CRISA Soap market segmentation using Clustering

Introduction:

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods.

It has both transaction data (each row is a transaction) and household data (each row is a household), and, for the household data, maintains the following information: -

Demographics of the households (updated annually) -Possession of durable goods (car, washing machine, etc.; updated annually) and a computed audience index" on this basis -Purchase data of product categories and brands (updated monthly).

CRISA has two categories of clients:

(1) Advertising agencies who subscribe to the database services; they obtain updated data every month and use it to advise their clients on advertising and promotion strategies.

(2) Consumer goods manufacturers who monitor their market share using the CRISA database

CRISA, which caters primarily to ad agencies and consumer product manufacturers, has been tasked with segmenting the consumer market. And in doing so, hopes to divide the market of potential customers into groups, or segments.

The segments created are composed of consumers who will respond similarly to marketing strategies and who are predicted to share traits such as similar interests, needs, or locations.

They would now like to segment the market based on **two key sets of variables more directly related to the purchase process and to brand loyalty which are buying behavior and the basis of purchase**. Doing so would allow CRISA to accumulate data about what demographic characteristics are associated with different purchasing behaviors and degrees of brand loyalty, and more effectively deploy promotion budgets. This allows their clients to precisely target a consumer with specific needs and wants. In the long run, this benefits the company because they can use their corporate resources more effectively and make better strategic marketing decisions. This would also result in a more cost-effective allocation of the promotion budget to

different market segments. It would also enable CRISA to design more effective customer reward systems and thereby increase brand loyalty.

Clustering:

Identifying clusters of households based on the customer demographics and purchase behaviors are primarily supported by the value of k in the k-means clustering algorithm. Two of the factors of concern are the **within-cluster distance** 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 are not clusters significantly smaller as compared to the other clusters.

Some data-transformations prior to applying algorithms were needed:

- Data was loaded as all continuous. However, due to business sense nominal variables are converted into factors. Ordinal variables are not changed because the values are discrete and maintain their order.
- Nominal variables are then converted to continuous variables by one-hot encoding. These variables were: **Food eating habits, child possession, gender, mother tongue and availability of television.**

a.) Clustering on the basis of Purchase behaviour (including brand loyalty):

We found out that there are certain **measures of brand loyalty** including **number of brands, number of transactions per brand, maximum number of times a household buys that brand** and also **how often a brand is switched by that household**.

Moreover, there are certain variables in the data consisting of brand codes which sum up to 100%. So, we can make a variable maximum brand by taking maximum of these variables. This is because if a customer is loyal to brand A and the second customer is loyal to brand B that means they both are considered to be loyal towards their respective brands.

We also made a variable, "**brnd_lylty**" which means brand loyalty with the help of above mentioned variables as:

Brand Loyalty = Proportion of maximum used brand - Proportion of switch to other brands + Number of transactions per brand - Total number of brands

Two variables are subtracted because they are negatively correlated with the brand loyalty.

Note: We first scaled the variables because this formula would become absurd if the scale of any variable is much bigger than other variables.

Also, we **dropped brand_runs** which means consecutive purchases of that particular brand. As the logic is flawed. For example, a customer buys only two brands with consecutive streaks of 5 for the first brand and then once for the second brand whereas some other customer buys 4 brands with consecutive streaks of 4,4,5 and 6. So, this "brand_runs" variable implies that the second customer is more loyal to the 4th brand than the first customer to the 1st brand which is wrong.

Moreover, we removed two more variables:

- **Value** - It was highly correlated with Number of transactions.
- **Volume per transaction** - We already have volume and transaction in our data which makes more sense than volume/transaction. For instance: if a customer buys 50 items from me 50 times per month and there is another customer who buys 2 items from me 2 times per month. Even though volume per transaction is the same for both the customers but still revenue generated from the first customer is way higher than the second one.

Now, variables regarding purchase behaviour are:

- Total Volume
- Number of transactions
- Average Price spent per purchase
- Brand Loyalty

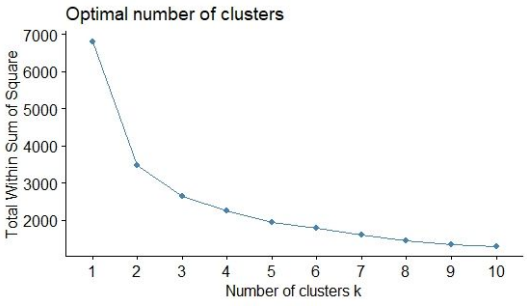
The above mentioned variables are then passed into the K-means algorithm.

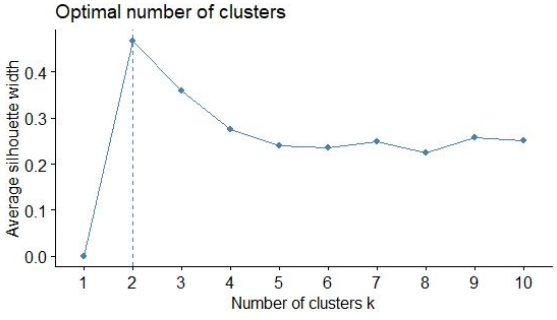
K-means algorithm has 2 parameters to play with: **# of clusters** i.e. k and the second is **nstart.**

Nstart is fixed to 30 as it suggests how many random sets should be chosen. Sometimes, a smaller number of nstart gives the wrong result as it doesn't allow the centroid to settle at the local optimum. So, it is better to fix "nstart" to a larger number such as 20,30 or 50.

We first set k to 3 according to business knowledge. We assumed that in the presence of purchase behavior variables, brand loyalty can be high, low or neutral.

We then find best k using **elbow-method** and **silhouette** method:





Optimum number of clusters using the elbow method is either **2** or **3** and using the silhouette method is **2.**

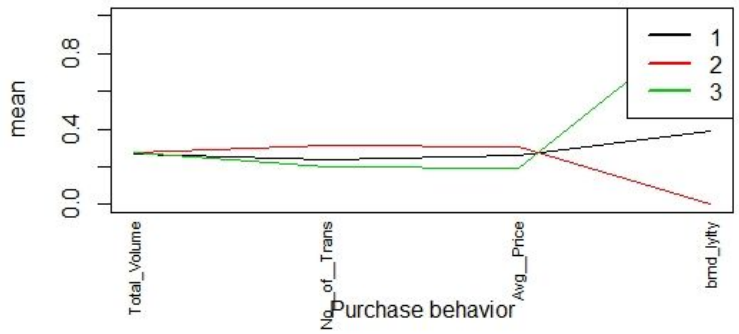
We then ran k means on 2, 3 as well as 4 to check other measures.

k	DB	size of clusters
2	1.045	157, 443
3	0.6377	212,307,81
4	0.4931	272,196,107,25

Here, we can see that Davies-Bouldien index for 2 is relatively too higher than 3 and 4. Moreover, it's second cluster is too generic. And for k = 4, there is a cluster which has only 25 observations which are too uneven.

Due to **DB index and cluster size trade-off**, k=3 seems a better clustering than both 2 or 4 which was our earlier assumption too according to the business knowledge.

It's silhouette width is also **0.32** which is decent enough.



