SMU

SINGAPORE MANAGEMENT UNIVERSITY

ISSS621 Data Science for Business

Group 3 Project Report

ASM

AG

MHS

WK

PRA

YY

Contents

1 Introduction and Problem Statement

This study would analyse a supermarket chain with outlets in Yangon, Mandalay, and Naypyitaw in Myanmar. This supermarket chain offers a total of six distinct product categories, including fashion accessories, food and beverages, electronic accessories, sports and travel, home and lifestyle, and health and beauty. Customers may pay with cash, a credit card, or an electronic wallet, and they can also join up for membership programmes; 50% of all transactions are completed by registered members.

We will analyse the existing data and formulate strategies for customer retention and revenue growth based on the following scope:

i. Revenue from customers

Understanding the customer and ensuring that the store is able to fulfil their needs and wants is one way to cut down on income loss. The supermarket will be able to modify the inventory and product placement in the various branches in order to guarantee a match after it has gained an understanding of the different customer segments.

ii. Revenue from payment mode

We would want to learn more about consumer purchasing behaviour, particularly the preferred mode of payment. After obtaining the knowledge, we would explore collaborations with banks to promote new payment modes or urge consumers who use cash to new payment modes, allowing the supermarket market to leverage and promote customer purchase.

iii. Retention and upselling

Today's customers have an abundance of choice; hence businesses strive to convert a first-time customer to a repeat customer and grow the wallet share spent. For example, the supermarket in this project has a membership programme, through which they can surprise and impress customers by extending membership benefit and discounts. Our study aims to recommend potential gaps in customer retention, and how the supermarket can plug these gaps to maximize or grow wallet share.

2 Data Sources

Dataset used in project

Our project uses supermarket sales data from Kaggle (Pyae, 2019). This dataset has 3 months of transaction data (January 2019 to March 2019) from three branches of the supermarket.

Description of variables used from this dataset:

Column Name	Description
Invoice ID	Computer generated– unique identifier
Branch	A - supermarket branch in city Yangon B - supermarket branch in city Mandalay C - supermarket branch in city Naypyitaw
City	Location of supermarket branch
Customer type	Member (with member card) Normal (without member card)
Gender	Male/Female
Product line	6 product lines- Electronic accessories, Health and beauty, Sports and travel, Home and lifestyle, Food and beverages, Fashion accessories
Unit price	Unit price of product in $
Quantity	Number of products purchased
Total	Total price
Date	Date of purchase
Time	Time of purchase
Payment method	cash/credit card/e-wallet
Rating	From 1 to 10 - Provided by customer on overall shopping experience

Other data that would be useful for a complete data ecosystem

The current dataset doesn't contain relevant columns such as Customer membership ID, product names in the transactional data as well as receipt number which could be used to perform market basket analysis. Also, if the Customer membership ID were to be captured, we could map it to additional customer information such as age stored in another table (using Customer membership ID as the foreign key). This mapping would be helpful in segmenting the customers for analysis.

Other data sources that could be relevant to achieve the business aim of increasing revenue include:

- Data from promotional campaigns or events
- External data from social media or third-party surveys to find out information on popular brands
- Competitor analysis (publicly available competitor data – analysts' reports, annual reports as well as data from credit card companies)
- Economic and population data: disposable income and population trend

3 Analysis and outcome

Analytical framework:

The two main objectives of the supermarket are to maximize revenue, maintain customer retention, and aim for higher customer acquisition numbers by ensuring customer satisfaction.

Descriptive analysis:

For this sub-topic, we focused our attention on product line analysis. The supermarket has 6 product lines, as detailed in the data sources. The aim is to identify what product line does better in terms of ratings, gender preference, by city, quantity sold and total sales. Such insights are valuable to making strategic, marketing, and analytical decisions.

Predictive analysis:

We performed linear and logistic regression, and clustering analysis. Regression is imperative to predict revenue indicators. Clustering analysis is extremely significant for us to comprehend what segments need more attention in which arena so that customized marketing, advertising, and other strategies can be leveraged to increase revenue and customer acquisition.

a. Confirmatory Data Analysis

The purpose of this confirmatory data analysis is to conduct correlation analysis between the desired variables in our dataset. This is a preliminary step that must be performed before moving on to more advanced forms of analysis such as grouping and regression. We will classify the desired variables in our dataset under two categories as shown below:

Categorical variables: City, Customer type, Gender, Product line, Payment

Numerical variables: Total, Rating, Tax 5%

i. Correlation between categorical variables: Chi square test

H_0: The 2 categorical variables are correlated.

H_1: The 2 categorical variables are not correlated.

Alpha level: 0.05

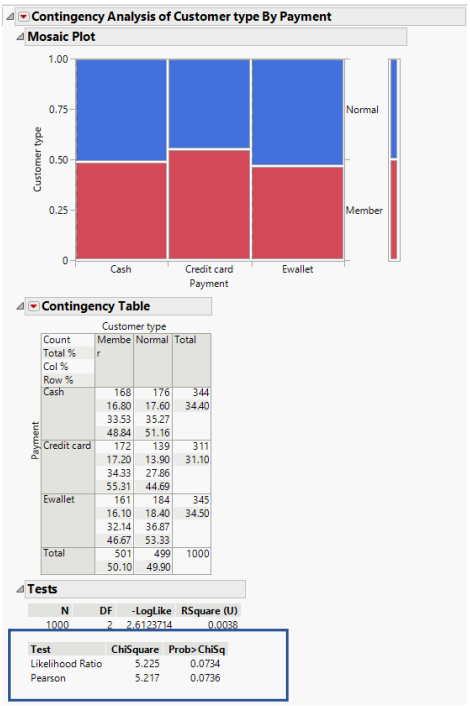


Figure 1: Chi Square test of Customer Type by Payment using JMP Pro

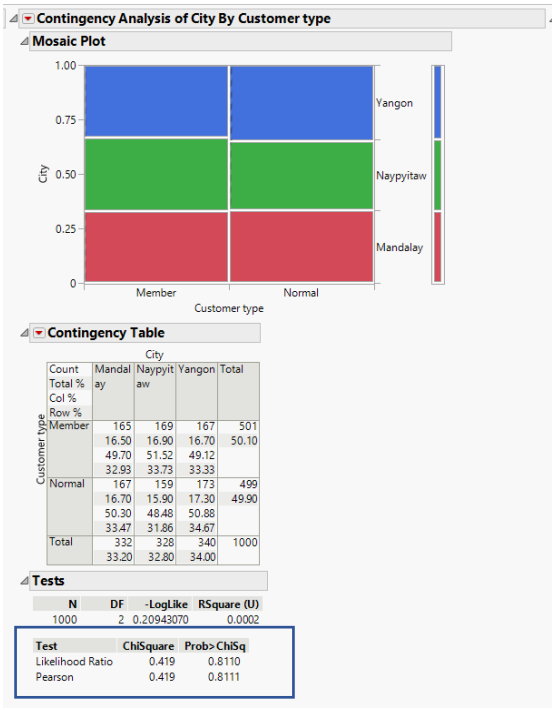


Figure 2: Chi Square test of City by Customer Type using JMP Pro

The p values for the chi-squared tests performed on all categorical pairs contained within the dataset range from the lowest value of 0.07 (relating to customer type and payment) to the highest value of 0.81 (relating to city and customer type). As a result, we are unable to reject the null hypothesis that all pairs of categorical variables within the dataset are correlated.

ii. <u>Correlation between the categorical variables and numerical variables:</u> Chi Square Test

Alpha level: 0.05

H_0: The pair of categorial and numerical variables are correlated.

