Executive Summary:

The current population for Portugal is 10.1 million and its improved economic environment has increased consumer confidence and expenditures and contributed to better retail sales. The Portuguese food industry sector has gone through a great deal of change over the past few years as new competitors are introduced. Hypermarkets and supermarkets remained the largest channel in Portugal as slow economic growth and weak purchasing power encouraged Portuguese consumers to look for private label brands and value deals, which resulted in high sales in the channel.

The purpose of this analysis was to break down transaction information of about 8,000 customers concerning the importance of 10 store characteristics in selecting their primary grocery store. This analysis uses k-means clustering to conduct the segmentation of Portuguese supermarket shoppers and stores. For a store to be successful, it must offer a good balance of variety of products and affordable prices to all consumers. Each customer was clustered on 5 variables: amount of product purchased, amount spent, number of items in a transaction, number of coupons used, and how much of a discount they received. From this, we were able to segment the customers into 4 segments - volume discounters, bargain hunters, average (need-based), and all-in-one-go buyers, and provide insight and recommendation consisting of product placement and customer targeting to improve the current marketing strategy of Pernalonga.

Background:

The Portuguese food industry sector has seen dramatic changes in the past few years driven by a rise in the introduction of e-commerce channels and demographic changes. Broadly, the food industry can be divided into the retail food sector, comprising sales of food items consumed at home, and the food service sector, comprising sales of prepared food that is consumed away from home. In such a climate, where food retailers are facing increasing competition not only from other retailers but also from the food service sector, the fight for the modest increase in consumer dollars spent on retail food is intense. Understanding consumer preferences and the drivers behind these preferences is crucial for any supermarket chain to differentiate itself.

Since the early days of commerce, brands have been asking important data questions. Initially, the big question revolved around “How do I learn more about my customers?”. That question has since been answered because that data now lives everywhere. From swiping your card at a retail store or restaurant to just visiting a brand’s website, or even geolocation through mobile apps. The well of

customer data sources is far from dry. But in a world where traditional bases of competitive advantages have dissipated, analytics-driven processes are one of the few remaining points of differentiation for firms in any industry, especially for major supermarket chains.

Businesses have understood that a unique customer experience is their competitive weapon that boosts conversions in the short term and builds customer loyalty in the long term. And as companies need to know their customers to create such experiences, customer data analytics has become more important. Customer segmentation affects a company's core marketing strategy, because your product, pricing, placement, and promotional strategies all depend on your client segmentation. Segmentation facilitates the company's ability to find the most valuable client types for its business and effectively market to them.

To this day, Portuguese residents seek out goods with the best quality-price ratio, and there are no signs of this behavioral trend going anywhere. A study published by in-Store Media and Netsonda, examined consumer shopping habits in Portugal and found that the country's shoppers tend to 'shop around', with 73% saying that they shop in more than one supermarket. We can see that the promotional environment has grown, and with this consumer expectations around private labels have also grown. Nielsen reported that own-brand growth was higher than mainstream-brand growth in 2017. As Portuguese consumers show a greater appetite for promotional campaigns, this analysis aims to help Pernalonga better understand their customer's needs to develop micro-marketing strategies to help them boost revenue and gain customer loyalty.

Method:

The Data

There are two datasets involved in the analysis of this project. The first dataset contains transaction history from 2016-2017 for approximately 8,000 customers. This dataset includes information like date of purchase, items purchased, number of items purchased, and dollar amount spent (before and after discount). Each row in the dataset represents one product purchased in a particular transaction by a particular customer. The second dataset contains information about the products stocked at the stores such as brand name, type of product, and product subcategory. Here one row in the dataset represents the description for one particular product_id. While the dataset is not missing any data, it is unbalanced across stores. There are some stores where we have only 2 transactions worth of information, while others have over 10,000 transactions. We decided to leave the store information as is because during our cleaning process most of this was taken care of.

Data Cleaning

We started the data cleaning process by looking at the structure and summary statistics of the data. We noticed a lot of the columns that were “ids” were imported as integers so we transformed them into factors as the values have no real significance other than being a unique identifier.

After initial exploration, we found that “Private Label bags” were bought with almost every transaction. Our preliminary understanding was that these were plastic bags available to take the products home, but were not sure if they were included in the store revenue or not. Upon some further research, we found that a plastic bag tax was implemented in February 2015 to reduce the consumption of plastic grocery bags in Portugal and in turn reduce the potential contribution to marine litter. The money collected from this has no impact on store revenue and therefore our marketing strategy, and so we removed this information from the dataset.

