Data Wrangling Report

Project Objectives:

The primary objective of this project is to perform data wrangling techniques on a dataset to prepare it for further analysis. This includes gathering, assessing, and cleaning data to ensure it is ready for use in a data analysis pipeline.

Gathering Data:

-We gathered data from supermarket in Myanmar country which in Asia this data is a collection on data from 3 branches of this market which located in different parts in Myanmar in an excel file.

-We used this Supermarket data and read this file as (Supermarket Sales.csv").

Assessing Data:

After gathering data, the next step involves assessing its quality. This may include:

- Checking for missing values
- Identifying duplicates
- Understanding the structure of the dataset (column types, value ranges)

- Detecting any anomalies or outliers

-We use common Python functions like info(), describe().

Cleaning Data:

This phase includes fixing or removing any issues identified during assessment.

1. Data tidiness issues:

Dataset	Observation	Solution
Supermarket Sales	1. There are multiple columns (e.g., Yangon, Naypyitaw, Mandalay) representing branches as separate columns	We melted this 3 columns in 1 column called city

2. Data Quality Issus:

Dataset	Observation	Solution
Supermarket Sales	The Tax 5% column has 9 missing values. The Total column has 3 missing values.	since that the missing values in columns like (Tax 5%,Total) which we can calculate by this formulas: • Tax 5% = Quantity x Unit price x 0.05 • Total = Quantity x Unit price + Tax 5%

Unit price and Total should be floating data type	We changed the datatype of unit price and recalculated total during missing value problem
There are 6 duplicate rows in the dataset	We removed them as they don’t give us a new information
Customer type has 27 Nulls represented as dashes (-) records.	We have decided to drop these records. As they doesn’t consume a large portion of the data. Imputing values for these records could potentially skew the analysis, so removing them will help maintain the accuracy and integrity of our results.
Rating columns have value equal 97	We consider this rating problem as wrong during inserting the data so we replace it with 9.7
The quantity column has negative values.	We decided to take the absolute value of this column so we can have all values in the same format as negative value can skew the analysis and give wrong information
The Time column has inconsistent formats (e.g., 8 - 30 PM).	This problem which could cause problems when performing time-based analyses we put all time in the same 24 h format
Customer type has 'memberr' instead of 'member'.	We consider this rating problem as wrong during inserting the data so we replace it with member
Unit price column includes prices in USD.	We first made sure that all values are in USD by calculating the total with the formula : Total = Quantity x Unit price + Tax 5% And found out that they have the same currency value After that we removed USD remark so that we have the same format in the coulmn

Result:

Before Data Wrangling:

```
Supermarket_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1006 entries, 0 to 1005
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Invoice ID     1006 non-null   object
 1   Branch         1006 non-null   object
 2   Yangon         1006 non-null   int64
 3   Naypyitaw      1006 non-null   int64
 4   Mandalay       1006 non-null   int64
 5   Customer type  1006 non-null   object
 6   Gender         1006 non-null   object
 7   Product line   1006 non-null   object
 8   Unit price     1006 non-null   object
 9   Quantity       1006 non-null   int64
 10  Tax 5%         997 non-null    float64
 11  Total          1003 non-null   object
 12  Date           1006 non-null   object
 13  Time           1006 non-null   object
 14  Payment        1006 non-null   object
 15  Rating         1006 non-null   float64
dtypes: float64(2), int64(4), object(10)
memory usage: 125.9+ KB
```

After Data Wrangling:

```
Supermarket_data_copy.info()

<class 'pandas.core.frame.DataFrame'>
Index: 973 entries, 0 to 1003
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Invoice ID     973 non-null    object
 1   Customer type  973 non-null    object
 2   Gender         973 non-null    object
 3   Product line   973 non-null    object
 4   Unit price     973 non-null    float64
 5   Quantity       973 non-null    int64
 6   Tax 5%         973 non-null    float64
 7   Total          973 non-null    float64
 8   Date           973 non-null    object
 9   Time           973 non-null    object
 10  Payment        973 non-null    object
 11  Rating         973 non-null    float64
 12  City           973 non-null    object
dtypes: float64(4), int64(1), object(8)
memory usage: 106.4+ KB
```

Now the data is ready for data analysis.