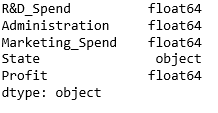
**NAME: WAIRIUKO SAMUEL**

**TITLE: PREDICTION USING LINEAR REGRESSION MODEL**

**1.0 BACKGROUND**

This project is about building a prediction model to predict profits. I downloded the dataset from [www.kaggle.com](http://www.kaggle.com). The dataset had 5 variables and 1000 observations. The 5 variables include the target variable Profits, other variables include R&D Spend, Administration, State and Marketing spend. During the data cleaning process, I dropped one attribute(state)

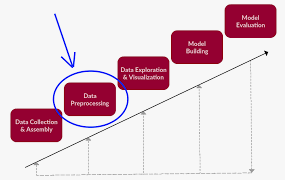
**PROPERTIES OF THE DATASET**



**OBJECTIVES**

* To build an accurate and precise model for predicting profits.
* To analyze the dataset and give conclusion and insights to the management.
* To check the accuracy of the model.

**2.0 DATA PREPARATION, CLEANING & ANALYSIS**

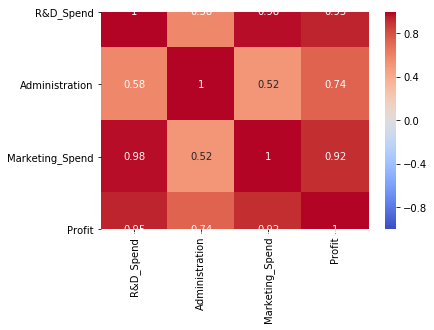


* Checked and removed duplicates
* Standardized the dataset
* Dropped irrelevant column (State).

**3.0 DATA ANALYSIS & SUMMARY FINDINGS**

|  | R&D d | Administration | Marketing Spend | Profit |
| --- | --- | --- | --- | --- |
| R&D | 1.0 | 0.58 | 0.98 | 0.95 |
| Administration | 0.58 | 1.0 | 0.52 | 0.74 |
| Marketing | 0.98 | 0.52 | 1.0 | 0.92 |
| Profit | 0.95 | 0.74 | 0.92 | 1.0 |
|  |  |  |  |  |

There is high correlation between R& D and marketing spend.



**3.2 R2, ADJUSTED R2, RMSE**

* R2 test score 0.85
* Adjusted R2 test score 0.80
* K fold score was 0.86
* RMSE score from model was $20230 compared to $37934 from Cross Validation.

**4.0. CONCLUSIONS AND RECOMMENDATIONS**

Comparing the R2 for K fold and test score, shows no significance difference, the final test score was 0.85.

The model is good for prediction but not the best.

It’s important to try other different models that can improve the test score.

**. 5.REFFERENCES**

-www.kaggle.com {open data set}

-Python Project Codes(attached)