IDS 572 Assignment 3
Data Mining for Business

Market Segmentation (Segmenting Consumers of Bath Soap)

Team: Viharika Bharti -655974244
Chaitra Srirama - 674121942
Aditya Gajula - 660252623

Question 1:

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

CRISA has traditionally segmented markets on the basis of purchaser demographics. They would now like to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty, doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and more effectively deploy promotion budgets.

We are using the clustering based approach for this case study because in the context of market 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.

The cluster's definitions change every time the clustering algorithm runs, ensuring that the groups always accurately reflect the current state of the data. The customers within each segment are very similar to one another and significantly different than those in other segments. In other words, each segment tells a different customer story. So our approach will be first, a cluster that describes purchase behavior, a second cluster that describes basis-for-purchase. A third clustering will then consider both sets of variables.

The better and more effective market segmentation would enable CRISA's clients to design more cost-effective promotions targeted at appropriate segments. Thus, multiple promotions could be launched, each targeted at different market segments at different times of a year. This would result in a more cost-effective allocation of the promotion budget to different market segments. It would also enable CRISA to design more effective customer reward systems and thereby increase brand loyalty.

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 that describe Purchase Behaviour:

[#brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, maxbr]

1) No_of_brands = Number of Brands Purchased.
2) Brand_Runs = Number of runs (streaks) of purchasing the same brand.
3) Total_Volume = Volume of product purchased (grams)
4) No_0f_Trans = Number of transactions
5) Value = Value in paise (100 paise = 1 rupee)
6) Avg_Price = Avg. price (rupees per 100 gram cake); computed from total volume and value

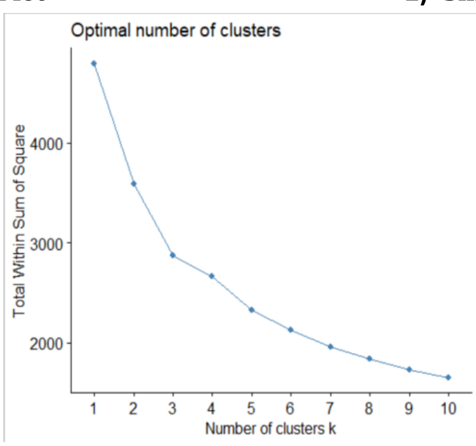
7) Others_999

8) Brand Loyalty is being evaluated on the fact that a customer would be most loyal to a brand if he/she would purchase a single brand more than the others. Hence based on this criteria, we evaluated brand loyalty as the maximum value(row-wise) out of all the different brand codes - Br. Cd. 57,144; Br. Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5. Others999 and assigned it to the variable **'MaxBr'.**

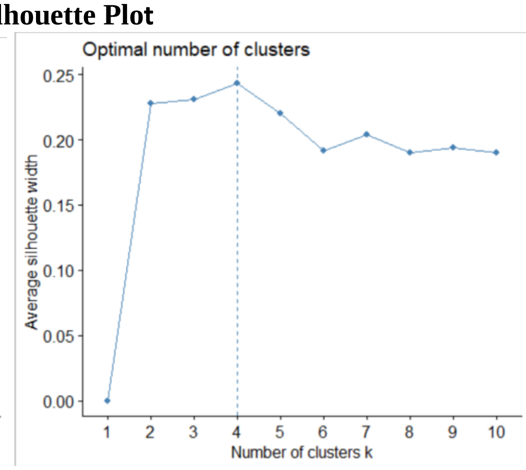
Another approach to evaluate the Brand Loyalty is we checked the number of brands purchased by members/customers.
Example : member id = 1047980 has a number of brands = 1, which means the customer buys only from a particular brand (in this case it is Br_cd_24).

Evaluating the optimal K-value for the purpose of clustering:

1) **Elbow Plot**

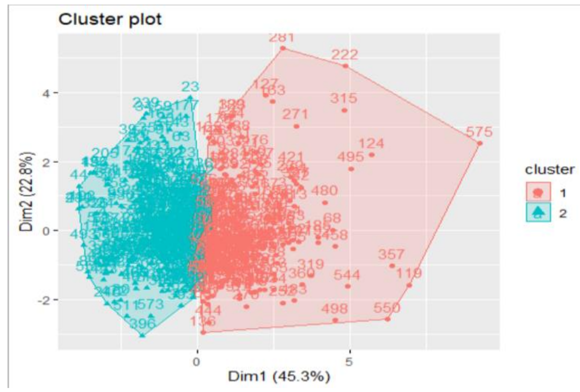
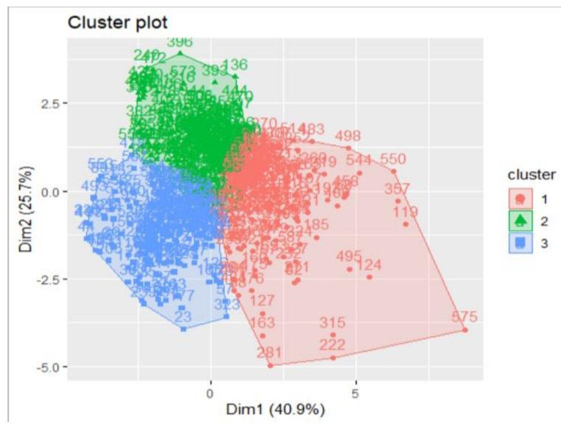
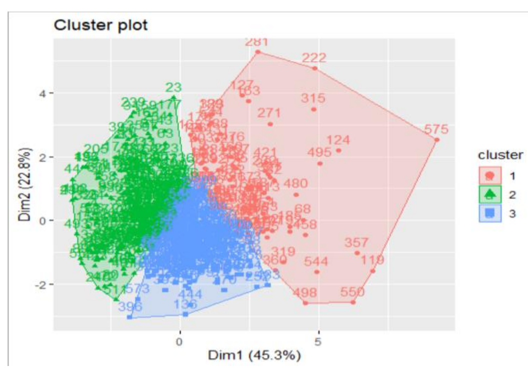
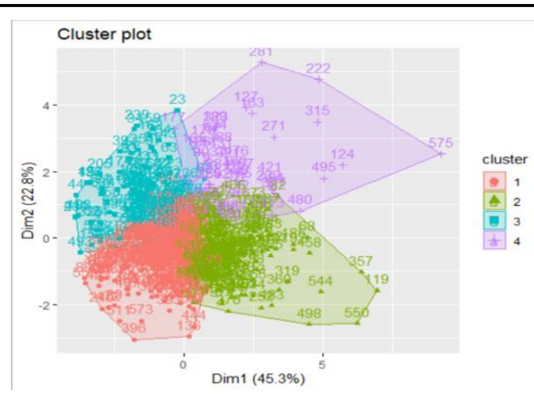


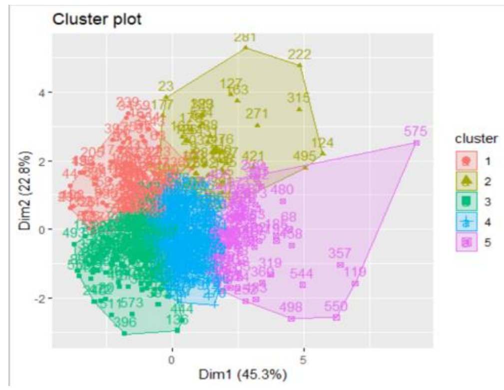
2) **Silhouette Plot**



Clustering based on Purchase Behavior variables

Centers	nstart	Data Points distribution	Cluster

2	30	2 clusters of sizes 330, 270 Within cluster sum of squares by cluster: [1] 1384.434 1661.896 (between_SS / total_SS = 27.3 %)	
3	20	3 clusters of sizes 253, 98, 249 Within cluster sum of squares by cluster: [1] 985.2814 782.0465 855.4295 (between_SS / total_SS = 37.4 %)	
3	60	3 clusters of sizes 259, 166, 175 Within cluster sum of squares by cluster: [1] 1242.739 1141.758 1585.147 (between_SS / total_SS = 33.7 %)	
4	30	4 clusters of sizes 188, 175, 191, 46 Within cluster sum of squares by cluster: [1] 875.1105 1170.8136 879.7605 502.2760 (between_SS / total_SS = 42.8 %)	

5	30	5 clusters of sizes 44, 74, 146, 117, 219 Within cluster sum of squares by cluster: [1] 224.1762 772.3575 868.4410 710.1273 461.5422 (between_SS / total_SS = 49.3 %)	

The elbow plot and the silhouette plot shows k=4 as the optimal number of clusters. However, among the above clusters, we observe the cluster with k=3, iter.max = 10 and nstart= 60 has a good cluster as we see decent suboptimal separation between the clusters along with convex shapes compared to the other K-means clusters. This cluster has a low ratio of between_SS / total_SS = 33.7 %.

