Marketing Segmentation Written Report

Pernalonga's Initial Insights: Marketing Segmentation

1. Business Context

Pernalonga is a supermarket chain which finds itself in a spot where it is reliant on promotions to drive sales, as over 30% of sales derive from promotions. Currently, the majority of the promotions at Pernalonga are in-store promotions. The problem associated with in-store promotions is that they offer temporary price reductions to customers regardless of their need. Hence, Pernalonga is losing part of its revenue by offering promotions to those who would be willing to purchase the items at full price. Because Pernalonga does not have a strong or thorough marketing campaign, they are leaving money on the table by not effectively targeting stores, customers, or products with their promotions. Thus, as the consulting team working with Pernalonga, we will help them better understand their customers and how to develop a strong marketing campaign that will boost revenue through the use of segmentation.

2. Data Preparation & Exploration

Data Understanding

The first step is to understand the data that Pernalonga has. We see that their data contains 7850 customers, 10770 products, 421 stores, which is generated from over 29 million transactions from the beginning of 2016 to the end of 2017. In order to gain a basic understanding of Pernalonga's customers, products and stores, we cleaned the data and calculated additional variables. This step involves multiple parts: merging the two given tables, standardizing all our data, and aggregating data to create new columns for the three segmentations.

Before adding extra columns to the dataset that joins both the product table and transaction table, there are a few special notices and assumptions we made. First of all, we noticed that although each row represents one product transaction made a customer in a certain store, the transaction ID given in the dataset is not unique. Also, we found that after joining product and transaction tables, there are 510 rows of transactions do not have corresponding product information, so some of the rows had missing columns. Yet considering the fact that we have over 29 million rows of data in total, we decided to simply remove the rows with missing data for clustering purpose.

In order to perform better clustering analysis using meaningful attributes, we created the following column grouped based on customer ID, product ID, and store ID. For customers, we created the following columns:

- Discount_freq - how frequently a customer uses a discount
- Discount_rate - the percentage of the money they save on these discounts
- Product_types - a count of unique products bought by a customer
- Total_spending - how much a customer spent in 2 years
- Total_freq - how often they went to the store
- Month - breaking down their number of visits by month
- Weekend - the percentage of their visits that came on the weekend

Likewise, we performed a similar thing for the products. These columns are:

- Mean_price - average price for that product
- Sum_sale_quantity - the sum of the quantity sold in 2 years
- Freq_of_discount_number - how often the product has a discount
- Discount_ratio - the percentage of money saved on the product
- Weekends - the percentage of the product being sold on the weekend

Lastly, we did the same thing for stores. The new columns are below:

- Discount_count - the number of discounts offered
- Discount_rate - the percentage of money that is saved from discounts
- Discount_freq - how often discounts are used at the store
- Total_sales - total sales for that store in 2 years
- Product_types - how many unique products each store offers

Data Exploration

Once we understood the structure of the data, we started looking more in-depth into customers, products, and stores to explore whether there are any trends we can find. For customers, we found interesting patterns in the top spenders. Below, we can see the spending breakdown by product category for these customers, as well as how often these top customers used a discount for their transactions.

