**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.**

**Inferences from the Data Set:**

Data Set talks about the profit of the 50-startups dependent variable is continuous with respect to around 5 Independent variables & 50 observations.

5 Independent variables are

R.D.Spend

Administration

Marketing.Spend

State

R.D.Spend

Along in 5 Independent variables one variable is categorical data i.e; states. We have to create the dummy variables for that variable

Measures of central tendency explains about mean, median and mode of the data set.

Measures of dispersion explains about the standard deviation variation and range of the data set.

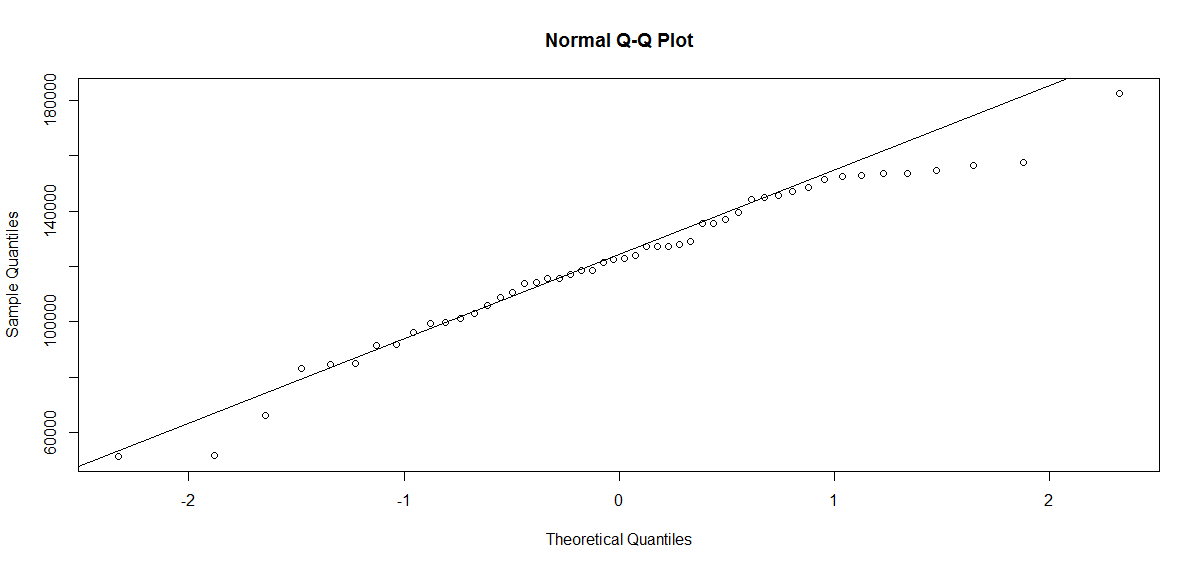
Third moment business decision and Fourth moment business decision explains about the skewness and kurtosis of the data set.

R.D.Spend f the data shows as positive skewness i.e; data skewed to the right and kurtosis is the degree of peakness of distribution.

