Assignment No.6

Problem Statement: Reading and writing different types of datasets.

Objective: The goal is to understand and implement data cleaning and error correction on a dataset to prepare it for analysis. This process improves data quality by handling inconsistencies, missing values, duplicates, and incorrect data.

Prerequisite :

1. **Basic Python and Pandas Knowledge**: Understanding how to use pandas for data manipulation.
2. **Data Types and Formats**: Familiarity with data types (int, float, string) and the types of errors common in datasets.

Theory :

1. Importance of Data Cleaning and Error Correction

Data cleaning is crucial because real-world data is often messy. It may contain:

- Missing values (e.g., NaNs or blank entries)
- Outliers or unrealistic values (e.g., age of -5)
- Inconsistent data formats (e.g., different date formats)
- Duplicate entries

Uncleaned data can lead to inaccurate results in analyses and machine learning models. Cleaning and correcting errors ensures reliability, consistency, and overall data quality.

2. Common Cleaning Steps

1. Handling Missing Values: Options include:
    - Dropping rows with missing values.
    - Filling missing values with a placeholder (e.g., mean or median for numeric data).
2. Removing Duplicates: Duplicate rows can cause skewed results. Identifying and removing duplicates helps maintain data integrity.
3. Outlier Detection: For example, ages in a heart disease dataset should be within a plausible range. Outliers can be identified with statistical methods or by setting reasonable limits.
4. Correcting Inconsistent Formats: For example, standardizing date formats, case sensitivity in strings (e.g., ‘Male’ and ‘male’).

Step-by-Step Process

1. **Identify Missing Values**: Use `isnull()` to locate and summarize missing data.
2. **Handle Missing Values**: Decide to drop or fill in these values.
3. **Detect and Remove Duplicates**: Use `duplicated()` and `drop_duplicates()`.
4. **Identify Outliers**: Define limits for numeric columns and adjust as necessary.
5. **Fix Data Format Issues**: Standardize string formats and convert columns to appropriate data types.

References :

Pandas Documentation on Data Cleaning

Conclusion

Data cleaning and error correction improve dataset quality, ensuring analyses and models are built on reliable, consistent data.