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

Purch.Vol.no promo = Percent of volume purchased not on promotion
Purch.Vol.promo 6 = Percent of volume purchased on promo code 6
Purch.Vol other promo = Percent of volume purchased on promo code other than 6
Price codelist
Proposition codelist.

Variables used to describe basis-for-purchase are:

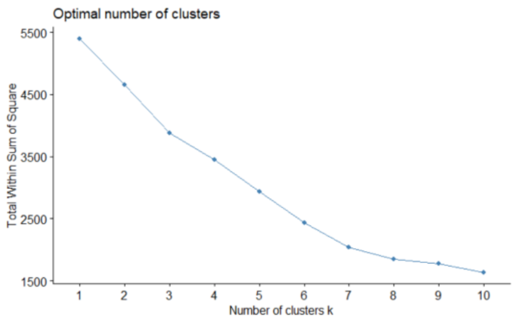
1) Promotion related variables:
    - We have derived the Purchase volume by promotions as follows:
      bsd$Pur_Promotion <- bsd$Pur_Vol_Other_Promo__ +bsd$Pur_Vol_Promo_6__
    - Variable Pur Vol No Promo - % has been dropped for the purpose of clustering.

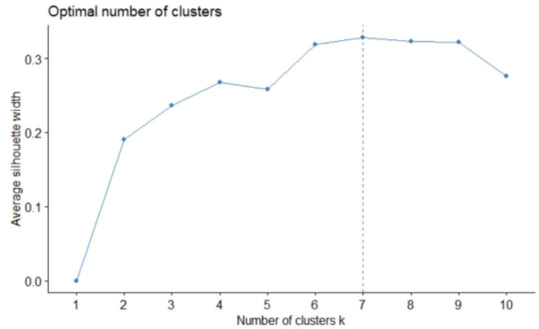
2) Price Categories: All the 4 price categories have been taken into consideration.- Pr_Cat_1, Pr_Cat_2, Pr_Cat_3, Pr_Cat_4,

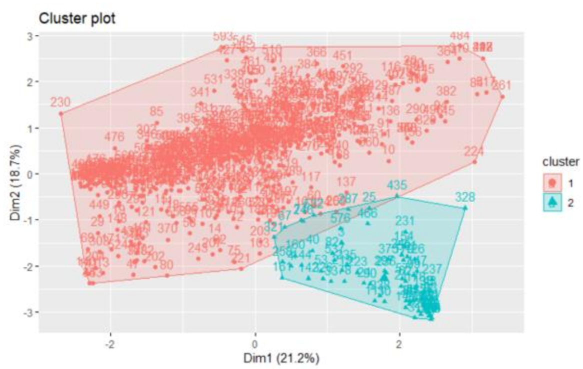
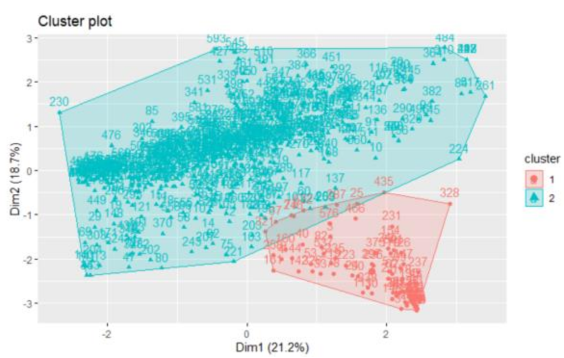
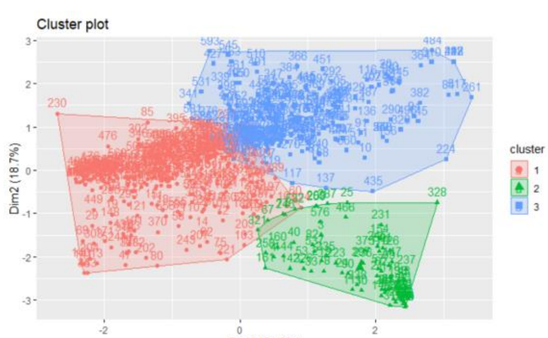
3) Selling Propositions: The selling proposition or unique selling point is a marketing strategy of making a unique proposition to customers that convinced them to switch brands. It was used in successful advertising campaigns.(USP, also seen as a unique selling point) is a factor that differentiates a product from its competitors, such as the lowest cost, the highest quality or the first-ever product of its kind. A USP could be thought of as “what you have that competitors don't.”
   The categories from 5 to 8 have been considered for the basis of purchase. Categories 9 to 15 did not show much distribution in terms of values, hence these have been dropped.

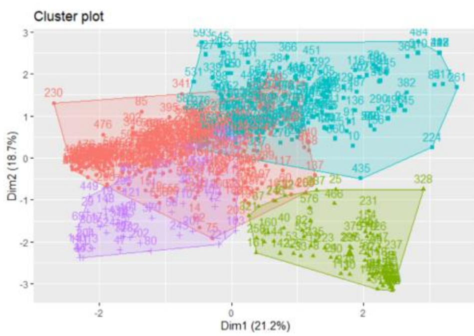
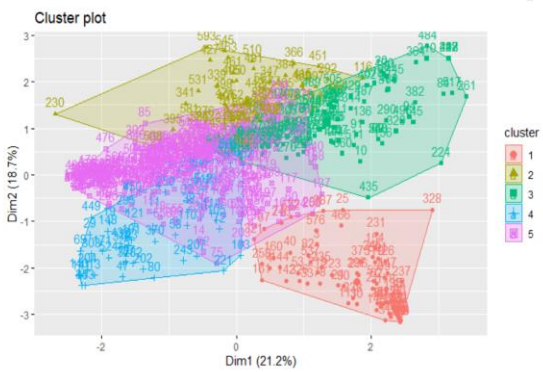
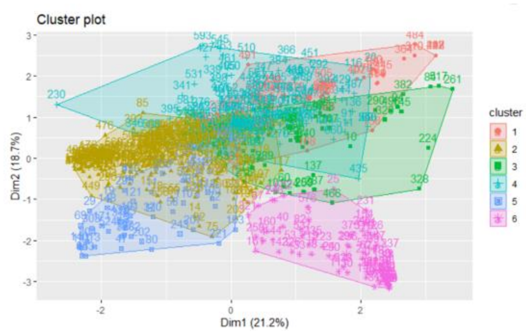
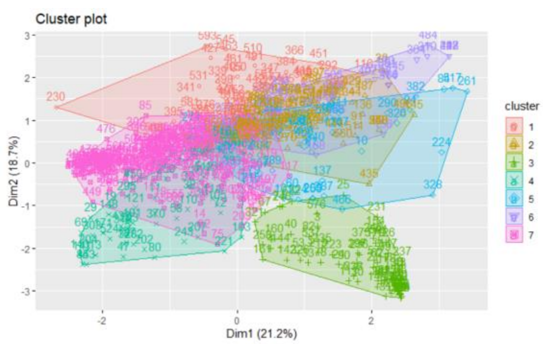
To find the optimal value of k as follows:

1)Elbow Plot **2) Silhouette plot**





K Value	nstart	Data Points distribution	Cluster
2	30	2 clusters of sizes 520, 80 Within cluster sum of squares by cluster: [1] 4370.7957 185.3537 (between_SS / total_SS = 15.5 %)	
2	60	2 clusters of sizes 80, 520 Within cluster sum of squares by cluster: [1] 185.3537 4370.7957 (between_SS / total_SS = 15.5 %)	
3	60	3 clusters of sizes 81, 215, 304 Within cluster sum of squares by cluster: [1] 194.8455 1551.5754 2083.7473 (between_SS / total_SS = 29.0 %)	

4	60	4 clusters of sizes 300, 79, 166, 55 Within cluster sum of squares by cluster: [1] 1347.2891 229.2883 176.6143 1471.1061 (between_SS / total_SS = 40.2 %)	
5	60	5 clusters of sizes 63, 55, 79, 106, 297 Within cluster sum of squares by cluster: [1] 416.8335 229.2883 176.6143 730.9184 1232.9868 (between_SS / total_SS = 48.3 %)	
6	60	6 clusters of sizes 50, 225, 60, 136, 53, 76 Within cluster sum of squares by cluster: [1] 992.4802 259.3452 203.2978 489.4181 151.2842 276.1565 (between_SS / total_SS = 56.0 %)	
7	60	7 clusters of sizes 49, 87, 76, 54, 57, 41, 236 Within cluster sum of squares by cluster: [1] 311.9959 253.0490 344.3188 551.2130 151.2842 224.5370 198.8737 (between_SS / total_SS = 62.2 %)	

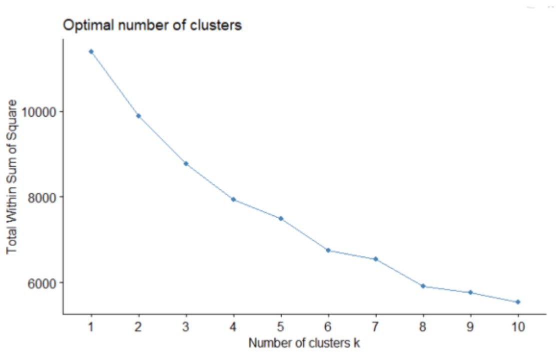
The elbow plot and the silhouette plot shows k=7 as the optimal number of clusters. However, among the above clusters, we observe the cluster with k=3 and nstart= 60 has a good cluster as we see decent suboptimal separation between the clusters along with convex shapes compared to the other K-means clusters. This cluster has a low ratio of between_SS / total_SS = 29.0 %.

c) The variables that describe both purchase behavior and basis for purchase.