H_1: The pair of categorial and numerical variables are not correlated.

Figure 3 Chi Square test of City by Rating

Chi square tests are also performed on all the numerical variables and on each of the categorical variables. The diagram that is displayed above shows a pair of categorical and numerical variables called "city and rating." Out of all the other pairs, this pair has produced the result with the lowest p value, which is more than 0.05. Therefore, we are unable to reject the null hypothesis or rule out the possibility that both the categorical and numerical variables are correlated. The same is true for the other pairs.

iii. Correlation between the numerical variables: Multivariate Analysis

Pairwise Correlations

Variable	by Variable	Correlation
Total	Tax 5%	1.0000
Tax 5%	Quantity	0.7055
Total	Quantity	0.7055
Tax 5%	Unit price	0.6340
Total	Unit price	0.6340
Quantity	Unit price	0.0108
Time	Unit price	0.0076
Time	Tax 5%	-0.0052
Time	Total	-0.0052
Time	Quantity	-0.0086
Rating	Unit price	-0.0088
Rating	Quantity	-0.0158
Rating	Time	-0.0261
Rating	Tax 5%	-0.0364
Rating	Total	-0.0364

Table 4: Pairwise Correlation Pre-removal using JMP Pro

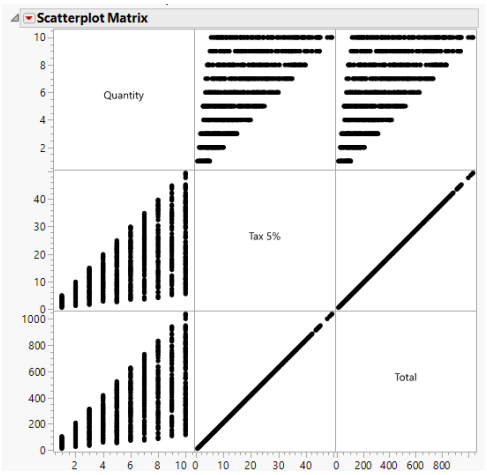


Figure 5: Scatterplot Matrix of selected variables using JMP Pro

As displayed in the table above, the variable "tax 5 %" has a perfect correlation with the total variable, and is the highest in the table. "Tax 5 %" is also shown to have some correlation with the variable quantity, however it is weak and does not reach 0.80. It can be inferred that the variable "tax 5%" is the main variable that contribute to the strong correlation and should be omitted from future analysis. When the "tax 5%" is removed from the pairwise correlation table, the highest value of correlation is just 0.7055, as seen in the table below. Visually, the scatterplot diagram on the left below validates that there are no pairs that display a strong linear correlation.

Pairwise Correlations

Variable	by Variable	Correlation
Total	Quantity	0.7055
Total	Unit price	0.6340
Quantity	Unit price	0.0108
Time	Unit price	0.0076
Time	Total	-0.0052
Time	Quantity	-0.0086
Rating	Unit price	-0.0088
Rating	Quantity	-0.0158
Rating	Time	-0.0261
Rating	Total	-0.0364

Table 6: Pairwise Correlation Post Removal using JMP Pro

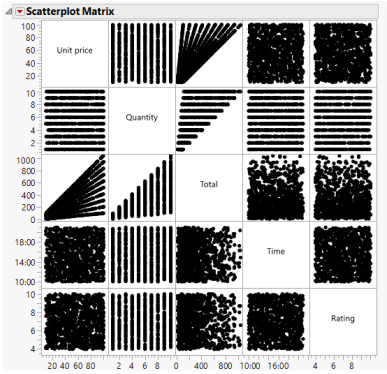


Figure 7: Scatterplot Matrix of remaining variables using JMP Pro

b. Identifying customer clusters using latent class analysis to tailor retention strategies

Given the different attributes about customer's demographics and purchase details, it boded well to conduct cluster analysis. Our objective of the cluster analysis is to segment customers on their behavioural aspects rather than only their demographics. Cluster analysis can help dissect customers and tailor strategies in future for each cluster to understand their purchase behaviour better and retain them for longer.

We performed latent class analysis (LCA) using JMP Pro tool offered by SAS. LCA is a type of cluster analysis that can be used when the attributes are mostly of categorical nature. The attributes considered for customer cluster analysis were customer type (member, non-member of the supermarket), payment type (e-wallet, cash, credit card), gender, and transaction amount (Under 250, between 250 and 500, and over 500). As LCA can only be performed on categorical variables, we converted the transaction amount to categorical variable from continuous variable using interactive binning feature of JMP Pro. We set the cut points at 250 and 500.

After running the latent class analysis three clusters were derived with minimum AIC and minimum BIC values. Although the cluster proportions are disbalanced, it does provide some insight into customer purchase preferences.

Loyalist – this cluster made up the biggest chunk of the supermarket customers. Around 66% of the customers in this cluster were members, who gender, and payment type was split almost evenly across each class of attributes. With respect to transaction amount, 46% of the cluster spent under 250 while the rest spent over 250 at the store. We classify this cluster as loyalists as they are members who spent fairly at the supermarket.

Low spenders – this cluster is predominantly made up of non-members. Close to 63% of the purchasers are males and spend 79% spent less than 250 at the supermarket. With respect to the payment type, while 43% opted for e-wallet payments, the remaining 57% were almost evenly spread across credit card and cash payments. This cluster is a low spending group with 78% spending under 250 at the supermarket.

Hot targets – this cluster makes up the smallest proportion of the three clusters. Close to 83% of the customers in this cluster are non-members. However, these can be considered as hot targets for conversion to supermarket members as 72% spend between 250 and 500 and 19% spend over 500 at the supermarket. These customers mainly used cash (54%) and e-wallets (42%) for making a purchase at the supermarket.

