MKT 680 Marketing Analytics Project 1

Exploratory Data Analysis

Team 6: Joyce Zhang, Nicole Santolalla, Justin Wood

2/13/2021

Executive Summary

To analyze Pernaloga's purchase transactions we used K-means clustering technique to segment products, customers, and stores. We included recency, frequency, and promotion percentage to segment our customers and label the variables as cherry pickers, normal and most valuable customers. From these methods, we were able to pull the necessary insights needed to answer our question of who the best customer is by segmentation. We found that our most valuable customers "loyal" are ones who on average have transactions that are less than 20% on promotion. Our cherry pickers are ones who on average have transactions that are more than 45% on promotion. We created this threshold to represent who the valuable customers and cherry pickers are.

To extract insights from the dataset and our analysis, we took a high level approach to the case and conducted market research. To understand the dataset, we needed to find any patterns or trends in the Portuguese market that would support our findings. Our assumption is that Pernagoga is in Portugal. According to a consumer report from Santander Trade, the Portuguese tend to buy whatever is cheapest and are "addicted to deals". This insight is what we explored further in our dataset and helped us in segmenting our cherry-pickers from valuable customers. "Good prices and sales" were other factors that influence purchasing behavior that this report mentioned. Another report came from consumergoodsforum.com, which included a study of 500 families across Portugal age range of 25-55. This report mentioned that "price is the number-one influencing factors on where they choose to shop" and "easy to find promotions" rank highly on which locations to choose. These initial insights provided us with better questions to answer on how we can apply our segmentation process in the dataset.

Statement of Assumptions

Our team started by examining the dataset to get a grasp on how many unique identifiers we had per column and a description of the dataset, which represents 2 years of purchase history. The following are our observations and how we approached correcting some anomalies that we found:

Transactions: Each row represents a single product type bought, meaning that there could be numerous rows for a single transaction and a single customer. We proceeded to identify that the minimum transaction amount for the whole dataset was negative. Our assumption is that these transactions represent returns.

Transaction Id: The transaction id didn't represent a unique identifier for the actual transaction, but instead, it was an elongated version of the transaction date. Moreover, there were some transaction ids that were incorrect. We proceed to create a unique identifier ('id') that would represent a single transaction per customer in each separate date. We assumed that a single customer will only go once a day to the grocery store.

Promotions: We created 2 new columns based on promotions. The first one is "total promo" that represents the absolute value of the total amount discounted per transaction. The second one is "promo proportion average" that represents the offer counts divided by the sales quantity, resulting in how many/much of the products bought were in promotion.

Profit: Since we didn't have the cost per item, we were not able to calculate the profit for each transaction. We decided to assume that the average profit per item in a grocery store is between 10% and 12%. We multiplied the transaction product paid amount by 12% and created a new column for profit.

Products and Categories

Products

We started by investigating the products and categories in all the transactions. There are a total of **10,767** unique products and **429** unique categories. We realized that the product units are distinguished by "count" and "kg", 75.53% of the products are in unit "count", and 24.47% of the products are in unit "kg". To prevent future confusion on the quantity sold and discount count, we decided to separate the transactions with these two types of product units. In addition, we created **revenue**, **volume**, **number of promotions**, **number of transactions**, and **average promotion proportion** as metrics to further perform data analysis on products.

First, we looked at the products with unit "count". Based on revenue, product "99951863" which belongs to the "FRESH UHT MILK" category generates the highest revenue, as shown in *figure 1* with the top 5 products.

prod_id	category_desc_eng	promo_count	volume	trans_frequency	revenue	promo_prop_avg
999951863	**FRESH UHT MILK**	66996	656033.0	149533	290502.57	0.160783
999345410	**OLIVE OIL**	26470	83213.0	67387	254840.49	0.350096
999512554	**STRAWBERRY**	46343	114216.0	91054	202374.04	0.440980
999958970	**FRESH UHT MILK**	43492	333713.0	75236	191127.24	0.225081
999421692	**OIL**	49506	163866.0	98863	190577.59	0.378616

Figure 1. Product (CT) sort by revenue descending order

As we performed similar groups for the products with unit "count", we found that product "999231999" which belongs to the "BAGS" category occurs in almost every transaction. However, with further research, we found out it is common practice for stores in Europe to charge for bags on each customer's transaction. Hence, we decided to exclude the bag category for further analysis since it did not add additional insights about the transactions.

