Market Segmentation

IDS 572 Assignment 4

Group Details

NagaShrikanth Ammanabrolu 676837954

Suresh Sappa 667192596

Sagar Kanchi 669850639

Table of Contents

Introduction

Catering primarily for Advertising agencies and consumer product manufacturers, CRISA has been tasked with segmenting the consumer market based on the demographics' purchase behavior and their basis of purchase, in addition to the consumer demographics. Such a market segmentation will help CRISA to target appropriate segments of customers with cost-effective promotions.

Question No. 1)
Solution.

Using k-means clustering to Identify clusters of households

Identifying clusters of households based on the customer demographics and purchase behaviors is primarily supported by the value of k in the k-means clustering algorithm. Two of the factors of concern are the within cluster distance from centroid and the between cluster centroid distance for optimal customer market segments. Reducing the within cluster distance of points from the centroid and maximizing the distance between clusters is how effective segments could be obtained. Also, obtaining an equally weighted distribution of the clusters is essential, so that there do not clusters significantly smaller as compared to the other clusters. For the purpose of Normalization, we've used Range Transformation.

Part (a) Solution

There are a total of 600 customers and supporting details about their purchasing behavior is described in the given dataset. Based on these attributes of the 600 customers, we will be trying to come up with a clustering model to adequately segregate the consumer market. Below listed are the variables that we will be using to identify the clusters of households based on their purchasing behavior and patterns.

1. **Total number of Brands**
2. **Brand Runs**
3. **Total Volume**
4. **Number of Transactions**
5. **Value**
6. **Average Price**
7. **Share to other Brands (Others999 attribute)**
8. **Brand Loyalty (MaxBrLoyalty) – Defined as the Maximum value among the attributes: Br. Cd. 57_144, Br. Cd. 55, Br. Cd. 272, Br. Cd. 286, Br. Cd. 24, Br. Cd. 481, Br. Cd. 352, and Br. Cd. 5. This would help us identify specific brands to which a customer is particularly loyal to.**

We now identify the clusters of households based on their purchasing behavior as described by the subset of variables mentioned above. The different values of K and their performance based on the clusters are described below.

K = 2

For k = 2, we obtain two clusters of households. From the centroid plot obtained for the clustering model when k = 2, we can say that the clusters are compact in nature. The clusters are adequately separable and although the data points are a bit spread out, they are clearly distinguishable from one another. Moreover, from the Cluster Model, it could be said the clusters are almost evenly distributed. From the Cluster model plot, it could be concluded that for k = 2, Cluster 1 exhibit a higher Brand Loyalty.

Performance Vector of Clusters	
Average within Centroid Distance	0.180
Average within Centroid Distance (Cluster 0)	0.187
Average within Centroid Distance (Cluster 1)	0.169
Davies Bouldin Index	1.062

Cluster Model	
Cluster 0	373 items
Cluster 1	227 items

K = 3

For k = 3, we obtain three clusters of households. From the centroid plot obtained for the clustering model when k = 3, we can say that the clusters are distinguishable from one another. Cluster 1 exhibits a more distributed nature as compared to the other two clusters. By comparing the cluster model plot to k = 2, we could see that the cluster 1 has been divided into two clusters, cluster 0 and cluster 1. Clusters 0 and 2 have many overlapping features, although cluster 0 exhibits a significantly higher Brand Loyal customer.

Performance Vector of Clusters	
Average within Centroid Distance	0.140
Average within Centroid Distance (Cluster 0)	0.142
Average within Centroid Distance (Cluster 1)	0.160
Average within Centroid Distance (Cluster 2)	0.119
Davies Bouldin Index	1.255

Cluster Model	
Cluster 0	197 items
Cluster 1	204 items
Cluster 2	199 items

K = 4

For k = 4, we obtain four clusters of households. From the centroid plot obtained for the clustering model when k = 4, we can say that the clusters 0 and 2 exhibit overlapping features, while the rest of the clusters are quite distinguishable from one another. By comparing this cluster model plot to when k = 3, we could see that the cluster 0 has been decomposed into two clusters, cluster 0 and cluster 2. Here, Cluster 3 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we can see that Cluster 0 consists of a significantly low number of cases as compared to the other evenly distributed clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.127
Average within Centroid Distance (Cluster 0)	0.200
Average within Centroid Distance (Cluster 1)	0.116
Average within Centroid Distance (Cluster 2)	0.125
Average within Centroid Distance (Cluster 3)	0.121
Davies Bouldin Index	1.345

Cluster Model	
Cluster 0	51 items
Cluster 1	191 items
Cluster 2	181 items
Cluster 3	177 items

K = 5

For k = 5, we obtain five clusters of households. From the centroid plot obtained for the clustering model when k = 5, we can say that the clusters 0, 2 and 3 exhibit features that are quite distinguishable from one another. Clusters 1 and 4 are more scattered and are comparatively more overlapping with other clusters. Here, Cluster 2 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we see that Cluster 4 consists of a significantly low number of cases as compared to other clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.110
Average within Centroid Distance (Cluster 0)	0.080
Average within Centroid Distance (Cluster 1)	0.116
Average within Centroid Distance (Cluster 2)	0.093
Average within Centroid Distance (Cluster 3)	0.113
Average within Centroid Distance (Cluster 4)	0.217
Davies Bouldin Index	1.251

Cluster Model	
Cluster 0	135 items
Cluster 1	145 items
Cluster 2	100 items
Cluster 3	180 items
Cluster 4	40 items

(*Please Refer Appendix for Further Reference*)

Part (b) Solution

Next, we have to identify clusters of households based on the Basis for Purchase variables. Below listed are the variables that we will be using to identify the clusters of households based on basis of purchase.

1. **Percent of volume purchased not on promotion (pur_vol_no_promo)**
2. **Percent of volume purchased on promo code 6 (pur_vol_promo_6)**
3. **Percent of volume purchased on promo code other than 6 (pr_vol_other)**
4. **All 4 Price Categories**
5. **Selling Propositions (PropCat5 to PropCat9 and PropCat14)**

Although we consider all of the attributes from the above mentioned set of variables, we further select only a few attributes from the Selling proposition variable. After exploring the dataset, we find that many of the Proposition categories have null (0.0%) values associated with them. Hence, we selected a threshold of 60% for these null values, above which all of the categories have been taken into consideration. This leaves us with selecting Proposition categories 5 through 9 and the Proposition category 14 attribute.

The different values of K and their performance based on the clustering are described below.

K = 2

For k = 2, we obtain two clusters of households. From the centroid plot obtained for the clustering model when k = 2, we can say that the cluster 0 is more compact in nature, while cluster 1 is more spread out and the clusters are adequately distinguishable from one another. Moreover, from the Cluster Model, it could be said the clusters are almost evenly distributed. From the Cluster model plot, it could be concluded that for k = 2, Cluster 1 exhibits a higher Brand Loyalty. The clusters could not be said to be evenly distributed as from the Cluster model, we see that cluster 1 contains significantly low number of cases as compared to cluster 0.

Performance Vector of Clusters	
Average within Centroid Distance	0.443
Average within Centroid Distance (Cluster 0)	0.484
Average within Centroid Distance (Cluster 1)	0.170
Davies Bouldin Index	0.844

Cluster Model	
Cluster 0	522 items
Cluster 1	78 items

K = 3

For k = 3, we obtain three clusters of households. From the centroid plot obtained for the clustering model when k = 3, we can say that the clusters are distinguishable from one another. Cluster 1 exhibits a more distributed nature as compared to the other two clusters. By comparing the cluster model plot to k = 2, we could see that the cluster 0 has been divided into two clusters, cluster 0 and cluster 2. All three of the Clusters have many overlapping features, although cluster 1 exhibits a significantly higher Brand Loyal customer.

Performance Vector of Clusters	
Average within Centroid Distance	0.348
Average within Centroid Distance (Cluster 0)	0.374
Average within Centroid Distance (Cluster 1)	0.174
Average within Centroid Distance (Cluster 2)	0.374
Davies Bouldin Index	1.287

Cluster Model	
Cluster 0	373 items
Cluster 1	79 items
Cluster 2	148 items

K = 4

For k = 4, we obtain four clusters of households. From the centroid plot obtained for the clustering model when k = 4, we can say that the clusters 2 and 3 exhibit overlapping features, while the rest of the clusters are quite distinguishable from one another. Here, Cluster 3 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we can see that Cluster 2 and 3 consists of a significantly low number of cases as compared to the other evenly distributed clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.293
Average within Centroid Distance (Cluster 0)	0.370
Average within Centroid Distance (Cluster 1)	0.300
Average within Centroid Distance (Cluster 2)	0.230
Average within Centroid Distance (Cluster 3)	0.174
Davies Bouldin Index	1.178

Cluster Model	
Cluster 0	141 items
Cluster 1	321 items
Cluster 2	59 items
Cluster 3	79 items

K = 5

For k = 5, we obtain five clusters of households. From the centroid plot obtained for the clustering model when k = 5, we can say that the clusters 0, 1 and 2 exhibit features that are quite distinguishable from one another. Clusters 0 and 4 are more scattered and are comparatively more overlapping with other clusters. Here, Cluster 0 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we see that Cluster 4 consists of a significantly low number of cases as compared to other clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.249
Average within Centroid Distance (Cluster 0)	0.159
Average within Centroid Distance (Cluster 1)	0.349
Average within Centroid Distance (Cluster 2)	0.142
Average within Centroid Distance (Cluster 3)	0.335
Average within Centroid Distance (Cluster 4)	0.218
Davies Bouldin Index	1.296

Cluster Model	
Cluster 0	75 items
Cluster 1	120 items
Cluster 2	174 items
Cluster 3	176 items
Cluster 4	55 items

(*Please Refer Appendix for Further Reference*)

Part (c) Solution

Next, we have to identify clusters of households based on the combined variables of Purchasing behavior and Basis for Purchase.

The different values of K and their performance based on the clustering are described below.

K = 2

For k = 2, we obtain two clusters of households. From the centroid plot obtained for the clustering model when k = 2, we can say that the cluster 0 is more compact in nature, while cluster 1 is more spread out and the clusters are adequately distinguishable from one another. From the Cluster model plot, it could be concluded that for k = 2, Cluster 1 exhibits a higher Brand Loyalty. The clusters could not be said to be evenly distributed as from the Cluster model, we see that cluster 1 contains significantly low number of cases as compared to cluster 0.

Performance Vector of Clusters	
Average within Centroid Distance	0.736
Average within Centroid Distance (Cluster 0)	0.667
Average within Centroid Distance (Cluster 1)	0.877
Davies Bouldin Index	1.876

Cluster Model	
Cluster 0	401 items
Cluster 1	199 items

K = 3

For k = 3, we obtain three clusters of households. From the centroid plot obtained for the clustering model when k = 3, we can say that the clusters are distinguishable from one another. Cluster 2 exhibits a more distributed nature as compared to the other two clusters. By comparing the cluster model plot to k = 2, we could see that the cluster 1 has been divided into two clusters, cluster 0 and cluster 2. All three of the Clusters have many overlapping features, although cluster 2 exhibits a significantly higher Brand Loyal customer.

Performance Vector of Clusters	
Average within Centroid Distance	0.581
Average within Centroid Distance (Cluster 0)	0.545
Average within Centroid Distance (Cluster 1)	0.729
Average within Centroid Distance (Cluster 2)	0.284
Davies Bouldin Index	1.637

Cluster Model	
Cluster 0	306 items
Cluster 1	221 items
Cluster 2	73 items

K = 4

For k = 4, we obtain four clusters of households. From the centroid plot obtained for the clustering model when k = 4, while the rest of the clusters are quite distinguishable from one another, cluster 2 exhibits a sparser distribution. Here, Cluster 2 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we can see that Cluster 2 consists of a significantly low number of cases as compared to the other clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.511
Average within Centroid Distance (Cluster 0)	0.559
Average within Centroid Distance (Cluster 1)	0.430
Average within Centroid Distance (Cluster 2)	0.290
Average within Centroid Distance (Cluster 3)	0.638
Davies Bouldin Index	1.662

Cluster Model	
Cluster 0	260 items
Cluster 1	143 items
Cluster 2	74 items
Cluster 3	123 items

K = 5

For k = 5, we obtain five clusters of households. From the centroid plot obtained for the clustering model when k = 5, we can say that the clusters 0, 2, 3 and 4 exhibit features that are quite distinguishable from one another. Cluster 1 is more scattered and comparatively more overlapping with other clusters. Here, Cluster 2 exhibits a significantly higher Brand Loyal customer. From the cluster model table described below, we see that Clusters 1 and 2 consists of a significantly low number of cases as compared to other clusters.

Performance Vector of Clusters	
Average within Centroid Distance	0.453
Average within Centroid Distance (Cluster 0)	0.476
Average within Centroid Distance (Cluster 1)	0.316
Average within Centroid Distance (Cluster 2)	0.284
Average within Centroid Distance (Cluster 3)	0.416
Average within Centroid Distance (Cluster 4)	0.625
Davies Bouldin Index	1.491

Cluster Model	
Cluster 0	233 items
Cluster 1	53 items
Cluster 2	73 items
Cluster 3	131 items
Cluster 4	110 items

(*Please Refer Appendix for Further Reference*)

Part (d) Solution

To begin with, we will be considering variables from the Part (a) of Question no. 1, i.e., the variables defining the purchasing behavior of customers. We considered this subset, since it gives the lowest within cluster distance from the centroid and the centroid plot shows that this subset performs the best under k-means clustering.