**Probability distributions of variables:**

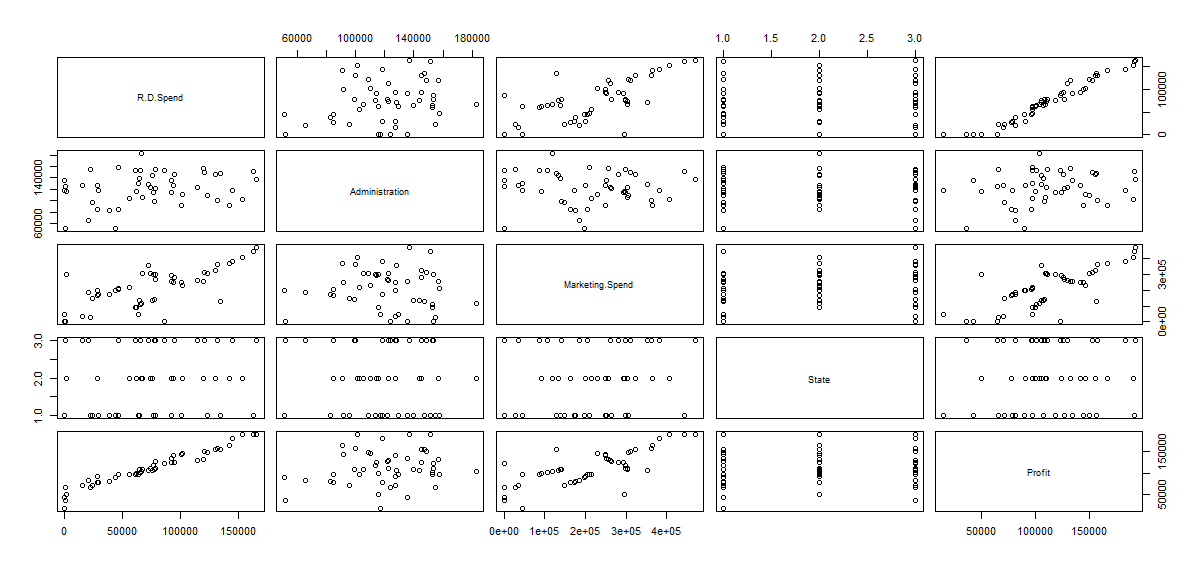
Normal Quantile-Quantile Plot: talks about the data is normally distributed or not if not have to apply transformation to get the data normally distributed.

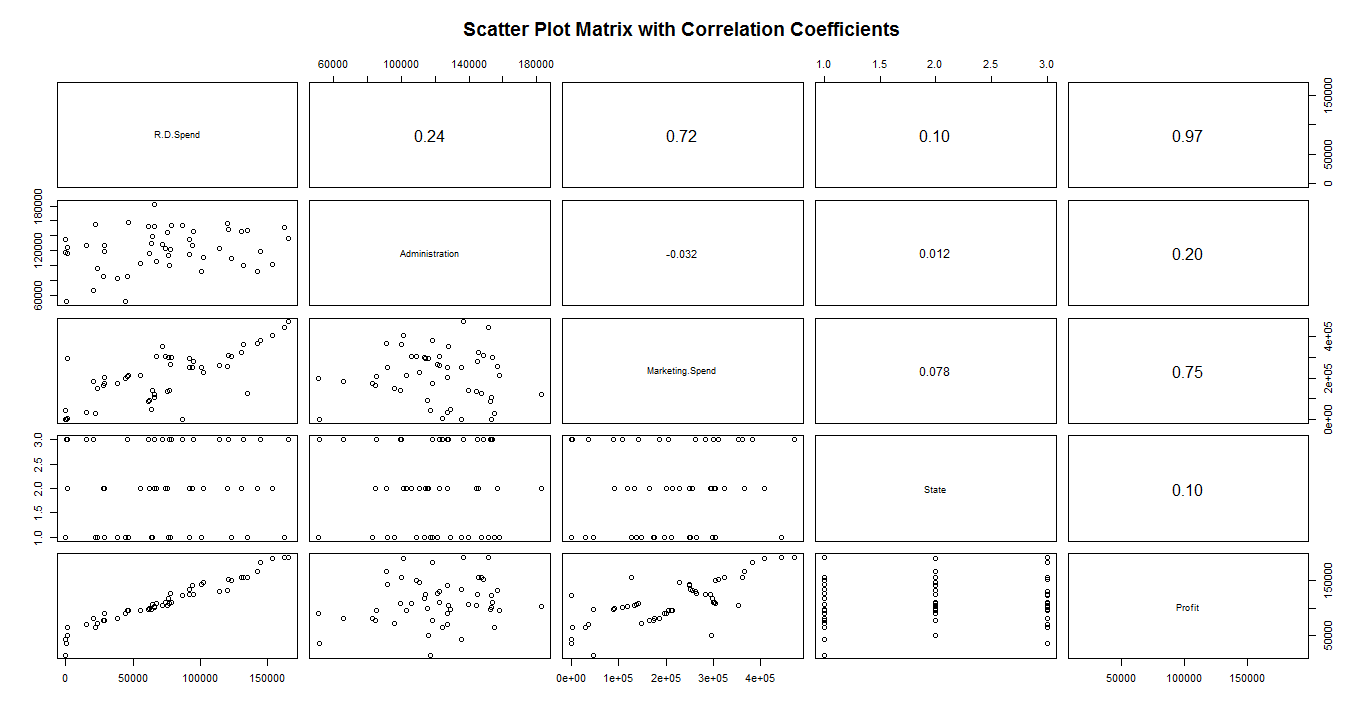
The data is normally distributed for **R.D.Spend**



**scatter plot:**

pairs(startup) # Scatter plot for all pairs of variables





The about scatter plot explains about the relationship about the input and output variables.

