Day -1

**🧹 Data Cleaning & Preprocessing Summary**

**Dataset:** *Mall Customer Raw Data*  
**Objective:** Prepare a clean and consistent dataset by addressing missing values, duplicates, inconsistent formatting, and outliers.

**Summary of Changes Made**

1. **Column Renaming**
   * Standardized all column names to snake\_case for consistency and ease of use.
   * Example:
     + CustomerID → customerid
     + Annual Income (k$) → annual\_income\_k
     + Spending Score (1-100) → spending\_score\_1100
2. **Missing Values**
   * Checked for missing/null values using .isnull().sum()
   * No missing values were found in any column.
3. **Duplicate Rows**
   * Checked for duplicates using .duplicated().sum()
   * No duplicate rows were found.
4. **Text Standardization**
   * Converted gender values to lowercase (Male, Female → male, female)
   * Removed any leading/trailing whitespace (if present).
5. **Data Type Consistency**
   * Ensured each column has an appropriate data type:
     + customerid, age, annual\_income\_k, spending\_score\_1100 → int64
     + gender → object (categorical)
6. **Outlier Detection**
   * Detected outliers in the annual\_income\_k column using the IQR method.
   * Found **2 potential outliers**, but they were not removed or capped, as they may represent high-income valid customers.

**Final Dataset Overview**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| customerid | Integer | Unique ID for each customer |
| gender | Categorical | Customer gender (male/female) |
| age | Integer | Age of the customer |
| annual\_income\_k | Integer | Annual income in thousands ($) |
| spending\_score\_1100 | Integer | Spending score from 1 to 100 |

**Conclusion**

The dataset has been cleaned, standardized, and validated with no missing or duplicate entries. It is now ready for analysis, visualization, or machine learning tasks.