IDS 572 - MARKET SEGMENTATION – CRISA CLUSTERING CASE STUDY

IDS 572 – Assignment 3 on Bath Soap Data

Abstract

Analyzing the bath soap data for CRISA to form various clusters

Submitted By:
Archana Singh - 668528470
Nikita Bawane - 661069000
Ritu Gangwal - 670646774

IDS 572 - MARKET SEGMENTATION – CRISA CLUSTERING CASE STUDY

Question 1: What is the business goal of clustering in this case study?

CRISA is a market research agency that follows consumer purchase behavior in consumer goods. In this project about market segmentation, CRISA tracked about 30 product categories and within each category, about 60 – 70 brands. CRISA has traditionally segmented markets based on purchaser demographics. But now they want to create segmentation based on Purchase Behavior, Basis of purchase and finally combining both. This way they can target at the customers who are more loyal to the brand and hence help in better marketing of the products.

Clustering is the process of grouping the observations of similar kinds into smaller groups within the larger population. Cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in such a way that the degree of association between the two objects is maximum if the objects belong to the same group and minimum otherwise.

Clustering can also be used for exploratory purposes - it may be useful just to get a picture of typical customer characteristics at varying levels of your outcome variable.

The objective of this assignment is to aid CRISA, a market research agency, to segment market based on Purchase Behavior and Basis of Purchase. CRISA aims at segmenting the market based on two key sets of variables namely Purchase Behavior and Basis of Purchase, to gain information about what demographic attributes are associated with different Purchase Behaviors and degrees of Brand Loyalty.

In this case study of CRISA, we look first at the clusters based on purchase behavior, then clusters based on the basis for purchase, then clusters based on both – purchase behavior and basis of purchase. The complexity of marketing to 5 segments would probably not be supported by clustering just based on purchase behavior, or clustering just based on basis for purchase, so we will look at 2-3 clusters for those variables, and more when we cluster using both sets of variables. In choosing the value of k, we would seek a k value that produces clusters that are distinct and separate from each another. The variables we have been asked to consider are those that relate to purchase behavior (volume and frequency of purchase, brand loyalty), and a separate set that relate to the basis for purchase (response to promotions, pricing, and selling proposition). Finally, we look at predictive models that classify customers into segments based on demographic data.

To help CRISA deploy their promotion budgets more effectively, we will create clusters using multiple clustering methods. These clusters will enable CRISA's clients to design more cost-effective promotions targeted at appropriate segments, hence reaching their business objective.

Data Exploration:

- The cluserdata.xsl contains data pertaining to the Bath Soaps product category. It has 46 variables captured for 600 households.
- The dataset is clean with no missing values, so no imputations were made.

Dummy Coding:

- The BathSoap dataset contains many nominal variables. These are the variables that describe a characteristic using two or more categories.
- These variables are very common in quantitative research but are not always useable in their original form.
- The common workaround for using these variables is dummy coding. It allows us to turn categories into something that can be treated as having a high (1) or low (0) score.
- We have transformed the below mentioned variables to dummy variables for better cluster interpretation.

New Variables formed:
After looking at the data, we felt the need to transform some of the fields. The following transformations are done by us:

1) Food Eating Habits (FEH):
We have categorized Pure Vegetarian as1, Vegetarian but serving eggs as 2 and Non-Vegetarians as 3 in binary forms.

Codes 1,2,3	FEH_1, FEH_2, FEH_3
Code 0	Deleted as not specified

2) Mother Tongue:

We have dropped the category “0” as it is not specified. Amongst the remaining categories we have retained the MT_4, MT_5, MT_10 and MT_17 categories under their original category names because they have a relatively high frequency as compared to the other languages. All other languages except the ones mentioned above have been retained under the MT_0 dummy column. The new dummy columns will each show whether each category was a household’s Mother Tongue.

3) CHILD:
We have dropped the category “0” as it is not specified. Every other category has been transformed to a new variable or column. These columns will each show the presence of children for each household. The new columns are named as CHILD_1, CHILD_2, CHILD_3 and CHILD_4.

4) Socio Economic Class (SEC):

We have retained all the categories of SEC and have transformed them to a respective new variable or column each. These columns will each show the whether each category was a household's Social Economic Class. The new columns are named as SEC_1, SEC_2, SEC_3 and SEC_4.

5) Education (EDU):

We have dropped the category "0" as it is not specified. In addition to this, we have modified the existing categories into a narrow range for better prediction. EDU is modified as follows:

Category	Description
1	Illiterate
2	Literate, but no formal schooling
3	Mid School (3rd and 4th category from the previous setup)
4	High School (5th category from the previous setup)
5	College (6th to 9th category from the previous setup)

These columns will each show whether each category was a household's education. The new columns are named as EDU_1, EDU_2, EDU_3, EDU_4 and EDU_5.

6) AGE

We have retained all the categories of AGE and have transformed them to a respective new variable or column each. These columns will each show the whether each category was a household's age. The new columns are named as AGE_1, AGE_2, AGE_3 and AGE_4.

7) SEX

We have dropped the category "0" as it is not specified. Every other category has been transformed to a new variable or column. These columns will each show whether each category was a household's sex. The new columns are named as SEX_1 and SEX_2.

8) Brand Loyalty:

We have created the maxBr attribute to explain Brand Loyalty. It is calculated as the maximum of the purchase by major brands including Br. Cd. 57, Br. Cd. 144, Br. Cd. 55, Br. Cd. 272, Br. Cd. 286, Br. Cd. 24, Br. Cd. 481, Br. Cd. 352 and Br. Cd. 5. We have not included Others_999 in this calculation as it gives the share of transactions towards other brands which indicates that a customer is not brand loyal.

After making all the modifications, the BathSoap dataset now consists of **67 variables**.

Question 2: 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)].**

Solution 2a:

To evaluate the Purchase Behavior (including Brand Loyalty) of a consumer, we have used the below listed variables. The table also includes the description for each variable taken into consideration.

Variable Name	Description
Brand Runs	Provides number of instances of consecutive purchase of brands which would help us to determine the frequency of purchases by a consumer for a specific brand
Avg. Price	Average price of a purchase allows us to understand the purchasing capacity of a consumer
No. of Trans	Number of purchase transactions, intuitively allows us to understand that if number of transactions for a brand is high, then the brand loyalty of the consumer for the brand is high.
No. of Brands	Number of brands purchased, indicates the loyalty of consumer towards specific brands.
Max Brand Purchase	The maximum purchases in a brand (other than other 999) indicates predilection of a household to a given brand.
Others 999	Indicates lack of loyalty since this is a group of households with no brand specificity.
Value & Total Volume	The purchase behavior can be understood from the total volume and the associated value of the transaction made by a consumer.
Trans Brand Runs	It indicates the average transaction per brand run which helps us to determine the average transaction for a single purchase of a brand
Vol Tran	This indicates the average volume per transaction.

Taking into consideration the business requirements of CRISA, we have planned to create models with k values through 2 to 4, thus specifying the number of clusters. We have created several k means clustering models by varying the below mentioned **input parameters**:

(a) **centers:** The number of clusters we wish to extract
(b) **nstart:** It attempts multiple initial configurations and reports the best one
(c) **iter.max:** It is the number of times the algorithm will repeat the cluster assignment and moving of centroids.

In addition to this, the parameters which will help us **evaluate the performance** of the clustering models are listed below:

a. **betweenss:** It is the between clusters sum of squares. In fact, it is the mean of distances between cluster centers. As we aim to have heterogeneity between different clusters, this value should be as high as possible

b. **withinss:** It is the within cluster sum of squares. So, it results in a vector with a number for each cluster. As we aim to have homogeneity within the clusters, this value should be as low as possible.

The table below depicts different clustering models with different combinations of the input parameters.

	Input Parameters		**Output Parameter**					
Model No.	**K value**	**nstart value**	**No. of Iteration**	**tot.withinss**	**betweenss**	**DB Index**	**Dunn Index**	**No. of Data Points in Each Cluster**
Model 1	2	25, 100, 500	1	4754.01	1235.99	1.9611	0.03938785	Cluster 1: 317 Cluster 2: 283 Total No. of Data Points: 600
Model 2	3	25, 100	4	3969.644	2020.355	1.703715	0.02855756	Cluster 1: 259 Cluster 2: 175 Cluster 3: 166 Total No. of Data Points: 600
		500	3					
Model 3 (best model)	4	25	3	3427.9606	2562.039	1.51476	0.04914977	Cluster 1: 191 Cluster 2: 188 Cluster 3: 175 Cluster 4: 46 Total No. of Data Points: 600
		100	5					
		500	6					

We perform cluster validity using Internal Cluster Validation, wherein the clustering result is evaluated based on the data cluster itself (internal information) without reference to external information. The two Internal Cluster Validity indices that we have used are listed below.

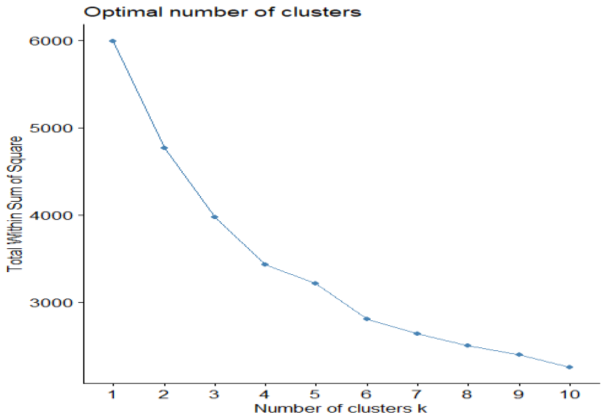
1. **Dunn Index:** It aims to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. Higher the Dunn index value, better is the clustering. The number of clusters that maximizes Dunn index is taken as the optimal number of cluster k.

2. **DB Index:** It validates the clustering using quantities and features inherent to the dataset. Lower the DB index value better is the clustering. The number of clusters that minimizes Dunn index is taken as the optimal number of clusters k.

Best K chosen:

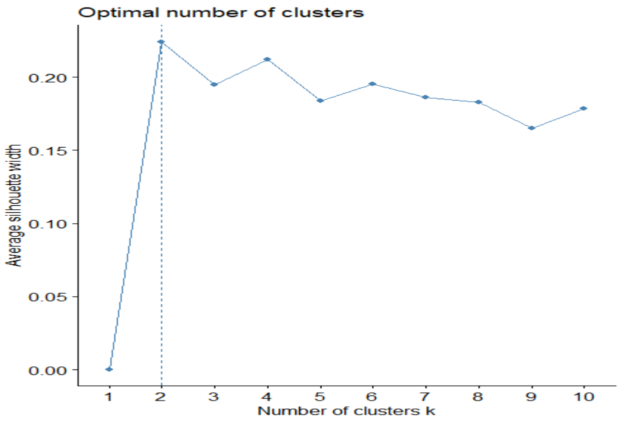
To create various clustering models, we cannot choose the number of clusters by visually inspecting the data points as this would introduce a lot of ambiguity in the process. Thus, one of the trickier tasks in clustering is identifying the appropriate number of clusters k. Therefore, we have used two different methods to choose some optimal values of k.

1. **Elbow Method:**



The Elbow method looks at the total WSS as a function of the number of clusters. The total WSS measures the compactness of the clustering and we want it to be as small as possible. Thus, we choose several clusters so that adding another cluster doesn't improve much better to the total WSS. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters. Here, K=4 is considered as best K value.

2. **Silhouette Method:**

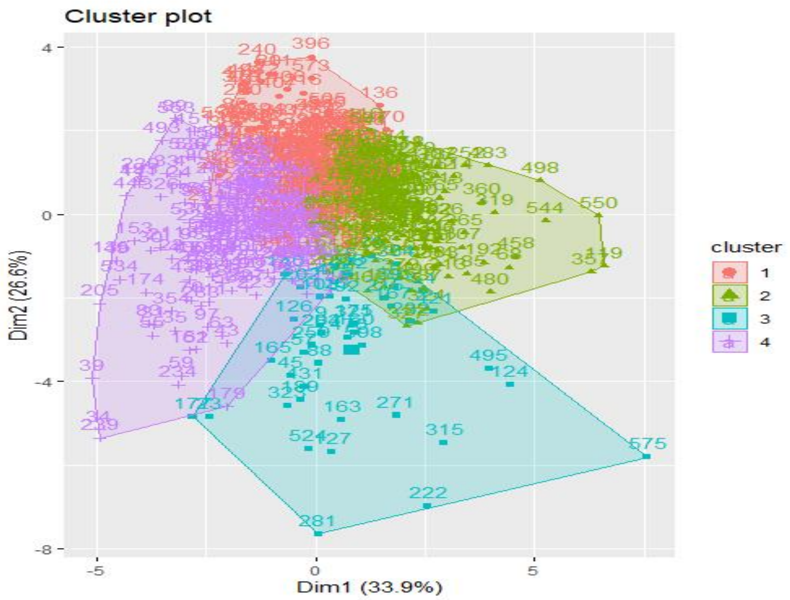


Silhouette approach measures the quality of a clustering. It determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. Below is the graph for optimal number of clusters using silhouette method. Here, best K is 2 but going by sub optimal value, we select K = 4 as best.

Best Model:

- Choosing the best value of k for the given purpose has both objective and subjective criteria. Objective criteria include consistent cluster sizes, larger inter cluster distance and smaller intra cluster distance.