After excluding bags from our analysis, the key value item “99951863” (“FRESH UHT MILK”) continued to be the product that generates the most revenue. Another finding is that the products driving the most revenue also have relatively high transaction frequency and number of promotions, for example, milk, oil, and strawberry products.

Then we looked at the products with the unit “kg”. These products typically will be priced based on weight, like vegetables, fruits, and meats. We followed the same grouping procedures as before and came up with the best products based on revenue, as shown below in *figure 2.* The products that make the most revenue are the meat products. However, the key value item that occurs in almost every transaction is the product “999956795”, which belongs to the category “banana”. Also, the traffic driver products are mainly fruits and vegetables, only the fresh pork products tend to be both the top traffic driver and revenue generator.

prod_id	category_desc_eng	promo_count	volume	trans_frequency	revenue	promo_prop_avg
999749469	**FRESH BEEF**	6251	98925.265	99996	602109.43	0.082805
999956795	**BANANA**	179372	566374.282	491580	546554.81	0.371202
999749894	**FRESH PORK**	9241	189019.014	132278	530179.25	0.071205
999455829	**FRESH POULTRY MEAT**	6717	122780.181	112245	482758.42	0.066901
999649801	**DRY SALT COD**	17449	58416.509	18474	416159.93	0.364968

Figure 2. Product (KG) sort by revenue descending order

Categories

After understanding the products in the transactions, we investigated higher level categories. There are **439** unique categories, and **1,476** unique subcategories. We did not differentiate the product units when analyzing the categories, because we wanted to compare the categories on the same level, thus we used the same metrics when exploring the products except volume. As shown in figure 3, products in the meat category generate the highest revenue. This finding aligns with the culture in Portugal that pork is the most consumed meat.

Categories like “green beans”, “chestnut”, “pomegranate”, “Christmas perfumery” have the highest promotion proportion average. This means that in these categories, a large proportion of the items are always on sale. The key value category that occurs in most transactions is “fine wafers", a very popular snack in Europe.

category_id	category_desc_eng	promo_count	total_promo	trans_frequency	revenue	promo_prop_avg
95890	**FRESH PORK**	54984	41694.53	632046	2.483963e+06	0.092641
95894	**FRESH BEEF**	43351	66386.99	408244	2.401663e+06	0.135631
95888	**FRESH POULTRY MEAT**	51467	64468.24	597559	2.379408e+06	0.070600
95971	**DRY SALT COD**	51627	458735.67	58819	1.344207e+06	0.353539
95797	**FINE WINES**	190138	575060.83	333855	1.296609e+06	0.435643

Figure 3. Categories sort by revenue descending order

Product Segmentation

With a basic understanding of the best products and categories in the transactions, we performed K-means clustering to further segment the products. We segmented the products based on three important metrics: **number of transactions**, **average promotion proportion** and **revenue**. The reason why we chose the average promotion proportion instead of the volume was due to different products units. We expected that this procedure would tell us which product clusters were seldom sold on discount, the “must buy” items, etc. After completing the transforming procedures in K-means clustering, we identified 3 clusters, with summary statistics of each cluster, as shown below in figure 4.

Cluster	trans_freq mean	avg_promo_prop mean	revenue mean	count
0	89595.5	0.3	180990.8	69
1	1568.2	0.7	4138.0	4469
2	2637.1	0.2	5013.4	6229

Figure 4. Product segmentation with 3 clusters from K-means clustering

The table shows that the products in cluster 1 are bought very frequently, and generate the most revenue, even though only around 30% are on sale. We can categorize this cluster as “must buy” products that the customers tend to buy the products under any conditions. Examples of products in cluster 0 will be fruits, napkins, etc. The products in cluster 1 follow the pattern that the majority are on sale, but customers do not buy them very often, thus not much revenue is made. Beyond cluster 0 and 1, cluster 2 represents products that are discounted the least, generate moderate revenue and account for moderate number of transactions, we may classify the products as premium brands among cluster 1.

