Problem Statement:-

2. Perform the following operations using R/Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles b) Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification) the following operations using R/Python on the data sets:

Library:

Pandas, Matplotlib.

Theory:

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

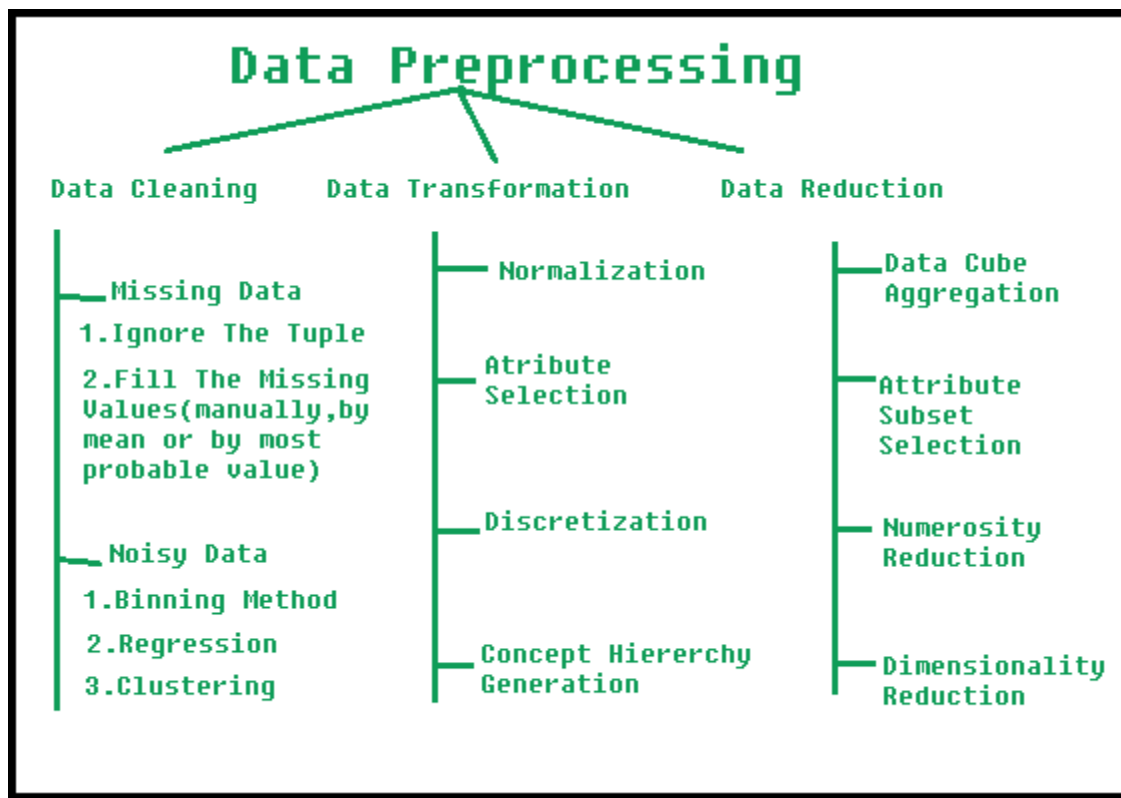**Some common steps in data preprocessing include:**

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

Diagram:

Data Preprocessing

- Data Cleaning
  - Missing Data
    1. Ignore The Tuple
    2. Fill The Missing Values(manually,by mean or by most probable value)
  - Noisy Data
    1. Binning Method
    2. Regression
    3. Clustering
- Data Transformation
  - Normalization
  - Atribute Selection
  - Discretization
  - Concept Hiererchy Generation
- Data Reduction
  - Data Cube Aggregation
  - Attribute Subset Selection
  - Numerosity Reduction
  - Dimensionality Reduction

Conclusion:

Data cleaning and transformation represent the critical processes that lay the foundation for accurate, reliable, and meaningful insights from raw datasets. The meticulous effort invested in identifying and rectifying inconsistencies, errors, and outliers ensures that data-driven decisions are based on trustworthy information.