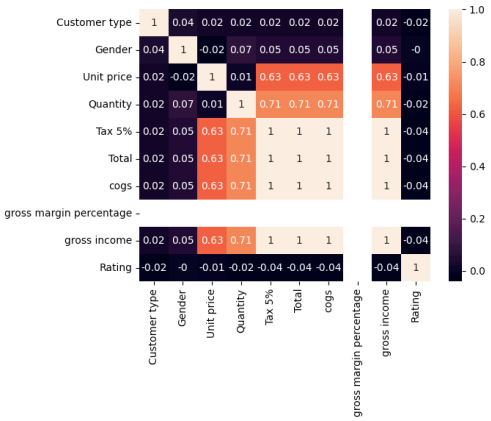
IST 707 Applied Machine Learning
Assignment 2
Kabir Thakur - 427482939

Report

Introduction

The data given to us shows the invoices of 3 supermarket stores in Myanmar, Yangon, Naypyitaw / Naypyidaw, Mandalay referred to as A,B and C respectively for the rest of the report. This data been collected between 1st January 2019 and 3rd March 2019.

The first step in our analysis will be to clean this data, removing the NA values and performing the necessary transformations. I have covered analysis based on gender and time in this report.



The heatmap above represents the correlation of the dataset. We can see that the tax, total, cogs and gross income are completely correlated with each other, this is because they are derived quantities. Gross margin percentage is just a constant value throughout.

Data cleaning

The data has no NA or missing values that need to be accounted for and appears to be clean. Few unnecessary columns like Invoice ID and city name were dropped. We can also drop total, cogs, and gross margin percentage for the rest of the analysis and just keep gross income.

Transformations

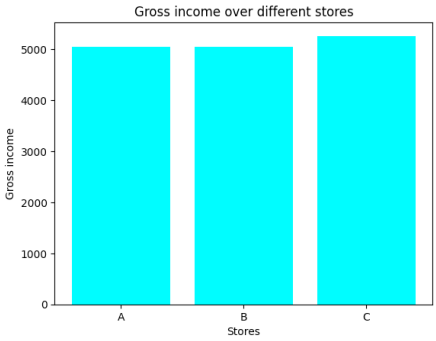
The date and time had been read as ‘object’ data type and have been converted to their proper format. The columns gender and customer type have been converted to binary variables. For gender 0 represents male and 1 represents female. For customer type 0 represents non-member while 1 represents members.

Exploratory Data Analysis

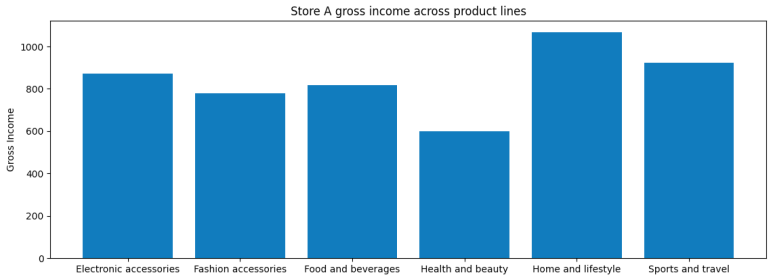
Store Based Analysis

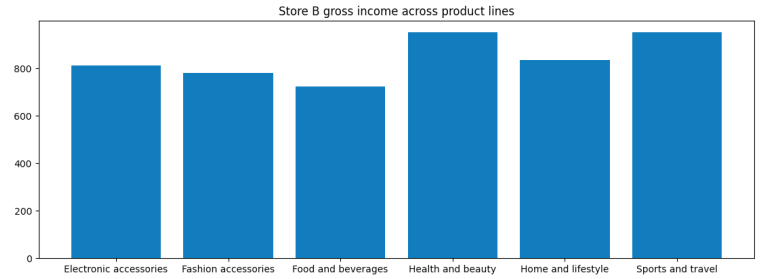
There are a 1000 invoices in the data frame. Store A has the most with 340 invoices followed by store B with 332 and store C has the least invoices 328. The difference however is not too significant.

