Pernalonga Exploratory Data Analysis Report

Harshit Arora, Patrick He, Halle Steinberg, Sean Yu

Background on Pernalonga

Pernalonga is a leading supermarket chain based in Lunitunia that offers a vast variety of over 10 thousand products among more than 400 categories. Its business is currently heavily reliant on promotions, both in-store and through third-party suppliers, with nearly 30% of sales being driven by promotions. While promotions are key in driving customer loyalty and maintaining satisfaction, in-store promotions, in particular, can be problematic to the business in that they offer a temporary price reduction regardless of whether or not the customer needs it. Some customers are happy to buy products regardless of whether a promotion is in place or not, so Pernalonga is losing money on these loyal customers that would purchase at full-price. Pernalonga is in need of a strategic marketing approach that can target customers with personalized promotions, and as the consulting team tasked with this problem, we hope to help the supermarket chain better understand how and what its customers are buying and where they are most often buying from. We hope to implement a thorough segmentation approach that will allow Pernalonga to optimize its promotional spending and in turn boost revenue from its most loyal customers.

Data Understanding and Anomalies

The dataset contains transactional data for 7,920 customers, 10,770 products, and 421 stores during the years 2016 and 2017. A second dataset that contains information such as product category, subcategory, brand name & description is also provided. Since product_id serves as a key to map the two datasets together, we merged the data together to get a holistic view of the data. In doing so, we realized that 510 transactions involved products that do not have data available in the products dataset. Thus, any product segmentation excluded part of these 510 transactions, which span across 3 product_id's

One of the biggest anomalies in the data comes from the fact that there are only 753 unique transaction_id's assigned. Given there are a lot more customers who made multiple transactions, and we

even have rows that have the same cust_id, tran_id, prod_id and store_id (39 records), it was obvious that transaction_id's were not captured correctly and had to be discarded.

We identified a reasonable workaround by defining new transaction_id's based on the following facts and assumption(s):

1. Each transaction can only be linked to one customer at a specific store
2. Each transaction_id is only valid for that particular purchase and should be linked to a specific date

Keeping these facts in mind, we used a combination of customer_id, store_id, and transaction date to define the transaction ids. Doing so leaves us with 2.83 million unique transactions with a varying number of products purchased across customers. We expect that most customers would only make one purchase a day at one store, but this may not be true at all times. Thus, the transaction count is likely underestimated, but should not strongly impact the insights from the data exploration.

While we had data on product unit price, quantity, discount, and the final amount paid for each transaction, there was no data that could help identify profit margins or product costs. However, in order to get directional insights into profits, we identified an approach based on certain industry benchmarks and experience. We analyzed sales for each product (using product id) and ranked products based on their velocity. For instance, high velocity products (those sold in higher quantities) that are being purchased by a relatively bigger customer set (of the 7,920 customers) are likely priced at a lower profit margin than products with a lower velocity and smaller customer base. Using this simple criterion, we categorized each product as High/Medium-High/Medium-Low/Low profit margin. We used relatively broad levels of profit categories as there would be certain exceptions to this rule, such as a high-velocity product priced at a higher margin due to lack of competition, or a high-velocity product priced at low margin because it is perishable (e.g. yogurt) or a low-velocity product priced at lower margins because of low demand (due to seasonality or poor marketing). For this same reason, a benchmark profit percentage (e.g. 10% as low profit) was not used. Customers were classified by their count of H/MH/ML/L product purchases. We took a count of the low profit margin transactions for each customer, as this gives insight

into the customer's contribution to Pernalonga's overall profit. Since high velocity products are typically priced at lower margins, customers that have a high majority of purchases involving low profit margin products will be classified as high profit customers.

Further, we noticed there are 3,927 transactions having a negative or a zero paid amount. While it is possible that these were customers with previous store credit before 2016, it seems more likely that these were data capturing errors. Occasionally, this happened when the discount amount was more than the sales amount, something that is not possible. Thus, we excluded these transactions from analysis.

We also noticed that the paid amount for each transaction was rounded up to two decimal places. This lead to minor inconsistencies in about 33% of the transactions where the actual transaction was slightly less or more than the amount (calculated as unit price * quantity - discount). The magnitude of this difference was *very small* (in order of 10^{-8}), so calculations are not likely to be impacted. However, it is probably best to use the precise transaction amounts as minor inconsistencies across millions of transactions can reveal incorrect results.

Finally, there were two minor issues that came across. First, a store with id 302 had only one transaction in all of 2016 and 2017. The possibility of this happening seems rare, so we excluded this store from the analysis as well. Customer with id 96879682 made a single purchase of 3,371 KG of Cod Fish, with 76 offers in the transaction totalling a discount of $1,400. Given the possibility this could be true, we kept this transaction as is, but it seemed worth noting because this transaction is a rare occurrence.

Looking at expenditure on specific products, we found that customers spend most of their money on foods. For example, we can see that top 3 categories are all fresh meat and sum of these three categories take more than 10% of the total sales.

Top selling product categories

(X axis shows the % of overall sales and Y axis shows the top product categories)

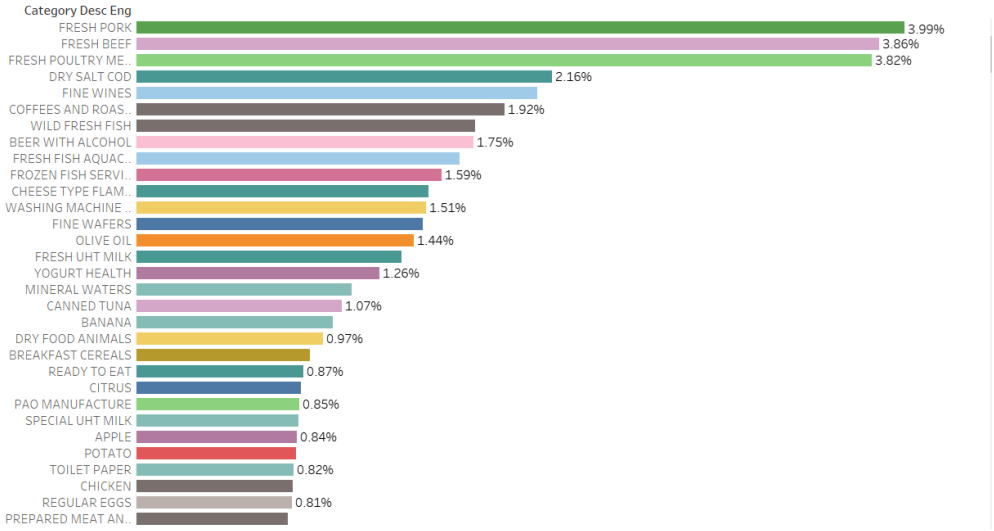


Figure 1.1 List of product categories by market share

Moreover, sales tend to differ by days of the week. For instance, Friday and Saturday saw the most sales across stores, while Monday saw the least.

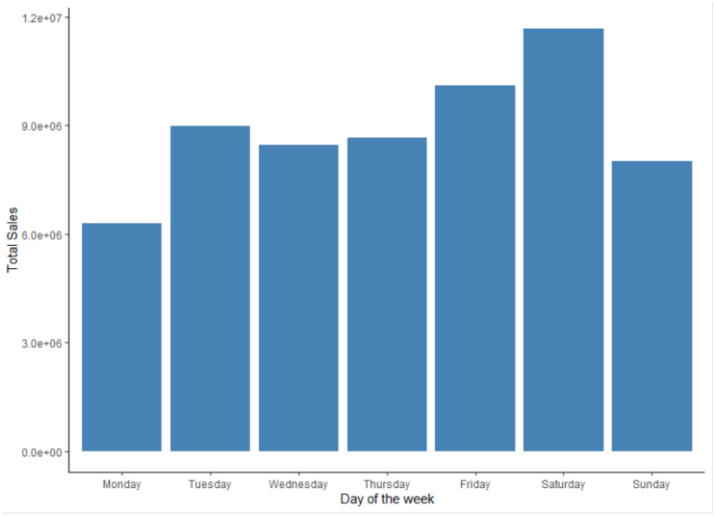


Figure 1.2 Total sales by day of the week

While this chart shows sales at an aggregate level, we found the trend to be similar for most product categories (especially the top ones).

For customers, from the below chart we can see that most of the customers have specific preferences in certain categories. For instance, the majority of customers have high proportions of their overall purchases concentrated in few, specific categories.

Customer spendings proportion on categories

(X axis is how each category takes part for a customer's purchases, Y axis is customer ID)

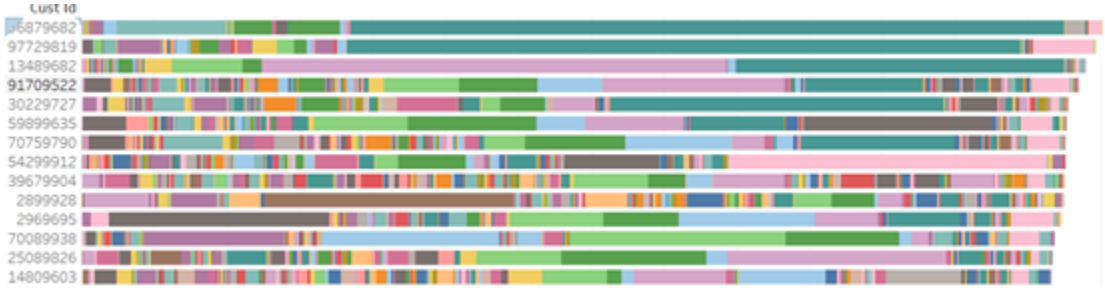
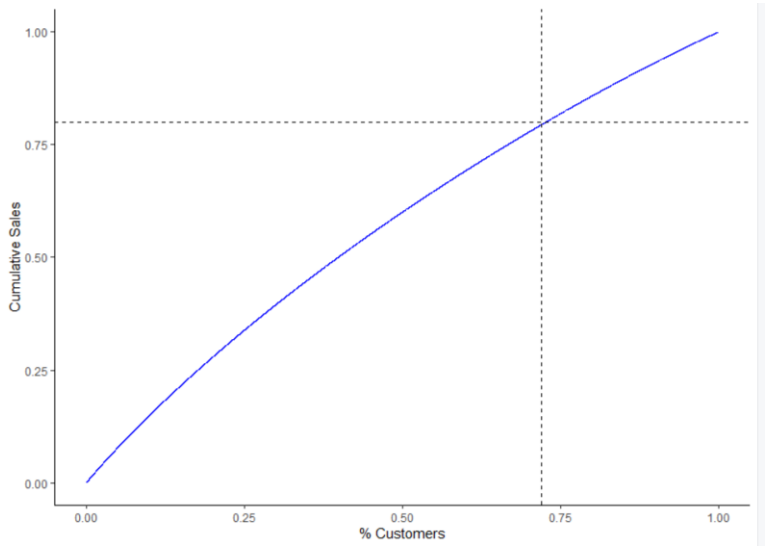


Figure 1.3 Customer spendings proportion on categories

Segmentation

Customer Segmentation



80% of Pernalonga's overall sales are concentrated in 73% of the customers, indicating that sales distribution is somewhat uniform across customers (i.e. no/few high value customers)

Figure 2.1 Cumulative Sales vs. % Customers

Before beginning any clustering of customers within the dataset, we first wanted to create new variables to best capture what the data is telling us about customer behavior at Pernalonga stores. We

created new features for each customer to help us answer questions about who the top customers are in terms of revenue, profit (using a count of low profit-margin products as a proxy), total dollar value of discounts used, number of products purchased, and number of transactions made. We were also interested in seeing how seasonality could factor into a customer's purchasing behavior, so we explored how much a customer spends in each quarter of the year, as well as the difference between a customer's spending on the weekend compared to during the week.

After crafting our features, we experimented with customer segmentation by using the k-means clustering algorithm. We first used the elbow method, which determines the optimal number of segment clusters to create based on minimizing the difference between customers within clusters. The optimal number of clusters is not extremely clear, so we tried methods using two, three, and four clusters. We ultimately decided to use four clusters.

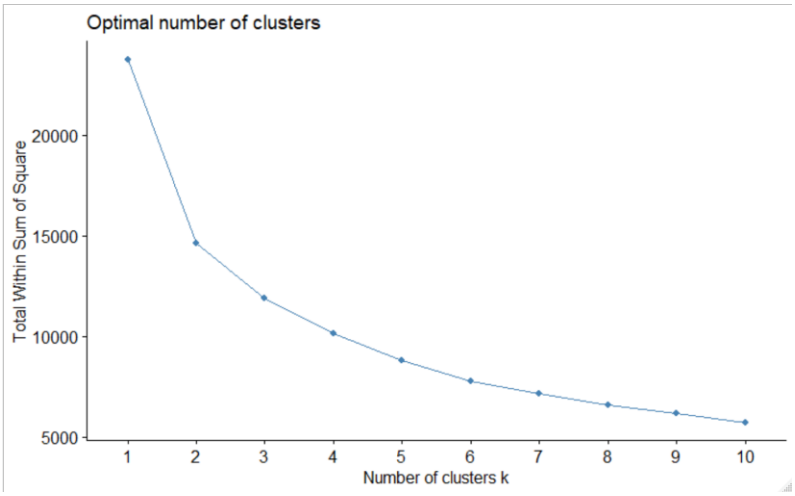


Figure 2.2 Optimal number of clusters for customer segmentation

We then applied the clustering technique to the data and found four customer segments by which to define Pernalonga's customer base. Through some trial and error and experimentation, we found that when we include a customer's revenue, low profit-margin product count, and total discount value, this will result in a segmentation with the smallest within cluster sum of squares, meaning we are optimizing the homogeneity within clusters as much as possible. In this case, the segmentation resulted in a within cluster sum of squares of 42.62%. These features give a clear view into a customer's purchase habits,

especially when our goal is determine the need of promotions among our customers. The graph below gives us a visual representation of the four customer segments. By using a technique to reduce the dimensionality of the data down to only two dimensions (in this case, we have three variables, but can only present two visually), we can see the segments in a two-dimensional space. The percentages on both the x- and y-axis show us the percent of variance that is explained by each of the dimensions.

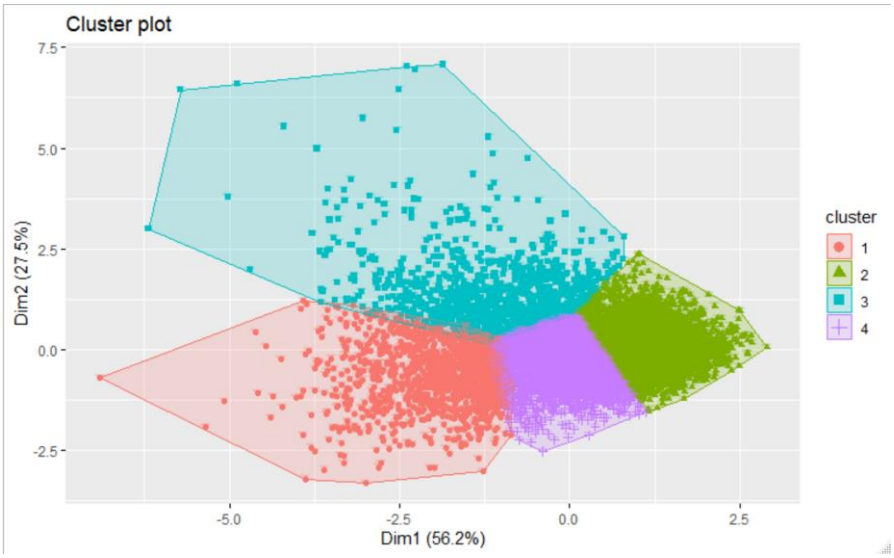


Figure 2.3 Customer segmentation result

After creating the four clusters and assigning each customer in the data to one of the clusters, we wanted to look at what characteristics define each of the three customer segments. Which group contains the high spenders? Is there a group that tends to use more discounts when making purchases? Is there a group that likes to make purchases on the weekend versus during the week? In which segment can we find Pernalonga's loyalty customers?

The table below gives us an idea of the behavior within each segment by looking at the average of various customer characteristics. We saw no significant differences in quarterly spending, and therefore no clear indications of seasonality among any of the segments. Below is an overview of the rest of the metrics. *(Values are averages for that segment)*

Cluster	Weekday Spend ($)	Weekend Spend ($)	Low-prof. margin products	Revenue ($)	Product types	Total discount ($)	Segment Definition
1	$7,142.88	$3,535.14	44	$10,678.02	5,206	$1,786.53	***Pernalonga loyals*** - high spend, high purchase volume
2	$4,390.55	$1,912.75	27	$6,303.31	2,931	$848.00	***Brand indifferent*** - low spend, little product variety
3	$6,072.40	$2,798.64	71	$8,871.04	3,959	$1,111.57	***Basic customers*** - likes Pernalonga, but shops elsewhere
4	$5,462.87	$2,573.69	31	$8,036.55	3,962	$1,294.89	***Frequent discount shoppers*** - high revenue, but also high discount

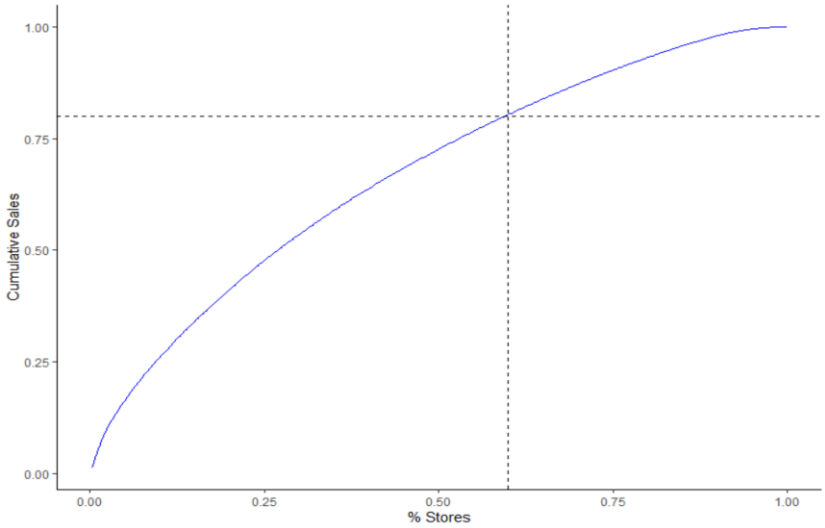
Figure 2.4 Summary of the cluster details

