#ASSIGNMENT-4

Market Segmentation - Segmenting Consumers of Bath Soap

-by
Akash Adlakha(UIN:678130796)
Shalakha Coutinho(UIN:651029194)
Srija Gupta(UIN:656331239)

CRISA has both transaction data and household data and, for the household data, maintains the following information: • Demographics of the households • Possession of durable goods and a computed "affluence index" on this basis • Purchase data of product categories and brands. We need to calculate the Brand Loyalty for the given dataset which has 600 examples,1 Special attribute and 45 regular attributes.

Question 1.a) Use k-means clustering to identify clusters of households based on a. The variables that describe purchase behavior (including brand loyalty). How will you evaluate brand loyalty – describe the variables you create/use to capture different perspectives on brand loyalty.
[Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)].

Answer 1:
We had in total 45 variables and 600 attributes.We went through the statistical distribution of each variable and tried to group a few of them as follows:

	Attribute Name	**Function expression**
	Gujarati	if(MT=="4",1,0)
	Marathi	if(MT=="10",1,0)
	Urdu	if(MT=="17",1,0)
	Hindi	if(MT=="5",1,0)
	Cast_Not_Specified	if(MT=="0",1,0)
	Edu_Low	if(EDU=="1"|| EDU=="2"||EDU=="3"||EDU=="4",1,0)
	Edu_Mid	if(EDU=="6"|| EDU=="5",1,0)
	Edu_High	if(EDU=="7"|| EDU=="8"||EDU=="9",1,0)
	TV	if(CS=="1",1,0)
	Max_Percent_Brand_Purchase	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],[Others 999])
	Purchase_without_Promotions	f(max([Pur Vol No Promo__],[Pur Vol Promo 6_],[Pur Vol Other Promo_])==[Pur Vol No Promo__],1,0)
	Price_Code	if(max([Pr Cat 1],[Pr Cat 2],[Pr Cat 3],[Pr Cat 4])==[Pr Cat 1],1, if(max([Pr Cat 1],[Pr Cat 2],[Pr Cat 3],[Pr Cat 4])==[Pr Cat 2],2, if(max([Pr Cat 1],[Pr Cat 2],[Pr Cat 3],[Pr Cat 4])==[Pr Cat 3],3,4)))

	Sex	if(SEX=="2",1,0)

Variable Transformation:

1)Purchase by promotions: Pur Vol NoPromo % , Pur Vol Promo 6%, Pur Vol Other Promo%

2)Price Category: Price cat 1-4

3)Selling propositions: Proposition cat 5-15

4)Price_code : Promotion code list categorized.

5)Purchase without promotion: Customers who buy the product even if there is no promotional discount on that particular product are the most loyal ones. So, we assigned:
>Pur Vol Promo 6%, Pur Vol Other Promo% as "0"
> Pur Vol NoPromo % as "1"

6)Brand Loyalty: For calculating brand loyalty, we gave half the weightage to brand run and almost 35% weightage to the average volume of products purchased and the rest weightage to percentage of a particular brand purchased.

7)Maximum percentage Brand Loyalty : Maximum percentages amongst the various brand codes given.

8)Education: Classified a low, medium and high.
9)MT: The mouther tongue were grouped as the ones with maximum data in it: Gujarati,Marathi,Urdu,Hindi and Cast Not specified

K-means Clustering

Variables used:

- Avg. Price
- Brand_Loyalty
- Max_Percent_Brand_Purchase
- Total Volume
- Value

Parameters

Parameters	Values
K	2,3,4,5
Max runs	10
Measure Types	Mixed measures
Mixed Measures	Mixed Euclidean Distance
Max Optimization Steps	100

Performance Vector

k	2	3	4	5	6
Avg. within centroid distance	0.764	0.633	0.539	0.480	0.434
Avg. within centroid distance_cluster_0	1.275	1.418	0.537	0.341	2.062
Avg. within centroid distance_cluster_1	0.629	0.539	1.486	0.532	0.925
Avg. within centroid distance_cluster_2	-	0.445	0.373	0.514	0.549
Avg. within centroid distance_cluster_3	-	-	0.369	3.066	0.297
Avg. within centroid distance_cluster_4	-	-	-	0.349	0.321
Avg. within centroid distance_cluster_5	-	-	-	-	0.346
Davies Bouldin	0.280	0.288	0.252	0.257	0.252

EXPLANATION : According to the performance vector, If we look at Davies Bouldin's distance, we would go for k=4 and k=6, Also amongst these We get the highest value of 2.062 (Within cluster 0) for K=6,while for k=4 we get the highest value of 1.486(within cluster 1), which is less amongst the two, So We Choose K=4 . Also, We straight away ignore k=5, because the distance within centroid 3 is too much to compromise i.e. 3.066.

Similarity Measure object (Data to similarity)

K=2	**Cluster**	**Cluster 1**			**Cluster 2**	
	Cluster 1				0.266	
K=3	**Cluster**	**C1**		**C2**		**C 3**
	Cluster 1			3.078		3.287
	Cluster 2					1.814
K=4	**Cluster**	**C1**	**C2**	**C3**		**C4**
	Cluster 1	0	3.86	2.211		2.5
	Cluster 2			3.18		3.37
	Cluster 3					1.88
K=5	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**
	Cluster 1		2.24	2.31	5.52	1.90
	Cluster 2			3.40	5.98	2.55
	Cluster 3				3.66	2.44
	Cluster 4					5.848
K=6	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
	Cluster 1		4.38	5.73	4.9	5.07	5.10
	Cluster 2			4.08	3.14	3.36	4.19
	Cluster 3				2.58	2.84	2.13
	Cluster 4					2.15	1.69
	Cluster 5						1.75

