ASSIGNMENT - 3

Market Segmentation (Segmenting Consumers of Bath Soap)

Ashok Bhatraju - 670248723

Shourya Narayan – 651193827

Vivek Kumar - 670460685

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

In the context of customer segmentation, cluster analysis is the use of a mathematical model to discover groups of similar customers based on finding the smallest variations among customers within each group. These homogeneous groups are known as Clusters. The goal of cluster analysis in marketing is to accurately segment customers in order to achieve more effective customer marketing via personalization.

In our case, we are going to find clusters based on purchase behaviour and basis of purchase. Later we compare how demographic attributes are associated with these two parameters. In doing so, it will give us in depth understanding of our target market behaviours. This will enable advertising agencies and consumer goods manufacturers to plan their market strategies accordingly. This way they will be able to create marketing campaigns, promotions, discounts and other techniques based on their target cluster wants, needs and preferences.

2. Use k-means clustering to identify clusters of households based on

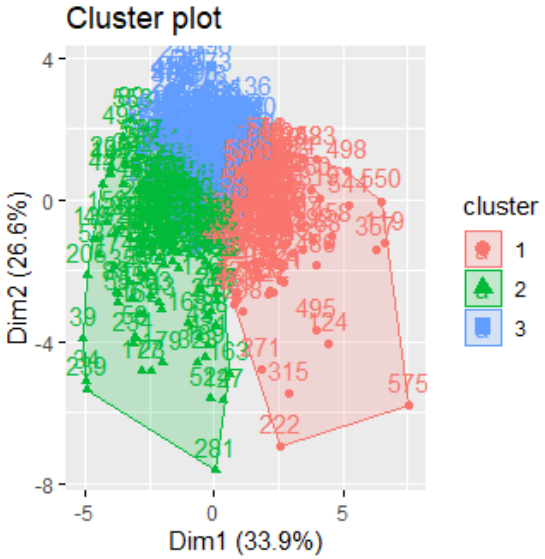
a. The variables that describe purchase behaviour (including brand loyalty). How will you evaluate brand loyalty – describe the variables you create/use to capture different perspectives on brand loyalty.

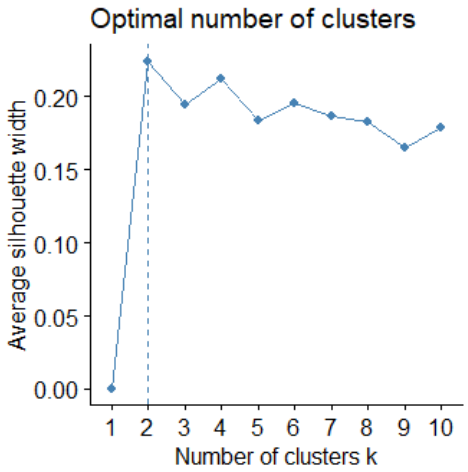
Answer:

Variables used:

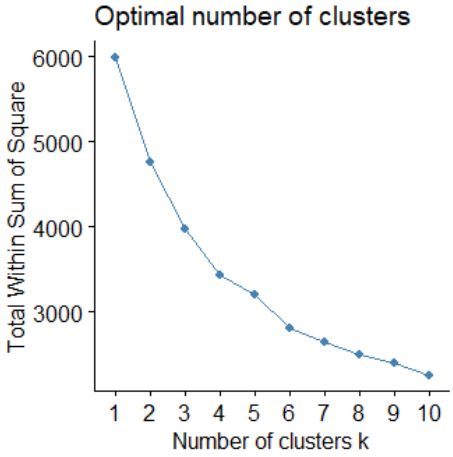
- Number of brands
- Brand runs
- Total volume
- Number of transactions
- Value
- Average price
- Share to other brands
- Brand loyalty

K=3





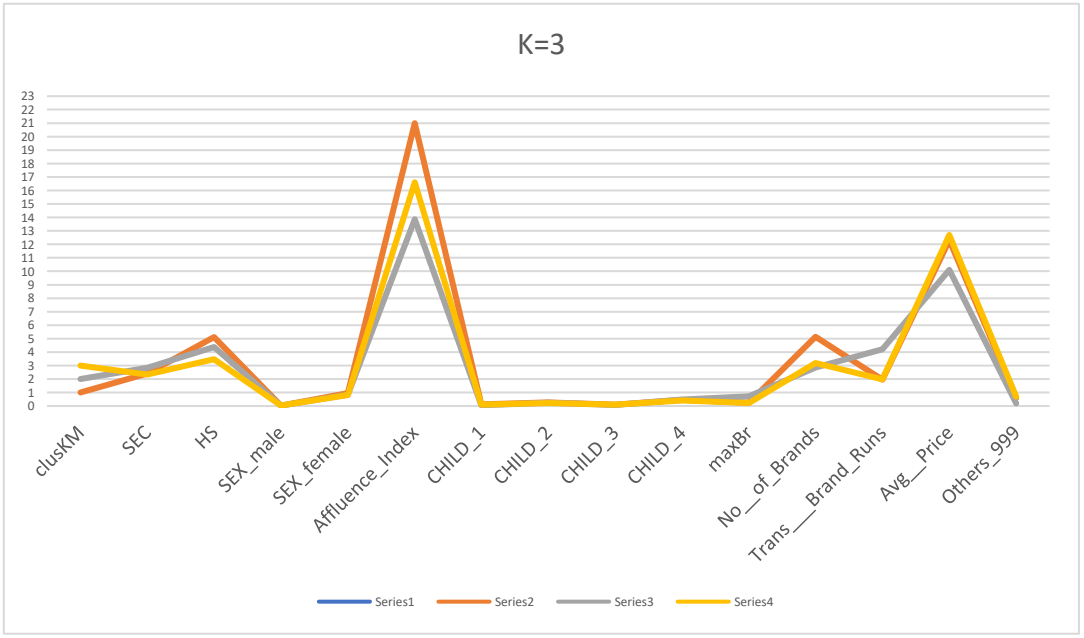
- silhouette plot suggests that the optimal K value for this data is K=2



Clusters 1 and 2 are well separated from each other but there are few similarities that intersect between these 3 clusters. From the graph below we can see that the parameters SEC, HS, SEX, Number of Brands, Number of Transactions are the attributes that provide us the better distinction between clusters. The other parameters like Child, maxBr and few other doesn't provide us with the information required for our business decisions.

Legend for all Excel plots:

Series 1 – Nil, Series 2 – Cluster 1, Series 3 – Cluster 2, Series 4 – Cluster 3, Series 5 – Cluster 4, Series 6 – Cluster 5

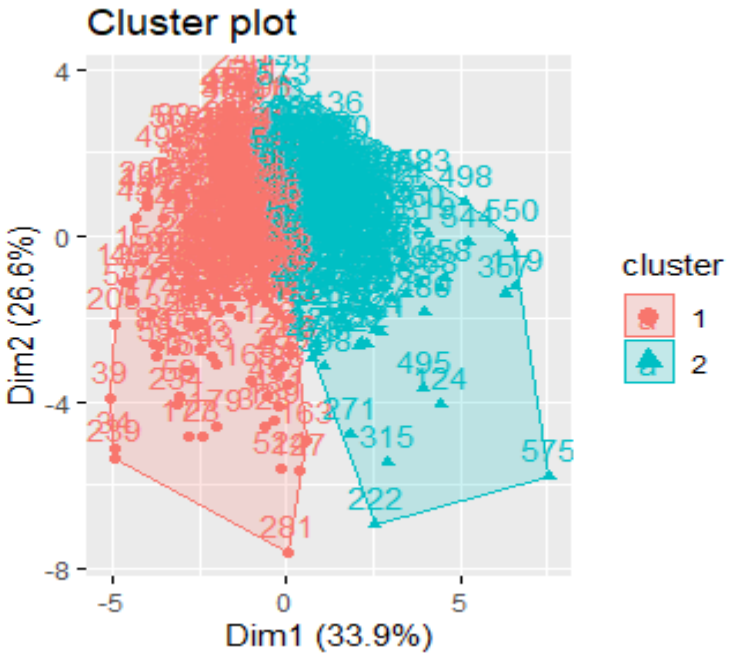


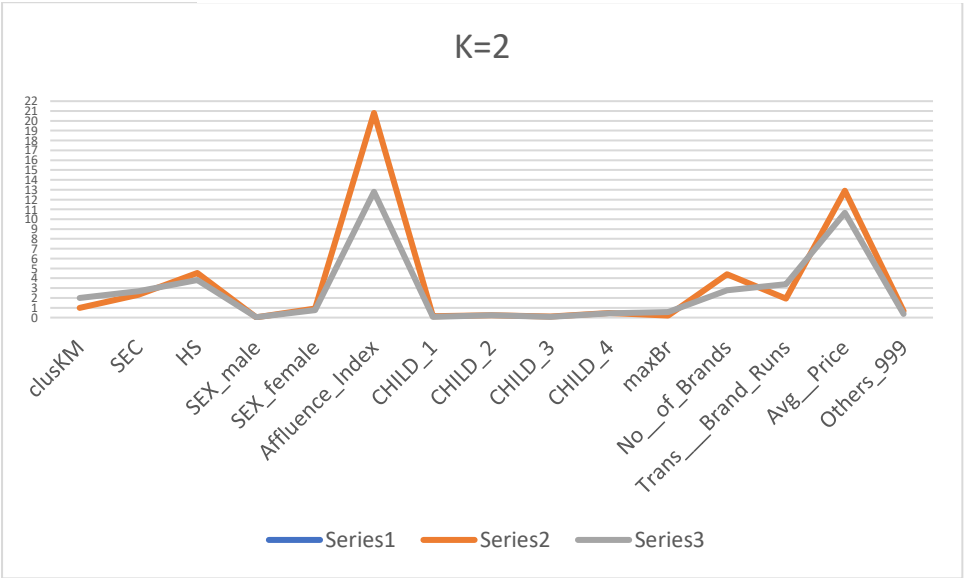
Cluster 1: This group has the maximum Affluence index and number of brands, suggest that these are least brand loyal.

Cluster 2: This group has the least affluence index and least average price when compared with other 2 clusters.

Cluster 3: This group has highest average price among all the clusters.

K=2





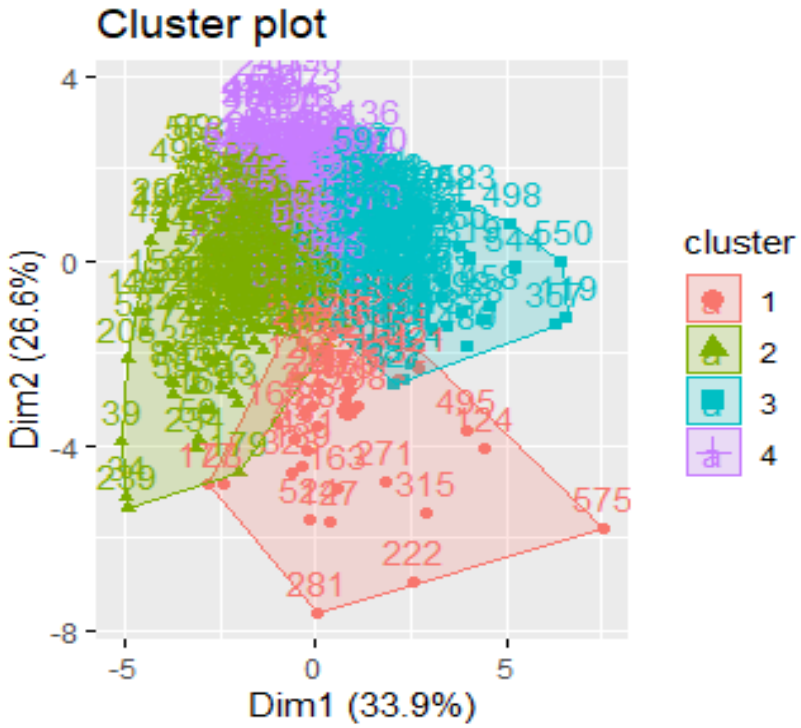
Cluster 1: This group has the highest affluence index as well as highest average price.

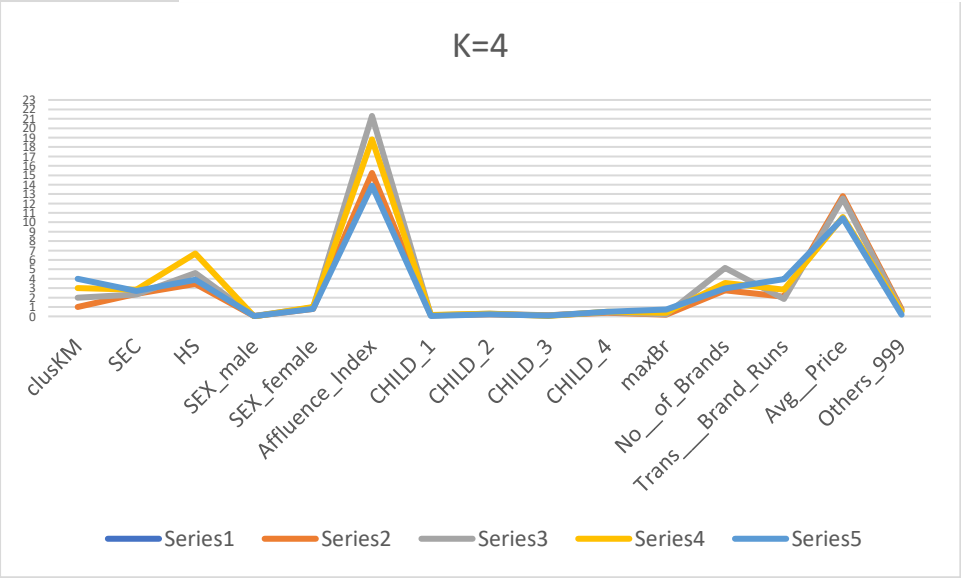
Cluster 2: This group has the least affluence and average price among these 2.

Both the clusters above are well separated and there are just few similarities between these clusters.

From the graph above we can see that the parameters SEC, HS, SEX, Number of Brands, Number of Transactions are the attributes that provide us the better distinction between clusters. The other parameters like Child, maxBr and few other doesn't provide us with the information required for our business decisions.

K=4



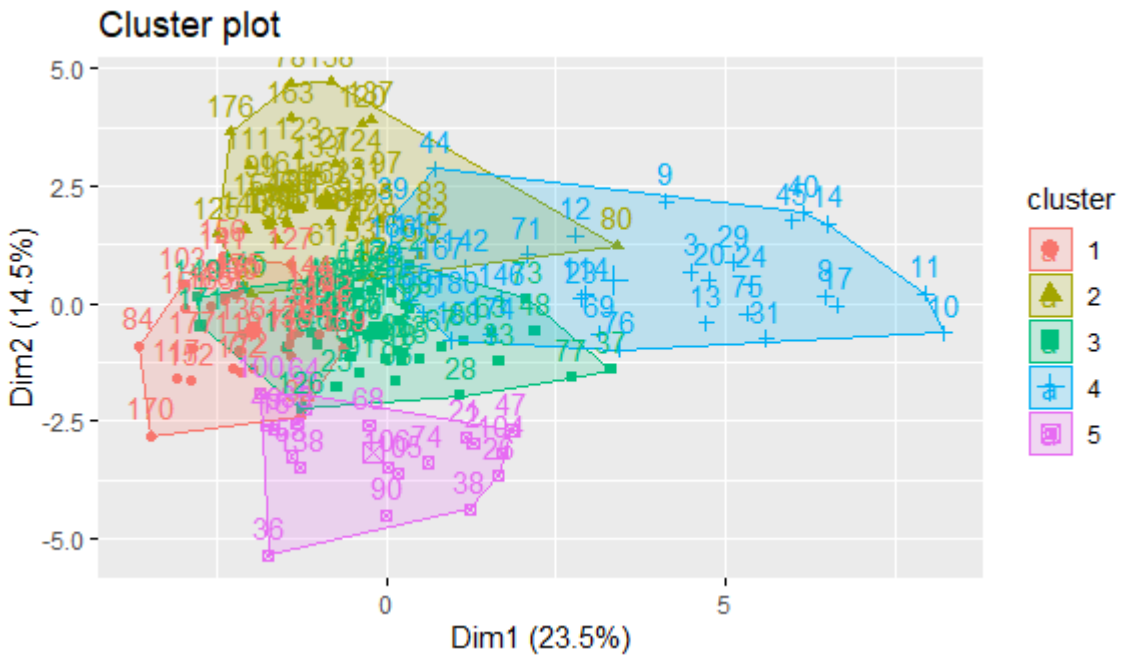


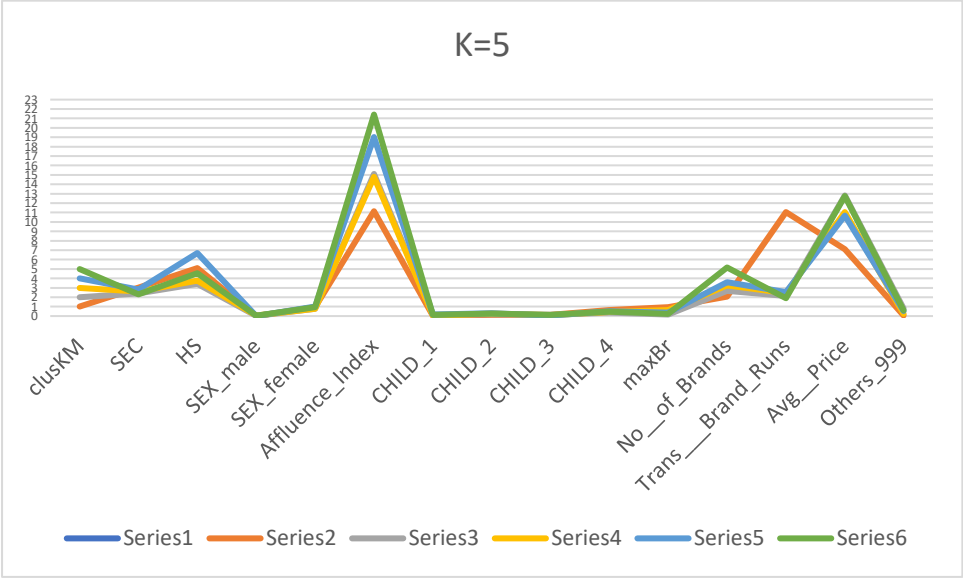
Cluster analysis: all the clusters above have a lot in common, most of them have the highest affluence index and highest average price.

Clusters 2 and 3 are well separated from each other and clusters 1 and 4 are well separated from each other. Overall there are few attributes that are similar among these 4 clusters.

These similarities can be explained from the graph above, we can see that attributes like maxBr and Child doesn't really show any difference among the 4 clusters. These are the attributes that created the overlapping of clusters and they are no useful for us to make any cluster-oriented business plans.

K=5





Cluster analysis:

Affluence index, number of brand runs and average price are distinct between all the above clusters. These attributes can be used to make our business decisions in order to have better outcomes.

Clusters 1 and 5 are very well separated from rest of the clusters. Clusters 2,3 and 4 have few similarities that caused the overlapping of these clusters. From the graph above we can see that the parameters SEC, HS, SEX, Number of Brands, Average price are the attributes that provide us the better distinction between clusters. The other parameters like Child, maxBr and few other doesn't provide us with the information required for our business decisions.

Summary:

	Within Cluster	Between Cluster	Cluster Size
K=3	**3970**	**2020**	**166,175,259**
K=2	**4754**	**1236**	**283,317**
K=4	**3428**	**2562**	**46,175,188,191**
K=5	**3037**	**2953**	**29,166,182,179,44**

The cluster plots for various K values have overlapping among clusters, so its not easy to decide which K value is best for our model.

