Global Pollution Analysis and Energy Recovery

Objective

The goal is to analyze global pollution data and develop strategies for pollution reduction and converting pollutants into energy. The dataset will be used for both **data preprocessing** and **building regression models** to predict energy recovery from pollution levels.

Phase 1: Data Collection and Exploratory Data Analysis (EDA)

Step 1 - Data Import and Preprocessing

1. **Datasets**
   Load the dataset (`Global_Pollution_Analysis.csv`).

2. **Handle Missing Values**
   Identify missing or inconsistent data, and handle them using appropriate imputation strategies.

3. **Data Transformation**

   - Normalize or scale pollution indices (air, water, and soil).
   - Encode categorical features such as `Country` and `Year` using label encoding or one-hot encoding.

Step 2 - Exploratory Data Analysis (EDA)

1. **Descriptive Statistics**
   Calculate descriptive statistics for numerical features like `CO2_Emissions` and `Industrial_Waste_in_tons`.

2. **Correlation Analysis**
   Visualize the correlation between pollution levels and other features like energy consumption using a heatmap.

3. **Visualizations**
   Create bar charts, line plots, and box plots to explore trends in pollution over time and across countries.

Step 3 - Feature Engineering

1. **Yearly Trends**
   Extract year-based trends to understand how pollution and energy recovery have evolved over time.

2. **Energy Consumption per Capita**
   Calculate energy consumption per capita for better analysis.

Phase 2: Predictive Modeling

Step 4 - Linear Regression Model (for Pollution Prediction)

1. **Model Objective**
   Predict energy recovery (in GWh) based on pollution levels, industrial waste, and other features.

2. **Model Building**
   Train a **Linear Regression** model to predict energy recovery using features like `Air_Pollution_Index`, `CO2_Emissions`, and `Industrial_Waste_in_tons`.

3. **Evaluation Metrics**
   Use **R^2**, **Mean Squared Error (MSE)**, and **Mean Absolute Error (MAE)** to evaluate model performance.

Step 5 - Logistic Regression Model (for Categorization of Pollution Levels)

1. **Model Objective**
   Classify countries into pollution severity categories (Low, Medium, High).

2. **Model Implementation**
   Use **Logistic Regression** to classify pollution severity based on features like `Air_Pollution_Index` and `CO2_Emissions`.

3. **Evaluation Metrics**
   Evaluate using metrics like **Accuracy**, **Precision**, **Recall**, **F1-score**, and plot the **Confusion Matrix**.

Phase 3: Reporting and Insights

Step 6 - Model Evaluation and Comparison

- Compare **Linear Regression** and **Logistic Regression** models based on their performance and metrics.
- Visualize results using **confusion matrices** and **classification reports**.

Step 7 - Actionable Insights

- Provide insights on how different pollution levels affect energy recovery and suggest countries that could benefit from improvement.
- Offer recommendations for **reducing pollution** and **improving energy recovery**.

Final Deliverables

1. **Jupyter Notebook (.ipynb)** containing the entire code and analysis.
2. **Data Visualizations** in image format or embedded in the notebook.
3. **Final Report** summarizing key findings, model evaluations, and actionable recommendations.