(Please refer the Appendix for further details of the parameters for the different algorithms, distance between the clusters, the 3D distribution and the cluster model plot.)

k-Medoids

The performance vectors and the cluster distribution for the model are as described below.

Performance Vector of Clusters	
Average within Centroid Distance	0.155
Average within Centroid Distance (Cluster 0)	0.128
Average within Centroid Distance (Cluster 1)	0.263
Average within Centroid Distance (Cluster 2)	0.121
Average within Centroid Distance (Cluster 3)	0.175
Average within Centroid Distance (Cluster 4)	0.128
Davies Bouldin Index	1.612

Cluster Model	
Cluster 0	74 items
Cluster 1	69 items
Cluster 2	141 items
Cluster 3	163 items
Cluster 4	153 items

For getting a model defined by k-medoids algorithm, we arrive at the said parameters after working through the different values of k from 2 to 5. As we increase the value of k from 2 to 5, we find that the distance within the cluster goes on reducing and hence we obtain the value of k to be 5 in the parameters defining the k-medoids model. Other parameters which we've tried but didn't yield us better results were obtained by varying the max runs and the max optimization steps to 5 and 50 respectively.

Although an optimum model could be obtained using k-medoids algorithm, we find that it still yields a poorer distribution of the clusters as compared to the k-means clustering algorithm.

Kernel k-means

The cluster distribution for the model are as described below.

Cluster Model	
Cluster 0	74 items
Cluster 1	69 items
Cluster 2	141 items
Cluster 3	163 items
Cluster 4	153 items

For getting a model defined by the kernel k-means algorithm, we arrive at the said parameters after working through the different values of k from 2 to 5. As we increase the value of k from 2 to 5, we find that the distance within the cluster goes on reducing. But as we increase k further above 5, we find that there is no clear segmentation between different clusters and reduces the distance between clusters parameter significantly. Since we require a model where the data points within a cluster are compact, while increasing the distance between clusters, we arrive at the value of k as 5.

Other parameters which we've tried but didn't yield us better results were obtained by varying the kernel type to radial, polynomial and sigmoid.

Kernel k-means does give us an optimal model, where there is a clear separation between clusters, while at the same time not compromising on the compactness of data points within a cluster.

Agglomerative Clustering

The cluster distribution for the model are as described below.

Cluster Model	
Cluster 0	20 items
Cluster 1	326 items
Cluster 2	13 items
Cluster 3	114 items
Cluster 4	127 items

Agglomerative clustering is a bottom-up approach of hierarchical clustering and here, as seen from the above table, we obtain poor results. The distribution of data points within clusters are not properly distributed and there are a few clusters with either too many or too little data points. Also from the 3d Clustering plot, we see that the model yields poorly distributed clusters and the within clusters distance is also significantly high.

For getting a model defined by agglomerative clustering algorithm, we arrive at the said parameters after changing the mode from Single link to Complete link. The other parameter that we've considered is the value of the k parameter in the 'Flatten cluster' operator and changed the values from 2 to 5. Again, since we require a model where the data points within a cluster are compact, while increasing the distance between clusters, we arrive at the value of k as 5.

Agglomerative clustering on the given dataset yields us poorer results as compared to the kernel k-means, k-medoids and the k-means clustering algorithms. Since the clusters obtained through agglomerative clustering are sparsely distributed and with a higher within cluster distance, we do not consider this model to be optimal for our analysis.

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a clustering algorithm that also takes into consideration the noise/errors present in the given dataset.

The cluster distribution for the model are as described below.

Cluster Model	
Cluster 0	8 items
Cluster 1	592 items

We've tried changing all of the different parameters of the DBSCAN clustering algorithm including epsilon, min points, and the different measure types. We've tried different values of epsilon ranging from 0.5 to 2.0 and we found that increasing the epsilon only gave us very low number of clusters. Hence we set our epsilon to 0.5. On different values of the min points attribute like 5 (default), 10, 12, 15, 20, 50 and 100; we found the min. points of 20 to give us a model where the clusters are not too sparsely distributed. On an overall, DBSCAN clustering algorithm yields us the poorest results for our given dataset.

Conclusion

From the above observations, we can conclude that different clustering algorithms with varied parameters yield us clusters that are comparatively different from one another.

Clustering is basically about assigning data points to a plane in a hyperspace and since different clustering algorithms define different ways to segregate these points, the models yielded from different algorithms yield us varied results. Moreover, since there are several different ways to measure distance in an n-dimensional hyperspace, we get varied results.

Considering the fact that we have to reduce the within cluster distance and maximize the distance between clusters, k-means is clearly yielding us the best possible results. Clear segmentation of different clusters and the compactness of data points within the cluster make k-means clustering algorithm our 'best' model as compared to the other models.

Question No. 2)
Solution.

Part (a) Solution

As observed in the Question no. 1 Part D, we can conclude that k-means clustering yield us the best possible results for performing market segmentation on our given dataset. As compared to other models like the k-medoids, kernel k-means, agglomerative clustering and DBSCAN clustering, a clustering model obtained through k-means gives us 'good' clusters. It is observed that k-means clustering yield us a model of clusters where the intra-cluster distance is low and the inter-cluster distance is high. Moreover, we find that the clusters obtained through k-means are evenly distributed, except for a single cluster.

We find that the k-means clustering model applied to the customer purchasing behavior attributes yield us our 'best' model. Such kind of model makes the clusters obtained significantly efficient to identify meaningful patterns in the customer behavior.

Now that we have identified k-means clustering as the 'best' fit model for our given business problem, we compare the average within cluster centroid distance of the different k-means models obtained through different values of k.

Comparison of Different k-means clustering models based on the value of k	
Value of k	**Average within Centroid Distance**
2	0.180
3	0.140
4	0.127
5	0.110

From the above table and the 3d plot for the model, we can clearly see that k-means clustering with k as 5 is our 'best' model for the given dataset. Now, with k as 5 for the k-means clustering, we find that there are 3 clusters (out of 5) of households which yield us a higher Brand Loyalty and hence our deeming of the k-means clustering as our best model is justified.

Part (b) Solution

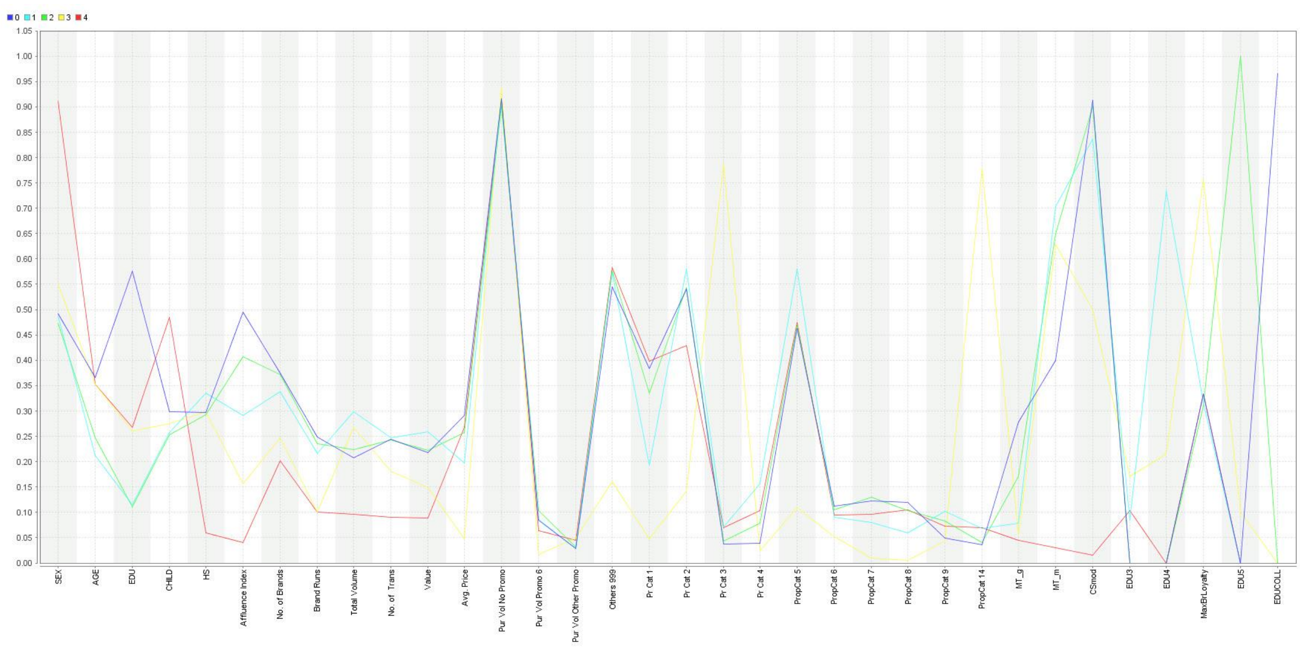


Fig 2.1 Centroid plot of the 5 clusters aggregated across the entire data set

From the previous part, we obtain five clusters of households which are clustered or segregated to the best possible level. The next step is now to take into consideration all of the factors, including the purchasing behavior, basis for purchase, and demographics to aggregate the data across all of the clusters.

While the required data transformation has already been performed on the purchasing behavior and the basis for purchase attributes previously, the demographics ought to be transformed before any further analysis could be performed. The centroid plot of the clusters thus obtained is illustrated above.

(Please refer the Appendix for details regarding the data transformation for demographics)

Cluster 0: 115 Households

Demographics

- More number of people are of higher ages i.e. Greater than 45 years olds
- Highest Affluence index
- Medium household size and highest with Gujarati as native language
- Highly educated with greater numbers possessing a college degree or higher

Purchasing Behavior

- Second-highest Brand Loyalty with highest average price, number of brands and brand runs

Basis for Purchase

- Highest percent of volume purchased under the Any health and Any freshness proposition category
- Second-highest percent of volume purchased under the Any premium soaps price category

Cluster 1: 165 Households

Demographics

- Smallest age group i.e. More number of people below 24 years of age
- Highest number of people with Marathi as their native language
- Medium Affluence index with second-highest educated cluster

Purchasing Behavior

- Low Brand Loyalty with highest number of transactions, volume and value

Basis for Purchase

- Highest percent of volume purchased under the Any popular soap and Any sub-popular price category
- Second-highest percent of volume purchased under the Any beauty and Any hair proposition category

Cluster 2: 182 Households

Demographics

- Medium-educated with most no. of people possessing 10-12 years of school
- Second-highest Affluence Index

Purchasing Behavior

- Lowest Brand Loyalty (second-highest support for Others 999) with second-highest number of brands, brand runs and value

Basis for Purchase

- Highest percent of volume purchased under Any Herbal proposition category
- Second-Highest percent of volume purchased under the Any Popular Soap price category

Cluster 3: 70 Households

Demographics

- Low Affluence Index with second-last in terms of educated cluster
- More number of people falling in the Median age group

Purchasing Behavior

- Highest Maximum Brand Loyalty with second-highest total volume
- Lowest Average price

Basis for Purchase

- Highest percent of volume purchased under Any Economy/Carbolic price category
- Highest percent of volume purchased under Any Carbolic proposition category

Cluster 4: 68 Households

Demographics

- Lowest Affluence Index with low Education
- Least number of people with either Marathi or Gujarati as native languages

Purchasing Behavior

- Medium Brand Loyalty with highest support for ‘Others 999’ brands
- Second-highest average price with lowest No. of brands, brand runs, total volume, number of transactions and volume

Basis for Purchase

- Highest percent of volume purchased under Any Premium soaps price category

- Second-Highest percent of volume purchased under the Any Beauty, Any Freshness and Any Carbolic proposition category

From the above characteristics, we can conclude that Clusters 0, 3 and 4 should be targeted since they yield the highest Brand Loyalty.

Question No. 3)

Solution.

Based on the best segmentation that we've obtained through the previous steps, we now implement a decision tree to further interpret the clusters and help us choose the 'best' clustering.

