Subtask 3: Data Cleaning and Preprocessing

Raw datasets often contain **missing values, duplicates, incorrect data formats, and inconsistencies** that can affect analysis. In this step, you will **clean and preprocess the data** to ensure accuracy and reliability. This involves handling missing values, removing duplicates, correcting data types, and preparing the dataset for further analysis.

By completing this step, you will ensure that the dataset is **structured, error-free, and ready for exploratory analysis**.

Subtask 3: Data Cleaning and Preprocessing

🛠 How You Can Perform This Task?

1 Handle Missing Values
- Identify missing values in critical columns like **Order ID, Customer ID, Revenue, and Date**.
- Decide on the best approach: **remove incomplete records or fill missing values with appropriate methods** (e.g., median, mean, or forward-fill techniques).

2 Remove Duplicate Records
- Check for **duplicate transactions** by examining key columns like Order ID and Customer ID.
- Remove redundant entries to avoid skewed results in the analysis.

3 Correct Data Types
- Convert **date columns** into a standard datetime format.
- Ensure **numeric fields (such as revenue and quantity) are in the correct numerical format**.
- Convert categorical variables (e.g., product categories) into a **standardized format**.

4 Fix Data Inconsistencies
- Standardize text-based columns like product categories to avoid variations (e.g., "Electronics" vs. "electronics").
- Correct negative or unrealistic values (e.g., negative prices or order quantities).

5 Handle Outliers
- Identify extreme values in **revenue and quantity sold** that could be errors or anomalies.
- Decide whether to cap, transform, or remove outliers based on business logic.

6 Create New Columns (if necessary)
- Generate new columns like **Total Revenue (Quantity × Price)** if not already available.
- Extract insights from the date column, such as **day of the week, month, or year** to analyze trends.

7 Save the Cleaned Dataset
- Store the cleaned dataset in a new CSV file to **ensure reproducibility**.
- Maintain a copy of the raw dataset for reference.

Tasks

- [ ] Identified and handled missing values appropriately.
- [ ] Removed duplicate records.
- [ ] Standardized and corrected **data types** (dates, numbers, text).
- [ ] Fixed **data inconsistencies and unrealistic values**.
- [ ] Identified and handled outliers.
- [ ] Created any additional **useful columns** for analysis.
- [ ] Saved the cleaned dataset for future use.