Data Wrangling Report

1. Data Gathering

- A structured CSV file containing transaction records.

2. Data Assessing

After loading the dataset, an assessment was performed to identify data quality and tidiness issues.

Quality Issues

- Missing values in critical columns such as Total and Tax 5%.
- Incorrect data types, such as Date stored as a string instead of datetime.
- Negative values in the Quantity column.
- Inconsistent naming in Customer type (e.g., "Memberr" instead of "Member").
- Unexpected characters (-) in some categorical fields.

Tidiness Issues

- Branch column could be derived from Yangon, Naypyitaw, and Mandalay columns.
- Tax 5% should be calculated as Total - (Unit price * Quantity).
- Some categorical variables were split across multiple columns instead of a single column.

3. Data Cleaning

To address the identified issues, the following steps were taken:

Fixing Quality Issues

- Missing values in Total and Tax 5% were recalculated using available information.
- Data types were corrected (e.g., Date converted to datetime format).
- Negative Quantity values were manually corrected based on logical assumptions.
- Standardized Customer type, replacing Memberr with Member.
- Replaced - values with "Not Registered" in categorical fields.

Fixing Tidiness Issues

- Branch column was reconstructed using the encoded values of Yangon, Naypyitaw, and Mandalay.

- Tax 5% was recalculated using the formula: Total - (Unit price * Quantity).
- Categorical variables were merged properly into a single structured format.

Combine Data Frames

- All cleaned datasets were merged into a single master DataFrame, ensuring consistency and completeness.

4. Data Storing

After cleaning, the final dataset was stored in a CSV file (Master_DataFrame.csv) for further analysis and visualization.

5. Data Visualization

The cleaned data was visualized to derive meaningful insights:

Sales Trends Over Time

- A time series plot was created to analyze Total sales trends over Date.
- This helped in identifying seasonal patterns and sales fluctuations.

Sales Distribution by Branch

- A bar chart was used to compare total sales across different branches.
- Helped in understanding which branch had the highest revenue.

Top-Selling Products

- A bar chart visualizing total sales by Product line.
- Provided insights into the most profitable product categories.

Customer Satisfaction Analysis

- Ratings were analyzed based on Customer type to determine which customers were most satisfied.
- Helped in identifying areas for customer experience improvement.

Conclusion

By performing a structured data wrangling process, we successfully cleaned and organized the dataset, making it suitable for analysis. The insights derived from the visualization provide valuable business intelligence to improve operations and customer satisfaction.
