MKT 680 – Marketing Analytics
Segmentation Project

Group 6: Josh Chen, Eric Gu, Meng Wang, Kami Wu, Andy You

INTRODUCTION

Pernalonga, a leading supermarket chain of over 400 stores in Lunitunia, sells over 10 thousand products in over 400 categories. Pernalonga, as our client, wants our team to develop a marketing campaign to experiment on personalized promotions. As the first step, our team used the transaction data and the product data to perform descriptive analysis of customers, products and stores, as well as conduct segmentation for these three key business elements.

Our goal of this project is to have some basic understanding of customers, products and stores, and being able to use segmentation insights to power personalized promotions.

DATA UNDERSTANDING AND DATA PREPARATION

There are two sets of data provided to conduct the project - product_table and transaction_table. The product table contains 10,767 rows and 7 columns of data. The tranaction_table contains 29,617,585 rows and 12 columns. In product_table, there are 10,767 unique prod_id, 429 unique category id with 425 unique category_desc, 1,476 unique subcategory_id with 1,430 unique sub_category_desc, and 1,169 unique brand_desc. We have found that several unique category_id have the same category_desc, which exactly explain the inconsistent numbers between category id and category description names. We filtered out those

pairs and decides to leave it the same way as they are, as we believe it is possible for several category ids to have the same category description name. In transaction table, there are 7,920 unique customers, 753 unique transaction id, 727 unique transaction dates, 421 store ids, 10,770 product ids. We found out there are a severe amount of cases that one transaction has zero amount of discount offers but non-zero value of the transaction discount amount.

We have noticed that the original transaction id is not unique, since one unique id can represent multiple customer ids or store ids. In order to solve that, we extract the first eight digit of transaction id, then paste it with customer id and store id. In such way, we are able to make sure each transaction can be uniquely identified.

After analyzing the business problems our clients held out, we thereby decided to split the datasets into three parts focusing on different perspectives: customer, products and stores. For each of the perspectives, we defined several key attributes based on our business understanding. The customer dataset contains 11 attributes: total revenues, and transactions each customer has made, total distinct products, total distinct stores each customer has visited, the distinct categories, brands each customer has purchased, the probability of discount purchases, and the average discount rate of each customer, and finally RFM (Recency, Frequency, Monetary). Recency describes the freshness of customer activity: how many days since customers' last visit (baseline is the last day of transaction date). Frequency describes the frequency of customers' visits: how many times have each customer visited for the past time period. Monetary describes the customers' willing to pay: how much customers have purchased. Figure 1 displays the detailed descriptive statistics of our defined customer dataset.

```
> summary(customer[,2:12])
 total_revenue   total_transaction  t_d_product       t_d_store       t_d_category     t_d_brand
 Min.   : 3867   Min.   :  174     Min.   : 113.0   Min.   : 1.000   Min.   : 64.0   Min.   : 29
 1st Qu.: 6292   1st Qu.: 2930     1st Qu.: 808.0   1st Qu.: 3.000   1st Qu.:192.0   1st Qu.:184
 Median : 7465   Median : 3585     Median : 975.0   Median : 5.000   Median :210.0   Median :219
 Mean   : 7855   Mean   : 3740     Mean   : 982.3   Mean   : 6.637   Mean   :208.4   Mean   :218
 3rd Qu.: 9138   3rd Qu.: 4418     3rd Qu.:1146.0   3rd Qu.: 9.000   3rd Qu.:227.0   3rd Qu.:252
 Max.   :14020   Max.   :10790     Max.   :1972.0   Max.   :76.000   Max.   :288.0   Max.   :414
 p_o_discount_purchase avg_discount_rate    recency         frequenci        monetary
 Min.   :0.1023        Min.   :0.03412   Min.   : 0.000   Min.   : 38.0   Min.   :  8.519
 1st Qu.:0.2568        1st Qu.:0.12709   1st Qu.: 0.000   1st Qu.:272.0   1st Qu.: 18.931
 Median :0.3100        Median :0.15865   Median : 0.000   Median :318.0   Median : 23.259
 Mean   :0.3163        Mean   :0.15943   Mean   : 1.418   Mean   :336.4   Mean   : 25.015
 3rd Qu.:0.3680        3rd Qu.:0.19034   3rd Qu.: 1.000   3rd Qu.:388.0   3rd Qu.: 29.174
 Max.   :0.7925        Max.   :0.37380   Max.   :30.000   Max.   :724.0   Max.   :226.124
```

