Customer Clustering and Product Recommendation

By Team Sherlock

Introduction

Our presentation contains the idea behind our working, the framework we conceptualised and the final clusters formed using the data provided to us. Using this, we constructed our recommendation system which heavily depends upon the clustering. We first deal with the clustering problem and then the recommendation one.

Since the data was very big and the names and product descriptions were also very diverse which needed manual intervention, we generalised into some very specific demographics only.

Understanding the problem

Customer clustering is the process of classifying customers into distinct groups based on the similarities they share with respect to any characteristics deemed relevant to the business.

Key components in developing proper, actionable segmentation-

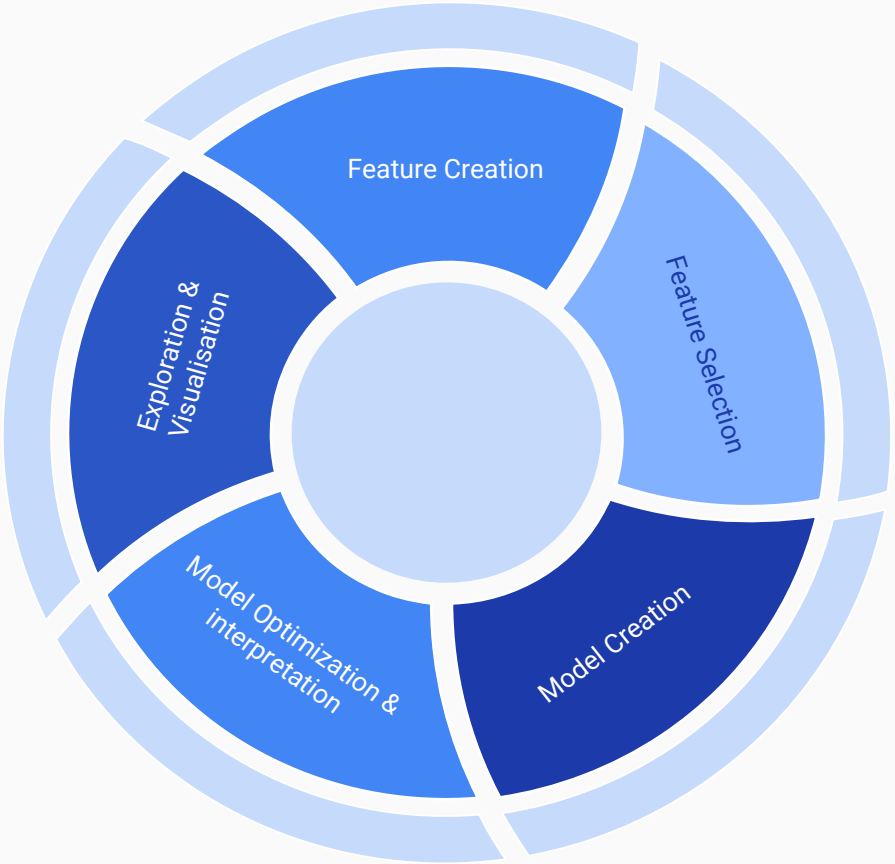
- **Understand business needs and objectives**

Customer satisfaction, Share of wallet, Market share, Loyalty, Promotions

- **Available customer information**

Demographical, Geographical, Spending pattern and frequency, Product description, Tender and Discount types

Customer Clustering


Feature Creation
Feature Selection
Model Creation
Model Optimization & interpretation
Exploration & Visualisation

Workflow

Approach

- Clustering using Frequency and Monetary value (both price after promo and price after discount)
- Verification using age demographics and spending frequencies by month and day of week
- Clustering using Recency(Months elasped since the Last Transaction)
- Clustering using payment using Credit Card(salaried) or payment using Cash or none
- Clustering using Discount used
- Clustering using religion, age and whether the family has a baby or not

Reasons behind the following steps

- Clustering using Frequency and Monetary value helps us understand who are high potential, mid potential and low potential households and who are more regular and more seasonal among the high and mid potential ones
- Verification using age demographics and spending frequencies by month and day of week helps realise how they spend by the weeks and months and helps understand if a person is self employed or has a employer
- Clustering using Recency helps understand who are the old and the new customers among households with lesser number of visits
- Clustering using the next 2 criteria is mainly for low frequency low recency based customers

Reasons behind the following steps

- Clustering using payment using Credit Card or Cash or none helps understand if a household has a salaried person or not, as for holding a credit card, one needs to be salaried or have a salaried person in the family and more use of cash helps us understand the tendency of the customer, whether he is inclined towards future purchases or not
- Clustering using Discount used helps understand whether the person is actively involved in the loyalty based rewards and payback programs or not
- Clustering using religion, age and whether the family has a baby or not is mainly for mid potential customers to understand why they shop seasonally and also how can they be attracted

Algorithm used

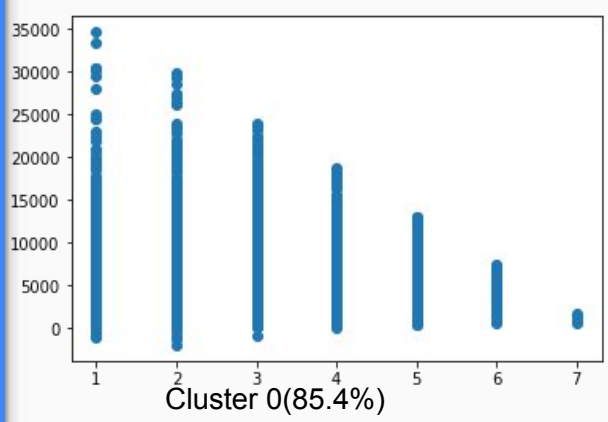
- Clustering algorithm used is k-means algorithm
- K-means algorithm works on the distance metric, and hence can only work using continuous variables
- Since it won’t work for categorical variables, we had to one-hot encode them
- Also, we needed to make sure that the feature set doesn’t increase a lot which might cause the ‘curse of dimensionality’ and prevent us to understand the actual significance of clusters for business purposes

Exploration and clustering of store 3692

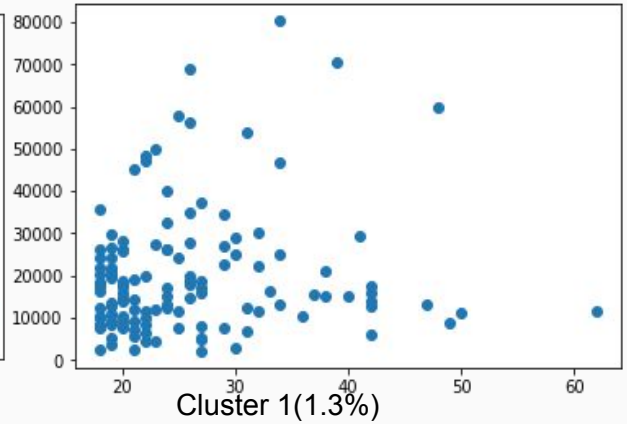
- This store is located in Ludhiana-Ferozepur, Punjab
- It has around 9.7k distinct customers.
- We follow the above approach to create 20 clusters of store 3692

Clustering using FM analysis

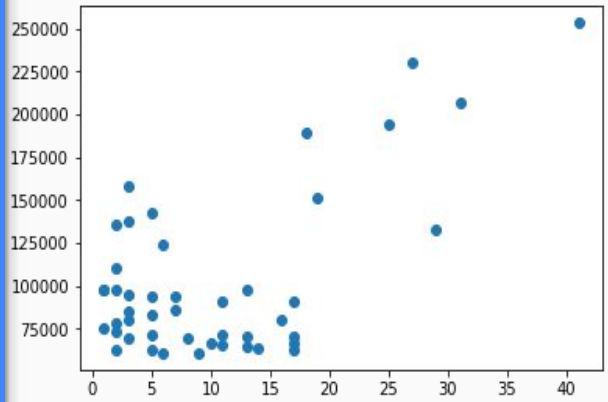
We create new variables frequency, recency and monetary value using the given data. Since we find some customers missing from tenders csv, we use the prices after promo in such cases and hence find the statistical outcome(mean, median, mode) which is almost similar, hence we can comment that their distributions are almost same, hence using any one as indicator for monetary value would work. After clustering, we get the following frequency-monetary value distribution.