Here, we can see that among purchase behaviour variables, **brand loyalty** is the influencer here. So, k=3 has segregated the observation into 3 clusters mainly on the basis of brand loyalty.

	clusKM	EDU	Affluence_Index	Total_Volume	No__of__Trans	Avg__Price	brnd_lylty
1	1	3.773585	15.05189	11531.56	26.87736	11.361748	0.9011784
2	2	4.472313	19.76221	12091.49	36.64495	12.785781	-2.0918747
3	3	3.123457	11.77778	12247.96	21.53086	9.466885	5.5698236

Interpretation of the clusters are:

- **Cluster 1 -** These households have medium education level, medium affluence index, medium number of transactions and the average price per purchase also lies between cluster 2 and cluster 3. However, in terms of brand loyalty, they are neutral.
- **Cluster 2 -** These households are well-educated possessing high affluence index. Their number of transactions and average price per purchase are also high. However, their brand loyalty score is negative which suggests that they switch to different brands more frequently.

- **Cluster 3 -** These customers are less-educated possessing low affluence index and their number of transactions are also low. They spend less money per purchase. However, they stick to their favorite brands as their brand loyalty score is quite high.

b. Clustering on basis-for-purchase.

There are certain measures which describe basis-for-purchase:

- **Promotions -** There are three variables - percent of volume purchase under no promotion, percent of volume purchased under banded offer or any other offer. On the basis of analysis, we figured out that more than 90% of the data were no promotion data which made this variable useless and promotion from other offers were less than 3%. Therefore, we dropped these two variables.
- **Price categories -** There are 4 variables which determine price category from the type of soap.
- **Selling Propositions -** There are certain proposition categories like beauty, herbal, baby, cabolic, glycerine, fairness, skincare, freshness, etc. We categorized these variables in only three categories: **beauty, health** and **others.** We figured out that Price category 3 is highly correlated with the newly made “health” category. This makes this health category redundant. “Others” category was not providing us any information too after exploring the data through EDA. Therefore, we dropped these two.

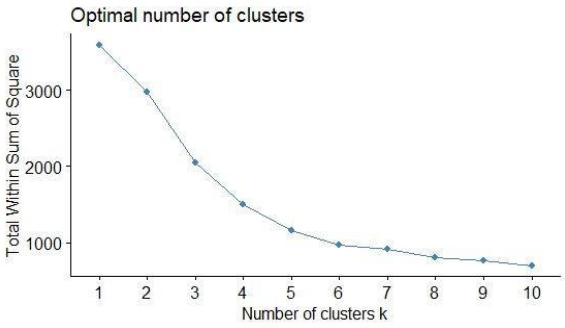
Now, variables regarding basis-for-purchase are:

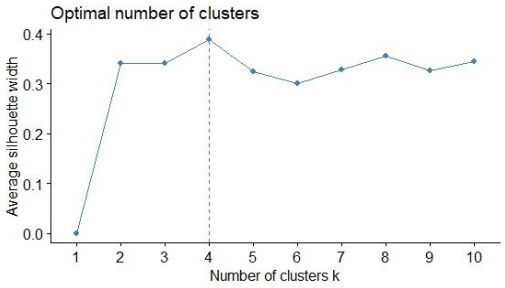
- Purchase volume with promo 6 (banded offer)
- Price category 1
- Price category 2
- Price category 3
- Price category 4
- Proposition of beauty soaps

The above-mentioned variables are then passed into the K-means algorithm

We first set k to 4 according to the business knowledge. We assumed that the price category might be impactful while choosing clusters.

We then find best k using **elbow-method** and **silhouette** method:





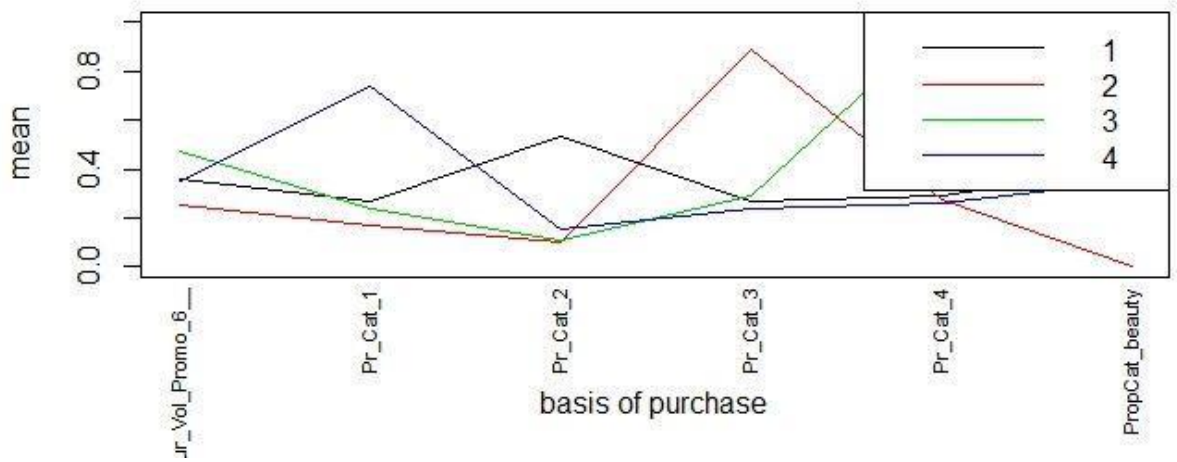
Optimum number of clusters using the elbow method is **4** and using the silhouette method is **4** too**.** This was exactly our assumption.

To further check other measures, we then ran k means on 2, 3, 4 and 5.

k	DB	size of clusters
2	2.952	513,87
3	1.3173	310,209,81
4	0.6905	332,128,83,57
5	0.4741	51,119,41,83,306

Here, we can see that Davies-Bouldien indexes for 2 and 3 are relatively too higher than 4 and 5. Also the cluster size is not acceptable in 2 and 5.

Therefore, due to **DB index, cluster size trade-off, elbow-method, silhouette method**, k=4 is a clear winner.



Here, we can see that there are 4 peaks at each of the price categories. That means k=4 has segregated the customers in such a way that they purchase any of these 4 categories.

	clusKM	EDU	Affluence_Index	AGE	Pur_Vol_Promo_6__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
1	1	4.355422	18.078313	3.234940	0.05513837	0.17901379	0.7318098	0.04826307	0.04091333	0.7248086
2	2	2.481928	9.144578	3.036145	0.01348648	0.05768982	0.1647835	0.74805954	0.02946710	0.1586160
3	3	3.350877	14.192982	3.263158	0.10044152	0.14571452	0.1724969	0.06788830	0.61390025	0.8592975
4	4	4.554688	20.640625	3.250000	0.05429327	0.74137063	0.2298029	0.01203571	0.01679078	0.6470410

Interpretation of the clusters are:

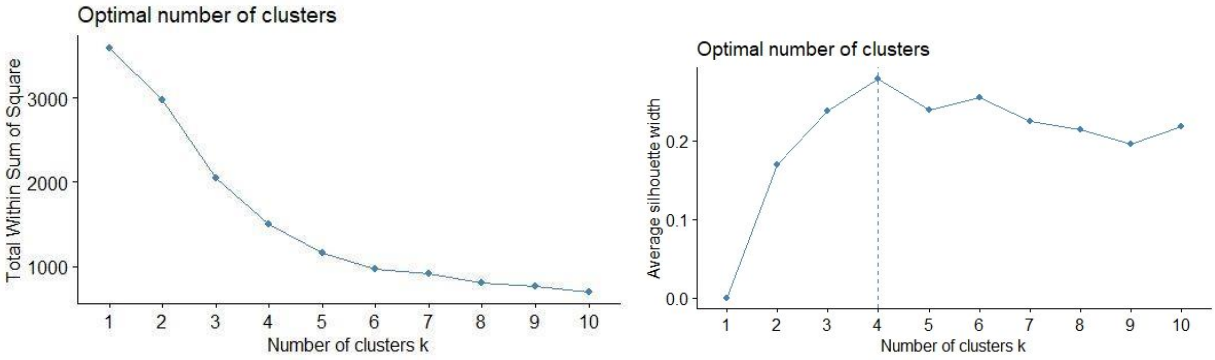
- **Cluster 1 -** These households are major buyers of price category 2 which are “popular soaps”. They also have a high education level, high affluence index and they buy a high proportion of beauty soaps which must be most popular in India. All these make sense.
- **Cluster 2 -** These households are major buyers of price category 3 which are “economical/herbal” soaps. They are less-educated possessing a low affluence index. They also buy a minimum number of beauty soaps which makes sense.
- **Cluster 3 -** These customers are major buyers of price category 4 soaps which are “sub-popular” soaps. However, those soaps must be related to the beauty segment as they usually buy 85% of them. They are medium-educated possessing a medium affluence index and surprisingly, they have way higher percentage of promotion 6 codes than any other cluster which means they can be targeted by “sub-popular” brands if they sell these customers banded soaps.
- **Cluster 4 -** These customers are major buyers of price category 1 that are “premium soaps”. They are highly educated with relatively higher affluence index than others. This suggests that they can afford premium soaps.

c. Clustering on both purchase behavior and the basis of purchase.

Here, we used the above mentioned variables of both segments and passed them into the K-means algorithm

We first set k to 4 according to the business knowledge.

We then find best k using **elbow-method** and **silhouette** method:



Optimum number of clusters using the elbow method is **4** or **5** and using the silhouette method is **4.**

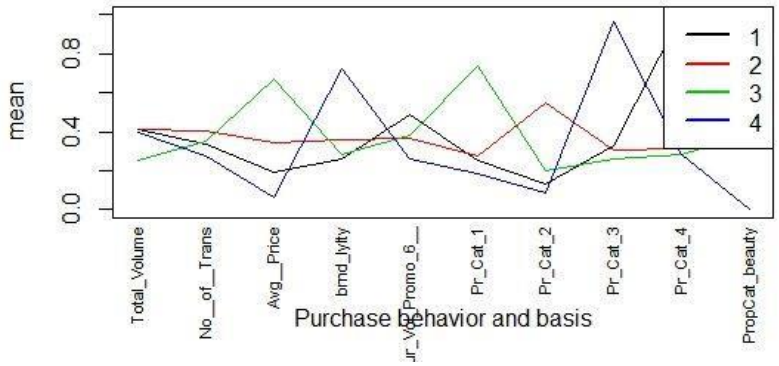
To further check other measures, we then ran k means on 2, 3, 4 and 5.

k	DB	size of clusters
3	1.8697	141,382,77
4	1.1982	329,141,70,60
5	0.9186	119,56,157,201,67
6	0.7429	55,185,139,108,46,68

Here, we can see that Davies-Bouldien index for 3 is relatively higher than 4 and 5. Also the cluster size is not acceptable in 2. However, when we took 6 clusters, we saw that it provides the same information as in 4 or 5.

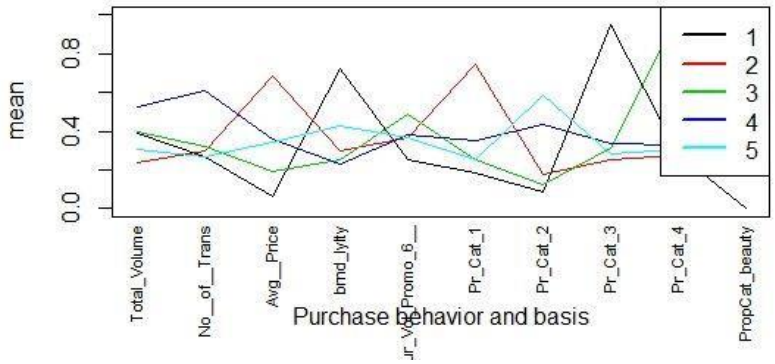
Therefore, due to **DB index, cluster size trade-off, elbow-method, silhouette method**, k=4 seems best. However, we cannot ignore 5 too. The reason is explained below:

For **K=4,**



Here, we can see that there are 4 segments made mainly on the basis of Price categories - 1,2,3 and 4. However, in price category 2, “popular soaps”, there are around 329 observations so it’s quite generic. So we had to do some pre-processing. We tried to figure out whether this problem was solved by k means with 5 clusters otherwise we needed to run a model on k=2 on only those segments. However, surprisingly, the problem got solved.

For **K=5**,



Here, we can see that there are 2 peaks in price category 2 - cluster 4 and cluster 5. So, now cluster 4 is broken into 2 decent sized clusters - 157 and 201. Moreover, we observed that cluster 4 segment customers buy in large volume with relatively high number of transactions.

This would be helpful in making strategies further.

	clusKM	EDU	Affluence_Index	Total_Volume	No__of__Trans	Avg__Price	brnd_lylty	Pur_Vol_Promo_6__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
1	1	2.223881	7.761194	12555.821	23.62687	6.842014	4.4653004	0.006330285	0.05554489	0.1077673	0.821305107	0.01538274	0.1217251
2	2	4.462185	20.025210	7427.311	25.96639	16.967704	-0.8888349	0.053869017	0.74314553	0.2344148	0.009101884	0.01333783	0.6772723
3	3	3.267857	13.321429	12854.464	27.58929	8.913597	-1.4059839	0.101688500	0.13776876	0.1664311	0.079448928	0.61635126	0.8488258
4	4	4.566879	19.757962	17353.981	49.59873	11.659778	-1.7083899	0.058668794	0.25829208	0.5866895	0.099615940	0.05540251	0.5993747
5	5	4.208955	17.218905	9847.498	23.31841	11.410073	0.7639230	0.051541903	0.13432472	0.7927305	0.036433221	0.03651154	0.7722175

Interpretation of the clusters are:

- **Cluster 1 -** These households are major buyers of price category 3 which are “economical/herbal” soaps. They are really less-educated customers possessing a low affluence index. Their average price per purchase is also quite low. However, they are the most loyal customers towards their favorite brands.
- **Cluster 2 -** These customers are major buyers of price category 1 that are “premium soaps”. Hence, their average price per purchase is the highest among all. They are highly educated with relatively higher affluence index than others. This suggests that they can afford premium soaps.
- **Cluster 3 -** These customers are major buyers of price category 4 soaps which are “sub-popular” soaps. However, those soaps must be related to the beauty segment as

they usually buy 85% of them. They are medium-educated possessing a medium affluence index and their average price per purchase is also low as they have possess a higher percentage of promotion 6 codes than any other cluster which means they can be targeted by “sub-popular” brands if they sell these customers banded soaps.

- **Cluster 4 -** These households are major buyers of price category 2 which are “popular soaps”. They also have a high education level, high affluence index and they buy a high proportion of beauty soaps which must be most popular in India. Additionally, their number of transactions and volume are the highest among similar sized clusters.
- **Cluster 5:** They possess the same traits as cluster 4 customers. However, they purchase less amount of soaps and their number of transactions is also quite lower than cluster 4 segmentation. They can be targeted by the popular brands by giving discounts in order to increase the sale among these customers.

Note: Volume is not making much effect here as the number of observations are different in different clusters.

3. For this task, we chose **Agglomerate Hierarchical Algorithm(agnes) and DBScan Algorithm**.

We performed the same data transformation for the above two algorithms as mentioned in the KMeans. The variables were then passed into both the algorithms.