Figure 1: Descriptive statistics of the customer dataset

The product dataset contains 6 attributes: total revenues, total transactions, total number of distinct customers (who has purchased) of the product, the total number of stores that the product has been purchased, and finally the average discount rate of the product and the percentage of times for the product being on-sale condition. Figure 2 displays the detail information about the descriptive statistics of product dataset.

```
 total_revenue      total_transact   total_distinct_customer  total_stores
 Min.   :   500.0   Min.   :     3   Min.   :   2.0          Min.   :  1.0
 1st Qu.:   939.5   1st Qu.:   307   1st Qu.: 189.0          1st Qu.:124.0
 Median :  1898.0   Median :   757   Median : 405.5          Median :220.0
 Mean   :  5802.4   Mean   :  2764   Mean   : 725.8          Mean   :221.9
 3rd Qu.:  4591.7   3rd Qu.:  2061   3rd Qu.: 882.8          3rd Qu.:324.0
 Max.   :602109.4   Max.   :769890   Max.   :7862.0          Max.   :419.0
 avg_discount_rate   percentage_discount_product
 Min.   :0.0000432   Min.   :0.0005537
 1st Qu.:0.0227892   1st Qu.:0.0362067
 Median :0.1095774   Median :0.0715926
 Mean   :0.1587641   Mean   :0.1185603
 3rd Qu.:0.2787683   3rd Qu.:0.1564843
 Max.   :0.6895196   Max.   :1.0000000
```

Figure 2: Descriptive statistics of the product dataset

The store dataset contains 8 attributes: total revenue, total transactions, total distinct customers, products, product categories, and product brands of each store, the average discount rate, and finally the percentage discount purchases of each store. Figure 3 displays the detailed descriptive statistics summary of store dataset.

```
> summary(store[,-c('store_id')])
 total_revenue      total_transact  total_distinct_customer total_distinct_products
 Min.   :   372.7   Min.   :   10   Min.   :  6.0           Min.   : 147
 1st Qu.: 88067.7   1st Qu.: 4038   1st Qu.: 75.0           1st Qu.:4929
 Median :123258.4   Median : 5802   Median :104.0           Median :5810
 Mean   :148472.2   Mean   : 6755   Mean   :125.5           Mean   :5681
 3rd Qu.:181250.3   3rd Qu.: 8359   3rd Qu.:150.0           3rd Qu.:6602
 Max.   :786520.9   Max.   :30357   Max.   :836.0           Max.   :9521
 total_distinct_categories total_distinct_brands avg_discount_rate percentage_discount_purchase
 Min.   : 92.0             Min.   :  53.0        Min.   :0.07593   Min.   :0.1677
 1st Qu.:345.5             1st Qu.: 679.0        1st Qu.:0.14545   1st Qu.:0.2877
 Median :364.0             Median : 765.0        Median :0.15796   Median :0.3089
 Mean   :356.0             Mean   : 740.3        Mean   :0.15741   Mean   :0.3102
 3rd Qu.:376.0             3rd Qu.: 834.5        3rd Qu.:0.17086   3rd Qu.:0.3323
 Max.   :418.0             Max.   :1084.0        Max.   :0.23856   Max.   :0.4727
```

Figure 3: Descriptive statistics of the store dataset

DESCRIPTIVE STATISTICS AND SEGMENTATION ANALYSIS

Product Segmentation

For product segmentation, we first ran some descriptive analysis to rank all the products in terms of volumes, revenue, transactions, customers, percentage of discounted products and average percentage rate. Detailed rankings can be generated by running the R script but a brief overview is provided here:

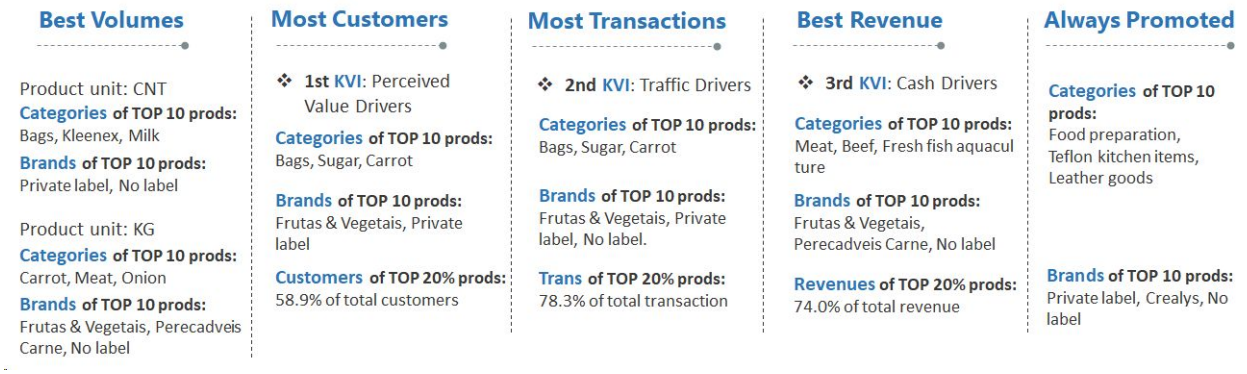

Figure 4: Descriptive Analysis of Products

Here are some key takeaways:

1. Fresh products appear to be the most important products for Pernalonga because most of the top-ranked products in terms of revenue, transaction and customer are fresh products.
2. If we rank the products by the amount of revenue each brings in, the top 20% of products contribute to almost 75% of total revenue generated from all the products. This result is very close to what 80/20 rule suggests.

We also defined a few KVIs here. The 1st KVI are perceived value drivers. Those are high-popularity products that have large customer base. The 2nd KVI are traffic drivers. Those are high-demand products that have high number of transactions. The 3rd KVI are cash drivers. Those are the products that bring in the most revenue. Marketing professionals at Pernalonga can decide where to set the cut-off line in the product ranking tables based on business needs. For example, we can label the top 20% of products that generate the most revenue as cash drivers because of the 80/20 rule.

The 4th KVI are the basket drivers. Those are the items which lead to buying additional products. We went through an iterative process to find the basket drivers. Due to the limiting computing power and memory, we randomly sampled 500K observations for the modeling process. For each unique transaction ID, we created a list of all purchased products (items in a basket) in this transaction. Then, we performed combinations on the list to find all the co-purchase pairs as shown below:

	tran_id	product_1	product_2
1	2016061536109982560	999412624	999495537
2	2016122369779596345	999262736	999339933
3	2016041242749616482	999380861	999545378
4	2016102875859738187	999247012	999679887
5	2016030388339693486	999231104	999252445
6	2016040929349622315	266417009	999344744
7	2016040929349622315	266417009	999680491
8	2016040929349622315	999344744	999680491
9	2017021359729769322	999172070	999504322
10	2016060564299942541	999182268	999944032

Figure 5: Co-purchase product pairs

For example, if we had n distinct purchased products in one transaction, we would have C(n,2) co-purchase pairs. Next, we calculated the frequency of all co-purchase pairs:

product1	product2	frequency(weight)
999231999	999956795	41
999231999	999951863	20
999231999	999361204	16
999231999	999712725	15
999231999	999747259	15
999231999	999455829	14
999231999	999749894	14
999356553	999361204	14
999361204	999712725	14

Figure 6: Co-purchase product pairs frequency

Then, with R's "igraph" package, we generated a co-purchase network. The first two columns in the above table are used as an edge list and the frequency is used as weight for each edge(tie). We computed the betweenness centrality for each node (product) and the resulting network is shown below:

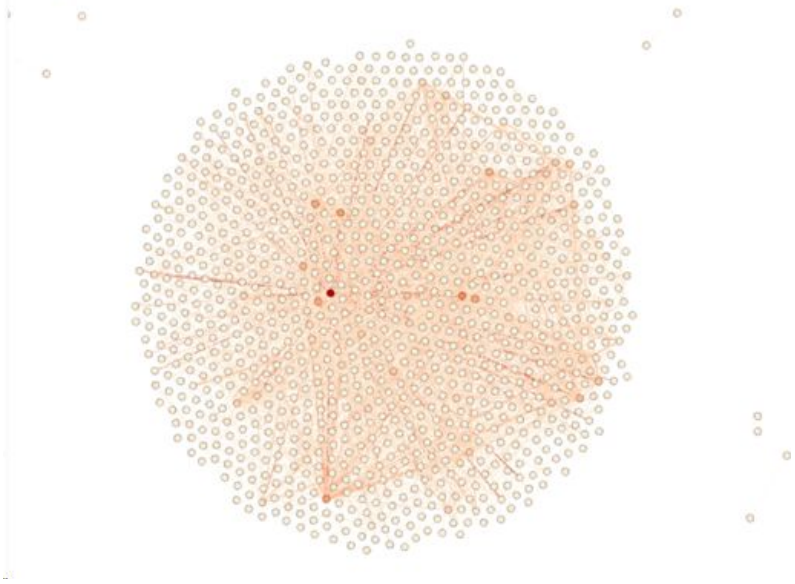

Figure 7: Co-purchase network showing high betweenness nodes

Each node in the network represents an unique product and they're color-coded in a way so that dots that are "redder" have higher betweenness centrality. Products have higher betweenness tend to be the bridges between different products so they can trigger additional purchases. We ranked products in descending order in terms of betweenness.

prod_id	betweenness	category_desc_eng
999231999	3653319.9	BAGS
999956795	2369328.9	BANANA
999361204	1313309.0	CARROT
999951863	1050182.7	FRESH UHT MILK
999401500	890386.4	MINERAL WATERS
999712725	875374.5	ONION
999356553	764250.1	SUGAR
999957158	703529.5	ZUCCHINI
999953571	698248.9	CITRUS
999401572	670645.8	MINERAL WATERS
999421692	656600.9	OIL
999749894	639577.6	FRESH PORK
999944034	636541.2	REGULAR EGGS
999749469	550738.8	FRESH BEEF
999967197	539150.8	TOMATO

Figure 8: High betweenness products and their corresponding category

We can see that most of products with high betweenness are fresh products or cooking ingredients. This is reasonable because when one decides to cook and start picking up fresh products or cooking ingredients, they'll very likely end up buying a lot of food-related items. And that's the reason those are basket drivers.

Next, we used k-means clustering to find if there are any natural groupings of products. Firstly we used min-max standardization to rescale every key attribute we selected. Then we determined the optimal number of k by using elbow graph, average silhouette graph and gap statistics graph. After comparing the three graphs, we decided to use the elbow graph as the result was more reasonable and accorded with our business understanding. The optimal value of k was five based on the elbow graph.

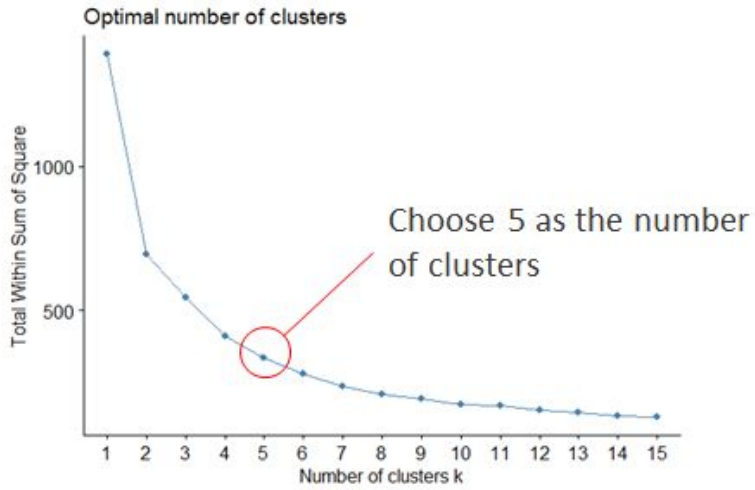

Figure 9: Elbow graph for choosing the optimal k on product data

We fitted k-means models several times to find the steady clusters. Then we used the values of centroids to interpret the meaning of each cluster.

