Multi Linear Regression

Problem Statement: – To predict the profit earned by a startup company provided the R&D, admin, marketing expenses along with the state location.

Dataset: – 50_Startups

Algorithm Used – Multi Linear Regression Model

Assumptions made while preparing a Multi linear regression is that:

1. All features have a linear relationship with the outcome or output variable (L).
2. Features should not be dependent on each other, if so the phenomenon of multi-collinearity will exist and there by impacting the o/p prediction or estimates.
3. Multiple regression assumes that the residuals are normally distributed.
4. This assumption states that the variance of error terms are similar across the values of the independent variables.

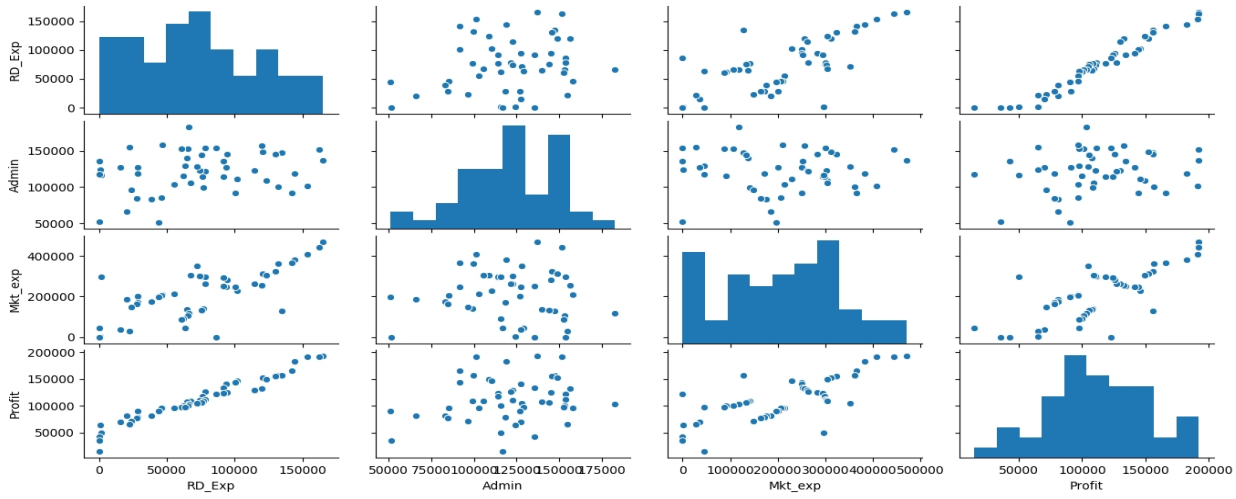
Feature Description: -

1. R&D Spend: - Research and develop spend in the past few years
2. Administration: - Spend on administration in the past few years
3. Marketing Spend: - Spend on Marketing in the past few years
4. State: -States from which data is collected (location of the company)
5. Profit: - Profit of each state in the past few years

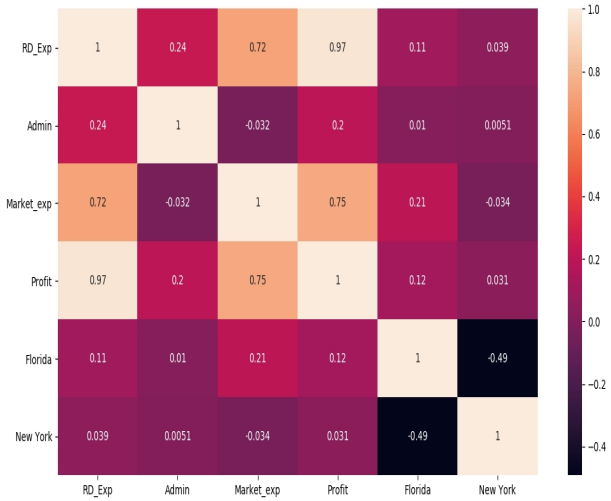
Modelling Process followed –

1. Reading the data (using pandas – pd)
2. EDA or Exploratory Data Analysis (Involves finding influencers, treating the outliers, removal of the same and data cleansing)
3. Model input – to test the data after cleansing the data
4. Splitting the data – To split the data into test and train
5. Checking the statistics and accuracy of the model.
6. Final model presentation

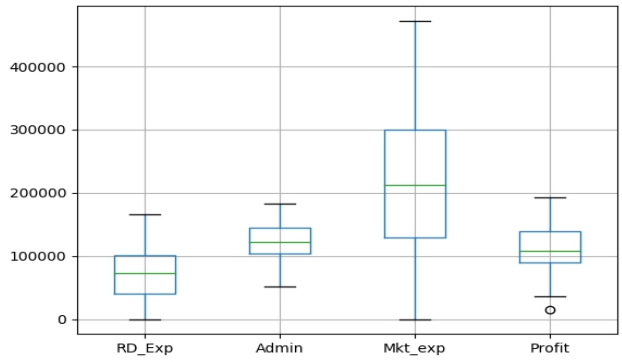
Pairplot – (sns.pairplot(dataframe_name)



Correaltion plot (heatmap using seaborn)



Boxplot – for checking the outlier presence in each feature.



After building the model taking raw input into account (without any data wrangling), we get the below features.

Regression Results:

R-squared: 0.951
Adj. R-squared: 0.945
F-statistic: 169.9
Prob (F-statistic): 1.34e-27
Log-Likelihood: -525.38
AIC: 1063.
BIC: 1074.