(Please refer the Appendix for additional details regarding Decision Tree parameters)

Describing the Decision Tree rules for the clusters developed:

Cluster 0

- IF Brand Loyalty <= 0.7 and Share to other brands <= 0.628 and Value > 157.750 and Total Volume <= 20137.5 and Brand Runs <= 19.5 THEN Cluster 0

Cluster 1

- IF Brand Loyalty <= 0.7 and Share to other brands <= 0.628 and Value > 157.750 and Total Volume <= 20137.5 and Brand Runs > 19.5 THEN Cluster 1
- IF Brand Loyalty <= 0.7 and Share to other brands > 0.628 and Number of Brands > 4.5 and Brand Runs > 15 THEN Cluster 1
- IF Brand Loyalty <= 0.7 and Share to other brands > 0.628 and Number of Brands <= 4.5 and Number of Transactions > 66 THEN Cluster 1
- IF Brand Loyalty <= 0.7 and Share to other brands > 0.628 and Number of Brands <= 4.5 and Number of Transactions <= 66 and Number of transactions > 50 THEN Cluster 1

Cluster 2

- IF Brand Loyalty > 0.7 and Value <= 2777.50 THEN Cluster 2
- IF Brand Loyalty <= 0.7 and Share to other brands <= 0.628 and Value <= 157.75 THEN Cluster 2

Cluster 3

- IF Brand Loyalty <= 0.7 and Share to other brands > 0.628 and Number of Brands > 4.5 and Brand Runs <=15 THEN Cluster 3
- IF Brand Loyalty <= 0.7 and Share to other brands > 0.628 and Number of Brands <= 4.5 and Number of Transactions <= 66 and Number of transactions <= 50 THEN Cluster 3

Cluster 4

- IF Brand Loyalty <= 0.7 and Share to other brands <= 0.628 and Value > 157.750 and Total Volume > 20137.5 THEN Cluster 4
- IF Brand Loyalty > 0.7 and Value > 2777.50 THEN Cluster 4

The performance vector of the Decision Tree created for the said clusters is as given below.

accuracy: 92.17%

	true cluster_0	true cluster_1	true cluster_4	true cluster_2	true cluster_3	class precision
pred. cluster_0	121	6	3	3	2	89.63%
pred. cluster_1	13	128	0	0	6	87.07%
pred. cluster_4	0	2	36	0	1	92.31%
pred. cluster_2	1	0	1	97	0	97.98%
pred. cluster_3	0	9	0	0	171	95.00%
class recall	89.63%	88.28%	90.00%	97.00%	95.00%	

Conclusion

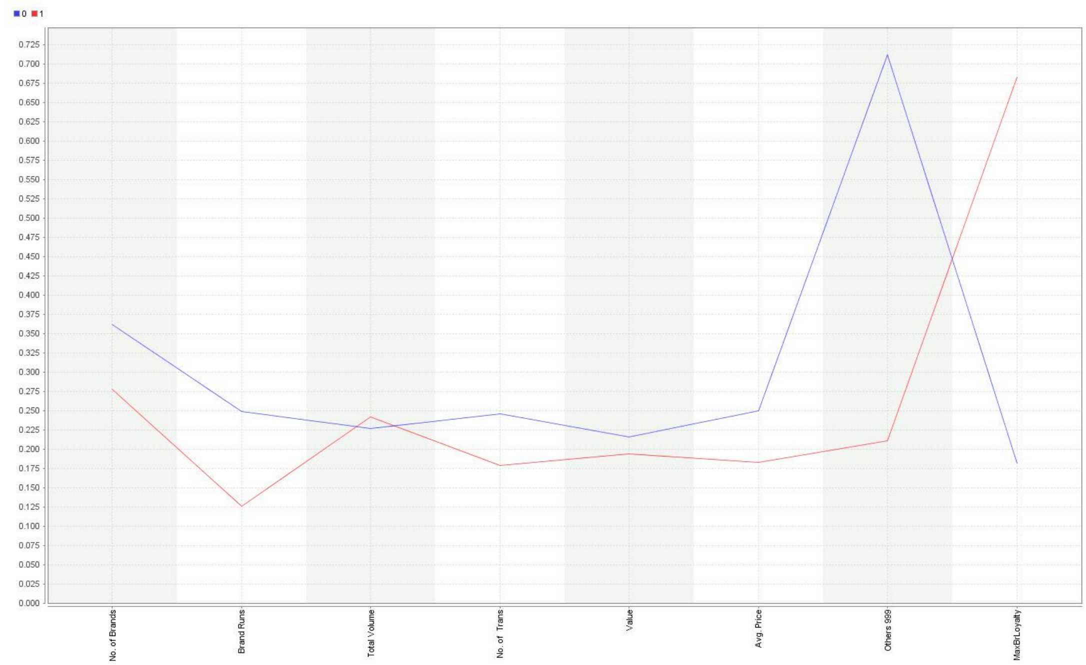
The Decision Tree thus obtained above could be used to target segments of customers who will yield us better results upon selective promotional targeting. Brand Loyalty and Share to other brands are two of the primary attributes upon which the selection of these household clusters must be made upon. It is intuitive that we need to have a maximum brand loyalty and a minimum share to other brands for the cluster of our primary concern. From the above analysis, we determine that Cluster 3 fits perfectly for our customer targeting. This cluster has a share of 12% in our entire group of households and also has a high accuracy with our Decision Tree model.

After performing the clustering analysis, it would be helpful to use a Decision Tree to further analyze these clusters and to narrow down the segments of customers that have to be targeted first. The Decision Tree model obtained thereof yields us great results and it is effective in identifying the different clusters as seen above. The Decision Tree does help us in interpreting the different clusters, with a set of rules which could be easily put into use when we score new incoming household information. Since the Decision Tree rules are much easier to comprehend and apply to real-world cases, it is of great use to interpret the different clusters created with the clustering algorithms.

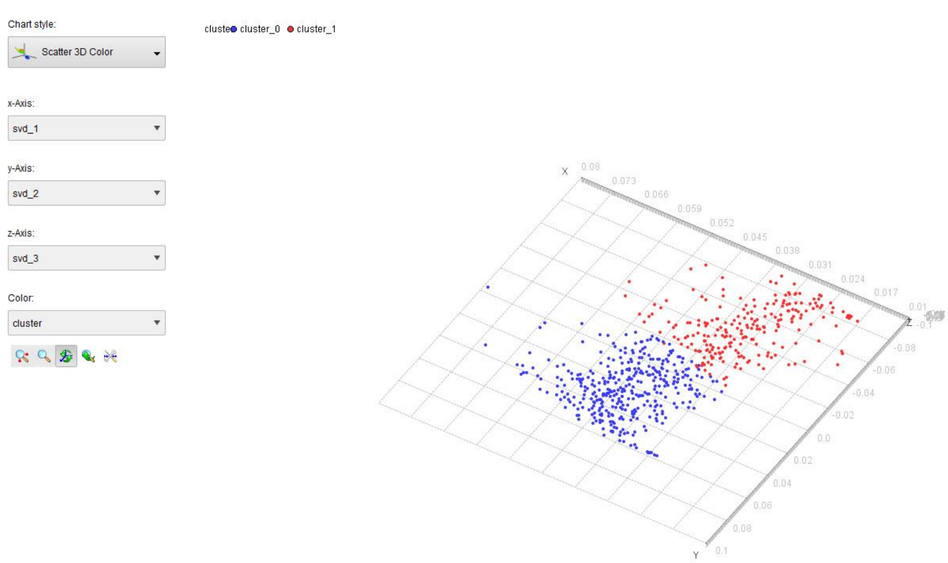
APPENDIX

Question No. 1 (a)