The above table provides us information about distance within clusters and between clusters. From table we can see that K=5 has the lowest within distance and highest between distance.

So, K=5 is the best model among the K values we tried with.

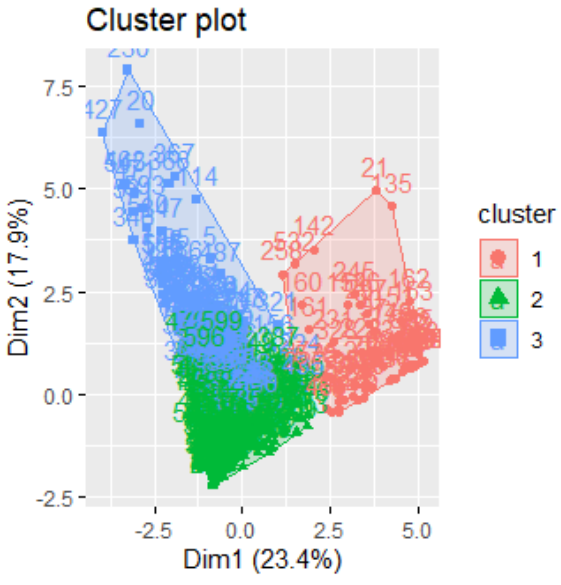
b. The variables that describe basis-for-purchase.

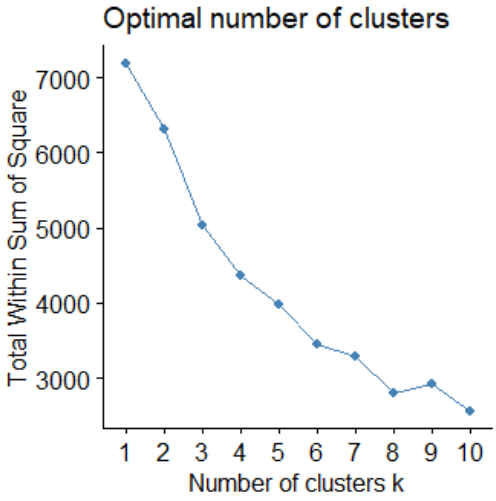
Answer:

Variables used

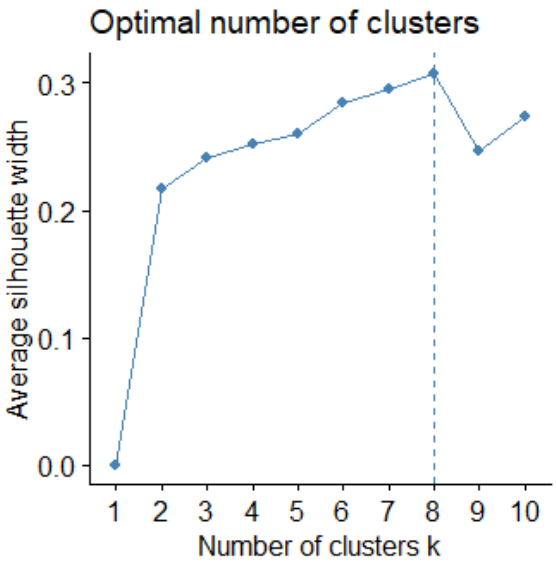
- Purchase by promotions
- Price categories
- Selling propositions

K=3



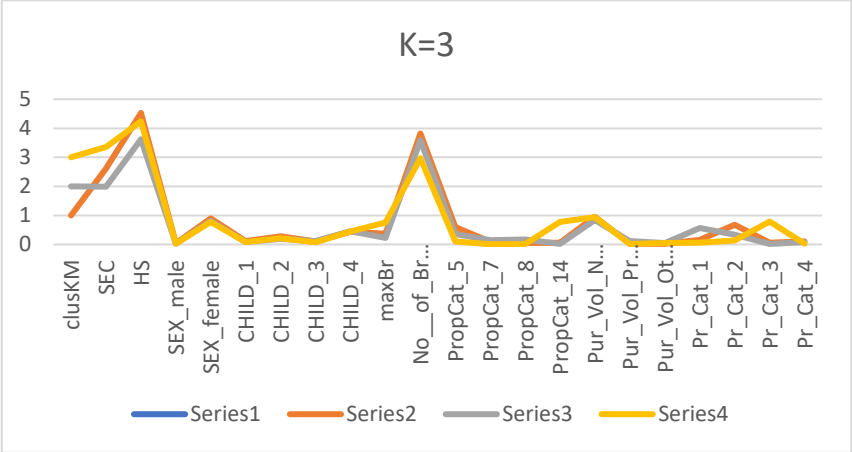


- Silhouette graph suggests that K=8 is the best number for this data set



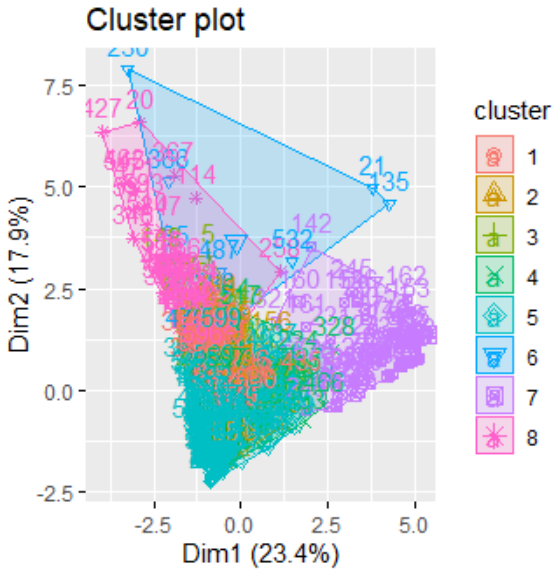
Clusters 2 and 3 are well separated from cluster 1. There is a bit of overlapping among clusters 2 and 3. This infers that there are few attributes that has no effect on these clusters.

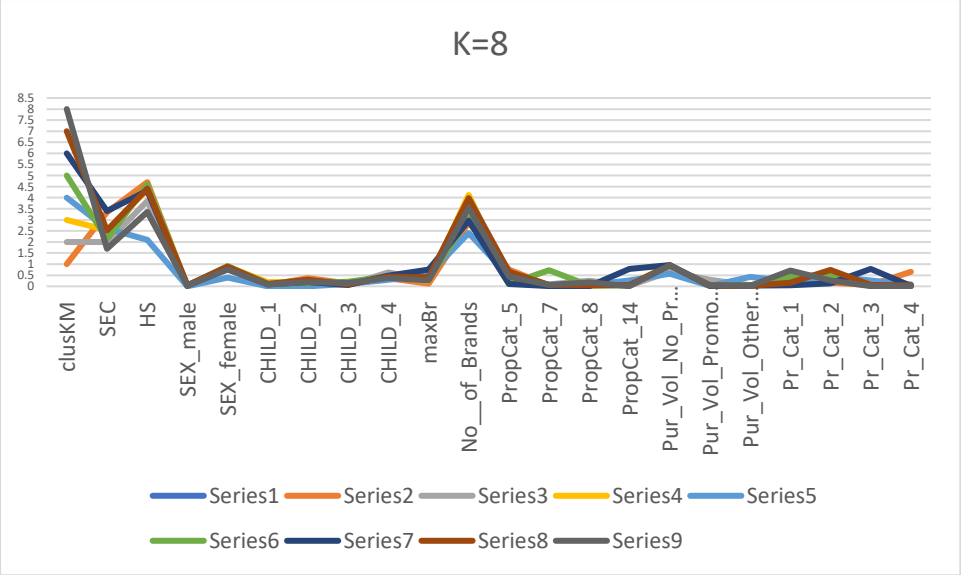
From the graph above we can see that the parameters SEC, HS, SEX, Number of Brands, purchase volume are the attributes that provide us the better distinction between clusters. The other parameters like maxBr, Child and few other doesn't provide us with the information required for our business decisions.



Cluster analysis: For the above clusters, number of brands is highest in all of them. It suggests that these group of clusters have a very least brand loyalty.

K=8

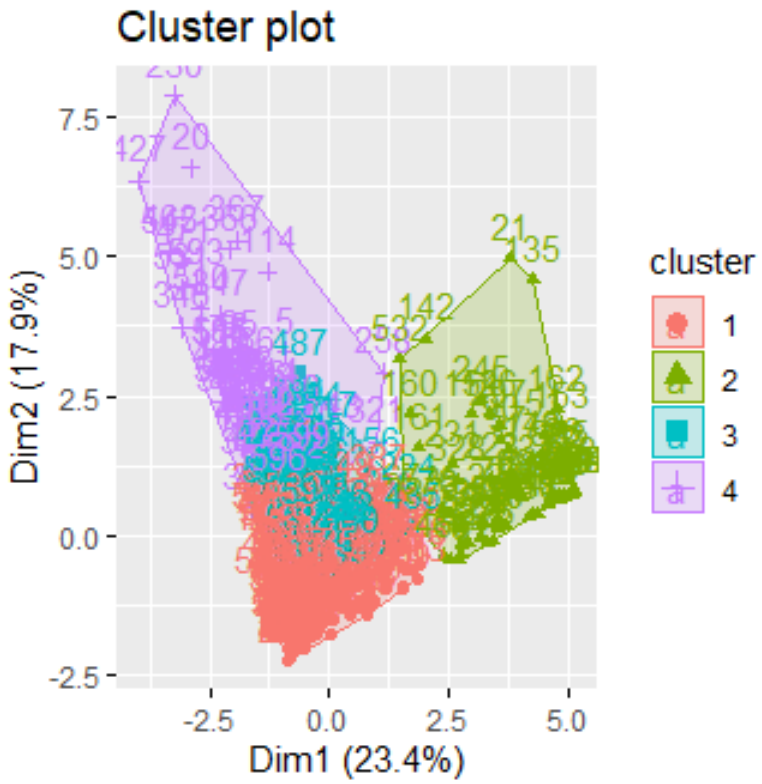


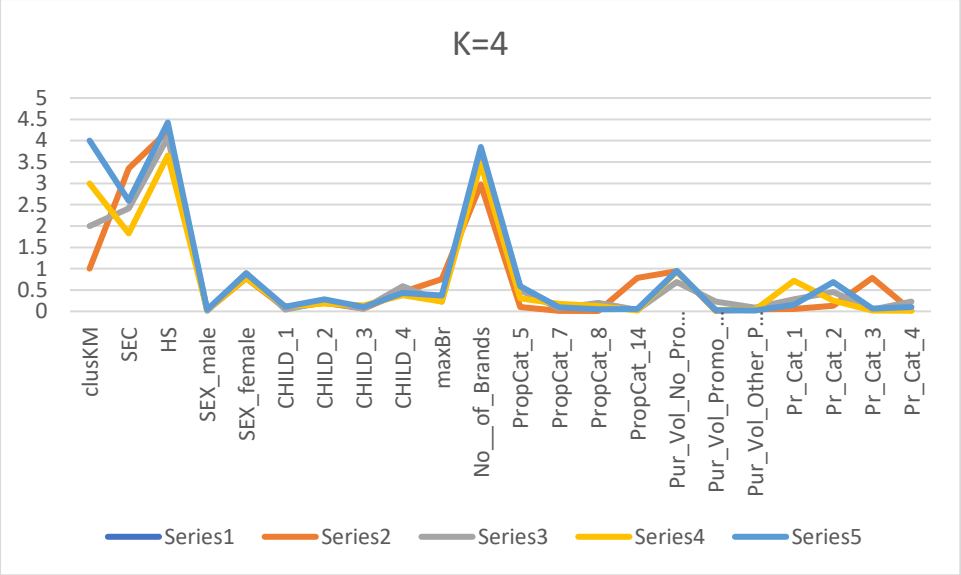


Apart from 6,7 and 8 clusters, rest all other clusters seem to have a lot of attribute values in common.

From the above graph we can see that variables SEC, HS, SEX, Number of Brands and Purchase volume are the variables that provide us with the required distinction. We can use these attributes for our business decisions to create cluster-oriented promotions and discounts.

K=4

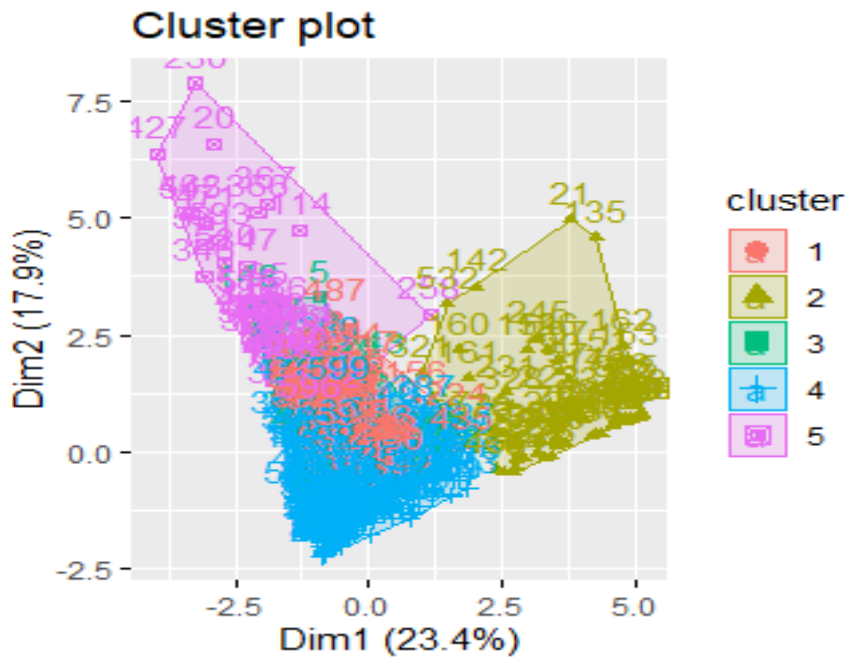


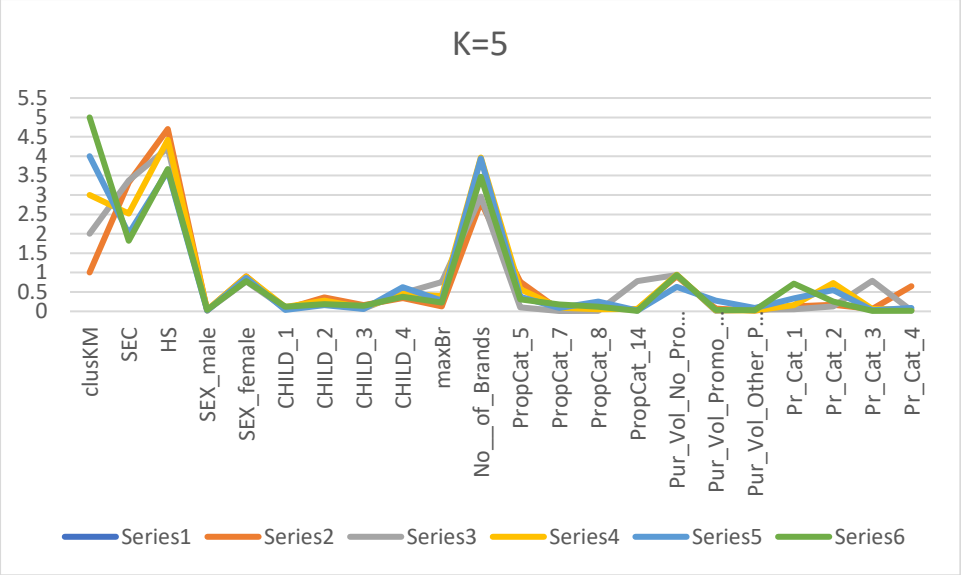


We can see that clusters 2 and 4 are well separated from each other and other 2 clusters. There is overlapping between clusters 1 and 3. This overlapping can be well explained by the graph above.

From the above graph we can see that variables like propcat_7, 8 and Child doesn't really provide us with the required distinction among our clusters. This is the main reason behind overlapping of our clusters.

K=5





Clusters 5 and 2 are well separated from each other, whereas the other 3 does have overlapping because of few variables that doesn't have much of difference in these clusters.

Summary:

	Within Cluster	Between Cluster	Cluster Size
K=3	**5029**	**2159**	**76,326,198**
K=8	**2782**	**4406**	**91,40,50,58,234,10,71,46**
K=4	**4364**	**2824**	**320,75,127,78**
K=5	**3813**	**3375**	**128,75,50,297,50**

The cluster plots for various K values have overlapping among clusters, so it's not easy to decide which K value is best for our model.

The above table provides us information about distance within clusters and between clusters. From table we can see that K=8 has the lowest within distance and highest between distance. But K value of 8 is too big and not feasible for business to have such a high number of cluster-oriented business plans.

So, the next best K value 5 is our best value for this model.

C. The variables that describe both purchase behaviour and basis of purchase.

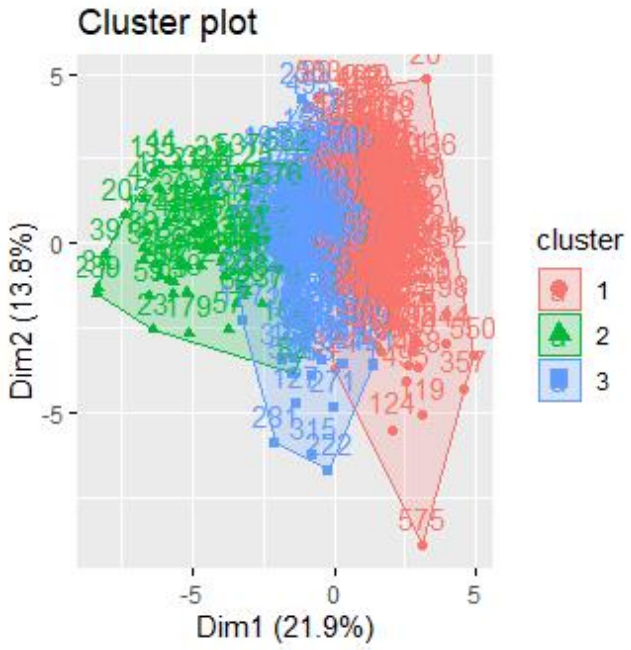
Answer:

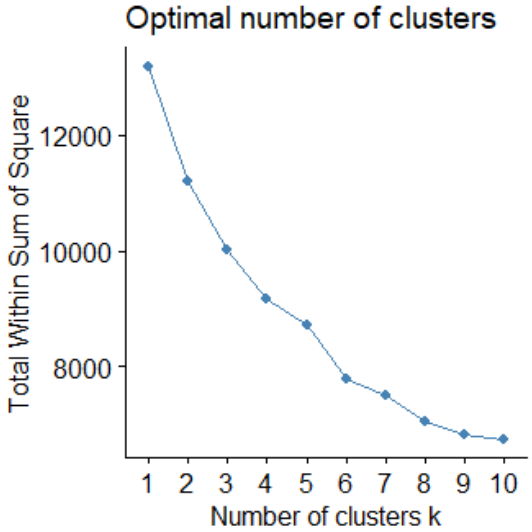
Variables used:

- Number of brands
- Brand runs
- Total volume
- Number of transactions
- Value
- Average price

