# Project Documentation: Supermarket Sales Data Cleaning and Analysis

## Project Overview

This project focuses on cleaning and preparing the "Supermarket Sales" dataset for analysis. The dataset contains detailed sales records across different branches of a supermarket, including information about sales transactions, customer demographics, and product categories.

### Goals:

* Identify and resolve quality and tidiness issues in the dataset.
* Prepare the data for further analysis and visualization by ensuring it is clean, consistent, and structured.

## Data Overview

The dataset contains the following fields:

* **Invoice ID:** A unique identifier for each transaction.
* **Branch:** The supermarket branch where the sale took place (A, B, or C).
* **City:** The city where the branch is located (Yangon, Mandalay, or Naypyitaw).
* **Customer Type:** Type of customer (Member or Normal).
* **Gender:** Gender of the customer.
* **Product Line:** The category of the product sold.
* **Unit Price:** The price per unit of the product.
* **Quantity:** The number of units sold.
* **Tax 5%:** The amount of tax applied to the sale.
* **Total:** The total amount for the sale, including tax.
* **Date:** The date of the sale.
* **Time:** The time of the sale.
* **Payment:** The method of payment used (Cash, Credit card, or E-wallet).
* **Rating:** The customer’s rating of their shopping experience.

## Data Cleaning Process

### Quality Issues

1. **Missing Values:**
   * The 'Tax 5%' and 'Total' columns contained missing values.
2. **Invalid Data Types:**
   * The 'Unit price' column had 'USD' included in the values and needed to be converted to float.
   * The 'Invoice ID' column contained hyphens and needed to be converted to integer.
3. **Inconsistent Values:**
   * The 'Customer type' column had inconsistent entries (e.g., 'Memberr' instead of 'Member').
4. **Incorrect Values:**
   * The 'Rating' column contained an incorrect value of 97, which needed correction.
   * The 'Quantity' column contained negative values.
5. **Duplicate Rows:**
   * Several duplicate rows were found and needed to be removed.

### Tidiness Issues

* **Multiple Columns for One Variable:**
  + The dataset had separate columns for each city ('Yangon', 'Naypyitaw', and 'Mandalay') instead of a single 'City' column.

## Data Cleaning Steps

1. **Remove Duplicates:**
   * Identified and removed duplicate rows to ensure data consistency.
2. **Fix 'Rating' Values:**
   * Corrected the outlier value in the 'Rating' column (changed 97 to 9.7 based on the scale of other ratings).
3. **Fix 'Quantity' Values:**
   * Changed negative values in the 'Quantity' column to positive values to maintain logical consistency.
4. **Clean 'Invoice ID':**
   * Removed hyphens from 'Invoice ID' and converted it to an integer type for ease of use.
5. **Clean 'Unit Price':**
   * Removed 'USD' from the 'Unit price' values and converted the column to a float type for numerical operations.
6. **Correct 'Customer Type':**
   * Standardized inconsistent entries in the 'Customer type' column to ensure uniformity ('Memberr' was corrected to 'Member').
7. **Fill Missing Values:**
   * Filled missing values in 'Tax 5%' and 'Total' columns based on existing data (e.g., recalculating values based on the 'Quantity' and 'Unit price').
8. **Clean 'Time' Column:**
   * Removed 'PM' and 'AM' from the 'Time' values and converted the time to a 24-hour format for easier time-based analysis.
9. **Create 'City' Column:**
   * Merged the 'Yangon', 'Naypyitaw', and 'Mandalay' columns into a single 'City' column by mapping the branch codes (A, B, and C) to their respective cities.

## Final Dataset

The cleaned dataset was saved as Cleaned\_data.csv, and it is now ready for further analysis. The dataset is free of duplicates, missing values, and inconsistencies, ensuring high-quality data for accurate insights and visualizations.

### File:

* **Cleaned\_data.csv**