| <pre>import pandas as pd import seaborn as sn import matplotlib.pyplot as plt import numpy as np import statsmodels.formula.api as smf</pre> In [2]: #importing the data set   |      |
|--|------|
| dataset=pd.read_csv('C:\\Users\\DELL\\Downloads\\50_Startups.csv')         Out[2]:       R&D Spend Administration Marketing Spend State Profit         0       165349.20       136897.80       471784.10       New York       192261.83         1       162597.70       151377.59       443898.53       California       191792.06   |      |
| 2       153441.51       101145.55       407934.54       Florida       191050.39         3       144372.41       118671.85       383199.62       New York       182901.99         4       142107.34       91391.77       366168.42       Florida       166187.94         5       131876.90       99814.71       362861.36       New York       156991.12  |      |
| 6 134615.46 147198.87 127716.82 California 156122.51  7 130298.13 145530.06 323876.68 Florida 155752.60  8 120542.52 148718.95 311613.29 New York 152211.77  9 123334.88 108679.17 304981.62 California 149759.96  10 101913.08 110594.11 229160.95 Florida 146121.95  |      |
| 11       100671.96       91790.61       249744.55       California       144259.40         12       93863.75       127320.38       249839.44       Florida       141585.52         13       91992.39       135495.07       252664.93       California       134307.35         14       119943.24       156547.42       256512.92       Florida       132602.65   |      |
| 15       114523.61       122616.84       261776.23       New York       129917.04         16       78013.11       121597.55       264346.06       California       126992.93         17       94657.16       145077.58       282574.31       New York       125370.37         18       91749.16       114175.79       294919.57       Florida       124266.90  |      |
| 19       86419.70       153514.11       0.00       New York       122776.86         20       76253.86       113867.30       298664.47       California       118474.03         21       78389.47       153773.43       299737.29       New York       111313.02         22       73994.56       122782.75       303319.26       Florida       110352.25  |      |
| 23       67532.53       105751.03       304768.73       Florida 108733.99         24       77044.01       99281.34       140574.81       New York 108552.04         25       64664.71       139553.16       137962.62       California 107404.34         26       75328.87       144135.98       134050.07       Florida 105733.54         27       72107.60       127864.55       353183.81       New York 105008.31  |      |
| 28       66051.52       182645.56       118148.20       Florida       103282.38         29       65605.48       153032.06       107138.38       New York       101004.64         30       61994.48       115641.28       91131.24       Florida       99937.59         31       61136.38       152701.92       88218.23       New York       97483.56  |      |
| 32 63408.86 129219.61 46085.25 California 97427.84  33 55493.95 103057.49 214634.81 Florida 96778.92  34 46426.07 157693.92 210797.67 California 96712.80  35 46014.02 85047.44 205517.64 New York 96479.51  |      |
| 36 28663.76 127056.21 201126.82 Florida 90708.19  37 44069.95 51283.14 197029.42 California 89949.14  38 20229.59 65947.93 185265.10 New York 81229.06  39 38558.51 82982.09 174999.30 California 81005.76  40 28754.33 118546.05 172795.67 California 78239.91  |      |
| 41       27892.92       84710.77       164470.71       Florida       77798.83         42       23640.93       96189.63       148001.11       California       71498.49         43       15505.73       127382.30       35534.17       New York       69758.98         44       22177.74       154806.14       28334.72       California       65200.33   |      |
| 45         1000.23         124153.04         1903.93         New York         64926.08           46         1315.46         115816.21         297114.46         Florida         49490.75           47         0.00         135426.92         0.00         California         42559.73           48         542.05         51743.15         0.00         New York         35673.41  |      |
