**MULTILINEAR RIGRESSION**

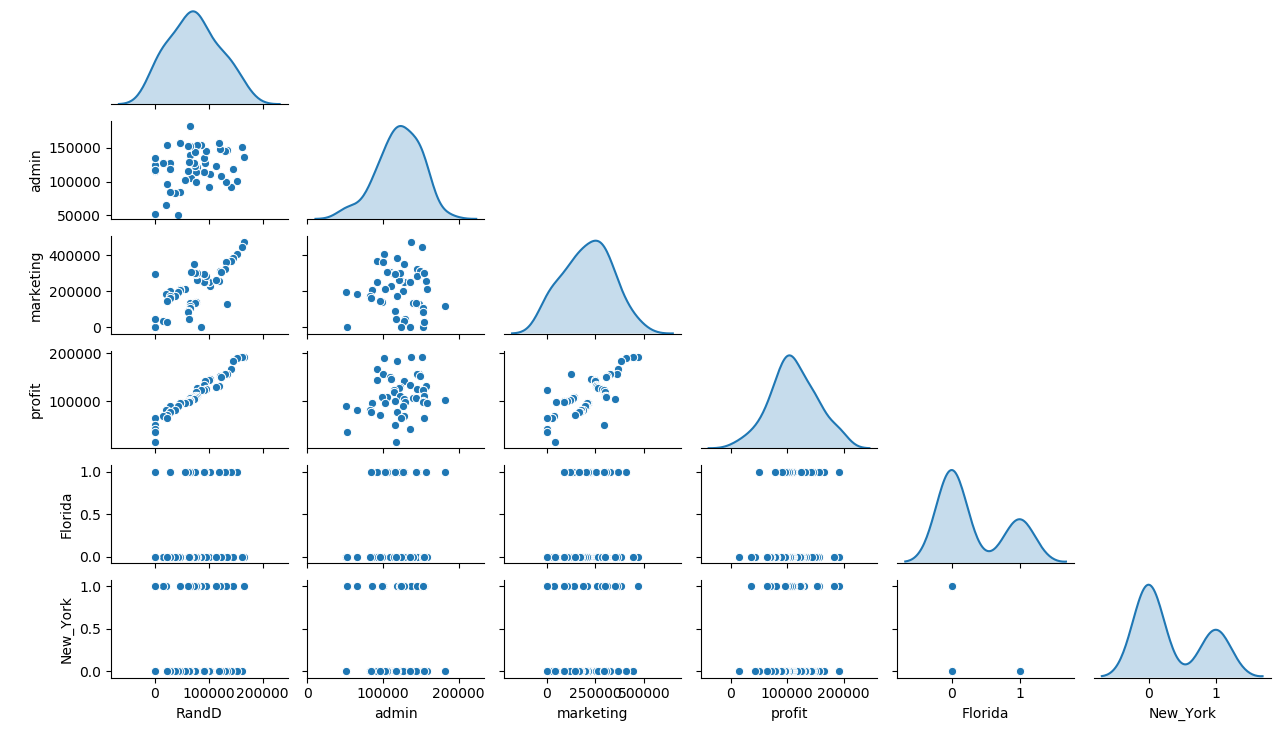
**Business Problem** = ﻿Prepare a prediction model for profit of 50\_startups.

* **Name of the File: -** 50\_Startups.csv
* **Size of the File: -** 2 KB
* **Data: -** 50 Observation, 5 Variable
* **Column Name: -** R&D Spend, Administration, Marketing Spend, State, Profit

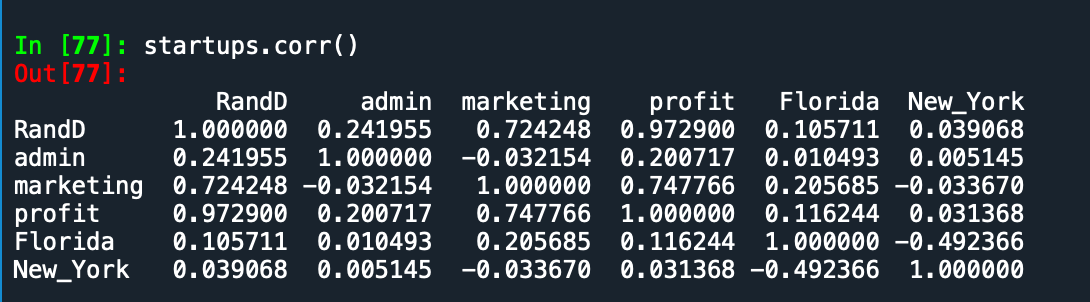
**Exploratory data Analysis** =

* **Outliers: -**  Profit variable having outliers.
* **Missing Value: -** Data don’t have Missing Values
* **Normality: -** Data are near normal
* **Transformation: -**  May be Required to improve accuracy

**Scatter plot =** From below scatter plot we can say that profit and R&D have strong correlation also Marketing and profit have good correlation.

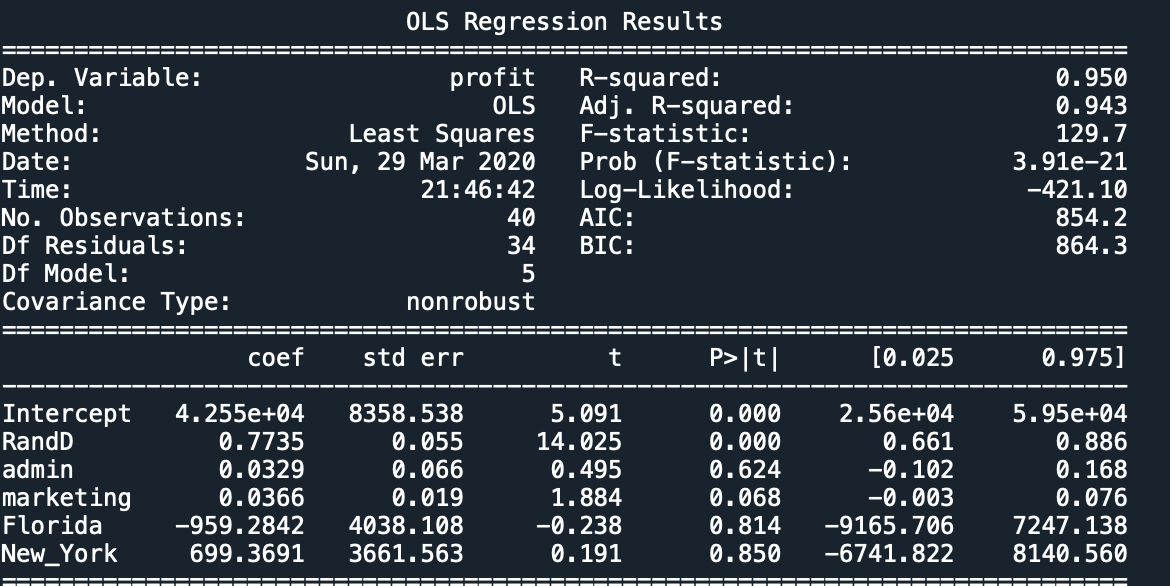


**Correlation Coefficient (r) =**  Correlation between marketing and R&D is quite high as compare to other it may causes collinearity problem.

****

**Model 1 Building =**

* **Summary: -**

****

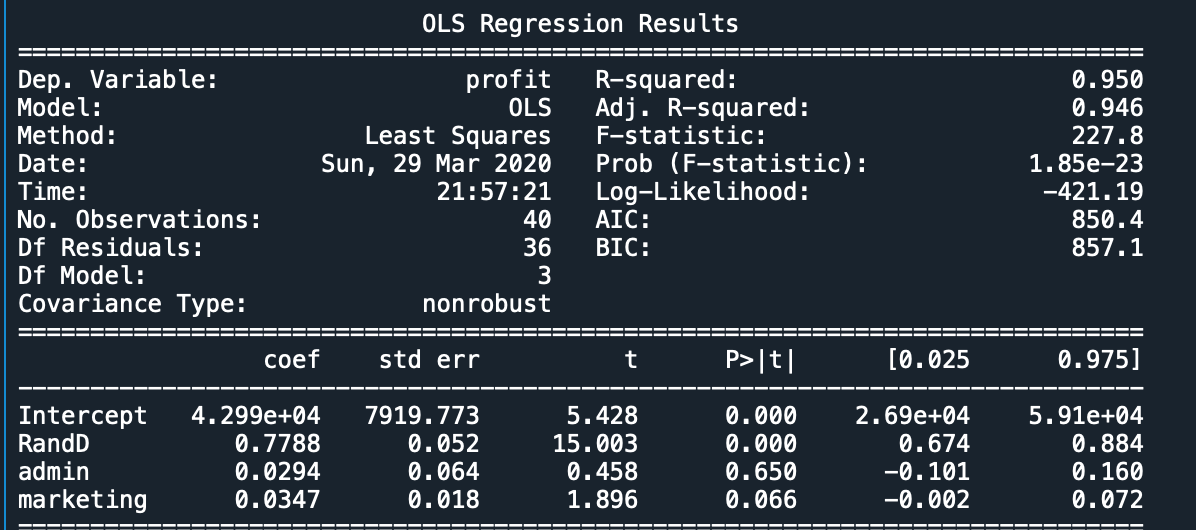
Form above summary we can observe that R2 is high but parameters admin and states are insignificant

So we are building model individually with price

After creating model of each we can conclude that only states is insignificant so we have to remove it and create new model.

