**Household Energy Consumption Analysis**

This documentation outlines the various stages and components of the Python script, designed to process, analyze, and model household energy consumption data. Each section of the code is detailed for clarity.

**1. File Extraction**

**Purpose**

To extract and load the energy consumption data from a zipped file.

**Key Steps**

* File Paths: Define the paths for the zip file and the extraction location.
* Extraction: Use the zipfile module to extract the contents of the zip file.
* Load Data: Read the extracted .txt file into a pandas DataFrame with pd.read\_csv().

**Output**

* DataFrame data containing the raw energy consumption data.

**2. Data Preprocessing**

**Purpose**

To clean, parse, and transform the dataset for analysis.

**Key Steps**

* Missing Value Handling: Replace ? with pd.NA and convert numeric columns to appropriate data types.
* Datetime Parsing: Combine Date and Time columns into a single Datetime column. Extract components such as year, month, day, hour, etc.
* Imputation: Fill missing numeric values with the column mean and drop rows with missing datetime values**.**

**Output**

* A cleaned and feature-engineered DataFrame.

**3. Feature Engineering**

**Purpose**

To enhance the dataset with new features for analysis and modeling.

**Key Steps**

* Daily Aggregates: Compute daily averages of energy variables.
* Rolling Averages: Add 1-day and 7-day rolling averages of Global\_active\_power.
* Peak Hour Identification: Determine the hour with the highest average consumption.
* Weekly Averages: Summarize energy usage by day of the week.
* Scaling: Normalize selected columns using MinMaxScaler.

**Output**

* Enhanced DataFrame saved as feature\_engineered\_data.csv.

**4. Exploratory Data Analysis (EDA)**

**Purpose**

To visualize and understand the relationships and trends in the dataset.

**Key Steps**

* Correlation Heatmap: Displays correlations between numeric features.
* Histograms: Visualize distributions of energy consumption variables.
* Time Series Trends: Analyze daily, monthly, and yearly trends in energy usage.

**Output**

* Multiple visualizations showcasing data relationships and trends.

**5. Modeling**

**Purpose**

To predict Global\_active\_power using various machine learning models.

**Models Implemented**

1. **Linear Regression**
   * Train-Test Split: 80-20 split.
   * Evaluation Metrics: RMSE, MAE, R².
2. **Random Forest Regressor**
   * Default Model: Basic model training and evaluation.
   * Hyperparameter Tuning: Performed with RandomizedSearchCV.
   * Feature Importance: Visualized and tabulated.

**Evaluation Metrics**

* Root Mean Squared Error (RMSE)
* Mean Absolute Error (MAE)
* R-Squared (R²)

**Output**

* Performance comparison table for the models.
* Best parameters for the Random Forest model.

**6. Visualization of Predictions**

**Purpose**

To compare the model predictions with actual values visually.

**Key Steps**

* Scatter Plot: Visualizes actual vs. predicted values for the best Random Forest model.
* Time Series Plot: Displays a time series comparison of actual and predicted energy usage for a sample.

**Output**

* Visual comparisons of model performance.

**7. Summary and Insights**

**Purpose**

To summarize the key findings and performance of the models.

**Outputs**

* A performance comparison table for Linear Regression and Random Forest models.
* Insights into the best model for predicting household energy consumption.

**Key Libraries Used**

* Data Manipulation: pandas, numpy
* Visualization: matplotlib, seaborn
* Modeling: scikit-learn