# Data Cleaning & Preprocessing Report

Task Name: Task 1 - Data Cleaning and Preprocessing

Dataset: Customer Personality Analysis (Kaggle)

Tool Used: Python (Pandas),Excel

Prepared By: Sanika Palaw

Date: May 26, 2025

## Objective

To clean and prepare a raw dataset by handling missing values, removing duplicate records, standardizing formats, and ensuring the data is structured and consistent for analysis.

## Initial Data Overview

- Rows: 2240  
- Columns: 29  
- Key Fields: Income, Year\_Birth, Dt\_Customer, Marital\_Status, Education, MntWines, NumDealsPurchases, etc.

## Cleaning Steps Performed

|  |  |  |
| --- | --- | --- |
| Step | Description | Result |
| 1 | Checked Missing Values using isnull().sum() | 24 missing in Income |
| 2 | Filled Missing Values in Income | Replaced with median income |
| 3 | Checked Duplicates using duplicated().sum() | 0 duplicate rows found |
| 4 | Standardized Text Data | Converted object columns to lowercase and removed spaces |
| 5 | Renamed Columns | Used lowercase and underscores |
| 6 | Converted Date Format | Dt\_Customer converted to datetime format |
| 7 | Fixed Data Types | Converted year\_birth to integer |

## Final Dataset Summary

- No missing values  
- No duplicate rows  
- Standardized column names and text values  
- Clean and uniform data types (int, float, datetime)  
- Ready for visualization, analysis, or modeling

## Files Generated

- cleaned\_data.csv – Final cleaned dataset  
- cleaning\_script.ipynb – Python script used for data cleaning  
- README.md – GitHub documentation  
- data\_cleaning\_report.docx – This report

## Key Learnings

- Importance of handling missing and duplicate data  
- Text standardization improves data consistency  
- Proper data types are crucial for downstream tasks  
- Clean data = better insights and modeling performance