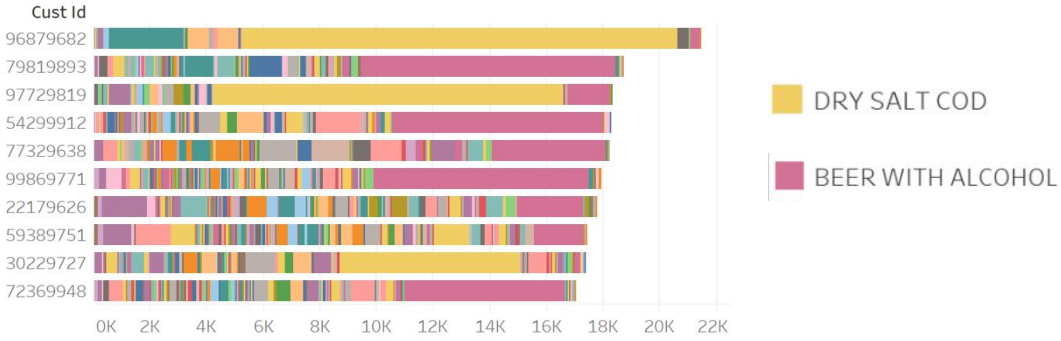


Figure 1. Top 10 customer total sales amount shown in product categories

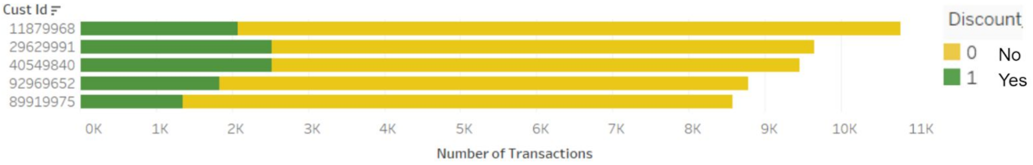


Figure 2. Top 5 customer number of transactions shown in discount frequency

Some preliminary takeaways from these two bar charts were that the big spenders typically have a very strong preference in a particular category, whether it be from buying in bulk or something similar, and do not necessarily prioritize discounts, as these ratios are less than the average for all customers.

When developing personalised marketing campaigns, we might not need to offer too much discount for these customers in order to attract them. Insights like these give us a glimpse into the behavior of Pernalonga's customers before we got into more complex segmentation methods.

Moving on to products, the first step we did was to distinguish the products counted in CT and the products counted in KG. They are different in nature because products counted by weight are generally fresh products instead of pre-packaged ones. From figure, below we can see that the top products counted by weight are actually fresh meats that have relatively high unit price and low sales quantity, such as pork and poultry. Also, our exploratory analysis revealed that there is a large discrepancy of sales, as 10% of product categories actually generated more than half of the sales revenue.

Looking into products' discounts, we found different trends for products sold by weight and products sold in units. Products sold in units typically have medium to high discount frequency (40% to 60%) whereas products sold by weight have a varying discount frequency. We can see that products like fresh meat and fish hardly have any discount but generate a lot of value for Pernalonga.

