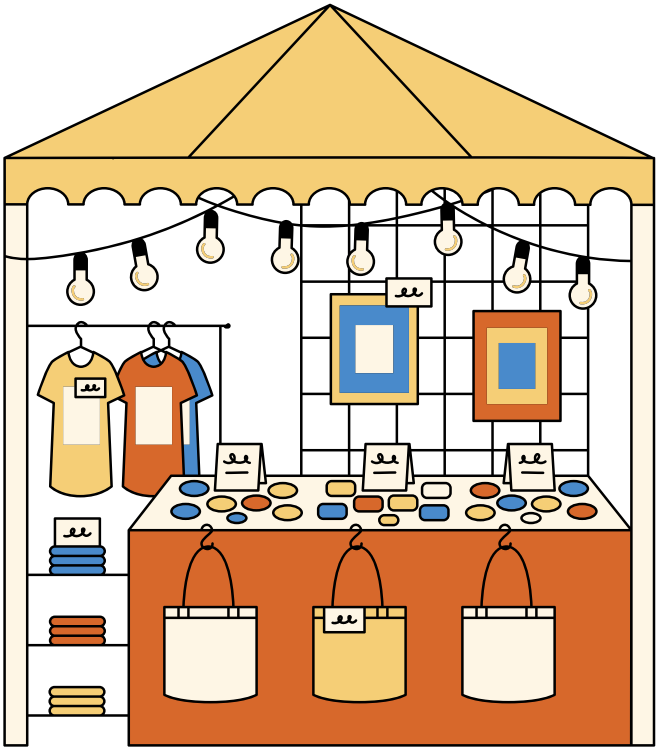
NTI Data Analysis Capstone Project

[Supermarket Sales]

Data Wrangling Project



By: Team El يحابيح

Introduction

The supermarket sales dataset contains detailed transactional records from multiple branches of a supermarket. The dataset provides valuable insights into customer purchasing patterns, including information on product categories, customer demographics, payment methods, and ratings of customer satisfaction.

This report outlines the data wrangling process undertaken to clean and prepare the dataset for further analysis. The purpose of this analysis is to gain insights into various aspects of supermarket sales, such as customer behavior, product popularity, and sales trends, to inform business decisions. The report covers the identification and treatment of missing data, outliers, inconsistencies, and duplicates. After cleaning, the dataset was used to explore potential insights into sales performance and customer preferences, focusing on customer type, sales by product line, and peak shopping hours.

Project index :

1. Data Gathering

The dataset provided contains 1,006 rows and 16 columns. This dataset includes information about customer transactions, covering:

- Invoice ID: Unique identifier for each transaction.
- Branch: Store location (Branch A, B, C).
- City: Cities where transactions occurred (Yangon, Naypyitaw, Mandalay).
- Customer Type: Normal or Member customers.
- Product Line: Categories of products purchased.
- Unit Price, Quantity, Total, Tax 5%: Pricing and quantity details.
- Payment: Payment method (Cash, Credit Card, E-wallet).
- Rating: Customer satisfaction rating.

Data Assessing

After gathering each of the above pieces of data, we need to assess them visually and programmatically for quality and tidiness issues.

1- Quality Issues

Quality issues are concerning:

1-Missing Values:

- The Customer Type column contains missing data. There are only two valid categories: "Normal" and "Member." Some values are incorrect, such as "memberr," which needs to be corrected.

- Missing values were also identified in the Total and Tax 5% columns. These values were recalculated based on the available unit price and quantity.

2-Outliers:

- The Rating column contains an incorrect value of "97" (should be "9.7").
- Negative values were found in the Quantity column, which is inconsistent with valid transaction data.

3-Duplicate Records:

- Six duplicate rows were found in the dataset and removed.

2- Tidiness issues

These issues comes from the concept of Tidy Data which means :

1- Each variable forms a column and contains values.

2- Each observation forms a row.

3- Each type of observational unit forms a table.

3. Data Cleaning

3.1 Fixing Quality Issues

- Missing Values:
  - Missing data in the Customer Type column was filled based on available information. Invalid entries such as "memberr" were corrected to "Member."
  - Missing values in the Total and Tax 5% columns were recalculated using the formula:
    - Tax = 5% of (Unit Price * Quantity)
    - Total = (Unit Price * Quantity) + Tax
- Outliers:
  - The incorrect rating value of "97" was replaced with "9.7."
  - Negative values in the Quantity column were corrected based on logical assumptions.

3.2 Fixing Tidiness Issues

- The city columns (Yangon, Naypyitaw, Mandalay) were merged into a single City column.

4. Data Transformation

No significant data transformations were needed except for standardizing the time format for one record where the value was "8 - 30 PM." It was converted to a proper time format (20:30).

5. Data Analysis

5.1 Customer Type Insights

- Using the mode function, it was determined that "Normal" customers are more frequent than "Member" customers in the dataset.

5.2 Product Line and Sales Analysis

- Sales were broken down by product lines, revealing which categories generated the highest revenue.
- Further analysis can show the relationship between product categories and the total cost of transactions.

5.3 Peak Shopping Hours

- The Time column was analyzed to identify peak shopping hours, providing insights into the busiest times for the supermarket.

6. Conclusion

The data wrangling process successfully identified and corrected several data quality issues, including missing values, outliers, and duplicates. The cleaned dataset is now ready for further analysis to explore customer behavior, product popularity, and other sales-related trends. This will help in making informed business decisions and optimizing sales strategies.

THANK YOU!