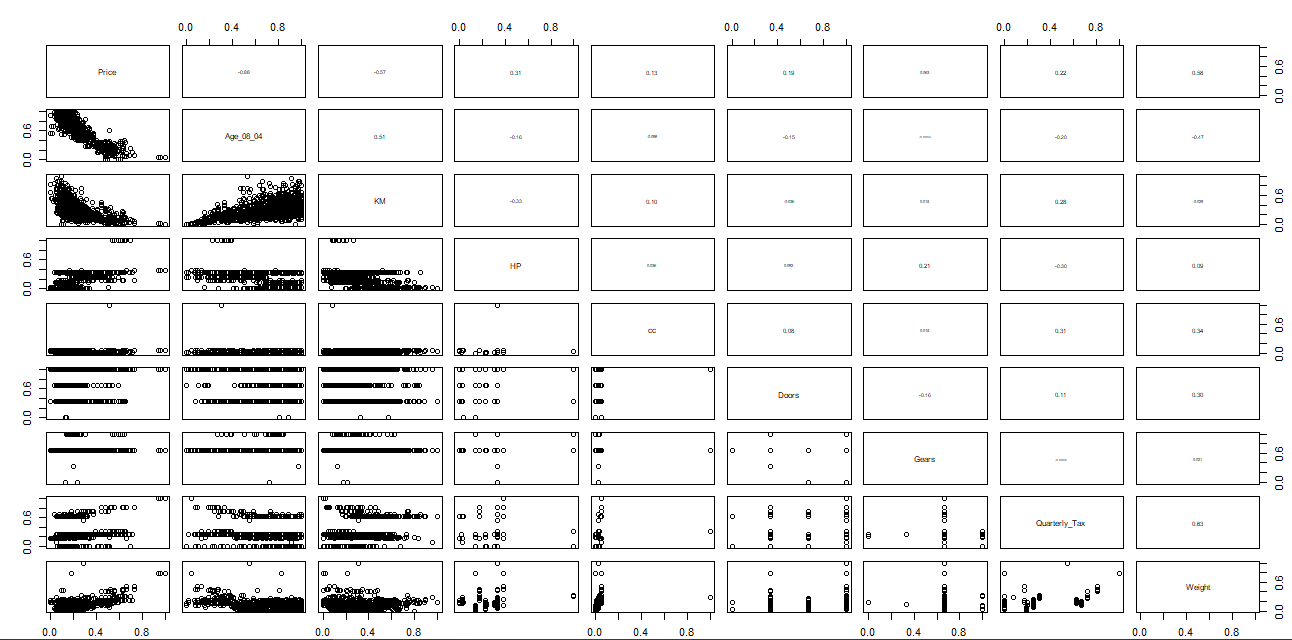
**Missing data analysis**

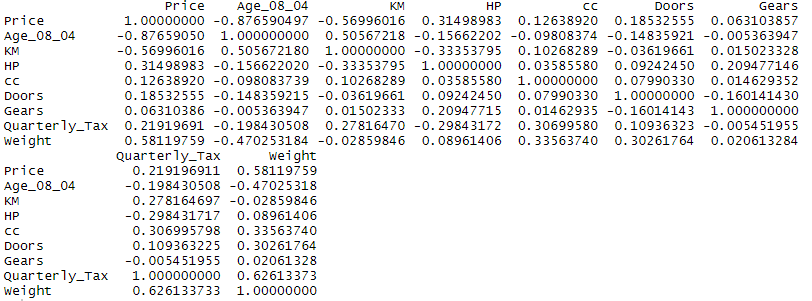
There are no missing data in the Datasets

**Normalization: -** After applying the normalization function data is unit less and scale free.foSummary of the data after normalization 

Graphical Representation

**4.) Scatter Plot and correlation matrix**





**5.) Model Building**

Model1

Model Summary

Call:

lm(formula = Price ~ ., data = price\_car)

Residuals:

Min 1Q Median 3Q Max

-0.33273 -0.02818 -0.00076 0.02841 0.22892

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.3835580 0.0154607 24.809 < 2e-16 \*\*\*

Age\_08\_04 -0.3414214 0.0073405 -46.512 < 2e-16 \*\*\*

KM -0.1796995 0.0108112 -16.622 < 2e-16 \*\*\*

HP 0.1384281 0.0123146 11.241 < 2e-16 \*\*\*

cc -0.0632389 0.0470450 -1.344 0.17909

Doors -0.0001723 0.0042635 -0.040 0.96777

Gears 0.0633378 0.0210006 3.016 0.00261 \*\*

Quarterly\_Tax 0.0370358 0.0122857 3.015 0.00262 \*\*

Weight 0.3704994 0.0233315 15.880 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04769 on 1427 degrees of freedom