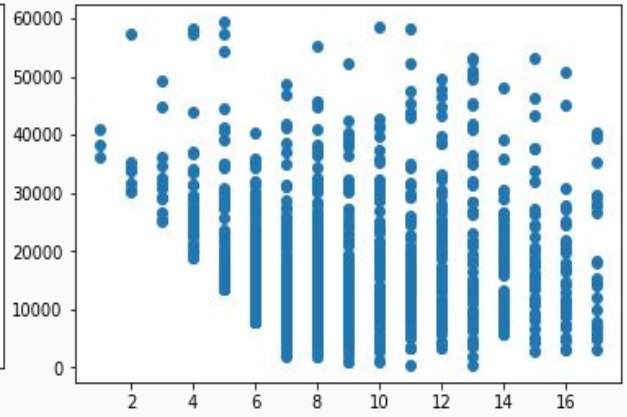
Cluster 0(85.4%)



Cluster 1(1.3%)



Cluster 2(0.47%)



Cluster 3(12.76%)

Discussion after FM clustering

So now we have 4 clusters, where-

- Cluster 0 has people spending in the bracket 0-35k and frequencies less than 7. The mean frequency is almost close to 2, which means these are not frequent buyers. Their monetary value is less as well, with mean around 4k. However they make up 85% of the customers, which is in line with Pareto's principle. However knowing this demographic will enable us differentiate between a frequent customer and a non-frequent one.
- Cluster 3 has people spending in the bracket 0-60k and frequencies upto 16. With mean frequency around 9 and monetary value of 17k, they are one signifiant bunch of customers.

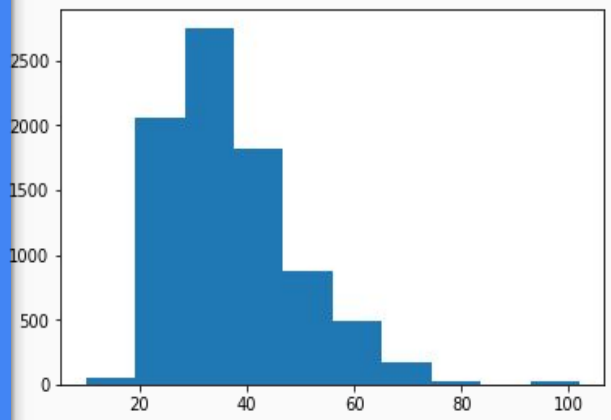
Discussion after FM clustering

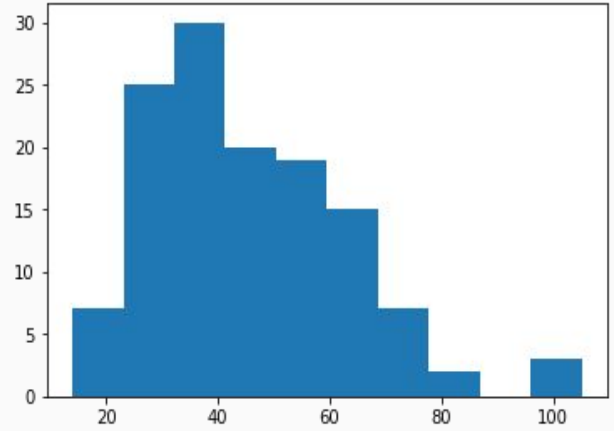
- Cluster 1 are even more frequent buyers, with monetary values having a mean of 20k, spending ranges of 0-80k and frequency mean of 25 with 75% of customers visiting with frequencies of 29.
- Final cluster, Cluster 2 are the high spending bracket customers, with monetary values having mean above a lakh and frequencies around 9-16.
- What clusters 1,2 and 3 mean for FG - Cluster 2 and 3 visit BB store maybe around once every 3 months, while cluster 1 seems to be a monthly visitor. Now it needs to be seen when people of these clusters visit usually.
- One more thing, unlike Pareto principle, the price after promo sum and price after discount sum for cluster 0 are over 50%, that is cluster 0 contributes almost half of the sales in store code 3692.

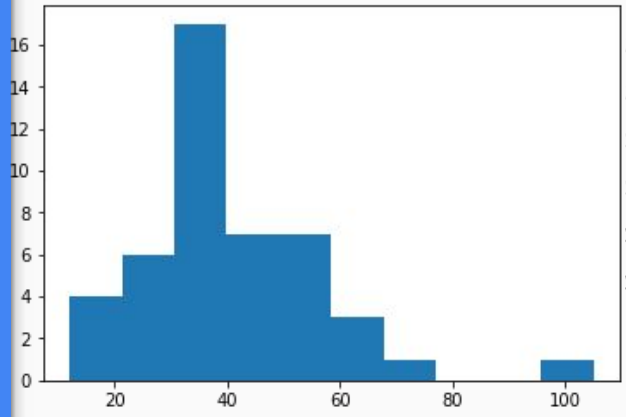
Verification using age demographics and spending frequencies

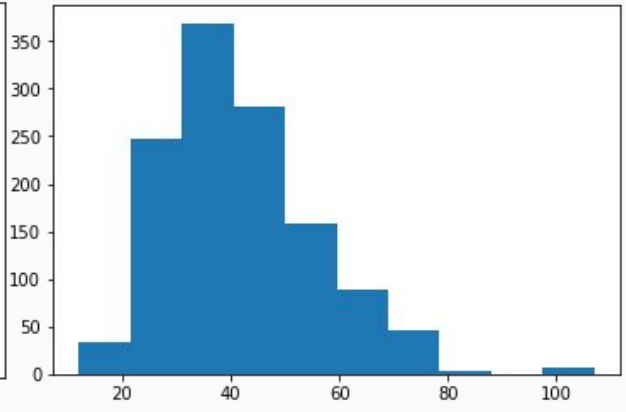
For us, cluster 1 are the loyal ones who seem to come mainly for groceries(will check in a bit). Cluster 2 and 3 needs to be targetted for more frequent visits as they are seemingly regular but we would love to have them in cluster 1. Also cluster 1 should be harvested for higher profits. Understanding cluster 0 will enable us to recognise the low loyalty people who visit occasionally and have monetary value less as well.

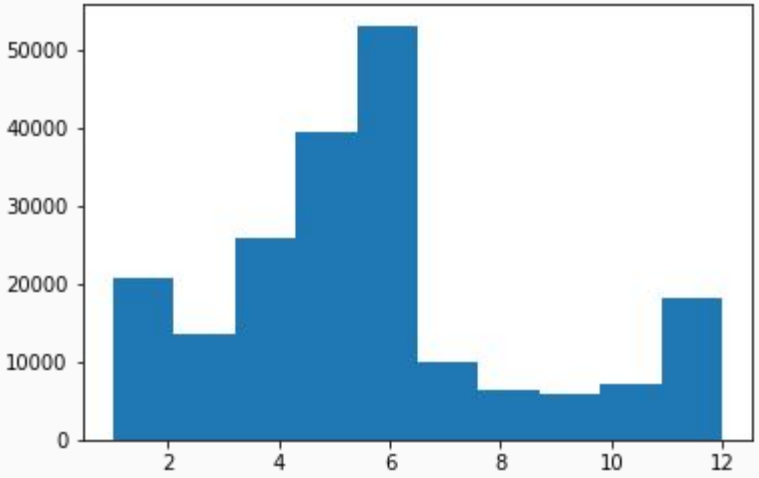
Here are the histograms of clusters using age.

