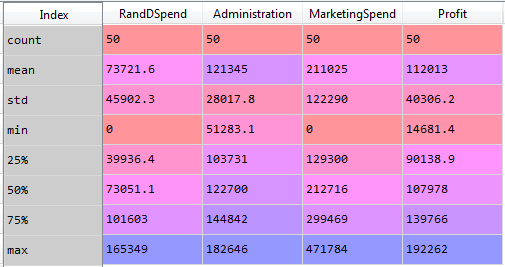
**Case Study of 50StartUps**

**Problem Statement:**

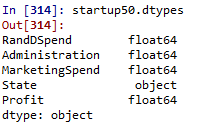
Prepare a prediction model for profit of 50\_startups data. Do transformations for getting better predictions of profit

**Exploratory data analysis:**

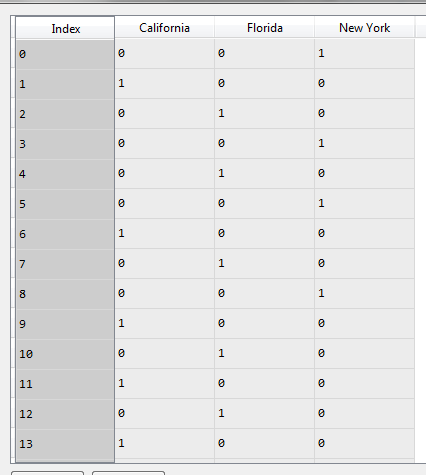


The above is the descriptive statistics of the 50 Startups dataset. There are 50 observations and 5 variables in the data frame. The missing variable in the above descriptive analysis is the “State” variables. Which is a categorical variable.

The following are the variable types:

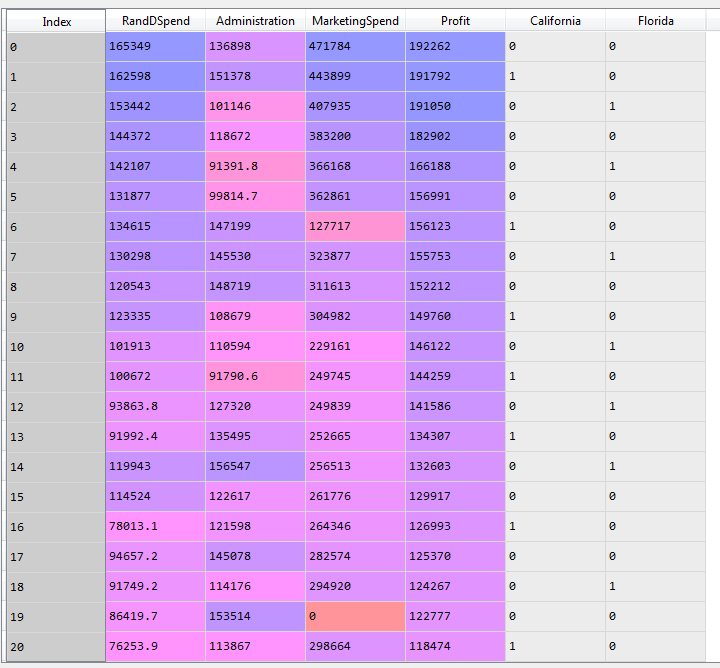


The categorical variable “State” is converted to the dummy variables of three named “California”, “Florida” and “New York” as follows:

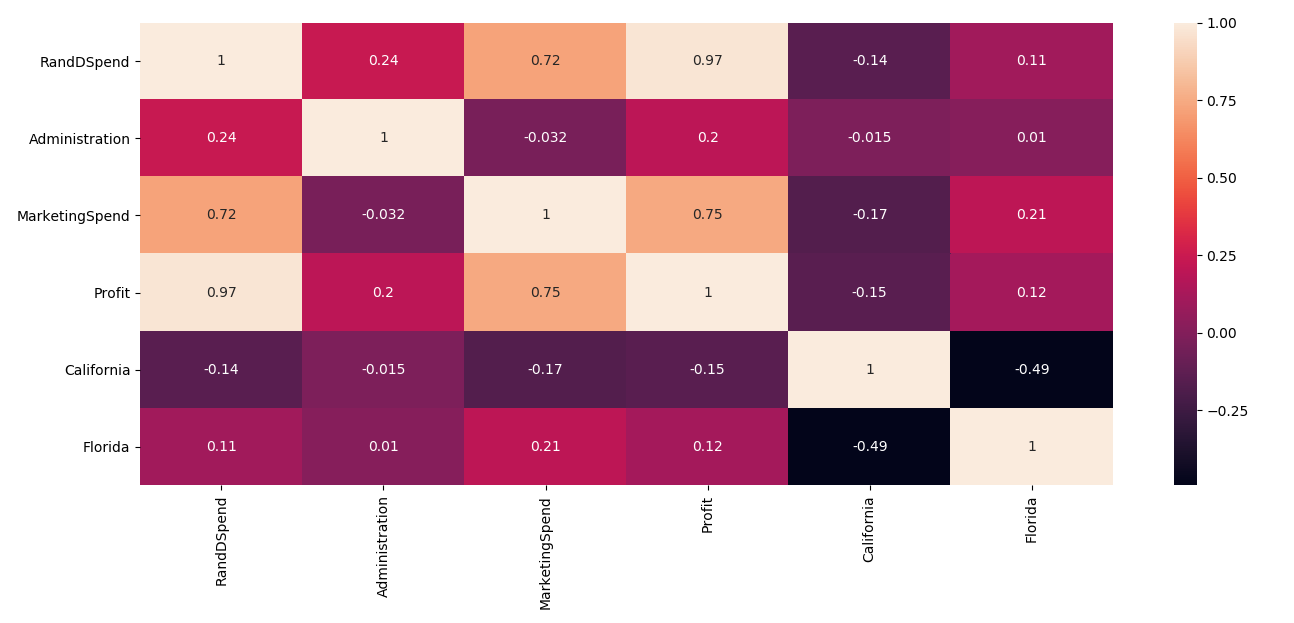


The variable “New York” is dropped including the categorical variable “State”. And the rest of the dummy variables “Florida” and “California” is concatenated to the main data frame of 50Startups.

Finally, the dataset looks like the following:



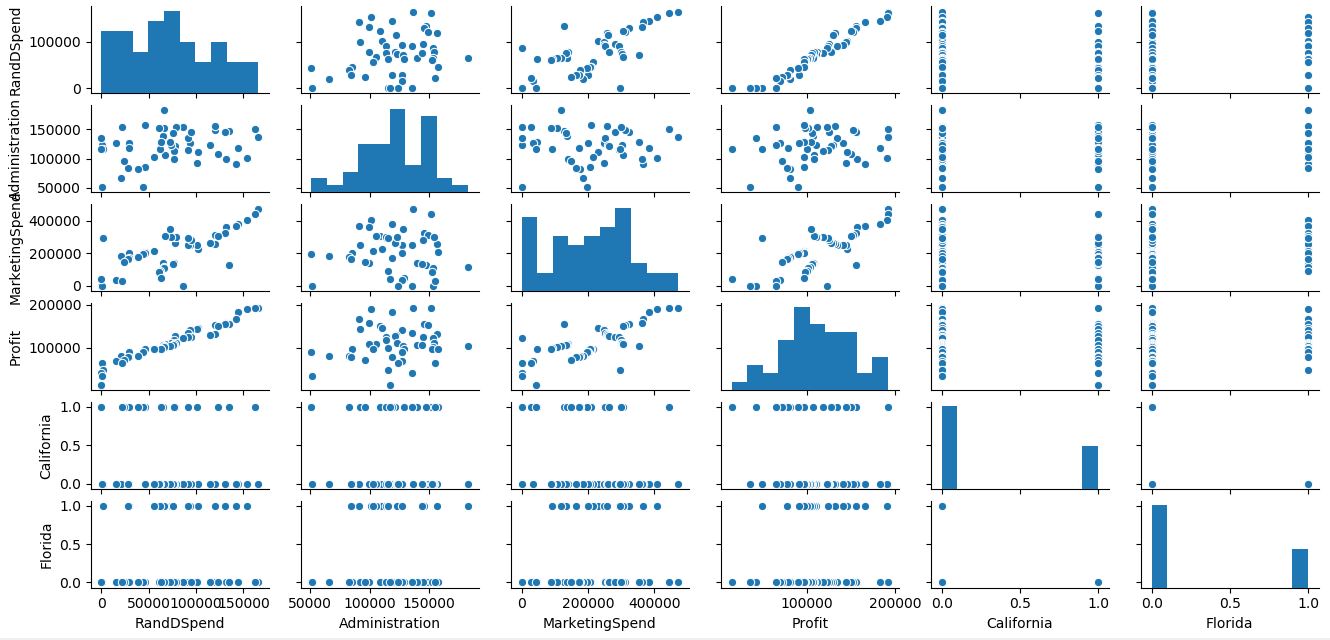
The heat map of the correlation between different variables