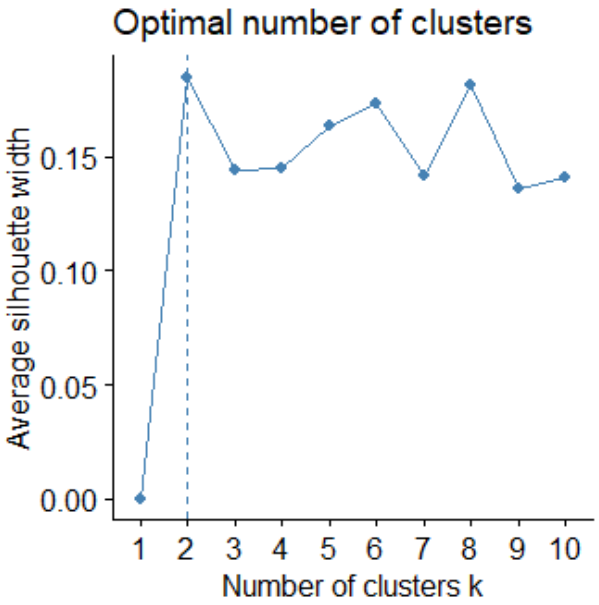
- Share to other brands
- Brand loyalty
- Purchase by promotions
- Price categories
- Selling propositions

K=3





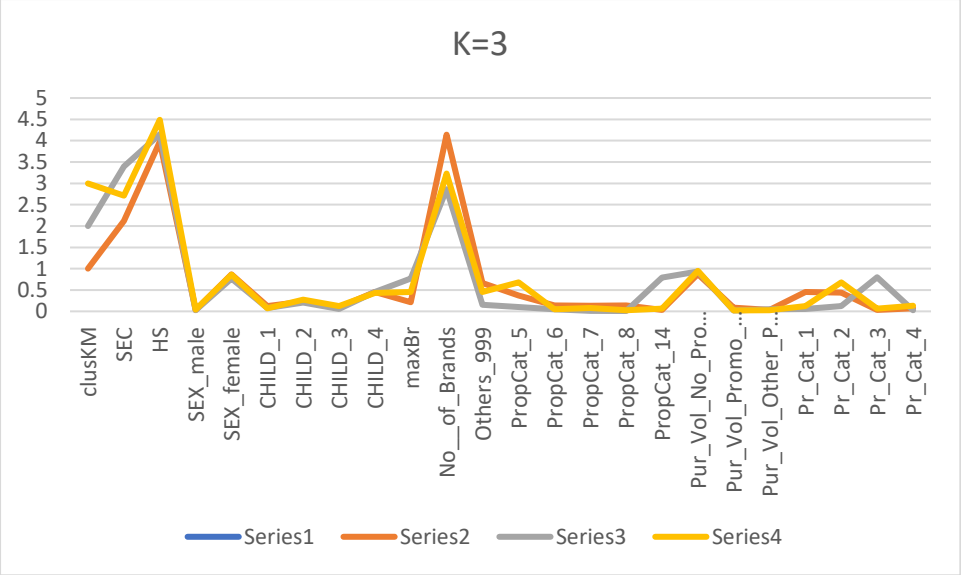
- Silhouette graph suggests that K=2 is the best value for these attributes



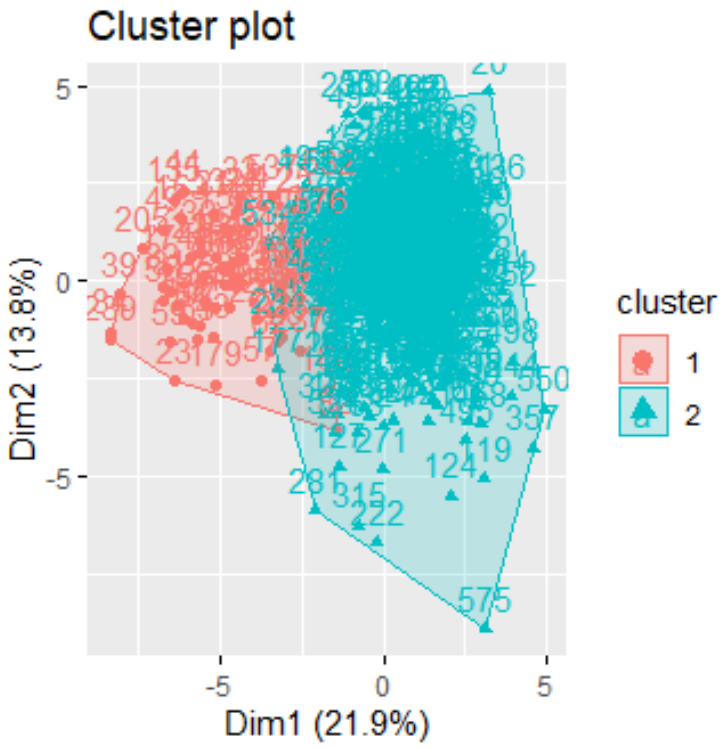
All the 3 clusters have few variables that overlap and that can be explained from the graph below

From the graph below we can see that the parameters SEC, HS, SEX, Number of Brands, purchase volume are the attributes that provide us the better distinction between clusters. The other parameters like maxBr, propcat_7, 8, Child and few other doesn't provide us with the information required for our business decisions.

Cluster analysis: Purchase volume with no promotion is highest for both the clusters, suggests that promotions doesn't have much of impact of buying behaviour of customers in these clusters.



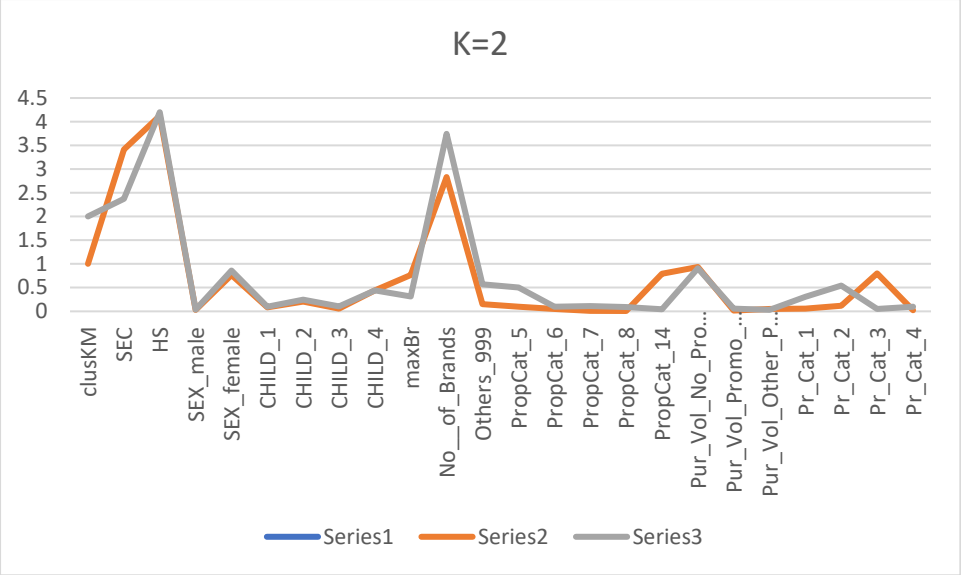
K=2



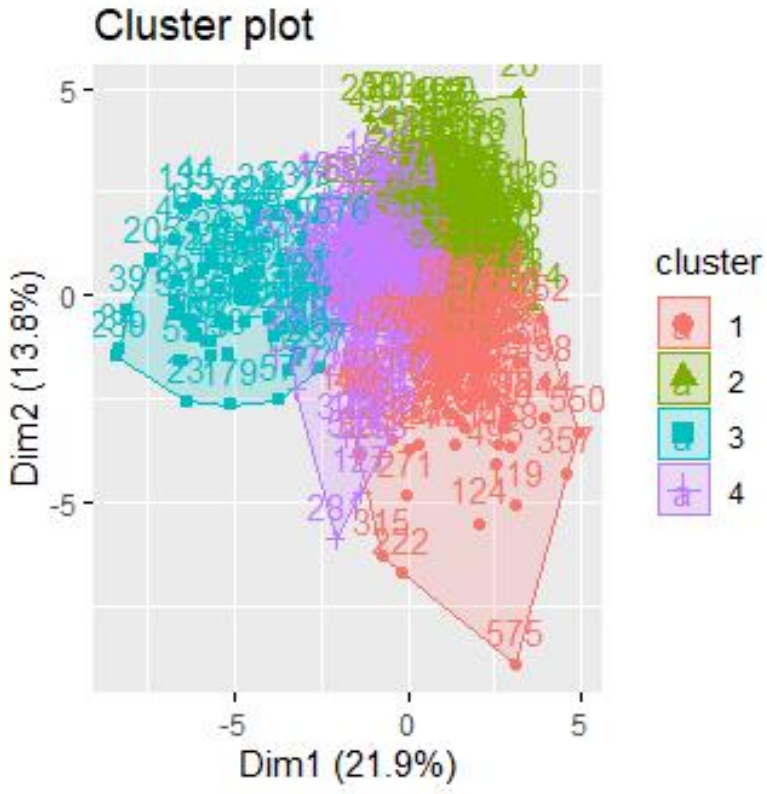
Both the clusters are well separated except for few overlapping caused by variables maxBr, propcat_7, 8 and Child that remained same across the clusters.

The graph below helps us in recognizing the attributes that are best suitable for each cluster and plan accordingly.

Cluster analysis: Purchase volume with no promotion is highest for both the clusters, suggests that promotions doesn't have much of impact of buying behaviour of customers in these clusters.

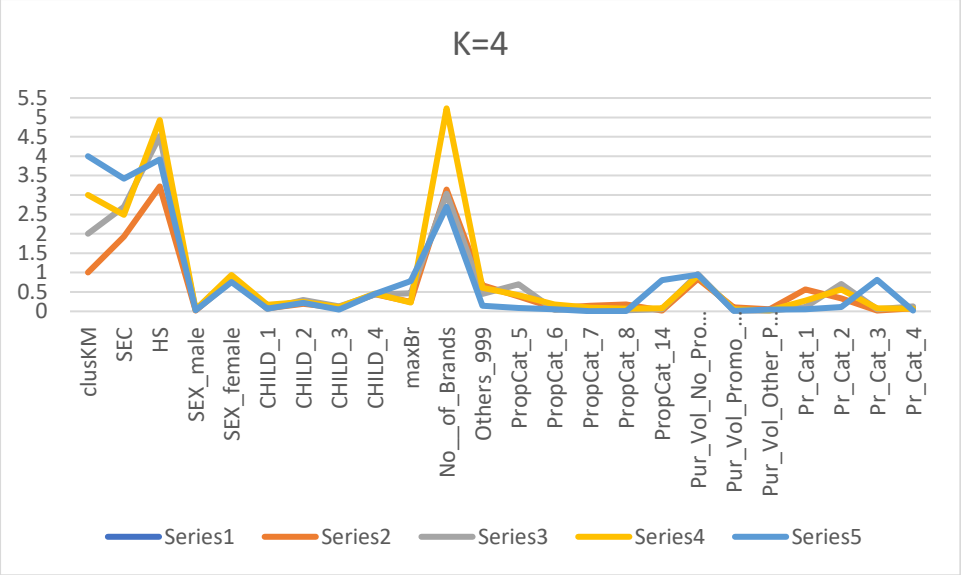


K=4



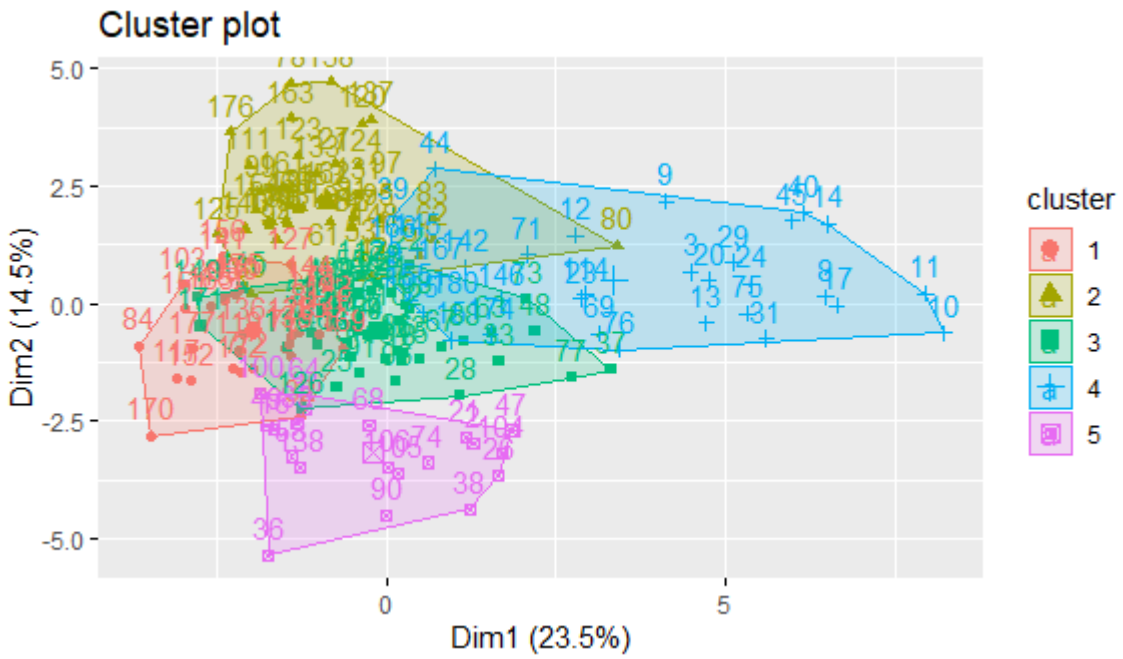
Clusters 1,2 and 3 are well separated from each other. Cluster 4 is the one that has overlapping with all other clusters.

The graph below helps us in recognizing the attributes that caused this overlapping.



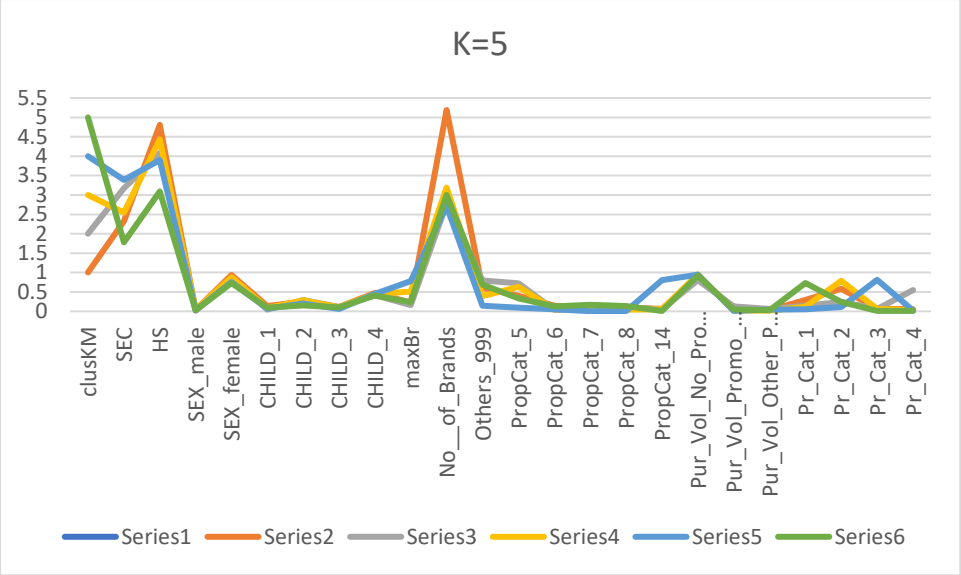
Cluster analysis: Purchase volume with no promotion is highest for both the clusters, suggests that promotions doesn't have much of impact of buying behaviour of customers in these clusters.

K=5



Clusters 1,3 and 5 are well separated from each other. Overall there is a bit of overlapping among clusters above.

The graph below gives us the significant variables that can help us in understanding the demographics of our clusters and plan accordingly.



Summary:

	Within Cluster	Between Cluster	Cluster Size
K=3	**10015**	**3163**	**298,73,229**
K=2	**11197**	**1981**	**72,528**
K=4	**9176**	**4002**	**163,171,69,197**
K=5	**8408**	**4770**	**176,62,69,182,113**

The cluster plots for various K values have overlapping among clusters, so it's not easy to decide which K value is best for our model.

The above table provides us information about distance within clusters and between clusters. From table we can see that K=5 has the lowest within distance and highest between distance.

So, K=5 is the best model among the K values we tried with.

3. Try two other clustering methods (for a single person team, try one other method) for the questions above - from agglomerative clustering, k-medoids, kernel-k-means, and DBSCAN clustering

We have applied the following two clustering techniques on the three segments in the given data set

1) Hierarchical Clustering
2) K-Medoids (PAM – partition around medoids)

Hierarchical Clustering:

Hierarchical clustering can be performed using either a distance matrix or raw matrix. Here we are using the distance matrix to plot the hierarchical cluster for the given dataset. Basically, in hierarchical clustering the process is performed in two steps :

- Identify the two clusters that are closest to each together
- Merge the two most similar clusters

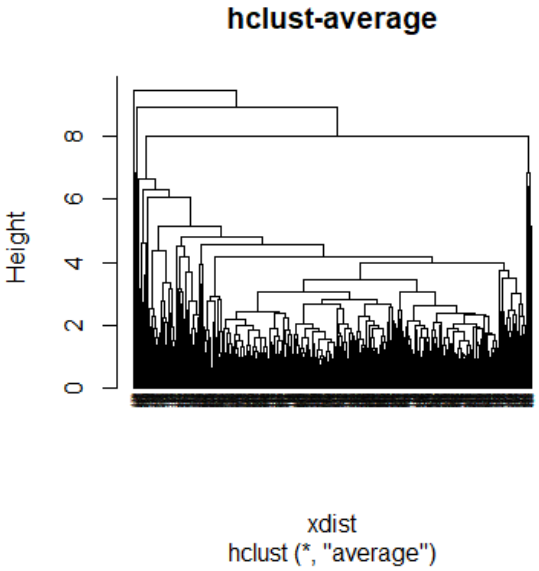
This process is done till all the clusters are merged together.

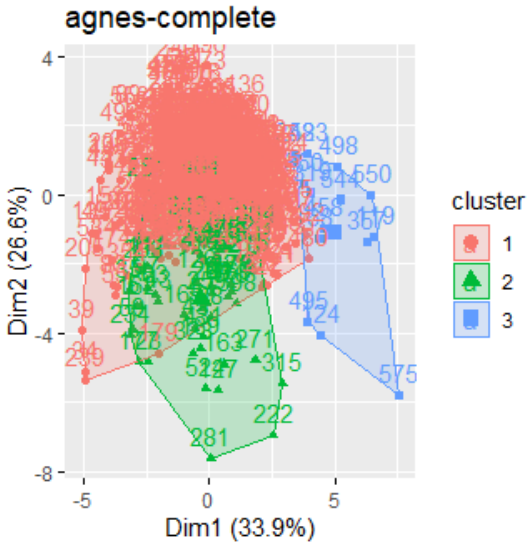
The distance metric here is calculated using the euclidean method between centres of the clusters (mean or average linkage). We are performing the agglomerative hierarchical clustering which typically works by sequentially merging similar clusters and obtaining the main end output dendrogram. The agglomerative hierarchical clustering module builds a cluster hierarchy which is commonly displayed as a tree diagram known as dendrogram.

