MKT 680 - Marketing Analytics

Project 2: Recommender System

Group 4: Carl Xi, Frank Fan, Jie Zhu, Yaping Zhang

Background

As a brief refresher, Pernalonga is an undisputed leader in the Lunitunia retail space with over 10,000 products in 400+ categories. They are currently trying to derive actionable marketing insights from their transaction and product data and have asked us for help. Their strategic direction is primarily focused on the customer, product and store categories. Here is the same excerpt from our previous project that speaks about Pernalonga's partnership and sales strategies:

" Pernalonga regularly partners with suppliers to fund promotions and derives about 30% of its sales on promotions. While a majority of its promotion activities are in-store promotions, it recently started partnering with select suppliers to experiment on personalized promotions. In theory, personalized promotions are more efficient as offers are only made to targeted individuals who require an offer to purchase a product. In contrast, most in-store promotions make temporary price reductions on a product available to all customers whether or not a customer needs the incentive to purchase the product. The efficiency of personalized promotion comes from an additional analysis required on customer transaction data to determine which customers are most likely to purchase a product to be offered in order to maximize the opportunity for incremental sales and profits."

Problem Definition and Scope

In our previous project, we conducted segmentation analysis on Pernalonga's customers, products and stores. In this project, we are now moving onto developing targeted and personalized promotions for the specific segments we outlined. When considering the three categories (products, stores, and customers), customer-level personalized promotions have always proved most effective, as they are the only ones with the freedom to move between product categories and store locations.

In short, our main objective in this project is to develop customized promotional campaigns for the different customer segments using the insights from our previous project and new insights outlined in the project below. The brand and product we are focusing on are Nesle's chocolate products and we are trying to increase their market share in the confectionery segment. With our understanding of Pernalonga's customers and Nesle's confectionery products, we are fully confident in the marketing campaign we have highlighted below.

Data Understanding and Exploratory Data Analysis

We will be using the same dataset as last time., which is split into two tables, with 10767 observations and 7 variables covering the transaction data and 29,617,585 observations and 12 variables covering the product data. We derive customer and store insight from the transaction table by matching the customer and store ID variables with the products in the product table. Some discrepancies were found between the IDs and descriptions of categories. Upon closer inspection, we found that some category IDs share the same description, which we decided to keep, as our dataset is large enough and that this may capture unique complementary goods effects. Our preliminary EDA helped us find the outliers, NAs, and bad data points within the data. We decided to not include summary statistics in this report as we have already outlined them in our previous report and that they are not helpful in our customer-level discriminative discounting strategy below. With that said, we would like to outline a few key numbers for your consideration below.

There are a total of 7848 customers who purchase chocolate items, with 5809 of them being existing Nesle customers. We have 279 different types of chocolate products and 31 types of Nestle products, with all of them being chocolate products luckily. We can see that Nestle already has a majority share in this sector, but we would like to further drive wallet share and customer loyalty with customized discount strategies.

Defining the Customer Groups

While our previous report segmented the customers into clusters based on all of their purchases, we wanted to focus on creating segments based on their purchasing patterns related to chocolate. From examining the business objective behind the promotional campaign, we think the 5 customer groups below make sense:

1. Cherry-Pickers: Customers who really like to buy on discount
2. Chocolate add-ons: Customers who buy chocolate as a complimentary item (e.g. those who love grabbing a chocolate bar at checkout)
3. Potential Customers: Customers who don't buy Nestle items but will probably like them.
4. Fans of competitors: Customers who love Nestle's competitor's chocolate products
5. Nestle Fans: Customers who love Nestle's products

By targeting these five customer groups and customizing different pricing strategies for each of them, we will maximize the benefits of our customized pricing strategy. With that said, we will have to conduct an individualized business analysis, data preparation, modeling and insight extraction.

Customer Data Insights

Being a readily available and relatively inexpensive treat, chocolate is a common confectionary. Common sense tells us that people should at most purchase chocolate once per day. Indeed, this is the case, with the biggest chocolate lover buying chocolate 505 times in the two years of our data. Interestingly, only 7 of those 505 purchases were Nestle, and the biggest Nestle fan who bought chocolate 141 times in our data's timeframe bought nestle 106 times. On average, customers who buy chocolate purchase 20 times per year. The same demographic purchase Nestle 3 times on average per year. This is a lot of room for improvement for both the chocolate category and Nestle!

We also noticed customers who purchase a significant amount of chocolate (1140 chocolate items in 2 years, that's 2 chocolate items consumed per day). We think these customers are likely restaurants/dessert shops who use chocolate for their own products, and/or public places. In short, not for personal use, as that would be very concerning. Alternatively, these customers could be wholesalers, who bulk purchase chocolate and later resell them in smaller packages.

