MLRAMAN

CONTENTS

QUESTION 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.

Answer:

Variables available: “R.D.Spend”, Administration, “Marketing.Spend” , State Profit

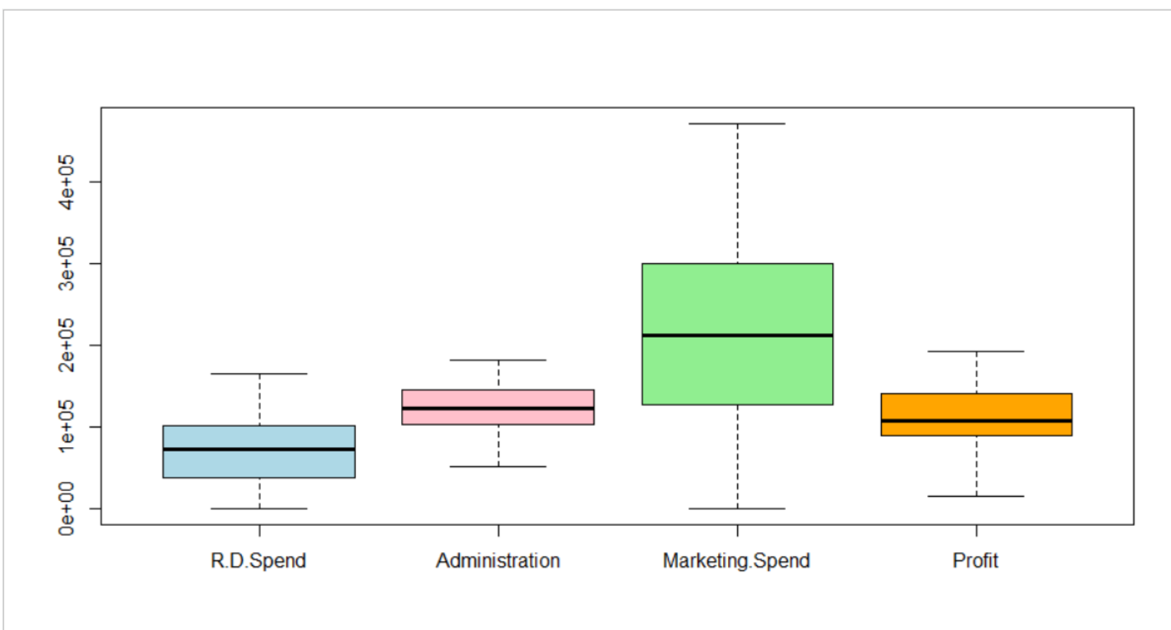
Target Variable: Profit

Let’s have a look on the Summary Statistics of the Variables

R.D.Spend	Administration	Marketing.Spend	State	Profit
Min.: 0	Min.: 51283	Min.: 0	California:17	Min.: 14681
1st Qu.: 39936	1st Qu.:103731	1st Qu.:129300	Florida :16	1st Qu.: 90139
Median: 73051	Median :122700	Median :212716	New York :17	Median :107978
Mean.: 73722	Mean :121345	Mean :211025		Mean :112013
3rd Qu.:101603	3rd Qu.:144842	3rd Qu.:299469		3rd Qu.:139766
Max.:165349	Max. :182646	Max. :471784		Max. :192262

Here we have 4 variables (including the Target Variable) of Continuous type. & one variable of Categorical type i.e. “State”

BOXPLOT:



Looking at the Boxplot we can say that no variables contains outliers.
Now lets move to the Correlation analysis between the continuous variables present in the data.

CORRELATION

Let’s have a look on the correlation coefficient between the variables:

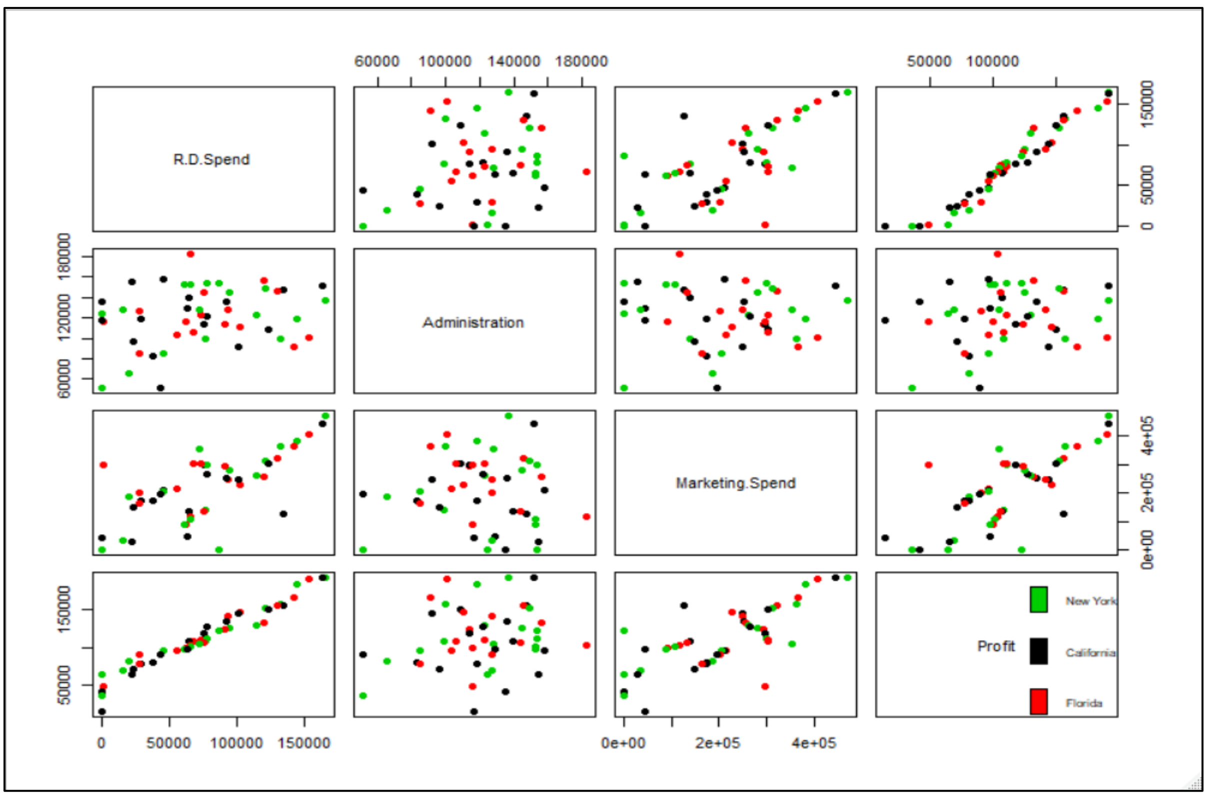
	R.D.Spend	Administration	Marketing.Spend	Profit
R.D.Spend	1	0.241955245	0.724248133	0.972900466
Administration	0.241955245	1	-0.032153875	0.200716568
Marketing.Spend	0.724248133	-0.032153875	1	0.747765722
Profit	0.972900466	0.200716568	0.747765722	1

Seems like Except the pair (“Administrative” and “Marketing Spend”) all the variables are positively correlated with each other.

And there may be no collinearity problem in our Independent variables .

PAIRS PLOT

Let’s Have a look on the Pairs Plot for visualizing the Scatterings of the data among themselves.



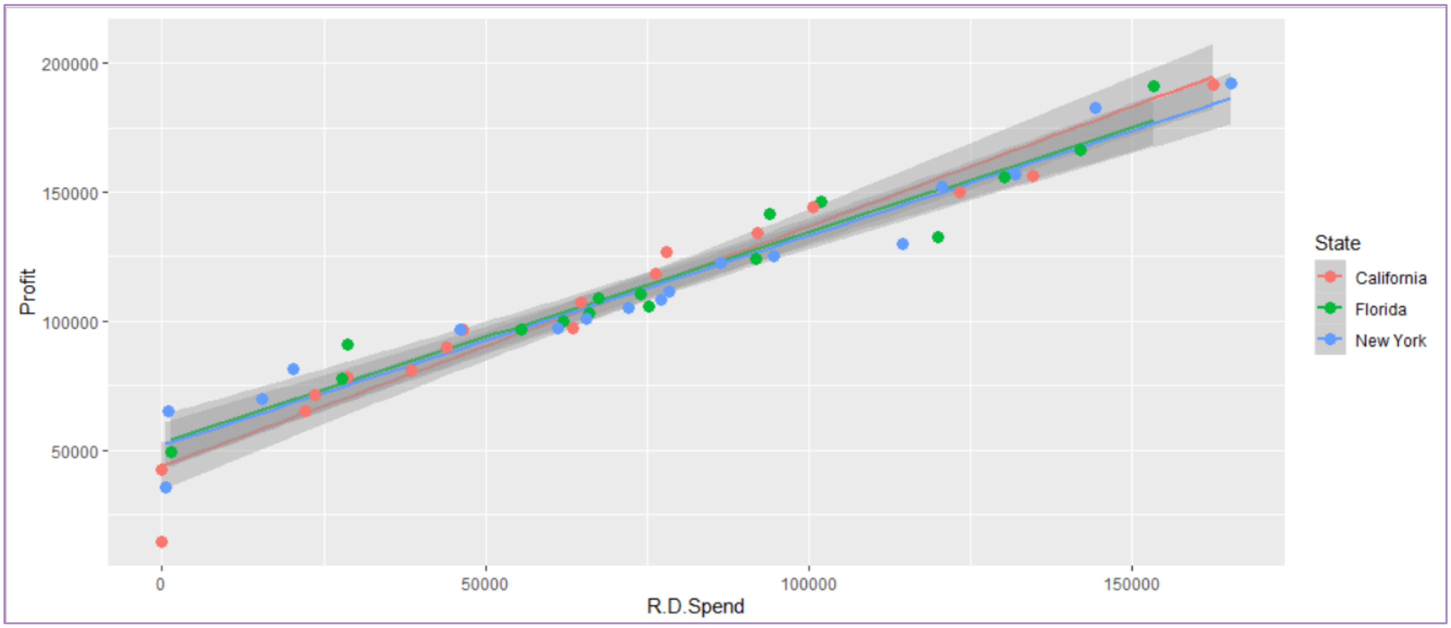
FITTING REGRESSION MODEL:

MODEL 1:

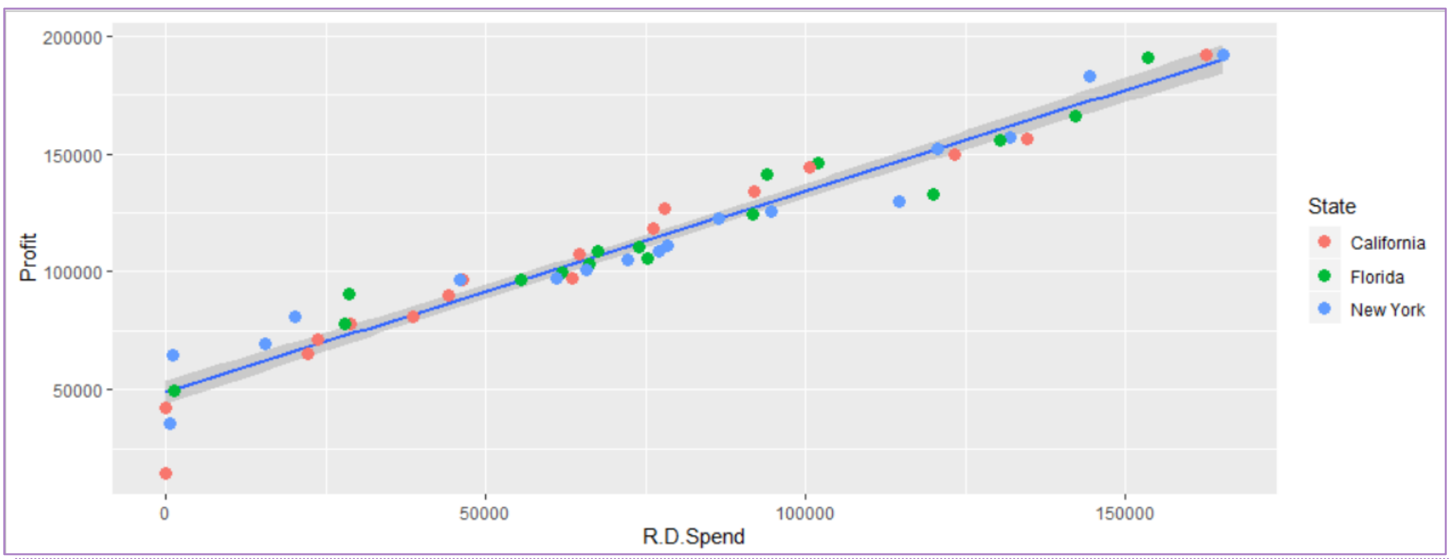
model.S <- lm(Profit~R.D.Spend+Administration+Marketing.Spend)

We get R^2 value as 0.9507. Which convey that 95% of variation in the "profit" is explained by the Independent variables in our model. Where we found that variable "Administration", is not significant in our model, where the variable "Marketing Spend" is somewhat significant with significance level of 0.1 i.e. 90% confidence level.

NOW LET'S HAVE A LOOK ON THE VARIABLE "STATE"

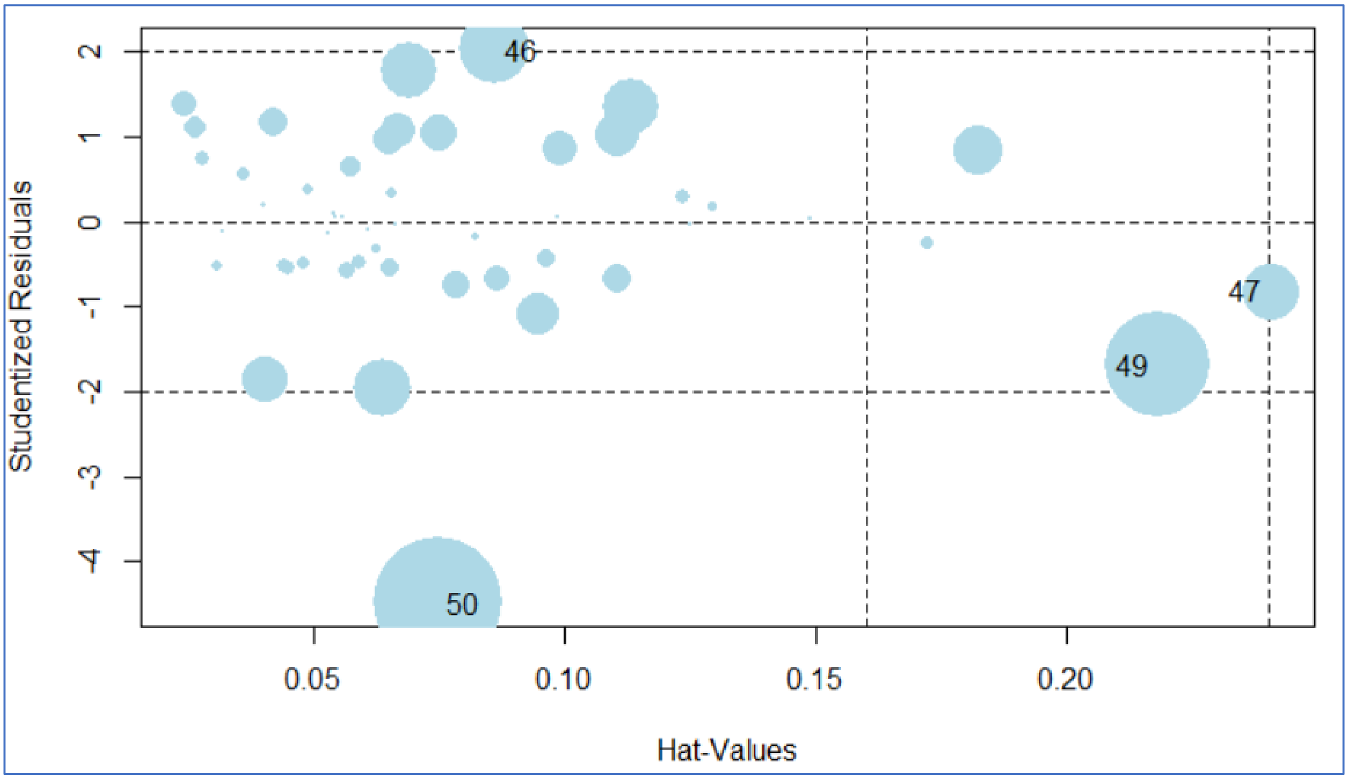


Here in the above plot we can see clearly that, If we plot 3 simple linear regression, taking the variable "R.D.Spend" as independent and "Profit" as dependent. And plotting 3 regression lines for the 3 individual "State" data i.e. "California", "Florida" and "New York" than we can see that all the plots are overlapping with each other in the same confidence belt. The difference between the plots is negligible. Which may be considered as one regression line over there. As from above plot we can say that if we are not considering the state variable in our plot, then also we are getting the similar accuracy. So, we may not require to consider the "State" variable in our model.



INFLUENCE PLOT:

Now we will look on the observations which are highly influencing our model fitting.



Here we can see that observation no 50.49,47,46 are our influence indexes. So, we may remove them for getting more accuracy in our model.

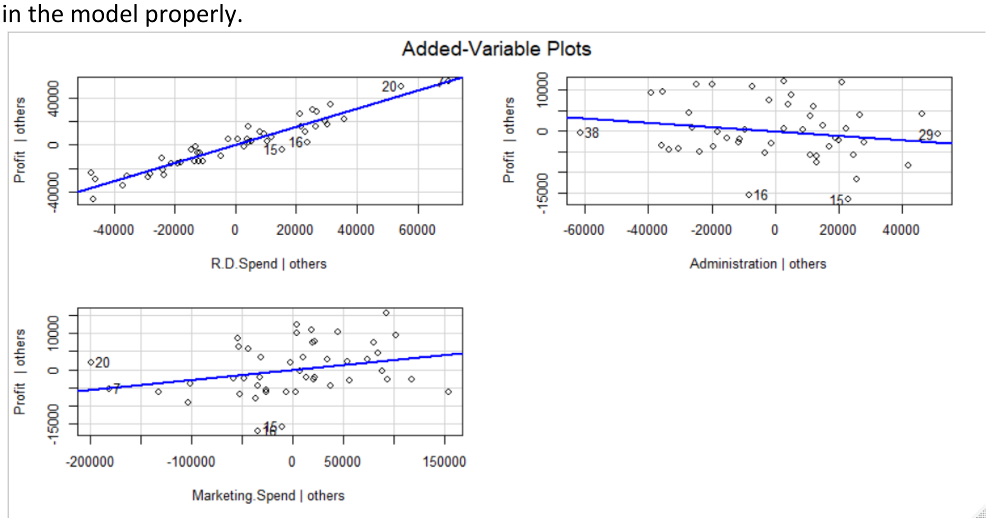
MODEL 2:

model.S.2 <- lm(Profit~R.D.Spend+Administration+Marketing.Spend,data = df_Startups)

In our model 2 we remove the influencing observations. We get R^2 as 9626 with RMSE 6774 and

Correlation between actual and predicted = 0.9748282

We can see that still “Administrative” is insignificant in our model with level of significance (α) probability of error 0.21 i.e. 79% confidence. But incase of “Marketing Spend “we can say that its significance level is 0.06 i.e. approx. to 0.05 so we can relay on this variable to explain the model properly.



