2022

Customer Segmentation for retail industry

MARKETING INSIGHTS BASED ON CUSTOMER ANALYTICS SKILLS: PYTHON, SQL, DATA ANALYSIS

Table of Contents

Executive Summary

The purpose of the report is to propose an RFM-based clustering technique and to classify customers of a national convenience store chain into segments. The proposed model helps prioritize the targets of future marketing campaigns. Recommendations made based on the insights of the analysis are presented for managerial strategy development.

The report utilizes a dataset derived from the loyalty card system in-store which includes transaction records of 3000
customers over a period of six months. The dataset was separated into four files. In this report, two files are used as the main sources to underpin the analysis.

The report combines traditional RFM clustering with newly added customer behavioural features to generate clearer images of customer groups. Apart from core features, which are Recency, Frequency as well as Monetary, the length of a customer engagement period, average spending per basket, and the count of unique product id were also taken into consideration. The k-means algorithm was employed to perform the clustering. An index of the quality of clustering was used to optimize the number of groups for K-means.

Five groups were generated which were "Low Contribution Loyal", "At Risks", "Low-Value Lapsed", "High Contribution Loyal" and "High Contribution Active". Each group was then given a customer persona. A statistical summary was also provided to distinguish the difference between groups. The results indicated that two groups of customers should be considered as priorities of upcoming marketing campaigns.

Feature Description

RFM is one of the most widely used methods to perform customer segmentation due to its simplicity and availability. R stands for recency, indicating how up-to-date the customer's last purchase is. This can be regarded as one of the indicators to evaluate the level of customer engagement. Also, the more recent the last transaction is, the more likely an individual will return for repurchase. F represents the total number of purchases in a particular period. Knowing the frequency helps rate the level of customer loyalty. Lastly, M refers to the monetary value, reflecting the total amount an individual has spent with the brand during a given period. Monetary value is the most direct way to weigh up an individual's contribution to the company.

The classic RFM model uses previously mentioned features to provide information about the level of customers' engagement and contribution. However, when considering customers' shopping habits, these three features only manage to paint a cursory picture of customer groups. Since customers may have large variations in their shopping patterns, three additional features were taken into consideration for this clustering model.

Inspired by recency, the Length of the customer journey was introduced. It was derived from calculating the number of days between the last and first transactions. The feature provides further information on the stage of the customer journey an individual is at. For example, a low Monetary and Frequency customer may be classified as low contribution, with the value of Length, further examination can be done to judge if the individual is a newcomer or a low contribution inactive user.

Moreover, average spending per basket and count of unique products id enables marketers to dig more into customers' shopping habits. The combination of these two features makes it easier to know if an individual prefers buying the same item repeatedly to trying new products with few quantities per time.

Feature Dictionary:

A feature dictionary is supplied here to provide brief definitions and explanations about how the features were obtained.

Recency: The count of days since the individual's last transaction. The feature was generated from "purchase_date" of "baskets_sample" dataset, for each unique customer number, counting the days between the last purchase date and the data snapshot date (the date after the end of the data collecting period).

Frequency: The count of times an individual has purchased from the brand in the six months. It was derived from "purchase_date" of "baskets_sample" dataset. For each unique customer number, each unique date of transaction was counted as one purchase.

Monetary: The sum of an individual's transactions amount. It was calculated by summing up "basket_spend" of each customer from "baskets_sample*"* dataset.