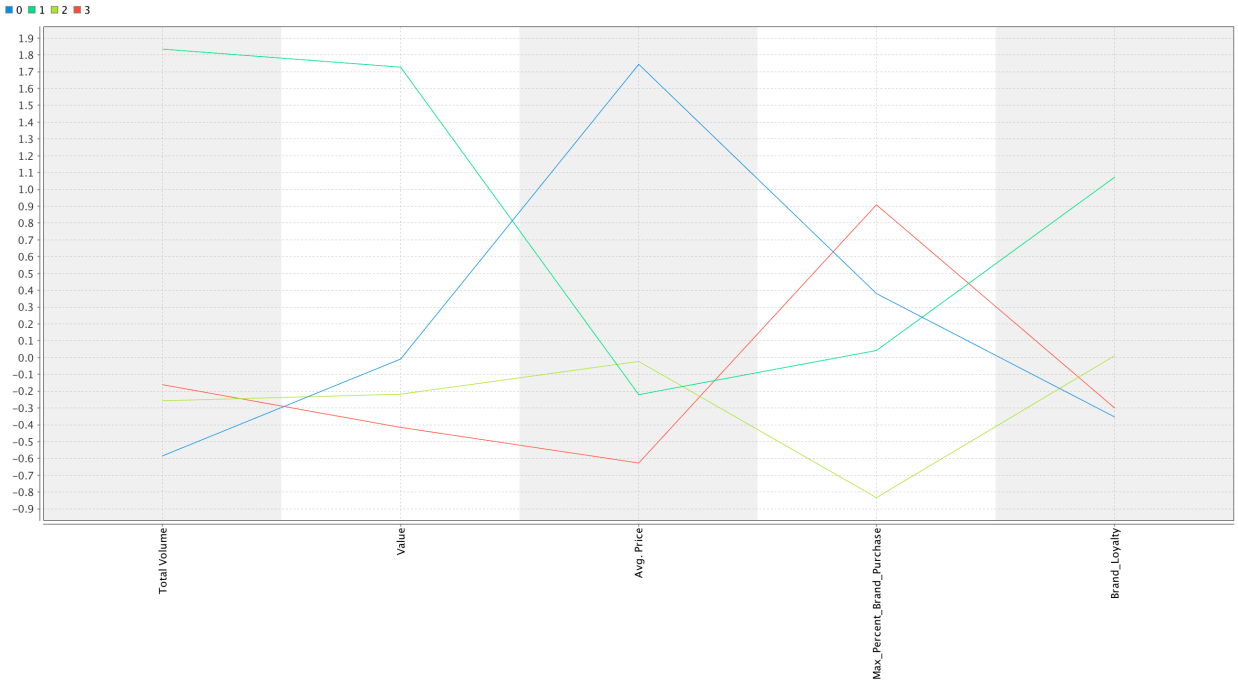
EXPLAINATION : This is the distance between clusters, so more the distance, more it is better. So by that We straight away can ignore K=2 (which has distance of0.266) and K=6 (which has distance of 1.75) also K=3(which has a distance of 1.81). We have a competition between k=4 and k=5. So from this model we Go for both K=4 and K=5

Distribution of data items

K=2	Cluster 0: 125 items Cluster 1: 475 items Total number of items: 600
K=3	Cluster 0: 87 items Cluster 1: 303 items Cluster 2: 210 items Total number of items: 600
K=4	Cluster 0: 82 items Cluster 1: 78 items Cluster 2: 249 items Cluster 3: 191 items Total number of items: 600
K=5	Cluster 0: 228 items Cluster 1: 79 items Cluster 2: 102 items Cluster 3: 18 items Cluster 4: 173 items Total number of items: 600
K=6	Cluster 0: 12 items Cluster 1: 52 items Cluster 2: 70 items Cluster 3: 171 items Cluster 4: 145 items Cluster 5: 150 items Total number of items: 600

EXPLANATION : distribution of K=5 is normal and good.

We will go for cluster 4 as from the performance vector, Data to similarity and the Distribution of items within each cluster.



Analysis of Clusters with K=4

1)Cluster 0: The group of customers in this group have a pretty good brand loyalty and their average price is the highest. But this group has the least brand loyalty and the total volume of bath Soap items bought.
2)Cluster 1: This group of customers have highest volume of items bought and they are the group of customers who have the highest brand loyalty and value of items bought.
3)Cluster 2: This group of customers have a pretty much midway statistics, neither too high, nor too low. Also, they have the least maximum percentage of brand purchase.
4)Cluster 3: This group of customers have the highest volume of items bought and the maximum percent brand purchase, but have the lowest average price of items bought .

Question 1.b:The variables that describe basis-for-purchase. [Variables: purchase by promotions, price categories, selling propositions] [Note – would you use all selling propositions? Explore the data.]

Answer:1-b

We selected the following variables for k-mean clustering:

VARIABLES
Purchase_without promotion
Price_code
PropCat 5
PropCat 6
PropCat 8
PropCat 15

Parameters: Same as above question

Performance Vector

K	2	3	4	5	6
Avg. within centroid distance	0.835	0.712	0.592	0.493	0.383
Avg. within centroid distance_cluster_0	0.511	4.632	0.795	0.348	0.327
Avg. within centroid distance_cluster_1	1.153	0.871	0.435	0.35	1.716
Avg. within centroid distance_cluster_2	-	0.382	1.086	0.804	0.258
Avg. within centroid distance_cluster_3	-	-	0.638	2.52	0.376
Avg. within centroid distance_cluster_4	-	-	-	0.359	0.662
Avg. within centroid distance_cluster_5					0.107
Davies – Bouldin Index	0.298	0.266	0.204	0.176	0.193

EXPLANATION : According to the performance vector, If we look at Davies Bouldin's distance, the least is for K=5. But the greatest value for avg. within the distance with k=5 cluster is more

