IDS 572 Assignment 3

Aditi Vishwasrao - 674849041
Mounica Sirineni - 652518967
Nishanth Reddy Konkala - 658824939

Question 1. Use k-means clustering to identify clusters of households based on
a) The variables that describe purchase behavior (including brand loyalty). How do you evaluate brand loyalty?

Answer:
Variables used for this process are:

- Average Price
- Brand Runs
- Number of transactions
- Number of brands
- Others999
- Total volume
- Value
- Maximum brand loyalty

Maximum brand loyalty is obtained by taking maximum values out of the variables - Br. Cd. 57,144; Br. Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5. Others999 gives the share of transactions towards other brands which indicates that a customer is not brand loyal. K-means algorithm is implemented on these variables and results are summarized in the table below.

Table 1.1 Variables that describe purchase behavior

K (Number of clusters)	Cluster model (no of items in each cluster)	Comment	View scatterplot
2	Cluster 0: 373 items Cluster 1: 227 items	Clusters are unevenly distributed.	Click here
3	Cluster 0: 197 items Cluster 1: 204 items Cluster 2: 199 items	Clusters are nearly equally distributed.	Click here
4	Cluster 0: 51 items Cluster 1: 191 items Cluster 2: 181 items Cluster 3: 177 items	Clusters are unevenly distributed with one of the clusters being very small	Click here
5	Cluster 0: 135 items Cluster 1: 145 items Cluster 2: 100 items Cluster 3: 180 items Cluster 4: 40 items	Clusters are unevenly distributed with one of the clusters being very small	Click here

Centroid plot view of clusters:

K = 2

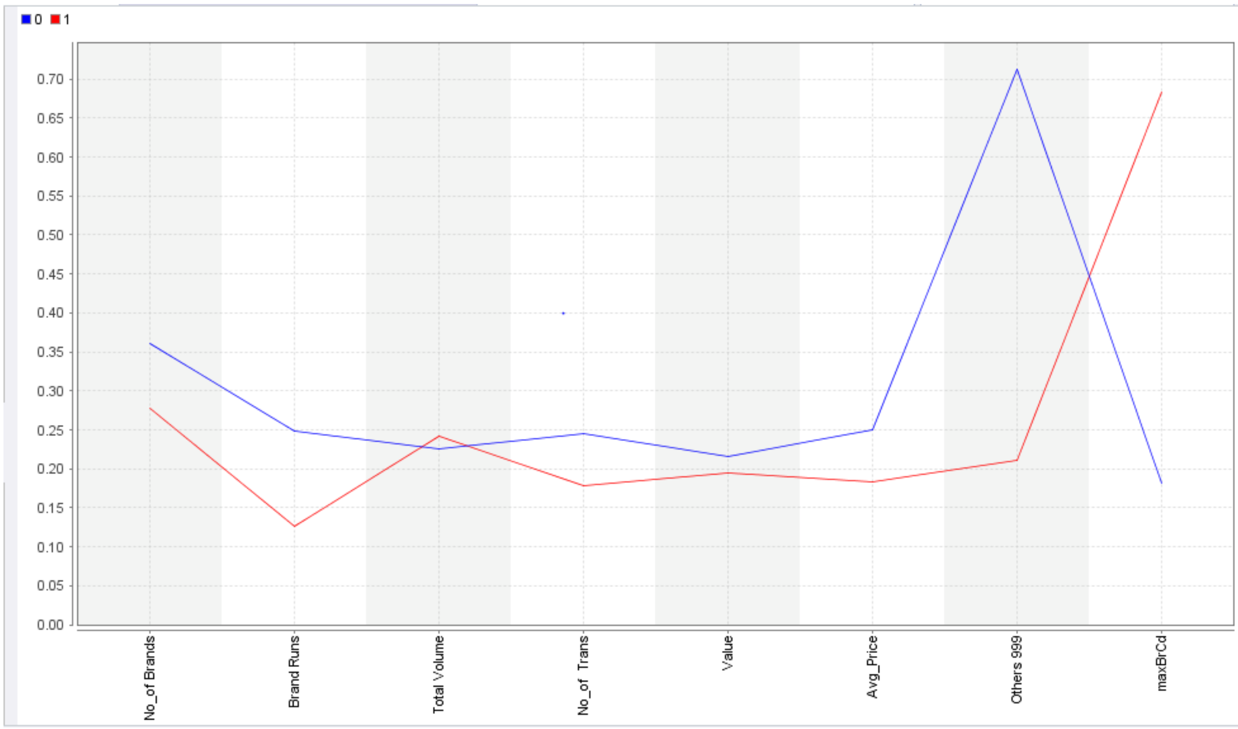


Figure 1.1.1 Centroid plot with k = 2

Description:

Cluster 0: Customers in this cluster appear to buy from Others999 brands which indicate they are not brand loyal customers. They buy the highest number of brands but the total volume of transactions is least for this cluster.

Cluster 1: Customers in this cluster are a bit opposite to that of cluster 0's. They have maximum brand loyalty, buys least number of brands with highest volume of transactions.

K = 3

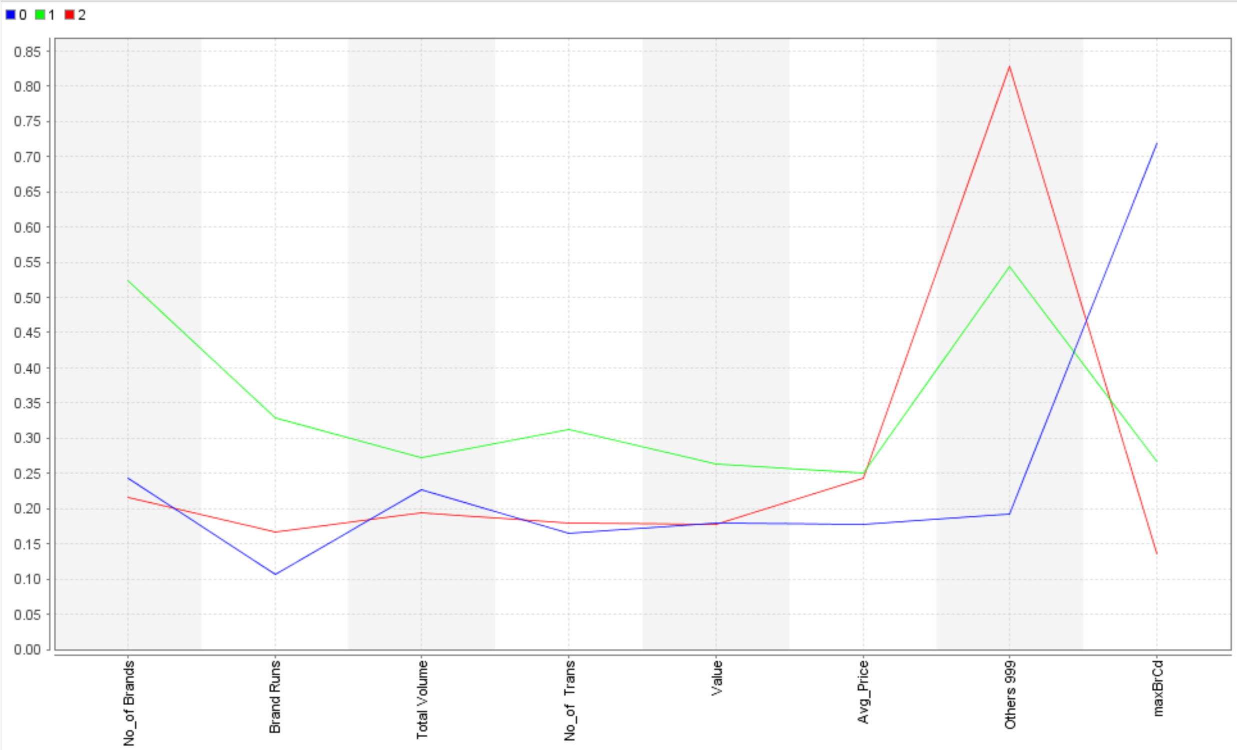


Figure 1.1.2 Centroid plot with k = 3

Description:

Cluster 0: This group has the maximum brand loyalty and also has the lowest number of brands and least average price of total transaction.

Cluster 1: This group purchase a large number of brands with highest brand runs. They have highest number and volume of transactions whereas their brand loyalty is between cluster 0 and cluster 2 customers.

Cluster 2: This group has least brand loyalty because they tend to buy from Others999 brands. They have least number of brands and total volume of transactions.

K = 4

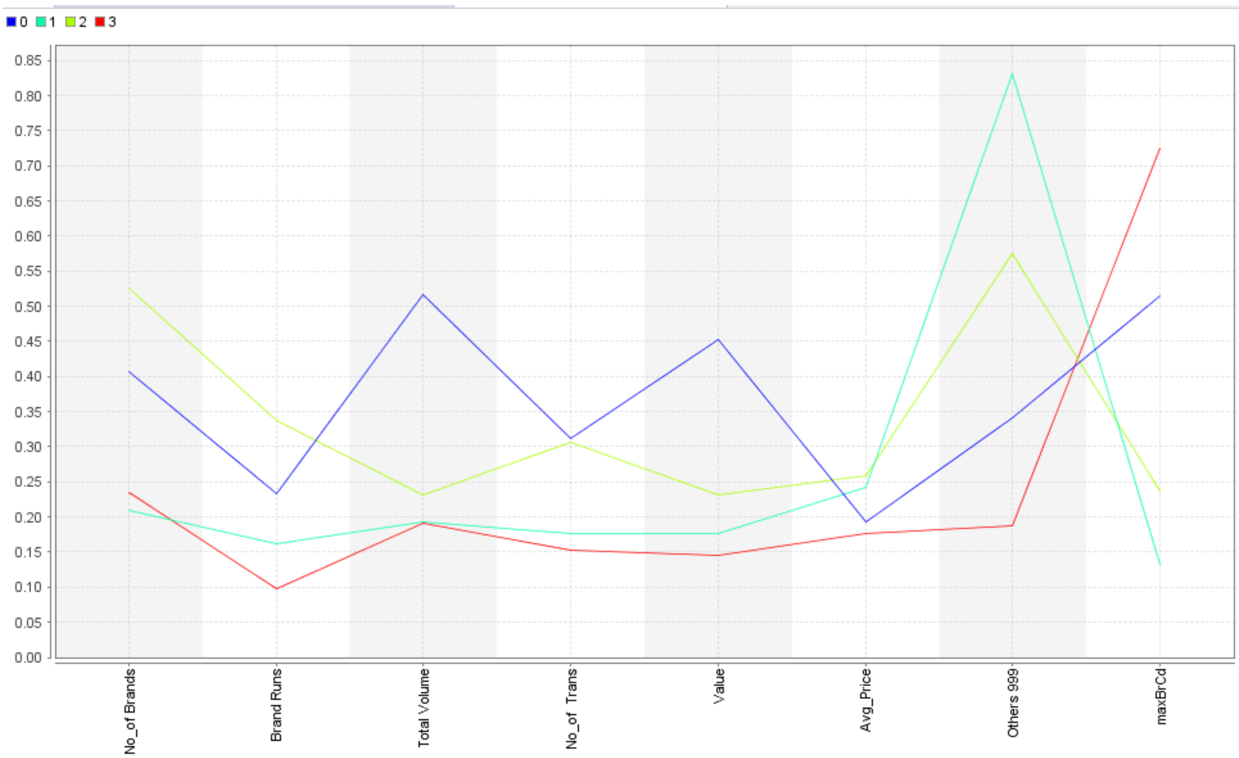


Figure 1.1.3 Centroid plot with k = 4

Description:

Cluster 0: This group has the highest total volume and number of transactions. They have a significant peak for brand loyalty but also has considerable purchases from Others999 brands.

Cluster 1: This group has the least number of brands and least brand loyalty. They have maximum brand share from Others999.

Cluster 2: This group has maximum number of brands, brand runs, average price of transactions. They have a significant peak for Others999 which show that they are not brand loyal.

Cluster 3: They are the most brand loyal customers as seen from the graph above. Their purchase in Others999 brand is the least which also adds on to prove their loyalty. They have a small number of brand runs and the least number of transactions and average price of good purchased.

