Customer Clustering in a Supermarket

Business Analytics with R – BUAN 6356.001

Group 4

Shrutika Pujari

Abhishek Yadav

Anantha Sameer Prashanth Patibandla

Siva Srinivas Narra

Shreya Devidas Amrutkar

Mitali Selot

Setting:

Clustering is a crucial task for supermarkets aiming to understand and cater to the diverse needs of their clientele. This data set contains historical sales records from three different branches of a supermarket, spanning three months, and their customer demographics. Supermarkets typically have a diverse customer base with varying preferences, buying behaviors, and demographic characteristics. By clustering customers based on their purchasing patterns, branch, and gender, the supermarket can segment its customer base into distinct groups with similar characteristics, enabling it to optimize its marketing strategies, inventory management, and overall customer experience. Therefore, our business question is: How can we optimize marketing strategies for targeted customer segmentation to improve sales?

Data Source & Description:

We have taken the data set from Kaggle, which covers sales data in 3 different supermarkets in Myanmar (https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales). The following table describes the dataset.

No	Attribute	Description
1	Invoice ID	A unique identifier of the invoice identification number.
2	Branch	Branch of supercenter (A, B, and C).
3	City	Location of supercenters in Myanmar.
4	Customer Type	Type of customers–normal customers and member customers.
5	Gender	Gender of the customer.
6	Product line	General item categorization groups.
7	Unit price	The price of each product in $.
8	Quantity	Number of products purchased by customers.
9	Tax 5%	5% tax fee for customers buying.
10	Total	Total price, including tax.
11	Date	Date of purchase.
12	Time	Time of purchase.
13	Payment	Type of payment used by the customer for purchase.
14	COGS	Cost of goods sold.
15	Gross Margin Percentage	Gross margin percentage.
16	Gross Income	Gross income.
17	Rating	The rate of their overall shopping experience on a scale of 1 to 10.

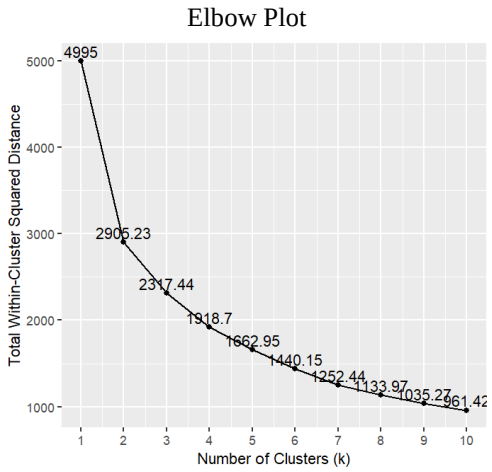
Challenges: Clustering algorithms produced complex and nuanced results that required careful interpretation. Understanding the meaning and implications of different clusters and determining actionable insights from the results was slightly challenging. Furthermore, part of our data were numeric variables, and the other half was categorical, so it was a challenge to normalize and factorize the data in order to conduct the necessary analysis.

Analysis & Discussion:

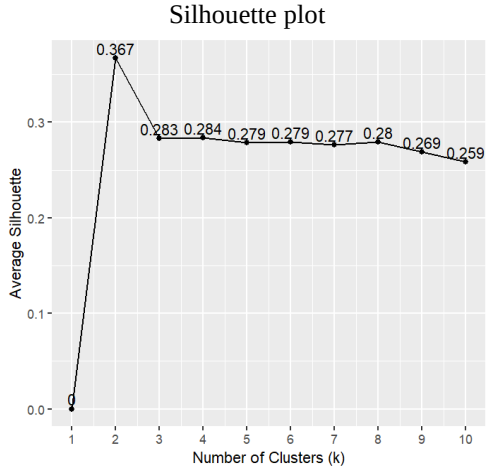
Overview of the analysis approach

The first step of the project was to convert the Branch, City, Customer, Gender, Product Line, and Payment variables into factor types using as.factor as there were no null values or data cleaning needed. Working with levels (categories) inside the data is made possible in R via the factor data type, which is helpful for categorical data. The next step was to normalize the numeric columns after converting them into factor types for categorical data analysis. A new data frame named Sales.norm was created by combining the output of the sapply function.

