[supermarket sales]

Data Wrangling Project

BY: ABDELAZIM MOHAMED

Introduction

Real-world sales data rarely comes clean. Using **Python** and its libraries, I gathered data from a supermarket sales dataset, assessed its quality and tidiness, and cleaned it to make it ready for analysis. This process is called **data wrangling**.

I documented the entire wrangling process in a **Colab Notebook** included in the project folder and showcased the results through analyses and visualizations using **Pandas, Matplotlib, and Seaborn**.

The dataset that I wrangled (and analyzed and visualized) contains transactional sales data from a supermarket. It includes details such as **Invoice ID, Branch, City, Customer Type, Gender, Product Line, Unit Price, Quantity, Tax, Total, Date, Time, Payment Method, and Rating**.

The goal of this project is to clean and transform the data, then perform exploratory data analysis (EDA) to answer key business questions, discover sales trends, compare branches and product lines, analyze customer behavior, and ultimately extract insights that can support **data-driven decision-making**.

BY:
ABDELAZIM
MOHAMED

Project index:

1- Data Gathering

    1-1-   **Load supermarket sales data (supermarket_sales.csv)**

    1-2-   **Combine multiple data sources if available (branches, transactions)**

    1-3-   **Prepare raw dataset for assessment**

2- Data Assessing

    2-1-   **Identify quality issues (missing values, duplicates, outliers, wrong data types)**

    2-2-   **Identify tidiness issues (column naming, formatting, multiple variables in one column)**

3- Data Cleaning

    3-1-   **Fix quality issues (handle missing data, remove duplicates, correct data types, replace outliers)**

    3-2-   **Fix tidiness issues (split combined columns, standardize text, restructure data)**

4- Data Visualization

    4-1-   **Sales performance insights (branch performance, product line sales, total revenue trends)**

    4-2-   **Customer insights (customer type contribution, payment method distribution, peak hours)**

    4-3-   **Product & rating insights (average rating by product line, quantity sold)**

Data Gathering

1- File in hand – supermarket_sales.csv

The main dataset used in this project is a transactional sales file
from a supermarket.
It contains detailed information about each transaction, including:
Invoice ID, Branch, City, Customer Type, Gender, Product line, Unit
Price, Quantity, Tax, Total, Date, Time, Payment Method, and Rating.
This file is manually loaded using pandas.read_csv () to create the initial
DataFrame.

2- Data Preparation

Once the raw data is loaded, the following preparation steps are
performed before assessment:
- Converting column names into a consistent format.
- Ensuring correct data types for Date, Time, and numeric columns.
- Creating new derived columns if necessary (extracting month, day
  from Date).

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 concern:

- **Completeness:**
  - Data is preferred to be complete.

- **Validity:**
  - Data must be valid.

- **accuracy:**
  - Data must be accurate.

- **Consistency:**
  - Data must be constant.

2- Tidiness issues

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

- **Each variable forms a column and contains values.**

- **Each observation forms a row.**

- **Each type of observational unit forms a table.**

Data Visualization

After cleaning and transforming the data, we now have a well-structured dataset that is ready for analysis.
Using *Matplotlib* **and** *Seaborn***, several visualizations were created to explore key business questions and extract actionable insights.**
Sales Performance Insights

Visualizations were created to analyze:

- Total sales by branch

- Daily sales trends

- Quantity sold by product line

- Average rating by product line

Customer Insights

Visualizations focused on customer behavior, including:

- Contribution of each customer type to total sales

- Distribution of payment methods

- Peak sales hours throughout the day

Product Insights

Additional plots were created to explore:

- Identification of product lines with the highest contribution to total revenue

These visualizations help decision-makers identify top-performing branches, understand customer preferences, and optimize product offerings.

Thank You

BY:
ABDELAZIM
MOHAMED