Summary of Data Preprocessing – Internship Task 1

After performing a detailed data cleaning and preprocessing operation on the Customer Personality Analysis dataset, the following transformations and improvements were made:

## 1. Duplicate Removal

Identified and removed all duplicate records to avoid data bias and repetition using df.drop\_duplicates().

## 2. Missing Values Handling

Detected missing values using df.isnull().sum(). Handled them by filling with mean or mode values for numerical and categorical columns, and dropped rows with excessive missing data.

## 3. Column Name Standardization

Renamed all column headers to a consistent format:  
- Converted to lowercase  
- Removed leading/trailing spaces  
- Replaced spaces with underscores  
Used: df.columns = df.columns.str.strip().str.lower().str.replace(' ', '\_')

## 4. Text Value Standardization

Standardized categorical values (e.g., gender, education, marital status) by converting all text to lowercase and removing whitespace.

## 5. Date Format Conversion

Converted the dt\_customer column from object to datetime format and formatted as dd-mm-yyyy using:  
df['dt\_customer'] = pd.to\_datetime(df['dt\_customer'], dayfirst=True, errors='coerce')  
df['dt\_customer'] = df['dt\_customer'].dt.strftime('%d-%m-%Y')

## 6. Data Type Corrections

Ensured numerical columns had correct data types. Casted float columns to integers (e.g., kidhome, teenhome) using .astype().

## 7. Feature Engineering

Added new columns:  
- age = 2025 - year\_birth  
- total\_kids = kidhome + teenhome  
- customer\_duration\_days = days since customer joined

## 8. Final Dataset Export

Saved the cleaned dataset using UTF-8 encoding with:  
df.to\_csv('cleaned\_data.csv', index=False, encoding='utf-8')