An elbow plot was created to understand the optimum number of clusters for the data set. From the graph, we can come to the conclusion that the optimum number of clusters is 2.



We cross-verified the optimum number of clusters by plotting a Silhouette plot, which also told us that the optimum number of clusters for this data set is 2.



We used various tools in our analysis approach, such as clustering and data visualization, to understand the variables that affect the three supermarkets. We specifically looked at how total revenue and gross income impacted sales according to customer demographics. On the basis of the train and test data, we assessed these models' performance and generated performance metrics. Additionally, we created visual representations of the data to find patterns and differences in the data that were crucial for finding insights.

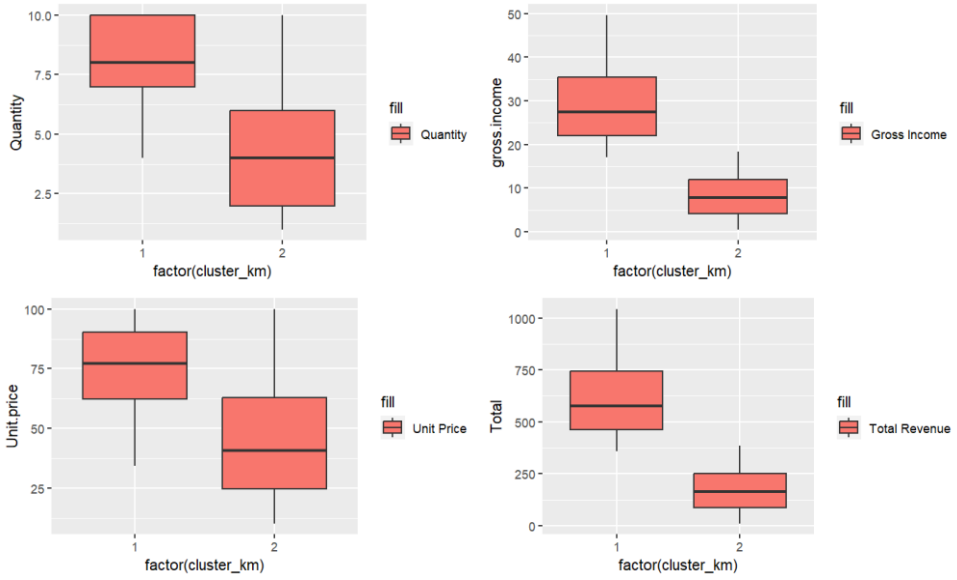
Aggregate Data Frame of Categorical Data

```
> aggdf <- aggregate(cbind(Branch1,City1,Customer1,Gender1, ProductLine1,Payment1,Rating) ~ cluster_km, data=survey, mean )
> aggdf
  cluster_km  Branch1     City1 Customer1   Gender1 ProductLine1 Payment1   Rating
1          1 2.023188 1.0028986 0.4782609 0.5304348     3.510145 2.008696 6.912464
2          2 1.969466 0.9862595 0.5099237 0.4854962     3.574046 1.996947 7.004427
```

We explored the data by creating an aggregate data frame to learn more about customer demographics. According to the aggregate demographics, it can be inferred that Cluster 1 has higher average values compared to Cluster 2, except for Rating and Customer 1. There are no major differences between the two; for example, in Cluster 1, the average Gender1 value is approximately 0.503. This value is close to 0.5, which suggests a nearly even distribution between males and females in this cluster. For Cluster 2, the average Gender1 value is about 0.485. Even though this number is less than 0.5, which suggests a slight male majority, the difference is very minimal, indicating a balanced mix of genders in this cluster, too.

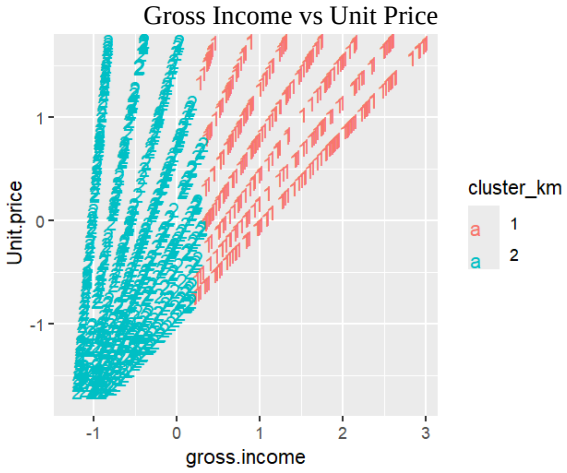
Additionally, the value of Customer1 in Cluster 1 is approximately 0.478, which is closer to 0 than 1. This indicates that there is a slightly higher proportion of members compared to normal customers since the code for members is 0 and 1 for normal customers. In Cluster 2, the average is 0.509, which is very close to the midpoint, suggesting that there is an equal mix of members and normal customers in Cluster 2.