From the above heat map we can estimate the following:

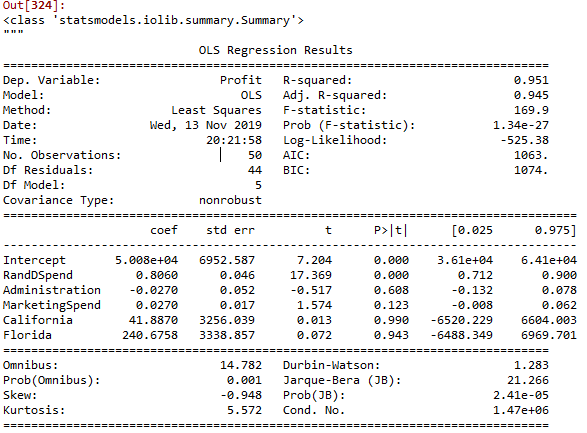
* High negative correlation between Profit and R&D Spend
* The correlation between independent variables are no significant enough to consider Multi Co linearity problem in the given dataframe.

The following is the pairplot of the dataframe:



Building the regression model:

The following is the model considering all the variables



From the above model summary.

* All the variables except R&DSpend are insignificant to the outcome

The above models have insignificant variables that needs to be treated.

# Preparing model based only on MarketingSpend

ml\_MarketingSpend=smf.ols('Profit~MarketingSpend',data = startup50).fit()

ml\_MarketingSpend.summary()



# Preparing model based only on Administration

ml\_Administration=smf.ols('Profit~Administration',data = startup50).fit()

ml\_Administration.summary()



# Preparing model based only on the states

ml\_states=smf.ols('Profit~California+Florida',data = startup50).fit()

ml\_states.summary()



# Preparing model based only on R&D Spend

ml\_RandDSpend=smf.ols('Profit~RandDSpend',data = startup50).fit()

ml\_RandDSpend.summary()



# Preparing model based only on RandDSpend & MarketingSpend

ml\_RandDSpend\_MarketingSpend=smf.ols('Profit~RandDSpend+MarketingSpend',data = startup50).fit()

ml\_RandDSpend\_MarketingSpend.summary()



# Preparing model based only on Administration & MarketingSpend

ml\_Administration\_MarketingSpend=smf.ols('Profit~Administration+MarketingSpend',data = startup50).fit()

ml\_Administration\_MarketingSpend.summary()



# Preparing model based only on Administration & RandDSpend

ml\_Administration\_RandDSpend=smf.ols('Profit~Administration+RandDSpend',data = startup50).fit()

ml\_Administration\_RandDSpend.summary()



#Dropping variables "California", "Florida" and "Administration" since they are insignificant to the outcome

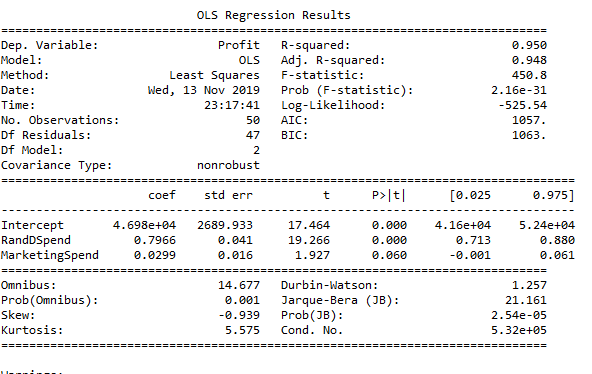
ml1\_v3 = smf.ols('Profit~RandDSpend+MarketingSpend',data = startup50).fit() # regression model

ml1\_v3.summary()

