Executive Summary

The customers of national convenience store chain were segmented successfully. To perform this segmentation, analysis was conducted in two separate sections and the results were combined. First, the customers were divided into four categories (loyalty) based quantile cutting their RFM (Recency, Frequency and Monetary) score. The other categorization was done using the behavioural aspect of the customer (Essential spenders, Free Spenders and travellers), These were derived by K means clustering the dimensionally reduced components extracted from categorical spends of the customer. These separate categories can be combined with each other to perform targeted marketing, or can be used separately to perform blanket marketing. The data provided was analysed, pre-processed, engineered and analysed to get the best customer archetypes.

Feature Description section

Data preparation and Feature selection/ Engineering -

Random samples of customer numbers were taken from different tables to check for appropriateness and consistency of data (few inconsistencies between tables were found which were dealt). As per analysis, the total spend of table 'lineitems_sample' and and 'baskets_sample' contained the same value justifying the appropriateness of the total spend. By using the 'categorical_spends' table we tried to find the amount spent on each category by the customer. These values were not in sync with lineitems data (eg-bakery section values missing) and thus required feature engineering. The amount categories in all the tables required to be pre-processed to be usable (removal of non-essential characters like '£' and ','). Total price column was added in the lineitems to further calculate the Monetary column. There were no NaN values present. Purchase time column was converted in the date time format. There were 114 duplicate rows in the lineItems table which were dropped. No columns had very high correlation making it suitable for clustering.

The items which were returned or customers winning the lottery (negative values) were dropped due to it being irrelevant (Returned/refund due to faulty/ damaged item and probability of winning the lottery being low)

Not all the features can be used due to the curse of dimensionality. Modifying the features available to best represent the whole dataset is preferred. RFM analysis is the preferred method for customer segmentation, It provides sufficient data to form customer archetypes with just a few columns. Thus, the features in the dataset were engineered to get the

- Recency- Derived by subtracting the last purchase date from the current selected date.
- Frequency-The sum of unique purchase date and time per customer number made up the frequency column
- Monetary- It's the sum of total cost per visit per customer number.

According to the company's Chief Data Officer suggestions/ final message these feature were considered essential for analysis and was proceeded with to determine the loyalty of customer. Quantile cutting was used on RFM values to give weightage. Lower RFM indicates high priority customer.

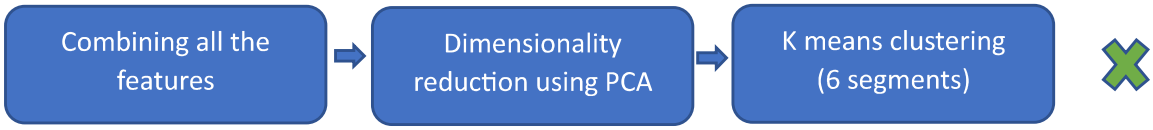
Due to categorical spends table provided had errors and was not consistent with line items it was recreated using SQL queries resulting into data consistency and accurateness. A customer's behavioural / favourable categories of items were found from this recreated data. All the NA and 0 values generated during this process were replaced with value of '1' as the data will undergo log

transformation in later stages. The data was heavily right skewed it needed to be normalised. Log transformation was performed to normalise the data. PCA was performed on the logged data to reduce dimensions. N-dimensions was chosen as 6 because it gave cumulative variance of 0.70. All the customer numbers were assigned a behavioural cluster according to clustering performed and the priority which was created by RFM analysis After finding the segments, it had to be reduced to 7 clusters according to the company's requirements.

Silhouette score decreased as the number of clusters were increased thus cluster size was chosen as 3. Also high number of clusters made the segments senseless with no big differences. It also lead to skewed number of customers in the clusters.

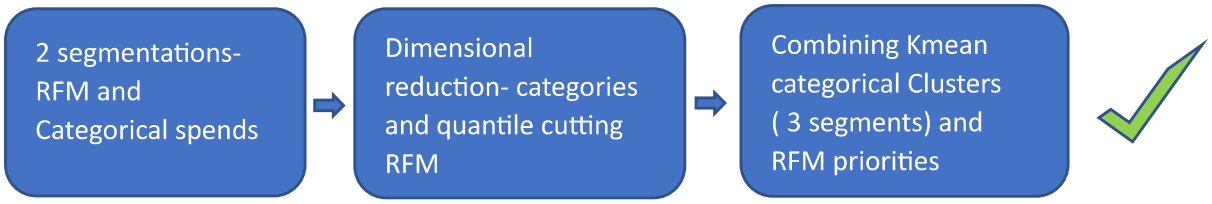
Segmentation Methodology

The tested methods for this use case were



This above method when implemented had a good silhouette score (0.34), but the clusters formed were highly imbalanced in number of customers and had no clear distinction between each other to form customer segments. This would have resulted into wrong segmentation in practise (even though good values) causing negative effect on the store sales.

