Assignment 1

Dataset: Cafe Sales – Dirty Data for Cleaning Training
https://www.kaggle.com/datasets/ahmedmohamed2003/cafe-sales-dirty-data-for-cleaning-training

You will work with a **café sales dirty dataset** that contains approximately 10,000 point-of-sale transactions from a café.

The dataset includes raw transactional data with realistic data quality issues commonly found in real-world sales systems.

The dataset intentionally includes:

- **Duplicate records** (from repeated transaction exports or system errors)
- **Null values** (missing data in customer names, products, timestamps, or payment information)
- **Wrong formats** (dates/times stored as strings, prices with inconsistent formatting, numbers stored as text)
- **Inconsistencies** (mixed capitalization in product names, varied spelling of payment methods, inconsistent category labels)
- **Outliers** (unrealistic transaction amounts or quantities due to data entry errors)

Your task is to **detect, clean, and preprocess** the data to make it ready for exploratory analysis and machine learning.

Requirements for Submission

You must submit the following deliverables:

1. **Cleaned dataset file** (cafe_sales_cleaned.csv)
2. **The original dataset file (.csv) before preprocessing**
3. **Python code with clear comments and explanations**

Your Python notebook or script should include:

- **Initial data inspection**: Display the first few rows, data types, summary statistics, and provide a written description of observed data quality issues.
- **Issue detection summary**: Document counts of null values per column, number of duplicate records, examples of format inconsistencies, and identified outliers.
- **Cleaning and preprocessing steps**: Show the code used for handling missing values, removing or merging duplicates, converting data types, standardizing text formats, and treating outliers.
- **Before/after examples**: Provide specific examples showing how particular rows or columns looked before cleaning and after your transformations.

Evaluation Criteria

Your work will be evaluated based on:

- Thoroughness of data quality issue detection
- Appropriateness of cleaning techniques applied
- Code quality and documentation
- Clear explanation of your preprocessing decisions
- Quality of the final cleaned dataset

Deadline for submissions: 28th of November at 12:00 afternoon