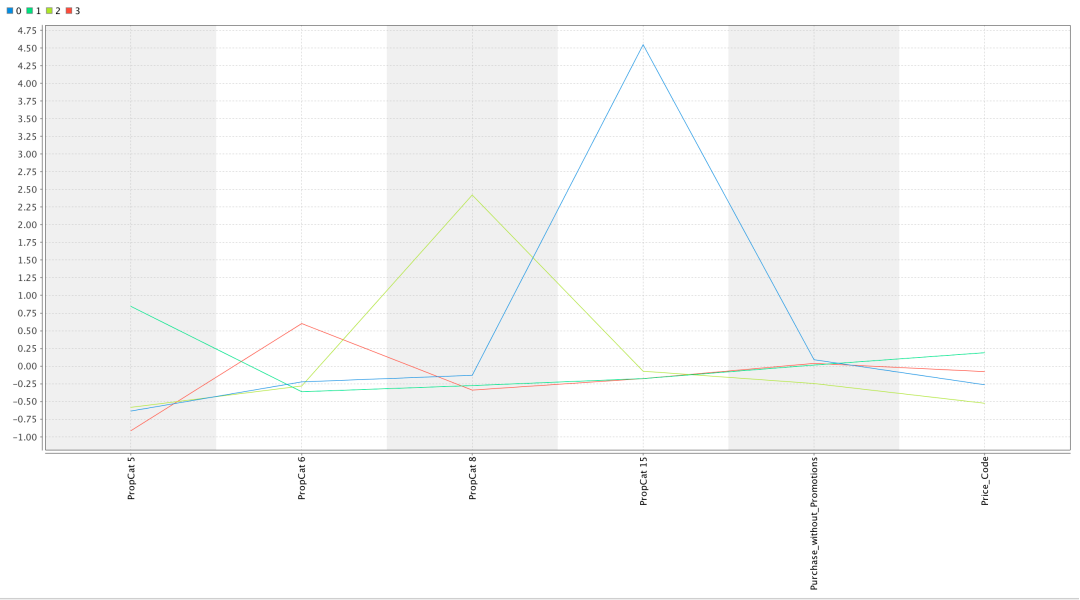
(2.52). Then the next smallest are with k=4 and K=6 and we ignore k=6 because again the average value within the cluster has a significantly high value for K=6,thus we finally stand for K=4

Similarity Measure object (Data to similarity)

K=2	**Cluster**	**Cluster 1**		**Cluster 2**		
	Cluster 1			1.977		
K=3	**Cluster**	**C1**		**C2**	**C 3**	
	Cluster 1			4.583	4.603	
	Cluster 2				2.080	
K=4	**Cluster**	**C1**	**C2**	**C3**	**C4**	
	Cluster 1	0	4.98	5.30	4.81	
	Cluster 2			3.14	2.03	
	Cluster 3				2.97	
K=5	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**
	Cluster 1		3.42	3.44	4.67	2.01
	Cluster 2			3.94	5.08	3.10
	Cluster 3				4.84	2.97
	Cluster 4					4.37
K=6	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
	Cluster 1		3.59	2.24	3.57	11.2	2.27
	Cluster 2			2.63	4.02	11.3	3.45
	Cluster 3				2.78	11.02	1.59
	Cluster 4					11.6	3.5
	Cluster 5						11.06

Explanation: We can ignore k=2 because it has significantly low values for the average distance between the clusters. Similarly in comparison to other number of clusters K=3 has the next low value of k=2.08. Also, We are not going in for cluster 6Because it has very high variations in the average distances from 2 to 11 . There is a close competition between K=4 and K=5.

We will go for cluster 4 as from the performance vector, Data to similarity within each cluster.



Analysis of Clusters with K=4

1)Cluster 0: This is the group of customers who are the most loyal ones since they the ones who buy items even without promotions also they have bought soaps more than anyone in the given good list.
2)Cluster 1: This is the group of customers who buy products in maximum volume also, the customers in this group are the ones who buy more beauty soaps than anyone else.
3)Cluster 2: These customers have bought bath soaps in a low volume and they have bought items with any freshness.
4)Cluster 3: These customers have bought More of health related Bath soaps have bought the least of freshness related bath soaps.

Question 1.c:The variables that describe both purchase behavior and basis of purchase.

Answer:1-c
We selected the following variables for k-mean clustering:

- Purchase_without promotion
- Price_code

- PropCat 5
- PropCat 6
- PropCat 8
- PropCat 15
- Total Volume
- Value
- Brand Loyalty
- Max_Percent_Brand_Purchase
- Avg Price

Parameters

Parameters	Values
K	2,3,4,5
Max runs	10
Measure Types	Mixed measures
Mixed Measures	Mixed Euclidean Distance
Max Optimization Steps	100

Performance Vector

K	2	3	4	5	6
Avg. within centroid distance	0.876	0.784	0.716	0.654	0.586
Avg. within centroid distance_cluster_0	0.811	0.615	0.556	0.633	0.437
Avg. within centroid distance_cluster_1	0.988	1.097	1.034	1.137	0.752
Avg. within centroid distance_cluster_2	-	0.942	1.051	1.223	1.045
Avg. within centroid distance_cluster_3	-	-	0.435	0.312	0.925
Avg. within centroid distance_cluster_4	-	-	-	0.499	0.491
Avg. within centroid distance_cluster_5	-	-	-		0.396
Davies Bouldin Index	0.119	0.094	0.089	0.157	0.140

EXPLANATION : According to the performance vector, If we look at Davies Bouldin's distance, the top 2 competitors are K=3 and K=4 .We go with K=4.Also, K=2 has fairly good distance within the cluster.