Hence, different non traditional methodology was used where the datasets were analysed separately and later combined using customer number as key to form distinct segments.



These segments allow two types of marketing to be performed 1. Blanket according to loyalty 2. Targeted according to loyalty and behavioural spends.

The silhouette score is low(0.20)in this case but the clusters formed are meaningful and not skewed. There is a clear distinction between the customer behavioural spends.

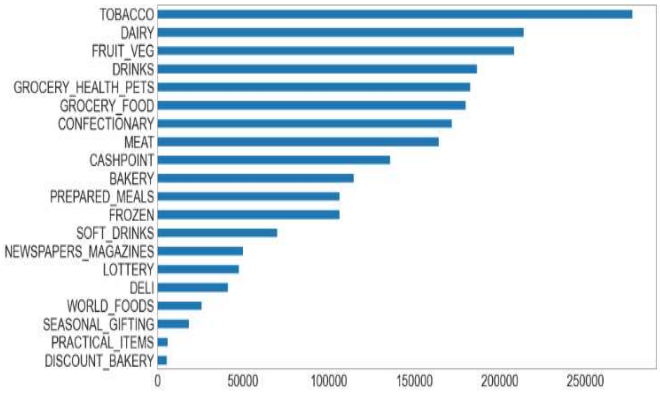
A good value of K was decided to be 3 after trial and error method on different values of K. K value of 3 gave segmentations in which you could see the difference in behaviours of customers. One cluster were high purchasers of essential products like grocery, meat etc, while other category had high purchases of tobacco, drinks, magazines and lottery etc. There was a clear segregation between the customers which is required for profitable marketing. These clusters formed when combined with the loyalty of customers gave the best representation of them in terms of purchasing power, visitations and behaviour. There can be a total of 12 segments formed (4 loyalty * 3 behavioural) but these will be reduced to 6-7 clusters as per the store's request.

They were successfully classified into respective archetypes for marketing purpose. The biggest classification criterion for the customers were the frequency of visits and Total amount of money they spent. Deeper analysis suggested different purchasing behaviours (categorical) by customers having similar RFM scores which was implemented in our methodology.

Results-

Overall Customer base and Statistical Summary of Clusters-

The customer base contained information of 3000 clients for the period of 6 months, the last date of purchase was 31/08/2007.

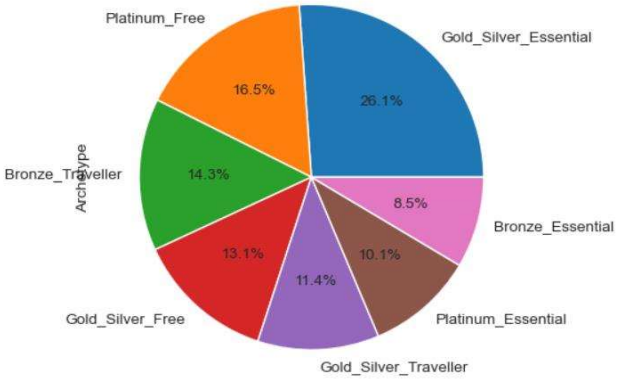


It was seen that customers who buy dairy products also buy Fruit and vegetables, confectionary, grocery food (high correlation of > 0.6).

Tobacco was the highest purchased category in this time period followed by Diary, fruits and vegetables. Discount bakery was the least produced item because of being available only during end of the day and in very limited quantities.

50% of the customers had very visited the store within the last two days. This indicates a very healthy business. On an average a customer has visited the store 65 times in the last 6 months approximately spending £1059 (per customer) which is also a very good indication. Overall the store has a very healthy customer base.

Gold silver essential has the highest number of customers as it makes up an average consumer which comes to the store. Gold and silver classes come under the middle 50% quartile range.