K = 5

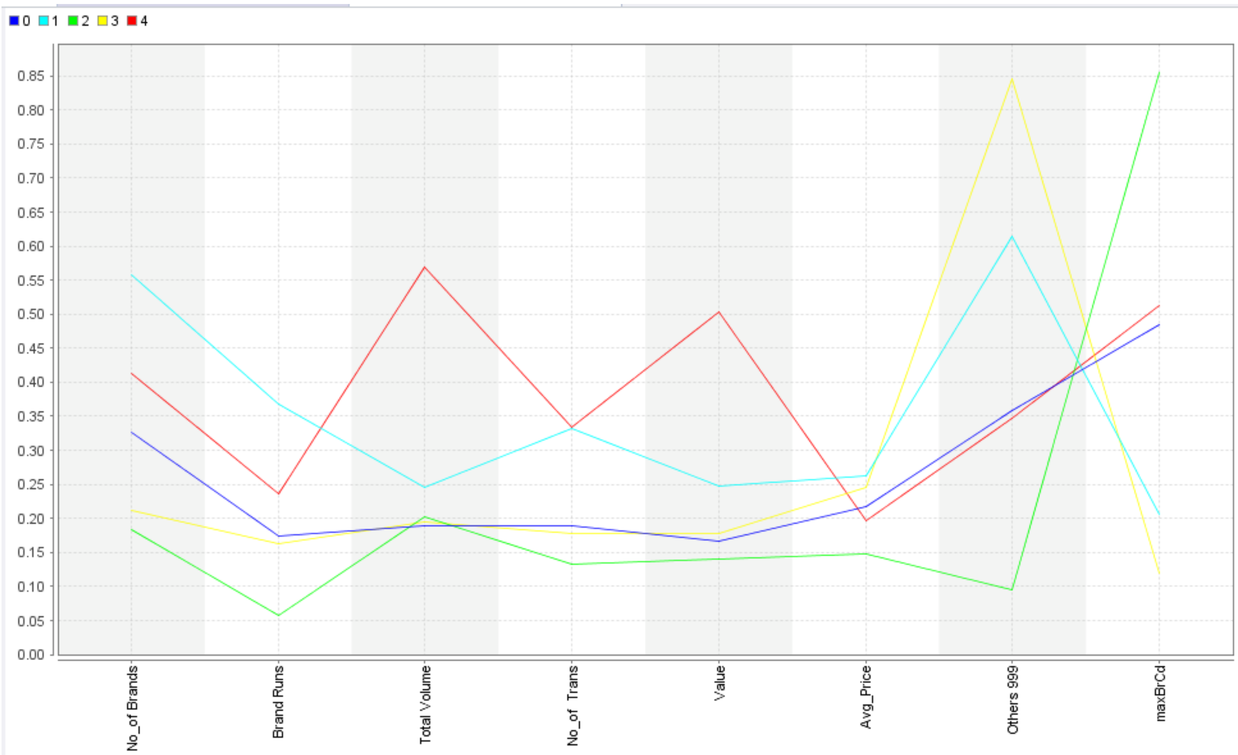


Figure 1.1.4 Centroid plot with k = 5

Description:

Cluster 0: This group has least total volume of transactions with a moderate peak in brand loyalty.

Cluster 1: This group has highest number of brands, brand runs, number of transactions, and average price for transactions. They have significant peak for share in Others999 brands.

Cluster 2: They have least number of brands, brand runs, number of transactions, value of goods purchased and average price of transaction. They have maximum brand loyalty.

Cluster 3: They are the least brand loyal customers. They are neither least nor highest in other characteristics when compared to other clusters.

Cluster 4: They have highest total volume and number of transactions, value of goods purchased. They have moderate peaks in both brand loyalty and share in Others999 brands.

b) The variables that describe basis-of-purchase

Answer: Variables used for this process are:

- All price categories
- Selling propositions
- Purchase volume with no promotion, promotion 6 and other promotion

We have plotted graphs for all selling propositions and observed that PropCat 9, PropCat 10, PropCat 11, PropCat 13, PropCat 14 have very less data points against them. We did not observe much distribution patterns for this variables. So we have considered only PropCat 5 – 8, PropCat 12, PropCat 15.

Table 1.2 Variables that describe basis of purchase

K (Number of clusters)	Cluster model (no of items in each cluster)	Comment	View scatterplot
2	Cluster 0: 370 items Cluster 1: 230 items	Clusters are nearly equally distributed.	Click here
3	Cluster 0: 149 items Cluster 1: 328 items Cluster 2: 123 items	Cluster 1 is very large and the other 2 clusters are almost equal in size.	Click here
4	Cluster 0: 134 items Cluster 1: 331 items Cluster 2: 78 items Cluster 3: 57 items	The cluster size in this case varies from large (331) to very small (57)	Click here
5	Cluster 0: 74 items Cluster 1: 120 items Cluster 2: 181 items Cluster 3: 170 items Cluster 4: 55 items	Clusters 1,2,3 are almost evenly distributed in size. The size of cluster 4 is very small.	Click here

Centroid plot view of clusters:

K=2

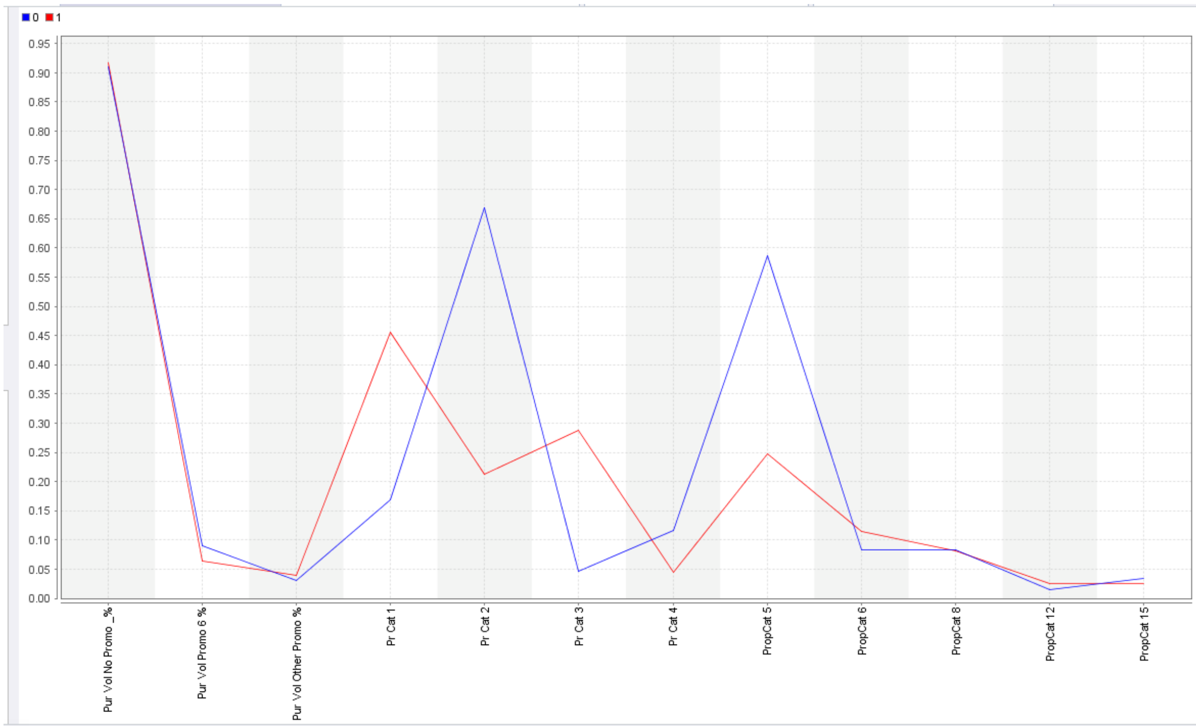


Figure 1.2.1 Centroid plot with k = 2

Description:

Cluster 0: Customers in this cluster are mainly those who have a high volume of purchases of items not on promotion. These customers also tend to have a higher volume of purchases then the product is promoted as a value added pack. They tend to buy a higher volume of beauty products

Cluster 1: Customers in cluster 2 also purchase a high volume of items not on promotion. They tend to purchase more items which are on price off, coupons or value added packs.

K = 3

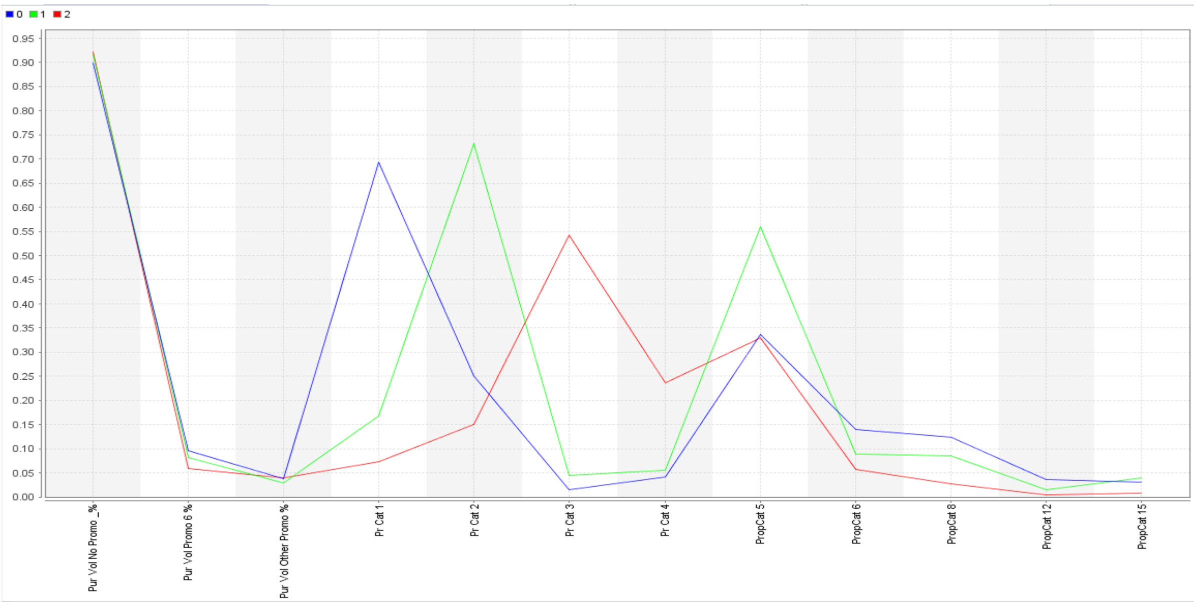


Figure 1.2.2 Centroid plot with k = 3

All clusters have display a common characteristic of the customers having a high volume of purchases when items are not on promotion and all three clusters show a preference for value added packs. The following are the distinctions between the 3 clusters:

Cluster 0: Prefers to buy more items when there are price offs

Cluster 1: Prefers to buy more products on exchange offer.

Cluster 2: Prefer to buy products with coupons.

K=4

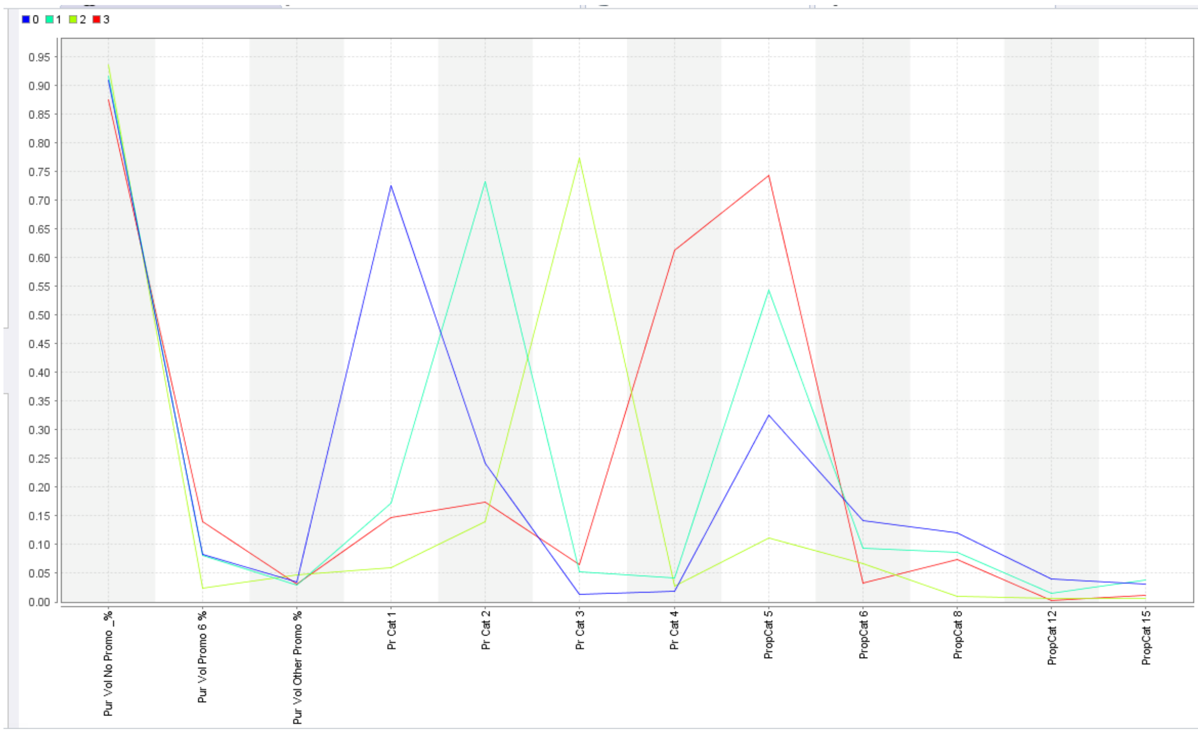


Figure 1.2.3 Centroid plot with k = 4

When the data is segmented in 4 clusters, we see the same trends in 3 of the clusters as we saw when K=3. This value of K reveals a new cluster which consists of customers that buy more products when they have an extra grammage along with value added packs associated with them.

