MKT 680 – Marketing Analytics
Recommender Systems Project

Group 1: Abinav Bharadwaj, Evan Kang, Meng Wang, Kami Wu

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. Specifically, Colgate-Palmolive is interested in a promotional campaign to boost the sales of Colgate toothpaste.

EXECUTION SUMMARY

We divide target customers into 5 different groups, then find more than three thousands specific target customers, and find a way to boost Colgate toothpaste sales by rearranging product placement. We estimate the final outcome of the campaign would be more than five thousand in extra sales and an increase in revenue by 10% for Colgate toothpaste.

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. We notice that the original transaction id is not unique, because one unique id can represent multiple customer ids or store ids. In order to solve that, we extract the first eight digits of transaction id, then paste it with customer id and store id. In such a way, we are able to ensure each transaction can be uniquely identified. After analyzing the business problems our client presented, we hereby decide to keep using the three datasets we defined in the previous project, focusing on different perspectives: customer, products and stores. The customer and product table have been largely used for this project. 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).

```
> 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 detailed information regarding descriptive statistics of the 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

CUSTOMER GROUPS DEFINITION

For this specific project, in order to boost the sales of Colgate toothpaste, we need to find types of customers who are likely to prefer Colgate toothpaste, then target each type of customer with a personalized promotion plans. After taking the business context behind Pernalonga and the nature of toothpaste into consideration, we finally settle down with the following 5 types of customer groups, which we believe should be the target customers of this campaign:

a) Colgate-lovers: People who love Colgate toothpaste.
b) Fans of competitors: People who buy toothpaste but not Colgate.

c) Cherry-pickers: People who love campaigns.
d) Toothpaste secondaries: People who buy Colgate toothpaste as an add-on.
e) Potential customers: People who didn't buy Colgate products but will probably like Colgate toothpaste.

We then tackle each type of customers differently, in terms of business understanding, data preparation, modeling construction and finally actionable promotion insights.

CUSTOMER FINDING

As is common sense, people usually use toothpaste two times a day and, on average, consume one tube of toothpaste every 2 – 3 months. In this project, given two years of transaction data, we can predict people normally consume 8-12 tubes of toothpaste during this period. In addition, some of them are likely to be family customers, causing the consumption of toothpaste to be multiplied two or three times. As a result, we conclude that people regularly buy 8- 36 tubes of toothpaste in a two year period.

However, what the data tells us is quite different from what we predict. For the total 7,920 registered customers, 945 (approximately 12%) customers purchased more than 36 tubes of toothpaste in this period. Moreover, 346 customers purchased more than 48 tubes. One customer even purchased 224 tubes, meaning he was buying toothpaste every 3 days!

Based on our assumption, two types of customers may consume toothpaste at an extreme volume. One of them are suppliers: they buy toothpaste to sell to other smaller retailers. They may also be small hotels, which buy toothpaste to meet the needs of their guests. We simply call this group, "company customers."

In order to incorporate the difference in demographics of toothpaste purchasers and exclude the effect it has on our predictions, we first want to determine whether this group of people are truly company customers. Under deep consideration, we find that company customers share some common features. First, for suppliers and hotels, toothpaste occupies a large portion of their purchases. In other words, their purchasing product categories will be fewer than "personal customers". Second, they will make fewer store visits than other customers and purchase as many tubes as possible at one time.

After we looked at the data again, and did some summary statistics, we can tell this group of customers are not real "company customers". First, we find that when they buy more toothpaste, they tend to buy more of other products in other categories. Second, this group of people actually paid slightly more visits than other customers who did not buy a large volume of toothpaste.

Moreover, the counts of tubes of toothpaste they bought each time are not significantly higher than those of other customers.

To conclude, based on rejection of our assumption, although 945 customers bought more than 36 tubes of toothpaste in the two years, they essentially share the same demographics as others customers and are unlikely to be suppliers or small hotels.

Colgate-lovers: People who love Colgate toothpaste
We want to look at people whose Colgate purchases make up most of their total toothpaste purchases. In particular, our target customers are people who have not bought much toothpaste in the two year period, but a high percentage of their toothpaste purchases are Colgate.

We first look at the general distribution of total toothpaste purchases during the period. The first quarter of people did not buy more than 8 tubes in the 2 years, which correspond with our prediction. We can conclude that these people may have other sources from which to purchase toothpaste. Thus, they are the customer to whom we want to promote to buy more. Second, we look at the ratio of Colgate to total toothpaste consumption.

