

MKT 680

Project Report

Exploratory Data Analysis of Customers, Products and Stores of
Pernalonga

Students:

Meng Cheng

Feifan Gu

Sidi Liu

Under supervision of

Prof. Alvin Lim Prof. David Sackin

Emory University

Goizueta Business School

CONTENTS

1. Introduction to the Transaction and Product Datasets
 - 1.1 Background
 - 1.2 Data preparation
2. Exploratory Data Analysis of Customers
 - 2.1 Exploration of best customers
 - 2.2 Customer clustering analysis
3. Exploratory Data Analysis of Products
 - 3.1 Exploration of best products
 - 3.2 Exploration of best product groups
 - 3.3 Advanced product and product group exploration
 - 3.2 Product clustering analysis
4. Identifying Natural Groupings of Stores
5. Conclusions

Chapter 1: Introduction to the Transaction and Product Datasets

Background

Pernalonga, our client, which is a leading supermarket chain of over 400 stores in Lunitunia, sells over 10 thousand products in over 400 categories. Our team is responsible for the development of a marketing campaign to experiment on personalized promotions for the client.

At this stage, our team is tasked to analyze and understand the data and report back initial insights to the client. To be more specific, based on product information and transaction history data, our team explores the basic understanding and conduct segmentation of products, customers and stores.

Data preparation

After checking the data carefully, we found several problems: first, in about 7 million records, the unit price multiplies the amount is not equal to sales (before discount); second, in about 3 million records, the sales(before discount) minus discount is not equal to sales(after discount); third, there are sales(after discount) less than or equal to 0; fourth, there are duplicated records with all the ids being the same but with different transaction amount and sales.

To deal with them, we decided to delete all those negative or zero sales records and those duplicated ids records. Also, we replace the unit price column with unit after-discount price, which is the sales(after discount) divided by sale_amount. We think this unit price makes

more sense than that inaccurate before-discount unit price. Last but not least, we made an assumption in order to calculate the cost of each product. First, we assumed that stores would never sell any products at a loss. Then, according to the 20/80 rule, which says that roughly 80% of the company's profits are usually coming from only 20% of the products, we made an assumption that 80% rather cheap products' cost would simply be the lowest price they ever sold at, while other 20% expensive products' cost would be the lower price between the lowest unit price and $0.7 * \text{average of the unit price}$. We think for those 20% of products, the cost should be no more than 70% of the average unit price according to the research we have done regarding the industry. With cost calculated, we also added a profit column which is the difference between revenue and cost.

Chapter 2: Exploratory Data Analysis of Customers

Exploration of best customers

In this part, we took four basic information of customers into consideration. The first one is the total revenue a customer created through transactions in stores and the second one is the total profit. Such two data represents the most direct benefit the customer has been giving to Pernalonga. Since we have revenue and profit data for each transaction, we could easily sum up the revenue and profit of each customer's transactions to calculate this two data information.

We consider the above two measurements in a cumulative way. However, we will explore the following two measurements in an average sense. We considered the frequency of store visit to measure how much he loves shopping. Here the frequency is calculated from the count of the transaction the customer had divided by the number of days, which is the latest date minus the earliest date in the dataset. The last one is the average number of products the customer would purchase per shopping. I counted the number of products of each transaction in the dataset and then calculated the mean number for each customer's all transactions. I put all those basic information in a table to record the basic information of customers.

	cust_id	revenue_sum	profit_sum	visitfreq	prodcnt_avg
1	139662	9747.21	5987.198	0.6135734	18.636085
2	799924	10107.31	6018.800	0.4842250	20.486684
3	1399898	7392.69	3293.592	0.8257888	11.499719
4	1749580	5375.40	2515.794	0.4370709	21.629045
5	1889991	9665.97	4004.742	0.4868966	15.181325
6	1979557	4385.18	2770.863	0.5424955	10.966728
7	2109544	6201.83	2335.947	0.4392060	32.093417
8	2559894	7314.18	3686.162	0.6049383	13.072351
9	2649945	5222.20	2152.506	0.3964335	10.489688
10	3249808	5455.65	3180.680	0.4087791	14.201828

FIG 1: The head of the customer information table. “revenue_sum” represents the total revenue the customer has created through shopping in stores; ‘profit_sum’ represents the total profit the customer has created; ‘visitfreq’ represents the frequency the customer visits stores; ‘prodcnt_avg’ represents the average number of products the customer purchased per transaction.

We can easily find the best customers in terms of any of the four measurements by sorting by any of them in descending order. To provide a more direct impression of who are the best customers, the following data visualization is plotted, with the best customers in each of the four senses pointed out.

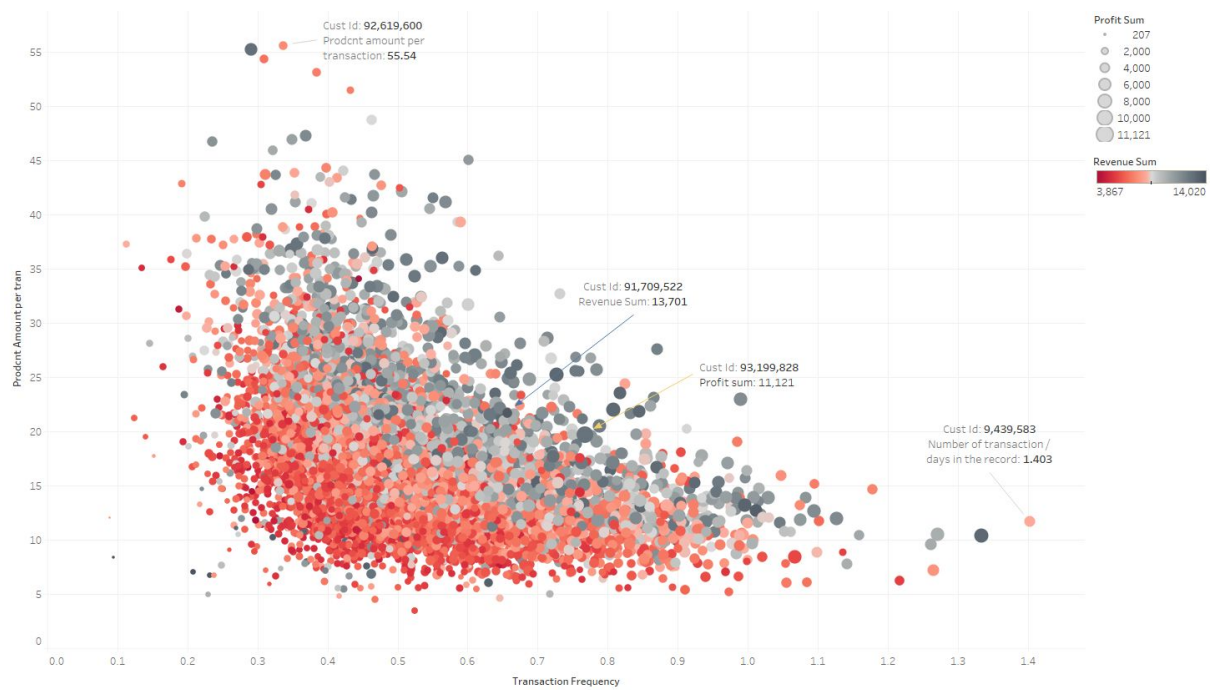


FIG 2: Data viz of customers' basic information. Each bubble represents a customer; X-axis represents shopping frequency; Y-axis represents average product amount per transaction; the size of the bubble represents the sum of profit the customer has created; the color of the bubble gradually changing from red to black represents the sum of revenue the customer has created changing from low to high.

Using the dashboard created in Tableau, you can easily check the best customers in terms of any one of the measurements using the sliders to filter. In the plot, four best customers in each sense are pointed out. They are cust_id 92619600 with highest average product amount per transaction 55.54, cust_id 91709522 with highest sum of revenue \$13,701, cust_id 93199828 with the highest sum of profit \$11,121, cust_id 9439583 with highest shopping frequency 1.403/day.

Customer clustering analysis

In this part, we had to explore some deeper attributes of customers to help us group customers in more dimensions. First, we would like to measure the extent that a customer is attracted by discounts. Since we had records of the count of discounts offered on each

product, it is easy to calculate the mean discount offered of a customer's product portfolio, which is a perfect measurement for a customer's "discount obsession".

