**MKT 680 - Marketing Analytics**

**Project 1: Segmentation**

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

**Background**

With over 10,000 products in 400+ categories, Pernalonga is an undisputed leader in the Lunitunia retail space. Our project aims to help Pernalonga conduct segmentation analysis on their customers, products and stores for marketing analytics purposes. Below is an excerpt 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**

As we can see from the excerpt above, Pernalonga has long conducted product-wide promotions for all their stores and all their customers. We are here to develop a marketing campaign that utilizes segmentation to create personalized promotions for Pernalonga’s customers. Before we jump into the analytics and actual segmentation, we would like to demonstrate our high level of understanding of the initial data given to us by conducting thorough exploratory data analysis below. With our understanding of Pernalonga’s stores, products and customers, we are fully confident in delivering a successful marketing campaign for Pernalonga.

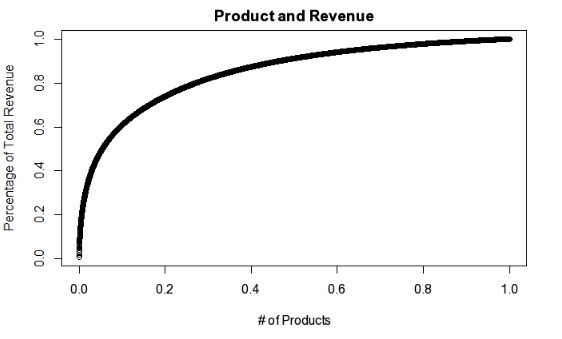
**Data Understanding and Exploratory Data Analysis**

The data that our client has given us is split into two table2, with 10767 observations and 7 variables covering the transactions data and 2,961,785 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.

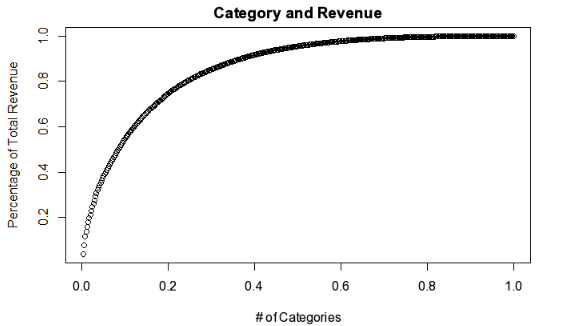
**1) Product**

We began with exploratory data analysis to determine the best products and product groups in terms of volumes, revenues, number of transactions and customers. While we presented our detailed results in R notebooks, below are some highlighted findings:

1. **Overall Trends:** There are 10770 distinct products in Pernalonga. However, the top 20% of products have brought in almost 75% of total revenue of Pernalonga’s 421 stores. The Pareto principle (also known as the 80/20 rule) states that roughly 80% of the effects come from 20% of the causes, which is once again proved here. Pernalonga should spend the majority of time and money on the small percentage of products that bring the most revenue.



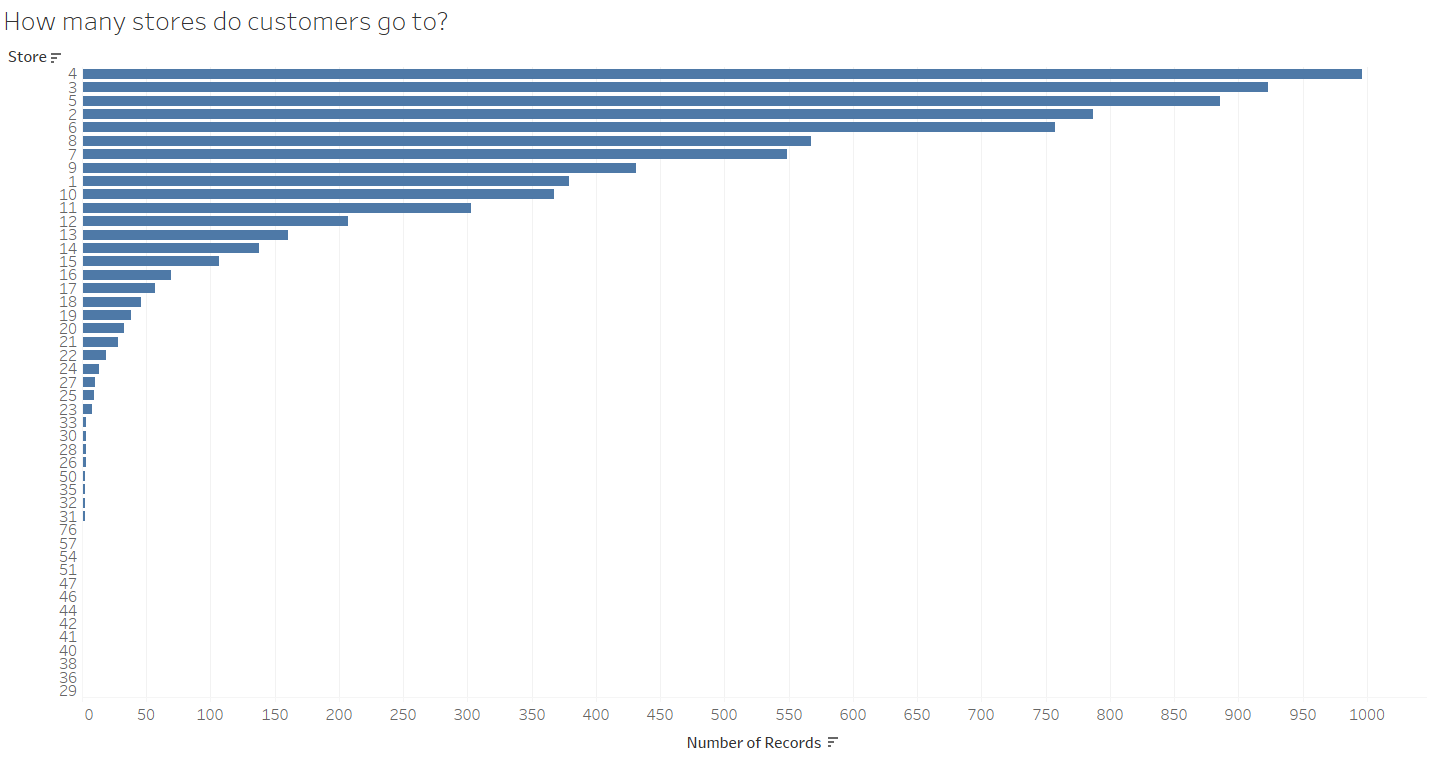
1. **Traffic-driving products:** By aggregating the count of transactions using product\_id, we identified the products that bring the most transactions. The top five categories of traffic-drivers are: Bags, Banana, Mineral Waters, Cheese Type Flamengo and Pao Manufacture. The top five products are: 999231999, 999956795, 999361204, 999951863 and 999746519. We observed that most of the products here are high-demand life necessities, which customers need to buy them frequently.
2. **Customer-magnet products:** By aggregating the count of distinct customers using product\_id, we identified the products that have been purchased by lots of customers. The result is different from the above traffic-drivers, because even though many customers have bought this type of product, they don’t necessarily need to buy them frequently. For example, the top one customer-magnet is rice, where most customers need but they don’t buy on a frequent basis. The top five customer-magnet categories are: Rice, Bags, Fine Wafers and Mineral Waters. The top five customer-magnet products are: 999231999, 999956795, 999361204, 999512554 and 999712725.
3. **Revenue-driving products:** As stated above, the top 20% products generate almost 75% of total revenue. If we look at this problem at a category-level perspective, we are going to get a similar result. There are 419 categories of products in Pernalonga, and the top 36 categories composes 50% of total revenue; the top 25% of categories composes top 80% of total revenue. The top five revenue-driving categories are Fresh Pork, Fresh Beef, Fresh Poultry Meat, Dry Salt Cold and Fine Wines. The top five revenue-driving products are: 999749469, 999956795, 999749894, 999455829 and 999649801.