Agglomerate Hierarchical Algorithm(agnes) uses two parameters namely, dissimilarity measure and linkage. Below are the results we found.

Dissimilarity parameters used: **Euclidean** and **Pearson**.

Linkage parameters used: **Average, Single, Complete and Ward**

For Purchase Behavior:

average	single	complete	ward
0.93539	0.850226	0.968542	0.992368

Dissimilarity measure = “Euclidean”.
Linkage = “Average”, “Single”, “Complete”, “Ward”

This is clear from the **Dissimilarity measure = “Euclidean”, Linkage method = “Ward”** is giving better results compared to rest.

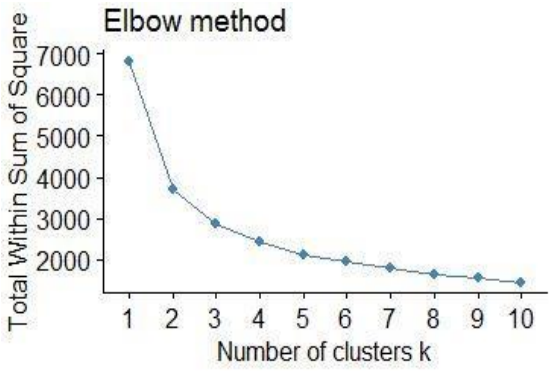
We then tried measuring correlation distance and thus chose **Dissimilarity measure = “Pearson”. Even here Linkage method = “Ward”** is giving better results compared to rest.

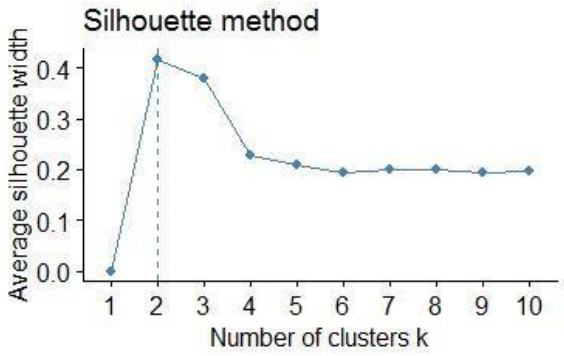
average	single	complete	ward
0.99711	0.969607	0.9977208	0.999829

Dissimilarity measure = "Pearson".
Linkage = "Average", "Single", "Complete", "Ward"

Among both the dissimilarity measures, **the "Pearson" correlation distance method is doing better**. So we chose it for our analysis.

We then find the best cut using the **elbow-method** and **silhouette method**.





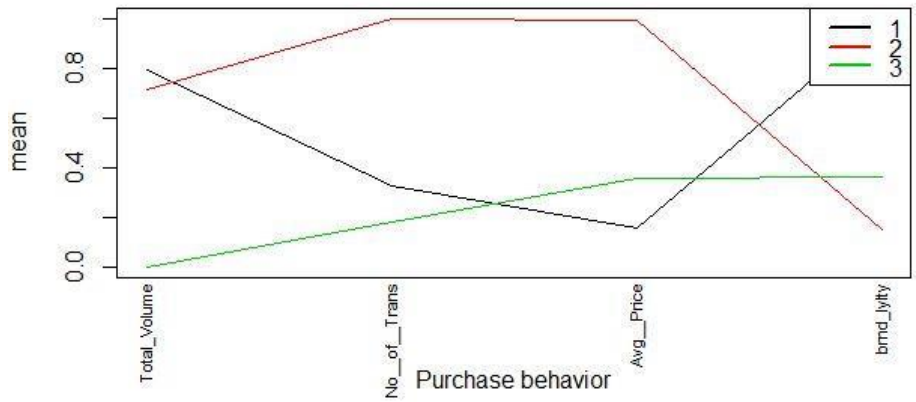
Davies-Bouldin's index

k	DB Index	Size of cluster
2	1.1287	257 343
3	1.48149	257,262,81
4	1.70333	196,262,61,81
5	1.63736	162,262,61,34,81

Here, we can see that Davies-Bouldien index for 4 is relatively too high than 2 and 3. But the silhouette width is relatively quite low for index = 4. For k=2 has better silhouette width = 0.38 but the second cluster is too generic.

Thus, due to **DB index and cluster size trade-off**, **cut=3** seems a better clustering than both 2 or 4.

It's silhouette width is also **0.19**.



Here, we can see that like Kmeans interpretation, **brand loyalty** is the influencer here. So, cut=3 has segregated the observation into 3 clusters mainly on the basis of brand loyalty.

Filter

	clus_aeg	EDU	Affluence_Index	Total_Volume	No__of__Trans	Avg__Price	brnd_lylty
1	1	3.509728	13.48249	11950.700	23.53307	10.49300	2.514054
2	2	4.530534	20.26718	13838.500	41.68702	11.76104	-2.051250
3	3	4.160494	17.74074	5578.333	21.25926	16.32901	-1.341784

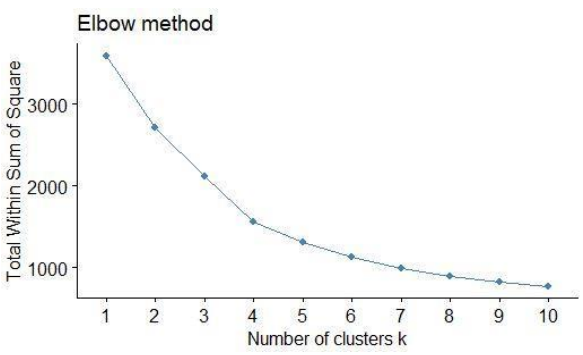
Interpretation of the clusters are:

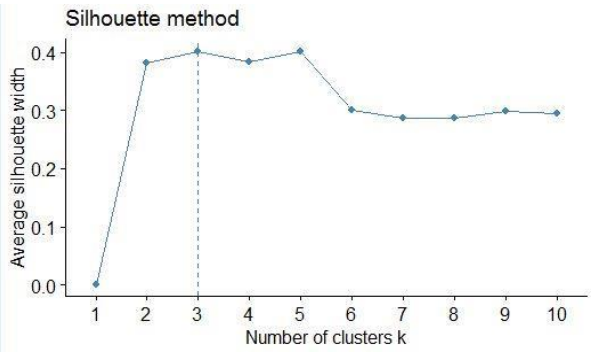
- **Cluster 1 -** This cluster is similar to cluster 3 found out in k means.
- **Cluster 2 -** This cluster is the same as cluster 2 of k means.
- **Cluster 3 -** This cluster is similar to cluster 1 of k means but it contains more observations than in cluster 1 in k means.

For Basis of Purchase:

Similar to the Purchase Behavior segmentation with Agnes, we used the **Dissimilarity measure = "Pearson" with Linkage method = "Ward"** as they gave us the best results.

We then find the best cut using the **elbow-method** and **silhouette method**.