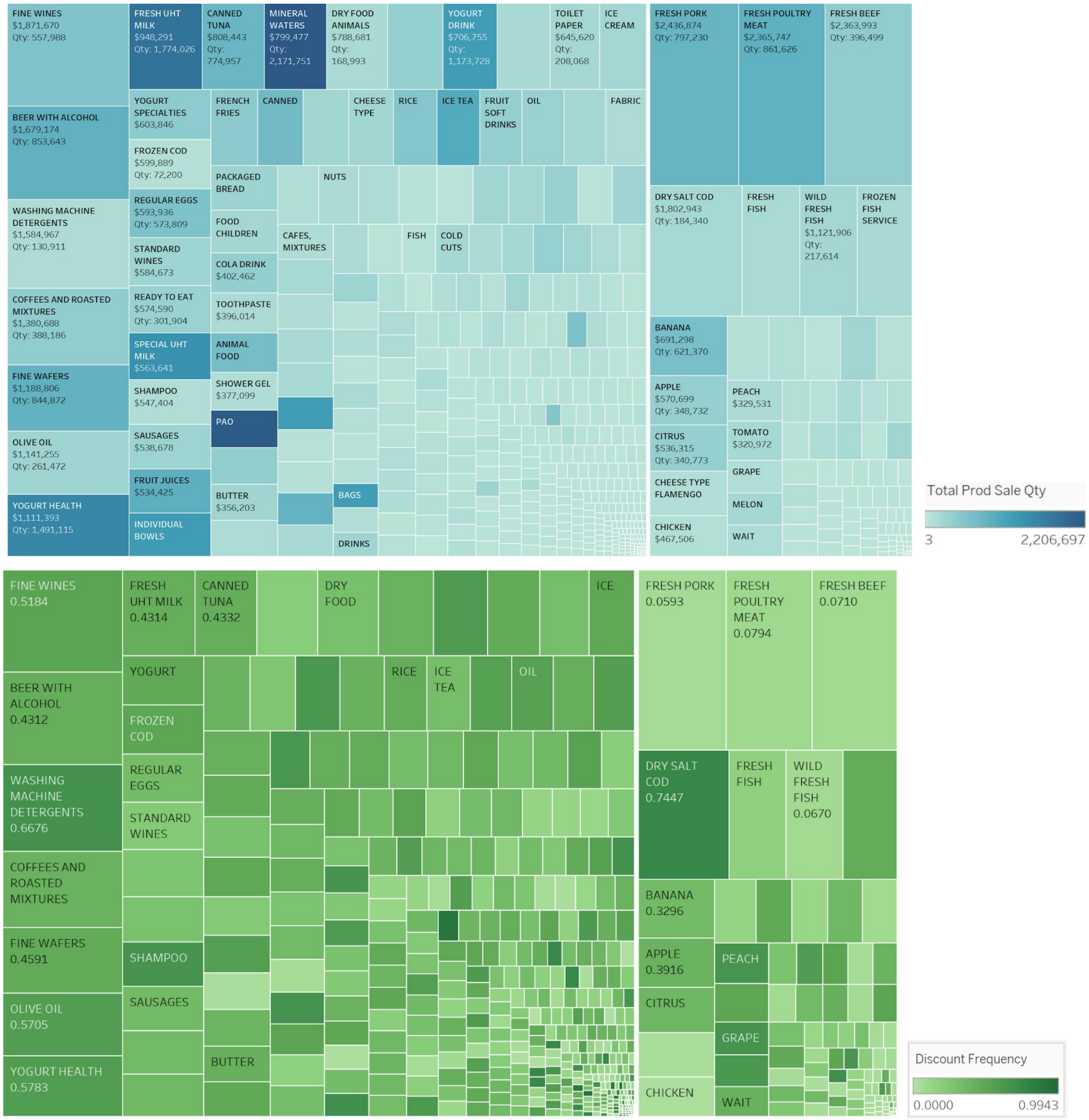


Figure 3. Product treemaps: Sales amount in size and quantity in color (top); discount frequency in color (bottom)

As for stores, we found that not a single store generates exceptionally large sales and those ranking high in revenue typically offer more discounts. Additionally, we see that the stores that offer the fewest discounts typically don't bring in a lot of revenue. Such trends are shown in the figures below.