Relation between R.D.Spend and profit there is strong positive correlation we can observe.

And also observe that R.D.Spend and Marketing.Spend have strong positive correlation and Administration, State have low correlation to output variables.

**# The Linear Model**

By applying the linear model on the dataset we can observe from the summary of linear model, all coefficients are insignificant. So **Coefficients are insignificant** then check the model for individual inputs to get the coefficients are statistically significant.

By applying linear model on profit to Administration observed as **Coefficients are insignificant.**

And apply linear model on Marketing.Spend observed as **Coefficients are significant.**

Apply combination of independent variables where the Coefficients are insignificant to dependent variable and observing as the **Coefficients are insignificant.**

**Partial Correlation matrix:**

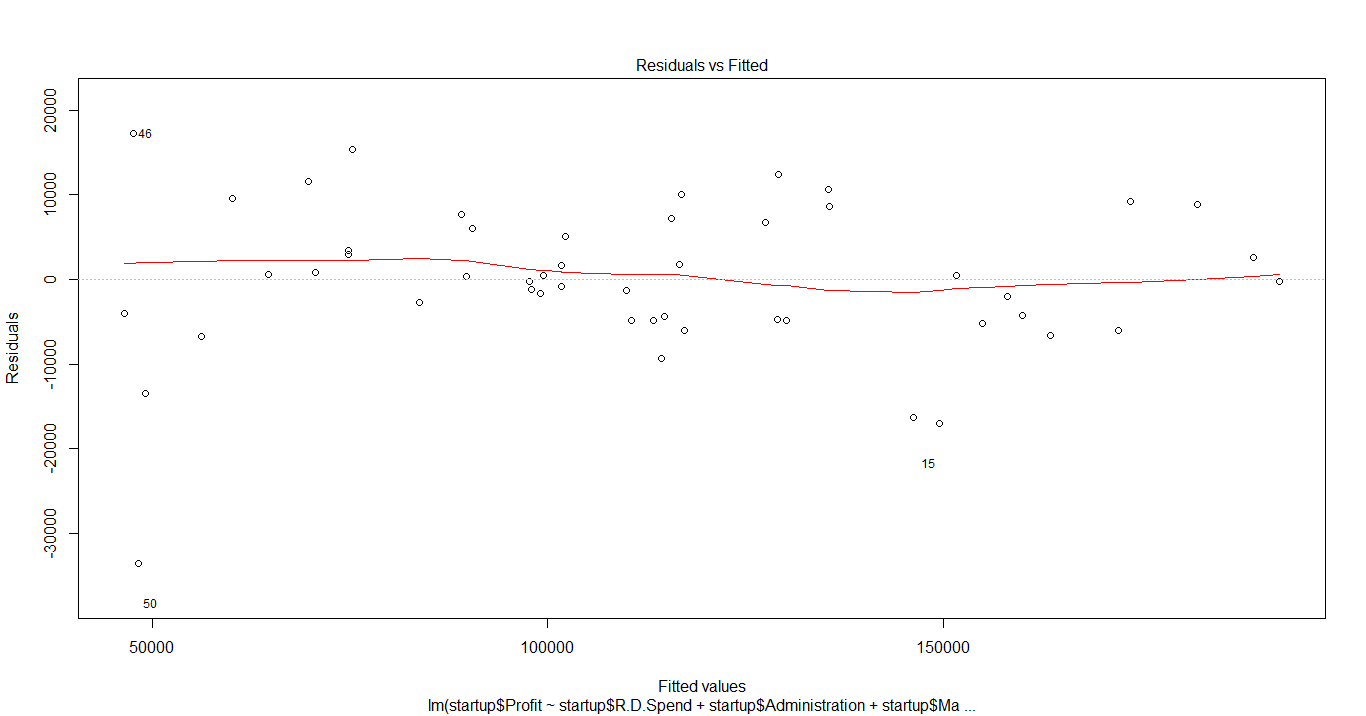
cor2pcor(cor(startup))