- Thus, we would seek a k that produces clusters that are distinct and separate from one another, in ways (variables) that are translatable into marketing actions.



- Moreover, we have also used the Dunn index and DB index as criteria to measure the overall cluster quality wherein the combination of high Dunn index and low DB index indicate better clustering.

- All these factors together have contributed towards our decision of choosing the **Model 3 (k=4) as our best model.**

The table below summarizes the cluster description of the best model in terms of broader set of variables

clusKM	SEC_1	SEC_2	SEC_3	SEC_4	HS	SEX_1	SEX_2	EDU_1	EDU_2	EDU_3to4	EDU_5	EDU_6to9	Affluence_Index	AGE_1	AGE_2	AGE_3	AGE_4
1	0.293	0.225	0.262	0.220	3.450	0.031	0.785	0.073	0.010	0.267	0.314	0.152	15.241	0.016	0.251	0.257	0.476
2	0.271	0.314	0.234	0.181	4.596	0.048	0.926	0.048	0.016	0.239	0.399	0.266	21.303	0.027	0.181	0.298	0.495
3	0.196	0.130	0.326	0.348	6.696	0.000	1.000	0.065	0.022	0.565	0.217	0.109	18.804	0.043	0.087	0.217	0.652
4	0.194	0.240	0.234	0.331	3.909	0.034	0.806	0.131	0.017	0.269	0.251	0.154	13.891	0.029	0.246	0.309	0.417

clusKM	CHILD_1	CHILD_2	CHILD_3	CHILD_4	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	0.105	0.215	0.120	0.377	0.178	2.780	22.314	12.031	7799.241	934.670	2.084
2	0.122	0.277	0.101	0.473	0.233	5.160	46.729	26.505	13197.138	1604.908	1.849
3	0.152	0.283	0.065	0.500	0.397	3.543	36.870	16.022	29981.087	3186.105	2.843
4	0.051	0.223	0.091	0.474	0.723	2.960	22.566	8.189	10280.086	1003.580	3.967

Explanation/ Inference of Clusters:

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

Cluster 2: This group has maximum number of brands, brand runs, no. of transactions and average price of transactions. They have a significant peak for Others999 which show that they are not brand loyal. These are the people with highest affluence index.

Cluster 3: This group has the highest total volume and high number of transactions. They have a significant peak for brand loyalty but also has considerable purchases from Others999 brands. This group contains all female homemakers with considerable affluence index.

Cluster 4: They are the most brand loyal customers as seen from the table above. Their purchase in Others999 brand is the least which also adds on to prove their loyalty. They have less no. of brands, a small number of brand runs and the least number of transactions and average price of good purchased. These people have the least affluence index.

2b. 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.]

Solution:
To evaluate Basis for Purchase with given set of variables are considered

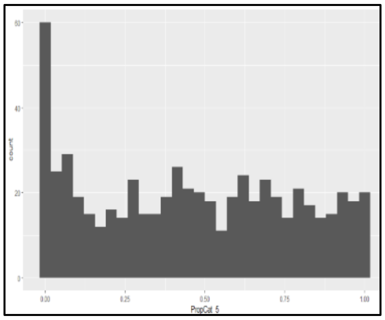
- Price Categories (1-4) - Provides us categories of price
- Pur Vol No Promo - Provides us percent of volume purchased under no-promotion
- Pur Vol Promo 6 - Provides us percent of volume purchased under Promotion Code 6
- Pur Vol Other Promo - Provides us percent of volume purchased with other promotions under other promotions
- Selling Proposition Categories- Provides us Percent of volume purchased under the product proposition. Only proposition categories - 5, 6, 7, 8 and 14 are considered as others have distribution tending to zeroes.

Selling Proposition Variables Analysis-

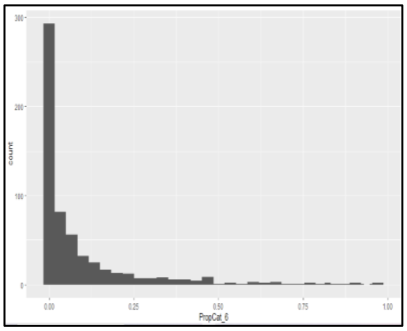
By analyzing the histograms for all selling proposition variables, it is seen that PropCat_5, PropCat_6, PropCat_7, PropCat_8 have good distribution pattern as there are data points apart from zero. While PropCat_10, PropCat_11, PropCat_12, PropCat_13, PropCat_14 and PropCat_15 mostly have zero value.

On further analyzing the quartiles for PropCat 10 to 15, it is seen that PropCat_14 has values in its third quartile, unlike the other variables which only have zero values. Hence, we retain PropCat_14.

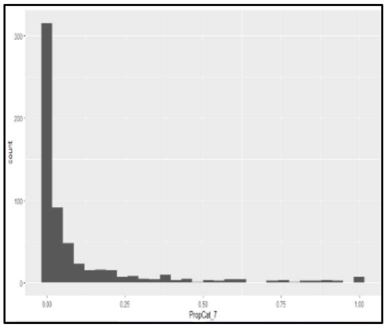
PropCat_5



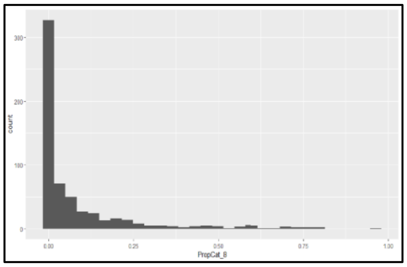
PropCat_6



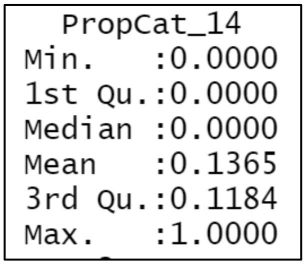
PropCat_7



PropCat 8



PropCat 14



```
   PropCat_14
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.1365
 3rd Qu.:0.1184
 Max.   :1.0000
```

Hence, the variables retained are - *PropCat_5, PropCat_6, PropCat_7, PropCat_8, PropCat_14*.

With the above variables, we have performed cluster analysis to analyze Basis of Purchase and tabulated below:

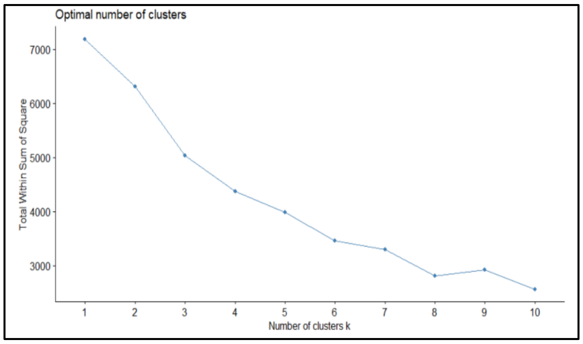
K-Means Results for different K-Values, N Start and iter.max

	Input Parameters			Output Parameters				
Model	**K Value**	**N Start**	**iter.max**	**tot.withinss**	**betweenSS**	**db index**	**dunn index**	**No. of objects in each cluster**
Model 1	2	25	10	5831.844	1356.156	1.1662	0.0438	Cluster 1 - 524 Cluster 2 - 76 Total No. of Data Points: 600
		100	10					
		500	10					
		500	25					
Model 2	3	25	10	5029.342	2158.658	2.0067	0.0394	Cluster 1 - 198 Cluster 2 - 326 Cluster 3 - 76 Total No. of Data Points: 600
		100	10					
		500	10					
		500	100					
Model 3	4	25	10	4363.87	2824.13	1.657	0.0256	Cluster 1 - 127 Cluster 2 - 75 Cluster 3 - 78 Cluster 4 - 320 Total No. of Data Points: 600
		100	10					
		500	10					
		500	100					

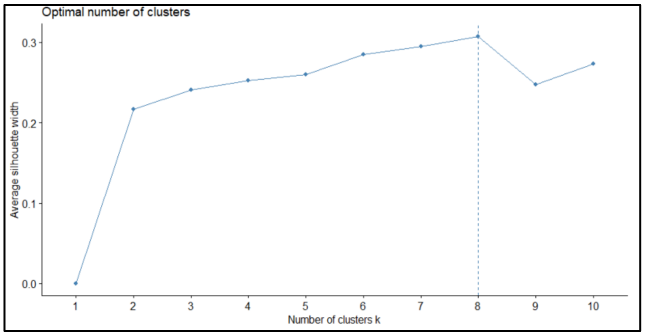
Best K chosen:

In order to determine the best value of K for the given set of variables, we can use Elbow and Silhouette method (as mentioned earlier):

Elbow Method-



Silhouette Method-

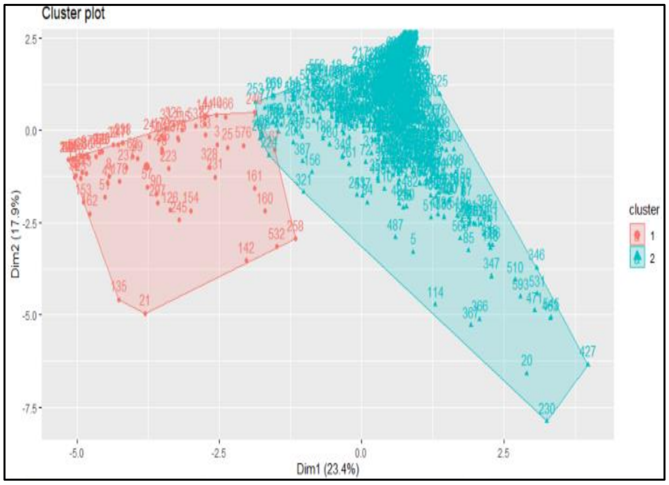


Elbow Graph – The optimal value of K is determined by bent (or elbow) of the graph, which happens at k = 2 and 6. Seeing the business considerations, we choose K = 2.

Silhouette Graph – Silhouette graph gives a sharp peak at k = 8, which is the best value of k for this variable set. K = 2 is the sub optimal value which is appropriate for this model.

Evaluation of the Best Model –

The models have been evaluated on different parameters such as DB Index, Dunn Index and analysis of between and within distances. We want a K value that produces clusters that are distinct and separate from one another.



a. Between and Within Distances – Models for K value 2 and 4 have high total within distance and low between distances, which makes the ratio of between/within distance low.

b. DB Index – An optimal cluster has low value of DB Index. Based on this, K=2 seems to be optimal K value.

c. Dunn Index – For a good cluster, Dunn Index should be high. Based on this, K=2, seems to be optimal.

Hence, we choose **K =2 as best model for Basis of purchase.**

The table below summarizes the cluster description of the best model in terms of broader set of variables:

clusKM	SEC_1	SEC_2	SEC_3	SEC_4	HS	SEX_1	SEX_2	EDU_1	EDU_2	EDU_3to4	EDU_5	EDU_6to9	Affluence_Index
1	0.28244275	0.2614504	0.2576336	0.1984733	4.188931	0.03625954	0.8625954	0.05916031	0.01526718	0.2729008	0.3416031	0.20610687	18.177481
2	0.02631579	0.1710526	0.1973684	0.6052632	4.210526	0.02631579	0.7763158	0.23684211	0.01315789	0.3421053	0.1315789	0.03947368	9.039474

AGE_1	AGE_2	AGE_3	AGE_4	CHILD_1	CHILD_2	CHILD_3	CHILD_4	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
0.02290076	0.2041985	0.2824427	0.490458	0.10114504	0.2480916	0.10687023	0.4427481	0.31143745	0.5449726	0.04566209	0.09792783
0.03947368	0.2894737	0.2763158	0.3947368	0.07894737	0.1973684	0.06578947	0.4605263	0.05564182	0.1357805	0.784152	0.02442566

PropCat_5	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__
0.5079922	0.09749477	0.10949966	0.090463148	0.0434703	0.9094443	0.05893418	0.03162151
0.106716	0.0566696	0.01010128	0.009030569	0.7777628	0.9375812	0.01603306	0.04638573

Cluster 1: Customers in this cluster are mainly those customers who have a high volume of purchases of items not on promotion. They tend to buy a higher volume of beauty products. These customers also tend to have a higher volume of purchases then the product is promoted as a value-added pack.

