# 50 Start up

1. Loading the Data set
2. Data Has no Outliers and NaN values in the data set
3. Histogram shows Data is not normally distributed
4. QQ plot shows Data is linear
5. Creating Dummy variables for States Column

6) Correlation Plot to understand multiple variable

7) Correlation Coefficient to find the strength and direction

8) Plot Used to find the relation between profit and other variables

Profit and RD data positively Linear and normal distribution

Profit and Marketing data is non normal distribution

Profit and Administration is Scatter across the plot and non normal distribution

9) Model Building

Model 1 : All Variables

Result : RSq Value is 0.9507

Found : Administration and Marketing.spend P value is greater than .05

Model 2 : Only Marketing.spend

Result : Rsq value is 0.5592

Found : P value is less than .05. Data is Significant

Model 3 : Only Administration

Result : Rsq Value is 0.04029

Found : P value is greater than .05

Model 4 : Only Administration and Marketing.spend

Result : Rsq value is 0.6097

Found : P value is less than .05, Both are Significant

Model 5 : Model based on Sq rt on all

Result : Rsq Value is 0.908

Found : Administration and Marketing.spend P value is Greater than .05

1. Scatter plot to understand the correlation between two variables.
2. Car library is used identify the influence Index Plot and VIF
3. In Influence Index Plot, observed that 47,49,50 entries as Influence entry

Model 6 : without entry 47,49,50

Result : RSq Value is 0.9614

Found : Administration and Marketing. Spend P value is Greater than .05

Model 7 : Without Administration Features

Result : RSq Value is 0.9601

Found : All the features P value is less than .05.

Variance Inflation Factors ( VIF) greater than 10 means col-linearity in the variables. All the values are less than 10

Model 6 without entry 47,49,50 has best RSq value.

**Final Model**

80% Train Data

20% Test Data

RSq Value: .7869

RMSE is 266.4478