Second, we want to know about a customer's "category loyalty". Here we counted the number of different categories a customer has purchased and then calculated the skewness of the numbers. If the distribution of numbers is right-skewed, which would result in a large skewness, it means the customer has many rather small counts of categories and fewer large counts of categories, indicating that he is very attracted to few categories, which means he has a high category loyalty ; if the distribution of count numbers is left-skewed, which would result in a smaller even negative skewness, it means the customer has many rather large counts of categories, indicating that he is attracted to many categories in a significant way, which means he has a low category loyalty.

We use the same way to measure a customer's store loyalty. The larger the skewness of the distribution of a customer's count of different stores, the more loyal he is to those stores that he visits a lot. A few records got store skewness as NA and we excluded those rows.

	cust_id	dispercent	cateskew	storeskew
1	139662	0.1627248	2.764834	0.749949811
2	799924	0.3060221	2.836385	0.384887919
3	1399898	0.3820834	3.720312	0.380936941
4	1749580	0.5232833	2.538427	1.298856945
5	1889991	0.5846710	1.957431	1.073117575
6	1979557	0.4284658	3.361191	0.384897307
7	2109544	0.4158455	3.754682	0.649728942
8	2559894	0.2618432	2.648629	3.722849762
9	2649945	0.4570743	2.145032	0.733704963
10	3249808	0.3846154	3.235394	4.097221072

FIG 3: The head of the customer deeper information table. "dispercent" represents the mean discount offered of a customer's product portfolio; "cateskew" represents a customer's category loyalty; "storeskew" represents his store loyalty.

After combining the customer information table and the customer deeper information table, we started our customer clustering using KMeans.

Since KMeans uses Euclidean distance to measure the distance between points, we had to scale the attributes first. We chose to use the 5 attributes: sum of visit count, average product count per transaction, average discount offered, category loyalty and store loyalty of customers in the KMeans model, so we used the min-max method to scale those attributes.

One important issue of using KMeans is to decide the best K. Here we tried to find the optimal K via the gap statistic. The Gap statistics (Malika Charrrd, 2014) compares the total within intra-cluster variation for different values of K with their expected values under null reference distribution of the data. The estimate of the optimal K will be the value that maximize the Gap statistic.

We drew the Gap ~ K plot to show directly how the Gap changes depending on K, therefore we determined the best K.

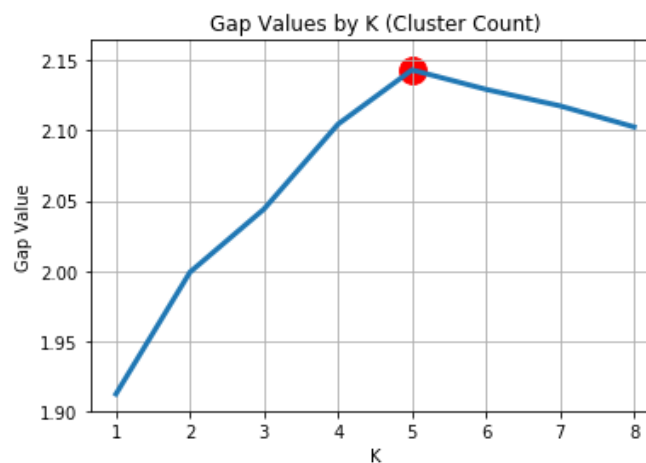


FIG 4: The plot of Gap value ~ K value. The maximum value of gap emerges when K = 5, indicating that clustering into 5 groups here is the optimal choice.

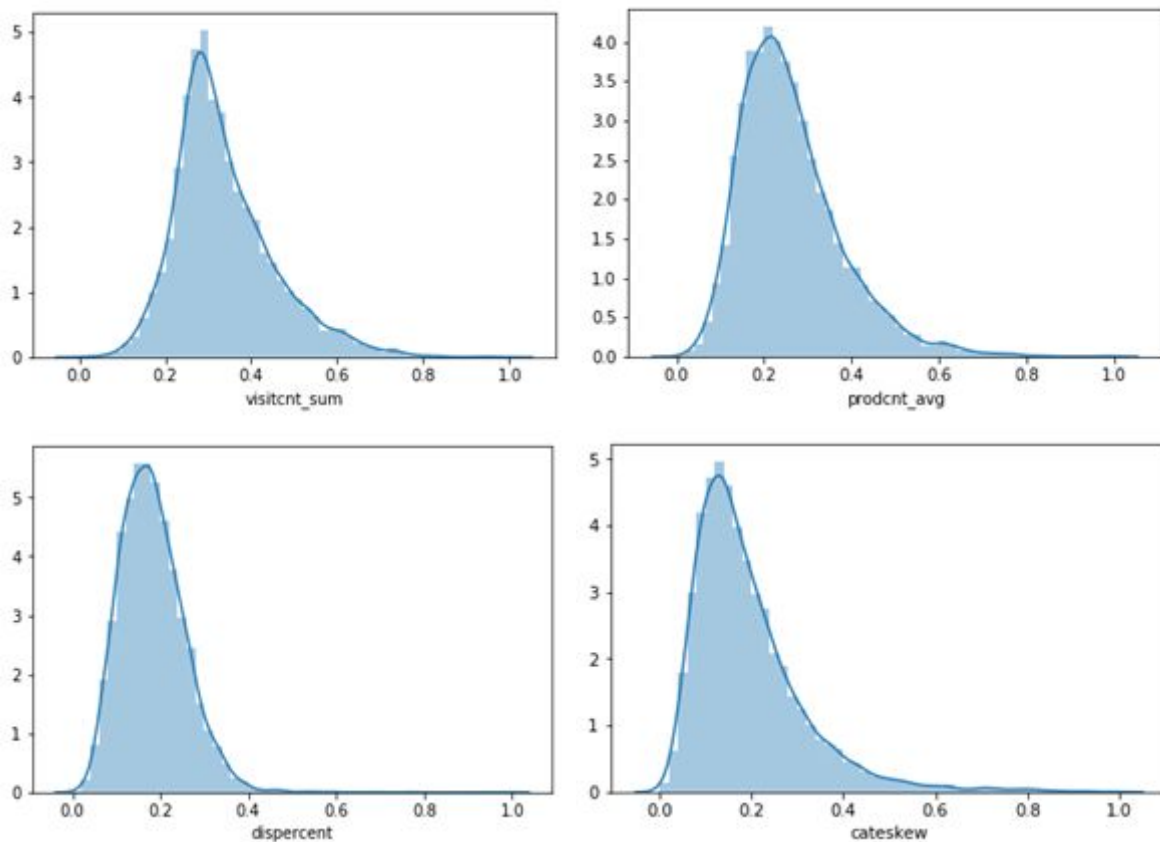
The best K is obviously 5. So we decided to divide the customers into five groups. After running a KMeans model setting number of clusters as 5, we got the labels of every customer ranging from 0 to 4, indicating which group he is in among those five groups. Also, we got

the 5-dimension coordinate for the five cluster centers. Then, we can explore the overall characteristics of those five groups from the five centers because they are perfect representatives.

According to the following coordinates of the clusters and distribution of the attributes, we can determine what level each cluster has in terms of all those measurements we used.

	visit count sum	product count mean	discount frequency	category loyalty	store loyalty
Cluster 1	0.294	0.233	0.184	0.161	0.209
Cluster 2	0.392	0.191	0.157	0.465	0.262
Cluster 3	0.516	0.179	0.152	0.208	0.247
Cluster 4	0.325	0.252	0.195	0.164	0.434
Cluster 5	0.253	0.453	0.200	0.138	0.275

TABLE 5: The coordinates of the five cluster centers, all of the figures are scaled to range from 0 to 1.