We then looked for any transactions with negative paid amounts. There were 8 transactions where the amount paid was negative (the discount amount was larger than the paid amount). Our assumption here is that this could be a return that was discounted to return the initial money paid. We mapped the return to the original purchase transaction and removed both the original transaction and the return from the dataset to prevent reporting higher sales than the actual amount.

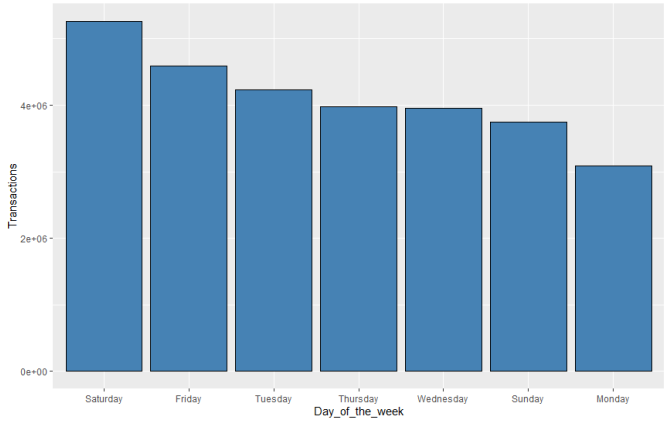
Upon grouping all the transactions by *transaction_id* we found that the *transaction_id* in the dataset does not uniquely identify each transaction due to it being imported as a large number. We created a new unique identifier for each transaction by concatenating the c*ustomer_id, store_id*, and the transaction date. Our assumption in doing so is that each customer only visits a store once during one day. This enables us to see purchases made in a single transaction giving us insights into units per transaction and dollars per transaction.

Lastly, we verified all the transactions were calculated correctly by subtracting the discount amount offered from the original transaction amount to ensure there were no data entry issues. Additionally, we also ensured the total amount was entered accurately by multiplying the unit price by the quantity purchased. All rows passed both the quality checks and so we could proceed with our analysis.

Data Exploration

To understand the data, we structured our exploration between customers, products, and stores. To begin, we studied the distribution of transactions across the week for Pernalonga stores. We

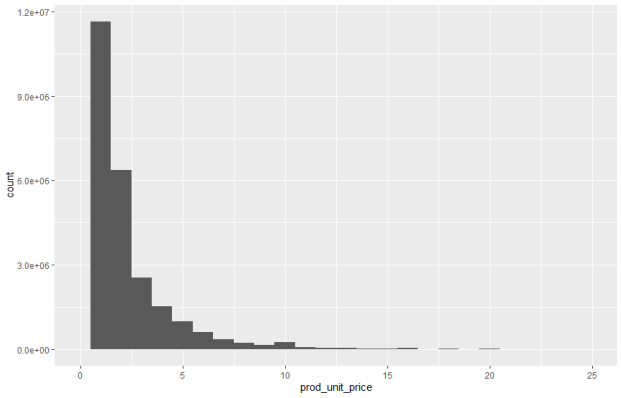
extracted the day of the week from the transaction date variable and discovered that Saturday was the busiest day of the week followed by Friday and Monday had the least foot traffic. The amount of revenue made by the store was correlated to foot traffic and thus, Monday was the slowest day in terms of revenue. This was very useful for our analysis as we could incentivize customers to come in on slow days by providing good offers on products.



We studied which customers had the most transactions, overall, and based on the store they shop at. This provided us with insight into what customers were loyal to a particular store in their purchases. However, we wanted to verify whether the customers who were most loyal to the store were also customers that provided the store with a lot of revenue. The union of these two groups of customers provided us with the desirable segment to target for loyalty programs. Additionally, we also identified the customers who made most of their purchases using coupons and offers. We recognized that there were two segments of customers that preferred to do this - those customers who mostly bought their products in bulk and those customers who preferred shopping for regular, price elastic goods such as bananas, milk, eggs, bread, beer, coffee, etc. We recognized these groups as ideal targets for coupons and offers.

To understand the stores, we explored the stores with the highest revenue and used them to understand the customer's preferences and needs. Additionally, we looked at what kind of discounts were being offered per store and ascertained that the median discount (per store) to all customers was about €23093 and the median discount per product offered was 0.79 cents. This indicated that most customers clubbed together multiple offers during each transaction. Lastly, we identified all the stores that offered the most discounts and identified what products they usually discounted during sales. The most discounted products sold at these stores were carrots, mineral water, bread, onion, fresh milk, citrus fruits, bananas, and zucchini. As we can see from this breakdown, these are regular products bought weekly for consumption and thus, are usually discounted and well-received by customers. It is also inferrable that these goods are price-elastic because the same customers buy these goods across stores and thus, the discounts play a big role in determining what stores the customers visit for these products.

