**Exploratory Data Analysis**

**COSC 5010 – 03**

**Isaac Baah**

## 1.0 Introduction

To be able to build a robust machine learning model, it is important that we reduce the noise found in data. One of those ways is through exploratory data analysis. The main goal of this exploratory analysis exercise is to look at the raw data to see any trend before performing any machine learning algorithms on the data. The exploratory analysis will make sure we are familiar with the data and then see any potential problems hidden within the dataset. For example, if there are a lot of missing values in a particular variable, we may decide to drop that variable. Again, if there is any multi-collinearity in the data, we may decide to drop one of the variables.

## 2.0 Data

The data used is from a machine learning tutorial I did on Udemy, the data consists of five variables. Four of the variables are numeric and one of the variables is categorical. The dependent variable is the "Profit". The data is trying to predict the profit made by startups based on their expenditure in the following area: "Research & Development", "Administration", and "Marketing". Regarding missing values, the data does not include any missing values.

| R&D Spend | Administration | Marketing Spend | State | Profit |
|---|---|---|---|---|
| 165349.2 | 136897.8 | 471784.1 | New York | 192261.8 |
| 162597.7 | 151377.59 | 443898.53 | California | 191792.1 |
| 153441.51 | 101145.55 | 407934.54 | Florida | 191050.4 |
| 144372.41 | 118671.85 | 383199.62 | New York | 182902.0 |
| 142107.34 | 91391.77 | 366168.42 | Florida | 166187.9 |
| 131876.9 | 99814.71 | 362861.36 | New York | 156991.1 |
| 134615.46 | 147198.87 | 127716.82 | California | 156122.5 |
| 130298.13 | 145530.06 | 323876.68 | Florida | 155752.6 |
| 120542.52 | 148718.95 | 311613.29 | New York | 152211.8 |
| 123334.88 | 108679.17 | 304981.62 | California | 149760.0 |

Figure 1. A snapshot of the data used in the project.

## 3.0 Results of the Analysis

In this section, we will look at the results obtained from our exploratory analysis which was done in python. The first thing done was to look at the data using ".head()". A snapshot is shown in Figure 1 above. Then move on to describe the data as shown in Figure 2. The description of the data gives the values when plotting using a boxplot.

| | R&D Spend | Administration | Marketing Spend | Profit |
|---|---|---|---|---|
| **count** | 50.000000 | 50.000000 | 50.000000 | 50.000000 |
| **mean** | 73721.615600 | 121344.639600 | 211025.097800 | 112012.639200 |
| **std** | 45902.256482 | 28017.802755 | 122290.310726 | 40306.180338 |
| **min** | 0.000000 | 51283.140000 | 0.000000 | 14681.400000 |
| **25%** | 39936.370000 | 103730.875000 | 129300.132500 | 90138.902500 |
| **50%** | 73051.080000 | 122699.795000 | 212716.240000 | 107978.190000 |
| **75%** | 101602.800000 | 144842.180000 | 299469.085000 | 139765.977500 |
| **max** | 165349.200000 | 182645.560000 | 471784.100000 | 192261.830000 |

Figure 2. A summary of the data.

The next task was to plot the profit according to the state in which the company is found. Figure 2 below shows the profit made from all the startup companies in a particular state. From the Figure 3 below, the data revealed that all the startup companies almost have the same profit based on their states.

