MKT 680 Marketing Analytics

Report for Project 1: Segmentation

Group 7: Myra Liu, Philipp Scherbel, Khasim Shaik, Jacky Yang, Jingwen (Kivi) Zuo

Date: 02/06/2019

1. Overview

This report is about segmentation of the customer base of Pernalonga, a leading supermarket chain with over 400 stores in Lunitunia. Pernalonga wants to experiment on personalized promotions to grow revenue. There are 7920 customers in the data, which have to be grouped and clustered for further direct personalized marketing. This section of the report deals with data exploration. In addition, a summary of the features will be conducted to provide an overview for subsequent topics like clustering.

There were two datasets provided, that needed to be merged:

- Transactional data consisting of 29617075 rows, each having information about one transaction line (item line)
- Product data with 429 categories and 10767 products

After merging the data, we have ~ 30 mio detailed transactional data on the transaction line level. Thus, one transaction consisting of multiple products will be displayed in multiple columns - with the same transaction ID. There are 6 numerical features: Sales Amount, Sales Quantity, Discount Amount, Discount yes/no, Amount paid per single item, unit price per item and 12 categorical features and IDs such as category ID, subcategory ID, store ID, transaction # etc.

The distribution of the numerical features is as follows:

tran_prod_sale_amt	tran_prod_sale_qty	tran_prod_discount_amt	tran_prod_offer_cts	tran_prod_paid_amt	prod_unit_price
Min. : 0.010	Min. : 0.001	Min. :-1400.2500	Min. : 0.0000	Min. : -1.41	Min. : 0.0075
1st Qu.: 0.900	1st Qu.: 1.000	1st Qu.: -0.2400	1st Qu.: 0.0000	1st Qu.: 0.84	1st Qu.: 0.7400
Median : 1.590	Median : 1.000	Median : 0.0000	Median : 0.0000	Median : 1.37	Median : 1.3900
Mean : 2.503	Mean : 1.668	Mean : -0.4028	Mean : 0.3409	Mean : 2.10	Mean : 2.0930
3rd Qu.: 2.790	3rd Qu.: 2.000	3rd Qu.: 0.0000	3rd Qu.: 1.0000	3rd Qu.: 2.32	3rd Qu.: 2.4900
Max. :3371.250	Max. :2112.000	Max. : 0.0000	Max. :76.0000	Max. :1971.00	Max. :399.0000

Figure 1: Summary statistics of numerical variables

Some interesting insights were gained from plotting the categories, stores and products against their contribution to overall sales of Pernalonga.

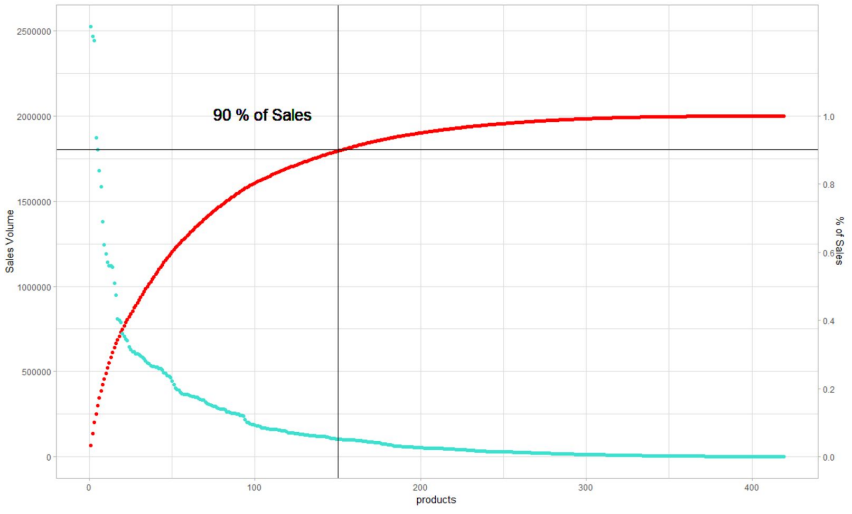


Figure 2: Contribution of products (here at the level of categories) to overall sales

The graph shows the very popular distribution of top sellers and sales contributors that account for a large amount of the sales volume. In addition, it shows the long tail of products, that need to be provided to the customers for the sake of completeness - they are barely sought after butt still need to be provided.

The data does not have any missing values. This initial exploration serves the purpose to define and analyze features. In this specific case, as k-means will be applied in subsequent steps, it is important to note that the continuous variables have to be normalized to create equal distances and thus weights for the algorithm. This will be explained in detail in the following sections.

