Objective

The goal of this assignment is to apply various data cleaning techniques to detect and correct data quality issues such as missing, noisy, inconsistent, and duplicate records.

Dataset:

https://www.kaggle.com/datasets/uciml/adult-census-income

Task 1: Load and Explore

1. Import the dataset into Python (Pandas) or R or other programming languages.
2. Display its shape, column names, and first few records.

Task 2: Handle Missing Data

1. Identify columns containing missing or placeholder values (e.g., "?" or NaN).
2. Apply appropriate handling methods:
    - **Deletion** (remove rows or columns with too many missing values)
    - **Imputation** (mean, median, mode, or predictive filling)

Task 3: Detect and Treat Noisy Data

1. Identify potential **outliers** in numeric columns
2. Handle them

Task 4: Identify and Fix Inconsistent Data

1. Standardize categorical values (e.g., “Male”, “male”, “M” → “Male”).
2. Correct inconsistent formats:
    - Dates (DD/MM/YYYY vs MM-DD-YYYY)
    - Categorical labels (spaces, capitalization, typos)

Task 5: Remove Duplicates

1. Identify duplicate rows.
2. Remove or merge duplicates appropriately.

Task 6: Handle Irrelevant or Redundant Data

1. Drop columns that are not useful for analysis (e.g., ID fields).
2. Merge or encode attributes where necessary.

Deliverables

Students must submit:

- A **notebook/report (.ipynb or .pdf)** containing:
  - Step-by-step data cleaning process
  - Before and after summaries (e.g., record counts, missing value counts)
  - Visual evidence (histograms, boxplots, correlation plots)
- A **cleaned dataset file**