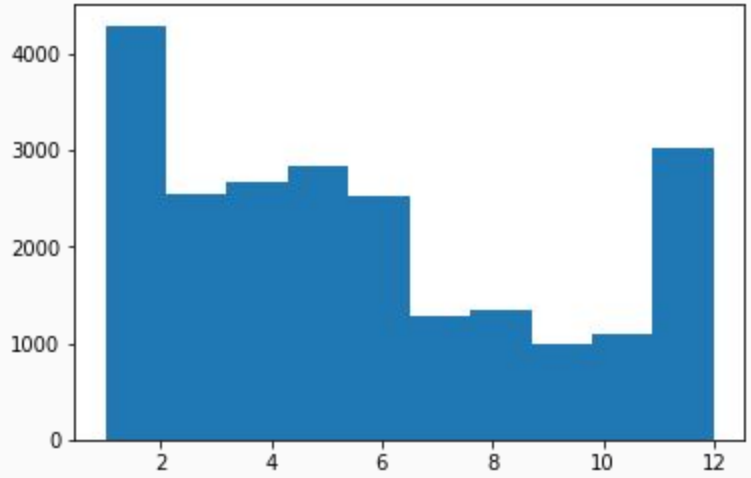


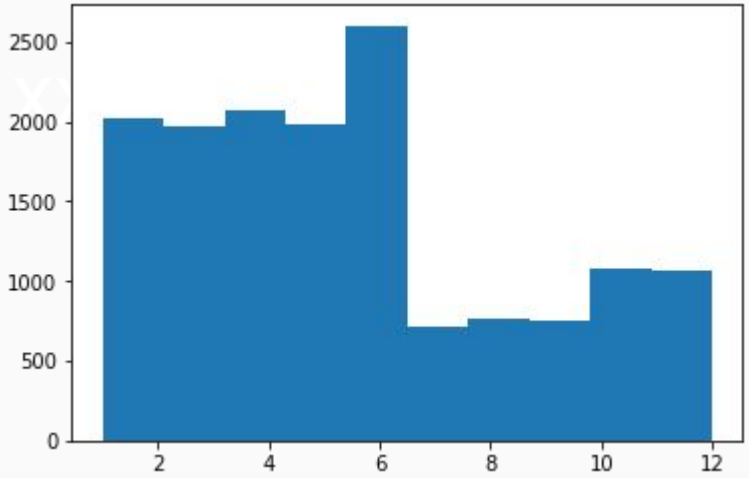


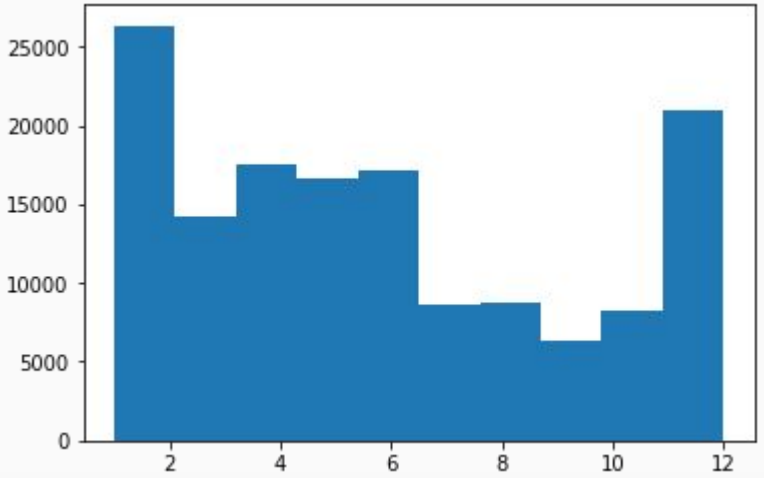




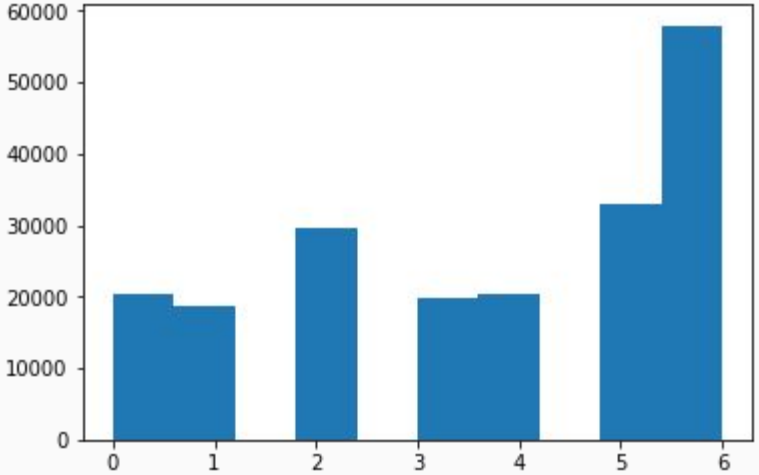


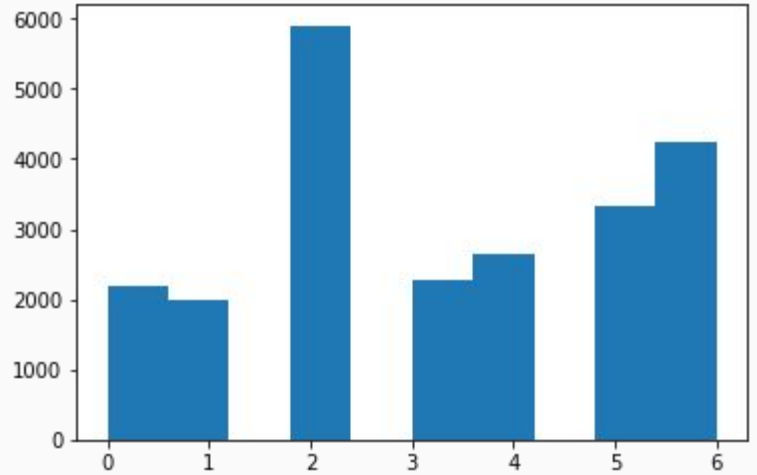


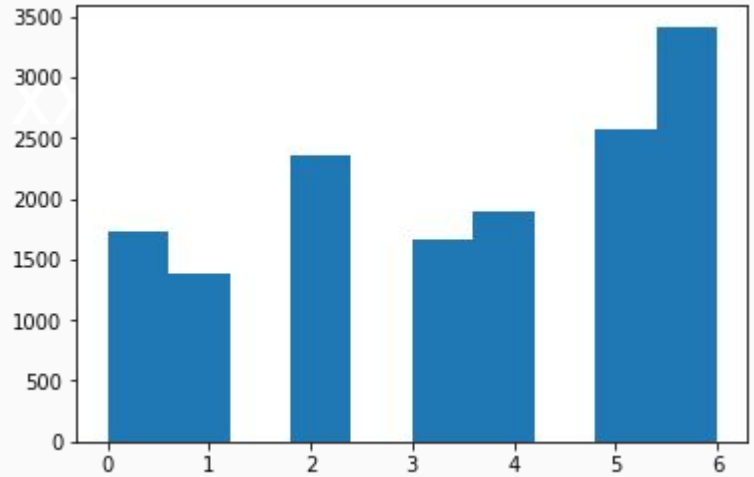


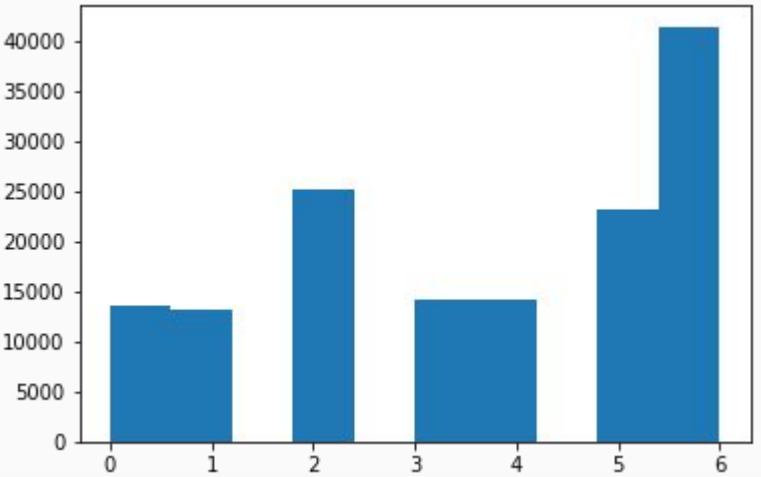


Histograms of clusters using shopping frequency by month









Histograms of clusters using shopping frequency by day of week

Discussion after verification with age

For cluster 2, it is clearly dominated with people in age group 30-40, that means they might be families mainly(since spending high, so probably married with one or more salaried persons). For the other three clusters, there are significant portions of people in age group 20-30 and 40-60. Studying them will give some more insights about the customer behaviour. Since cluster 1 has only 128 people, with decent spending amounts, these can be attributed to again single or married households(salaried as well). Same can be said for cluster 3 with 1236 members mostly in the range 20-60. So these seem to be salaried people. Cluster 0 needs to be explored some more.

Discussion after verification with spending frequency by month

There is a dip in purchase in the second half of the year, but for clusters 1 and 3, there has been a surge in purchase in Jan and December. For cluster 0, there has been a slight surge in Jan and Dec but more surge in June and July. For cluster 2, there has been a consistent purchase in first months with surge in June. So in the perspective of business, we can say-

1. Cluster 0 with surge in June and July.
2. Cluster 2 with consistent first half, surge in July and then slight surge in December.
3. Clusters 1 and 3 are similar, with surge in Jan and Dec.

Discussion after verification with spending frequency by day of week

Wednesday being a day with offers has in general higher sales than the rest of the weekdays, but cluster 1 shows unexpectedly high sales on that day. Maybe that explains why with even high frequencies, they have lower monetary value.
In general sales on weekends is higher than weekdays which is justified for cluster 1 and 3 as they have more salaried people as argued above. Cluster 0 needs more analysis.
Now we work using cluster 0 only. We find that tender used isn't that understandable in business terms so we avoided that and instead looked into the payment used. If credit card used, that means person or his family members are employed.