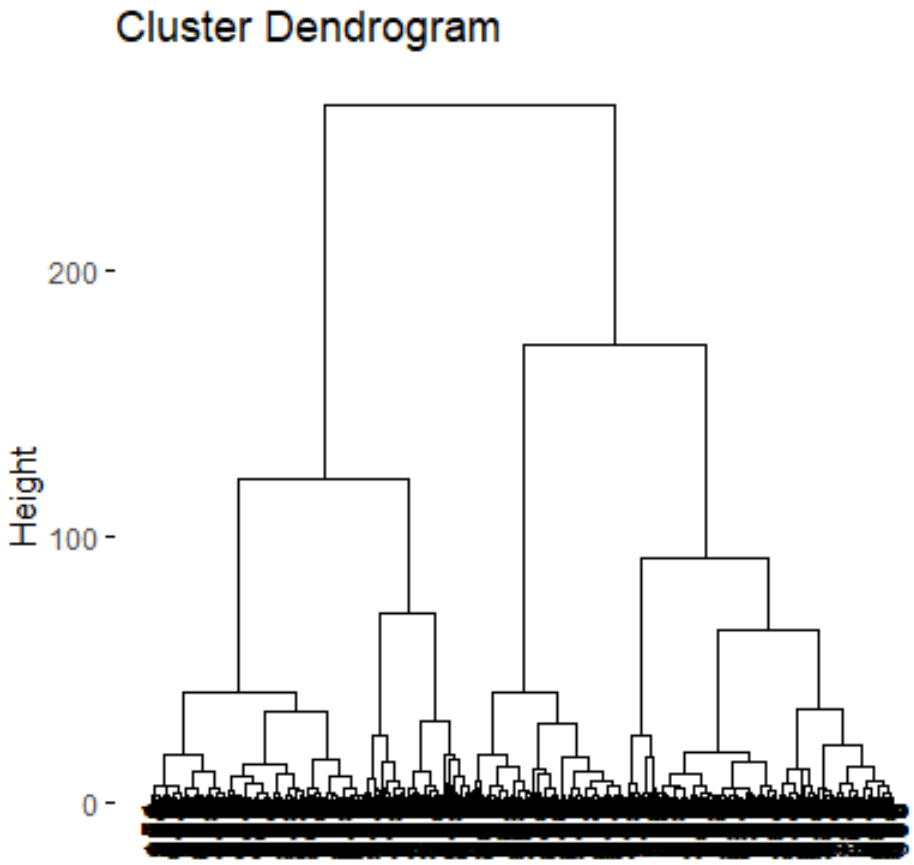
Dendrogram: In the given representation of the dendrogram the horizontal axis represents the distance between the clusters and the horizontal axis represents the objects and clusters. Each joining of two clusters is represented by splitting of a vertical line into two vertical lines. The vertical position of the split gives you the distance between two clusters.

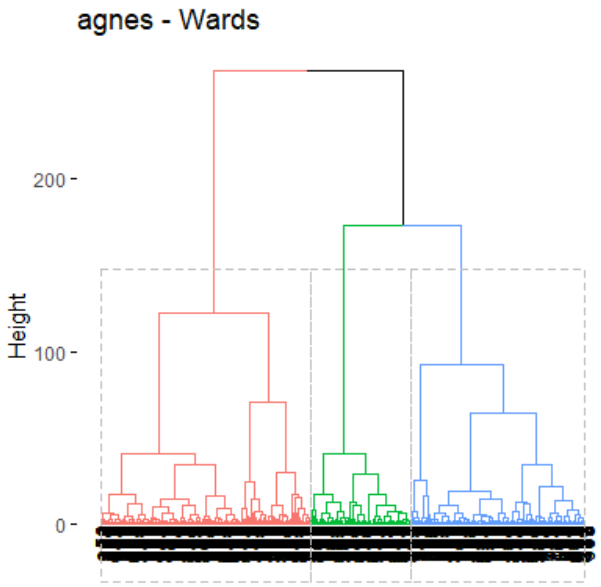
Circular dendrogram: In circular dendrogram representation, nodes of the dendrograms are radially distributed. In this the dimension of the representation is done by diameter than use of width. The radius will correspond to the extension of the drawing area. The 360-degree representation expresses a lap angle in which all the terminal nodes of the dendrogram will be distributed along the circumference.

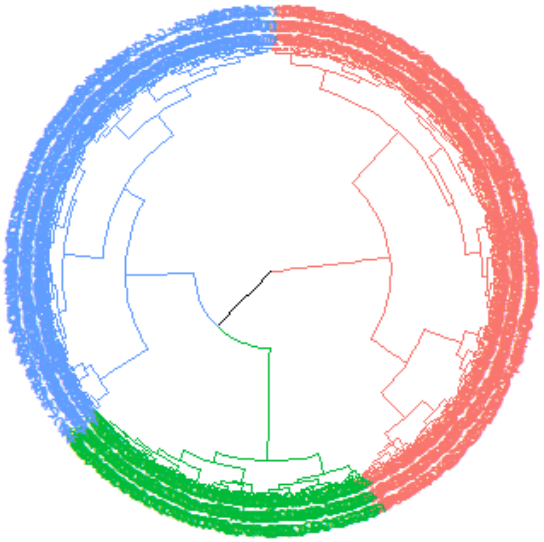
Purchase behaviour



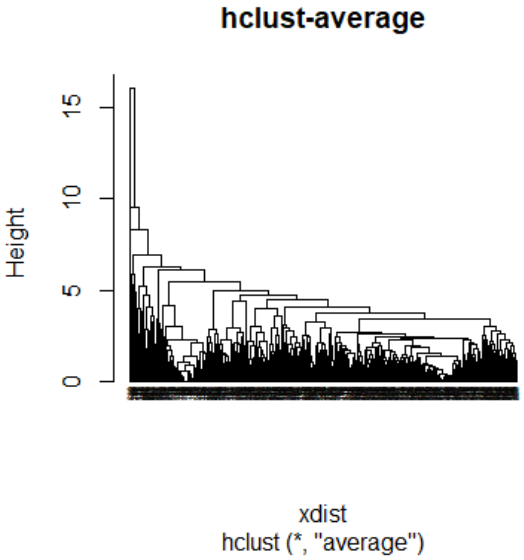


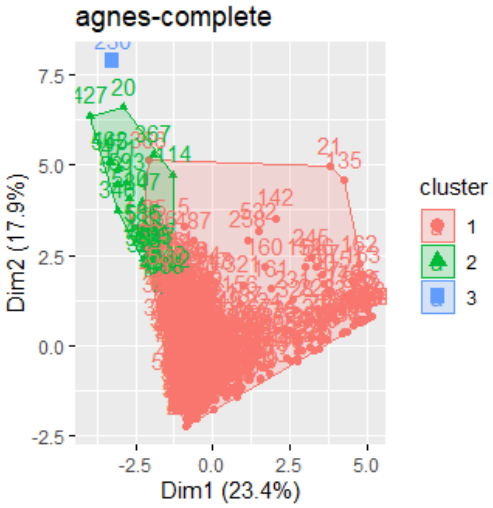


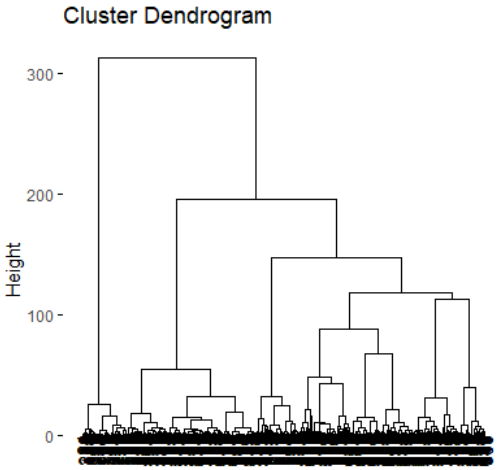


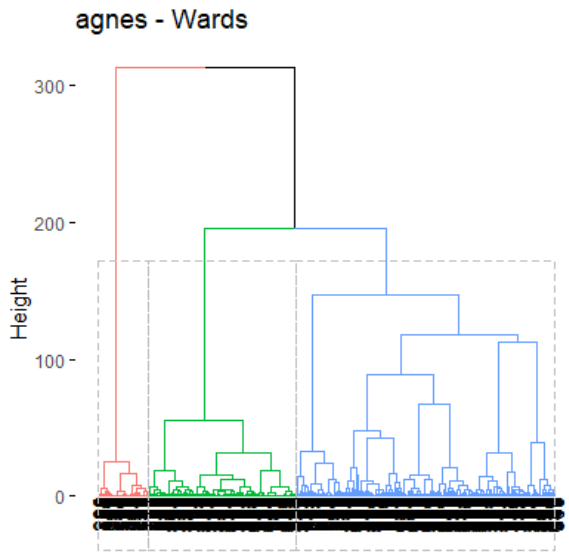


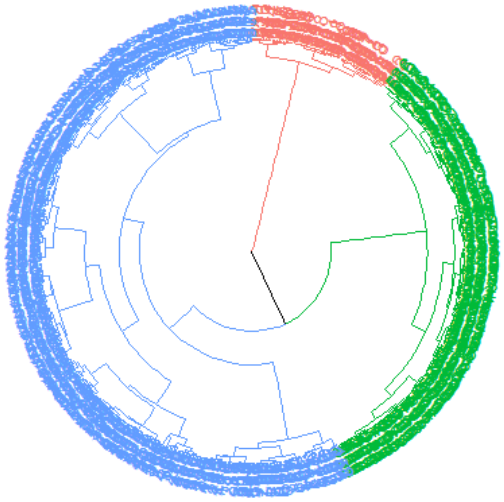
Basis for purchase



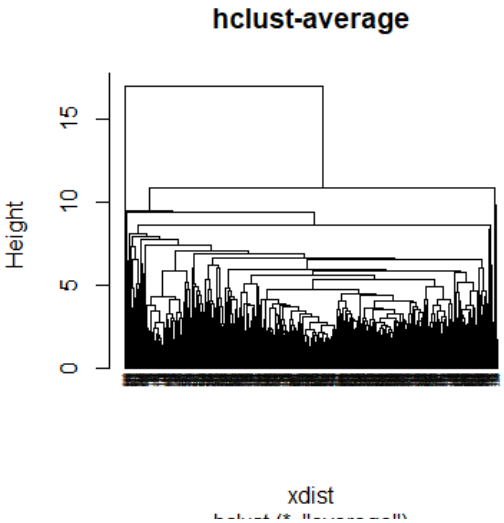


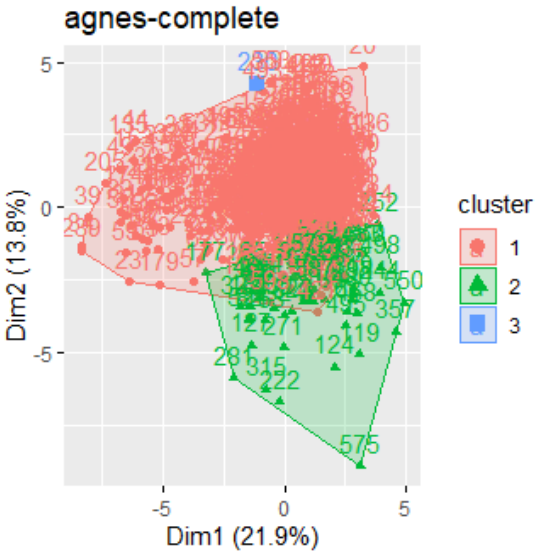


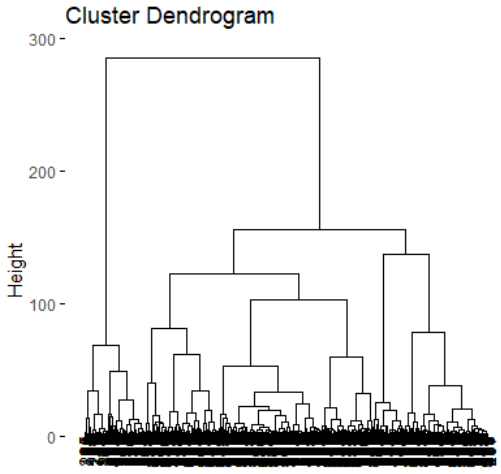


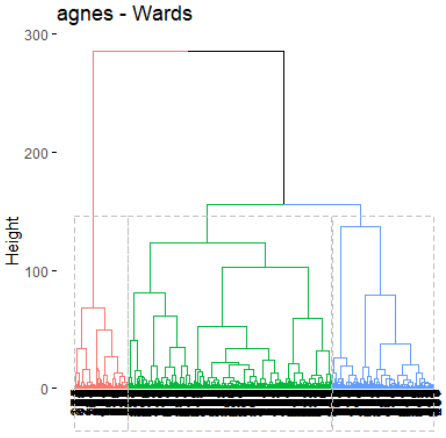


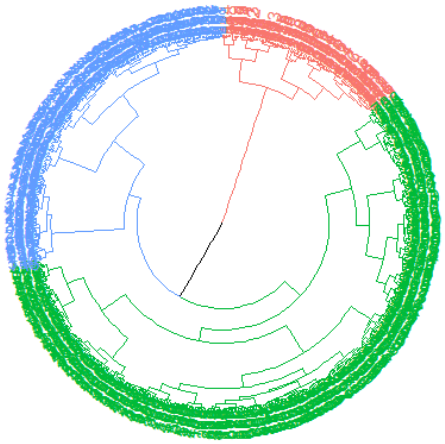
Purchase Behaviour and Basis for purchase











Summary:

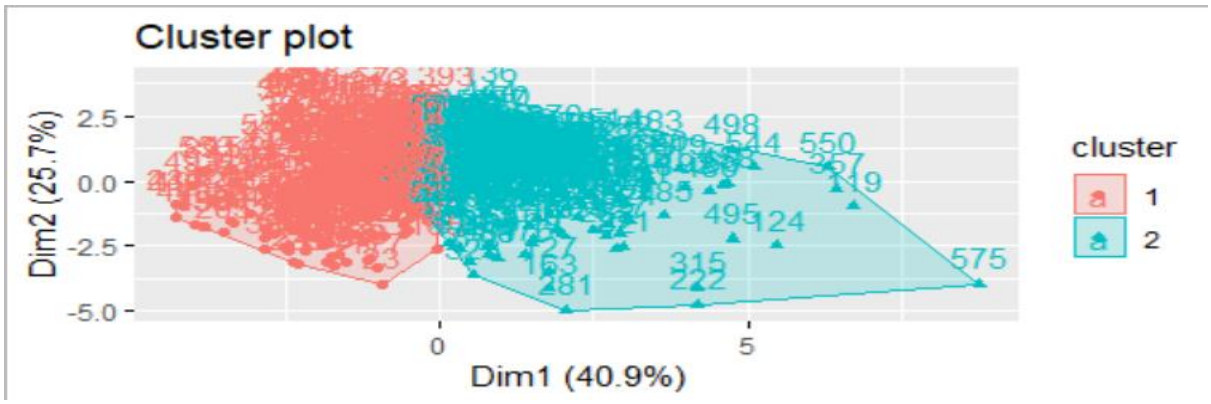
	Accuracy	Initial Cluster Split	Final Cluster Split
Purchase Behaviour	**0.93**	**533,53,14**	**215,260,125**
Basis for Purchase	**0.94**	**567,32,1**	**195,340,65**
Both	**0.89**	**540,59,1**	**170,342,88**

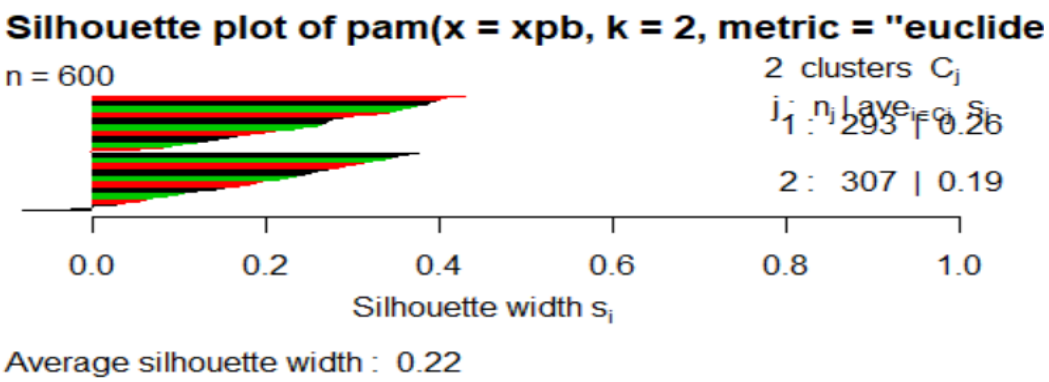
K-Medoids (PAM – Partition around medoids)

Purchase Behaviour

K=2

Euclidean

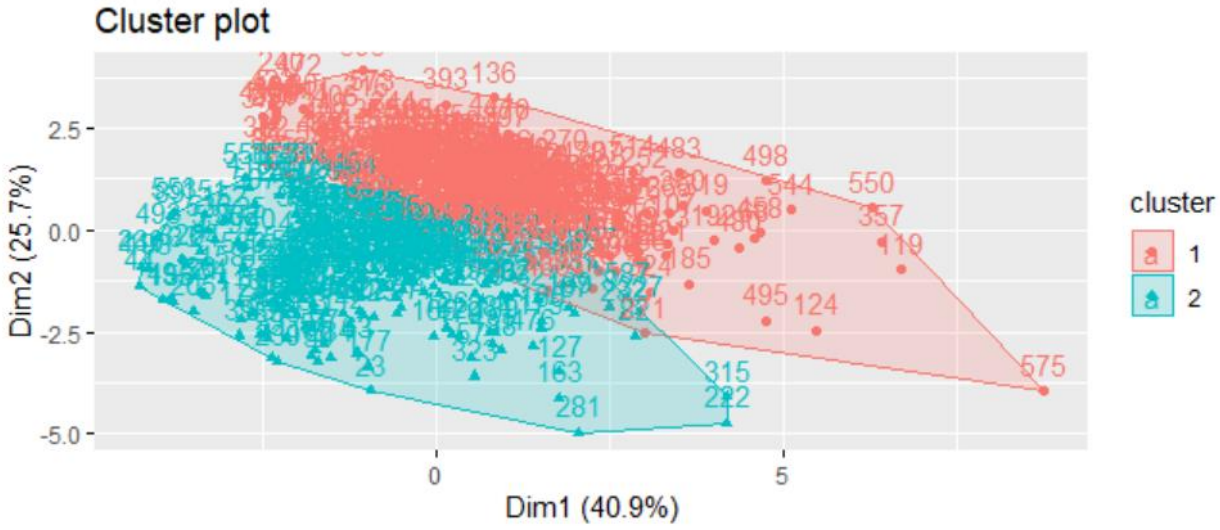


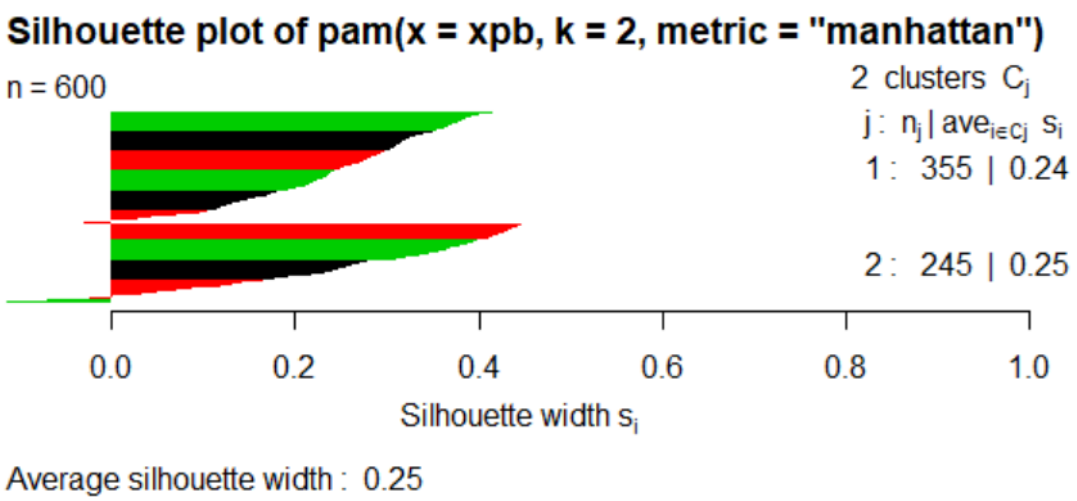


The Davies Bouldin Index for this cluster is 1.734

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	273	-0.4030277	-0.7456049	-0.6421531	-0.8121377	-0.6718935	-0.29146557	0.419804	-0.1528894
[2,]	477	0.863028	0.6971911	0.1525319	0.4502525	0.2231924	-0.03228384	-0.3914631	0.2148692