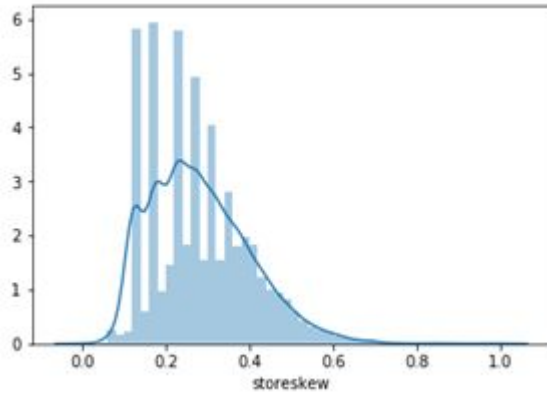


FIG 6: The distribution of the five measurements taken into consideration in KMeans, all after scaling.

After a simple judgment of where the centers are on each of the distribution graphs, we could qualify those groups in each sense.

	visit count sum	product count mean	discount frequency	category loyalty	store loyalty	Typical customers
Cluster 1	Middle	Middle	Middle	Middle	Low	Ordinary people
Cluster 2	Middle	Middle	Low	High	Middle	Category addicts
Cluster 3	High	Middle	Low	Middle	Middle	Rich People
Cluster 4	Middle	Middle	High	Middle	High	Frugal people
Cluster 5	Low	High	High	Middle	Middle	Shopaholics

TABLE 7: Attribute levels of five clusters according to the center coordinate and each attribute distribution.

The following table records the average revenue and profit that customers created in each cluster. Cluster 2 is the most profitable cluster, whose typical customers are rich people; customers in Cluster 4 created the most revenue, they acted like shopaholics. It makes sense that those two kinds of customers created the most value for the stores.




label 	revenue_mean 	profit_mean 
0	7320.038	3287.122
1	7559.313	3462.365
2	8501.043	4439.752
3	7753.541	3529.022
4	8733.093	3930.431

TABLE 8: Average revenue and average profit that customers created in each cluster.

Chapter 3: Exploratory Data Analysis of Products

Exploration of best products

In this part, we want to evaluate a product with several metrics:

- total revenue a product gained through transactions in stores
- total profit a product earned
- total volume a product being purchased, the unit could be count or kilogram, we will consider those two kinds separately
- time of transactions a product being purchased
- unique store count of a product being purchased in
- unique customer count of a product being purchased in
- the average discount rate of a product calculated by dividing the total discount amount by the total sale amount
- promotion frequency of a product, calculated by below formula¹:

$$\frac{\Sigma(\text{count of stores with promotion transactions in a day} / \text{count of store with transactions in a day})}{\text{amount of days a product being available in Pernalonga stores}}$$

The first two metrics focus on the monetary benefit a product brings to Pernalonga and the middle four metrics measure the popularity of a product. The last two metrics show us the promotion characteristic of a product.

	prod_id	prod_unit	subcategory_id	category_id	brand_desc	volume	revenue	tran	storecnt	customercnt	dayon	profit	avgdiscount	salecnt	promotion
1	999231999	CT	90512	95677	PRIVATE LABEL	1062992.0	105883.67	769801	419	7862	729	72872.3502	0.003909060	33.23948	0.04559600
2	999680491	CT	90726	95072	NO LABEL	676506.0	57381.26	79665	182	3706	729	6930.6492	0.005819932	33.53686	0.04600392
3	999951863	CT	94003	95861	PRIVATE LABEL	656001.0	290502.57	149529	417	6494	729	56095.4041	0.068471605	272.72794	0.37411240
4	999956795	KG	94414	95934	FRUTAS&VEGETAIS	566287.0	546554.81	491498	405	7573	729	476961.0244	0.075791108	233.45224	0.32023627
5	999401572	CT	93726	95809	PRIVATE LABEL	553851.0	89934.83	119587	415	6560	729	18421.6959	0.007085274	36.52081	0.05009713
6	999232655	CT	90726	95072	NO LABEL	491784.0	58794.61	75017	182	3767	729	8716.7160	0.004328351	29.75190	0.04081193
7	999974203	CT	93998	95860	PRIVATE LABEL	377541.0	81312.92	54592	417	3230	729	11298.8801	0.020621495	77.86631	0.10681250
8	999958970	CT	94003	95861	MIMOSA	333713.0	191127.24	75236	418	5827	729	40590.8427	0.124041831	228.65514	0.31365588
9	999302933	CT	90756	95073	NO LABEL	315627.0	47762.25	56281	398	4764	729	8708.1974	0.011307534	35.37030	0.04851893
10	999255351	CT	93366	95719	PRIVATE LABEL	299076.0	18680.27	17137	413	3860	729	679.8478	0.032814954	113.23134	0.15532420

Table 9: The head of the product information table. metrics above refer to column 'revenue', 'profit', 'volume', 'tran', 'storecnt', 'customercnt', 'avgdiscount', 'promotion' in order.

¹ Amount of days a product being available in Pernalonga stores is calculated by the day difference between the first and last day it has a transaction history

The statistic summary of product information table is provide below:

```
> summary(prod_info_summary[,6:15])
```

volume		revenue		tran		storecnt		customercnt		dayon		profit		avgdiscount	
Min. :	3	Min. :	500.0	Min. :	3	Min. :	1	Min. :	2.0	Min. :	9	Min. :	0.0	Min. :	0.00000
1st Qu. :	339	1st Qu. :	933.9	1st Qu. :	301	1st Qu. :	123	1st Qu. :	187.0	1st Qu. :	550	1st Qu. :	295.2	1st Qu. :	0.02185
Median :	908	Median :	1891.5	Median :	751	Median :	220	Median :	402.0	Median :	721	Median :	672.4	Median :	0.10785
Mean :	4588	Mean :	5819.5	Mean :	2754	Mean :	221	Mean :	723.6	Mean :	614	Mean :	3532.8	Mean :	0.15762
3rd Qu. :	2766	3rd Qu. :	4569.8	3rd Qu. :	2038	3rd Qu. :	323	3rd Qu. :	877.0	3rd Qu. :	729	3rd Qu. :	1882.2	3rd Qu. :	0.27625
Max. :	1062992	Max. :	602109.4	Max. :	769801	Max. :	419	Max. :	7862.0	Max. :	729	Max. :	683411.9	Max. :	0.68952

salecnt		promotion	
Min. :	0.00	Min. :	0.00000
1st Qu. :	23.83	1st Qu. :	0.04078
Median :	48.25	Median :	0.08529
Mean :	67.67	Mean :	0.11689
3rd Qu. :	94.50	3rd Qu. :	0.16202
Max. :	521.81	Max. :	0.91111

Fig 10: Statistic summary of product information table

All metrics show large differences between products.

To have a more direct overview of products in Pernalonga, the visualizations of some key product information with Tableau are plotted below, the best product in each dimension and metrics is labeled.

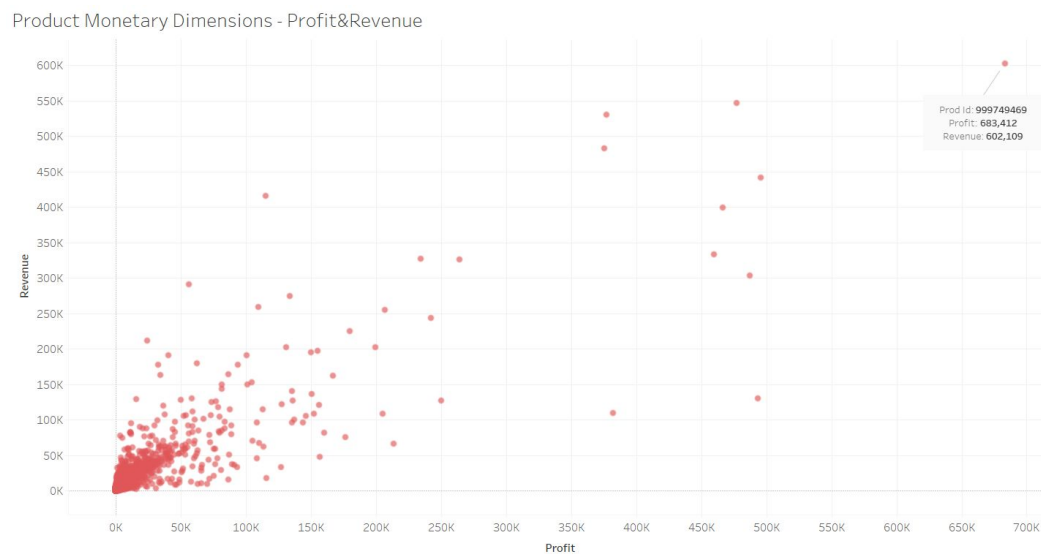


Fig 10: Profit and Revenue by products.