**2) Customer**

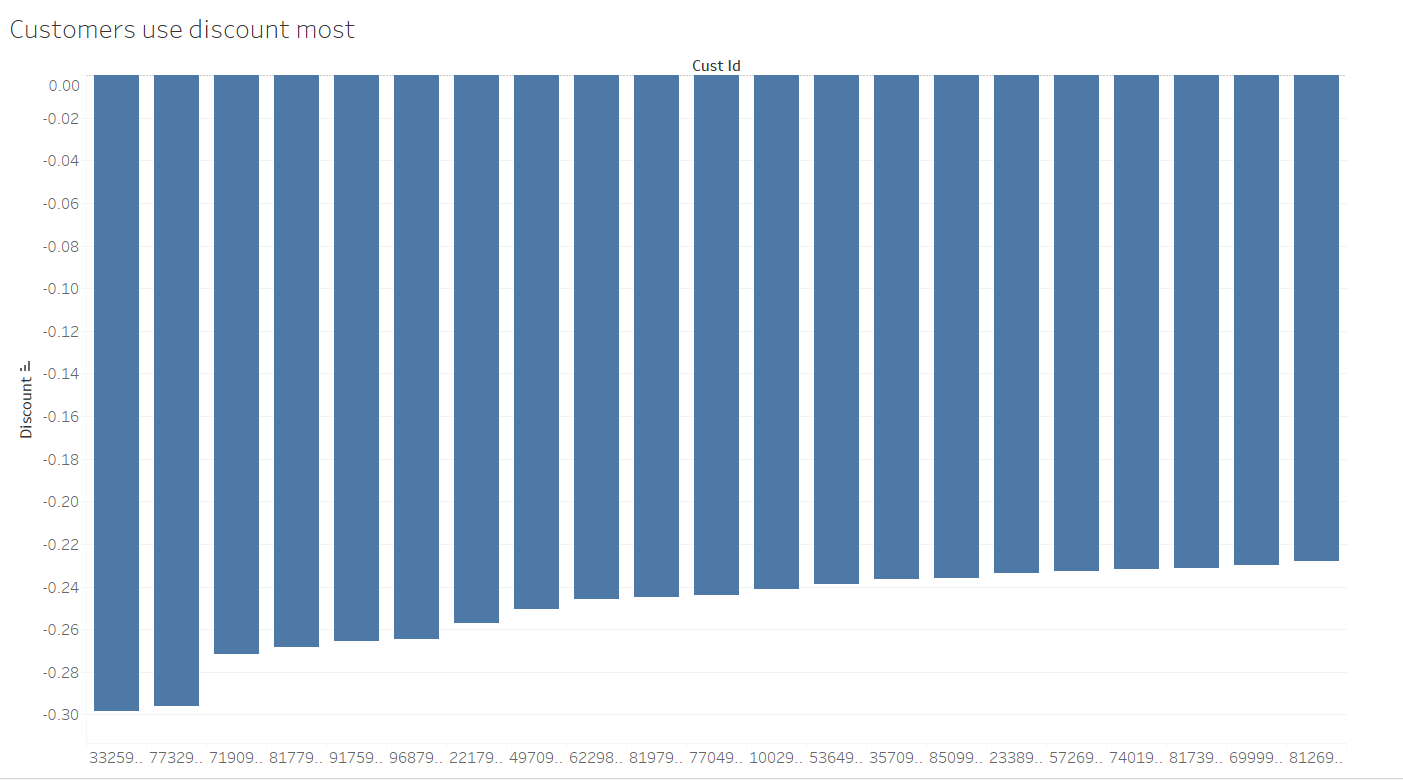
We decided to take a different approach that incorporates more human elements when analysing customer trends. In order to understand customer purchasing behavior, we aggregated features by customer level and explore human factors that can be complicated to express with a single variable:

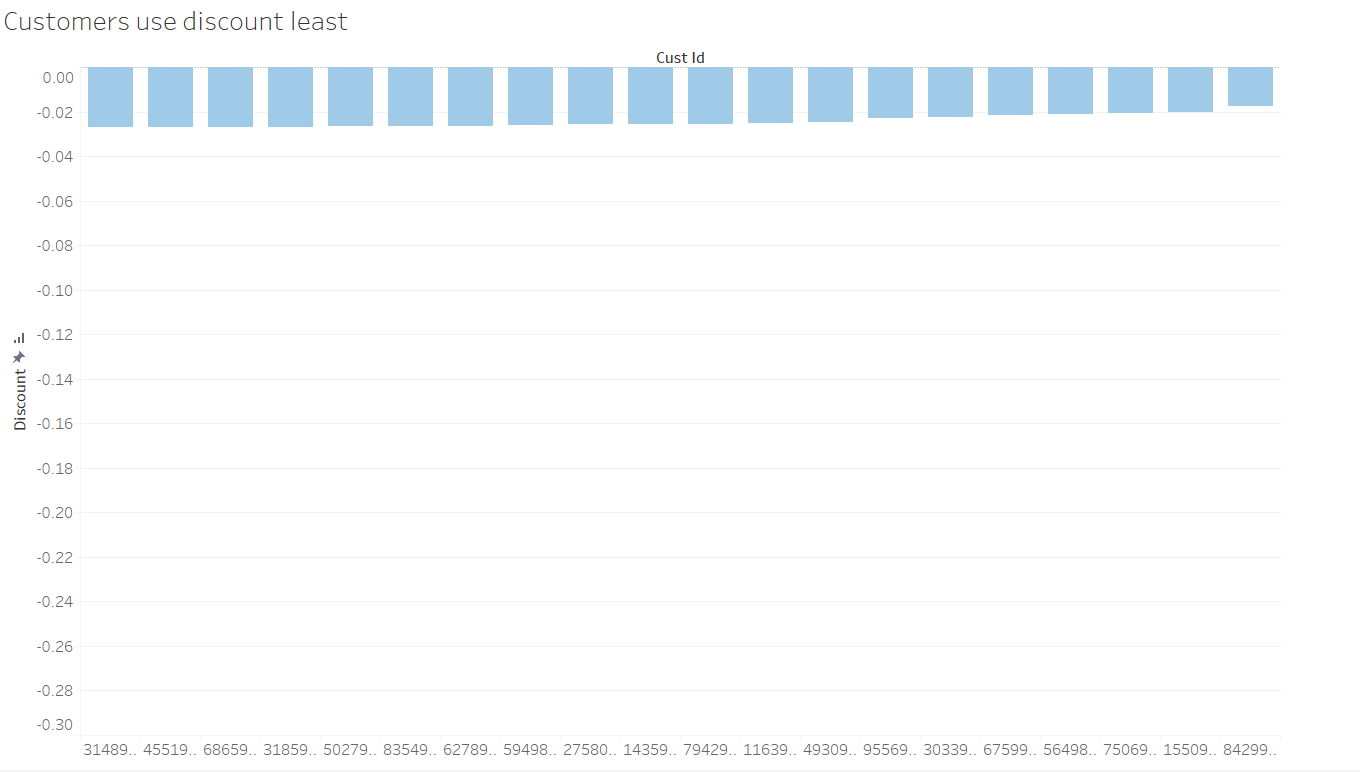
1. **Customers’ loyalty to store locations:** In order to find out customer loyalty (loyal to store), we calculated the unique store count that a customer has visited. Based on the result, we found that the majority of people only went to less than 10 stores (6642), taking up 83.86% of the total customer population. The number of people visiting over 30 stores is really small (23), taking up less than 0.3% of the whole customer population. And the extremely loyal customers, those who only visit a single store, takes up 4.79% of the total population.