K = 2

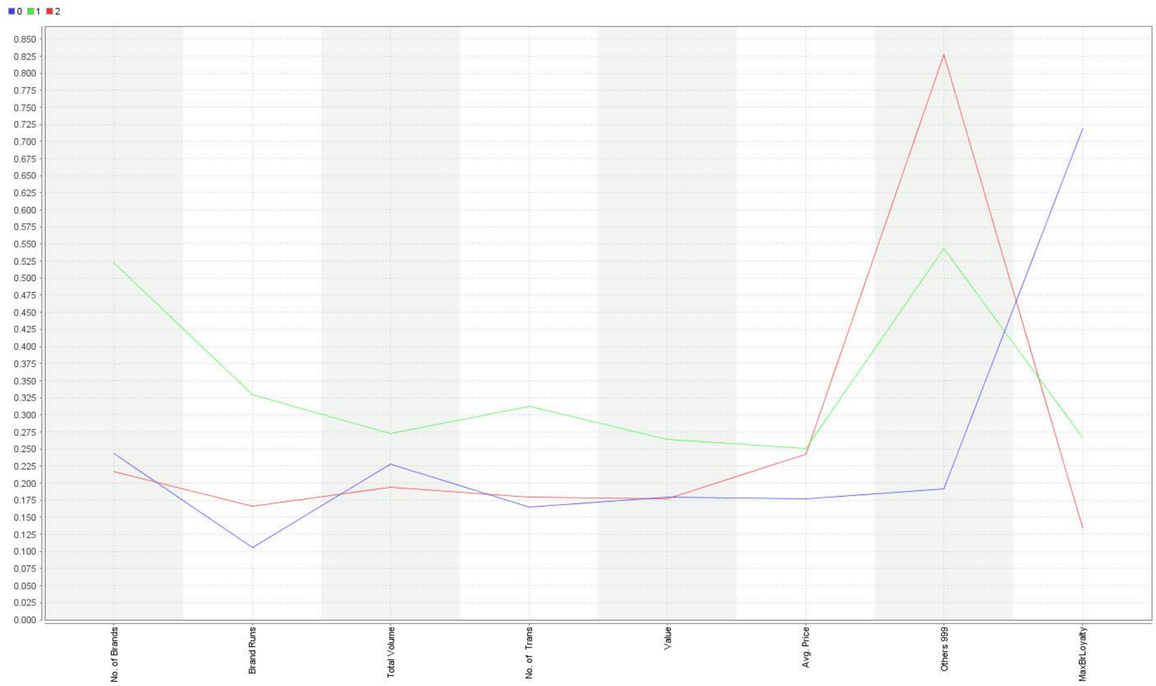


Centroid Plot

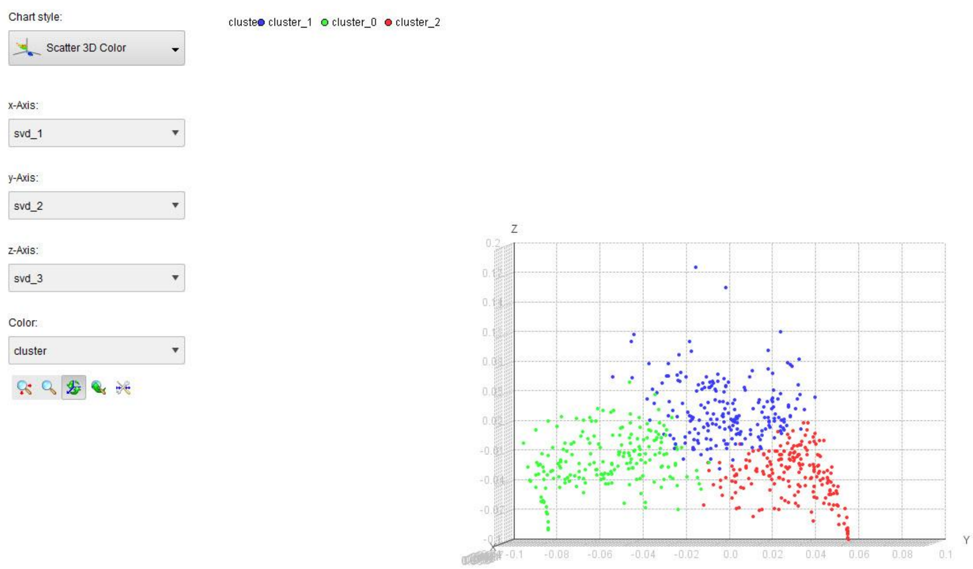


3D Centroid Distance plot

K = 3

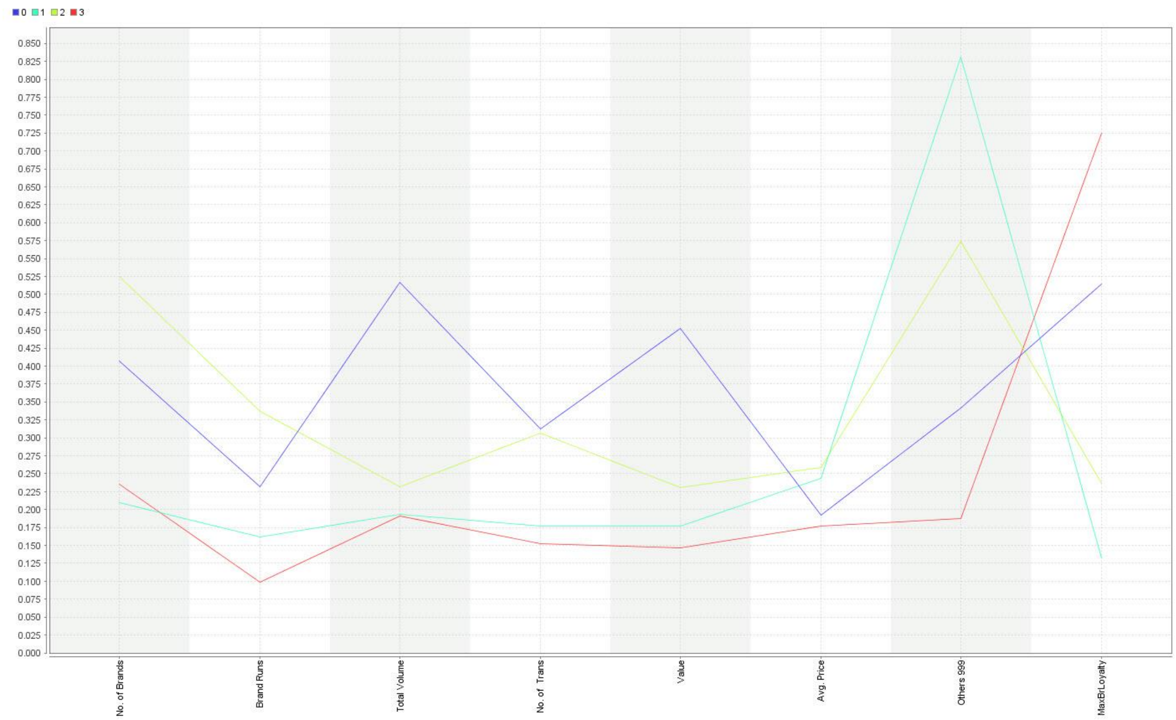


Centroid Plot

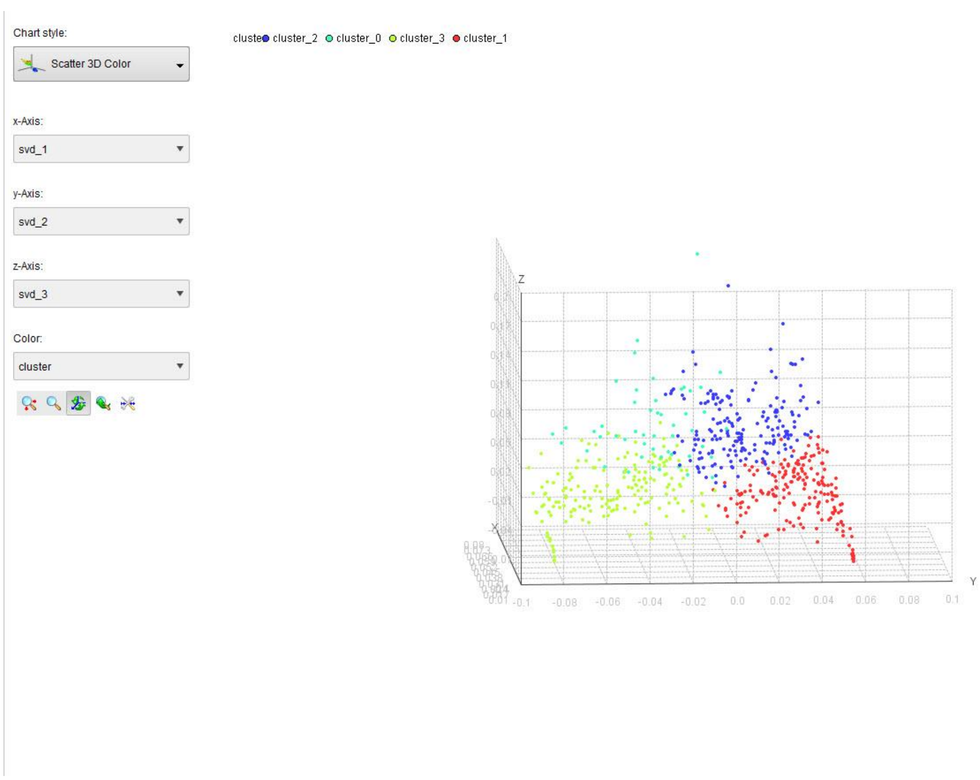


3D Centroid Distance plot

K = 4

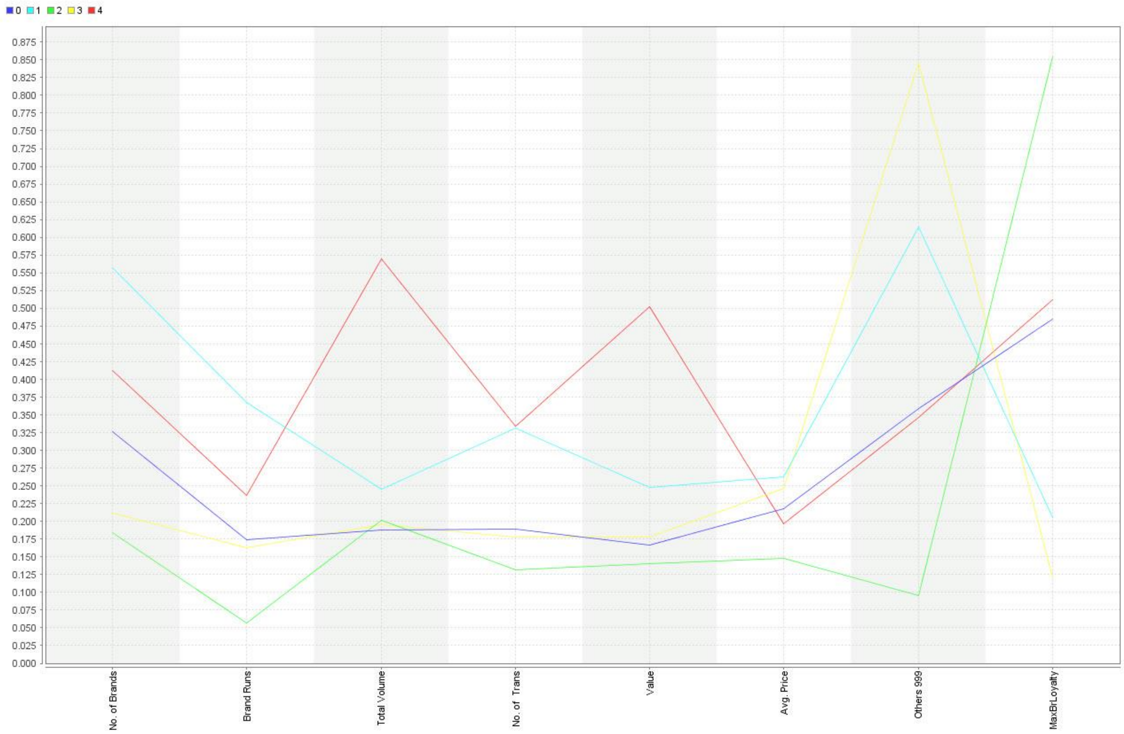


Centroid Plot

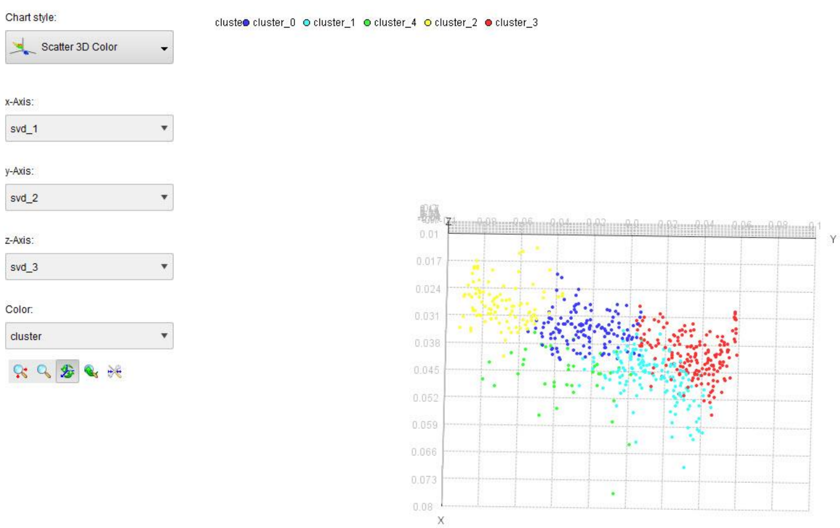


3D Centroid Distance plot

K = 5



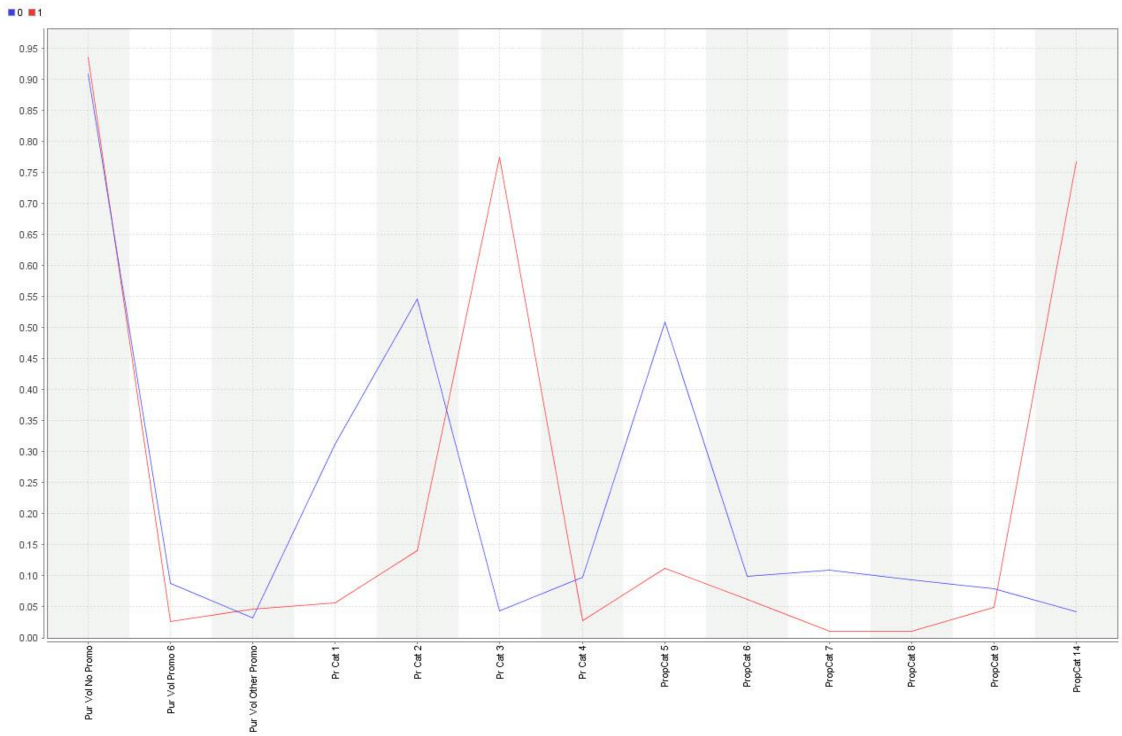
Centroid Plot



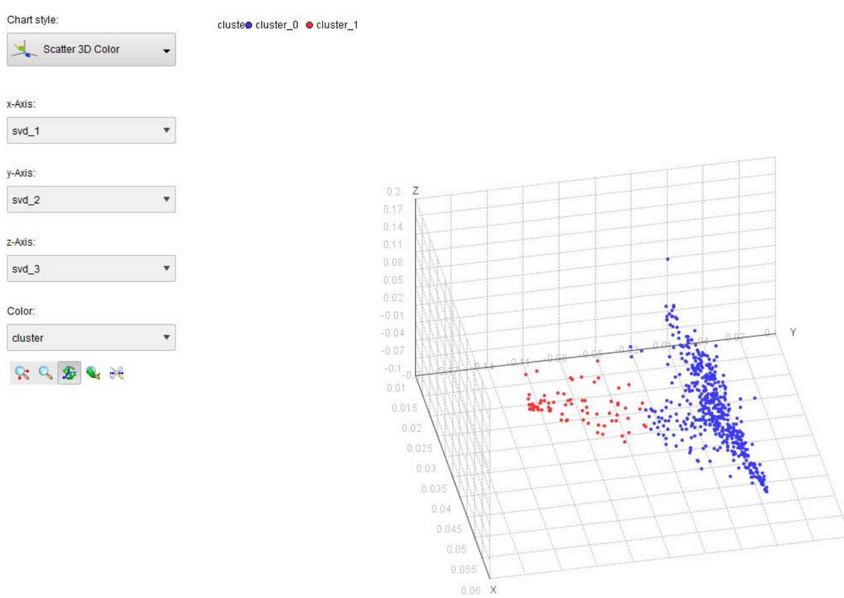
3D Centroid Distance plot

Question No. 1 (b)