Manhattan



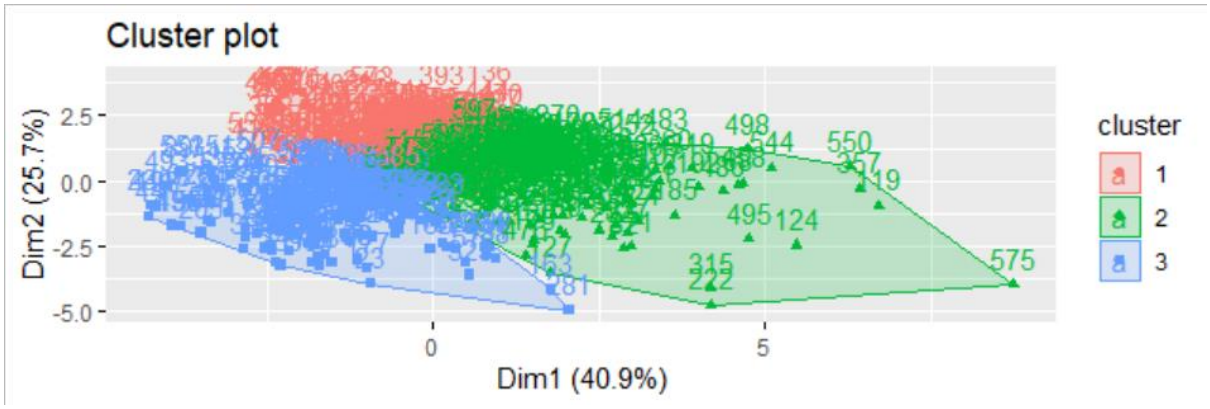


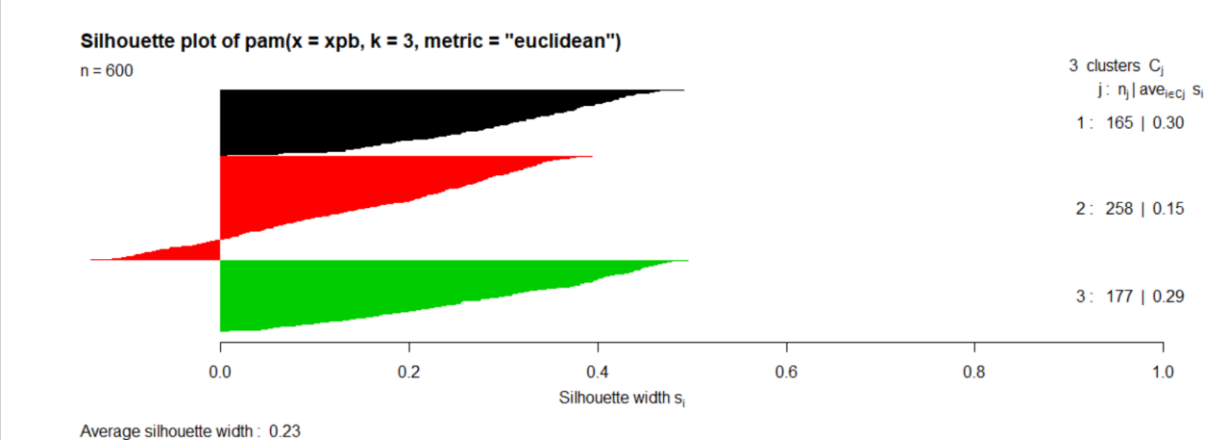
The Davies Bouldin Index for this cluster is 1.867

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	225	0.2300001	0.1200727	-0.4040693	-0.238324	-0.3786269	-0.1080408	-0.8618399	0.9273672
[2,]	268	-0.4030277	-0.7456049	-0.0147702	-0.3530867	-0.1640554	-0.4618607	0.9267515	-0.8720451

K=3

Euclidean

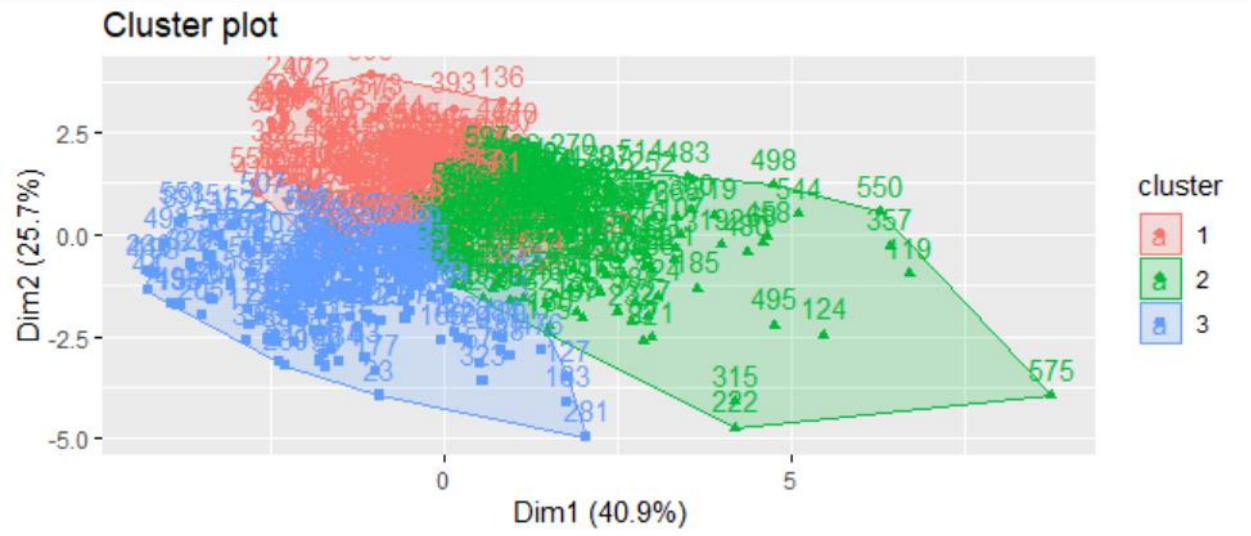


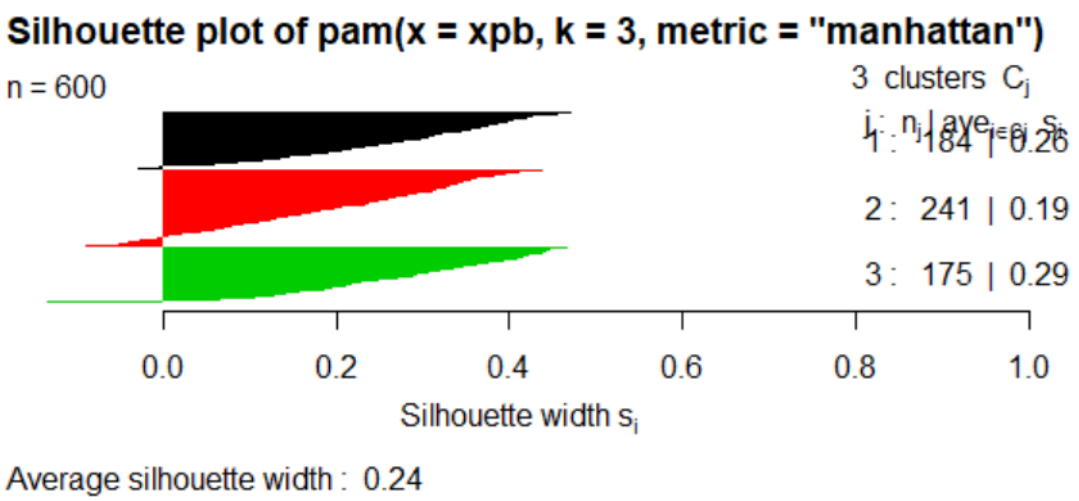


The Davies Bouldin Index for this cluster is 1.492

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	512	-0.4030277	-0.2646729	-0.696848	-0.8121377	-0.6905765	-0.1716011	-0.8159	0.7409946
[2,]	477	0.863028	0.6971911	0.1525319	0.4502525	0.2231924	-0.03228384	-0.3914631	0.2148692
[3,]	200	-0.4030277	-0.8417913	-0.1016386	-0.5252308	-0.1595262	-0.28842668	1.503065	-1.2190108

Manhattan



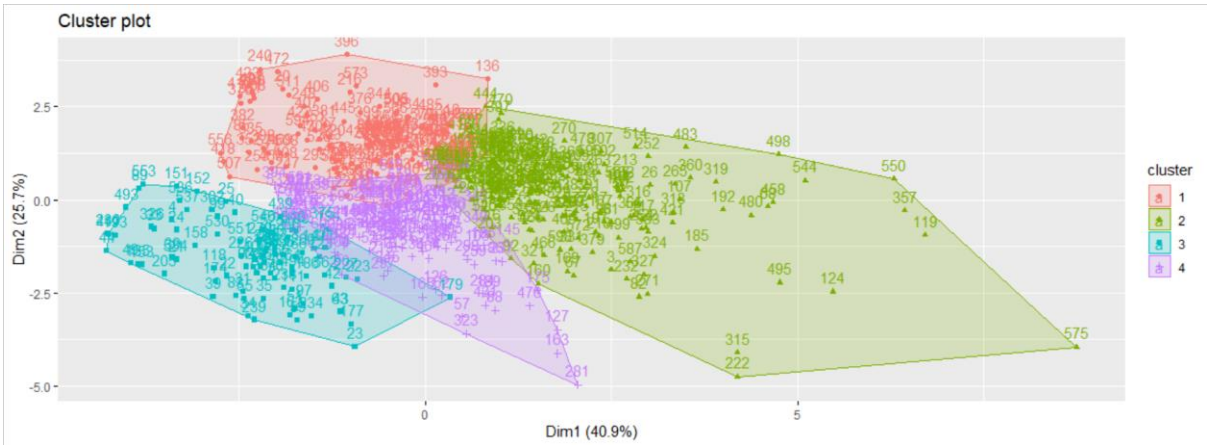


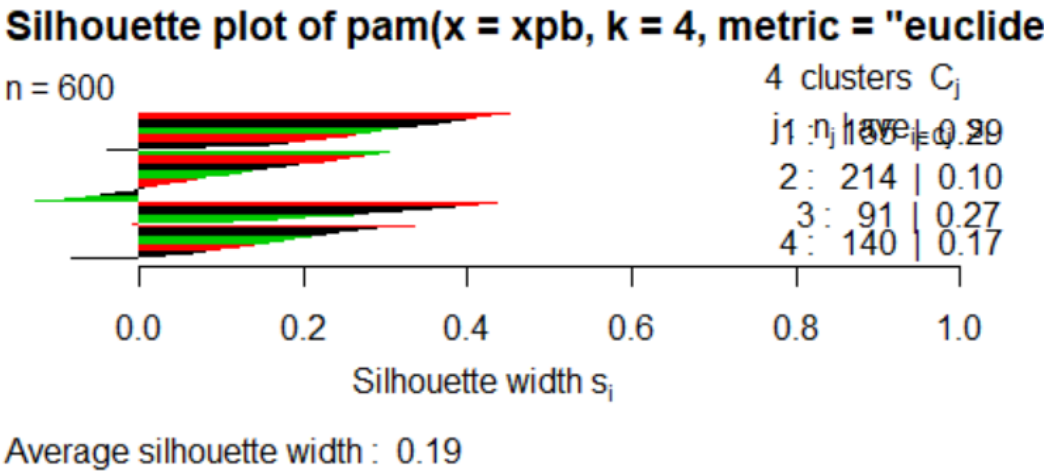
The Davies Bouldin Index for this cluster is 1.568

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	512	-0.4030277	-0.2646729	-0.696848	-0.8121377	-0.6905765	-0.1716011	-0.8159	0.7409946
[2,]	477	0.863028	0.6971911	0.1525319	0.4502525	0.2231924	-0.03228384	-0.3914631	0.2148692
[3,]	200	-0.4030277	-0.8417913	-0.1016386	-0.5252308	-0.1595262	-0.28842668	1.503065	-1.2190108

K=4

Euclidean

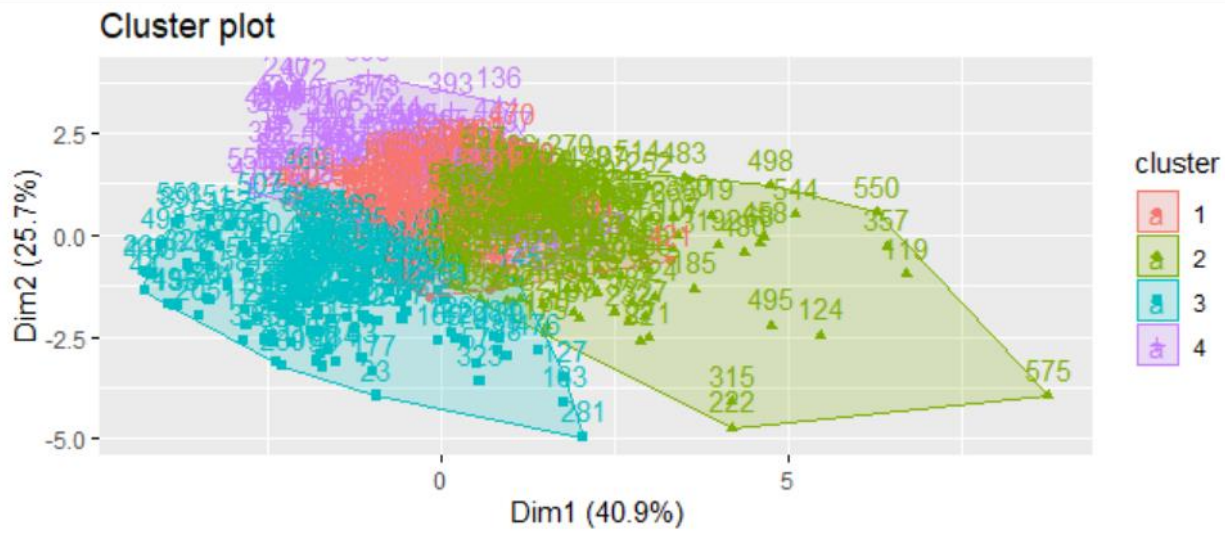


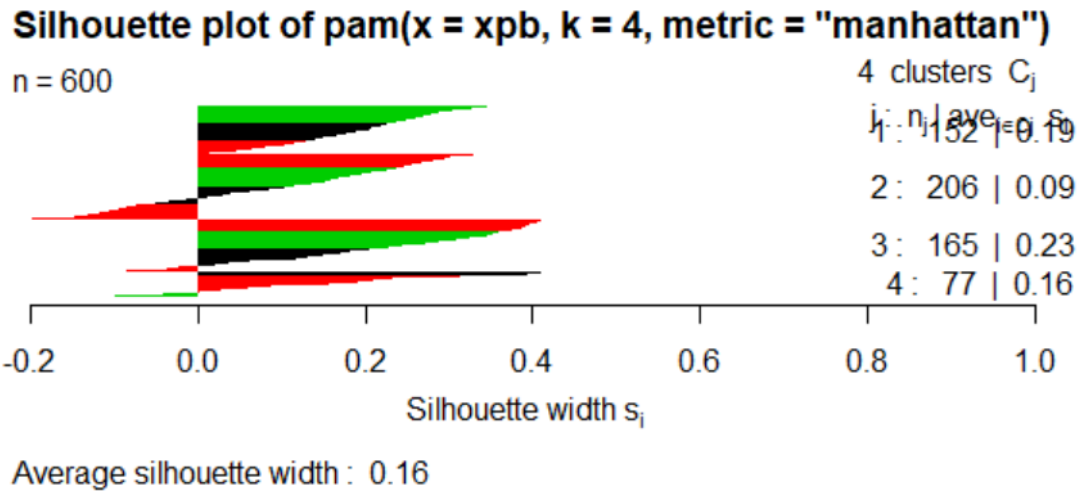


The Davies Bouldin Index for this cluster is 1.727

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	512	-0.4030277	-0.2646729	-0.69684801	-0.8121377	-0.69057653	-0.1716011	-0.8159	0.7409946
[2,]	2	0.863028	0.8895639	0.26513909	0.5076339	0.38964099	0.05280279	-0.7933249	0.5968948
[3,]	436	-1.0360556	-1.2265369	0.04957676	-0.6399936	-0.14027704	-0.52600526	1.8051574	-1.372814
[4,]	328	0.2300001	-0.3608593	-0.03407429	-0.1809426	0.01541469	-0.06359057	0.4589637	-0.9905826

Manhattan





The Davies Bouldin Index for this cluster is 2.034

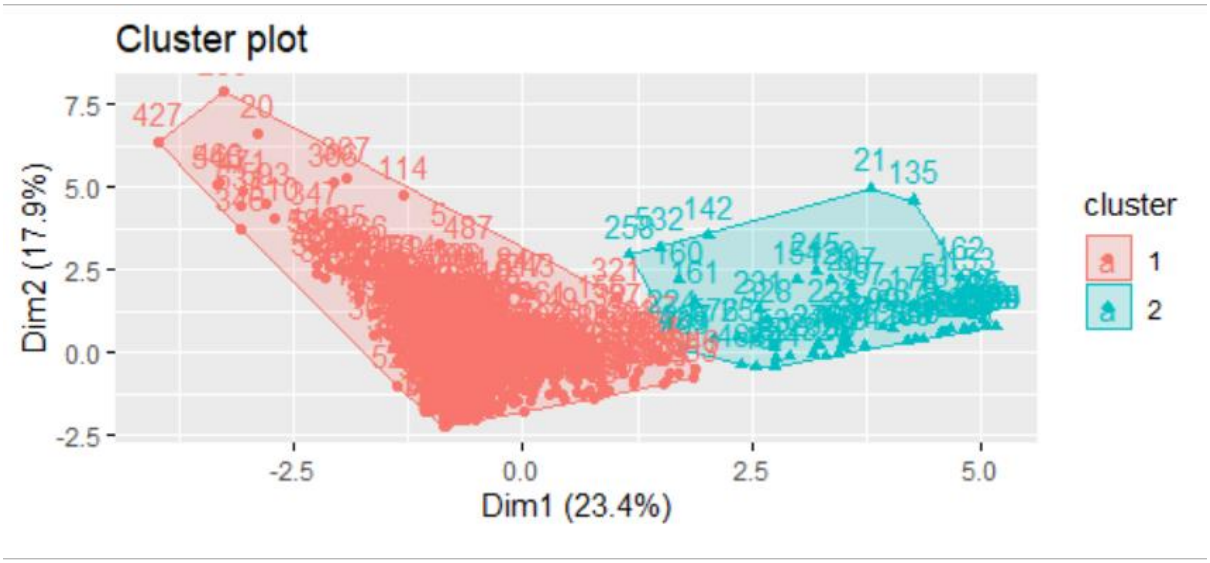
	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999
[1,]	503	-0.4030277	-0.07230011	-0.2689407	-0.238324	-0.3406948	-0.3433159	-0.3022669	0.4951468
[2,]	477	0.863028	0.69719106	0.1525319	0.4502525	0.2231924	-0.03228384	-0.3914631	0.2148692
[3,]	200	-0.4030277	-0.84179128	-0.1016386	-0.5252308	-0.1595262	-0.28842668	1.503065	-1.2190108
[4,]	445	-1.0360556	-1.13035047	-0.7193694	-0.8121377	-0.6600043	0.02519048	-1.2181324	1.5279364

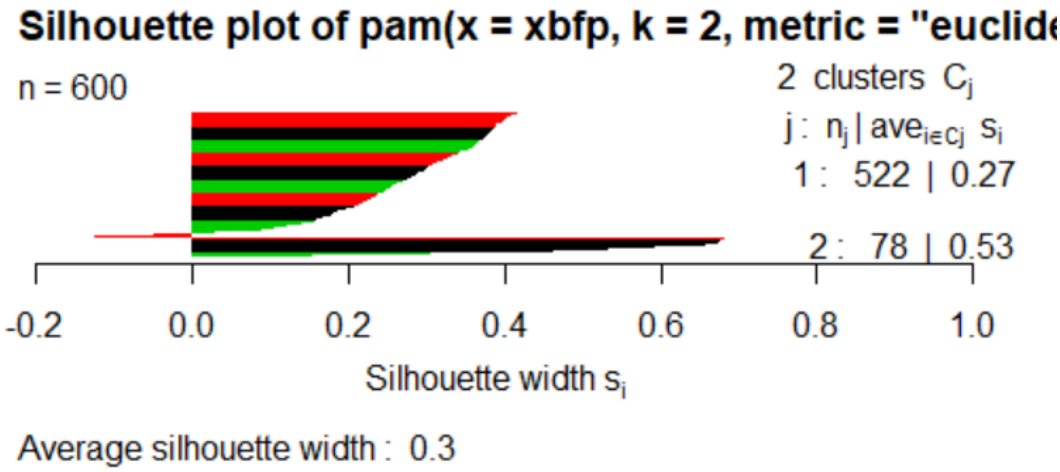
We have used the average silhouette method to evaluate the clusters. From the plots above, clusters with K=2 and distance measure = manhattan, provide the best average silhouette width of 0.16.