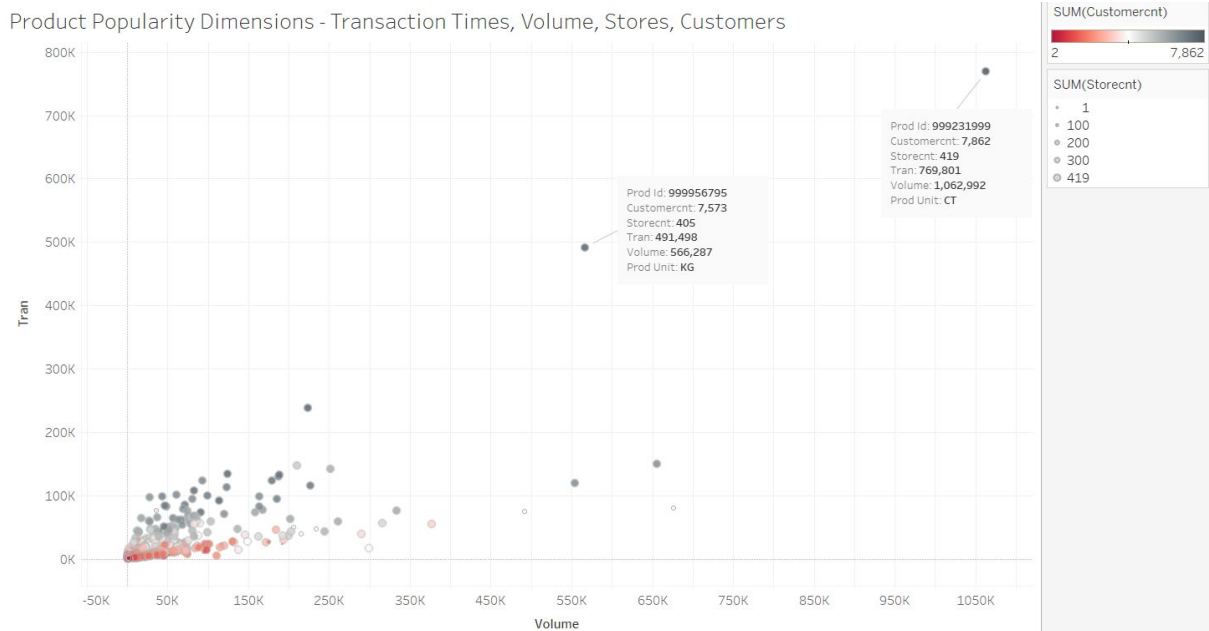


Fig 11: Transaction Times, Volume, Unique Store Count and Unique Customer Count by products.

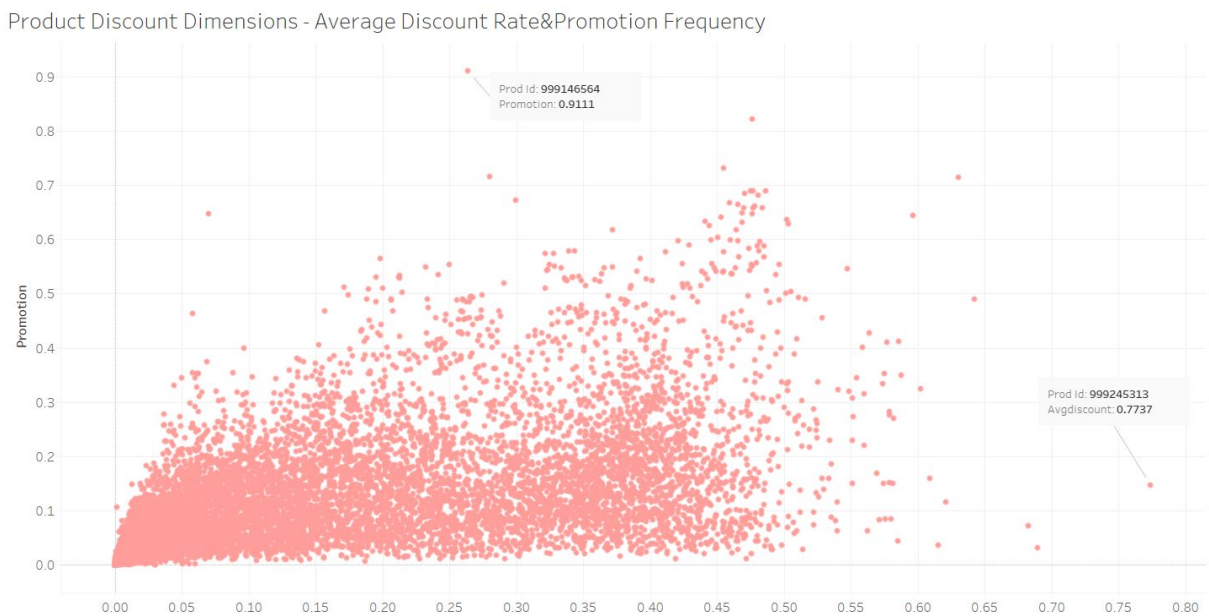


Fig 12: The promotion frequency and the discount rate of a product by product.

To respond to the question of the best product, the answer is obvious with the above exploration. The best product with the best revenue and profit is product with id 999956795(revenue: 546,554.81, profit: 476,961.0244). The best product with the best

transactions, customers and stores is product with id 999231999(transactions: 769,801 times, customers: 7,862, stores: 419) and the best product with the best volumes is product 999231999 with 1,062,992 counts and product 999956795 with 566,287 kg.

Exploration of best product groups

In this part, we want to advance our analysis on the group(category) level. We evaluate the performance of a category based on²:

- total revenue a category gained through transactions in stores
- total profit a category earned
- time of transactions a category related to
- unique store count of a category being purchased in
- unique customer count of a category being purchased in

The first two metrics focus on the monetary benefit a category brings to Pernalonga and the last three metrics measure the popularity of a category.

	category_id	category_desc_eng	revenue	tran	storecnt	customercnt	profit
1	95677	BAGS	159454.57	831712	419	7896	56767286101
2	95072	PAO MANUFACTURE	527313.92	483069	230	6446	17953474289
3	95861	FRESH UHT MILK	859381.87	382556	418	7823	12607744137
4	95934	BANANA	636698.82	542437	418	7868	237007773407
5	95809	MINERAL WATERS	697257.86	514412	419	7896	12285789439
6	95860	SPECIAL UHT MILK	525038.38	227922	419	7374	1342604437
7	95073	FROZEN BREAD	267699.62	299571	414	7443	2764031374
8	95719	KLEENEX	46550.02	48891	419	6863	52754963
9	95963	INDIVIDUAL BOWLS	448772.74	257830	417	7680	2158326110
10	95831	REGULAR EGGS	505660.35	322938	419	7758	18004535856

Table 13: The head of the category information table. metrics above refer to column 'revenue', 'profit', 'tran', 'storecnt', 'customercnt' in order.

The statistic summary of category information table is provide below:

² We don't use volume as a metric because the unit is not uniformed in one category.

revenue	tran	storecnt	customercnt	profit
Min. : 504.1	Min. : 3	Min. : 3.0	Min. : 3	Min. : 0
1st Qu.: 8862.0	1st Qu.: 2548	1st Qu.: 322.2	1st Qu.: 1127	1st Qu.: 2408843
Median : 42066.3	Median : 15982	Median : 409.0	Median : 3500	Median : 42082568
Mean : 144678.1	Mean : 60652	Mean : 347.0	Mean : 3839	Mean : 3265826331
3rd Qu.: 151085.3	3rd Qu.: 77690	3rd Qu.: 416.0	3rd Qu.: 6690	3rd Qu.: 693929300
Max. : 2483963.4	Max. : 831712	Max. : 420.0	Max. : 7898	Max. : 237007773407

Fig 14: Statistic summary of category information table.

All metrics show large differences between categories.