K=5

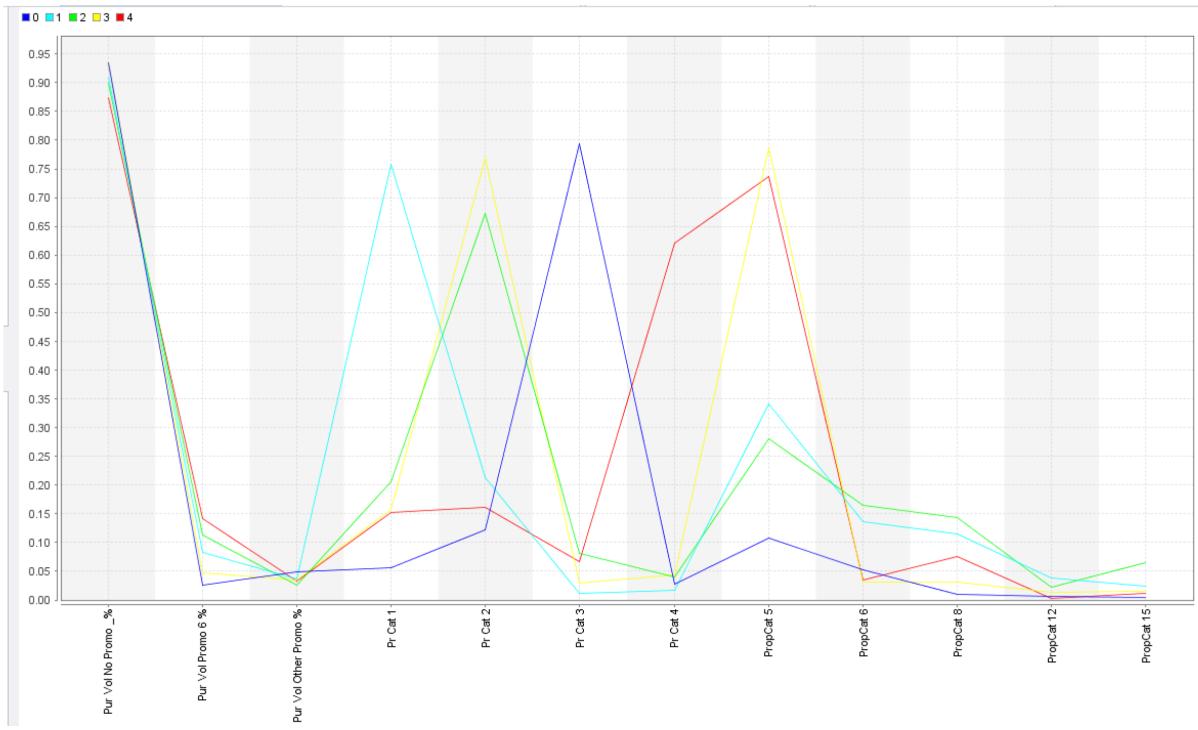


Figure 1.2.4 Centroid plot with k = 5

Apart from the clusters formed when K=4, the centroid plot for K=5 does not reveal any new segments of customers whose preferences are different from those in the K=4 model.

c) The variables that describe both purchase behavior and basis of purchase:

Answer:

Table 1.3 Variables that describe both purchase behavior and basis of purchase

K (Number of clusters)	**Cluster model (no of items in each cluster)**	**Comment**	**View scatterplot**
2	Cluster 0: 366 items Cluster 1: 234 items	Clusters are nearly equally distributed.	Click here
3	Cluster 0: 79 items Cluster 1: 283 items Cluster 2: 238 items	Clusters are unevenly distributed with one of the clusters being very small	Click here

4	Cluster 0: 252 items Cluster 1: 77 items Cluster 2: 123 items Cluster 3: 148 items	The cluster sizes varied from very small ones to bigger ones.	Click here
5	Cluster 0: 223 items Cluster 1: 54 items Cluster 2: 74 items Cluster 3: 130 items Cluster 4: 119 items	Like in 4 clusters, the size of cluster 0 is very big when compared to others and the rest have different sizes	Click here

Centroid plot view of cluster:

K=2:

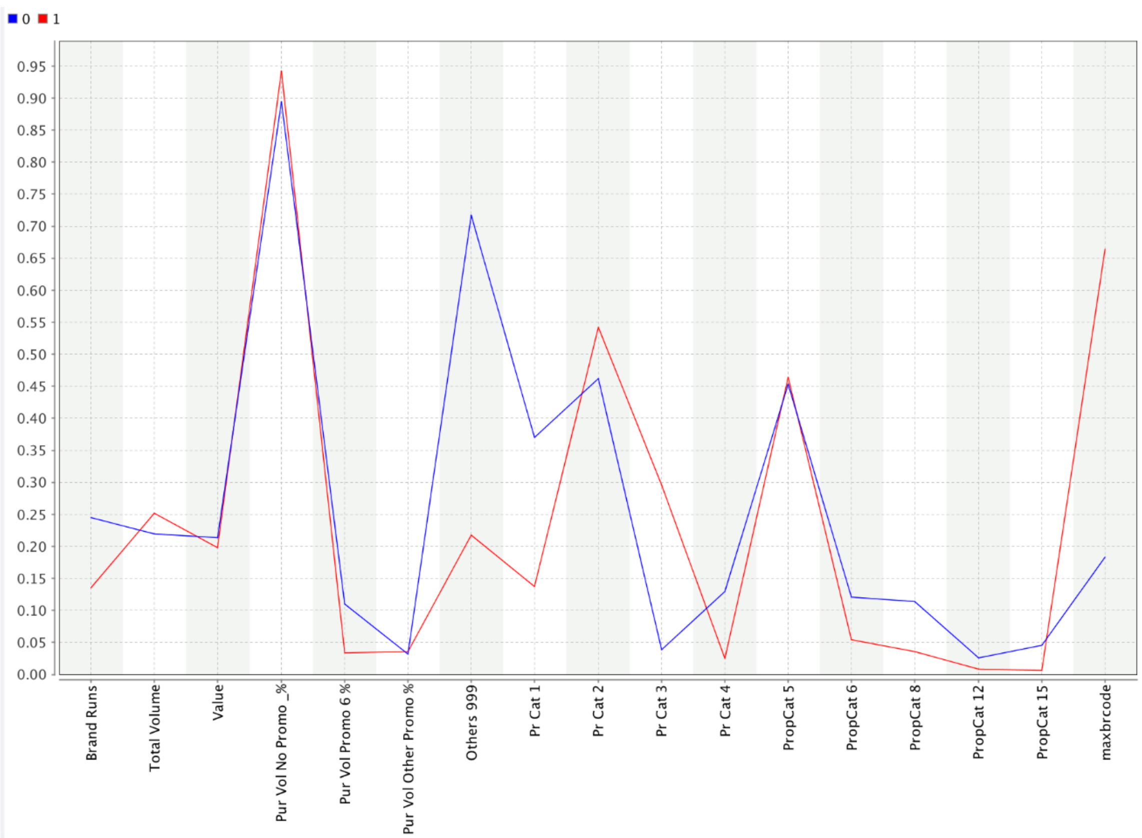


Figure 1.3.1 Centroid plot with k = 2

Description:

Cluster 0: Customers in this cluster appear to have slightly higher total purchases and do not particularly consider promotions while purchasing items. They are highly brand loyal and do not buy products of other brands.

Cluster 1: Customers in this cluster are a bit opposite to that of cluster 0's. They have very less brand loyalty and buy products from any brand, which offers promotion, especially in freshness soaps.

K = 3

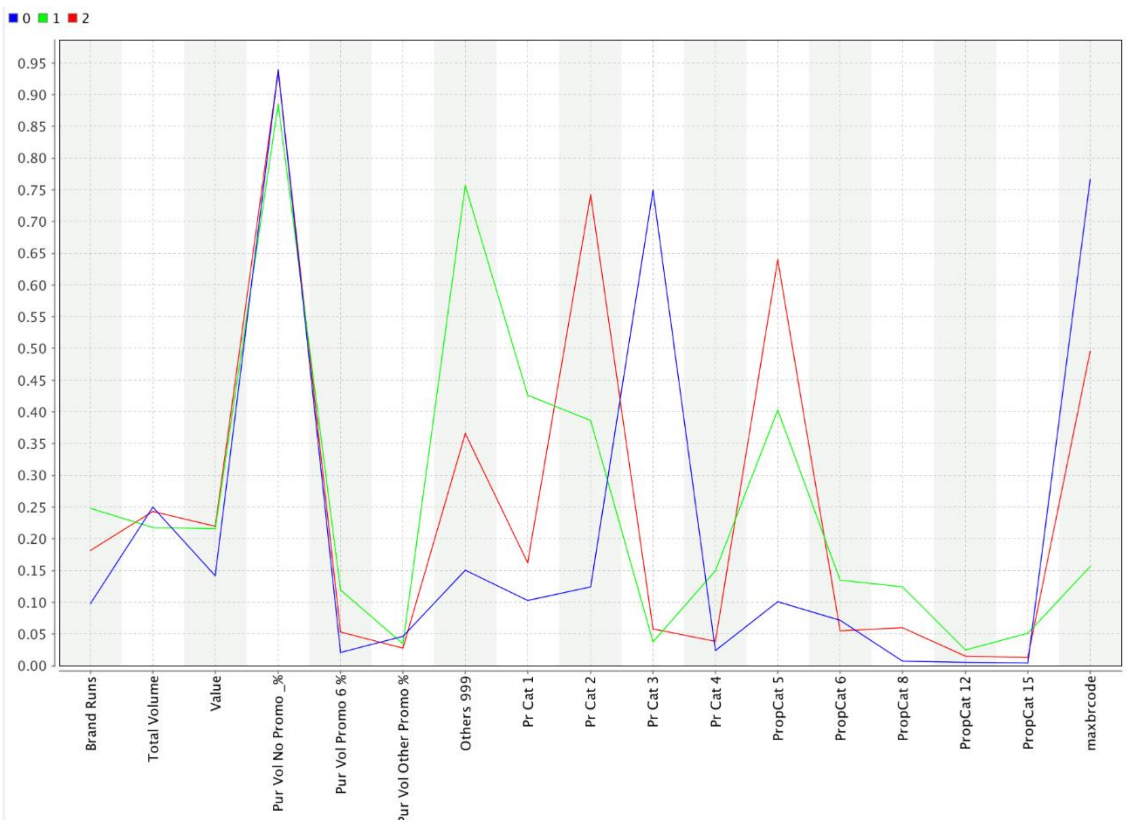


Figure 1.3.2 Centroid plot with k = 3

Description:

Cluster 0: Customers in this cluster are highly brand loyal and buy products irrespective of the promotions. One interesting fact is that they use coupons frequently for any purchases

Cluster 1: Customers in this cluster are least brand loyal and buy products from any brand. They majorly buy beauty and freshness soaps.

Cluster 2: Customers in this cluster have avg. brand loyalty and buy products from other brands not very often. They mainly make use of exchange offers for purchasing and they majorly purchase beauty soaps.

K = 4

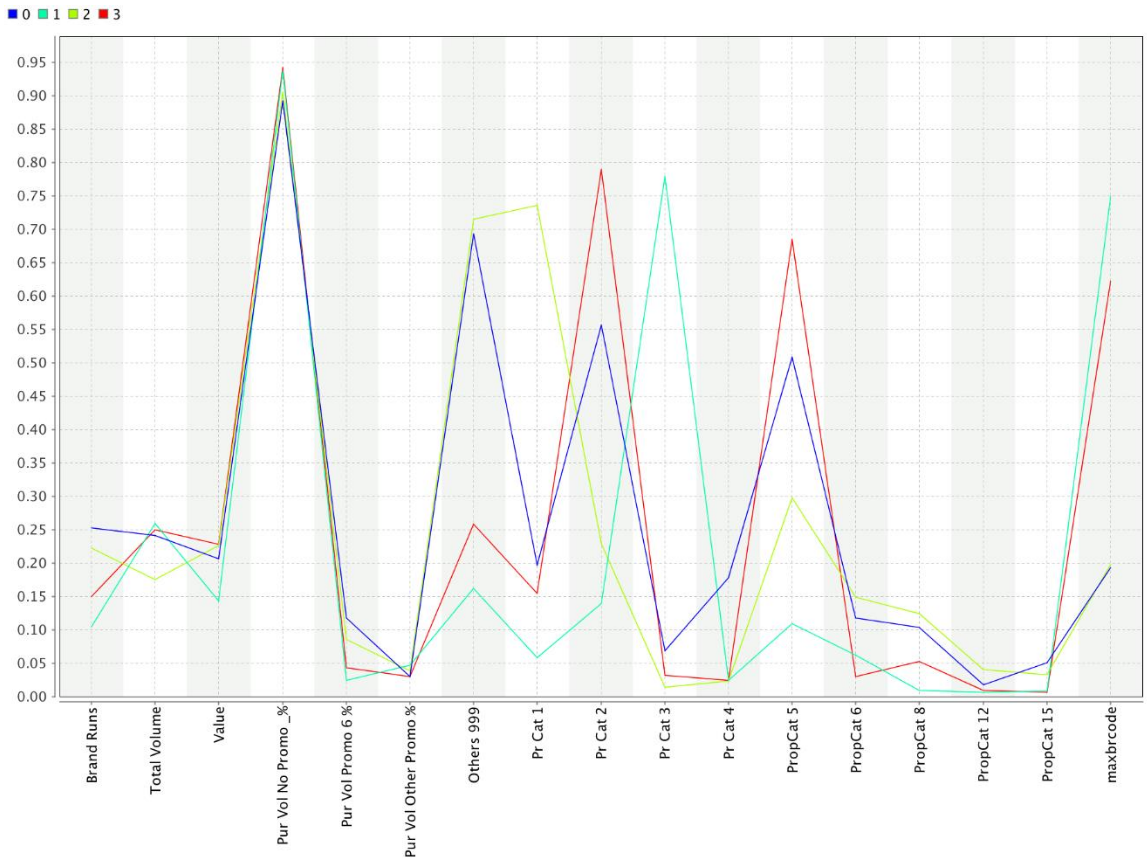


Figure 1.3.3 Centroid plot with k = 4

Description:

Cluster 0: Customers in this are least brand loyal and buy products from different brand very often and the other characteristics of this cluster are pretty average and nothing significant.