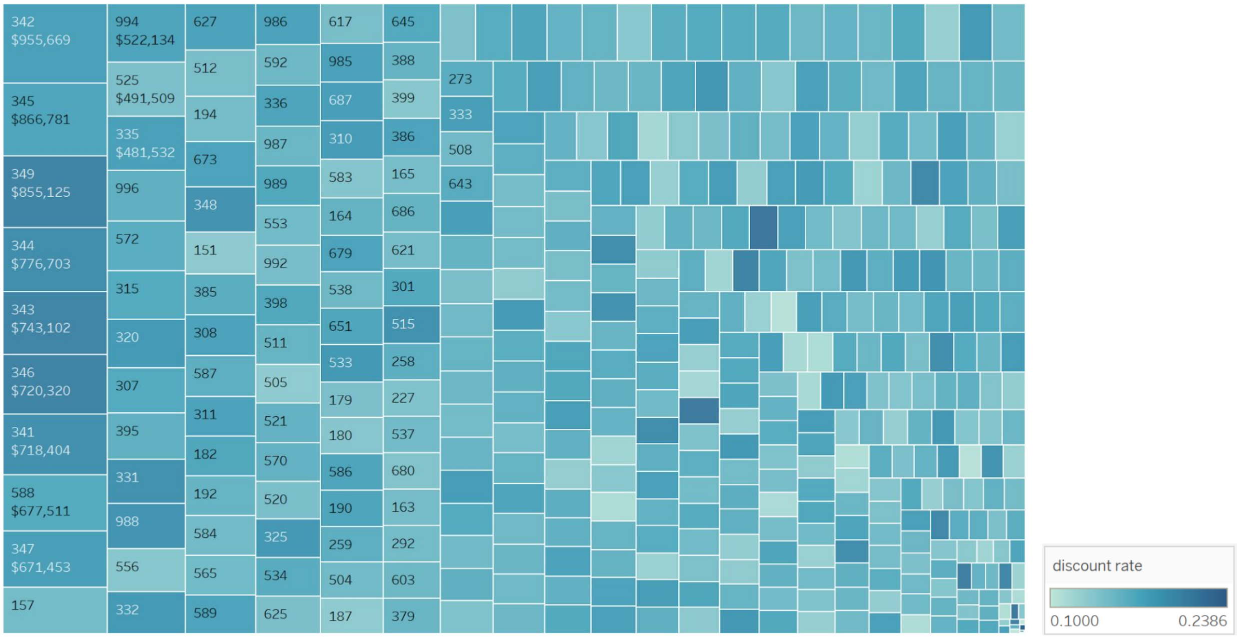


Figure 4. Store treemap: Sales amount in size and discount rate in color

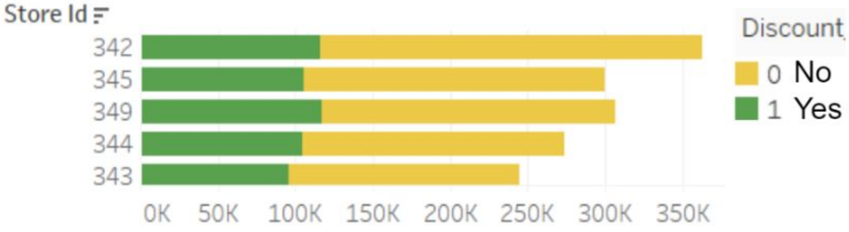


Figure 5. Top 5 store total sales shown in discount frequency

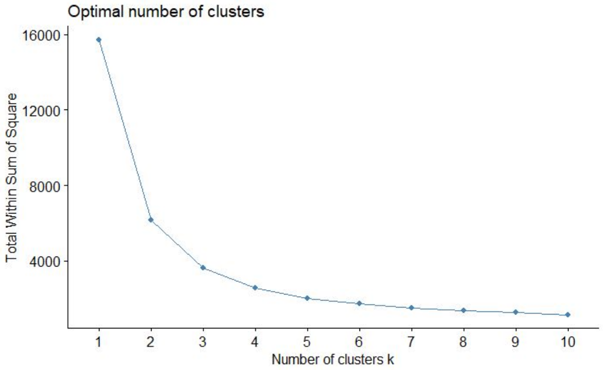
We believe these explorations provided guidance for our segmentation as they helped us better understand Pernalonga's business model and the data they have.

3. Segmentation

For customer, product, and store segmentations, we decided to use k-means clustering method for easy understanding of our results and for the depiction of the business insights. We used the elbow method to identify the optimal k value based on the graphs produced, and we used the ratio between sum of squares within clusters and sum of squares for the whole dataset to determine what attributes we should use for the best clustering. In most cases, the smaller this ratio is, the better the clustering result is as smaller ratio shows that the clusters are more homogeneous within their cluster and heterogeneous across clusters. For each clustering analysis we conducted, we tested different combinations of attributes before arriving at the final results as discussed below.

Customer Segmentation

After experimenting, we found that when only two features are included, discount_freq and discount_rate, the segmentation worked the best and gives the lowest sum of squares ratio at 23%. While there is an argument to be made that more variables should be included, we felt comfortable using only these two as they both are very important factors to consider regarding customer behavior, especially when we are focusing on distinguishing the need of promotions among different customer groups. Below is our elbow graph, showing us our optimal value for k=3, and a visual representation of the customer segments.



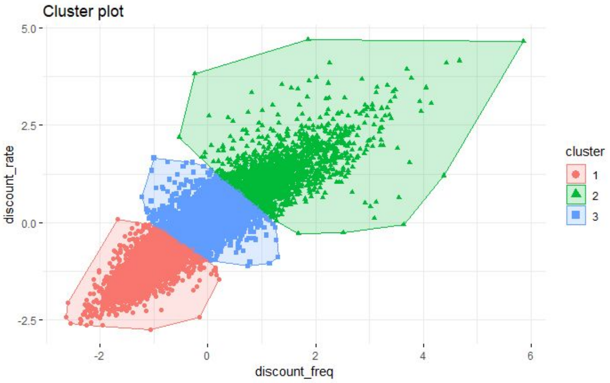


Figure 6. Customer clustering elbow graph **Figure 7. Customer clusters**

Now that we have our 3 customer clusters, we can look deeper and examine some kstatistics about them. Below are the mean statistics of all the calculated attributes at the cluster level.