Basis of Purchase

K=2

Euclidean

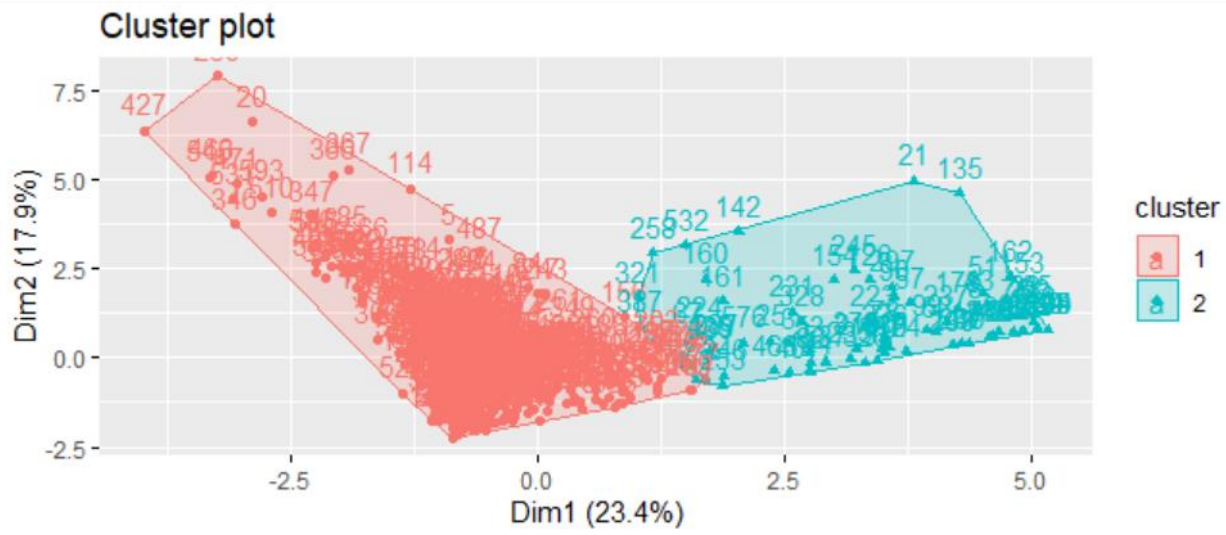


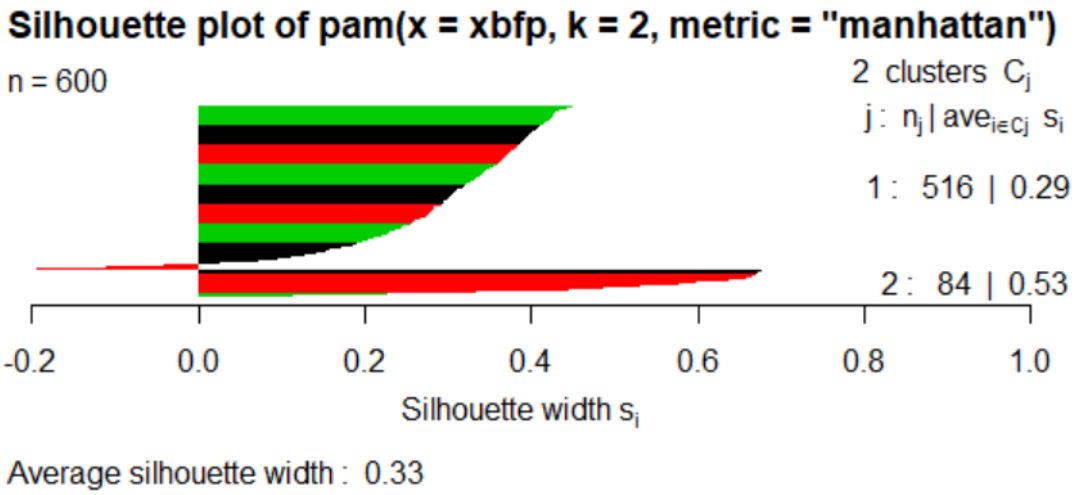


The Davies Bouldin Index for this cluster is 1.204

	ID	PropCat_5	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo___	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	575	0.2334012	-0.140224525	0.09752144	0.1235658	-0.3551302	0.2886584	-0.01075452	-0.46525557	-0.2119594	0.4697202	-0.3626554	0.0541638
2	62	-1.2272596	-0.005079551	-0.45663258	-0.4842747	2.4827055	0.4973048	-0.5753469	-0.08227662	-0.663273	-1.2448511	2.4541188	-0.4361558

Manhattan



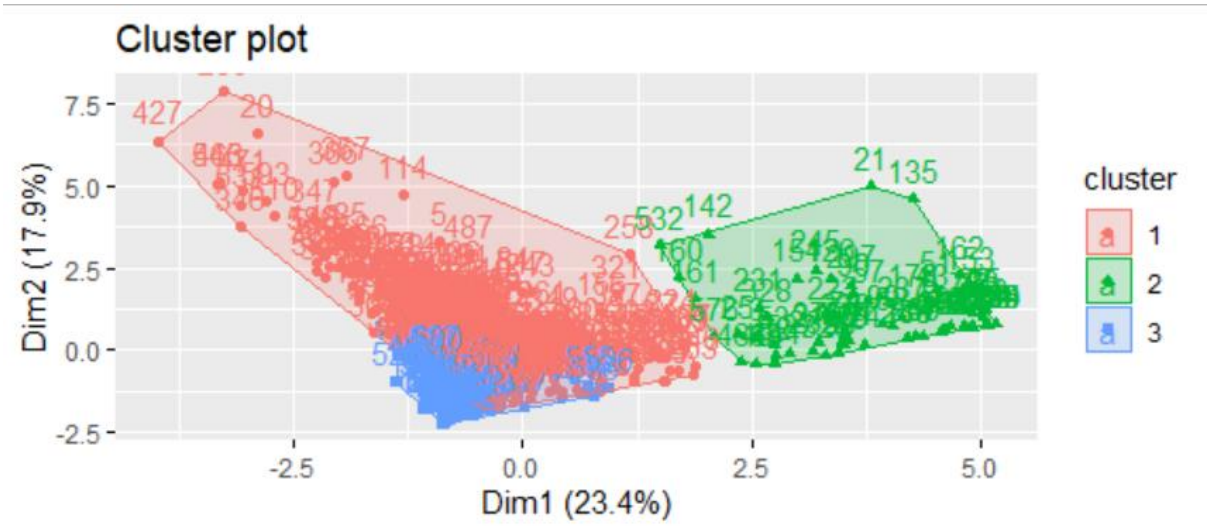


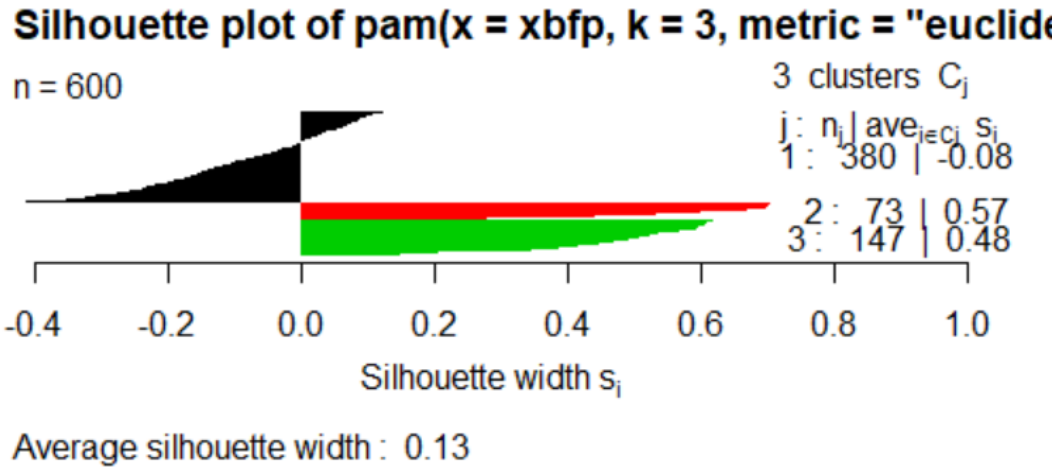
The Davies Bouldin Index for this cluster is 1.260

	ID	PropCat_5	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	151	0.6622533	-0.5550619	-0.4950408	-0.5253439	-0.5129981	0.7280271	-0.5753469	-0.4652556	0.1933042	0.5569085	-0.5193517	-0.4623055
[2,]	238	-0.9905019	-0.5145306	-0.4606029	-0.5253439	2.6291565	0.7280271	-0.5753469	-0.4652556	-0.9454101	-1.0994704	2.5994829	-0.4623055

K=3

Euclidean

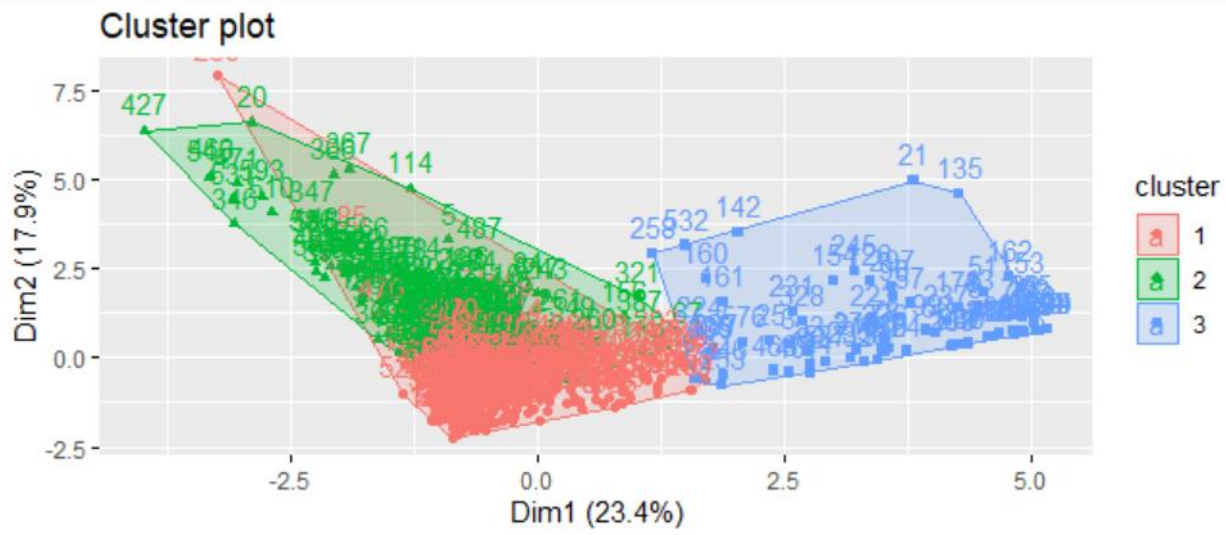


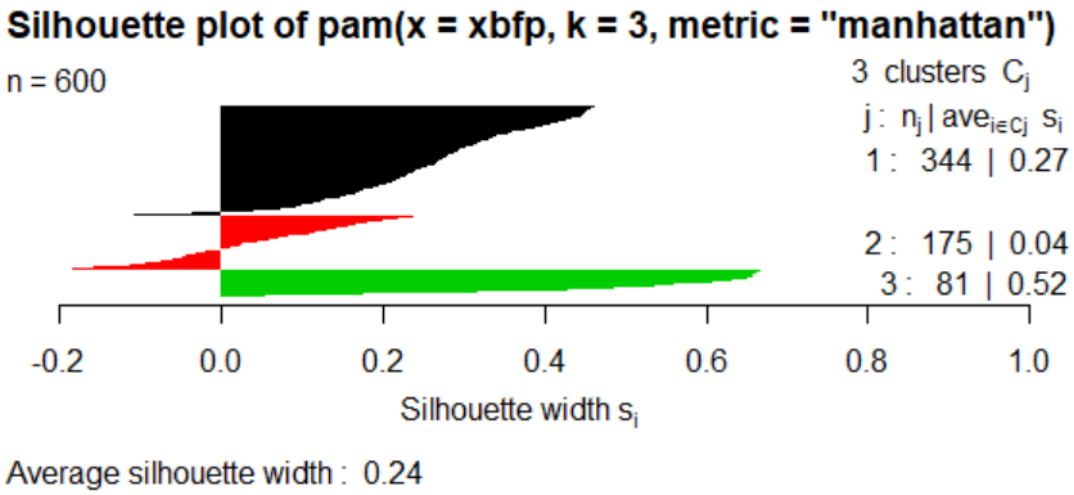


The Davies Bouldin Index for this cluster is 1.910

	ID	PropCat_5	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo___	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	443	-0.0712774	0.02302808	0.3497926	0.0040696	0.0074455	0.08426341	0.0037222	-0.1446781	-0.2813823	0.3429495	-0.0027706	-0.1412679
2	62	-1.2272595	-0.0050795	-0.4566326	-0.4842747	2.4827055	0.49730477	-0.5753469	-0.0822766	-0.6632731	-1.2448511	2.4541188	-0.4361558
3	28	1.2262004	-0.5550619	-0.4333713	-0.5253439	-0.5129981	0.67751043	-0.5104326	-0.4652556	-0.4418397	1.0003244	-0.5193517	-0.2523714

Manhattan

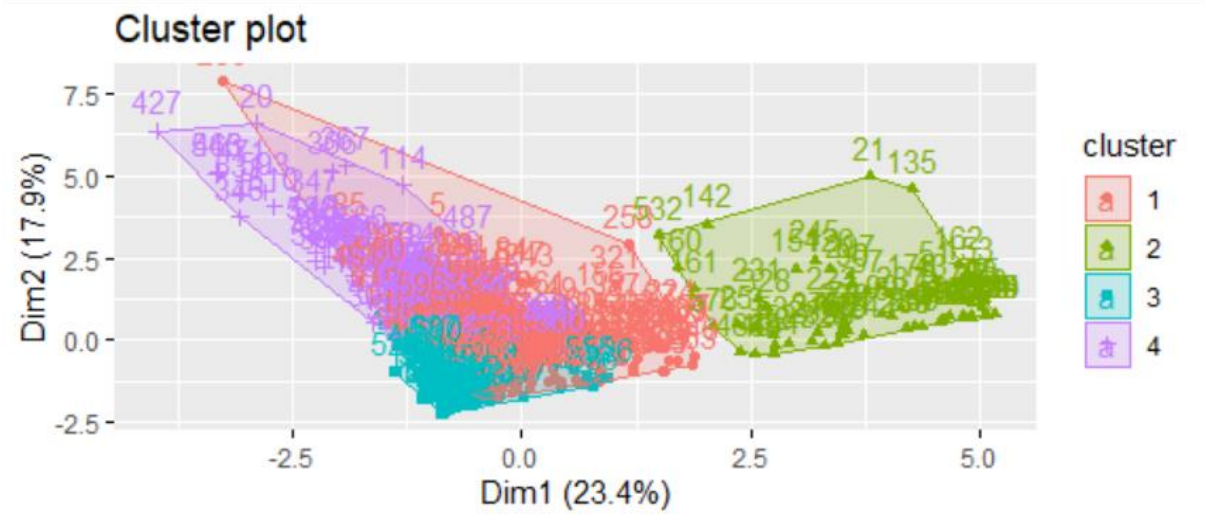


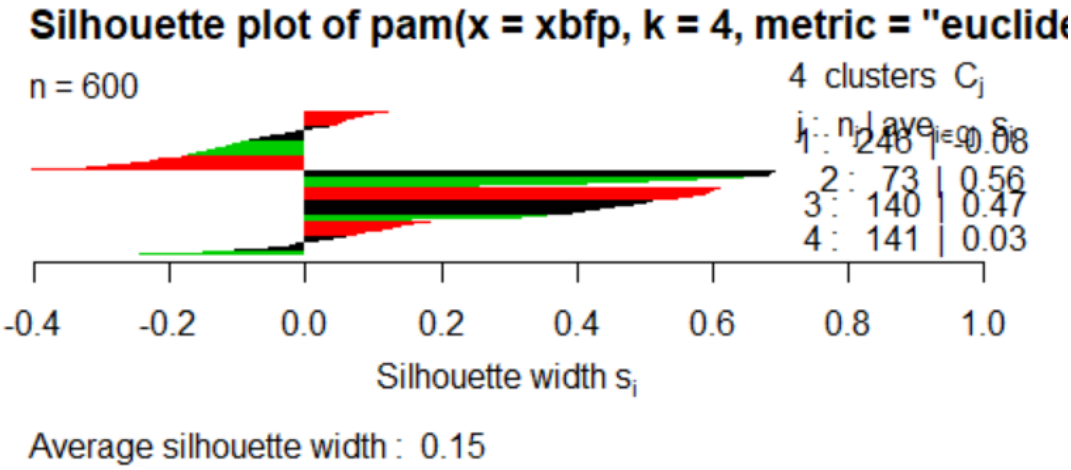


The Davies Bouldin Index for this cluster is 2.180

	ID	Prop Cat_5	Prop Cat_6	Prop Cat_7	Prop Cat_8	Prop Cat_14	Pur_Vol_No_Promo___	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	151	0.6622533	-0.5550619	-0.4950408	-0.5253439	-0.5129981	0.7280271	-0.5753469	-0.4652556	0.1933042	0.5569085	-0.5193517	-0.4623055
[2,]	409	-0.4468947	-0.1437052	-0.4950408	-0.0768683	-0.5129981	-0.5493356	1.0660746	-0.4652556	0.3182217	0.4442985	-0.5193517	-0.4623055
[3,]	238	-0.9905019	-0.5145306	-0.4606029	-0.5253439	2.6291565	0.7280271	-0.5753469	-0.4652556	-0.9454101	-1.0994704	2.5994829	-0.4623055