MODEL 3:

model.S.3 <- lm(Profit~R.D.Spend+Marketing.Spend,data = df_Startups)

In model 3 we get $R^2 = 0.9612$ and RMSE = 6899.99 and correlation between the actual and predicted is 0.9748121. In our model 3 we can see that After removing the variable "Administrative" we get probability of error (α) for considering the variable "MarketingSpend" is 0.01 i.e. less than 0.05, now we can say that it's a significant variable in our model.

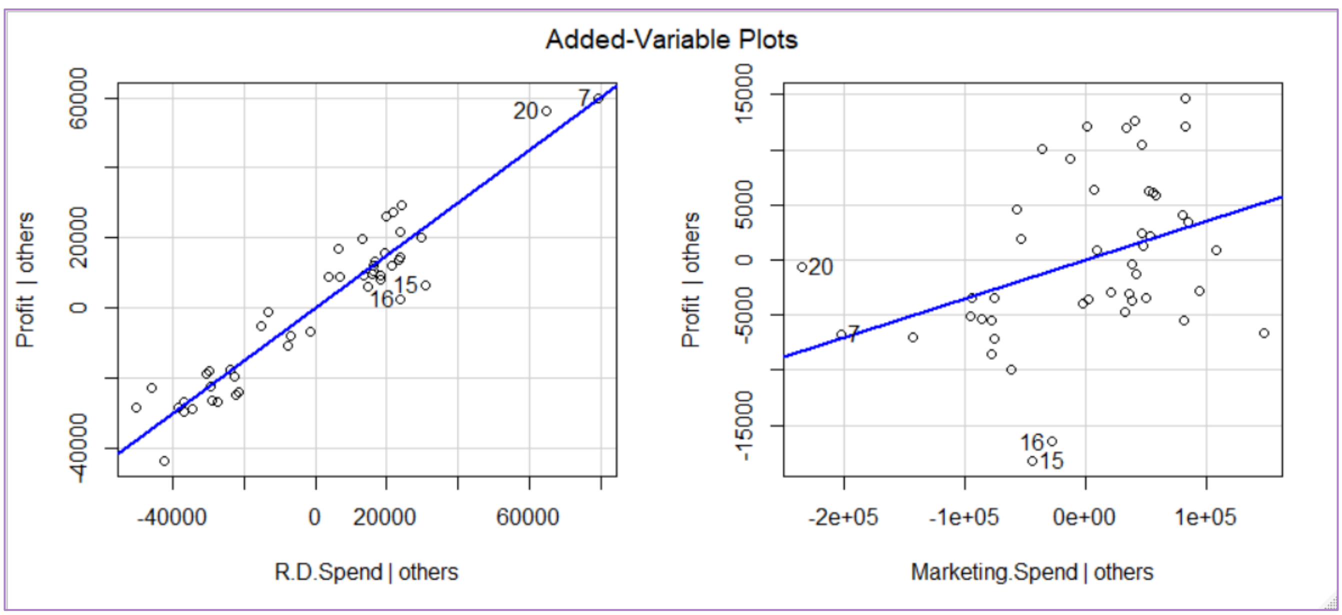
TABULATION:

	Variable	R^2	RMSE	Cor (Y, Predicted)
Model 1	All variable except State	0.9507	8855.344	0.975062
Model 2	Removed the Obs. # 50,49,47,46	0.9626	6774.245	0.9748282
Model 3	Removed variable "Administration: from model 2	0.9612	6899.99	0.9748121

CONCLUSION:

So Here My best fit will be either Model 2, as having the least RMSE value as well as higher R^2 value as well as correlation between the Actual and predicted values.

But in certain case if we go for considering only the variable which are significant enough to explain our model properly. Then I might go for the Model 3, as in model 2 we are considering the insignificant variable "Administration", so may be this is the possible reason for slight increase in our R^2 values, which may not matter as already in model 3, its explaining our 97% variation in our target variable. So, I may rely on my third model also.



QUESTION 2 :

PREDICT PRICE OF THE COMPUTER

Answer:

Available variable: "X", "price", "speed", "hd", "ram", "screen", "cd", "multi", "premium", "ads", "trend"

Target Variable : price

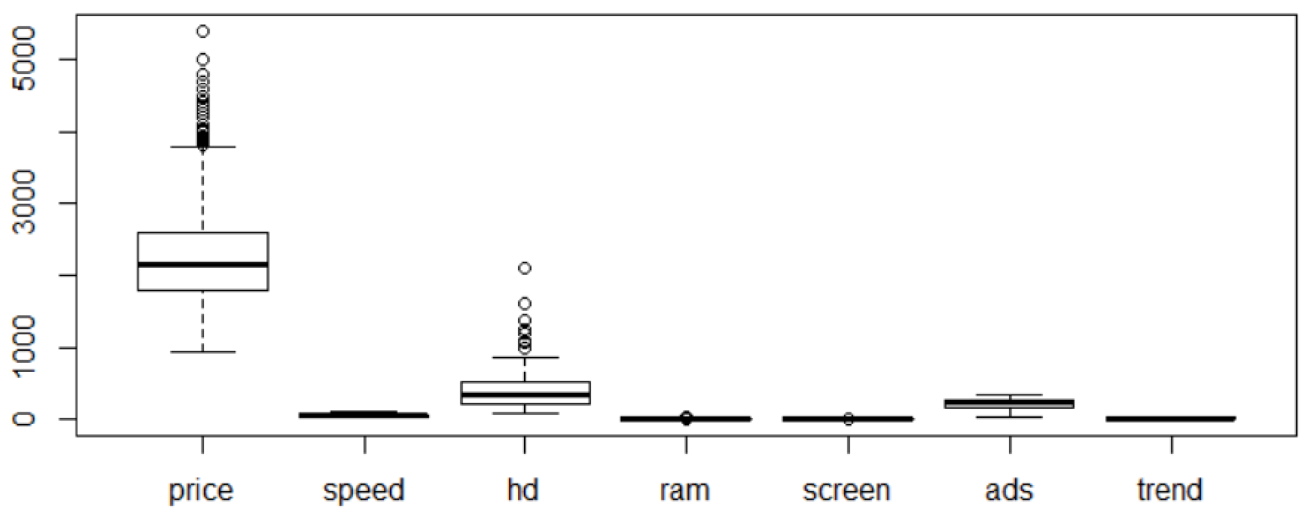
Lets have a look on the summary Statistics of the variables

X	price	speed	hd	ram	screen	ads	trend
Min. : 1	Min. : 949	Min. : 25.00	Min. : 80.0	Min. : 2.000	Min. :14.00	Min. : 39.0	Min. : 1.00
1st Qu.:1566	1st Qu.:1794	1st Qu.: 33.00	1st Qu.: 214.0	1st Qu.: 4.000	1st Qu.:14.00	1st Qu.:162.5	1st Qu.:10.00
Median :3130	Median :2144	Median : 50.00	Median : 340.0	Median : 8.000	Median :14.00	Median :246.0	Median :16.00
Mean :3130	Mean :2220	Mean : 52.01	Mean : 416.6	Mean : 8.287	Mean :14.61	Mean :221.3	Mean :15.93
3rd Qu.:4694	3rd Qu.:2595	3rd Qu.: 66.00	3rd Qu.: 528.0	3rd Qu.: 8.000	3rd Qu.:15.00	3rd Qu.:275.0	3rd Qu.:21.50
Max. :6259	Max. :5399	Max. :100.00	Max. :2100.0	Max. :32.000	Max. :17.00	Max. :339.0	Max. :35.00