To have a more direct overview of categories in Pernalonga, the visualizations of some key category information with Tableau are plotted below, the best category(product group) in each dimension and metrics is labeled.

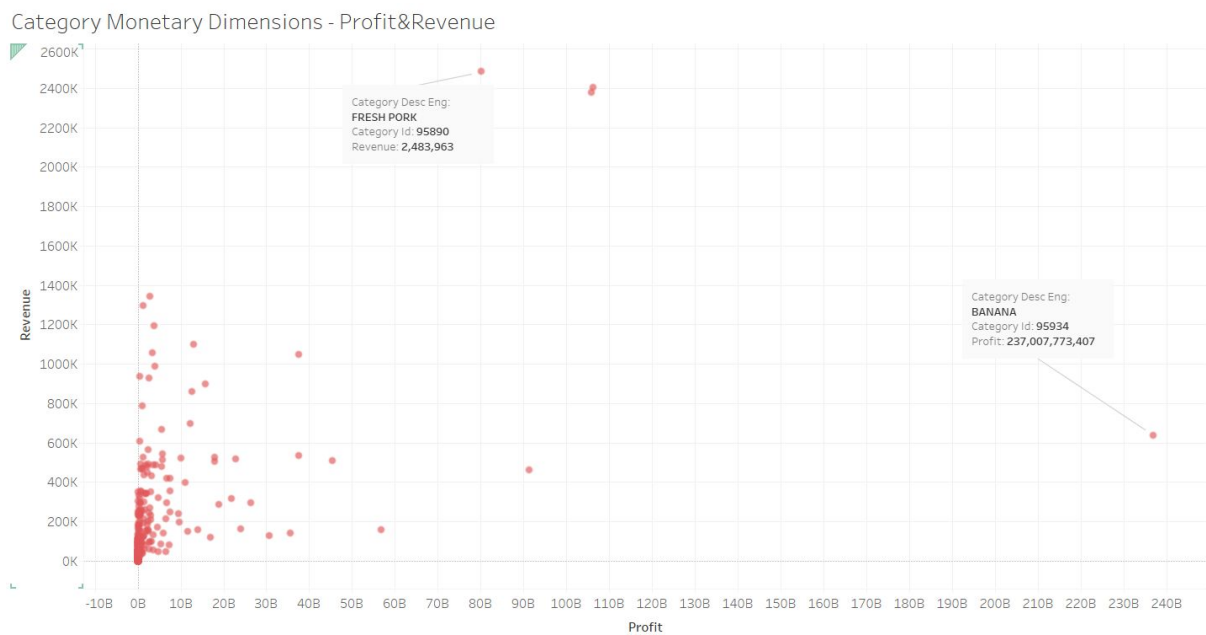


Fig 15: Profit and Revenue by category.

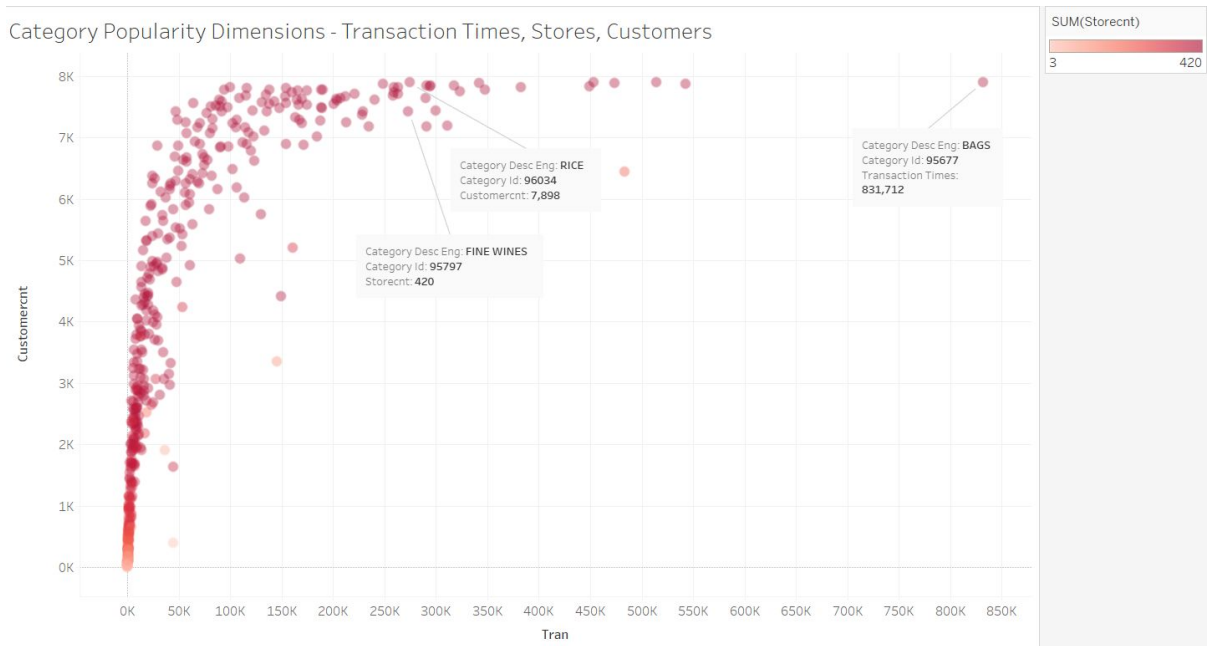


Fig 16: Transaction Times, Unique Store Count and Unique Customer Count by category.

To respond to the question of ‘What are the product groups with the best volumes, revenues, profits, transactions, customers, etc.’, the answer is obvious with the above exploration. The best product group with the best revenue is Fresh Pork(revenue: 2,483,963), with the best profit is Banana(profit: 237,007,773,407), with the best transactions is Bags(transaction: 831,712 times), with the best customers is Rice(customers: 7,898) , with the best stores is Fine Wines(stores: 420).

Advanced product and product group exploration

To know more about the characteristics of products, we analyse the following aspects of products and list the top 10 products of each.

1. Key Value Items & Key Value Categories

KVI 1 revenue			KVI 2 profit			
id	brand	category	id	brand	category	
1	999749469	PERECÃVEIS CARNE	FRESH BEEF	999749469	PERECÃVEIS CARNE	FRESH BEEF
2	999956795	FRUTAS&VEGETAIS	BANANA	999443828	PRIVATE LABEL	CHEESE TYPE FLAMENGO
3	999749894	NO LABEL	FRESH PORK	999696393	NO LABEL	FRESH FISH AQUACULTURE
4	999455829	NO LABEL	FRESH POULTRY MEAT	999956795	FRUTAS&VEGETAIS	BANANA
5	999649801	NO LABEL	DRY SALT COD	999747259	NO LABEL	CHICKEN
6	999747259	NO LABEL	CHICKEN	999557956	NO LABEL	FRESH POULTRY MEAT
7	999557956	NO LABEL	FRESH POULTRY MEAT	999662852	TERRA NOSTRA	CHEESE TYPE FLAMENGO
8	999955966	NO LABEL	FRESH FISH AQUACULTURE	999749894	NO LABEL	FRESH PORK
9	999749460	PERECÃVEIS CARNE	FRESH BEEF	999455829	NO LABEL	FRESH POULTRY MEAT
10	999696393	NO LABEL	FRESH FISH AQUACULTURE	999749460	PERECÃVEIS CARNE	FRESH BEEF

KVC 1 revenue		KVC 2 profit	
category	amount	category	amount
1 FRESH PORK	2483963.39	BANANA	636698.82
2 FRESH BEEF	2401663.24	FRESH BEEF	2401663.2
3 FRESH POULTRY MEAT	2379408.02	FRESH POULTRY MEAT	2379408
4 DRY SALT COD	1344207.42	CHEESE TYPE FLAMENGO	460217.14
5 FINE WINES	1296608.99	FRESH PORK	2483963.4
6 COFFEES AND ROASTED MIXTURES	1191705	BAGS	159454.57
7 WILD FRESH FISH	1097216.16	CHICKEN	507734.26
8 BEER WITH ALCOHOL	1057405.95	CITRUS	533708.17
9 FRESH FISH AQUACULTURE	1046854.07	FRESH FISH AQUACULTURE	1046854.1
10 FROZEN FISH SERVICE	988971.26	CARROT	140836.73

Table 17: Criterias for KVI(Key Value Items) and KVC(Key Value Categories): revenue and profit.