K=4

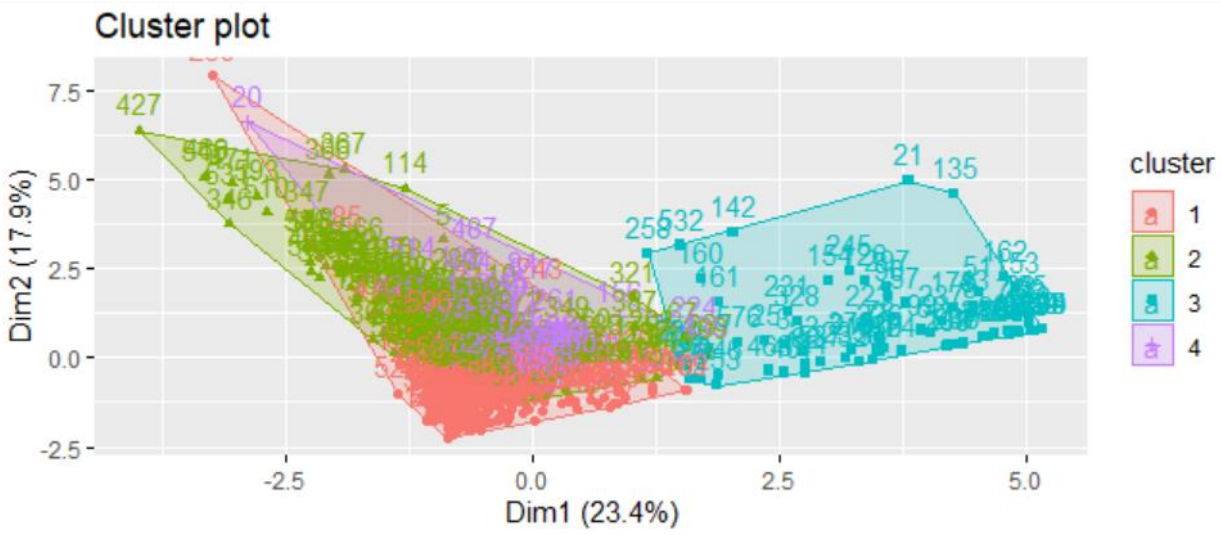


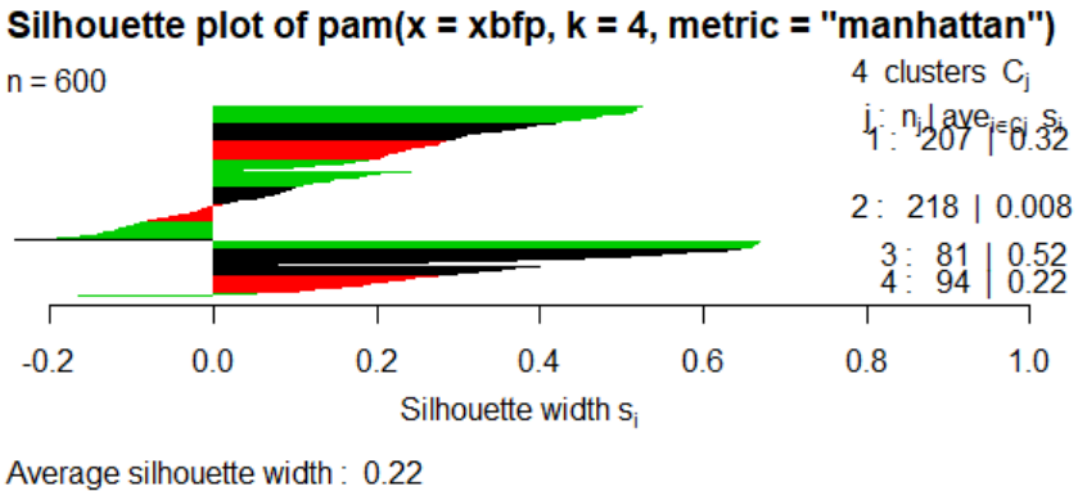


The Davies Bouldin Index for this cluster is 2.272

	ID	PropCat_5	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo__	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	443	-0.0712774	0.02302808	0.3497926	0.00406964	0.00744583	0.08426341	0.003722206	-0.14467809	-0.2813823	0.3429495	-0.00277059	-0.1412678
2	62	-1.2272595	-0.00507955 1	-0.456632 6	-0.48427468	2.48270548	0.49730477	-0.575346904	-0.08227662	-0.663273	-1.244851 1	2.45411879	-0.4361558 3
3	28	1.2262004	-0.555 06188 7	-0.43 33713	-0.525 34387 3	-0.512 99813 7	0.67751043	-0.510432 575	-0.46525557	-0.44 18397	1.00 03244	-0.519 35172	-0.252 37144
[4,]	563	-0.0824603 7	-0.464 65383 2	-0.39 26189	1.199 56225 7	-0.512 99813 7	-0.2787612 2	0.718384 298	-0.46525557	1.06 77268	-0.48 87556	-0.519 35172	-0.043 91058

Manhattan





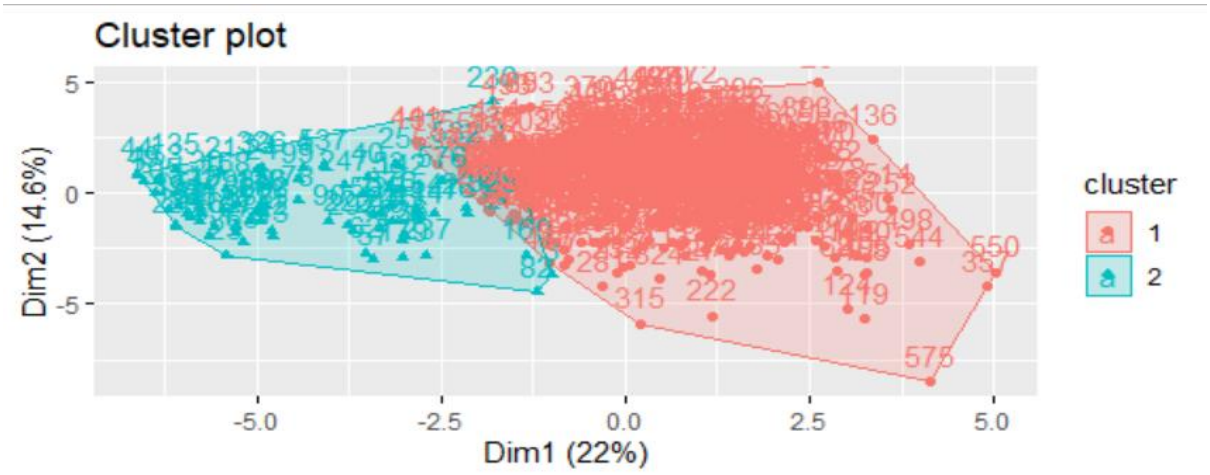
The Davies Bouldin Index for this cluster is 2.237

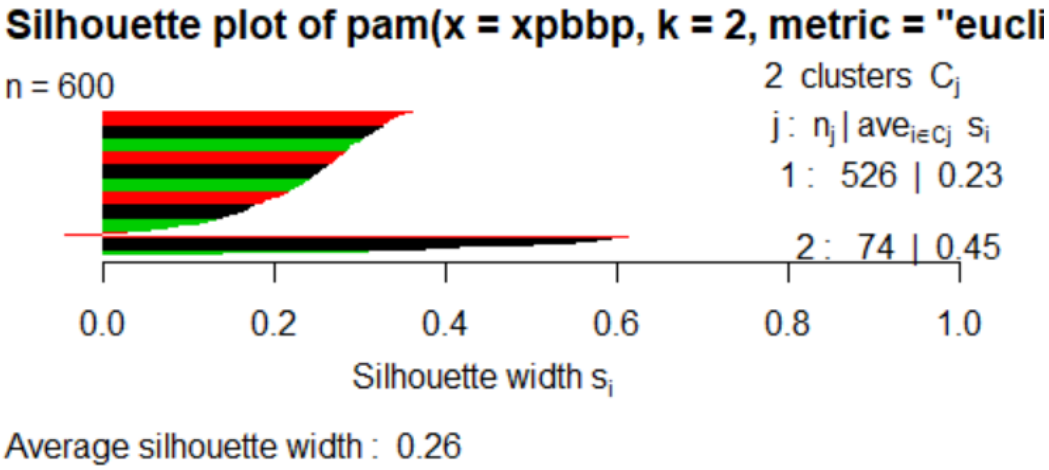
	ID	Prop Cat_5	Prop Cat_6	Prop Cat_7	Prop Cat_8	Prop Cat_14	Pur_Vol_No_Promo___	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	52	0.9841682	-0.5550619	-0.4950408	-0.5253439	-0.5129981	0.72802709	-0.5753469	-0.4652556	-0.7915769	1.4447539	-0.5193517	-0.4623055
[2,]	265	-0.4146103	-0.1350679	-0.4686072	0.1190971	-0.5129981	-0.09479117	0.3011225	-0.2316308	0.2132619	0.4890949	-0.461434	-0.4623055
[3,]	238	-0.9905019	-0.5145306	-0.4606029	-0.5253439	2.6291565	0.72802709	-0.5753469	-0.4652556	-0.9454101	-1.0994704	2.5994829	-0.4623055
[4,]	44	-1.0452333	-0.5550619	-0.4950408	-0.2884291	-0.5129981	0.72802709	-0.5753469	-0.4652556	1.9876765	-1.0606727	-0.5193517	-0.4623055

We have used the average silhouette method to evaluate the clusters. From the plots above, clusters with K=3 and distance measure = euclidean, provide the best average silhouette width of 0.13.

Purchase Behaviour and Basis of Purchase

K=2 (Euclidean)



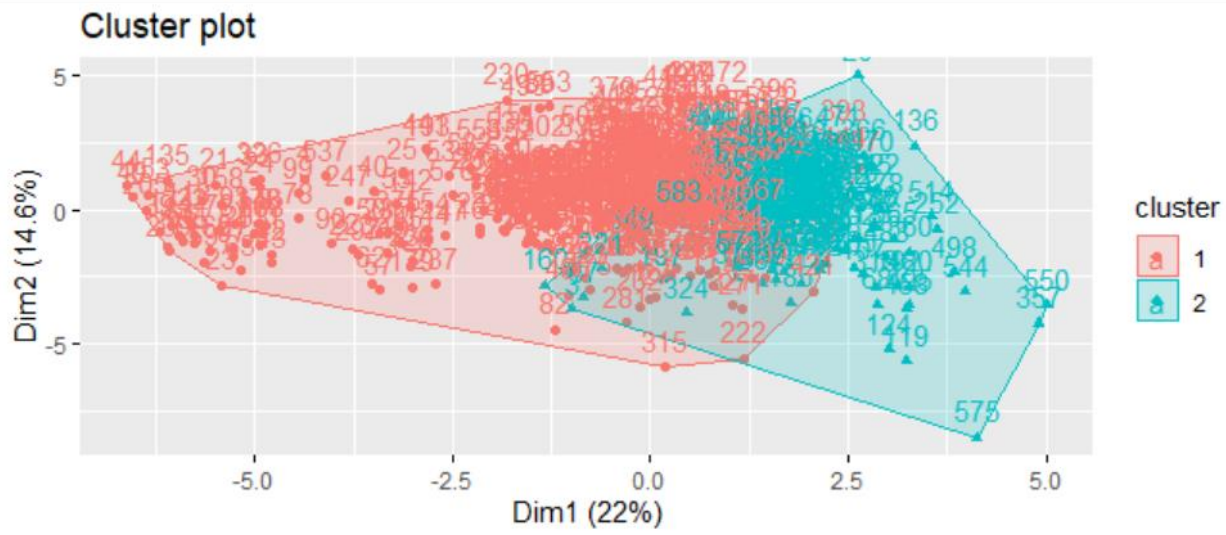


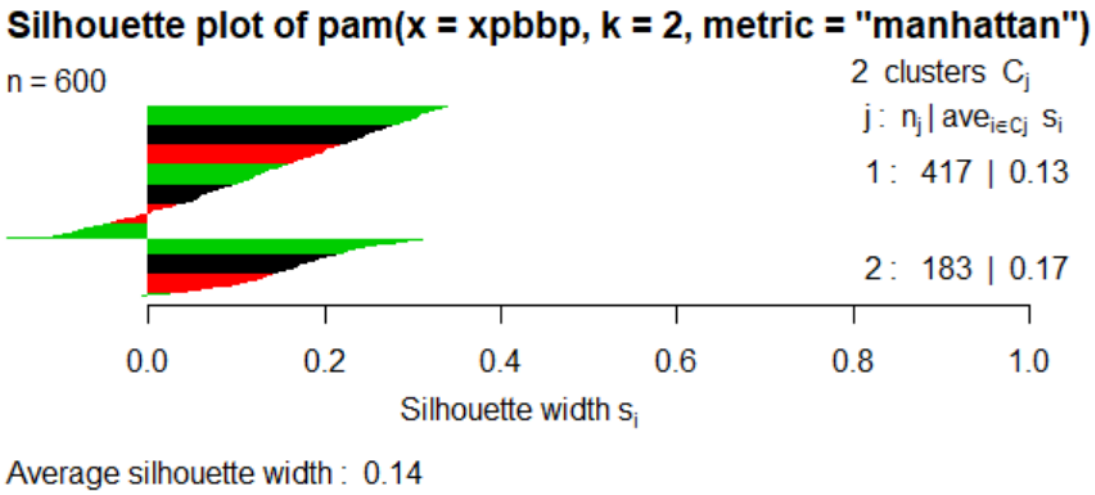
The Davies Bouldin Index for this cluster is 1.413

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5	PropCat_6
1	350	0.2300001	0.2162591	-0.2432019	-0.4104681	-0.0791326	0.2161018	0.2372703	-0.1620338	0.6675086	0.8392738
2	8	-0.4030277	-0.7456049	-0.336505	-0.3530867	-0.8694848	-1.5258518	1.4691417	-1.195077	-1.2411878	-0.5550619

	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	-0.4440852	-0.1820833	-0.5129981	0.2271436	0.06829243	-0.4652556	0.05421278	0.3621578	-0.5193517	0.0580797
2	-0.4950408	-0.5253439	2.8516985	0.1880963	-0.5753469	0.4309826	-0.8498577 2	-1.444643	2.8203733	-0.3501149

Manhattan



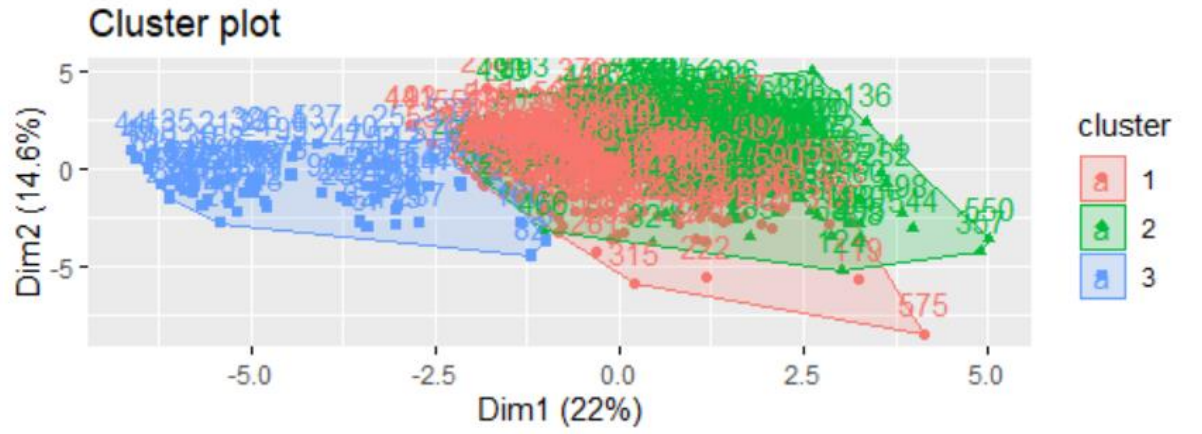


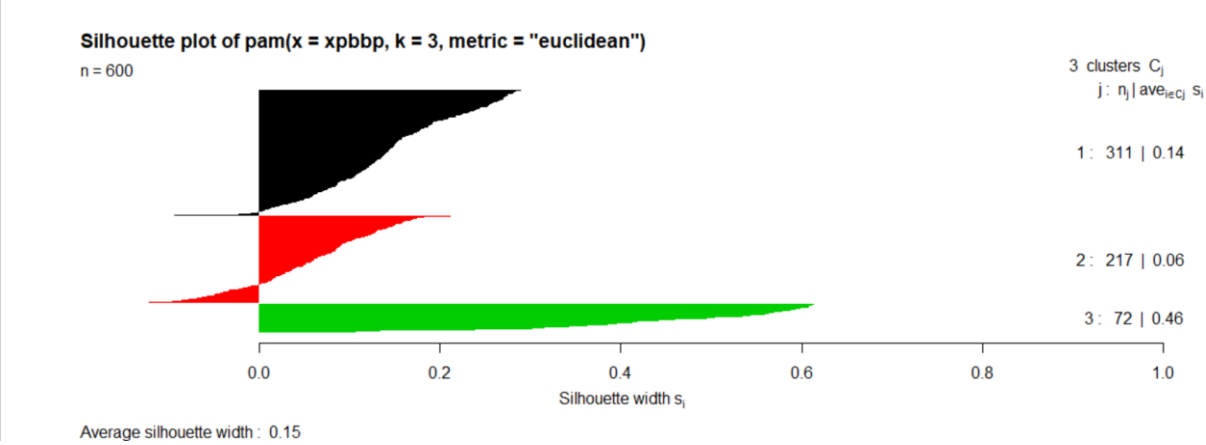
The Davies Bouldin Index for this cluster is 1.413

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5
[1,]	1	-0.4030277	0.1200727	-0.50058977	-0.4104681	-0.5881031	-0.4385586	0.02009	-0.1001607	0.1403318
[2,]	459	0.863028	0.7933775	0.02383798	1.0814476	0.2730137	0.3235257	-0.910165	0.7807644	-0.3870832

	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	-0.5550619	-0.4950408	-0.5253439	-0.0211974	0.7280271	-0.5753469	-0.46525557	-0.1616016	0.2169737	-0.0312013	-0.0722597
[2,]	0.1157136	-0.3050629	-0.5253439	-0.4508987	-0.9837991	1.2244163	0.05137756	0.8455078	-0.1371319	-0.4576845	-0.3760764

K=3 (Euclidean)