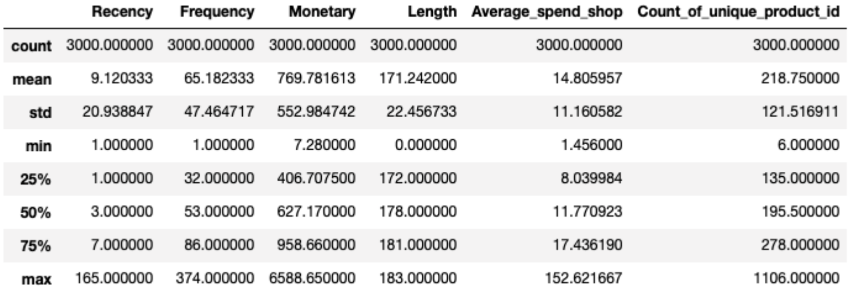
Length: The duration between one's first and last transaction date. In the six months, each date was labeled with a number starting from 1 to 183. The feature was generated by deducting the max date by min date. It was derived from "purchase_date" of baskets_sample dataset.

Average_spend_shop: The average value of one's spend per transaction. The feature was engineered by dividing Monetary by Frequency.

Count_of_unique_product_id: The count of unique product id an individual has bought in six months. It was generated through "product_id" from "lineitems_sample" dataset.

Customer Base Summary

Table 1-1 shows the statistical summary of the six selected features of all the samples collected. The customers demonstrated high frequency, long engagement length, and relatively recent transactions. This could reflect the fact that people tended to have closer relationships with convenience stores because of their convenience and accessibility. Also, people tended to have greater variations in spending (Monetary, Average_spend_shop) and visiting patterns (Recency, Frequency). These could potentially be key indicators when performing group partitioning.



	Recency	Frequency	Monetary	Length	Average_spend_shop	Count_of_unique_product_id
count	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
mean	9.120333	65.182333	769.781613	171.242000	14.805957	218.750000
std	20.938847	47.464717	552.984742	22.456733	11.160582	121.516911
min	1.000000	1.000000	7.280000	0.000000	1.456000	6.000000
25%	1.000000	32.000000	406.707500	172.000000	8.039984	135.000000
50%	3.000000	53.000000	627.170000	178.000000	11.770923	195.500000
75%	7.000000	86.000000	958.660000	181.000000	17.436190	278.000000
max	165.000000	374.000000	6588.650000	183.000000	152.621667	1106.000000

Table 1-1

Figure 1-1 presents the distribution of data in each feature. Except for Length, data points in the rest of the features show a tendency of right-skewed distribution, meaning that there are more large value outliers leading the distribution has a long tail to the right-hand side.

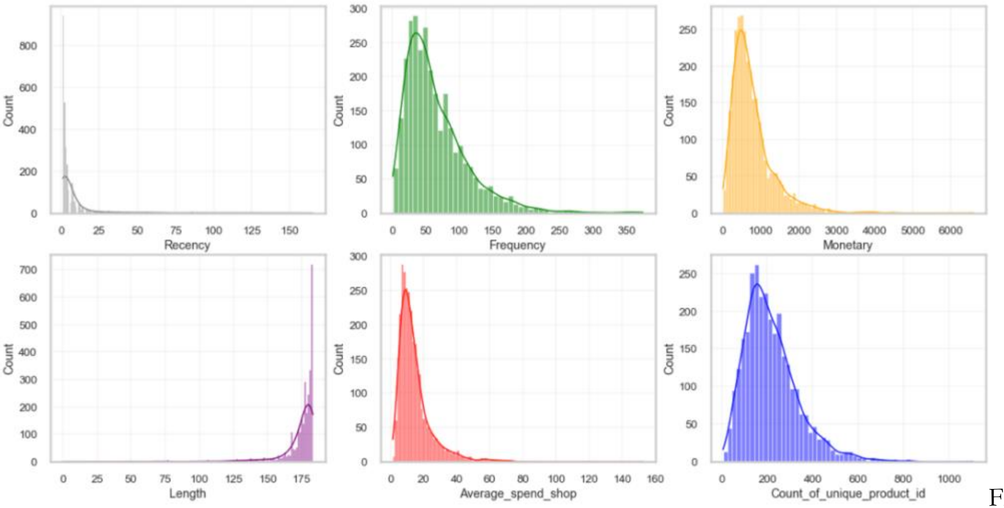


Figure 1-1

Figure 1-2 shows the result of the investigation on the correlation between features. There is a strong correlation between Recency and Length; Monetary and Count_of_unique_product_id. There are also mild to moderate correlations between other features.