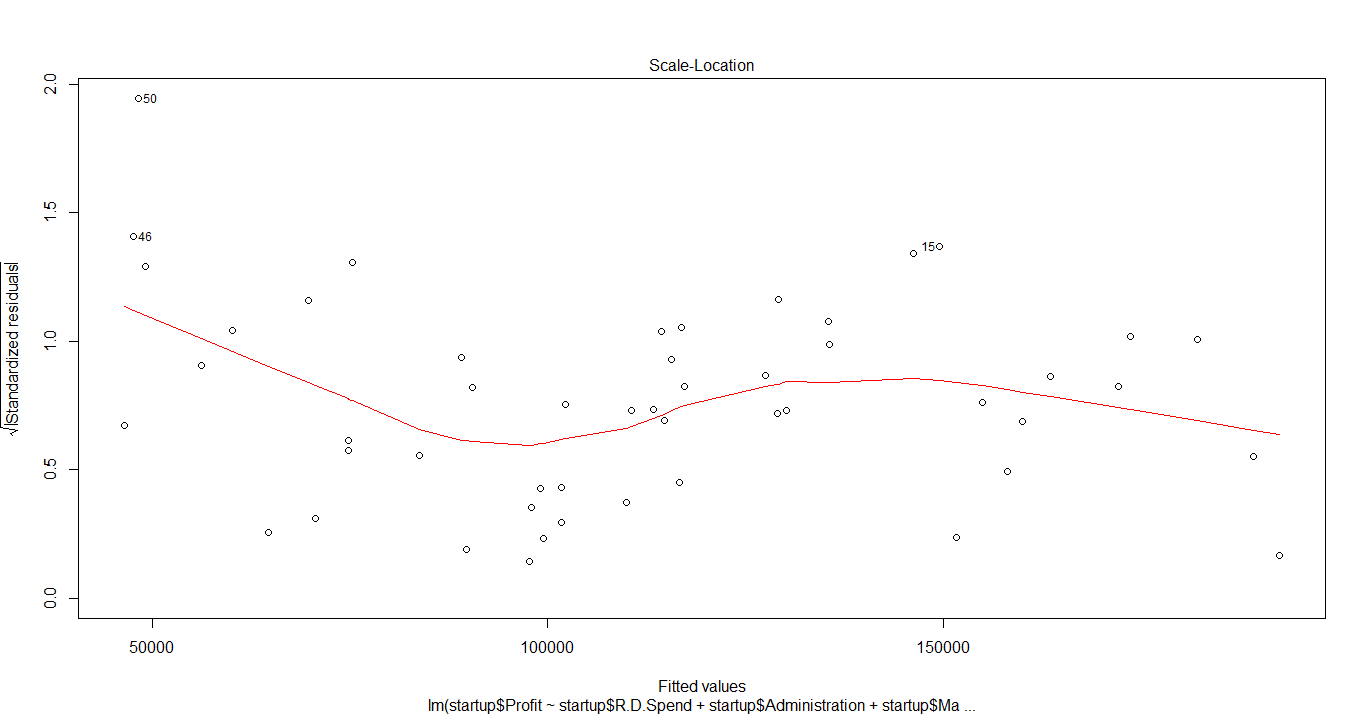
partial correlation gives what is the effect of variable in entire equation which has the least effect on output variable identifying. In this partial correlation observed as **Administration and state** has least effect on output variable.