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

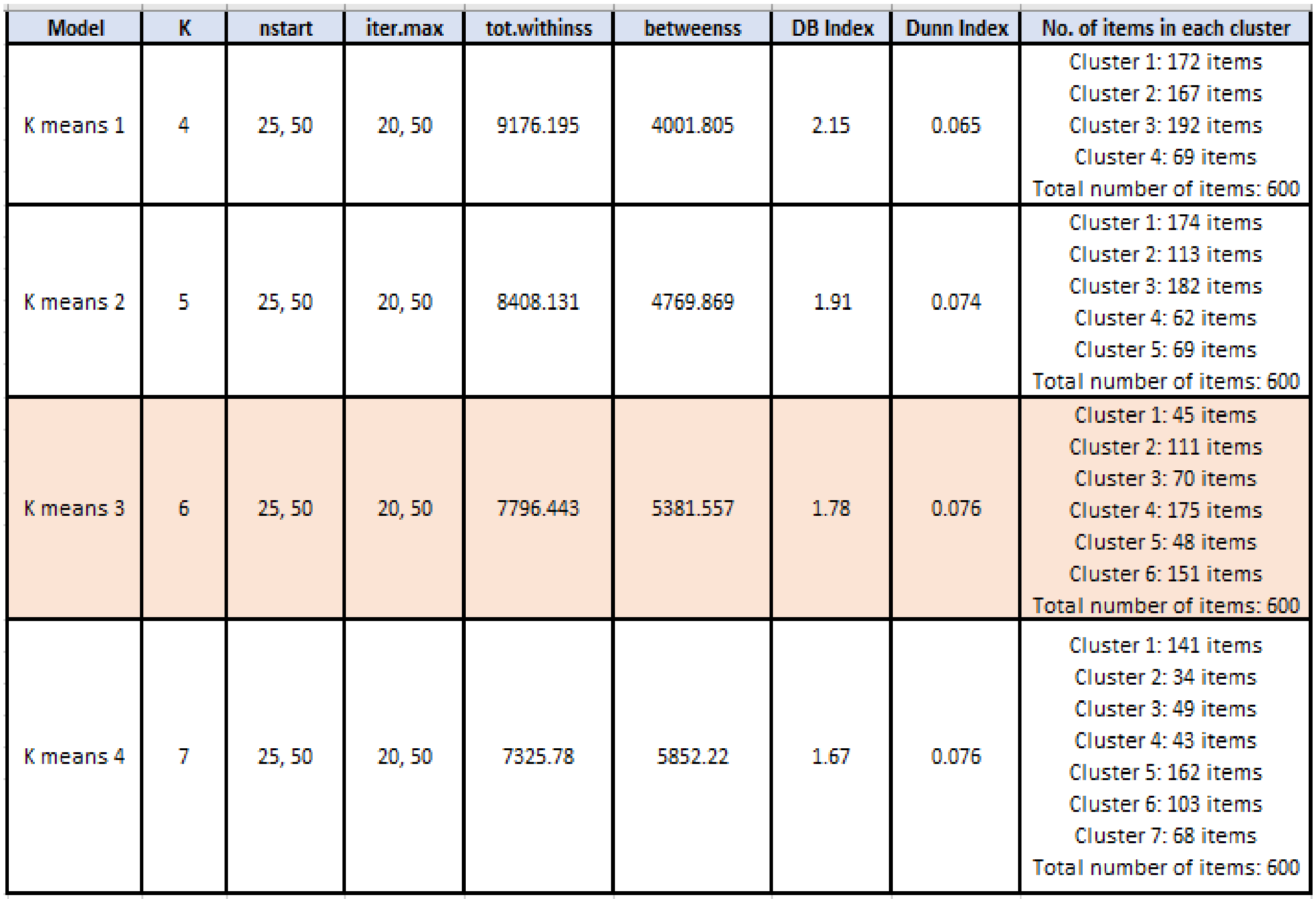
2c. The variables that describe both purchase behavior and basis of purchase.

We combined variables from Question 2(a) and 2(b) as these variables describes both purchase behavior and basis of purchase, the variables are tabulated below:

Avg. Price	Others 999	Price Categories (1-4)
Brand Runs	Total Volume	Proposition Categories (5-8,14)
No. of Trans	maxPurchase & Value	Pur Vol No Promo
No. of Brands	Pur Vol Promo 6	Pur Vol Other Promo

With the above variables, we have performed cluster analysis to analyze Basis of Purchase & Purchase Behavior. The table below depicts different clustering models with different combinations of the input parameters.

The input and output variables are same as explained in above parts.

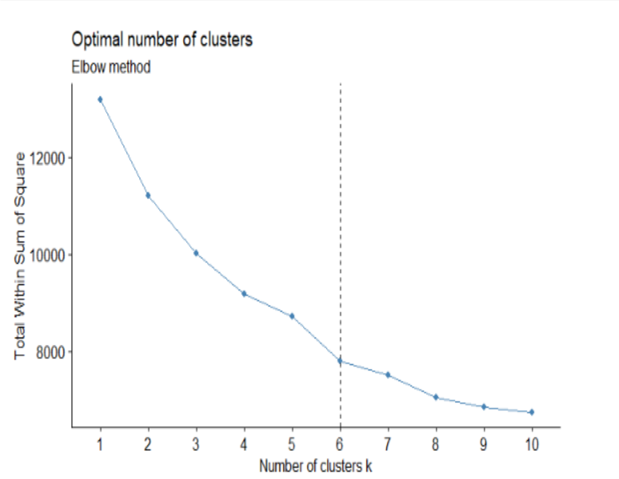


Model	K	nstart	iter.max	tot.withinss	betweenss	DB Index	Dunn Index	No. of items in each cluster
K means 1	4	25, 50	20, 50	9176.195	4001.805	2.15	0.065	Cluster 1: 172 items Cluster 2: 167 items Cluster 3: 192 items Cluster 4: 69 items Total number of items: 600
K means 2	5	25, 50	20, 50	8408.131	4769.869	1.91	0.074	Cluster 1: 174 items Cluster 2: 113 items Cluster 3: 182 items Cluster 4: 62 items Cluster 5: 69 items Total number of items: 600
K means 3	6	25, 50	20, 50	7796.443	5381.557	1.78	0.076	Cluster 1: 45 items Cluster 2: 111 items Cluster 3: 70 items Cluster 4: 175 items Cluster 5: 48 items Cluster 6: 151 items Total number of items: 600
K means 4	7	25, 50	20, 50	7325.78	5852.22	1.67	0.076	Cluster 1: 141 items Cluster 2: 34 items Cluster 3: 49 items Cluster 4: 43 items Cluster 5: 162 items Cluster 6: 103 items Cluster 7: 68 items Total number of items: 600

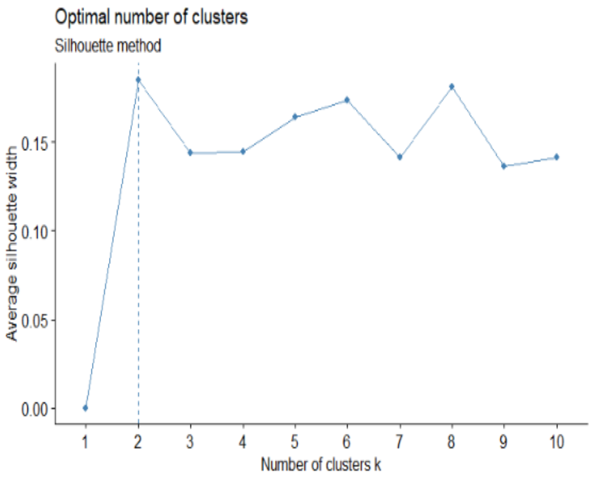
Best K chosen:

In order to determine the best value of K for the given set of variables, we can use Elbow and Silhouette method (as mentioned earlier):

Elbow Method-



Silhouette Method-

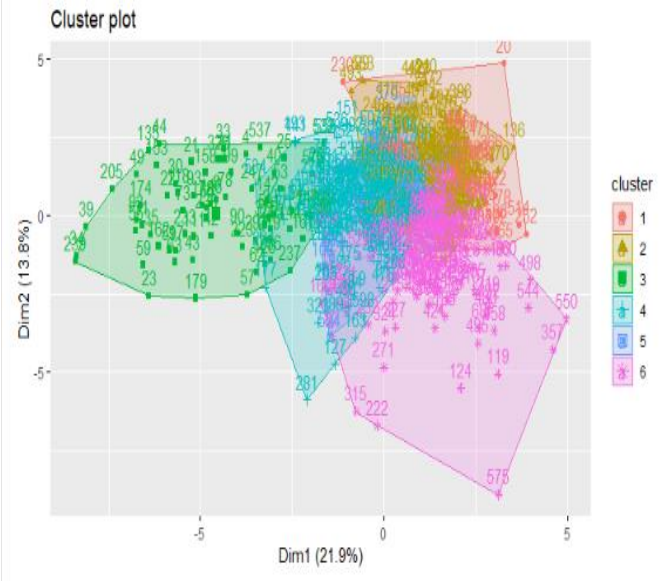


Elbow Graph – The optimal value of K is determined by bent (or elbow) of the graph, which happens at k = 6. Seeing the business considerations, we choose K = 6.

Silhouette Graph – Silhouette graph gives a sharp peak at k = 2,6 and 8 which is the best value of k for this variable set. K = 6 is the sub optimal value which is appropriate for this model.

Hence by seeing elbow and silhouette graphs, dunn index and DB index we choose K = 6 as the best model for variable set having both purchase behavior and basis for purchase.

Best Model Evaluation: Cluster Plot-



Cluster 1: One noteworthy feature about customers in this cluster is that they buy soaps majorly in exchange offer.

Cluster 2: Customers in this cluster are least brand loyal and buy products from other brands very often. They majorly buy products in extra gram mage and buy beauty soaps.

Cluster 3: Customers in this cluster have high brand loyalty and they use coupons majorly for their purchases.

Cluster 4: The characteristics of customers in this cluster are average except that they buy products in exchange offer majorly.

Cluster 5: Customers in this cluster have less brand loyalty and buy products from any brand which are being offered at price off

3. Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering. How do different parameter values for the different techniques affect the clusters obtained?
a). K-medoids:

The k-medoids algorithm is a clustering algorithm related to the k-means algorithm and the medoid shift algorithm. K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers.

A medoid of a finite dataset is a data point from this set, whose average dissimilarity to all the data points is minimal i.e. it is the most centrally located point in the set. It could be more robust to noise and outliers as compared to k-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances.

For K-medoids, we have used **Partitioning Around Medoids (**PAM) algorithm. Also, the input variables are values of K and distance types with out put parameters as DB index, Dunn index and Average Silhouette width.

Silhouette analysis: Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

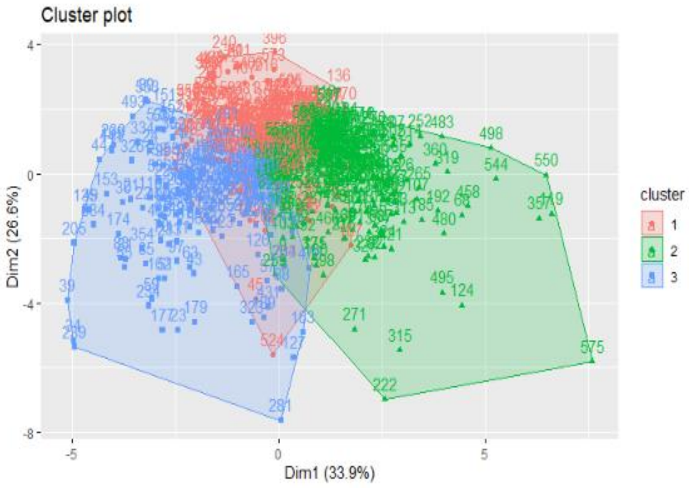
Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

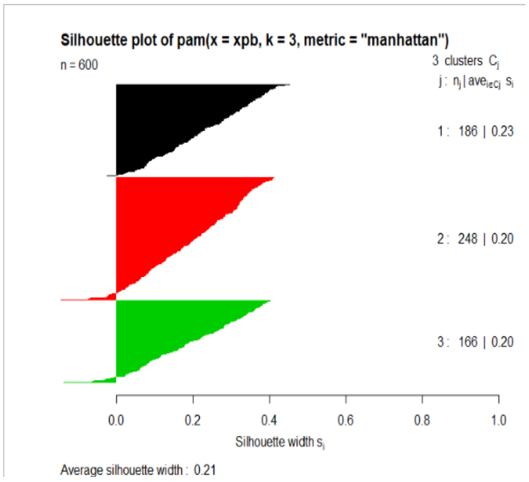
Table of various k values and distance metrices:

a) Purchase Behavior:

	Input Parameters		**Output Parameters**		
Model	**K**	**Distance**	**Avg Silhouette width**	**DB Index**	**Dunn Index**
1	2	Euclidean	0.2	2.08	0.037
2	2	Manhattan	0.24	2.01	0.026
3	3	Euclidean	0.19	1.94	0.035
4	3	Manhattan	0.21	1.9	0.031
5	4	Euclidean	0.21	1.77	0.038
6	4	Manhattan	0.17	2.41	0.033

Cluster plot and the silhouette plot for the best model in purchase behavior segment:

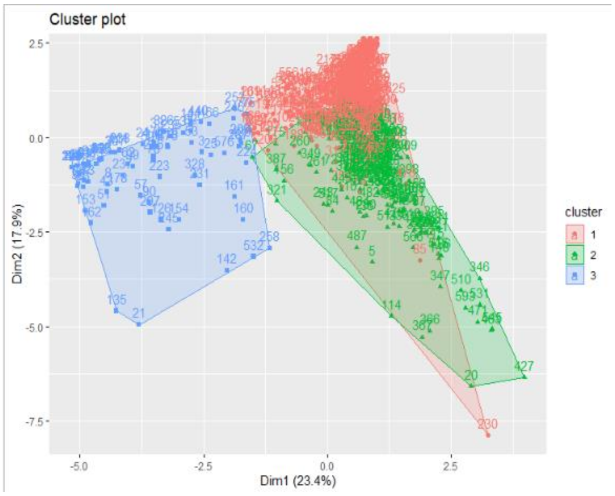


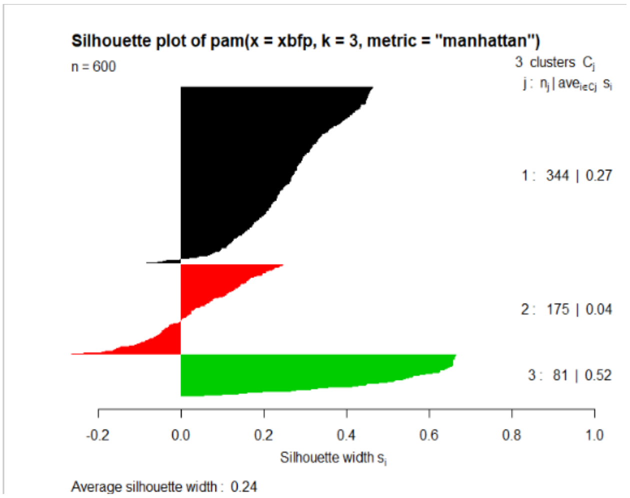


b) Basis for Purchase:

	Input Parameters		Output Parameters		
Model	**K**	**Distance**	**Avg Silhouette width**	**DB Index**	**Dunn Index**
1	2	Euclidean	0.3	1.16	0.04
2	2	Manhattan	0.33	1.17	0.036
3	3	Euclidean	0.13	2.14	0.027
4	3	Manhattan	0.24	2.07	0.035
5	4	Euclidean	0.15	2.27	0.027
6	4	Manhattan	0.22	2.37	0.027

Cluster plot and the silhouette plot for the best model in basis for purchase segment:

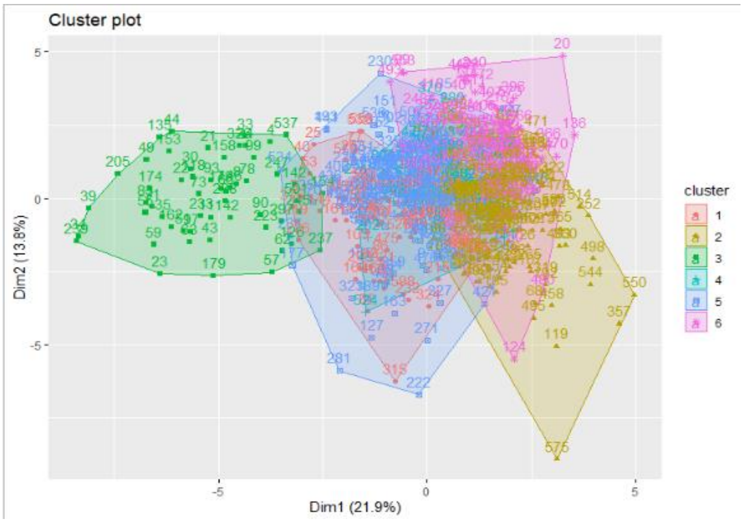




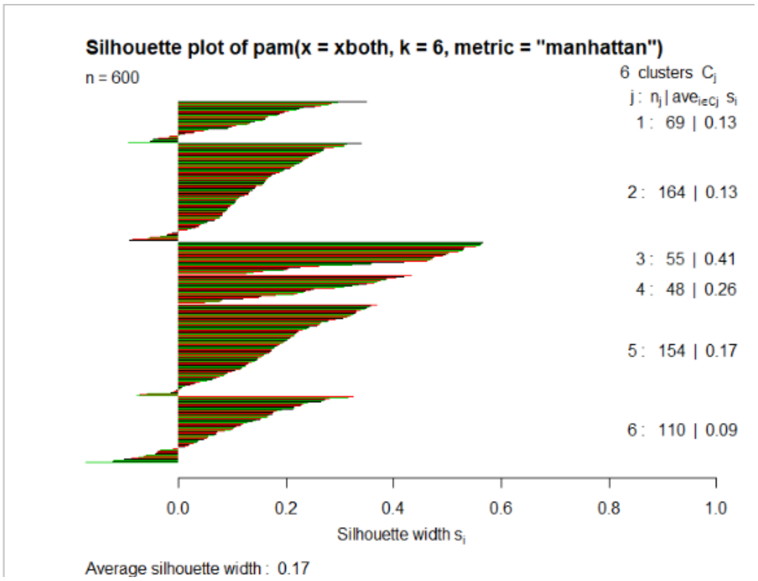
c) Purchase Behavior and Basis for Purchase:

	Input Parameters		Output Parameters		
Model	**K**	**Distance**	**Avg Silhouette width**	**DB Index**	**Dunn Index**
1	4	Euclidean	0.14	2.18	0.06
2	4	Manhattan	0.13	2.41	0.056
3	5	Euclidean	0.15	2.02	0.06
4	5	Manhattan	0.17	2.32	0.053
5	6	Euclidean	0.16	2.16	0.064
6	6	Manhattan	0.17	1.89	0.067
7	7	Euclidean	0.15	2.05	0.063
8	7	Manhattan	0.16	2.07	0.066

Cluster plot and the silhouette plot for the best model in both segments:



This cluster plot with K = 6 for both the segments is comparitively not good as the clusters overlap each other. The distance between the clusters in this case with very low.



The silhouette plot also shows a bad performance as the value of silhouette width is very low.

b). Hierarchical Clustering:

Hierarchical clustering is a type of clustering algorithm that produces 1 to n clusters, where n is the number of observations in the data set. As you go down the hierarchy from 1 cluster (contains all the data) to n clusters (each observation is its own cluster), the clusters become more and more similar (almost always). There are two types of hierarchical clustering: divisive (top-down) and agglomerative (bottom-up). We will be experimenting with Agglomerative Hierarchical Clustering on all three variable sets – Purchase Behavior, Basis for Purchase and Both combined.

Agglomerative Hierarchical clustering: In this technique, initially each observation is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed. The basic algorithm of Agglomerative is –

- Compute the dissimilarity matrix
- Consider each observation to be a cluster.
- Repeat: Merge the two closest clusters and update the matrix
- Perform the above step, until a single cluster remains

Agglomerative Clustering has been implemented through *hclust* and *agnes* functions in R. To perform hierarchical clustering, first the dissimilarity matrix is calculated by dist() function. We have experimented with distance matrix *Euclidean* and *Manhattan.* Post this, the matrix is passed to hclust() and agnes() functions with specified agglomeration method – *complete, average, single and ward.D*.
Dendrogram is used as graphical tool to get insights from the cluster solutions. For cluster analysis, dendrograms are plotted by passing the resultant cluster to plot or fviz_dend functions.

This algorithm works by grouping the data one by one based on the distance measure of all the pairwise distance between the data point. This is done through the below

- *single* - distance between two clusters is the distance between their two closest objects.
- *complete* - distance between two clusters is the distance between their two most distant objects.
- *average*- average distance or average linkage.
- *ward's method* - the distance between clusters is a weighted squared Euclidean distance between the centroids of each cluster

Cutting the Dendrogram

In a dendrogram, every observation is a leaf. Similar observations are combined to form a branch as we move up the tree and the branches merge at a higher height. The vertical axis in the dendrogram indicates the dissimilarity between the two observations. Hence, higher the height, less similar are the observations. To understand the (dis)similarity between the observations, we must see the height of the tree (vertical axis).
Height of the dendrogram controls the number of clusters formed. It is same as k in k-means. Using hclust and agnes, we can cut the tree using different values of h and k. We have experimented with values of k.

Input Parameters

Based on the above-mentioned parameters, we have experimented with the following –

1. Distance matrix – Euclidean and Manhattan

2. Linkage method – complete, average, single, ward.D. (Note that, ward.D2 method used in hclust yields the same results as Agnes ward.D)
3. Functions to implement Hierarchical clustering – hclust and agnes.
4. Cutting the tree at different heights based on the clusters needed.

Evaluation Parameters:

1. Choosing the best method in agnes- Agglomerative coefficients are used to evaluate the clustering structures. It measures the amount of clustering structure found i.e, values closer to 1 implies strong clustering structure. While experimenting with different methods for agnes, we will use the method which gives the highest Agglomerative Coefficient.
2. Choosing best clusters from agnes and hclust- As mentioned earlier, the clusters will be evaluated based on DB index, Dunn Index and cluster formation. We must determine the best cluster based on good cluster formation and good DB and Dunn index values. We will see ahead that certain methods (such as Single and Average) have good DB and Dunn index values but form bad clusters which cannot be used for market segmentation. Keeping the business goal in mind, we will select the model with sub-optimal index values and cluster formation.

Clustering Results and Interpretation for different Variable Sets

1. **PURCHASE BEHAVIOR –** Hierarchical clustering has been applied to purchase behavior variable set for different input parameters. The dendrograms formed by different methods are cut for different values of K and the results have been tabulated. The dendrogram of the most optimal cluster has been shown later.

Hclust Results - Clusters have been formed using methods such as Complete, Single, Average and Ward. To evaluate the clusters, we evaluate cluster formations, DB index and Dunn Index. Below are the observations –

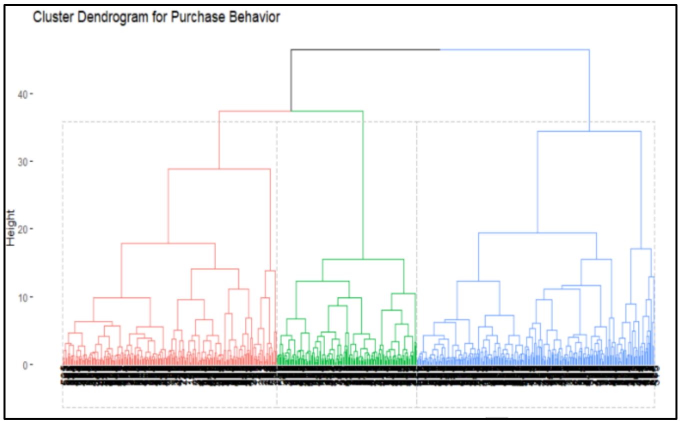
a. Single and Average Method clusters – The clusters formed by these methods have the lowest DB index and high Dunn index. But on evaluating the cluster formations, we notice that high number of observations (more than 580 out of 600) are concentrated in a single cluster and the remaining clusters have less than 10 observations, some have single observations. Such clusters, if used to implement any marketing campaign will not be cost effective or useful. Hence, we disregard these clusters.
b. Complete method clusters – We can see slightly better clusters under this method, but a similar problem persists here as well, as the clusters distribution is very high in one cluster.
c. Ward method clusters – Cluster formation is fairly good and the observations are not concentrated in a single cluster. The sum of squares is zero in the beginning as observation is its own cluster and then grows as we merge clusters. Ward's method keeps this growth as small as possible.

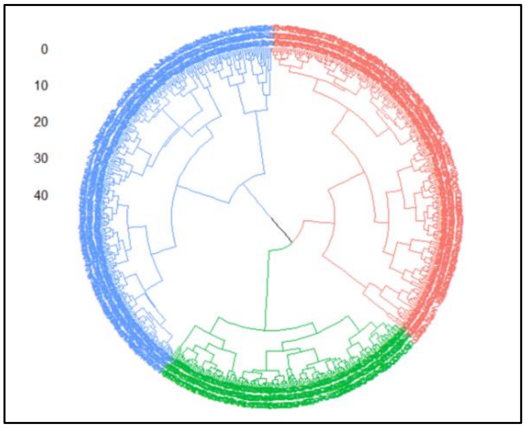
Agnes Results- In agnes, we get agglomerative coefficient which measures the strength of the clustering. This helps us to find the clustering methods that can identify strong clustering structures. Analyzing the results below, we can see that ward method has the highest coefficient. Hence, we used form method to derive clusters.