Box Plots of Quantity, Gross Income, Unit Price, and Total Revenue



The median total for Cluster 1 is significantly higher than Cluster 2 in Quantity, Gross Income, Unit Price, and Total Revenue, which tells us that Cluster 1 buys more items at a higher price than Cluster 2. For example, the length of the box plot for Cluster 1 for Total is wider, indicating more variability in the total revenue than in Cluster 2. The "whiskers" of the box plot show the data range; in this case, Cluster 1 exceeds 1000, while the upper whisker in Cluster 2 only reaches 375. Moreover, the box plot for the unit price is higher in Cluster 1 than in Cluster 2, indicating that customers in Cluster 1 are buying more expensive products than in Cluster 2. Therefore, when comparing the two clusters, it's apparent that Cluster 1 generally has higher total figures than Cluster 2, as evidenced by the higher total and unit prices and larger range of values.

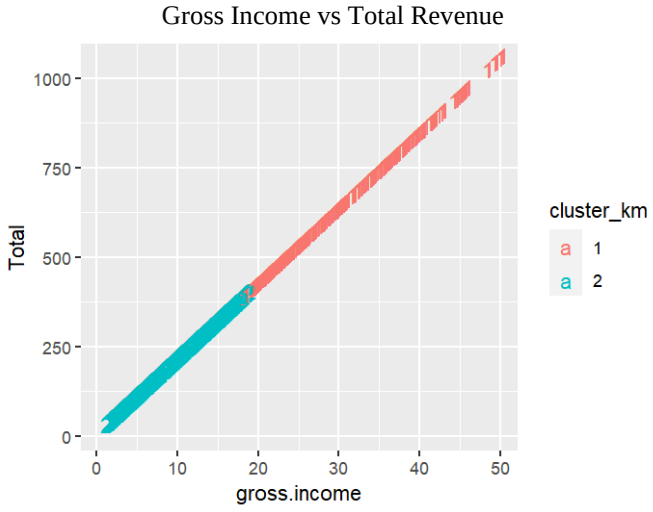
Data Visualization



The scatter plot depicts a link between the two clusters' gross income and unit price. Lower gross income levels correlate with lower unit prices for Cluster 2. This implies that this group is making lower-priced purchases. Conversely, Cluster 1 shows that the unit price grows in tandem with an increase in gross revenue. Therefore, Cluster 1 is more likely to buy expensive goods, increasing the supermarket's gross income.

The second scatter plot shows the correlation between each cluster's gross revenue and the number of things purchased. Generally speaking, Cluster 2 spends less money overall, which is consistent with their lower gross income. By contrast, Cluster 1 exhibits a higher gross income and makes more purchases. This implies that clients in Cluster 1 contribute significantly to the sales volume and revenue of the supermarket, suggesting that they are either frequent or bulk buyers compared to customers in Cluster 2.

The supermarket can adjust its marketing efforts based on these insights. Promotions on more reasonably priced or smaller products may be more appropriate for Cluster 2, while promoting more premium products and bulk buy discounts may be more effective for Cluster 1. Knowing these trends can also help the supermarket manage its inventory, making sure that the correct assortment of goods is on hand to satisfy the various needs of each client group.



There's a clear positive linear relationship between gross income and total revenue. As gross income increases, total revenue also increases. The graph reveals that Cluster 1 has a higher gross income and contributes more to the supermarket's total revenue as they have more disposable income compared to Cluster 2.

The dot plot depicts the spending analysis of the two clusters concerning the unit price of the products. Cluster 1 can be said to be purchasing products with higher unit prices or products labeled 'Premium' when compared to those purchased by Cluster 2. As mentioned earlier in the plot that compared the relationship between the unit price and the clients' net income, it can also be inferred that perhaps the customers that prefer the premium products could have memberships with the supermarket and might be offered membership advantages as well. This analysis can further help the management in creating niches for their target customers and design marketing strategies to bring in new potential members under the ‘Premium Membership’ umbrella.

