**Multiple Linear Regression on 50\_Startups Dataset**

**1. Introduction**

This project aims to build a multiple linear regression model to predict the profit of startups based on various expenditure features. The objective is to identify the most significant factors influencing profit and optimize the prediction accuracy.

**2. Problem Statement**

Predict the profit of startups using features like R&D Spend, Administration Spend, Marketing Spend, and State.

**3. Dataset Description**

* **R&D Spend:** Research and development expenses (numeric)
* **Administration:** Administrative expenses (numeric)
* **Marketing Spend:** Marketing expenses (numeric)
* **State:** The location/state of the startup (categorical)
* **Profit:** The target variable (numeric)

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

* Checked for missing values and data types.
* Visualized distributions of variables.
* Analyzed correlations among numeric features.
* Investigated the impact of categorical variable (State) on Profit.

**5. Data Preprocessing and Transformations**

* Converted categorical variable “State” into dummy/indicator variables.
* Performed necessary feature scaling if required.
* Checked for multicollinearity and outliers.
* Applied transformations to variables where needed to improve linearity and model performance (e.g., log transformations).

**6. Model Building and Selection**

* Built an initial multiple linear regression model using all predictors.
* Evaluated model performance with R² value.
* Iteratively refined the model by removing insignificant variables and applying transformations.
* Selected the best model based on highest R² and adjusted R² values.

**Table: R² Values of Models Prepared**

| **Model Description** | **R² Value** |
| --- | --- |
| Initial Model (all variables) | 0.950 |
| Model after removing insignificant variables | 0.958 |
| Model with transformations applied | 0.965 |

*(Note: The above values are illustrative; actual values depend on the dataset analysis.)*

**7. Solution Architecture**

Data Collection--> Data Cleaning & Preprocessing--> EDA & Feature Engineering --> Model Building

--> Model Evaluation --> Prediction

* Data is collected and loaded into the environment.
* Cleaned and preprocessed for missing values and categorical encoding.
* Explored through EDA to understand data characteristics.
* Built multiple regression models to predict profit.
* Evaluated and selected the best performing model.

**8. Methodology**

1. **Data Understanding:** Load and inspect dataset.
2. **EDA:** Visualize data, check distributions and correlations.
3. **Preprocessing:** Encode categorical variables, handle missing data.
4. **Feature Engineering:** Create new variables or transform existing ones if necessary.
5. **Model Training:** Use statsmodels or sklearn to build regression models.
6. **Model Evaluation:** Use R² and residual plots to check performance.
7. **Optimization:** Refine model by variable selection and transformation.
8. **Final Model Deployment:** Predict profit on test/new data.
9. **Time Taken**

| **Phase** | **Time Taken** |
| --- | --- |
| **Data Understanding & Loading** | **4 hours** |
| **Exploratory Data Analysis** | **6 hours** |
| **Data Preprocessing** | **4 hours** |
| **Model Building** | **5 hours** |
| **Evalation & Selection** | **3 hours** |
| **Documentation** | **2 hours** |

**Total Time: Approx. 2 days (24 hours)**