Cluster 1: Customers in this cluster has the highest total volume of purchase and they are highly brand loyal which makes them stick to one brand they rarely buy products from other brands. They make use of coupons for their purchases.

Cluster 2: Customers in this cluster have very low brand loyalty and they buy products from other brands very often. They buy products that are being offered with price off.

Cluster 3: They have average brand loyalty and they majorly buy soaps in exchange offer. Their major proportion of purchase are beauty soaps.

K = 5

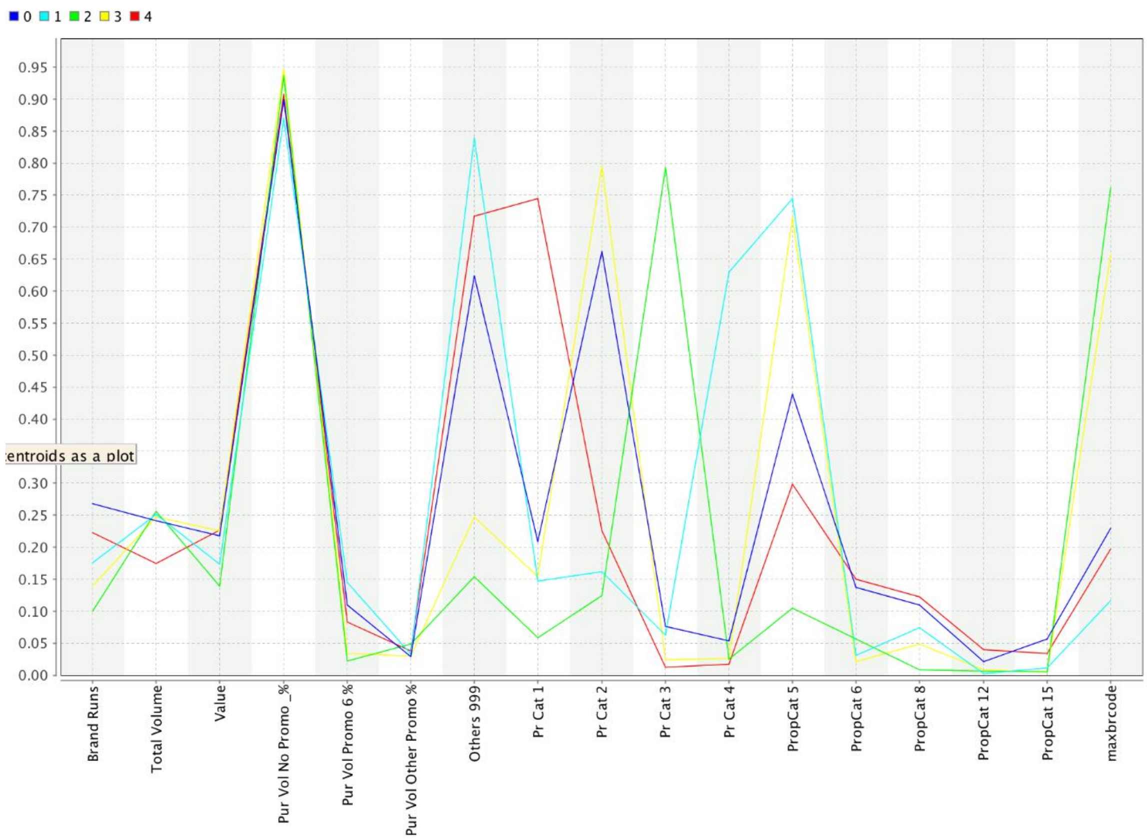


Figure 1.3.4 Centroid plot with k = 5

Description:

Cluster 0: One notable feature about customers in this cluster is that they buy soaps majorly in exchange offer

Cluster 1: Customers in this cluster are least brand loyal and buy products from other brands very often. They majorly buy products in extra gram mage and buy beauty soaps.

Cluster 2: Customers in this cluster are have high brand loyalty and use coupons majorly for their purchases

Cluster 3: The characteristics of customers in this cluster are pretty average except that they buy products in exchange offer majorly.

Cluster 4: Customers in this cluster have less brand loyalty and buy products from any brand which are being offered at price off

How should k be chosen?

Ans. The value of 'K' should be chosen in such a way that:
1) The Intra cluster distances are minimum in all clusters
2) The clusters are well apart. That is, the inter cluster distances are maximum.

The value ok K can be chosen in the next question.

How should the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is?

Ans. The percentages of total purchases should not be considered individually as they increase the inter cluster distances and the effectiveness of the clustering drops. Instead, consider MaxBrCode(Max proportion of purchase) which give the brand loyalty of the customer.

d. Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering too. You do not need to try all techniques on all combinations in (a)-(c) above; you may pick one set of variables in (a) thru (c) that you feel may be best suited for addressing the segmentation need. How do different parameter values for the different techniques affect the clusters obtained? Are the clusters obtained from the different procedures similar? What might be some reasons for differences in clusters obtained using different procedures? Which would you pick as your 'best' and why?

Ans.

For best clustering technique analysis, we have decided to choose the following set of variables:
We have chosen the variables in **part (c)**:
#brands, brand runs, total volume, #transactions, value, Avg. price, share to other brands, (brd_loyalty), Pur_vol_no_promo, Pur_vol_promo_6, Pur_vol_other, all price categories, PropCat 5,6,8,12,15.

I. **K-medoids**

The major difference between the k-means and k-medoids algorithm is the centroid. In the k-means algorithm the centroid of a cluster will frequently be an imaginary point, not part of the cluster itself, which we can take to mark its center. Whereas, in K-medoids, the centroid of a cluster is a point for which each attribute value is the average of the values of the corresponding attribute for all the points in the cluster. The centroid of a cluster will always be one of the points in the cluster.

1.4.1 Performance Vector for K-medoids:

K	Avg. within centroid distance_cluster_0	Avg. within centroid distance_cluster_1	Avg. within centroid distance_cluster_2	Avg. within centroid distance_cluster_3	Avg. within centroid distance_cluster_4
2	-1.072	-0.860			

3	-0.927	-0.789	-0.786		
4	-0.727	-1.401	-0.430	-0.692	
5	-0.471	-0.661	-0.417	-0.608	-0.700

1.4.2 Performance Vector for K-means:

K	**Avg. within centroid distance_cluster_0**	**Avg. within centroid distance_cluster_1**	**Avg. within centroid distance_cluster_2**	**Avg. within centroid distance_cluster_3**	**Avg. within centroid distance_cluster_4**
2	-0.542	-0.617			
3	-0.286	-0.559	-0.391		
4	-0.466	-0.218	-0.451	-0.334	
5	-0.366	-0.260	-0.203	-0.331	-0.447

Conclusion: Comparing average within centroid distance and cluster between K-means and K-medoids, we observe that K-means has less distance within clusters. Therefore, K-means performs better than K-medoids.

II. Kernel K-means:

For each value of k, the k-means (kernel) algorithm gives the cluster sizes which are almost similarly distributed as in the k-means algorithm:

K=2

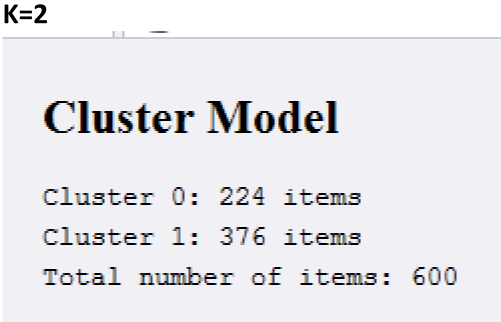


Figure 1.4.1 Cluster model with k = 2

K=3

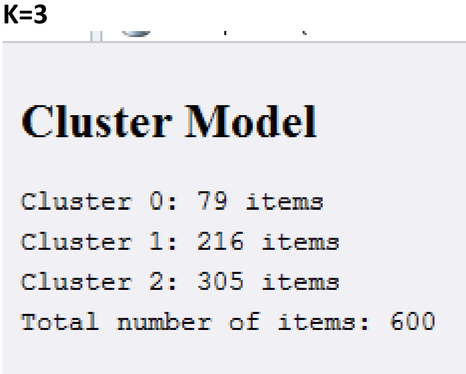


Figure 1.4.2 Cluster model with k = 3

K=4

Cluster Model

Cluster 0: 74 items
Cluster 1: 123 items
Cluster 2: 144 items
Cluster 3: 259 items
Total number of items: 600

Figure 1.4.3 Cluster model with k = 4

PerformanceVector

PerformanceVector:
Avg. within cluster distance: -169.912
Avg. within cluster distance for cluster 0: -53.256
Avg. within cluster distance for cluster 1: -126.966
Avg. within cluster distance for cluster 2: -115.233
Avg. within cluster distance for cluster 3: -254.039

Figure 1.4.4 Performance vector with k = 4

However, the average within cluster distance is very large in case of k-means (kernel). This means that data points within a cluster are far apart from each other than the data points in the clusters given by K-means. Thus, we will not be considering the K-means (kernel) model.

III. **Agglomerative clustering:**

We used default parameters and implemented the algorithm.

Figure 1.4.5 Hierarchical cluster model

Agglomerative clustering gave more number of clusters than the data points, which is not possible. Therefore we discarded this technique from further evaluations.

IV. **DBSCAN clustering:**

We changed parameters and implemented DBSCAN algorithm.

For Epsilon = 0.01 and Min points = 5, result is shown below

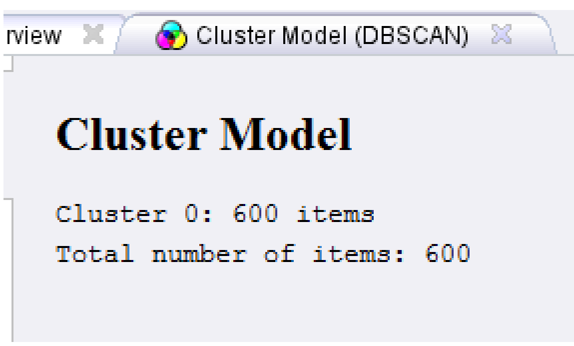


Figure 1.4.5 Cluster model with epsilon = 0.01 and min points =5

For Epsilon = 0.1 and Min points = 5, result is shown below

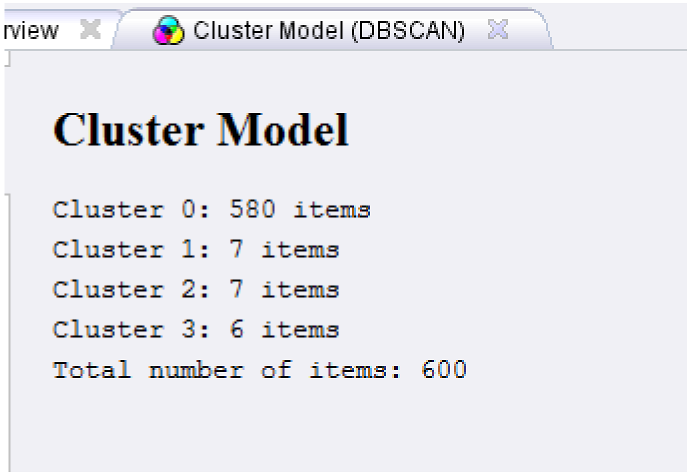


Figure 1.4.6 Cluster model with epsilon = 0.1 and min points =5

Conclusion: DBSCAN clustering clusters all data points in a single cluster. When the parameters are changed i.e. epsilon value and min points is changed, the formation of clusters does not show a good segmentation size of the data. Hence we can cannot consider this model for this dataset.

2. a) Select what you think is the 'best' segmentation - explain why you think this is the 'best'.

Answer: Based on the process implemented in question 1, the following tables gives the performance of K-means algorithm for different K values on each variable data set.

