Purpose of the Dataset

The dataset represents supermarket sales data collected from three cities (Yangon, Naypyitaw, and Mandalay). It includes transactional details such as unit prices, quantities, payment methods, customer ratings, and dates/times of purchase. The purpose of this dataset is to analyze sales trends, customer behavior, and city performance to inform business decisions.

Objective of the Data Wrangling Process

The primary objective of the data wrangling process is to clean and transform the dataset to ensure accuracy, consistency, and readiness for analysis. This includes handling missing values, correcting data types, and identifying and resolving any inconsistencies or outliers.

Data Overview

Description of the Dataset

- **Number of Rows:** 1006 rows
- **Number of Columns:** 16 columns
- **Data Types:**
    - Numeric: Unit price, Quantity, Total, Tax 5%, Rating
    - Categorical: Product line, Payment, Gender
    - Date/Time: Date, Time

Key Issues Identified During Initial Exploration

1. **Missing Values:** Missing entries in columns like Time and Tax 5%.
2. **Data Types:**
    - Unit price was incorrectly stored as a string (object) instead of float.
    - Date and Time columns needed consistent formatting.
3. **Duplicate Records:** Identified rows with duplicate data.
4. **Unnecessary Columns:** Columns for individual cities (Yangon, Naypyitaw, Mandalay) were redundant after creating the City column.
5. **Outliers:** Extreme values in numeric columns like Rating and Total.

Wrangling Process

Steps Taken to Clean and Transform the Data

1. **Handling Missing Values:**

    - Filled missing Tax 5% values using the formula: Tax 5% = Total * 0.05.
    - Filled missing Total values using the formula: Quantity* Unit price + Tax 5%
2. **Data Type Corrections:**
    - Converted Unit price from object to float.
    - Reformatted Date and Time columns to consistent formats (YYYY-MM-DD for Date, HH:MM for Time).
3. **Creating a Consolidated Column:**
    - Combined individual city columns into a single City column with values "Yangon", "Naypyitaw", and "Mandalay".
4. **Removing Duplicates:**
    - Identified and dropped duplicate rows to ensure each record was unique.
5. **Removing Unnecessary Columns:**
    - Dropped the redundant columns: Yangon, Naypyitaw, and Mandalay.
6. **Handling Outliers:**
    - Reviewed numeric columns (Total, Rating) for extreme values and ensured they were valid.

Justifications for Decisions:

- **Imputation of Missing Values:** Ensures no loss of data while maintaining consistency and logical correctness.
- **Type Corrections:** Ensures proper numerical and temporal operations can be performed.
- **Removing Duplicates:** Avoids bias caused by repeated entries.
- **Column Removal:** Simplifies the dataset for analysis without loss of information.
- **Outlier Handling:** Retains realistic data for accurate analysis.

Visualizations

Before-and-After Comparisons

1. **Missing Data:**
    - **Before:** Visualization (e.g., heatmap) showing missing values in Time and Tax 5%.
    - **After:** Visualization showing no missing values post-imputation.
2. **Distribution of Ratings:**

- **Before:** Histogram showing skewed or irregular distribution of Rating values.
- **After:** Histogram showing cleaned and consistent distribution.

3. **City Representation:**
    - **Before:** Pie chart showing three individual city columns.
    - **After:** Pie chart showing consolidated City column.

Conclusion

Summary of the Cleaned Data

The dataset has been successfully cleaned and transformed:

- Missing values have been appropriately handled.
- Data types have been corrected to enable proper analysis.
- Duplicate rows and unnecessary columns have been removed.
- The data is now consistent, complete, and ready for visualization and analysis.

Readiness for Analysis

The cleaned dataset is now ready for exploratory data analysis, trend identification, and insights extraction to support business decision-making.

Tools Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn