Task 1: Data Cleaning and Preprocessing

Objective

To clean and prepare a raw dataset by handling missing values, removing duplicates, standardizing text formats, and converting data types to ensure data consistency and readiness for analysis.

Dataset Used

Customer Personality Analysis
Source: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Tools Used

- Python 3.x
- Pandas
- Jupyter Notebook / Google Colab

Data Cleaning Steps Performed

1. Handling Missing Values

- Identified missing values using `df.isnull().sum()`
- Replaced missing values in the `Income` column with the column mean using:
  df['Income'] = df['Income'].fillna(df['Income'].mean())

2. Removing Duplicates

- Removed duplicate rows using:
  df.drop_duplicates(inplace=True)

3. Standardizing Text Formats

- Unified inconsistent values in categorical columns:
  - Converted '2n Cycle' → 'Master', 'Basic' → 'Undergraduate' in `Education`
  - Merged rare categories into broader ones in `Marital_Status` (e.g., 'YOLO' → 'Single')

4. Date Formatting

- Converted `Dt_Customer` from string to datetime format using:
  df['Dt_Customer'] = pd.to_datetime(df['Dt_Customer'], format='%d-%m-%Y')
- Extracted year and month into separate columns.

5. Renaming Columns

- Standardized column headers to lowercase and replaced spaces with underscores using:
  df.columns = df.columns.str.lower().str.replace(' ', '_')

6. Fixing Data Types

- Ensured proper types for key fields:
  - Converted `year_birth` to integer
  - Derived an `age` column as: 2025 - year_birth

Output

- Cleaned dataset saved as `cleaned_customer_personality.csv`
- Ready for EDA, modeling, or dashboard development