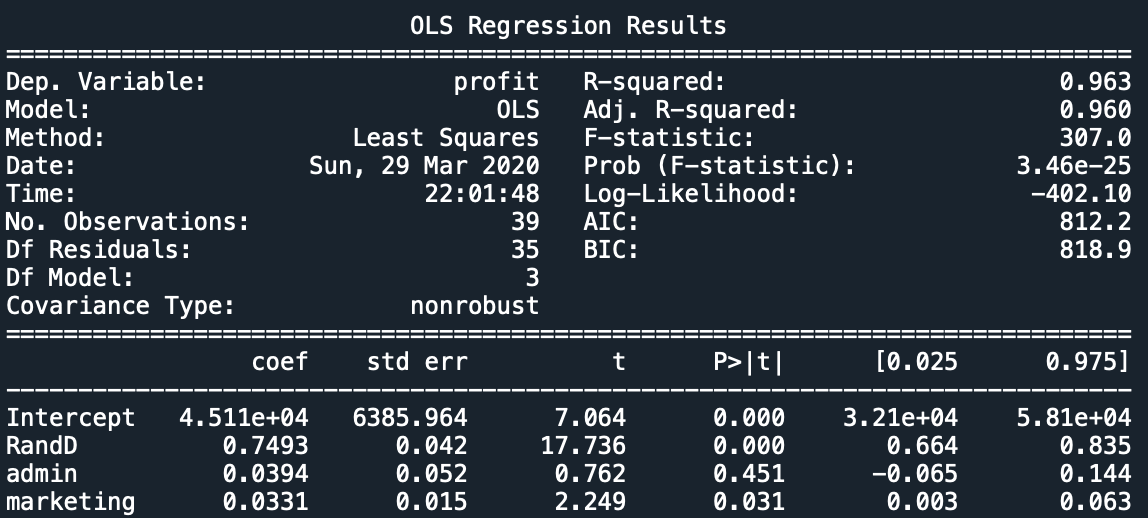
**Model 2 Building =** Building model after removing States

* **Summary: -**

****

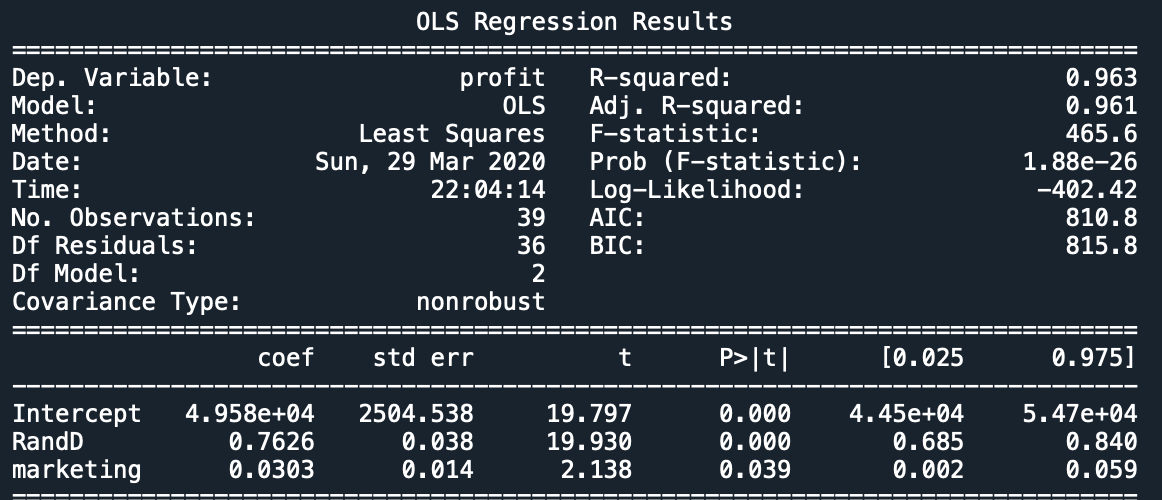
We can see that Marketing as well as Admin are not significant so we have check whether there are any influencer in data. Building model after removing influencer