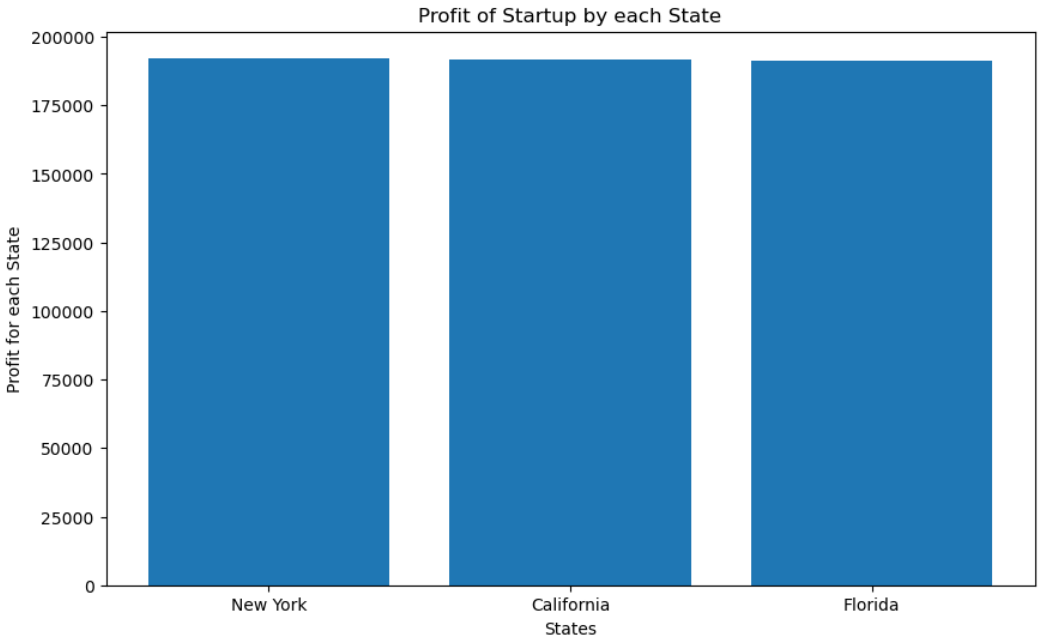


Figure 3. Profits made by Startups in their states.

The distributions of both the dependent and the independent variables were then looked at. When we are fitting a parametric algorithm like regression analysis on a data, we need to obey the assumptions of the linear regression model. We can do some transformations on the data, for example, the distribution of the “R&D Spend” does not look normal, hence we can do some transformation on the variable to make the distribution normal. This will ensure that when we are fitting a linear regression, we will have robust performance output.

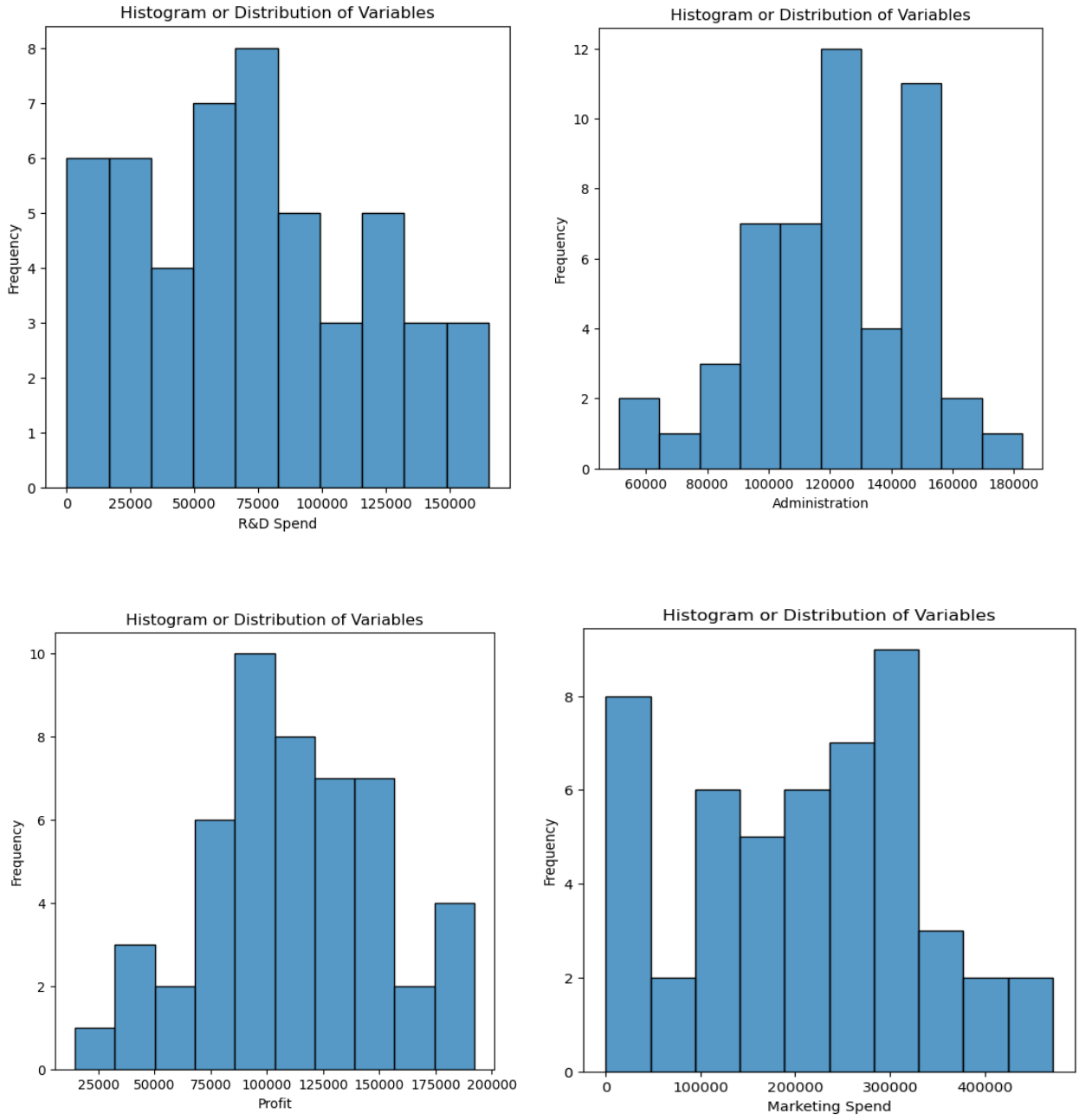


Figure 4. Distribution or Histograms of the Variables

Correlation between the independent variables or predictors and the dependent variable was also assessed. The correlation plot showed that "R&D Spend" and "Marketing Spend" have a high correlation with the dependent variable "Profit". By using only these two independent variables in our model, we are likely going to have a good performance compared to when are three independent variables are used.