```
> summary(customer_buy$col_ratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.2500  0.5000  0.5036  0.7600  1.0000 
```

Figure 3: The Summary Statistics for the ratio of Colgate to total toothpaste consumption

The result shows a uniform distribution of the "loyalty" ratio. Therefore, we can say the 4th quarter of people are loyal to Colgate brand. We combine these two filters and finally find 442 customers belong to this group: people who do not purchase toothpaste very often but are loyal to Colgate when they do.

Fans of competitors: People who buy toothpaste but not Colgate
Next, we want to target people who are loyal customers of other brands, meaning people who bought tubes of other brands more than those of Colgate. In particular, we would like to look at customers who bought toothpaste regularly or more than regularly (including the 945 customers who bought a lot), but only a small portion of them are Colgate.

We use the same statistics as "Colgate lovers" and decide that people who bought more than 8 tubes in the two year period are regular toothpaste customers. Combing the filter of Colgate's ratio, less than 25% (the first quarter of customer), we find 1464 people are "fans of competitors."

Cherry-Pickers: People who love campaigns

During the previous project, we defined cherry pickers as the type of the customer who tends to purchase discounted products. Therefore, we believe the average percentage of discount of customers' purchases and the percentage of purchases that include at least one discount are good measures to infer a customer's "cherry-picker" level. For this specific promotional plan, we need to target the cherry-pickers who are interested in toothpaste. Thus, we decide to include the percentage of transactions including toothpaste product, and the percentage of transactional value including toothpaste among all transactions for each customer. The following figure 4 displays the summary statistics for the generated customer table.

```
> summary(cpcs)
    cust_id           p_o_discount_purchase avg_discount_rate  tp_amt_ratio      tp_value_ratio       cluster_3
 Min.   :   29568   Min.   :0.0000        Min.   :0.0000    Min.   :0.00000   Min.   :0.00000   Min.   :1.00
 1st Qu.:25009812   1st Qu.:0.2239        1st Qu.:0.2737    1st Qu.:0.05762   1st Qu.:0.04341   1st Qu.:1.00
 Median :50389851   Median :0.3010        Median :0.3666    Median :0.11358   Median :0.08658   Median :2.00
 Mean   :50261305   Mean   :0.3101        Mean   :0.3689    Mean   :0.13744   Mean   :0.10617   Mean   :2.07
 3rd Qu.:75739898   3rd Qu.:0.3851        3rd Qu.:0.4599    3rd Qu.:0.19129   3rd Qu.:0.14681   3rd Qu.:3.00
 Max.   :99999776   Max.   :1.0000        Max.   :1.0000    Max.   :1.00000   Max.   :1.00000   Max.   :3.00
```

Figure 4: The Summary Statistics for the Defined Customer Table

We decide to use the k-means clustering method to find the clusters. Before approaching the clustering modeling process, we use Silhouette Analysis to find the optimal k to run the k-means model, which is displayed in figure 5. It is clearly showing that 3 is the optimal k to begin with.

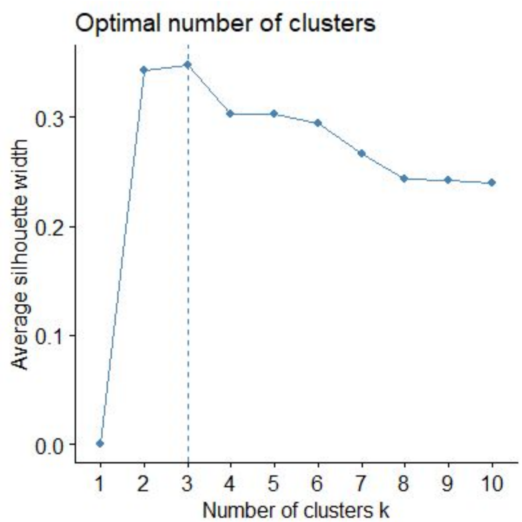

Figure 5: The Visual Result of the Silhouette Analysis

After incorporating Silhouette result, we run the k-means model and generate three different clusters. The following table 1 clearly shows the three clusters with its own attribute values. From the table, we can tell that all three cluster have been clearly differentiated. Also, the customers of each cluster are relatively homogeneous within each group. Cherry-pickers like discounts, both in terms of amount-level and value-level. Our target customers also need to love to purchase toothpaste. Cluster 2 has the highest tp_amt_ratio, and tp_value_ratio, and a relatively high p_o_discount_purchase and avg_discount_rate. Even though cluster 1 has the highest discount purchases, we still decide to define cluster 2 as the cherry-pickers as it has a considerably high level of both toothpaste purchase amount and value. Therefore, we define cluster 2 as the cherry-pickers for toothpaste, which make up our target customers. There are totally 1496 target cherry-picker customers.

Cluster	p_o_discount_purchase	avg_discount_rate	tp_amt_ratio	tp_value_ratio	
1	0.4174304	0.4931449	0.10839023	0.08491742	
2	0.3069617	0.3699543	0.29998016	0.22478482	Cherry Picker
3	0.2211934	0.2638520	0.09218143	0.07319522	

Table 1: The Clustering Result in terms of Cherry Picker and Toothpaste

Toothpaste secondaries: People who buy Colgate toothpaste as an add-on