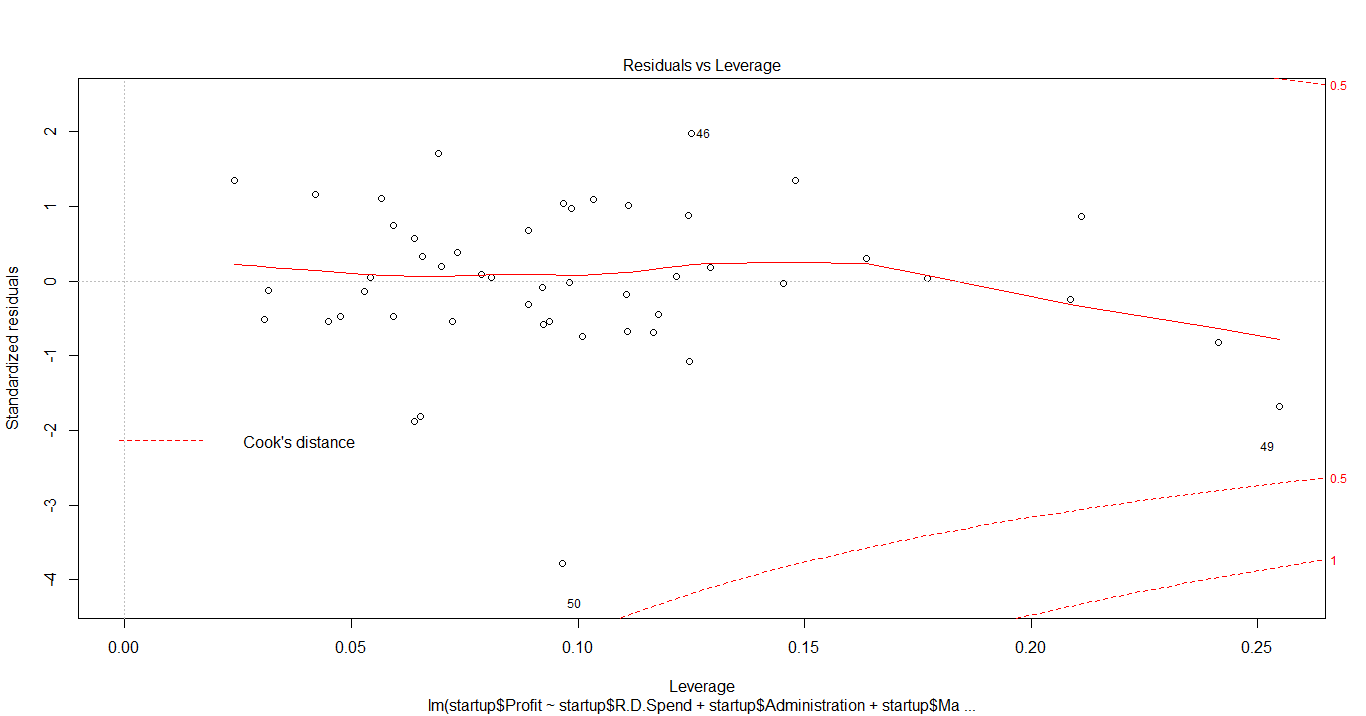
**# Diagnostic Plots**

plot(model.startup)

Plotting the linear model get the Residual Plots, Std. Residuals vs Fitted, Cook's distance.





