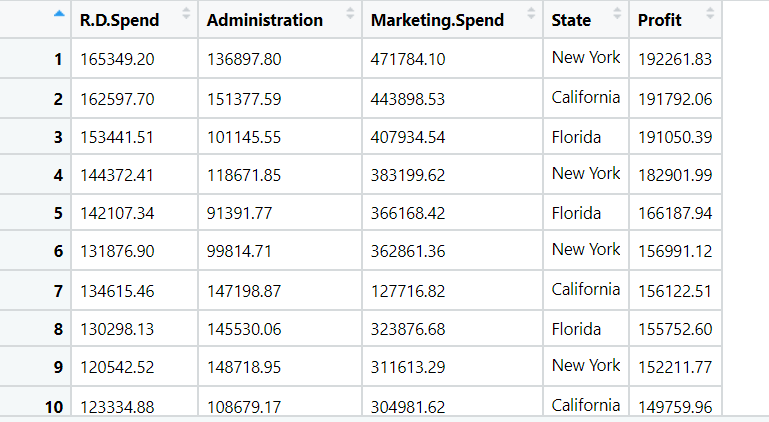
# Multiple Linear Regression (Module -7)

## Q1) Prepare a prediction model for profit of 50\_startups data.

Note: Do transformations for getting better predictions of profit and make a table containing R^2 value for each prepared model.



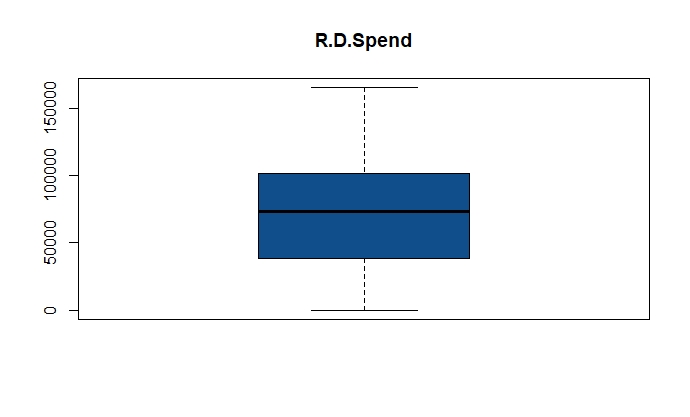
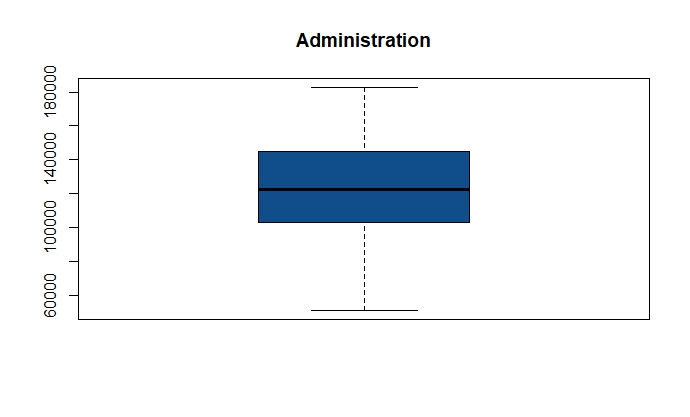
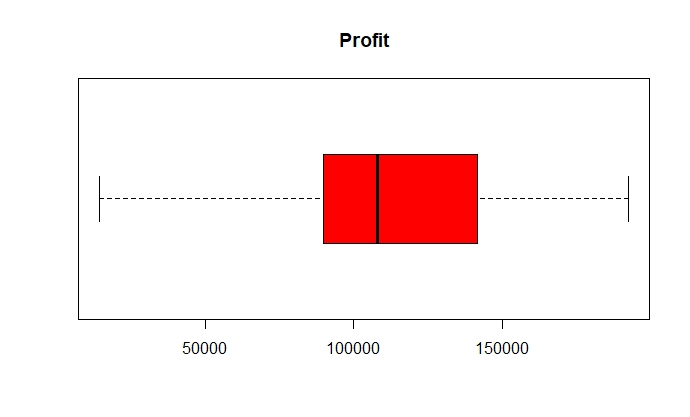
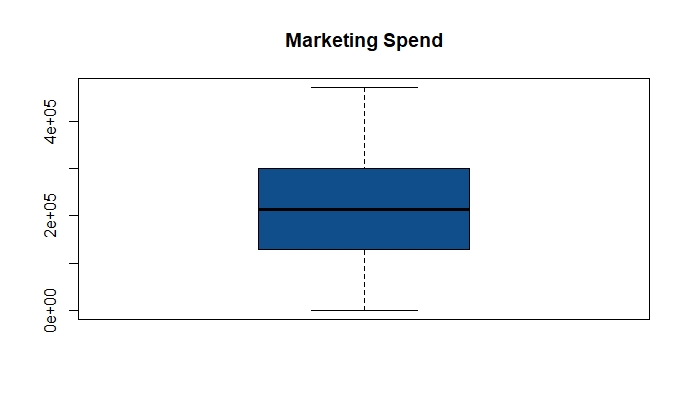
**Ans:**

Output Variable(y) = Profit

Input Variables (x) = R.D Spend, Administration, Marketing Spend, State.