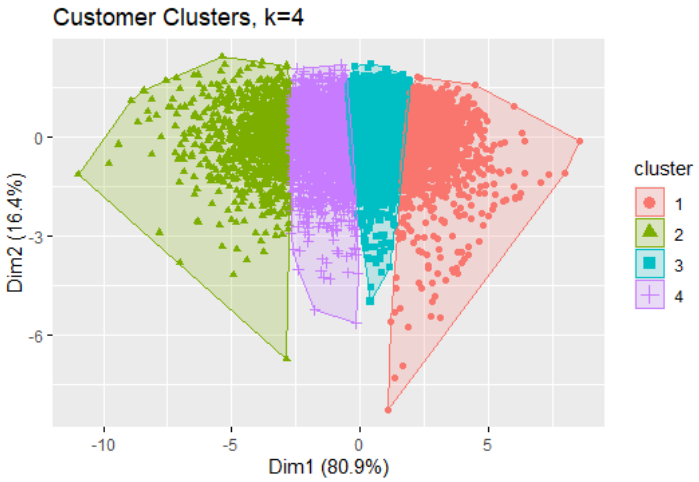
Lastly, we explored all the products stocked in the stores. We started by identifying the most popular products purchased by customers. This gave us similar insights from what we explored with stores and customers and weighed heavily in the direction of price-elastic goods. We also studied the distribution of prices per product to identify the skewness in prices and recognize the most expensive products. Some of the grocery stores also sold kitchen electronics and appliances such as mixers which led to a right-tail in the distribution. The exploration process gave us considerable insight into customers, stores, and products, and also provided us with a path towards segmentation to allow for better marketing practices. We proceeded with this insight to create variables that could be grouped and would provide us with clusters in the market to better direct Pernalonga's marketing strategy.



Data Analysis

To gain a better understanding of the customers, stores, and products we decided to use kMeans Clustering to segment the population to give us more homogenous groups. We started by engineering new variables based on which we wanted to cluster our customers. We segmented customers on 5 different variables:

1. How many units of a product they purchased: as there are two different unit types (Count and Kilogram) we consider the total units of each respectively
2. How much money they spent (after discount)
3. How many items they purchased in 1 transaction
4. How many coupons they use per transaction
5. How much of a discount are they getting

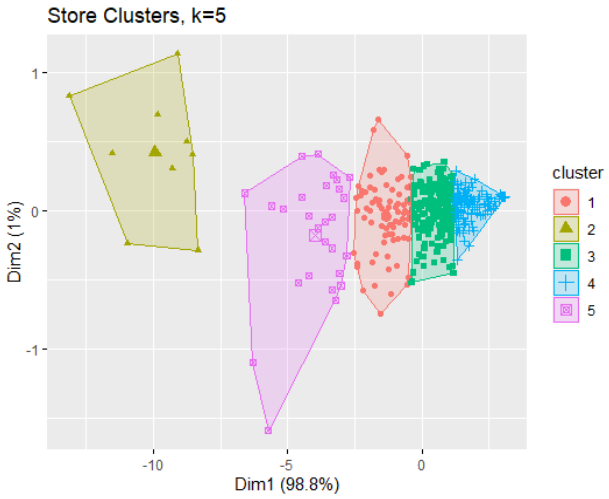


We then scaled the data and ran an initial clustering with 2 centroids. This clustering was far from ideal as it was stretching the

boundaries to include some outliers. We ran an Elbow Method analysis and an Average Silhouette analysis to see what would be the ideal number of clusters to consider. Upon this analysis, it was clear that 4 clusters would be ideal and so our final analysis includes 4 customer segments.

We also segmented the stores so we could get a better understanding of what services the store could provide. Similar to the customer segmentation, we performed data engineering to get some variables to segment on:

1. Number of transactions processed at the store
2. Number of total units being told: again, we have split out count and kilogram to account for the difference in the unit types
3. How much revenue they are generating: this helps us understand if they're a low/medium/high volume store
4. How many discounts they typically offer



Similar to the customer segmentation, we scaled the data and began our analysis with 2 centroids. After running Elbow Method and Average Silhouette, we found 5 clusters were ideal.

Customer Segmentation

The kMeans clustering analysis provided us with four clusters for the customers in the Pernalonga dataset. On studying these four clusters, we created the following strategy to target each customer segment based on their purchasing pattern.

Volume Discounters

The first cluster of customers consists of those who buy large quantities of produce in weight and account for about 22% of the total customers. Their item quantity count is not very high but they prefer to shop in bulk. This is also paired along with a large discount value they avail for every transaction since their cart value is rather high. These are not customers who use a lot of offers but the limited offers on their bulk purchase reduce the transaction amount greatly, thereby making them the cluster that spends the least amount of money out of all the clusters.