Customers

Customer Segmentation Based on Wine Product

The dataset contains **7,920** unique customer identifiers. To segment our customers, we identified some characteristics of interest that are relevant to our goal of targeting customers with more personalized advertisement. These characteristics include how frequently customers go to the stores (**frequency**), how many products they buy on promotion (**promotion percentage**), how recently they made a purchase (**recency**). By analyzing and then segmenting customers with these characteristics we were able to create clusters that help identify most valuable customers and learn more about their purchasing behaviors.

We decided to use our clustering approach with customers that had purchased products in certain categories. These procedures can also be applied to other interested categories. In our example we used Fine Wines (*According to Wine Intelligence, Portuguese wine drinkers spend per bottle by occasion has increased since 2017*). We chose wine because the "FINE WINE" from above category analysis turned out to be the #5 revenue driver. We continued our analysis by clustering customers in the "wine" category to see what attributes they had in common regarding recency, frequency, and average promotion proportion. We performed K-means clustering and identified 3 clusters with corresponding summary statistics, as shown in *figure 5*. Also, we graphed each cluster based on the three attributes to represent the commonality in these wine customers in *figure 6*.

	Recency	Frequency	PromotionPercentage	
	mean	mean	mean	count
Cluster				
0	4.6	1.9	0.3	2417
1	1.9	4.3	0.5	3817
2	4.3	1.9	0.8	1192

Figure 5. Summary statistics of customer clusters on "FINE WINE" category

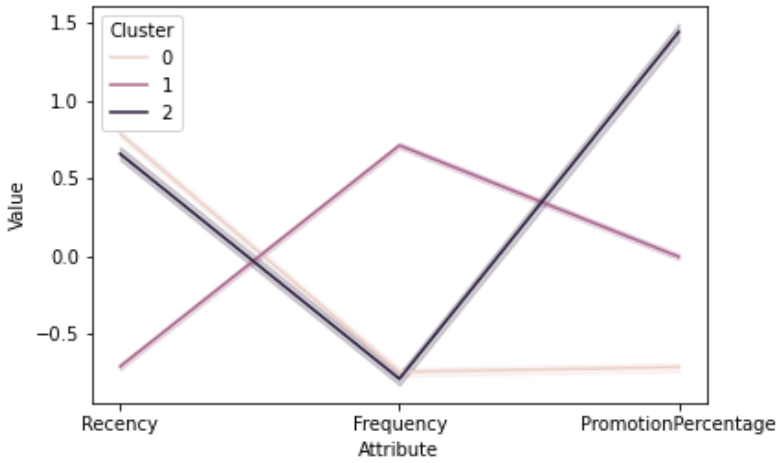

Figure 6. Snake plot of customer clusters on "FINE WINE" category

From the above plots, we can characterize the three clusters. Cluster 0 customers bought wine not long ago, did not buy wine frequently, and did not buy the wine based on if it was discounted or not. Cluster 0 could represent potential customers. Cluster 1 customers bought wine a while ago, they bought wine frequently, and bought them on promotions periodically. Cluster 1 could be our loyal wine customers. Cluster 2 customers bought wine recently, they did not buy frequently, and were very sensitive to promotions. This group of customers tend to only buy wine when they are on sale. They are cherry pickers of the wine products. If we are going to customize wine promotion, cluster 2 could be our potential targets.

With this analytical procedure we are able to recreate the clustering technique for each category of interest, and not only target our customers with personalized advertisement, but also build assumptions on why it is that they visit our stores and buy certain categories.

Customer Segmentation on All Products

Moreover, we wanted to learn customers' behavior across all products in general to identify the groups of customers who always buy promotions (cherry-pickers) and who buy everything at any price (most-valuable). We implemented the average promotion percentage metric to segment the customers. This measures among all transitions for the customer that the average discounted rate the customer received. If the average rate is under a certain threshold, we can identify this customer as "loyal", otherwise as "cherry picker". We found that the average promotion percentage among the customers was 31%, with a minimum 5% and maximum 87%. We calculated the 5 percentile and 95 percentiles of the average as thresholds. If the average promotion percentage was below 20%, we classified the customers as "most valuable customers"; if the average promotion percentage was above 45%, we classified the customers as "cherry pickers"; and the customers in between will be labeled as "normal".