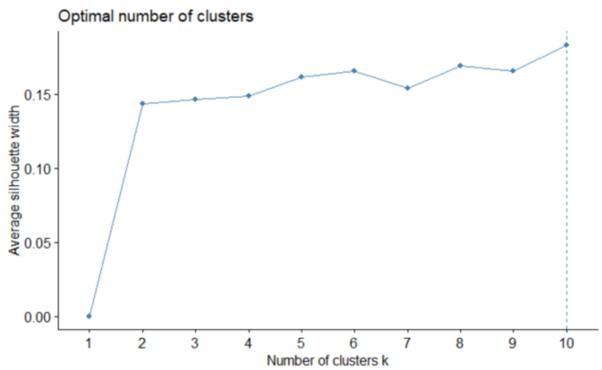
We built clusters with the combined set of variables for purchase behavior as well as for basis of purchase which are : No__of_Brands, Brand_Runs, Total_Volume, No__of__Trans, Value, Avg__Price, Trans___Brand_Runs, Vol_Tran, maxBr, Others_999, Pr_Cat_1, Pr_Cat_2, Pr_Cat_3, Pr_Cat_4, Pur_Promotion, PropCat_5, PropCat_6, PropCat_7 and PropCat_8

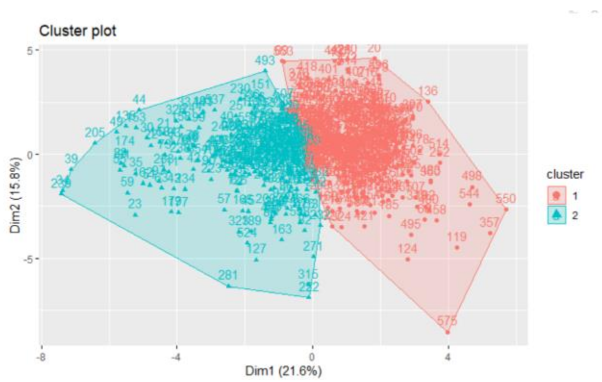
Evaluating the optimal value of k based on the above variable set.

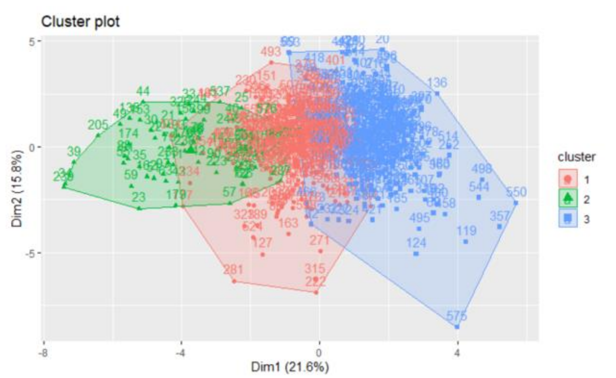
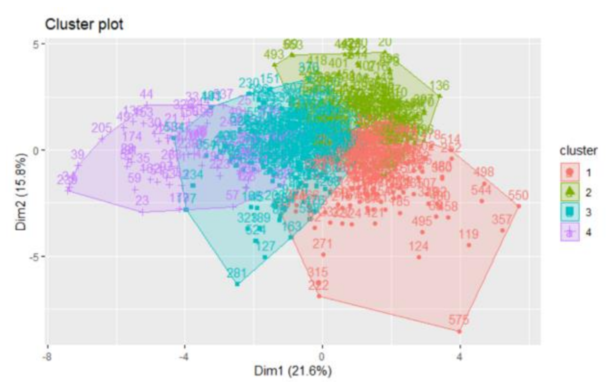
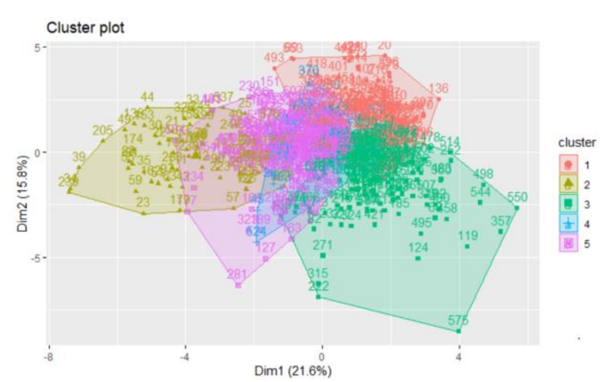
1)Elbow Method:

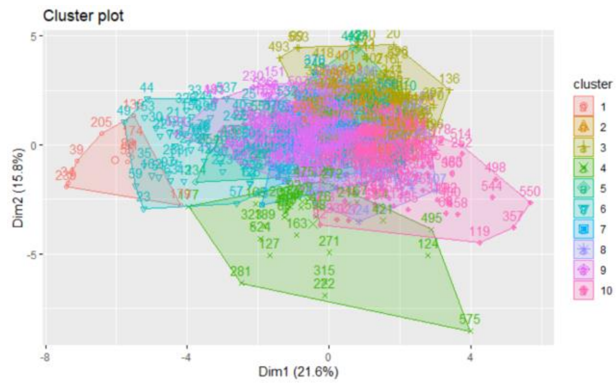


2)Silhouette method:



K Value	nstart	Data Points distribution	Cluster
2	30	2 clusters of sizes 360, 240 Within cluster sum of squares by cluster: [1] 5628.967 4165.342 (between_SS / total_SS = 13.9 %)	

2	60	2 clusters of sizes 360, 240 Within cluster sum of squares by cluster: [1] 5628.967 4165.342 (between_SS / total_SS = 13.9 %)	
3	60	3 clusters of sizes 257, 68, 275 Within cluster sum of squares by cluster: [1] 3844.5876 697.6663 4230.7620 (between_SS / total_SS = 22.9 %)	
4	60	4 clusters of sizes 170, 131, 231, 68 Within cluster sum of squares by cluster: [1] 1980.5333 1955.6474 3302.6337 697.6663 (between_SS / total_SS = 30.3 %)	
5	60	5 clusters of sizes 129, 69, 167, 49, 186 Within cluster sum of squares by cluster: [1] 1943.9389 710.8762 1906.6200 486.6451 2133.3348 (between_SS / total_SS = 36.9 %)	

10	60	10 clusters of sizes 10, 37, 67, 33, 28, 59, 50, 52, 137, 127 Within cluster sum of squares by cluster: [1] 81.71114 363.41562 683.58642 512.54226 382.95159 380.08268 415.65877 [8] 493.15440 968.72016 1009.61459 (between_SS / total_SS = 53.5 %)	

The elbow plot and the silhouette plot shows k=10 as the optimal number of clusters. However, among the above clusters, we observe the cluster with k=2 and nstart= 60 has a good cluster as we see decent suboptimal separation between the clusters with least overlap along with convex shapes compared to the other K-means clusters. This cluster has a low ratio of (between_SS / total_SS = 13.9 %).

Question 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. Show how you experiment with different parameter values for the different techniques, and how these affect the clusters obtained.

