**Project Report: Predicting Profit**

**Objective:**

The primary objective of this project is to predict the profit of an organization based on the money they spend on R&D, Administration, Marketing, and the city in which they operate.

**Methodology:**

**Data Preparation:**

* Import necessary packages in Jupyter Notebook.
* Extract and load the dataset and copy the data.

**Data Preprocessing:**

* Identify and handle incorrect 0 values in the dataset.
* Impute the incorrect data using the median of the respective columns.

**Exploratory Data Analysis (EDA):**

* Analyze the data to gain insights.
* Use box plots to identify outliers in the "State" variable.
* For the New York City, identify 3 outliers.

**Outlier Imputation:**

* Filter the data for New York City and impute the outliers using upper and lower threshold values.
* Merge the cleaned data with the original copied dataset.

**Data Visualization:**

**Create scatter plots for:**

* Administration vs. Profit: The correlation is 0.1884, indicating a weak relationship.
* Marketing Spend vs. Profit: The correlation is 0.6907, showing a stronger relationship.
* R&D Spend vs. Profit: The correlation is 0.8819, indicating a strong relationship.

**Categorical Variable Analysis:**

* Analyze the impact of the "State" variable on profit using one-way ANOVA.
* F-Statistic: 0.5989
* P-value: 0.5535
* The analysis suggests that "State" is not a significant factor in profit variation.

**Feature Selection:**

* Choose R&D Spend and Marketing Spend as independent variables for predicting profit.

**Model Building:**

* Create a multiple linear regression model using Scikit-Learn.
* Split the data into training and testing sets with an 80:20 ratio.
* Train the model on the training data and test it on the testing data.

**Model Evaluation:**

* Assess the model's performance using the R-squared (r2\_score).
* Achieve an r2\_score of approximately 90.35%, indicating a strong model fit.

**Conclusion:**

This project aimed to predict profit based on R&D Spend and Marketing spending, as these variables showed significant correlations with profit. The analysis revealed that the "State" variable had a limited impact on profit. The multiple linear regression model yielded a high R-squared value, indicating that it can effectively predict profit.

The findings from this project can help businesses make data-driven decisions and allocate resources efficiently to maximize profit.