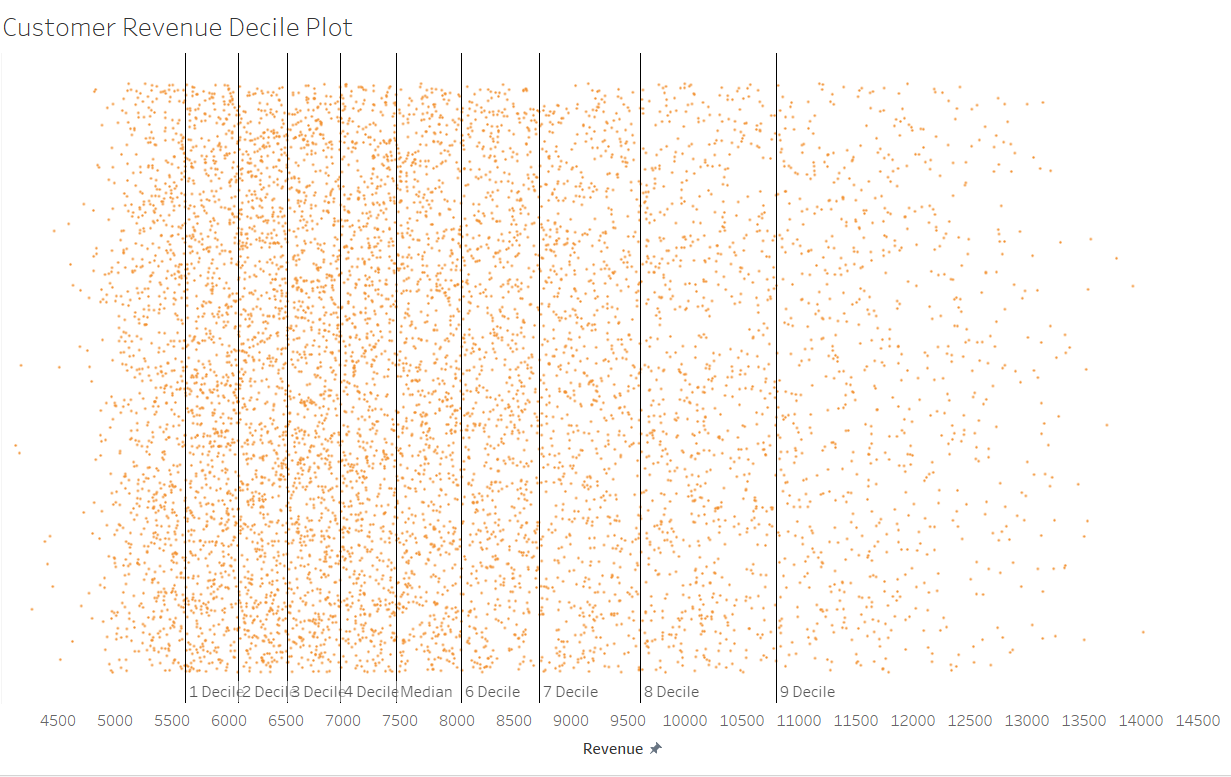
·

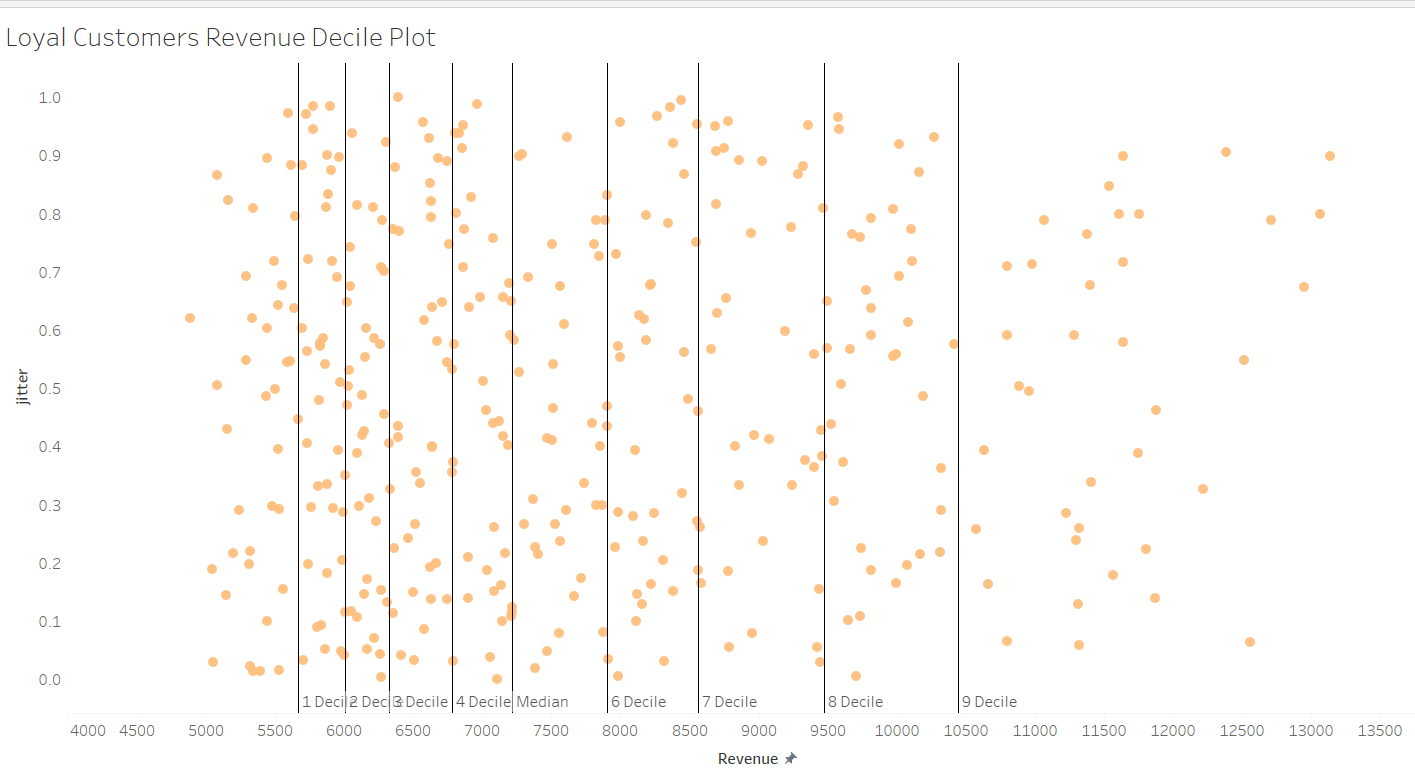
1. **Customers’ purchasing behavior on discounted products:** In order to calculate how the percentage of discount money as the total value (price should be paid before sale and discount) of a product on individual customer level, we used tran\_prod\_discount\_amt/tran\_prod\_sale\_amt to calculate the discount percentage of money should be spent for each customer. Here are two visualizations, one is the top 20 people who are really good at saving money and the second is the bottom 20 customers who are really not good at buying discount products. And based on the comparison of two plots, we can see the gap between the two groups of people is really huge.





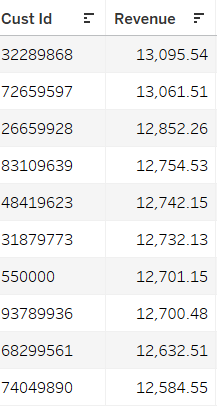
1. **Customer revenue analysis:** In order to analyze customers’ contribution to revenue, we calculated the aggregated tran\_prod\_sale\_amt for each customer. And we visualized its distribution using decile plot. From the plot above, we can see that the first 8 deciles are gradually widened in terms of revenue gap, however, the 9th and 10th deciles’ gap become even wider. This indicates that when it comes to a group of people who contribute the most revenue, their consumption difference becomes bigger. And when we filter customers to only leave those extremely loyal ones(only go to a single store), we can see their distribution is very similar to that of the whole customer population, indicating they share the same pattern in terms of revenue contribution: revenue contribution gap between people becomes wider and wider as the amount of revenue contribution goes up.