We have used the following clustering methods:

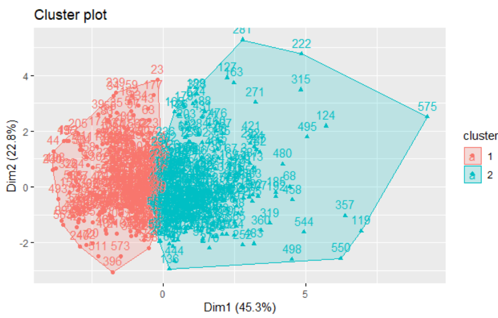
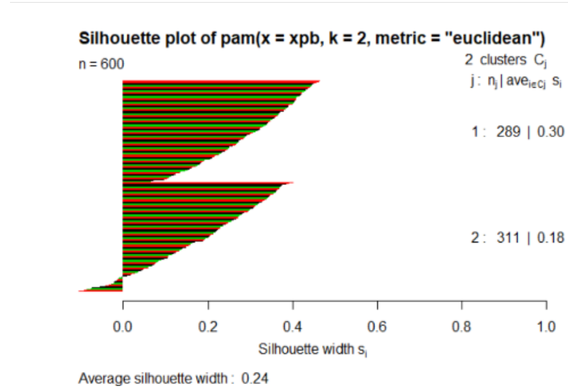
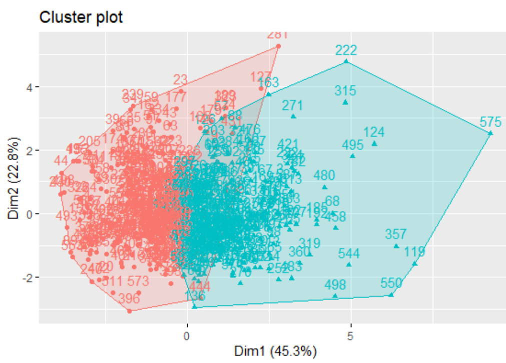
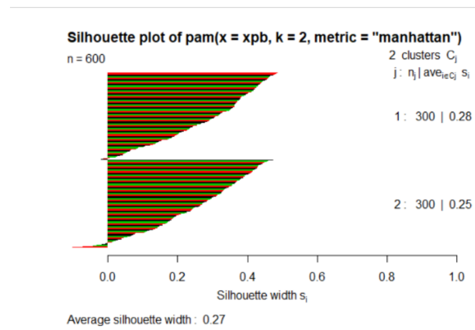
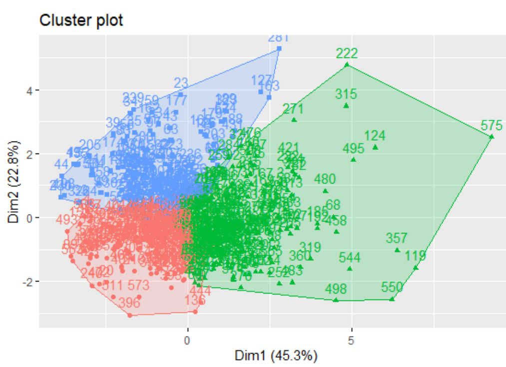
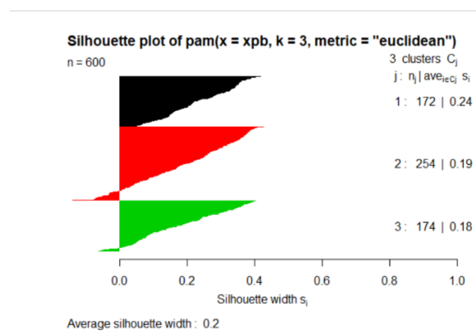
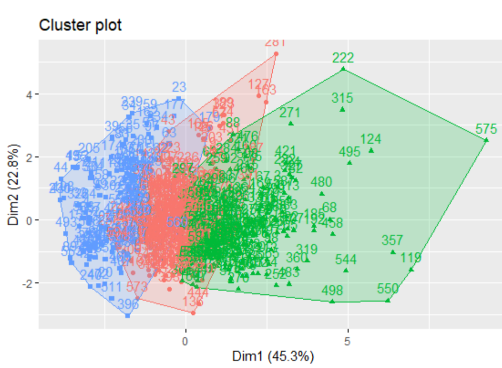
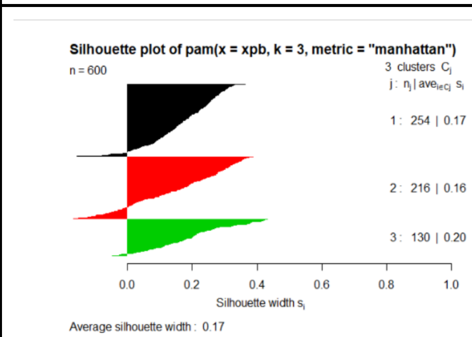
1) K-Medoids (PAM)
2) Hierarchical clustering (Agglomerative method)
3) Kernel-K means.

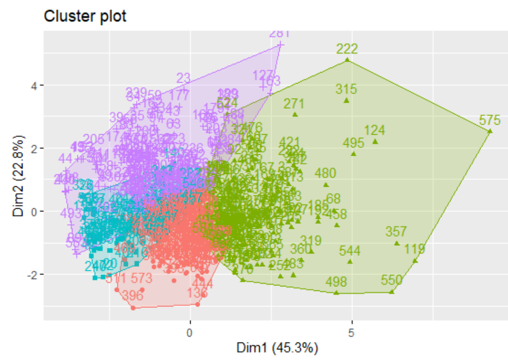
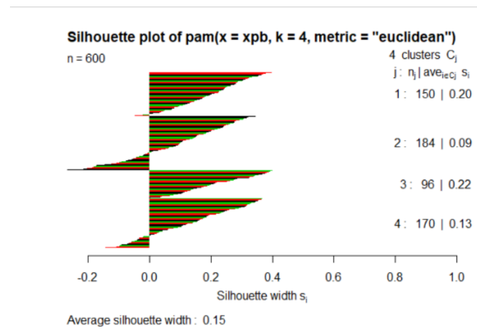
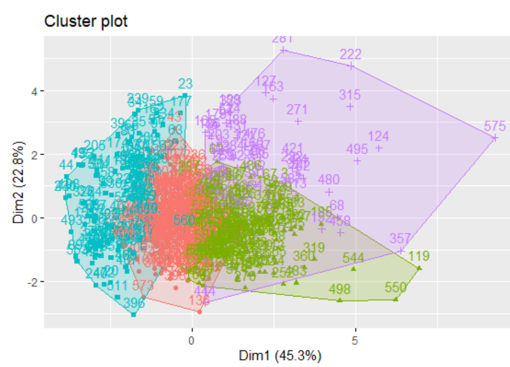
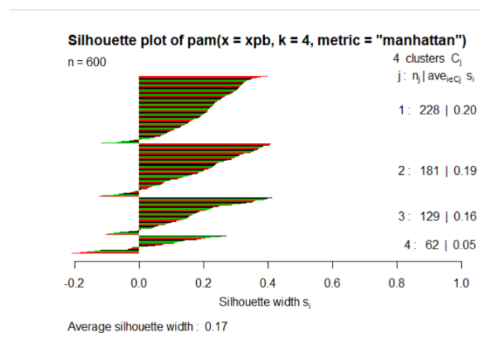
1)Partitioning around medoids (PAM) (K-medoids):

In this method, K- Medoids are used instead of means to avoid outliers.

a) Using the Purchase Behavior variables

K value	metric	cluster	Silhouette plot (showing details of cluster size and silhouette with)

<table>
<tr><td>2</td><td>euclidean</td><td>
</td><td>
</td></tr>
<tr><td>2</td><td>manhattan</td><td>
</td><td>
</td></tr>
<tr><td>3</td><td>euclidean</td><td>
</td><td>
</td></tr>
<tr><td>3</td><td>manhattan</td><td>
</td><td>
</td></tr>
</table>

4	euclidean		
4	manhattan		

We use the Average Silhouette Width to evaluate the clusters. From the clusters above, clusters with k=2 with distance measure= manhattan provide the best average silhouette width of 0.27.

Below are the cluster statistics to compare for a better cluster. Clusters with k=2 with distance measure= manhattan has highest separation and lowest(smallest) negative silhouette width and hence can help in choosing a better cluster.