Similarity Measure object (Data to similarity)

K=2	**Cluster**	**Cluster 1**		**Cluster 2**			
	Cluster 1			2.417			
K=3	**Cluster**	**C1**		**C2**		**C 3**	
	Cluster 1			3.09`		2.47	
	Cluster 2					3.64	
K=4	**Cluster**	**C1**	**C2**	**C3**		**C4**	
	Cluster 1	0	2.27	3.23		2.46	
	Cluster 2			3.80		3.62	
	Cluster 3					3.83	
K=5	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**	
	Cluster 1	0	3.62	3.96	2.43	3.80	
	Cluster 2			4.28	2.40	3.41	
	Cluster 3				3.46	4.33	
	Cluster 4					3.31	
K=6	**Cluster**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
	Cluster 1	0	5.23	4.22	3.69	3.99	2.48
	Cluster 2			5.98	5.10	5.65	4.99
	Cluster 3				4.50	4.58	3.68
	Cluster 4					3.48	2.38
	Cluster 5						3.430

Explanation : The Average within centroid distance for k=4 and k=2 is the least and between the cluster is the most.

Question 2.a: (a) Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering. 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?

Answer:2-a

We go with 1-c) and thus selected the same variables for k-mediods clustering:
As we got K=2 from 1-c) but from rest of the above parts we were getting the best clustering with K=4, Therefore, we decided to record values for both K=2 and K=4

K medoids: Performance Vector within the cluster

K	2	4
Avg. within centroid distance	1.282	1.109
Avg. within centroid distance_cluster_0	1.335	1.146
Avg. within centroid distance_cluster_1	1.182	1.397
Avg. within centroid distance_cluster_2	-	0.924
Avg. within centroid distance_cluster_3	-	1.087
Davies Bouldin Index	0.116	0.131

EXPLANATION : According to the performance vector, If we look at Davies Bouldin's distance, there isn't much difference between K=2 and K=4,still k=2 has a lower value. But if we see the average value within the cluster both have the same performance.

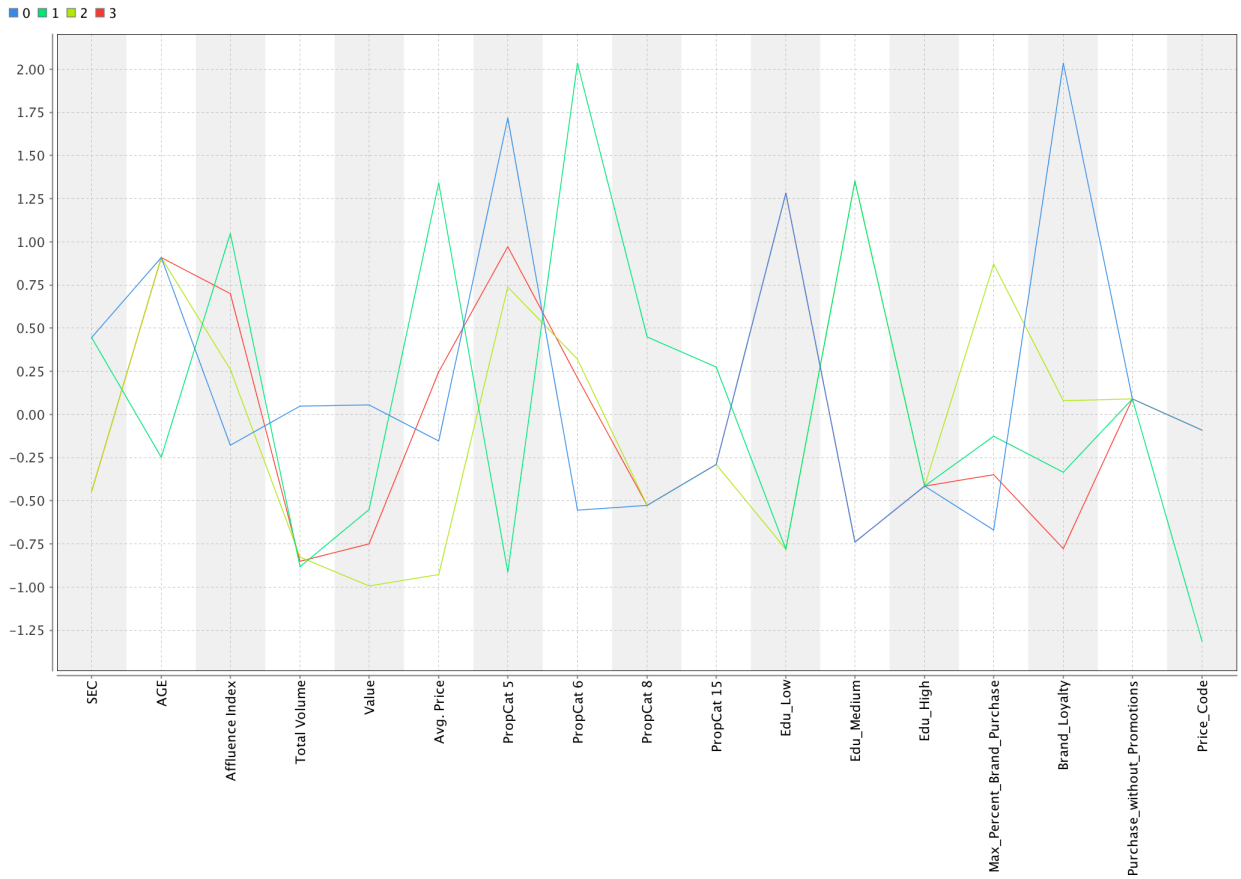
Avg. Distance between the cluster for K=2

AVG DISTANCE	**Cluster 1**	**Cluster 2**
Cluster 1	0	4.470

Avg. Distance between the cluster for K=4

AVG DISTANCE	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Cluster 1	0	6.101	4.470	3.513
Cluster 2		0	4.207	4.689
Cluster 3			0	3.541

EXPLANATION : According to the Average distance between the clusters, we see that K=4 has an upper hand with higher values.



Analysis of Clusters with K=4

1)Cluster 0: The socio-economic status of the customer in this range is the lowest, and these people buy beauty products, but these group people are the one who have the low education level and highest brand loyalty.

2)Cluster 1: The socio-economic status of the customer in this range is the lowest, with highest Affluent Index, and they buy Bath soaps with highest average price and mostly health related bath soaps. These group of people have mid-level education level and they buy any sub-popular items.