We can gather insights about four customer categories from the table above, including: loyalists who will always go to Pernalonga to shop (segment 1), those customers who might go to Pernalonga if it's convenient or they need something specific, but might shop elsewhere (segment 2), customers who probably do most of their shopping at Pernalonga but aren't high spenders or high discount users (segment 3), and then shopper who frequently use discounts with their purchases (segment 4). Segment 1 has by far the highest values for revenue and product variety, and because these are likely the Pernalonga loyalists, they also in turn have the highest value in terms of discount use. These customers will shop at Pernalonga no matter what. The second segment, on the other hand, spends very little at Pernalonga, likely only shopping there for specific products when it's convenient. Interestingly, segment 2 has the lowest count of low-profit margin product purchases, indicating that they are buying cheap products that won't add too much to Pernalonga's overall profit. These customers might be stopping at Pernalonga for a gallon of milk or toothpaste when they need it, but they aren't doing the majority of their shopping there. Segment 3 are the basic customers - they have middle-tier spending and discount usage, with nothing that uniquely stands out. They shop at Pernalonga, but probably frequent other neighborhood supermarkets as well. Lastly, segment 4 has Pernalonga's discount shoppers. These are the ones that love to shop at

Pernalonga, but will jump at the opportunity of a promotion when they get it. They spend a lot there, but also use a lot of discount as well.

These segments give us a view into customer behavior and help us and Pernalonga to answer the problem of who to target with personalized promotions. These insights allow us to target customers based on the segment they fall into and their current purchase behavior with Pernalonga. With these segments, we can begin to design a tactical marketing strategy and boost revenue and profit from all customers.

Store Segmentation



80% of Pernalonga's overall sales are concentrated in 60% of the stores, indicating that some stores have significantly higher sales than others but a huge disparity doesn't exist

Figure 3.1 Cumulative Sales vs. % Stores

In continuing our analysis, it is important to understand how stores differ from each other. For instance, are there groups of stores that tend to sell products that have higher discounts vs. stores that sell large numbers of products despite lower discounts? To do this, we created relevant store attributes that can add insights, which would be used for grouping stores into segments via k-means clustering. Some of these store-level features are: total sales (quantity and revenue), total number of transactions, average number of items in each transaction, average transaction amount, product portfolio (number of unique products sold), percent of products sold at a discount, average discount amount per transaction, and discount to revenue ratio (ratio of total discount offered and total sale amount).

The discount-related features seemed to be very useful in segmenting stores. While we calculated many useful store attributes, only three features were used in the segmentation process: average transaction amount, percent products sold at a discount, and discount to revenue ratio. All other relevant attributes were used to see the similarity/difference between stores after segmenting by comparing average feature value across segments.

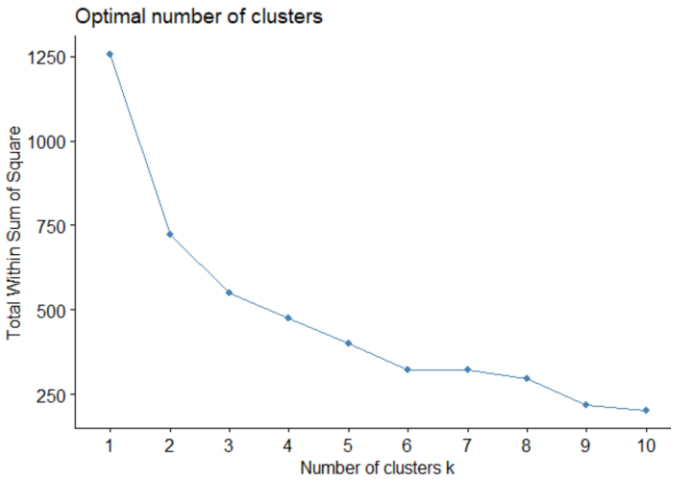


Figure 3.2 Optimal number of clusters for customer segmentation

The elbow method (a technique that helps identify the ideal number of clusters/segments by minimizing the variance of stores within clusters) revealed that there should be three or four natural groupings. We finally decided on using three clusters as this revealed the best insights from groupings.