Table 1. Customer cluster statistics (Mean values of each cluster)

Cluster	**Discount Frequency**	**Product Types**	**Discount Rate**	**Total Spending**	**Total Frequency**	**Cluster Size**	**Comments**
1	23.20%	892	10.98%	8011	49.19%	2504	Loyal customers– lowest discount rates
2	42.40%	1050	21.70%	7605	43.92%	1901	Cherry-pickers- high proportion of discount users
3	31.87%	1016	16.40%	7888	46.10%	3445	Middle of the road consumer (nothing special)

Based on the table above, we can see that cluster 1 and 2 are significantly different as cluster 1 has a very low discount frequency and discount rate (meaning these customers are loyal to Pernalonga), whereas cluster 2 has almost twice higher discount frequency and rate (meaning that cluster 2 contains cherry-pickers who are always looking for a deal). Cluster 3 doesn't have any standout feature, bur represents the "middle of the road" customer because they have an average discount frequency, discount rate, and average total spending. These insights are very helpful when building a marketing campaign for personalized promotions, as it allows to target customers in a certain segment based on their usage of promotions to help boost our profits. Lastly, the box plots of the customers that show the mean, IQR and spread of discount frequency and rate within clusters are shown below. This is done to again depict the differences among the clusters. They provide a clearer visualization on customer clusters differentiate from each other in terms of discounts applied.

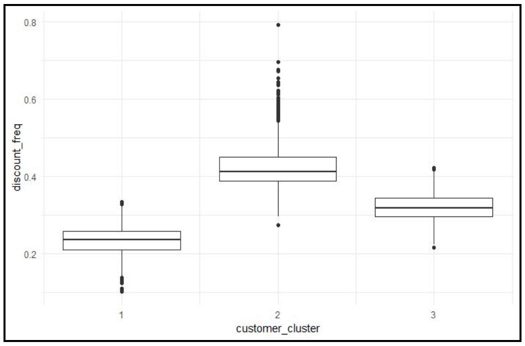


Figure 8. Boxplot of discount frequency

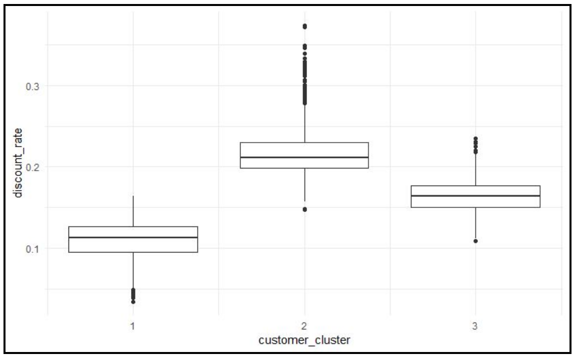


Figure 9. Boxplot of discount rate

Product Segmentation

In order to segment the products using k-means in a way other than simply using product categories and subcategories, we first selected the best-fit attributes which can explain the characteristics and distinctions among the products. Furthermore, it is important to note CT and KG are different units, with KG measuring the weight of vegetables, meats and fruit while CT just represents each countable unit in a pre-packaged item. That is to say, we have to differentiate products between each unit in individual clusters. To begin, we calculated the mean unit price. Additionally, it's meaningful to take product discounts into consideration. If one product has large discount ratio as well as high discount frequency, it could reveal that this product is either a perishable item close to its expiration date or just an unpopular product that doesn't sell well at full price. Last but not least, we include the dates in the analysis to measure the ratio of transaction that take place on the weekend. If the product sales on weekends occupied a large proportion compared to weekdays, it could turn out that these products are preferred in leisure time on the weekends.

In the "CT" unit data set, we found that weekend was an unnecessary addition to the model and removed it. By excluding the "weekend" attribute, we can get the plot (x-axis is k clusters, y value is WCSS) drawn by elbow method as well as silhouette_score. This shows us that our best k for CT products is k=4. On the other hand, the inclusion of "weekend" attributes worked better in clustering model for products sold in KG. Here, the best k is k=3. Below are the elbow plots previously mentioned.