To identify whether these customers are truly wholesalers or use them for commercial purposes, we further examined their purchase behavior across other products and stores. We first looked at their purchase frequency (these customers should buy more with fewer store visits), which unfortunately did not support our hypothesis, as these customers visited our stores around the same number of times, and sometimes even more than normal customers. We then looked at their purchase patterns of other items, which shows that they bought everything more, but not items that you would typically expect from wholesalers (e.g. paper).

With mixed signals, we think these customers could be a mix of wholesalers, commercial purchasers and just really fanatic chocolate lovers. It is hard to put a single hat that defines the entire group, and we would love to analyze them in the future.

Customer Groups In-Depth Analysis

Nestle Fans:

This group of customers is people who have Nestle as one of their most purchased chocolate brands. We define Nestle Fans as customers who spend a higher percentage of their "chocolate money" on Nestle chocolates than average chocolate buyers do.

On average, customers who bought chocolates in the past spent 10.67% of their chocolate budget on Nestle products, while the median is around 6.36%. It also appears that the top quartile of customers spent more than 15% of their chocolate budget on Nestle products. Statistically, it makes sense to view the top quartile as our Nestle Fans, because they spent more heavily on Nestle chocolates than

the rest 75% of chocolate buyers. However, one in every seven dollars is not enough for us to call them “Nestle Fans” in real life, given the fact that there is still plenty of money spent on the competitors. Thus, we decide to be conservative here and use 25% as our bar for Nestle Fans. From the histogram below, it also confirms that only a very small portion of the chocolate buyers spent more than 25% on Nestle products.

Percentage of Total Chocolate Budget Spent on Nestle Products					
Min	**25th Percentile**	**Median**	**Mean**	**75th Percentile**	**Max**
0.0000	0.0000	0.0636	0.1067	0.1495	1.0000

Figure 1. Percentage of Total Chocolate Budget Spent on Nestle Products - Summary Statistics

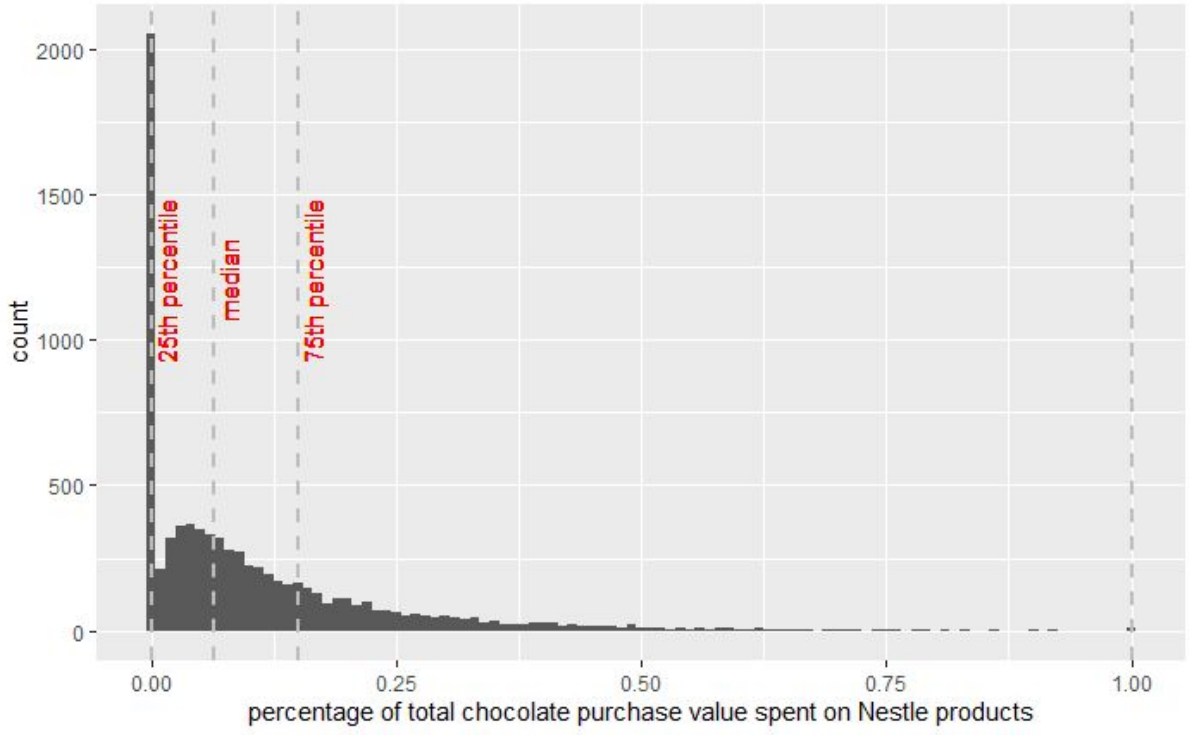

Figure 2. Percentage of Total Chocolate Budget Spent on Nestle Products - Histogram

