Pernalonga

Exploratory Data Analysis

MKT 680 - Marketing Analytics

Eleanor Lee, Yifei Ren, Julie Wang, Xiao Chen

I. Introduction & Business Background

The supermarket industry is composed of six main types of supermarkets:

1. Traditional/Conventional Supermarkets
    - Have annual sales of $2 million or more
    - Carry between 15,000-60,000 SKUs
    - Tend to have multiple service departments (floral, bakery, etc.)
    - Examples: Publix, Kroger
2. Limited Assortment Supermarkets
    - Carry 4,000 SKUs or less
    - Few to no service departments
    - Higher percentage of SKUs are private labels (store brands)
    - Examples: Aldi, Save-A-Lot
3. Superettes/ “Mom & Pop” Stores
    - Sell mostly packaged and perishable food items
    - Narrow selection of SKUs
    - Tend to have a few service departments
    - Carry limited private label/store brand products
4. Supercenters
    - Large stores (average over 170,000 square feet)
    - As much as 40% of space devoted to groceries
    - Combination of store plus discount store services
    - Example: Walmart, Target
5. Natural/Gourmet Stores
    - Specialty stores focused on healthy living
    - Limited selection of general merchandise
    - Example: Whole Foods, Sprouts
6. Convenience Stores
    - Small outlets

- Carry 800-3,000 SKUs
- Mainly sells dry, perishable, and prepared foods

The top 10 firms occupy 57.1% of the industry in the United States. This highly fragmented picture of the supermarket industry is reflected across in other countries with **a few firms taking a majority of the industry share** (ex. UK has 8 firms occupying 92.9% of the industry share). Based on our analysis outlined below and the facts above, Pernalonga is a smaller **traditional/conventional** supermarket chain that tends to carry **a large amount of SKUs and have extra services** (as seen by the bakery and frozen seafood products).

II. Data Understanding & Preparation

Two datasets are provided for this project – product_table and transaction_table. Transaction Table has 29,617,585 rows and 12 columns. In transaction data, the granular level is item purchased per transaction. In the transaction table, there are 7,920 unique customers, 753 unique transaction id, 727 unique transaction dates, 421 store ids, 10,770 product ids.

To have a quicker and more comprehensive of our data, we do preliminary EDA through summary() and create_report() function in DataExplorer package. We found that there are **no missing values** in our data but some features have **very extreme values**. Please see the analysis results and detailed data manipulation in Exhibit 1,2.

Data Preprocessing for major issues we found from abnormal data are:

(1) Transaction_id is not unique and that all id left with 11 0s after six-digit figures that look like transaction date. Therefore, we rename each transaction id by assigning "trans_date", three-digit "store_id", and eight -digit "cust_id" filled with leading zeros.
(2) Two records occurred even if we group by (cust_id, trans_date, store_id). We found that these records share the same transaction id and prod id but measured in different units. Since only 39 entries have this problem, we simply remove these records.
(3) Remove 3927 records that have Total_prod_paid_amt to be negative or zero.
(4) Leave transaction amount information the same as the differences between (sale_amt, paid_amt-dicsount_amt) are all less than 0.01, meaning the imbalance is due to rounding.

(5) Relabel the tran_prod_offer_cts as 1,841 records that have been mislabeled.

Those records have not used any discounts but labeled as use discounts.

	cust_id	tran_id	tran_dt	store_id	prod_id	prod_unit	tran_prod_sale_amt	tran_prod_sale_qty	tran_prod_discount_amt	tran_prod_offer_cts	tran_prod_paid_amt	prod_unit_price
1:	48129784	20171223000000000000	2017-12-23	167	152761010	CT	1.29	4	0	1	1.29	0.3225
2:	48129784	20171223000000000000	2017-12-23	167	152761012	CT	1.29	4	0	1	1.29	0.3225

One thing we noticed and left room for improvements was that the price difference for one product greatly differs for some product, and we would like to check the patterns behind that.

Looking at the product table, there are 10,767 products, 429 categories, and 1476 subcategories. There is no missing data in the product table either. The only inconsistencies we found immediately were for 429 categories, there were 419 english descriptions and for 1476 subcategories, there were 1430 subcategory descriptions. That means that there are **descriptions being repeated across categories and across subcategories.** We needed to determine if these were due to error, or just repeated usage.

After digging into the 11 descriptions being repeated across categories, we noticed that we need to make the English category description more descriptive based on differences in the Portugeuse category description. We also consolidated some of the categories down that had the same descriptions in English, and no difference in the Portugeuse category description, for the sake of clarity and easier cluster analysis later on.

Looking into the 23 descriptions being repeated across subcategories, we saw **many subcategories appearing across different categories**. Most of these were correct/possible, so we elaborated the subcategory by concatenating the category name to the beginning to add clarity. There was only one item that had a mislabeled subcategory. We also had to change a single "BABY FOOD" subcategory product to the more frequent "FOOD CHILDREN"; this seemed to be a new category due to a translation issue that did not need to be a new category. There was a "ROAST PORK" item under "PASTEURIZED MILK" due to a translation issue as well.

