### **Linear Regression and Multiple Linear Regression Analysis**

### 

### **1. Business Problem**

The goal of this assignment is to analyze how different company expenditures impact profit. Specifically, we aim to assess if increasing spending in areas like R&D, Administration, and Marketing drives profitability, and to what extent each category contributes to overall profit.

### **2. Data Dictionary**

## **Dataset Link:** [**50\_Startups**](https://drive.google.com/file/d/13fdu7lupRF5EaVOlnlDuJ5VNDLkKiYTu/view?usp=drive_link)

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **DataType** | **Description** |
| **R\_D\_spent** | Float | Amount spent on Research and Development |
| **Administration** | Float | Administrative expenses |
| **Marketing\_Spend** | Float | Marketing expenditures |
| **State** | Categorical | Location of company operations |
| **Profit** | Float | Profit generated by the company |

### **3. Data Pre-Processing**

#### **3.1 Data Inspection and Summary Statistics**

#### **Load the Dataset**: Import the dataset and review its basic structure, including column names, data types, and a few initial records.

#### **Generate Summary Statistics:** Calculate key statistics (mean, median, min, max, standard deviation, etc.) to understand the primary characteristics of each column.

#### **3.2 Data Cleaning and Feature Engineering**

**Missing Values**: Verify and handle any missing or inconsistent values.

#### **3.3 Outlier Treatment**

**Outlier Detection**: Identify outliers in R\_D\_spent, Administration, and Marketing\_Spend using box plots and Z-scores and If possible handle outliers.

### **4. Exploratory Data Analysis (EDA)**

#### **4.1 Summary Statistics**

**Summary Table**: Calculate the mean, median, variance, and standard deviation for each numerical feature to understand the data’s distribution.

#### **4.2 Univariate Analysis**

**Numerical Data**: Visualize distributions with histograms and box plots to check for skewness and outliers.

**Categorical Data**: Use bar charts to observe the distribution of categorical data.

#### **4.3 Bivariate Analysis**

**Numerical-Numerical Relationships**: Create scatter plots to observe relationships between pairs of numerical variables.

**Categorical-Numerical Relationships**: Use box and violin plots to analyze how categorical variables impact numerical distributions.

**Correlation Analysis**: Generate a heatmap of the correlation matrix to identify potential relationships.

### **5. Model Building**

#### **5.1 Feature Engineering and Scaled Data**

**Encoding Categorical Variables**: Convert the "State" variable into dummy variables to incorporate it into the regression model.

**Data Scaling**: Apply standardization to ensure all features are on a similar scale.

**Model Selection**: Start with simple linear regression, then progress to multiple linear regression to capture more insights.

#### **5.2 Multiple Linear Regression Model**

**Multicollinearity Check**: Assess multicollinearity by calculating the Variance Inflation Factor (VIF) for each feature.

**Influence Analysis**: Use AvPlots and Influence Index Plots to identify any influential data points that could affect model accuracy.

#### **5.3 Train-Test Split**

**Data Split**: Divide data into training and testing sets.

**Performance Metrics**: Calculate R-squared and RMSE for both training and testing sets to evaluate the model’s performance.

#### **5.4 Model Interpretation and Tuning**

**Interpretation**: Examine model coefficients to understand the impact of each feature on profit.

**Model Tuning**: Explore advanced techniques such as polynomial and interaction terms, and consider using Ridge or Lasso regularization to improve model performance.