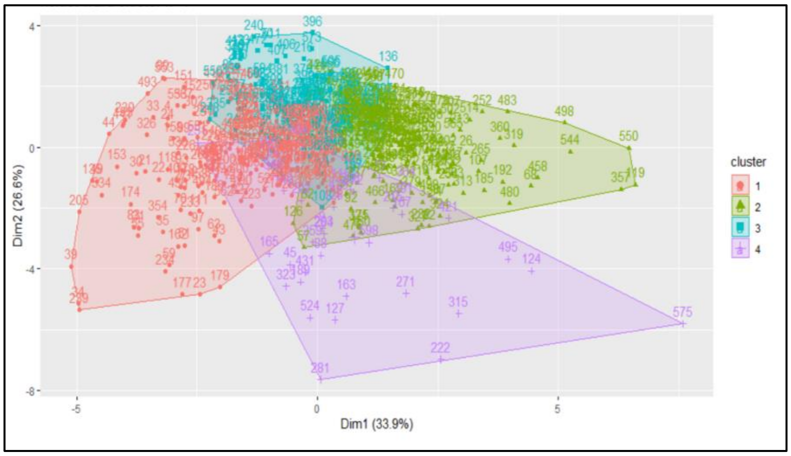
Agglomerative Coeeficient		
Method	Euclidean	Manhattan
complete	0.9250937	0.9363268
average	0.8871901	0.8752767
single	0.8198055	0.800362
ward	0.976708	0.9786299

VARIABLE SET -> PURCHASE BEHAVIOUR							
HCLUST							
Input Parameters		**Euclidean Distance**			**Manhattan Distance**		
Methods	**K Value**	**DB Index**	**Dunn Index**	**No. of Items in the clusters**	**DB Index**	**Dunn Index**	**No. of Items in the clusters**
Complete	2	0.972261469	0.087214693	Cluster 1 - 586 Cluster 2 - 14	0.937052778	0.127157916	Cluster 1 - 588 Cluster 2 - 12
	3	1.52609415	0.090231235	Cluster 1 - 533 Cluster 2 - 53 Cluster 3 - 14	1.246773579	0.050431926	Cluster 1 - 515 Cluster 2 - 73 Cluster 3 - 12
	4	1.445877592	0.101436255	Cluster 1 - 520 Cluster 2 - 53 Cluster 3 - 13 Cluster 4 - 14	1.534306775	0.065290327	Cluster 1 - 515 Cluster 2 - 46 Cluster 3 - 27 Cluster 4 - 12
Average	2	0.314581364	0.186001154	Cluster 1 - 596 Cluster 2 - 4	0.65860847	0.230125996	Cluster 1 - 595 Cluster 2 - 5
	3	0.97195803	0.191230723	Cluster 1 - 594 Cluster 2 - 4 Cluster 3 - 2	0.622098306	0.230125996	Cluster 1 - 594 Cluster 2 - 5 Cluster 3 - 1
	4	1.364899198	0.175000444	Cluster 1 - 587 Cluster 2 - 4 Cluster 3 - 7 Cluster 4 - 2	0.874772922	0.22524941	Cluster 1 - 591 Cluster 2 - 3 Cluster 3 - 5 Cluster 4 - 1
Single	2	0.225080509	0.37122809	Cluster 1 - 599 Cluster 2 - 1	0.244550606	0.344955437	Cluster 1 - 599 Cluster 2 - 1
	3	0.22591451	0.354975675	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1	0.280634965	0.354975675	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1
	4	0.243841208	0.277338345	Cluster 1 - 597 Cluster 2 - 1 Cluster 3 - 1 Cluster 4 - 1	0.32219678	0.277338345	Cluster 1 - 597 Cluster 2 - 1 Cluster 3 - 1 Cluster 4 - 1
Ward.D	2	2.573747675	0.053460523	Cluster 1 - 340 Cluster 2 - 360	2.595028469	0.047941899	Cluster 1 - 219 Cluster 2 - 381
	3	2.311028193	0.053460523	Cluster 1 - 215 Cluster 2 - 260 Cluster 3 - 125	2.251278208	0.049827967	Cluster 1 - 219 Cluster 2 - 95 Cluster 3 - 286
	4	2.00113775	0.053460523	Cluster 1 - 215 Cluster 2 - 173 Cluster 3 - 87 Cluster 4 - 125	1.927205623	0.054090154	Cluster 1 - 219 Cluster 2 - 95 Cluster 3 - 54 Cluster 4 - 232
AGNES							
Input Parameters		**Euclidean Distance**			**Manhattan Distance**		
Model	**K**	**DB Index**	**Dunn Index**	**No. of Items in each clusters**	**DB Index**	**Dunn Index**	**No. of Items in each clusters**
Ward.D	2	2.578064165	0.053460523	Cluster 1 - 359 Cluster 2 - 241	2.483464059	0.047083087	Cluster 1 - 375 Cluster 2 - 225
	3	2.231739151	0.053460523	Cluster 1 - 217 Cluster 2 - 241 Cluster 3 - 142	2.16453278	0.050812634	Cluster 1 - 253 Cluster 2 - 122 Cluster 3 - 225
	4	**1.946171598**	**0.055001739**	**Cluster 1 - 217** **Cluster 2 - 202** **Cluster 3 - 142** **Cluster 4 - 39**	2.105330515	0.064524974	Cluster 1 - 253 Cluster 2 - 122 Cluster 3 - 54 Cluster 4 - 171

Best Model -
From the analysis of the clusters above and considering a tradeoff between a good value of DB index, Dunn index and Cluster formations, a sub-optimal cluster is obtained from Ward method in Agnes (highlighted in orange). The dendrogram below shows the tree is cut to form 4 clusters. The circular dendrogram is like the dendrogram on the left but arranges everything radially.







2. **BASIS FOR PURCHASE-** The results for hclust and agnes for basis for purchase variables have been tabulated in the table below.

Hclust Results - Clusters have been formed using methods such as Complete, Single, Average and Ward. To evaluate the clusters, we evaluate cluster formations, DB index and Dunn Index. Below are the observations –

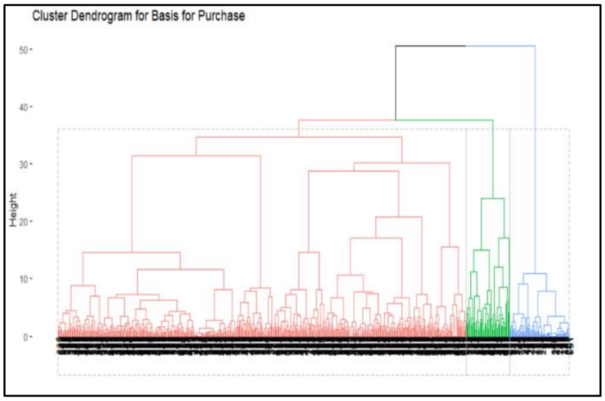
a. Single and Average Method clusters – Similar to Purchase Behavior, the clusters formed by these methods have the lowest DB index and high Dunn index. But on evaluating the cluster formations, we notice that a high number of observations (more than 580 out of 600) are concentrated in a single cluster and the remaining clusters have less than 10 observations, some have single observations. Such clusters, if used to implement any marketing campaign will not be cost effective or useful. Hence, we disregard these clusters.
b. Complete method clusters – The clusters formed through this method has a cluster which has just one observation in it. Similar to Single and Average method, the cluster formation is not useful to create market segments to effectively implement campaigns.
c. Ward method clusters – Considering the cluster formation and indices (DB and Dunn), the clusters formed are sub-optimal. The DB index is high (>=2) except for cluster k = 3 for Euclidean and k = 2 for Manhattan.

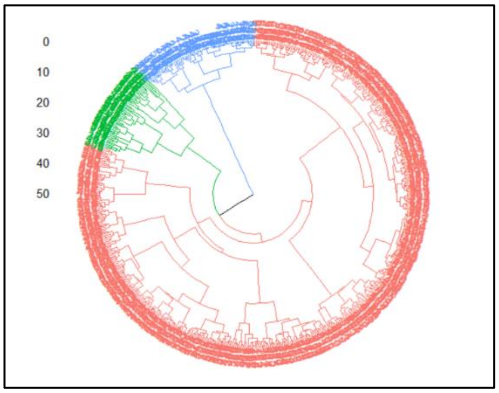
Agnes Results- In agnes, we get agglomerative coefficient which measures the strength of the clustering. This helps us to find the clustering methods that can identify strong clustering structures. Analyzing the results below, we can see that ward method has the highest coefficient. Hence, we used ward method to form clusters. The cluster formation is good and DB index is lower when compared to the clusters from hclust Ward method.

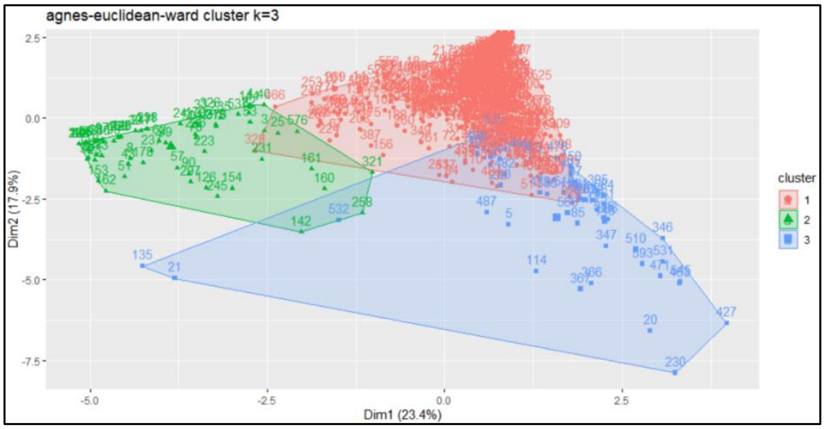
Agglomerative Coeeficient		
Method	Euclidean	Manhattan
complete	0.9414433	0.9343055
average	0.9348039	0.9148875
single	0.8888392	0.8488117
ward	0.9789986	0.9812461

VARIABLE SET -> BASIS OF PURCHASE							
HCLUST							
Input Parameters		Euclidean Distance			Manhattan Distance		
Methods	K Value	DB Index	Dunn Index	No. of Items in the clusters	DB Index	Dunn Index	No. of Items in the clusters
Complete	2	0.125741864	0.689548087	Cluster 1 - 599 Cluster 2 - 1	1.313557833	0.114223933	Cluster 1 - 594 Cluster 2 - 6
	3	1.074763189	0.108615181	Cluster 1 - 567 Cluster 2 - 32 Cluster 3 - 1	2.671269969	0.114223933	Cluster 1 - 579 Cluster 2 - 15 Cluster 3 - 6
	4	1.959204521	0.120686303	Cluster 1 - 555 Cluster 2 - 32 Cluster 3 - 12 Cluster 4 - 1	3.439543949	0.046222254	Cluster 1 - 410 Cluster 2 - 169 Cluster 3 - 15 Cluster 4 - 6
Average	2	0.125741864	0.689548087	Cluster 1 - 599 Cluster 2 - 1	0.195898553	0.689548087	Cluster 1 - 599 Cluster 2 - 1
	3	0.224593219	0.35192184	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1	0.281397682	0.35192184	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1
	4	0.274465723	0.323196166	Cluster 1 - 597 Cluster 2 - 1 Cluster 3 - 1 Cluster 4 - 1	0.494640528	0.296827229	Cluster 1 - 596 Cluster 2 - 1 Cluster 3 - 2 Cluster 4 - 1
Single	2	0.125741864	0.689548087	Cluster 1 - 599 Cluster 2 - 1	0.195898553	0.689548087	Cluster 1 - 599 Cluster 2 - 1
	3	0.224593219	0.35192184	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1	0.281397682	0.35192184	Cluster 1 - 598 Cluster 2 - 1 Cluster 3 - 1
	4	0.311321742	0.350839668	Cluster 1 - 597 Cluster 2 - 1 Cluster 3 - 1 Cluster 4 - 1	0.332858204	0.350839668	Cluster 1 - 597 Cluster 2 - 1 Cluster 3 - 1 Cluster 4 - 1
Ward.D	2	1.099032908	0.05734946	Cluster 1 - 535 Cluster 2 - 65	1.036806366	0.05734946	Cluster 1 - 532 Cluster 2 - 68
	3	1.430166354	0.040844529	Cluster 1 - 195 Cluster 2 - 340 Cluster 3 - 65	2.261978587	0.03426861	Cluster 1 - 284 Cluster 2 - 68 Cluster 3 - 248
	4	2.06158078	0.041057329	Cluster 1 - 195 Cluster 2 - 277 Cluster 3 - 65 Cluster 4 - 63	2.191306387	0.03426861	Cluster 1 - 284 Cluster 2 - 68 Cluster 3 - 162 Cluster 4 - 86
AGNES							
Input Parameters		Euclidean Distance			Manhattan Distance		
Model	K	DB Index	Dunn Index	No. of Items in each clusters	DB Index	Dunn Index	No. of Items in each clusters
Ward.D	2	1.158351167	0.05734946	Cluster 1 - 530 Cluster 2 - 70	1.066662952	0.05734946	Cluster 1 - 527 Cluster 2 - 73
	3	**1.443081361**	**0.06113192**	Cluster 1 - 479 Cluster 2 - 70 Cluster 3 - 51	1.812793733	0.039911183	Cluster 1 - 415 Cluster 2 - 73 Cluster 3 - 112
	4	1.794261921	0.040499594	Cluster 1 - 252 Cluster 2 - 227 Cluster 3 - 70 Cluster 4 - 51	2.000926568	0.039007693	Cluster 1 - 313 Cluster 2 - 73 Cluster 3 - 102 Cluster 4 - 112

<u>Best Model</u> -
From the analysis of the clusters above and considering a tradeoff between a good value of DB index, Dunn index and Cluster formations, a sub-optimal cluster is obtained from Ward method in Agnes (highlighted in orange). The dendrogram below shows the tree is cut to form 3 clusters.







<u>BOTH VARIABLES COMBINED</u>

The results for hclust and agnes for both, basis for purchase and purchase behavior, variables have been tabulated in the table below.

<u>Hclust Results -</u> Clusters have been formed using methods such as Complete, Single, Average and Ward. To evaluate the clusters, we evaluate cluster formations, DB index and Dunn Index. Below are the observations –

a. Single and Average Method clusters – Similar to previous cases, the clusters formed by these methods have the lowest DB index and high Dunn index. But on evaluating the cluster formations, we notice that a high number of observations (more than 580 out of 600) are concentrated in a single cluster and the remaining clusters have less than 10 observations, some have single observations. Such clusters, if used to implement any marketing campaign will not be cost effective or useful. Hence, we disregard these clusters.
b. Complete method clusters – The clusters formed through this method has a cluster which has just one observation in it. Like Single and Average method, the cluster formation is not useful to create market segments to effectively implement campaigns.

c. Ward method clusters – Considering the cluster formation and indices (DB and Dunn), the clusters formed are sub-optimal. The DB index is high (>=2).

Agnes Results- In agnes, we get agglomerative coefficient which measures the strength of the clustering. Analyzing the results below, we can see that ward method has the highest coefficient. Hence, we used ward method to form clusters. The cluster formation is good and DB index is the least and Dunn index is the highest when compared to the clusters from hclust Ward method for k = 6.