K = 2

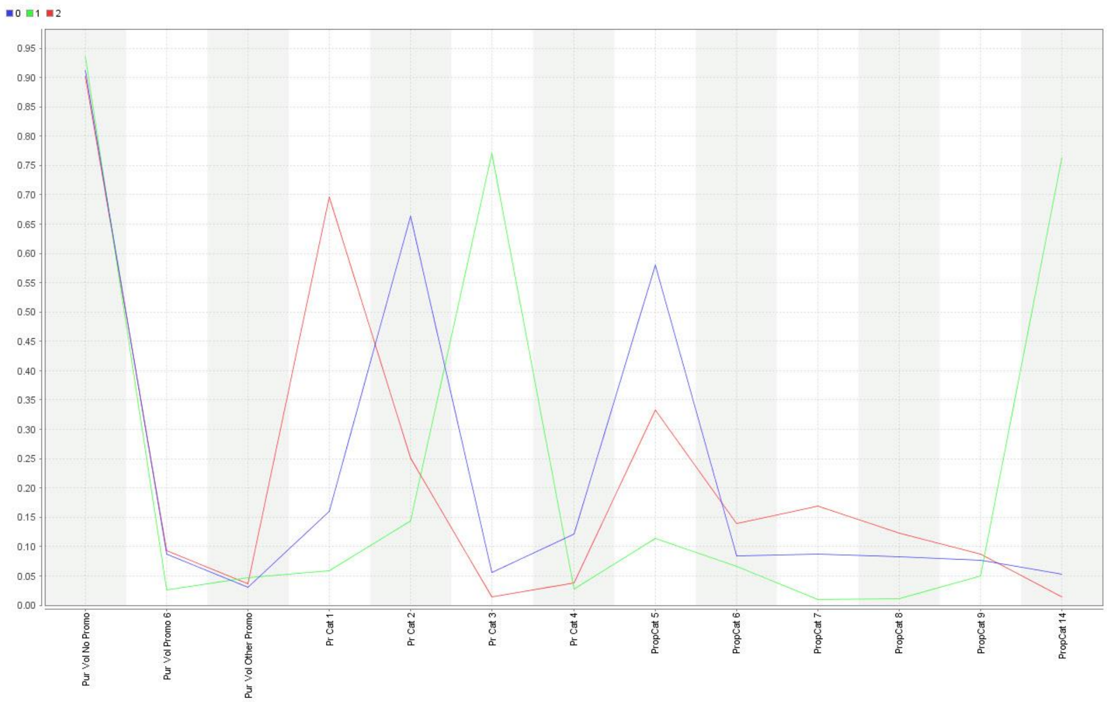


Centroid Plot

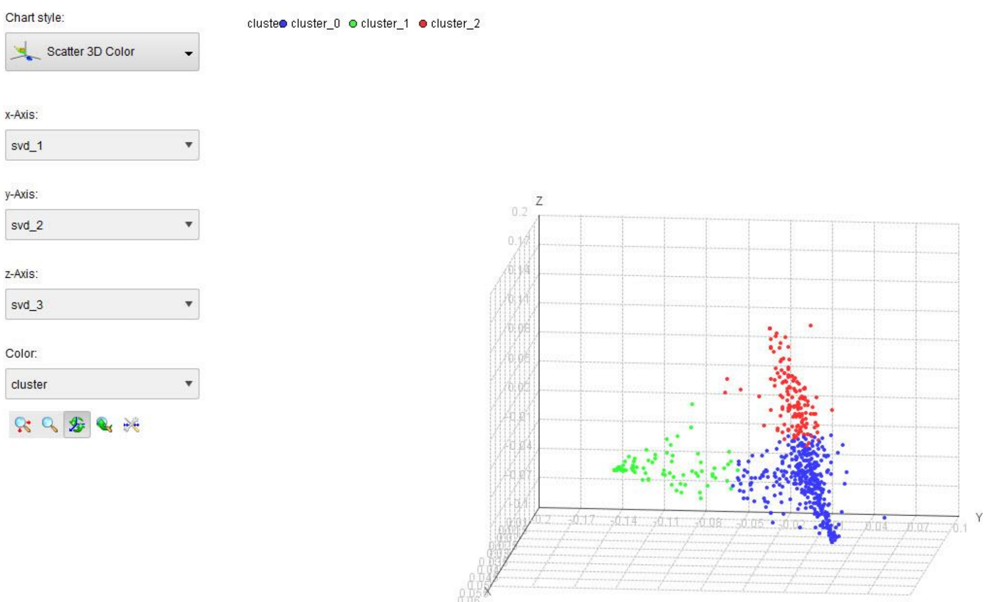


3D Centroid Distance plot

K = 3

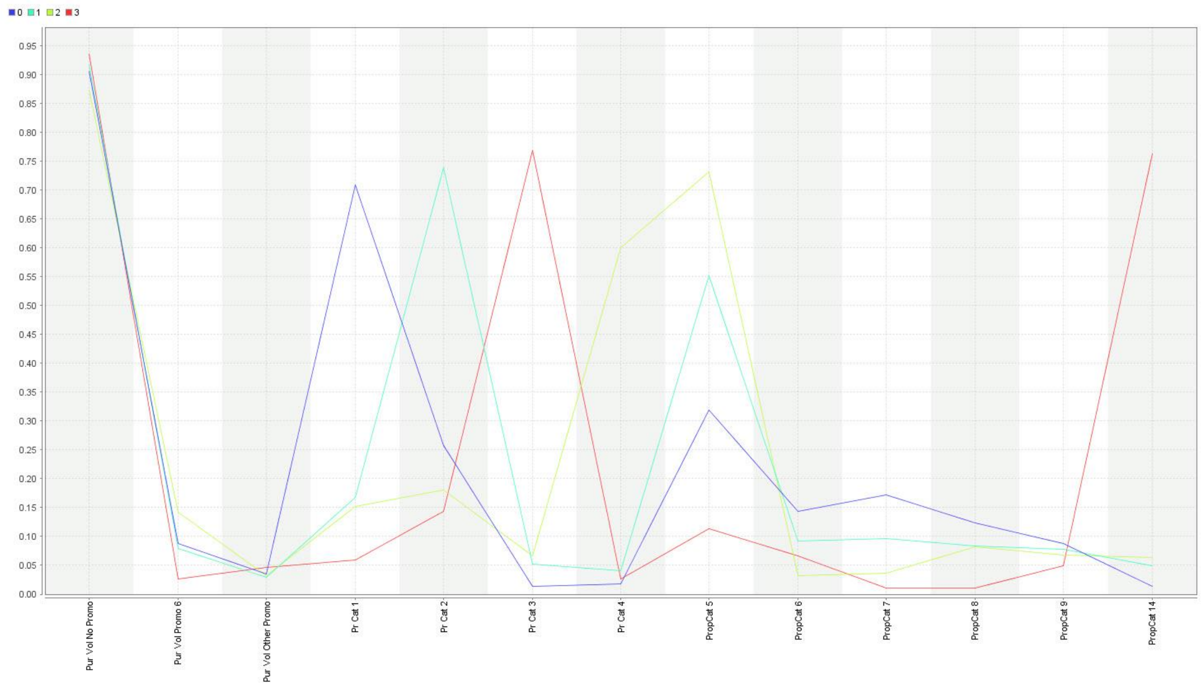


Centroid Plot

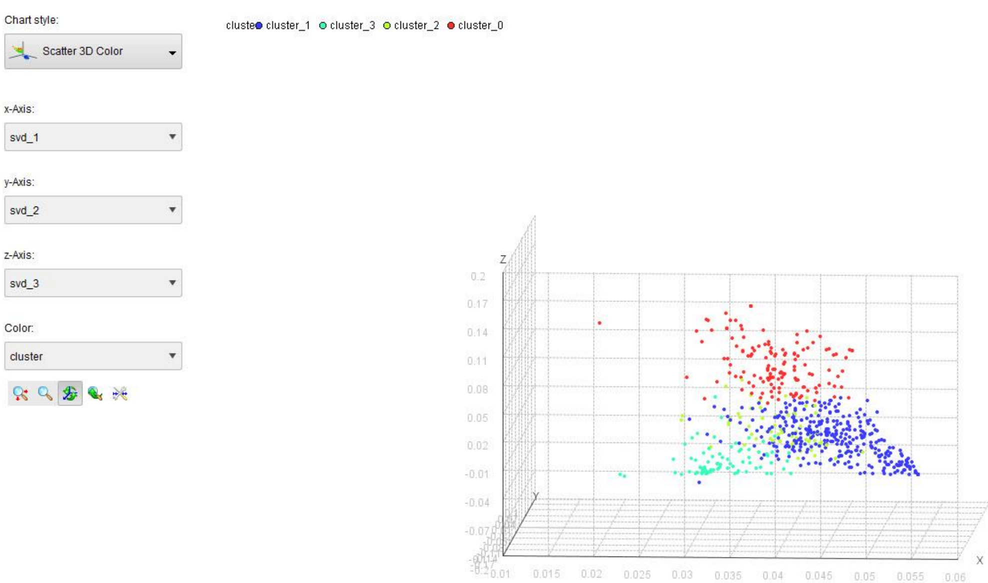


3D Centroid Distance plot

K = 4

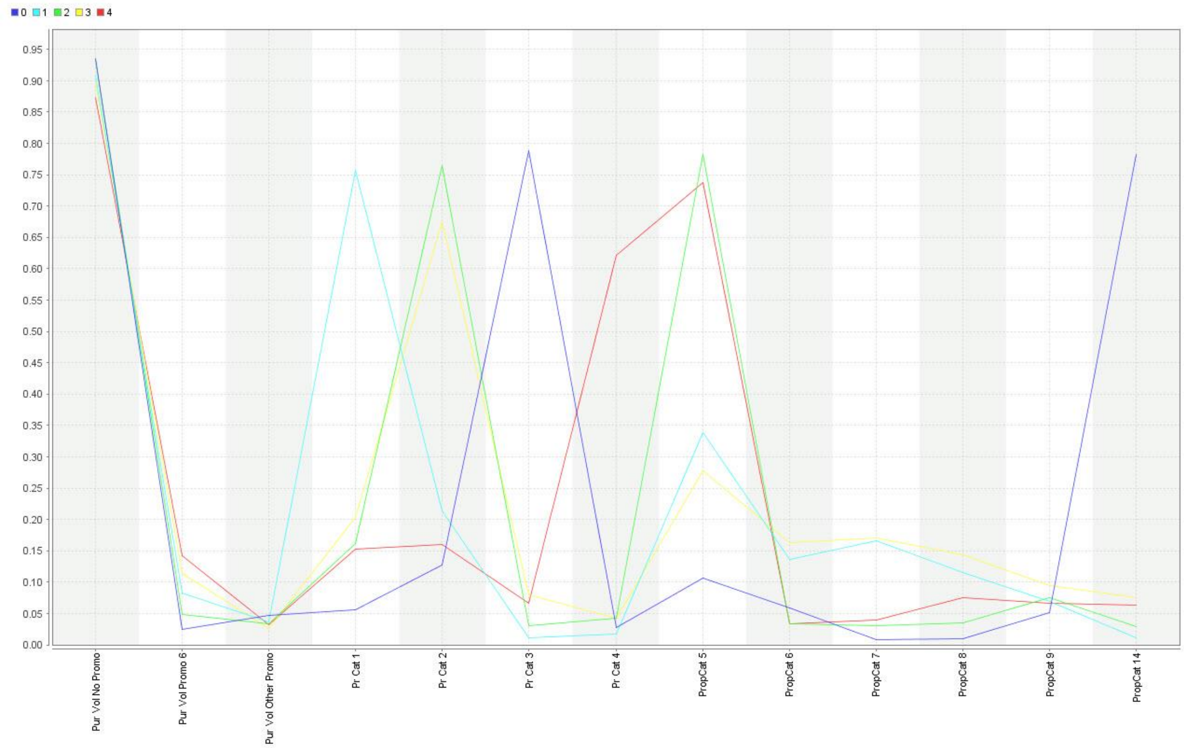


Centroid Plot

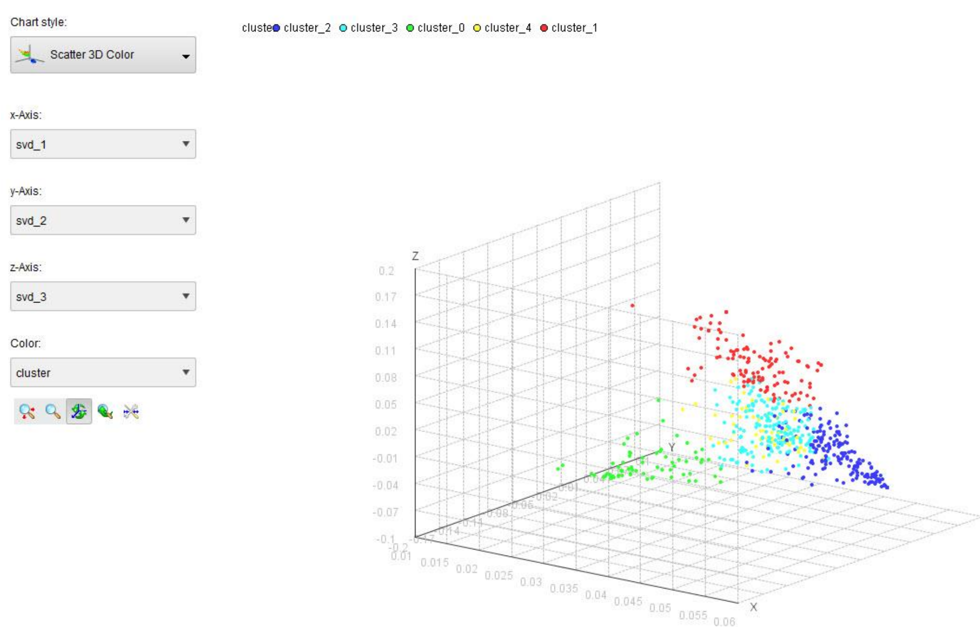


3D Centroid Distance plot

K = 5



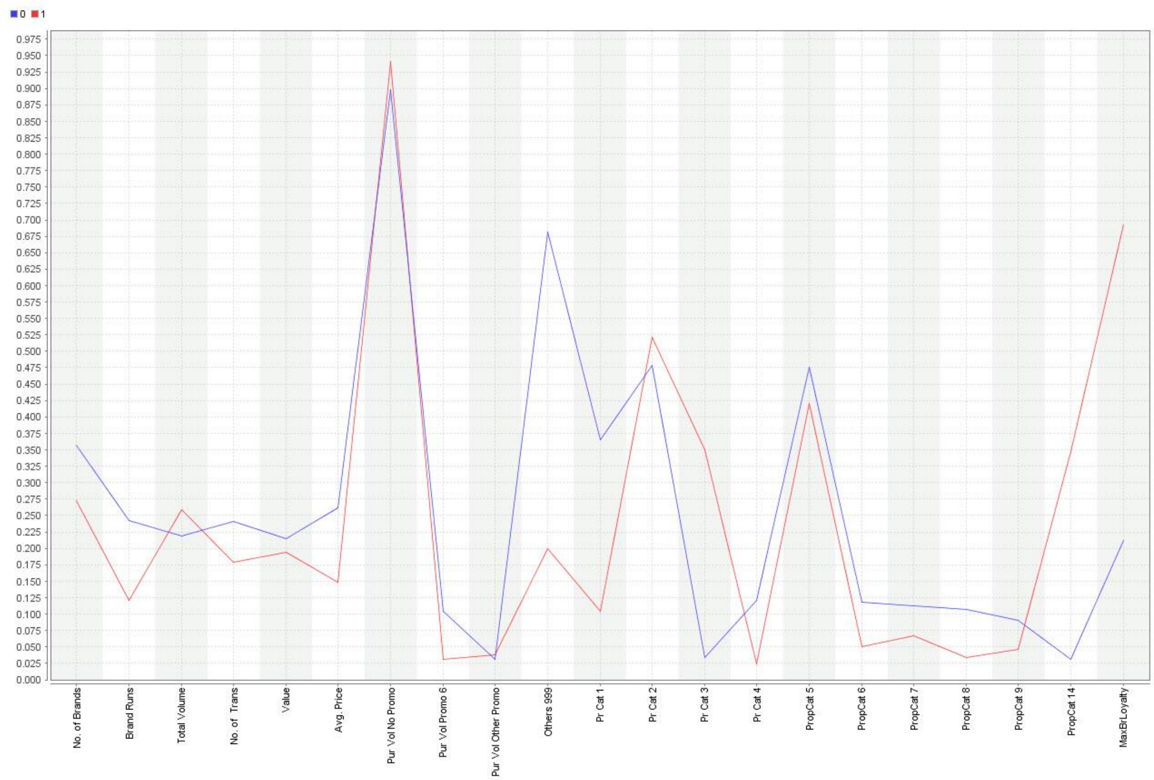
Centroid Plot



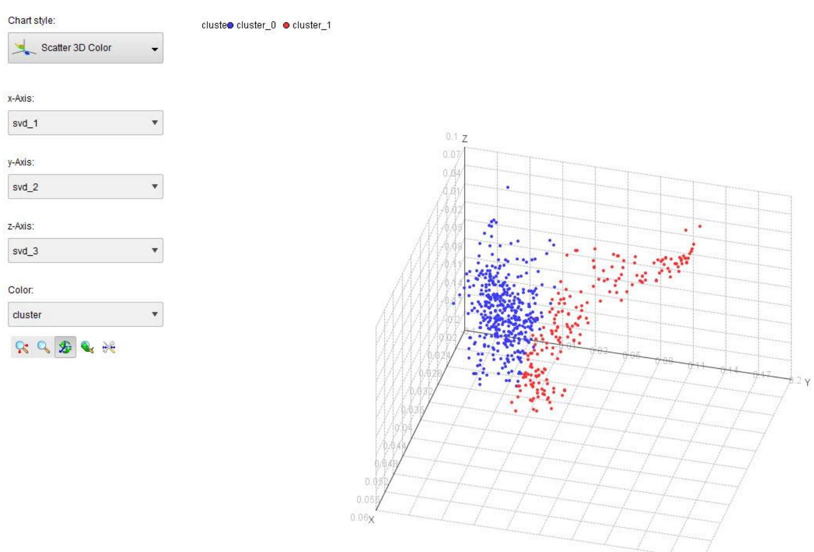
3D Centroid Distance plot

Question No. 1 (c)