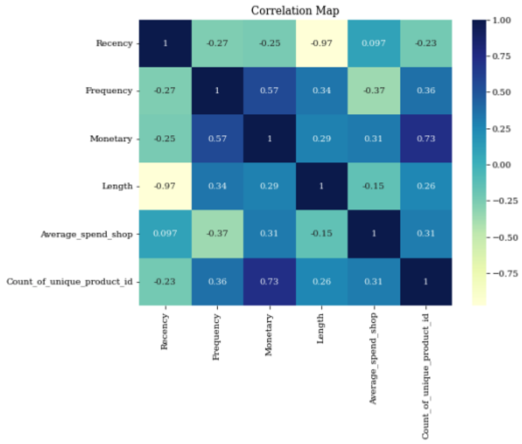


Figure 1-2

Methodology

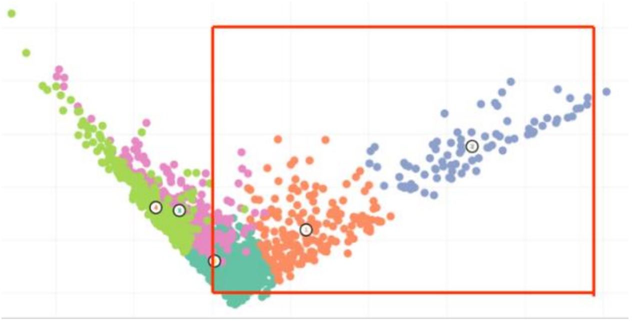
The section focuses on the approaches taken for clustering.

Before applying clustering, standardization was performed to prevent data points with large scales from dominating how clusters were formed. Most of the distributions were right-skewed, with small unit values, standardization was applied instead of normalization. Secondly, PCA was then employed to combat feature correlations. With the introduction of PCA, four components were generated with 95% data variance retained.

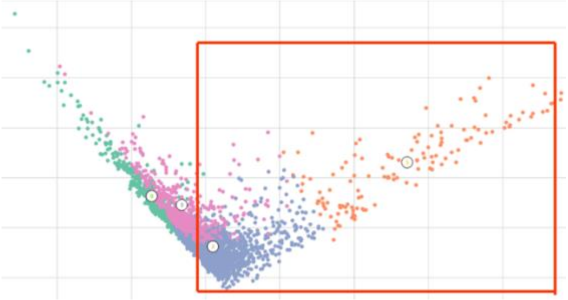
One of the most known and popular clustering methods, K-means was applied for the segmentation task. After executing K-means with different values of groups, ranging from 4 to 7, an index called Silhouette score was employed to find the optimal number of clusters(k). Silhouette Analysis helped judge the homogeneity in clusters and heterogeneity between clusters. With the value of 4 and 5, the score was 0.37 and 0.39, the score started to plunge when the value of k exceeded 5.

The potential value of k was narrowed down to two options, k= 4 or 5. Thus, K-means was executed with these values. Scatterplots were utilized to visualize the distribution of clusters. On a closer look, when k = 5, it gave a more precise partitioning between clusters 0,1 and 2. (area in the red square frame)

Furthermore, considering the final interpretation, more specific customer profiles could be given with five clusters. Therefore, a good value of k would be 5 as it separates the clusters the best.



k = 5



k = 4

Segmentation Result

Five segments were generated after the deployment of the K-means algorithm. In the section, an overall summary is presented to create a general picture of the customer base. A group-based statistical summary is then provided to illustrate the differences between segments. Finally, the pen profile for each group is detailed to create vivid images of the clusters.

Overall Customer base summary

Figure 1-2 indicates the composition of customers. 61% of the sample population belongs to Cluster 0, cluster 4 occupies 17%, followed by cluster 3 with 13% of the population. Clusters 1 and 2 have 6% and 3% respectively.

