

Task 1: Data Cleaning Using Pandas

Task Description

You are provided with a raw dataset that contains various data quality problems such as missing values, duplicates, inconsistent formats, and incorrect data types. Your task is to clean the dataset using Pandas and prepare it for further analysis.

Tasks to Perform

1. Load the Dataset

- Import the dataset using Pandas.
- Display the first few rows and understand the structure of the data.

2. Explore the Data

- Check the number of rows and columns.
- Inspect column names and data types.
- Generate basic summary statistics.

3. Handle Missing Values

Identify missing or null values in the dataset.

Decide whether to:

- Remove rows/columns with missing values, or
- Fill missing values using appropriate methods (mean, median, mode, or constant values).
- Justify your choice.

4. Remove Duplicate Records

- Detects duplicate rows in the dataset.
- Remove duplicates and explain how many rows were affected.

5. Fix Data Types

- Identify columns with incorrect data types (e.g., numbers stored as strings, dates as text).
- Convert them to appropriate data types.

6. Standardize and Clean Text Data

Clean text columns by:

- Removing extra spaces
- Converting text to lowercase or uppercase
- Fixing inconsistent category names (e.g., “Male”, “male”, “M”)

7. Rename Columns

- Rename columns to be clear, consistent, and Python-friendly (e.g., no spaces, lowercase).

8. Final Clean Dataset

- Display the cleaned dataset.
- Save the cleaned dataset to a new CSV file.