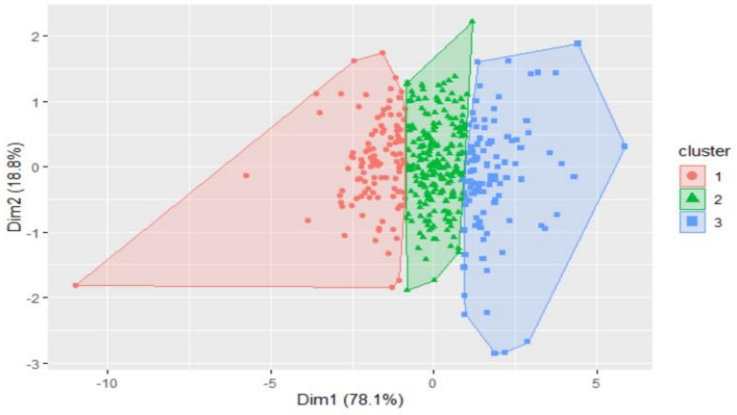
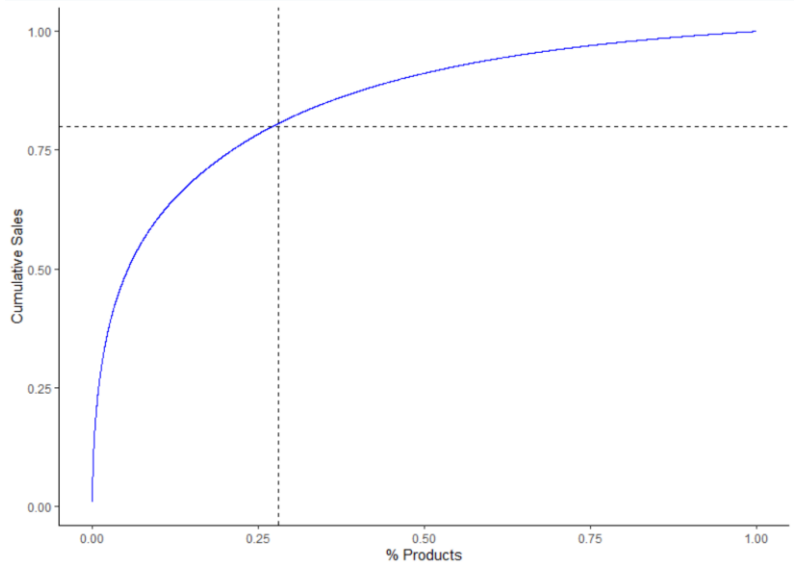


Figure 3.3 Store segmentation result

Given that there were 3 features used in k-means, dimensional reduction (principal component analysis, a technique used to reduce the dimensions of store attributes from three to two) was performed to plot the clusters in just two dimensions (X & Y axes). The table below gives us an idea of the store characteristics within each segment by looking at the average of various store KPIs.

Cluster	Types of products	Average transaction amount ($)	Average discount/ transaction ($)	Percent Products on discount	Discount/ Revenue ratio (%)	Segment Definition
1	4,984	17.9	5.6	27	13	***For Pernalonga loyals*** - Stores visited by loyal customers despite lower discounts and low product variety (which can explain lower spend)
2	5,950	22.1	7	31	16	***The Versatile Stores*** - Highest variety of products, relatively high discounts
3	5,836	26.8	10.7	35	18	***For Cherry Pickers***- Stores visited by customers targeting high discounts & variety. Spend the most and probably buy in bulk from these stores

Product Segmentation



80% of Pernalonga's overall sales come from only 28% products in their entire portfolio, indicating that some products sell much more than others (as seen earlier)

Figure 4.1 Cumulative Sales vs. % Products

Similar to the analysis used for customers and stores, we first created new variables to better measure the varying sales & discount patterns of the products. To find which products and product groups have the highest volumes, revenues, profits, and transactions, we created variables like product sale quantity, product revenue, number of transactions. percentage of discounted transactions, average discount rate, and mean sale price. We also wanted to measure the profitability of a product and a product group, so we used the criteria defined earlier to approximate profit based on logical assumptions of product quantity sold and penetration.

We noticed that there are two product units, CT and KG, which represent two different grocery products. To ensure the accuracy of the segmentation, we decided to divide the products into two groups based on their unit type and further segments the product within the two unit types. To find the optimal number of clusters, we again used the elbow method and also tried to adjust the variables used to determine the best segmentations for Pernalonga's products. We ultimately decided to create three clusters each for both the CT group and the KG group.

