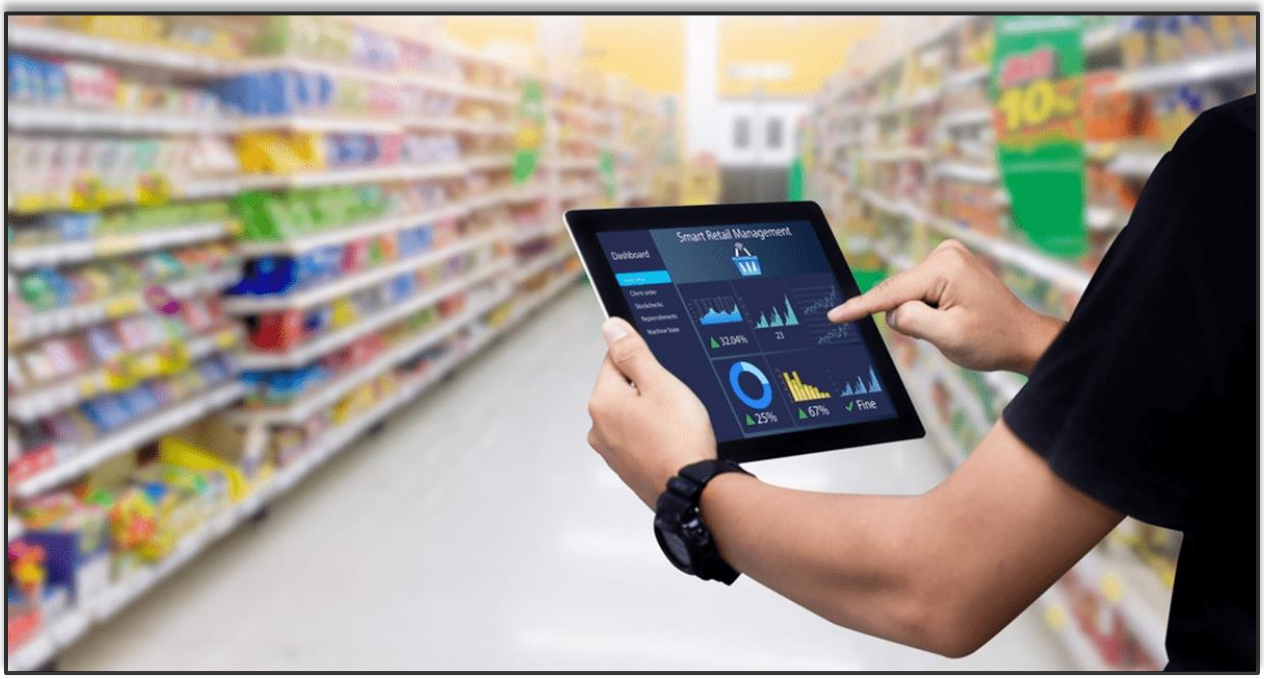


NTI Capstone Project

Supermarket Sales

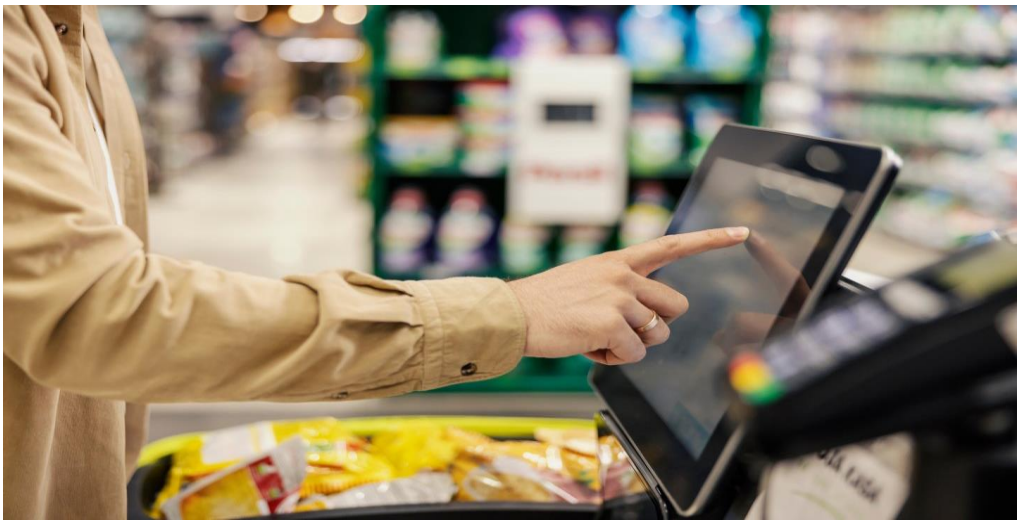
Data Wrangling Report

Introduction

Data wrangling involves gathering, assessing, and cleaning data to prepare it for analysis. In this report, I will document the data wrangling process for a supermarket dataset, including data gathering, assessment, cleaning, and analysis steps. The cleaned dataset will be used to generate insights into sales patterns, customer behavior, and inventory management to support business decision-making.

The dataset contains detailed transactional data from a supermarket, capturing information such as the date and time of purchase, product details, quantities, prices, payment methods, and customer IDs. Analyzing this data provides valuable insights into purchasing trends, popular products, and inventory optimization, allowing for more efficient operations and improved customer satisfaction.

Supermarket transactional data is extensive and varied, reflecting diverse customer preferences and purchasing behavior. Through this data wrangling and analysis, we aim to uncover key insights that can help in understanding purchasing patterns, identifying top-selling products, optimizing inventory levels, and enhancing overall business strategy.



Project index:

Data Gathering

1.1 File in Hand: Supermarket_Sales.csv

The main dataset for this analysis is a CSV file, Supermarket_Sales.csv, containing transactional data from the supermarket. It includes key details such as transaction ID, date, time, product category, unit price, quantity sold, total amount, store location, and customer demographics (e.g., age, gender). This data will provide the foundational information needed to analyze sales trends, customer preferences, and revenue generation.

1.2 File with Additional Columns:

- City (Yangon, Naypyitaw, Mandalay): Derived from the Branch column to represent transactions by city.
- Day of the Week: Extracted from the Date to analyze weekly sales patterns.
- Customer Loyalty Score: A calculated score to segment customers based on purchase frequency and Rating.
- Month: Derived from the Date column to analyze sales trends on a monthly basis.