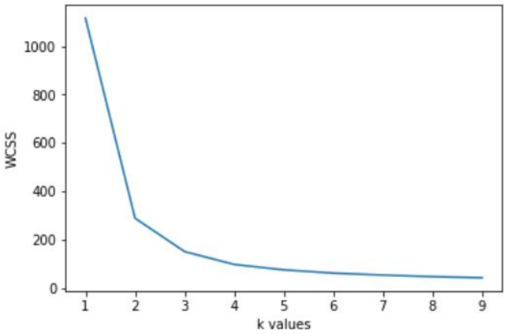


Figure 10. Product in CT elbow graph

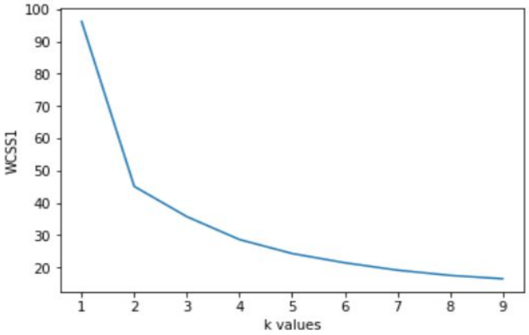


Figure 11. Product in KG elbow graph

The two clustering plots show the sample dots drawn by each k means in CT unit and KG unit. The left graph shows the 4 clusters created in CT plot while the right one shows the 3 final clusters of the KG plot.

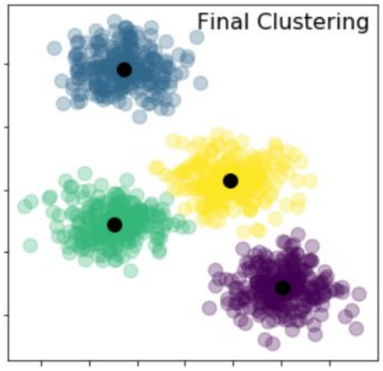


Figure 12. Product clusters in CT

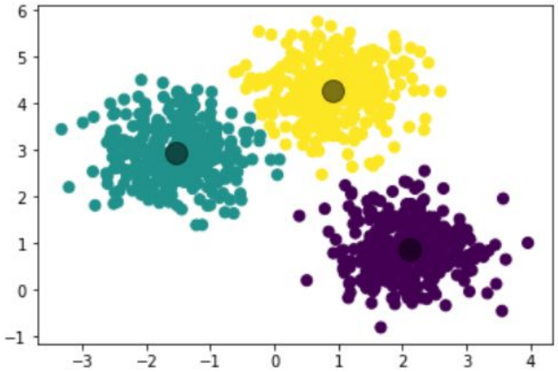


Figure 13. Product clusters in KG

Table 2. Product in CT cluster statistics (Mean values of each cluster)

Cluster	Weekend Visit	Mean Price	Discount Rate	Quantity Sold	Discount Frequency	Cluster Size	Largest Category	2nd Largest Category
1	31.89%	2.90	1.94%	5899	9.70%	3236	Pao Manufacture	Mineral Waters
2	29.50%	8.25	37.64%	2488	77.10%	1939	Yogurt Health	Yogurt Drink
3	31.87%	3.75	9.78%	5530	34.57%	2274	Fresh Uht Milk	Individual Bowls
4	30.35%	5.37	22.15%	2478	58.17%	2293	Mineral Waters	Fresh Uht Milk

From the statistics drawn by the CT plot, there are some interesting points worth the attention. For the second clustering, the mean price of the products is expensive compared to others while the discount rate and frequency are very high. We can see that the main products of second cluster are expensive dairy products that could spoil after a short time. Also, second cluster shows us that customers are more willing to purchase more expensive products with high discounts rather than the original-price products at their original cost. On the other hand, the first cluster is definitely the daily basic necessity with low price, low discount rate, and low discount frequency instead of the high sales quantities. It means that the customers seldom put discount into consideration while purchasing these high-demand products.

Table 3. Product in KG cluster statistics (Mean values of each cluster)

Cluster	Weekend Visit	Mean Price	Discount Rate	Quantity Sold	Discount Frequency	Cluster Size	Largest Category	2nd Largest Category
5	41.24%	11.19	1.71%	820	9.79%	205	Fresh Baked Seafood	Seafood Frozen Service
6	31.38%	4.29	19.44%	22166	56.77%	157	Banana	Apple
7	29.42%	4.99	1.59%	5309	7.08%	673	Fresh Poultry Meat	Fresh Pork

For the KG statistics table, the patterns of the categories are also very remarkable and match the characteristic of each cluster. For the cluster 5, the mean price is the highest with the lowest discount rate and frequency. These represent special products with high value. Judged by the most popular categories, it makes sense if the products are fresh frozen fish and seafood. By looking through frequency of weekend transactions, we can come up with the idea that the weekend sales on these products could be boosted if more promotions take place on weekends. In the 6 cluster, we can assume these products are fresh food like fruit because of the relatively low mean price, high discount rate and frequency with an extremely high demand. These are the “Traffic drivers”. Last but not least, the final cluster shows the same pattern as meet in widey daily consumption.