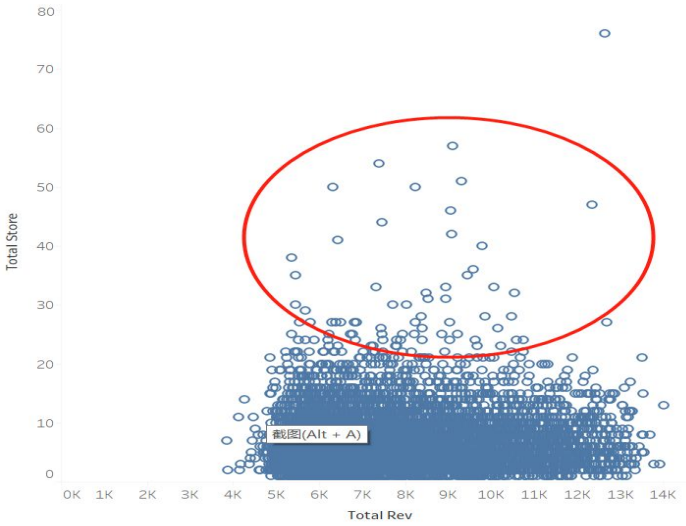
After these clarifying corrections and checks, the product data was ready for analysis together with the transaction table.

III. Descriptive Analysis

★ **Best Customers**

In order to find our best customers, we grouped customers based on their unique customer_id and ranked them based on each person's revenue contribution, number of stores he has visited and number of transactions he has made. We found that there is no clear evidence showing the minority of customers produce the majority of revenue. According to the comparison graph, there is **no strong hint for the correlation between revenue and number of stores visited**. Instead, we noticed that there is a group of people ,who generate very random revenue (ranging from 5k to 12k), favoring visiting lots of different stores. That phenomenon probably could be a good direction when we actually put hands on clustering.

The detailed rank of customers based on different factors, please refer to our attached code. We had very well documented notes for each line of code. And here we have a sample rank based on revenue generated by each customer.



customer_to... Cust Id	customer_total.... Total Rev	customer_total.csv Total Store	customer_total.csv Total Trans
96879682	14,019.71	13	67
97729819	13,929.00	3	150
13489682	13,785.17	3	164
91709522	13,701.00	8	260
30229727	13,559.72	5	234
59899635	13,535.48	8	371
70759790	13,530.19	15	371
54299912	13,519.10	5	301
39679904	13,501.64	21	738
2899928	13,501.51	8	395
2969695	13,448.93	6	457

★ Best Products

Regarding defining our best products, we decided to explore product groups' certain features at category level first. After fully understanding our product data and some possible patterns within the dataset, later we would segment products in more detailed manner (like sub-category level or even product level).

We ranked all the categories in terms of volumes(both in Count and Kilogram), revenue, transactions, and customers. Detailed rankings can be generated by running the R script but a brief overview is provided here:

Sorted by # of Customer bought

category_total.csv Category Id	category_total.csv Category Des...	category_total.c... Total Rev	category_total.csv Total Cust
96034	RICE	419,739.05	7,898
95991	FINE WAFERS	928,409.36	7,896
95809	MINERAL WATERS	697,257.86	7,896
95677	BAGS	159,454.57	7,896
96017	CANNED VEGETA...	477,213.40	7,884
95888	FRESH POULTRY ...	2,379,408.02	7,881
96013	CANNED TUNA	664,737.94	7,876
95934	BANANA	636,698.82	7,868

Sorted by Revenue Generated

category_total.csv Category Id	category_total.csv Category Des...	category_total.c... Total Rev	category_total.csv Total Cust
95890	FRESH PORK	2,483,963.39	7,827
95894	FRESH BEEF	2,401,663.24	7,776
95888	FRESH POULTRY ...	2,379,408.02	7,881
95971	DRY SALT COD	1,344,207.42	6,612
95797	FINE WINES	1,296,608.99	7,426
96026	COFFEES AND RO...	1,191,705.00	7,185
95977	WILD FRESH FISH	1,097,216.16	7,527
95800	BEER WITH ALCO...	1,057,405.95	7,531
95978	FRESH FISH AQU...	1,046,854.07	7,117
95974	FROZEN FISH SE...	988,971.26	7,553

Some Important Takeaways:

1.The most prominent traffic drivers for Pernalonga all belong to the food category. Rice, water, meat and vegetables are most popular among customers of Pernalonga, and that phenomenon is reasonable. However, bags have an incredibly large buying group so we may need to dig more into that.

2.The most important revenue drivers for Pernalonga turned out to be fresh products. Those fresh products include but are not limited to: fresh pork, fresh beef, cod, fresh fish, etc. These products have the highest sale volume, relatively large buying group and provide the highest revenue boost.

★ Best Stores

We set 3 basic criterions to value our 421 different stores: revenue, number of customers visited and volume of products sold (both in Count and in Kilogram). After exploring the store data, we noticed that several stores have very few visits. For example, store 302 only has 1 customer visited over the 2-year period. We assumed that such stores were newly opened, so not enough data was collected at the time when it was retrieved from the database.

