# Project 1 Report

Team 6 | Shuwen Huang, Firay Geyik, Gaows Mohammad, Cecilia Wang

## Background

ACSE Supermarket, a company that sells everything, has over 40 stores in Lunitunia and sells over 100 thousand products in over 100 categories. ACSE customers can opt to join the Lunie Rewards program to avail of weekly sales and promotions. ACSE regularly partners with suppliers to fund promotions and derives a significant portion 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.

Your analytics consulting firm is being considered by ACSE (the client) to develop a marketing campaign to experiment on personalized promotions. While the details of specific partnerships with suppliers to fund the experimental personalized promotions are still being negotiated, you have started to receive data from the client. You have two weeks to analyze and understand the data and report back initial insights to the client. In order to be selected as the sole-developer of the marketing campaign, your team needs to demonstrate that you know the data very well, i.e., you need to show the client that you know the profiles of their stores, products and customers better than they do and are ready to take on the task of developing the marketing campaign.

From the client's point of view, they need to be confident that you know the answers to the following key questions:

- Who are the best customers in terms of revenues, profits, transactions/store visits, number of products, etc.?
- What are the products and product groups with the best volumes, revenues, profits, transactions, customers, etc.?
- Which stores rank the highest in volumes, revenues, profits, transactions, customers, etc.?

- Are there interesting groupings of customers, e.g., most valuable (buy everything at any price) or cherry-pickers (buy mostly on promotions), defined by certain categories (buy baby products or never buy milk), etc.?
- Other than product categories and sub-categories, are there other product groupings, e.g., Key Value Items (KVI) and Key Value Categories (KVC), traffic drivers, always promoted versus seldom/never promoted, etc.?
- Are there natural groupings of stores, e.g., stores frequented by cherry-pickers versus stores visited by most loyal customers?

**Data Preperation**

There are two datasets, transactions dataset and products dataset.

1.transactions.csv contains transaction history in 2017, 2018, 2019 and 2020 for over 9 million customers

- cust_id – Customer ID: Format of 1######### represents a Lunie Rewards member
- store_id – Store ID
- prod_id – Product ID
- trans_id – Transaction ID
- trans_dt – Transaction Date
- sales_qty – Quantity/units of the product in the transaction
- sales_wgt – Weight of the product in the transaction if sold by weight
- sales_amt – Sales amount for the product before discounts in the transaction

2. products.csv contains the product to subcategory and category mapping and descriptions for over 100,000 products

- prod_id – Product ID
- prod_desc – Product description
- prod_section – Product section description
- prod_category – Product category description
- prod_subcategory – Product subcategory description
- prod_type – Product type description
- prod_mfc_brand_cd – Code representing the Product manufacturer/brand
- prod_unit_qty_count – Count per unit quantity of the Product

- prod_count_uom – Unit of measure (UOM) for a count of the Product
- prod_uom_value – Value UOM per count of the Product

As for data preparation, after importing both datasets, since the original transactions dataset has over 9 million rows and is therefore too large and could not be efficiently processed in Jupyter Notebook, we decide to use a random sample of this dataset and for the rest of the business questions, we should also use the random sampled subset of transactions. After random sampling, the new transactions dataset now has 1,757,241 rows and 9 columns.

Then, we check if there are NA values in both datasets and  remove all NA values in. Next, we want to check if all the columns make sense. For the transactions dataset, we want to make sure that sales amount and sales quantity are not negative, and remove those negative rows.

After data cleaning, the products dataset has 152,578 rows and 10 columns. The transactions dataset has 1,733,239 rows and 9 columns.

In order to solve the business problems, including stores analysis, customers analysis, and product analysis, the next step is to merge the transactions dataset and products dataset based on “prod_id”. Now the new merged dataset has 1,703,298 rows and 18 columns.

**Business Questions**

**1. Who are the best customers in terms of revenues, profits, transactions/store visits, number of products, etc.?**

```
cust_id
1127597307        4956.39
1133063688        1204.24
1127617494        1007.64
1127804456         846.94
1006885219         801.83
```

In terms of revenue/profit, we want to sort the “sales amount” column by descending order. The top 5 customers’ IDs are shown above.

```
cust_id
1045022989        118
1147458804         82
1143554806         57
1126912017         44
1135252713         44
```

The top 5 customers are shown above in terms of highest transactions.

```
cust_id
1045022989        118
1147458804         82
1143554806         57
1126912017         44
1135252713         44
```

The top 5 customers are shown above in terms of store visit.

```
cust_id
1045022989        107
1147458804         76
1143554806         53
1126912017         44
1135252713         43
```

```
cust_id
1045022989        118
1147458804         82
1143554806         57
1126912017         44
1135252713         44
```

As for the number of products, the right chart shows the Top 5 customers that bought the most number of unique products; the left charts shows the Top 5 customers that bought the most number of products, however, not unique.

**2. What are the products and product groups with the best volumes, revenues, profits, transactions, customers, etc.?**

| prod_id | |
|---|---|
| 20189092 | 7147 |
| 20175355001 | 1606 |
| 21097012001 | 799 |
| 20028593001 | 682 |
| 20070132001 | 631 |

The Top 5 product IDs are shown above in terms of best volume.

| prod_id | |
|---|---|
| 20175355001 | 40137.21 |
| 20027156 | 34961.00 |
| 20159690001 | 31246.00 |
| 20252014 | 30588.61 |
| 20055266001 | 29993.88 |

The Top 5 product IDs are shown above in terms of highest profit/revenue.

| prod_id | |
|---|---|
| 20189092 | 66946 |
| 20175355001 | 27698 |
| 20055266001 | 7945 |
| 20070132001 | 7925 |
| 20812144001 | 7581 |

| prod_id | |
|---|---|
| 20189092 | 66946 |
| 20175355001 | 27481 |
| 20070132001 | 7887 |
| 20055266001 | 7555 |
| 20812144001 | 7446 |

The Top 5 product IDs are shown above in the left chart in terms of highest transactions (not unique).

The Top 5 product IDs are shown above in the right chart in terms of highest unique transactions.