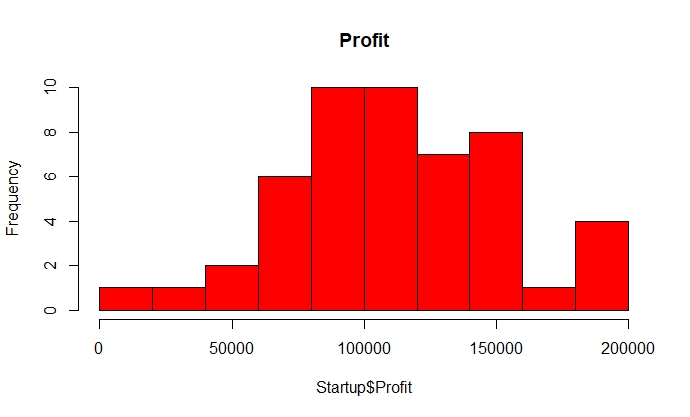
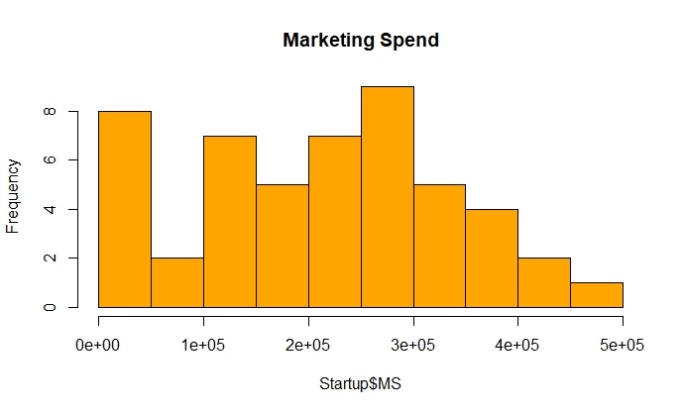
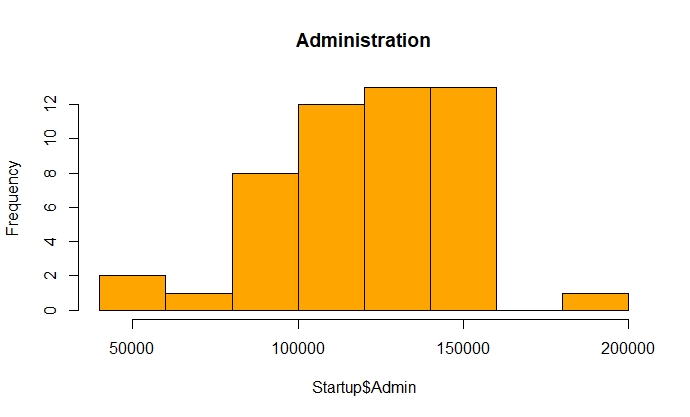
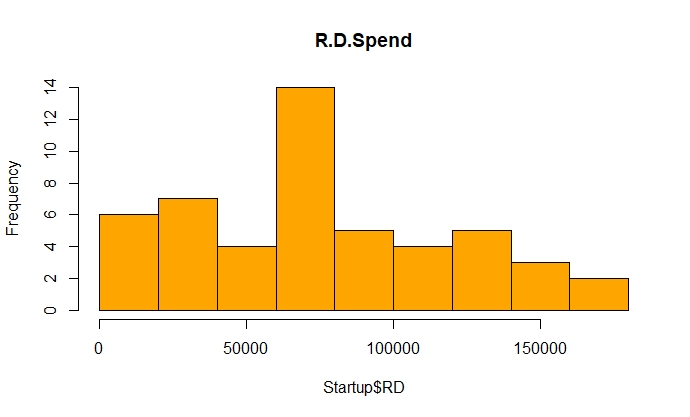
* As the State (x) is a categorical variable, need to create a dummy variable so the type casting will take place and string data will be converted to integers.