Figure 1-3 illustrates the sales contribution from each segment. What is worth noticing is that more than half of the sales revenue came from clusters 3 and 4 even though they represented only 30% of the population. On the other hand, Cluster 0 with the most population contributed 44% of the sales revenue.

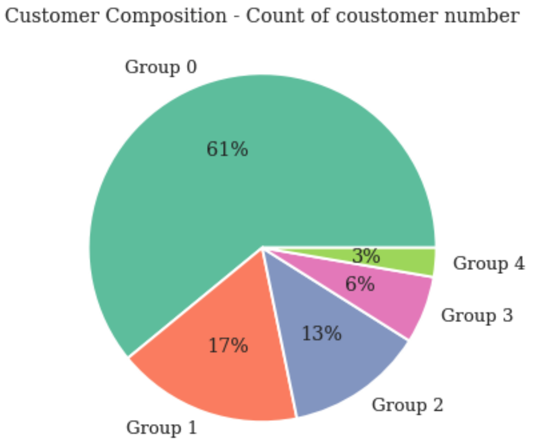


Figure 1-2

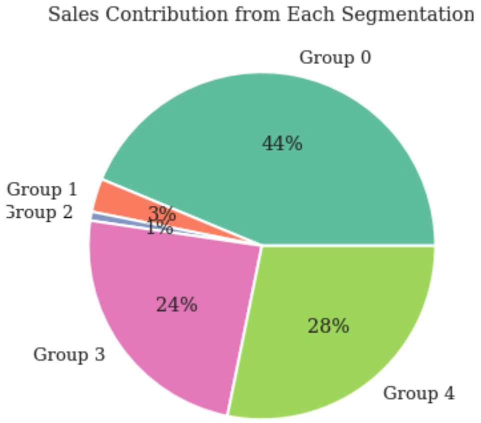


Figure 1-3

Individual Statistical Summaries of Clusters

Box plots of each feature are utilised to illustrate the differences between groups. Besides, a comparison table of the mean value of each group is also provided. Figures 1-4 and Table 1-2 show the analysis result. Bullet-pointed descriptions are given to catalogue the main attributes of each segment.

- Cluster 0:
  - Slightly lower Frequency, a high value of Length, low Recency value.
  - Less Monetary value and average spend per basket.
- Cluster 1:
  - Greater Recency than average, low Frequency but relatively long Length.
  - Low Monetary value, but with a high average spend per basket.
- Cluster 2:
  - Lowest Frequency. Shortest Length. And the highest Recency value.
  - Lowest total monetary value but with a high average spend per basket.
- Cluster 3:
  - Long Length, low Recency. Slightly fewer Frequency compared with average.

    - Highest Monetary Value and average spend per transaction.
    - The most diverse basket content.
- Cluster 4:
    - Highest Frequency, longest Length, and most recent Recency.
    - High Monetary value and count of unique product id.
    - Lowest average spending per transaction.