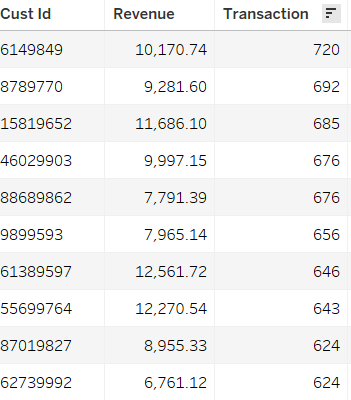
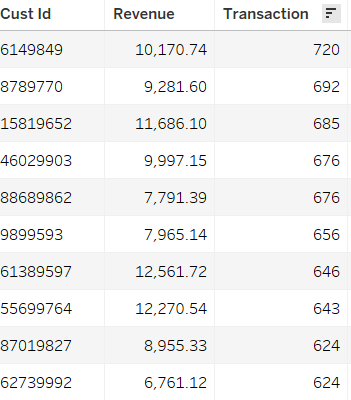
****

**Customer Summary:**

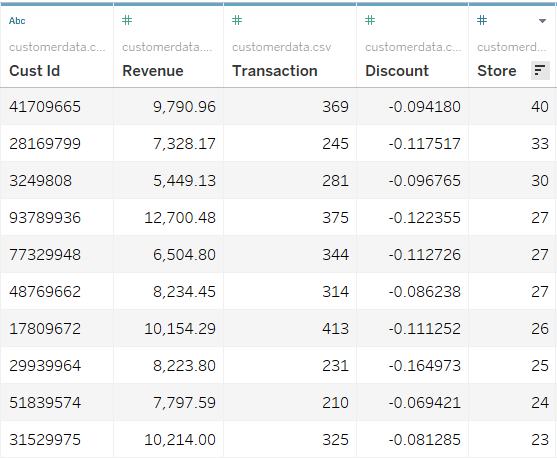
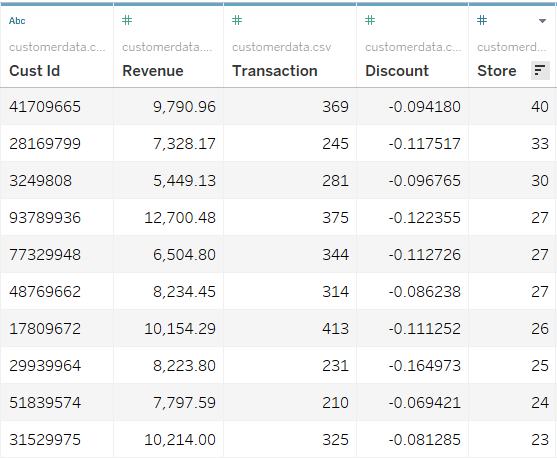
The top 10 revenue-driving customers are:



Customers with most transactions are:



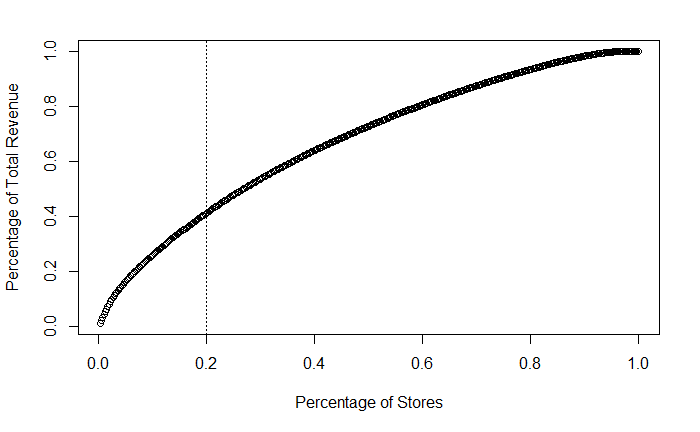
Customers visit the most stores are:



**3) Store**

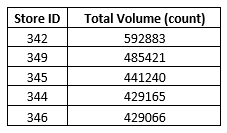
Currently, Pernalonga operates 421 stores, which have disparate distribution of total revenue. Store 302, for example, has a total revenue of $4.15 and only one transaction. In comparison, store 342 has a total revenue of $786,521 and 30,352 transactions. After carefully observing the performance between different stores, we find there are some stores that have extremely low total revenue and transactions. This could be due to the fact that they were just opened and had been in operation for very few days. It is also possible that these records are mistakes that happened during the data entry. Thus, we decide to exclude these stores from our segmentation analysis.

Moreover, it appears that stores, unlike products, don’t follow the 80/20 rule, but rather demonstrate a more diversified pattern. The top 20% of the stores, in this case, generate about 40% of the total revenue.

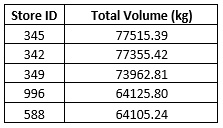


In order to find top stores within Pernalonga, we also examined their volumes, revenues, transactions, and customers.

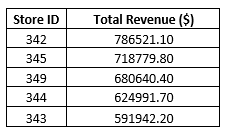
1. **Stores with highest volumes (count)**



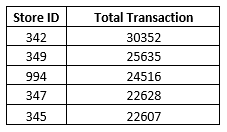
1. **Stores with highest volumes (kg)**



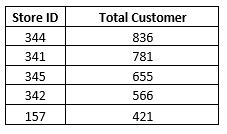
1. **Stores with highest total revenue**



1. **Stores with most transactions**



1. **Stores with most customers**



**Segmentation Analysis**

For segmentation analysis, we utilized k\_means modeling, which is intuitive and simple and perfectly compliments. Similar to our EDA above, we have again divided our K-Means analysis into product, customer and stores.

**a) Product segmentation using k\_means model**

Apart from examining the product categories and sub-categories, we also used K-means clustering to identify natural groupings based on product attributes. Here we took the following steps:

1. Identify the attributes used for clustering, for each unique product\_id, we calculated:

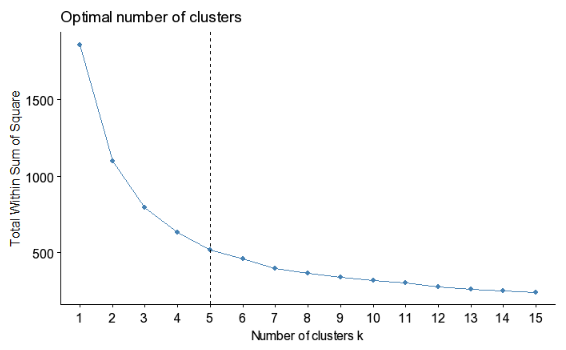
* Total Revenue
* Total Transactions
* Total Distinct Customers
* Total Stores
* Average Discount Rate
* Number of Discount Products
* Percentage of Discount Products