**3. Which stores rank the highest in volumes, revenues, profits, transactions, customers, etc.?**

| store_id | |
|---|---|
| 1212 | 385040.85 |
| 1050 | 360811.26 |
| 1004 | 328369.59 |
| 1007 | 323747.74 |
| 1066 | 321932.86 |

The Top 5 stores are shown above in terms of highest revenue/profits.

| store_id | |
|---|---|
| 1212 | 92974 |
| 1050 | 86073 |
| 1007 | 81734 |
| 1004 | 76996 |
| 1066 | 75939 |

The Top 5 stores are shown above in terms of highest volume.

| store_id | |
|---|---|
| 1212 | 73436 |
| 1050 | 67461 |
| 1007 | 65814 |
| 1004 | 60922 |
| 1066 | 58956 |

| store_id | |
|---|---|
| 1212 | 67003 |
| 1050 | 62060 |
| 1007 | 61713 |
| 1004 | 55173 |
| 1066 | 53895 |

The Top 5 stores are shown above in the left chart in terms of highest transactions (not unique).

The Top 5 stores are shown above in the right chart in terms of highest unique transactions.

| store_id | |
|---|---|
| 1212 | 73436 |
| 1050 | 67461 |
| 1007 | 65814 |
| 1004 | 60922 |
| 1066 | 58956 |

The Top 5 stores are shown above in terms of the most customers.

## 4. Customer Grouping

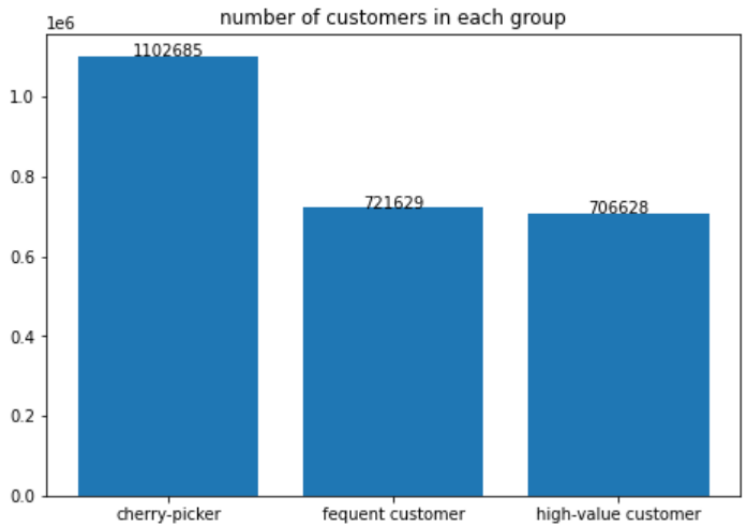
In this part, we applied K-means model to understand the customer pattern and potential groupings. Based on that, customers can be provided with discounts, offers, promo codes etc. Considering the information provided in the dataset, we decided to deploy FRM model because it includes three dimensions of customer value. In addition to FRM model, we added a new measure avg_discount to evaluate the customer preference for discounts.

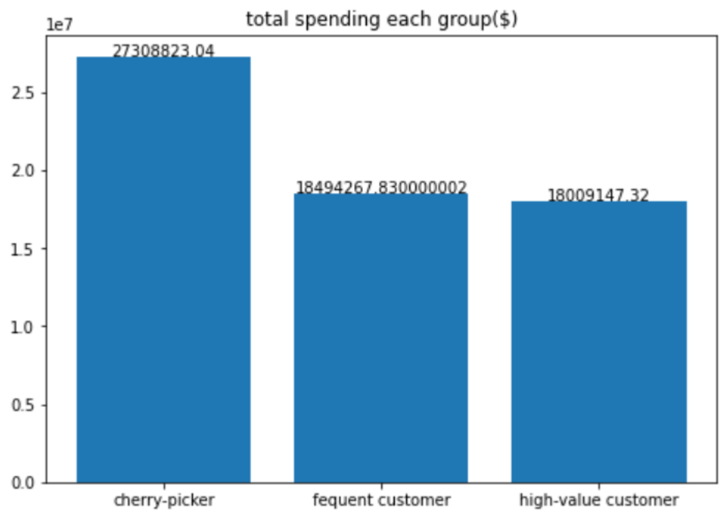
- Frequency: the frequency is calculated by the time customer have transactions. In other word, we group by customer ID to count the distinct value of transaction ID.
- Recency: the recency is calculated by the difference of last time customer had the transaction and the last date in our dataset.
- Monetary: the monetary is calculated by average total money spent on each transaction by each customer.
- Avg_discount: Based on the unit price of each product, the discount rate is defined as (1- current price)/max price in history. We take average discount rate for each customer and and product to build this feature.

After scaling the 4 features, we build K-Means model and decide cluster size of 3 based on the elbow method (appendix 4.1). And the results shows that we have three different groups of customers: cherry-pickers, frequent customers and high-value customers (appendix 4.2).

| **Customer group** | **Avg_discount(%)** | **Frequency(times)** | **Recency(days)** | **Monetary($)** |
|---|---|---|---|---|
| cherry-pickers | High (36.35) | Medium (4.24) | Low (916.64) | Low (4.56) |
| frequent customers | Medium (34.08) | High (4.3) | Medium (443.34) | Medium (4.78) |
| high-value customers | Low (30.97) | Low (4.14) | High (83.97) | High (5.10) |

The distribution of three groups are listed blow:




**Cherry-Picker:** this group is the most sensitive to discount with the highest avg_discount rate. They also share the lowest recency and monetary. Hoever, they are the largest group of our customers and also revenue drivers.

**Frequent Customers:** this group is in the middle for discount sensitivity, recency and monetary. However, they visit the stores more frequently compared with other groups. Thus it drives the second highest total revenues to the stores.