Latent Class Analysis

Cluster Comparison

NCluster	-LogLikelihood	BIC	AIC	Best
3	3516.41	7170.98	7072.83	Smallest BIC Smallest AIC
5	3512.31	7259.49	7092.63	
4	3512.82	7212.14	7079.63	

Latent Class Model for 3 Clusters

Model Summary

Measure	
-LogLikelihood	3516.414
Number of Parameters	20
BIC	7170.984
AIC	7072.828

Parameter Estimates

		Customer type		Gender		Payment			Total Groups 2		
Cluster	Overall	Member	Normal	Female	Male	Cash	Credit card	Ewallet	Under 250	250-500	Over 500
Cluster 1	0.68601	0.6633	0.3367	0.5370	0.4630	0.3315	0.3615	0.3071	0.4565	0.2977	0.2458
Cluster 2	0.21452	0.1387	0.8613	0.3709	0.6291	0.2899	0.2803	0.4298	0.7818	0.0352	0.1830
Cluster 3	0.09947	0.1645	0.8355	0.5343	0.4657	0.5464	0.0303	0.4234	0.0912	0.7156	0.1932

Cluster	Overall	Customer type	Gender	Payment	Total Groups 2
Cluster 1	0.68601				
Cluster 2	0.21452				
Cluster 3	0.09947				

Effect Sizes

Column	Effect Size	LR LogWorth
Customer type	0.4796	53.904
Gender	0.1361	4.057
Payment	0.2335	13.404
Total Groups 2	0.4329	42.897

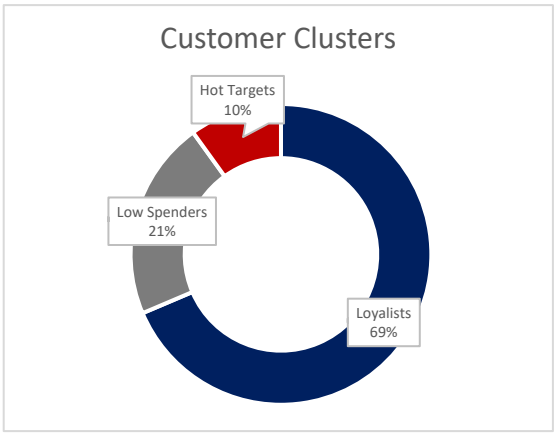


Figure 8. Three clusters were derived from latent class analysis using JMP Pro

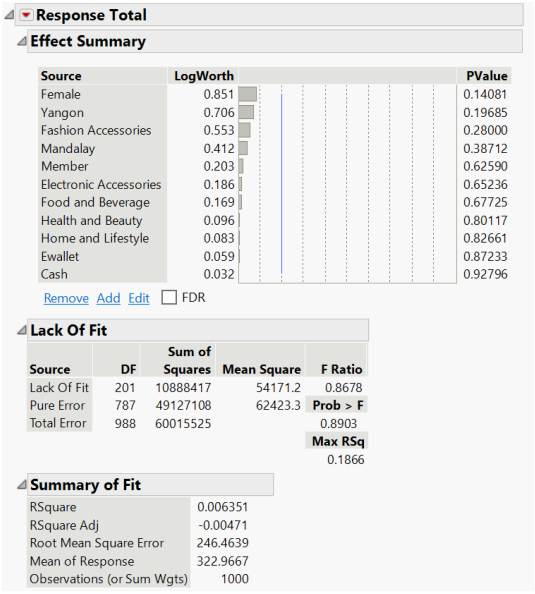
c. Most customer attributes are not significant indicators of supermarket's revenue

We wanted to analyze what factors most affect the revenue of the supermarket. For this we used two regression methods – linear regression and logistic regression.

Since most variables that we wanted to analyze for regression were categorical, we created dummy variables. For example, we created the variable female with female gender recoded as 1 and male gender as 0. However, after regressing dummy variables of city, gender, payment type, and product category, we found no attributes that were significant to predicting the revenue of the supermarket. The R-square value for this regression model was under 1%, with only intercept being significant to the model.

Next, we used logistic regression to analyze attributes that resulted in purchases at supermarket of over 500. For this purpose, we created a dummy variable for the revenue wherein income over 500 was encoded as 1 and under 500 as 0. Against this variable, we regressed the same variables – city, gender, payment type, and product type. Using forward stepwise regression with minimum AICc indicator, we found inconclusive results. None of the regressors were significant indicators of the revenue of the supermarket. (We also performed backward and mixed logistic regression using minimum AICc as well as minimum BIC indicators. The results remained inconclusive.

Figure xx. Regressing variables to predict revenue using linear and logistic regression using JMP Pro



Response Total

Effect Summary

Source	LogWorth	PValue
Female	0.851	0.14081
Yangon	0.706	0.19685
Fashion Accessories	0.553	0.28000
Mandalay	0.412	0.38712
Member	0.203	0.62590
Electronic Accessories	0.186	0.65236
Food and Beverage	0.169	0.67725
Health and Beauty	0.096	0.80117
Home and Lifestyle	0.083	0.82661
Ewallet	0.059	0.87233
Cash	0.032	0.92796

Remove Add Edit ☐ FDR

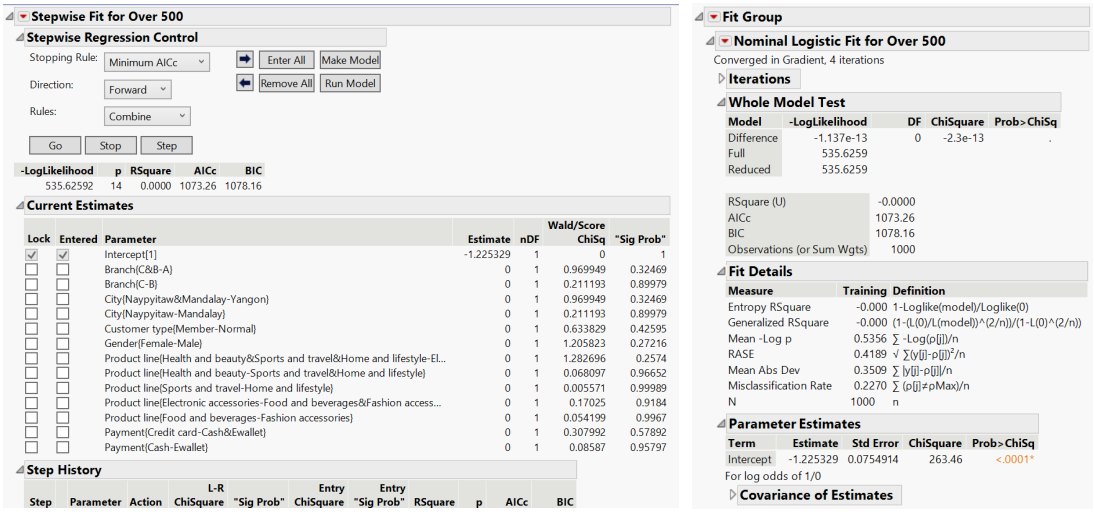
Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	201	10888417	54171.2	0.8678
Pure Error	787	49127108	62423.3	Prob > F
Total Error	988	60015525		0.8903
				Max RSq
				0.1866

Summary of Fit

RSquare	0.006351
RSquare Adj	-0.00471
Root Mean Square Error	246.4639
Mean of Response	322.9667
Observations (or Sum Wgts)	1000