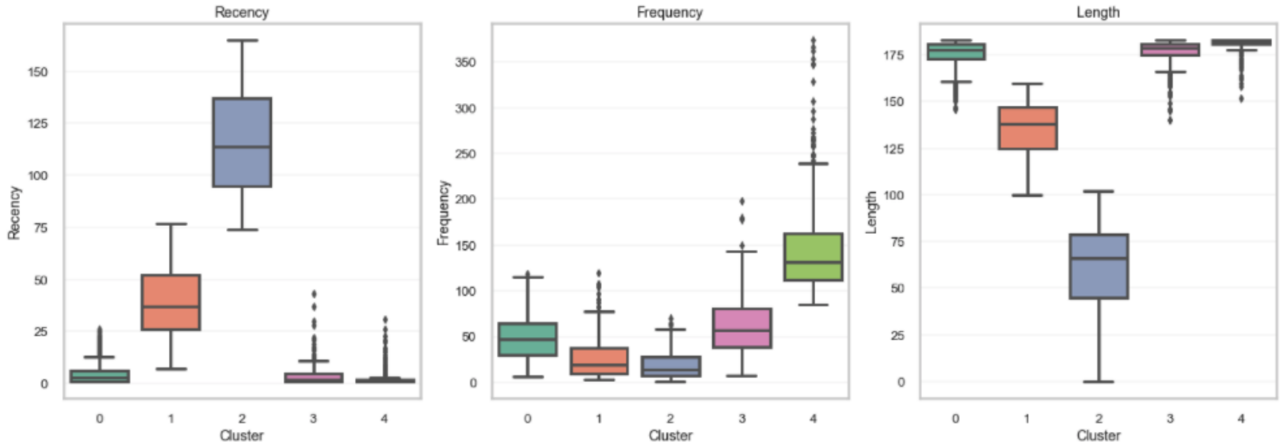


Figure 1-4

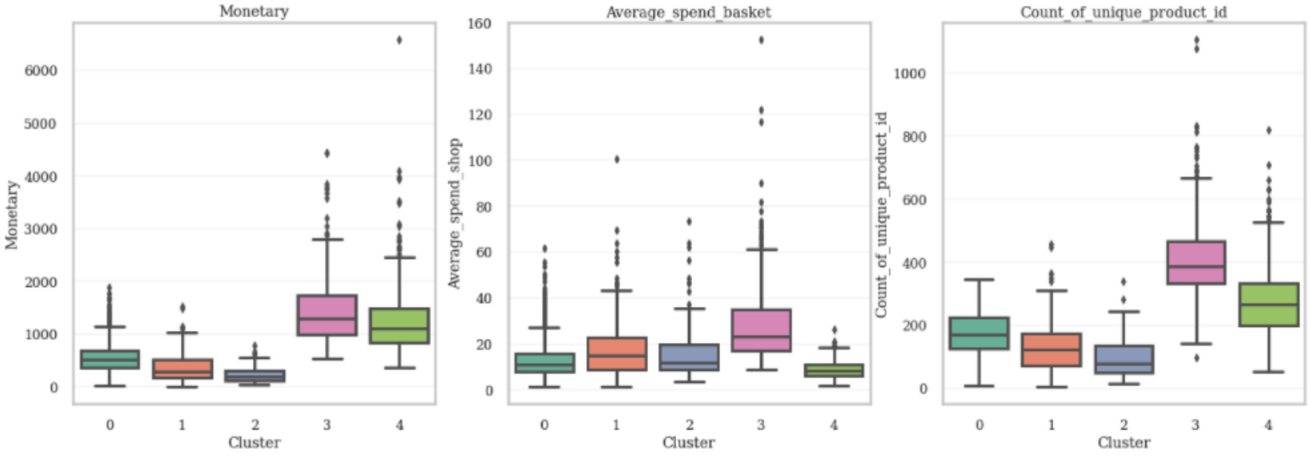
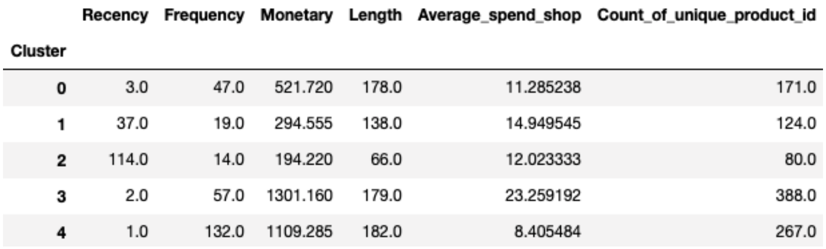


Table 1-2. Median of each cluster



Cluster	Recency	Frequency	Monetary	Length	Average_spend_shop	Count_of_unique_product_id
0	3.0	47.0	521.720	178.0	11.285238	171.0
1	37.0	19.0	294.555	138.0	14.949545	124.0
2	114.0	14.0	194.220	66.0	12.023333	80.0
3	2.0	57.0	1301.160	179.0	23.259192	388.0
4	1.0	132.0	1109.285	182.0	8.405484	267.0

Pen Portraits of Clusters

Group 0 – Low-Contribution loyal

The group is made up of loyal customers who have not spent much in your shops. Overall, they are not as engaged as those top customers, however, they have started their customer journey in your company long since.