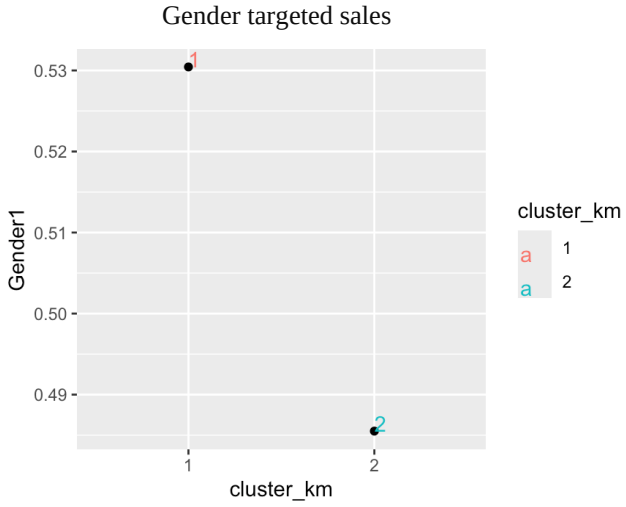
Apart from formulating marketing strategies, the analysis helps in one of the most crucial parts of any business: its supply chain. The decisions can be made on how and how much to order, when to restock, where to store, and when to release offers. The supply chain teams can focus on inventory management, involving inventory handling, holding costs, and maintaining an optimum inventory level. This analysis further gives rise to various inventory management techniques, such as Vendor Managed Inventory (VMI), where the vendors keep an eye on the inventory levels, know when and how to restock, and are aware of the customer trends involved with the store.

The dot plots solidify the claim that the clients belonging to Cluster 1 contribute to higher total profits for the supermarket as their contribution to sales and quantity purchased is comparatively higher than those belonging to Cluster 2. While Cluster 1 contributes as the major profit driver, Cluster 2 presents opportunities for improvement and optimization. The various reasons for this disparity can possibly be and are not limited to the ones mentioned so far, such as lower sales volumes and sales by Cluster 2, lower net income, and not being involved in memberships and membership-related promotions.

This analysis helps to assess not only the situation of the customers but also major business processes such as logistics. The need to optimize logistical costs to inculcate smarter spending practices is very crucial. The management may attract more customers to become members through marketing strategies such as same-day delivery and an easier return policy. This helps the

management think about involving third-party logistics (3PL) providers for transportation from wholesalers, warehouses, and other suppliers to the store and to the clients (as a part of the delivery process). The involvement of 3PL providers helps the management reduce the costs of maintaining and handling fleet, logistics, and manpower. All these activities are taken care of by the 3PL provider at an annual fee.

This plot depicts the proportion of customers in each of the clusters. The proportion of customers in both clusters is almost the same, which could mean there is less differentiation in the services provided to the customers in Cluster 1 or those with the premium membership and those in Cluster 2 or those without any membership. If not addressed, situations like these can lead to poor customer retention and extreme customer dissatisfaction. One way to ensure differentiation is through marketing. Having various target markets, such as member/non-member, gender, age, etc, can help analyze and attract more customers accordingly.



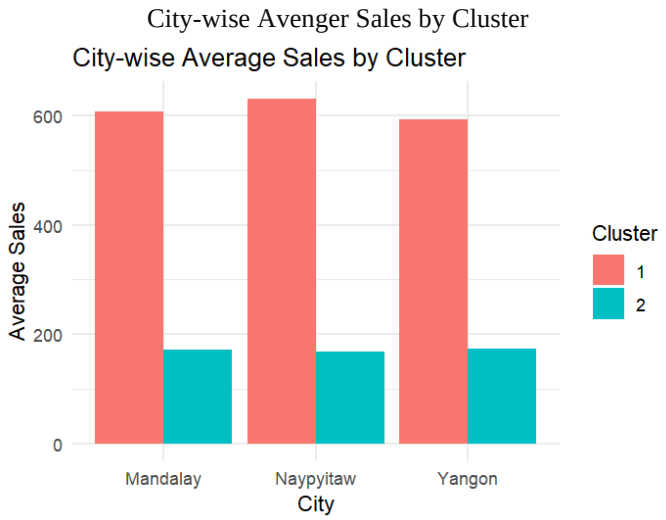
This plot tells us that when considering gender as a differentiator, the sales are almost the same for both clusters, irrespective of gender. Therefore, this highlights the need for better marketing techniques and strategies for products that are gender specific in order to sell to the appropriate customers.