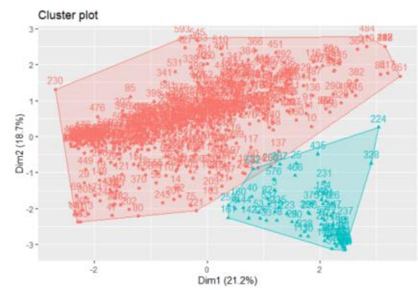
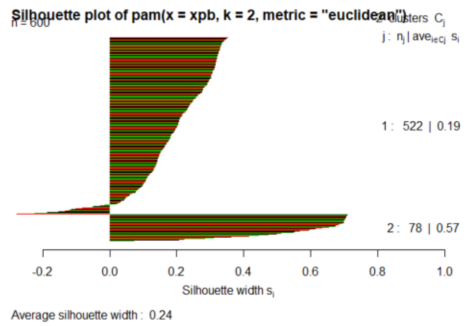
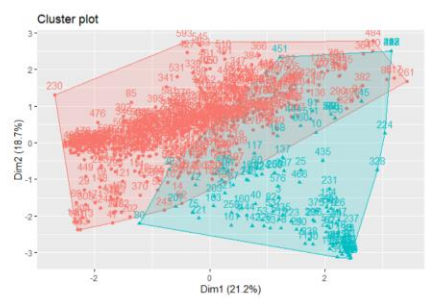
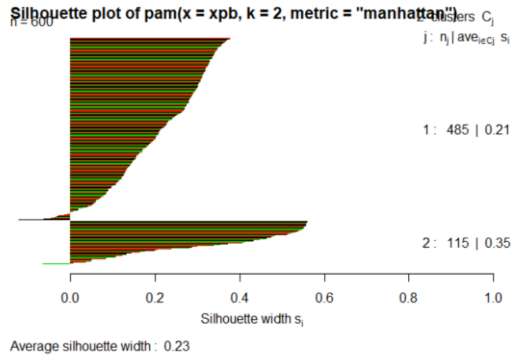
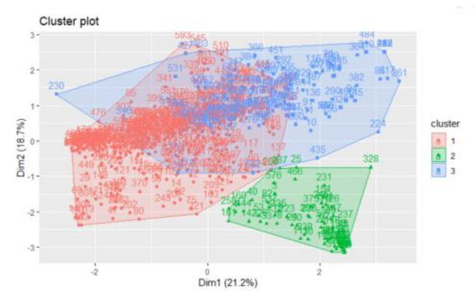
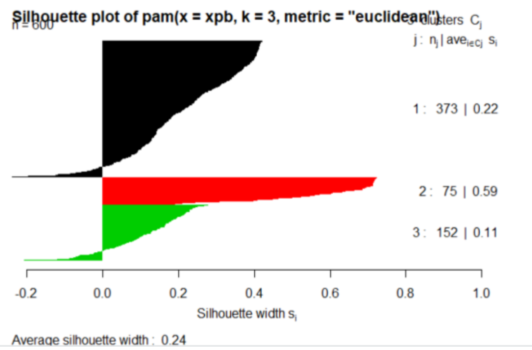
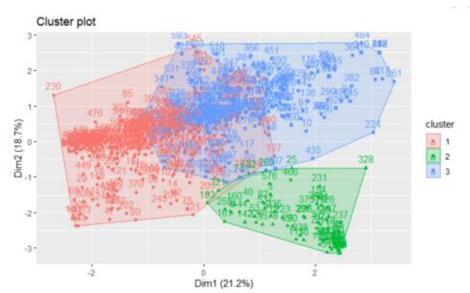
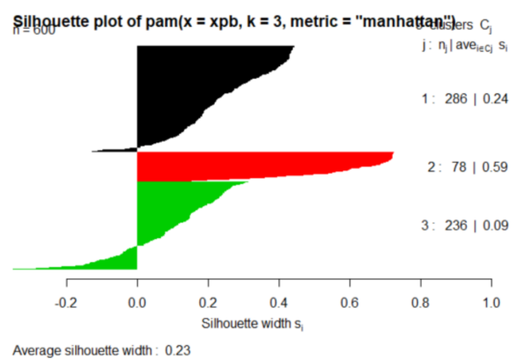
```
     size max_diss  av_diss  diameter separation
[1,]  165 6.421770 1.827404  7.901533  0.6450011
[2,]  258 8.751108 2.188601 10.690945  0.5967998
[3,]  177 6.797455 1.933117  9.825481  0.5967998
```

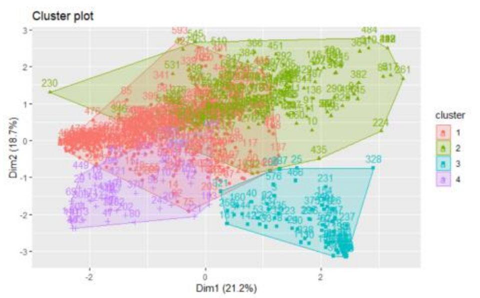
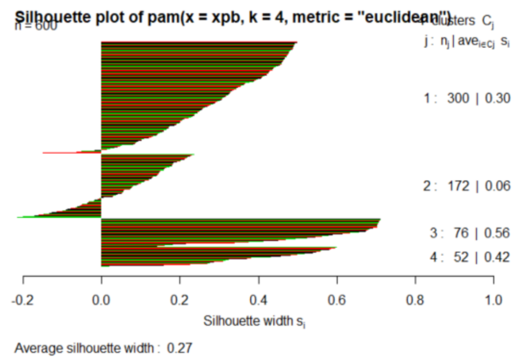
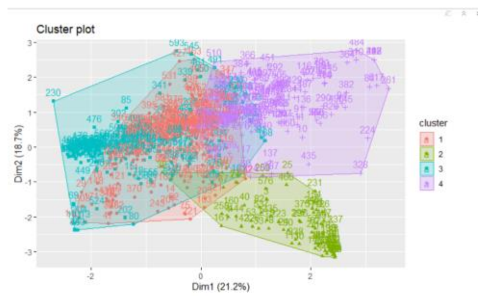
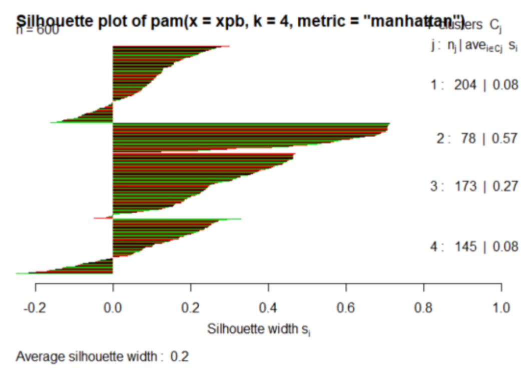
```
     size max_diss  av_diss diameter separation
[1,]  355 23.45453 5.368154 33.23515  0.9374531
[2,]  245 14.83649 5.023942 27.01899  0.9374531
```

But for a marketing approach, if we are to pick k=3, then we would choose the resulting cluster with distance= euclidean with an average silhouette width of 0.2.

b) Using the Basis for Purchase variables:

K value	metric	cluster	Silhouette plot

<table>
<tr><td>2</td><td>euclidean</td><td>
</td><td>
</td></tr>
<tr><td>2</td><td>manhattan</td><td>
</td><td>
</td></tr>
<tr><td>3</td><td>euclidean</td><td>
</td><td>
</td></tr>
<tr><td>3</td><td>manhattan</td><td>
</td><td>
</td></tr>
</table>

4	euclidean		
4	manhattan		

We use the Average Silhouette Width to evaluate the clusters. From the clusters above, clusters with k=2 with distance measure= euclidean provide the best average silhouette width of 0.24.