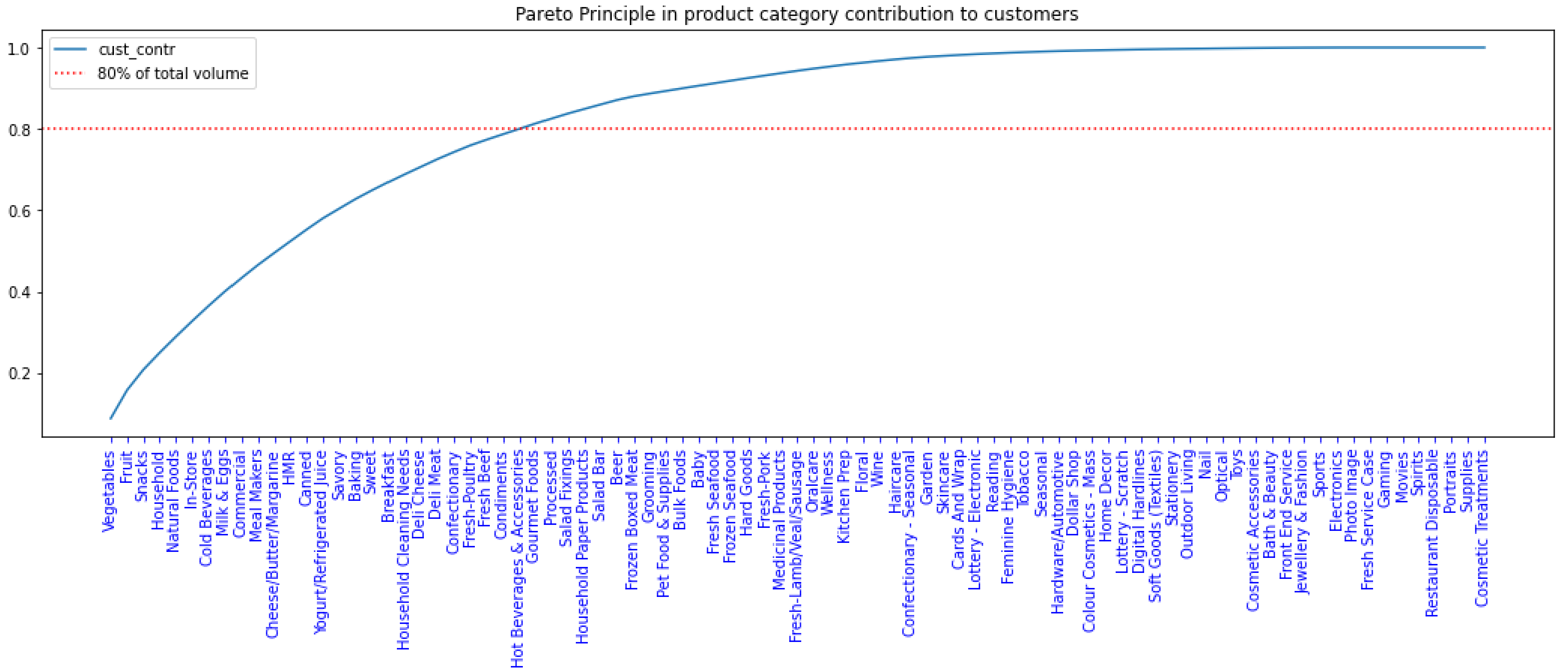
**High-value Customers:** this group is in the least sensitive to discount rates. And they are the most recent customers, with highest average spendings on a single transaction. From frequency and recency we can find that high value customers are probably our new customers. Although they counted as third place as for total spendings, they have the highest potential to our revenue because of high monetary.

### 5. Product Grouping

In order to find other natural grouping in products, we first followed the distribution of products that contribute relatively more than others based on a few metric:

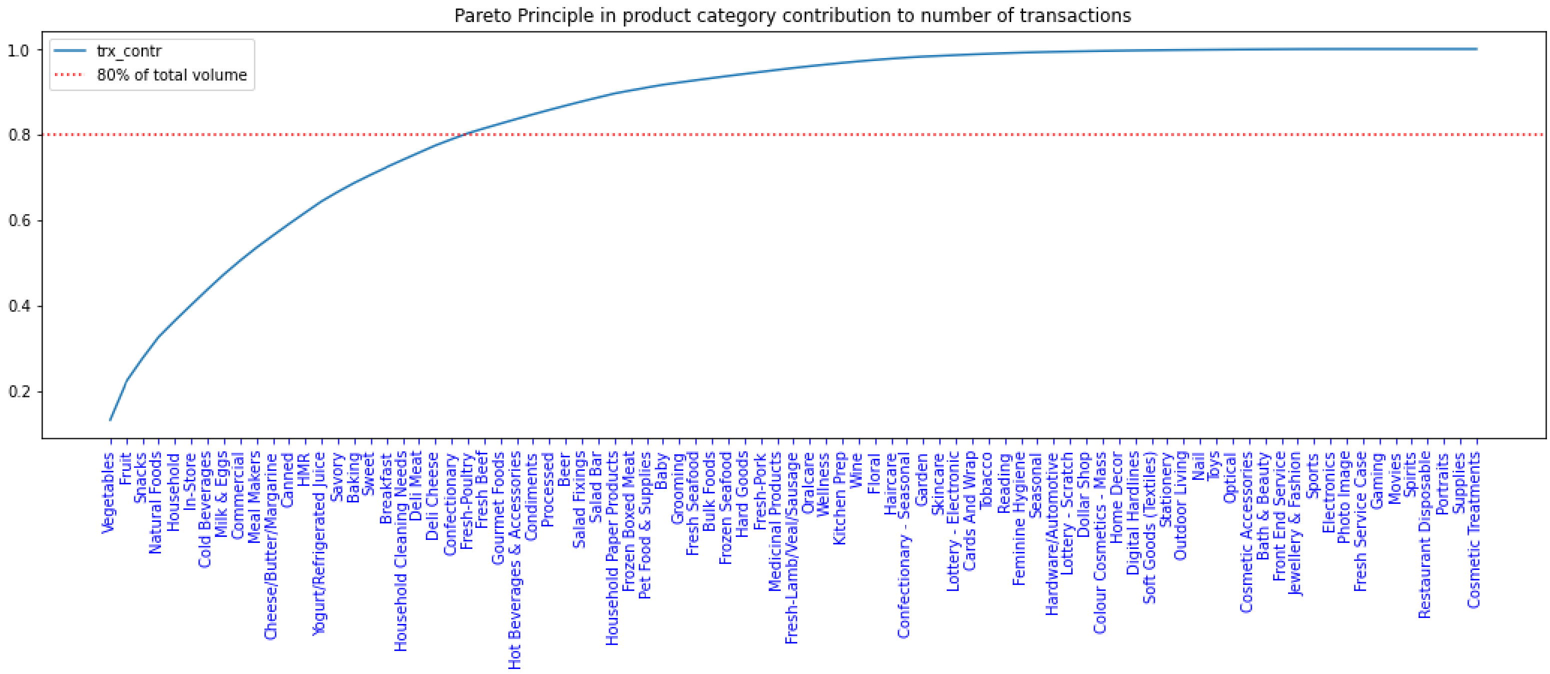
a. Customer traffic or Key Traffic Drivers
b. Sales volume or Key Value Items