cd	multi	premium
no :3351	no :5386	no : 612
yes:2908	yes: 873	yes:5647

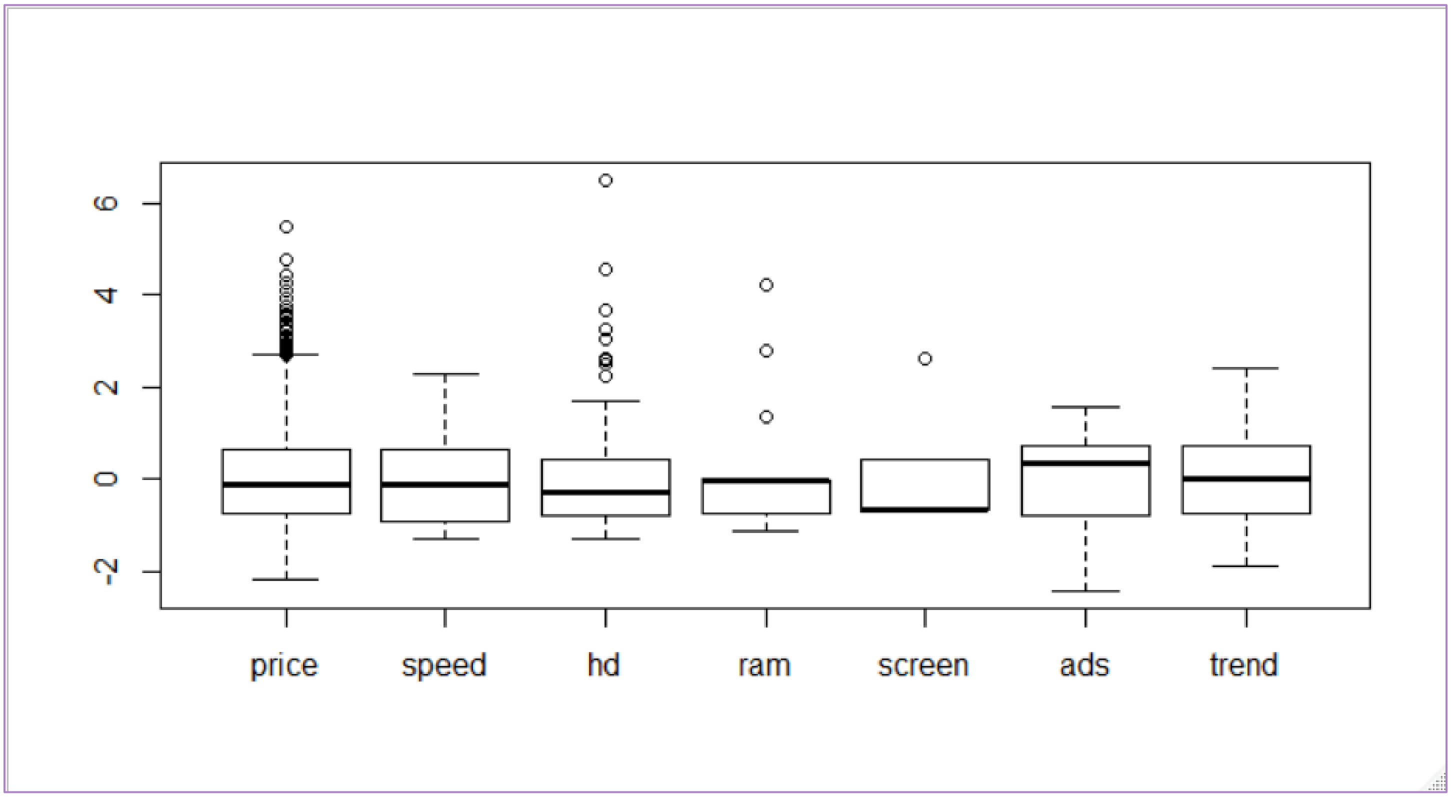
All except cd, multi, premium are discrete type. Where as cd, multi, premium is of factor type.

BOXPLOT:



We can see there is lots of outlier in the variable price and hd.

Lets make the plot unitless and scale free



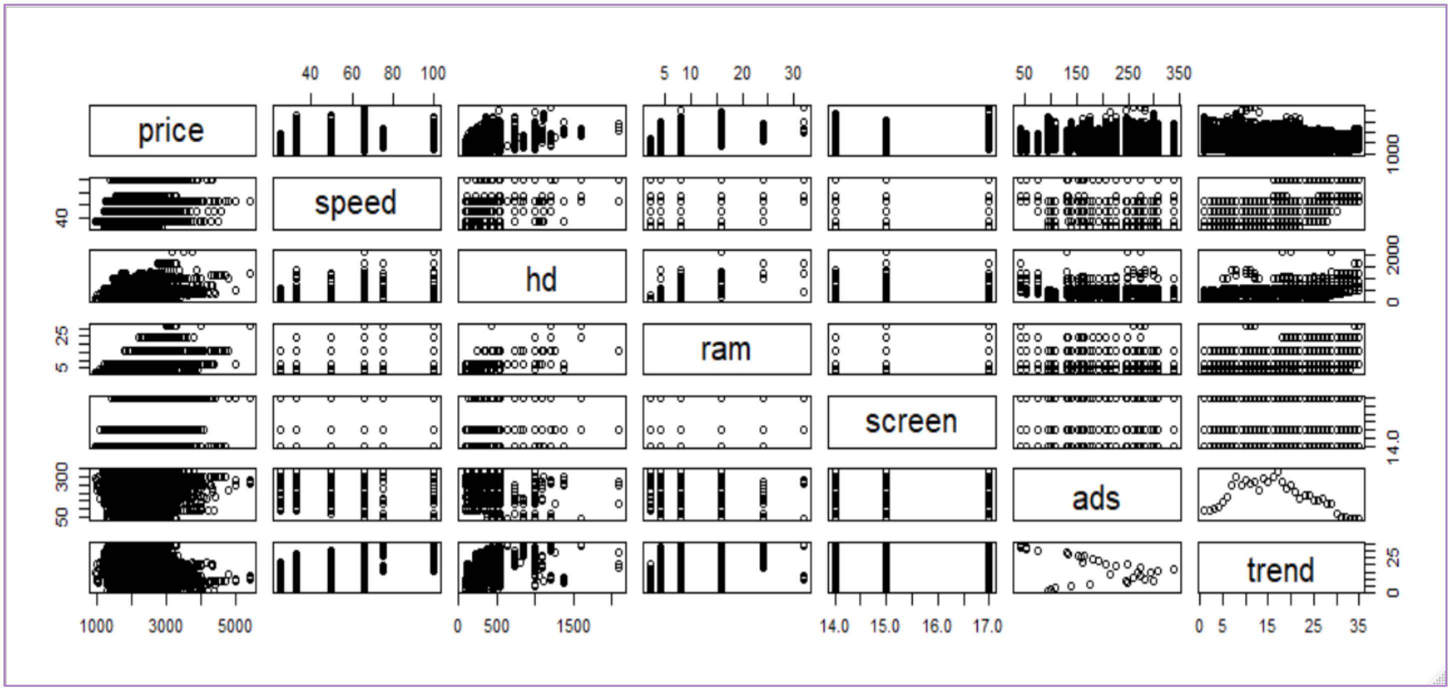
Now we have a clear view in each and every variable here

CORRELATION:

	price	speed	hd	ram	screen	ads	trend
price	1	0.300976	0.430258	0.622748	0.296041	0.05454	-0.19999
speed	0.300976	1	0.372304	0.23476	0.189074	-0.21523	0.405438
hd	0.430258	0.372304	1	0.777726	0.232802	-0.32322	0.57779
ram	0.622748	0.23476	0.777726	1	0.208954	-0.18167	0.276844
screen	0.296041	0.189074	0.232802	0.208954	1	-0.09392	0.188614
ads	0.05454	-0.21523	-0.32322	-0.18167	-0.09392	1	-0.31855
trend	-0.19999	0.405438	0.57779	0.276844	0.188614	-0.31855	1

None of the variables are strongly correlated as seen from the data bars.

PAIRS PLOT:

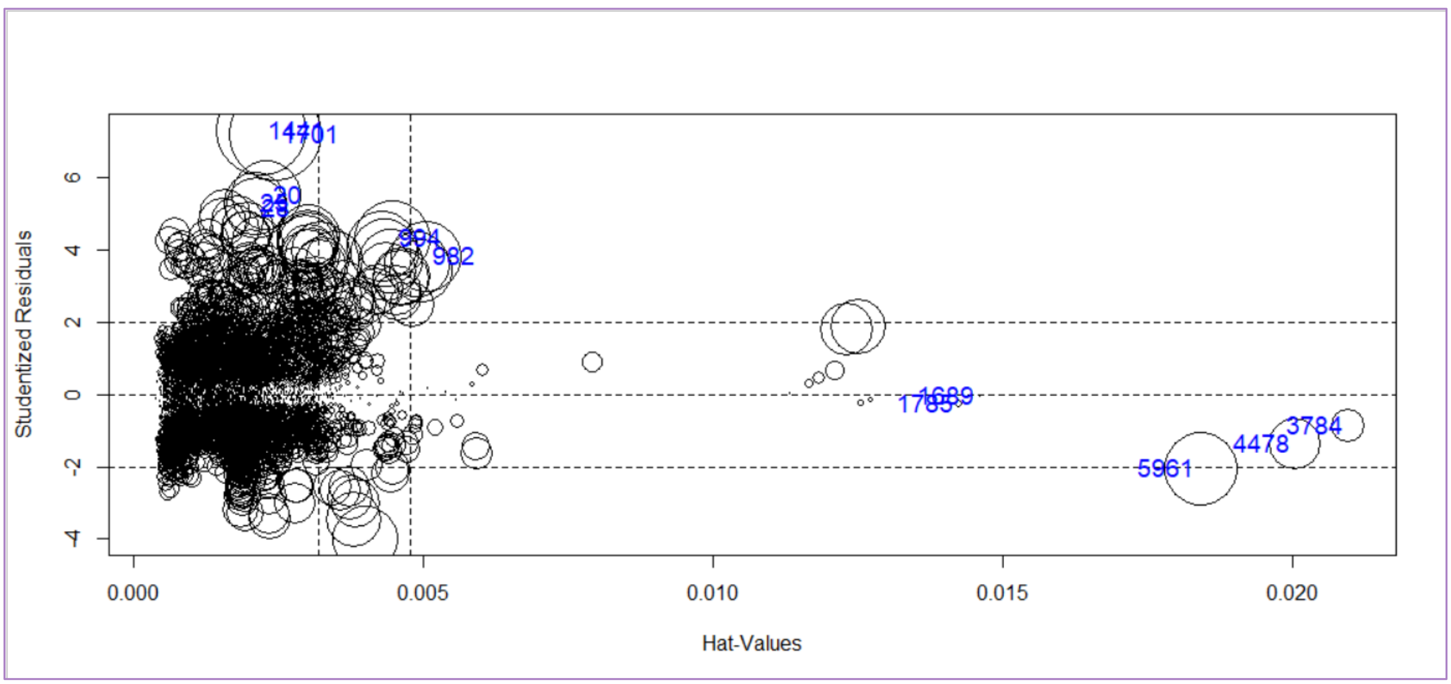


FITTING REGRESSION MODEL:

MODEL 1:

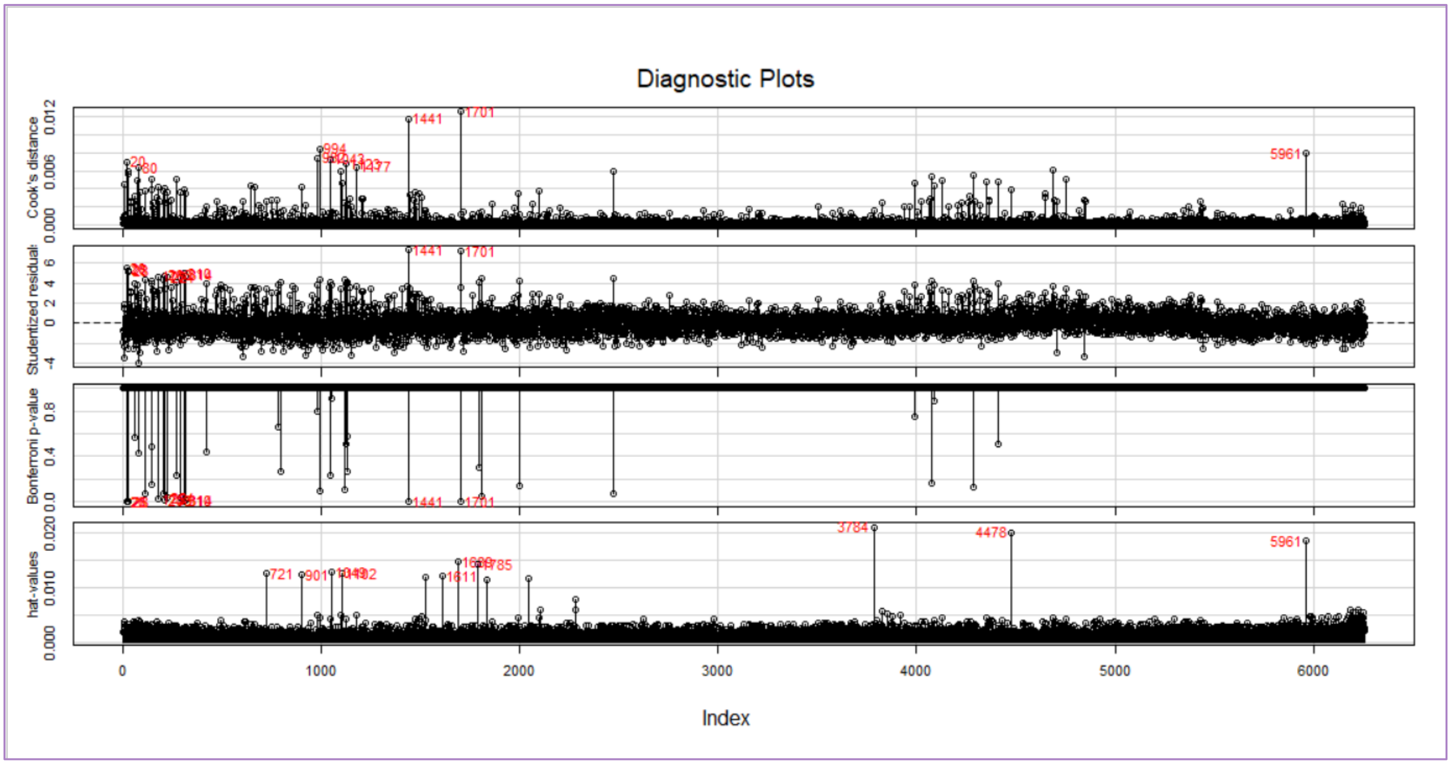
model_Comp_1 <-lm(price~speed+hd+ram+screen+cd+multi+premium+ads+trend,data = df_comp)

In model 1 I simply consider all the variables and fit the model and come of with no insignificant variables, lots of influencing indexes, with coefficient of determination 0.7756 , RMSR value 275.1298 and finally correlation between the predicted and actual value to be 0.8806631, which was not that much bad for me.



From above plot we can see the dispersion of the points.

Looking at this plot we can say that there are pretty large numbers of influencing observations in our model.



Than I make the data scale free and unit less for performing my next model.

MODEL 2:

df_comp2 <- data.frame(scale(log(Comp[,-c(1,2,7,8,9)])),"price" = df_comp$price,"cd" = df_comp$cd,"premium" = df_comp$premium,"multi" = df_comp$multi)
model_Comp_2 <- lm(price~.,data=df_comp2)

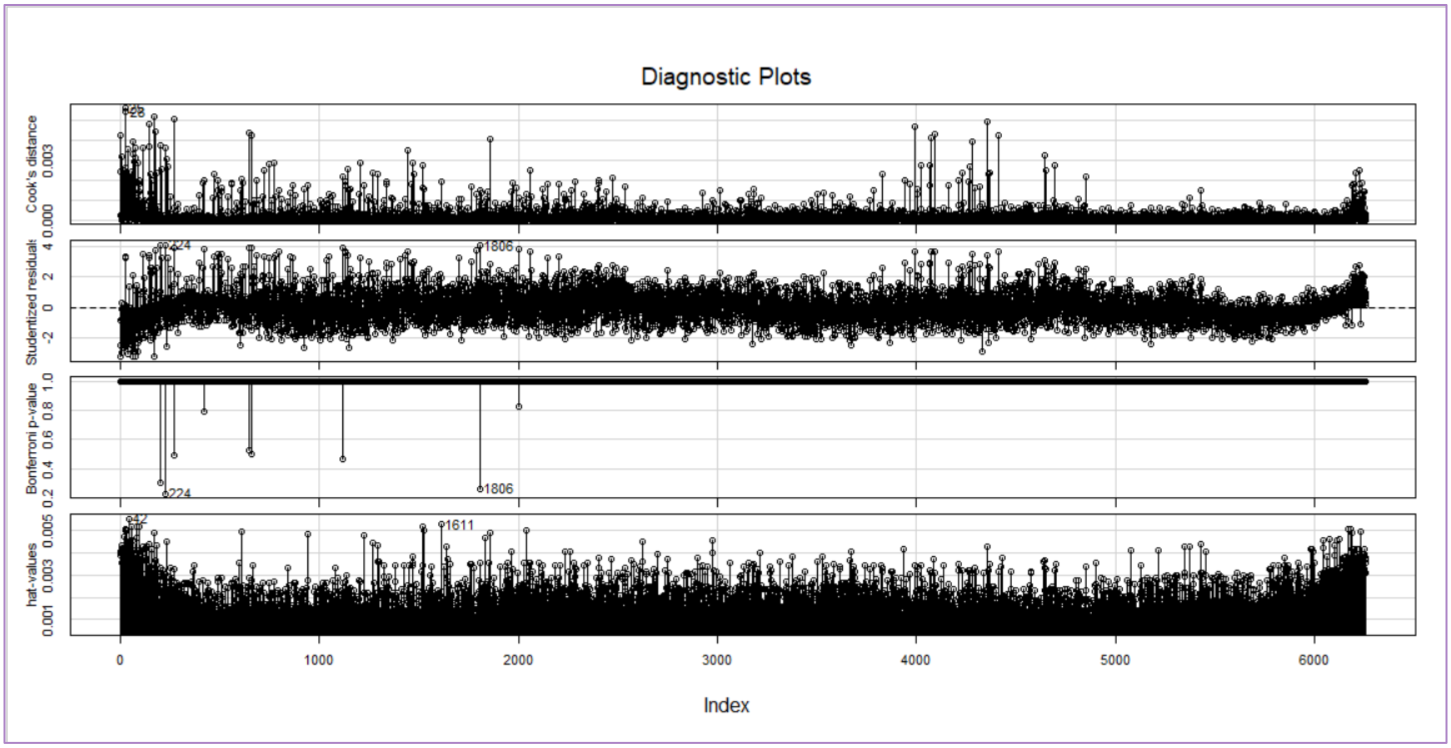
In model 2 I make the whole data (which columns are numerical) to unitless and scale free i.e. standardized the data, along with log transformation. Here I come up with coefficient of determination 0.7426, which was less than the previous model. So, I think better removing the influencing index from my data for my next model.

MODEL 3:

influ_comp <- as.integer(rownames(influencePlot(model_Comp_2,id = list(n=20,col="blue"))))
df_comp3 <- df_comp2[-c(influ_comp),]#head(df_comp2)
model_Comp_3 <- lm(price~.,data=df_comp3)

In model 3 I removed 20 influencing observations. And fit my model again. Now I come up with slight improvement in my coefficient of determination as 0.7508, RMSE as 281.3819 and finally correlation between the actual and predicted value as 0.8664

As we now our data set still contains lots of influencing index over there with count 291. So I thought removing 3% of my data in my model 4.