Total Chocolate Purchase in Dollars					
Min	**25th Percentile**	**Median**	**Mean**	**75th Percentile**	**Max**
0.45	28.34	54.03	74.16	96.00	1138.68

Figure 3. Total Chocolate Purchase in Dollars - Summary Statistics

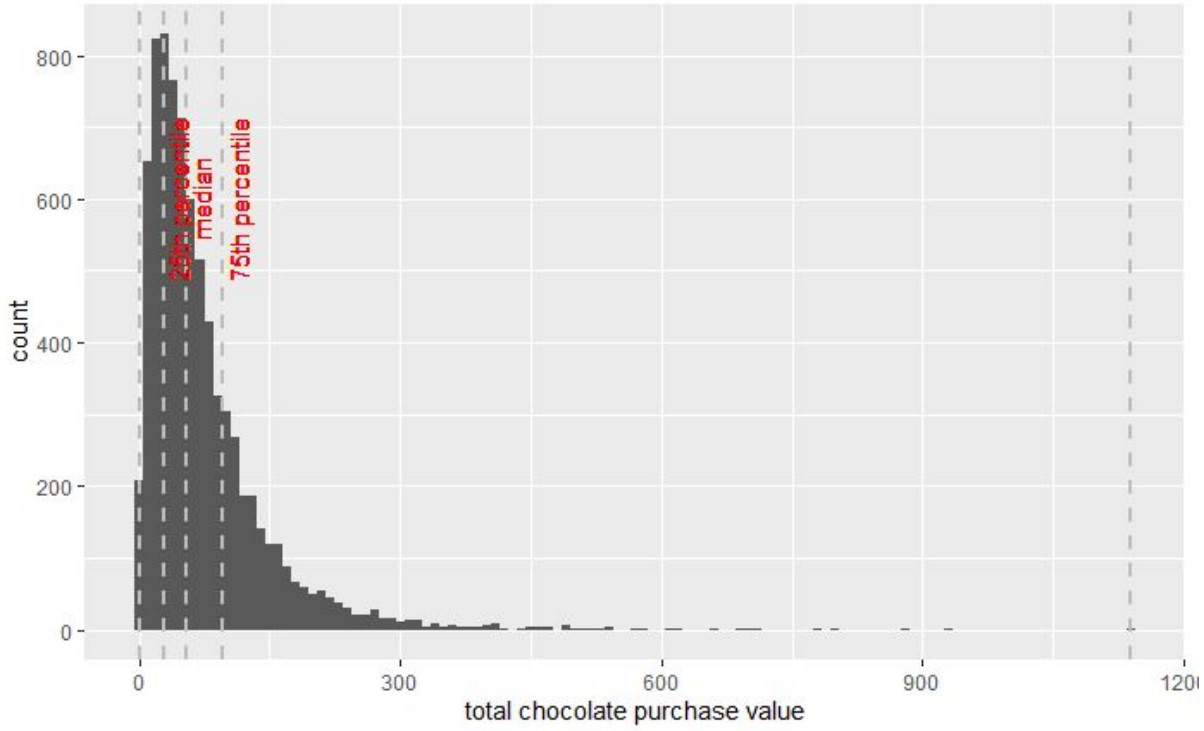

Figure 4. Total Chocolate Purchase in Dollars - Histogram

Given the fact that this group of customers already spent a big portion on Nestle chocolates, we decided to target Nestle Fans who didn't buy as much as the median chocolate buyer in dollar amount. This is because while they obviously prefer Nestle chocolates, they are probably purchasing from other sellers. After excluding Private Label buyers, there are 204 such customers. A discount here can attract these Nestle Fans who are not buying primarily from Pernalonga.

Fans of Competitors:

On the opposite side of Nestle fans is the fans of competitors. These customers spent only a very small portion of their chocolate budget on Nestle. Instead of targeting the subgroup with less spending, we decide to target those with more chocolate spending among fans of competitors. Since they are already spending more than the median, they are more likely to be attracted by the promotion and decide to give Nestle products another try. In particular, the fans of competitors that we are looking for spent less than 6% on Nestle and purchased more than $54 chocolates in the past. After excluding Private Label buyers, there are 230 such customers.

Cherry-Pickers:

In the previous project, we defined cherry pickers as the type of customers who tend to buy discounted products. And to define them, we can use the average discount percentage of a customer 's

purchase and the percentage of purchases with at least one discount as measure a customer ’s “cherry picker” level.

