PROFIT ESTIMATION AND COST ANALYSIS FOR STARTUPS
BAN 210 – GROUP 9-PREDICTIVE ANALYTICS

Marketing	Administration	R&D	California	Florida	New York	Actual profit	Predicted profit	R	R2
182,645.56	66,051.52	118,148.20	1	0	0	103,282.38	103,015.20	267.18	71,384.30
91,790.61	100,671.96	249,744.55	0	0	1	144,259.40	132,582.28	11,677.12	136,355,187.36
110,594.11	101,913.08	229,160.95	1	0	0	146,121.95	132,447.74	13,674.21	186,984,061.46
84,710.77	27,892.92	164,470.71	1	0	0	77,798.83	71,976.10	5,822.73	33,904,201.97
101,145.55	153,441.51	407,934.54	1	0	0	191,050.39	178,537.48	12,512.91	156,572,861.34
127,864.55	72,107.60	353,183.81	0	1	0	105,008.31	116,161.24	-11,152.93	124,387,898.93
65,947.93	20,229.59	185,265.10	0	1	0	81,229.06	67,851.69	13,377.37	178,953,972.01
152,701.92	61,136.38	88,218.23	0	1	0	97,483.56	98,791.73	-1,308.17	1,711,318.55
122,782.75	73,994.56	303,319.26	1	0	0	110,352.25	113,969.44	-3,617.19	13,084,029.71
91,391.77	142,107.34	366,168.42	1	0	0	166,187.94	167,921.07	-1,733.13	3,003,724.68
								Sum	**835,028,640.31**

Table 1 :- Demonstrating the profit predicted compared against the actual profit with residuals calculated. The categorical data(state) is converted to binary temporary variables to fit the regression model. It can be observed from the above table that predicted profit is very similar to the actual profit. R-square score of actual test results and predicted test results is 0.935 which means that 93.5% of the observed variation can be explained by the model.

Actual profit	Predicted profit	Variance from actuals
103,282.38	103,015.20	0.3%
144,259.40	132,582.28	8.1%
146,121.95	132,447.74	9.4%
77,798.83	71,976.10	7.5%
191,050.39	178,537.48	6.5%
105,008.31	116,161.24	-10.6%
81,229.06	67,851.69	16.5%
97,483.56	98,791.73	-1.3%
110,352.25	113,969.44	-3.3%
166,187.94	167,921.07	-1.0%

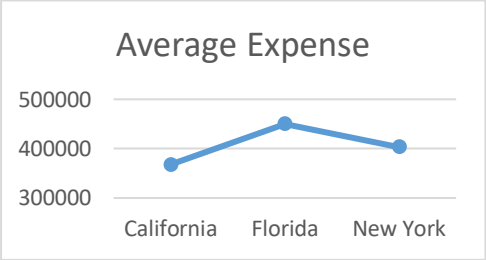
Table 2:- We can conclude that the model can be used further, based on the above table where the variance from actuals are considerably insignificant.

	Administration	R&D Spend	Marketing Spend	Profit
Administration	1	0.241955	-0.03215	0.200717
R&D Spend	0.241955	1	0.724248	0.9729
Marketing Spend	-0.03215	0.724248	1	0.747766
Profit	0.200717	0.9729	0.747766	1

Table 3: - demonstrates that correlation between the variables.

Key Findings

1. The prediction model was best fit and did not under or over fit as per the error statistics explained below.
2. As seen on **table 3**, Profit is highly positively correlated to R&D and marketing spend.
3. Administration cost has a weak correlation to profit
4. Interestingly, we can see that R&D cost and Marketing cost have a high correlation, indicating that if one cost increases, the other increases and vice versa. This an area that requires attention and further analysis to improve profits.
5. Profit is the highest in Florida and the average costs are the lowest in California which is 22% lesser than Florida and 10% lesser than New York