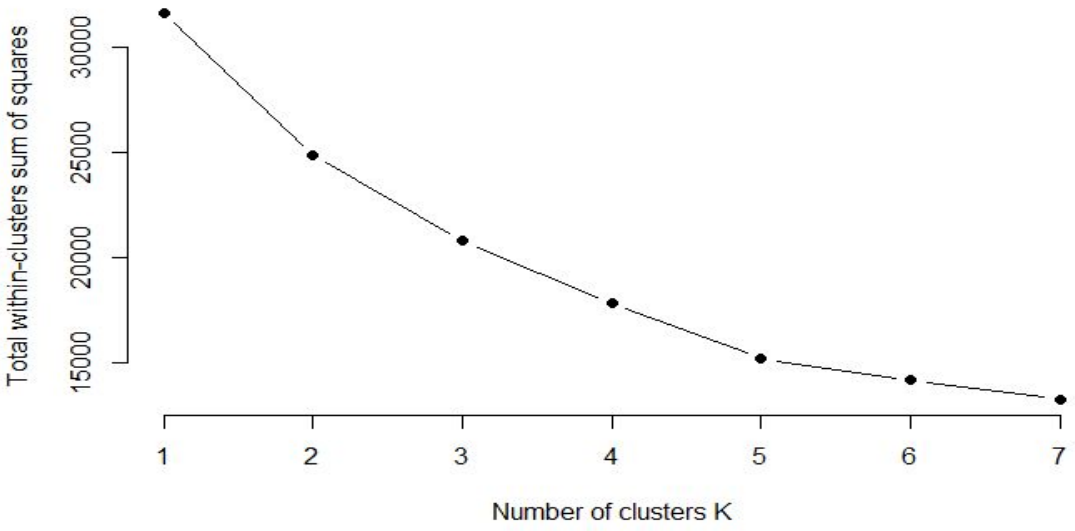
By ranking the stores based on different measurements, we were able to identify the top performing stores from the descriptive results. One interesting finding was that a few stores ranked very high when evaluated by almost every measure. These stores,

including store 345 and 342, are the most valuable stores, and the company should pay close attention to their performances.

store_total.csv Store Id	store_total.csv Total Rev	store_total.csv Total Cust
342	786,521.11	566
345	718,779.77	655
349	680,640.41	416
344	624,991.68	836
343	591,942.18	367
341	580,609.28	781
346	572,839.05	206
588	564,905.19	327
347	551,868.55	180
157	476,098.85	421

IV. Customer Segmentation Analysis

To identify natural grouping of customers, we used the following attributes to form a clustering analysis: total number of unique stores visit by customer (tot_unique_store), total number of transactions conducted (tot_unique_transct), average discount percentage across all purchases (avg_discount), and finally the revenue generated (revenue_sum). We used these attributes with k-means clustering and used the elbow method to pick the best number of k to determine the appropriate number of clusters. All numeric attributes have been scaled prior to clustering since k means clustering is distance-sensitive.



Using the elbow method we determined that five clusters seems to be appropriate for our clustering algorithm.

Cluster_ID	Unique Store Visited	Total Transactions Made	Average Discount Applied	Total Revenue	Number of Customers In cluster
1	16.29	352.44	11.1%	7522.09	834
2	5.42	539.31	8.9%	8508.21	1205
3	5.99	321.81	14.8%	6917.84	1637
4	6.03	343.00	10.1%	10555.22	1552
5	4.93	307.23	8.7%	6678.96	2692
Total:	38.67	1863.79	5.3%	40182.32	7920

Cluster 1 (Professional Buyer):

For the 834 customers in the first cluster, we observed that they visited on average 16.29 unique stores, much higher than the average among four other clusters. Combining with the fact that on average there is a 11.2% discount, we think customers in cluster 1 are professional buyers who would explore deals in a variety of different stores.

Cluster 2 (Frequent Buyers):

For the 1205 customers in the second cluster, we observed that even though they are not the largest cluster, they contributed most in terms of number of transactions made, and it looks like they do not care as much if there is a discount available.

Cluster 3 (Cherry Picker):

For the 1637 customers in the third cluster, we observed that they have the highest average discounts applied across all transactions. They do not generate very high revenue nor do they make a lot of purchases, but whenever they do, they make sure it is the sweetest deal available.

Cluster 4 (Cash Cow):

For the 1552 customers in the fourth cluster, we observed that they do care about discounts, not as much as the cherry pickers, and they do make a decent amount of purchases. They are around the average in terms of the cluster size, but they contributed the most compared with other groups (highest revenue per transaction). Thus, they are our cash cow for the store.

Cluster 5 (Loyalty Candidates):

For the 2652 customers in the fifth cluster, we observed that they make the least amount of purchases, they do not care that much about discounts compared with other groups, but they are the largest group. We believe that this cluster has a huge potential for profit if we can incentivize them properly and convert them into loyal customers.

V. Product Segmentation Analysis

In order to do product segmentation analysis, we split our products into 2 groups based on units: products that are measured by kilograms, and products that are measured in counts. We created several continuous variables for the clustering model to make use of.

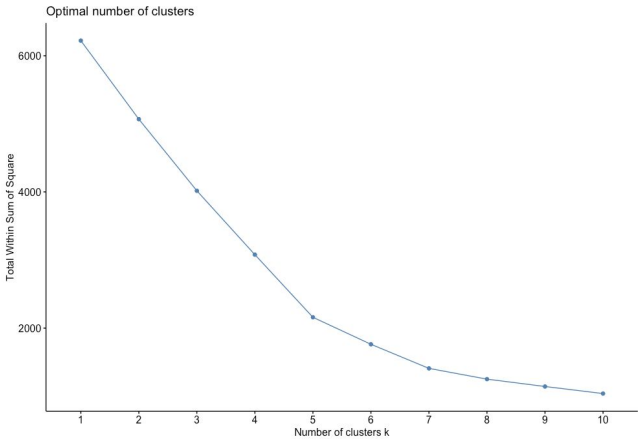
Variable	Explanation
discount_per_unit	● tran_prod_discount_amt/tran_prod_sale_qty ● Represents how much discount was received for each unit sold
paid_per_unit	● tran_prod_paid_amt/tran_prod_sale_qty ● Represents how much was paid for each unit sold
original_price_per_unit	● tran_prod_sale_amt/tran_prod_sale_qty ● Represents how much the unit would have originally cost
discount _percentage	● discount_per_unit/ original_price_per_unit ● Represents percentage off the original price that was discounted off
discount_per_offer	● tran_prod_discount_amt/ tran_prod_offer_cts or 0 if there were no offers ● Represents how much discount was given on a per offer basis