K = 2

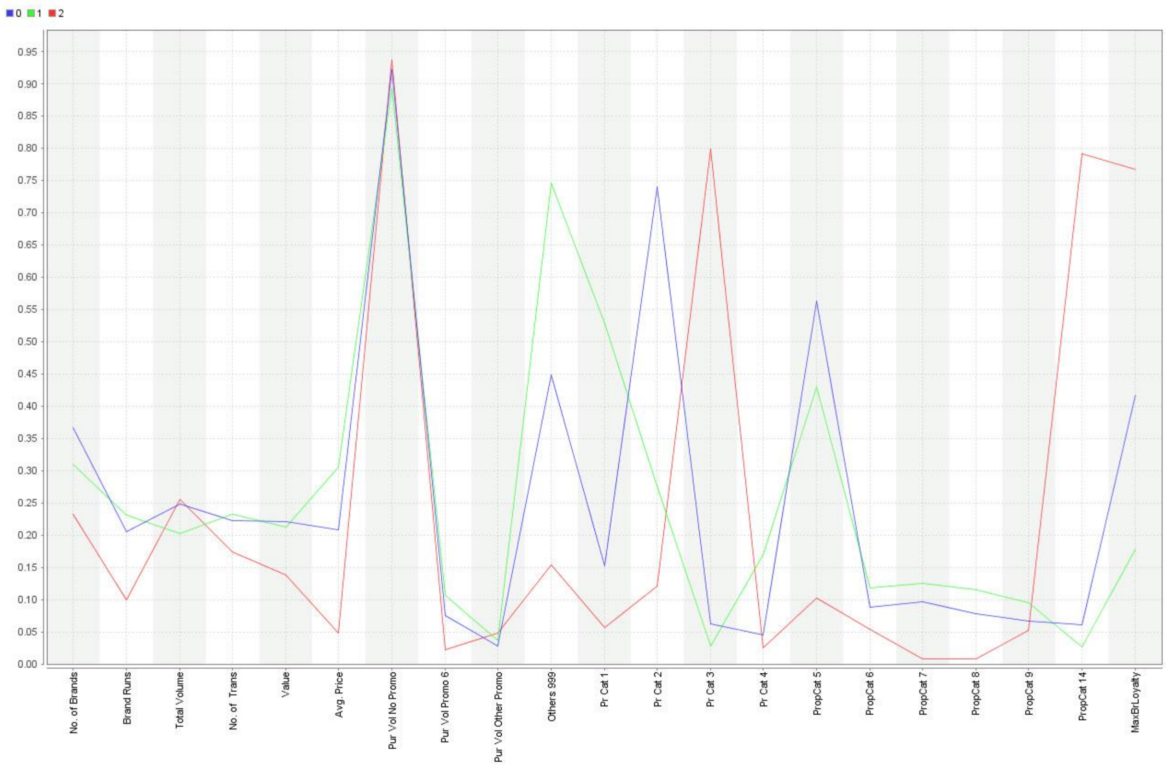


Centroid Plot

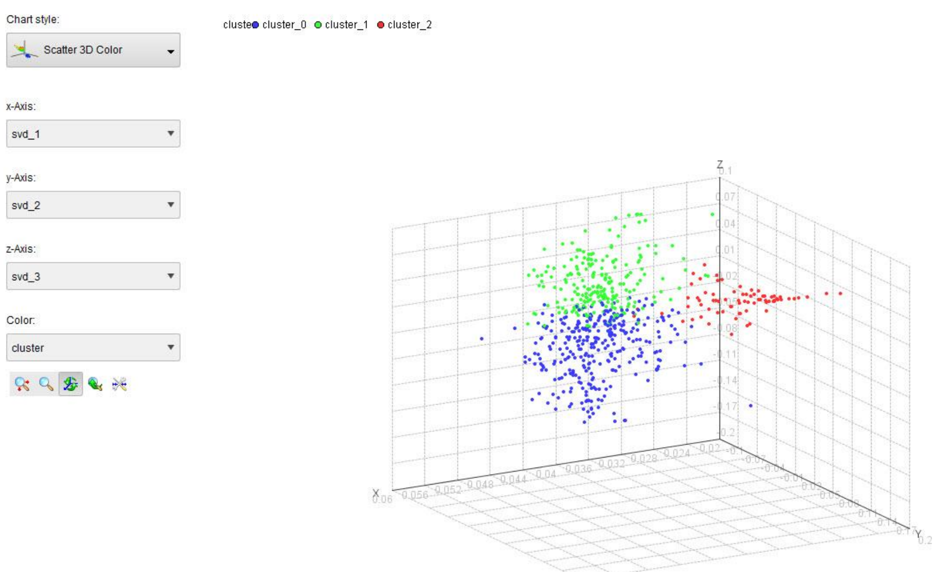


3D Centroid Distance plot

K = 3

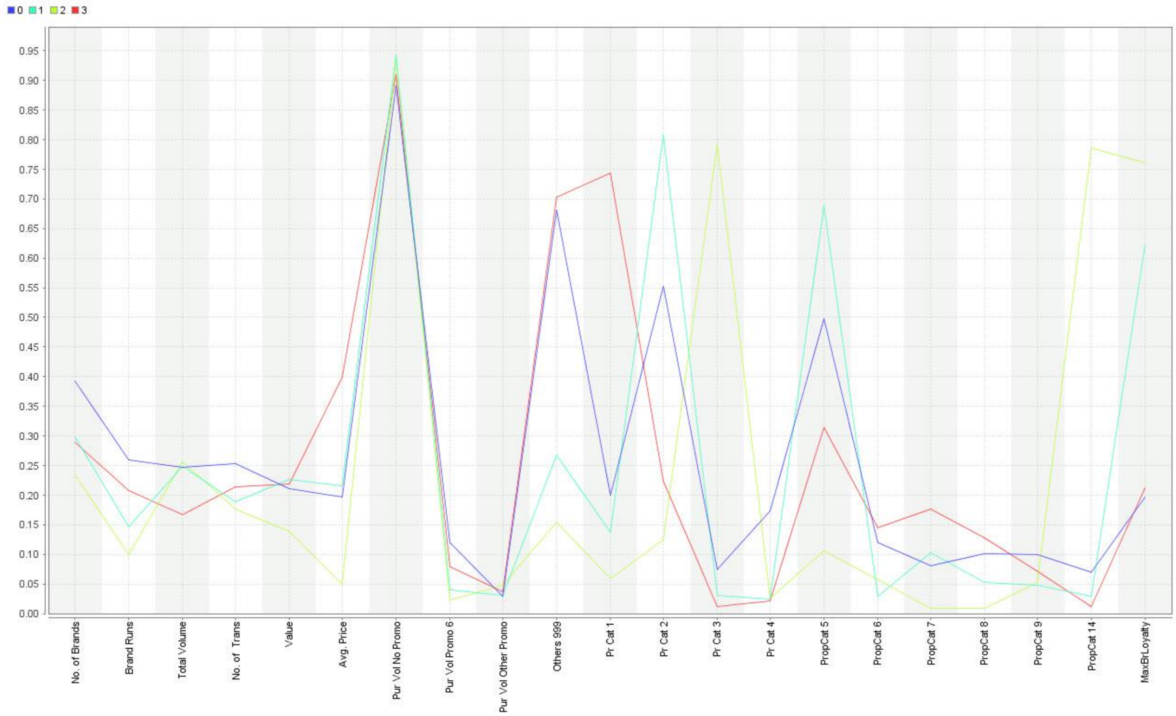


Centroid Plot

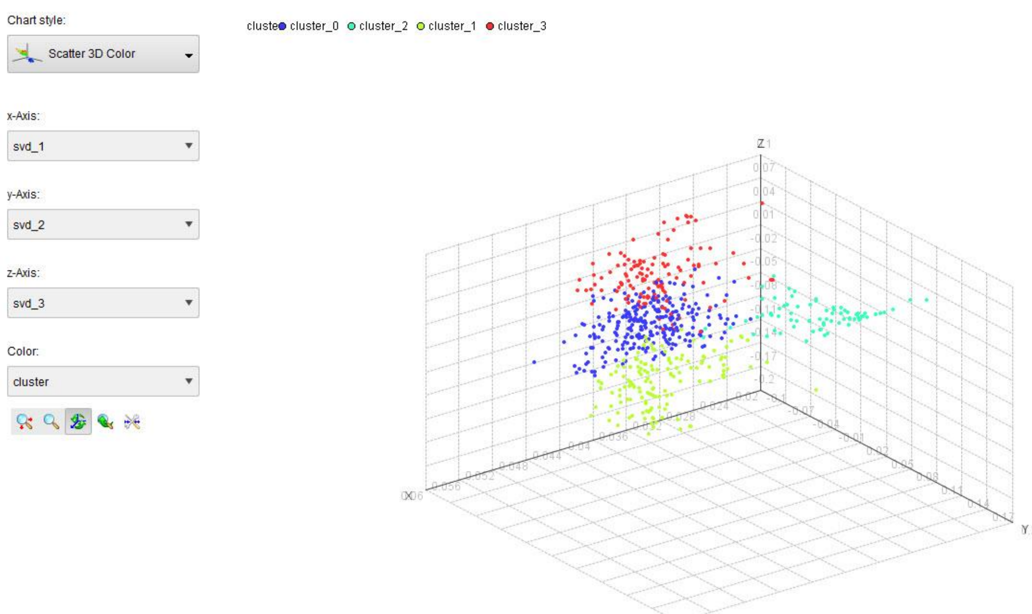


3D Centroid Distance plot

K = 4

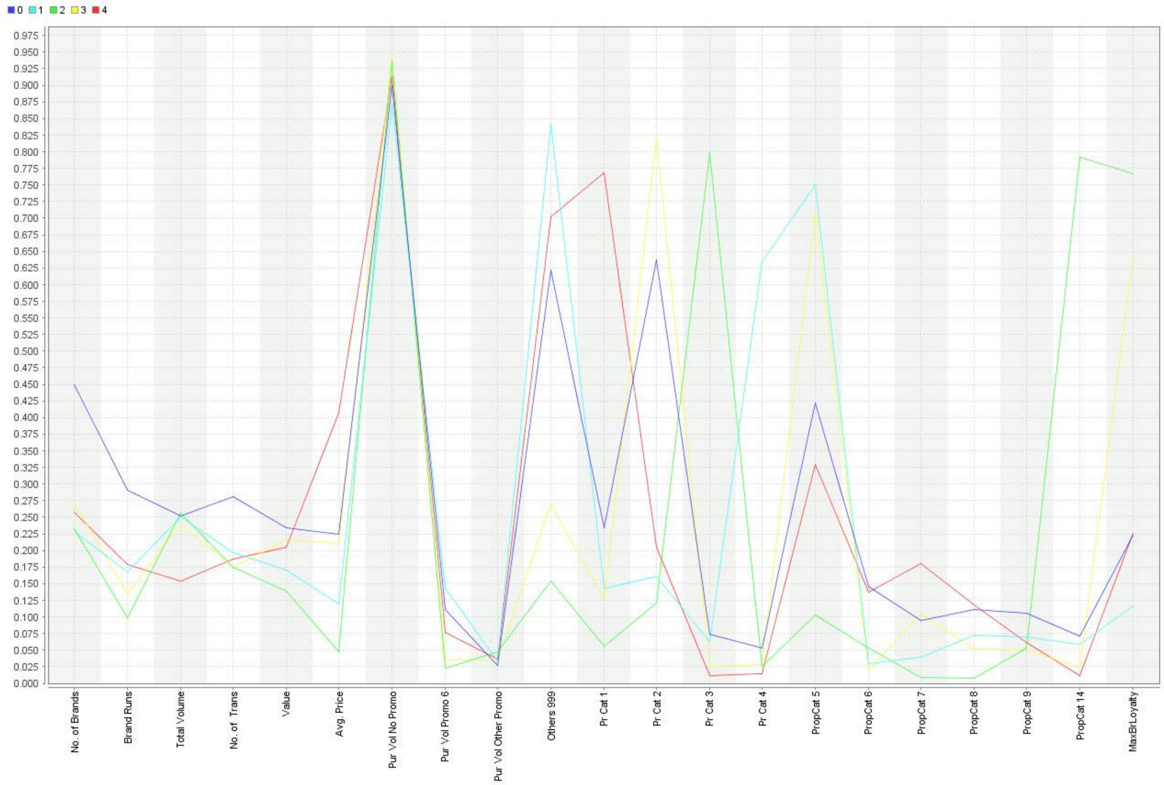


Centroid Plot

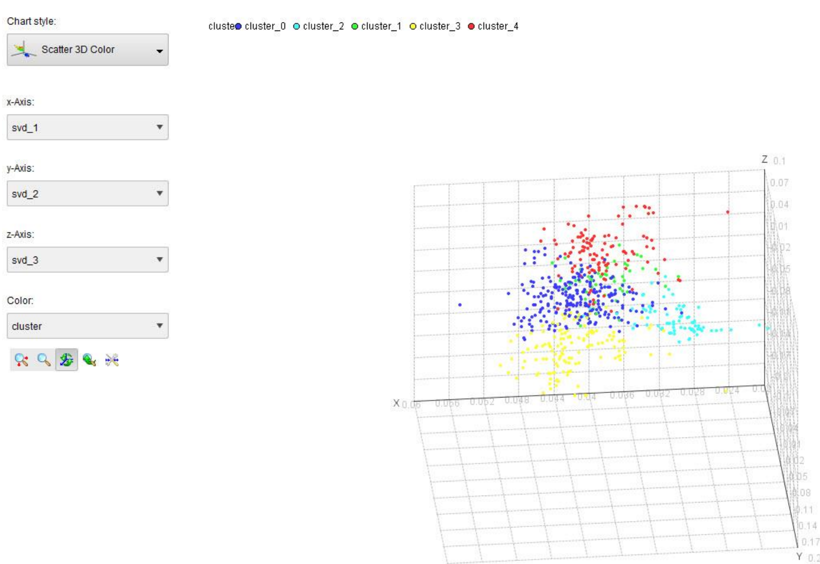


3D Centroid Distance plot

K = 5



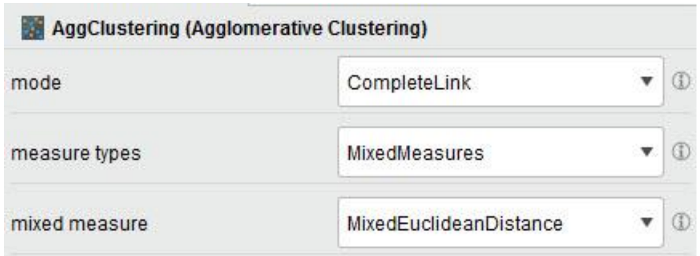