Store segmentation

To arrive at the optimum number of store segments, we used an elbow graph which showed that the optimal number of clusters was k=4. We can visualize the clusters on a 2-dimensional plane where the dimensions represent total sales revenue and no. of product types.

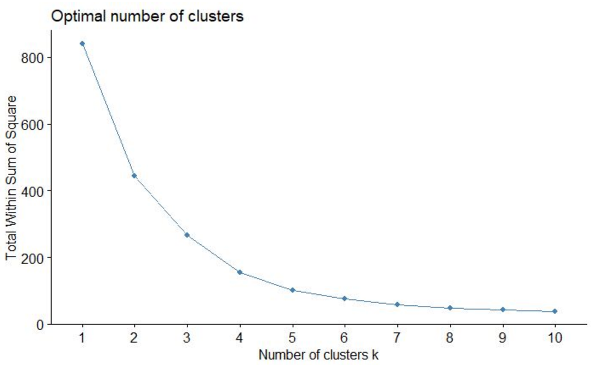


Figure 14. Store clustering elbow graph

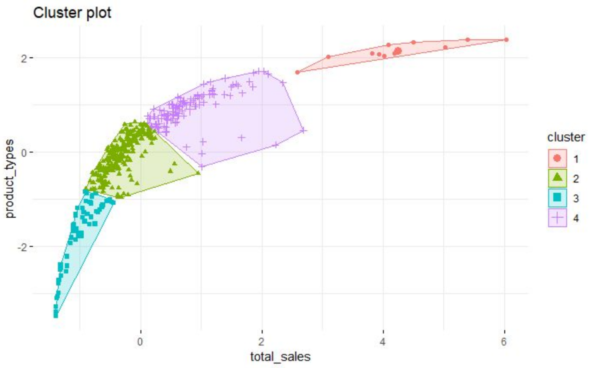


Figure 15. Store clusters

As we can see from the graphs above, the four distinct clusters are distinguished by the average amount of revenue they make and product types they have. We see that there is a large cluster of stores which has the highest average sales and large product base. The next cluster is the biggest which seems to have the average sales and product types, followed by the remaining two clusters that continually trend towards less sales/products offered. The summary statistics of the clusters are as follows:

Table 4. Store cluster statistics (Mean values of each cluster)

Cluster	Discount Rate	Discount Frequency	Total Sales	Product Types	Weekend Frequency	Cluster Size	Comments
1	18.02%	34.90%	597318	9141	40.52%	11	Largest stores frequented by cherry pickers
2	15.72%	30.96%	114515	5539	30.06%	240	“Typical” stores with average sales, product types and discount rates
3	14.42%	29.41%	42569	2822	25.20%	61	Small stores, frequented by loyal customers
4	16.01%	31.09%	234504	7143	31.95%	109	Large stores with average sales

As seen above, cluster 1 has the highest average discount rate and also the highest discount frequency. Moreover, the product assortment is the highest along with a high weekend frequency. This means the customers who frequent these stores, plan their visits with large grocery lists. These are also the customers who are looking for large purchases with typically high discounts (cherry pickers).

Cluster 2, on the other hand, contains the bulk of the stores that have an average product assortment and offer an average number of discounts. They don't generate very high amount of sales because they neither offer the best discounts, nor the best product variety.

Next, cluster 3 represents the small stores which are mostly seeing steady sales every day, hence no spikes in the weekends. They have loyal customers, who are not really looking for discounts, but more convenience. These seem to be the small neighborhood grocery stores who are visited by their customers almost everyday.