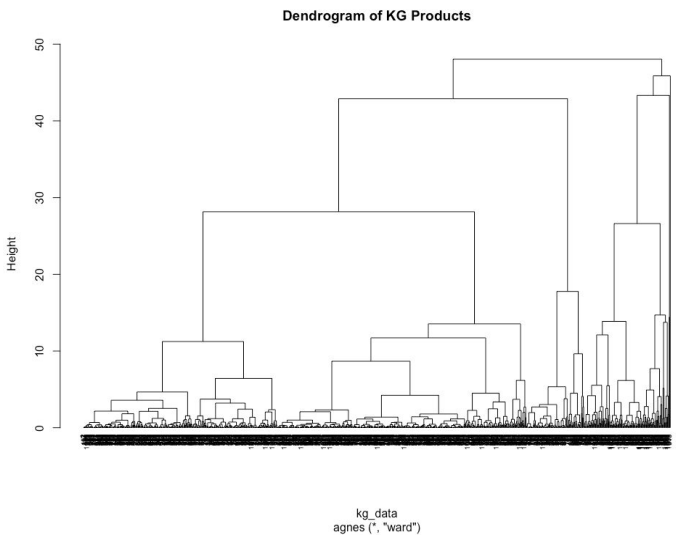
Using these calculated columns and the original columns for each transaction, we were able to aggregate them into product-level measures that were input into both models for kg and count products. The variables we choose to include into our models were at the product level. We used the minimum prod_unit_price, the average prod_unit_price, the total tran_prod_sale_qty, the total tran_prod_offer_cts, the average discount_percentage(ratio; it is shown in decimal form not percentage) and the maximum discount per offer. We believe that these are variables that will help

distinguish the groups due to how these variables can fluctuate greatly and be indicators of sales drivers, traffic, and key value to Pernalonga's revenue operations.

Kilogram Products

For the products that are measured in units of kilograms, we found that we only have 1038 products that fall into this category. Due to the smaller volume of items that needed to be clusters, we ran a hierarchical clustering model. To do this model we optimized 2 parameters, the number of clusters, and the agglomeration method. We found that the method with the highest agglomeration coefficient was 'Ward' and that the optimal number of clusters was 6 based on the elbow method.





These are statistics of the clusters based on their transaction-level that we will use to examine and determine each cluster's description, as well as how much revenue they contributed.

Cluster	Total Sales	Average Discount Ratio	Minimum Paid/Unit	Maximum Paid/Unit	Average Paid/Unit	Average Discount/Unit	Average Sale Quantity	Number of Products
1	3430645	-0.016	0.0163	99	7.2	-0.14	0.604	343
2	451763	-0.0136	0.781	50	13.5	-0.198	0.477	102

3	8840089	-0.0217	0.00462	49	3.66	-0.0822	0.985	440
4	8225541	-0.152	0.00524	18	1.8	-0.359	1.09	150
5	723313	-0.103	0.0117	2	0.813	-0.082	1.08	2
6	760	-0.206	9.99	16.5	12.3	-3	8.46	1

cluster	1	2	3	4	5	6
Revenue	$3,340,093	$443,833	$8,609,416	$6,657,987	$652,378	$599

Cluster 1 (Expensive But Willing):

This is the second largest cluster of the kilogram products, with 343 products. We see that these products are discounted the least, which makes sense as they are also the products that are paid the most per unit, and also one of the higher paid on average per unit. The products are the more expensive products that are not frequently discounted, yet still bought enough to contribute over 3 million to revenue. Customers are willing to pay higher prices for these products despite there being infrequent discounting. Items in this cluster include wild fresh fish and other fresh seafood.

Cluster 2 (Smaller Amount Purchases):

While these products may feature some decent discounts, they are the 2nd most expensive cluster per unit on average. They are also bought in lower quantities. This may be large items that may only require one purchase here and there, that are relatively expensive compared to the usual purchase at a supermarket. Categories of items in this cluster include salads and fish dishes.

Cluster 3 (Stable Staple KVI/KVC):

This is the largest cluster of the kilogram products, which is also where most of the revenue for kilogram items is contributed from. Products in this cluster also are discounted very infrequently, like cluster 5. However if they do get discounted, it is for very little. These items are key value items in that they reliably sell even without discounts, because of necessity. Items in this cluster include nuts and fresh pork.

Cluster 4 (Discounted Often Traffic Drivers):

The products in this cluster are more heavily and often discounted, looking at the average discount percentage and discount per unit. Outside of cluster 6, these items have the highest quantity sold on average. These products are often discounted and then bought in bulk. Yet, these items also bring in the 2nd most revenue, suggesting that because of these deals, these products will sell. Items in this cluster include a lot of fruits and vegetables.This suggests that these products are often marked higher in the sale amount, yet usually sold at a discount, so that customers are used to always getting some sort of deal when buying in bulk.

