Assignment 2: Predicting Bike Sharing Demand Using Linear Regression

Dataset Overview:

Bike-sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people can rent a bike from one location and return it to a different place on an as-needed basis.

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed are explicitly recorded. In this assignment, you're asked to combine historical usage patterns with weather data to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Columns description:

Datetime: hourly date + timestamp

Season: 1 = spring, 2 = summer, 3 = fall, 4 = winter

Holiday - whether the day is considered a holiday

working day - whether the day is neither a weekend nor holiday

Weather:

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Temp - temperature in Celsius

Atemp - "feels like" temperature in Celsius

Humidity - relative humidity
Windspeed - wind speed

Casual - number of non-registered user rentals initiated

Registered - number of registered user rentals initiated

Count - number of total rentals

Tasks:

1. **Preprocessing the Data:**
    1. **Inspect the Dataset**:
        - Load and display the dataset.
        - Check for missing values and duplicates.
    2. **Handle Missing Data**:
        - Address any missing values appropriately.
    3. **Encode Categorical Variables**:
        - Convert categorical variables into a numerical format using encoding techniques (e.g., one-hot encoding).

2. **Exploratory Data Analysis (EDA):**

    1. **Analyze the Target Variable**
    2. **Analyze Relationships**:
        - Use scatter plots or bar charts to analyze relationships between features and the target variable.
    3. **Correlation Analysis**:
        - Create a correlation heatmap to identify features strongly correlated with the target variable.

3. **Feature Engineering:**

    1. **Create New Features**:
        - Generate at least one new feature from the existing data
    2. **Explain Feature Importance**:
        - Briefly discuss why the new feature(s) might improve model performance.

4. **Build and Evaluate a Linear Regression Model:**

1. **Train-Test Split**:
2. **Train the Model**:
    - Train a Linear Regression model on the training data.
3. **Evaluate the Model**:

○ Use metrics like **MAE**, **RMSE**, and **R² Score** to assess performance.
○ Plot and interpret residuals to evaluate how well the model fits the data.

5. Short Report:

In a Markdown cell within the notebook, write a brief report (2–3 paragraphs) summarizing:

- Key findings from the EDA.
- How feature engineering impacted the model.
- Model performance and challenges encountered during the process.

Bonus Task (Optional):

- Visualize the **actual vs. predicted values** for the test dataset.
- Discuss patterns or trends observed and any areas where the model might have performed poorly.

Submission Requirements:

1. A (Only One) Ipynb file containing:
   - Clear code for preprocessing, EDA, feature engineering, model training, and evaluation.
   - Visualizations and brief explanations for each step.
   - The short report is written in a Markdown cell.
2. A link to **a Public Github Repository** containing the data and the ipynb file.