**Box Plot (Univariate Plot):**

****

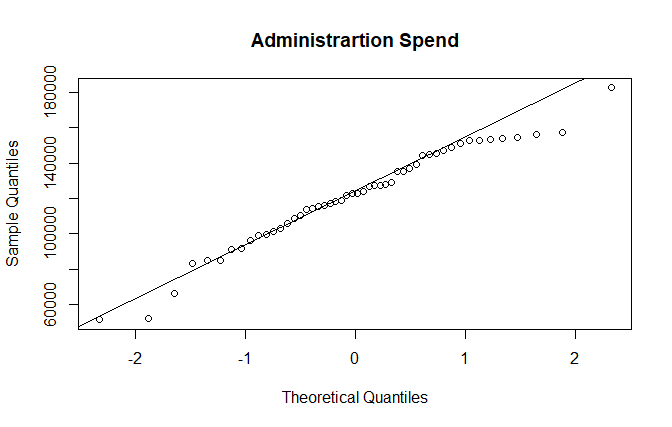
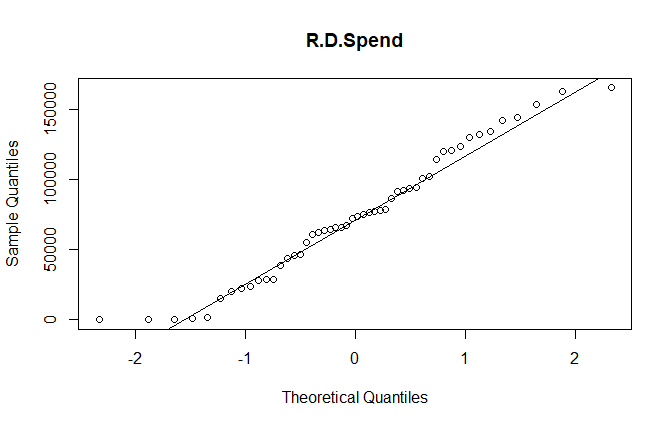
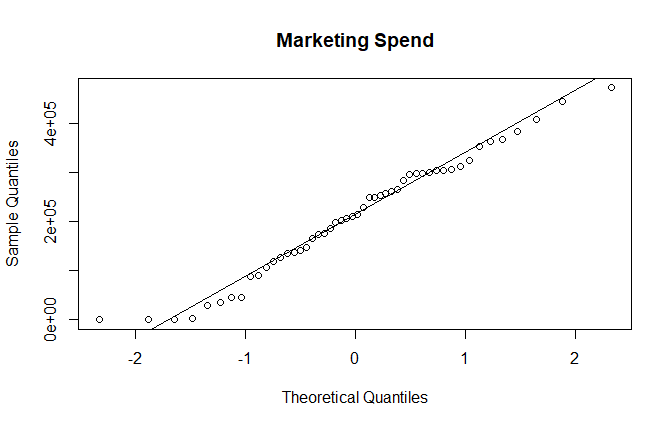
* From the above box plots it shows that there are no outliers exist for individual variables

**Histogram:**

****

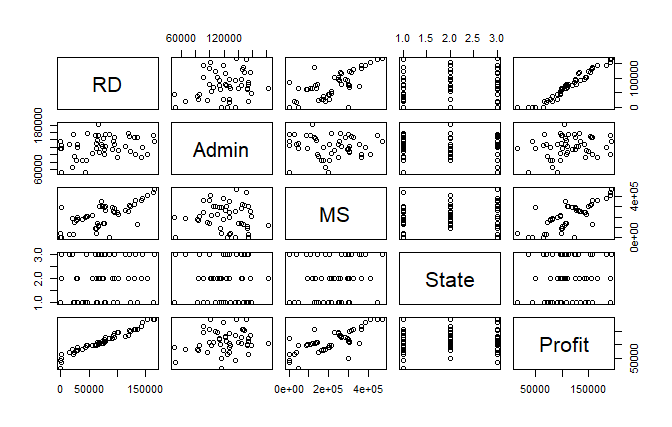
* The graphs indicate that the data is normally distributed for the individual variables

**Plotting:**



* From the above given plots, we can that the data is normally distributed and there are some outliers exist

**Scatter Plot:**



* From the above scatter plot, we can state that there exists a strong relation between RD & Profit and moderate relation between RD and MS

**Model Building:**

* RD 2e-16 \*\*\*
* Admin 0.606
* MS 0.109
* State 0.989

From the above p values of linear model states that the p value is insignificant for Admin, MS and State.

Over all R square is 0.9507 and p is less than 0.05

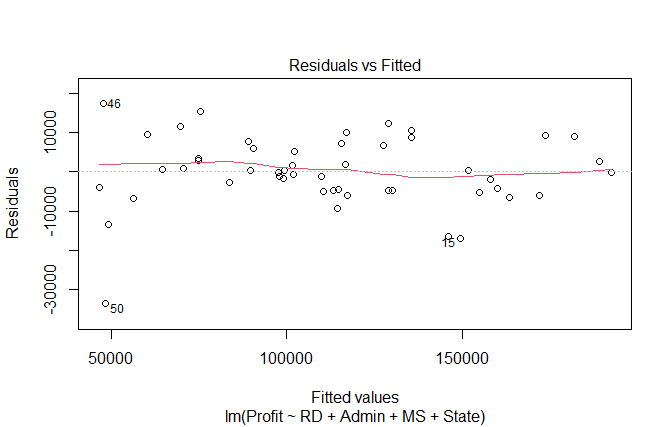
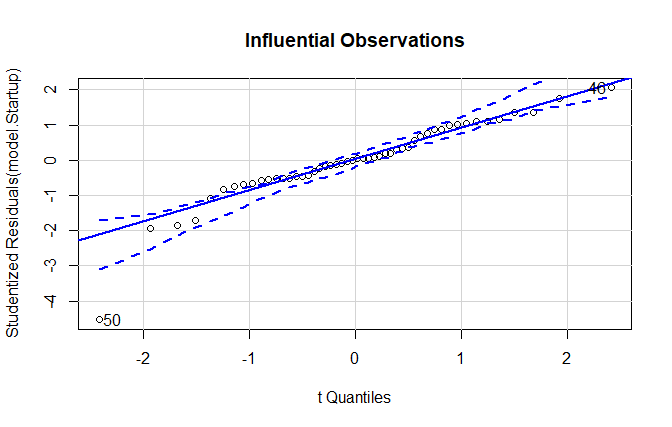
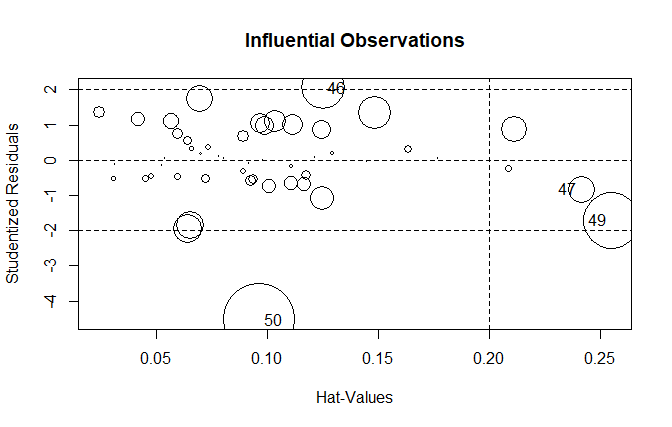
Checking the influence of other variables on the targeted variable:

**For Profit ~ Admin**: P value is 0.1622 which is greater than 0.05 and r square value is also less 0.04029 which makes this insignificant

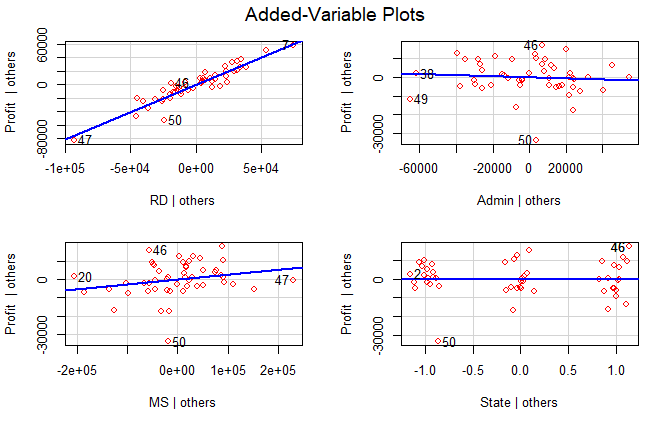
**For Profit ~ MS**: Overall P values are significant so we can consider MS

**For Profit ~ State**: P values making state insignificant

* Identifying influential observations from Diagnostic Plots

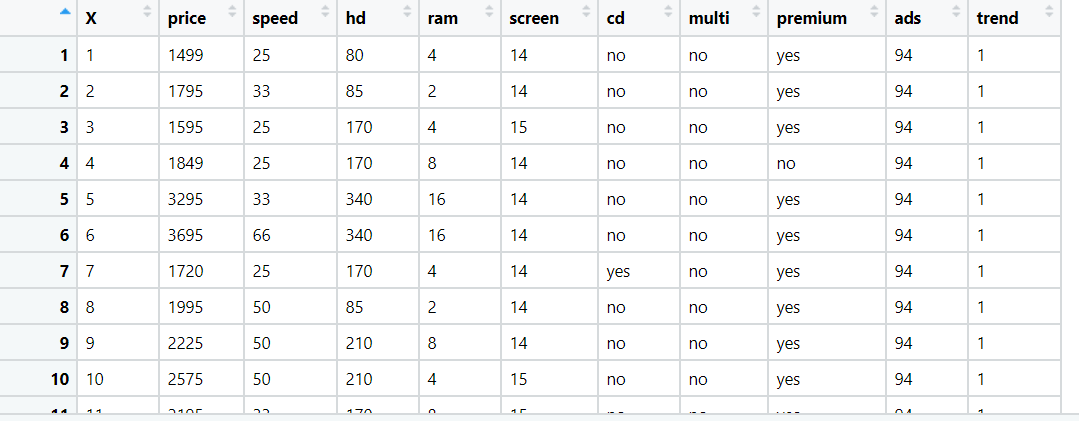


* From the above graphs the influential observations are at 49 and 50th index, build a model after eliminating the influential observations
* Still the variables show insignificant and as there is difference checking with Variance Inflation Factor (VIF is > 10 => collinearity)
* There is no VIF great10 so there is no collinearity problem in this model
* Regression model to check R^2 on Independent variables, still the state shows insignificant



* After removing the ‘state’ variable still the ‘admin’ variable shows insignificant, so build a final model with RD and MS and over all P value is 0.95
* The test rmse = 10843.54 and train rmse = 8101.32 which shows the model overfit and need to do regularization techniques to make the model as right fit.