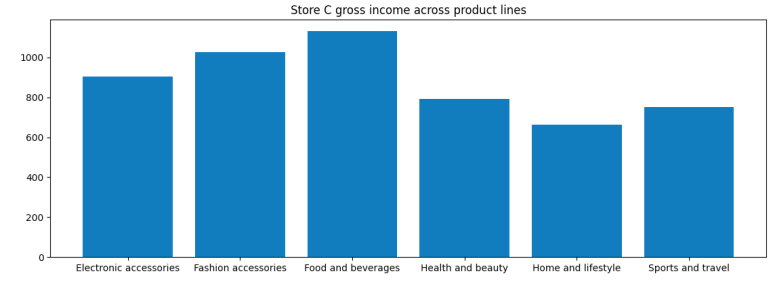
The average rating of customer experience across all stores is 6.97/10. This can be used as a metric to identify which stores are performing best in terms of customer satisfaction. Store A has an average rating of 7.02, store B is 6.81 and store C is 7.07. So we can see that store A is the best followed by store C and then store B.

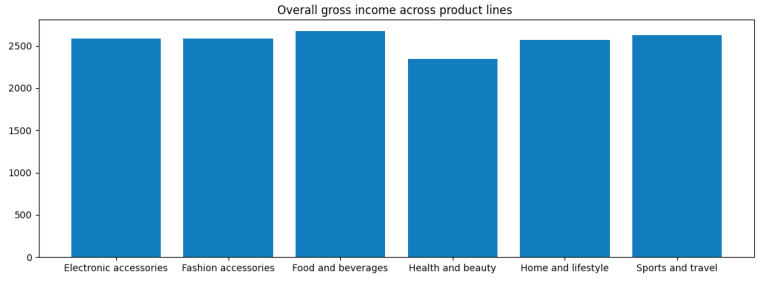


We can see that Store A sells the least health and beauty products and Store B sells the most. Health and beauty products are the least sold category overall. Home and lifestyle is the least sold product category at store C and the most sold product category at store A. Overall food and beverages sells the most. The number of invoices show that store C is the least popular but the gross income graph shows store C is making the most money.