As discussed earlier, it can be noticed in the following dot-plots that the customer satisfaction rates, particularly for those belonging to Cluster 1, are lower than those customers belonging to Cluster 2.

The lack of differentiation for customers who opted for the premium membership has resulted in lower or almost the same satisfaction rates as those belonging to Cluster 2, the normal customers. Thus, it is necessary for the management to brainstorm where things are going wrong and how things need to be improved in order to increase the quality of service provided to the customers who have paid the extra penny for membership. This improvement in customer experience can come from various aspects such as better marketing strategies (as discussed above), better implementation of reverse logistics, better customer service and quality assurance practices, and more member perks and benefits like discounts, cash backs, reward points, simpler exchange and return policies, same day home delivery services and much more.

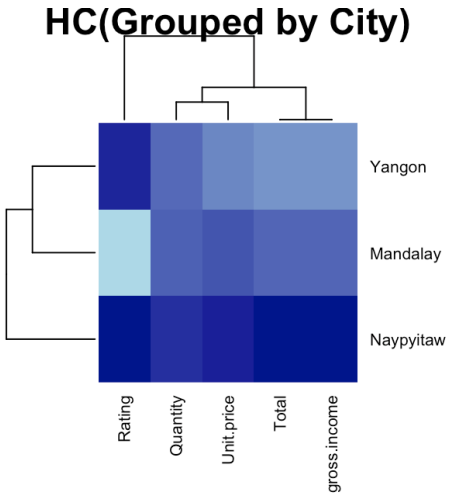
Furthermore, a feedback mechanism can be incorporated by the management where the store collects regular feedback from the customers who are members to address the issues, they face with respect to the customer experience and what more can be done in order to resolve them. This process can be performed on a regular basis to keep the management up to date with customer trends and cater to customer needs appropriately. Another way to be more customer-centric is by leveraging data analytics and customer purchase history, providing personalized product recommendations tailored to premium customers' purchasing behavior, and making them feel more valued.

We have discussed the quantity, sales, and which cluster is buying more goods, but to create insights on what the supermarket should do to increase sales, we must take a look at customer demographics.



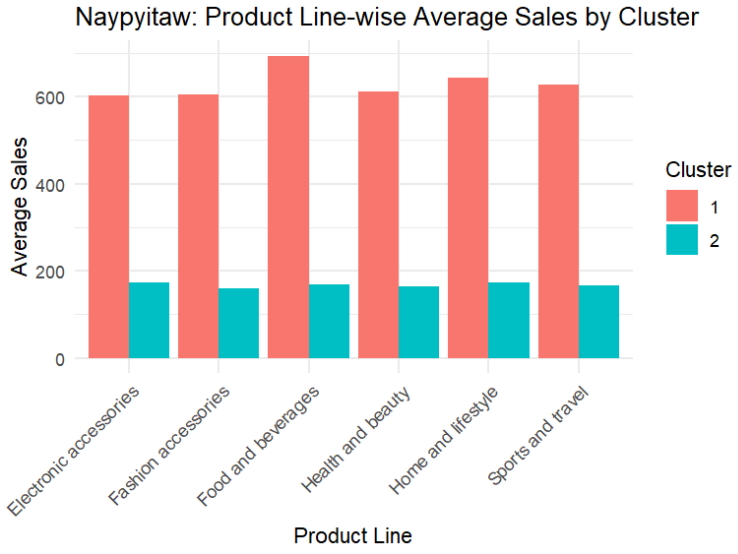
According to the bar chart, the average sales figures for both clusters are similar within each city, meaning that the clusters are likely determined by factors other than just the city. This could suggest that the segmentation is based on product categories or gender. However, Naypyitaw has the highest average sales in Cluster 1 by a small amount.

Furthermore, hierarchical clustering was used to cross-verify if the insights through k-means clustering were correct.