2. Modelling method

The following section is about the modelling techniques used for the segmentation, data preparation and how to determine optimal number of clusters. K-means is used as an unsupervised learning method to segment customers, products and stores following the following steps:

- Standardize all features in the data set;
- Use the elbow method to find the best number of clusters;
  The elbow method measures the sum of squared errors (SSE) against the number of clusters used in the model. Although smaller SSE leads to better results, the graph shows that the SSE decreases converges to null as k increases. SSE becomes null when k equals the number of observations. Therefore, the goal is to choose a relatively small k with a relatively low SSE, and the “elbow” usually represents where the return begins to diminish by continuously increasing k.
- Use silhouette score to find the best number of clusters;
  The score ranges from -1 to 1, and the higher the score the more sparse the segmented clusters. A higher score indicates better results.
- As a final step, apply k-means clustering with the optimal number of clusters.

To analyze the cluster features, we find the observations that are the closest to the centroids and analyze the differences among them since they are the most representative subjects (customers, stores, products) of their clusters.

3. Segmentation Analysis

In this chapter clustering and segmentation analysis is discussed in detail. The data is clustered three ways: the customers, stores and products. This approach attempts to gain holistic insights on all levels. The analysis is used to develop strategies for customized promotions, direct marketing and driving sales.

3.1 Clustering of Customers

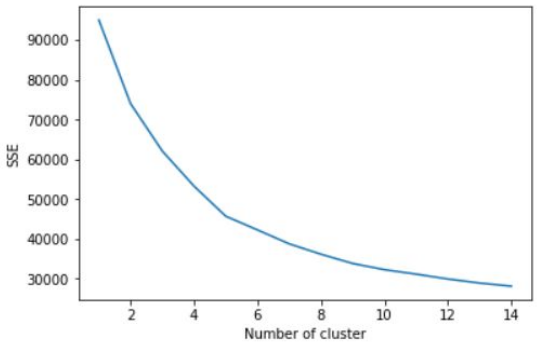
Based on the attributes provided, there are four categories that we would look into for customers. Customers are assessed on their basket diversity, promotion sensitivity, store loyalty and purchase behaviors.

1. At first, the basket is captured from the customer's two-year purchase data, and basket diversity is calculated by the number of unique products purchased.
2. Secondly, customers' promotion sensitivity is measured by how sensitives customers are to promotion from three perspectives. These measures capture how much a customer seeks out and buys products that are on discount.
    - the ratio of discounted dollar amount
    - The ratio of discounted products,
    - discounted percent on products (mean and quartiles). Both mean and quartiles are used to better capture different distributions for each customer's transactions.
3. Thirdly, store loyalty is calculated based on how many unique stores that a customer has purchased from.
4. The last feature is measuring purchase behaviors on the sales amount per transaction, also using mean and quartiles to capture various distributions for customers.
   Transaction is calculated based on unique combination of customer ID and transaction ID.

Feature	Attribute
Basket Diversity	Number of unique products
Promotion Sensitivity	Total discounted amount/sales # discounted products/total discount % per product avg discount % on products Q2 discount % on products Q3 discount % on products Q4 discount % on products
Store Loyalty	# of unique stores visited
Purchase Behavior	Avg sales amount per transaction Q2 sales per transaction Q3 sales per transaction Q4 sales per transaction

Figure 3: Features and attributes selected for customer segmentation

Subsequently, all features are implemented to the k-means model to find the optimal segmentations. The optimal number of clusters is chosen by evaluating both the SSE and Silhouette score at 5.




```
For n_clusters = 2 The average silhouette_score is : 0.2507608926323644
For n_clusters = 3 The average silhouette_score is : 0.26104959710096304
For n_clusters = 4 The average silhouette_score is : 0.26215148854665193
For n_clusters = 5 The average silhouette_score is : 0.27151152696715924
For n_clusters = 6 The average silhouette_score is : 0.23430865586926228
For n_clusters = 7 The average silhouette_score is : 0.1972048385086853
For n_clusters = 8 The average silhouette_score is : 0.18600984815810426
For n_clusters = 9 The average silhouette_score is : 0.19097618833967028
For n_clusters = 10 The average silhouette_score is : 0.1896125264508456
For n_clusters = 11 The average silhouette_score is : 0.17680452681254766
For n_clusters = 12 The average silhouette_score is : 0.17376438233027658
```

Figures 4 & 5: Result of "Elbow" method and Silhouette scoring

To better illustrate the result of clustering, five centroids' features are presented below with a color scaling on their values. The highlighted values will be used to assess the characteristics of each cluster later.