Clustering using Recency

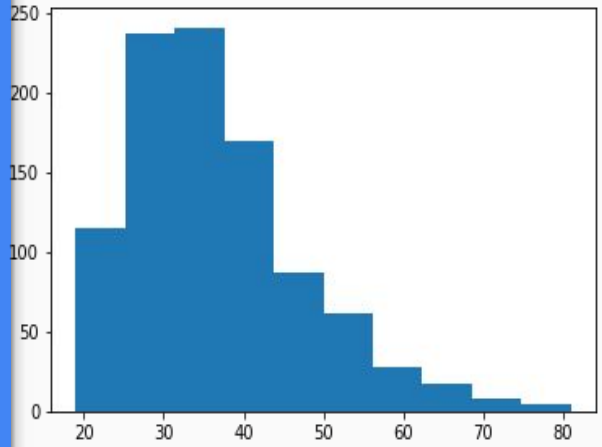
- First we create features corresponding to whether a person used credit card or cash for payment and how many times they used. Then we created features corresponding to discounting used like payback, bbprofit club, bbsavings club, t24 club, fg shopping festa and elders club. Then we cluster using recency variable first.
- Now one of the clusters of the two has mainly those people who have visited back a long time ago so they aren't of our interest. The ones in our business interest lies in the other cluster as they have visited recently.

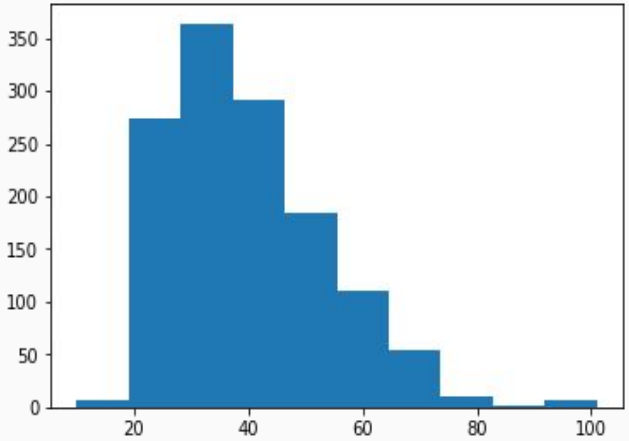
Clustering using credit card used or cash used or none of them

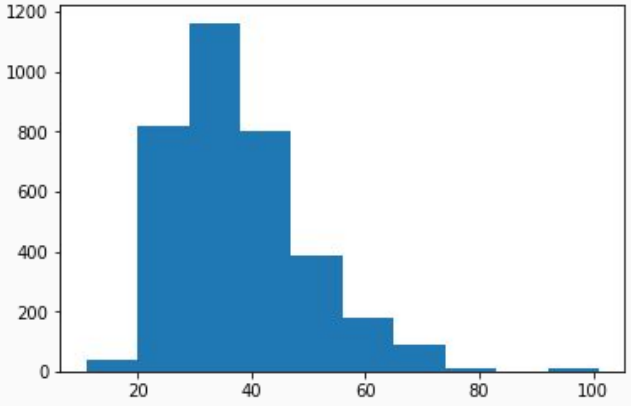
Now we deal with second clustering using payment mode used. Since cash and credit cards are somewhat different as the ones with credit card probably have a sound earning for sure, we cluster using these to check the results.

Clustering them into 3 parts gave the best silhouette score.

These are the age demographics of the 3 clusters.







Discussion after clustering using whether credit card or cash used

Cluster 0 is credit card dominant and cluster 1 is cash dominant. Cluster 0 has a slightly more younger generation than cluster 1. Monetarily, cluster 2 seems the one with lower number of salaried people. Cluster 0 has slightly higher monetary value than 1, with both having significantly higher value than cluster 2. So we focus on cluster 0 and 1, who need to be further focussed upon for better future sales. Cluster 2 seems to have got some offers which they have encashed and seem unlikely to be among the households that should be focussed upon.

Clustering using discount used

Since the number of clusters is slowly increasing, we now focus to cluster only the cluster 0 from last clustering. Cluster 1 will show some results that can be argued as below.

	Cluster 0	Cluster 1	Cluster 2
Payback	27	24	53
BBProfitClub	0	65	2
BBSavingsClub	18	5	24
T24Club	0	3	49
FGShoppingFest	0	0	0
EldersClub	0	0	0

Clustering using discount used(contd.)

Thus, from here we can see that cluster 0 can be attracted by providing discounting options via BBSavings club offers, while cluster 1 can be attracted using BBProfit club offers and cluster 2 can be attracted using BBSavings and T24 club offers. Paybacks are high for all three, with almost double for cluster 2. Payback points based discounting is something that directly reflects loyalty of customers. So cluster 2 can be said to be more loyal, or will be more loyal in future so targeting them is of utmost importance. Other 2 clusters have significant loyalty as well. Payback points are associated with cards as well, so targeting such customers with offers directly related to the cards they use to pay can be beneficial.

Clustering using discount used(contd.)

"Members of the Big Bazaar PROFIT CLUB will also receive the benefit of existing offers on their Payback cards and T24 mobile services." - It can be used to promote BBProfit club services to cluster 2.

"The Big Bazaar PROFIT CLUB Card can be used as a Gift Card for your friends and family, like children living away from home or parents residing in other cities etc." - Maybe that is why is has been used less frequently.

"The Big Bazaar PROFIT CLUB Card can also serves as an excellent Employee Incentive program for your employees . It can also be extended to your business partners as a gift." - Can be harnessed for Cluster 1

Clustering using religion, age and whether the family has a baby or not

- Finally we analyse cluster 3, a cluster having mid potential households and cluster then using religion, age and whether the family has a kid/baby or not
- Product descriptions were too varied and too many in numbers to study any other product category that well
- Finding religion manually wud have been very tough, so we scraped the data using Ethnea(http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py)
- Also distinguishing Hindus from Sikhs and Christians from Jews using the above, so we clubbed up (Hindus,Sikhs) and (Christians,Jews)
- All in all, we had 3 religions(above ones and Muslims), age and has a kid or not to cluster. Creating 11 clusters gave the best silhouette score.

	Freq	MVP	MVD	R	Age	Kid	H/S	C/J	M
0	9	17k	16k	8	29	1	1	0	0
1	8.7	20k	19k	8.5	61	1	1	0	0
2	8.8	12k	10k	8.4	27	0	~1	0	0.005
3	9	20k	19k	8.3	50	0	0	~1	0.02
4	9.4	16k	15k	8.5	60	0	~1	0	0.006
5	8.8	17k	15k	8.5	42	0	~1	0	0.003
6	9.25	21k	20k	8.4	41	1	~1	0	0.012
7	9	15k	12k	8.6	31.25	0	0	1	0
8	8	20k	18k	8	36	1	0	~1	0.017
9	9.125	15k	13k	8.8	100	0.375	0.875	0.125	0
10	10	17k	15k	8	68	0.333	0	1	0

Cluster description using mean

Discussion after clustering using religion, age and whether family has a baby or not

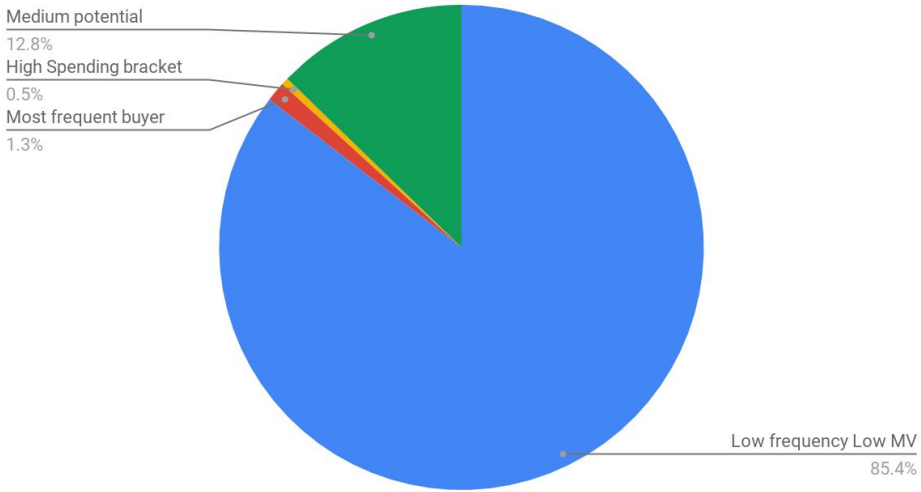
- Cluster 0 is exclusively Hindu/Sikh religion based cluster having pretty young age. Most probably they have a small child and they are newly married.
- Cluster 1 is also exclusively Hindu/Sikh religion based cluster but have old age Most probably the kid is their grandson/granddaughter.
- Cluster 2 has some Muslim population but the whole cluster seem to have no child, hence might be married or not, since they have young age.
- Cluster 3 has a dominant Christian/Jew population with some Muslims with no kids and age mean around 50. This again seems natural, the person might be married or not, if married must have kids in their adolescence.
- Cluster 4 has a dominant Hindu/Sikh with some Muslims, of mean age near 60, means they are around their retirement years, with no grandkids.