<u>Key Traffic Drivers / Product categories contributing to customers:</u>




This graph plots the cumulative sum of customers as the y axis with the product categories as the x-axis

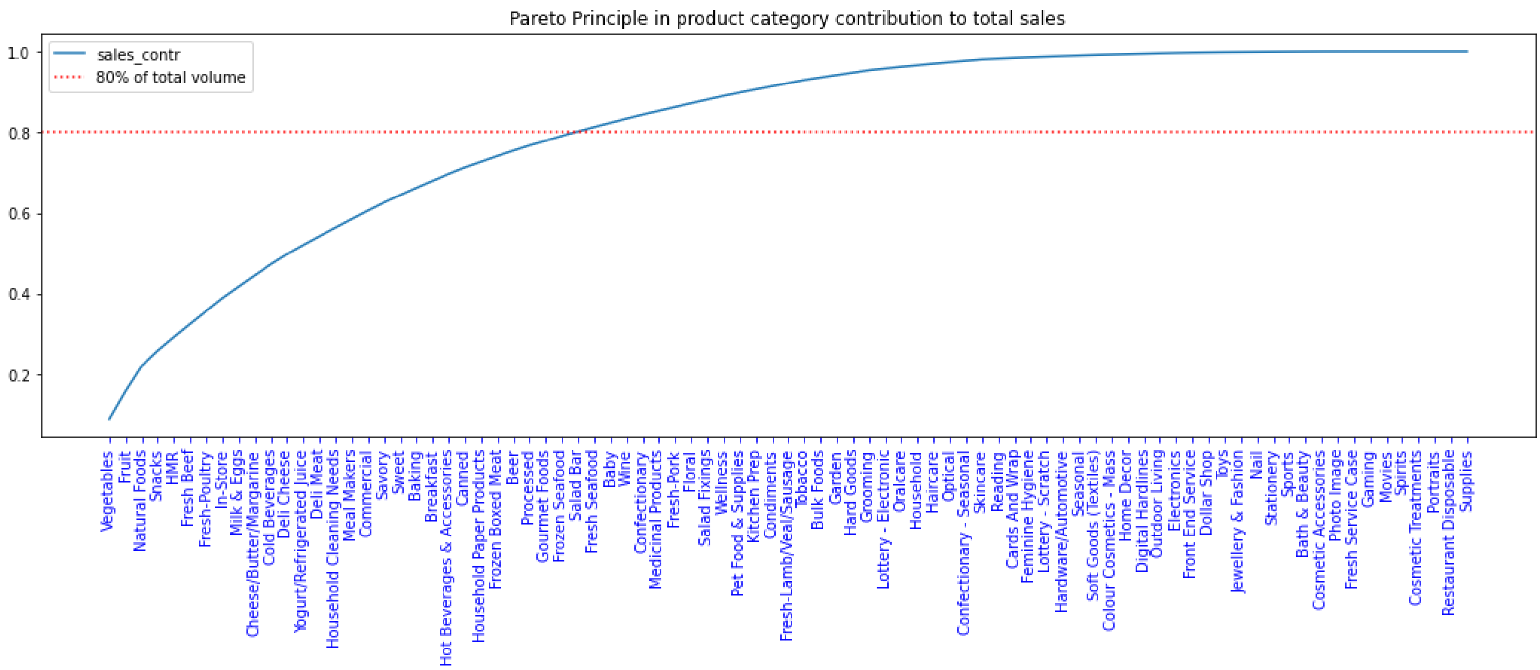
As we can see from the above distribution, in our sample of transactions, the distribution of customers to the product categories, follow a pareto principle, a large chunk of the customers are buying a fewer selection of products. The product categories at the left bring more customers than the ones at the right. from the data, 80% of the customers are contributed by the leftmost 26 categories starting from “Vegetables” to “Hot Beverage and Accessories”




This graph plots the cumulative sum of transactions as the y-axis with the product categories as the x-axis.

A similar distribution was found with the transactions for each category as well. However, here we see that only 23 product categories contribute to 80% of the transactions. These important categories start from “Vegetables” to “Fresh-Poultry”.