Cluster 5 (Never Discounted Produce):

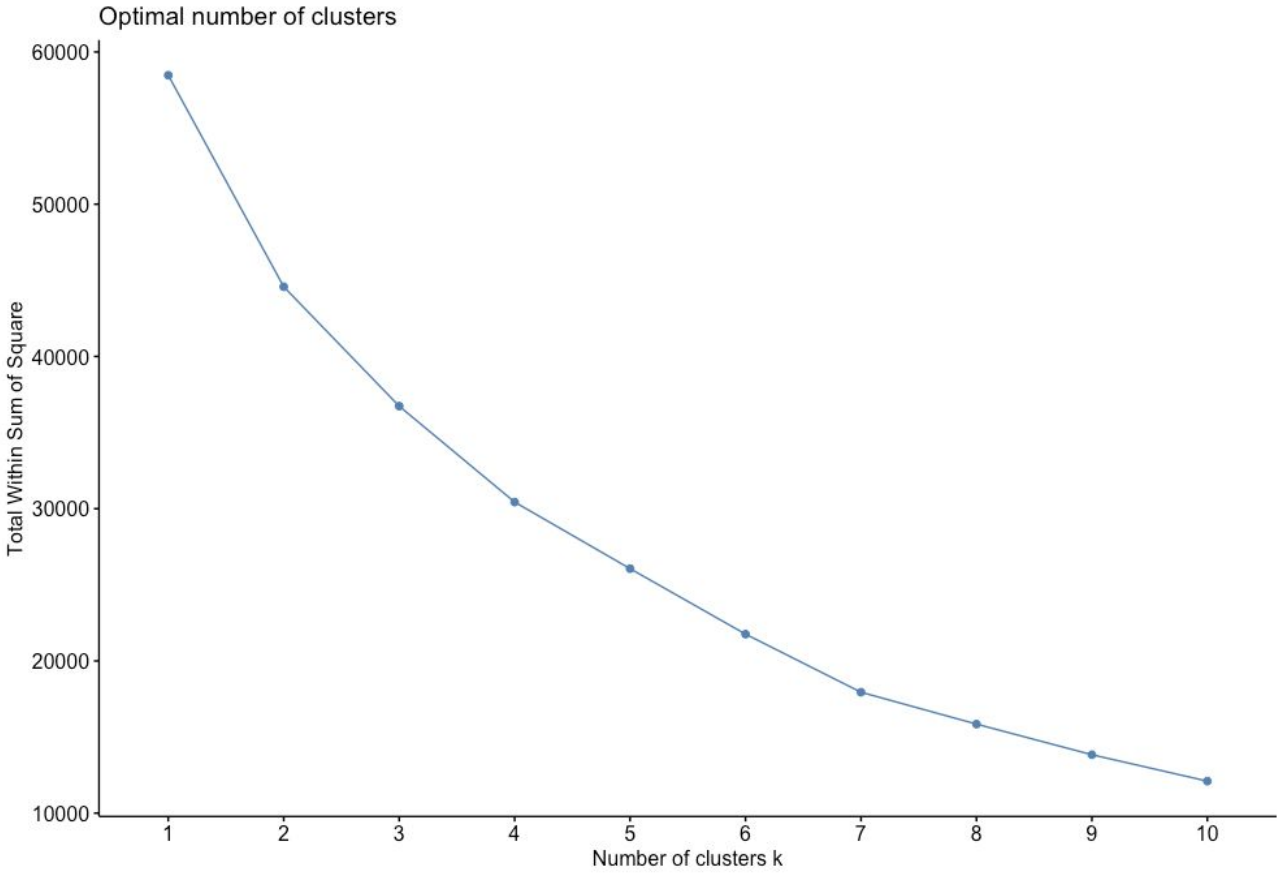
This cluster only has 2 products (bananas and carrots). These are both products from the produce section. It also has the lowest average discount and the lowest paid per unit. We see that these products are not likely to have a lot of discounts. These products are bought year round regardless of prices. These are key value items, with just these 2 items accounting for $652,378 of revenue.

Cluster 6 (Seasonal Sale):

This cluster contains only one product. A product related to the frozen fish service category and some kind of clam subcategory. All transactions related to this product occurred from late November to mid December. This product was only bought 6 different times, and yet the average amount bought is incredibly high at 8.46 kg on average. This product stands as its own cluster because of how steeply discounted it is and how much in bulk it was bought. These are products that are discounted due to seasonality and issues with freshness, especially in the winter. They are discounted to encourage mass buying from customers to move as much of the product as possible before it spoils.

Count Products

Unlike the kilogram products, the count products had 9746 unique products. Because this data is much larger, we had to use k-means instead of hierarchical clustering to handle the clustering of the count products. For k-means, we had to determine the optimal amount of clusters using the elbow method, which determined 7 clusters. The data was prepared the same way as it was with the kilogram products, and the same variables were used to perform the clustering.



After assigning clusters to each product, we examined the cluster statistics like before.