Customer Archetype	Number of Customers
Gold_Silver_Essential	784
Platinum_Free	495
Bronze_Traveller	428
Gold_Silver_Free	392
Gold_Silver_Traveller	342
Platinum_Essential	303
Bronze_Essential	256

Behavioural Statistics-

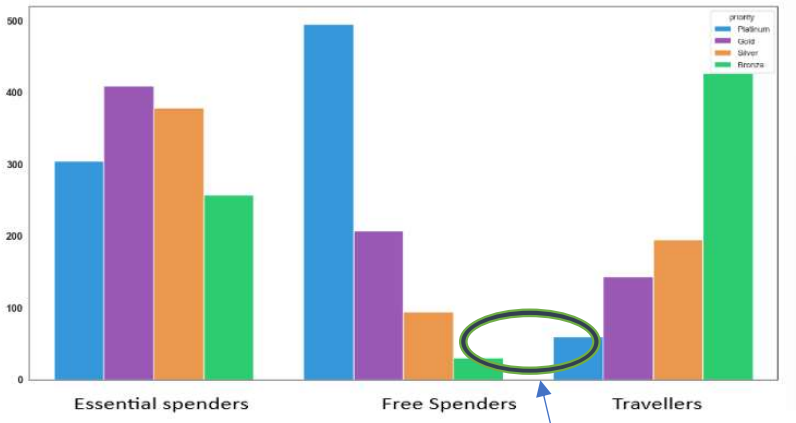
- Free spenders purchased 116 times more tobacco than Essential spenders and 10 times more than travellers, 3 times lottery spend than other segments 2 times more newspapers, magazines, soft drinks, drinks, 10 times more cashpoint usage than essential spenders and 4 times more than travellers
- Top spenders in Traveller category only spent £30 on food groceries whereas free spenders and essential purchasers spent approximately £100
- Customers who visited at end of the day were able to grab items from discount bakery.

Platinum category was assigned to top 0.25 percentile of the RFM values making the customers in this section the most important for profits of company. Similarly, Bronze was assigned to the bottom 0.25 percentile.

Loyalty	Mean Spend	Mean frequency	Mean Recency
Platinum	£1768	115 visits	0.46 days
Gold	£1089	65 visits	2.55 days
Silver	£792	43 visits	6.26 days
Bronze	£423	25 visits	24.99 days

Travellers had less overall interaction with all categories compared to Free Spenders and Essential spenders but had similar seasonal gifting, newspaper and magazine, cashpoint usage, lottery purchases.

Essential Spenders have high purchases in -Groceries, fruits vegetables and less in bakery, prepared meals, gifting, magazines, soft drinks, drinks, cashpoint, tobacco.



As per the graph, It is noted that there are high number of platinum loyalty people in the free spenders and less in travellers category. The essential spending category is very balanced mainly due to an average customer requiring essential products.

Bronze free Spenders and Platinum Travellers had very few customers so they were combined with Gold Silver Free spenders to reduce the number of archetypes to 7.

Customers have certain pattern in their visitation according to time of the day and and day of the week. It is seen that customers prefer to go shopping on Fridays and the most footfall is seen during non-working hours.

As per the recency frequency and monetary graph in appendix it is seen that most of the customers having recency >25 are categorised as bronze except few who had really high frequency before partial churning. Recency vs frequency data is skewed

Monetary vs frequency graph shows that most of the customers have spend between 0 and £5,000 and frequency between 0-200 except a few outliers which have very high monetary and frequency. These outliers were not removed from the data as data is true.

Monetary vs Recency is also very skewed and most of the customer base has have a recent visit.

For the dimensions created by PCA which was input for the clusters- Grocery food, tobacco and Dairy makes up for the maximum variance, dimension 1 is balanced.

Pen portraits of customers-

The customer base was divided into 7 customer archetypes on recommendation by the company. The pen portraits of these customers are-

Platinum Free Spenders- The category has the highest loyalty ie- the most recent visits, regular visitation and a high spending capacity. They will be more susceptible to trying on new products offered by the company due to their spending habits in multiple categories. It is observed they have a exponentially high affinity towards Tobacco products compared to other groups. They also have a high indulgement/ non essential purchasing compared to other segments.

Platinum Essential Spenders- The customers in these categories also have high loyalty to the store, but they are limited to certain categories which are essential. Customers in this segment are probably family spenders and wouldn't explore different categories. The customers will be very interested in the trying out different essential products if on a sale.