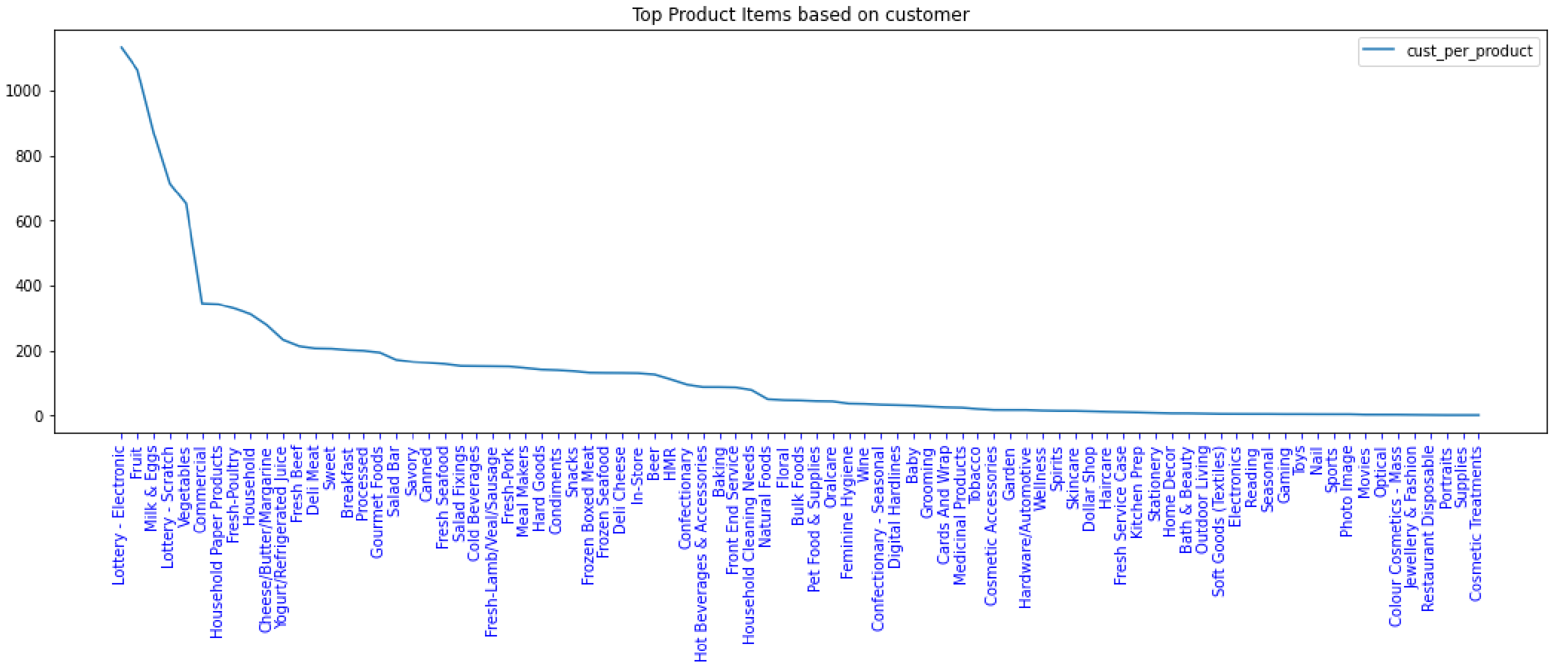
Key Value Items / Product categories contributing to sales:



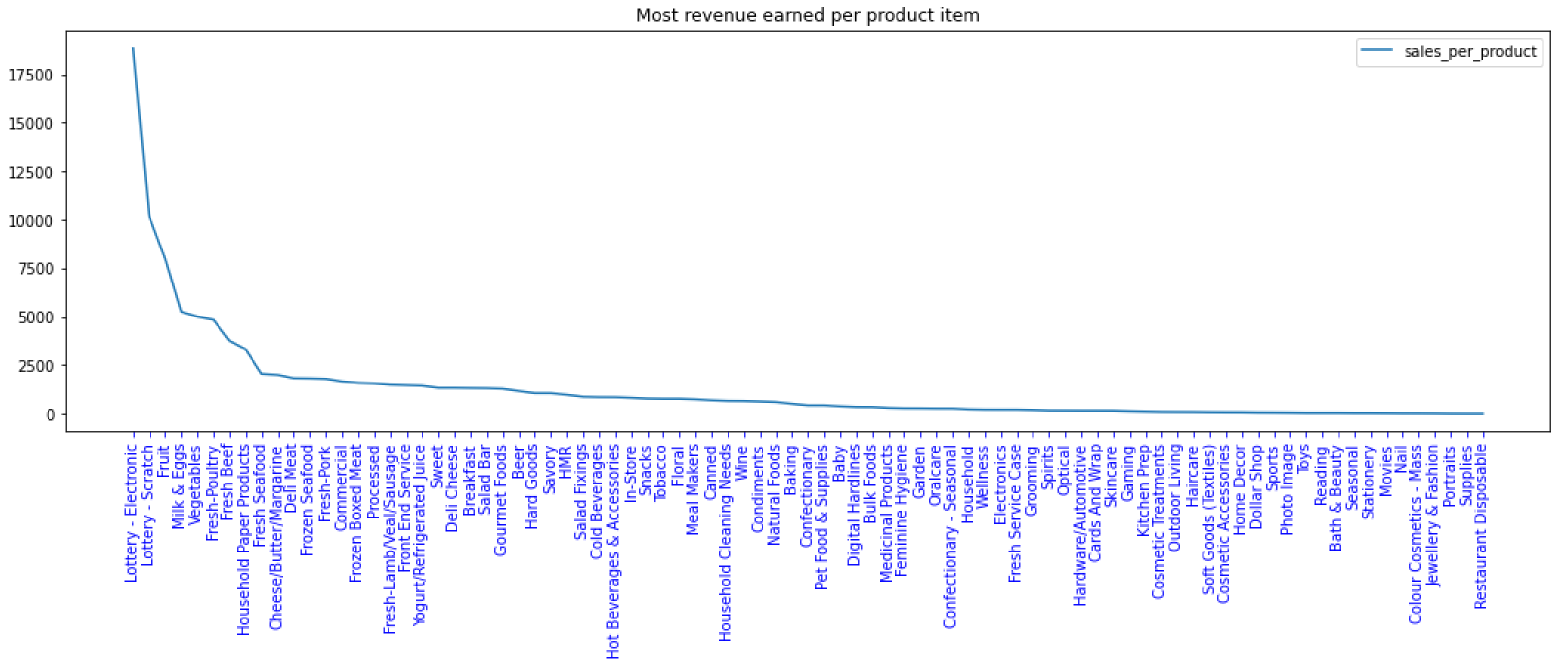
In the above distribution, cumulative sales is plot as the y-axis with product categories as the x-axis. 80% of the total sales volume is driven by the leftmost 30 product categories, starting from “Vegetables” to “Salad Bar”

At this point, what we have not considered in the above partitioning/segmenting the product is the number of product items that each of these categories have. Some of the high contribution of the above could be because of the high number of product items. Hence we looked into these categories normalized by their number of product items.

<u>Per Product Item Performance:</u>




We have plot the average number of customers per product id for each of the categories, and its more obvious now, that the left-most 6 categories draw in way more customers than any of the other categories. These 6 (like “Lottery-Electronic”) are not in the top 80% customer categories, however, since it has few product items, when we normalized the categories, it stood out. This is important in understanding which product id or individual product items are important drivers for the business.