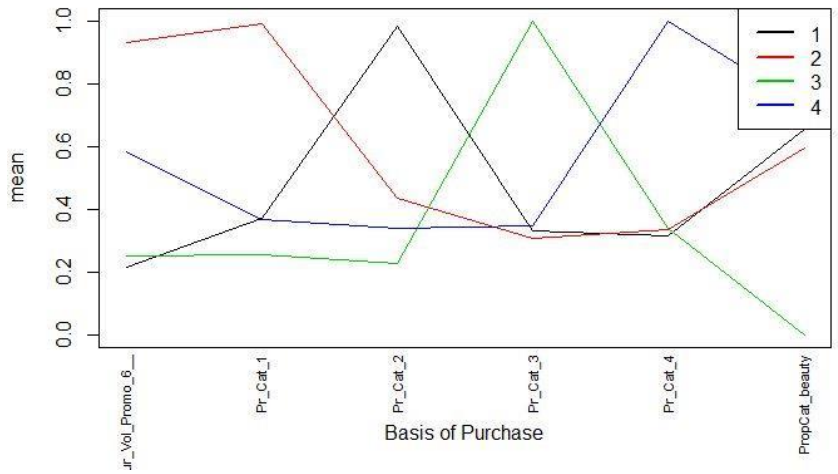
Davies-Bouldin's index

k	DB Index	Size of cluster
2	1.88578	210, 390
3	1.64032	210, 213, 177
4	1.3172	210,213,99,78
5	1.21907	210,105, 99, 78,108

Here, we can see that Davies-Bouldien indexes for 2 and 3 are relatively too higher than 4 and 5. Also the cluster size is not acceptable in 2.

Thus, due to **DB index and cluster size trade-off**, cut=3 and cut=4 seem better clustering. However, 4 clusters did a better job in segmenting the clusters in each price category.

It's silhouette width is also **0.3** which is quite decent.



Here, we can see that there are 4 peaks at each of the price categories. That means cut=4 has segregated the customers in such a way that they purchase any of these 4 categories.

	clus_aeg	EDU	Affluence_Index	AGE	Pur_Vol_Promo_6__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
1	1	4.257143	17.03810	3.242857	0.01662183	0.14132623	0.8000498	0.03992846	0.01869550	0.7401047
2	2	4.629108	21.22066	3.253521	0.10244714	0.56603981	0.3837042	0.01723598	0.03301999	0.6912260
3	3	2.757576	10.32323	3.030303	0.02082671	0.06359245	0.2228047	0.67836528	0.03523755	0.2046237
4	4	3.500000	14.00000	3.256410	0.06059427	0.13950460	0.3088187	0.05523182	0.49644485	0.8037423

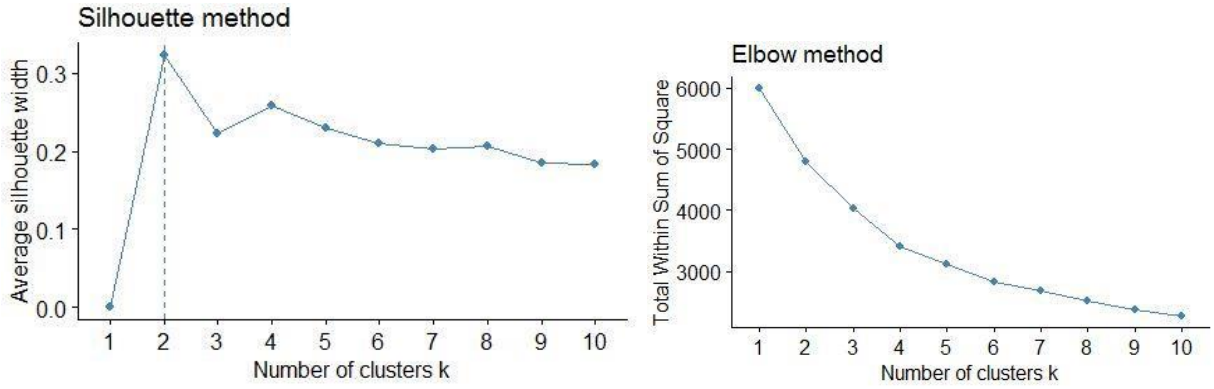
Interpretation of the clusters are:

- **Cluster 1 -** This cluster is similar to cluster 1 of k means.
- **Cluster 2 -** This cluster is the same as cluster 4 of k means. However, it also contains some observations which should be in cluster 1. That means, k means did a better job here.
- **Cluster 3 -** This cluster is similar to cluster 2 of k means.
- **Cluster 4 -** This cluster is the same as cluster 3 but with more observations which should be in cluster 1 according to business acumen.

For clustering on both purchase behavior and the basis of purchase.

We again chose **Dissimilarity measure = "Pearson" and Linkage = "Ward"** for the analysis.

We then find best k using **elbow-method** and **silhouette** method:



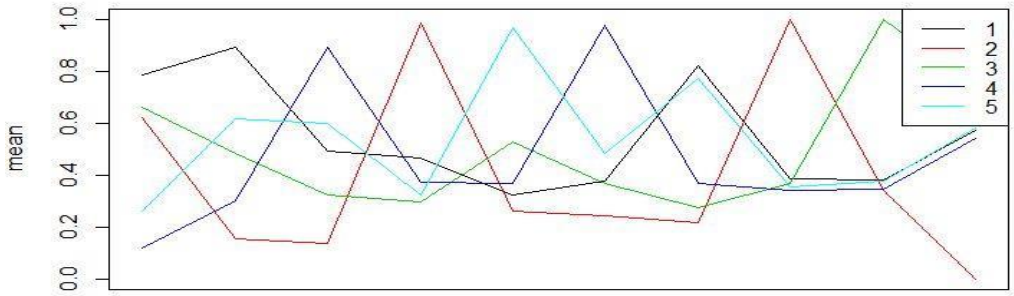
Optimum number of clusters using the elbow method is **4** or **5** and using the silhouette method is **2.** To further check other measures, we then tested on custers - 2, 3, 4 and 5.

Davies-Bouldin's index

# of clusters	DB Index	Size of cluster
2	1.07336	534, 66
3	1.55291	408, 66, 126
4	1.42931	352, 66, 56, 126
5	1.47235	300, 66, 56, 126, 52

Here, we can see that Davies-Bouldien index for 2 is relatively lower than others however, almost all the observations fall in 1 cluster only which is wrong.

Therefore, due to **DB index, cluster size trade-off, elbow-method, silhouette method**, cut =4 and 5 makes more sense. As seen in the above table, one of the bigger clusters of 408 observations in cut=3 is further divided into 352 and 66 in cut=4 and then again the bigger cluster is divided into 300 and 52 in cut=5.



Purchase behavior and basis

For **K=5**, one cluster is segregated whose customers do a large number of transactions. This would be helpful in making strategies further.

	clus_aeg	EDU	Affluence_Index	Total_Volume	No_of_Trans	Avg_Price	brnd_lylty	Pur_Vol_Promo_6_	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
1	1	4.346667	18.063333	14019.233	35.59667	11.331147	-0.08298621	0.027302913	0.17262442	0.7059771	0.07381989	0.04757860	0.6943897
2	2	2.287879	8.106061	12390.000	23.33333	6.820472	4.47520080	0.005235692	0.05014205	0.1155608	0.82373185	0.01056531	0.1235145
3	3	3.428571	14.982143	12774.554	28.76786	9.201830	-1.55817200	0.096911600	0.16282485	0.1696422	0.05384359	0.61368940	0.8710802
4	4	4.285714	18.966254	7508.135	25.80159	16.459459	-0.89922573	0.042113871	0.70899355	0.2604690	0.01761917	0.01291831	0.6641512
5	5	4.596154	19.788462	8922.154	30.98077	12.731799	-1.34437142	0.246734510	0.26680615	0.6566499	0.03413384	0.04241011	0.7051993

Interpretation of the clusters are:

Please refer to the Kmeans interpretation because the interpretation in this case is similar to that of KMeans for Both purchase behavior and basis of purchase. You can see the segmentation into the cluster is the same. The cluster sequence has changed. However, this does not affect the business interpretation and is also still the same.