Multiple R-squared: 0.8638, Adjusted R-squared: 0.863

F-statistic: 1131 on 8 and 1427 DF, p-value: < 2.2e-16

R-squared = 0.8637

RMSE =0.04754

All variables are significant accept Doors and cc

**6.) Regression Line**

Price= 0.3835579906 - 0.3414214472Age\_08\_04 - 0.1796995230KM + 0.1384281143 HP -0.0632388781 cc - 0.0001722886Doors + 0.0633378263 Gears + 0.037035784Quarterly + 0.3704994167 Weight

**7.) VIF**

Age\_08\_04 KM HP cc Doors Gears Quarterly\_Tax Weight

1.884620 1.756905 1.419422 1.163894 1.156575 1.098723 2.311431 2.516420

VIF < 10 for all variables, so there is no collinearity problem.

**8.) Evaluation**

Now removing doors variable, as the variable is not significant for predicting price another

model is built

**Model 2**

Model Summary

Call:

lm(formula = Price ~ ., data = price\_car[, -6])

Residuals:

Min 1Q Median 3Q Max

-0.33259 -0.02815 -0.00076 0.02846 0.22900

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.383362 0.014678 26.118 < 2e-16 \*\*\*

Age\_08\_04 -0.341421 0.007338 -46.528 < 2e-16 \*\*\*

KM -0.179712 0.010803 -16.636 < 2e-16 \*\*\*

HP 0.138393 0.012279 11.270 < 2e-16 \*\*\*

cc -0.063208 0.047022 -1.344 0.17909

Gears 0.063499 0.020611 3.081 0.00210 \*\*

Quarterly\_Tax 0.037073 0.012247 3.027 0.00251 \*\*

Weight 0.370263 0.022576 16.401 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04767 on 1428 degrees of freedom

Multiple R-squared: 0.8638, Adjusted R-squared: 0.8631

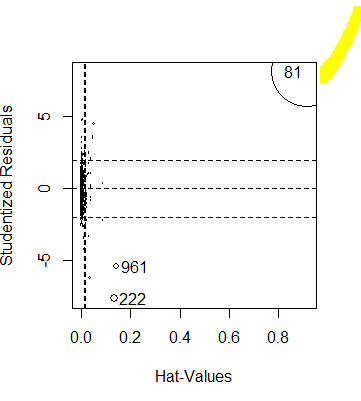
F-statistic: 1293 on 7 and 1428 DF, p-value: < 2.2e-16

R-squared = 0.8638,

RMSE = 0.04754029

There is no much improvement in R-squared and RMSE value, And cc is not significant also

Lets finding the Influencing factor



We find 81 as the influencing factor. Lets builid a model removing the 81th record.

**Model 3**

Summary

Call:

lm(formula = Price ~ Age\_08\_04 + KM + HP + cc + Gears + Quarterly\_Tax +

Weight, data = price\_car[-81, ])

Residuals:

Min 1Q Median 3Q Max

-0.40399 -0.02700 -0.00094 0.02641 0.24076

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.370961 0.014435 25.699 < 2e-16 \*\*\*

Age\_08\_04 -0.338052 0.007188 -47.031 < 2e-16 \*\*\*

KM -0.154465 0.011010 -14.029 < 2e-16 \*\*\*

HP 0.171104 0.012663 13.512 < 2e-16 \*\*\*

cc -1.309096 0.159878 -8.188 5.83e-16 \*\*\*

Gears 0.058586 0.020165 2.905 0.00373 \*\*

Quarterly\_Tax 0.085117 0.013353 6.374 2.47e-10 \*\*\*

Weight 0.436121 0.023514 18.547 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04662 on 1427 degrees of freedom

Multiple R-squared: 0.8693, Adjusted R-squared: 0.8687

F-statistic: 1356 on 7 and 1427 DF, p-value: < 2.2e-16

Here all the variables are significant

R-square value= 0.8693

RMSE = 0.0464

This is the acceptable value and the model is the best model among all the three models.

AIC=-8790.61

Model is

**Price = 0.37096076 -0.33805179Age\_08\_04 -0.15446458KM + 0.17110438HP-1.30909601cc + 0.05858591Gears + 0.08511707Quarterly\_Tax +0.43612070 Weight**

**8.) AV plots**

