Analyzing the sales trends in supermarket data

There is an increase in the number of supermarkets in cities. In this experiment, we analyze the dataset of historical sales data of a supermarket chain that has stores in 3 cities. The data has been collected over a period of three months, and are available in the public domain in https://www.kaggle.com/aungpyaeap/supermarket-sales. Analyzing such datasets can be fruitful for supermarkets to understand the sales patterns of the supermarkets in order to anticipate and prepare for future sales.

We start by downloading and importing the dataset from the public domain. Then we perform some basic Exploratory Data Analysis on the dataset to understand the columns on the dataset. We see that the dataset has the following columns -

- Invoice ID
- Branch
- City
- Customer type
- Gender
- Product line
- Unit price
- Quantity
- Tax 5%
- Total
- Date
- Time
- Payment cogs
- gross margin percentage
- gross income
- Rating

We see that all the columns in the dataset have multiple unique data embedded in it except the “gross margin percentage”, which has 1 unique value across all the rows. We would not consider the column for further analysis because of the lack of variation in the data.

We, then, check the number of null values in the data. There are no rows with any null values in any column. This implies that the POS software used across the supermarket chains has recorded the data consistently.

We will start by checking the count of entries that have been recorded in each of the three supermarket chains.

Fig.1 shows that the distribution of sales in the three stores is roughly the same. This shows that all the cities that the supermarket serves are of equivalent importance. There is no reason to assume that the distribution of sales of products in the three cities would be different.

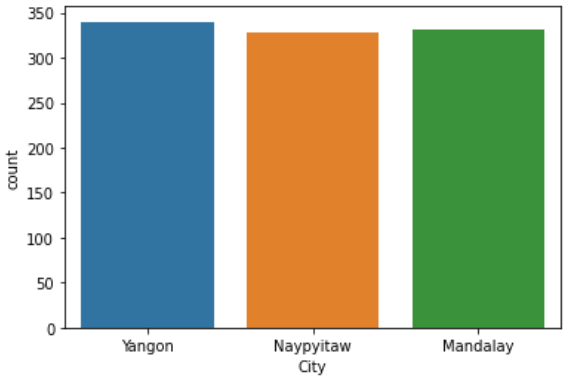

Fig 1: Count of sales in the 3 cities

However, we would look into the distribution of products in segmented across the three cities. Fig 2. shows the categorical plot segmented by cities. We see that, contrary to our assumption, there is a variation in the

category-wise sales of items across the three cities.

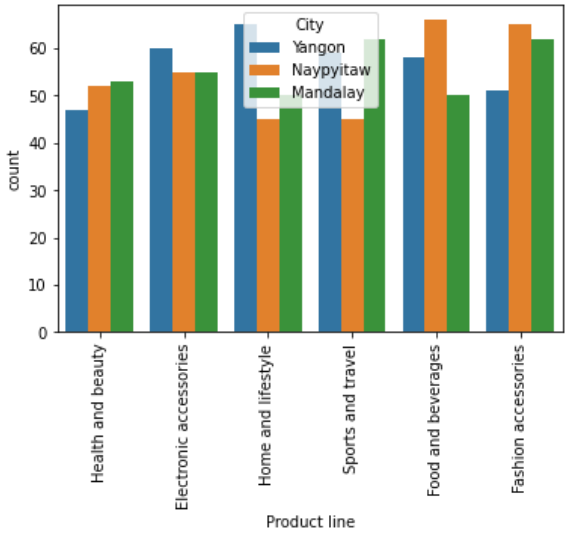

Fig 2: Categorical sales by cities

We see that although the cities had roughly the same amount of sales, there is a distinctive difference between how the categories of products in the different cities differ. For example, we see that there is a far lower sales of ‘Health and beauty’ products in Yangon compared to the city of Mandalay. On the other hand, the sales of ‘Home and lifestyle’ is highest in Yangon, followed by Mandalay and with a significantly less amount of sales in Naypyitaw.

A domain expert who understands the socio-political scenarios of the three cities may be able to understand the reasons for these trends. We can use this analysis to decide which products must be made available in which city in a higher quantity.

We would finally try to understand if there is a correlation between the different attributes that are recorded. We see that Unit Price, Quantity, Gross Income, and Rating are the numerical data in the dataset.

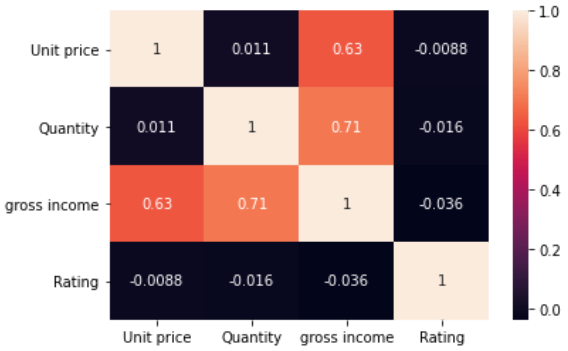

Fig 3: Heatmap showing correlation

We see that there is a high correlation between Unit Price and Gross Income. Similarly, there is a high correlation between Quantity and Gross Income. We can use this information to estimate that the items which have a high price and are usually sold in a higher quantity would result in a higher gross income. This follows from basic arithmetic, and unfortunately, without learning more about the precise items and their sale patterns, not much analysis could be made into this matter.

Reference -

https://www.kaggle.com/aungpyaeap/supermarket-sales