DBscan Algorithm:

For this task, We had two **parameters** to play with **eps**(size of the epsilon neighborhood) and **minPTs**(number of minimum points in the eps region (for core points). Default is 5 points)

For Purchase Behavior:

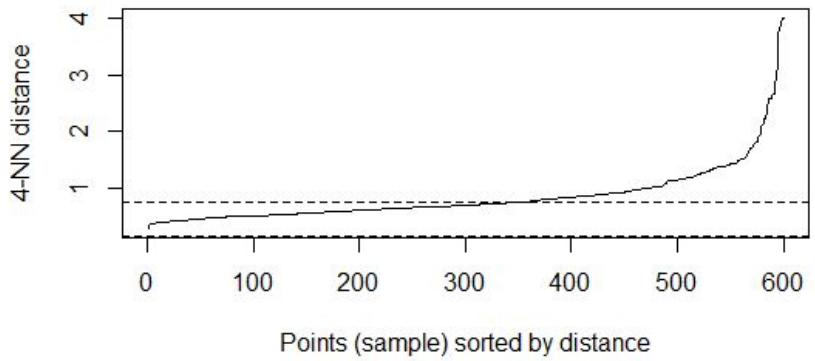
We found below results with trying out different values for parameters:

eps	minPTs	size of clusters
0.8	5	149, 427, 18, 6
0.9	5	115, 463, 22
0.75	5	174, 403, 8, 5, 5, 5
0.8	6	166, 410, 8, 6, 6, 4
0.9	6	121, 452, 14, 7, 6
0.75	6	204, 388, 8

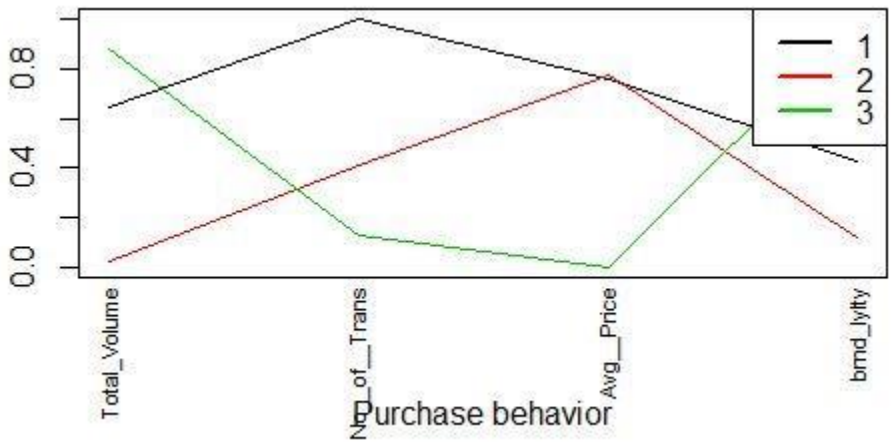
Above table shows how sensitive DBscan is to the changes in parameters.

This is evident from the above table that with **eps = 0.75 and minPTs = 6** we get a decent cluster size. If we aggregate cluster 3 with cluster 1, we would be able to derive better business interpretation. Comparing the same with other parameter values, either the size of the cluster is too big or most data points get classified into one cluster.

We find the best optimal eps value using **knndisplot(**the average of the distances of every point to its k nearest neighbors. The value of k will be specified by the user and corresponds to *MinPts).* A knee corresponds to a threshold where a sharp change occurs along the k-distance curve.



DBscan created here different clusters altogether from k means and agnes as shown below:



Here, DBscan is not able to segregate the customers who keep on switching brands more frequently.

Filter

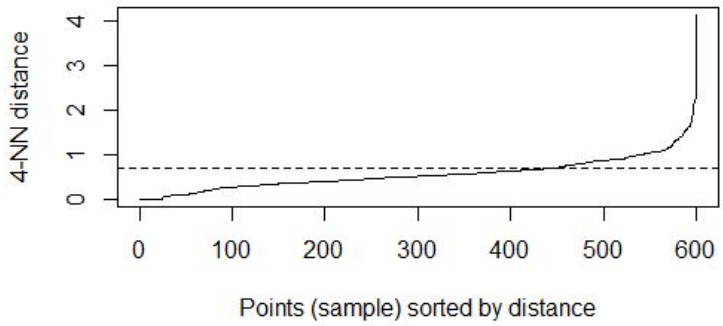
	clus_dbs	EDU	Affluence_Index	Total_Volume	No__of__Trans	Avg__Price	brnd_lylty
1	1	3.595349	16.13953	15589.023	36.24651	11.82982	1.1574061
2	2	4.326912	17.69761	9693.825	28.38727	11.95458	-0.7613091
3	3	2.625000	8.75000	17831.250	24.62500	6.30714	4.7714013

For Basis of Purchase:

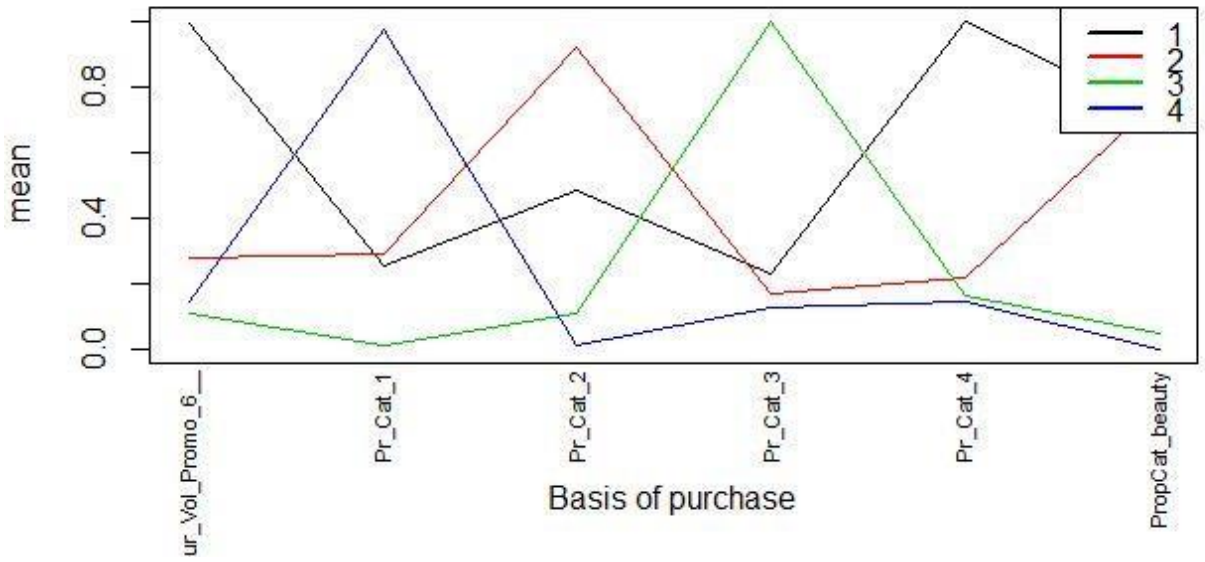
eps	minPTs	size of clusters
0.8	5	81, 488, 18, 5, 8
0.75	5	94, 481, 18, 7
0.69	5	106, 470, 12, 6, 6
0.8	8	110, 473, 17
0.75	8	118, 469, 13
0.69	8	155, 370, 63, 12

This is evident from the above table that with **eps = 0.69 and minPTs = 8** we get a decent cluster size but we figured out after observing the clusters that some customers from cluster 2 should be in cluster 1 and cluster 4.