Figure 8: Linear Regression



Stepwise Fit for Over 500

Stepwise Regression Control

Stopping Rule: Minimum AICc — Enter All — Make Model

Direction: Forward — Remove All — Run Model

Rules: Combine

Go Stop Step

-LogLikelihood	p	RSquare	AICc	BIC
535.62592	14	0.0000	1073.26	1078.16

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[1]	-1.225329	1	0	1
☐	☐	Branch(C&B-A)	0	1	0.969949	0.32469
☐	☐	Branch(C-B)	0	1	0.211193	0.89979
☐	☐	City(Naypyitaw&Mandalay-Yangon)	0	1	0.969949	0.32469
☐	☐	City(Naypyitaw-Mandalay)	0	1	0.211193	0.89979
☐	☐	Customer type(Member-Normal)	0	1	0.633829	0.42595
☐	☐	Gender(Female-Male)	0	1	1.205823	0.27216
☐	☐	Product line(Health and beauty&Sports and travel&Home and lifestyle-El...	0	1	1.282696	0.2574
☐	☐	Product line(Health and beauty-Sports and travel&Home and lifestyle)	0	1	0.068097	0.96652
☐	☐	Product line(Sports and travel-Home and lifestyle)	0	1	0.005571	0.99989
☐	☐	Product line(Electronic accessories-Food and beverages&Fashion access...	0	1	0.17025	0.9184
☐	☐	Product line(Food and beverages-Fashion accessories)	0	1	0.054199	0.9967
☐	☐	Payment(Credit card-Cash&Ewallet)	0	1	0.307992	0.57892
☐	☐	Payment(Cash-Ewallet)	0	1	0.08587	0.95797

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC

Fit Group

Nominal Logistic Fit for Over 500

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	-1.137e-13	0	-2.3e-13	.
Full	535.6259			
Reduced	535.6259			

RSquare (U)	-0.0000
AICc	1073.26
BIC	1078.16
Observations (or Sum Wgts)	1000

Fit Details

Measure	Training	Definition
Entropy RSquare	-0.000	1-Loglike(model)/Loglike(0)
Generalized RSquare	-0.000	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5356	∑ -Log(p[j])/n
RASE	0.4189	√ ∑(y[j]-p[j])²/n
Mean Abs Dev	0.3509	∑ |y[j]-p[j]|/n
Misclassification Rate	0.2270	∑ (p[j]≠pMax)/n
N	1000	n

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-1.225329	0.0754914	263.46	<.0001*

For log odds of 1/0

Covariance of Estimates

Figure 9: Forward Stepwise Logistic Regression

d. Analyzing product lines to improve sales and stock quantity in branches

To cater for high demand for certain product lines by branches we need to analyze the supermarket sales data to discover meaningful insights and patterns. By doing so, the supermarket will be able to effectively serve the demands of popular products. By doing such an analysis we can help the supermarket to react effectively and efficiently to improve its performance, cash flow and thus increase its profit margin.

We also aim to improve the revenue of the supermarket. For this, we need to look at the sales data available to gauge how well the different branches are performing for the different product lines. We need to analyze the sales with other attributes such as rating to check which attributes matter the most to the customers.

This analysis will be done using Python. Firstly, let us look at the different branches and product lines available in this supermarket to get an idea of what data we will be working with.

```
supermarket_sales_data['Product line'].unique()
```

```
array(['Health and beauty', 'Electronic accessories',
       'Home and lifestyle', 'Sports and travel', 'Food and beverages',
       'Fashion accessories'], dtype=object)
```

Figure 10: Product lines

There are 6 different product lines, and these product lines are being sold in 3 different branches ‘A’, ‘B’ and ‘C’ as can be shown below:

```
supermarket_sales_data['Branch'].unique()
```

```
array(['A', 'C', 'B'], dtype=object)
```

Figure 11: Branches

Let us look at which gender contributes more to the number of purchases of products in the different product lines.

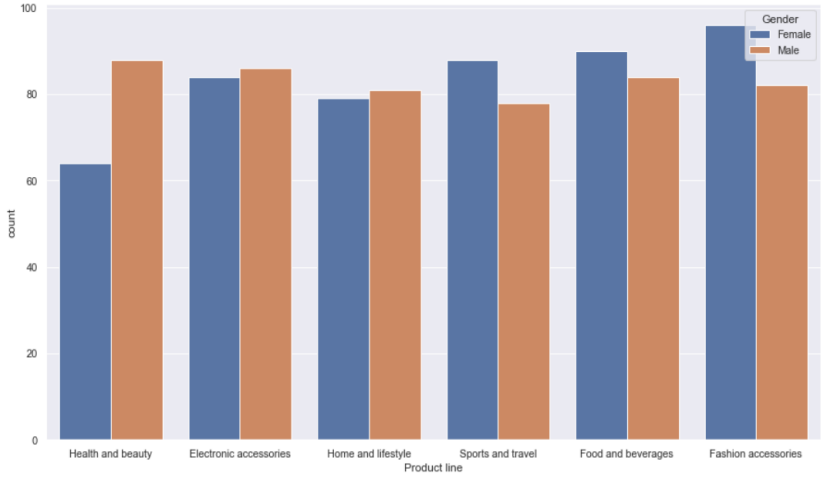


Figure 12: Number of purchases by product line and gender

In this plot we can see that Females tend to purchase products from product lines such as Fashion accessories, Food and Beverages and Sports & Travel more than Males. An interesting insight we can notice

here is that Males have a higher count of purchases for Health and Beauty whilst we may assume that the alternative might be true.

Next, we will also look at the different ratings of the product lines by branch.

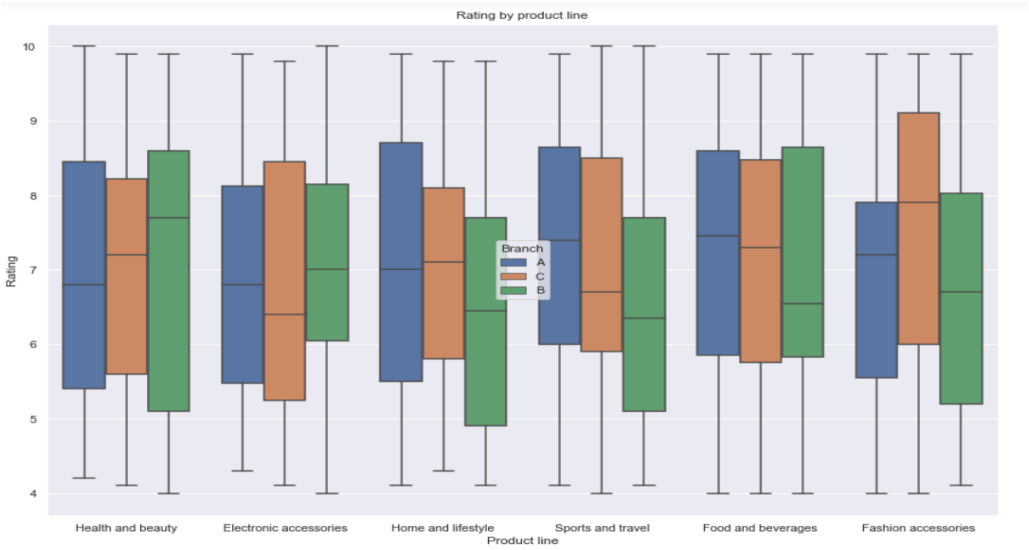


Figure 13: Rating of product lines by branches

Here, we can see that fashion and accessories has the highest rating in Branch C compared to Branch A and B. If we look at Health and Beauty Branch B has a higher average rating compared to Branch A and Branch C.