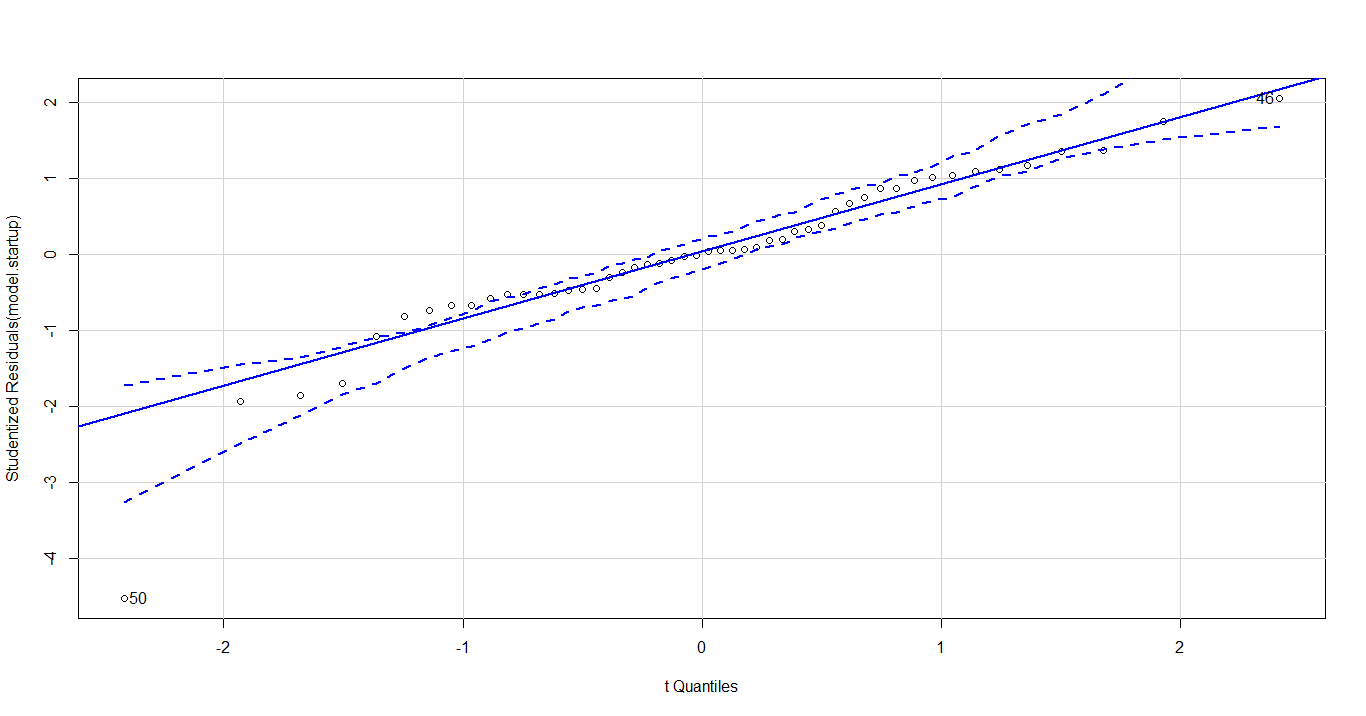


The above plots explains about the Residual/errors of the independent variables and residuals vs leverage plots shows that the value having the above the line 0.5 have the outlier and have to remove outlier.

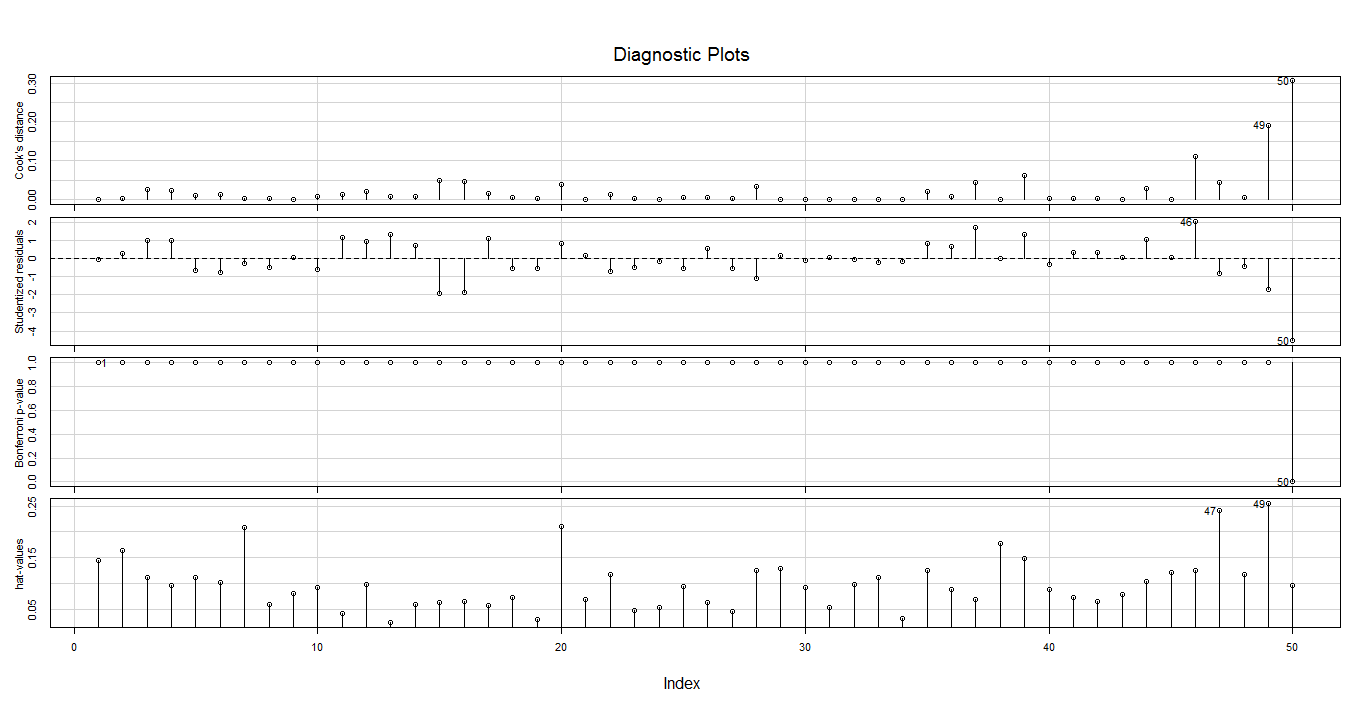
And identifying influential variables means that there is any outliers or not.