Purchase behavior

Table 2.1.1 Performance for Purchase behavior

K (Number of clusters)	Average distance between clusters	Average distance within clusters
2	0.731	centroid distance: -0.180 centroid distance_cluster_0: -0.187 centroid distance_cluster_1: -0.169
3	0.690	centroid distance: -0.140 centroid distance_cluster_0: -0.142 centroid distance_cluster_1: -0.160 centroid distance_cluster_2: -0.119
4	1.008	centroid distance: -0.127 centroid distance_cluster_0: -0.200 centroid distance_cluster_1: -0.116 centroid distance_cluster_2: -0.125 centroid distance_cluster_3: -0.121

5	1.391	centroid distance: -0.110 centroid distance_cluster_0: -0.080 centroid distance_cluster_1: -0.116 centroid distance_cluster_2: -0.093 centroid distance_cluster_3: -0.113 centroid distance_cluster_4: -0.217

Basis of Purchase

Table 2.1.2 Performance for basis of Purchase

K (Number of clusters)	Average distance between clusters	Average distance within clusters
2	0.687	centroid distance: -0.383 centroid distance_cluster_0: -0.322 centroid distance_cluster_1: -0.481
3	0.81185	centroid distance: -0.303 centroid distance_cluster_0: -0.299 centroid distance_cluster_1: -0.263 centroid distance_cluster_2: -0.418
4	0.961636	centroid distance: -0.240 centroid distance_cluster_0: -0.284 centroid distance_cluster_1: -0.258 centroid distance_cluster_2: -0.120 centroid distance_cluster_3: -0.190
5	0.934286	centroid distance: -0.198 centroid distance_cluster_0: -0.107 centroid distance_cluster_1: -0.274 centroid distance_cluster_2: -0.266 centroid distance_cluster_3: -0.115 centroid distance_cluster_4: -0.187

Both purchase behavior and basis of purchase

Table 2.1.3 Performance for Purchase behavior and basis of purchase

K (Number of clusters)	Average distance between clusters	Average distance within clusters
2	0.808	centroid distance_cluster_0: -0.542 centroid distance_cluster_1: -0.617
3	1.047	centroid distance_cluster_0: -0.286 centroid distance_cluster_1: -0.559 centroid distance_cluster_2: -0.391
4	1.0405	centroid distance_cluster_0: -0.466 centroid distance_cluster_1: -0.218 centroid distance_cluster_2: -0.451 centroid distance_cluster_3: -0.334

5	1.0904	centroid distance_cluster_0: -0.366 centroid distance_cluster_1: -0.260 centroid distance_cluster_2: -0.203 centroid distance_cluster_3: -0.331 centroid distance_cluster_4: -0.447

We have considered 3 criteria to choose K:

- Minimum distance within cluster
- Maximum distance between clusters
- Information from centroid plot of clusters

In all 3 segmentations, we observe that for K = 4 or 5, distance within cluster is minimum and distance between clusters is maximum. But when we look at centroid plots in question 1, we notice that algorithm gives similar information for both k = 4 and 5. Since we are getting similar information at 4 with minimum distance within cluster and maximum distance between clusters, we conclude that **K-means algorithm with K = 4 is the best model.**

b) Comment on the characteristics (demographic, brand loyalty and basis-for-purchase) of these clusters.

Demographics:

SEC

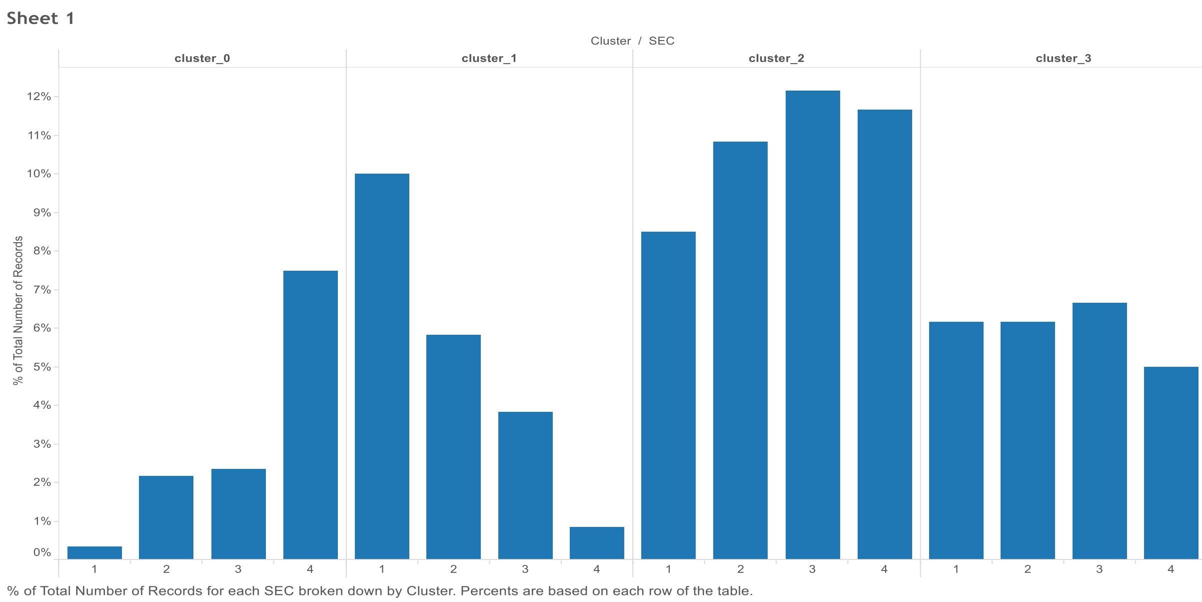
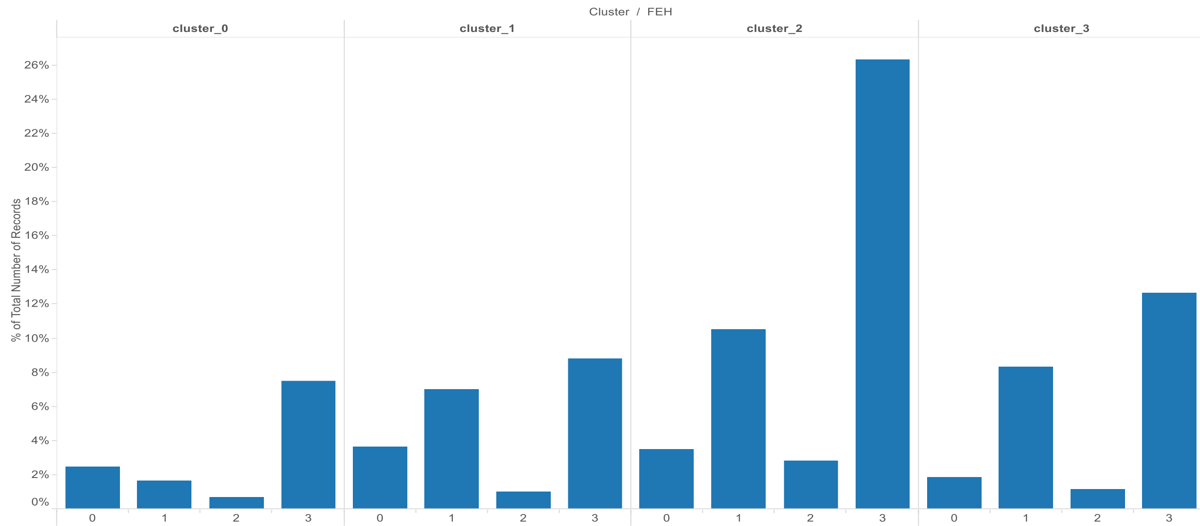


Figure 2.2.1 Socio economic status

Cluster 2 includes the people who show a high tendency to buy premium soaps. From the distribution of socio economic status of the customers buying these soaps, it looks like people buy premium soaps irrespective of the socio-economic status. Another interesting fact is that there is a high percentage of customers of A and B socio economic status in cluster 1, indicating that they prefer to buy any kind of soap but the MaxBrCd is also maximum in cluster 1, which indicates that customers with a high socio economic status do not care about premium or popular soaps, but their brand loyalty is high.

FOOD EATING HABITS

Sheet 1



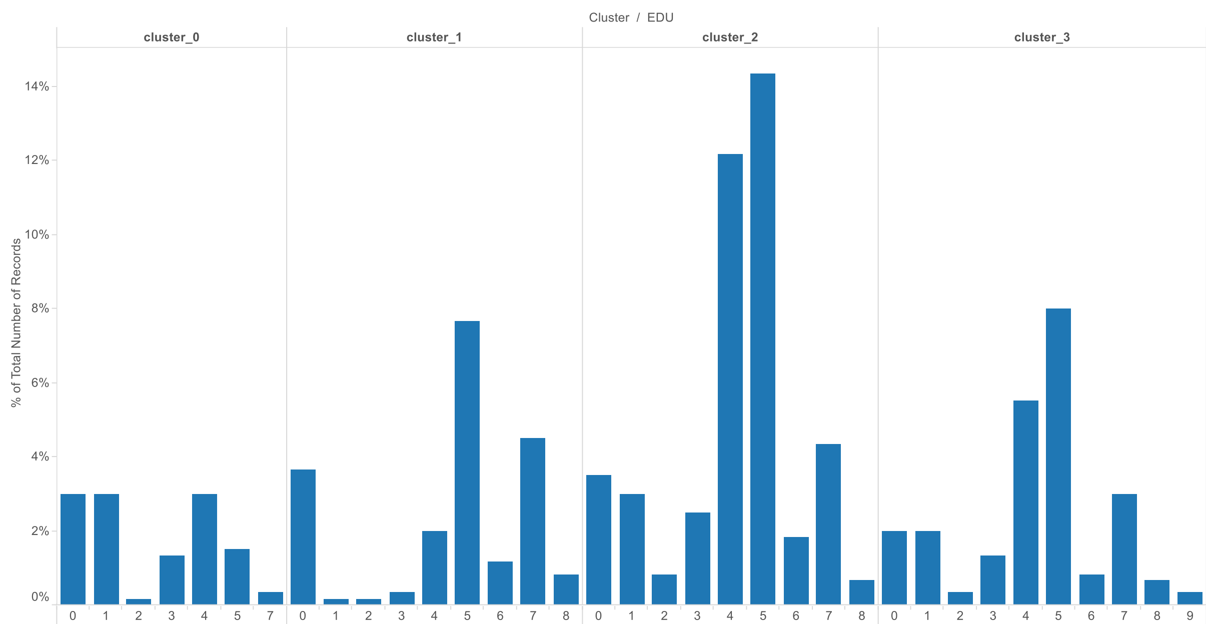
% of Total Number of Records for each FEH broken down by Cluster. Percents are based on each row of the table.

Figure 2.2.2 Food eating habits

Not much trend can be observed in the food eating habits. There is a high percentage of non-vegetarians in cluster 2 which buys value added packs and premium soaps but do not show much brand loyalty. However, we cannot relate the non-vegetarian customers with certainty with purchase.

EDUCATION:

Sheet 1



% of Total Number of Records for each EDU broken down by Cluster. Percents are based on each row of the table.

Figure 2.2.3 Education

Considering education as demographics, it looks like most of the customers in this sample had between 5-12 years in school. There is a high proportion of college graduates in cluster 3 which buys value added packs and premium soaps but do not show much brand loyalty. The proportion of college graduates is also more in cluster 1 which tends to buy with coupons and show a high level brand loyalty.

GENDER:

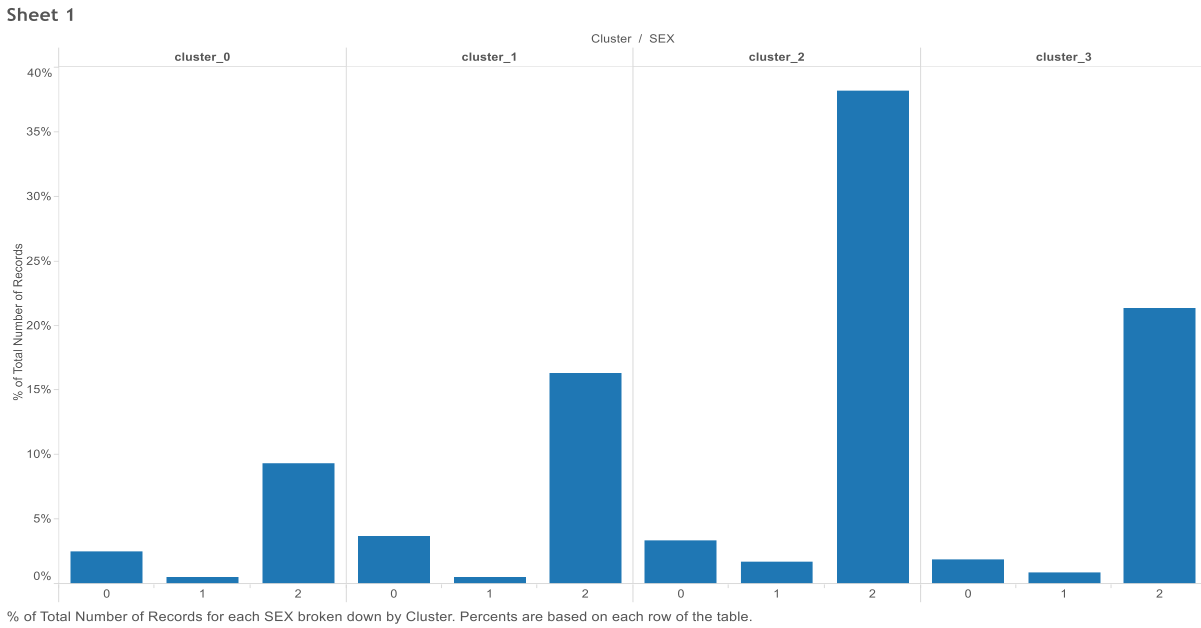


Figure 2.2.4 Gender

A maximum number of customers in each cluster are women. There are maximum women in cluster2 which buys value added packs and premium soaps but do not show much brand loyalty.