Cluster	Total Sales	Average Discount Ratio	Minimum Paid/Unit	Maximum Paid/Unit	Average Paid/Unit	Average Discount Paid/Unit	Average Sale Quantity	Number of Products
1	22565601	-0.0493	0.0025	25.0	1.38	-0.0894	1.61	5775
2	17613119	-0.305	0.0075	20.0	2.08	-1.05	1.58	3159
3	969736	-0.0212	0.0025	4.14	0.216	-0.0109	2.82	7
4	120419	-0.353	5.98	399	64.7	-33.3	1.04	28
5	2394	-0.251	299	299	299	-100	1.2	1
6	8522717	-0.0989	0.00167	8.99	0.862	-0.157	2.58	156
7	2665720	-0.319	0.580	85.2	13.1	-6.71	1.11	620

cluster	1	2	3	4	5	6	7
Revenue	$21,059,780	$11,548,065	$910,888	$79,338	$1,794	$7,150,932	$1,754,753

Cluster 1 (Small Quantity Purchase and Rare Discount KVI/KVC):

This is the largest cluster of the kilogram products, with 5775 products, and it also contributes the most revenue of the counts products. It is, on average, discounted less than most other count products, and is on the slightly more expensive side in terms of final sales. Customers usually buy only 1 or 2 of these products at a time. Categories included in this cluster are special yogurts, health yogurts, and fruit juice. These items are the key value items and key value categories that contribute most to the revenue and are steady in terms of expected amount paid due to the infrequent discounts.

Cluster 2 (Cheap and Discounted Traffic Drivers):

This is the second largest cluster with 3159 products.The products here have relatively higher discounts on average and are on the cheaper side as well in terms of final paid per unit. This cluster is responsible for $11,548,065 of revenue, suggesting that these products are reliably bought like in cluster 1, but often have to be sold at a discount to be reliably bought, making them traffic drivers. Categories included in this cluster are fine wafer and fine wines.

Cluster 3 (Cheap and Never Discounted):

This cluster only features 7 products. These products are rarely discounted and cheap to begin with, with a max unit price of $4.14 seen. However, customers tend to buy these in large quantities, with an average of 2.86 counts in each transaction. This clusters includes products in categories like mineral waters and fresh milk.

Cluster 4 (Expensive and Always Discounted):

This cluster of 28 products features the highest discounted (other than cluster 5) and also the most expensive products on average. Customers also usually buy these products 1 at a time. These products contribute to the revenue at a much lower amount and should not be the core product offerings of Pernalonga. Categories included are teflon kitchen items and soil & conservation.

Cluster 5 (Discount Luxury Appliances):

This cluster features only one product. This product belongs to the category ‘food preparation’ and appears to be an appliance. It has only ever been bought when it was discounted for (on average)

25% of its original unit price. This seems to be a luxury appliance that is not quite worth the price unless it is discounted, possibly due to similar items that are available at a cheaper price.

Cluster 6 (Sometime Discounted Smaller Indulgences):

This is also a cluster where customers will tend to buy 2 or more at a time. However these items are discounted slightly more and more expensive than those in cluster 3. This cluster also contributes more than 7 times that of cluster 3, at $7,150,932. This cluster includes categories like preserved fruits and chocolate bars. These are unnecessary products that customers will buy when given an incentive (the discount), and the customers feel the need to buy in bulk to take advantage of the discount.

Cluster 7 (Heavily Discounted Non-Necessities):

The products in this cluster are discounted heavily on average and are discounted the most on a per unit basis. These products are also on the slightly expensive side, with an average paid per unit of $13.1. These items are usually bought in quantities of 1 or 2 and are not necessities. Products in this category include ice cream, girl toys, and dry food animals.

VI. Store Segmentation

To segment the 420 stores without graphical and descriptive information, we try to analyze in four aspects : (1) **sales performance**, (2) **stability**, (3) **discount frequency**, and (4) **customer outlook**, and created several continuous variables to improve the demonstrability of the clustering model. To prepare for the clustering table, we firstly create a transaction level table:

Categorical identifier: New_transaction_id, store_id, customer_id, customer_cluster_segment
Aggregate: total_rev_per_transaction, total_product sales_vol_per_transaction,
Ratio: discount_ratio = sum(tran_prod_offer_cts)/sum(tran_prod_sale_qty)
Day_interval_between_next_pay : visit = as.numeric(-difftime(tran_dt, lag(tran_dt)), units="days")
grouped by customer

Then, we aggregate these records again to a store-level table with created features listed below:

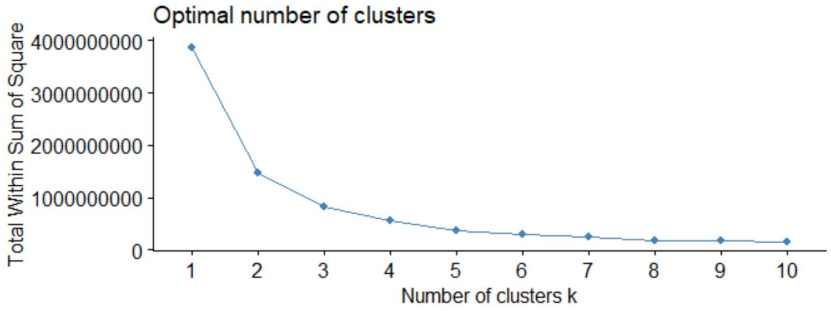
Aspects	Created Variables	Formula/ Meaning

sales performance	(1) Total_revenue	• sum(total_rev)
	(2) total_sales_volume	• sum(total_sales_vol)
stability	(3) Volume_volatility	• sd(total_sales_vol) • Represents how the quantity of the sold product varies through all transactions in the store
	(4) visit_volatility	• sd(visit) • Represents how varies to take a customer to come back and buy again in day differences
discount frequency	(5) promotion_count	• mean(discount_ratio) • Represents the average discount given ratio out of the total transactions happened in the store
customer outlook	(6) total_customer	• uniqueN(cust_id)
	(7) average_visit_interval	• -mean(visit)
	(8) Customer per segment	• cust#=cust# /total_cust (uniqueN in that cluster)