Here we note that Branch B has a lower average rating in these 4 product lines: Home and Lifestyle, Sports and Travel, Food and beverages and lastly, Fashion Accessories.

Quantity sold: Our first analysis will look at the quantity sold by different branches by product lines.

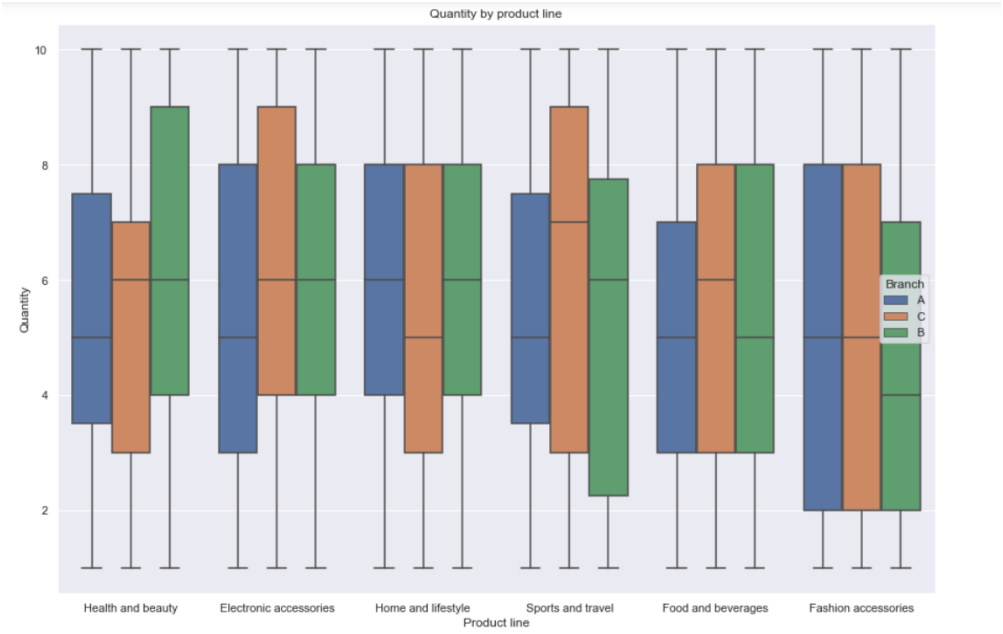


Figure 14: Quantity by product line and branch

Looking at the graph above, we can make the following inferences:

1. Branch C: higher average quantity sold for Food and Beverages as well as Sports and travel. It has the least average quantity sold for Home and lifestyle.
2. Branch A: least average quantity sold for Electronic accessories.
3. Branches B and C have a higher average of quanity sold for both Health and beauty as well as Electronic Accessories.
4. Branches A and C have a higher average quanity sold for Fashion accessories product line.

Sales: Our second analysis will be to look at the total sales of the supermarket

Firstly, we will look at the total sales by product lines.

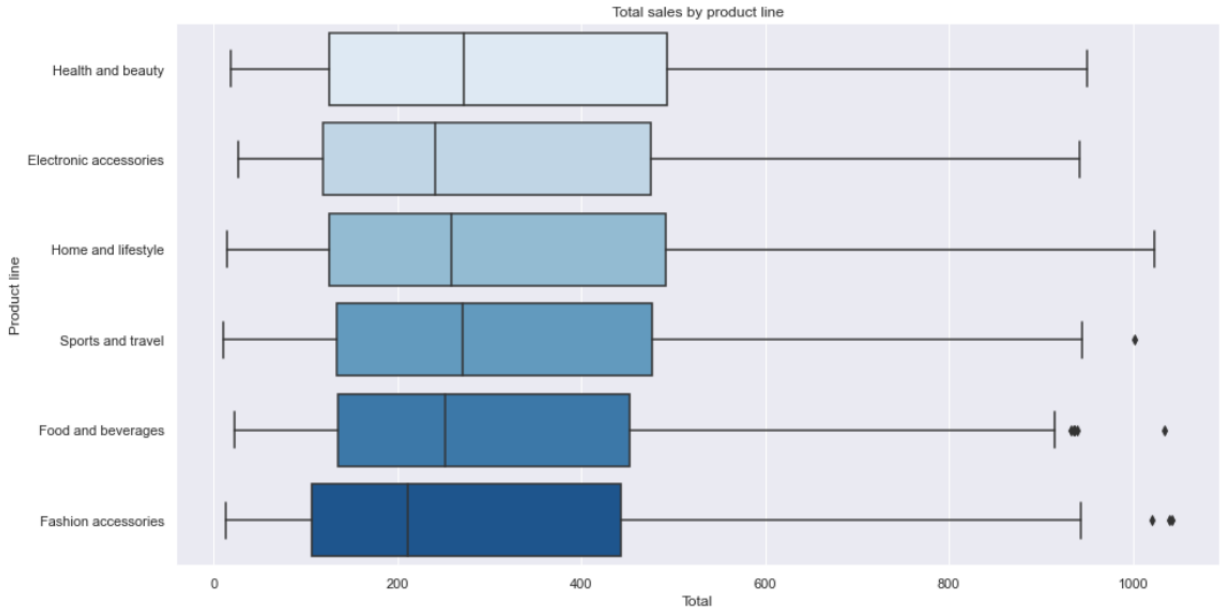


Figure 15: Sales by product line

From the above graph we can see that the Health and Accessories has a higher average sale whereas Fashion Accessories has the least average sales.

Secondly, we will break down the previous analysis to make it more granular by looking at the different branches.

