**Assignment 2**

**Name: Ashutosh Shinde**

**Roll no: 282006**

**Batch: B1**

**Problem Statement**

Perform the following operations on a dataset:

1. **Summary Statistics:** Compute and display summary statistics for each feature (min, max, mean, range, std, variance, percentiles).
2. **Data Visualization:** Create histograms for each feature to illustrate distributions.
3. **Data Preprocessing:** Perform data cleaning, integration, and transformation.
4. **Model Building:** Build a classification model (e.g., Linear Regression).

**Objective**

1. To analyze and summarize dataset features.
2. To visualize feature distributions.
3. To preprocess data for modeling.
4. To build and evaluate a predictive model.

**Resources Used**

* **Software:** Google Colab
* **Libraries:**
  + Pandas (Data Manipulation)
  + Matplotlib (Visualization)
  + Scikit-learn (Model Building)

**Methodology**

**1. Data Loading & Exploration**

* Loaded the dataset (admission.csv).
* Displayed the first few rows to understand structure.

**2. Summary Statistics**

* Computed descriptive statistics (mean, std, min, max, percentiles).

**3. Data Visualization**

* Created histograms for key features (GRE Score, TOEFL Score, University Rating, SOP).

**4. Data Cleaning**

* Checked for missing values (none found).
* Filled missing values (hypothetical step, though none existed).

**5. Model Building (Linear Regression)**

* **Features:** GRE Score, TOEFL Score, University Rating, SOP, CGPA.
* **Target:** Research (binary: 0 or 1).
* Split data into training (75%) and testing (25%) sets.
* Trained a Linear Regression model.
* Evaluated using:
  + Mean Absolute Error (MAE)
  + Mean Squared Error (MSE)
  + R² Score

**Conclusion**

1. **Summary Statistics:** Provided insights into feature distributions.
2. **Visualization:** Confirmed normal distributions for most features.
3. **Data Cleaning:** No missing values were found.
4. **Model Performance:**
   * Linear Regression achieved an **R² score of 0.26**, indicating moderate predictive power.
   * Future work: Try classification models (Logistic Regression) for better accuracy.

This assignment demonstrated **data analysis, visualization, preprocessing, and modeling** using Python.