Cluster	Total revenue	Total transactions	Distinct customers	Distinct stores	% of discount products	Discount rates	Group	Typical categories
1	High	High	High	High	Medium	Low	Cash-cow	Garlic, Apple
2	Low	Low	Low	Low	Low	Low	Hibernation	Oral hygiene
3	Medium	Medium	Medium	Medium	High	High	"Save for later"	Personal deodorant
4	Medium	Medium	Medium	Medium	Low	Low	Necessity	Drinks
5	Low	Low	Low	Low	High	High	Lost Kids	Games

Figure 10: Result for k-means modeling on product dataset

We ended up with 5 different product groups. The first cluster is called “cash-cow”. Those are popular products that can bring in a lot of revenue and have large customer base. Those are high-value products that customers would buy them at any price even when the discount rates are low. Typical product categories are garlic and apple.

The second cluster contains “hibernation products”. Those are the products that have relatively smaller customer base and people could use them for a long time after one purchase. But whoever need those products would buy them at any price even when there’s no promotion. A typical category is oral hygiene. For example, not everybody uses dental floss but those who think these are necessities would always buy them. However, a unit of dental floss is not expensive and one can use it for a very long time before another purchase. These factors explain why “hibernation products” have low revenue and low number of transactions.

We named the third cluster “save for later”. These are daily/common products that people buy all the time, but usually when they are on promotion. A typical category is personal deodorant. Customers could buy and stack a few of them to “save for later” when those products are being promoted.

The fourth segment’s products are necessities. They are very comparable with “save for later” group in terms of revenue, number of transactions and customer base. However, only a small portion of products in this cluster are promoted and they usually have low discount rates. A typical category is drinks. When someone is thirsty and would like to get something to drink,

they would not wait until there is a discount.

Products in the fifth segment are “lost kids”. Like “hibernation products”, they also have a small customer base and do not bring in much revenue. However, they are hard to sell even though most of the products in this group are on promotion and they usually have high discount rates. A typical category is games. Most people download games nowadays instead of buying physical copies and game-lovers would prefer to visit an electronics or gaming store for game shopping. Games in a supermarket are like “lost kids” that do not get a lot of attention and care.

Customer Segmentation

To understand the customer behaviour, we explored the customer-level distributions for each continuous attributes. We found that there is no clear evidence showing the minority of customers produce the majority of revenue. Top 10 % of customers only cover 15 % of the revenue, 16% of transactions/store visits (we assume each customer only visit one store), 17% of the amount of products sold in count and 19% of products sold in kilograme. In order for a further understand our customer behavior, we conduct the basic RFM value to classify our customer. The first cluster is called “Champions” (covers 21 customers among all 7920 customers), who comes recently (less than 1) and frequently (above 75% of the frequency), and spend a large amount of money (above 75% of the monetary) purchasing products. We can reward those “Champions” and promote new products. The second cluster is called “loyal customer” (covers 28 customers among all 7920 customers), who comes frequently (above 75%

of the frequency) and spend a large amount of money (above 75% of the monetary) purchasing products. We are able to provide more discounts and promotions to further engage them. The

third type of customers is potential loyalists (covers 250 customers among all 7920 customers), who comes recently (less than 1) and buy a large amount value of products (above 75% of the frequency). We can recomende membership program and valuable products to invite them in. The next cluster is called “At risk” (covers 0 customers among all 7920 customers), who used to purchase frequently (above 75% of the frequency) with a large value of purchased (above 75% of the monetary) but have not come back to purchase for a long while (more than two weeks). We can offer renewals and send personalized emails or calls to retain those customers. The last group of customers is called “already lost” (covers 0 customers among all 7920 customers), who does not come recently (more than two weeks) and frequently (below 25% of the frequency) and spend less money (below 25% of the monetary) on purchasing products. We have already lost them and better to ignore them for the time and money interest.

Before coming to the data modeling part, we did the min-max normalization and excluded attributes with high correlation. We then used k-means to classify customers by defining the distinguishing attributes among all different clusters. After several iterations with k-means on the customer dataset, we decided to find the possible best k by plotting elbow graph, average silhouette graph and gap statistics graph. After comparing the three graphs, we decided to use the gap statistics graph as it make more business sense. Figure 11 displays the gap statistics graph for choosing the optimal K. We finally are able to clearly state that 5 is the optimal number of clusters to choose.