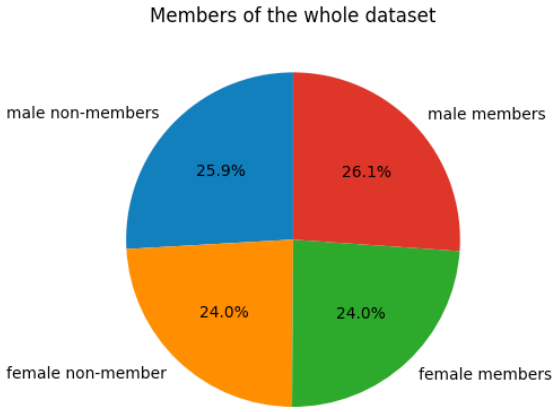






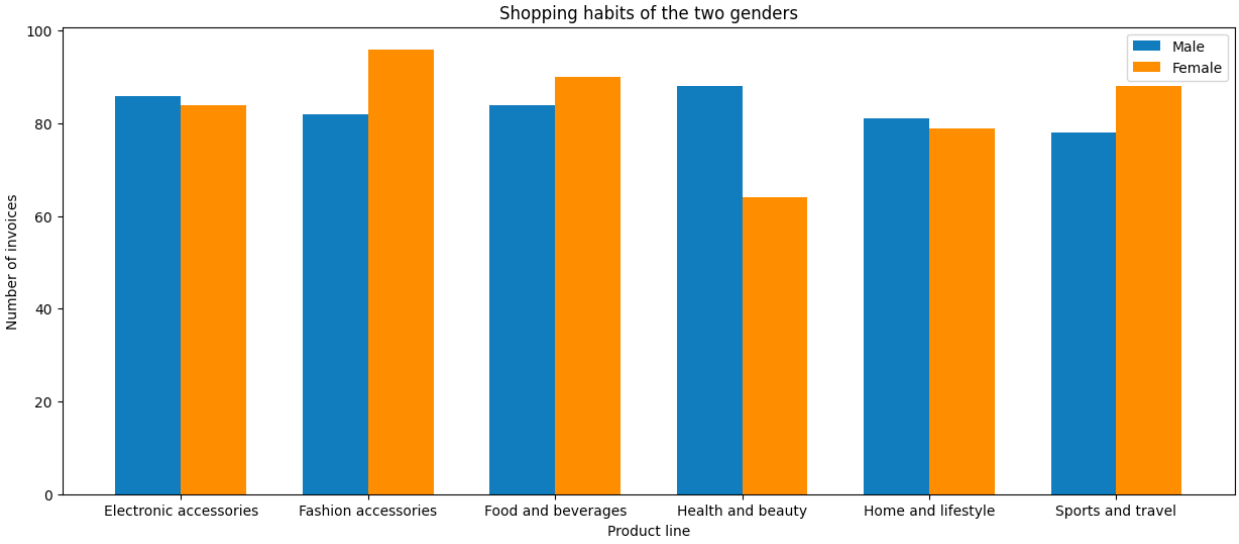
Gender Based Analysis

There 501 females shopping across all three stores and 499 males so the dataset is spread even among shoppers of both genders. The pie chart shows the distribution of male and female shoppers who are members/non-members of the supermarket chain.



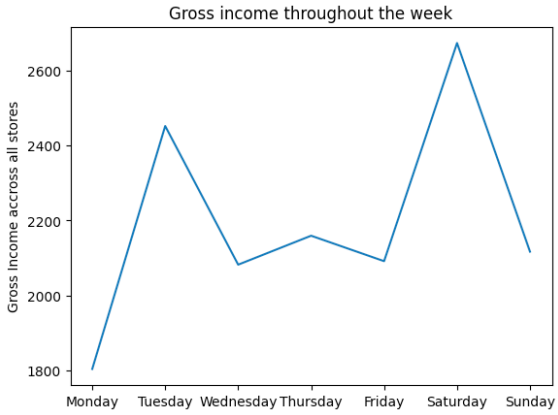
The grouped bar graph shows the total number of purchases made by the two genders on each product line. Key insights here are

1. Men spend more on health and beauty
2. Men spend more on health and lifestyle
3. Women spend more on sports and travel

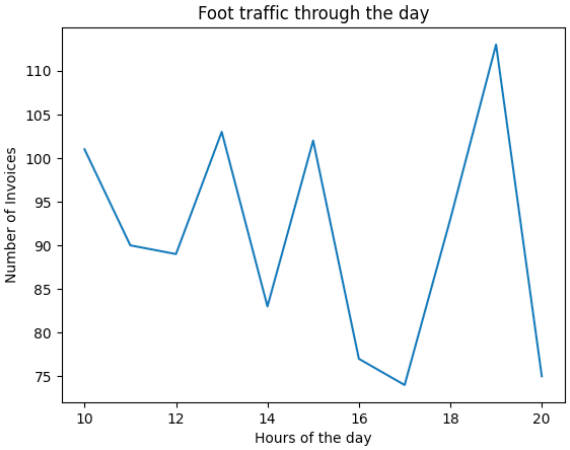


Analysis based on time

The dataset spreads 3 months in duration. From this we can do several kinds of analysis.

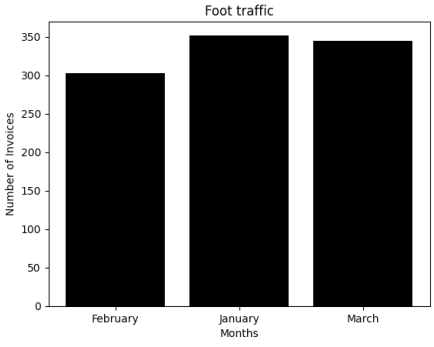


We can see that based on days of week, all stores see peak traffic on Saturdays after which traffic falls to it's lowest till Tuesday when it gets back up.



From the above plot we can see that the stores run from 10am to 8pm. The foot traffic in the stores dies down right before it reaches its peak. The peak operation hours for the stores is between 5 and 8 PM.

The bar graph on the right shows that Feburary saw this least amount of traffic between the three months and Janurary saw the most.



Conclusion

In conclusion we saw that store C is the highest rated store among the three and makes the most profit while store A and B need improvement.

Men spent more on health, beauty and lifestyle than women. So one can introduce more products in this category for men to increase profit. Women spent more on sports and travel.

On Tuesdays and Saturdays the stores see the most traffic between the hours of 5 and 8PM.

January was the most profitable month for all stores and February was the least profitable.