Q2 Predict the sales of the computer



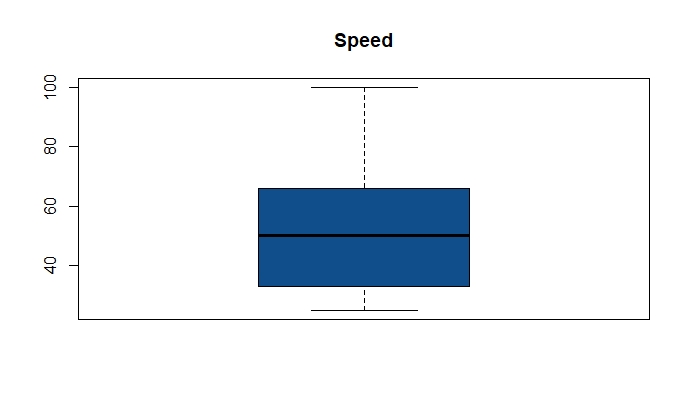
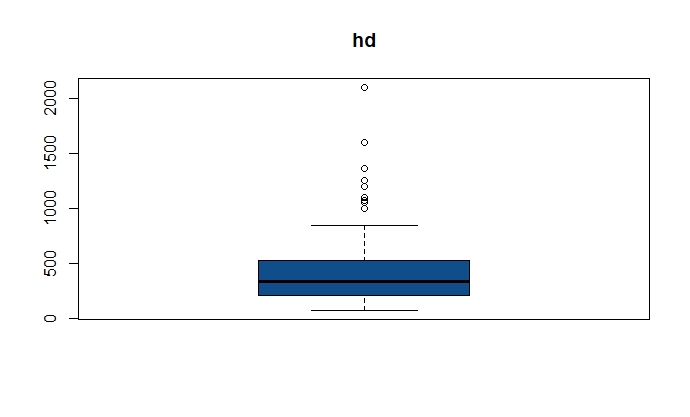
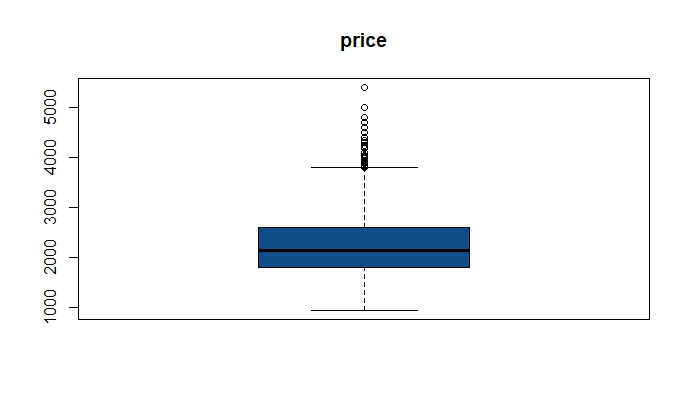
**Ans:**

Input Variables (x) = speed, hd, ram, screen, cd, multi, premium, ads, trend

Output Variable(y) = Sales Price

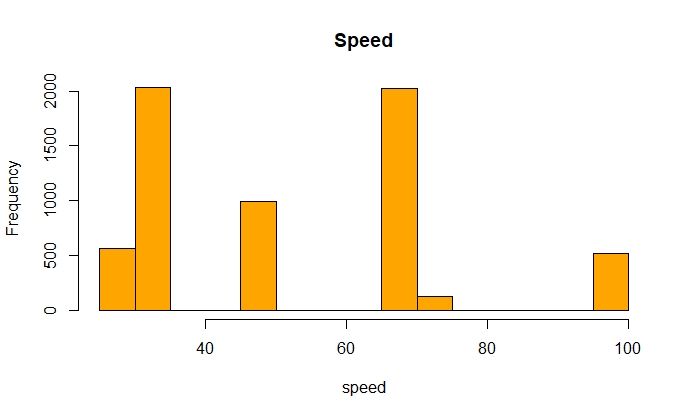
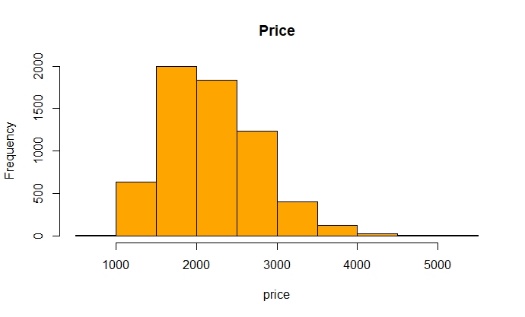
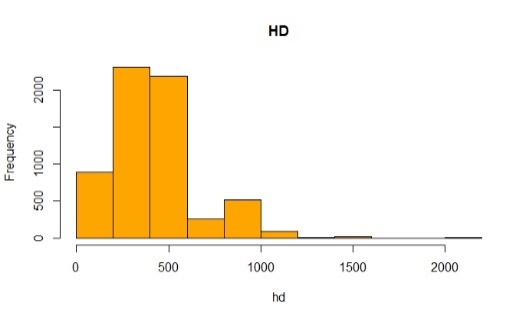
* Removing unnecessary columns so that it should not impact on output of the model
* As there are 3(cd, multi, premium) categorical variable, need to create a dummy variable so the type casting will take place and string data will be converted to integers.

