Mini Project: Data Preprocessing, EDA, and Simple Linear Regression

Files to Submit

1. **startups_clean.csv – cleaned dataset**
2. **EDA_Preprocessing.html – exported HTML version of your Jupyter Notebook**
3. **regression_analysis.R – your R script**
4. **Assignment_Presentation.pptx – presentation slides**
5. **Recorded presentation video – upload to Google Drive and share the link with editing access**

Part 1 – Python (Data Preprocessing & EDA)

Perform all steps inside a Jupyter Notebook and export it as HTML when finished.

Step 1: Import Dataset
Import the file Startup_Dataset.csv.

Step 2: Data Preprocessing
Perform the data-including the following:

- Handle missing values appropriately.
- Remove duplicate rows.
- Rename columns to consistent, clear names.
- Arrange columns in a logical order.
- Standardize text values (consistent capitalization).
- Encode categorical variables where needed.
- Standardize or scale numerical columns where relevant.

Step 3: Handle Missing Values
Follow these rules:

- If one expense (R&D_Spend, Administration, or Marketing_Spend) is missing and Total_Spend exists, calculate the missing expense.
- If Total_Spend and Profit_Margin are missing but all expenses are available, calculate both.
- For other missing values, apply a justified imputation or removal method.

Step 4: Visualization

- Visualize the distributions of all numeric variables (use histograms).
- Create a correlation heatmap to show relationships between numeric features.

Step 5: Dependent Variable

- Decide the most suitable dependent variable based on EDA and business logic.
- State your justification briefly in markdown.

Step 6: Export Cleaned Data

Save your cleaned dataset as startups_clean.csv.

Part 2 – R (Simple Linear Regression)

Perform your regression analysis using R.

Step 1: Import Data

Import startups_clean.csv.

Step 2: Build Model

- Use a simple linear regression model.
- The dependent variable should be the one you selected from EDA.
- Choose one independent variable that logically influences the dependent variable.

Step 3: Interpret Results

Interpret:

- Coefficients
- R^2
- p-values
- Business meaning

Step 4: Visualization

- Plot the regression line.
- Plot residuals and comment on the model fit.

Part 3 – Presentation (PPTX)

Prepare slides summarizing:

- Background of the dataset

- Issues identified and how they were resolved/justifications

- Key EDA results and findings

- Regression results (output summary and plots)

- Final remarks, conclusions, and business interpretation

Part 4 – Recorded Presentation

Record a short presentation summarizing your work:

- Duration: 10-20 minutes

- Include your screen, video and voice

- Every member needs to present

- Present the slides and key findings

Upload it to Google Drive and share the **editable link**, including it on the **last slide** of the PowerPoint presentation.

Grading Rubic

Criteria	Description	Marks
1. Data Preprocessing (Python)	Correct handling of missing values, duplicates, column naming, data types, encoding, and standardization. Evidence of logical steps and justification for choices.	**30**
2. EDA and Visualization	Appropriate use of visualizations (distributions, correlation heatmap). Insightful summary of data patterns and relationships. Clear choice and justification of dependent variable.	**20**
3. Regression Analysis (R)	Correct model syntax and execution. Interpretation of coefficients, R^2, significance, and business implications. Accurate and well-labeled visualizations (regression line, residuals).	**20**
4. Presentation (PPTX) & Recorded Presentation Video	Professional and logical structure. Covers background, identified issues, data cleaning steps, EDA findings, regression results, and business recommendations. Clear explanation, confident delivery, and logical flow.	**30**
Total		**100**