**Supermarket Sales Data Cleaning and Analysis**

**1 - Data Gathering**

**-- Loading the data from CSV file**

**Data Overview**

The dataset contains information about supermarket sales, including:

* **Invoice ID:** Unique identifier for each transaction.
* **Branch:** The branch where the sale occurred (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:** Category of the product sold.
* **Unit price:** Price of a single unit of the product.
* **Quantity:** Number of units sold.
* **Tax 5%:** Tax applied to the sale.
* **Total:** Total amount of the sale.
* **Date:** Date of the sale.
* **Time:** Time of the sale.
* **Payment:** Payment method used.
* **Rating:** Customer rating of the shopping experience.
* **Sale with taxes:** Total amount of the sale with tax.

**Data Cleaning Process**

**Tidiness Issues**

* **Multiple columns represent one column:**  The issue with the data is that locations (Yangon, Naypyitaw, Mandalay) are represented as columns, which creates redundant columns for each location.

**Quality Issues**

**Dataset have duplicated rows**

**- `Unit price` -  This columns are not in correct data tybe**

**- `Quantitiy` -  have nagitive values**

**- Two columns containing missing data**

**1.  `Tax 5%` : 9 missing data**

**2.  `Total`  : 3 missing data**

**- `Customer type`  column due to inconsistent data entries**

**1- " - "**

**2. " memberr "**

**- `Time` column have value is not in  time standared format**

**- `Rating` column have value is not in  between 1 and 10**

**Cleaning Steps**

1. **Remove duplicates:** Duplicate rows were dropped.
2. **Fix 'Rating' values:** Replaced the incorrect value 97 with 9.7.
3. **Fix 'Quantity' values:** Converted negative values to positive.
4. **Clean 'Invoice ID':** Removed hyphens and converted to int.
5. **Clean 'Unit price':** Removed 'USD' or other non numeric char and converted to float and remove 0 value.
6. **Clean 'Customer type':** Corrected inconsistent values.
7. **Fill missing values:** Filled missing values in 'Tax 5%' and 'Total' columns based on other values.
8. **Clean 'Time':** converted from 24-hour format to 12 h.
9. **Convert 'Customer type' column :** from ( - ) to unkown.

**Data Tidying**

* **Combined columns:** The 'Yangon', 'Naypyitaw', and 'Mandalay' columns were combined into a single 'City' column by creating a function to map branch codes to city names.