Key Value Items (KVI) and Key Value Categories are products and categories that make a significant contribution to the overall sales of a business. In most cases, there are one or more products and categories that make up most of a company's sales. Top 10 products as KVI and Top 10 categories as KVC are shown above. Under both criterias, food, especially fresh food, such as fresh meat, contributes the most. Also, as for brand, it seems not to be an important factor for KVI. However, PERECÃVEIS CARNE is outstanding.

2. Traffic Drivers

transaction_time			
	id	brand	category
1	999231999	PRIVATE LABEL	BAGS
2	999956795	FRUTAS&VEGETAIS	BANANA
3	999361204	FRUTAS&VEGETAIS	CARROT
4	999951863	PRIVATE LABEL	FRESH UHT MILK
5	999746519	NO LABEL	DRINKS
6	999401500	PRIVATE LABEL	MINERAL WATERS
7	999712725	FRUTAS&VEGETAIS	ONION
8	999749894	NO LABEL	FRESH PORK
9	999953571	FRUTAS&VEGETAIS	CITRUS
10	999356553	PRIVATE LABEL	SUGAR

Table 18: Criteria for traffic driver: time of transaction..

Traffic drivers are products that make customers visit the stores. They do not need to be profitable or high priced. The function of traffic drivers is to let customers in and buy other profitable products. We assume transaction time is a good indicator to pick traffic drivers. The top 1 product belongs to the bags category which is an exception as the high transaction time is due to the nearly necessary demand of shopping bags. Other products mostly belong to groceries and consumer goods that people consume daily and purchase frequently.

3. Known Value Items

customer_count			
	id	brand	category
1	999231999	PRIVATE LABEL	BAGS
2	999956795	FRUTAS&VEGETAIS	BANANA
3	999361204	FRUTAS&VEGETAIS	CARROT
4	999512554	FRUTAS&VEGETAIS	STRAWBERRY
5	999712725	FRUTAS&VEGETAIS	ONION
6	999953571	FRUTAS&VEGETAIS	CITRUS
7	999998053	PRIVATE LABEL	SALT
8	999967197	FRUTAS&VEGETAIS	TOMATO
9	999356553	PRIVATE LABEL	SUGAR
10	999944034	PRIVATE LABEL	REGULAR EGGS

Table 19: Criteria for Known Value Items: unique customer count.

Known value items refer to products that disproportionately drive the price value perception. So, in a grocery store it would include eggs and an automotive store might include motor oil,

and a convenience store it might include cigarettes. Here we assume unique customer count is a good indicator for Known Value Item as high unique customer count help us pick products that everyone buy and may have a sense of the price. Not surprisingly, Pernalonga has REGULAR EGGS as one of its top Known Value Items. As for pricing strategy, Pernalonga may give discounts on key fruit and vegetables for lower price perception.

4. Always Promoted, Strongly Promoted, Never Promoted

Highest Average Discount Rate			
id	brand	category	Average Discount Rate
1	999171880	NO LABEL GIRL TOYS	0.69
2	999236394	PORTAL DE SÃO BRÁ FINE WINES	0.68
3	999169848	M&M'S SEASONAL CHRISTMAS CHOCOLATES	0.64
4	999146004	ENCONSTAS DE ALQI FINE WINES	0.63
5	999262372	ENCOSTAS DE PIAS FINE WINES	0.62
6	999943438	FRUTAS&VEGETAIS CABBAGE	0.62
7	999234009	FERRERO BONBONS	0.61
8	999271417	PORTAL DE SÃO BRÁ FINE WINES	0.60
9	999146840	TAPADA DOS GAMA FINE WINES	0.60
10	999180965	FINISH DISHWASHING MACHINE DETERGENTS	0.59

Table 20: Top 10 products with highest average discount rate

Highest average discount rate evaluate the overall discount rate of a product. At pernalonga, a high discount rate seems to always appear with labeled non-essentials, such as toys, fine wines and chocolates. Those products get promotions to attract purchasing(as non-essentials) especially for the holiday season.

Highest Promotion Frequency			
id	brand	category	
1	999146564	REGUENGOS	FINE WINES
2	999146571	ERMELINDA FREITAS	FINE WINES
3	999245601	SAGRES	BEER WITH ALCOHOL
4	999649801	NO LABEL	DRY SALT COD
5	999146004	ENCONSTAS DE ALQUEVA	FINE WINES
6	158284005	CORPOS DANONE	YOGURT HEALTH
7	158284006	CORPOS DANONE	YOGURT HEALTH
8	153701007	ACTIVIA	YOGURT HEALTH
9	152576009	ACTIVIA	YOGURT HEALTH
10	153701006	ACTIVIA	YOGURT HEALTH

Table 21: Top 10 products with highest promotion frequency

At pernalonga, labeled alcohol and yogurt products are always promoted. This may due to the highly competitive and disperse market within each product category.

There are 61 products never promoted which have a zero average discount rate and promotion frequency. Those products belong to 28 brands and 36 categories. Compared to top 10 products with the highest average discount rate, nearly half of products never promoted are non labeled, which might indicate products without brand pay less attention to promotions. Children and baby food, medicine take up significant places in never promoted products. It may be due to the high willingness to pay to these specific categories as they are necessary and the quality matters.

Product clustering analysis

In this part, we want to find if there are other product groupings other than product categories we already mentioned above. For this purpose, we conducted product clustering using KMeans. Since KMeans uses Euclidean distance to measure the distance between points, we had to scale the attributes first. We chose to use the 7 attributes: sum of revenue, sum of profit, sum of transaction times, unique store count, unique customer count, average discount rate and promotion frequency in the KMeans model, so we used min-max method to scale those attributes.

One important issue of using KMeans is to decide the best K. Here we tried to find the optimal K via the gap statistic. The Gap statistics (Malika Charrrd, 2014) compares the total within intra-cluster variation for different values of K with their expected values under null reference distribution of the data. The estimate of the optimal K will be the value that maximize the Gap statistic.

We drew the Gap ~ K plot to show directly how the Gap changes depending on K, therefore we determined the best K.

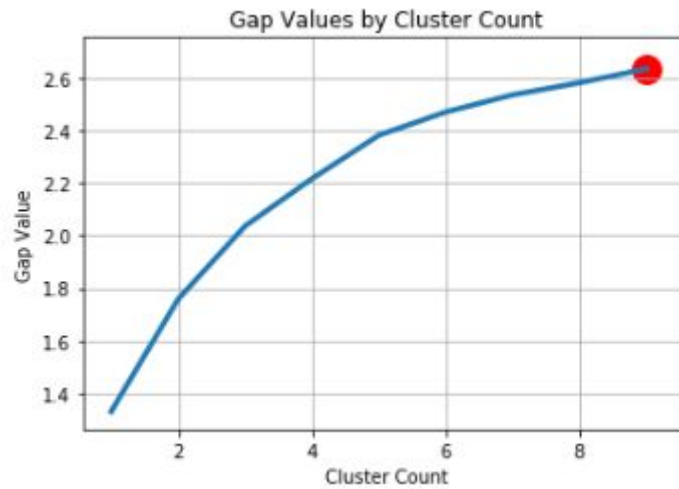


FIG 22: The plot of Gap value ~ K value. The value of gap constantly increase as the k increase, so we choose the elbow point where K = 5, indicating that clustering into 5 groups here is the optimal choice.

We divide the products into five groups. After running a KMeans model setting number of clusters as 5, we got the labels of every product ranging from 0 to 4, indicating which group it is in among those five groups. Also we got the 5-dimension coordinate for the five cluster centers. Then, we can explore the overall characteristics of those five groups from the five centers.

According to the following coordinates of the clusters and distribution of the attributes, we can determine what level each cluster has in terms of all those measurements we used.