Store Segmentation

Lastly we explored store groupings to find any patterns related to stores and our customers. There are **421** unique stores in the dataset. First, we grouped the stores by **volume**, **revenue**, **frequency**, **number of unique products**, and **average promotion proportion**. Then we performed K-means clustering to segment the stores based on the customers categories we defined above (cherry picker, most valuable, normal).

Next, we ran some summary statistics for grocery stores. Store "342" had the highest revenues of $786,520, volume of 670,290, transaction frequency of 363,011 and number of unique products of 9,521, as shown in figure 7.

store_id	volume	revenue_by_store	profit	tran_frequency	number_unique_product	promo_avg
342	670290.662	786520.95	94382.5140	363011	9521	0.302046
345	518803.903	718779.77	86253.5724	299550	9519	0.345247
349	559421.824	680640.41	81676.8492	306796	9273	0.373387
344	491200.375	624991.67	74999.0004	274151	9431	0.363338
343	458829.587	591942.18	71033.0616	244279	9056	0.370381

Figure 7. Stores sort by revenue descending order

An assumption can be made that this store is in a heavily populated area. Lisbon is the capital of Portugal and the largest populated city in Portugal. It would make sense that this store would be located here. It could also explain why this store has the greatest number of unique products, as many international tourists may fly or travel into Lisbon and use this store to find what they need.

We found that the top 5 stores (342, 349, 345, 344, 347) by revenue were also ranked top in other metrics. According to our prior assumption of store "342" located in a large city like Lisbon, all 5 of these stores may be located near each other in the same city. One inference

is that each of these top store numbers start with a “3”, which could reflect that all these stores are relatively close to each other.

Store “469” is the best store based on an average promotion of 70%. This store also has “165” unique products, which is very low. It could be assumed that the reason that the average promotion is so high is that this store is advertised as a mostly discounted store. With the limitation of items, this store may be in areas of Portugal where essential items are needed and where possibly cherry pickers live.

After conducting initial insights and creating certain assumptions on different stores, we wanted to find out which stores our cherry pickers, normal, and most valuable customers visited the most. These insights can shed light on to which store or groupings of stores we can focus our personalized promotions on. Store “349” has the highest frequency of visits by cherry pickers, shown in *figure 8*.

store_id	num_cherry_picker	num_normal	num_most_valuable
349	43818.0	256617.0	6361.0
344	36544.0	232496.0	5111.0
331	33373.0	115834.0	3694.0
343	33118.0	206633.0	4528.0
347	27412.0	231344.0	11573.0

Figure 8. Best stores most frequently visited by cherry pickers

To further our research, we then clustered the stores into a K-means clustering based on the number of customers in each category. We identified three clusters with summary statistics as shown in *figure 9*.

	num_cherry_picker	num_most_valuable	num_normal	
	mean	mean	mean	count
store_cluster				
0	2283.6	2138.0	43928.2	293
1	4262.7	9211.4	91839.9	114
2	24137.6	10005.2	211914.1	14

Figure 9. Summary statistics of store clusters on customer category

Cluster "0" has the lowest cherry pickers, valuable customers and normal customers. This cluster is the group of stores that most likely does not offer promotions, have limited items, and are not in a city/populated area. This cluster has the greatest number of stores "293". Cluster "1" and "2' are better and carry the customers we want. Cluster 2 is the best as it has the highest number of most valuable customers in the smallest number of stores "14". These stores are most likely located in the city or heavy populated areas. This category also has the greatest number of cherry pickers as well, nearly **6x** from cluster "1" and **12x** from cluster "0". This is neither good or bad, but these clusters of stores might have items on promotion at average 30% of items or higher and most likely have more unique items available.

Sources

Santander Trade

https://santandertrade.com/en/portal/analyse-markets/portugal/reaching-the-consumers

The Consumer Goods Forum

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

Wine Intelligence

https://www.wineintelligence.com/and-suddenly-portugal-was-trendy/