Percent of products on promotion	Number of store visited	Number of unique products	Percent of discounted sales	Avg discount % on products	Q2	Q3	Q4	Avg sales per transaction	c.Q2	c.Q3	c.Q4	Size
-0.084216	0.284907	0.276555	-0.268788	-0.05134	-0.315679	0.242068	-0.055389	-0.260052	-0.011237	-0.129099	-0.139651	3598
1.25287	0.075762	0.049569	0.73415	-0.423345	-0.602359	-0.532225	-0.435178	1.19046	-0.011237	-0.129099	1.129305	1789
-1.233346	-0.133383	-0.756622	-1.035505	-0.439313	-0.236408	-0.318141	-0.325032	-1.130717	-0.011237	-0.129099	-1.402859	1568
-0.451724	-0.551672	0.045656	-0.564883	1.301058	1.21002	1.511191	1.222378	-0.590341	-0.011237	-0.129099	-0.484162	416
0.528207	-0.342527	0.433097	0.40203	2.020416	1.620493	2.326506	2.033633	0.632881	-0.011237	-0.129099	0.715728	549

Promotion Sensitivity	Discount Rate on Products	Basket Diversity	Store Blindness	Purchase Volume	Cluster	Size	Population %	Sales per Cluster	Sales Contribution per Cluster
Medium	Medium	Medium	High	Medium	Mediocres	3,598	45%	33,871,992	46%
High	Low	Medium	Medium	High	Promotion-driven customers	1,789	23%	15,801,344	21%
Low	Low	Low	Medium	Low	Aliens (potential customers)	1,568	20%	13,089,421	18%
Low	High	Medium	Low	Low	Cherry-pickers	416	5%	4,588,186	6%
High	High	High	Low	Medium	Khasim and his brothers	549	7%	6,790,330	9%

Figure 6: Result summary of centers of customer clusters

There are five clusters in total that the model has developed. The following section explains these segments in detail:

Segment 1: "Mediocres"
Defining characteristic: A Medium score across all attributes

People falling in this category do not have a high/low score on any of the attributes, except for Store blindness which means they are likely to buy from multiple stores and are not so particular about which store they visit. A score of medium on all attributes means they are regular buyers who represent the 'average' shopper.

Segment 2: "Promotion Driven-Customers"
Defining Characteristic: High Promotion sensitivity and high purchase volume

These are customers who are very highly influenced by promotions. This can be said because the whole cluster has a promotion sensitivity that is 1.2 Standard deviations or 40% higher than the rest of the population. The purchase volume of these customers is also high which means these people want to maximise their purchases on promotions as they believe they get a better deal.

Segment 3: "Aliens"
Defining characteristic: Low across all attributes; bulk shopping happens in competitor stores

These are people who do not buy a lot nor do they visit the store a lot. Their basket diversity is also low which means when they do come into the store, they look for specific products. It is very likely that a bulk of their shopping happens at other stores. These people visit pernalonga only in times of necessity. These people should be the subject of our conquesting strategy.

Segment 4: "Cherry Pickers or Treasure Hunters"
Defining characteristic: High Discount rate on products

These people have a low promotion sensitivity; which means they were not driven to the store because of a promotion campaign. Nor are they after products that are on promotion. The fact that they have a low purchase volume and a medium basket diversity suggests that they are looking for specific products and spend a lot of time looking for discounts on products.

Segment 5: "Khasim and his brothers"
Defining characteristic: Same as cherry pickers but a higher basket diversity and purchase volume than cherry pickers

These customers are similar to cherry pickers. The only differentiating factor is the high basket diversity and higher purchase volume than cherry pickers. These people are driven by discounts; but they also have a need of buying across different product categories. Perhaps it is their high diversity that drives higher purchase volume.

3.2 Clustering of Products

To investigate on product groups other than category and subcategory, several attributes on three dimensions were created in order to describe product features: value driver, traffic driver and promotion undertook. These attributes were developed to measure products on each feature as follows:

Feature	Attribute
Value driver	total revenue
Traffic driver	number of transactions avg unit per transaction Q2 unit per transaction Q3 unit per transaction Q4 unit per transaction
promotion	# discounted transactions / total transactions avg discount Q2 discount Q3 discount Q4 discount

Figure 6: Features and attributes selected for customer segmentation

A data set for analysis on a product basis was created. Then the data set was analyzed with the k-means model, which is evaluated using the elbow method to find the best number of clusters. The results showed three and six as optimal clusters. Using silhouette score finds out that six clusters reach the highest score. Finally, the optimal number of six clusters are used as parameter for the k-means model.