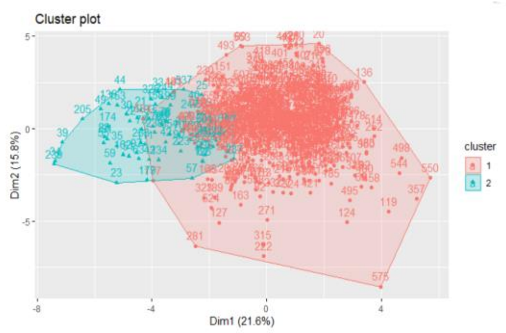
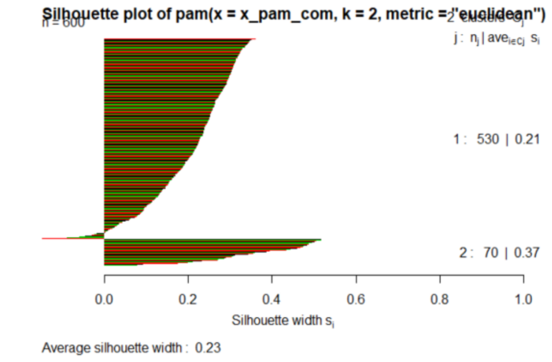
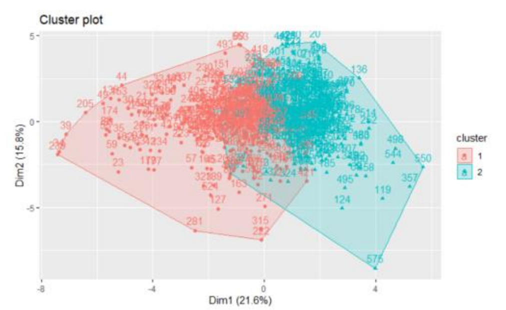
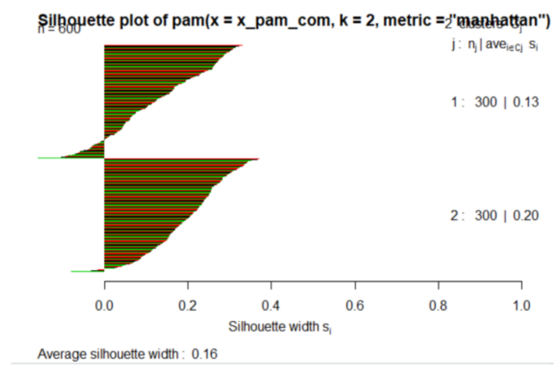
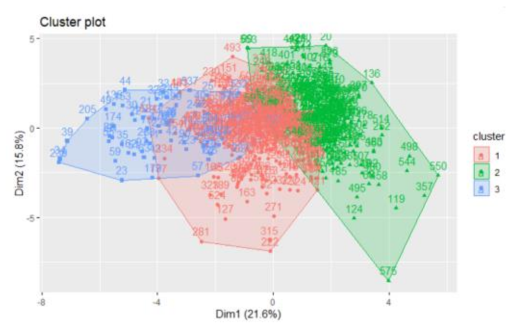
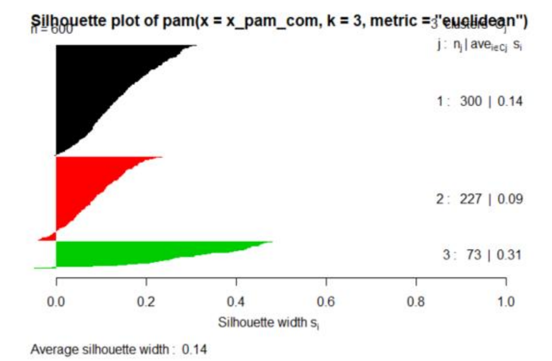
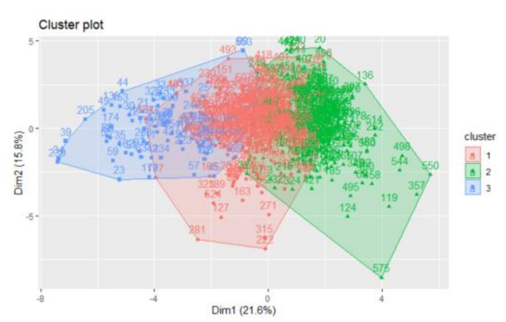
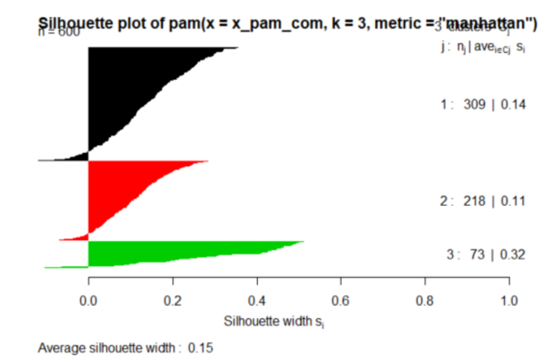
Below are the cluster statistics to compare for a better cluster. Clusters with k=2 with distance measure= manhattan have the highest separation and lowest(smallest) negative silhouette width and hence can help in choosing the better cluster.

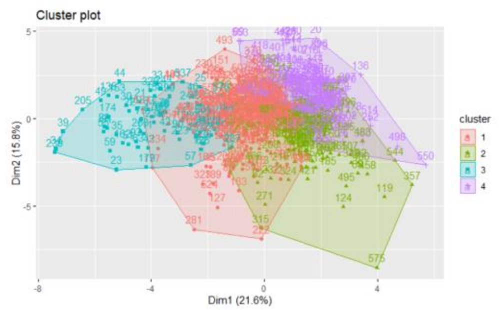
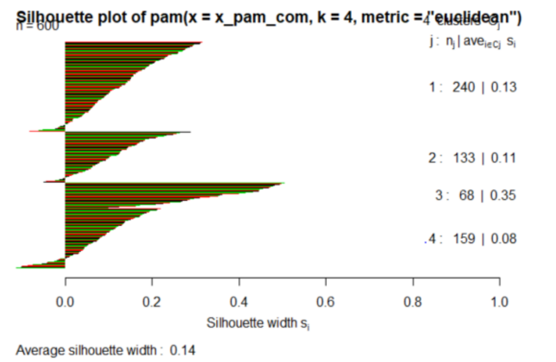
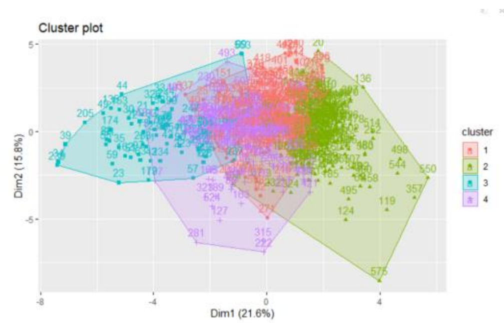
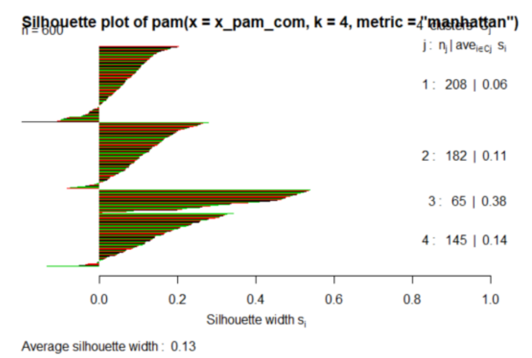
```
     size max_diss  av_diss  diameter separation
[1,]  373 6.921316 2.311887  9.996845  0.5444283
[2,]   75 3.432694 1.299409  5.095618  0.6481742
[3,]  152 8.151696 2.703246 11.362254  0.5444283
```

```
     size max_diss  av_diss diameter separation
[1,]  485 13.56811 5.494903 23.16923  0.8630616
[2,]  115 13.42676 4.375352 19.28822  0.8630616
```

c) The variables that describe both purchase behavior and basis for purchase.

K value	metric	cluster	Silhouette plot

2	euclidean		
2	manhattan		
3	euclidean		
3	manhattan		

4	euclidean		
4	manhattan		

We use the Average Silhouette Width to evaluate the clusters. From the clusters above, clusters with k=2 with distance measure= manhattan provide the best average silhouette width of 0.23.

Below are the cluster statistics to compare for a better cluster. Clusters with k=2 with distance measure= Euclidean have the highest separation and lowest(smallest) negative silhouette width and hence can help in choosing the better cluster.

```
     size  max_diss  av_diss diameter separation
[1,]  303  8.585378 3.357052 12.12440   1.045291
[2,]  228  9.081886 4.316412 15.77438   1.045291
[3,]   69 16.695708 2.659542 17.94937   1.533391
```

```
     size  max_diss  av_diss diameter separation
[1,]  531  9.979055 4.046454 15.77438   1.533391
[2,]   69 16.695708 2.661793 17.94937   1.533391
```

Davies-Bouldin's index for all sets of variables:

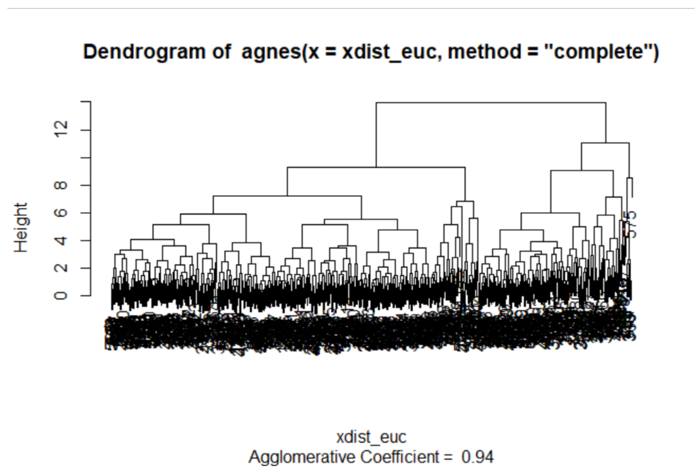
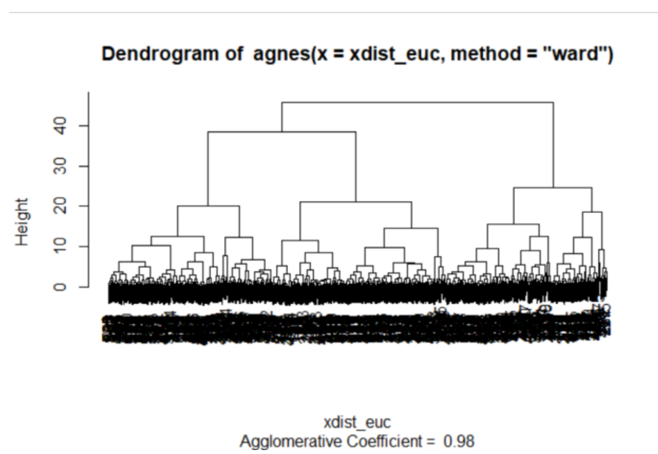
Purchase of behavior		Basis for purchase		Combined Variables	
k=2	1.734174	k=2	1.27682	k=2	2.220139
k=3	1.492116	k=3	1.859276	k=3	2.523514
k=4	1.727159	k=4	1.620845	k=4	2.590322
k=5	1.513672	k=5	1.781159	k=5	2.378192