```
       lhs            rhs          support     confidence  lift     count
[1]  {999956795} => {999302196} 0.002023913 0.005818791 1.545371 2056
[2]  {999956795} => {999222814} 0.002001272 0.005753698 1.461595 2033
[3]  {999956795} => {999180064} 0.001877238 0.005397099 1.407255 1907
[4]  {999956795} => {999223643} 0.002142040 0.006158409 1.524749 2176
[5]  {999956795} => {999433934} 0.002241464 0.006444255 1.517483 2277
[6]  {999956795} => {999177170} 0.002575173 0.007403676 1.444967 2616
[7]  {999361204} => {999180321} 0.001830972 0.009166038 1.549311 1860
[8]  {999956795} => {999180321} 0.003115605 0.008957429 1.514050 3165
[9]  {999361204} => {999331572} 0.002144993 0.010738063 1.512521 2179
[10] {999956795} => {999331572} 0.003502472 0.010069678 1.418375 3558
```

Table 2: The Association Rule Model results: products purchased with Colgate toothpaste

In order to find people who purchase Colgate toothpaste as an "add on" item, we seek to find purchase pairs with Colgate toothpaste. 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. The table above shows the results of the Association Rules model. The right hand side (rhs) shows product ids for Colgate toothpaste and the left hand side (lhs) displays product ids for various other products. These results provide us with insight into which products customers first buy, then purchasing Colgate toothpaste as an "add on" item.

Potential customers: People who didn't buy Colgate products but will probably like Colgate toothpaste

To find potential customers - those who didn't buy Colgate products but will probably like Colgate toothpaste, we decide to use collaborative filtering.

We built a table with 3 columns - customer ID, product ID, which a customer has bought correspondingly, as well as the amount a customer has paid for it. We use the amount paid by a customer for a product as the implicit rating because it takes both purchase frequency and purchase quantity into consideration.

After examining the frequency of the paid amount, we find the value was extremely right-skewed. We use logarithm to transform the value and make it more normally distributed.

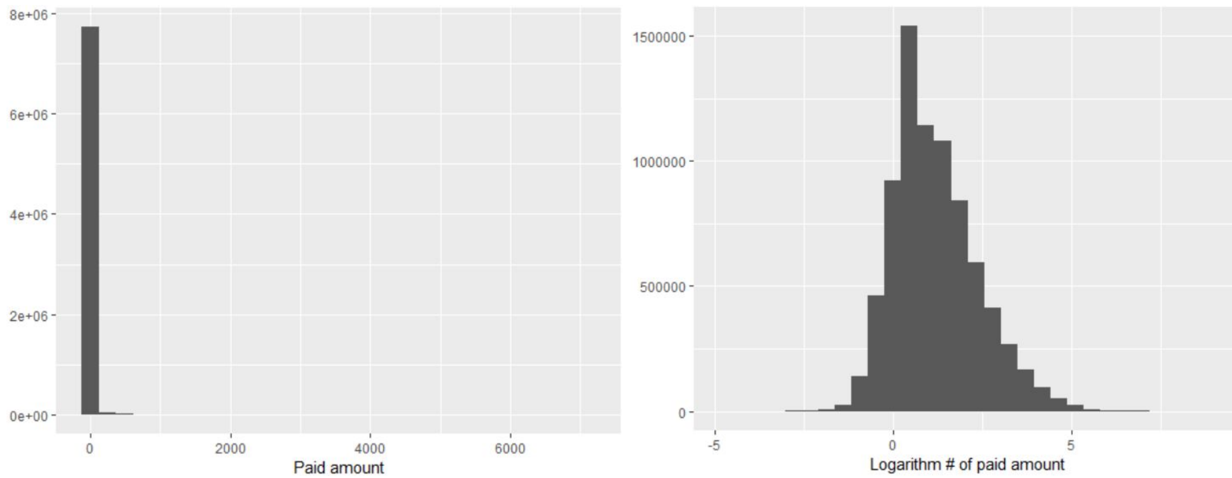

Figure 6: Distribution of paid amount before and after transformation

Then, we use min-max standardization to make the range of the rating fall between 0 and 1. We examine the basic statistics of the rating, and find that the value of the third quartile was 0.4836. This number will be used as the cut-off value of "good rating" in our models.

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

Figure 7: Basic statistics of paid amount

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 similar a 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 Colgate toothpaste, 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	0.05816879	0.003383608	0.04318315
IBCF.cosine	0.05810988	0.003376758	0.04314522
IBCF.pearson	0.05809613	0.003375161	0.04315047

Figure 8: Error evaluation matrix of different models

We then use the Pearson model to find which customers would be most likely to buy which Colgate toothpaste. We are able to construct a table of 3 columns with customer ID, Colgate 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 Colgate toothpaste.

PROMOTION DESIGN

Colgate-lovers: People who love Colgate toothpaste

