Q4 Recommendation and Business Analysis

To get a better and comprehensive insight into the retail supermarket, we should first do some Data Exploration to provide a summary of the dataset, including its basic statistics.

Data Exploration

From the table and the summary information, we can find that there are some missing values in the dataset.

	Transaction_id	Product_id	Description	Quantity	Date	Price	Customer ID
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12.0	2009/12/1 7:45	6.95	13085.0
1	489434	79323P	PINK CHERRY LIGHTS	12.0	2009/12/1 7:45	6.75	13085.0
2	489434	79323W	WHITE CHERRY LIGHTS	12.0	2009/12/1 7:45	6.75	13085.0
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48.0	2009/12/1 7:45	2.10	13085.0
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24.0	2009/12/1 7:45	1.25	13085.0
...	...	...	...	...	...	...	...
1048570	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1048571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1048572	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1048573	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1048574	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 1 Overview of the data

We simply drop all the samples with missing values and get a complete dataset. And we can also find there are some outliers in the data, for example: the quantity and the price should not have negative value.

	Quantity	Price	Customer ID
count	1000000.000000	1000000.000000	774502.000000
mean	10.053332	4.669050	15325.209913
std	136.005474	125.621421	1694.663838
min	-74215.000000	-53594.360000	12346.000000
25%	1.000000	1.250000	13969.000000
50%	3.000000	2.100000	15262.000000
75%	10.000000	4.150000	16794.000000
max	74215.000000	38970.000000	18287.000000

Figure 2 Outlier Detection

After remove all the anomalies, we can get a normal one as below:

	Quantity	Price	Customer ID
count	756531.000000	756531.000000	756531.000000
mean	13.378881	3.231976	15332.616078
std	115.017471	29.972501	1693.957022
min	1.000000	0.001000	12346.000000
25%	2.000000	1.250000	13979.000000
50%	5.000000	1.950000	15272.000000
75%	12.000000	3.750000	16798.000000
max	74215.000000	10953.500000	18287.000000

Figure 3 Normal Data

Visualization and BA

Customer Side

For a supermarket or any business, customer loyalty is essentially important. We can analysis the relationship of number of transactions per customer and their corresponding frequency and draw a bar plot.

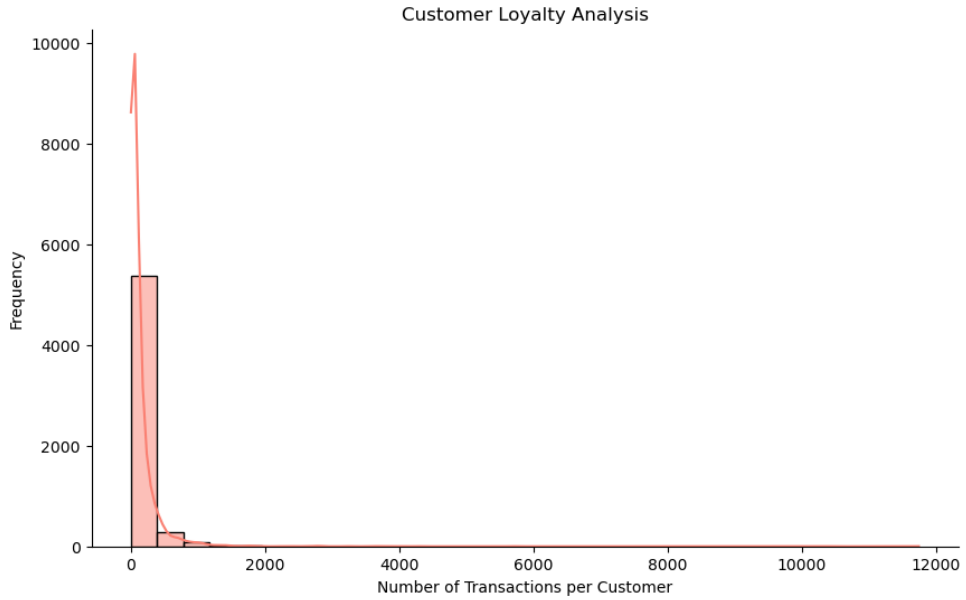

Figure 4 Customer Loyalty

Repeat customer and big spenders are also very important to a supermarket because they can contribute a significant portion of the profits.