The above distribution plots the revenue earned per product category for each of its product items

Similar trend from the per item customer distribution is observed here. The leftmost 9 categories contribute to significantly more sales per item than the others. So these also need further consideration in future product placement, campaigning and supplier negotiations.

<u>Algorithmic Clustering among product items.</u>

From the dataset we have, we found there to be two distinct clusters of product items in the data.

| | total_sales | Total_qty | cust_count | prod_id_count |
|---|---|---|---|---|
| Cluster | | | | |
| 0 | 57,296,808 | 13,435,316 | 2,048,993 | 91,060 |
| 1 | 5,448,558 | 1,275,230 | 467,477 | 52,696 |

The two clusters can be distinguished based on their sales volume, qty sold and logically on the customers they bring and the number of product items they contain. Although cluster 0 contains roughly twice the number of product items than cluster 1, its contribution to sales and customer traffic is many times over.

| | avg_price_item | avg_qty_item | avg_discount |
|---|---|---|---|
| Cluster | | | |
| 0 | 5.42 | 1.27 | 33.13 |
| 1 | 5.40 | 1.26 | 34.52 |

Surprisingly the average price of items are very similar between the two clusters, and so are the number of product items and even more surprisingly the amount of discount offered as well.

We believe that there are intrinsic differences between these two clusters that draw out customers differently. A snapshot of the tabulation of these clusters are available in the appendix.

## 6. Store Grouping

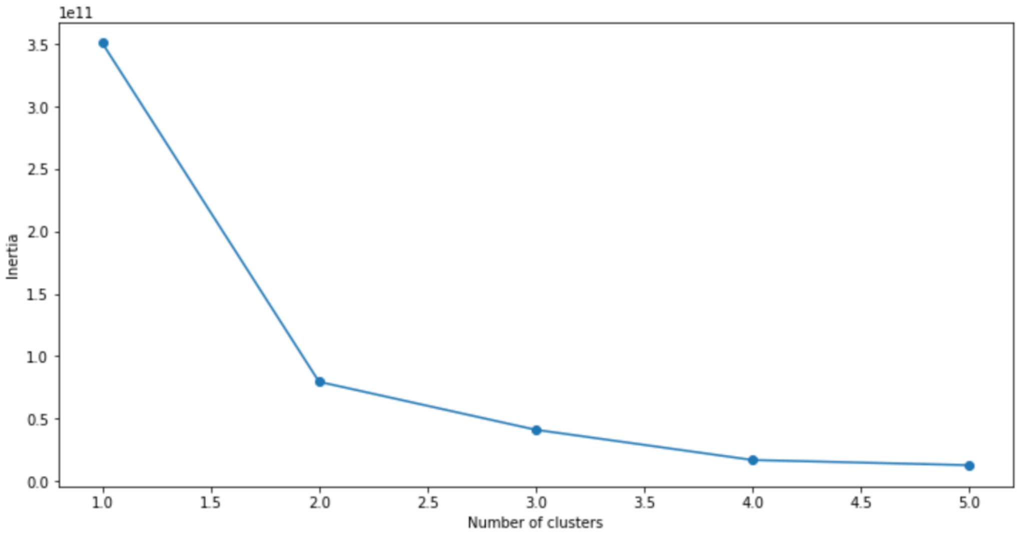
There are 58 stores in our dataset. In these stores, we first study the top 5 most frequently visited stores by each group of customers(appendix 6.1).

| Ranking | cherry-pickers | frequent customers | high-value customers |
|---|---|---|---|
| #1 | 1212 | 1212 | 1007 |
| #2 | 1050 | 1007 | 1212 |
| #3 | 1007 | 1050 | 1050 |
| #4 | 1004 | 1004 | 1004 |
| #5 | 1066 | 1066 | 1066 |

From the result we can find store 1212, 1050, 1007, 1004 and 1066 are most popular among all groups of customers, which indicates that the 5 can be accounted as ‘flag-stores’.They attracted the most traffics from most customers. The only slight differences are 1007, 1212 and 1050, where high value customers prefer store 1007 more than the other 2 groups. With plots of distributions (Appendix 6.2), we can find that customers don’t show significant different patterns in store selection.

# Appendix

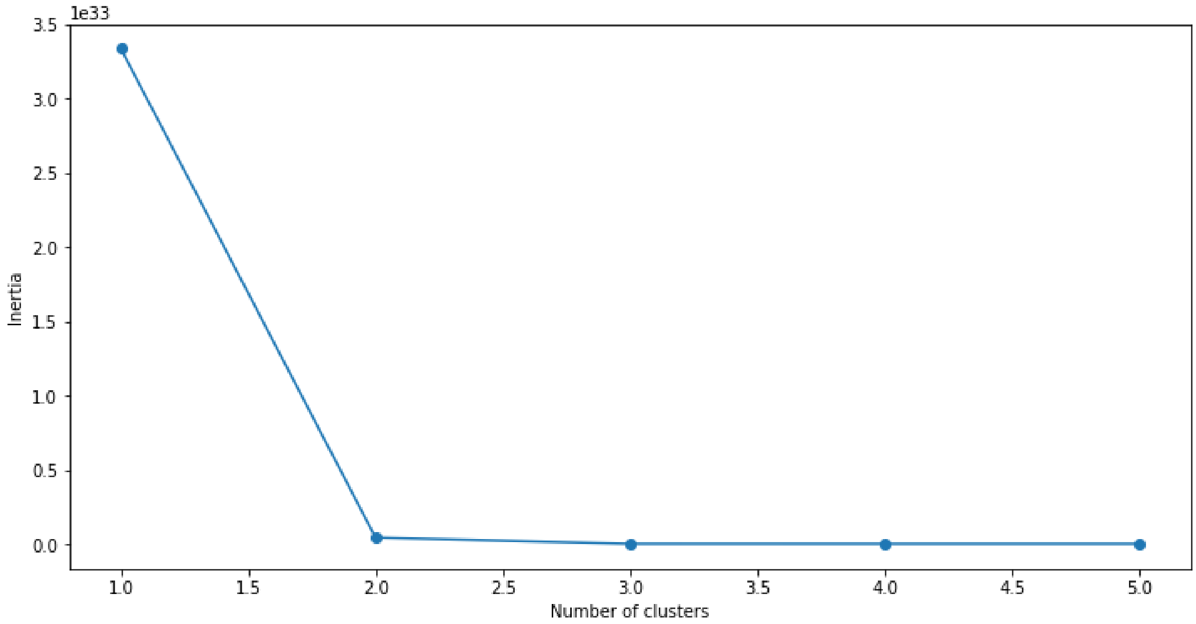
4.1 SSE measure of cluster size for question 4.