```
> summary(customer_col_lover$avg_disc_rate_col)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.08375 0.22083 0.21964 0.33728 0.57776 
> summary(customer_buy$avg_disc_rate_col)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.1178  0.2469  0.2265  0.3341  1.0000 
```

Figure 9: Summary statistics for Colgate-lovers

Although Colgate-lovers have slightly lower discount rate than all buyers, the statistics do not show a significant difference. That means, this group of 442 people who purchase a fewer amount of toothpaste are not more likely to be tempted by price markdown or other kinds of pricing promotions.

But not surprisingly, we find that these customers paid significantly fewer visits than the ordinary customers. A proper guess is that these people visit *Pernalonga* less because they have other better choices when shopping. In order to capture the most value out of them, in other words, in order to make them buy more toothpaste at their limited visits to *Pernalonga*, we would recommend the “add-on item” strategy. We would try to place Colgate toothpaste together with some of the most frequently bought products on the shelf. As they are loyal to Colgate, when seeing a Colgate toothpaste, their instinct will push them to think do they need a tube toothpaste now or in the near future. The other promotion plans include letting *Pernalonga* attract this group of 442 customers to pay more visits by promoting other items. We will not dig deeper in this direction because this is out of Colgate’s control.

Fans of competitors: People who buy toothpaste but not Colgate