We analysis top10 customers by transaction amount and number of repeat purchases and give the plots as below:

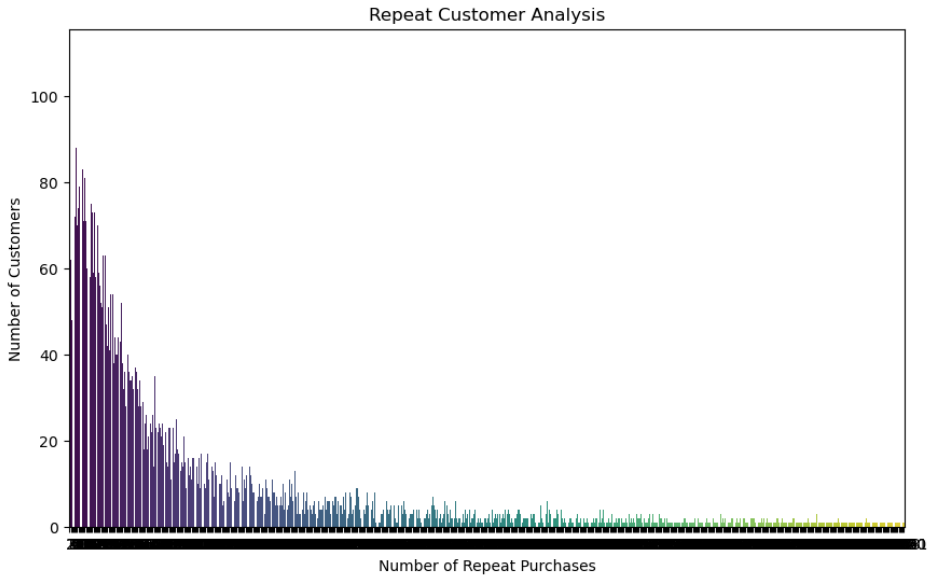

Figure 5 Repeat Customer Analysis

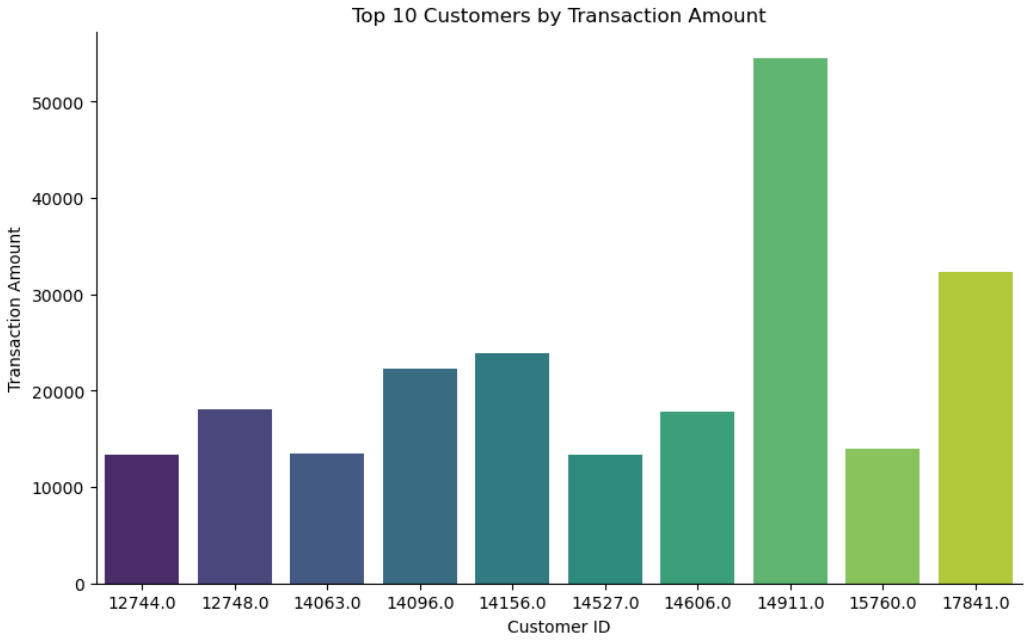

Figure 6 Top10 customers by transaction amount

Product side

Products are the key to the success of a business, supermarkets are no exception, how to introduce goods according to the sales volume and customer preferences is also an important factor in the success of running a good supermarket.