4.2 Clustering result for question 4.

| cluster | avg_discount | frequency | recency | monetary |
|---|---|---|---|---|
| 0 | 36.345994 | 4.240099 | 916.643463 | 4.567942 |
| 1 | 34.084752 | 4.308512 | 443.340675 | 4.784780 |
| 2 | 30.966602 | 4.139560 | 83.965092 | 5.106690 |

5.1 SSE measure of cluster size for question 4.




5.2 Snapshot of clusters and their constituent categories' sales performance:

| Cluster | prod_category | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Baby | 71449.0 | 8.650610 | 33.001056 | 0.15 | 1.99 | 3.49 | 7.00 | 4956.39 |
| | Baking | 226169.0 | 4.695041 | 3.310515 | 0.04 | 2.59 | 3.99 | 5.98 | 129.87 |
| | Bath & Beauty | 2658.0 | 5.766881 | 3.837864 | 0.22 | 2.94 | 4.99 | 7.99 | 49.96 |
| | Beer | 104205.0 | 7.223133 | 5.943823 | 1.19 | 2.79 | 5.32 | 10.80 | 161.28 |
| | Breakfast | 193337.0 | 5.295391 | 4.310358 | 0.24 | 3.49 | 4.99 | 5.99 | 1397.20 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1 | Toys | 307.0 | 10.040684 | 10.150448 | 0.10 | 3.38 | 6.99 | 12.98 | 79.97 |
| | Vegetables | 121724.0 | 3.650451 | 2.287993 | 0.02 | 1.99 | 2.99 | 4.99 | 83.86 |
| | Wellness | 4552.0 | 12.641852 | 10.334323 | 0.20 | 6.49 | 10.49 | 15.49 | 122.99 |
| | Wine | 3275.0 | 15.467255 | 10.417840 | 2.08 | 10.62 | 13.05 | 16.59 | 192.72 |
| | Yogurt/Refrigerated Juice | 24756.0 | 4.617860 | 2.409523 | 0.00 | 3.00 | 3.99 | 5.68 | 63.84 |

6.1 Store ranking for different groups for question 6.

```
store_grouping.loc[store_grouping['cluster']==0].head(5)
```

| | cluster | store_id | number_of_transactions |
|---|---|---|---|
| **47** | 0 | 1212 | 197742 |
| **21** | 0 | 1050 | 182910 |
| **5** | 0 | 1007 | 181679 |
| **3** | 0 | 1004 | 165992 |
| **24** | 0 | 1066 | 159329 |

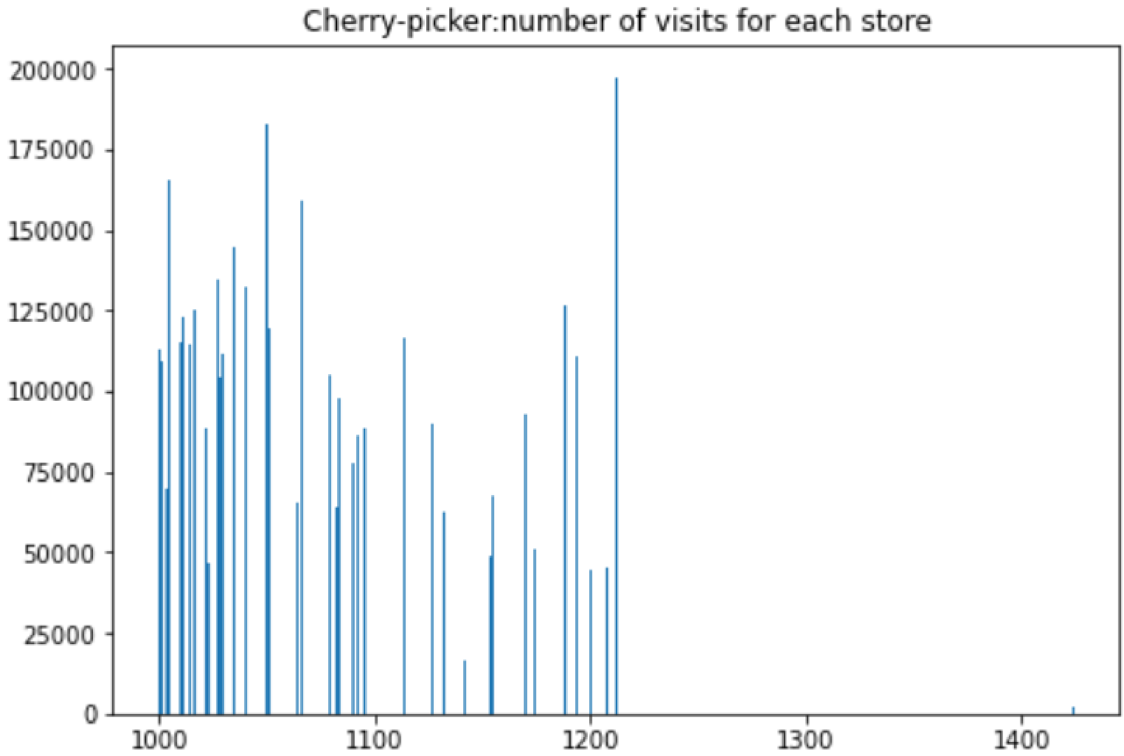
```
store_grouping.loc[store_grouping['cluster']==1].head(5)
```