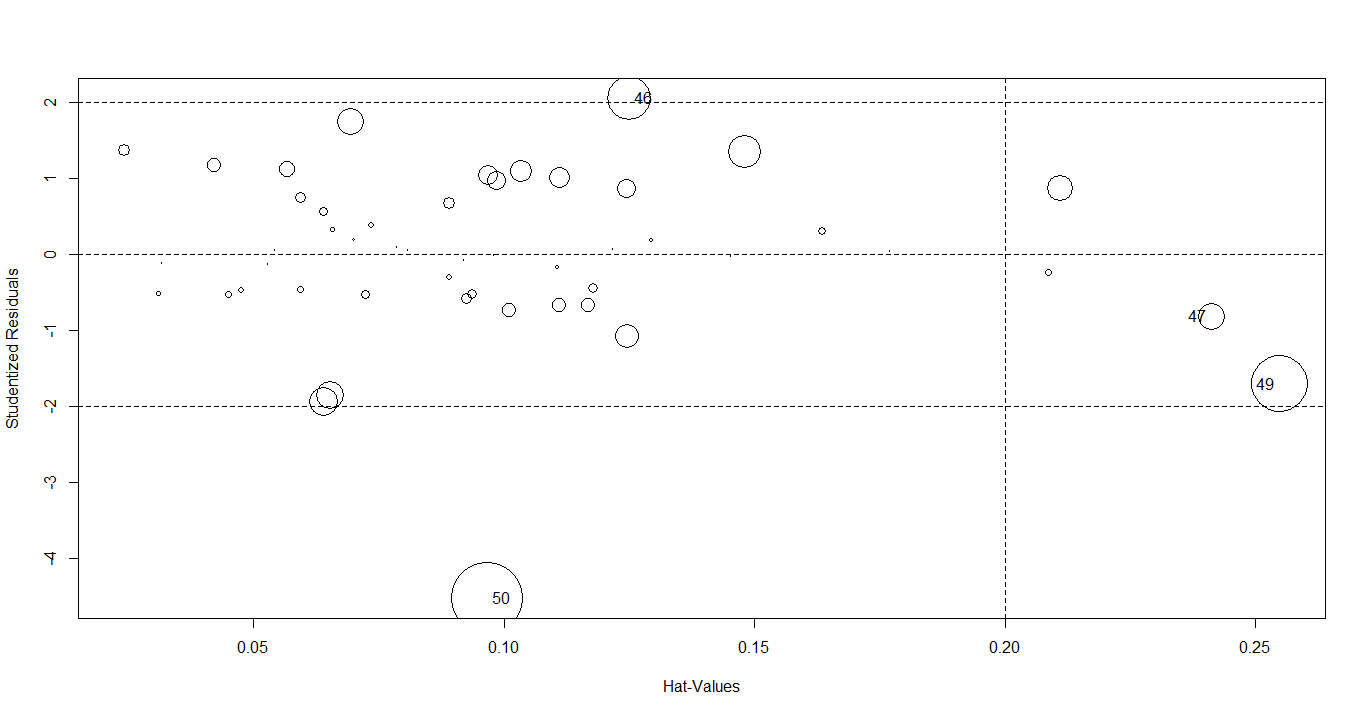
**QQ plots:**

QQ plots explains/shows as there in any outliers in the dataset or not. We can assume by seeing the below plot there are 46th and 50th records in the dataset are outliers. So we can remove the outliers from the dataset to get the coefficients of model are significant.



