Coffee Sales Project - Data Cleaning Process

This document outlines the step-by-step process I followed to clean and prepare the dataset for analysis.

The goal was to ensure accuracy, consistency, and reliability of insights derived from the data.

1. Data Import

- Imported the dataset into **Excel** from https://mavenanalytics.io/data-playground/coffee-shop-sales?pageSize=10

- Created a backup copy to preserve raw/original data.

2. Initial Inspection

- Checked the dataset shape (rows and columns).

- Scanned for missing, inconsistent, or duplicate values.

- Identified data types (numeric, categorical, date).

3. Handling Missing Data

- Removed rows with completely empty fields.

- For partially missing values:

  - Used **mean/median imputation** for numerical columns.

  - Used **mode imputation** for categorical fields.

  - Where imputation wasn't appropriate, marked as `N/A`.

4. Removing Duplicates

- Used Excel's *Remove Duplicates* function on key identifiers.
- Confirmed that no customer/transaction/product was counted more than once.

5. Standardizing Data Formats

- Standardized **date formats** to `YYYY-MM-DD`.
- Converted all categorical values to consistent text case (e.g., "Yes" vs "YES").
- Ensured all numeric fields were properly formatted (no text/number mismatches).

6. Outlier Detection

- Checked for unusually high or low values using:
  - Conditional formatting in Excel.
  - Simple statistical checks (mean ± 3 standard deviations).
- Outliers were:
  - Retained if valid (e.g., high-value transaction).
  - Corrected/removed if identified as entry errors.

7. Data Validation

- Applied Excel **Data Validation rules** (drop-downs, restricted ranges).
- Cross-checked totals against known benchmarks for accuracy.

8. Final Clean Dataset

- Exported the clean dataset into a new Excel sheet titled `Cleaned_Data`.
- Ensured it was analysis-ready, with no blanks, inconsistencies, or duplicates.