Discussion after clustering using religion, age and whether family has a baby or not

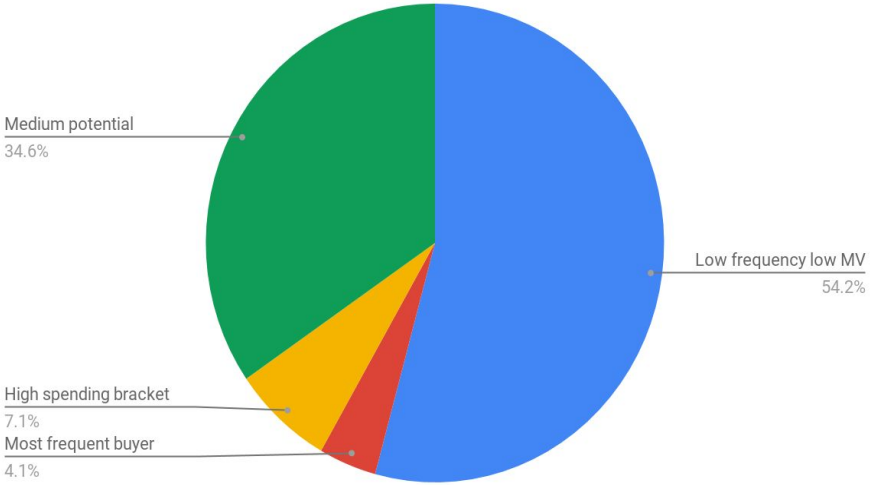
- Cluster 5 has a dominant Hindu/Sikh with some Muslim population in their 40s with no kids. Probably either they are unmarried or married by their late 20s or do not have second kids.
- Cluster 6 has similar demographics as 5 but have kids, so most probably they are married in late 30s or have second kids.
- Cluster 7 has Christian/Jew dominated population with no kids and age around early 30s. Can be married/unmarried.
- Cluster 8 is also Christian dominated with some Muslims, are in their mid 30s and have kids. They are married and might have their first or second kid.
- Last 2 clusters have mostly old aged people from Christian/Jew & Hindu/Sikh community with 1/3rd probability of having kids, so most probably will be grandkids.

Final clustering of all customers under store 3692

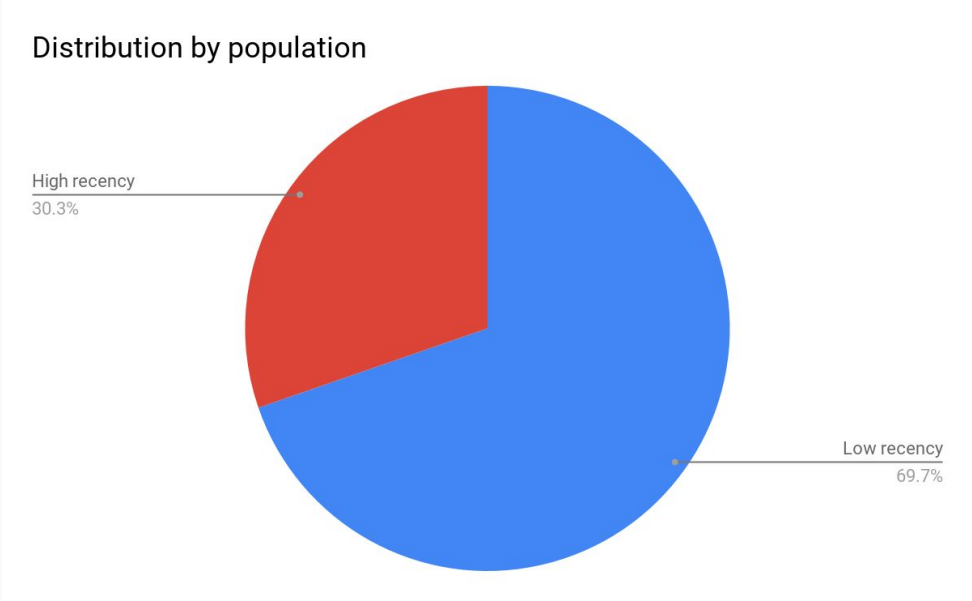
Distribution by population



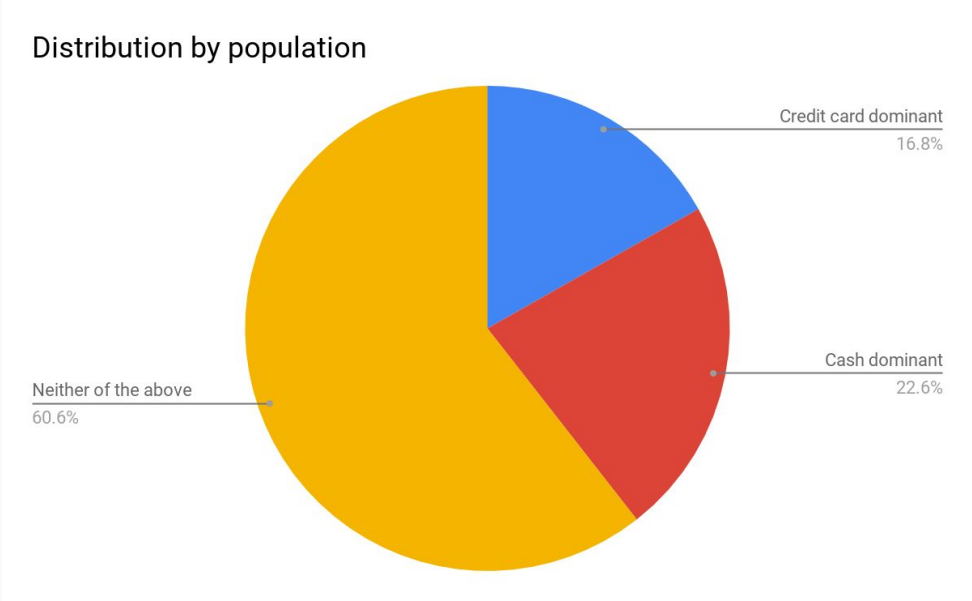
Distribution by earnings



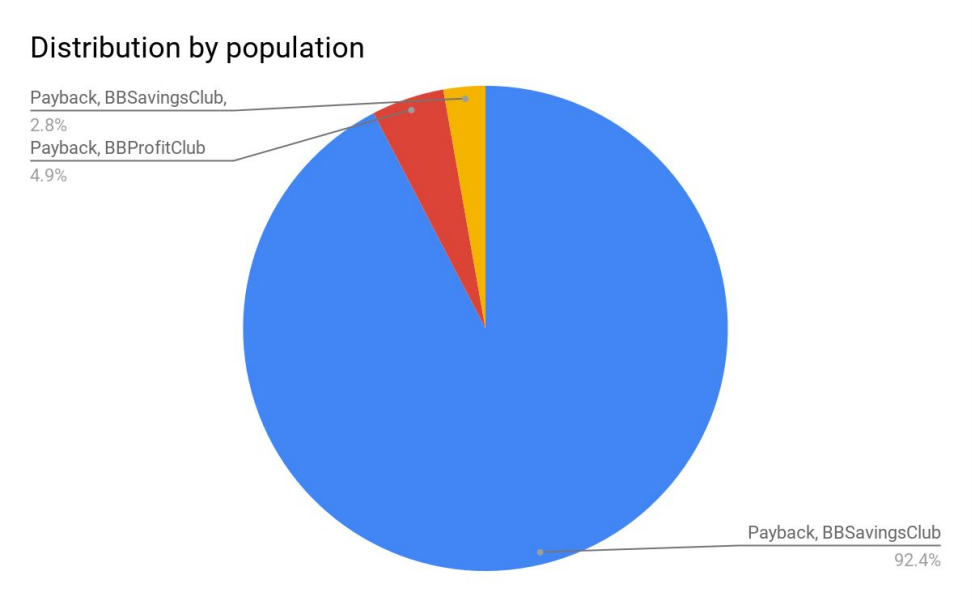
Final clustering of low frequency low MV cluster