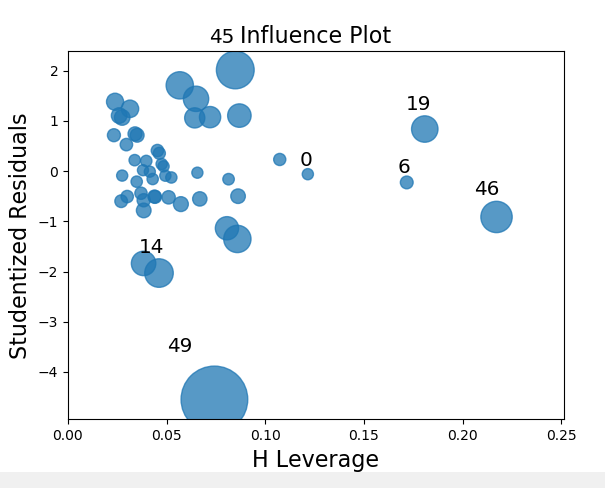


**The variables R&DSpend and MarketingSpend both has a high statistical significance in the outcome of the output variable “Profit” and hence there is an improvement in the overall R-squared and Adjusted R-squared value of the model**

Preparing final model with variables R&DSpend and Marketing Spend we have the following results:



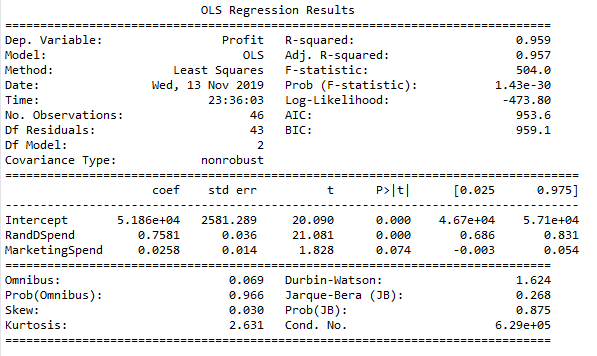
# influence index plots



The Observations 19,49, 46 and 45 are highly influential to the outcome of Profit

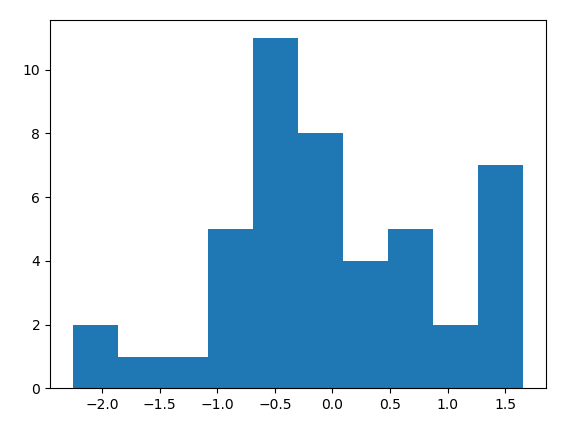
Building the final model after dropping the influential points.

We get the following results:

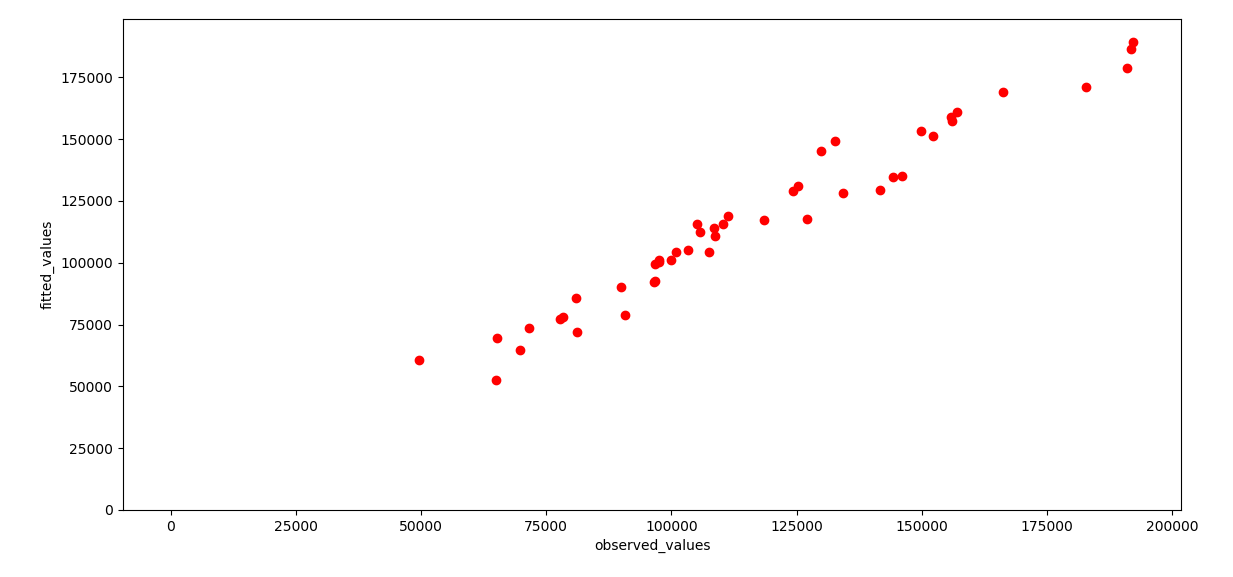


The model has high R and Adjusted R values with very good F-statistics score. The Durbin-Watson with the score 1.624 indicates that the auto correlation in the model is at permissible limit.

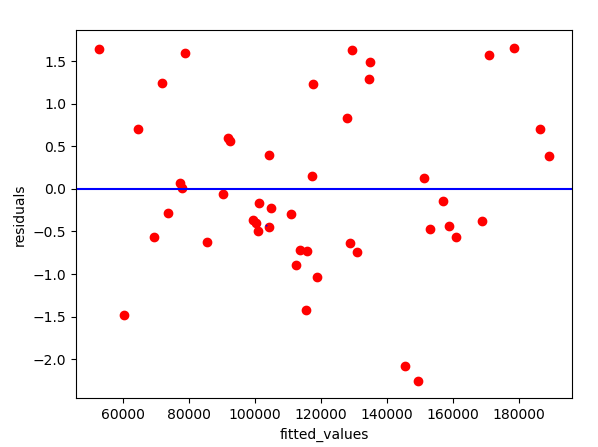
**Residual Analysis:**



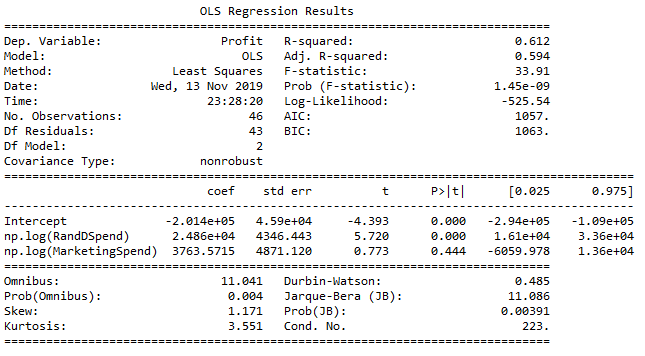
**Checking for Fitted Vs Observed Values**



Checking for Residuals Vs Fitted values



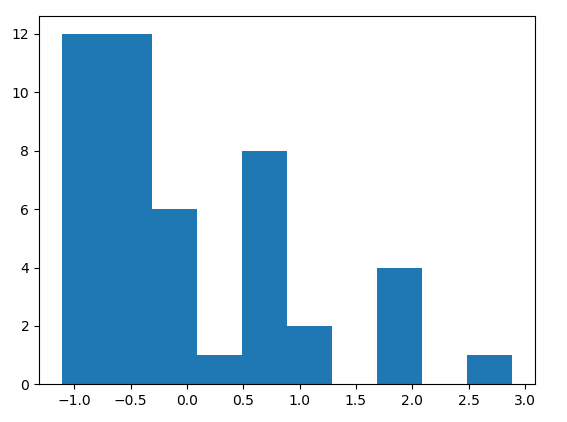
**Applying Log Transformation:**



* The model is deemed to be unfit since there the R and Adjusted R values are low and the variable “MarketingSpend” is insignificant to the outcome.

