Task 1: Data Cleaning and Preprocessing Report

Dataset Used: Customer Personality Analysis (Kaggle)

Objective: To clean and prepare the dataset for analysis by handling missing values, removing duplicates, standardizing values, and formatting columns properly.

Steps Performed:

1. **Handled Missing Values:**
    a. Checked for missing values using filters in Excel.
    b. Found and reviewed missing values in the `Income` column.
    c. Replaced missing values with appropriate logic (e.g., average or "N/A").
2. **Removed Duplicate Rows:**
    a. Used Excel's "Remove Duplicates" feature.
    b. Ensured all columns were selected to check for full-row duplicates.
3. **Standardized Text Columns:**
    a. Created a new column `marital_status_cleaned`:
        i. Combined values like "Together" into "Married".
        ii. Combined values like "Single", "Alone", "Widow", "Divorced", and "YOLO" into "Single".
    b. Cleaned the `Education` column:
        i. Ensured consistent formatting like "PhD", "2nd Cycle", "Graduation", etc.
        ii. Used formulas like `PROPER()` and `TRIM()` in Excel.
4. **Converted Date Formats:**
    a. Standardized the `Dt_Customer` column to the `dd-mm-yyyy` format using Excel's date formatting.
5. **Renamed Column Headers:**
    a. Ensured column headers are lowercase and free of spaces (e.g., `Year_Birth` to `year_birth`).

    b. Added new derived columns like `marital_status_cleaned`.
6. **Checked and Fixed Data Types:**
    a. Created Age using formula: `=YEAR(TODAY()) - Year_Birth`.
    b. Ensured `Income` is numeric and `Dt_Customer` is in date format.
7. **Non-Applicable Columns:**
    a. No `Gender` or `Country` columns existed in this dataset.

Final Output:

- Cleaned dataset with consistent text values, fixed types, no missing values or duplicates.
- Dataset is now ready for analysis or visualization tasks.

New Columns Added:

- `marital_status_cleaned`
- (Optionally) Age

Tools Used: Microsoft Excel

Prepared By: Yogesh Kumar

Date: 07-04-2025

<u>Learning Outcomes:</u>

By completing this task, I have:

- Gained hands-on experience in identifying and fixing common data issues like missing values, duplicates, and inconsistent formatting.
- Learned to use Excel functions for real-world data cleaning.
- Improved my understanding of data pre-processing, which is a critical step before data analysis or visualization.
- Built confidence in handling raw datasets independently.
- Created a clean, structured dataset that is ready for analysis or modelling.