```
> summary(customer_other_lover$avg_disc_rate_col)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.1736  0.1877  0.3350  0.6003 
> summary(customer_buy$avg_disc_rate_col)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.1178  0.2469  0.2265  0.3341  1.0000 
```

Figure 10: Summary statistics for Fans of Competitors

To the fans of competitors, the statistics gives us a more clear perspective. Compared with the median of 24.69% and mean of 22.65%, other brands' loyal buyers have a significant lower discount rate regarding Colgate toothpaste purchasing (median of 17.36%, mean of 18.77%). That could explain why those customers prefer other brands: they seldom receive a good promotion with Colgate.

Therefore, for this group of customers, giving a higher discount rate more frequently is a feasible approach. Although discounts will eat some of our profits, considering the large amount these customers purchase toothpaste, our total revenue will likely increased.

Cherry-Pickers: People who love campaigns

There are four sub-categories in Colgate toothpaste: brandq, tradic, infant, medici. A clear descriptive statistics table for each type of Colgate is shown in table 3. If we want to target cherry-pickers for Colgate toothpaste, the sub-brand of Colgate with the highest discount rate and value is our priority choice. From the table, we see that both brandq and tradic have a relatively high discount rate and discount value. Tradic has a much more broad coverage than brandq, as tradic has twice as many transactions, customers, revenues, transactions, and stores. Therefore, we promote tradic and then promote both tradic and branq as supplements, depending on the store's transaction situation. We also coordinate with the store manager to adjust the promotion plan, depending on the number of transactions for each sub-category, to decide whether to promote both brandq and tradic, only tradic, or only brandq.

Promoted Products

Colgate Types	% of Colgate	Discounted transt CNT	Total revenue	Total transact	Total distinct customs	Total stores	Avg discount rate
BRANDQ	11.26	**96.29**	2880.1	974.6	657.3	197.9	**0.3006**
TRADIC	85.63	**147.3**	5824.3	2241.2	1206	288.3	**0.24750**
INFANT	19.6	65	877.2	396.7	282.3	181.3	0.15819
MEDICI	1.15	93	2964	689	412	230	0.176

Table 3: Descriptive statistics of 4 Sub-category Colgate Toothpaste

Toothpaste secondaries: People who buy toothpaste as an add-on

Initial Customer Purchase	Add On
Bananas	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (BRANQ)
Bananas	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (TRADIC)
Carrot	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (TRADIC)
Carrot	Colgate Toothpaste (TRADIC)
Bananas	Colgate Toothpaste (TRADIC)

Table 4: Item initially purchased before Colgate Toothpaste is added on

Pulling the names of the products from the lhs and rhs pairs mentioned above, we are able to see which items are purchased before Colgate toothpaste is added on. Bananas weigh very heavily in this table. As shown, there are a few pairs of carrot purchases with Colgate toothpaste, as well, but the primary good are bananas. Additionally, the subcategory of the Colgate toothpaste is almost unanimously tradic. Therefore, we recommend that a small number of our “star” products, Colgate toothpaste with sub-category tradic, be placed next to the fruits and vegetables, encouraging more joint purchases of these two items.

Potential customers: People who will probably like Colgate toothpaste

As was mentioned earlier, by using different cut-off values of ratings, Pernalonga is able to choose an optimal number of customers, as the target, for promoting Colgate toothpaste. Based on our understanding of the situation and our investigation of the data, for the predicted ratings, a cut-off value of 0.465 gave us 231 observations and 204 unique customers. That being said, some customers will have high ratings for several Colgate products. The number of customers to target seems rational number to us.

After examining the product ID in the 231 observations, as well as the corresponding subcategories, we find that there are only 4 products have high ratings from customers. They are from the branq, medici and tradic subcategories.

Colgate Types	Product ID	# of customers to target
BRANQ	999152257	114
MEDICI	999344671	95
TRADIC	999164127	11
TRADIC	999331572	11
		Total: 231

Table 5: Product recommendation summary

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 a single Colgate toothpaste tube to a customer. Such precision marketing can save plenty of marketing spend and will have a higher ROI.

EXPECTED OUTCOME

At last, we have 4 different groups of customers. Obviously, there are some overlaps between different groups, as shown in the Venn diagram below.

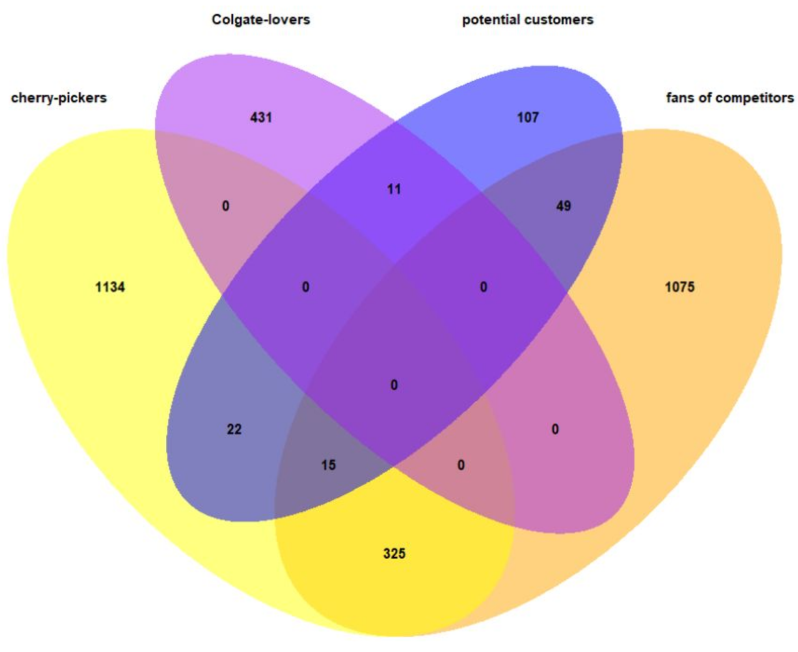

Figure 11: Venn diagram of different groups

Based on the Venn diagram, the total number of specific customers we will target is 3,169. For the groups of customers that fall in the overlaps, we will set different levels of discount rate. For example, for Colgate-lovers, since they already love Colgate toothpaste very much, a small discount will likely be able to trigger their willingness to purchase. Thus, the discount level for them should be the lowest. However, for the customers in the overlap of cherry-pickers and fans of competitors, because they love discounts and they are fans of other brands, we need to set a higher discount than competitors in order to attract them to buy Colgate toothpaste. The detail discount level design is shown below.

Customer Groups	Discount Level
Colgate-lovers	1
Potential customers	2
Cherry-pickers + fans of competitors + potential customers	3
Cherry-pickers + potential customers	3
Cherry-pickers	4
Fans of competitions + potential customers	4
Fans of competitions	5
Cherry-pickers + fans of competitors	6

Table 6: Discount level for different groups of customers

Altogether, with discounts to those 3,169 customers, and rearranging the store placement of Colgate toothpaste, we are confident to say this campaign will strongly trigger the sales of Colgate toothpaste. Approximately, this campaign will trigger more than 5,000 extra transactions of Colgate toothpaste, and boost the revenue of Colgate toothpaste by 10%.