We found the best optimal eps value using **knndisplot.**



The clusters are shown below:



Here, in cluster 1, some of the customers are wrongly placed who should be in cluster 2. Otherwise, business interpretations of the DBscan algorithm are somewhat similar to k-means and agglomerative algorithms.

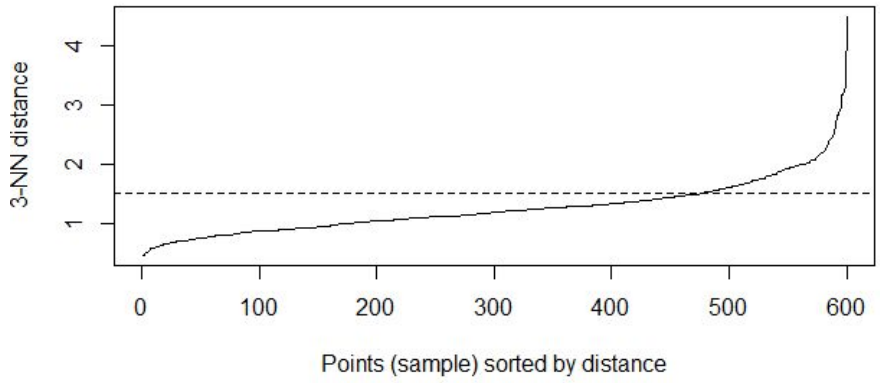
	clus_dbs	Pur_Vol_Promo_6__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
1	1	0.136419786	0.2689969	0.35070881	0.1003676608	0.279926599	0.68625091
2	2	0.028819660	0.3022942	0.63153242	0.0411968003	0.024976551	0.73332790
3	3	0.002909535	0.0396249	0.11515570	0.8368057921	0.008413610	0.11199464
4	4	0.009031620	0.9485202	0.05027615	0.0003264773	0.000877193	0.07057467

For Both purchase behavior and basis of Purchase:

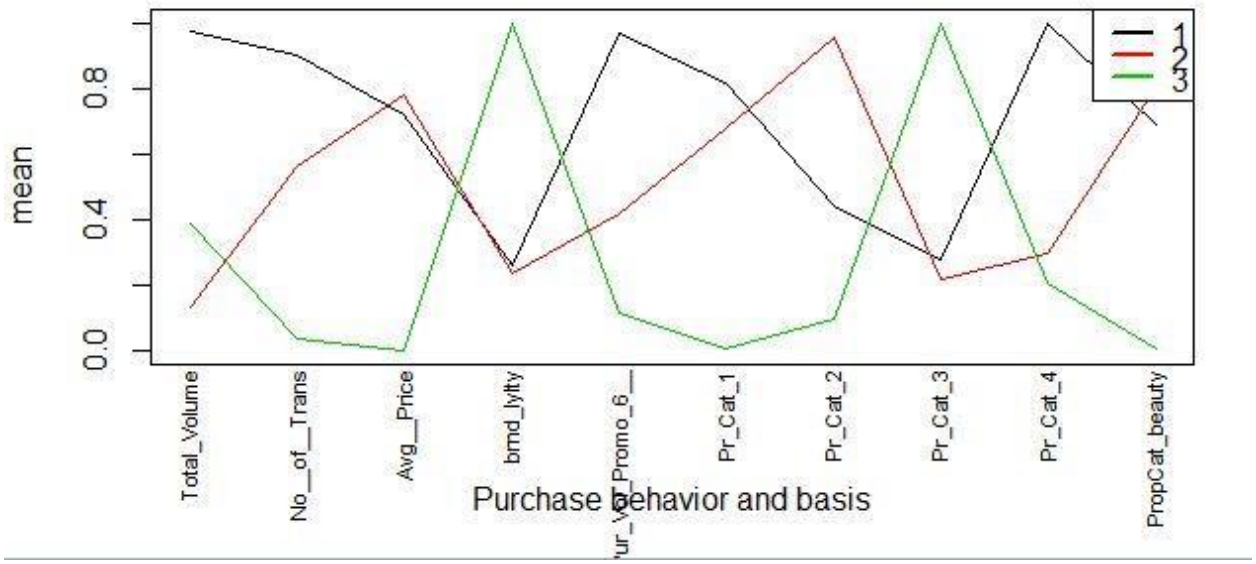
eps	minPTs	size of clusters
1.2	15	422, 29, 127, 22
2	15	42, 558
1.5	15	199, 346, 55
1.2	11	352, 206, 42
2	11	32, 568
1.5	11	162, 377, 61

Here, the best clustering happened when we chose **eps= 1.5 and minPTs = 15.**

We found the best optimal eps value using **knndisplot.**



The clusters are shown below:



The results here are quite absurd as the 2nd cluster here contains the observations which should be in cluster 1 and cluster 2. And they are not able to segregate less loyal customers with neutral customers.

clus_dbs	EDU	Affluence_	Total_Volume	No__of__Trans	Avg__Price	brnd_lylty
1	3.962963	17.265432	14901.48	35.66049	12.077016	-0.3551257
2	4.381963	18.514589	10629.4	30.5756	12.541185	-0.4941847
3	2.163934	7.131148	11926.89	22.7541	6.823596	3.9973444

Pur_Vol_Promo_6_	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_beauty
0.104000941	0.33888455	0.32035	0.10504864	0.23571682	0.6380914
0.039712283	0.2902095	0.6287105	0.04318324	0.0378968	0.7282713
0.004595709	0.05104464	0.1141712	0.82335287	0.01143131	0.1257982

Important: The results provided by DBscan are not that reliable because it is very sensitive to clustering parameters eps and minPTs.

We are thus not considering the results from DBscan as the clustering was far better in k means and hierarchical clustering.

4a) Clusters obtained from k-means and hierarchical clustering are almost similar in all the segments. However, DBscan has different clusters. For example, comparing them in combination of both the segments.

K-means vs Agglomerative:

		Agnes				
		1	2	3	4	5
Kmeans	1	0	113	7	4	33
	2	0	0	3	50	3
	3	0	0	66	0	1
	4	0	0	0	1	118
	5	103	46	2	4	46

As shown in the confusion matrix between k-means clusters and agglomerative clusters. Majority of the clusters are the same however in different order.

For example:

- Cluster 1 of agnes is cluster 5 of k-means
- Cluster 2 of agnes is cluster 1 of k-means with some mismatch.
- Cluster 3 of agnes is almost similar to k-means 3rd cluster.
- Cluster 4 of agnes is cluster 2 of k-means
- Cluster 5 of agnes is almost similar to cluster 4 of k means,

Conclusion: 80% of the observations are clustered in similar way in both k-means and agglomerative clustering.

K-means and DBscan:

		Dbscan	
		1	2
Kmeans	1	8	149
	2	6	50
	3	2	65
	4	14	105
	5	2	199

It is evident that DBscan clusters are way different than k-means or agglomerative clustering. Most of the observations have gone into cluster 2 of Dbscan. Hence, we are discarding DBscan due its low reliability.

4b) Best segmentation if we consider brand loyalty as the measure is definitely Purchase behaviour as shown in above questions. However, the best segmentation, if we consider the type of product being purchased, is basis-for-purchase.

However, when we clustered the customers according to both the segments, we got the same interpretations surprisingly which is in fact cost-effective.

That means while measuring purchase behaviour and basis for purchase of different thresholds, we got the same information while modeling both the segments together as we were getting when we model the respective segments separately.

Therefore, the **best segment** according to us is a **combination of both** modeled together.

According to this segmentation, we got 5 clusters as already explained in the above questions.