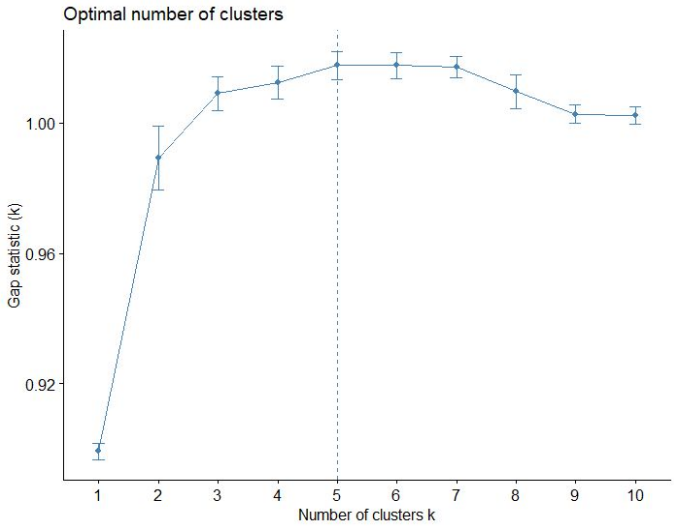

Figure 11: Gap-Statistics graph for choosing the optimal k on customer data

```
> l[,-c('cust_id','t_d_category')][,lapply(.SD,mean),cluster][order(cluster)]
   cluster total_revenue total_transaction t_d_store avg_discount_rate
1:       1      6102.666          318.4701  6.991761         0.1981292
2:       2     11027.336          385.6438  6.701936         0.1477790
3:       3      7990.633          537.1621  6.044101         0.1365200
4:       4      8403.485          331.2382  7.928780         0.1812973
5:       5      6679.015          313.6899  5.353033         0.1201924
```

Figure 12: Five centrics generated from k-means

Figure 12 shows clearly the five centrics we got. The first identified cluster is cherry-picker. They produce low-level revenue and the average percentage of discount on their purchases is high. Cherry-pickers are discount-lovers, and they only search for the low-valued discount products. The second identified cluster is high value customer. They produce most revenue and keep in medium level on other attributes. They are the best customers who buy the products they like at any price. We named the third identified cluster as “nearby turtle”. They produce the median-level of revenue and the average percentage of discount on their purchases. They also have a low level of total distinct store purchases and extreme high level of total

transactions. Nearby turtles are customers who live near the stores, visit stores and shop frequently, even they don't buy much items or make much revenue for every transaction. The four identified cluster is wholesale lover. They don't buy frequently, but they buy in many different stores and at high discount rate. Wholesale lovers are professional buyers who visit different stores to buy a large amount of discounted products. The last identified cluster is lost customer or passer. They keep a low level for each attributes. They seldom shop at our shops, even the products are on sale.

Group	total_revenue	total_transaction	total_store	avg_discount_rate	Description
1	Low	Low	Medium	High	Cherry Picker
2	High	Medium	Medium	Medium	High Value Customer
3	Medium	High	Low	Medium	Nearby Turtle
4	Medium	Low	High	High	Wholesale Lover
5	Low	Low	Low	Low	Lost Customers /Passer

Figure 13: Result for k-means modeling on customer dataset

We provide promotion suggestions for each type of customers. For cherry-pickers, we can put promotion coupons in our stores for customers to pick up. This is a price discrimination method, and cherry-pickers have the motivation to pick up coupon and save it until next time they use. For high value customers, we can offer membership to build long-term relationship. For nearby turtles, we don't need to change a lot because our stores are rigid demand for them. We can even maintain the original price of daily necessities to gain more profit. For wholesale lover, we can offer quantity-based promotion plan. If they buy more than certain threshold, they are automatically provided a high discount. For lost customers and passer, we can just ignore them for the time and money interest.