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Total Revenue** | **Total Transactions** | **Total Distinct Customers** | **Total Stores** | **Average Discount Rate** | **Percentage of Discount Products** |
| **Minimum** | 500.0 | 3.0 | 2.0 | 1.0 | 0.0000432 | 0.0005537 |
| **1st Quartile** | 939.5 | 307 | 189.0 | 124.0 | 0.0227892 | 0.0362067 |
| **Median** | 1898.0 | 757 | 405.5 | 220.0 | 0.1095774 | 0.0715926 |
| **Mean** | 5802.4 | 2764 | 725.8 | 221.9 | 0.1587617 | 0.1185600 |
| **3rd Quartile** | 4591.7 | 2061 | 882.8 | 324.0 | 0.2787683 | 0.1564843 |
| **Maximum** | 602109.4 | 769890 | 7862.0 | 419.0 | 0.6895196 | 1.0000000 |

A few things can be observed from the above summary table: the top quarter of products attracts far more customers and contributes much more revenue than the bottom three quarters of products, which again confirms the 80/20 rule; average discount rate for all products is about 10% while the max discount rate is as high as 69%; for most of the products, only a small percentage of times the products are on discount, but there are certain products have discounts all the time.

1. Normalize every key attribute above using min-max standardization, before we use distance-based clustering technique.
2. Use both elbow graph and average silhouette graph to choose the optimal number of k for clustering. While the elbow method looks at the percentage of variance explained as a function of the number of clusters, the silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). Both methods are useful for determining the number of k.

After comparing the results of both methods, we decided to use a cluster number of five to get the natural groupings of products.



1. We fitted the k-means models with different random seeds to get a steady clustering result. Then we can evaluate the attributes of each centroids to get the business understanding of each cluster. Below is table of attributes for each cluster centroid (all of the attributes have been scaled to between 0 and 1) :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Total Revenue** | **Total Transactions** | **Total Distinct Customers** | **Total Stores** | **Average Discount Rate** | **Percentage of Discount Products** |
| 1 | 0.035112 | 0.0155 | 0.281966 | 0.914481 | 0.123283 | 0.031123 |
| 2 | 0.008388 | 0.002594 | 0.115067 | 0.723884 | 0.4954 | 0.109235 |
| 3 | 0.001473 | 0.000348 | 0.023003 | 0.291503 | 0.479064 | 0.305507 |
| 4 | 0.004511 | 0.001809 | 0.073499 | 0.619741 | 0.094985 | 0.064234 |
| 5 | 0.001839 | 0.000733 | 0.0236 | 0.232097 | 0.074576 | 0.097621 |

We can then translate the attributes into business understanding:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster** | **Total Revenue** | **Total Transactions** | **Total Distinct Customers** | **Total Stores** | **Average Discount Rate** | **Percentage of Discount Products** | **Example** |
| **1** | High | High | High | High | Low | Low | “Apple” |
| **2** | Middle | Middle | High | High | Middle | Middle | “Tea” |
| **3** | Low | Low | Low | Low | High | High | “Toys” |
| **4** | Middle | Middle | Middle | Middle | Low | Low | “Soaps” |
| **5** | Low | Low | Low | Low | Low | Low | “DVD” |

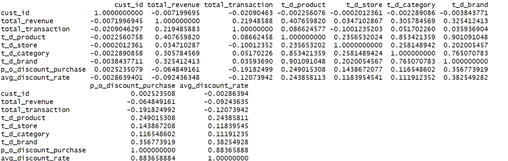
* Cluster 1: the first cluster is what we call “cash-drivers”. They are essentials that have large customer groups and high purchasing frequencies. Customers would buy them regardless of promotion. A typical example in cluster one is apple.
* Cluster 2: the second cluster is “buy with promotion”. Compared to essential cash drivers, these products have many alternatives and they are bought less frequently than necessary items. They need to have discounts to facilitate purchasing behavioral. For example, iced tea has a great fan base. But iced team also have many substitutes and people will choose iced tea over other drinks if they are on promotion.
* Cluster 3: the third cluster is “constantly needs help”. These products are always on promotion. For example, toys and gifts. Without promotion, they are too expensive. And they are only needed by a small proportion of customers.
* Cluster 4: the fourth cluster is “necessities”. Compared to “cash-drivers”, they are purchased less often, but they are still irreplaceable for almost everyone. Because of the importance and the typically low price of these products, they are almost never on discounts. A typical example would be soaps.
* Cluster 5: the fifth fluster we named “High-shelf” products. They only have a small customer base and contribute a small proportion of revenue. And promotion doesn’t work well for them too. An example would be CD/DVD. Almost everyone in modern society chooses to watch videos and listen to music online. Even these products are on promotion, they are barely noticed and less favored by customers.

**b) Customer segmentation using k\_means model**

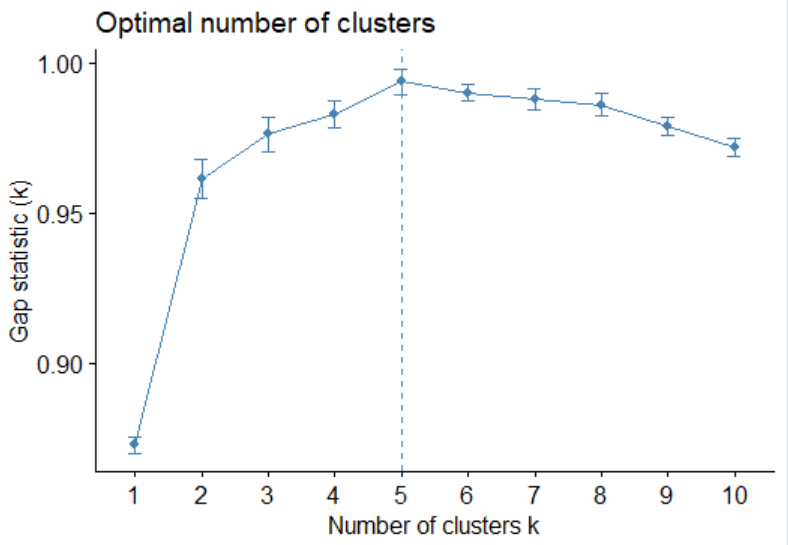
In order to build a customer segmentation model using k-means, we followed several steps:

1. Identify the attributes used for clustering, for each unique cust\_id, we calculated:

* Total Revenue
* Total Transactions
* Total Distinct products
* Total Stores visited
* Average Discount Rate
* Number of Discount Products
* Percentage of Discount Products

1. Then we need to normalize all features used to make sure different scaling won’t interfere with results. We used min-max normalization to accomplish this goal (all of the attributes have been scaled to between 0 and 1).
2. Besides this, we also calculated correlations between each feature and excluded features with high correlation. 
3. We built several k-means models by defining different numbers of clusters(k), and then used

the elbow method, gap statistics graph to choose the optimal k for clustering. After assessing the two plots (the actually show the same result), we finally are able to clearly state that 5 is the optimal number of clusters to choose.



1. Having decided that 5 clusters should be classified, we fitted the k-means models with different random seeds to get a steady clustering result. The resulting 5 clusters individually have 1894, 1889, 1069, 1677,1391 customers.

We then move to the model understanding phase. By Then calculating the attributes of each centroids we can get the business understanding of each cluster. Below is table of attributes for each cluster centroid :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Revenue** | **Total Transactions** | **Total categories** | **Total Stores** | **Average Discount Rate** |
| 1 | 0.2261363 | 0.3806635 | 0.6183601 | 0.08059838 | 0.4921419 |
| 2 | 0.2720808 | 0.3684890 | 0.5738651 | 0.06192342 | 0.2578444 |
| 3 | 0.3949973 | 0.6687389 | 0.6233128 | 0.05552853 | 0.3076482 |
| 4 | 0.4533388 | 0.3925081 | 0.7159229 | 0.09516995 | 0.4211302 |
| 5 | 0.7088462 | 0.4709178 | 0.7069522 | 0.07672178 | 0.3359648 |

We can then translate the attributes into business understanding:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster** | **Total Revenue** | **Total Transactions** | **Total categories** | **Total Stores** | **Average Discount Rate** |
| 1 | Low | Middel | Middle | Low | Middle |
| 2 | Low | Low | Low | Low | Low |
| 3 | Middle | High | Middle | Low | Low |
| 4 | Low | Middle | High | Low | Middle |
| 5 | High | Middle | High | Low | Low |

Cluster 1:

The first identified cluster is cherry-picker. They are discount lovers and they produce low-level revenue and the average percentage of discount on their purchases is high. And the category they buy is not that much, meaning they only purchase what they are interested in.

Cluster 2:

The second cluster is passers, people who seldoms shop or buy. They don’t buy many products even on sale and they nearly shop.

Custer 3:

The third cluster is customers who live nearby stores. They contribute a lot to revenue, they shop often, buy products they need, they are super loyal to stores and they don’t pursue discount products.

Cluster 4:

The fourth cluster is very like the first cluster, they are cherry-pickers at a wholesale level. They are discount lovers and they produce low-level revenue and the average percentage of discount on their purchases is high. The only difference between them and the first cluster is that they buy various products.

Cluster 5:

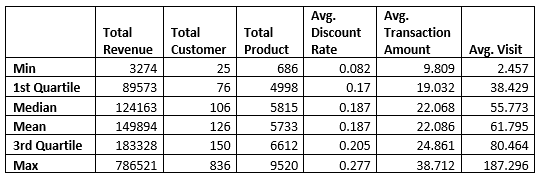
The fifth and last cluster are high-valued customers. They contribute the most to revenue, they shop often, buy a large variety of products, they are super loyal to stores and they seldom pursue discount products.

**c) Store segmentation using k\_means model**

As mentioned before, some stores have very low total revenue and transactions compared to their peers, so we decide to only include stores with more than 100 transactions. 1. For store segmentation analysis, we used K-means clustering to map stores into their natural groupings based on following store attributes:

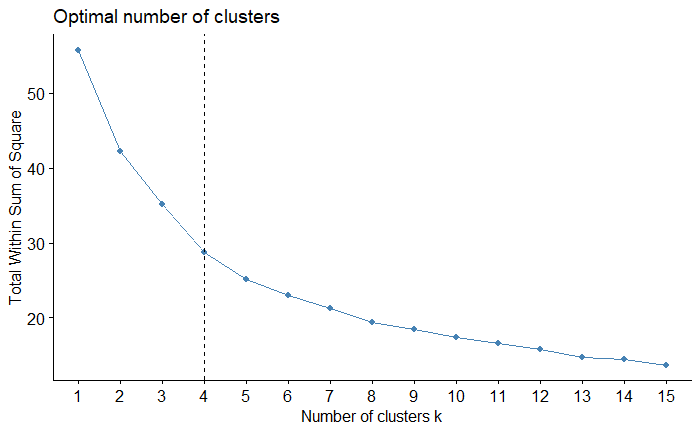
* Total Revenue
* Total Customer
* Total Product
* Average Discount Rate
* Average Transaction Amount
* Average Visit per Customer

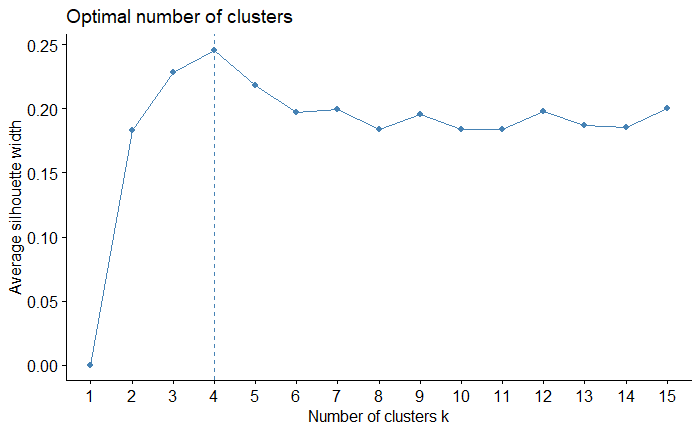
Following are descriptive statistics of these attributes:



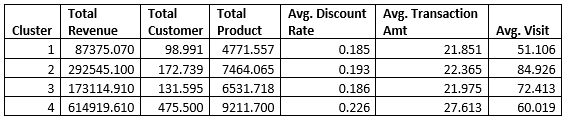
2. In order to give these attributes the same level of impact on the K-means clustering, we used the min-max standardization method.