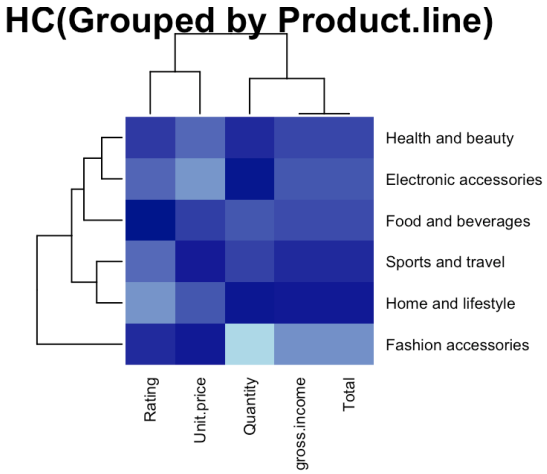


A heat map was plotted for the three cities using hierarchical clustering. From the map, it can be concluded that Naypyitaw is buying the most number and most expensive products, therefore confirming our results with k-means clustering.

Product Line Average Sales by Cluster in the City with the Highest Sales

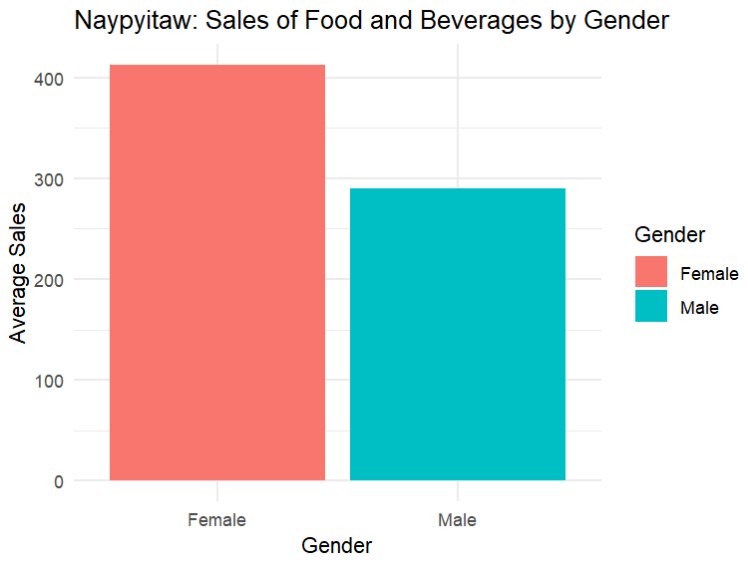


To go further in customer demographics, we're now going to focus on Naypyitaw—the city with the highest sales. In Naypyitaw, food and beverages are selling the most compared to other categories, suggesting that more resources should go into promoting food and beverages in Naypyitaw for Cluster 1 customers. There is no particular category that Cluster 2 prefers as the average sales are nearly identical in every category.

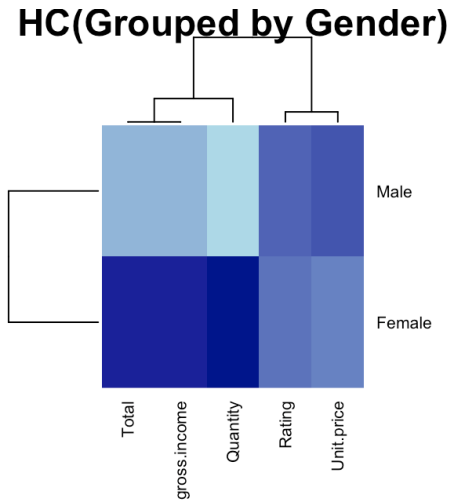


In hierarchical clustering, similar results are shown where food and beverages are the most bought category, with home and lifestyle in second place.