Steps for Prediction

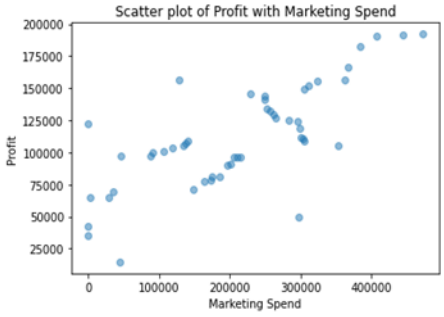
1. Import 'NumPy' and 'pandas' libraries
2. Import dataset – "50_Startups.csv" and display first 5 observations:

	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

Table 1 – Raw dataset showing variables

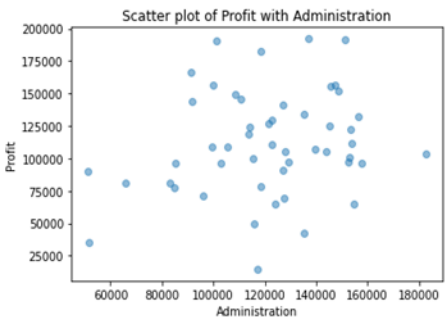
First 3 variables related to costs and 4th variable which defines location are identified as independent variables while 'Profit' is identified as dependent variable.

3. Plot scatter-plots to visualize possible correlations between independent and dependent variables:



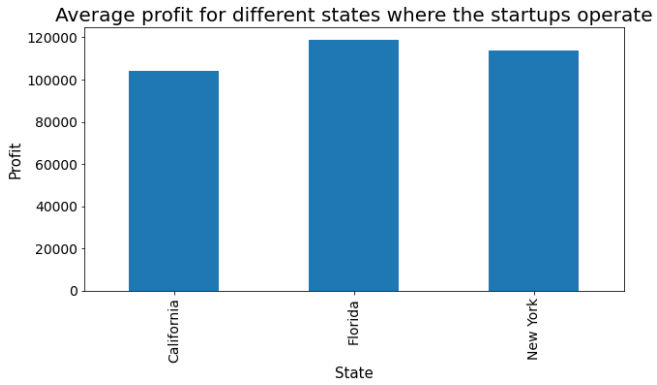
Strong positive correlation Between profit and marketing

Strong positive correlation between profit and R&D spend



No visible correlation between profit and Admin costs

4. Plot Bar Chart to illustrate profit per state:



Florida showed highest profits followed closely by New York and then California. Not a significant difference observed.

5. Convert categorical variable 'state' to numeric using 'One-hot encoding' technique:

	R&D Spend	Administration	Marketing Spend	Profit	NewYork_State	California_State	Florida_State
0	165349.20	136897.80	471784.10	192261.83	1	0	0
1	162597.70	151377.59	443898.53	191792.06	0	1	0
2	153441.51	101145.55	407934.54	191050.39	0	0	1
3	144372.41	118671.85	383199.62	182901.99	1	0	0
4	142107.34	91391.77	366168.42	166187.94	0	0	1

Table 2 – Using one-hot encoding to convert categorical to numeric for analysis

6. Create data of independent variables and dependent variable
7. Randomly split dataset into training data (80%) and test data (20%)
8. Execute multiple linear regression (due to presence of many variables) on the training set
9. Predict the test results using the multi regression model in the training dataset. Prediction results and code attached separately
10. Check for heteroscedasticity indicate a noticeable pattern to prove presence of heteroscedasticity. Therefore, assumption of linear regression is valid, and the regression model contains sufficient predictor variables to explain the performance of the dependent variable.
11. Check for under-fitting or over-fitting of the model:

Statistics	Training	Test	Variance
MSE	81,571,001.80	83,502,864.03	2%
RMSE	9,031.67	9,137.99	1%
MAE	6,341.54	7,514.29	18%

After calculating the Mean square error (mean of the difference of errors) and RMSE (Root) and MAE (Mean average error). We can conclude that the model is the best fit since the variance is not much between the test and the training datasets.