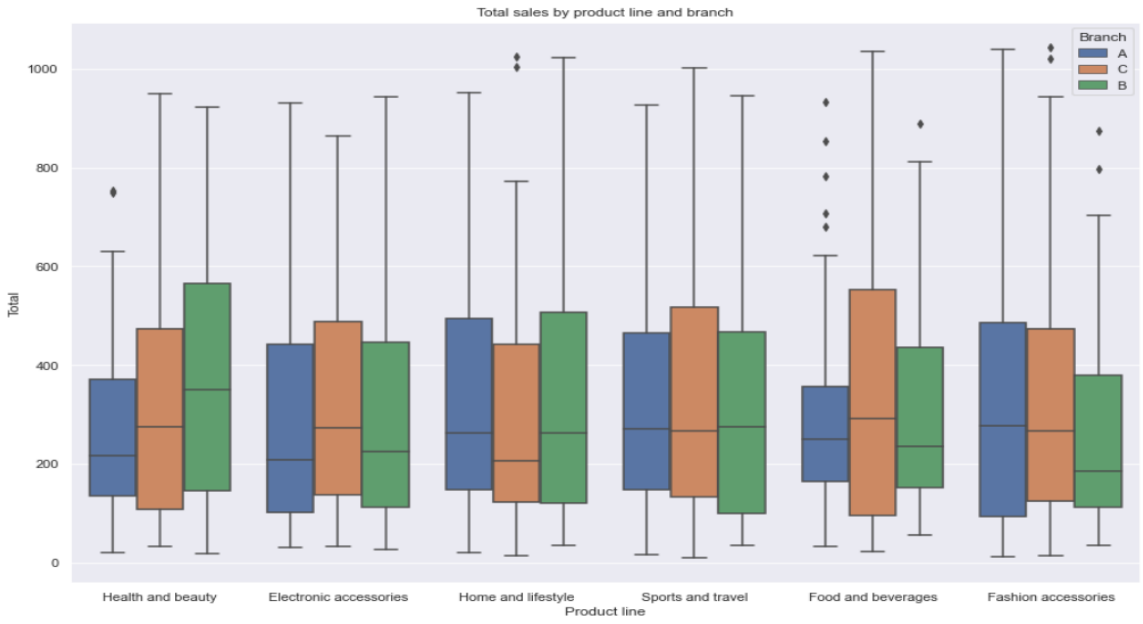


Figure 16: Total sales by branch and product line

From the above graph, we can see that Branch B has higher average total sales for Health and Beauty and least for Fashion accessories.

Branch C has higher average sales for Electronics accessories and Food and beverages and least for Home and lifestyle.

For Branch A has the highest average sales for Fashion Accessories followed by Branch C. Branch A has the least average sales for Health and Beauty and Electronic accessories.

Finally, we will further break down the above analysis by day of week to check which day of the week contributes more to the sales.

Let us look at branch B's Health and Beauty and Fashion accessories product lines.

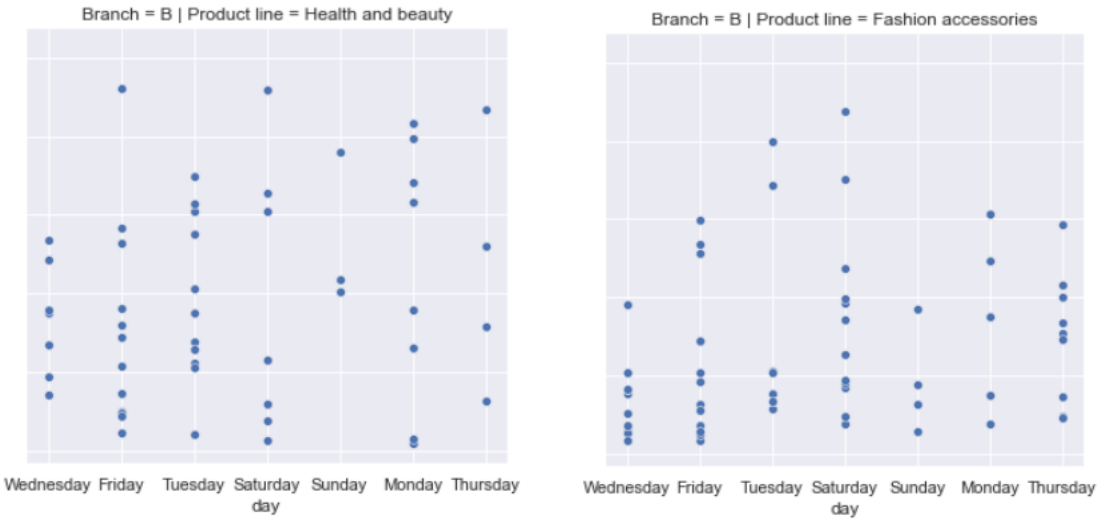


Figure 17: Sales by Branch B on different days of week

For the Health and Beauty product line, Branch B has highest total sale on Friday and Saturday and the least sale on Wednesday.

For the Fashion Accessories product line, Branch B made the highest sale on Saturday.

Next lets look at Branch C:

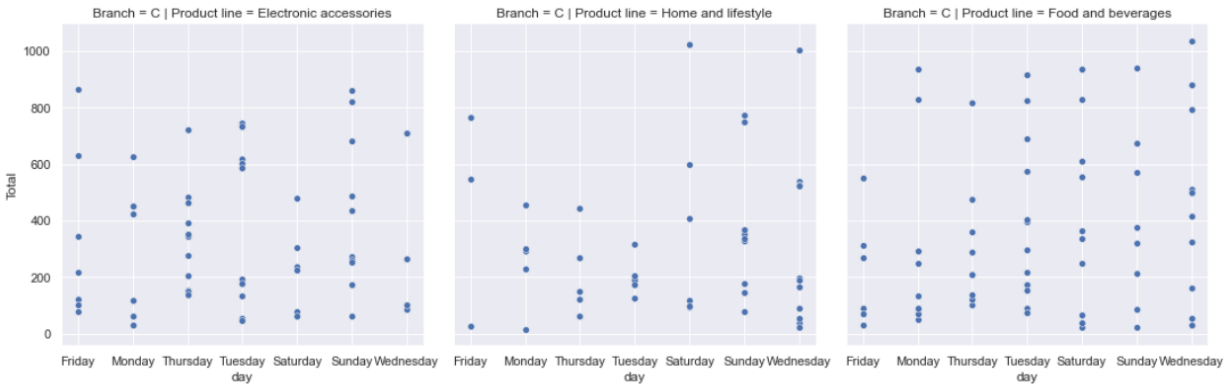


Figure 18: Total sales by Branch C by day of week

For Home and Lifestyle, the highest sale for this line were made on Saturday and the lowest sale were made on Tuesday.

For Electronic accessories, the highest sale for this line were made on Friday and Sunday and the lowest sale were made on Saturday.

For Food and beverages, the highest sale for this line were made on Wednesday and the lowest sale were made on Friday.