Therefore, we analyze the product metrics from different perspectives as follows:

Firstly, we analyze the distribution of the product prices and find that most of the product price are below 100, which consistent with our common sense.

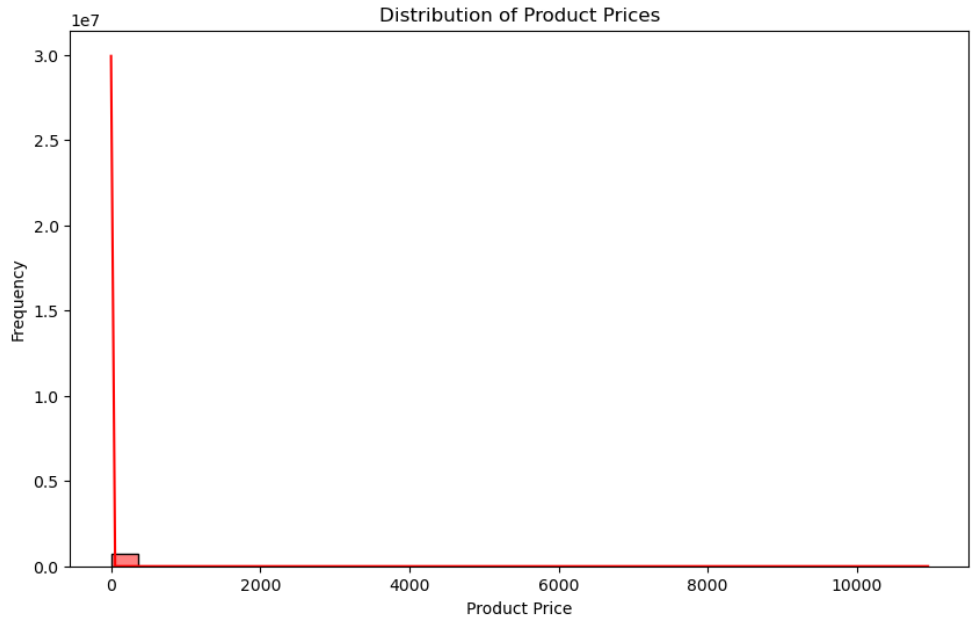

Figure 7 Distribution of Product Price

We also analyze the best selling products by quantity and their contribution to the sale