Bargain Hunter

This segment of customers buys the most number of products and also utilizes the highest quantity of offers for these goods. Despite using the most offers, their total discount amount is the lowest. This is an indication that they usually buy affordable and regular-priced products that do not offer large discounts. These are also the customers who spend the most out of all the segments and thus, could be regulars that visit the stores weekly for their essential purchasing of price-elastic products and thus, prefer to use discounts for them. This group of customers accounts for about 12% of the total customers.

The Average (need-based) Customer

This cluster includes the average customer that visits the store on a need-based basis and makes up the majority of customers (38%). They do not purchase very high quantities of products and this affirms that they only visit the store when they require something. Their purchasing behavior cannot be predicted easily as it is erratic and they only buy the products they need at the price allotted with little to no discount. We can affirm this as their transaction amount for the order is barely different from the amount they pay to the store, thereby indicating no discount amount.

All-at-one-go Buyers

The last cluster of customers is cyclical but the most infrequent. They purchase very large quantities of produce at long, spread out intervals. They account for 28% of the total customer base and are the second most common. It appears that they stock up on their produce for a long amount of time before heading back to the store and thus, prefer to purchase everything they require in one transaction. This could be an indication that they probably do not purchase many perishable goods, however, we cannot confirm that yet. They also do not avail any big offers or discounts on products. They do spend more than the average, need-based customer, and thus, are not as erratic to predict.

Store Segmentation

The kMeans analysis provided us with five clusters for store segmentation. These cluster characteristics provided from the analysis simply clustered stores on the basis of number of transactions and ranked them from one to five such that stores with the most transactions produced the highest revenue and also ended up accruing the most discount. In this case, the five attributes we used to generate clusters can be linearly predicted from each other. Due to this segmentation, the analysis did

not provide us with enough insight to target each segment individually as each variable appeared to be correlated to the other variables in the analysis.

Insights:

Using our exploration and interpretation from the kMeans clustering analysis, we can focus our marketing strategy to target Pernalonga's customer segments highlighted above. Since the volume discounters purchase the largest amount and spend the least amount from all the customer groups, it is important to direct our strategy towards them so they spend on other products as well that are not discounted. Pernalonga could utilize the concept of "loss leaders" by continuing the discounts on these bulk, essential goods but change product placement in their stores such that in order to reach the aisles that contain the loss leader products, they must pass by all the other attractive, higher-priced products. This can tempt the customers into picking up other products that are not heavily discounted, thereby increasing their cart value.

For the more infrequent customers such as the need-based and all-in-one, Pernalonga can utilize their purchasing history to run predictive analytics and send newsletters and personalized recommendations. These can serve as regular reminders about ongoing offers to incentivize the customers to shop more. Additionally, for the need-based customers who buy fewer products, the stores should ensure to accommodate a quicker check-out line or self-checkout options for smaller carts to not hinder their shopping experience. Since the all-in-one-go customers shop in large quantities but not very regularly, it could be possible that they prefer to shop at other supermarkets or stores for their daily essentials. We have already confirmed from our data exploration that Pernalonga stores heavily discount these daily products such as eggs, milk, bread, vegetables, etc., so they should target these customers better by bringing such discounts to their awareness.

We also discovered from data exploration that weekends were the most popular time for shopping whereas Monday was the least busy. Stores can choose to provide exclusive discounts on Mondays to spread foot traffic across the week and take away pressure from lower stock on weekends. Additionally, the stores can choose to withhold certain discounts on the weekend since they do not need to provide an additional incentive on busy days. The price-sensitive customers will change their schedule to adopt the new routine for offers, however, the rest of the customers unable to shop on weekdays may be unaffected by this change and would have to pay a higher value for the products.

Conclusion:

After cleaning the data to keep only valid transactions, we performed exploratory data analysis on Portuguese supermarkets using the product and transaction datasets. We discovered days with the most food traffic and also segmented the customers into four groups, which will allow Pernalonga to utilize analysis and product placement to aid store sales.

From the current consumer trends in the food industry in Portugal and the growth in personal disposable income and consumer demands, Pernalonga can devise a desirable marketing strategy that differentiates itself from its competitors by retaining their current customers and directing them towards newer products while acquiring new customers who can benefit from the vast array of offers. To conclude, Pernalonga should capitalize its strength in discounted price-elastic goods and engage in smart product placement with its loss leaders to direct consumer traffic through its stores for higher cart values.

Additional Resources:

https://www.theconsumergoodsforum.com/blog/portugal-retail-snapshot/

https://www.nielsen.com/wp-content/uploads/sites/3/2019/04/global-private-label-report.pdf

https://web3.cmvm.pt/sdi/emitentes/docs/FR69547.pdf