**Optimizing Energy Consumption in Smart Homes Using Machine Learning**

**1. Introduction**

In this phase of the project, we focus on collecting, cleaning, and performing exploratory data analysis (EDA) on the dataset. The goal is to prepare the data for building a predictive model to optimize energy consumption in smart homes. This report documents the progress made in understanding the data, selecting key predictor variables, and reducing the dataset to improve modeling efficiency.

**2. Data Collection**

The dataset used for this project is the **"Individual Household Electric Power Consumption"** dataset from the UCI Machine Learning Repository. It includes minute-level data over several years, providing detailed information on household energy usage. Key variables include:

* **Global\_active\_power**: Active power consumption (kW).
* **Global\_reactive\_power**: Reactive power.
* **Voltage**: Voltage across appliances.
* **Global\_intensity**: Current intensity (amps).
* **Sub\_metering\_1, 2, 3**: Energy consumption in different household areas (kitchen, laundry, air conditioning, etc.).

**3. Data Cleaning**

During data cleaning, several issues were addressed:

* **Missing Values**: 25,979 rows with missing data were removed.
* **Outliers**: Outliers in key variables like Global\_active\_power were capped or removed to prevent distortions.
* **Date-Time Conversion**: The Date and Time columns were combined into a single DateTime column for easier time-series analysis.
* **Duplicate Rows**: No duplicates were found in the dataset.

After cleaning, the dataset was reduced to approximately 2 million rows, ready for exploratory analysis and modeling.

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

EDA provided key insights into the structure and relationships between variables:

**Distributions:**

* **Global\_active\_power**: Displays two peaks, likely representing periods of low and high consumption.
* **Global\_reactive\_power**: Most values are near zero, with occasional spikes.
* **Voltage**: Normally distributed around 240 V.
* **Sub\_metering\_3**: Contributes significantly to total power usage, indicating high-energy appliances.

**Correlation Analysis:**

A correlation matrix was used to analyze relationships between variables:

* **Global\_active\_power** and **Global\_intensity** were perfectly correlated (1.00), indicating redundancy. As a result, Global\_intensity was dropped.
* **Sub\_metering\_3** had a strong correlation with Global\_active\_power (0.74), suggesting that appliances tracked by this meter significantly contribute to energy consumption.
* **Voltage** showed a weak negative correlation (-0.28) with Global\_active\_power.

**5. Choices for Predictor Variables and Data Reduction**

Based on the EDA and correlation analysis, several choices were made to streamline the dataset and optimize model performance:

**Predictor Variables:**

The following variables were selected as potential predictors for the modeling phase:

* **Global\_active\_power**: The target variable for predicting energy consumption.
* **Global\_reactive\_power**: Contributes to understanding total power usage patterns.
* **Voltage**: While weakly correlated with Global\_active\_power, it provides essential information on household electrical behavior.
* **Sub\_metering\_3**: This sub-meter tracks high-energy appliances and is strongly correlated with overall energy consumption.

**Dropped Variables:**

* **Global\_intensity**: Dropped due to perfect correlation with Global\_active\_power, which avoids multicollinearity in the model.
* **Sub\_metering\_1** and **Sub\_metering\_2**: These sub-meters displayed weak correlations with overall power consumption and may not be significant for the predictive model. Further analysis during the modeling phase will confirm whether they can be safely excluded.

**Data Reduction Methods:**

* **Outlier Removal**: Outliers in key variables were capped or removed to ensure that extreme values do not distort the results.
* **Redundant Feature Removal**: Global\_intensity was removed to prevent multicollinearity and simplify the model.

**A table with text and numbers

Description automatically generatedSchedule**