**DOCUMENT : MEASURE ENERGY CONSUMPTION**

**DATA PREPROCESSING STEP:**

**By the developments on the measurement and communication infrastructures in power systems, it has become possible to collect data from more points and with higher resolutions compare to the past. Increasing data volume, on the one hand, increases the quality of the information possessed, on the other hand, it has made the processing of data more complicated [16]. With the increase in data volume, the size and variety of data quality problems has also increased. The success of data analysis is closely related to data quality. In order to obtain consistent results, missing or outlier data must be determined and removed from the data sets, and the data should be formatted in accordance with the study.**

**Data integration:**

**In general, electrical energy consumption clustering studies are based on consumption data only. However, in some studies, various data affecting electricity consumption can also be included in the analysis. In such multivariate studies, different data sets should be combined and analyzes should be performed on a single data set.**

**Data Cleaning:**

**In data analyses, it is not feasible to use raw data directly. Quality problems in the raw data may cause problems in the implementation of the analyses or in obtaining consistent results after the analysis. Some of the reasons leading to quality problems in electricity consumption.**

**Outlier Data:In its most general definition, it is the values that are far from the general data distribution and are statistically inconsistent with other data.**

**Noisy Data:It is low-quality data that is not possible to be used with the help of any software or device to make sense of the information it contains**

**Missing Data:Missing data are empty or meaningless sections in the data set as the result of problems in the phase of measurement, transfer, or storage processe.**

**Data Reduction:**

**Datasets may have more features or instances than required. Working with an unnecessarily crowded data set increases the computational effort and time in the analysis.**