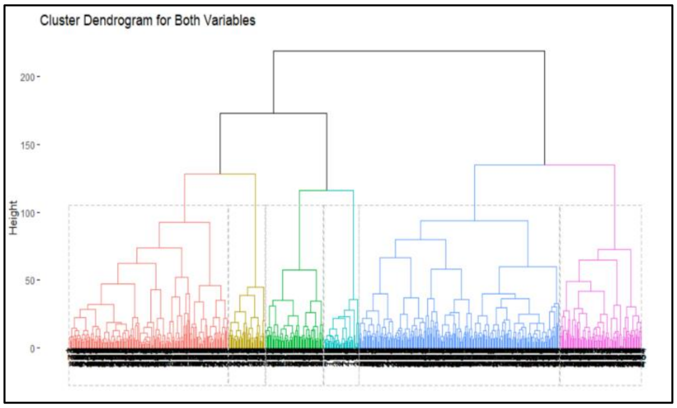
Agglomerative Coeeficient		
Method	Euclidean	Manhattan
complete	0.8904667	0.8838798
average	0.8672723	0.8324098
single	0.7744702	0.6946303
ward	0.9594398	0.9669323

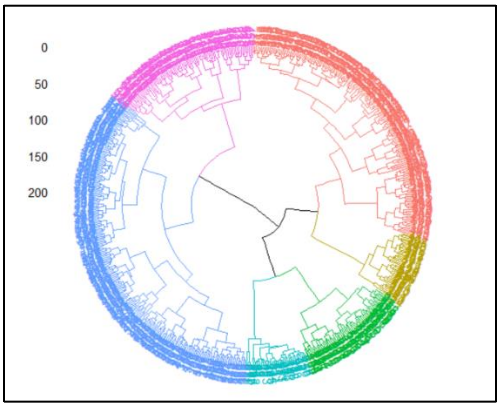
VARIABLE SET -> BOTH (PURCHASE BEHAVIOUR + BASIS FOR PURCHASE)							
Input Parameters		Euclidean Distance			Manhattan Distance		
Methods	K Value	DB Index	Dunn Index	No. of Items in the clusters	DB Index	Dunn Index	No. of Items in the clusters
Complete	4	1.616539073	0.074478135	1 2 3 4 480 59 60 1	1.808127929	0.095398683	1 2 3 4 357 205 37 1
	5	1.719918388	0.079524132	1 2 3 4 5 480 45 60 14 1	1.818322678	0.103949034	1 2 3 4 5 256 101 205 37 1
	6	1.659646483	0.089636975	1 2 3 4 5 6 471 45 60 9 14 1	1.800545424	0.107683559	1 2 3 4 5 6 256 101 30 37 175 1
	7	2.131638263	0.094098491	1 2 3 4 5 6 7 65 406 45 60 9 14 1	1.82120074	0.114660458	1 2 3 4 5 6 7 221 101 30 37 175 35 1
	8	2.389134609	0.095518602	1 2 3 4 5 6 7 8 65 406 45 48 9 12 14 1	1.904806405	0.121098479	1 2 3 4 5 6 7 8 221 101 30 37 155 35 20 1
Average	4	0.407008851	0.167977224	1 2 3 4 593 2 4 1	0.38331734	0.213717664	1 2 3 4 597 1 1 1
	5	0.417097638	0.167977224	1 2 3 4 5 592 2 4 1 1	0.727472048	0.207836249	1 2 3 4 5 594 1 3 1 1
	6	0.42292307	0.167977224	1 2 3 4 5 6 591 2 4 1 1 1	0.787787793	0.120669049	1 2 3 4 5 6 521 73 1 3 1 1
	7	0.517251996	0.167977224	1 2 3 4 5 6 7 588 2 4 3 1 1 1	0.843761499	0.120669049	1 2 3 4 5 6 7 517 73 1 3 4 1 1
	8	0.81749135	0.099538859	1 2 3 4 5 6 7 8 580 8 2 4 3 1 1 1	0.881403911	0.126402502	1 2 3 4 5 6 7 8 504 73 1 13 3 4 1 1
Single	4	0.284479421	0.339573141	1 2 3 4 597 1 1 1	0.390548255	0.290611205	1 2 3 4 597 1 1 1
	5	0.406420981	0.339363466	1 2 3 4 5 596 1 1 1 1	0.471551475	0.28442282	1 2 3 4 5 596 1 1 1 1
	6	0.415574704	0.324890181	1 2 3 4 5 6 595 1 1 1 1 1	0.533058376	0.27131635	1 2 3 4 5 6 595 1 1 1 1 1
	7	0.40193147	0.323066772	1 2 3 4 5 6 7 594 1 1 1 1 1 1	0.521430568	0.268498993	1 2 3 4 5 6 7 594 1 1 1 1 1 1
	8	0.409698948	0.321725282	1 2 3 4 5 6 7 8 593 1 1 1 1 1 1 1	0.541135232	0.268145127	1 2 3 4 5 6 7 8 593 1 1 1 1 1 1 1
Ward.D	4	2.428336535	0.064796969	1 2 3 4 125 342 88 45	2.433158384	0.070287596	1 2 3 4 296 103 108 93
	5	2.468084122	0.064796969	1 2 3 4 5 125 232 88 45 110	2.336156116	0.070287596	1 2 3 4 5 264 103 108 93 32
	6	2.660354272	0.064796969	1 2 3 4 5 6 125 154 88 45 110 78	2.164747026	0.070287596	1 2 3 4 5 6 264 103 71 37 93 32
	7	2.64848602	0.064796969	1 2 3 4 5 6 7 125 154 88 45 75 78 35	2.536015529	0.077973668	1 2 3 4 5 6 7 156 103 71 37 93 108 32
	8	2.424907686	0.064796969	1 2 3 4 5 6 7 8 100 154 88 45 25 75 78 35	2.465303236	0.077973668	1 2 3 4 5 6 7 8 156 103 71 37 93 50 32 58

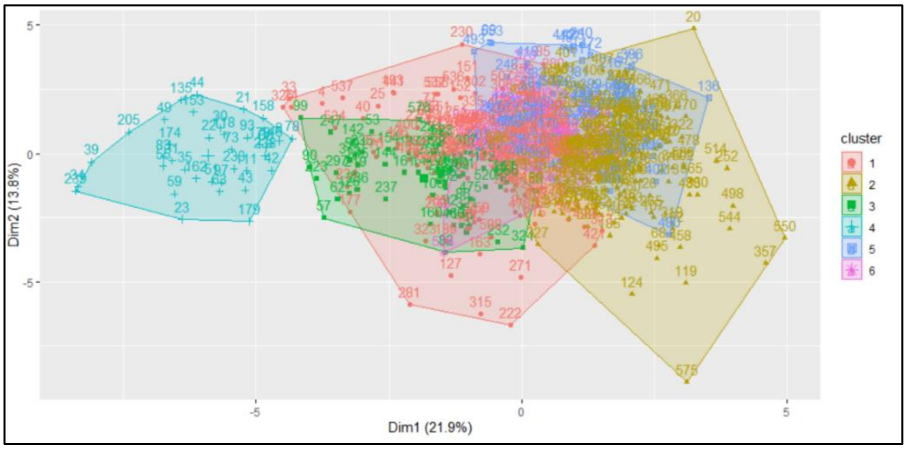
AGNES							
Input Parameters		Euclidean Distance			Manhattan Distance		
Model	K	DB Index	Dunn Index	No. of Items in each clusters	DB Index	Dunn Index	No. of Items in each clusters
Ward.D	4	2.374324584	0.061664722	1 2 3 4 209 310 44 37	2.403678897	0.077058392	1 2 3 4 206 210 98 86
	5	2.460024861	0.061664722	1 2 3 4 5 209 232 44 37 78	2.231673161	0.081326548	1 2 3 4 5 167 210 98 86 39
	6	2.795094629	0.070180318	1 2 3 4 5 6 209 157 44 37 75 78	**2.128143763**	**0.081326548**	**1 2 3 4 5 6** **167 210 61 37 86 39**
	7	2.639835108	0.070180318	1 2 3 4 5 6 7 125 157 84 44 37 75 78	2.38763824	0.082654524	1 2 3 4 5 6 7 167 120 61 90 37 86 39
	8	2.555345513	0.070180318	1 2 3 4 5 6 7 8 125 157 84 44 37 75 35 43	2.40927258	0.100448298	1 2 3 4 5 6 7 8 127 120 61 90 37 40 86 39

<u>Best Model</u> –

From the analysis of the clusters above and considering a tradeoff between a good value of DB index, Dunn index and Cluster formations, a sub-optimal cluster is obtained from Ward method in Agnes (highlighted in orange). The dendrogram below shows the tree is cut to form 6 clusters.







<u>Best Models for different Variable sets -</u> The most optimal models for the three variables have been collated and tabulated below -

Input Parameters				Cluster Evuaulation Parameter		
Variable Set	**K**	**Distance**	**Method**	**DB Index**	**Dunn Index**	**No. of Items in each clusters**
Purchase Behavior	4	Euclidean	Ward.D	1.946171598	0.055001739	Cluster 1 - 217 Cluster 2 - 202 Cluster 3 - 142 Cluster 4 - 39
Basis for Purchase	3	Euclidean	Ward.D	1.443081361	0.06113192	Cluster 1 - 479 Cluster 2 - 70 Cluster 3 - 51
Both Combined	6	Manhattan	Ward.D	2.128143763	0.081326548	1 2 3 4 5 6 167 210 61 37 86 39

DBSCAN Clustering:

DBSCAN is a partitioning method that can find out clusters of different shapes and sizes from data containing noise and outliers. The key idea is that for each point of a cluster, the neighborhood of a given radius must contain at least a minimum number of points. Thus, clusters are dense regions in the data space, separated by regions of lower density of points.

Parameter estimation

We have created several DBSCAN models by experimenting with the below mentioned **input parameters**

1. **Eps:** The parameter epsilon define the radius of neighborhood around a point x. It's called the ϵϵ-neighborhood of x. The value for ϵϵ can be chosen by using a k-distance graph, plotting the distance to the k = MinPts nearest neighbor. If ϵϵ is too small, sparser clusters will be defined as noise. If ϵϵ is too large, denser clusters may be merged. Good values of ϵϵ are where the plot shows a strong bend.

2. **MinPts:** This parameter is the minimum number of neighbors within "eps" radius. Any point x in the data set, with a neighbor count greater than or equal to MinPts, is marked as a core point. We say that x is border point, if the number of its neighbors is less than MinPts, but it belongs to the ϵϵ-neighborhood of some core point z. Finally, if a point is neither a core nor a border point, then it is called a noise point or an outlier. Thus, the choice of MinPts is very essential as it directly impacts the classification of a point as core, border or noise.

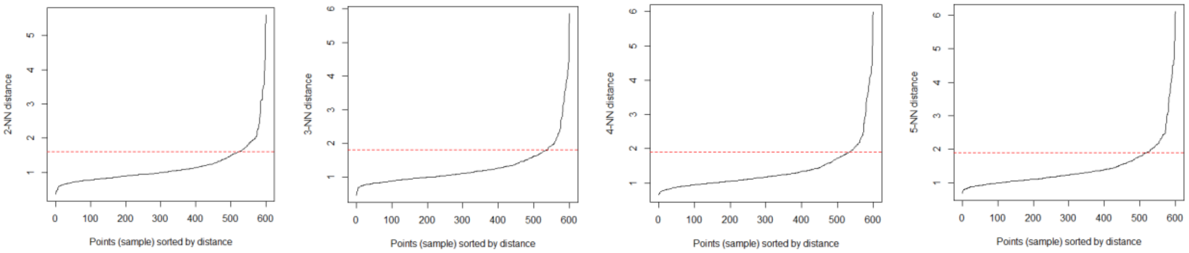
Method for determining the optimal eps value

- The method that we have used to determine the optimal eps value, consists of computing the k nearest neighbor distances in a matrix of points.
- The idea is to calculate, the average of the distances of every point to its k nearest neighbors. The value of k which we have considered are different for different set of variables.
    1. Purchase Behavior: k value 2 through 5
    2. Basis of Purchase: k value 2 through 5
    3. Both: k value 4 through 7
- Next, these k-distances are plotted in an ascending order. The aim is to determine the "knee" which corresponds to a threshold where a sharp change occurs along the k-distance curve. The knee gives a close approximation of the optimal eps value for DBSCAN clustering.

Part 1: Purchase Behavior

Using the approach defined above, we have created graphs for different values of k. Below are the graphs of best eps values for different values of k.