Average Sales of Highest Selling Product Line by Cluster in the City with the Highest Sales

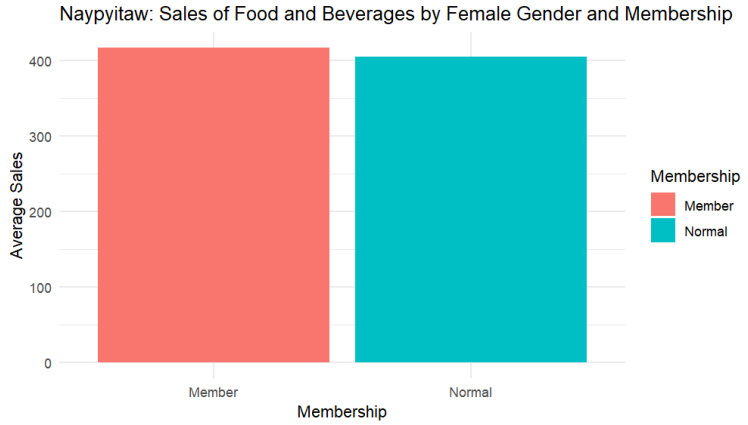


Female customers have a significantly higher sales figure for food and beverages in Naypyitaw, suggesting that there is a gender-based difference in purchasing behavior regarding food and beverages, with females, on average. Therefore, supermarkets should focus on women who are buying food and beverages in Naypyitaw to keep increasing sales, but to increase sales for men, supermarkets could engage with male customers through channels they frequent more, such as sports events, gaming communities, or tech-related forums. Additionally, they could sell products that complement male-oriented activities or interests, such as snacks for watching sports or high-protein health foods.



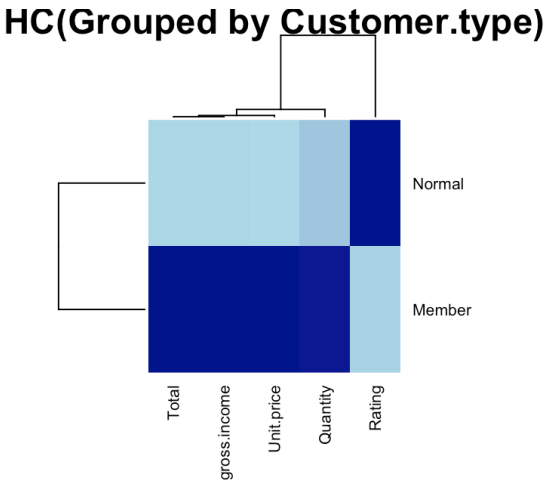
Like the graph plotted by k-means clustering, the hierarchical clustering graph confirms that women buy more products than men as the quantity box is darker for women than men.

Average Sales of Food & Beverages by Membership for Women in Naypyitaw

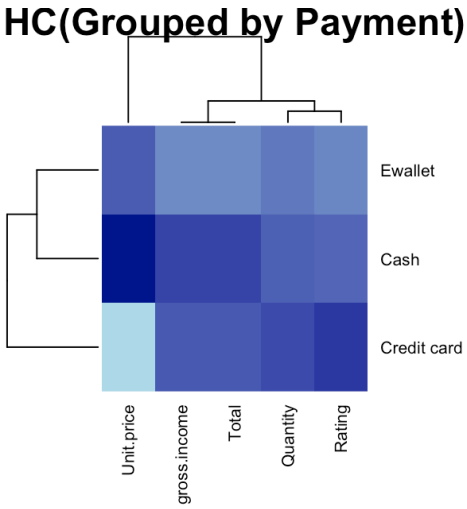


The chart compares the average sales of food and beverages between two types of female customers in Naypyitaw: members and normal customers. Female customers who are members have higher average food and beverage sales than non-members, indicating that membership may be associated with higher spending, more frequent purchases, or both. It also tells us that membership effectively contributes to sales as female customers who are not members have lower average sales than members. This might suggest that non-member customers either spend less per visit, visit less frequently, or purchase less expensive items. Therefore, we can say that women who have memberships in Naypyitaw purchase food and beverages the most out of every demographic group.