Lastly, cluster 4 is the middle of the road segment of stores. It has a third of its customers coming in during the weekends, large enough stores to gather footprint and sales without giving a lot of discounts. These are large stores where people are looking for their product of choice even though they might not get large discounts. This cluster is similar to cluster 2, but contains many more high earning stores, as the only differentiators are a) more products and b) over double the total sales.

These insights are also evident from the relationships between customer clusters and store clusters. As shown in the figure below, customers in cluster 2, who are identified as cherry pickers, actually take up a much larger portion of sales in store cluster 1 comparing to that in store cluster 3.

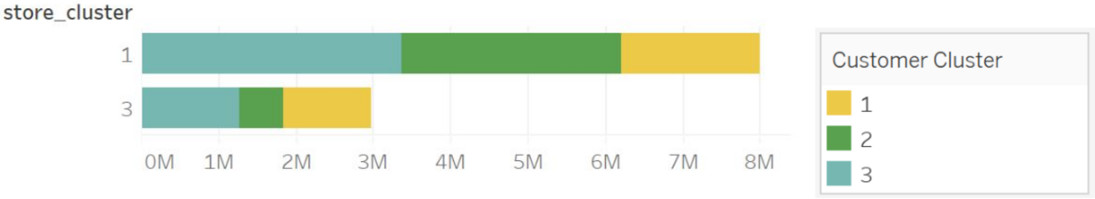


Figure 16. Store cluster 1 and 3 shown with customer cluster distributions

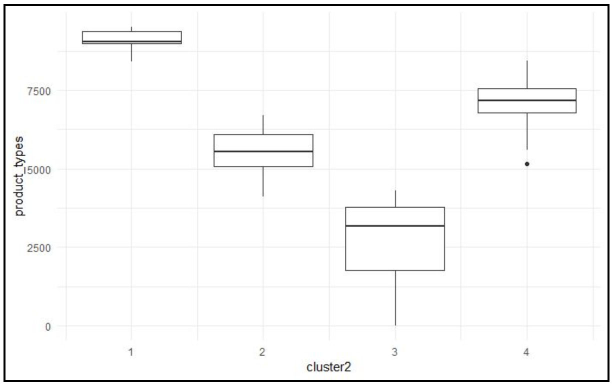


Figure 17. Boxplot of discount frequency

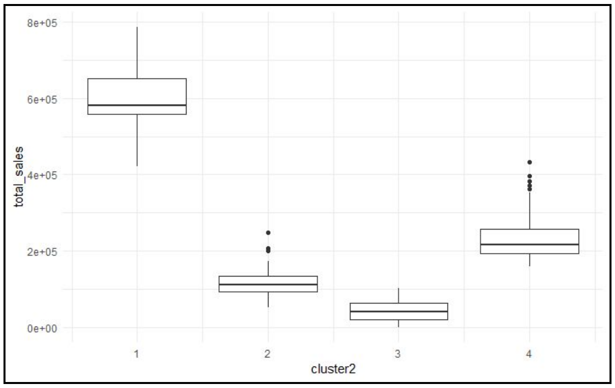


Figure 18. Boxplot of discount rate

Additionally, we have included box plots for these clusters as they depict the mean and spread of product types and total sales for each cluster. As we can see, cluster 1 has a similar number of product types but cluster 3 has a varied range of product types, showcasing the large stores and the small stores. The right box plot actually shows the stores mapped on total sales revenue. Here, we can see that the variance in sales for cluster 1 (large stores) is higher than that of cluster 3 (smaller stores). This could be due to the size of the cluster.

4. Next steps

There are a couple of next steps that we can do after completing this segmentation process. First of all, we can collected more data on users and stores - the data that we obtained from the transaction table was more on behavioral data, and we would like to overlap it with demographic data on the customers to understand their age segments, occupation, income levels etc. We could also obtain store level information like location and how much people spend time there on an average. This could help provide more information for better segmentation. Additionally, there is room for improvement when comparing the price in different categories (e.g., comparing the price between one apple with one electronic device), as we would like to be able to only have a single group of product segments, opposed to two.

Once we have chosen the customer to target to (who), the product to target (what) and the store to target at (where), we can come up with the personalized pricing and promotion (when). This way the approach is holistic and has the highest probability of success. In addition, using these insights, we can successfully design the desired marketing campaign which would boost Pernalonga’s revenue and profits.