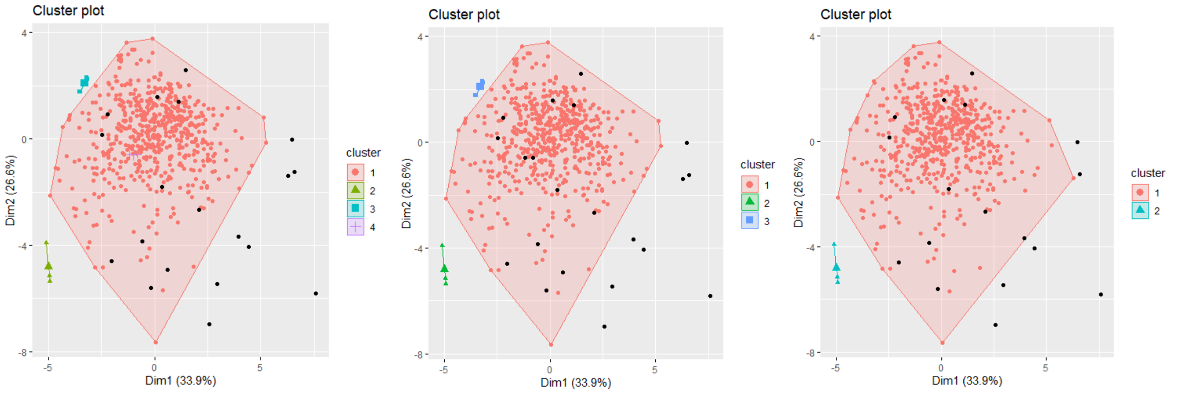
Below are the eps graphs for K = 2,3,4,5



As can be seen from the above graphs, the optimal eps values range near 2. Thus, we will be experimenting with eps values in this range with different combinations of MinPts to get good clusters using DBSCAN clustering. The table lists down the different models that we have created for Purchase Behavior.

	Input Parameters		Output Parameters																		
Model No.	EPS Value	Min Point	No. Of Clusters	Noise Points	Cluster Distribution															
1	1.5	2	15	53	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
					53	516	2	2	2	2	2	2	4	2	2	2	3	2	2	2
2		3	3	77	C0	C1	C2	C3												
					77	516	4	3												
3	1.6	2	11	39	C0	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11				
					39	540	2	2	3	2	2	2	2	2	2	2				
4		3	2	57	C0	C1	C2													
					57	540	3													
5	1.7	2	8	32	C0	C1	C2	C3	C4	C5	C6	C7	C8							
					32	552	3	3	2	2	2	2	2							
6		3	3	42	C0	C1	C2	C3												
					42	552	3	3												
7	1.8	2	7	27	C0	C1	C2	C3	C4	C5	C6	C7								
					27	559	3	3	2	2	2	2								
8		3	3	35	C0	C1	C2	C3												
					35	559	3	3												
9	1.9	2	7	24	C0	C1	C2	C3	C4	C5	C6	C7								
					24	562	3	3	2	2	2	2								
10		3	3	32	C0	C1	C2	C3												
					32	562	3	3												
11	2	2	5	21	C0	C1	C2	C3	C4	C5										
					21	569	3	3	2	2										
12		3	3	25	C0	C1	C2	C3												
					25	569	3	3												
13	2.1	2	4	20	C0	C1	C2	C3	C4											
					20	572	3	3	2											
14		3	3	22	C0	C1	C2	C3												
					22	572	3	3												
15	2.2	2	4	19	C0	C1	C2	C3	C4											
					19	573	3	3	2											
16		3	3	21	C0	C1	C2	C3												
					21	573	3	3												
17	2.3	2	2	18	C0	C1	C2													
					18	579	3													
18	2.4	2	2	17	C0	C1	C2													
					17	580	3													

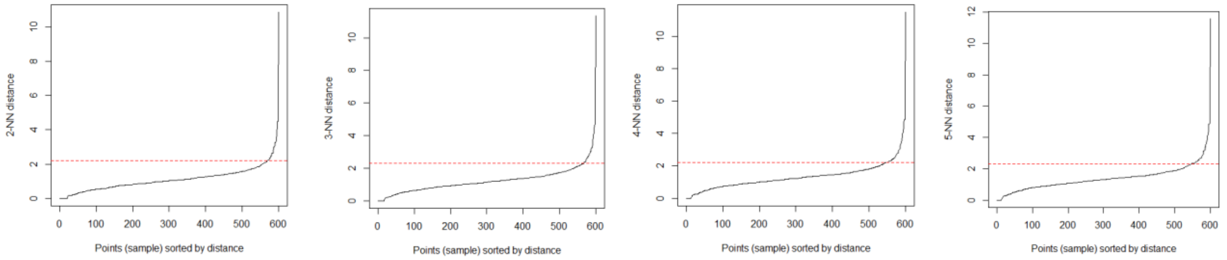
Best Model for Purchase Behavior



- ➢ Having ran DBSCAN clustering with different models, we infer from the table and graphs that the different combinations of the input parameters seem to have no impact on the output.
- ➢ As can be seen in the table and graphical results, most of the data points are going into one cluster and the other clusters have very minimal data points.
- ➢ The reason behind the inappropriate clusters could be the varying densities that DBSCAN is not able to handle.
- ➢ Thus, clusters with minimal data points prove out to be insignificant as such cluster segmentation will not help the company to create marketing strategies. The inappropriate cluster partitioning leads to the decision of rejecting DBSCAN clustering method.

Part 2: Basis of Purchase

We now proceed ahead with the second variable set. We repeat the process of creating graphs for different value of k. The interpretation from these graphs help us decide the narrow optimal range for the eps value. We have used the values of k from 2 through 5 to get the optimal eps values.

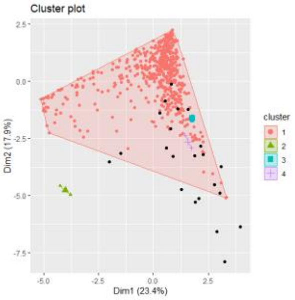


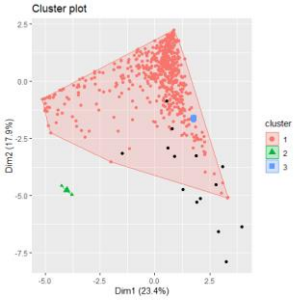
The graph for Basis of Purchase indicate that the optimal eps value will lie approximately close to 2. Thus, we play with eps value close to 2 and several MinPts to create different DBSCAN clustering models. We have stored the performance of different models in a tabular format which includes the different combinations of the input parameters and their corresponding output clusters. The table will help us to evaluate different clusters, briefly.

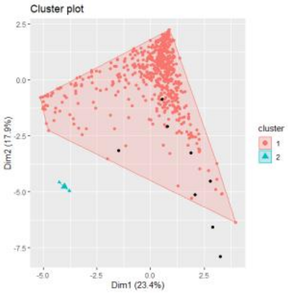
	Input Parameters		Output Parameters										
Model No.	EPS Value	Min Point	No. Of Clusters	Noise Points	Cluster Distribution								
1	1.8	2	7	42	C0	C1	C2	C3	C4	C5	C6	C7
					42	543	2	3	2	3	2	3
2		3	4	48	C0	C1	C2	C3	C4			
					48	543	3	3	3			
3	1.9	2	7	33	C0	C1	C2	C3	C4	C5	C6	C7
					33	552	2	3	2	3	2	3
4		3	4	39	C0	C1	C2	C3	C4			
					39	552	3	3	3			
5	2	2	7	27	C0	C1	C2	C3	C4	C5	C6	C7
					27	558	2	3	2	2	2	4
6		3	3	35	C0	C1	C2	C3				
					35	558	3	4				
7	2.1	2	4	22	C0	C1	C2	C3	C4			
					22	572	2	2	2			
8	2.2	2	3	18	C0	C1	C2	C3				
					18	578	2	2				
9	2.3	2	3	15	C0	C1	C2	C3				
					15	581	2	2				
10	2.4	2	3	14	C0	C1	C2	C3				
					14	582	2	2				
11	2.5	2	2	12	C0	C1	C2					
					12	586	2					
12	2.6	2	2	10	C0	C1	C2					
					10	588	2					
13	2.7	2	2	9	C0	C1	C2					
					9	589	2					
14	2.8	2	2	9	C0	C1	C2					
					9	589	2					
15	2.9	2	2	8	C0	C1	C2					
					8	590	2					
16	3	2	2	8	C0	C1	C2					
					8	590	2					

Best Model for Basis of Purchase

- ➢ A close look at the above table helps us to infer that DBSCAN clustering for Basis of Purchase is clustering all data points in a single cluster.
- ➢ Like the clustering for Purchase Behavior, the clustering in this case is also highly skewed towards one cluster such that a cluster alone contains the 99% data points.
- ➢ A single cluster containing most of the data points proves out to be of no use to the company that aims at creating approaches for handling different clusters. We therefore reject the DBSCAN clustering method for this variable set.

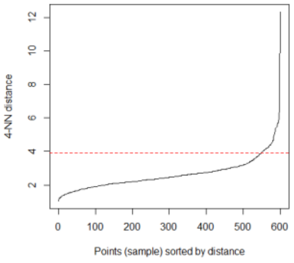


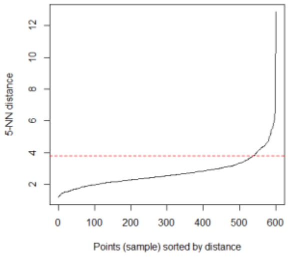


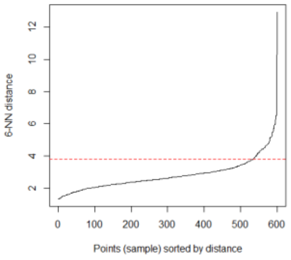


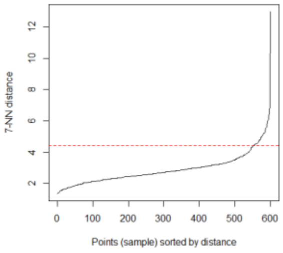
Part 3: Both (Purchase Behavior + Basis of Purchase)

This variable set includes both the Purchase Behavior variables and Basis of Purchase variables. In this case, as we are including a greater proportion of variables as compared to the other two variable sets, we have used a higher range of k values = 4 to 7.







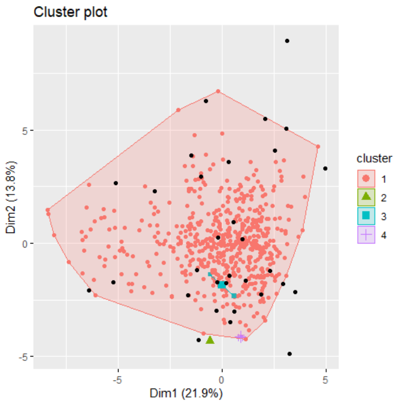


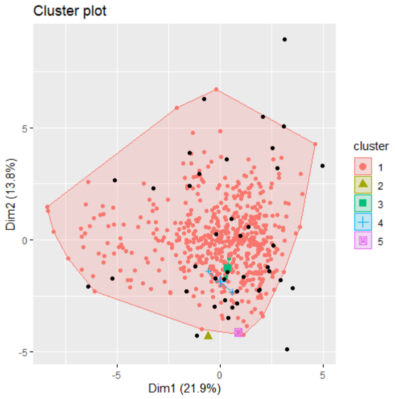
The graphs help us examine the optimal range for the eps values. Typically, the optimal eps value will lie roughly close to 4 for this set of variables. Using combinations of eps values in the range of 4 and divergent MinPts values, we have designed multiple DBSCAN clustering models.

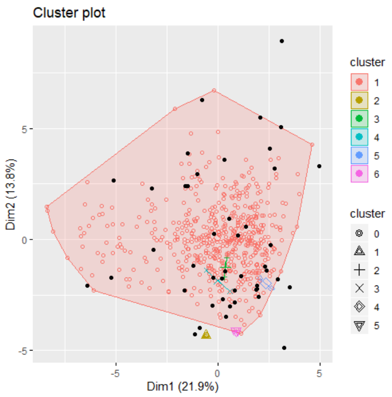
	Input Parameters		Output Parameters										
Model No.	EPS Value	Min Point	No. Of Clusters	Noise Points	Cluster Distribution								
					C0	C1	C2	C3	C4	C5	C6	C7	C8
1	3	2	8	50	50	532	2	2	2	2	2	5	3
2		3	3	60	60	532	3	5					
3	3.1	2	6	45	45	544	2	2	2	2	3		
4		3	2	53	53	544	3						
5	3.2	2	5	40	40	551	2	2	2	3			
6		3	2	46	46	551	3						
7	3.3	2	4	38	38	555	2	2	3				
8		3	2	42	42	555	3						
9	3.4	2	4	34	34	559	2	2	3				
10		3	2	38	38	559	3						
11	3.5	2	4	31	31	562	2	2	3				
12		3	2	35	35	562	3						
13	3.6	2	3	30	30	565	2	3					
14		3	2	32	32	565	3						
15	3.7	2	3	29	29	566	2	3					
16		3	2	31	31	566	3						
17	3.8	2	2	25	25	573	2						
18	4	2	2	19	19	579	2						
19	4.1	2	2	17	17	581	2						
20	4.2	2	2	15	15	583	2						
21	4.3	2	2	14	14	584	2						
22	4.4	2	2	12	12	586	2						
23	4.5	2	3	10	10	586	2	2					

Best Model for Both (Purchase Behavior + Basis of Purchase)

- Aligned with the results from the other two variable sets, the combined variable set of Purchase Behavior and Basis of Purchase also does not perform well with DBSCAN clustering.
- The reason behind the disability of DBSCAN clustering method to create distinct clusters lies in its weakness of not identifying clusters of varying density.
- In addition to this, DBSCAN clustering does not work well high dimensional datasets.
- Therefore, we can conclude that the DBSCAN method does not function as desired with any of our variable sets and hence will not be useful for CRISA to create their cluster segmentation for business requirements.
- Thus, we have rejected DBSCAN as a clustering method for CRISA.