Centroid Plot



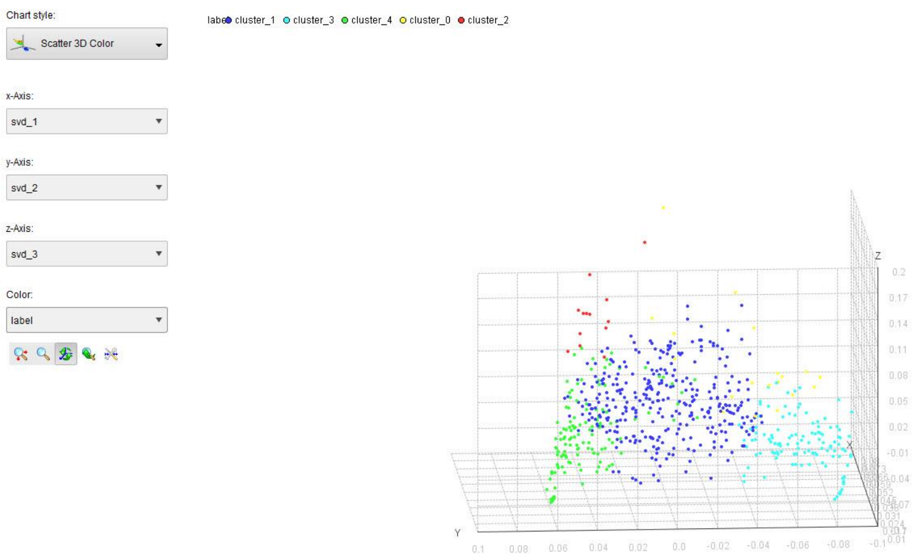
3D Centroid Distance plot

Question No. 1 (d)

Agglomerative Clustering



Agglomerative Clustering Parameters

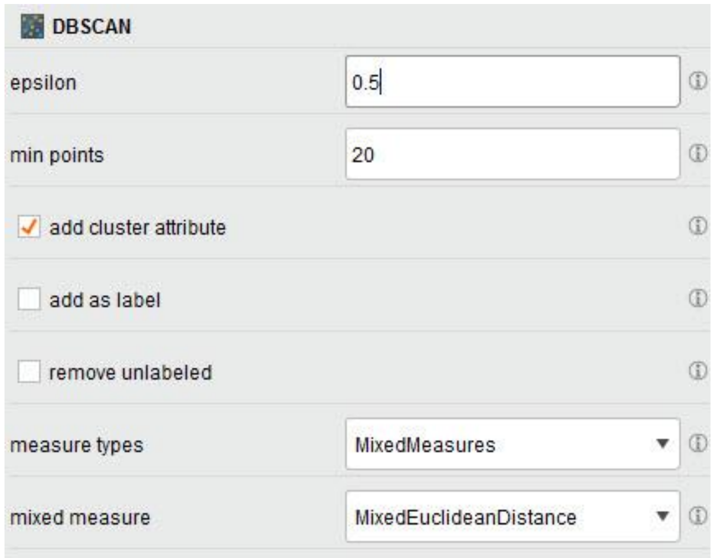


Agglomerative Clustering 3D Plot

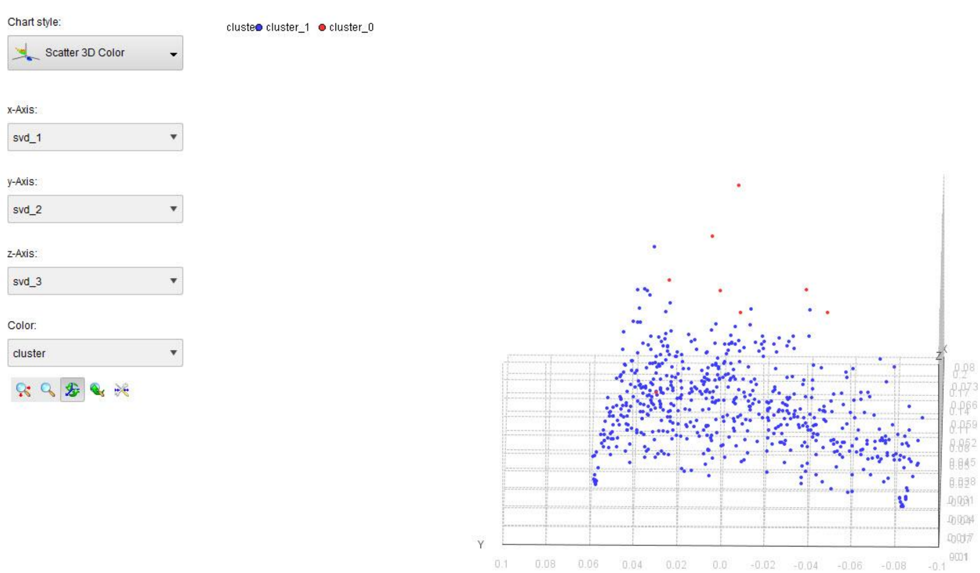
First	Second	Distance
1.0	2.0	1.240
1.0	3.0	1.442
1.0	4.0	1.276
1.0	5.0	1.398
2.0	3.0	1.234
2.0	4.0	1.236
2.0	5.0	1.107
3.0	4.0	1.684
3.0	5.0	1.290
4.0	5.0	1.471