Gold Silver Essential Spenders- They are regular family spenders and make up most of the client base. They only buy essential products which are required. Occasional discounts can be provided for retention.

Gold Silver Free Spender- These customers have a decent loyalty to the store or have a slightly lower RFM score. They try different categories of products and will be willing to spend more if found the right deal. They can be easily converted to Platinum loyalty by giving regular discounts for a time period. They are more important clients because of their exhibited behaviours displays a spending capacity

- Platinum Travellers- These Customers had just a few visits recently of significant cost, they are potential platinum clients but the data isn't enough to classify them as platinum clients yet. Hence, they are merged with Gold Silver free class.
- Bronze Free- These customers have less data on, They are potential bronze clients but have spendings on indulgents too, so can be a gold_silver client. As the number of customers on these segments are limited they are merged with Gold silver free class

Gold Silver travellers- They are new customers with little data, but can develop into a customer as time permits. Marketing on them can be proved to be useful. They have a low spend throughout all categories due to less visitations.

Bronze Essential Spenders- They purchase only essential products and have had enough visits. They spend less and only on essentials probably due to financial instability. They are not a good marketing segment

Bronze Travellers- They have less overall spending in the store, but according to their spending habits it can be that they are purchasing products from different stores. They are an ideal customer base for marketing and may have some potential platinum level customers. They are on the verge of churning and should be given some promotional offers.

Summary-

The customer base can be divided to four priority segments and three behavioral segments which can be combined with each other for marketing. But, for our particular case as the company requires 5-7 customer archetypes few categories are combined which will be based on similarity.

The two most important segments for our business is the **Bronze Travellers** and the **Platinum Free Spenders.**

Bronze Travellers- They have less spending in store, sometimes in a particular section, This may be due to unavailability of certain items in other store. Some customer's spending habits suggest potential premium clients. They should be heavily marketed to and given discounts for possible conversion. They are also on the verge of churning and that should be avoided.

Platinum Free Spenders-

'It is often said that 20% of a customer base accounts for 80% of company's profits'

~ Chief Data Officer

They are the top most clients of the store and are willing to try new introduced products. They should be given loyalty discounts to encourage even more spending and encourage retention. They have the spending capacity and the visiting time.

Insights and recommendations-

- In general, all the platinum customers should be introduced to the loyalty discounts to avoid the chance of sudden churn. Loyalty programme will have a threshold spend above which you get a particular discount.
- The Gold and Silver free spenders should be introduced to discounts on new products as they will be willing to try it.
- Gold and Silver essential spenders should be given basic low value discounts to have customer interaction, as they make up for most of the customer base.
- Bronze essential spenders can be notified about the discount of soon to be expiring product to help them with managing finance and making the customer loyal.