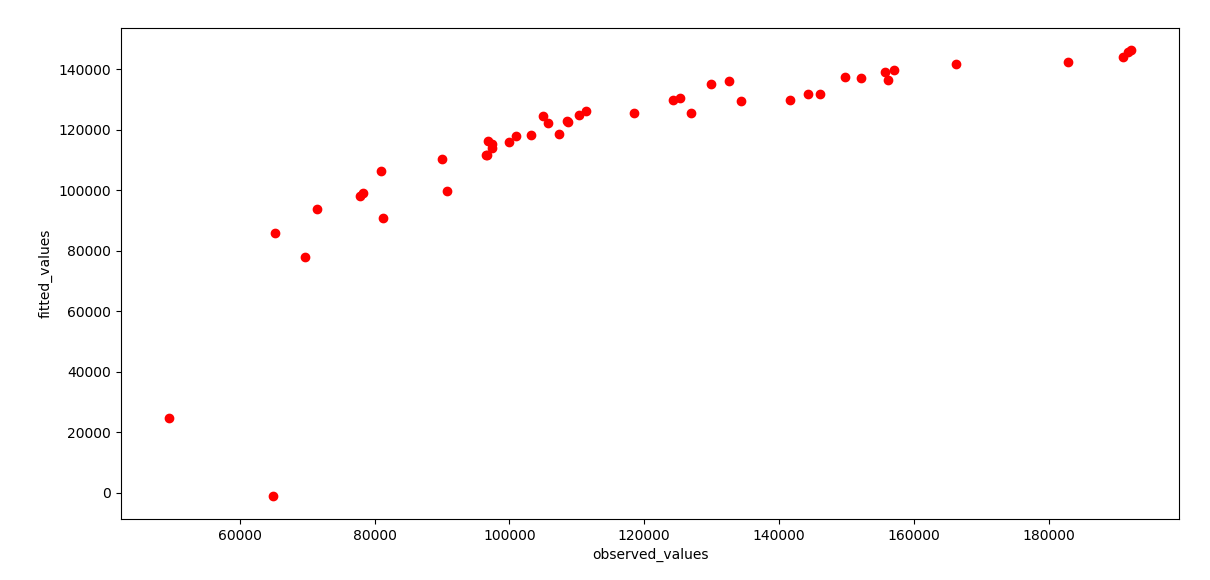
**Residual Analysis:**

Checking the normality of the residuals or errors:



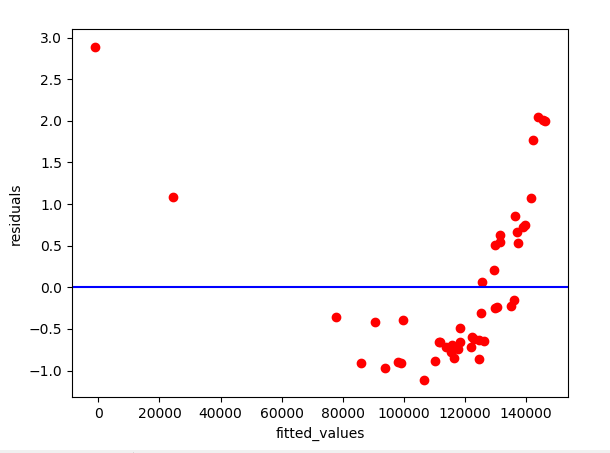
Heavily skewed toward right.

**Checking for Fitted Vs Observed Values**



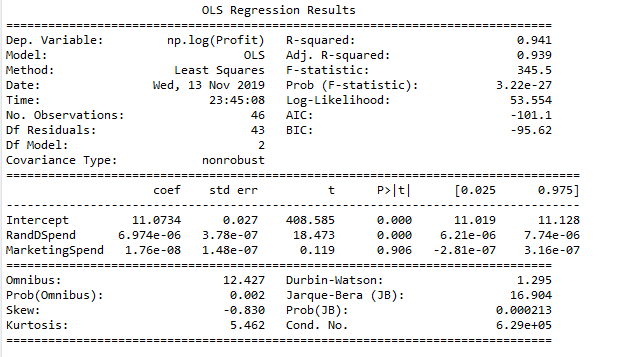
The above graph shows the non-conformance of the fitted and observed values

Checking for Residuals Vs Fitted values



The above graph shows there is unsimilar variances of the error from their means. This indicates that it is hetroscedastic in nature.

**We apply exponential transformation to the model:**



The variable in the above model “MarketingSpend” becomes insignificant.

We use the model without any transformations as the final model with variables R&D Spends and Marketing Spends as shown previously.