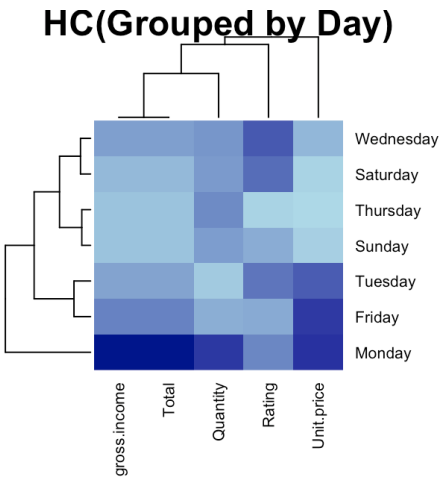
Supermarkets should think about launching marketing campaigns to entice non-members to become members by highlighting the advantages and special offers they could receive. This would help boost sales among female non-members and improve the benefits of membership to maintain the high spending levels of member customers. Supermarkets might also hold previews, tastings, or events exclusively for members, increasing the allure of membership and raising average sales even further. Supermarkets can implement tier-based membership tiers with escalating perks to entice members to 'level up' by increasing their expenditure in order to maintain membership. A few advantages are free shipping, longer return policies, and access to special merchandise. Supermarkets could also provide new members an instant advantage upon enrollment, such as a free product, a one-time discount, or a complementary service, in an effort to increase membership. Another recommendation is to consider offering a trial membership that gives users access to member advantages for a set period.



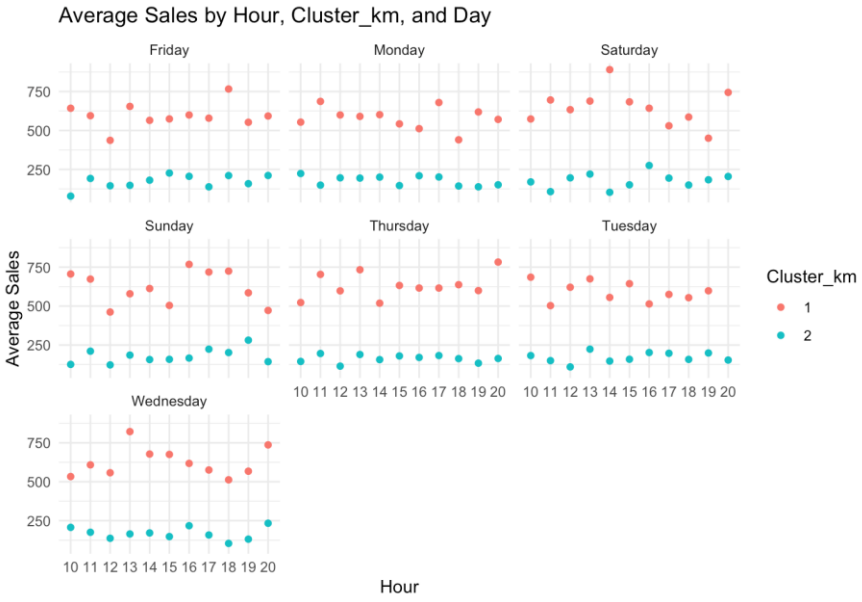
The graph plotted by hierarchical clustering matches the one plotted by k-means clustering, showing that members buy more expensive products in a greater quantity. However, their customer rating experience is less than that of normal members.



Additionally, a heat map regarding the payment methods was created using hierarchical clustering. From this graph, it can be said that credit card customers enjoy their customer experience at the supermarket as they rate their experience higher than the others and purchase a higher quantity of products. Moreover, it is curious to note that expensive items are bought using cash. Since credit card customers are buying a greater number of products, supermarkets can reward them by offering rewards or cash-back on credit card transactions.



This graph represents the sales-related metrics across the different days of the week. Customers buy more goods on Monday and the least on Thursday. To take advantage of this insight, supermarkets can schedule events, discounts, or offers on days that have lower sales, like Thursday or Sunday, to draw in customers. Furthermore, these insights also help inform the supply chain operations of the supermarkets, ensuring that inventory is replenished before high-sales days like Monday and Friday.



This graph shows the average sales achieved, divided by different days of the week and hours of the day, for the two clusters. Sales consistently rise around 2 pm with a slight dip afterward, and then there is a sharp increase in sales in the evening, the time when people leave work. To increase sales in the afternoon and morning, supermarkets can implement happy hour specials to boost sales.

In conclusion, the data provides strong evidence that targeted marketing, an emphasis on membership benefits, and a focus on high-performing product lines could enhance sales performance in Naypyitaw as well as the other two cities. Recognizing the role of gender in purchasing patterns can help tailor product offerings and marketing messages to optimize sales across all customer segments. Supermarkets can make informed decisions to improve sales with these insights and recommendations.