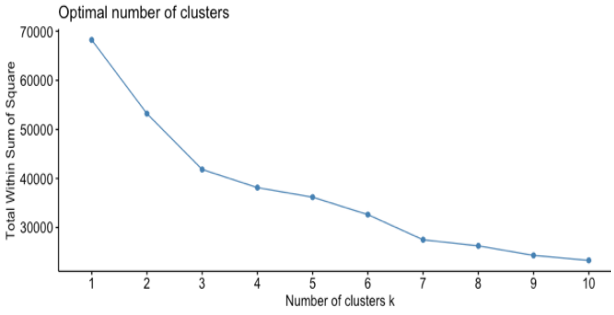


Figure 4.2 Optimal number of clusters for CT group

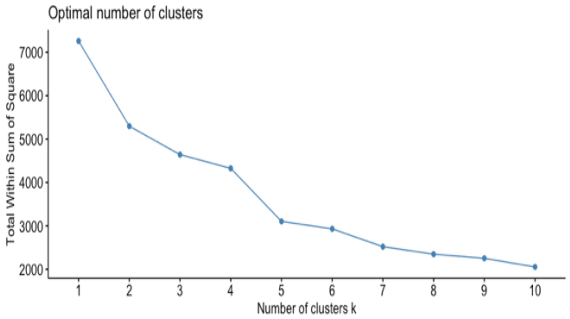


Figure 4.3 Optimal number of clusters for KG group

Once we determined an optimal number of clusters, we applied k-means clustering and found three segments for each product group, ensuring that products are homogeneous within clusters and heterogeneous across different segments. From the results shown below, we can see that there are some outliers in both groups, and we wanted to see what might be the reason behind these extremities.

We first reviewed the top ten products by sale quantity in segment 2 of the CT group and found that the top-selling product has a sale quantity that is much higher than others. We found that this product is a top-selling paper shopping bag produced by SACOS. We believe these shopping bags are placed at the check-out counter as one of the bagging options for the customer, and to promote the green lifestyle, customers will pay for the reusable shopping bags that can be used again in the future.

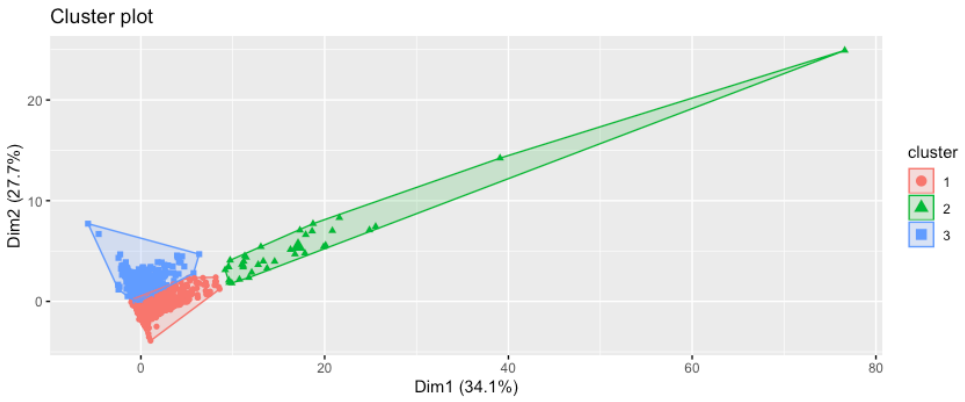


Figure 4.4 Clustering result for CT group

In the second cluster of the KG group, we can see another product with a significant quantity sold. This product is an imported banana produced by FRUTAS&VEGETAIS. Bananas have been top-sellers in the United States for a very long time, and it makes sense to include it in this cluster since its sale pattern is unique compared to the other two product clusters.