For this project, we need to find out cherry-pickers who are interested in chocolate. Therefore, we have decided to include the percentage of transactions with chocolate products and the percentage of transaction value contained in chocolate in all transactions for each customer. Figure 5 below shows the summary statistics for the cherry picker customers.

cust_id	p_o_discount_purchase	avg_discount_rate	cl_amt_ratio	cl_value_ratio
Min. : 29568	Min. :0.0000	Min. :0.0000	Min. :0.000000	Min. :0.000000
1st Qu.:25009812	1st Qu.:0.2239	1st Qu.:0.2737	1st Qu.:0.000000	1st Qu.:0.000000
Median :50389851	Median :0.3010	Median :0.3666	Median :0.007967	Median :0.007727
Mean :50261305	Mean :0.3101	Mean :0.3689	Mean :0.023285	Mean :0.023345
3rd Qu.:75739898	3rd Qu.:0.3851	3rd Qu.:0.4599	3rd Qu.:0.028163	3rd Qu.:0.026986
Max. :99999776	Max. :1.0000	Max. :1.0000	Max. :1.000000	Max. :1.000000

Figure 5. Cherry-Pickers - Summary Statistics

To run a segmentation, we decide to use the k-means clustering method to find the clusters.

To find out the best K, we use Gap Analysis and use the optimal k detected to run the k-means model, which is displayed in figure 6. We can see that our optimal k is 3.

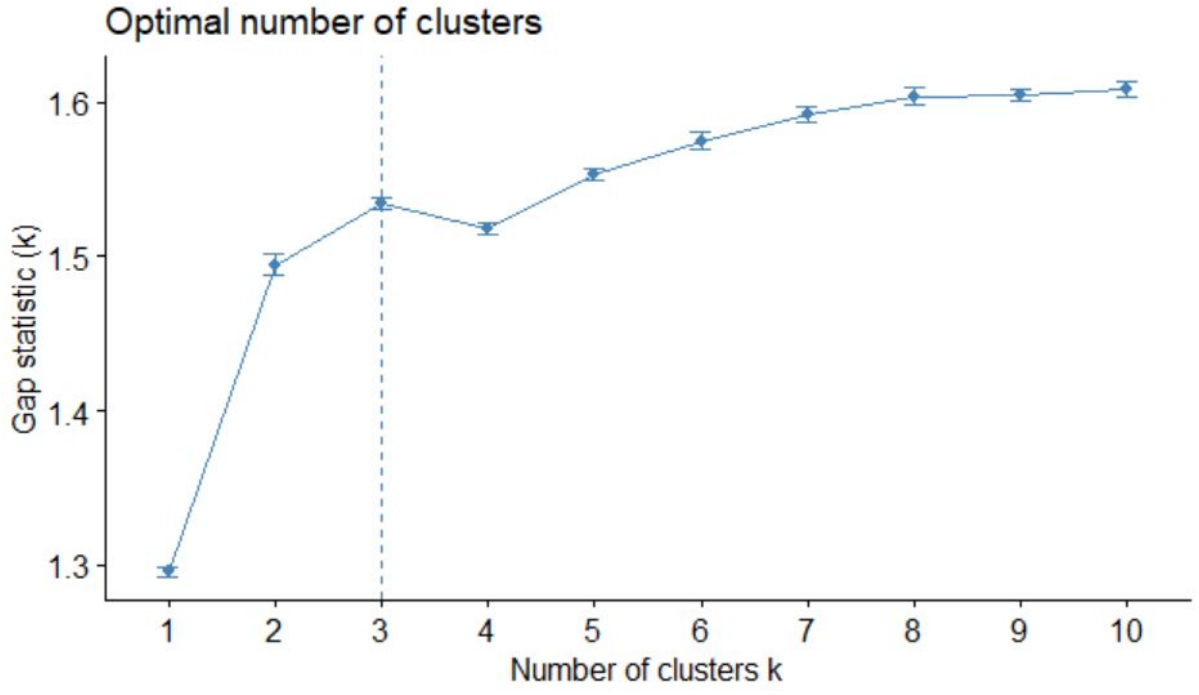

Figure 6. K-Means Clustering Method

After merging the GAP results, we run the k-means model and generate three different groups. Table 1 below clearly shows three groups with their own attribute values. As you can see in the table, the three groups are clearly distinguished. Cherry pickers love discounts, either the quantity level or the value level. Our target customers also love to buy chocolate. Cluster 2 has the largest tp_amt_ratio

and tp_value_ratio, while p_o_discount_purchase and avg_discount_rate are relatively high. This indicates that Cluster 2 is a cherry picker because it has a high purchase and purchase value of chocolate. Therefore, we define category 2 as cherry and chocolate pickers, which constitutes our target customers. There are a total of 1921 selective clients selected.

	p_o_discount_purchase	avg_discount_rate	cl_amt_ratio	cl_value_ratio
#1	0.1881072	0.2206440	0.02122495	0.02169495
#2	0.4647039	0.5399674	0.02320734	0.02324895
#3	0.3129384	0.3814931	0.02481333	0.02458643

Figure 7. K-Means Segmentation - Summary Statistics

Chocolate Add-Ons:

In order to find people who purchase Nestle chocolates as "add on" items, we seek to find purchase pairs with Nestle chocolates. Accomplishing this first requires selecting tran_id, the new transaction id described above, as well as prod_id. Using these two features we find unique pairs. Once we have the unique pairs, we run them through an Association Rules model. Our findings will be demonstrated in the next "Promotion Strategies" section.

