Data Science Final Project

Introduction

The primary goal of this project is to analyse and enhance retail sales performance by Looking into the insights derived from the data. Through a structured approach that involves **data processing**, **cleaning**, and **exploration**, the project aims to uncover patterns and trends that can inform strategic decisions. By applying a basic machine learning model, the analysis not only supports data-driven insights but also demonstrates the potential of predictive analytics in optimizing retail operations.

Data Description

The dataset used in this project consists of transactional retail data collected from various cities. It includes key information such as Transaction ID, Product Category, Store Location, Unit Price, Quantity Sold, Total Sale, Payment Method, Customer Age, and Refunded status. These features provide a comprehensive view of sales activity and customer behaviour, forming the foundation for in-depth analysis and the development of a predictive model aimed at improving retail sales performance.

Exploratory Data Analysis

We worked with a small dataset of 200 retail transactions from five cities in Egypt: Giza, Cairo, Asyut, Mansoura, and Alexandria. Each transaction includes details like product type, store location, price, quantity, customer age, payment method, and if the item was refunded.

Key Findings

- Total sales before refunds: $3.2 million

- Refunded amount: $310,177

- Final sales after refunds: $2.9 million

- Total items sold: 1,176

Top Cities

- Giza had the most transactions and the highest sales: $787,289

- Mansoura and Alexandria followed with strong sales.

- Cairo and Asyut had lower sales, but still made a good contribution.

Best-Selling Products

- Furniture made the most money: $813,032

- Groceries had the most units sold: 254 items

- Toys, Electronics, and Clothing also performed well.

Customer & Payment Insights

- Older customers (around 47) bought more furniture.

- Younger customers (around 38) mostly bought groceries.

- Bank Transfers were the most used payment method.

- Mobile payments were most popular in Mansoura.

- Cash and credit cards were also used often.

Refunds

Only 16 transactions were refunded, but they had a big impact, cutting down total revenue by about 10%.

Handling Missing Data

To handle the missing data, we first tried dropping the rows that had missing values, but that didn't work well since it reduced the dataset to just 41 rows---only about 20% of the original data. So instead, we replaced the missing numeric values with their average, and the categorical values with their mode (the most frequent value). After that, we rechecked the data and explored the insights again.

Machine Learning

In the machine learning part, we used 80% of the data for training and 20% for testing. The goal was to predict the payment method based on the product category and the total sale amount. This helped us see if we could find patterns in how customers choose to pay, depending on what they buy and how much they spend.

Comparison

We explored both the complete dataset and the version with missing values, and after analyzing and comparing the insights from each, we reached the following conclusions:

1. Replacing numerical data with their average impacts several key metrics, including total sales, units sold, number of refunds, and total refund amount.

2. Replacing categorical data with their mode (i.e., the most frequent value) influences the number of transactions per city, the count of payment methods, and the distribution of payment methods across cities.

3. Machine learning models generate different predictions when trained on the complete data versus the imputed data.

Payment Method	**Complete Data Prediction**	**Missing Data Prediction**
Bank Transfer	28	33
Mobile Payment	6	4
Credit Card	5	3
Cash	1	0

Conclusion

After conducting a thorough analysis, we concluded that retail sales operations can be improved by leveraging data-driven trends. Based on the data, we recommend opening new branches in Giza, Mansoura, and Alexandria. Additionally, forming partnerships with banks to offer special promotions could help attract more customers. From a data perspective, we found that handling missing values appropriately can produce results that closely resemble those obtained from complete datasets