Data Analysis

Report

Oludare Fasure

Contents

Introduction

The retail industry is one of the largest industries globally, making significant contributions to the GDP of many countries. In the United States of America, the retail industry is said to account for a third of all consumer spendings, indicating the important role this sector plays in driving economic activities.

The infusion of technology into this sector has resulted in the rise of several e-commerce giants. While companies like Amazon, eBay and Alibaba have built large retail businesses focusing primarily online, there are several other businesses that have maintained a combination of both physical and online marketplaces.

Amongst other things, one of the major advantages the introduction of technology into the retail sector has ushered in is the amount of data that is available for making business decisions. The amount of data that can be collected and processed within a short period can help business make informed decisions that help improve their profitability.

In this analysis, the dataset is a collection of retail data from various retailers that was got from Kaggle. It exemplifies the complexities of real life data in the retail industry, and it is intended to be used for training. There are 200000 rows and 19 columns of data, and there are no missing values.

DATA CLEANING

Upon inspection of this dataset, it was observed to be quite clean, which is a deviation from what is obtainable in many real life data. In many cases, it is quite common to see datasets with missing values, outliers, data wrongly formatted, etc. However, in this instance, the dataset appears to have the right data in the right places.

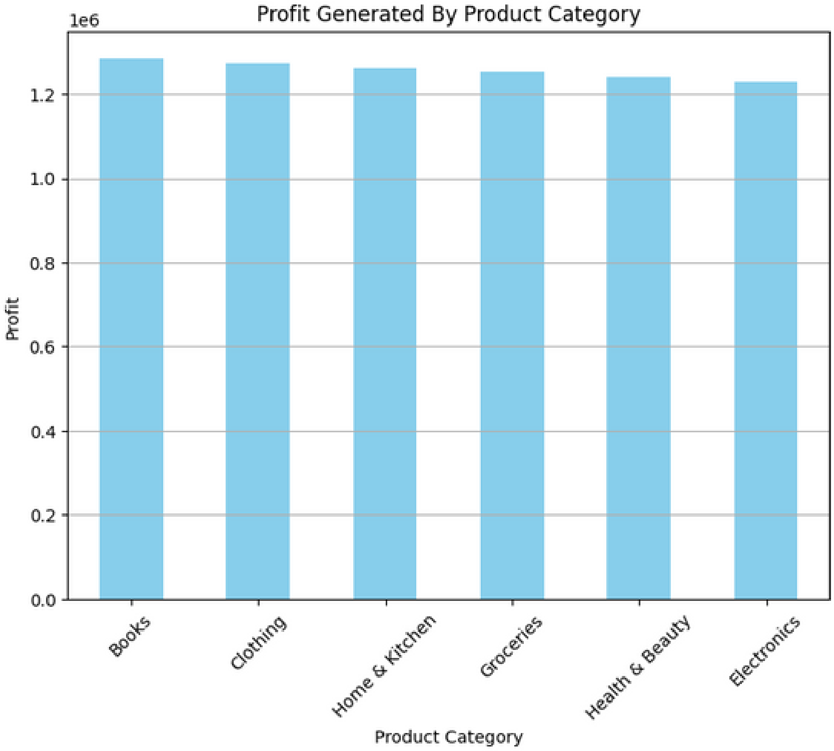
EDA

From the exploratory data analysis, I discovered that the total number of stores considered in the dataset is 100, located in 10 different states across the United States. All products sold were classified into 6 categories, and the data covered a two-year period.

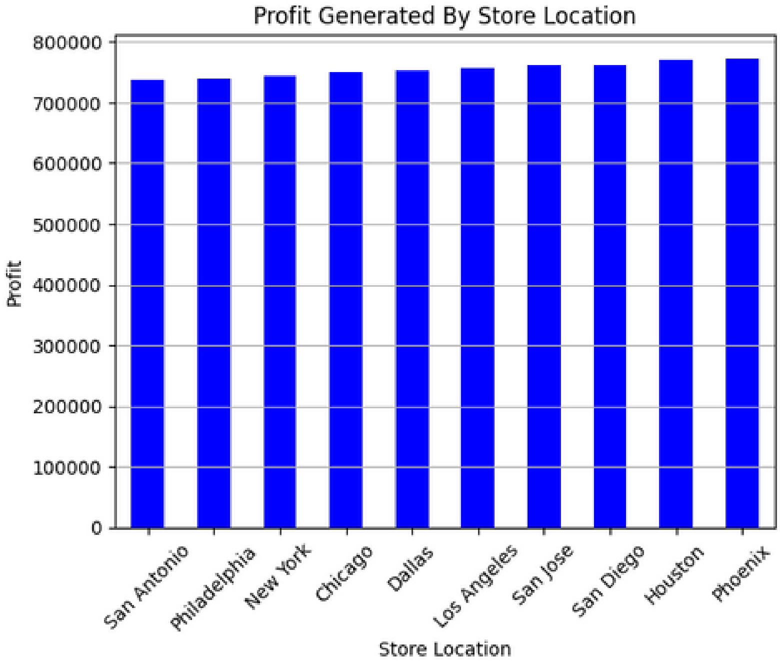
VISUALISATION

The following relationships were considered using different kinds of plots:

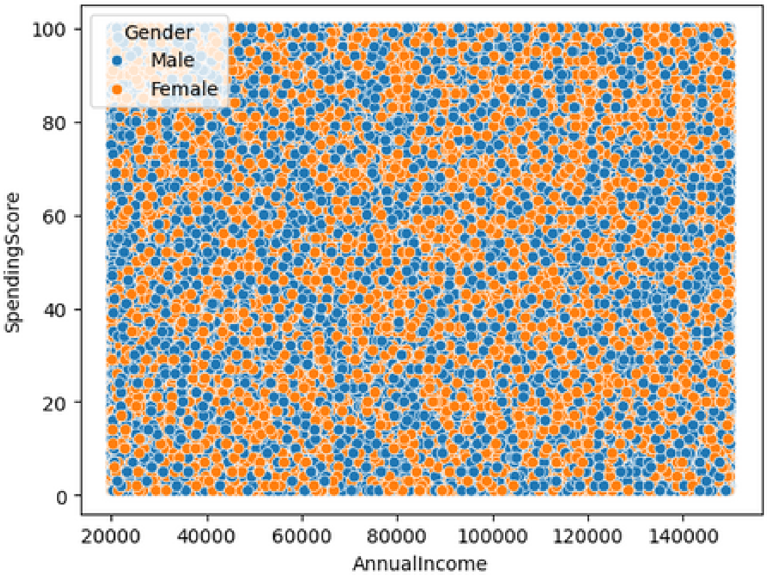
- Best performing product category



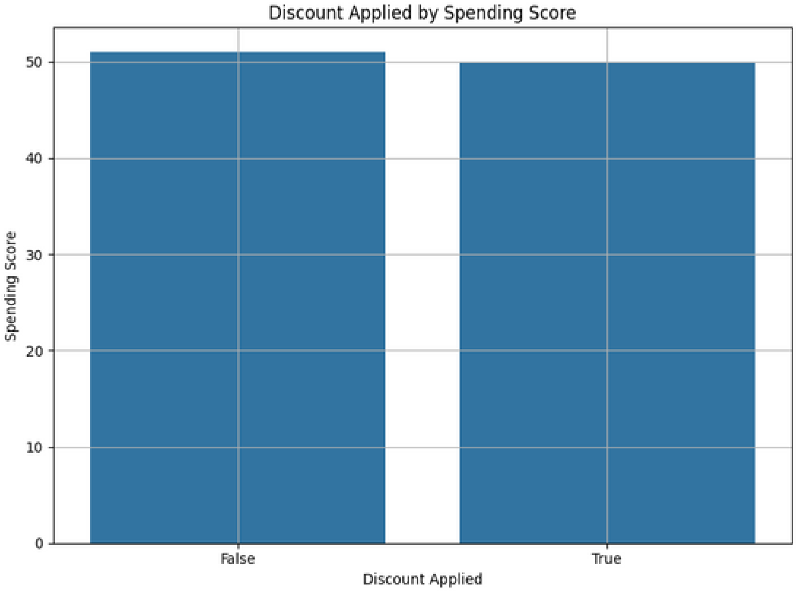
- The most profitable store location



- The spending pattern according to gender



- Influence of discounts on spending score



- Profit trend over time across all store locations
- The relationship between profit and inventory levels
- The correlation matrix
- Distribution of spending score by age

INSIGHTS

- According to our data, books is the best performing product category, while health and beauty is the least performing category. However, the difference between the performance across all categories is almost evenly distributed, as the difference between the best and least performing category is 307 transactions.
- In regard to profit generation, books remains the best performing category, while electronics is the least performing product category.
- Based on both insights above, the profit margin across is not the same across all categories. Hence, the fact that a product sold more quantities doesn't mean it generated more profit in comparison with another category.
- While all store locations generated above $700,000 in profit, Phoenix generated the most profit, while San Antonio generated the least profit.
- Interestingly, discount did not significantly impart customer spending. Customers spent more on products without discount.

SUGGESTIONS

- The market basket analysis would be an insightful area of study in future analysis. This would aid the understanding of the products that are bought together. That way, we know if the discount is being used effectively.
- The location population and see how that influences store performances
- Gather information on when the store was opened and if that has any impact on customer traffic in the store.