**Data Collection and Preprocessing Phase**

|  |  |
| --- | --- |
| Date | 20 July 20255 |
| Team ID | SWTID1750360304 |
| Project Title | Power consumption analysis for households |
| Maximum Marks | 6 Marks |

**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

|  |  |
| --- | --- |
| **Section** | **Description** |
| Data Overview | The household power consumption dataset contains 2,075,259 measurements collected over 47 months (December 2006 - November 2010). It includes 9 key attributes: date, time, global active/reactive power, voltage, current intensity, and three sub-metering measurements (kitchen, laundry, and HVAC systems). Approximately 1.25% of values are missing, represented by empty entries between semicolon separators. We will examine the data structure, handle missing values, and analyze energy consumption patterns across different household systems. |
| Univariate Analysis | The univariate analysis will profile each variable's distribution through histograms (power measurements), boxplots (consumption outliers), and time plots (temporal trends). Key metrics like mean/median power usage and missing value rates will be calculated. Derived energy calculations will be validated via scatterplots against source variables. |
| Bivariate Analysis | The bivariate analysis will examine relationships between power variables using scatterplots (active vs. reactive power), correlation matrices (voltage/current intensity), and time-based heatmaps (consumption patterns by hour/day). Key interactions between submetered systems and total consumption will be analyzed through grouped bar charts and regression plots. |
| Multivariate Analysis | The multivariate analysis will explore complex relationships between power consumption, time, and weather variables using heatmaps (for time-appliance correlations), 3D plots (power-weather-time interactions), and dimensionality reduction techniques (PCA) to identify dominant household energy usage patterns and hidden dependencies in the data. |
| Outliers and Anomalies | Outliers in power consumption data will be detected using IQR and boxplot analysis. We'll assess their impact on patterns and models, then determine appropriate handling through capping, transformation, or removal to ensure robust analysis while preserving critical consumption anomalies. |
| **Data Preprocessing Code Screenshots** | |
| Loading Data | We will load the dataset into a Jupyter Notebook using pandas.read\_csv(). |
| Handling Missing Data | We will check for missing values using isnull().sum() and handle them through imputation or removal, depending on their proportion and impact. |
| Data Transformation | Categorical variables will be encoded using Label Encoding or One-Hot Encoding. Numerical variables will be scaled or normalized if required. |
| Feature Engineering | We may create new features or combine existing ones if they improve model accuracy — for example, grouping similar power consumption levels. |
| Save Processed Data | The cleaned dataset will be saved in .csv format for training and model deployment. |