3)Cluster 2: These group of people have the highest socio-economic status , and they buy the lowest value Bath soaps, also they are the ones who aren't much educated and have low education levels, with maximum percentage of brand purchase items as their history of bought items.

4)Cluster 3: These group of people are mostly old age people who buy less volume of Bath soaps and are loyal customers who buy products even within promotions

Kernel K- means

Parameter	k	Max Optimization Steps	Kernel type	Kernel Gamma	Kernel A	Kernel B	Kernel Degree	Use Weights
Value 1	2	100	radial	1.0	-	-	-	No
Value 2	4	100	Dot	-	-	-	-	No
Value 3	4	100	Polynomial	-	-	-	3	No
Value 4	4	100	Sigmoid	-	1	2	-	No

Distribution of data items

Cluster Distribution	Value 1	Value 2	Value 3	Value 4
Cluster 0	300	89	5	128
Cluster 1	300	229	8	197
Cluster 2		67	4	94
Cluster 3		215	583	181

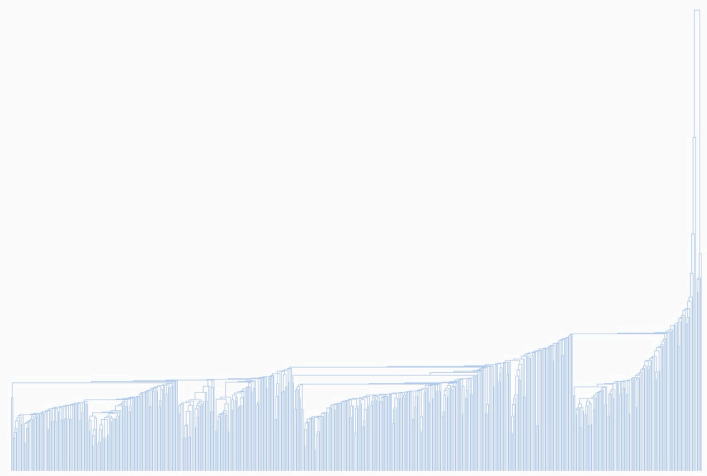
EXPLANATION: The best results were obtained with the kernel type Sigmoid because the best and even distribution of items were with the clusters of value 4 , while the rest has very uneven distribution.

Agglomerative clustering

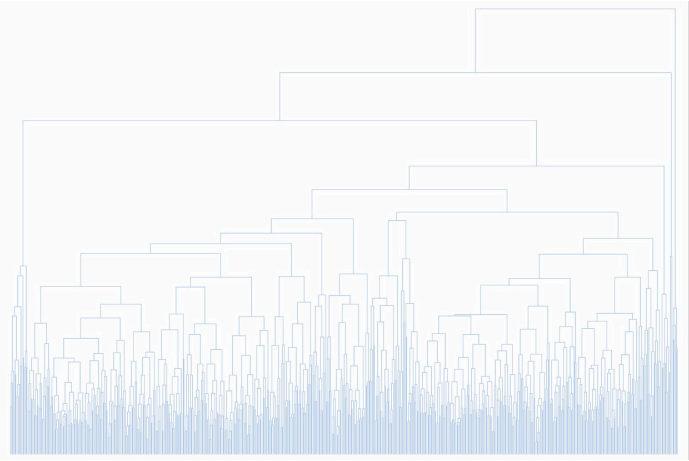
Parameter	**Values 1**	**Values 2**	**Values 3**
Mode	Single Link	Complete Link	Average Link
Measure Types	Numerical Measures	Numerical Measure	Mixed Measures
Numerical Measure	Euclidean Distance	Euclidean Distance	Mixed Euclidean Distance
Number of clusters	1199	1199	1199

So, here we see the number of clusters as 1199 which is not possible as we have just 600 items as our items to be clustered, so our maximum number of clusters can be 600,thus we can't trust the agglomerative clustering methods for our data items.
We then used Flatten Clustering. But one cluster took a lot of data ,so the number of clusters in the parameter was increased to attain good results ,but this method is again inefficient for the data analysis.

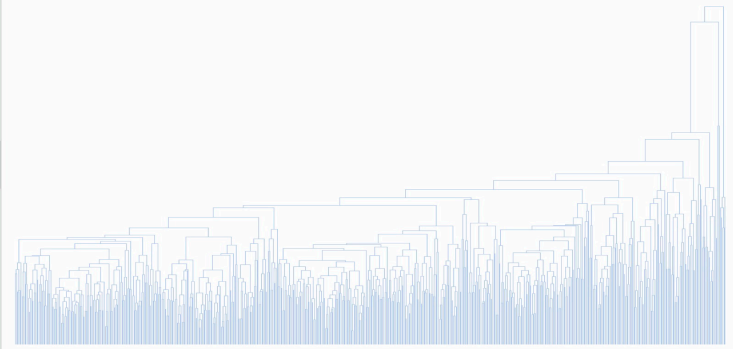
SINGLE LINK:



COMPLETE LINK:



AVERAGE LINK:



EXPLANATION: From the above dendrograms , the complete and average linkages are better than the single linkage. And, amongst complete and average linkages, the complete linkage is better one.