Potential Customers:

In order to identify prospective customers – those who haven't purchased Nestle chocolates before but have a high potential of purchasing them, my team performed collaborative filtering. First of all, since we are not allowed to target customers who purchase PRIVATE LABEL chocolate products, we filtered them out.

Then our next step was to build a table containing three columns: customer ID, product ID, which matches the customer's shopping history correspondingly, and the amount of money a customer has paid for the product. We used the amount paid by a customer for a product as the implicit rating for collaborative filtering, since it took both frequency(F) and monetary(M) into consideration.

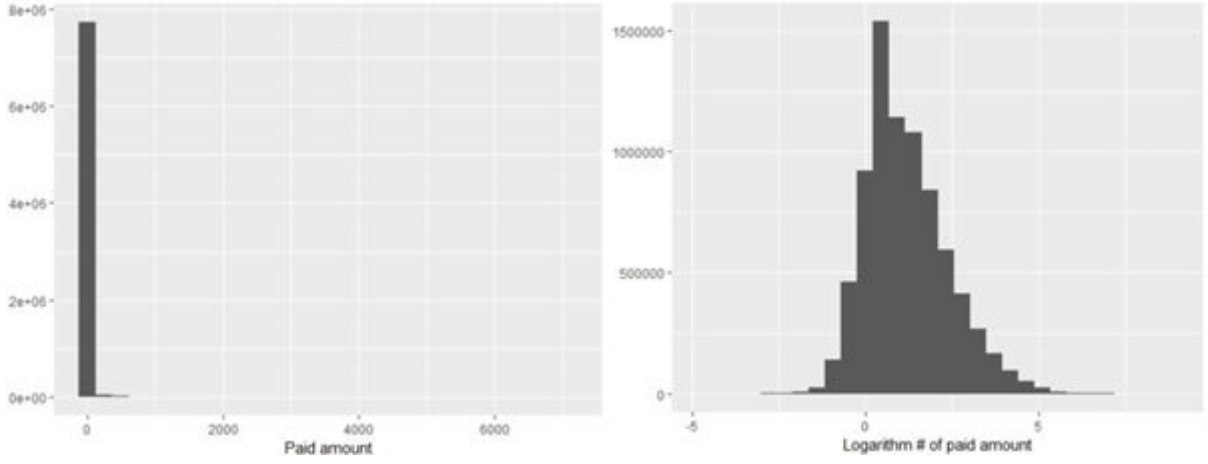

Figure 8. Paid Amount - Log of Paid Amount

Then the next step was to decide what should be categorized as a "good rating". We used ggplot2 to examine the histogram of paid amount and found out that this value (which we plan to use as rating) is very right-skewed. So, we performed the following steps:

- Use logarithm to transform the paid amount column
- Use min-max standardization to make the range of paid amount fall in the range of 0 and 1.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.3714  0.4206  0.4315  0.4836  1.0000
```

As the above table shows, the third quartile of our "rating" is 0.4836. We then used this number as our cut-off value in the item based collaborative filtering model,

After that, we build an affiliation matrix with product ID as rows, customer ID as columns and the standardized and transformed paid amount as implicit rating. This matrix is then fed into collaborative filtering models. Two customers with a similar paid amount on certain products suggests that they might have similar preferences in other products. Thus, we will be able to find those who didn't buy Nestle chocolate products, but are most likely to buy it in the future, by using collaborative filtering techniques.

We build 3 item-based collaborative filtering models based on various similarity distances, including Jaccard, Cosine and Pearson. The outcome of the 3 models did not differentiate a lot, but we find that the Pearson model was the best one in terms of RMSE and MSE.

	RMSE	**MSE**	**MAE**
IBCF-Jaccard	0.058168	0.003383	0.043183
IBCF-Cosine	0.058109	0.003377	0.043145
IBCF-Pearson	0.058096	0.003375	0.043150

We then use the Pearson model to find which customers would be most likely to buy which chocolate products from Nestle. We are able to construct a table of 3 columns with customer ID, Nestle product ID and the corresponding predicted ratings. It is worth mentioning that the ratings from active customers are not included in this table, only the predicted ratings of customers who had never bought the product are recorded. By using different cut-off values

for ratings, Pernalonga is able to choose an optimal number of customers, as the target, to promote Nestle chocolate products.

Promotional Strategies:

Nestle Fans:

Currently, Nestle chocolates have an average discount rate (total discount amount / total sale amount) of 32.50%, which ranks 13th among all 48 chocolate brands. Nestle also has the second largest customer base with 5809 customers each spent an average of $11.09. The only brand with a larger customer base is Pernalonga's Private Label with 6048 customers and average spending of $13.43. In other words, Nestle products already have very good market penetration and are heavily discounted. More discounts won't help much to attract these Nestle fans who are not buying a lot from Pernalonga. Instead, we should focus more on how to get them into stores. For example, offering coupons on volume-driven products like Coca-Cola will cause Nestle lovers to come to Pernalonga stores more often and thus purchase more Nestle chocolates. An alternative strategy will be placing Nestle chocolates with complements like marshmallow, so that Nestle lovers will see more times of Nestle chocolates every time they visit the store, increasing the probability of purchase.

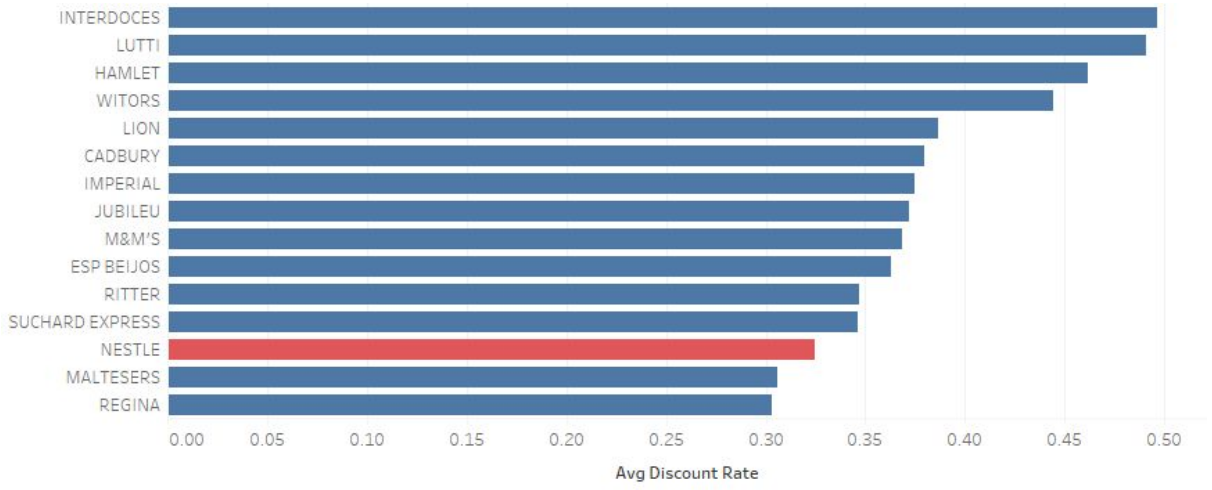

Figure 9. Top 15 Brands in Average Discount Rate

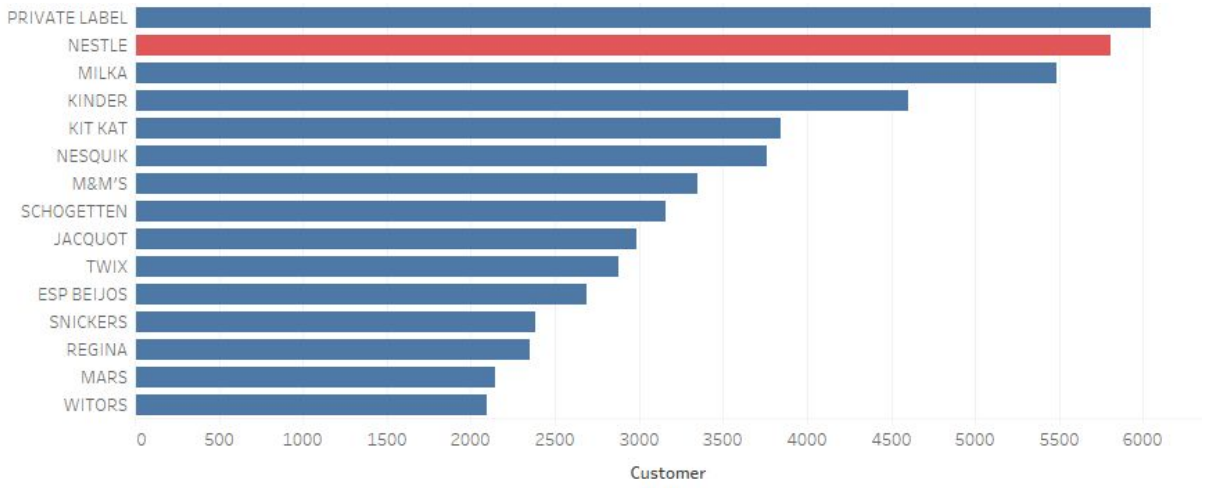

Figure 10. Top 15 Brands in Customer Base

Fans of Competitors:

Why didn't these fans of competitors buy Nestle chocolates? It could be that they hate the taste of Nestle chocolates or they have never tried Nestle products. Among the 230 targeted fans of competitors, 66 of them have never bought one single Nestle chocolate product. Although Nestle has good market penetration, there are still some customers who have never tried Nestle chocolates. Thus, for these fans of competitors, a good strategy is to offer free chocolate samples at Pernalonga stores. Some of them will hopefully start to purchase Nestle chocolates after trying them and finally become Nestle Fans.