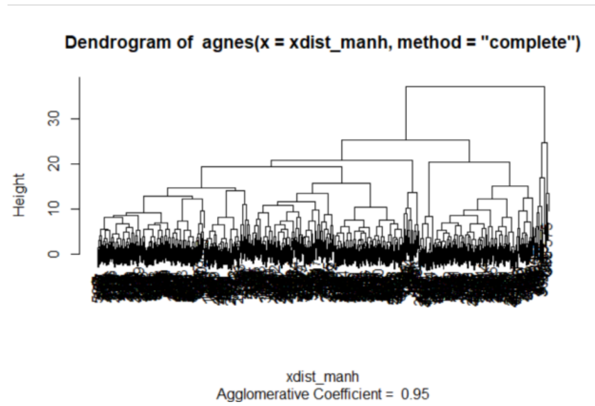
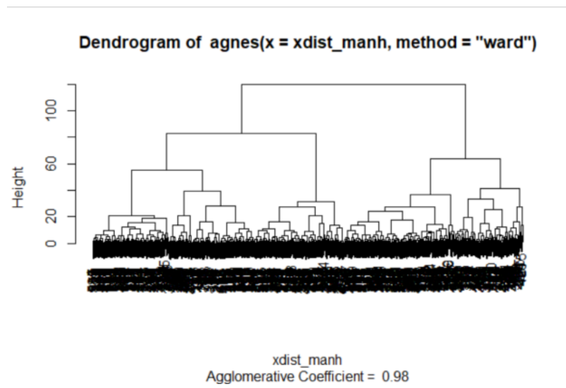
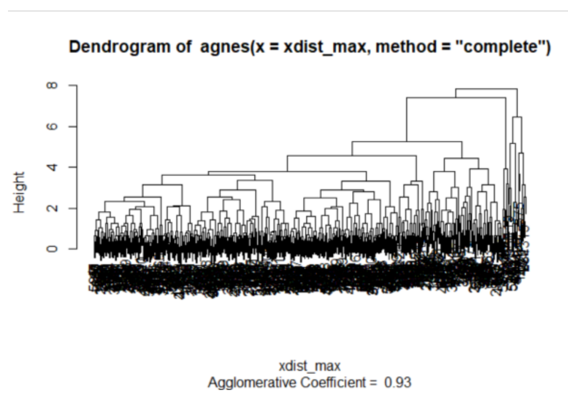
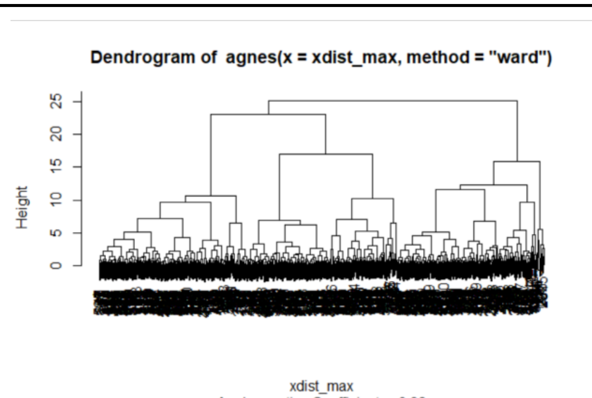
We have calculated the Davies-Bouldin Index for all sets of variables. DB Index evaluates intra-cluster similarity and inter-cluster differences. Lower value of db index indicates good clustering. We see that the db index is lower for purchase of behavior variable set (k=3) and basis of purchase variables (k=2). Based on this we can say that these combinations will give us better clustering.

2. Hierarchical Agglomerative Clustering (HAC) or AGNES

Agglomerative methods are good for identifying small clusters. Before clustering is performed, it is required to determine the distance matrix that specifies the distance between each data point using some distance function (Euclidean, Manhattan, Minkowski, etc.). We have used three matrices here for our analysis, i.e. Euclidean, Manhattan and Maximum.

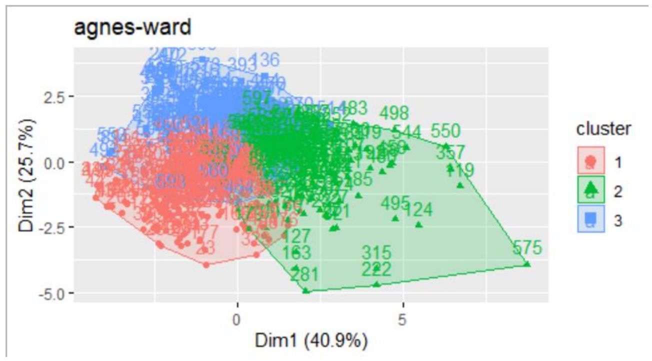
a) Purchase_Behavior

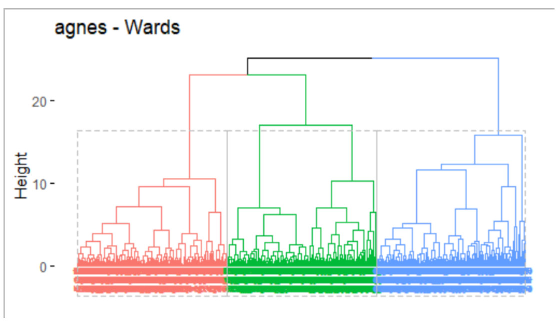
agglomerative /agnes clustering		agglomerative coeff given by agnes
euclidean	complete	
	ward	

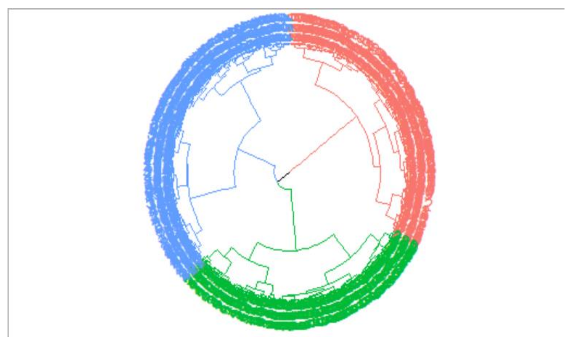
<table>
<tr><td rowspan="2">manhattan</td><td>complete</td><td>
</td></tr>
<tr><td>ward</td><td>
</td></tr>
<tr><td rowspan="2">maximum</td><td>complete</td><td>
</td></tr>
<tr><td>ward</td><td>
</td></tr>
</table>

Davies-Bouldin's index = 1.492116
From the above figures, we see that dendrograms obtained from the ward method are the best as it's agglomerative coefficient is better than the complete method. We get agglomerative coefficient as 0.98 for all three distancing methods analysed.

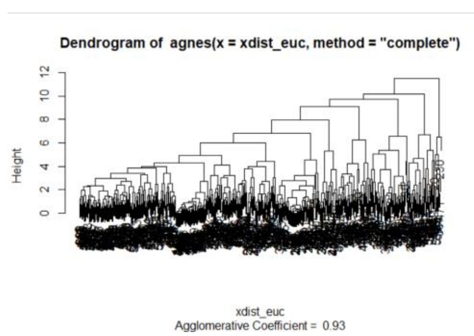
Now, we need to cut our dendrogram as everything is under one cluster and we need to have more numbers of clusters. We are cutting our dendrogram at cluster 3.