Have a look on this influencing index of model 3 , where I come up with more than 200 influencing index for Cook's distance

---

MODEL 4:

influencing_obs <- which(rowSums(influence.measures(model_Comp_1)$is.inf) > 0);influencing_obs # These are the influencing observations

HERE I GET 294 INFLUENCING INDEX

influence_obs <- as.integer(rownames(influencePlot(model_Comp_1,id=list(n=90,col="red"))))

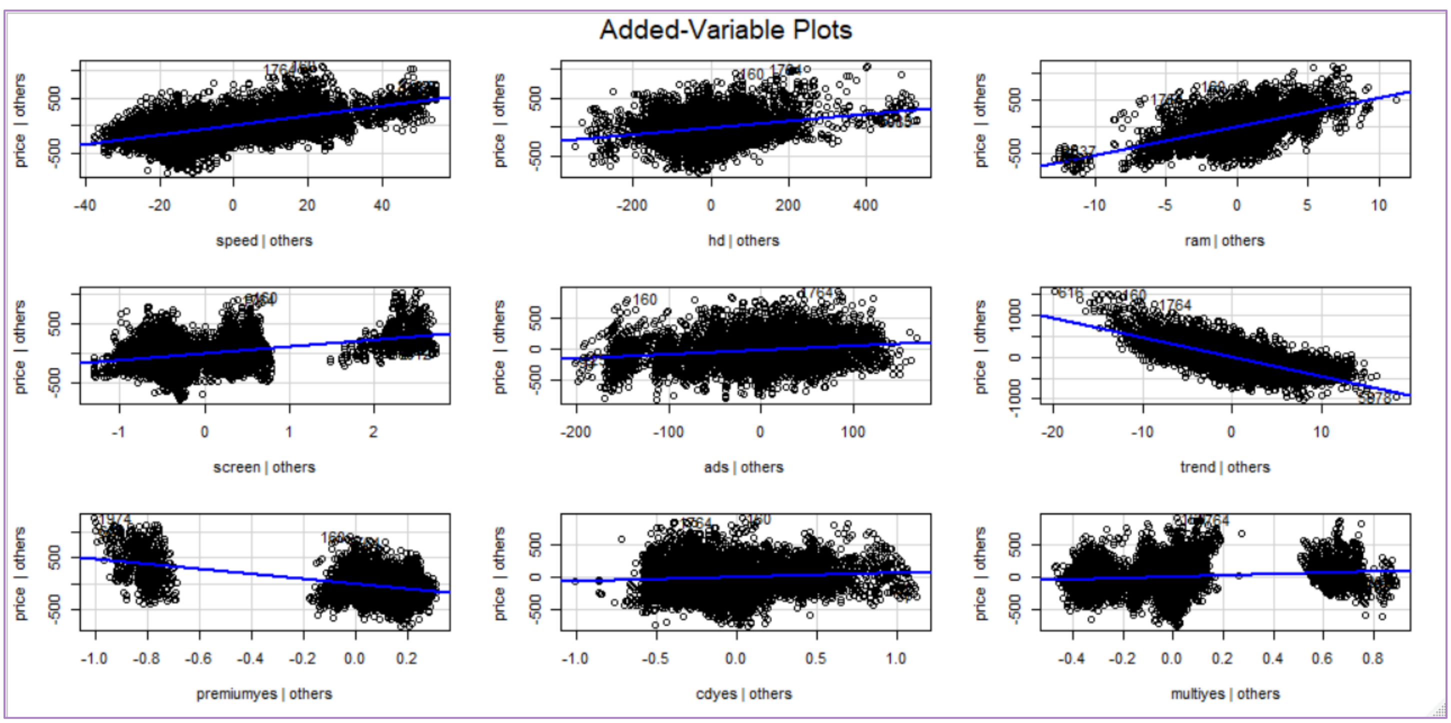
HERE I GET 186 INFLUENCING INDEX

df_Comp_scale <- data.frame(df_comp[,-c(6,7,8)],"premium"=df_comp$premium,"cd"=df_comp$cd,"multi"=df_comp$multi)

df_Comp_scale <- df_Comp_scale[-c(influence_obs),]

model_Comp_4 <- lm(price~.,data=df_Comp_scale)

ADDED VARIABLE PLOT



In my 4th model I prefer not to do any transformation with our data as it seems useless (observing the above two models) so in this model I prefer to focus more on my influencing observations and come up with 186 influencing index (defaulters) i.e. approx. 3% of my data set. So I build the model without those influencing factors but I predict my model with all the data set given to me. Now I come up with coefficient of determination as 0.804, RMSE is 238 and finally correlation between the actual and predicted values is 0.879

TABULATION:

Model No	Modeled with	Predicted with	Transformation	R^2	RMSE	cor
1	All Observations	All Observations	NA	0.7756	275.1298	0.880663
2	All Observations	All Observations	NA	0.7426	-	-
3	99.4% data	99.4% data	log, standard scalar	0.7508	281.3819	0.86647
4	97.02% data	All Observations	NA	0.804	238.0004	0.879204

CONCLUSION:

It's obvious that I will go for my 4th model which is explaining 80% of variation in our target variable due to the observations, along with least RMSE value among all fitted models.

QUESTION 3 :

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")]

Answer:

Available variable: "Price","Age_08_04","KM","HP","cc","Doors","Gears","Quarterly_Tax","Weight"

Target Variable : price

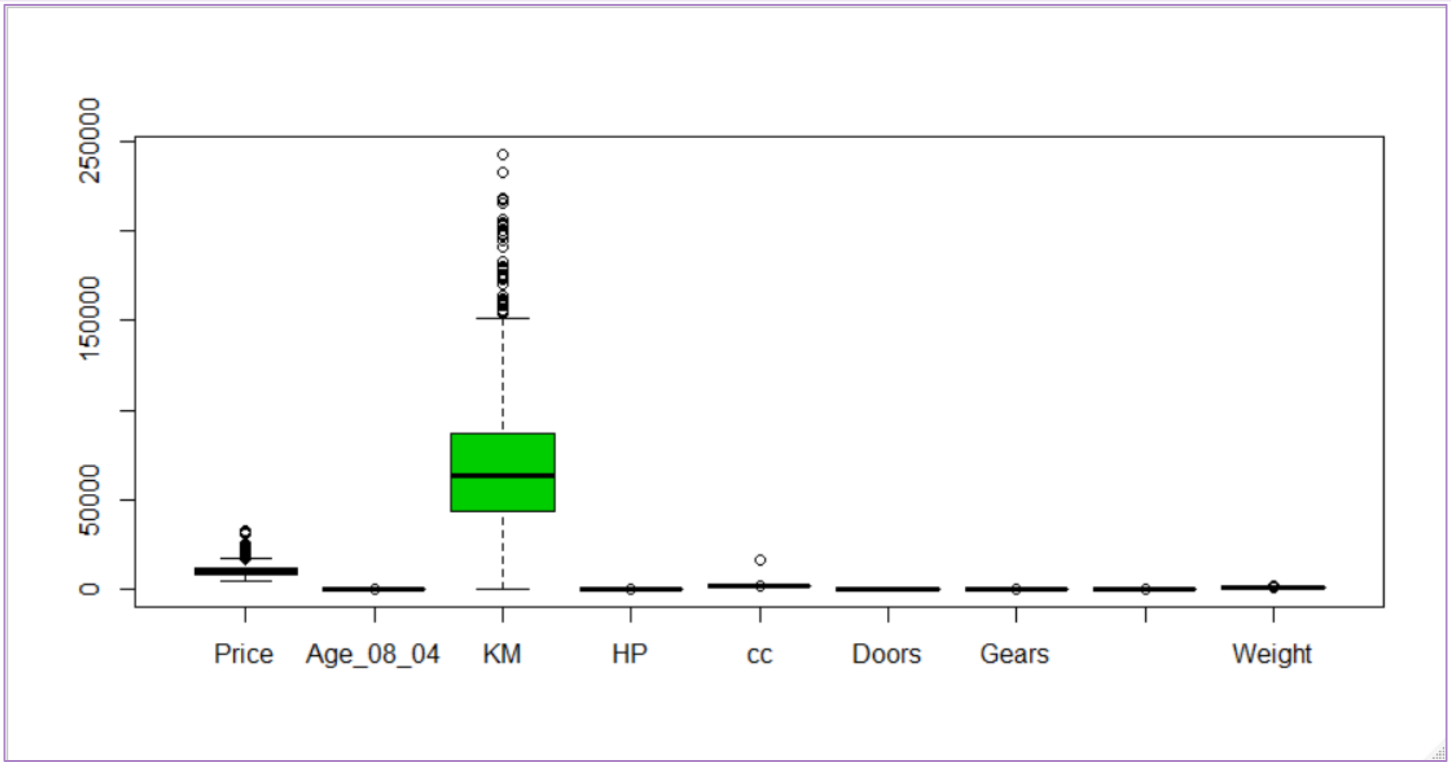
Lets have a look on the summary Statistics of the variables

SUMMARY:

Price	Age_08_04	KM	HP	cc	Doors	Gears	Quarterly_Tax	Weight
Min. : 4350	Min. : 1.00	Min. : 1	Min. : 69.0	Min. : 1300	Min. :2.000	Min. :3.000	Min. : 19.00	Min. :1000
1st Qu.: 8450	1st Qu.:44.00	1st Qu.: 43000	1st Qu.: 90.0	1st Qu.: 1400	1st Qu.:3.000	1st Qu.:5.000	1st Qu.: 69.00	1st Qu.:1040
Median : 9900	Median :61.00	Median : 63390	Median :110.0	Median : 1600	Median :4.000	Median :5.000	Median : 85.00	Median :1070
Mean :10731	Mean :55.95	Mean : 68533	Mean :101.5	Mean : 1577	Mean :4.033	Mean :5.026	Mean : 87.12	Mean :1072
3rd Qu.:11950	3rd Qu.:70.00	3rd Qu.: 87021	3rd Qu.:110.0	3rd Qu.: 1600	3rd Qu.:5.000	3rd Qu.:5.000	3rd Qu.: 85.00	3rd Qu.:1085
Max. :32500	Max. :80.00	Max. :243000	Max. :192.0	Max. :16000	Max. :5.000	Max. :6.000	Max. :283.00	Max. :1615