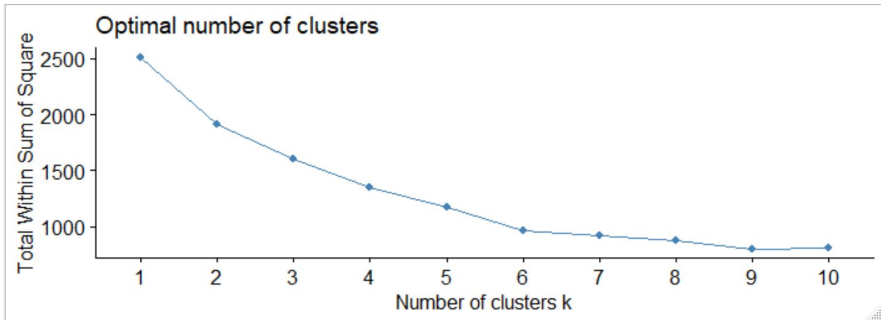
We ran three clustering models to test the interpretability of different combinations of feature selection, and we found that the second one is the most representable.

From model 1 to model 2, we removed total_customer_count because it shares high correlation with either total sales revenue or total sales quantity. Even though the clustering algorithm or linkage method should not be influenced by the correlation, having two variables that are highly correlated is like giving them more, double the weight in computing the distance between two points. (As all the variables are normalised the effect will usually be double). Therefore, we removed the total customer counts. Instead, we added features showing how often a store offers discounts, and the mean of visit day intervals. By doing so, the difference between each group becomes much clearer and obvious, so as the turning point on the elbow plot.

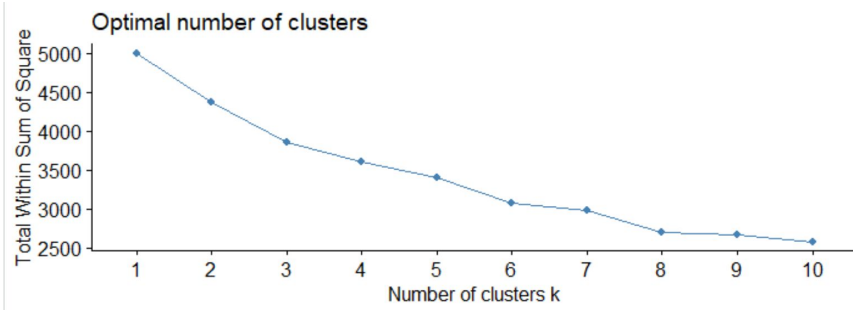
Elbow's Method - Model 2 (the one we picked)



Elbow's Method - Model 1



Elbow's Method - Model 3



	store_id	total_rev	total_ sales_vol	volume_ volatility	total_ cust	visit_ volatility	Size
1	207	0.162	0.202	2.755	0.062	0.486	22
2	186	-0.381	-0.369	-0.310	-0.336	-0.406	134
3	124	-0.462	-0.500	-0.162	-0.233	2.372	38
4	212	-0.324	-0.341	-0.012	-0.215	-0.224	145
5	81	4.659	4.471	0.479	4.011	0.263	10
6	158	0.935	0.983	-0.229	0.621	-0.237	70

	store_id	total_rev	total_ sales_vol	volume_ volatility	visit_ volatility	avg_visit _interval	promotion _count	size	Description
1	0.268	2.754	2.705	0.789	-0.031	0.144	0.355	28	Central Business District (best performance,
2	0.093	-0.193	-0.121	-0.150	-0.368	-0.662	-0.499	229	Residential area (lowest vloatility) (lowest promotion counts)
3	-0.178	-0.203	-0.297	0.076	0.525	0.911	0.644	162	Wholeseller (surburban area)

	store_id	total_rev	total_ sales_vol	volume_ volatility	visit_ volatility	avg_visit _interval	promotion _count	Professional Buyer	Frequent Buyers	Cherry Picker	Cash Cow	Loyalty Candidates	size
1	-0.930	2.112	0.864	9.283	0.421	0.718	-0.009	-0.334	-0.043	0.131	0.295	-0.131	3
2	0.098	-0.189	-0.242	0.179	4.490	2.852	0.538	0.281	-0.579	0.135	0.105	0.149	12
3	-0.297	-0.348	-0.295	-0.328	-0.056	-0.067	-0.022	1.376	-0.255	-0.319	-0.154	-0.297	85
4	-0.119	-0.239	-0.216	-0.121	-0.062	0.220	-0.019	-0.503	-0.602	-0.073	-0.509	1.589	68
5	-0.044	0.410	0.136	0.217	-0.032	0.497	0.642	-0.401	-0.438	1.388	-0.244	-0.392	67
6	0.268	-0.235	-0.215	-0.259	-0.289	-0.834	-0.185	-0.196	1.373	-0.477	-0.537	-0.263	97
7	1.592	3.627	4.877	1.126	0.164	0.346	-0.919	-0.448	-0.190	0.662	0.325	-0.457	11
8	-0.075	-0.037	-0.025	0.059	-0.241	-0.026	-0.246	-0.464	-0.434	-0.323	1.491	-0.371	74