**Model 3 Building =**

****

Still Admin is not Significant for accurate model we have to remove admin

**Model 4 Building =**

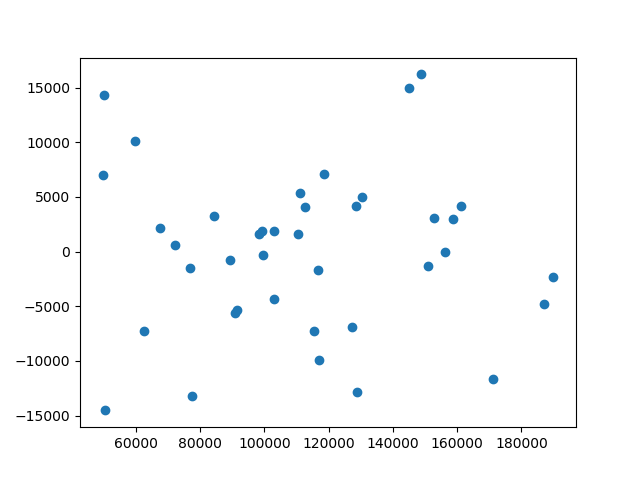
****

In above model R2 value is high , Adj. R2 is also high and all variable now significant we can used this model as final model.

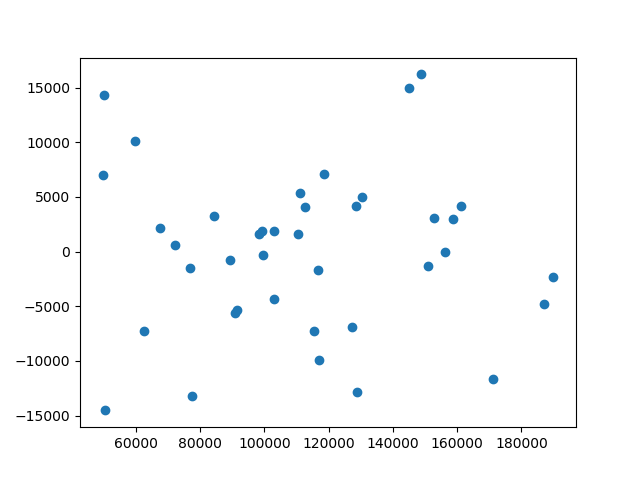
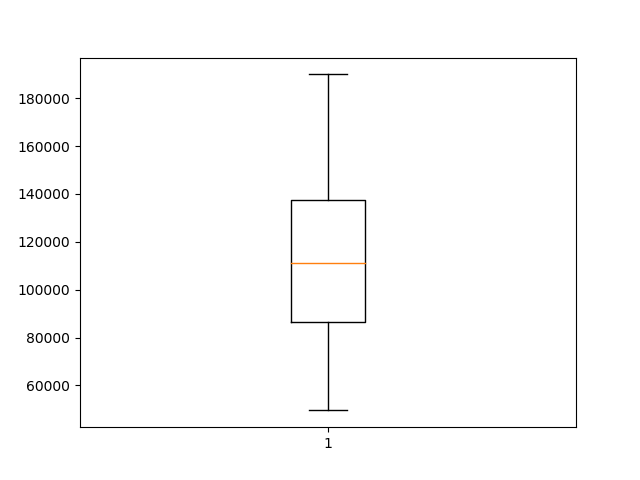
* **Intersect(B0): -** 49580
* **Slope(B1): -** 0.7626,0.0303
* **Coefficient of**

**Determination (R2): -** 0.96 - Strong Correlation

* **Pvalue: -** P value less than 0.05 so we can use B1, B2 and B0 for model building.
* **Model: - ﻿Profit =** 49580 + 0.7626(R&D) + 0.0303(Marketing)
* **RMSE:** - 7329

**Residual Scatter Plot**: -

Error do not have any pattern. It means errors are independent and from below box plot we can say that errors are nearly normal distributed.

****

**Final Models with 95% confidence interval: -**

* **﻿Profit =** 44500 + 0.685(R&D) + 0.002(Marketing)
* **Profit =** 54700 + 0.840(R&D) + 0.059(Marketing)

**Python code file**: - [50 Start up Analysis.py](https://github.com/nilaydeshmukh0/Multiple-Linear-Regression-With-EDA/blob/master/50%20Start%20up%20Analysis/50%20Start%20up%20Analysis.py)