AGE:

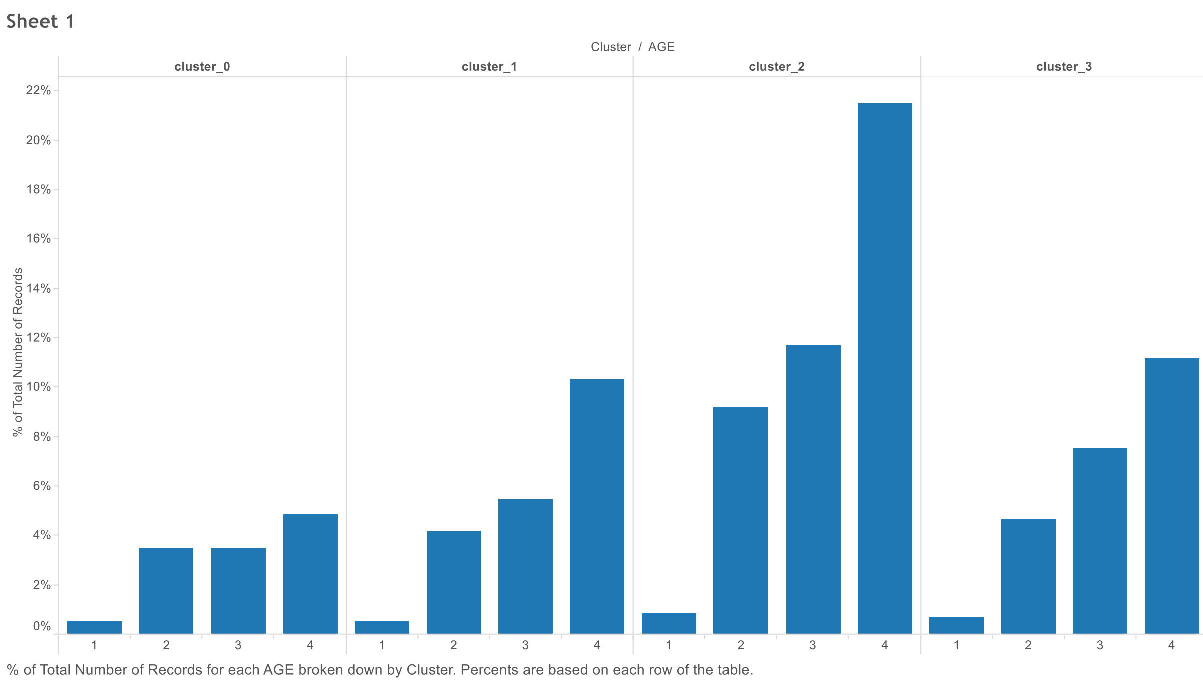


Figure 2.2.5 Age

Across all clusters, the number of people in age group 45+ are more. No significant trend can be observed across the age groups of the customers.

NUMBER OF PEOPLE IN HOUSEHOLD:

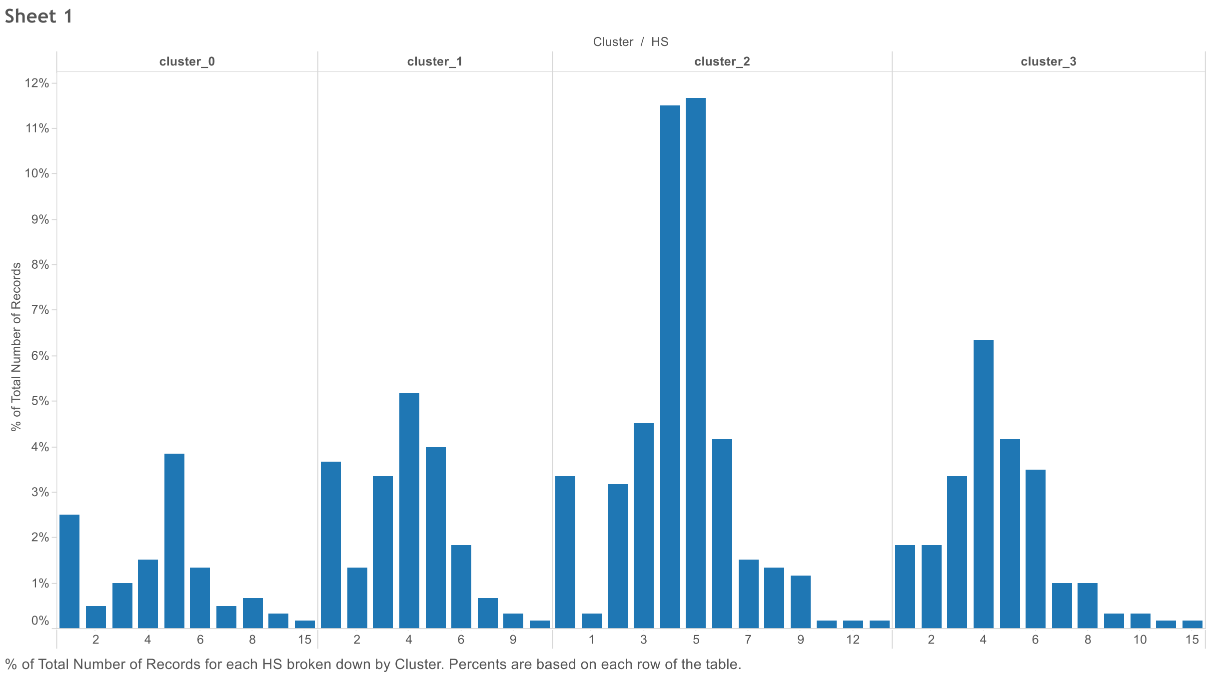


Figure 2.2.6 Household size

Most customers seem to have 4-5 people in the household. It looks these people do not care about brands and that they prefer to buy value packs and premium soaps.

MOTHER TONGUE:

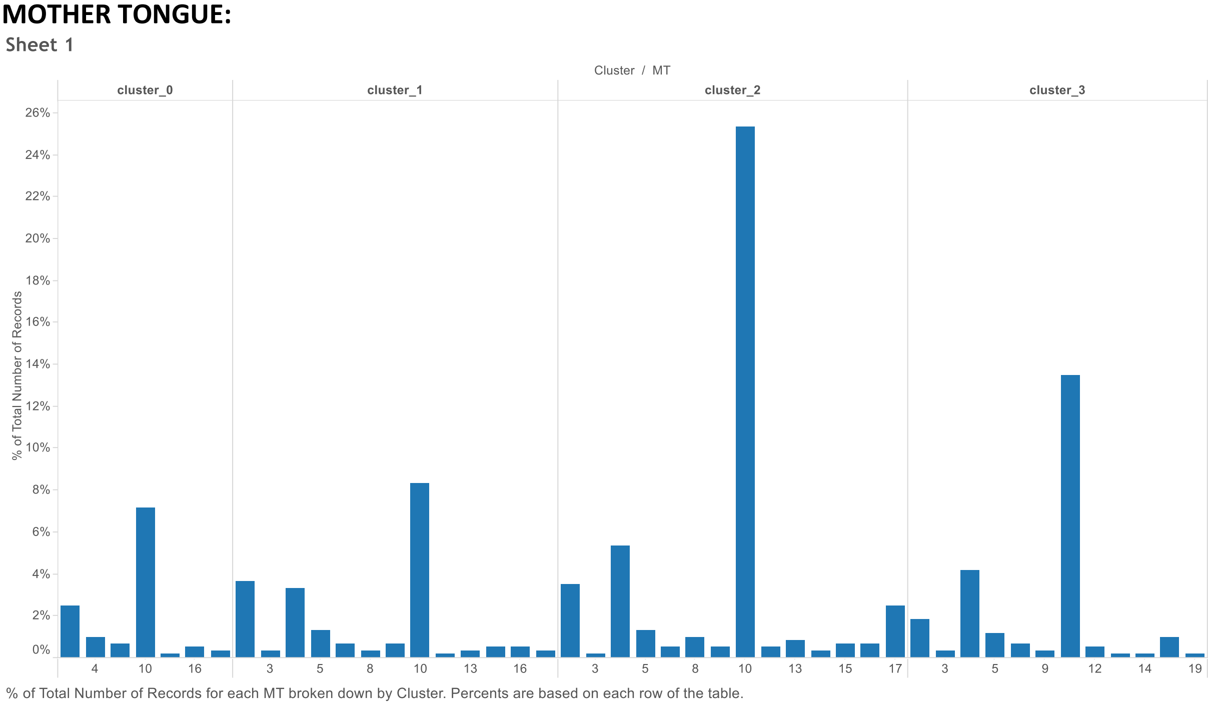
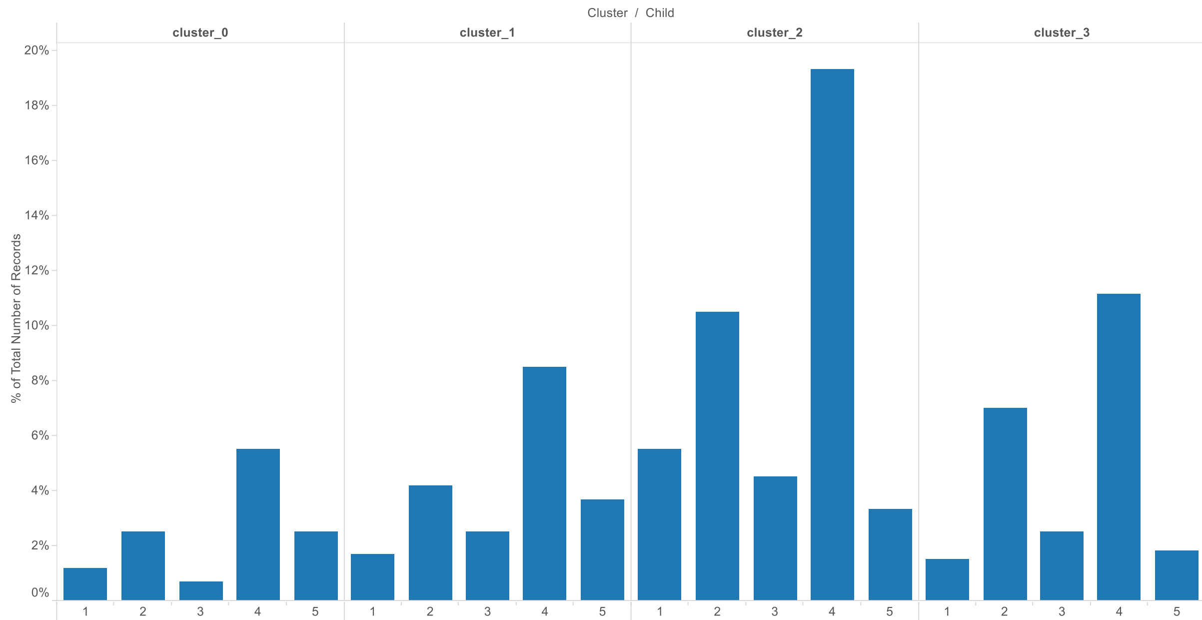


Figure 2.2.7 Mother Tongue

This demographic does not appear to be significant as all clusters are dominated by customer base whose mother tongue is Marathi. Looks like this sample data was gathered from a locality with a predominantly Marathi speaking population.

CHILD:

Sheet 1



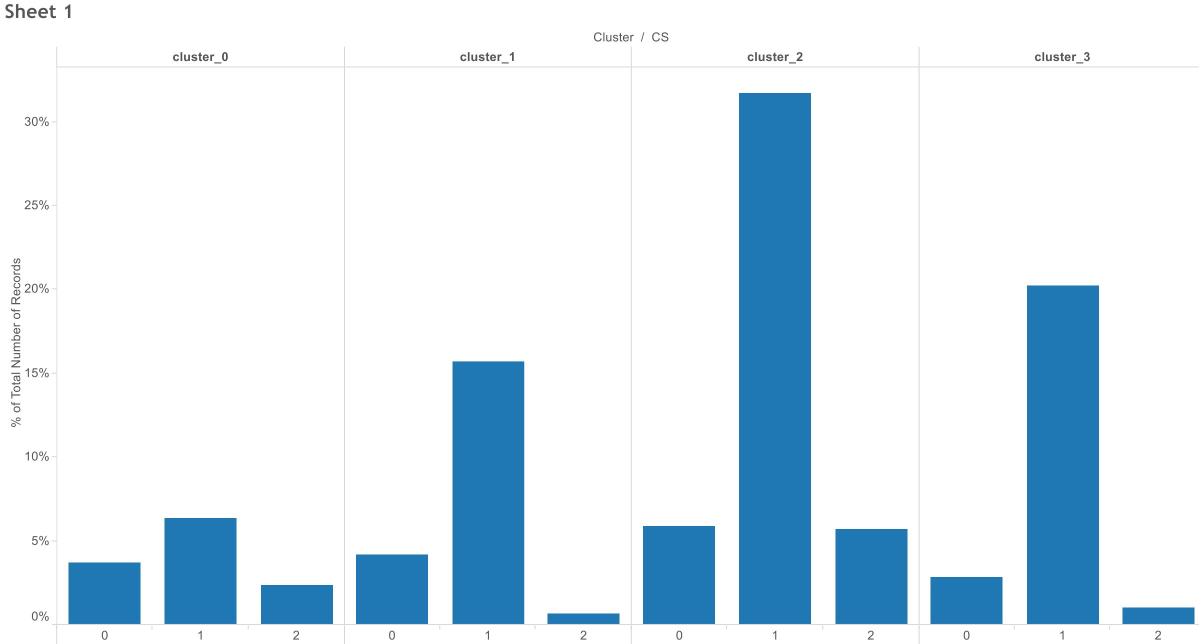
% of Total Number of Records for each Child broken down by Cluster. Percents are based on each row of the table.