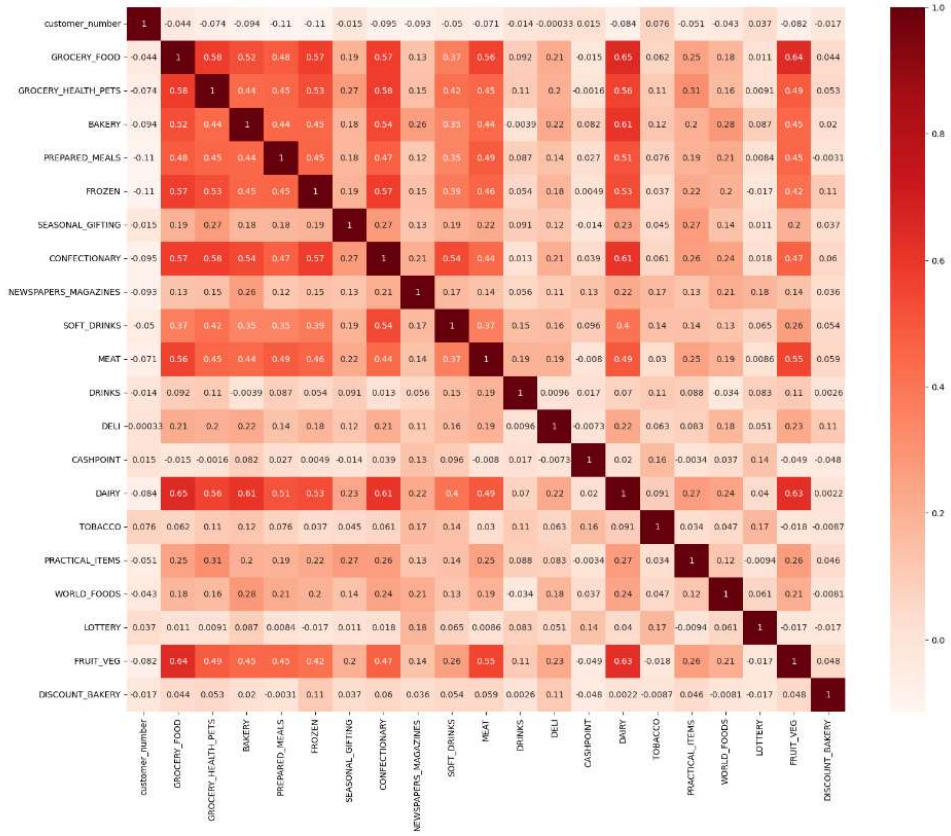
Data recommendations-

- Geographical information can help understand the climate and specific categories which will increase in sales during fixed time of day and year.
- Survey based discounts- For a short survey the clients can get a few pounds off on their purchase. This will help company understand their client base and take better data driven decisions.
- The number of clusters were reduced to match the company's segment criteria. More segments can lead to better personalised marketing.

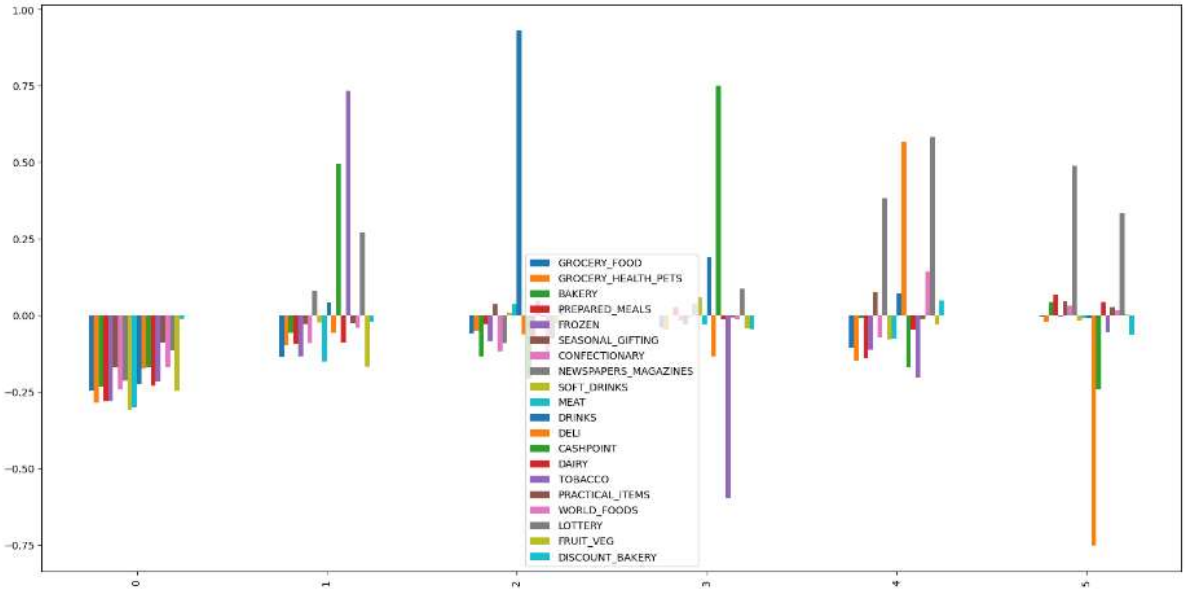
For detailed interpretation please look at the python notebook attached with this submission. It contains more details and visualizations. Also, please find attached the final customer segmentation csv which can be used for marketing.