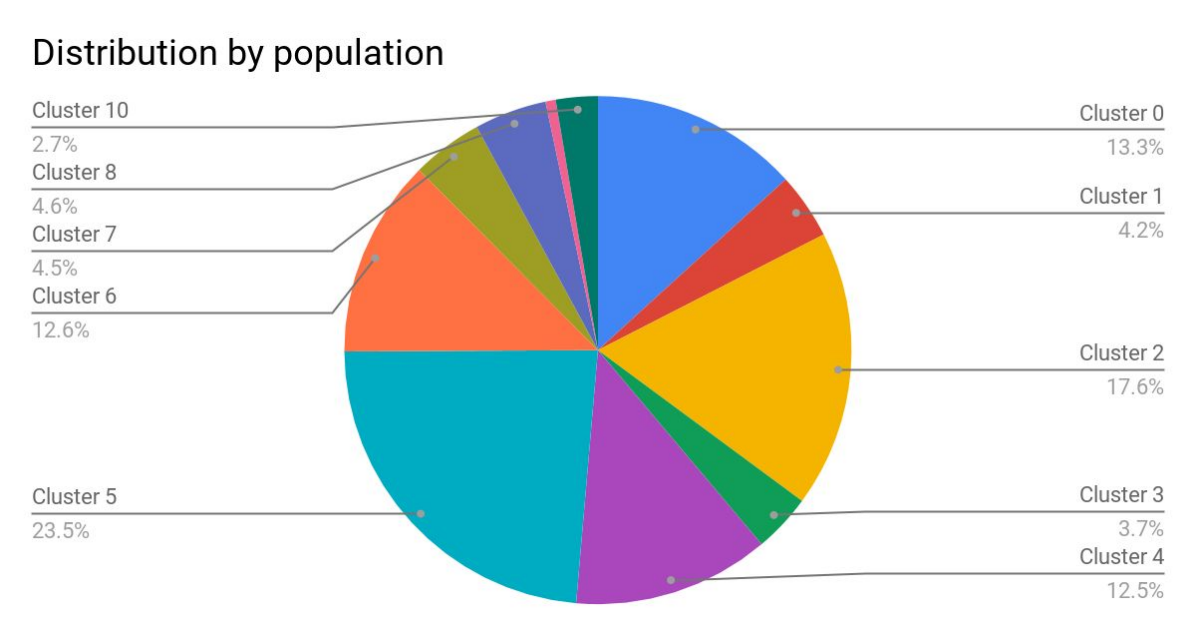
Final clustering of low recency cluster found above using credit card or cash used



Final clustering of credit card dominant cluster



Final clustering of medium potential cluster using age, religion and has a kid or not



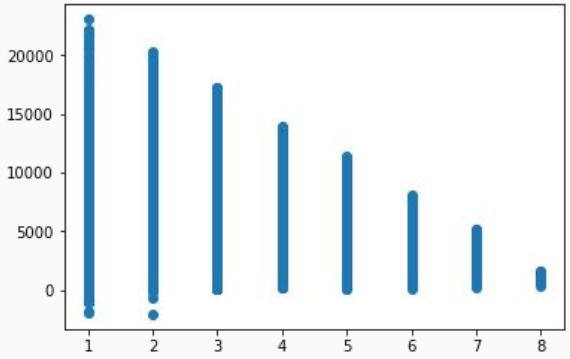
Similarly we have done for all other stores, as well as for the clusters formed in those stores. Some of them are up in the slides as follows(using mainly RFM analysis). Most of them showed similar trends for all forms of clustering mentioned above.

Exploration and clustering of store 2655

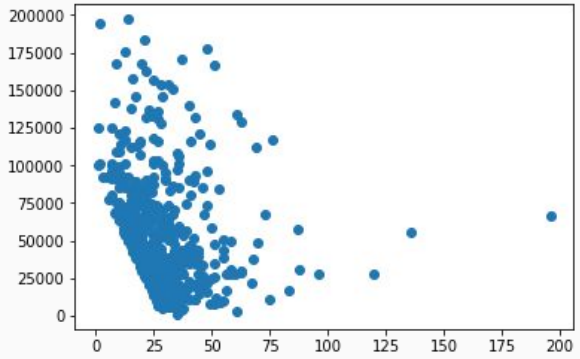
- This store is located in Indore-Malahar, Madhya Pradesh
- It has around 46k distinct customers.
- We follow the above approach to create some clusters of store 2655

Clustering using FM analysis

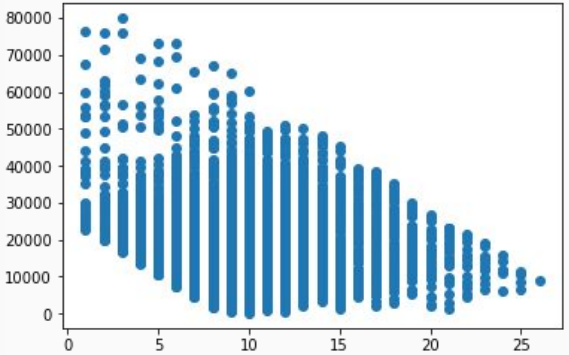
We create new variables frequency, recency and monetary value using the given data. Since we find some customers missing from tenders csv, we use the prices after promo in such cases and hence find the statistical outcome(mean, median, mode) which is almost similar, hence we can comment that their distributions are almost same, hence using any one as indicator for monetary value would work. After clustering, we get the following frequency-monetary value distribution.



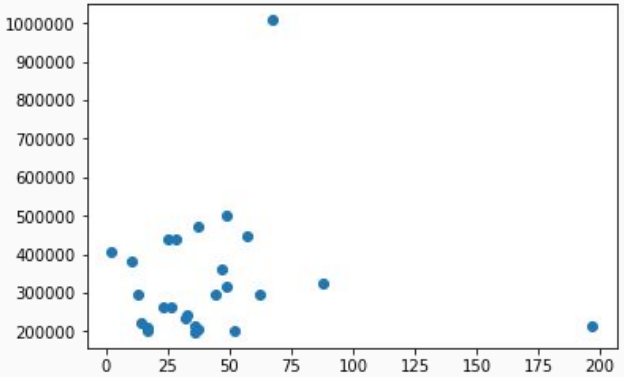
Cluster 0(86.1%)



Cluster 1(1.26%)



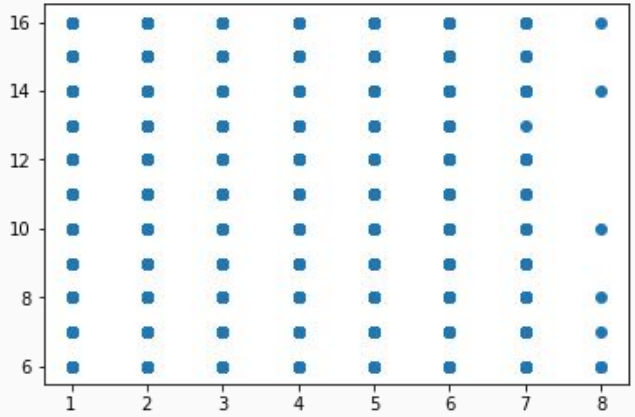
Cluster 2(12.52%)

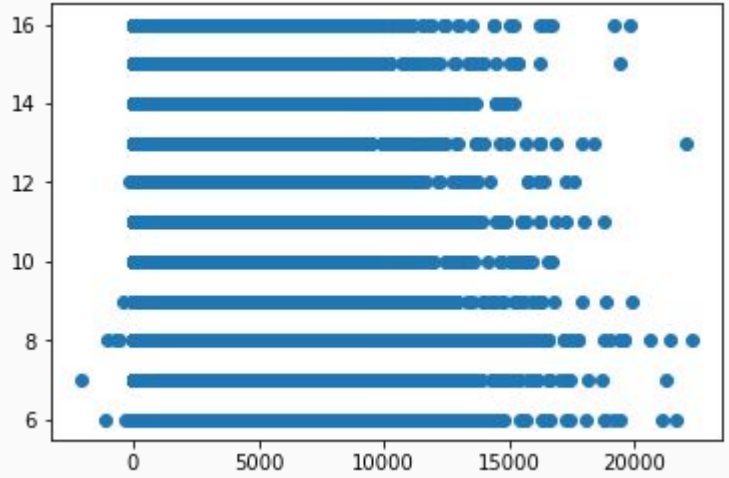


Cluster 3(0.05%)

Clustering cluster 0 using recency

Cluster 0 has the less frequent and low spending customers.Dividing this large cluster into 2 parts based on recency,we will focus only on the cluster which has lower recency,since they can be potential customers which FG may target.The following are the frequency vs recency plot(top) and the amount vs recency plot(bottom) for the low recency cluster.