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

The percentage of total purchases comprised by various brands should be considered alike while analyzing the clusters. For example, if a household is making purchases of high value with less no. of brands but high percentage of total purchases of other brands, we will consider this household to be less loyal in our case. A customer who buys all brand A is not just as loyal as a customer who buys all brand B because he may have had high brand runs before/after purchasing this brand.

Hence, variables such as brand run and share with other brands should also be considered while deciding the loyalty of a household. Many variables including brand share are in percentage format which need to be changed prior to running k-mean algorithm because this will increase the within cluster distance and the dimensionality of the clusters. Therefore, these variable needs to be normalized to allow all the variables to equally contribute to the clustering.

Question 4. (a) Are the clusters obtained from the different procedures similar/different? Describe how they are similar/different.
All the four clustering methods work on different concept:

- K- means – Based on centroids
- K medoids – Based on medoids and take care of outliers
- Agglomerative – Based on connectivity
- DBSCAN – Based on density

Hierarchical clustering can't handle big data well but K Means clustering can. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.

The difference in models is seen from the type of clusters they give. None of the above models in DBSCAN gave good performance in terms of well-defined clusters. There seem to be too much overlap of data points. DBSCAN clustering clusters all data points in a single cluster. When the parameters are changed i.e. epsilon value and min points is changed, the formation of clusters does not show a good segmentation size of the data.

Hence, we cannot consider this model for this dataset. So, we are rejecting these DBSCAN models as best models.

Comparing the best models of k-medoids, k-means and agglomerative models for all 3 variable sets:

Purchase Behavior: We observe that DB index is minimum for K -means and Dunn index is maximum for Agglomerative clustering.

Clustering Method	Value of K	DB Index	Dunn Index	No. of items in each cluster
K means	4	1.51	0.049	Cluster 1: 191 Cluster 2: 188 Cluster 3: 175 Cluster 4: 46 Total No. of Data Points: 600
K medoids	3	1.90	0.031	Cluster 1: 186 Cluster 2: 248 Cluster 3: 166 Total No. of Data Points: 600
Agglomerative	4	1.95	0.055	Cluster 1: 217 Cluster 2: 202 Cluster 3: 142 Cluster 4: 39 Total no. of Data Points: 600

Basis for Purchase: We observe that DB index is minimum for K -means and Dunn index is maximum for Agglomerative clustering like purchase behavior.

Clustering Method	Value of K	DB Index	Dunn Index	No. of items in each cluster
K means	2	1.17	0.044	Cluster 1: 524 Cluster 2: 76 Total No. of Data Points: 600
K medoids	3	2.07	0.035	Cluster 1: 344 Cluster 2: 175 Cluster 3: 81 Total No. of Data Points: 600
Agglomerative	3	1.44	0.061	Cluster 1: 479 Cluster 2: 70 Cluster 3: 51 Total No. of Data Points: 600

Both purchase behavior and Basis for Purchase: We observe that DB index is minimum for K -means and Dunn index is maximum for Agglomerative clustering as above.

Clustering Method	Value of K	DB Index	Dunn Index	No. of items in each cluster
K means	6	1.78	0.076	Cluster 1: 45 Cluster 2: 111 Cluster 3: 70 Cluster 4: 175 Cluster 5: 48 Cluster 6: 151 Total No. of Data Points: 600
K medoids	6	1.89	0.067	Cluster 1: 69 Cluster 2: 164 Cluster 3: 55 Cluster 4: 48 Cluster 3: 154 Cluster 4: 110 Total No. of Data Points: 600
Agglomerative	6	2.13	0.081	Cluster 1: 167 Cluster 2: 210 Cluster 3: 61 Cluster 4: 37 Cluster 5: 86 Cluster 6: 39 Total No. of Data Points: 600

For the bath soap data set, we found that method K-means with both purchase behavior and basis for purchase gives good results when compared based on DB index and Dunn index. It also has better formed clusters as shown in the graph enclosed in that section.

Hence, K-means method with K = 6 for clustering done with both variable sets is our best model.

4(b) 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.)

Our best model is **k-means with K=6** and we have done clustering on both purchase behavior and basis of purchase. We end up with 6 clusters.

We have tried to explain these clusters based on few demographic and other indicators which we feel render unique characteristics to each cluster.

cluster_1: Households here have, on average, 4 members per household and buy products from very few brands, but their average transactions per brand run is highest. Thus, we can say they have low or even no brand loyalty. Most of the purchases for these households seem to be purchases not on promotion. One point to note about customers in this cluster is that exchange offer drives their soap buying decision. These households probably try different brands in others999 or are loyal to some other brand listed in others999. Their major purchases comprise of the popular soaps.

cluster_2: Households here buy from a higher number of brands; highest amount of volume is those which has low brand runs. This again points to no brand loyalty. These households have the least brand loyalty for our selected brands and have the highest brand loyalty for the brands in others999. Premium soaps are their major purchases. They mostly purchase from brands that are not under promotion. They have highest number of household members which would account for their high volume of transactions. They majorly buy products in extra gram mage and buy beauty soaps.

cluster_3: Households here buy a decent number of brands, have a low number of brands runs but for a highest amount of volume. All the purchases for these households seem to be purchases not on promotion. They have high brand loyalty towards our selected brands and a low brand loyalty for all the other brands. They mostly buy popular soaps. They have highest average value. Customers in this cluster have high brand loyalty and they use coupons majorly for their purchases.

cluster_4: Households here buy moderately low number of brands, have lowest brand runs for the lowest amount of volume per transaction. Most of the purchases for these households seem to be purchases not on promotion. They have the highest brand loyalty for the selected brands and are least loyal towards the others. Customers in this cluster can be characterized as average except that they buy products in exchange offer majorly.

cluster_5: Households here have low affluence index, which means they purchase less and when they do, they could prefer their favorite brands. Number of transactions shows that. Their average price purchases are also average. All of their purchases are from brands not on promotion. Customers here have less brand loyalty and buy products from any brand which are being offered at price off**.**

cluster_6: Households for this cluster are like cluster_5 as they do not have multiple purchases from same brand, volume of their purchases and number of times they buy from separate brands. They differ from cluster_5 on average price. In one purchase, they buy more amount of product from same brand. Customers here probably prefer less volume but higher priced items. Their affluence index is higher as compared to cluster_5, also they prefer costlier items in the same brand. They do not seem to buy products from any kind of promotion but can be targeted for pricier variation of the product from our selected brands.

CONCLUSION:

- Most of the consumers are females, hence most of the ads by CRISA should be targeted for women.
- Also, most customers are in the segment who are not particularly brand loyal but they prefer to buy value added packs and try premium soaps.
- Most of the people have a TV/cable, so advertisements can be broadcasted on the television as an effective means of promoting these products.
- As evident from clusters 1 and 3, brand loyalty comes in effect when people have an option of coupons or exchange offers.
- Also, not many people care about the price offs. People buying on basis of price offs are comparatively very small across all the clusters.
- Thus, in order to promote brand loyalty, CRISA manufacturers should promote their brands by gifting coupons or exchange offers.
- From cluster_0, we can say that customers who care about the price offs, are those who will buy products on discounts irrespective of the brand and they are generally illiterate or do not have any formal schooling.
- Keeping all these points in mind and cost effectiveness, CRISA should make their marjeting strategy.

4(c) For one 'best' segmentation, obtain a description of the clusters by building a decision tree to help describe the clusters. How effective is the tree in helping to explain/interpreting the cluster(s)? (explain why/why not). (You may use a decision tree to help choose the 'best' clustering).

Decision tree is used here to describe the clusters. We run the decision tree on the clustering done for both which has six segments of customers. With the help of decision trees, we can find define the customer's purchase behavior and basis of purchase and how they are classified into a particular customer segment.

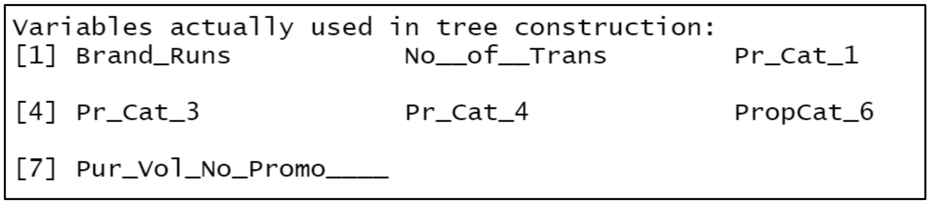
Parameters used for making a decision tree are:

- Criterion= gini
- Maximal depth= 10
- Confidence = 0.01
- Minimal size for split = 5

First, we ran the decision tree on our best model – K mean with k=6 for both segmentation:

<u>**Decision Tree for Both Variables**</u>

The rpart result for Purchase Behavior shows the variable used for the construction of the tree -

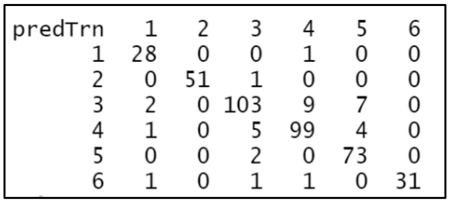


```
Variables actually used in tree construction:
[1] Brand_Runs            No__of__Trans        Pr_Cat_1

[4] Pr_Cat_3              Pr_Cat_4             PropCat_6

[7] Pur_Vol_No_Promo____
```

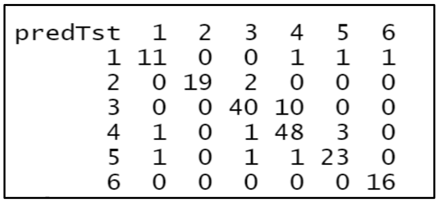
For the combined set of variables, rpart utilizes 7 variables (as shown above) to build the decision tree. All the variables are a subset of the variables which were used for cluster formations.

The decision tree seems to be capturing the segmentation of 6 households with test accuracy of 88.21%. The training accuracy came out to be = 92.7%. The accuracy is high which shows that decision tree is predicting majority clusters as true clusters.

The confusion matrix for train and test data below shows that the decision tree correctly classifies the clusters accurately.



```
predTrn   1   2   3   4   5   6
      1  28   0   0   1   0   0
      2   0  51   1   0   0   0
      3   2   0 103   9   7   0
      4   1   0   5  99   4   0
      5   0   0   2   0  73   0
      6   1   0   1   1   0  31
```

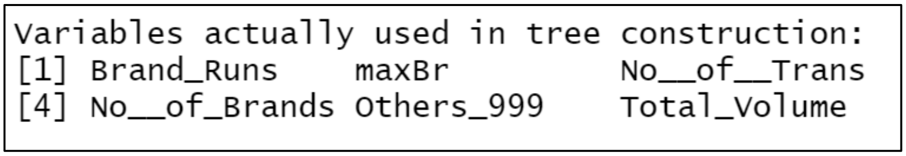


```
predTst  1  2  3  4  5  6
      1 11  0  0  1  1  1
      2  0 19  2  0  0  0
      3  0  0 40 10  0  0
      4  1  0  1 48  3  0
      5  1  0  1  1 23  0
      6  0  0  0  0  0 16
```

Secondly, we have also ran decision tree for our second-best model – k=4 for purchase behavior:

Decision Tree on Purchase Behavior variables

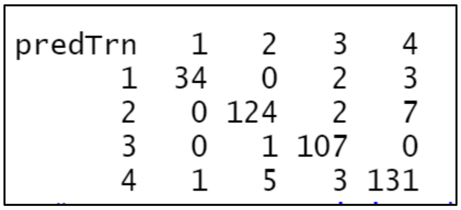
The rpart result for Purchase Behavior shows the variable used for the construction of the tree –



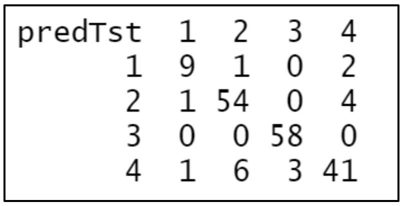
```
Variables actually used in tree construction:
[1] Brand_Runs    maxBr         No__of__Trans
[4] No__of_Brands Others_999    Total_Volume
```

Out of the 10 variables which are used to form clusters for Purchase Behavior, six of them are used to construct the tree and the training and test accuracies are – 0.94.3% and 90% respectively.

The confusion matrix for train and test data below shows that the decision tree correctly classifies the clusters accurately.



```
predTrn   1   2   3   4
      1  34   0   2   3
      2   0 124   2   7
      3   0   1 107   0
      4   1   5   3 131
```



```
predTst  1  2  3  4
      1  9  1  0  2
      2  1 54  0  4
      3  0  0 58  0
      4  1  6  3 41
```

Thus, we can see that the decision trees do help identify the clusters correctly.

REFERENCES:

1. http://www.learnbymarketing.com/tutorials/k-means-clustering-in-r-example/
2. https://www.displayr.com/what-is-hierarchical-clustering/
3. https://www.stat.berkeley.edu/~s133/Cluster2a.html
4. https://online.stat.psu.edu/stat555/node/85/
5. https://stats.stackexchange.com/questions/195446/choosing-the-right-linkage-method-for-hierarchical-clustering
6. https://rdrr.io/cran/mdendro/man/linkage.html
7. https://bradleyboehmke.github.io/HOML/hierarchical.html
8. https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf
9. https://towardsdatascience.com/introduction-hierarchical-clustering-d3066c6b560e