DB SCAN:

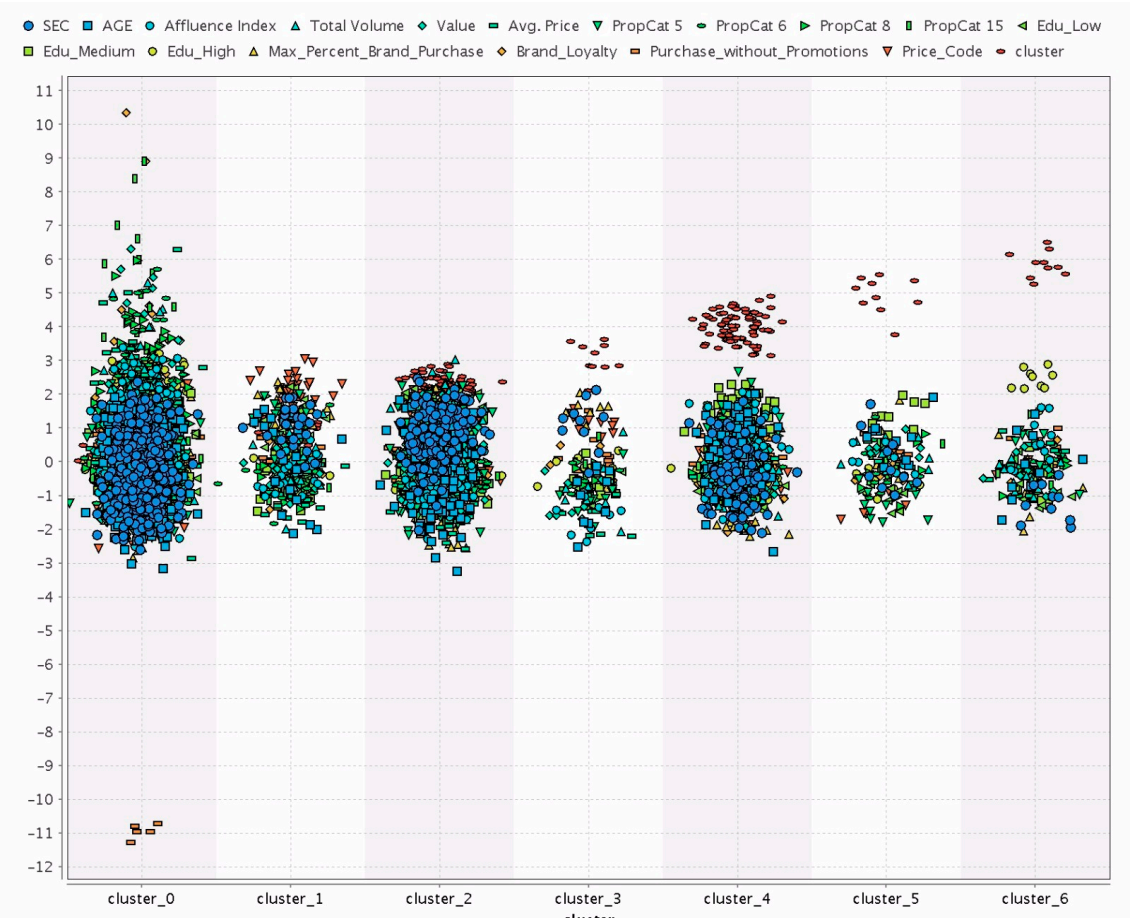
Parameters	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7
Epsilon	1.5	1.6	1.6	1.7	1.9	1.8	2.1
Min Points	10	15	10	10	10	15	10
Measure Type	Mixed Measures	Mixed Measures	Numerical Measures	Numerical Measures	Numerical Measures	Numerical Measures	Mixed Measures
Mixed Measure	Mixed Euclidean distance	Mixed Euclidean distance	Euclidean Distance	Euclidean Distance	Euclidean Distance	Euclidean Distance	Mixed Euclidean Distance

Distribution of data items

Clusters	Value 1	Value 2	Value 3	Value 4	Value 5	Value 6	Value 7
Cluster 0	590	600	571	514	455	562	346
Cluster 1	10		19	31	66	38	24
Cluster 2			10	24	30		124
Cluster 3				20	49		10
Cluster 4				11			76
Cluster 5							10
Cluster 6							10

EXPLANATION: We see that the distribution of Value 1 &2 is very bad that is the one with low epsilon value. As the epsilon value increases, it leads to better distribution and we get the best distribution of items clustered. And the best distribution is obtained at the highest value of epsilon (2.1).

Considering Epsilon =2.1 we made this clustering graph.



We see that the clusters are uniformly distributed.

Question:2 (b) Are the clusters obtained from the different procedures similar/different? Describe in what ways they are similar/different. What might be some reasons for differences in clusters obtained using different procedures?

Answer:2- No, not all the techniques gave the same results. The following are the results from each type of clustering algorithms:

1) **K medoids:** The number of cluster= 4 gives the best clusters.
2) **Agglomerative clustering:** This clustering techniques doesn't give a good clustering.
3) **Kernel K- means:** In this the sigmoid function give the best clustering.
4) **DB SCAN:** It gives a good clustering at higher epsilon values.

So we are using different clustering algorithms for getting results. We are using Partitioning and hierarchical clustering techniques.

The partitioning Algorithm like K-medoids has better distribution than the K-means which is a hierarchical distribution.

(c) Which would you pick as your 'best' and why?

Answer c:

Within the partitioning algorithm, the k-means works better than k-mediods because there is less distance within the clusters but more distance amongst the cluster.

Also, K-mediod is more flexible, Robust ,But much more expensive. Also,the medoid as used by k-medoids is roughly comparable to the *median* **the median is more robust to outliers than the arithmetic mean**.

3. (a) Select what you think is the 'best' segmentation - explain why you think this is the 'best'. You can also decide on multiple segmentations, based on different criteria -- for example, based on purchase behavior, or basis for purchase,....(think about how different clusters may be useful.

Answer:3 a) DAVIES BOULDIN INDEX: It is the ratio of the within cluster scatter and the between cluster separation, a lower value will mean that the clustering is better

With K-means.