- Rating Category: Categorized Rating into groups (e.g., Low, Medium, High) to better analyze customer satisfaction levels.

Data Assessing

After gathering the data, we assess it for both quality and tidiness issues:

2.1 Quality Issues

- Completeness: Ensure all transactions have complete information such as product details, quantities, and prices.
- Validity: Validate data types and ranges (e.g., price should be a positive number).
- Accuracy: Cross-check prices and quantities against the product catalog to ensure they match.
- Consistency: Ensure consistent use of measurement units, time formats (e.g., YYYY-MM-DD), and data representations.

2.2 Tidiness Issues

- Each variable forms a column: Ensure each column represents a distinct variable (e.g., product, price, quantity).
- Each observation forms a row: Each row should represent a single transaction.
- Each type of observational unit forms a table: For instance, separate tables for transactions, customer reviews, and inventory.

Data Cleaning

The data cleaning process will involve:

3.1 Fixing Quality Issues

- Handling missing data by filling in or dropping incomplete records.
- Correcting invalid data entries (e.g., negative prices or quantities).
- Ensuring accurate data by comparing against a validated source, like the supermarket's product catalog.

3.2 Fixing Tidiness Issues

- Restructuring the dataset to ensure each variable is in a separate column.
- Combining or splitting tables where necessary to maintain tidy data principles.

Data Storing

Once the dataset is cleaned, it will be saved in a .CSV file or a database format. This ensures the data is well-organized, easily accessible, and prepared for visualization and further analysis.

Data Visualization

With a clean, tidy, and stored dataset, we proceed to visualizations to derive insights:

5.1 Sales Insights

Create visualizations like bar charts, line graphs, and heatmaps to analyze sales trends across different products, times of day, and locations.

5.2 Customer Behavior Insights

Use histograms and scatter plots to examine customer purchasing patterns and gain insights into their preferences and spending habits.

5.3 Inventory Insights

Track inventory levels and detect trends in stock movement with line graphs and area charts to enhance supply chain management.