**Box Plot (Univariate Analysis):**

****

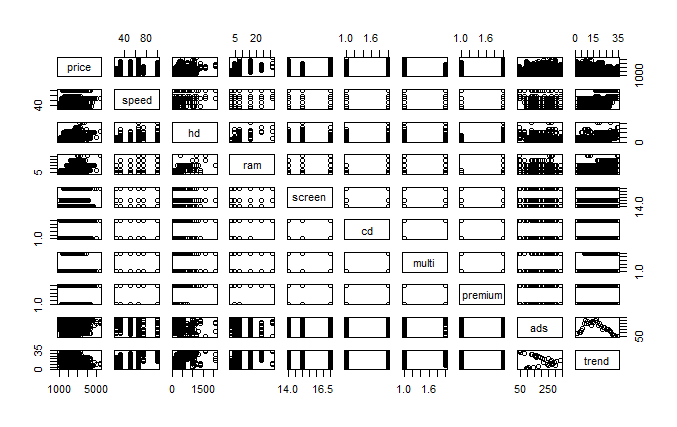
* From the above graphs price and hd having outliers and speed doesn’t have any outliers

**Histogram:**

****

* The graph of HD shows that the data is not normally distributed and it is right skewed in natures and for speed the data is uneven and in case of price the data is normal but there exist outliers

**Scatter Plot:**



* From the above graph there is strong correlation between ads and trends.

**Model Building:**

speed < 2e-16 \*\*\*

hd < 2e-16 \*\*\*

ram < 2e-16 \*\*\*

screen < 2e-16 \*\*\*

cd 1.65e-10 \*\*\*

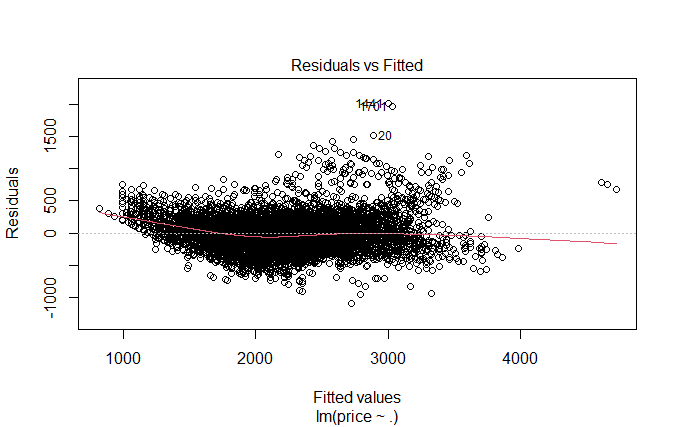
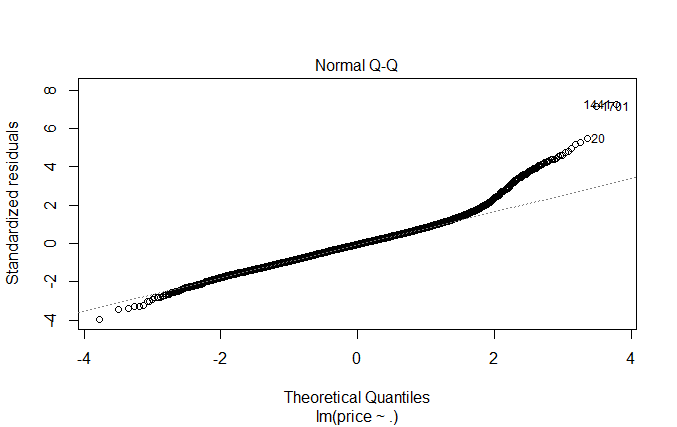
multi < 2e-16 \*\*\*

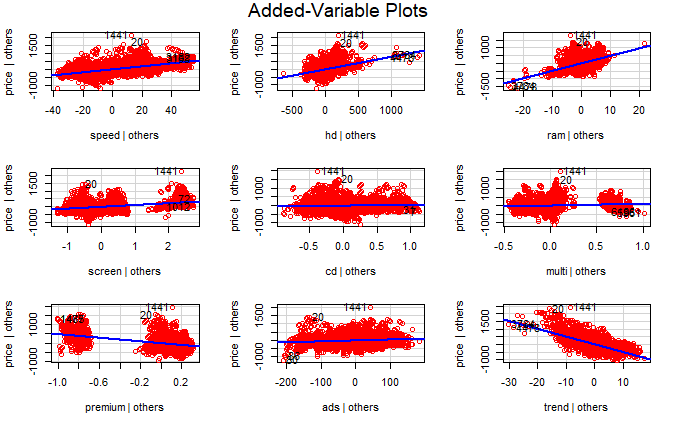
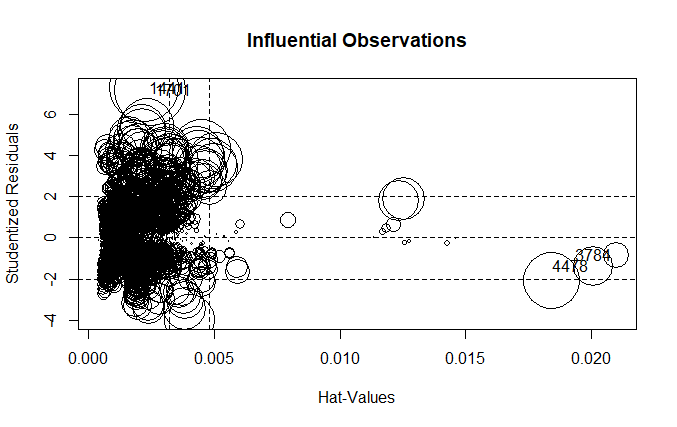
premium < 2e-16 \*\*\*

ads < 2e-16 \*\*\*

trend < 2e-16 \*\*\*

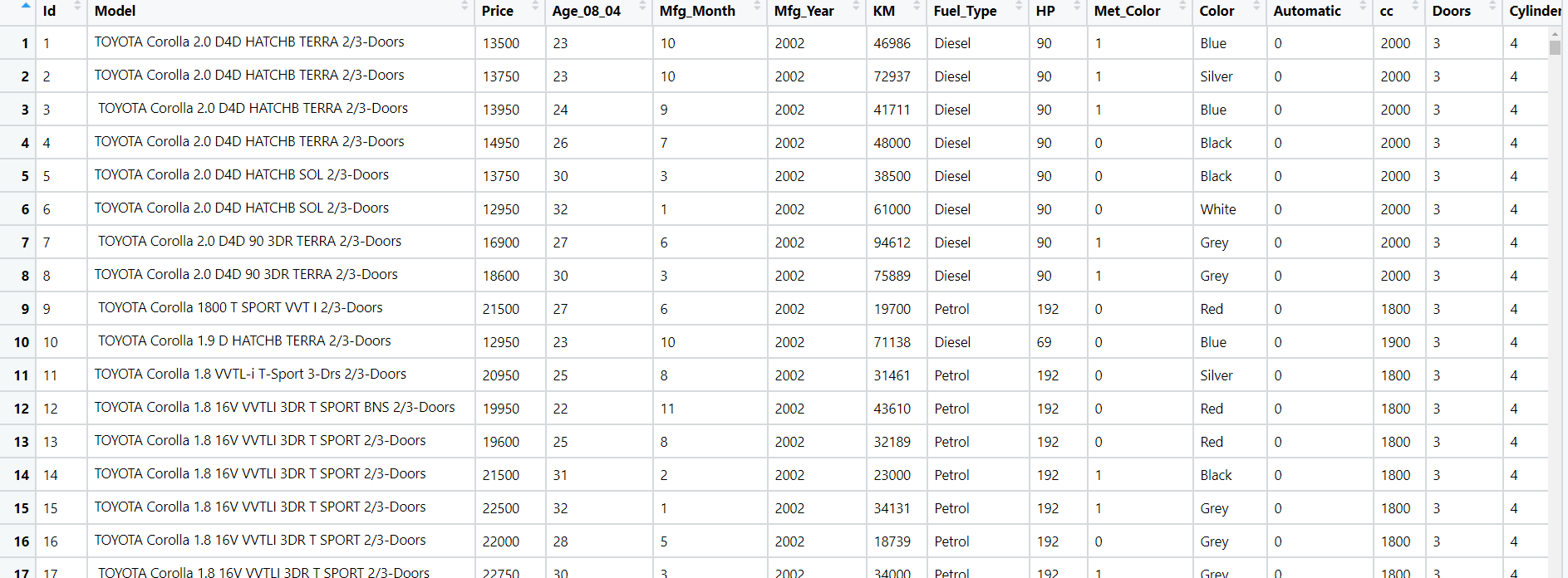
* P- value shows significant for all Input variables
* Over all R square is 0.77 and p is less than 0.05
* As there are no insignificant variables the linear model is the best to build the final model
* Identifying influential observations from Diagnostic Plots





* From the above graphs the influential observations are at 1441 and 1701 index, so building a model after removing the influential observations.
* The residual and fitted graph shows slight variation between the points.
* The final model is built after removing the influential values and the overall R^2 = 0.77 and P- Value < 0.05.
* The test rmse = 277.94 and train rmse = 274.09 and the model shows as right fit.

Q3) 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")]