Store Segmentation

After exploring the store data, we noticed that several stores have very few numbers of transactions. For example, store 351 only has 10 unique transactions over the 2-year period. We infer that such stores were newly opened, so no enough data was collected at the time when it was retrieved from the database. Our concern is that a few transactions might not be representative enough to demonstrate the performance of a store. Consequently, the cluster might be biased by these observations. Therefore, we decided to remove the store with less than 50 transactions. By ranking the stores based on different measurements, we were able to identify the top performing stores from the descriptive results. One interesting finding is 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. In addition, unlike products that follow a 80/20 rule, stores holds a 40/20 rule where top 20% of the stores generates 40% of the revenue and transactions.

Besides general descriptive insight about the stores, we would also like to identify stores most visited by cherry pickers and with most loyal customers. By understanding the client based at the store level, each store can implement targeted promotion strategies to increase revenue.

While the characteristics of cherry pickers are not clearly defined, we believe they tend to purchase discounted products. Therefore, the average percentage of discount of customers' purchases at the store level and the percentage of purchases that include at least one discount are good measures to infer a store's popularity among cherry pickers. To simplify the problem, we decided to aggregate these two measures while allocating equal importance on them. We

generated the cherry score measure that ranges from 0 to 1 as an indicator to identify cherry picker's favorite stores. The score was calculated by normalizing the average discount rate and the percentage of discounted purchase and taking the mean of the two measures. By ordering stores based on cherry score, we can infer which stores are most visited by cherry pickers. The average discount rate of the top 20 cherry picker's favorite stores is 20.2% while the all store is 15.7%. The average percentage of discounted purchase for the top 20 stores is 39.6% while the all store average is 31%. It suggests that the frequency of discount might be more attractive for cherry pickers than the actual discount amount.

To identify stores with a high percentage of loyal customers, we first would like to define loyal customers as those who repetitively visit the same store regardless of the dollar amount of their purchase. To identify loyal customers, we decided to analyze a customer's purchases frequency relative to the distribution generated by all the customers visiting the same store. We believe the number of transactions of each customer can capture this information. Implementing the idea, we classified customers as loyal if their transactions are above the 75% threshold on the distribution. The 75% threshold is set based on the observation that a large portion of the customers would become loyal if the threshold was lowered to 60%. The result suggests that the average stores have a loyal customer base of 3.9% while 27 out of 419 stores have a loyal customer base higher than 10%. For these stores, each customer makes purchases two times more frequently than customers visiting other stores. Thus, the stores with high share of loyal customers might want to apply strategies to increase the dollar amount of each purchase to increase revenue.

After categorizing stores based on descriptive measures, we also implemented k-means clustering to explore whether unobvious clusters of stores exists. First, we used the "elbow approach" to determine the optimal k for clustering.

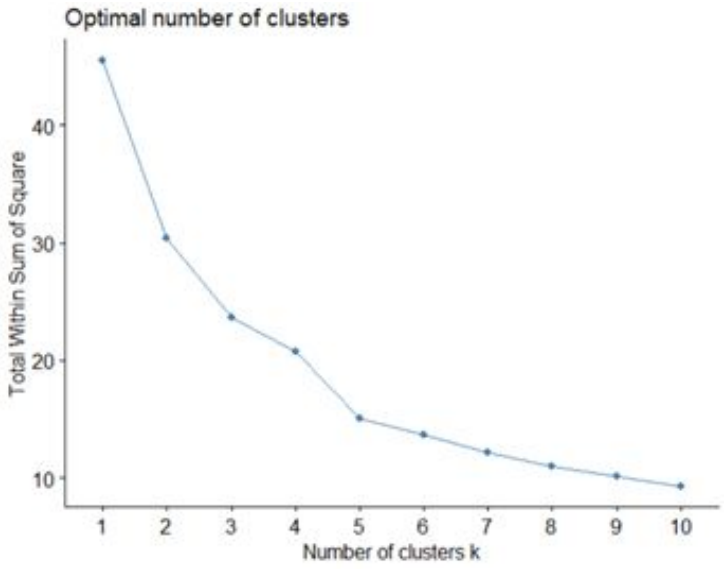

Figure 14: Elbow graph for choosing the optimal k on store data

Then we implemented the k-means clustering algorithm with the optimal 5 clusters as illustrated above by the "elbow" graph. However, looking at the store measures by cluster, we noticed three clusters have minimal differences in their total revenue, total number of transactions, and other measures. Since the observed differences are negligible, we decided to apply the k-means clustering again with only three centroids to generate more meaningful insight. The new segmentation resembles the characteristics compared to the cluster generated by simply merging the three similar segmentations. It suggests that reducing the clusters to three is still appropriate for our analysis. The result of the store segmentation and their corresponding based on k-means clustering is presented below.

