TASK -2

Problem-Solving Scenario

● Scenario: Imagine you are tasked with analyzing a dataset of patient records but find that a large portion of the data is missing or inaccurate. What steps would you take to clean and handle this data?
● Deliverables: A brief written response (250-300 words) explaining your approach.

Answer: A methodical approach is essential when studying a dataset that has a large number of missing or erroneous data in order to guarantee accurate and noteworthy conclusions. Here is the detailed procedure:

1. **Evaluate the Data Quality:** To determine the amount of missing and erroneous data, begin by looking through the dataset. To evaluate the distribution of missing values and identify trends, use visualization tools and descriptive statistics. For example, do missing values impact specific rows or are they clustered in particular columns?

2. **Remove Duplicates:** To prevent skewed analysis, remove duplicate records. Tools like Excel's "Remove Duplicates" feature or Python programming libraries like pandas can be used for this.

3. **Dealing with Missing Values:**

Imputation: Use statistical metrics such as the mean, median, or mode to fill in missing numerical data. Replace missing values in categorical data with the most prevalent category or according to associated characteristics (e.g., age averages by department).
Removal: If the loss of data has little effect on analysis, remove rows or columns with a high percentage of missing values.
Prediction: Based on other variables, forecast missing numbers using sophisticated methods like machine learning algorithms.
Fix Inaccurate Information:
Establish data validation guidelines.

Use domain expertise or outside references to cross-check errors.
Standardize the formats of data:
Make sure that dates, text capitalization, and terminology are all formatted consistently

TASK -3

Multiple-Choice Questions (MCQs)

1. Which of the following is NOT a typical step in data cleaning?
a) Removing duplicate rows
b) Filling missing data with random values
c) Standardizing formats
d) Identifying outliers

Answer: B is correct

2. What is the purpose of normalization in data analysis?
a) To reduce the size of the data
b) To ensure all variables are on a similar scale
c) To remove duplicates from the data
d) To convert data into categorical variables

Answer: B is Correct