[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:

BY:
ABDELAZIM
MOHAMED

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 Cleaning

Fix quality issues

- **Handle missing values in numerical and categorical columns using appropriate methods (imputation, removal).**
- **Remove duplicate records to avoid bias in the analysis.**
- **Correct inconsistent or incorrect data types (converting dates and times to proper datetime formats, converting numerical columns stored as strings).**
- **Detect and replace outliers in numerical columns (replacing extreme values in Rating or Total with the mean or median).**

Fix tidiness issues

- **Split combined columns into separate variables (separating contact information into phone number and email).**
- **Standardize text fields (converting names or categories to lowercase for consistency).**
- **Restructure the dataset into a tidy format where each variable has its own column, each observation has its own row, and each type of observational unit forms a separate table.**

Data Storing

After gathering, Assessing and cleaning data, we need to store the data into a .CSV file to make it easy for access. You can find it in the project folder.

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