Result of centers of the clusters are shown below:

Segment	# Products	Category of center product	Total sales	# Transactions (Traffic)	Avg units / transaction	Q2 units / transaction	Q3 units / transaction	Q4 units / transaction	% Discounted transaction	Avg discount	Q2 discount	Q3 discount	Q4 discount	Key value items	Traffic drivers	Cherry-picker's faverate	Not popular
1	8	Beer with alchohol	980.95	126	24.29	20	20	20	0.96	0.49	0.50	0.50	0.50			😉😉😉	💔💔
2	67	Oil	190577.59	98863	1.66	1	1	2	0.45	0.09	0	0	0.16	😎😎	😻😻		
3	585	Dry salt cod	2044.47	51	5.23	3.12	3.77	5.14	0.27	0.12	0	0.11	0.25				💔
4	880	Ice cream	7532.67	3045	1.09	1	1	1	0.74	0.40	0.43	0.50	0.50	😍	😻	😉😉	
5	3372	Breakfast cereals	2116.97	758	1.21	1	1	1	0.53	0.24	0	0.32	0.41			😉	
6	5858	Foreign cheese	4067.46	2760	1.20	1	1	1	0.16	0.04	0	0	0				

Segment	# Products	# Categories	Top 5 categories				
1	8	2	BEER WITH ALCHOHOL	KLEENEX			
2	67	39	FRESH PORK	FRESH POULTRY MEAT	FRESH BEEF	DRY SALT COD	FRESH FISH AQUACULTURE
3	585	47	YOGURT HEALTH	YOGURT DRINK	BEER WITH ALCOHOL	FRUIT JUICES	MINERAL WATERS
4	880	129	WASHING MACHINE DETERGENTS	HAIR CONDITIONERS	SHAMPOO	SOLAR	PERSONAL DEODORISERS
5	3372	268	FINE WINES	FINE WAFERS	SHAMPOO	WASHING MACHINE DETERGENTS	BREAKFAST CEREALS
6	5858	387	FINE WINES	DRY FOOD ANIMALS	FRESH PORK	FINE WAFERS	PROD. SPECIAL FEEDING

Figure 7: Result summary of centers of product clusters

Major columns are colored in blue, with dark representing the highest number and a lighter blue indicating a lower number respectively..

Below is the analysis on the six segments:

Segment 1: Few products and transactions but many discounts

Contains least products, with highest and most frequent discount but few transactions. This segment is cherry-picker's favorite, but it is fairly unpopular (even the cherry pickers don't like them that much!)

Segment 2: Best sellers
This segment contains 67 products from 39 categories. Products in the segment drives the highest sales and transactions. Although about half of the time they are on promotion, the discount is relatively low, with an average of 10%. Therefore, this is a segment of the key value items and traffic drivers! A good promotion strategy is to cross-sell these products with "teasers" of new-to-market products, which is a good way to attract brand and product awareness.

Segment 3: Bundle products
The main feature of this segment is that the occurrence of transactions is very low but always in bundle. It drives sales at lower level and very likely only sells to a small group of customers. If we look at the top categories, they are yogurt, juice and water, which individual customers would buy in single unit but a small portion of people would buy in bundles, i.e. long-distance drivers would buy a pack of 6 bottles of water.

Segment 4: Sales driver but only with discounts
Products in this segment is Tier 2 value drivers and traffic drivers. The major difference from segment 2 is the higher likelihood of being bought at discount and the discount is generally at a higher level, around 40-50%.

Segment 5: Question marks
This segment contains products that are tier 2 cherry picker's favorite. It is not the main sales driver, nor the traffic driver, however around half of those products are bought with promotion and the promotion is at a moderate level, around 20-30%.

Segment 6: The non-fast-movings, basic, but needed now and then
What makes this segment special is that the products included are not special compared to others. Moreover, they are not bought with promotion very often, and the discount on it is very low. Therefore, Contains mainly non-fast-moving products meeting various customer needs. Customers don't need promotions as incentives, nor would customers buy them as frequent as buying daily necessities.

3.3 Clustering of Stores

Three clusters at the store level were developed using k-means to help users understand how to make strategies towards stores. As the graph below shows, four normalized features were the main input for our k-means model: 1. Promotion, 2. Traffic, 3. Customer Loyalty, and 4. Average amount the customers spent.