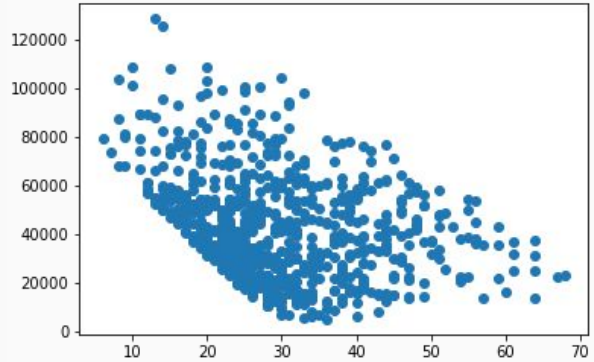


Exploration and clustering of store 4986

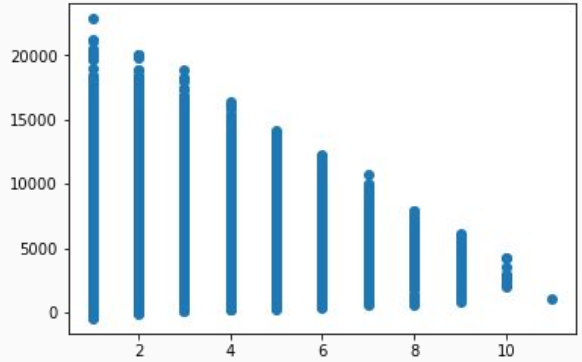
- This store is located in Madurai-Kochadai, Tamil Nadu
- It has around 18k distinct customers.
- We follow the above approach to create some clusters of store 4986

Clustering using FM analysis

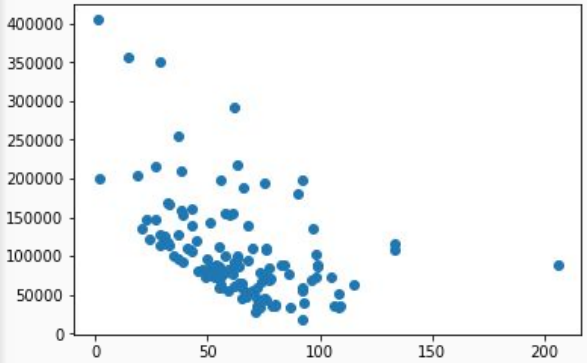
We create new variables frequency, recency and monetary value using the given data. Since we find some customers missing from tenders csv, we use the prices after promo in such cases and hence find the statistical outcome(mean, median, mode) which is almost similar, hence we can comment that their distributions are almost same, hence using any one as indicator for monetary value would work. After clustering, we get the following frequency-monetary value distribution.



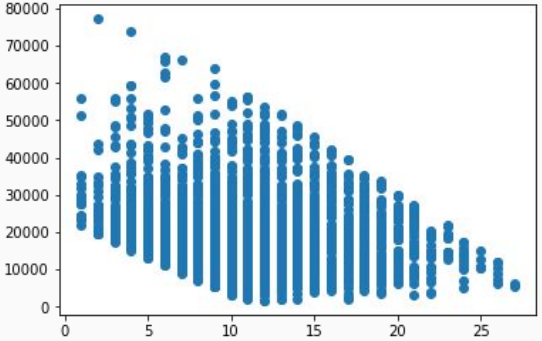
Cluster 0(3.96%)



Cluster 1(78.65%)



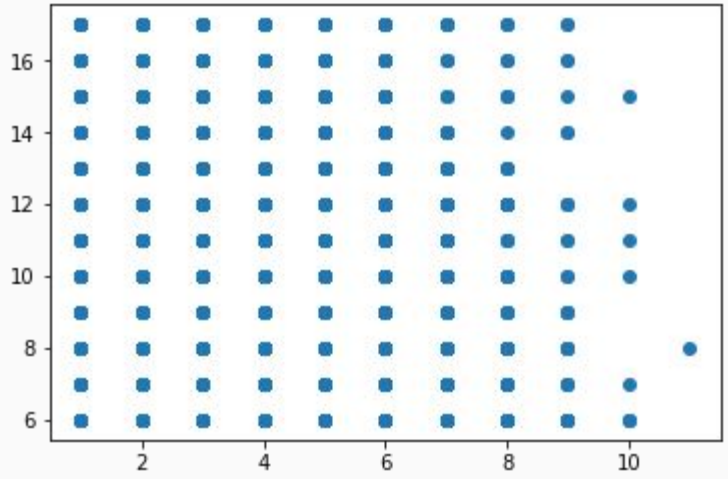
Cluster 2(0.67%)

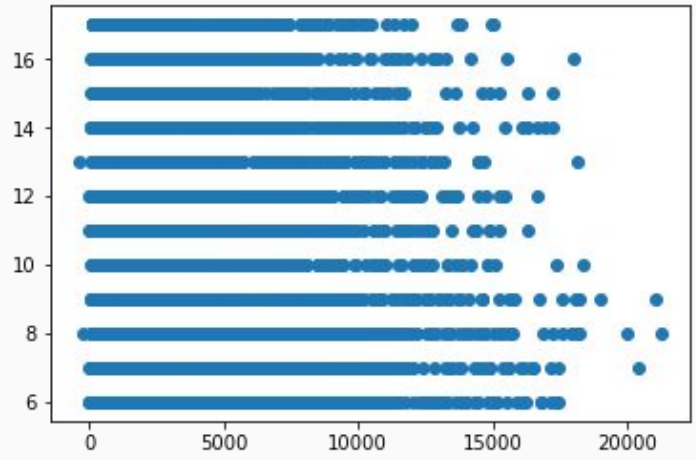


Cluster 3(16.71%)

Clustering cluster 1 using recency

Cluster 1 has the less frequent and low spending customers.Dividing this large cluster into 2 parts based on recency,we will focus only on the cluster which has lower recency,since they can be potential customers which FG may target.The following are the frequency vs recency plot(top) and the amount vs recency plot(bottom) for the low recency cluster.



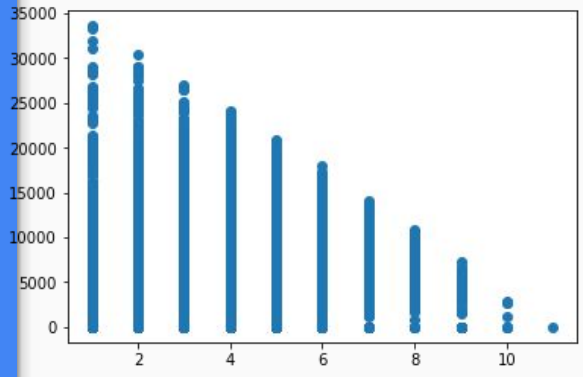


Exploration and clustering of store 4796

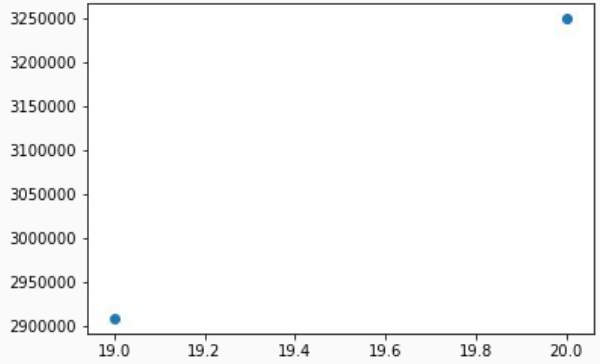
- This store is located in Hubli-Gokul, Karnataka
- It has around 32k distinct customers.
- We follow the above approach to create X clusters of store 4796

Clustering using FM analysis

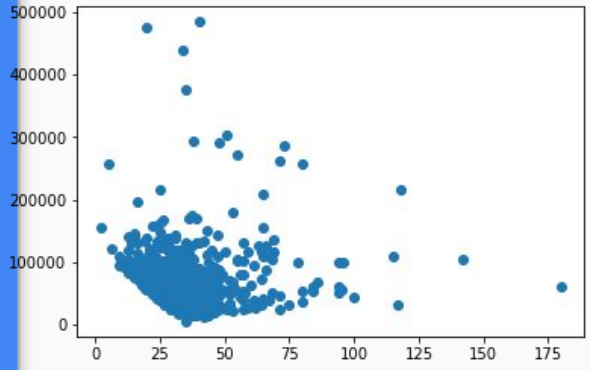
We create new variables frequency, recency and monetary value using the given data. Since we find some customers missing from tenders csv, we use the prices after promo in such cases and hence find the statistical outcome(mean, median, mode) which is almost similar, hence we can comment that their distributions are almost same, hence using any one as indicator for monetary value would work. After clustering, we get the following frequency-monetary value distribution.