| 49 0.00 116983.80 45173.06 California 14681.40  In [3]: dataset.info() <class 'pandas.core.frame.dataframe'=""> RangeIndex: 50 entries, 0 to 49 Data columns (total 5 columns):</class>  |      |
| # Column Non-Null Count Dtype  O R&D Spend 50 non-null float64  Administration 50 non-null float64  Marketing Spend 50 non-null float64  State 50 non-null object  Profit 50 non-null float64  dtypes: float64(4), object(1)   |      |
| memory usage: 2.1+ KB  In [4]: dataset.head(10)  Out[4]: R&D Spend Administration Marketing Spend State Profit  O 165349.20 136897.80 471784.10 New York 192261.83   |      |
| 1       162597.70       151377.59       443898.53       California       191792.06         2       153441.51       101145.55       407934.54       Florida       191050.39         3       144372.41       118671.85       383199.62       New York       182901.99         4       142107.34       91391.77       366168.42       Florida       166187.94   |      |
| 5       131876.90       99814.71       362861.36       New York       156991.12         6       134615.46       147198.87       127716.82       California       156122.51         7       130298.13       145530.06       323876.68       Florida       155752.60         8       120542.52       148718.95       311613.29       New York       152211.77         9       123334.88       108679.17       304981.62       California       149759.96   |      |
| In [5]: #EDA on dataset  In [6]: #histogram sn.distplot(dataset['Profit'])   |      |
| C:\Users\DELL\anaconda3\lib\site-packages\seaborn\distributions.py:2551: FutureWarning: `distplot` is a deprecated function and will be removed in a future version. Please a your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).  warnings.warn(msg, FutureWarning)  out[6]: <axessubplot:xlabel='profit', ylabel="Density">  le-5  12</axessubplot:xlabel='profit',>   | dapt |
|  |      |
| 0.4<br>0.2<br>0.0<br>0.0<br>0.0<br>0.0<br>0.0<br>0.0<br>0.0<br>0.0<br>0.0  |      |
| -50000 0 50000 150000 250000 250000 Profit  In [7]: #pairplot  sn.pairplot(dataset)  Out[7]: <seaborn.axisgrid.pairgrid 0x209f3453070="" at=""></seaborn.axisgrid.pairgrid>  |      |
| 150000 - Pundy 100000 |      |
| 175000   |      |
| 150000<br>75000<br>50000   |      |
| 400000 - Big 200000  |      |
| 150000   |      |
| 50000  |      |
| b 50000 100000 150000 50000 100000 150000 b 200000 400000 50000 100000150000200000  R&D Spend Administration Marketing Spend Profit  #boxplot sn.boxplot(x="State", y="Profit", data=dataset)  Out[8]: <axessubplot:xlabel='state', ylabel="Profit"></axessubplot:xlabel='state',>   |      |
| 200000<br>175000 -<br>150000 -   |      |
| 75000 -<br>50000 -<br>25000 -  |      |
| New York California Florida State  In [9]: #barplot sn.barplot(x="State", y="Profit", data=dataset)  |      |
| Out[9]: <axessubplot:xlabel='state', ylabel="Profit">  140000</axessubplot:xlabel='state',>  |      |
| 80000 -<br>60000 -<br>40000 -<br>20000 -   |      |
| New York California State    10]: #heatmap to get positive and negative relationship between all dataset sn.heatmap(dataset.corr())  |      |
| AxesSubplot:>  R&D Spend   |      |
| Administration 0.6  Marketing Spend 0.4  - 0.2   |      |
| R&D Spend - Administration - Profit - OO O  |      |
| #building a model   #rename the column   df=pd.DataFrame(dataset)  |      |
| df  ut[12]:  |      |
| 2       153441.51       101145.55       407934.54       Florida       191050.39         3       144372.41       118671.85       383199.62       New York       182901.99         4       142107.34       91391.77       366168.42       Florida       166187.94         5       131876.90       99814.71       362861.36       New York       156991.12         6       134615.46       147198.87       127716.82       California       156122.51   |      |
| 7 130298.13 145530.06 323876.68 Florida 155752.60 8 120542.52 148718.95 311613.29 New York 152211.77 9 123334.88 108679.17 304981.62 California 149759.96 10 101913.08 110594.11 229160.95 Florida 146121.95   |      |