Cherry-Pickers:

There is only one sub-category in Nestle chocolate: Saborisantes. A clear descriptive statistics table for this type is shown below. If we want to target cherry-pickers for Nestle chocolate, we need to consider what is the average discount rate to offer.

discounted_tran_cnt	total_revenue	total_transact	total_distinct_customer	total_stores	avg_discount_rate
Min. :46	Min. :2829	Min. :1243	Min. :915.0	Min. :315.0	Min. :0.1958
1st Qu.:57	1st Qu.:3073	1st Qu.:1297	1st Qu.:923.8	1st Qu.:318.5	1st Qu.:0.2288
Median :68	Median :3316	Median :1350	Median :932.5	Median :322.0	Median :0.2619
Mean :68	Mean :3316	Mean :1350	Mean :932.5	Mean :322.0	Mean :0.2619
3rd Qu.:79	3rd Qu.:3559	3rd Qu.:1404	3rd Qu.:941.2	3rd Qu.:325.5	3rd Qu.:0.2949
Max. :90	Max. :3803	Max. :1458	Max. :950.0	Max. :329.0	Max. :0.3280

Therefore, we compare Nestle's descriptive statistics with other brands:

COLA CAO:

discounted_tran_cnt	total_revenue	total_transact	total_distinct_customer	total_stores	avg_discount_rate
Min. :49.00	Min. :2642	Min. : 930	Min. :614.0	Min. :284.0	Min. :0.007756
1st Qu.:56.25	1st Qu.:2707	1st Qu.: 984	1st Qu.:615.8	1st Qu.:287.8	1st Qu.:0.099410
Median :63.50	Median :2772	Median :1038	Median :617.5	Median :291.5	Median :0.191064
Mean :63.50	Mean :2772	Mean :1038	Mean :617.5	Mean :291.5	Mean :0.191064
3rd Qu.:70.75	3rd Qu.:2838	3rd Qu.:1092	3rd Qu.:619.2	3rd Qu.:295.2	3rd Qu.:0.282718
Max. :78.00	Max. :2903	Max. :1146	Max. :621.0	Max. :299.0	Max. :0.374372

NESQUIK:

discounted_tran_cnt	total_revenue	total_transact	total_distinct_customer	total_stores	avg_discount_rate
Min. : 14.0	Min. : 1001	Min. : 238	Min. : 200	Min. :142.0	Min. :0.04804
1st Qu.: 43.0	1st Qu.: 1027	1st Qu.: 253	1st Qu.: 218	1st Qu.:168.0	1st Qu.:0.12693
Median : 98.0	Median : 2053	Median : 850	Median : 677	Median :247.0	Median :0.17246
Mean :117.8	Mean : 7855	Mean : 2920	Mean :1127	Mean :262.2	Mean :0.15382
3rd Qu.:193.0	3rd Qu.:11231	3rd Qu.: 3018	3rd Qu.:1253	3rd Qu.:343.0	3rd Qu.:0.20946
Max. :241.0	Max. :23963	Max. :10239	Max. :3287	Max. :411.0	Max. :0.21220

PRIVATE LABEL:

discounted_tran_cnt	total_revenue	total_transact	total_distinct_customer	total_stores	avg_discount_rate
Min. :105	Min. :9730	Min. :3866	Min. :1198	Min. :385	Min. :0.01522
1st Qu.:105	1st Qu.:9730	1st Qu.:3866	1st Qu.:1198	1st Qu.:385	1st Qu.:0.01522
Median :105	Median :9730	Median :3866	Median :1198	Median :385	Median :0.01522
Mean :105	Mean :9730	Mean :3866	Mean :1198	Mean :385	Mean :0.01522
3rd Qu.:105	3rd Qu.:9730	3rd Qu.:3866	3rd Qu.:1198	3rd Qu.:385	3rd Qu.:0.01522
Max. :105	Max. :9730	Max. :3866	Max. :1198	Max. :385	Max. :0.01522

SUCHARD EXPRESS:

discounted_tran_cnt	total_revenue	total_transact	total_distinct_customer	total_stores	avg_discount_rate
Min. :95	Min. :5400	Min. :1856	Min. :795	Min. :322	Min. :0.346
1st Qu.:95	1st Qu.:5400	1st Qu.:1856	1st Qu.:795	1st Qu.:322	1st Qu.:0.346
Median :95	Median :5400	Median :1856	Median :795	Median :322	Median :0.346
Mean :95	Mean :5400	Mean :1856	Mean :795	Mean :322	Mean :0.346
3rd Qu.:95	3rd Qu.:5400	3rd Qu.:1856	3rd Qu.:795	3rd Qu.:322	3rd Qu.:0.346
Max. :95	Max. :5400	Max. :1856	Max. :795	Max. :322	Max. :0.346