They may come for certain exclusive products which are only accessible in your company. Alternatively, they are more motivated to purchase something when there is a discount. The group of customers values price-quality relationships more than others. Thus, it is likely they would research before they shop to obtain the best products for them at an acceptable level of price.

Group 1 – At-Risk Customers

Customers who had their last purchase a while ago and made less frequent transactions and low monetary spending. They have a relatively long engaging period with your company but showing signs of churn. Despite this, they have a high average spending per visit.

The group might have encountered difficulties or negative customer experiences during shopping in your company. As another option, they might find the offers from competitors were more appealing to them, which caused their fewer visits and lower spending in your stores.

Group 2 – Low-Value Lapsed

The segment comprises those who have the least monetary spending and the fewest visit. They have not shown transaction activity for a long time. Besides, they do not demonstrate a strong connection with your company.

Based on the above, the customers in the group may have already exited from the customer base.

Group 3 – High-Contribution loyal

Customers in the segment spend the most among all groups and they have built a long-term relationship with the company. In general, they prefer to make less frequent trips to stores but whenever they shop, they tend to spend much more than average people do.

It is highly possible that they are satisfied with the services and products provided by your company. Your company may be one of their main channels for their big grocery shopping. Furthermore, they have shown the highest willingness to pay and to purchase diverse products from your stores. Products from your company are embedded in their life, they would buy things from your company whenever they have a chance.

Group 4 – High-Contribution Active

The group is composed of highly active loyal customers who pay the most visits to your company and contributed more than average on the sales revenue. Instead of having main grocery shopping, they are in favour of multiple top-up shops.

They enjoy the shopping experience in your company so they would not be bothered to visit the shops every one to two days. Although they tend to spend less per transaction, they are willing to try different products.

This group of customers cares more about efficiency and convenience. Shopping with minimum customer effort, such as time waiting for a check-up, looking for the items they need, etc., would be their priority.

Conclusion & Recommendations

Two groups were selected as the priority of the next marketing campaign considering their potential and influence on the overall business. Also, the proposals for marketing strategy were included.

1. Low-Contribution loyal (Group 0)

Reason: The majority of the customers belong to this segment and act as the base of your business revenue. Since they tend to have lower expenditures, the goal is to raise their spending in stores. With an average of 10% increase in their total spending, it would boost the company revenue growth by 3%.

Recommendations:

- The group of customers cares more about price, to stimulate their spending, promotion with conditions can be used. For example, three items for the price of two, or £5 off for every £100 spent, etc.
- Further analysis of the basket contents will also help identify their needs, and based on that we can reach out to them with more appealing offers. In addition, if it involves providing a discount after a certain amount of purchase, a predictor for the average spending of each customer will enable you to keep your appropriate profit while attracting customers.

2. At-Risk Customers (Group 1)

Reason: Though the sales revenue from the cluster is not high, these customers are still worth drawing for since they have a higher average spend per shop and longer relationships with your firm, which means they are the potential for turning into high-contribution loyal customers in the future. What's more, knowing what makes them tend to visit less frequently can also help spot and address the existing problems in an early stage.

Recommendations:

- If customers' contact is available, send them e-mails/pop-up notifications of new campaigns or provide them with a little number of coupons as an incentive to draw their attention.
- A customer feedback survey is also helpful to get to know more about their dissatisfactions. After knowing that, inform them what improvements have been done to make them feel valued and more willing to return to the stores.