Agglomerative Clustering Distance between Clusters

DBSCAN



DBSCAN Clustering Parameters

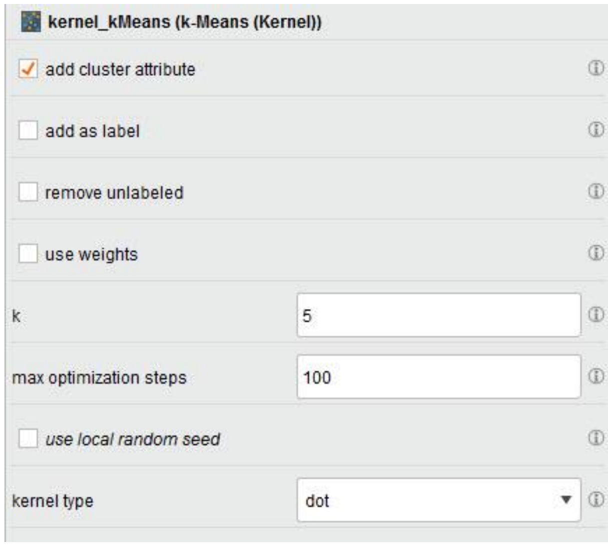


DBSCAN Clustering 3D Plot

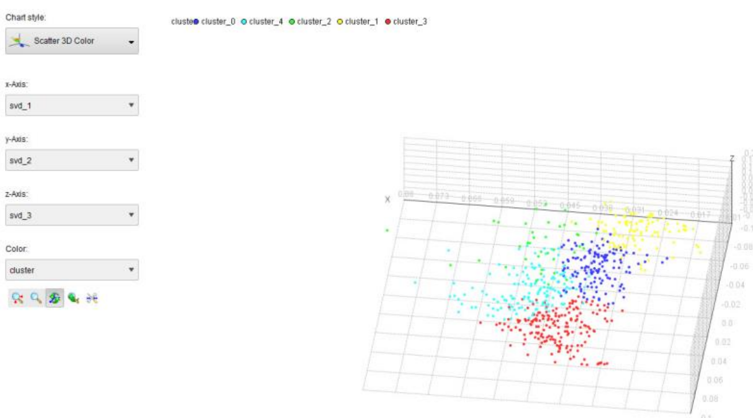
First	Second	Distance
1.0	2.0	1.308

DBSCAN Clustering Distance Between Clusters

Kernel K-means Clustering



Kernel k-means clustering Parameters

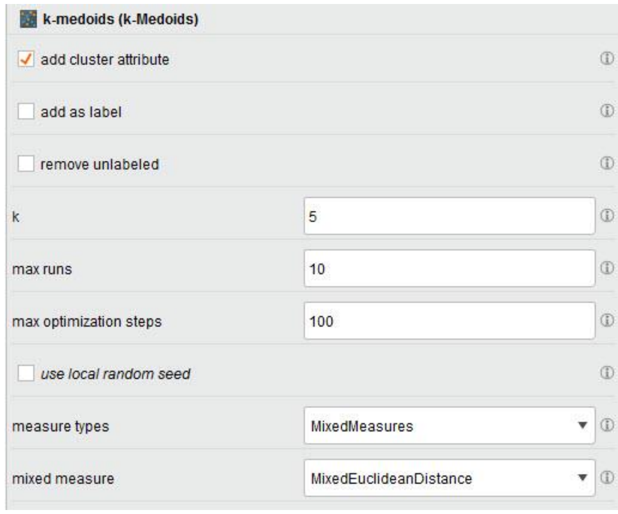


Kernel k-means clustering 3D Plot

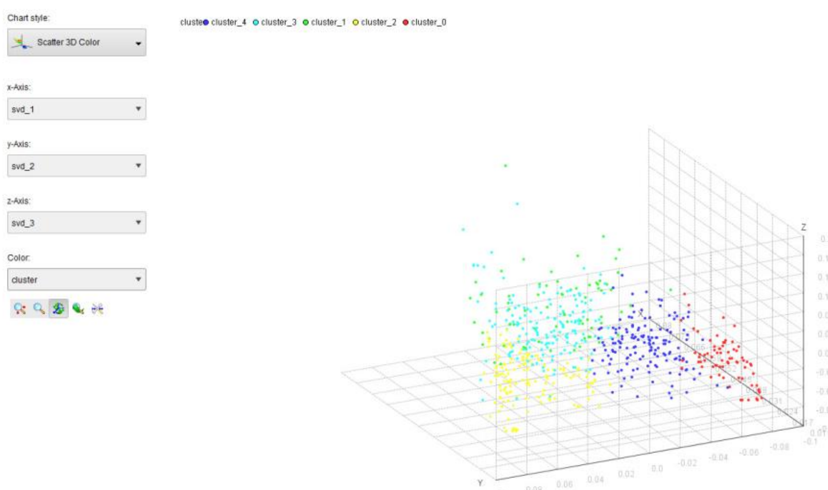
First	Second	Distance
1.0	2.0	1.118
1.0	3.0	1.136
1.0	4.0	1.177
1.0	5.0	1.125
2.0	3.0	1.254
2.0	4.0	1.459
2.0	5.0	1.412
3.0	4.0	1.312
3.0	5.0	1.174
4.0	5.0	1.119

Kernel k-means clustering Distance between clusters

k-Medoids Clustering



k-Medoids Clustering Parameters

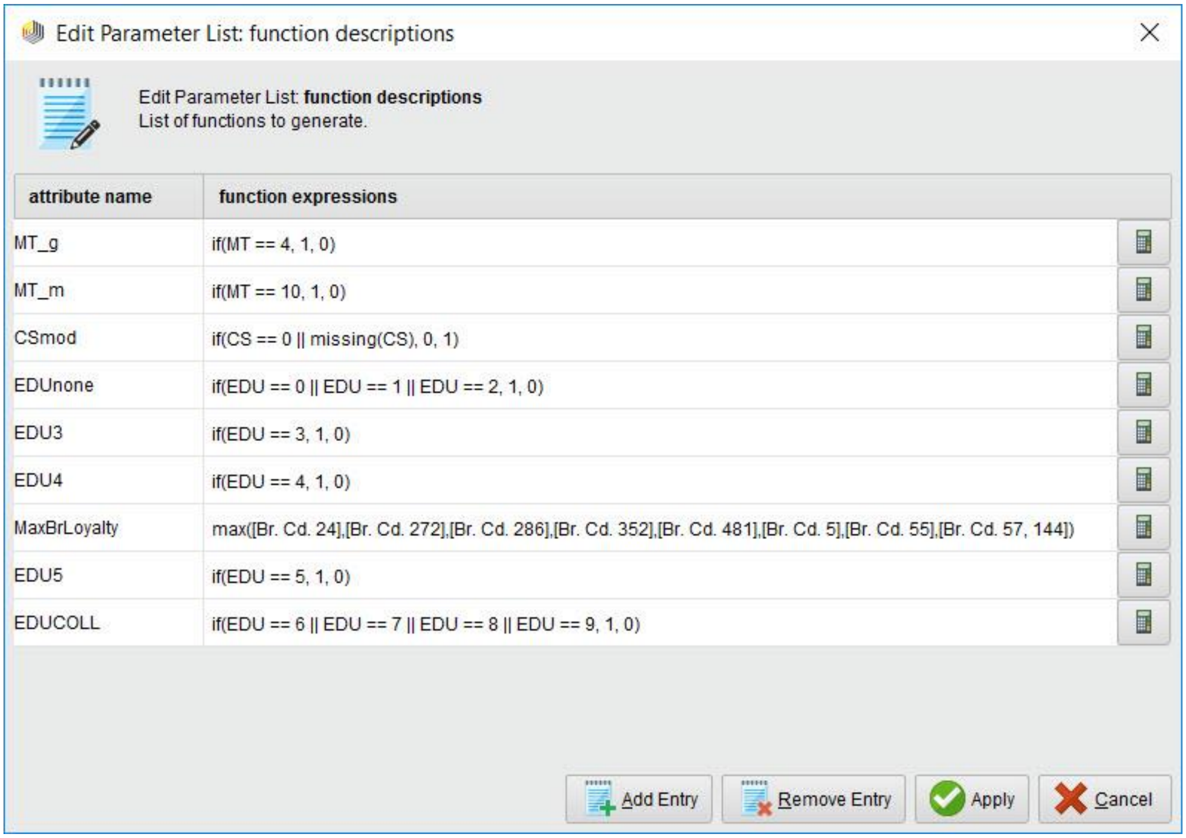


k-Medoids Clustering 3D Plot

First	Second	Distance
1.0	2.0	0.998
1.0	3.0	1.159
1.0	4.0	1.079
1.0	5.0	0.370
2.0	3.0	0.749
2.0	4.0	0.683
2.0	5.0	0.699
3.0	4.0	0.459
3.0	5.0	0.857
4.0	5.0	0.802

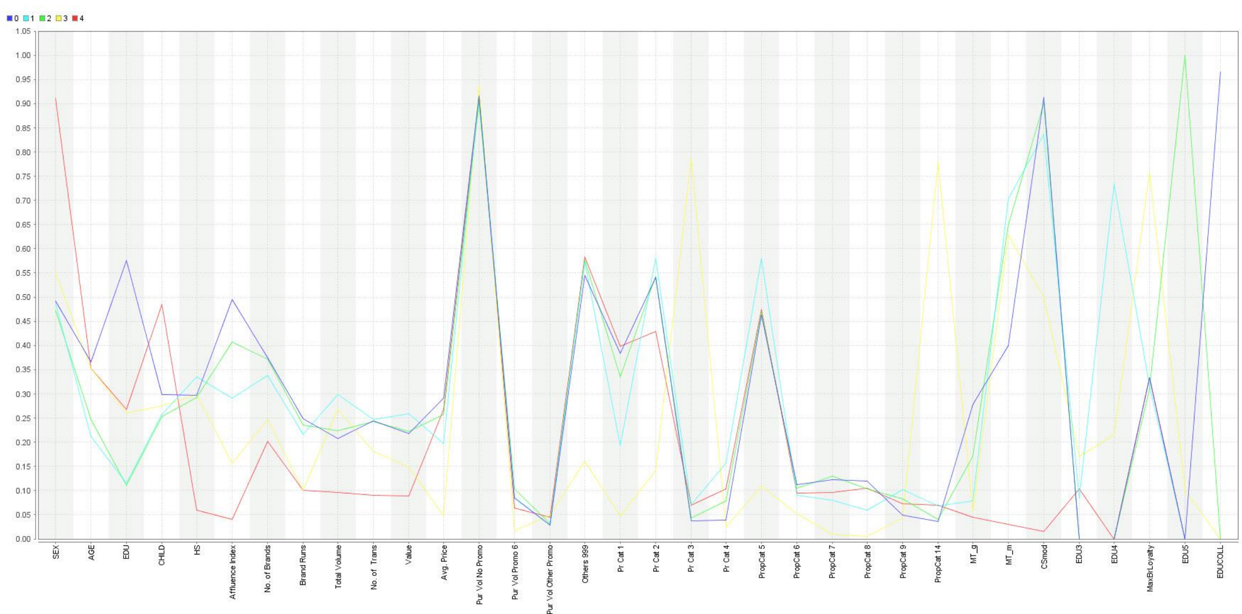
k-Medoids Clustering Distance between clusters

Question No. 2



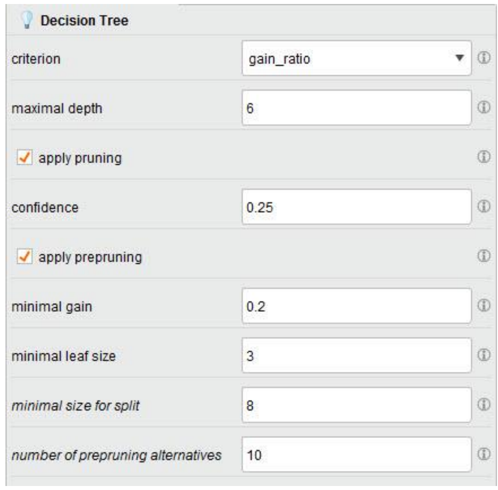
attribute name	function expressions
MT_g	if(MT == 4, 1, 0)
MT_m	if(MT == 10, 1, 0)
CSmod	if(CS == 0 || missing(CS), 0, 1)
EDUnone	if(EDU == 0 || EDU == 1 || EDU == 2, 1, 0)
EDU3	if(EDU == 3, 1, 0)
EDU4	if(EDU == 4, 1, 0)
MaxBrLoyalty	max([Br. Cd. 24],[Br. Cd. 272],[Br. Cd. 286],[Br. Cd. 352],[Br. Cd. 481],[Br. Cd. 5],[Br. Cd. 55],[Br. Cd. 57, 144])
EDU5	if(EDU == 5, 1, 0)
EDUCOLL	if(EDU == 6 || EDU == 7 || EDU == 8 || EDU == 9, 1, 0)

Data Transformation for the Demographics Attributes



Centroid plot of the 5 clusters aggregated across the entire data set

Question No. 3

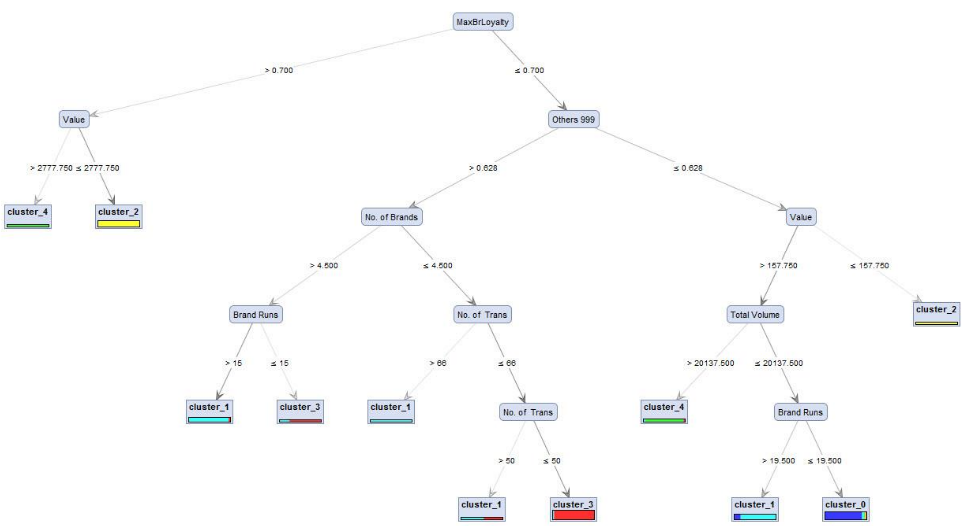


Decision Tree Parameters

accuracy: 92.17%

	true cluster_0	true cluster_1	true cluster_4	true cluster_2	true cluster_3	class precision
pred. cluster_0	121	6	3	3	2	89.63%
pred. cluster_1	13	128	0	0	6	87.07%
pred. cluster_4	0	2	36	0	1	92.31%
pred. cluster_2	1	0	1	97	0	97.98%
pred. cluster_3	0	9	0	0	171	95.00%
class recall	89.63%	88.28%	90.00%	97.00%	95.00%	

Decision Tree Performance Vector



Decision Tree