Multi-collinearity between the independent variables were also assessed. The correlation plot also revealed a strong relationship between "R&D Spend" and "Marketing Spend". Hence we can drop one of the two variables and only fit the machine learning with only one predictor.

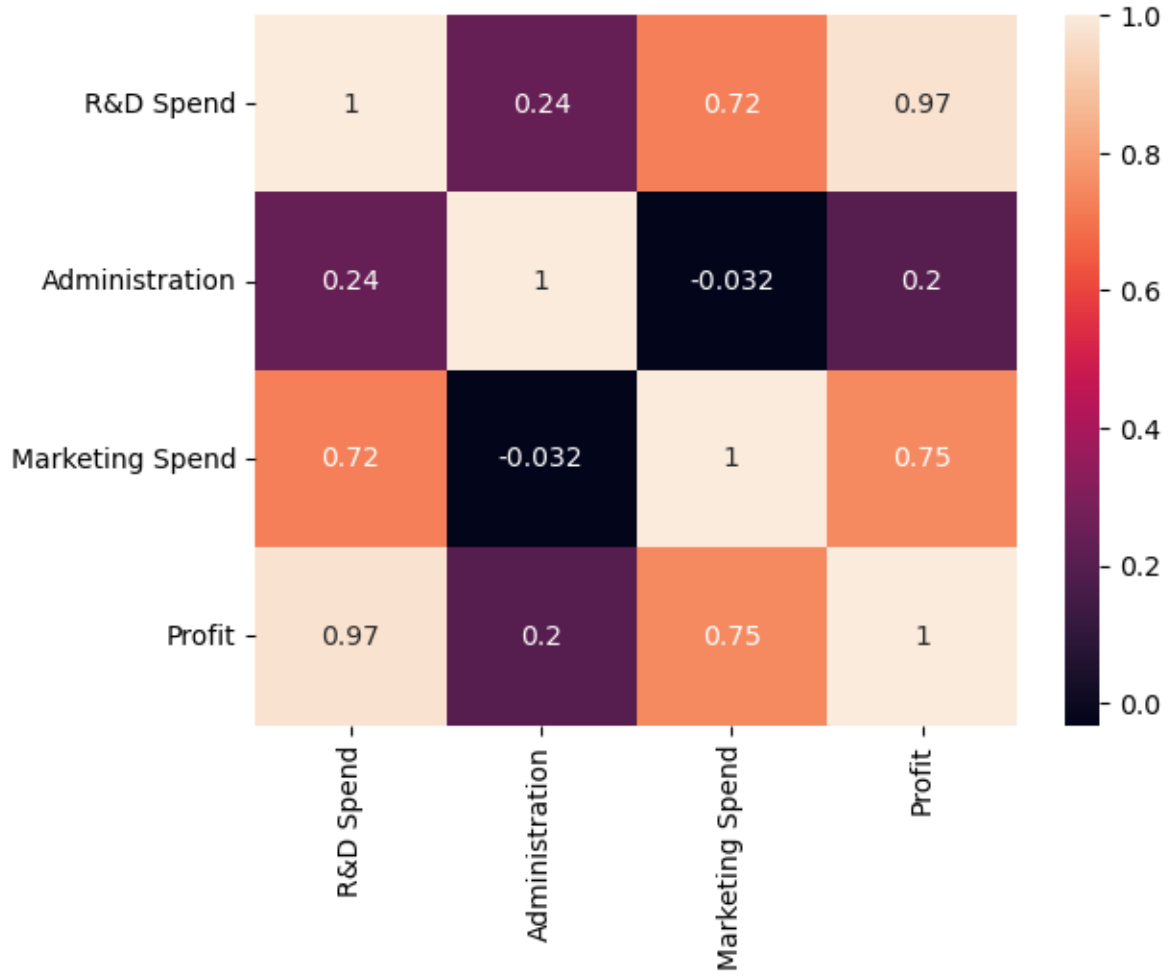


Figure 5. Correlation of the features

**4.0 References**

1. Hands-On Machine Learning with Python and R. https://www.udemy.com/course/machinelearning/. [Accessed, October 2023].