Branch A and C:

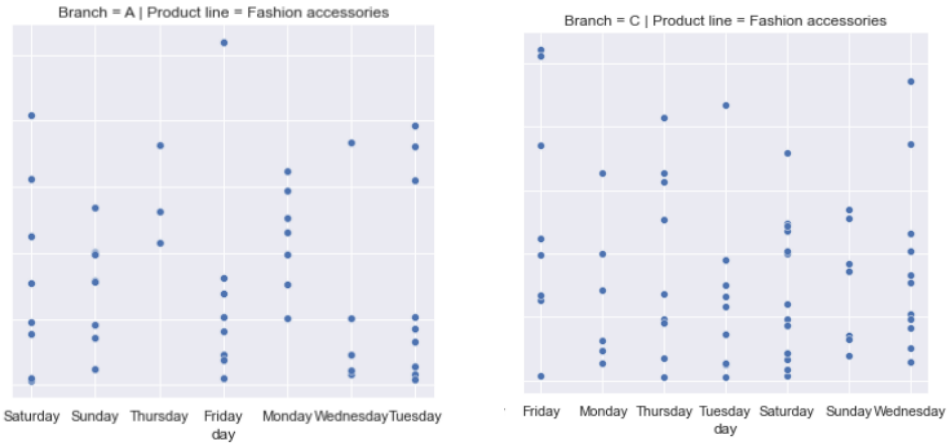


Figure 19: Sales for Branches A and C by day of week

We can see that the highest sale for this line were made on Friday for both Branch A and C.

4 Solutions and actionable intelligence

4.1 Recommendations based on product line analysis

Based on the analysis above, to improve overall customer satisfaction and sales, the following are recommended:

Gender:

We saw an interesting insight that males tend to purchase more Health and Beauty products than females. So, to improve the sales, more female customers can be provided with discounts or promotions to incentivize them to spend more on this product line.

Similarly, males can be provided more discounts and coupons for other product lines like Sports & Travel and Food & Beverages.

Rating:

Branch B has a lower average rating in 4 product lines: Home & lifestyle, Sports & Travel, Food & beverages and lastly, Fashion Accessories. To improve customer satisfaction and thus improve the sales, Branch B needs to provide promotions for these four product lines. Targeted advertising on these product lines for Branch B can help to bring in more customers and motivate existing customers to purchase more of these products.

Quantity sold:

Branch C has a higher average quantity sold for all product lines except Home and lifestyle. So, to cater for the high demand for the required products, Branch C needs to increase its stock. To increase the quantity of products sold for Home and Lifestyle it can introduce offers where these products can be bought along with the popular products for a lower price. This can help to maximize the sales for both the popular as well as the less popular products.

Similarly Branch A has the lesser average quantity sold for Electronic accessories and higher average quantity sold for Fashion accessories product line. The same strategy as above can be applied here.

Finally, since Branch B has a higher average quantity sold for both Health and beauty as well as Electronic Accessories, the stock for these products needs to be increased to effectively cater to the demand.

Sales:

The lowest average total sales for Branch A is for Health and Beauty as well as Electronic accessories whereas for Branch B is for the product line Fashion Accessories and finally for Branch C is for Home and lifestyle.

The sales for these branches can be improved by introducing discounts and improving the reach for these products through relevant advertising and marketing channels. Using social media influencers to promote purchase from the supermarket can spread huge awareness in the digital age and can also be cost effective.

Furthermore, we also looked at the day of the week where each product line is popular and least popular.

Branches A and C make the highest sale for food and beverages on Wednesday. For Branch C, for Home and Lifestyle, the highest sale for this line were made on Saturday. For the health and beauty product line, Branch B has the highest total sale on Friday and Saturday.

To make sure that the demand meets the sales, the above products need to be kept in stock. This will help to maximize the sales and help with customer satisfaction.

4.2 Shopping coupon for members to encourage spending

With the aim of stimulating higher spending in each receipt from the members, we can design coupons to be given to them. We illustrate in detail one potential coupon design.

Nature of coupon:

Cash coupon to offset payment; can be used when spending before coupon reaches a minimum spend criterion.

The minimum spend is required so that members will have to increase their purchase in the receipt to the minimum spend amount if they want to use the receipt. Revenue can thus be increased.

Distribution of coupons:

During coupon distribution period, one coupon (with a validity period) will be given to each member after their purchase at the supermarket. Members will have their details entered when they check out and hence it can be traced who has already received the coupon.

The coupons should aim to reach as many members as possible – to ensure their spending behaviour will be as predicted by the existing data, upon which the coupon strategy is derived.

Variables to decide on:

- Coupon face amount: how much is each coupon worth
- Minimum spend requirement

Criterion for optimizing the variables:

Increase in profit net of coupon cost, where the increase in profit is calculated as the increase in revenue * gross profit margin. Coupon cost is the coupon face amount * total number of coupons used.

Assumptions needed:

- When will a member be inclined to increase the purchase to the minimum spend so as to use the coupon, in terms of equivalent discount implied by the coupon.

  For example, when a member's original planned spending was $80, and a coupon of $10 is given with a minimum spend of $100, the member would only need to top up $10 in order to receive $20 extra worth of goods. This represents a 10/20 = 50% discount. If the member sees 50% discount as worthwhile, he/she is likely to top up spending and use the coupon.

- Gross profit margin as a percentage of revenue. We need to have an average gross profit margin (revenue – cost of goods sold) assumption to determine whether the profit has increased more than the coupon cost.

Sample Python code to find profit increase net of coupon costs:

The function below assumes that a member will increase his/her purchase to the minimum spend amount if the coupon represents a 20% or more discount for the additional purchase amount.

The gross margin assumed is 50%, i.e., gross profit = revenue * 50%.

```
def coupon_benefit(coupon_value, min_spend):
    required_discount = 0.2
    profit_margin = 0.5
#to find the minimum original purchase amount such that using a voucher is considered worthwhile
    purchase_min = min_spend - 10/2*coupon_value

    profit_inc = 0
    for i in l:
        if i >= purchase_min:
            new_purchase = max(min_spend, i)
            profit_inc = profit_inc + (new_purchase - i)*profit_margin - coupon_value

    return profit_inc
```

Note that the coupon is assumed to not be used by members with original purchase amount less than the purchase_min. No coupon cost will be incurred in those cases.

With this function, we can now test out potential coupon strategies, with different coupon face amount and minimum spend requirement combinations.

For example, in the code below, given a coupon face amount of $100, we test out several minimum spend requirements to see which one will give the maximum profit benefit to the supermarket.

```
import pandas as pd
sales = pd.read_csv('supermarket_sales - Sheet1.csv',encoding = "ISO-8859-1")
```

```
l = sales['Total'][sales["Customer type"] == 'Member']
```

```
for min_spend in [100,300,500,700,900,1000]:
    print(min_spend,':', coupon_benefit(100, min_spend))
```

```
100 : -48150.34824999998
300 : -28526.175
500 : 4952.451249999997
700 : 6921.0520000000015
900 : 6308.225000000002
1000 : 5984.734250000003
```

As can be seen above, $700 is the optimal minimum spend to set in this case. The profit increase amount of $6921 can be scaled up to the number of coupons that are intended to be given out, to form the expected increase in earnings from the coupon strategy.