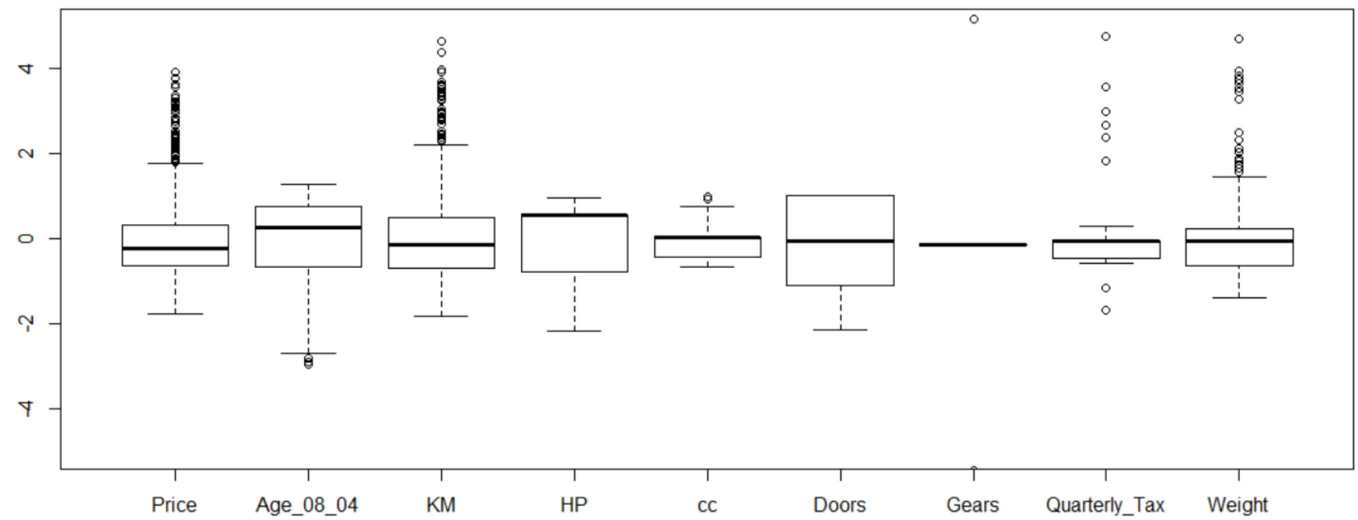
Price, “Age_08_04”, “KM” have significant difference in their mean and median. So, we can say they may contain outliers.

BOXPLOT:



This Plot cant convey as much as we desire.

So lets go for a scale free plot



Even if I have considered only y-limit from -4 to 4 , but originally there is lot more points beyond this plot limit. So, from this plot we can see that Weight , Quarterly_Tax also contains outliers in a larger amount.

CORRELATION:

	Price	Age_08_04	KM	HP	cc	Doors	Gears	Quarterly_Tax	Weight
Price	1.000000	-0.876590	-0.569960	0.314990	0.126389	0.185326	0.063104	0.219197	0.581198
Age_08_04	-0.876590	1.000000	0.505672	-0.156622	-0.098084	-0.148359	-0.005364	-0.198431	-0.470253
KM	-0.569960	0.505672	1.000000	-0.333538	0.102683	-0.036197	0.015023	0.278165	-0.028598
HP	0.314990	-0.156622	-0.333538	1.000000	0.035856	0.092424	0.209477	-0.298432	0.089614
cc	0.126389	-0.098084	0.102683	0.035856	1.000000	0.079903	0.014629	0.306996	0.335637
Doors	0.185326	-0.148359	-0.036197	0.092424	0.079903	1.000000	-0.160141	0.109363	0.302618
Gears	0.063104	-0.005364	0.015023	0.209477	0.014629	-0.160141	1.000000	-0.005452	0.020613
Quarterly_Tax	0.219197	-0.198431	0.278165	-0.298432	0.306996	0.109363	-0.005452	1.000000	0.626134
Weight	0.581198	-0.470253	-0.028598	0.089614	0.335637	0.302618	0.020613	0.626134	1.000000

From the above Tabulation we can see, only Age_08_04 and Price is highly (negatively) correlated with correlation coefficient -0.876

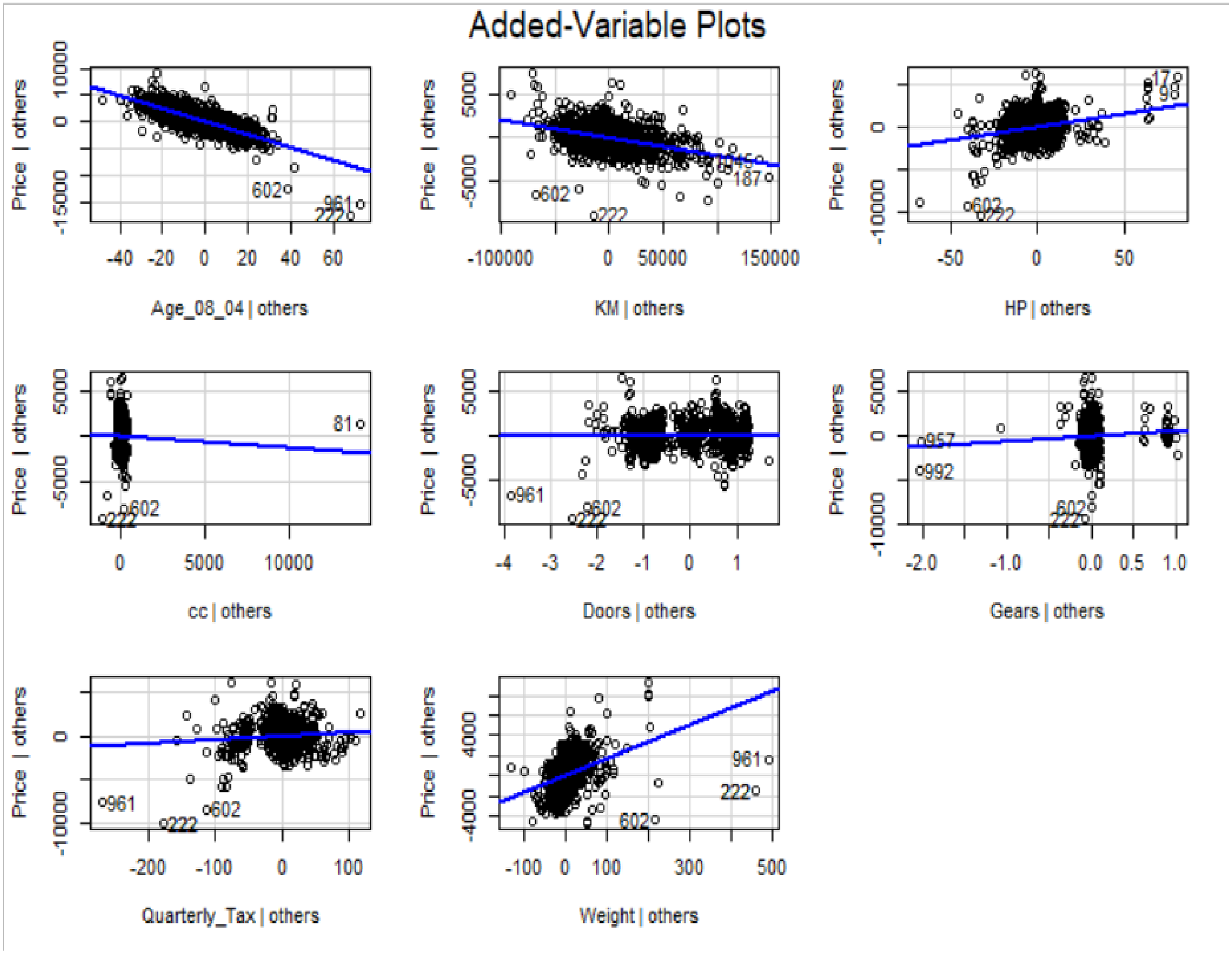
Its useless to go for pairs plot as we know that maximum variables are discrete with less unique values so those data may not convey much for finding out relationships, i.e. seen by our correlation table with a mare look.

FITTING REGRESSION MODEL:

MODEL 1:

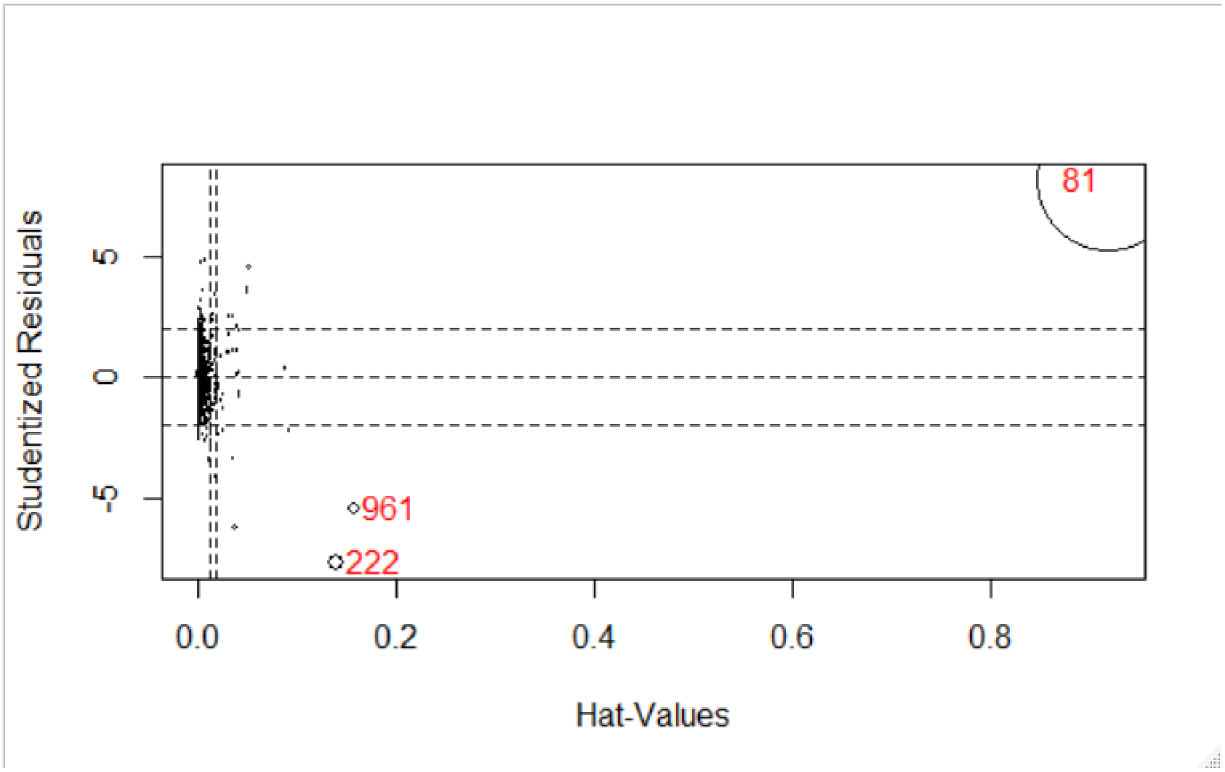
model_T_1 <- lm(Price~.,data = Corolla)

This is my simple model, considering all the observation as well as all the variables. In this model I get coefficient of determination as 0.8698, RMSE as 1338.25 and finally correlation between the actual and predicted as 0.929 .