Here, we summarize those clusters once again with proposed solutions to guide the development of promotion and advertising campaigns.

Cluster order might be different in k-means and agglomerative clustering but the clusters made are similar. For the sake of simplicity, we are taking k-means to explain here.

Cluster 1 - Major buyers of “Economy/Carbolic” soaps. They are less-educated customers with low affluence index. Average price per purchase is low. Most loyal customers.

Proposed Solution - As they tend to spend less and prefer economy soaps, so the new products which are economical should be advertised to them. As they are loyal, they would help in generating a decent revenue as they do not change their brands frequently and they buy soaps in bulk.

Cluster 2 - Major buyers of “Premium soaps” which are costlier. They are highly educated with relatively higher affluence index. They keep on switching the brands.

Proposed Solution - They should be targeted for the sale of high quality, costlier soaps. Advertising strategy would be different here. They need to feel that the product or service is a physical manifestation of luxury rather than a normal one.

Cluster 3 - Major buyers of “Sub-popular soaps”. They are basically related to beauty segments. The main thing is that these customers are attracted to promo code 6 which means banded-soaps.

Proposed Solution - If a new company wants to sell new products which are not popular yet, these are the customers to target. Simple thing the firm has to do is sell the products in banded offers. For example, a pack of 4 soaps at a price of 3.

Cluster 4 - Major buyers of “Popular soaps”. They usually buy popular products that are either shown on television or newspaper. As the number of customers are highest in this segment, their number of transactions and volume purchased in a month is the highest.

Proposed Solution - They are the major proportion of customers which every company wants to target. They are the major source of revenue generation. As they are least loyal towards their preferred brands, a proper advertising strategy is important here which makes the product popular so that they would tend to purchase that product.

Cluster 5 - These customers are similar to cluster 4 customers in every trait except their purchase in a month is less than the cluster 4 segments. We could have merged them in cluster 4 too but the promotional campaign should be different for this cluster, hence keeping it separate.

Proposed Solution - They are around 10% in number who buy popular brands. However, to increase the sale among these customers, proper promotions and discounts should be offered to these customers.

4c) Best segmentation is K-means.

We can make a decision tree and check the accuracy of the models in all 3 algorithms to find out which is the best technique in clustering.

Here, we split the data in train and test in 70:30 proportion and check the accuracy of test data which would give us an estimate of how good the technique is.

K Means

Confusion Matrix of test data:

Testing					
	true				
pred	1	2	3	4	5
1	26	0	0	0	2
2	1	17	0	0	0
3	1	0	21	0	0
4	5	0	0	25	0
5	16	1	1	5	59

Accuracy :0.8222222

Agglomerative:

Confusion Matrix of test data:

Testing					
	TRUE				
pred	1	2	3	4	5
1	29	9	0	1	13
2	4	28	3	0	8
3	0	1	22	0	0
4	1	0	0	15	0
5	2	3	0	0	41

Accuracy: 0.75

DBscan:

As we know that DBscan's clusters were absurd because there are 568 observations in 1 cluster which is of no use. Hence, there is no point checking accuracy for this algorithm.

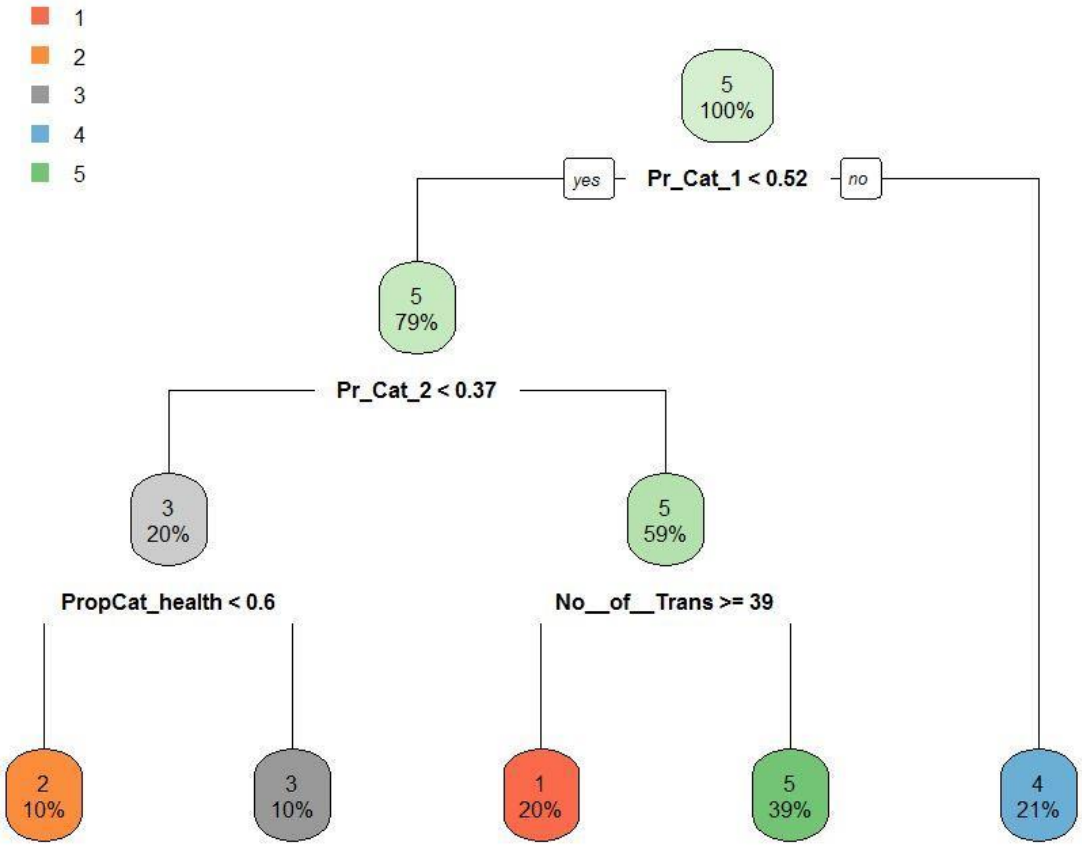
Dbscan clusters	Count
1	32
2	568

From checking test accuracy, we can conclude that k-means is the best technique among all clustering techniques which we explored.

Important Variables came out as:

Important Variables	Importance
Avg__Price	104.555061
Pr_Cat_1	87.457732
No__of__Trans	66.335197
Pr_Cat_3	61.43791
Pr_Cat_2	60.44187
Pr_Cat_4	52.604935
PropCat_health	47.08315
PropCat_beauty	44.408004
brnd_lylty	42.880265
Total_Volume	8.822028
FEH_2	3.742271
Affluence_Index	2.598781

We can also confirm the interpretations of our clusters by building a decision tree on our best segmentation and k-means algorithm.



Interpretations:

Cluster 1: Major buyer of “Popular soaps” with a number of transactions on the higher side.

Cluster 2: Other remaining clusters which should be major buyers of “sub-popular soaps”.

Cluster 3: Major buyer of “Economy/Carbolic” soaps.

Cluster 4: Major buyer of “Premium Soaps”

Cluster 5: Major buyer of “Popular soaps” with a smaller number of transactions in a month.

These interpretations are almost similar in 4 clusters which means the 5th cluster is by default interpreted as similar to our earlier interpretations too.

This is because the important variables which come out of decision trees are either the same or having similar correlation as what we observed in the above line graphs in k-means sections.

However, decision trees cannot distinguish between correlated variables and they can have equally higher importance which is not the case when we make the clusters using business acumen.