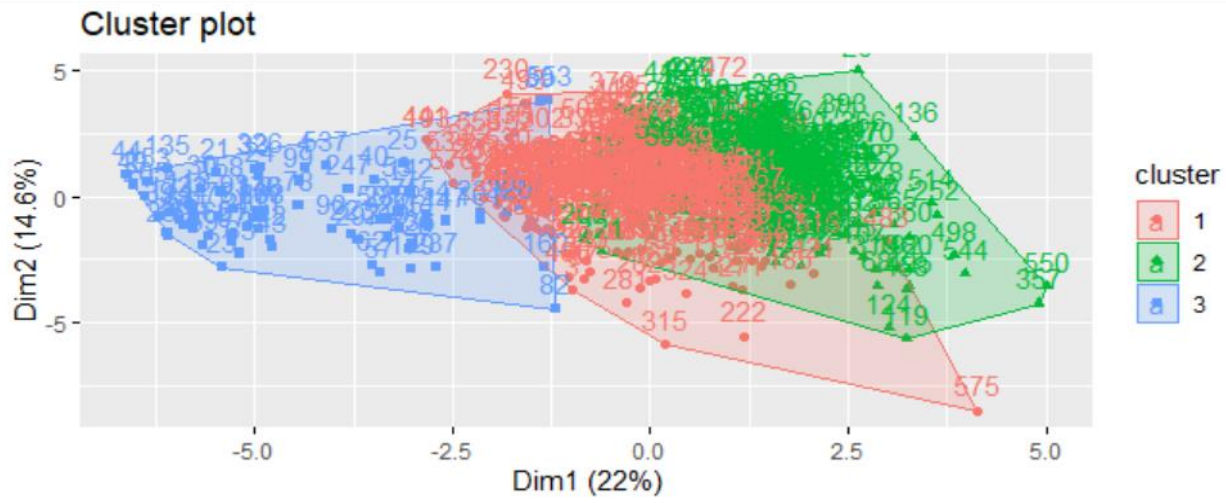


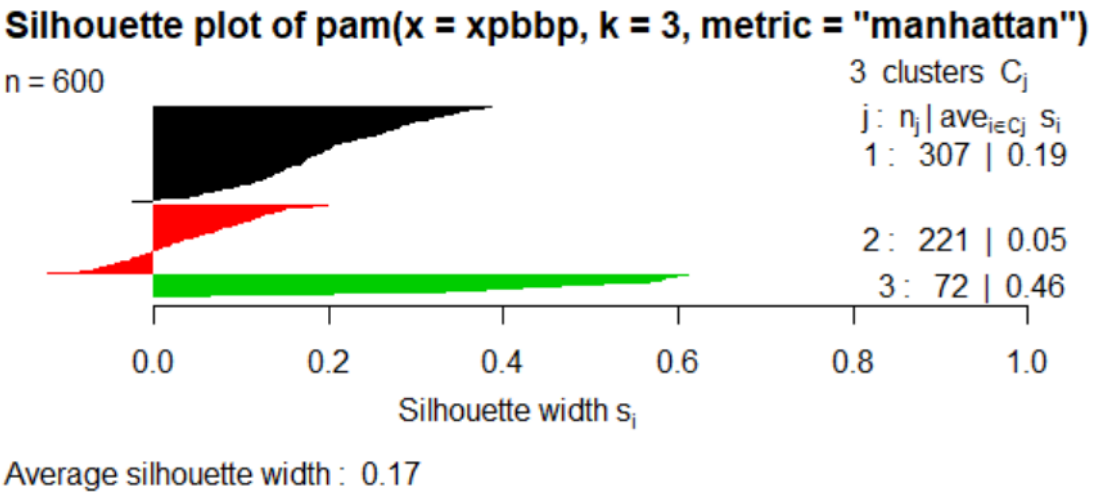
The Davies Bouldin Index for this cluster is 2.737

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5	PropCat_6
1	414	-0.4030277	-0.07230011	0.1557492	0.04858289	0.1184543	-0.2265446	0.2477936	-0.4423314	1.13189827	-0.5550619
2	305	0.2300001	0.12007268	-0.3236356	-0.5252308	0.1994144	1.1399284	-0.2384536	0.1049341	0.06815115	1.0758898
3	8	-0.4030277	-0.7456048	-0.336505	-0.3530867	-0.8694805	-1.5258518	1.4691417	-1.195077	-1.24118782	-0.5550619

	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	0.088764361	-0.4004935	-0.5129981	0.1063353	0.22353211	-0.4652556	-0.4034446	0.8503379	-0.5193517	-0.06483033
2	-0.005948712	0.4160145	-0.5129981	0.3273869	-0.06052102	-0.4652556	1.3358343	-0.4730535	-0.5193517	-0.46230552
3	-0.495040812	-0.5253439	2.8516985	0.1880963	-0.5753469	0.4309826	-0.8498577	-1.444643	2.8203733	-0.35011494

Manhattan



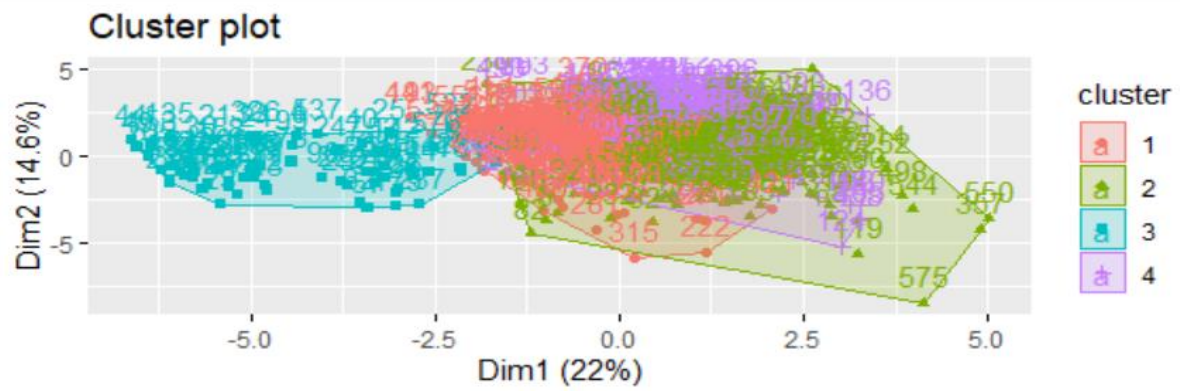


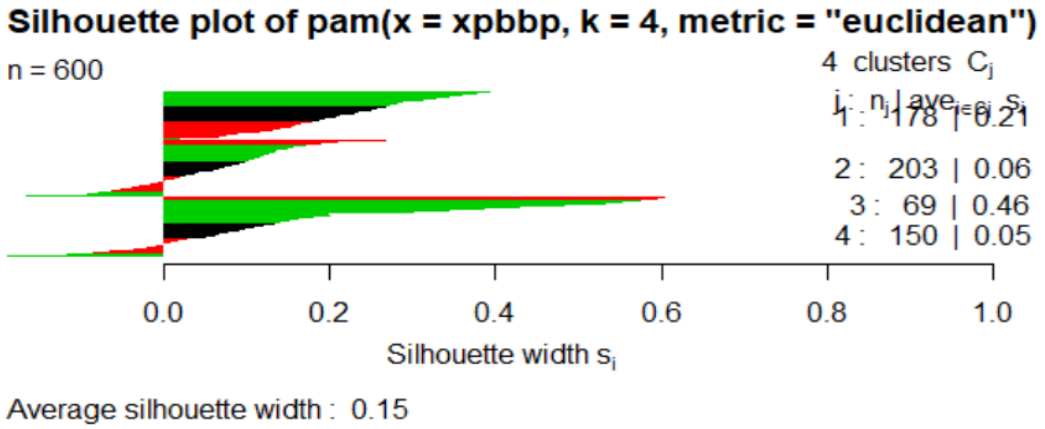
The Davies Bouldin Index for this cluster is 2.393

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5
[1,]	332	0.2300001	0.1200727	-0.2592887	-0.06617985	-0.3112549	-0.2945005	0.5045536	-0.6600015	0.989536
[2,]	437	0.2300001	0.2162591	0.1506015	-0.69737494	0.579868	0.6145	-0.7788686	0.8173099	-0.7566255
[3,]	233	-0.4030277	-1.1303505	0.5707872	-0.52523082	-0.2784181	-1.3783181	1.8182305	-1.4779259	-1.271124

	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	-0.5550 6189	-0.340 244	-0.5253 4387	-0.2282 436	0.7280271	-0.5753469	-0.4652556	-0.525 9178	0.913 5068	-0.236 7105	-0.383 2622
[2,]	0.0996 7968	-0.319 3639	0.0382 0016	-0.5129 981	-0.6150947	1.1505755	-0.4652556	0.900 2552	-0.227 5532	-0.519 3517	-0.223 0914
[3,]	-0.2241 1864	-0.495 0408	-0.5253 4387	2.8319 536	0.7280271	-0.5753469	-0.4652556	-0.797 4406	-1.406 0177	2.800 7749	-0.462 3055

K=4 (Euclidean)



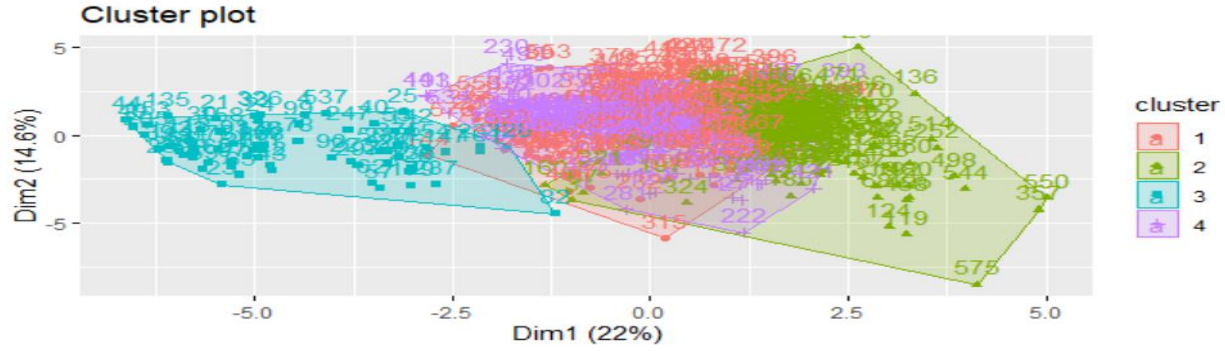


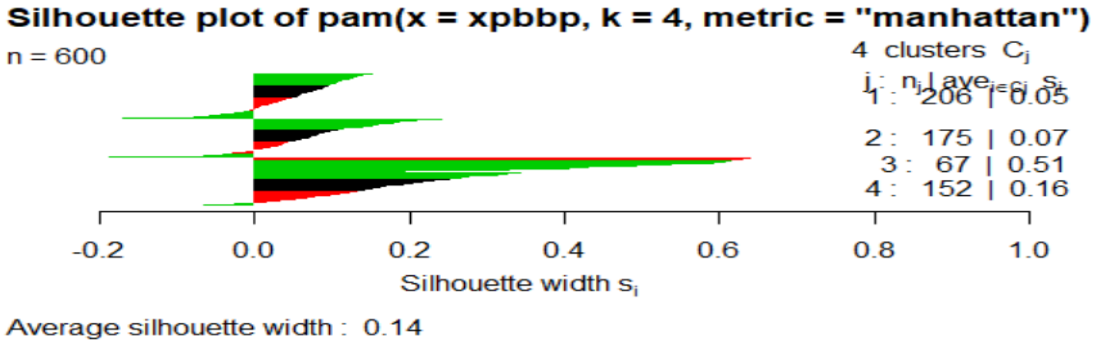
The Davies Bouldin Index for this cluster is 2.266

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5	PropCat_6
1	217	-0.4030277	-0.4570457	0.1525319	-0.2957053	-0.0344066	-0.4962824	0.4648208	-0.5489408	1.1247511	-0.3255911
2	338	0.863028	0.6971911	-0.3332877	0.5650152	-0.2433166	0.0542277	-0.2862929	0.2281021	0.7752496	-0.4583519
3	8	-0.4030277	-0.7456049	-0.336505	-0.3530867	-0.8694804	-1.5258517	1.4691417	-1.1950777	-1.2411878	-0.5550619
4	305	0.2300001	0.1200727	-0.3236356	-0.5252308	0.1994144	1.1399283	-0.2384536	0.1049341	0.0681511	1.0758898

	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
1	-0.4950408	-0.5253439	0.0034727	0.7280271	-0.5753469	-0.4652556	-0.952647	1.1489704	-0.0067139	-0.4623055
2	-0.4128698	-0.1035974	-0.5129981	-0.6630548	0.750901	0.1306347	0.5146403	0.0607302	-0.5193517	-0.1266361
3	-0.4950408	-0.5253439	2.8516984	0.1880963	-0.5753469	0.4309826	-0.8498577	-1.444643	2.8203732	-0.3501149
4	-0.0059487	0.4160145	-0.5129981	0.3273869	-0.0605210	-0.4652556	1.3358343	-0.4730534	-0.5193517	-0.4623055

Manhattan





The Davies Bouldin Index for this cluster is 2.846

	ID	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Avg__Price	maxBr	Others_999	PropCat_5
[1,]	1	-0.4030277	0.1200727	-0.5005898	-0.4104681	-0.5881031	-0.43855865	0.02009	-0.1001607	0.1403318
[2,]	459	0.863028	0.7933775	0.023838	1.0814476	0.2730137	0.3235257	-0.910165	0.7807644	-0.3870832
[3,]	233	-0.4030277	-1.1303505	0.5707872	-0.5252308	-0.2784181	-1.37831807	1.81823	-1.4779259	-1.271124
[4,]	211	-0.4030277	-0.4570457	0.0527941	-0.238324	0.1263804	-0.02083462	1.038047	-0.7186619	1.5620581

	PropCat_6	PropCat_7	PropCat_8	PropCat_14	Pur_Vol_No_Promo____	Pur_Vol_Promo_6__	Pur_Vol_Other_Promo__	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4
[1,]	-0.5550619	-0.4950408	-0.5253439	-0.0211977	0.7280271	-0.5753469	-0.46525557	-0.1616016	0.2169737	-0.0312013	-0.0722598
[2,]	0.1157136	-0.3050629	-0.5253439	-0.4508699	-0.9837991	1.2244163	0.05137756	0.8455078	-0.1371319	-0.4576845	-0.3760764
[3,]	-0.2241186	-0.4950408	-0.5253439	2.8319536	0.7280271	-0.5753469	-0.46525557	-0.7974406	-1.4060177	2.8007749	-0.4623055
[4,]	-0.4087219	-0.4950408	-0.5253439	-0.5129981	0.660125	-0.5753469	-0.35254408	-0.4951357	1.1775197	-0.5193517	-0.4623055

We have used the average silhouette method to evaluate the clusters. From the plots above, clusters with K=2 and k=4 and distance measure = manhattan, provide the best average silhouette width of 0.14, as per the standard practice we shall go ahead with K=2.

4. (a) Are the clusters obtained from the different procedures similar/different? Describe how they are similar/different.

Purchase Behaviour	Cluster Size
K=3	166,175,259
K=2	283,317
K=4	46,175,188,191
K=5	29,166,182,179,44

Basis for purchase	Cluster Size
K=3	76,326,198
K=8	91,40,50,58,234,10,71,46
K=4	320,75,127,78
K=5	128,75,50,297,50

Both	Cluster Size
K=3	298,73,229
K=2	72,528
K=4	163,171,69,197
K=5	176,62,69,182,113

Hierarchical	Initial Cluster Split	Final Cluster Split
Purchase Behaviour	533,53,14	215,260,125
Basis for Purchase	567,32,1	195,340,65
Both	540,59,1	170,342,88

the cluster sizes for all the models performed above are different from each other. This difference is maintained even with different K values as well as using hierarchical clustering.

From these values we can say that the clusters obtained from different procedures are different from each other.

(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 behaviour, or basis for purchase,....(think about how different clusters may be useful.

Purchase Behaviour	Within Cluster	Between Cluster
K=3	3970	2020
K=2	4754	1236
K=4	3428	2562
K=5	3037	2953

Basis for Purchase	Within Cluster	Between Cluster
K=3	5029	2159
K=8	2782	4406
K=4	4364	2824
K=5	3813	3375
Both	Within Cluster	Between Cluster
K=3	10015	3163
K=2	11197	1981
K=4	9176	4002
K=5	8408	4770

With the information obtained from Q2 and Q3, we can observe that within cluster distance is lowest for K=5 and between cluster distance is highest for K=5.

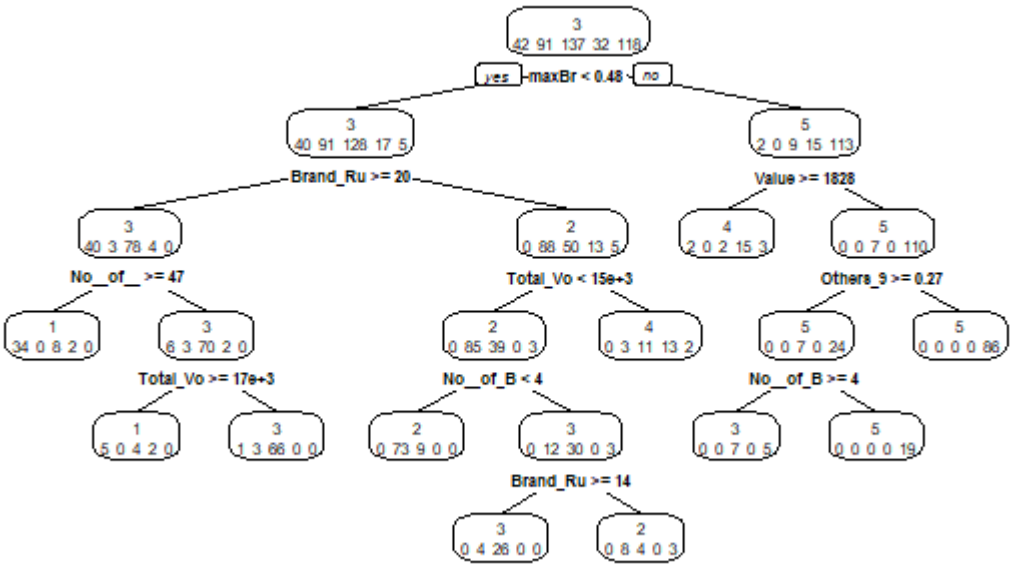
From this information we can conclude that **K=5 is the best model**

(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/interpret the cluster(s)? (explain why/why not). (You may use a decision tree to help choose the 'best' clustering).

The best segmentation that we are taking is for the clustering obtained on purchase behaviour using kernel k-mean method for **K value 5**. Decision tree is helpful in making clustering interpretable as its interpretation is a critical and non-trivial task for the end- user. Decision trees use a form that is intuitive and easy to understand. A decision tree with a cluster as its target variable is used in order to explain why an element is assigned to the cluster.

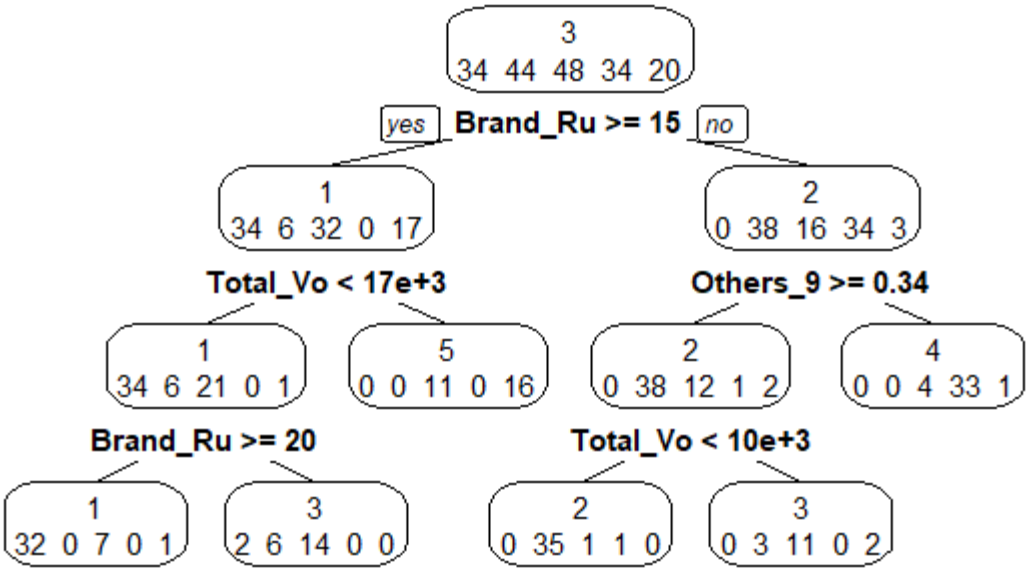
Firstly, we have categorized the dataset based on purchase behaviour, then we have separated dataset into training and test dataset in 70:30 ratio. Then the decision tree is built for two datasets/

Decision tree on Training Data:



Accuracy calculated for the tree is 84%.

Decision tree on Test data:



The accuracy calculated for the model is about 84%