	Transaction			Unique	Unique	Average Promotion	
	Revenue	Profit	Times	Customer Count	Store Count	Discount Rate	Frequency
Group1	0.004	0.002	0.002	0.064	0.573	0.097	0.089
Group2	0.002	0.001	0.001	0.021	0.207	0.085	0.047
Group3	0.002	0.001	0.000	0.026	0.321	0.501	0.152
Group4	0.030	0.020	0.013	0.256	0.898	0.100	0.137
Group5	0.011	0.005	0.003	0.127	0.747	0.487	0.271

Table 23: The coordinates of the five cluster centers, all of the figures are scaled to range from 0 to 1.

The above clusters can also be interpreted as below:

	Revenue	Transaction Times	Unique Store Count	Unique Customer Count	Profit	Average Discount Rate	Promotion Frequency
Group1	Low	Low	Middle	Middle	Low	Low	Low
Group2	Low	Low	Low	Low	Low	Low	Low
Group3	Low	Low	Low	Low	Low	High	Middle
Group4	High	High	High	High	High	Low	Middle
Group5	Middle	Middle	High	Middle	Middle	High	High

Table 24: The coordinates of the five cluster centers with each figures interperated as low, middle and high level.

As shown above, there are five groups with different characteristics.

Group 1 can be concluded as ‘necessity’ as it has relatively high popularity, low monetary benefit to Pernalonga. People consume those products daily and people may purchase necessities together as a family.

Group 2 can be concluded as a ‘niche product’ with low monetary benefit, low popularity and low promotion. Those products attract only a small amount of customers, however, those customers are willing to pay at no discount.

Group 3 can be concluded as ‘dogs’ with low monetary benefit to Pernalonga and low popularity. Customers have a low tendency to buy those products even if they do promotions at a high discount rate. However, with the limited benefit those products bring to Pernalonga, those products still take up resources like cash flow and space in store. Pernalonga may consider decreasing the product stock of this group or even stop selling them.

Group 4 can be concluded as ‘cash cow’ with high monetary benefit to Pernalonga and high popularity. Customers purchase those products frequently no matter of the promotion frequency and discount rate. The unit profit may not be large, but overall, with the high amount of purchase, it brings significant profit.

Group 5 can be concluded as ‘promotional goods’ with high discount rate and promotion frequency, not so high monetary benefit to Pernalonga. Those products are often at promotions and the promotions are the main reason for customer purchasing as they are not essentials.

Chapter 4: Identifying Natural Groupings of Stores

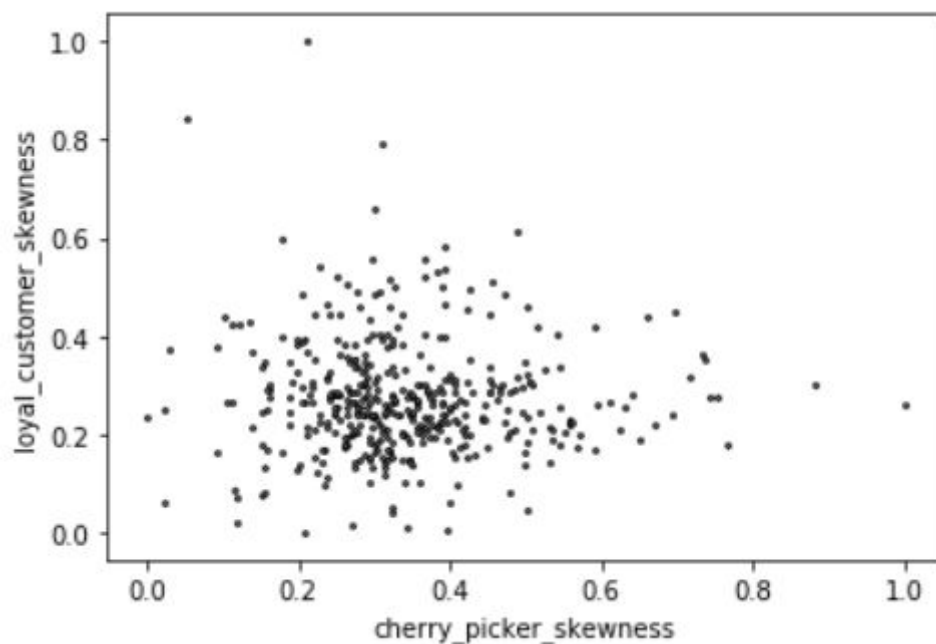
In this part, we would like to find out the natural groupings of stores. In particular, we would like to segment all the stores based on two attributes. First, the stores that are frequently visited by cherry-pickers, which we defined as people that tend to buy mostly on promotion. Second, the stores that are frequently visited by loyal customers, which we defined as people that tend to visit a few of their favorite stores very frequently.

Before actually starting the segmentation, we would like to calculate the two attributes that are mentioned above. Using the 'dispercent' variable we calculated earlier, which measures the average amount of offer a person would take among all of its transactions, we would be able to come up with an index that measures to what degree, a person is a cherry-picker. By comparing the skewness of the distribution of customers 'cherry-picker index' across stores, we would be able to figure out stores frequented by cherry-picker. In other words, the larger the skewness, the less frequent the store is in terms of visiting by cherry-pickers.

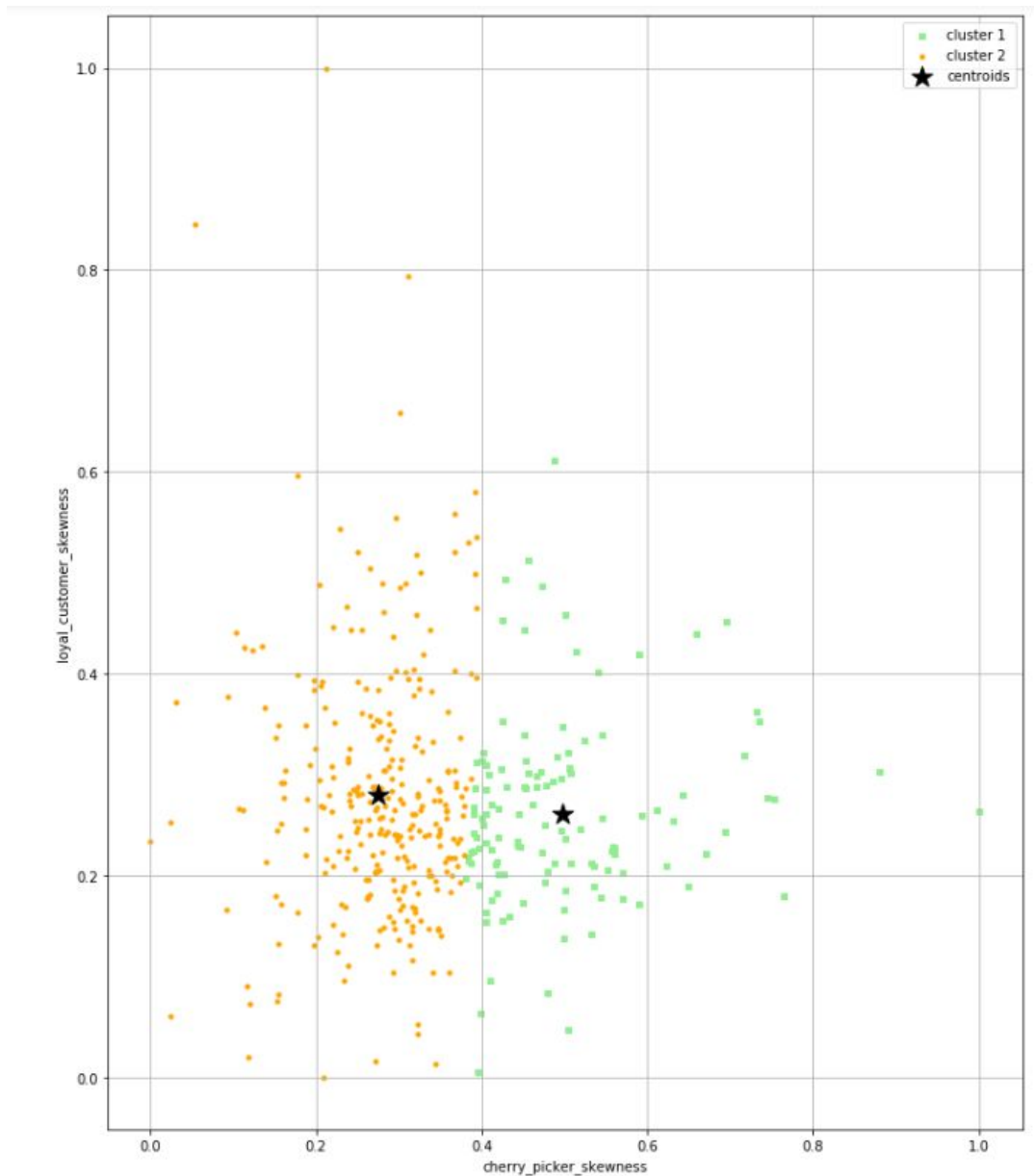
We can calculate the other attribute in a similar way. First, we calculated the skewness of the distribution of each customer's frequency in visiting each store. A larger skewness would mean the customer has many rather small counts of visited stores and fewer large counts of visited stores, indicating that he is very attracted to few stores, which he visits very frequently. These customers are considered to be loyal customers since they mostly only visit a few of their favorite stores. On the contrary, if the distribution of count numbers is left-skewed, which would result in a smaller or even negative skewness, it means the customer has many rather large counts of visited stores, indicating that he is attracted to many

