**50 Startup's Data Report**

* In the data set we have 5 feature’s 4 are Independent Features and 1 Dependent Feature
* State feature is Object data type which contains 3 Categories of data Florida, California, New York
* The Dimensions of the data set is (50,5)

Pre-Processing The data Set

* It is not a perfect format for Names in data Set Features Names are modified.
* There are no Missing Values and NaN values in the dataset
* Checking for outliers in Boxplot Profit which is Dependent Variable has a Outlier for a Reason which is acceptable
* Plotted Graphs Distplot for the data set
* Created Dummy Variables for the State Feature

Correlation

* Checking Correlation with the Data Set
* Correlation between the Independent and Dependent Features are partially Correlated

Creation of Best Fit

* Checking with all Variables Created a model and fit with the data
* By that model has some in-significant variables which are Admin and Marketing Spend whose p-values are greater than 0.05 which are in-Significant

Here we have Multiple in-Significant Feature’s which means Which are not fit for you model.

features are (Administration and Market Spend) based on P value which are greater than 0.05 value

Now check for is there any problem called Multicollinearity by using (Simple Linear Regression)

Independently Checking in-Significant Features

* here checking for the First In-Significant feature's Independently for Multicollinearity==> (0.162) greater than 0.05 Again it becomes In-Significant Feature checking with Admin
* here checking for the Second In-Significant feature's independently for Multicollinearity==> It Became Significant Feature whose p value is 0 which is less than 0.05 Checking with MS

Checking with Both Significantly and In-Significantly

* Now check Both first and Second In-Significant variable by Combining ==> Here we go again One Significant and Another one In-Significant feature
* Whose p-value is Admin🡺0.017
* Whose p-value is MS🡺0.00

Now Check Removing One by One feature with Metrics of R^2 and AIC values

* Removing the Admin Feature
* Check by Removing one Feature by another and check the R^2 and AIC Values based on these choose the Best fit Model Excluding Administration Spend
* Removing Admin Feature
* Whose p -Values are less than 0.05
* Whose R^ 2 value is 🡺0.950
* Whose AIC value is🡺1061

Removing the MS (Marketing Spend)

* Check by Removing one Feature by another and check the R^2 and AIC Values based on these choose the Best fit Model Excluding Marketing Spend
* Removing MS Feature
* Whose p -Values are greater than 0.301
* Whose R^ 2 value is 🡺0.948
* Whose AIC value is🡺1064

Here We Don’t Have the Multi-Collinearity Problem that’s why we are did not finding the VIF value

Model Validation techniques:

* QQ-plot which are useful for Detecting the Outliers and Influential Variables
* Standardized Data and plot the Scatter plot we have to check is there are any patterns in the data or not, if there is no patterns in data then it is consider as best fit model
* Residuals Vs Regressor Plots which are checking for difference between Actual values and Predicted values

Model Deletion Techniques

**Cook’s Distance**: cooks Distance and Influence Plot which helps in Detecting the outliers and Influential observations

**High Influence Plots**: By calculating the Leverage value we can know the influential observations in the data set by influence plot we found that 49,14,46 are the Highly influential plots

**Here Improving the Final Model**

* Again, Loading the dataset and start preprocessing the data set
* Again, change the Names of the features
* Drop or Remove the Influential rows in the data set
* Create a model after dropping the rows
* Here we are Creating the model with out using the Admin Feature because we are getting the High R ^2 and AIC values
* Final model R^2 values are 🡺0.966
* Final model AIC values are 🡺975

**Cross Checking for the R Squared Values with Models Using Final Model has the Highest R-Squared Value**

| **Models** |  | **R\_Squred\_values** |
| --- | --- | --- |
| **0** | all\_variables | 0.950752 |
| **1** | Admin & MS | 0.609723 |
| **2** | with out Admin | 0.950453 |
| **3** | With out MS | 0.947980 |
| **4** | Final Model | 0.965783 |

In [178]:

**AIC Values Cross Checking Whatever the AIC value is Low value that is Best Fit Model**

|  | **Models** | **Aic\_values** |
| --- | --- | --- |
| **0** | All variabels | 1062.764910 |
| **1** | with admin & Ms | 1160.264804 |
| **2** | without Admin | 1061.067742 |
| **3** | without Ms | 1063.503444 |
| **4** | Final Model | 975.526281 |

**Adj- R^2 values Cross Checking with all the Models Created**

Normally R-Squared values will be Fluctuating when the feature Variables are added or deleted but the Adjusted R-Square will not be Fluctuated When the Variables added or deleted

Adjusted R-Squares will only be increased when the Feature Variable is Significant and that feature variable will be used for the Dependent Variable

|  | **Models** | **Adj\_rsquares** |
| --- | --- | --- |
| **0** | All variabels | 0.945 |
| **1** | with admin & Ms | 0.593 |
| **2** | without Admin | 0.966 |
| **3** | without Ms | 0.963 |
| **4** | Final Model | 0.966 |