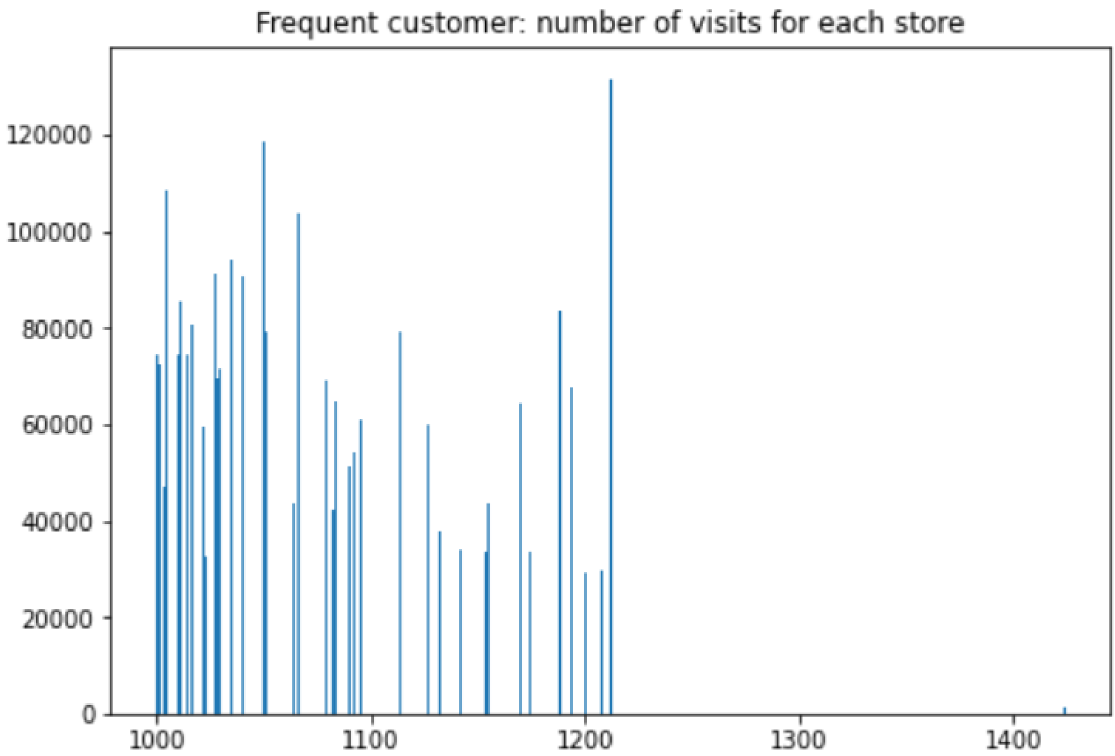
| | cluster | store_id | number_of_transactions |
|---|---|---|---|
| **105** | 1 | 1212 | 131740 |
| **63** | 1 | 1007 | 125354 |
| **79** | 1 | 1050 | 118651 |
| **61** | 1 | 1004 | 108246 |
| **82** | 1 | 1066 | 103811 |

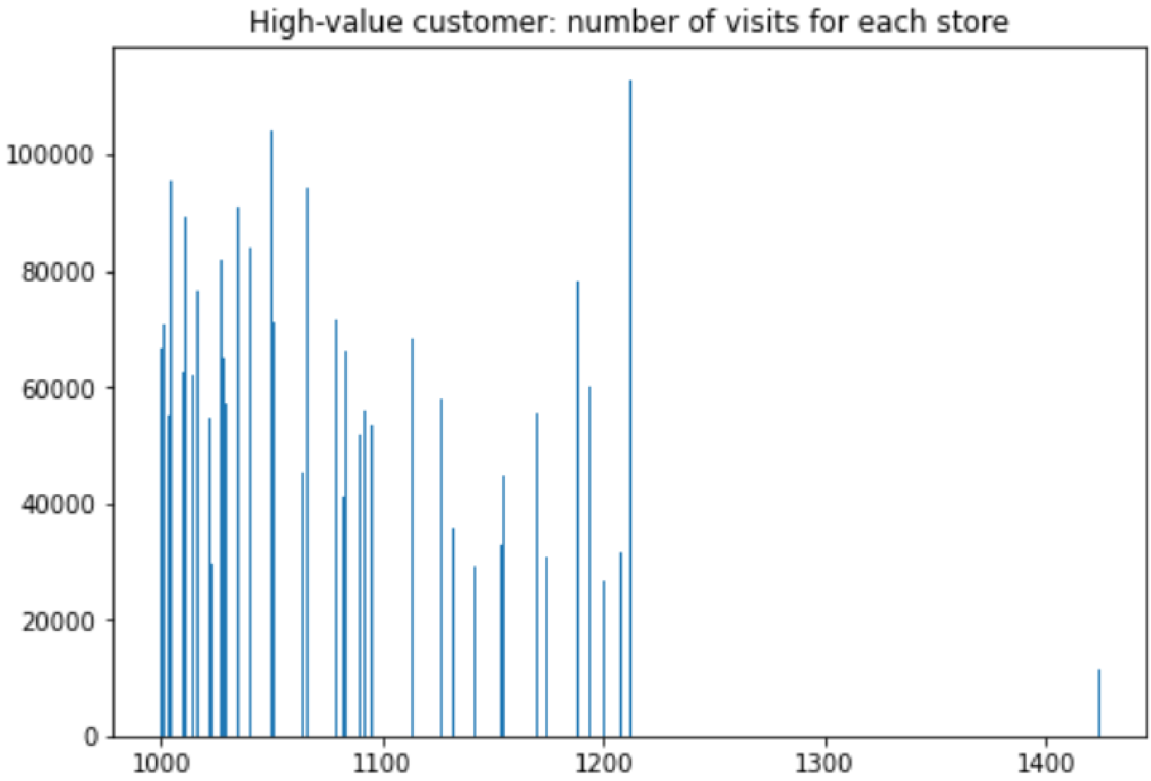
```
store_grouping.loc[store_grouping['cluster']==2].head(5)
```

| | cluster | store_id | number_of_transactions |
|---|---|---|---|
| **121** | 2 | 1007 | 113011 |
| **163** | 2 | 1212 | 112802 |
| **137** | 2 | 1050 | 104013 |
| **119** | 2 | 1004 | 95454 |
| **140** | 2 | 1066 | 94287 |

6.2 store visits distribution by customer group for question 6.




High-value customer: number of visits for each store
0
20000
40000
60000
80000
100000
1000
1100
1200
1300
1400