Data feedback and improvement cycle:

The effectiveness of the coupon strategy can be monitoring by looking at the coupon usage rate, and the average spending from members during the coupon validity period. The increase in profits should be estimated from the new data and compared against the original profit increase prediction produced by the model. Explanations for the gaps should be explored and hence the model can be improved to design better coupon strategies. For example, it could be due to the assumptions of equivalent discount to stimulate spending or the gross profit margin being inaccurate. These fixed inputs to the model should be constantly monitored and revised if necessary.

Over time, revised spending habits from the members should be constantly feedback to the coupon campaign to design more attractive coupons. For example, coupons targeted at a particular product line can be considered if members have a tendency to spend high amounts on those categories of goods.

4.3 E-wallet

In this section, we target to grow revenue stream through considering ePayment methods in which our supermarket can leverage on. This is illustrated through our targeting of *Hot Targets* customer segment.

Following the preliminary customer segmentation, we conduct further partitioning for *Hot Targets* to understand how targeting can be differentiated through payment methods. Through partitioning, we observe that *Time* of purchase was a variable which differentiates mode of payment in this customer segment.

Transactions from *Hot Targets* were hence differentiated into the following sub-segments:

		Total Revenue	Total Qty	% on eWallet	Avg Price
Health and beauty	Before 2pm	3,360	54	43%	62.22
Fashion accessories	After 2pm	3,768	68	49%	55.42
Electronic accessories	After 2pm	4,002	65	14%	61.57
Fashion accessories	Before 2pm	4,199	71	14%	59.14
Food and beverages	After 2pm	4,785	80	63%	59.81
Home and lifestyle	After 2pm	5,092	80	74%	63.65
Others	After 2pm	6,149	96	-	-
Others	Before 2pm	8,303	148	-	-

Breakdown of Hot Targets transaction by Product Line and Time

From 4.2, we assumed that customers need to perceive a discount of 20% before a coupon or voucher is sufficiently attracted to incentivize purchase. Similarly, to encourage payment channel swing from cash to e-wallet, we make the same assumption. Based on our understanding from the table above, we focus our resources on lines where e-wallet penetration is low. For example, electronic accessories purchased after 2pm and fashion accessories purchased before 2pm.

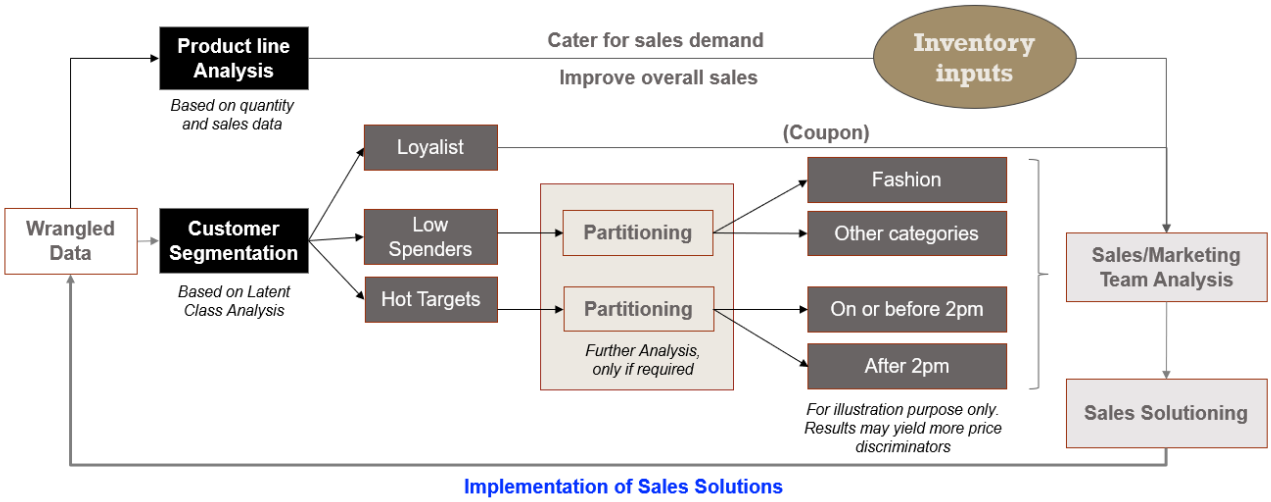
We apply the algorithm from 4.2 on the sub-set of data on fashion accessories purchased before 2pm (highlighted in amber), we observe that a minimal spend of $550 is required before the promotion breaks even.

```
In [23]:
1 for min_spend in [450,500,550,600]:
2     print(min_spend,":",coupon_benefit(50,min_spend))
```

```
450 : -25.86500000000001
500 : -0.8650000000000091
550 : 24.13499999999999
600 : 49.13499999999999
```

There are key differences between a coupon campaign to reward existing members and a campaign to encourage payment channel swing. Considering the mid- to long-term benefits that it brings (e.g. customer engagement, promotion of supermarket's payment ecosystem), the management may have a bigger appetite for revenue risk, or even adopt a strategy of paying to encourage customers to adopt the e-Wallet.

5 Closed-loop data ecosystem



This is a summary of the data analysis done and how they contribute to the closed-loop data ecosystem for the supermarket.

The loop starts from having the transaction data at hand, gathered over a period of time. Analysis with various data techniques was done on the data to explore the possible insights related to the business goal of revenue and profit maximization. Segmentation was also done to enable targeted strategy setting. This represents the data-driven feature of the loop as everything is based on observations of the data.

Business strategies are then devised for different segments, with joint analysis from the sales and marketing teams for the strategy design details. The sales solutions are then implemented, changing parts of the operations of the supermarket. Included in the solutions will be the additional data to be collected for the monitoring of the strategies' effectiveness, for example as explained in section 4.2 for the coupon strategy.

Thereafter, with new data collected over time, the loop goes back to the first stage. With consistent data analysis effort, there will be new insights found. Strategies will be improved or new strategies can be designed. The loop will continue to evolve, with the overall business objective of maximizing profits.

6 Considerations for future works

- Currently, there seems to bias in the data. Based on the confirmatory analysis, we see that many attributes are correlated. In the future, we can consider better data capture to contain the biases and multicollinearity among attributes.
- The regression analysis is inconclusive. To analyze predictors of the revenue in the future, we can consider capturing other attributes such as frequency of purchase, actual products purchased as well as more demographic attributes of the customers.
- Capturing additional data – both in terms of attributes and time period – can also subsequently help enhance customer clusters from the latent class analysis.

- Enhance clustering can be used in the future to tailor coupons by customer's respective categories.

Reference

Pyae, A. (2019, May 27). *Supermarket sales*. Kaggle. Retrieved July 3, 2022, from https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

supermarket_sales - Sheet1.csv

After reasonability checks on the data, columns “cogs”, “gross margin percentage” and “gross income” are not used in this project as their contents are not consistent with their names. For example, “cogs” is equal to unit price * quantity, i.e. gross revenue, instead of cost of goods sold.