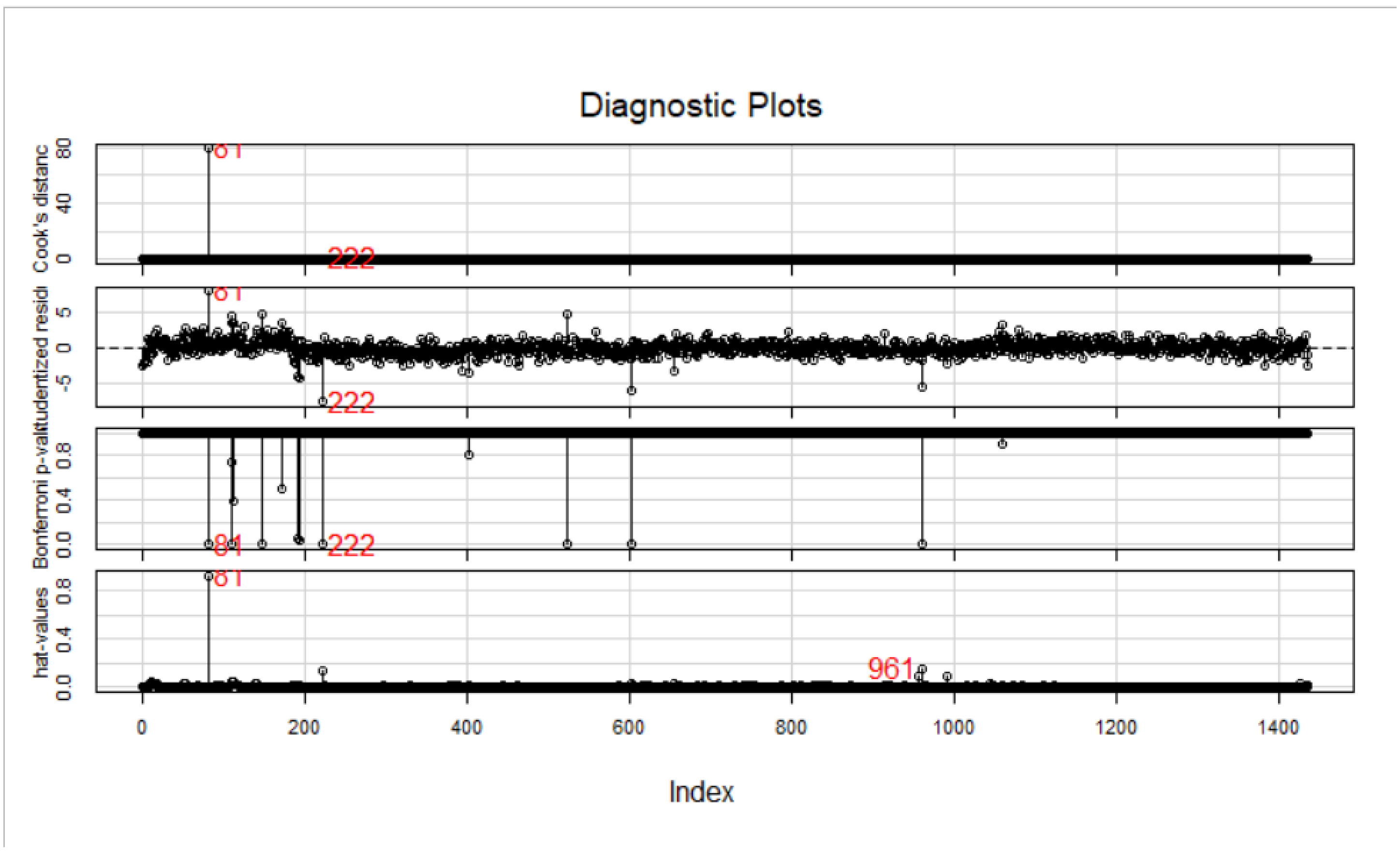


From the above model we find that the variable cc and Doors are insignificant. But looking at the plot we can also say that Gears may also be insignificant for our model.

INFLUENCE PLOT:



It seems like observation 81,961, and 222 are influencing more in our model.



MODEL 2:

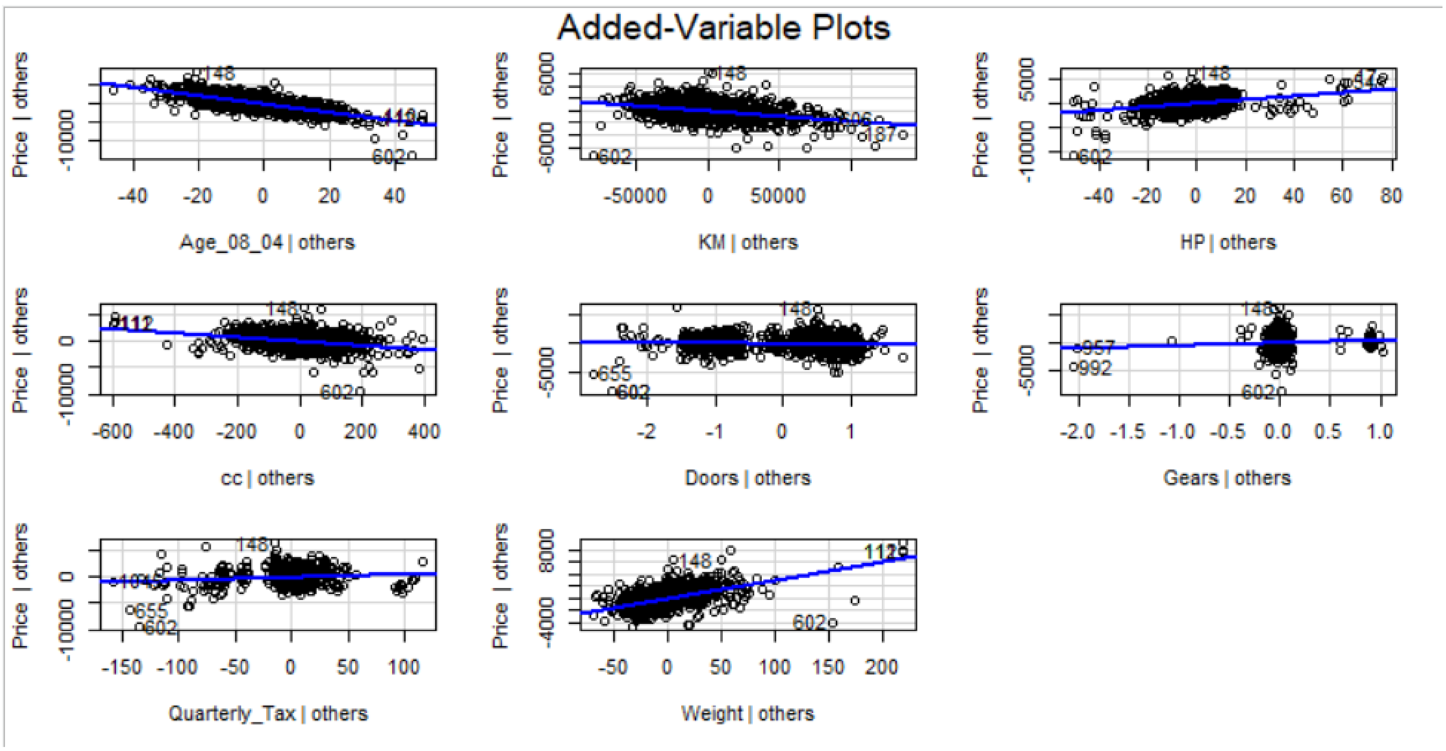
In our second model we are removing the three influencing observations.

df_Corola <- Corolla[-c(influence_index),]

model_T_2 <- lm(Price~.,data = df_Corola)

In our second model we come up with all significant variables. We can see our R^2 value little bit increased i.e. 0.8852 with RMSE decreased 1227.474 and finally correlation between the actual and predicted value is found out to be 0.9408425

ADDED VARIABLE PLOT:



From this plot we can see that now cc is showing somewhat significant behavior, but even if Doors and Gear are not showing that much significance, but we may consider them as per our |t| statistics value in our model.

TABULATION

Model No	R^2	RMSE	Cor
1	0.8698	1338.25	0.929
2	0.8852	1227.474	0.9408425

CONCLUSION:

Its Obvious that I will go for my second model with higher R^2 value i.e. this model can able to explain 88% of variation in the price with the help of all the given independent variable in the data set. This is enough for predicting price. This model also with lower RMSE and higher Correlation between the actual and predicted variable as compare to the previous model.