Cluster	Total revenue	Total transaction	Distinct customer	Discount products	Discount score	Group	Store Type	Number of store
1	High	High	High	High	Medium	Large Scale	Supermarket	68
2	Medium	Medium	Medium	Medium	Medium	Middle Size	Drug store	226
3	Low	Low	Medium	Low	Low	Small or distanced	Convenience store	122

Figure 15: Result for k-means modeling on store dataset

Based on our business understanding, the segmentations reflect three types of stores: large supermarket, medium size pharmacy, and small convenience store. Cluster 1 captures stores with high revenue, high traffic, and diversified products. It is typical for supermarkets as it fulfills the demands for all types of customers. Cluster 2 represents a medium size store for a specific group of products since it generates medium revenue and has a medium customer base but has fewer products than supermarkets. It resembles the characteristics of a pharmacy or food store that mainly provides products in a certain categories. The last cluster shares the feature of a small convenience store as it offers limited discounts on products. The store segmentation will provide more insight once it is analyzed with the customer and product segmentations.

BUSINESS INSIGHTS AND DEPLOYMENT

a. Business Insights

After determining the segmentation of customers, products and stores, we examined correlation tables among three perspectives (customer, product, store), in order to reconfirm the rightness of the segmentations and draw insights from the combination results. The outcome is shown below:

Customer \ Product	Cash-cow	Hibernation	Save for later	Necessity	Lost kids
Cherry picker			😍	😐	
High value	😍	😍	😍	😍	😍
Nearby turtle				😍	😐
Wholesale lover		😐			😍
Lost customer		😍	😐		

Figure 16: Correlation between customer clusters and product clusters

Customer \ Store	Large	Medium	Small
Cherry picker		😍	😐
High value	😍		
Nearby turtle		😐	😍
Wholesale lover		😍	😐
Lost customer	😐		😍

Figure 17: Correlation between customer clusters and store clusters

Product \ Store	Large	Medium	Small
Cash-cow			😍
Hibernation	😍		
Save for later		😍	
Necessity	😍		
Lost kids	😍		😐

Figure 18: Correlation between product clusters and store clusters

The definition of the correlation here is that for each cluster the proportion of the occurrence of the cluster from another object. For example, in the customer-product correlation table, we calculated the proportion of the cash-cow products bought by each segmentation of

customers, and compared the proportion on the customer dimension. In other words, the proportion will be summed into one vertically, but not horizontally.

Here are some key takeaways:

a. Cherry picker customers love "save for later" products because these products generally have higher discounts, while they are not that interested in necessity products because these products generally have lower discounts.
b. High value customers share similarly high interests in each segmentation of products. That says high value customers would love to buy each kind of products. That probably is one of the reasons why they are high value customers.
c. Nearby turtles prefer necessities the most, which is corresponding to their characteristics.
d. For the customer-store correlation, medium stores are prefered by cherry pickers and wholesale lovers, which seems to be a very interesting outcome. We think that might be related to the site selection process that makes medium size stores locate in neighborhood with higher proportion of cherry pickers and wholesale lovers.
e. Cash-cow products are the most popular products in small scale stores which makes perfect sense since small scale stores tend to stock more valuable products in order to hit higher ROI.
f. Large scale stores have a higher proportion of hibernation products and lost kids products. Large scale stores would stock any kinds of products no matter how much value they can drive.

b. Future Deployment

Based on the previous analysis, we have a deeper understanding of Pernalonga's business. For future deployment of the segmentation results, there are several possible implementations:

a. Calculate customer value and recognize key customers
b. Develop suitable pricing strategy for each product segmentation
c. Optimize in-store product placement and product distribution
d. Personalize loyalty program for different groups of customers or stores
e. Personalize promotion campaign by having the knowledge of which customers to target for which products at which stores

At the early stage of deployment, we advise to implement segmentation-related business strategies at limited scale to examine their validity.