Task 11

CLEANING DATA

Cleaning data in Pandas involves several steps to ensure that the dataset is accurate and consistent for analysis. Here's a general outline:

- **Identify Missing Values:**

Check for missing values in the dataset and decide how to handle them. Options include removing rows or columns with missing values or filling them with appropriate values.

- **Handle Duplicate Entries:**

Look for duplicate rows in the dataset and decide whether to remove them to avoid redundancy.

- **Standardize Data Types:**

Ensure that data types are appropriate for each column. For example, numeric data should be represented as integers or floats, and dates should be in datetime format.

- **Rename Columns:**

Rename columns if they are not descriptive or if you prefer a different naming convention for clarity.

- **Clean Text Data:**

Remove any unnecessary characters or whitespace from text data.
Convert text data to lowercase for consistency.

- **Handle Outliers:**

Identify and decide how to handle outliers in the data. Options include removing them or transforming them to reduce their impact.

- **Encode Categorical Variables:**

Convert categorical variables into a numerical format suitable for analysis, such as one-hot encoding or label encoding.

- **Normalize or Standardize Data:**
- Scale numerical data to a similar range to prevent certain features from dominating the analysis due to their scale.
- **Check Data Integrity:**

Ensure that the data is consistent and free from errors, such as inconsistent units or encoding issues.

- **Address Data Imbalance:**

If dealing with classification problems, address any imbalance in the distribution of target classes to prevent bias in the analysis.

- **Document Changes:**

Document all the changes made to the dataset during the cleaning process for transparency and reproducibility.

- **Verify Cleanliness:**

After cleaning, verify that the dataset meets the necessary quality standards for analysis.