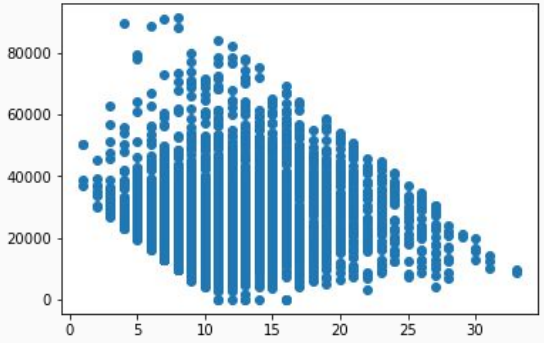
Cluster 0(83.5%)



Cluster 1(0.006%)



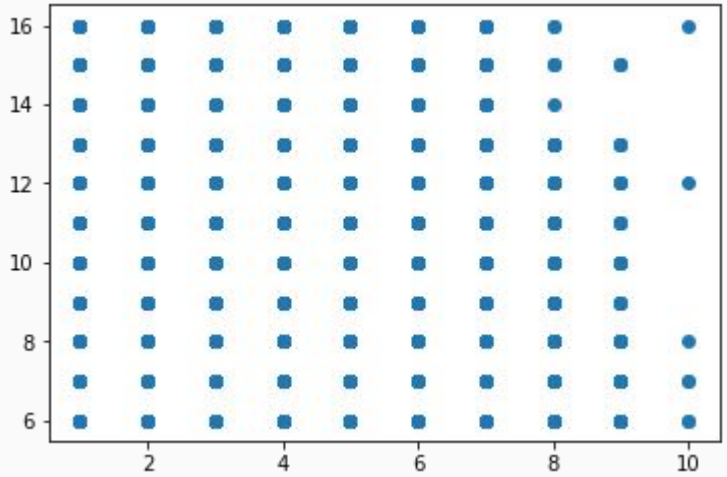
Cluster 2(2.28%)

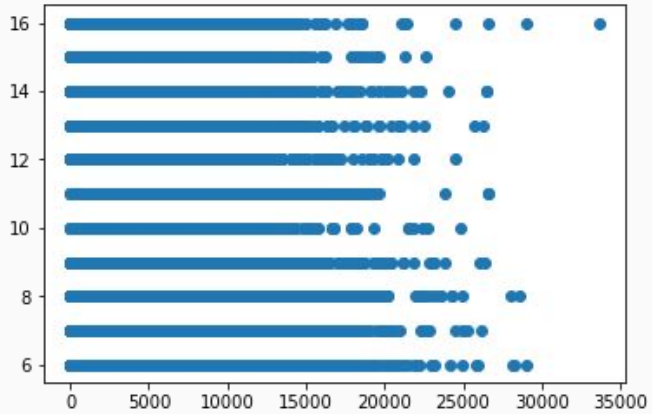


Cluster 3(14.17%)

Clustering cluster 0 using recency

Cluster 0 has the less frequent and low spending customers.Dividing this large cluster into 2 parts based on recency,we will focus only on the cluster which has lower recency,since they can be potential customers which FG may target.The following are the frequency vs recency plot(top) and the amount vs recency plot(bottom) for the low recency cluster.



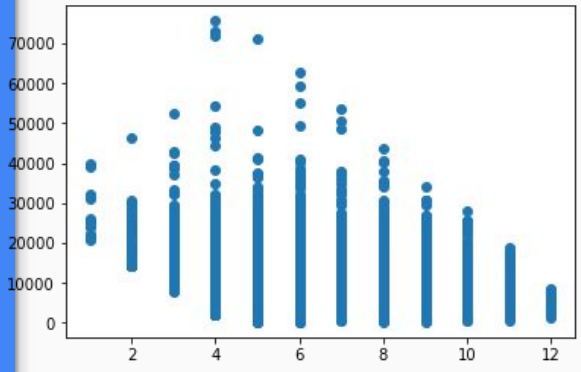


Exploration and clustering of store 2906

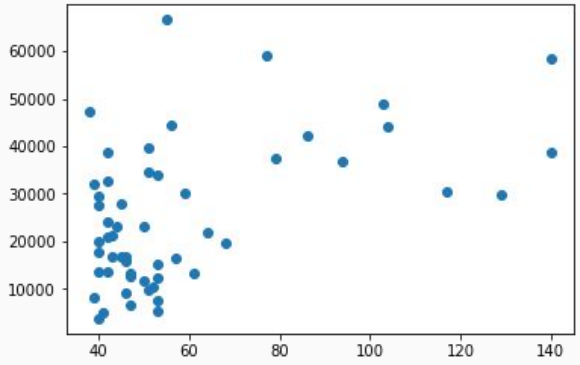
- This store is located in Indore-Treasure Island, Madhya Pradesh
- It has around 33k distinct customers.
- We follow the above approach to create X clusters of store 2906

Clustering using FM analysis

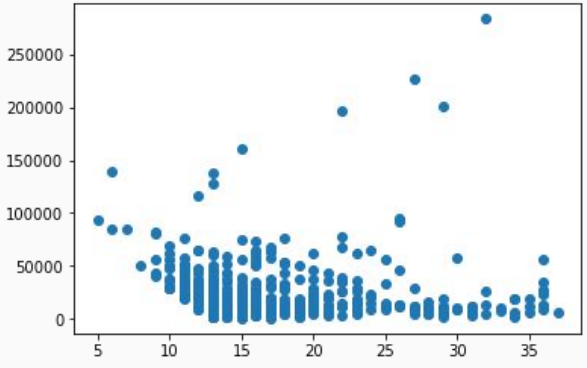
We create new variables frequency, recency and monetary value using the given data. Since we find some customers missing from tenders csv, we use the prices after promo in such cases and hence find the statistical outcome(mean, median, mode) which is almost similar, hence we can comment that their distributions are almost same, hence using any one as indicator for monetary value would work. After clustering, we get the following frequency-monetary value distribution.



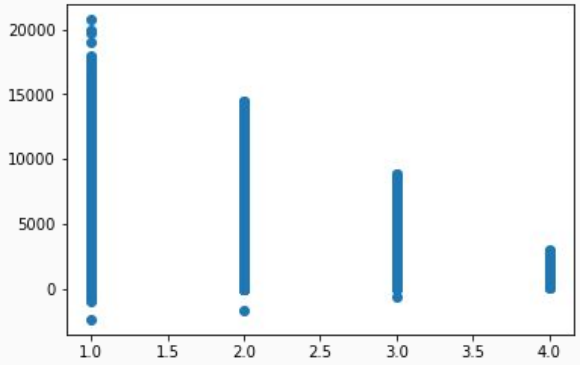
Cluster 0(13.25%)



Cluster 1(0.161%)



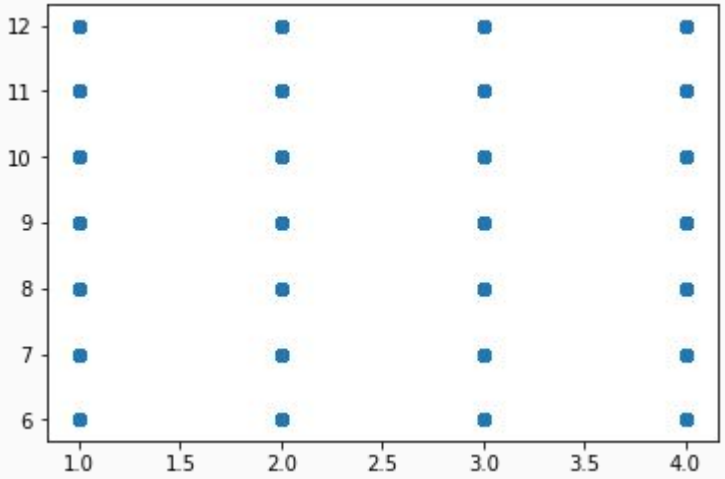
Cluster 2(1.54%)

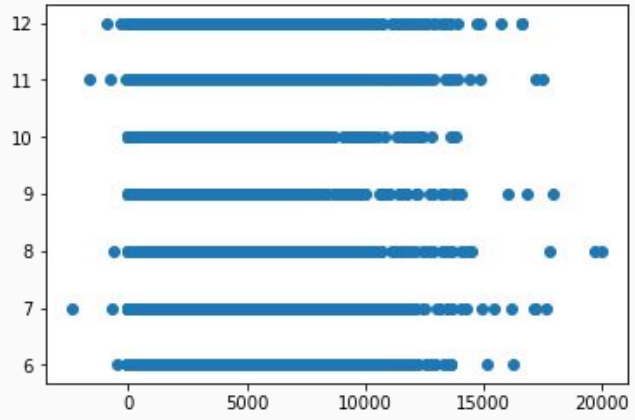


Cluster 3(85.04%)

Clustering cluster 3 using recency

Cluster 3 has the less frequent and low spending customers.Dividing this large cluster into 2 parts based on recency,we will focus only on the cluster which has lower recency,since they can be potential customers which FG may target.The following are the frequency vs recency plot(top) and the amount vs recency plot(bottom) for the low recency cluster.





Product Recommendation

Approach

- We utilize the fact the we have scarce knowledge of low potential customers and decent knowledge about high and medium potential customers.
- For low potential customers cluster, we suggest the past top 4 products bought by the customer only.
- For high and medium potential customers clusters, we apply cluster wise collaborative filtering.
- We take top four products recommended by the above collaborative filter and the rest as none, which gave us the best score on leaderboard.

Thank you!