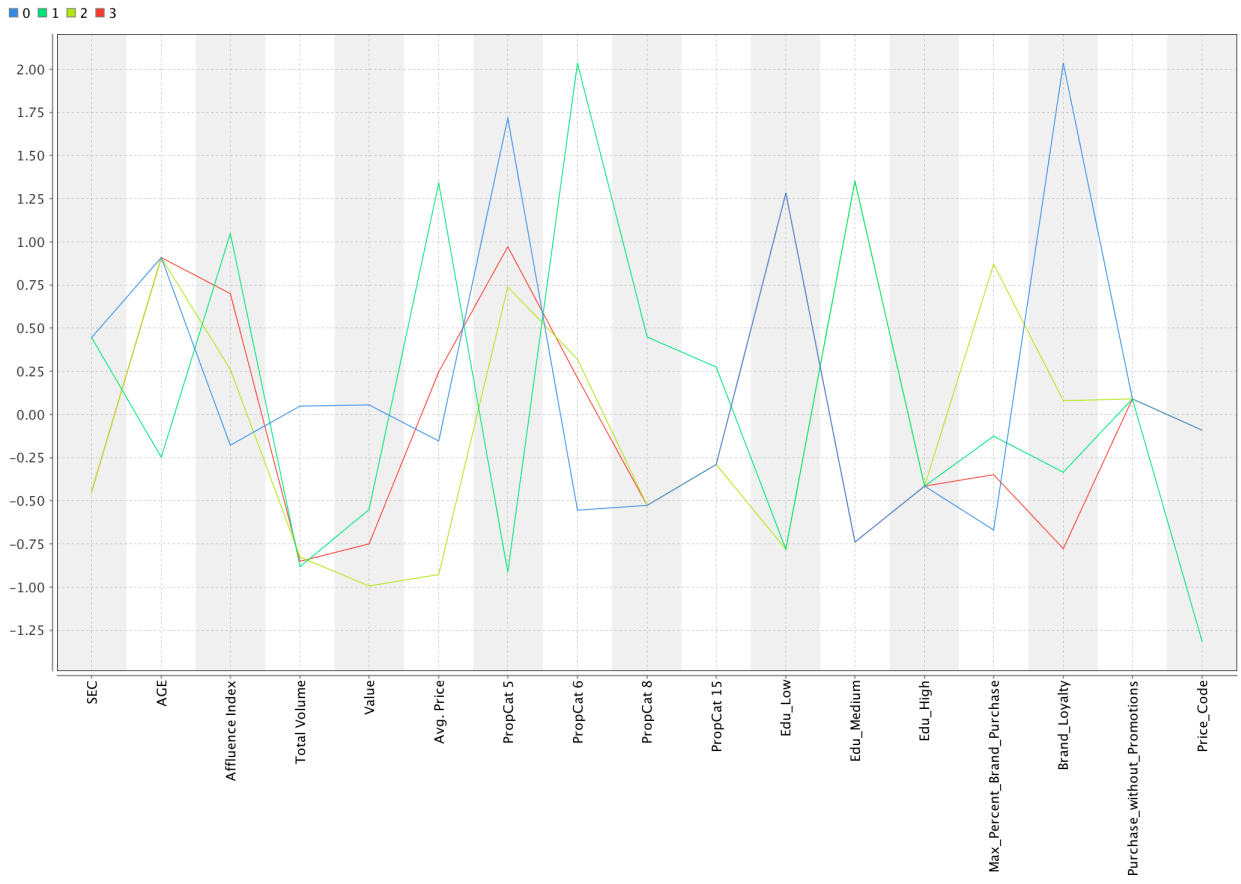
Purchase Behavior	**K=2**	0.280
	K=3	0.288
	K=4	0.252
	K=5	0.257
	K=6	0.252
Basis-for-purchase	**K=2**	0.298
	K=3	0.266
	K=4	0.204
	K=5	0.176
	K=6	0.193
Purchase Behavior + Basis-for-purchase	**K=2**	0.119
	K=3	0.094
	K=4	0.089
	K=5	0.157
	K=6	0.140

With K-mediods

Purchase Behavior + Basis-for-purchase	K=2	0.116
	K=4	0.131

Explanation: The best cluster with K-means :Davie's Bouldin index is with Purchase Behavior and Basis for purchase for the number of clusters=4.
With the K-mediods method the best cluster comes out with K=2.
In general, It would be the best to consider both the variables for market segmentation and considering our target customers.

(b) Comment on the characteristics (demographic, brand loyalty and basis-for-purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)
Answer: b)
The best segmentation chose is k-medians using variables basis of purchase and purchase behavior with the number of clusters as 4.



Analysis of Clusters with K=4

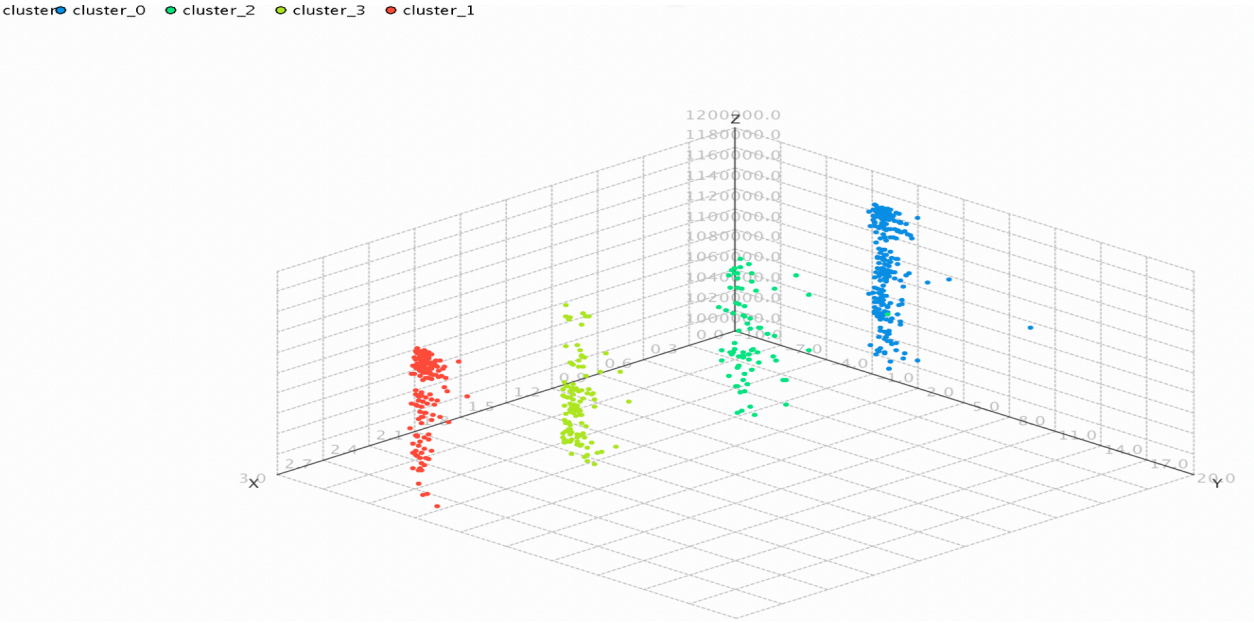
1)Cluster 0: The socio-economic status of the customer in this range is the lowest, and these people buy beauty products, but these group people are the one who have the low education level and highest brand loyalty.