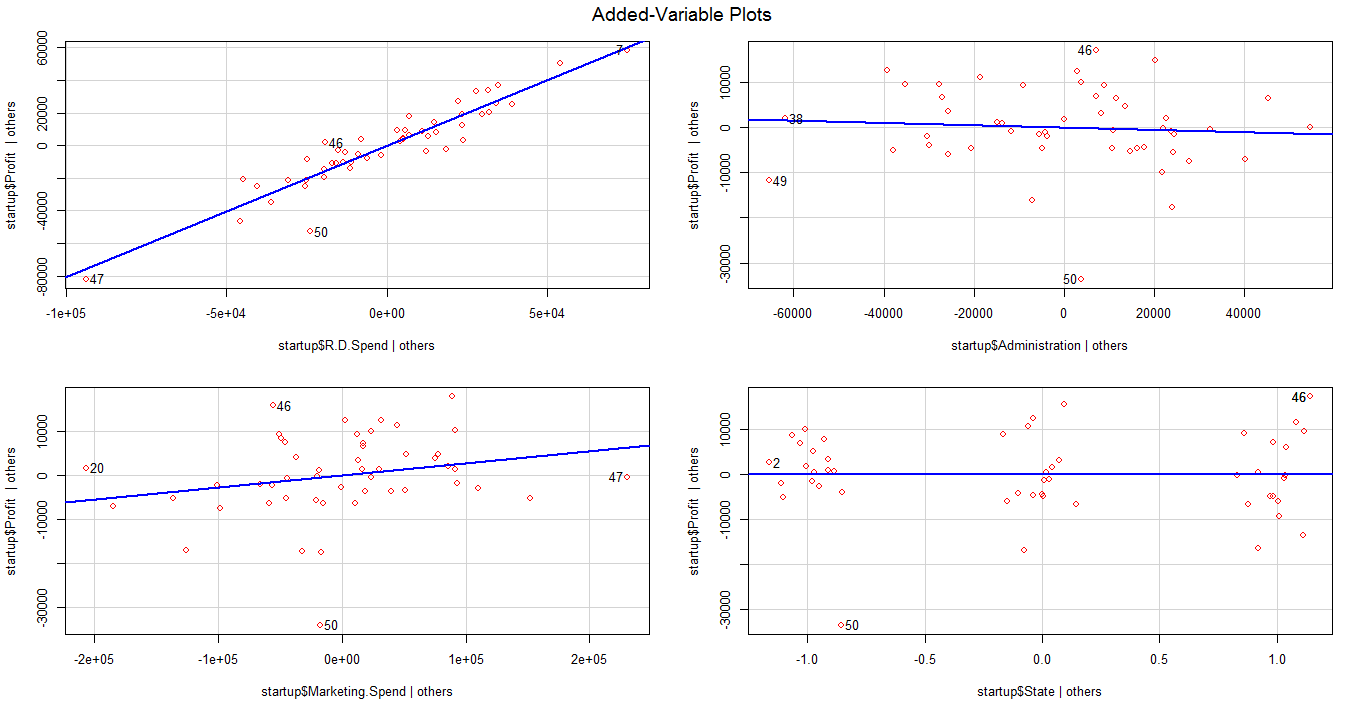
**Deletion Diagnostics for identifying influential variable:**

****

****

By seeing above plot we can clearly assume as 46, 47, 49 and 50th records are outliers so we can remove the outliers.

**Added Variable Plots**

****

After removing of outliers as the values are insignificant s there is no option to remove one variable.

So we can apply **VIF o**n that variables which tells us co-linearity in-between the independent variables i.e; how the i/p variable is getting effected by the other i/p variables and the thumb rue says any **VIF >10** will have co-linearity among all variables. Here no co-linearity among all variables because VIF values are below 10.

So by the above plot we can remove the states variable. We can get the final model with input variables

R.D.Spend and Marketing.Spend.

**Final model:**

From the final model we get the all coefficients are statistically significant so we can consider this model as final. And have **R^2 value is 0.95.** by applying transformations on to input and output variables we can get the better R^2 values which model is having highest R^2 value and by comparing train RMSE test RMSE values are very near to each other we can consider that model.