From model 2, we could see distinct behaviors among the three groups:

Cluster 1(Central Business District):

This group contributes both sales revenue and sales volume (calculated per store) the most, and the deviation of total sales quantity per transaction is the highest. The visit volatility is relatively small and the average visit interval is low, which means that people visit often in a quite regular norm. Promotion matters to these stores' customers but not very severely. Lastly, there are only 28 stores that have this pattern. Therefore, we conclude this to be in more metro areas, named as CBD (central business district). Their **typical customer** types would be : (1) **business man/ woman** who buys small amounts of goods regularly and frequently (2) **urban resident** that lives in the city who tends to often visit but relatively less things (3) **random and infrequent visitors/ travelers.**

Cluster 2(Residential Area):

This group of people are price-insensitive. Promotions can't give them big incentives. Both of their visit volatility and average visit interval are the lowest. That is, they go to the stores very often and in a very consistent pattern. The sales volume per transaction is the lowest, so is the volatility. Thus, we can interpret that these customers tend to buy small quantities of goods and probably mostly

the same items with similar amounts. Therefore, we conclude that these stores are mostly likely to be located in residential areas . For their **typical customers** would be local **residents and housewives/ house husbands** who are more likely to be **more loyal customers.**

Cluster 3(Suburban Wholesalers):

Transactions in this group are greatly triggered by discounts. Their visit volatility and average visit interval are the highest, which means people visit these stores once in a while and might not on a regular basis but on an on-demand basis. Therefore, we concluded that these types of stores are more like wholeseller offering discounted products and located in more suburban areas. **Typical customers** are **cherry-pickers**, who go after promotions. An interesting point we could reach is that the current locations of product offerings of these stores might have big room for improvements as the total sales quantity per store and the total sales revenue per store are the lowest. This phenomenon is not normal nor ideal and needs to be adjusted.

VII. Future Improvements

First, we would like to add the average sales volume for each transaction. By knowing this, if we see the figure in wholesalers group are greatly higher, we could further validate our hypothesis.

Second, split the date data into Year, Season, and Month, and Weekdays to see the performance change of its sales performance (both revenue and volume) . In addition, we could calculate the sales growth over the years. By doing so, we could evaluate each stores' potential and stability.

Third, add the product cluster into analysis. Even though adding the customer clusters make the model output become less explainable, analyzing the stores from a totally different angle (top_selling_product in product level instead of sales performance in transaction level) might provide us with brand new findings. The possible clustering outcomes might tell us which type of stores are more popular with fresh meats, fruits, beverages, frozen foods, or supplies. By knowing this, decision makers could tailored each store's current offerings. Despite that someone might argue inventory optimization could be realized by demand forecasting without segmentation, customers' unmet need could only be known and learned by their past behaviors or behaviors of people share same motives/ habites with. For example, if we could know the five out of ten top selling products are high-end imported fruits and the store currently only offers high-end apples, melons, and peaches, we could suggest those stores to import high-end strawberries, grapes, or even organic vegetables.

To achieve that, we are thinking to break down our current 429 categories into less groups in less amount but higher level categories (e.g. Fresh Meats, Vegetables, Fruits, Frozen Food, etc.) . One possible solution is to find lexicons/ dictionaries to test whether each product lies in which categories and then we could replicate the steps we did for adding customer clusters in store clusters.

Appendix

Exhibit 1

```
> summary(trans)
    cust_id           tran_id           tran_dt            store_id        prod_id          prod_unit          tran_prod_sale_amt
 Min.   :   29568   Min.   :2.016e+18   Length:29617585   Min.   :102.0   Min.   :145519008   Length:29617585   Min.   :   0.010
 1st Qu.:25009862   1st Qu.:2.016e+18   Class :character   1st Qu.:294.0   1st Qu.:999247000   Class :character   1st Qu.:   0.900
 Median :50259604   Median :2.017e+18   Mode  :character   Median :393.0   Median :999362424   Mode  :character   Median :   1.590
 Mean   :50168995   Mean   :2.017e+18                      Mean   :445.8   Mean   :979749581                      Mean   :   2.503
 3rd Qu.:75689897   3rd Qu.:2.017e+18                      3rd Qu.:588.0   3rd Qu.:999679887                      3rd Qu.:   2.790
 Max.   :99999776   Max.   :2.017e+18                      Max.   :999.0   Max.   :999999656                      Max.   :3371.250
 tran_prod_sale_qty tran_prod_discount_amt tran_prod_offer_cts tran_prod_paid_amt prod_unit_price
 Min.   :   0.001   Min.   :-1400.2500     Min.   : 0.0000     Min.   : -1.41     Min.   :  0.0075
 1st Qu.:   1.000   1st Qu.:    -0.2400    1st Qu.: 0.0000     1st Qu.:  0.84     1st Qu.:  0.7400
 Median :   1.000   Median :     0.0000    Median : 0.0000     Median :  1.37     Median :  1.3900
 Mean   :   1.668   Mean   :    -0.4028    Mean   : 0.3409     Mean   :  2.10     Mean   :  2.0930
 3rd Qu.:   2.000   3rd Qu.:     0.0000    3rd Qu.: 1.0000     3rd Qu.:  2.32     3rd Qu.:  2.4900
 Max.   :2112.000   Max.   :     0.0000    Max.   :76.0000     Max.   :1971.00    Max.   :399.0000
```

Exhibit 2