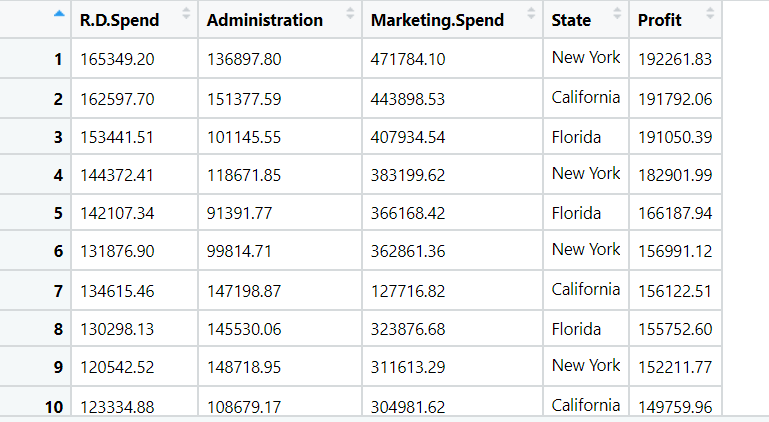
# Multiple Linear Regression (Module -7)

**Problem Statement: -**

An Analytics Company has been tasked by a crucial job of finding out what factors does affect a startup company and will it be profitable to do so or not. For this they have collected some historical data and would like to apply supervised predictive learning algorithm such as Multilinear regression on it and provide brief insights about their data. Predict Profit, given different attributes for various startup companies.



**Business Objective- build a ML model that predict Profit** .

**Python code details :**

Data Frame name is startup . It has 50 entries and 5 features.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance . we have changed Columns name to 'rd\_spend', 'Administration', 'm\_spend', 'State', 'Profit'.

Then we check for unique value in each columns

:-

rd\_spend 49

Administration 50

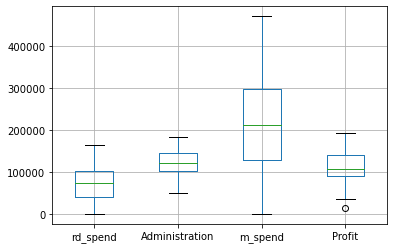
m\_spend 48

State 3

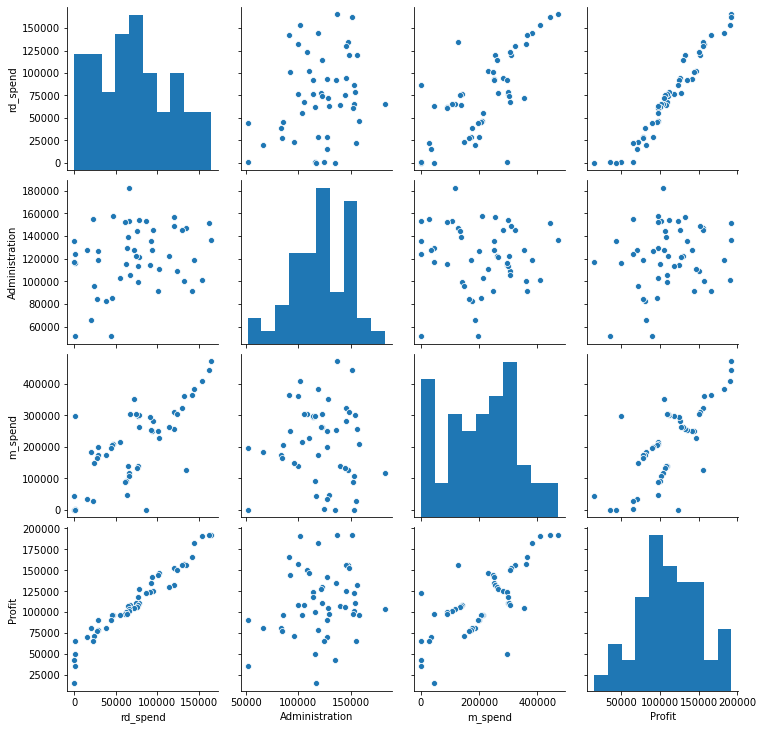
Profit 50

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .there is no outliers present except Profit.

Boxplot:- 

In Profit column we have low value outlier ,so we replaced with lower limit .

Histogram and Scatter plot:- 

If we take a look at our Dataset we can clearly see that State is a String type variable and like we have discussed,We cannot feed String type variables into our Machine Learning model as it can only work with numbers.To overcome this problem we use the one hot encoding object and create Dummy Variables using the grt dummy .

Using seaborn (pairplot) in python we can check for distribution and correlation between each other. According scatter plot strong positive correlation between profit and rd\_spend and also some relation between profit and m\_spend. no correlation of Profit with State and Administration. according to correlation coefficient no correlation of Administration & State with Profit.

**Correlation**

**rd\_spend Administration m\_spend Profit**

**rd\_spend 1.000000 0.241955 0.724248 0.972900**

**Administration 0.241955 1.000000 -0.032154 0.200717**

**m\_spend 0.724248 -0.032154 1.000000 0.747766**

**Profit 0.972900 0.200717 0.747766 1.000000**

**Model Building**

**Build the model on the scaled data (try multiple options)**

**Perform Multi linear regression model and check for VIF, AvPlots, Influence Index Plots.**

**Train and Test the data and compare RMSE values tabulate R-Squared values , RMSE for different models in documentation and provide your explanation on it.**

**Briefly explain the model output in the documentation.**

Using library statsmodels.formula.api as smf # for regression model

For ml1 p-values for State, Administration are more than 0.05

**Model name R^2 value Adj. R-squared: Intercept**

**Ml1 0.951 0.946 5.014e+04**

**Ml\_new 0.96 0.958 5.332e+04**

**Ml2 0.95 0.948 5.012e+04**

**Final Ml3 0.961 0.959 5.012e+04**

Model ml1 have some problem administration (0.606)

Are high . and then Checking whether data has any influential values . index 49 is showing high influence so we can exclude that entire row and build new model ml\_new but issues are not solved .

So we Check for Colinearity to decide to remove a variable using VIF

**Variables VIF**

**0 rd\_spend 2.495511**

**1 Administration 1.177766**

**2 m\_spend 2.416797**

**3 State 1.030434**

According to correlation coefficient no correlation of Administration & State with Profit

So , we build ml2 without column State but p value of Administration is high .

Then we build ml3 without Administration columns . model ml3 has no issue . that’s why we build our final model on this . residual is normally distributed .

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a Simple linear regression on training data set

R-squared: 0.94

Adj. R-squared: 0.94

,then test on test data ,

Evaluation on Test Data as result root mean square error=7201

Evaluation on Train Data also as result root mean square error=9191

Used library –

pandas for data manipulations

Numpy for Numerical Calculatations

Sklearn for Data mining / Machine learning

Matplotlib for Data visualization

Seaborn for Advance data visualization

Scipy for Advance data visualization

Statsmodels for Regression models