Feature	Attribute
Promotion	# Item Discounted / total items sold
Customer diversity (loyalty)	Avg # Transactions in this store / total transactions for each customer Q2 # Transactions in this store / total transactions Q3 # Transactions in this store / total transactions Q4 # Transactions in this store / total transactions
Traffic	# of transaction in this store/ total number of transactions in all store
Avg $ customer spent	# discounted transactions / total transactions

Figure 8: Features and attributes for stores segmentation

As indicated by the graph below, the optimum number of clusters is three.

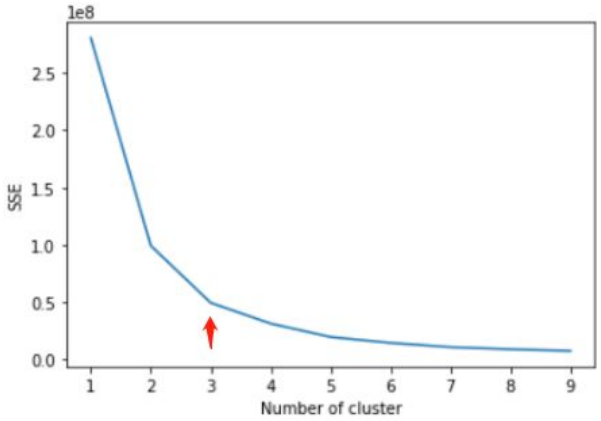


Figure 9: Result of "Elbow" method for clustering stores

Thus, running k-means clustering model with setting k to three resulted in three centroids. There are 208 for cluster number 1, 158 for cluster number 2 and 55 for cluster number 3. Using these three centroids, we found out that three stores (shown below) that are the closest to these centroids, which are the three most representative store in each cluster. These, in terms of segmentation "perfect", stores should be used for promotions and sample testing.

store_id	avg_spend	traffic	items_disc_percent	loyal_avg	Q2	Q3	Q4
643	1583.336991	0.002691	0.000959	769.345133	1.0	6.0	97.0
621	2639.120875	0.002814	0.001036	1306.350000	1.0	3.0	232.5
540	740.609359	0.002505	0.000570	345.320513	1.0	5.0	20.0

Figure 10: Result summary of centers of store clusters

As shown below, three clusters came out to be very different (at least one standard deviation away from each other) for every feature except the traffic feature. The reasons might be that the distribution (as shown below) of traffic feature is very narrow in the first place and it is hard to really distinguish stores based on traffic feature.

	avg_spend	traffic	items_disc_percent	loyal_avg
count	421.000000	421.000000	421.000000	421.000000
mean	1326.833421	0.002375	0.000746	633.023913
std	730.744615	0.000570	0.000563	355.838932
min	4.150000	0.000004	0.000001	2.000000
25%	818.671280	0.002223	0.000427	369.583815
50%	1194.953971	0.002497	0.000613	558.551724
75%	1727.802905	0.002763	0.000932	824.623116
max	3835.154815	0.002989	0.003965	1770.488889

Figure 11: Result summary of centers of store clusters, continued.

Clusters	Avg. Spend	Traffic	Discount Percent	Loyal Customers
1	Mid	Mid	Mid	Mid
2	High	High	High	High
3	Low	Low	Low	Low

Figure 12: Result summary of centers of store clusters, continued.

In addition, features were ranked and compared to each other by three levels: low, middle and high. These three clusters turned out to have the same level of features, which was a good indicator and measure to validate that the analysis conducted resulted in three distinguished groups.

4. Conclusion

Using these three clusters, the stores being visited by the same group of customers (always buying similar products) can be determined. Personalized promotion is arguably the most effective to them. Pernalonga can then start giving out more personalized promotions to product segment 4 in store cluster 2. Starting the promotion “experiment” with product segment 4 is recommend because these are the products that are the most promotion sensitive for customers.

Besides the direct effect on better promotions, this will also help the supermarket in further operational areas. For example, this will result in the reduction of inventory fluctuation and increase in the accuracy of demand planning. Furthermore, Pernalonga will improve and develop their understanding of popular items in those stores.

Note:
Clustering products, customers and stores results in a high number of individual segments, however, this approach attempts to gain a holistic picture. As the paper lines out every segmentation helped gaining understanding and recommending different strategies for this specific case. Knowing which customers are mainly receptive to promotions in combination to also having clusters of stores, helps to introduce specific promotions for every store. In addition, this can be taken one step further, having (at least partially) customized stores for the prevalent customer segment visiting that store.