And we can see that the average discount rate of Nestle is really high, only lower than Suchard Express. But the average discounted transaction count for Nestle is really low. The second lowest in all 4 categories. Therefore, we recommend uplift the discount rate for Nestle in promotion to evoke cherry pickers' purchasing intention.

Chocolate Add-Ons:

Pulling the names of the products from the lhs and rhs pairs mentioned above, we are able to see which items are purchased before Nestle chocolate products are added on. Yogurts weigh very heavily in the results, which indicate that Nestle chocolate are often purchased together with yogurt products. So a powerful campaign might be product bundling of yogurt and Nestle chocolate products.

Initial Purchases	Add on products	Subcategory
145519008	Nestle Chocolates	Yogurt Health
145519009	Nestle Chocolates	Yogurt Health
145519010	Nestle Chocolates	Yogurt Health
145519011	Nestle Chocolates	Yogurt Speacialties
145519012	Nestle Chocolates	Yogurt Speacialties

Potential Customers:

As discussed earlier, by using different cut-off values of ratings, Pernalonga is able to choose an optimal number of customers, as the target, for promoting Nestle chocolate products. Based on our understanding of the situation and our investigation of the data, for the predicted ratings, a cut-off value of 0.5 gave us 15322 records and 7899 unique customers. That being said, some customers will have high ratings for several Nestle products. The number of customers to target seems like a rational number to us.

Nestle Chocolate Product ID	Number of Target Customers	Subcategory
999398484	7498	Seasonal Christmas Chocolate
999609655	7824	Seasonal Christmas Chocolate

After examining the product ID in the 15322 observations, as well as the corresponding subcategories, we find that there are only 2 products that have high ratings from customers. They are all from Seasonal Christmas Chocolate subcategory. With this information in mind, we are able to conduct precision marketing for customers at the product level. We can offer coupons on receipts and push ads through digital channels of these products to our potential customers. Such precision marketing can save plenty of marketing spend and generate more revenue.

Next Step: Hierarchical Promotion for Different Customer Groups

In the end, we have successfully identified four different groups of customers. It is obvious that some customer groups will overlap with others. We have used the Venn diagram to demonstrate it:

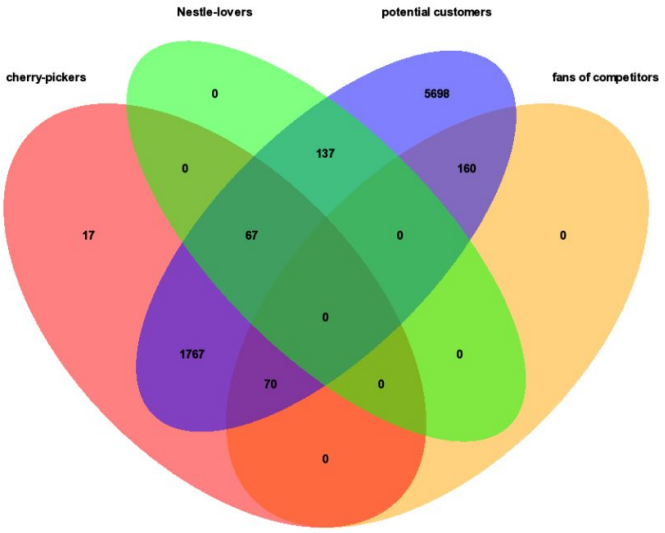

Figure 11. Venn Diagram

From the Venn diagram, there are a total of 7916 customers that should be targeted in our campaign. We figured for the groups of customers that fall into the overlaps, we can come up with specific discount levels. For example, Nestle lovers already have a tendency to purchase Nestle products, a smaller discount is likely to trigger purchase. Another example will be the overlapped 70 customers of "fans of competitors", "cherry-pickers" and "potential customers". They like promotion, and they are likely to switch from Nestle competitors to Nestle, hence we need to provide them with a greater level of discounts. Detailed hierarchical promotion plan is shown below:

Customer group	Discount Level
Nestle Fans	1
Potential-customers + Nestle Fans	2
Potential customers	2
Potential customers + Fans of competitors	3
Cherry-pickers	4
Cherry-pickers + Potential Customers	5
Cherry-pickers+Potential Customers + Fans of competitors	6

With data-driven promotion of six different discount levels to seven target customer groups, Nestle will be able to save costs and secure a high ROI.