| 11       100671.96       91790.61       249744.55       California       144259.40         12       93863.75       127320.38       249839.44       Florida       141585.52         13       91992.39       135495.07       252664.93       California       134307.35         14       119943.24       156547.42       256512.92       Florida       132602.65   |      |
| 15       114523.61       122616.84       261776.23       New York       129917.04         16       78013.11       121597.55       264346.06       California       126992.93         17       94657.16       145077.58       282574.31       New York       125370.37         18       91749.16       114175.79       294919.57       Florida       124266.90         19       86419.70       153514.11       0.00       New York       122776.86  |      |
| 20       76253.86       113867.30       298664.47       California       118474.03         21       78389.47       153773.43       299737.29       New York       111313.02         22       73994.56       122782.75       303319.26       Florida       110352.25         23       67532.53       105751.03       304768.73       Florida       108733.99  |      |
| 24       77044.01       99281.34       140574.81       New York       108552.04         25       64664.71       139553.16       137962.62       California       107404.34         26       75328.87       144135.98       134050.07       Florida       105733.54         27       72107.60       127864.55       353183.81       New York       105008.31  |      |
| 28 66051.52 182645.56 118148.20 Florida 103282.38  29 65605.48 153032.06 107138.38 New York 101004.64  30 61994.48 115641.28 91131.24 Florida 99937.59  31 61136.38 152701.92 88218.23 New York 97483.56  32 63408.86 129219.61 46085.25 California 97427.84   |      |
| 33 55493.95 103057.49 214634.81 Florida 96778.92 34 46426.07 157693.92 210797.67 California 96712.80 35 46014.02 85047.44 205517.64 New York 96479.51 36 28663.76 127056.21 201126.82 Florida 90708.19   |      |
| 37 44069.95 51283.14 197029.42 California 89949.14 38 20229.59 65947.93 185265.10 New York 81229.06 39 38558.51 82982.09 174999.30 California 81005.76 40 28754.33 118546.05 172795.67 California 78239.91 41 27892.92 84710.77 164470.71 Florida 77798.83   |      |
| 41       27892.92       84710.77       164470.71       Florida       77798.83         42       23640.93       96189.63       148001.11       California       71498.49         43       15505.73       127382.30       35534.17       New York       69758.98         44       22177.74       154806.14       28334.72       California       65200.33         45       1000.23       124153.04       1903.93       New York       64926.08  |      |
| 46       1315.46       115816.21       297114.46       Florida       49490.75         47       0.00       135426.92       0.00       California       42559.73         48       542.05       51743.15       0.00       New York       35673.41         49       0.00       116983.80       45173.06       California       14681.40  |      |
| df.rename(columns={'R&D Spend': 'RnD_Spend', 'Marketing Spend': 'MarketingSpend'}, inplace=True)  ut[13]: <bound dataframe.rename="" method="" of<="" td=""><td></td></bound>  |      |
| 2 153441.51 101145.55 407934.54 Florida 191050.39<br>3 144372.41 118671.85 383199.62 New York 182901.99<br>4 142107.34 91391.77 366168.42 Florida 166187.94<br>5 131876.90 99814.71 362861.36 New York 156991.12<br>6 134615.46 147198.87 127716.82 California 156122.51<br>7 130298.13 145530.06 323876.68 Florida 155752.60<br>8 120542.52 148718.95 311613.29 New York 152211.77<br>9 123334.88 108679.17 304981.62 California 149759.96  |      |
| 10 101913.08   |      |
| 17 94657.16 145077.58 282574.31 New York 125370.37 18 91749.16 114175.79 294919.57 Florida 124266.90 19 86419.70 153514.11 0.00 New York 122776.86 20 76253.86 113867.30 298664.47 California 118474.03 21 78389.47 153773.43 299737.29 New York 111313.02 22 73994.56 122782.75 303319.26 Florida 110352.25 23 67532.53 105751.03 304768.73 Florida 108733.99 24 77044.01 99281.34 140574.81 New York 108552.04   |      |
