**Multiple Regression**

50\_Startups Dataset

**Objective**  
To predict a startup’s profit based on R&D spend, administration, marketing spend, and location using multiple linear regression.

**Dataset Information**

* Size: 50 rows × 5 columns
* Features:
  + R&D Spend
  + Administration
  + Marketing Spend
  + State (Categorical)
  + Profit (Target Variable)

**Solution Architecture**

**Stepwise Architecture:**

1. **Data Ingestion** → Read CSV using pandas
2. **EDA** → Seaborn/Matplotlib for visual exploration
3. **Preprocessing** → Missing values, encoding
4. **Transformation** → Scaling, power transforms
5. **Model Building** → LinearRegression from sklearn
6. **Evaluation** → R² score
7. **Interpretation** → statsmodels OLS summary
8. **Model Comparison** → R² table with transformations

**Methodology**

| **Step** | **Description** |
| --- | --- |
| Data Collection | Loaded from 50\_Startups.csv |
| EDA | Used seaborn pairplot, boxplots, and correlation heatmap |
| Preprocessing | Categorical encoding (State), feature-target split |
| Transformations | Raw, Standard Scaled, and Power Transformed |
| Modeling | Linear Regression using sklearn |
| Evaluation | Compared models using R² Score |
| Interpretation | OLS summary from statsmodels |
| **Time Taken**   | **Phase** | **Estimated Time** | | --- | --- | | Data Collection & Cleaning | 15 mins | | EDA | 30 mins | | Preprocessing | 15 mins | | Modeling | 20 mins | | Evaluation & Comparison | 10 mins | | Report Generation | 30 mins | | **Total Time** | **~2 hours** |   **Challenges Faced**   * Handling categorical columns in visualizations * Fixing ValueError: could not convert string to float * Ensuring correct order of operations: EDA before encoding * Avoiding object types in regression models   **Complexity**  Level: Beginner to Intermediate  This project involves standard linear regression and is not computationally intensive. However, selecting and transforming features requires a clear understanding of statistical assumptions and preprocessing.  **Insights / Business Interpretation**  R&D Spend is the most influential factor for predicting profit.  Marketing Spend shows a moderate correlation.  Administration has minimal effect.  State introduces minor variations but is less significant after encoding.  **R² Score Comparison Table**   | **Model** | **R² Score** | | --- | --- | | Raw | 0.950 | | Standard Scaled | 0.950 | | Power Transformed | 0.957 |   **Conclusion**  The Power Transformed model achieved the highest R² score, indicating the best performance. Future work can involve polynomial regression, interaction terms, or regularization methods to further enhance accuracy. |  |