Significance stats:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
St[T.Florida]	198.7888	3371.007	0.059	0.953	-6595.030	6992.607
St[T.New York]	-41.8870	3256.039	-0.013	0.990	-6604.003	6520.229
RD_Exp	0.8060	0.046	17.369	0.000	0.712	0.900
Admin	-0.0270	0.052	-0.517	0.608	-0.132	0.078
Mkt_exp	0.0270	0.017	1.574	0.123	-0.008	0.062

As it can be observed that State variable is highly insignificance we wouldn't be considering the same for our final model preparation.
So after checking the influenced plot it was observed that there are few observations which are influencing the output (Profit).

Hence we will be removing few observations i.e. Row index – 46, 49.
Building the model after removing the above said values, we receive the below stats.

Regression Results:

R-squared:	0.960
Adj. R-squared:	0.958
F-statistic:	355.5
Prob (F-statistic):	7.48e-31
Log-Likelihood:	- 495.51
AIC:	999.
BIC:	1007.

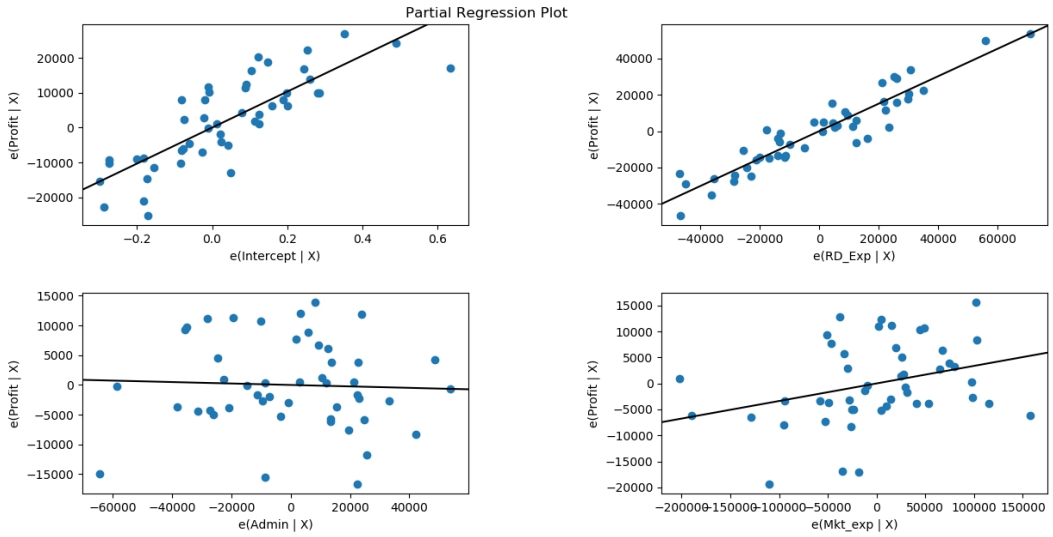
Significance stats:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.158e+04	5529.944	9.328	0.000	4.04e+04	6.27e+04
RD_Exp	0.7563	0.043	17.601	0.000	0.670	0.843
Admin	-0.0122	0.043	-0.282	0.779	-0.099	0.075
Mkt_exp	0.0338	0.015	2.229	0.031	0.003	0.064

As it can be observed that Admin variable is highly insignificance we wouldn't be considering the same for our final model preparation after we check for partial regression plot.

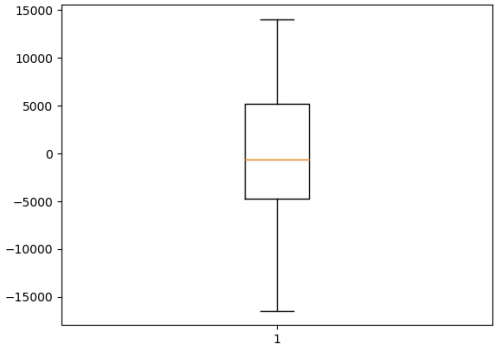
A partial regressio n plot is a graphical plot which is plotted taking the depende nt and individua

l features
into account.



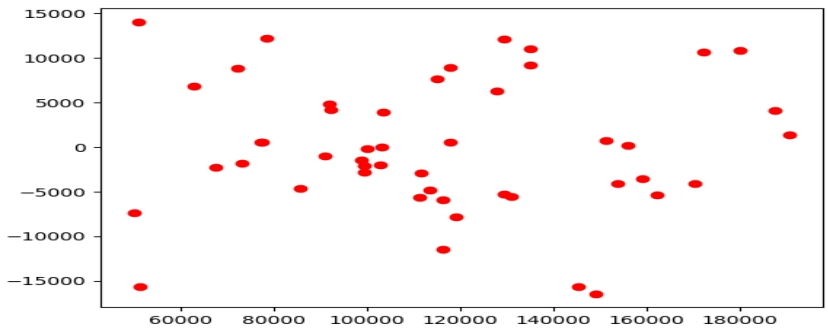
From here it can be observed that admin feature has no linear relationship with the output and there are not much significant changes in the output variable a well.

Error Distribution box plot – The plot shows that the error distribution is normal and the same can be identified by the mean/median/mode.



We got to know that the mean of the residue (y-y_pred) is almost equal to 0 (-2.531427e-11)

Now, after plotting the residuals we observed that the values are scattered all over the hyper plane.



Plotting the actual vs fitted
values in the hyperplane