| 25 64664.71 139553.16 137962.62 California 107404.34<br>26 75328.87 144135.98 134050.07 Florida 105733.54<br>27 72107.60 127864.55 353183.81 New York 105008.31<br>28 66051.52 182645.56 118148.20 Florida 103282.38<br>29 65605.48 153032.06 107138.38 New York 101004.64<br>30 61994.48 115641.28 91131.24 Florida 99937.59<br>31 61136.38 152701.92 88218.23 New York 97483.56  |      |
| 32 63408.86 129219.61 46085.25 California 97427.84 33 55493.95 103057.49 214634.81 Florida 96778.92 34 46426.07 157693.92 210797.67 California 96712.80 35 46014.02 85047.44 205517.64 New York 96479.51 36 28663.76 127056.21 201126.82 Florida 90708.19 37 44069.95 51283.14 197029.42 California 89949.14 38 20229.59 65947.93 185265.10 New York 81229.06 39 38558.51 82982.09 174999.30 California 81005.76   |      |
| 40 28754.33 118546.05 172795.67 California 78239.91<br>41 27892.92 84710.77 164470.71 Florida 77798.83<br>42 23640.93 96189.63 148001.11 California 71498.49<br>43 15505.73 127382.30 35534.17 New York 69758.98<br>44 22177.74 154806.14 28334.72 California 65200.33<br>45 1000.23 124153.04 1903.93 New York 64926.08<br>46 1315.46 115816.21 297114.46 Florida 49490.75<br>47 0.00 135426.92 0.00 California 42559.73  |      |
| 47 0.00 135426.92 0.00 California 42559.73 48 542.05 51743.15 0.00 New York 35673.41 49 0.00 116983.80 45173.06 California 14681.40>  model=smf.ols("Profit~RnD_Spend+Administration+MarketingSpend+State", data=dataset).fit()  #coefficient model.params   |      |
| ut[16]: Intercept 50125.343832<br>State[T.Florida] 198.788793<br>State[T.New York] -41.887019<br>RnD_Spend 0.806023<br>Administration -0.027004<br>MarketingSpend 0.026980   |      |
| <pre>dtype: float64  n [17]: #t and p values model.tvalues, '\n', model.pvalues  ut[17]: (Intercept</pre>  |      |
| RnD_Spend 17.368580 Administration -0.517012 MarketingSpend 1.573889 dtype: float64, '\n', Intercept 4.444178e-09 State[T.Florida] 9.532429e-01  |      |
| State[T.New York] 9.897941e-01 RnD_Spend 2.578772e-21 Administration 6.077373e-01 MarketingSpend 1.226769e-01 dtype: float64)  # r squared and r squared adjecent values model.rsquared, model.rsquared_adj  |      |
| ut[18]: (0.9507524843355148, 0.945156175737278)  n [21]: model.summary()  ut[21]: OLS Regression Results   |      |
| Dep. Variable:         Profit         R-squared:         0.951           Model:         OLS         Adj. R-squared:         0.945           Method:         Least Squares         F-statistic:         169.9           Date:         Sat, 17 Apr 2021         Prob (F-statistic):         1.34e-27           Time:         01:39:55         Log-Likelihood:         -525.38  |      |
| Time: 01:39:55 Log-Likelihood: -525.38  No. Observations: 50 AIC: 1063.  Df Residuals: 44 BIC: 1074.  Df Model: 5  Covariance Type: nonrobust  |      |
| 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           State[T.Florida]         198.7888         3371.007         0.059         0.953         -6595.030         6992.607           State[T.New York]         -41.8870         3256.039         -0.013         0.990         -6604.003         6520.229  |      |
| State[T.New York]       -41.8870       3256.039       -0.013       0.990       -6604.003       6520.229         RnD_Spend       0.8060       0.046       17.369       0.000       0.712       0.900         Administration       -0.0270       0.052       -0.517       0.608       -0.132       0.078         MarketingSpend       0.0270       0.017       1.574       0.123       -0.008       0.062         Omnibus:       14.782       Durbin-Watson:       1.283   |      |
| Omnibus:       14.782       Durbin-Watson:       1.283         Prob(Omnibus):       0.001       Jarque-Bera (JB):       21.266         Skew:       -0.948       Prob(JB):       2.41e-05         Kurtosis:       5.572       Cond. No.       1.45e+06  |      |
| Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 1.45e+06. This might indicate that there are strong multicollinearity or other numerical problems.   |      |
| strong multicollinearity or other numerical problems.  In []:  |      |
|  |      |
|  |      |
|  |      |
|  |      |

In [1]: #importing the libraries