stores in a significant way. These customers are considered customers with low store loyalty, since they frequently visit different stores without just sticking with a few.

Next, after successfully identifying those two attributes, we can start finding natural groupings of stores. First, we standardised both attributes using min-max scaling. Then, we plotted all the stores in terms of their relative position under the two attributes.

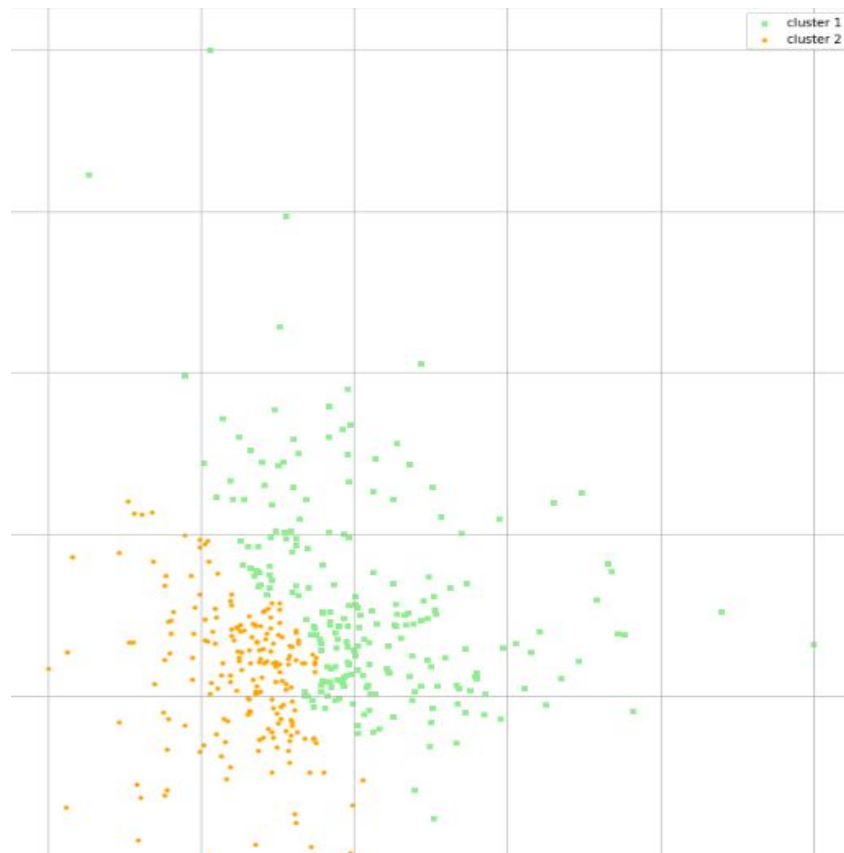
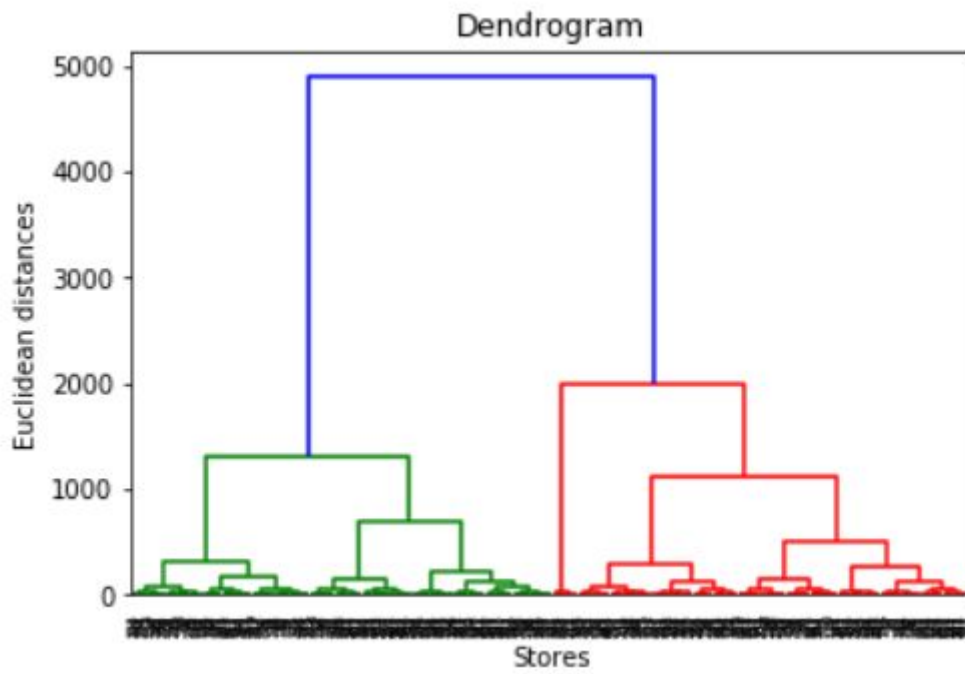


The first clustering method we chose is k-means clustering. The best number of clusters we identified is 2.



We can identify two groups of stores, with relatively the same level of customers in terms of their loyalty to the stores. On the contrary, the two groups of stores have different levels of customers in terms of the proportion of customers that are cherry-pickers. However, the resulting clusters are not that distinctive from the graph. So we tried using another clustering method.

The second method we tried is hierarchical clustering. Best number of clusters based on silhouette score is still 2. The resulting clusters are similar to that of k-means.



The natural groups of stores we have identified will definitely be beneficial when we start developing the marketing campaign.

Chapter 5: Conclusions

In this report, we did a complete exploratory data analysis for Pernalonga's customers, products and stores, and did reasonable clustering for each of them.

In terms of customers, we built an efficient way to check the best customers in the sense of revenues, profits, store visits and number of products using Tableau. We identified five customer clusters: ordinary people, category addicts, rich people, frugal people and shopaholics, and illustrated that the rich people and shopaholics are most valuable to the company.

In terms of products, we evaluate the best products and product groups based revenues, profits, volume, transaction times, unique customer count and unique store count. Furthermore, we pick the KVI, KVC, traffic drivers, known value items, always promoted items, never promoted items and strongly promoted items. We identified five product clusters: necessity, niche product, dogs, cash cow and promotional goods. The segmentation result suggests that dogs are the least valuable product group while cash cow is the most valuable product group.

In terms of stores, we identified two natural groupings of stores in terms of two attributes: stores frequented by cherry-pickers versus stores visited by mostly loyal customers. We were able to see that even though the segmentation is not that obvious, stores can still be grouped into two clusters based on the proportion of their customers being cherry-pickers. This information can be beneficial for our further development of personalized promotions based on the stores in the future.

In conclusion, we had a quite thorough understanding of the dataset received so far from Pernalonga and we are able to know the profiles and segmentations of the stores, products and customers. With the understanding in hand, we are confident to design personalized promotions for your customer groups after being provided with further details. We hope the initial insights we provided are helpful to your company and we are ready to take on the task of developing the marketing campaign for Pernalonga.