Figure 2.2.8 Number of children

Most people in cluster 2 and 3 do not have children. These people , interestingly, seem to have high brand loyalty.

CS:

Sheet 1



% of Total Number of Records for each CS broken down by Cluster. Percents are based on each row of the table.

Figure 2.2.9 Availability of TV

Most people across clusters have a cable TV available.

CONCLUSION:
Most consumers are females, thus most of the ads should be targeted for women. Also, most customers fall in the segment who are not particularly brand loyal but prefer to buy value added packs and premium soaps. As most people have a TV/cable, advertisements can be broadcast on television as an effective means of promoting products.
As evident from cluster 1 and 3, brand loyalty comes in the case when people have an option of exchange offers or coupons. Not many people care about price offs. People buying on basis of price offs are very small across all clusters. Thus, in order to promote brand loyalty, manufacturers should promote their brands by gifting coupons or exchange offers.
From cluster0, we can say that customers who do care about price offs, are those who will buy products on discount irrespective of the brand and they are generally illiterate or do not have any formal schooling.

3. For the best segmentation, obtain a description of the clusters using a decision tree – how effective is the tree in identifying the different clusters? Does the tree help in explaining/interpreting the different clusters?

Answer:

The parameters used to build decision tree are:

Criterion= gain_ratio
Maximal depth= 8
Apply pruning = checked
Confidence = 0.25
Apply prepruning = checked
Minimal gain = 0.1
Minimal leaf size = 2
Minimal size for split = 4
Number of prepruning alternatives = 3

Tree:

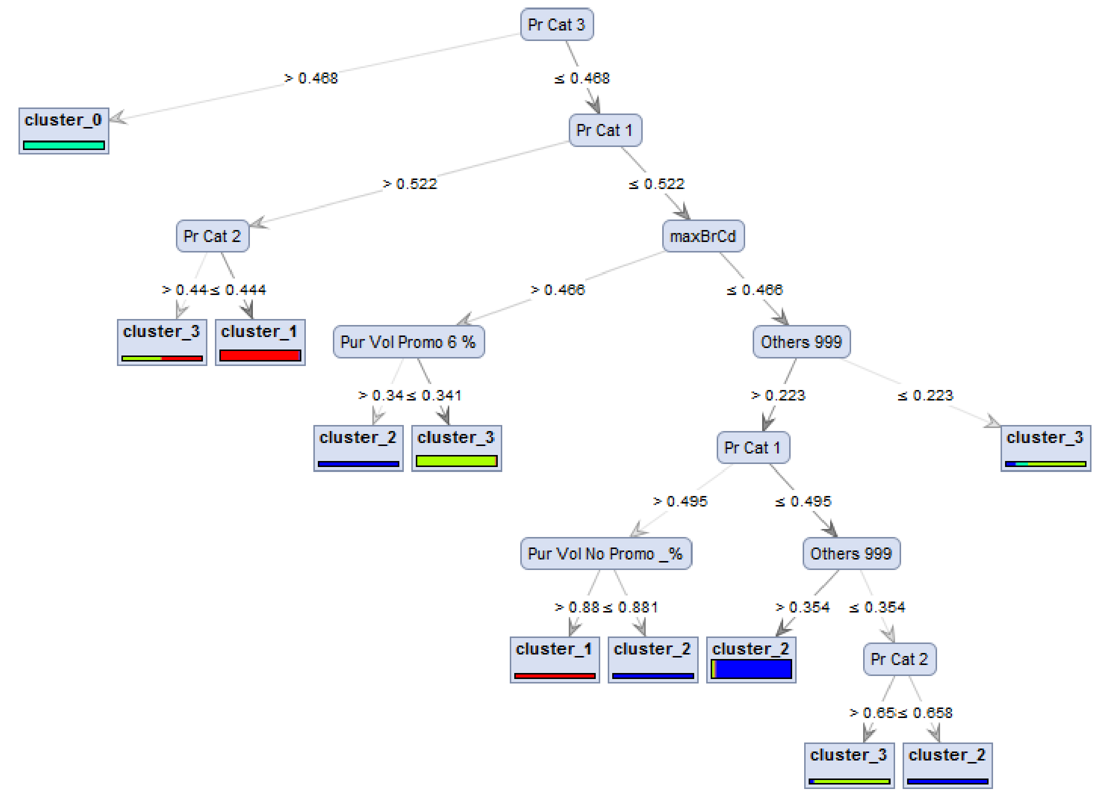


Figure 3.1 Decision tree with k = 4

Performance:

accuracy: 95.83%

	true cluster_2	true cluster_0	true cluster_3	true cluster_1	class precision
pred. cluster_2	259	0	13	4	93.84%
pred. cluster_0	0	73	0	0	100.00%
pred. cluster_3	2	1	127	2	96.21%
pred. cluster_1	1	0	2	116	97.48%
class recall	98.85%	98.65%	89.44%	95.08%	

From the tree, we can say that decision tree model was quite effective in classifying data. Most of the end leaves are dominated by individual clusters. So therefore, the company can target those cluster of customers by following the tree rules. We can also see that the model is not 100% accurate. It has 95.83% accuracy.

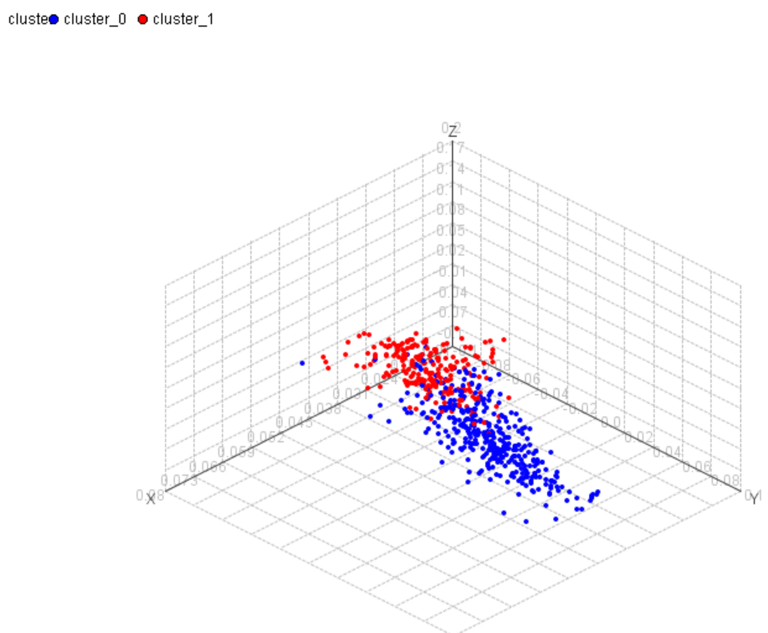
When we compare tree is figure 3.1 and centroid plot at figure 1.3.3, the variables at the top of the tree and the variables with lots of variation in centroid plot are same. Pr Cat 1, Pr Cat 2, Pr Cat 3 are some of the important variables in both decision tree and centroid plot with k = 4.

4. Appendix:

4.1 Scatterplots for variables that describe purchase behavior:

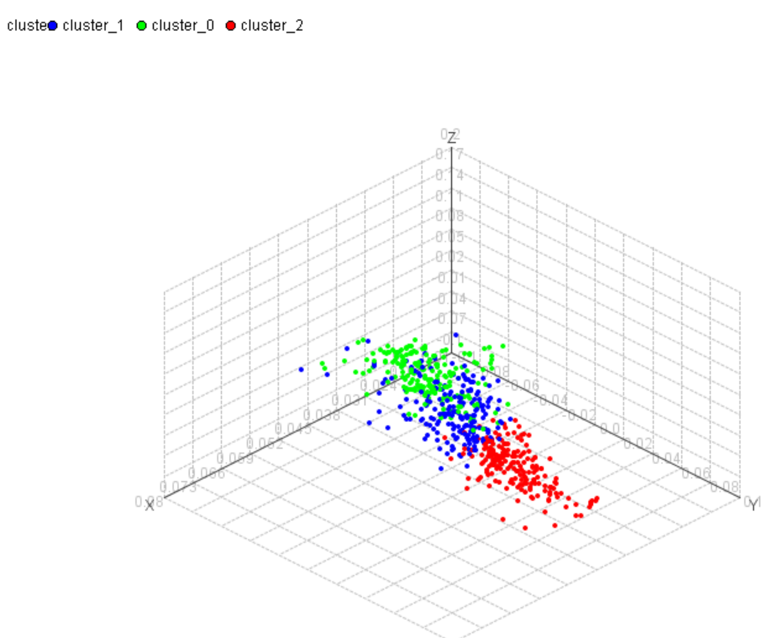
4.1.1 K = 2

[Back to the top]

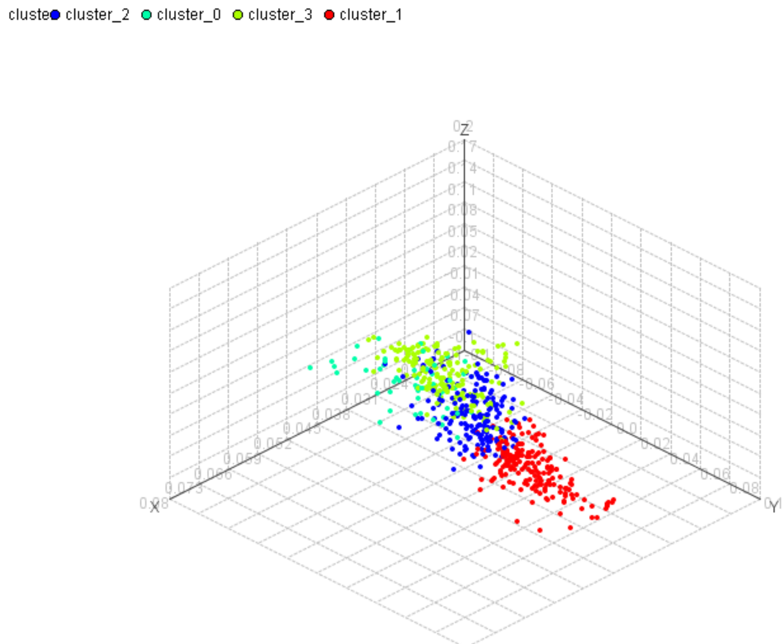


4.1.2 K = 3

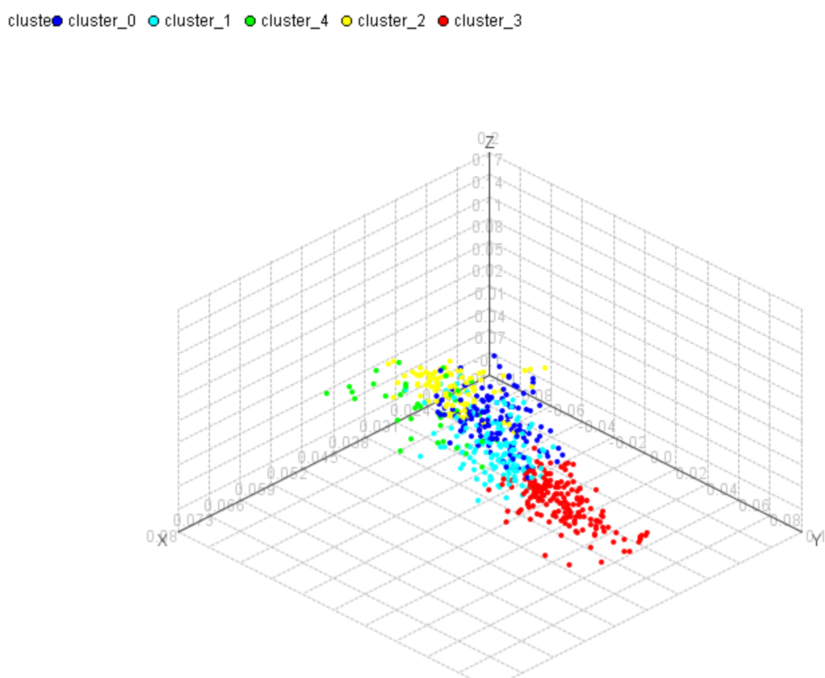
[Back to the top]



4.1.3 K = 4 [Back to the top]



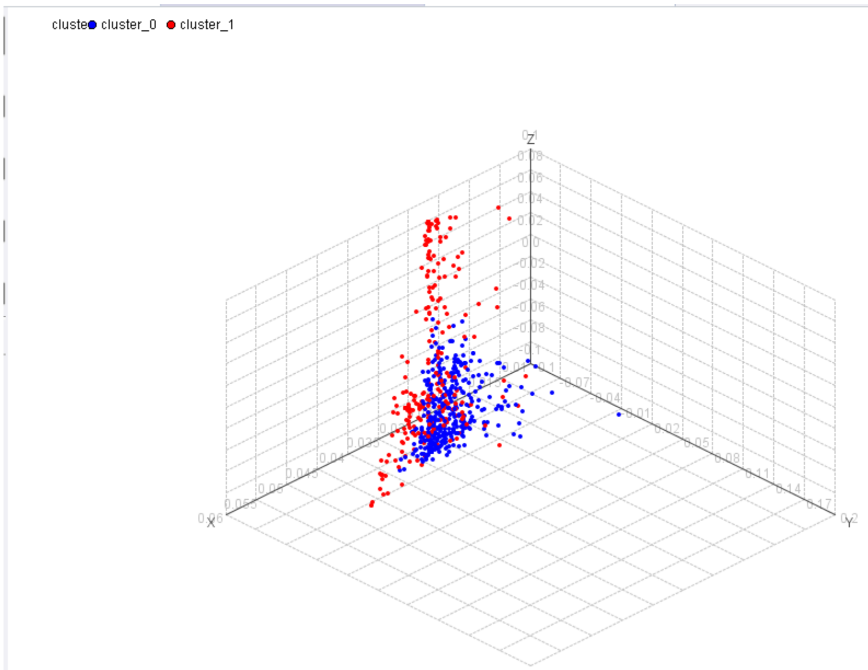
4.1.4 K = 5 [Back to the top]