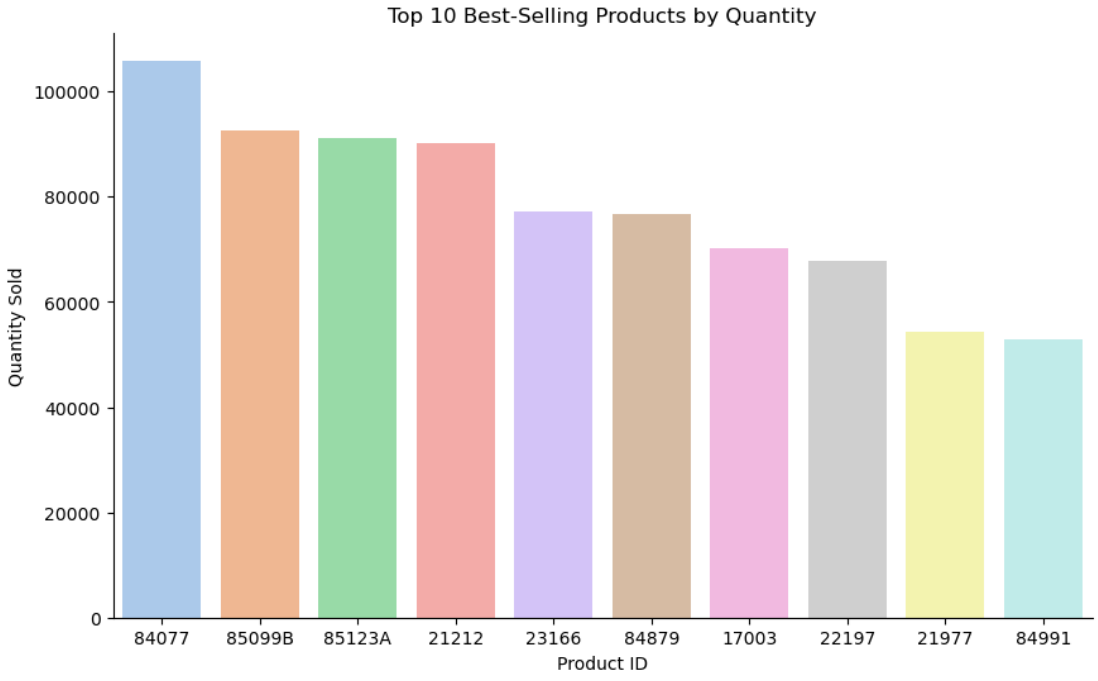

Figure 8 Best-selling products

The best selling products is product with id: 84077. But product 84077 is cheap, it is not the most major contribution to the total sale.

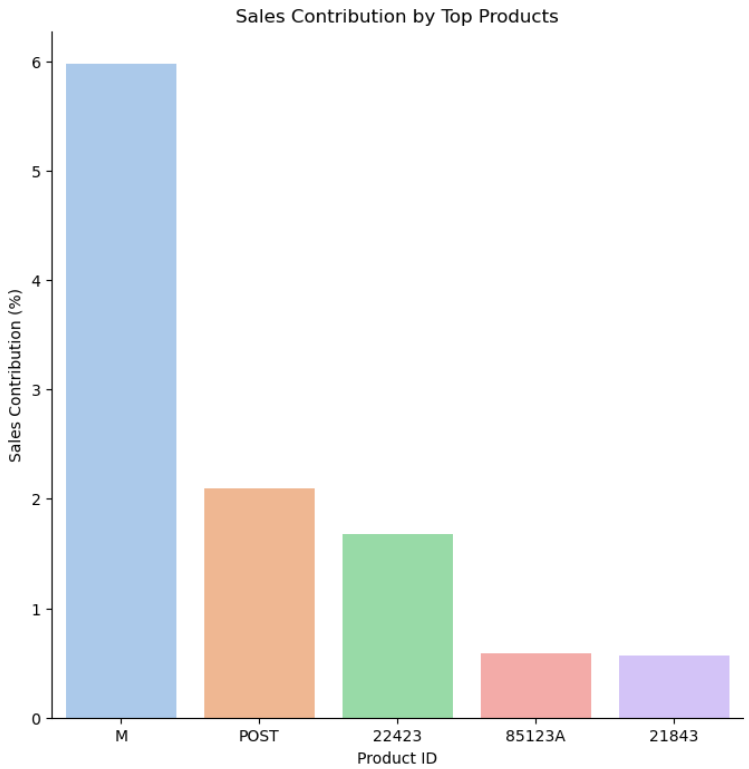

Figure 9 Sales Contribution by Top Products

The sales volume of a product is also related to the season, and we analyzed the relationship between some products and the season.