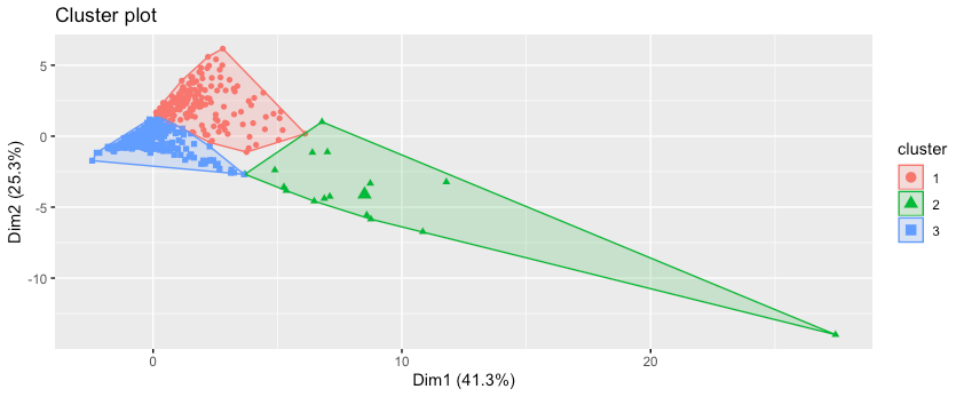


Figure 4.5 Clustering result for KG group

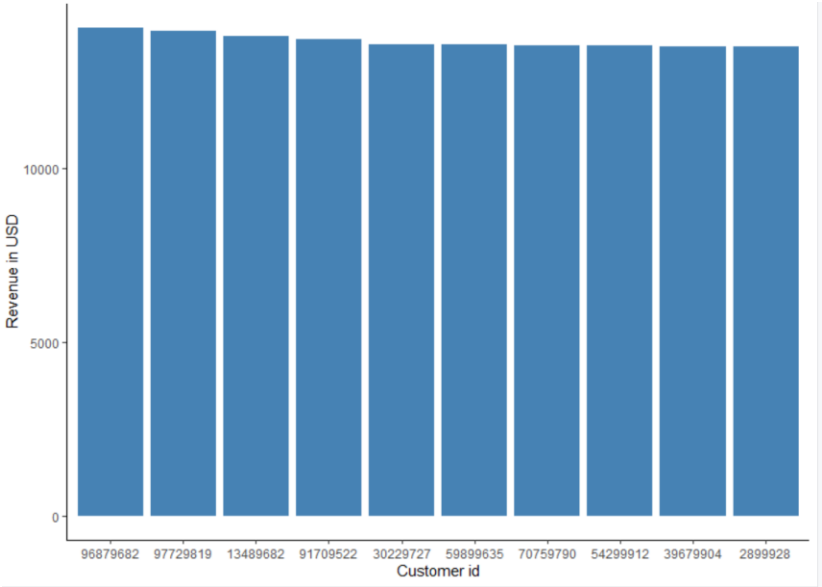
For the six clusters across the two product groups, we defined cluster types as Always Promoted, Seldom Promoted, Preserved Value Driver, and Traffic Driver, based on the average values of the key features in each cluster. The features of Always Promoted and Seldom Promoted clusters are obvious. We can observe from the result that the distinction is that products in the Always Promoted cluster are constantly on sale and offered with great discounts. Top selling products in this cluster include iced tea, bakery beard, etc. Alternatively, products that are commonly used and always popular with shoppers are concentrated in the Seldom Promoted/Preserved Value Driver cluster. Products in this cluster are rarely on sale and tend to remain at a constant sale price regardless of economic factors or differing sale patterns. Such products include beer, yogurt drink, mineral water, etc. Finally, we also have a Traffic Driver cluster which consists of products that are always in high demand. The price of the Traffic Driver products are usually very low, but their sale quantity and number of transactions are significantly larger than products in the other two clusters. Bread, croissants, and bananas are identified as Traffic Drivers in Pernalonga's business.

Cluster	Product Unit	Product Sale Quantity	Number of Transaction	Product Revenue ($)	Percent Products on discount	Average discount rate	Mean price after discount	Profit Score (1-5)	Segment
1	CT	4301.90	2,472	$4,207.91	20%	5%	$2.88	2.84	***Preserved Value Driver***
2	CT	265388.74	90,693	$109,292.48	27%	6%	$0.92	1.00	***Traffic Driver***
3	CT	2217.85	1,343	$3,796.2	66%	29%	$4.73	3.19	***Always promoted***
4	KG	15663.99	15,560	$30,161.18	55%	19%	$3.33	1.95	***Always promoted***
5	KG	145048.71	12,3244	$331,803.41	23%	7%	$3.53	3.53	***Traffic Driver***
6	KG	2762.88	3,195	$11,147.24	7%	1%	$6.55	3.41	***Preserved Value Driver***

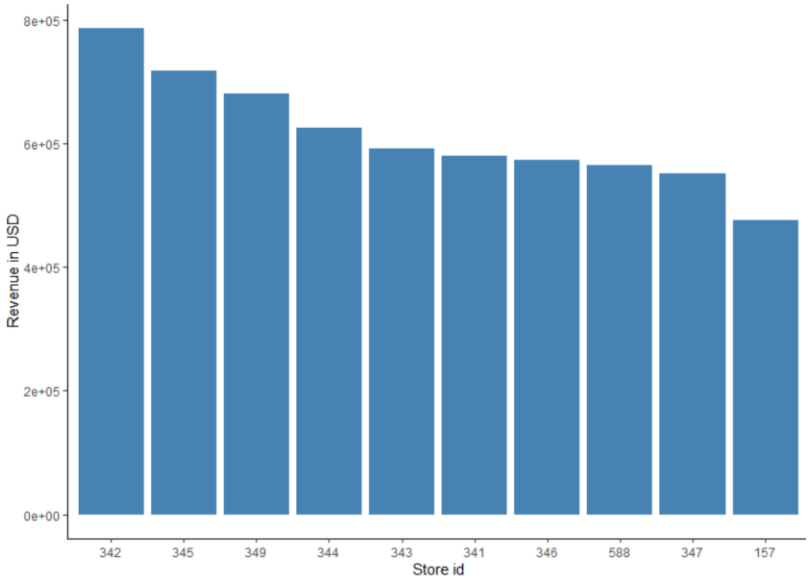
Figure 4.6 Summary of cluster details

Appendix

Top 10 Customers by Revenue (Total paid amount)



Top 10 Stores by Revenue (Total paid amount)



Top 10 Products by Revenue (Total paid amount)