3. Within Sum of Square (WSS) graph and Silhouette graph are plotted to decide how many clusters to separate these stores. Both methods agree on 4 to be the optimal number of k in this case.

****

****

4. Following is the numeric result of the k-means model for stores:



We can then translate the result into qualitative terms:



* Cluster 1: “7-11”. This group of stores is the lowest in total revenue, total customer, total product, and average visit per customer, indicating it is similar to “7-11” convenience store. These stores are small in size but large in quantity, making them easy to find. Most transactions there are small orders covering limited kinds of products. There are 228 such stores.
* Cluster 2: “Small-town Kroger”. This group of stores has a mid level of total revenue and high average visit per customer, indicating it is similar to a small-town Kroger, which has no or few competitors. People go there frequently but don’t buy a lot because there isn’t much traffic. There are 46 such stores.
* Cluster 3: “Large-town Kroger”. This group of stores is similar to cluster 2 in the sense that they cover similar amounts of products. What is different here is that cluster 3 has lower total revenue and total customer. This can be explained by more competitors, which attract some of its customer base. There are 131 such stores.
* Cluster 4: “Wal-Mart Supercenter”. This type of stores attracts the most customers, generates the highest revenue, and offers the highest discount rate. This group is similar to Wal-Mart supercenter in the United States. There are only 10 such stores, which also supports this theory.

**Results and Conclusion**

With the products, customers and stores clustered and analyzed, we lastly take a look at the implications of these clusters and how they interact with other clusters. The correlation is calculated by looking at the percentage of each green category that the blue categories have bought. Thus, each vertical column should sum up to 100% the green category.

Correlation between customer and product clusters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Customer\Product** | **Cash-Drivers** | **Promotional Purchases** | **Constantly Needs Help** | **Necessities** | **High-Shelf** |
| **Cherry Picker** |  | ✔ |  | ✘ |  |
| **Passers** |  | ✘ | ✔ |  |  |
| **Lives Nearby** |  |  |  | ✔ | ✘ |
| **Wholesale Cherry Pickers** |  |  | ✘ |  | ✔ |
| **High-Value** | ✔ | ✔ | ✔ | ✔ | ✔ |

Correlation between customer and store clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Customer\Store** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Cherry Picker** | ✘ | ✔ | ✔ |  |
| **Passers** | ✔ |  |  | ✘ |
| **Lives Nearby** | ✔ | ✘ | ✘ |  |
| **Wholesale Cherry Pickers** | ✘ | ✘ | ✔ |  |
| **High-Value** |  |  |  | ✔ |

Correlation between product and prstoreoduct clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Product\Store** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Cash-Drivers** | ✔ | ✘ |  |  |
| **Promotional Purchases** |  | ✘ | ✔ |  |
| **Constantly Needs Help** |  |  |  | ✔ |
| **Necessities** |  |  | ✘ | ✔ |
| **High-Shelf** | ✘ |  |  | ✔ |

Immediately, there are some interesting patterns that we can see. Looking at the first table, we can see that customers within the cherry picker cluster love promotional purchase products. At the same time, they are not particularly interested in necessities, as these items typically have a low base price and seldomly go on sale as we have discussed above. People who live nearby pay close attention to necessities, as to be expected as any other type of purchase they would probably make a deliberate trip to specific stores. High value customers will literally buy every type of product, which is to be expected.

Moving onto the correlation of the store clusters to customer clusters, the most prominent correlation is between store clusters 2 and 3 and cherry pickers. This suggests that cherry pickers like to shop at stores that are medium in size. If we think of the real world where small stores the size of corner grocery stores while large stores are the size of Costco. It would make sense for cherry pickers to lurk around medium sized stores that provide the sweet spot in quantity, selection and promotions typically found at medium sized stores. Looking at the ‘living nearby’ customers and ‘wholesale cherry picker’ customer segments, it intuitively makes sense why the former likes the smallest type of stores (neighborhood corner stores) and dislike larger stores while the latter dislikes smaller stores and prefers larger stores instead. Interestingly, we did not find a correlation between wholesale cherry pickers and the cluster 4 of stores (the largest), which again hints at medium stores having that perfect balance of selection and discounts.

Lastly, the two prominent correlations that we wanted to highlight from the last table are the ones between cash-driver products and the small store cluster and ‘constantly needs help’ products and the large store cluster respectively. Smaller stores typically need a higher profit margin compared to larger stores that more so rely on quantity to drive revenue, so it would make sense for smaller stores to favor cash-driver products. Larger stores can balance out the poor performance of some products with other products because of their sheer size, so they typically can afford to carry products that don’t sell too well like the ones that constantly need promotional help and the ones that are high-shelf (i.e. seldomly bought).

**Next Steps**

With our product, customer and store details dissected, we can now move onto suggesting and implementing targeted marketing tactics that would’ve previously been ineffective if deployed to the entire product, customer or store groups. Below are some ideas that we think are good places to start:

1. Developing targeted marketing campaigns for each customer segment based on their purchase pattern and affinity to certain clusters of products.
2. Rank customers by their total value and maximize the revenue from the top customers with specialized discounts.
3. Rearrange store layouts to promote products with lower sales volume (e.g. constantly needs help) and to highlight cash-drivers.
4. Create loyalty programs to reward high-value customers, lock in cherry pickers and encourage all categories to purchase more.
5. Create collaborations between certain products and store locations (e.g. product demonstration stands, sampling, etc.) based on store category to further drive sales of the popular product segments or promote circulation of the lesser-popular product segments.

All in all, we are confident that our exploratory data analysis and careful segmentation analysis will lay the perfect foundation for future successful marketing strategies.