4.2 Scatterplots for variables that describe basis of purchase:

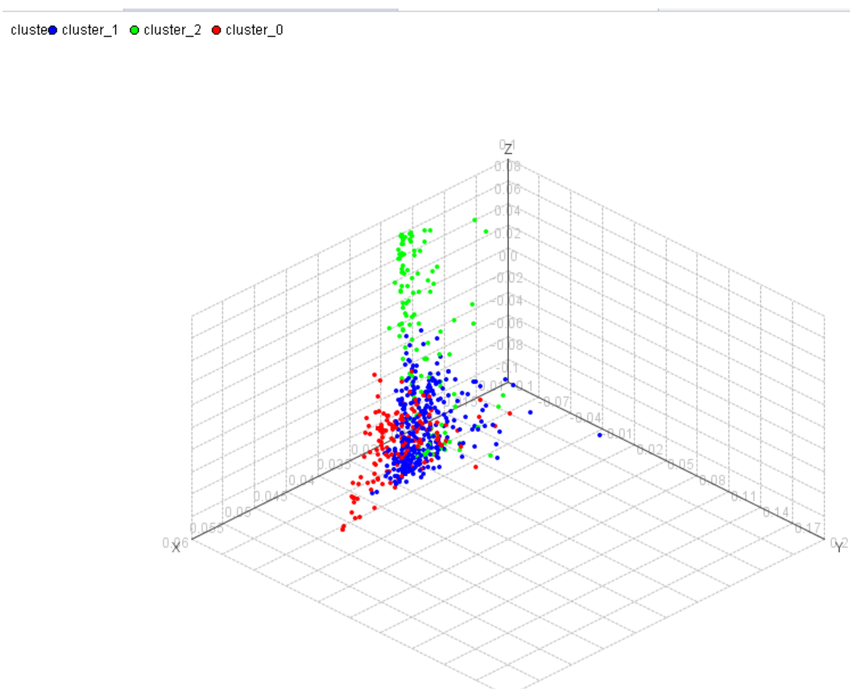
4.2.1 K = 2

[back to the top]



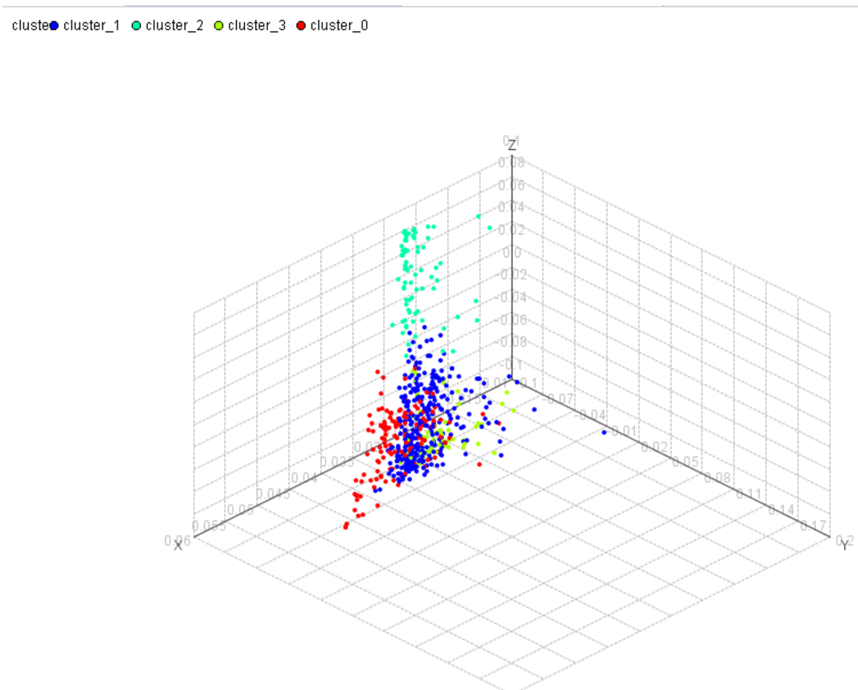
4.2.2 K = 3

[back to the top]



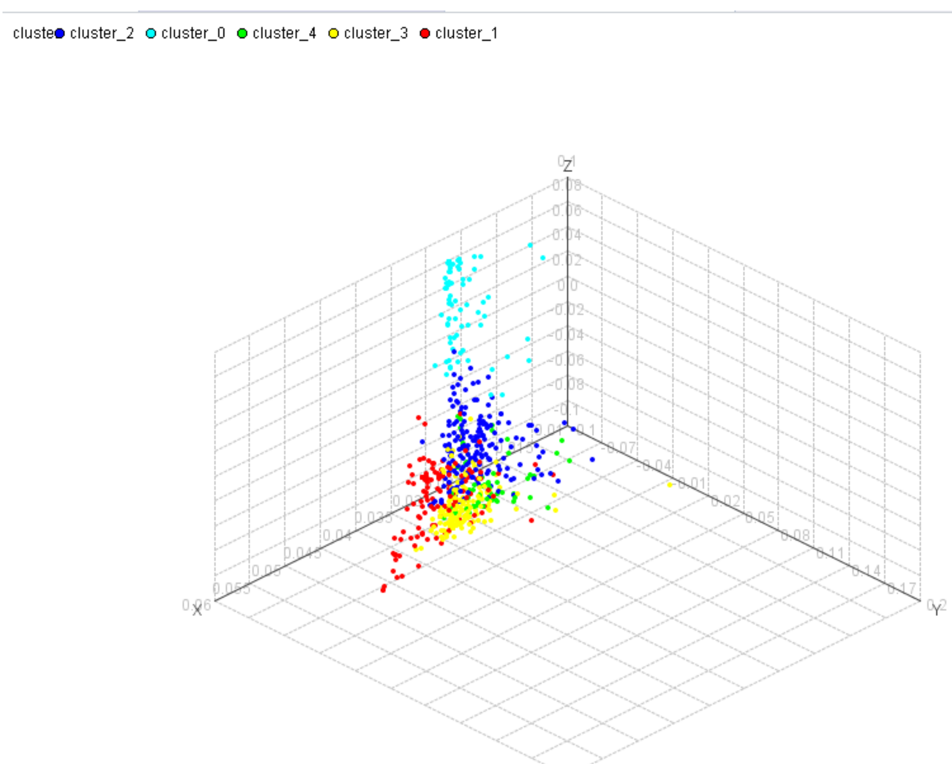
4.2.3 K = 4

[back to the top]



4.2.4 K = 5

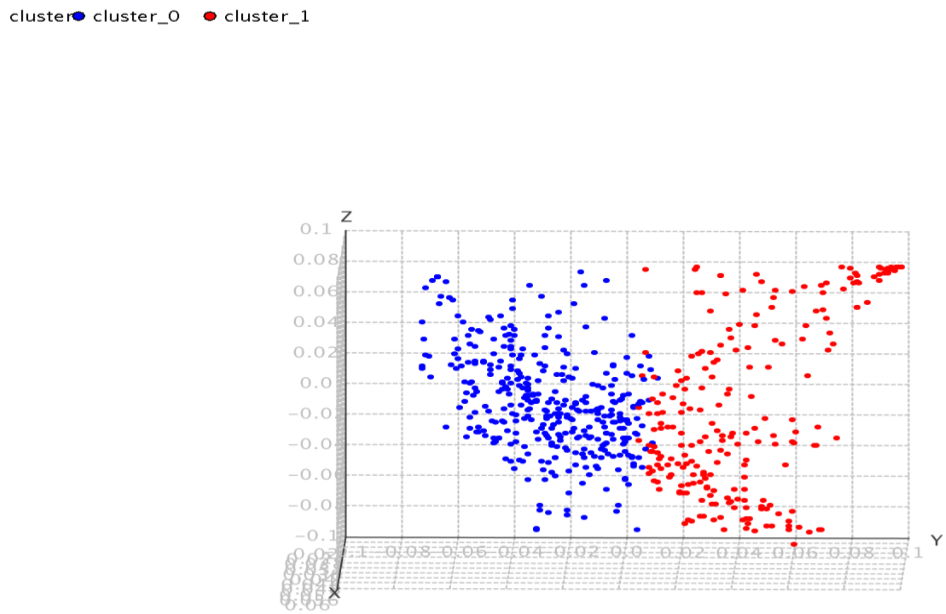
[back to the top]



4.3 Scatterplots for variables that describe both purchase behavior and basis of purchase:

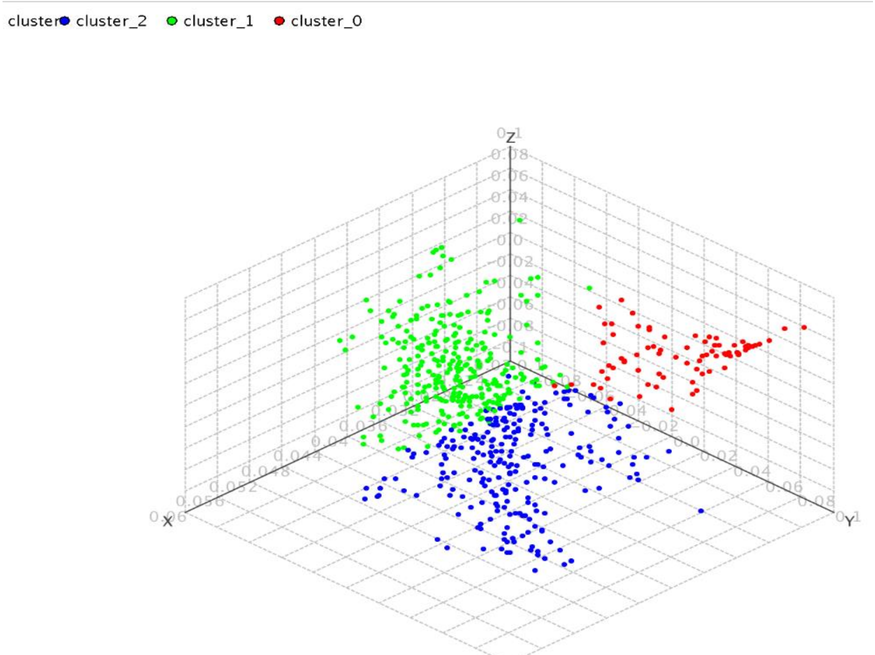
4.3.1 K = 2

[back to the top]



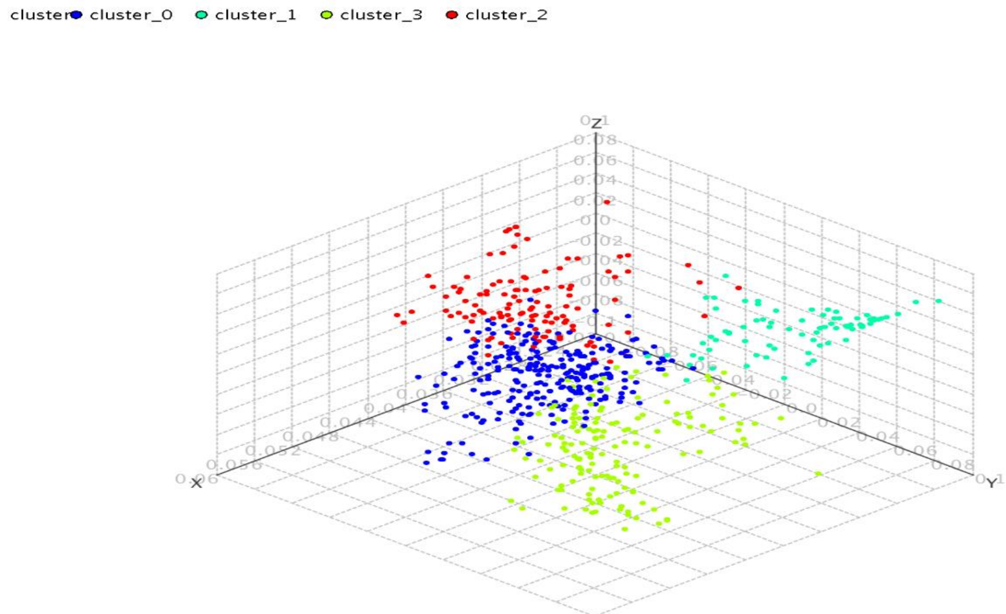
4.3.2 K = 3

[back to the top]



4.3.3 K = 4

[back to the top]



4.3.4 K = 5

[back to the top]