**Ans:**

Input Variables (x) = Other Variables

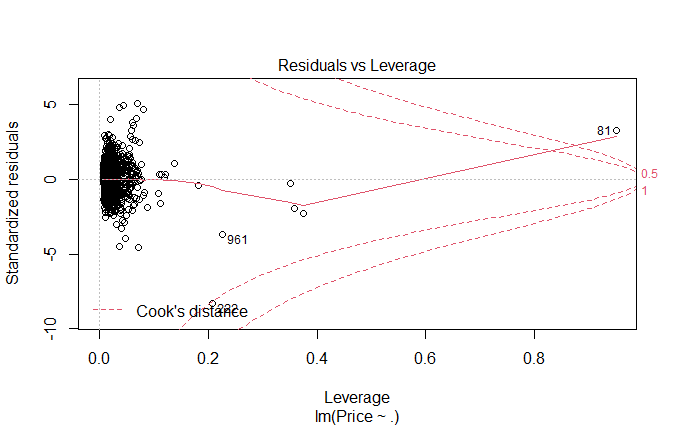
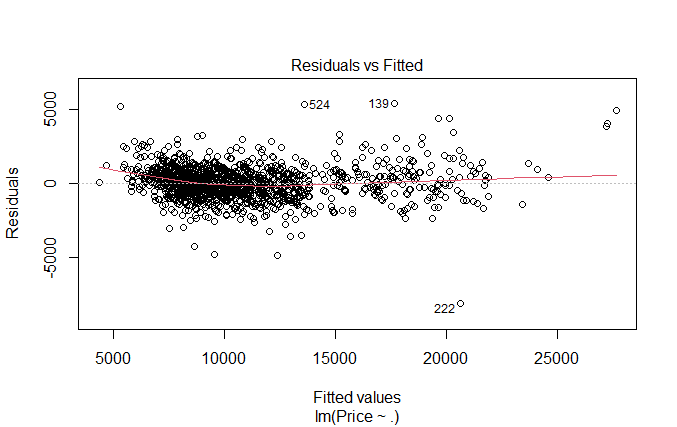
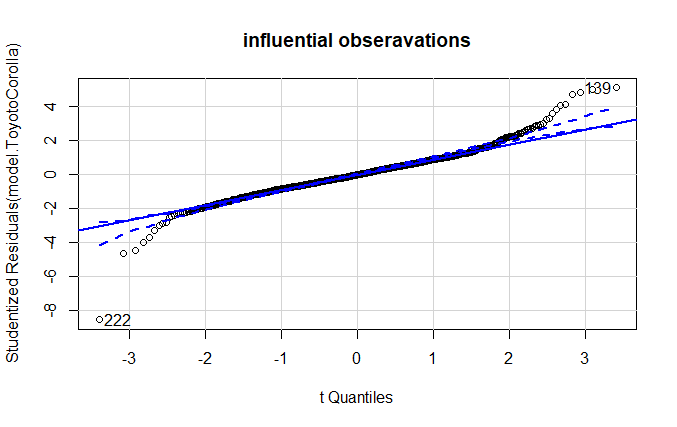
Output Variable(y) = Sales Price

* Removing unnecessary columns so that it should not impact on output of the model
* As there are 3(cd, multi, premium) categorical variable, need to create a dummy variable so the type casting will take place and string data will be converted to integers.
* From the correlation matrix we can say that there exists strong correlation between price and mfg\_year, Hp, Automatic, CC, and many other variables

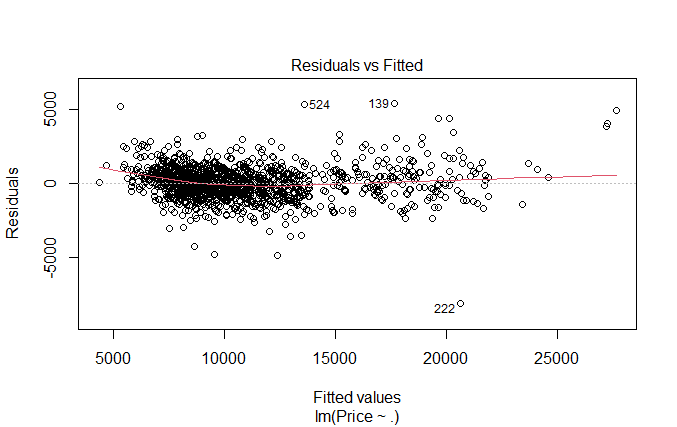
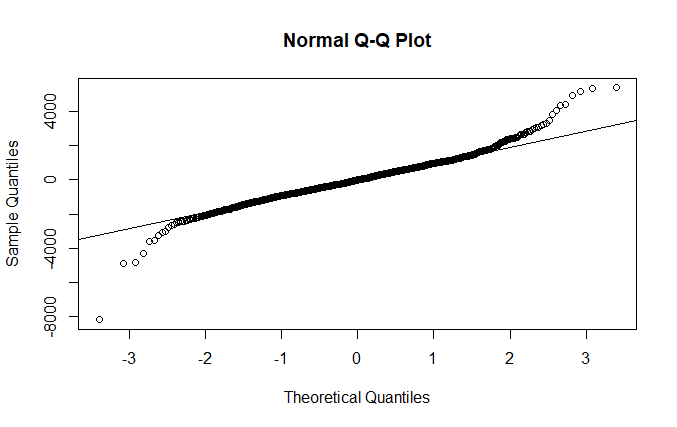
**Model Building:**

* As there are many insignificant values in the model summary need to check variation influence factor and make the model significant but the overall R^2 = 0.909 and P- value < 0.05

**Identifying influential observations from Diagnostic Plots:**

****

* From the above graphs the influential observations are at 139 and 222 indexes, build a model after eliminating the influential observations
* After removing the influential observations, the overall R^2 = 0.914 which has been increased slightly and P- value < 0.05
* Still the variables show insignificant and as there is no difference, checking with Variance Inflation Factor (VIF is > 10 => collinearity)
* After running the VIF the Mfg\_year and cylinder shows insignificant, by eliminating these 2 variables built another model but still the values are insignificant so again running VIF on new built model
* After running the VIF on new model the radio and Radio cassettes shows insignificant, by eliminating these 2 variables built the final model as it is advised not to remove many variables as the data is important
* There is no VIF great10 so there is no collinearity problem in this model
* Plotting the graphs for final model



* From the above graph there is a strong correlation and only slight variation between the residual and the fitted values.
* The model gives the over all R^2 = 0.91 and P – value < 0.05
* The test rmse = 1243.39 and train rmse = 1105.34 which makes the model as right fit.

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without error
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.