APPENDIX

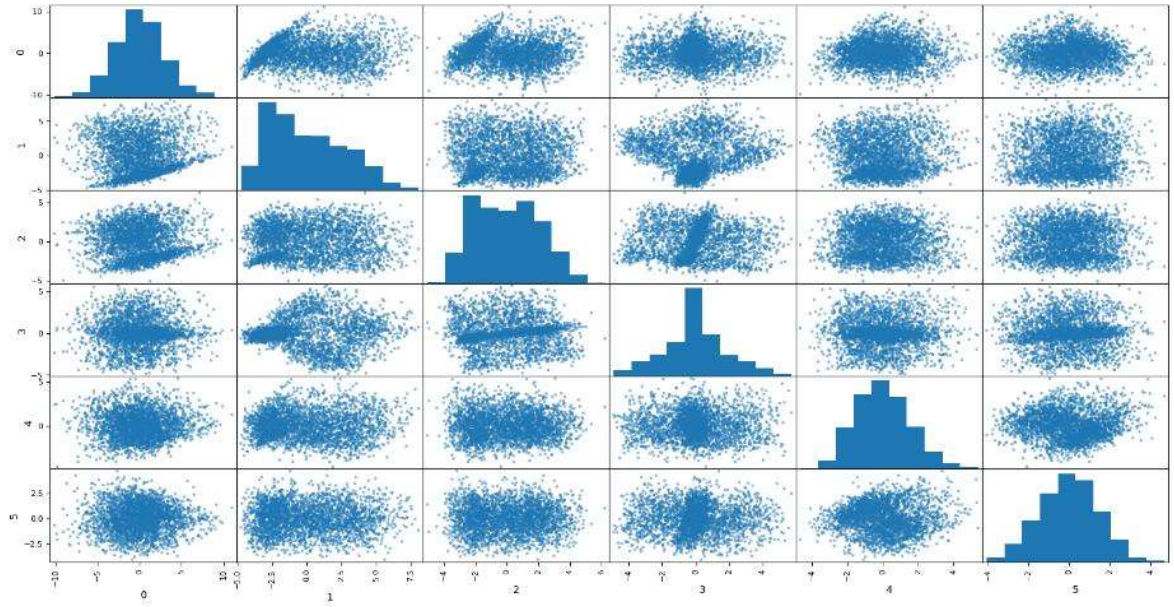
Correlation Graph- Correlation between features



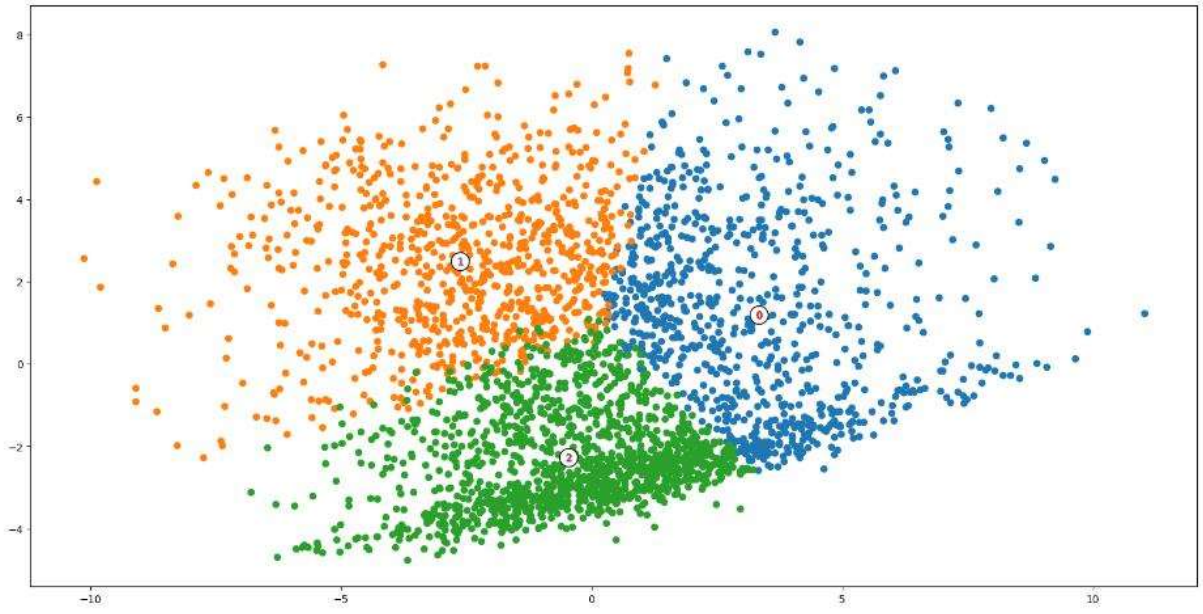
PCA dimensions- Distribution of features in in components used for clustering



Scatterplot after data conversion (log transformation)-



Formed clusters



Recency vs frequency vs Monetary graph