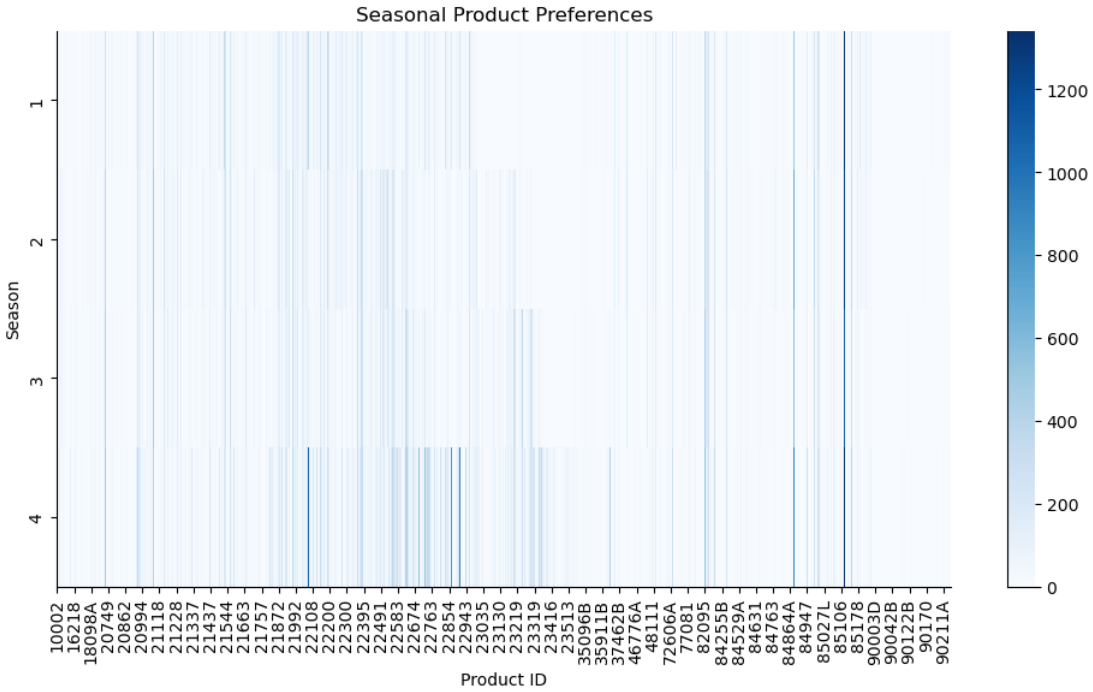

Figure 10 Seasonal Product Preferences

We can see products like 22108 is higly related to the season which means we don’t need to introduce it in the other seasons.

Sales and Price

Here we show the relationship between the quantity of a good purchased and its price

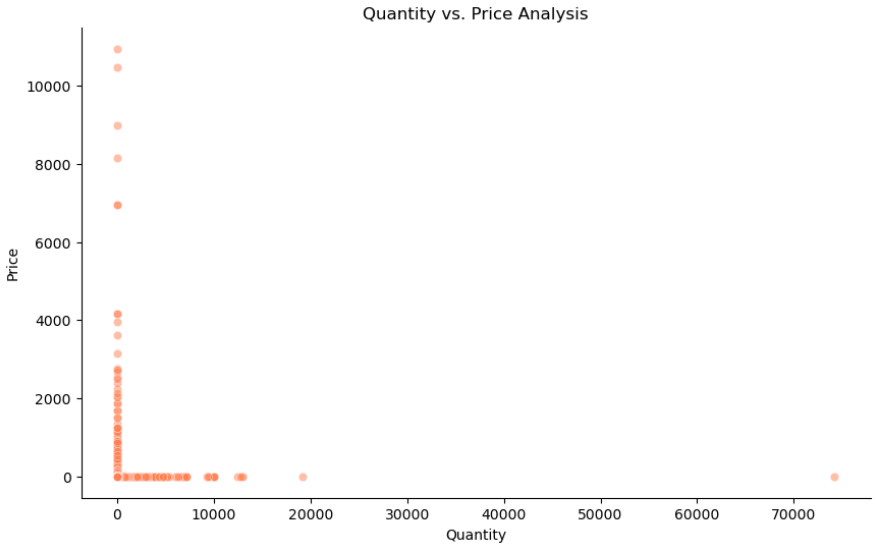

Figure 11 Quantity and Price Analysis

It can be seen that items purchased in high volumes are usually less expensive, items purchased in low volumes are more expensive, and there are also items purchased in low volumes that are less expensive, and these items are usually not essential for everyday life

We also show how this store's sales relate to the time of day as below:

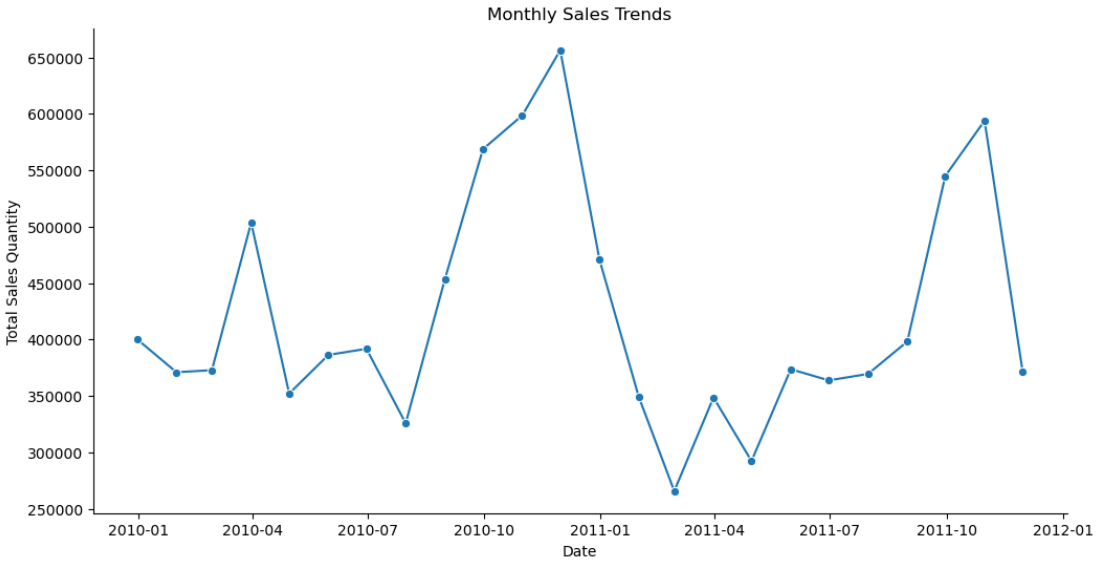

Figure 12 Monthly Sales Trends

In 2010-10 to 2011-3, there were large fluctuations in sales

Association Rule Analysis

In this task, I use FP-Tree to analyze the transaction sets. Unlike Apriori, FP-Growth does not generate candidate itemsets explicitly, which reduces the number of passes over the data. And also FP-Growth is known for its efficiency, especially in scenarios with large datasets and a large number of frequent itemsets. This is really important and let me choose this algorithm. With the **minimum support 0.01**, we can get the result as below:

```
(bioinfo) C:\Users\lenovo\Desktop\HKUSTGZ-PG\Course-project\DSAA-5002\Final-Project>D:/Program/Anaconda3/envs/bioinfo/python.exe c:/Users/lenovo/Desktop/HKUS
TGZ-PG/Course-project/DSAA-5002/Final-Project/src/Q4/main.py
     support  itemsets
41  0.104429  (85123A)
374 0.080496   (22423)
56  0.076746  (85099B)
80  0.059783   (21212)
57  0.059135   (20725)
6   0.052735   (84879)
224 0.051655   (47566)
0   0.049947   (21232)
337 0.045667   (22383)
111 0.044725   (21931)
       antecedents     consequents   support  confidence       lift
534        (22745)         (22748)  0.010268    0.794833  58.250382
535        (22748)         (22745)  0.010268    0.752518  58.250382
570        (21239)         (21240)  0.010091    0.699320  47.053040
571        (21240)         (21239)  0.010091    0.678996  47.053040
228        (21094)         (21086)  0.012506    0.658058  40.480095
229        (21086)         (21094)  0.012506    0.769324  40.480095
173  (22699, 22697)         (22698)  0.013606    0.700708  35.512289
176        (22698)  (22699, 22697)  0.013606    0.689552  35.512289
172  (22699, 22698)         (22697)  0.013606    0.896507  34.384558
177        (22697)  (22699, 22698)  0.013606    0.521837  34.384558
========================================================================
```

When arranging items, these items can be placed together:

- (22745) (22748)
- (21239) (21240)
- (21094) (21086)
- (22699, 22697) (22698)

At the same time, introduce more of the following products:

(85123A, 22423, 85009B, 21212)