2)Cluster 1: The socio-economic status of the customer in this range is the lowest, with highest Affluent Index, and they buy Bath soaps with highest average items and mostly health related items. These group of people have mid-level education level and they buy any sub-popular Bath soaps.

3)Cluster 2: These group of people have the highest socio-economic status , and they buy the lowest value Bath soaps, also they are the ones who aren't much educated and have low education levels, with maximum percentage of brand purchase items as their history of bought items.

4)Cluster 3: These group of people are mostly old age people who buy less volume of Bath soaps and are loyal customers who buy products even within promotions.

Cluster	High in	Low in
Cluster 0	Age, Total volume of Bath soaps ,Value of items, Buying beauty products, Brand loyalty.	Socio-economic status, Affluent Index, Buying health related Bath soaps ,Education level
Cluster 1	Affluent Index, Eductaion-medium,	Socio-economic status, price code,
Cluster 2	Socio-economic status, Brand loyalty, Maximum percentage brand loyalty	Value of Bath soaps
Cluster 3	Age, Maximum percentage brand loyalty	Brand loyalty



We see that all the four clusters are neatly segregated.

(c) For the best segmentation, obtain a description of the clusters. You may base this on attributes describing the clusters (not restricted to attributes used for clustering). You should also build a decision tree to help describe the clusters – how effective is the tree in identifying the different clusters? Does the tree help in explaining/interpreting the different clusters? (explain why/why not). (You may use a decision tree to help choose the 'best' clustering).

Answer:3-c:

The best segmented cluster obtained was with K-median with K=4 and Davies bouldin index as 0.0089.

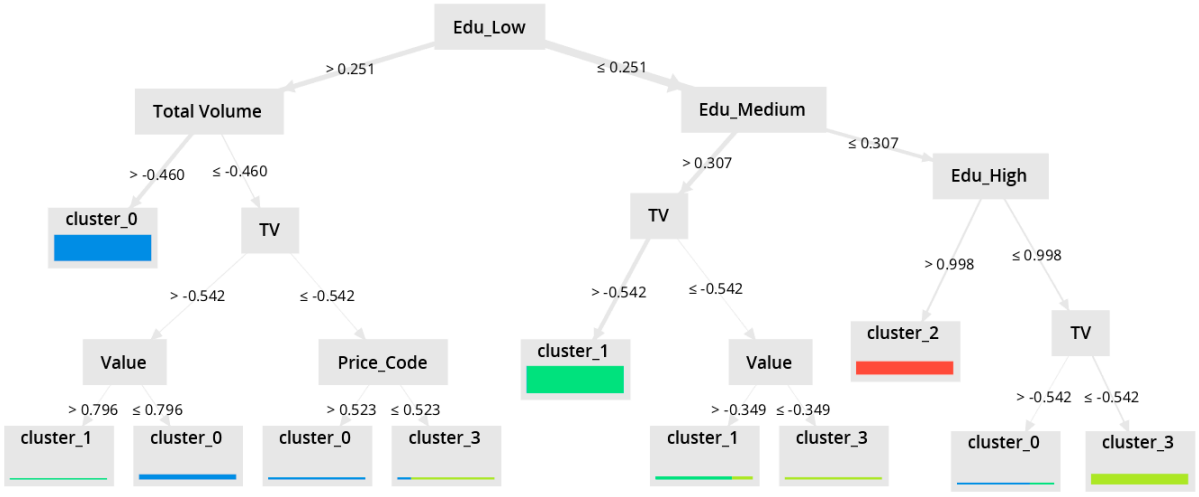
Variables used:

- SEC
- Price Code
- Brand Loyalty
- TV
- Maximum percentage of brand purchase

- Low, medium and high education
- value
- Total Volume
- Average Price.

Decision Tree:

Parameters	Value
Criterion	Information_gain
Maximal depth	7
Apply Pruning	Yes
Confidence	0.254
Apply Prepruning	No



accuracy: 99.00%

	true cluster_0	true cluster_1	true cluster_3	true cluster_2	class precision
pred. cluster_0	222	1	0	0	99.55%
pred. cluster_1	1	201	3	0	98.05%
pred. cluster_3	1	0	83	0	98.81%
pred. cluster_2	0	0	0	88	100.00%
class recall	99.11%	99.50%	96.51%	100.00%	

According to the decision tree the education level, Total volume of items bought and Owning a TV are the important variables in determining a cluster.

1)Cluster 0: If Education_low>0.251 and total volume of bath soaps purchase is >-0.460 then cluster 0.
2)Cluster 1: If Education_medium>0.307 and TV >-0.542 then cluster 1.
3)Cluster 2: If Education_high<=0.998 and Tv owners<=-0.542 then cluster 2.
4)Cluster 3: If Education_high>0.998, then cluster 1.

Overall the decision tree gave us the important factors and their grouping, they should be used along with the clustering techniques.