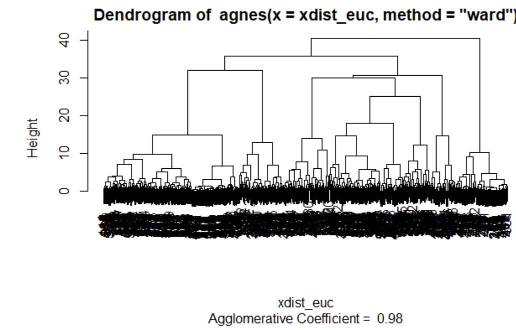
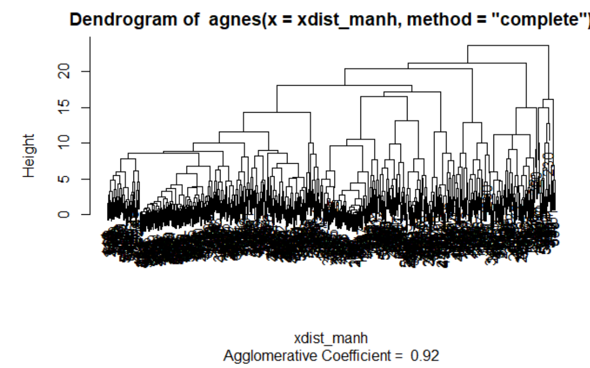
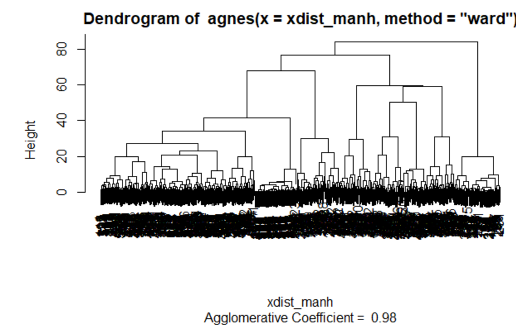
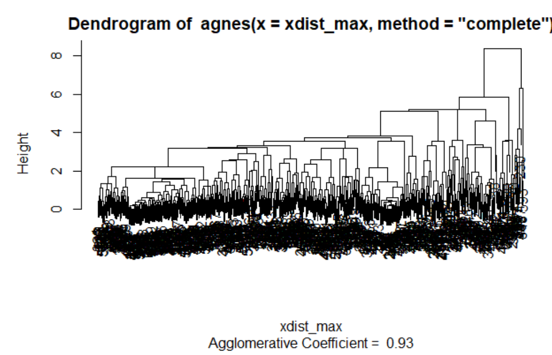


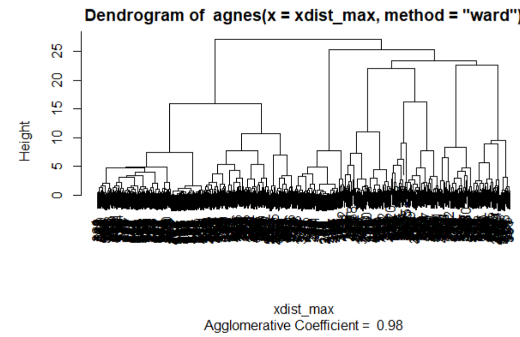




b) **<u>Basis for Purchase:</u>**

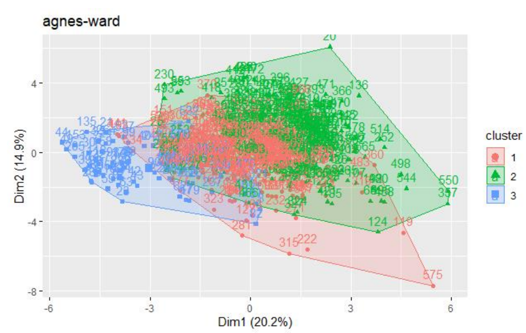
agglomerative/ agnes clustering		agglomerative coeff given by agnes
euclidean	complete	

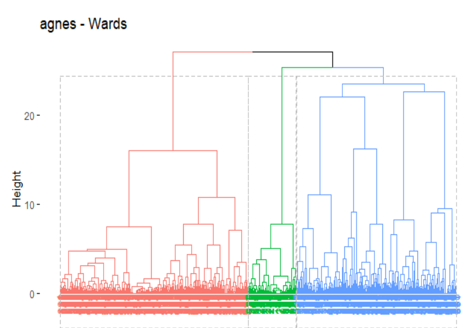
<table>
<tr><td></td><td>ward</td><td>
</td></tr>
<tr><td>manhattan</td><td>complete</td><td>
</td></tr>
<tr><td></td><td>ward</td><td>
</td></tr>
<tr><td>maximum</td><td>complete</td><td>
</td></tr>
</table>

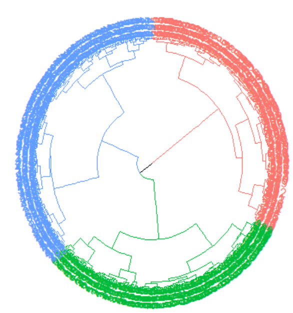
	ward	

Davies-Bouldin's index = 2.973925
For the basis of purchase behavior as well dendrograms obtained from the ward method are the best as it's agglomerative coefficient if better than the complete method. We get agglomerative coefficient as 0.98 for all three distancing methods analysed.

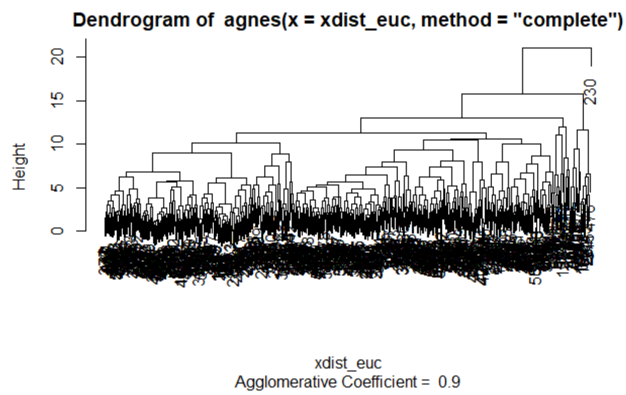
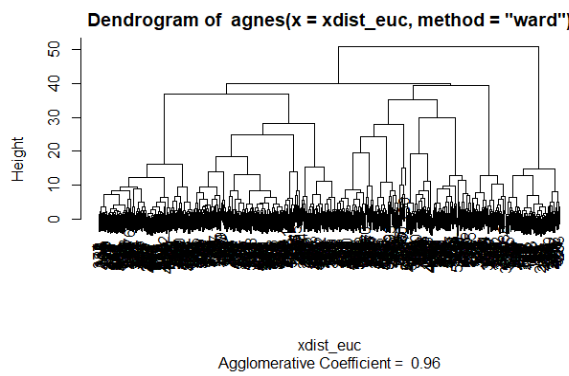
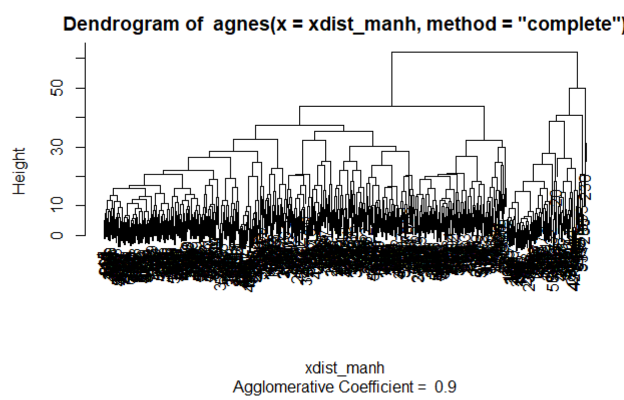
We are cutting our dendrogram at cluster 3.

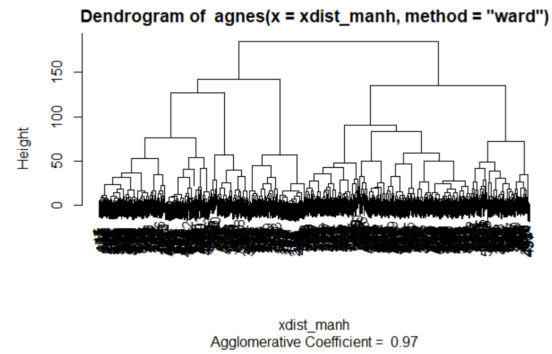
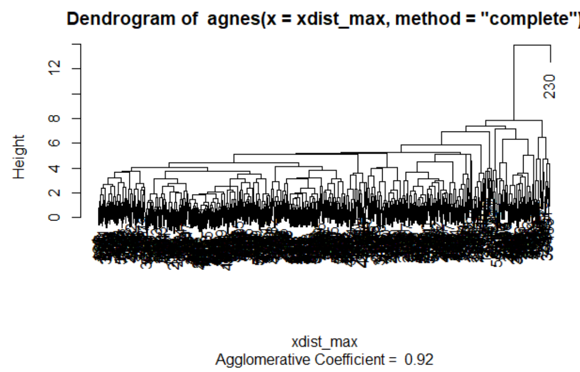
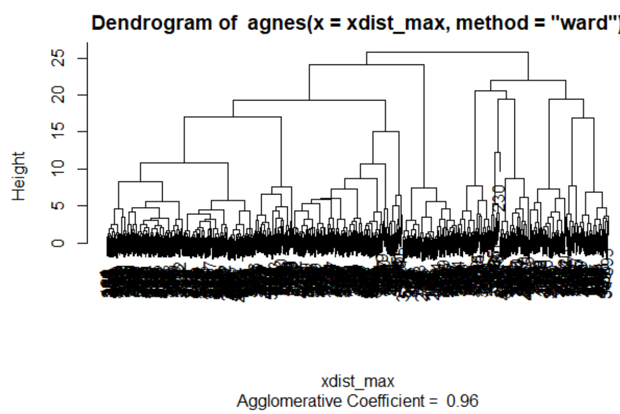






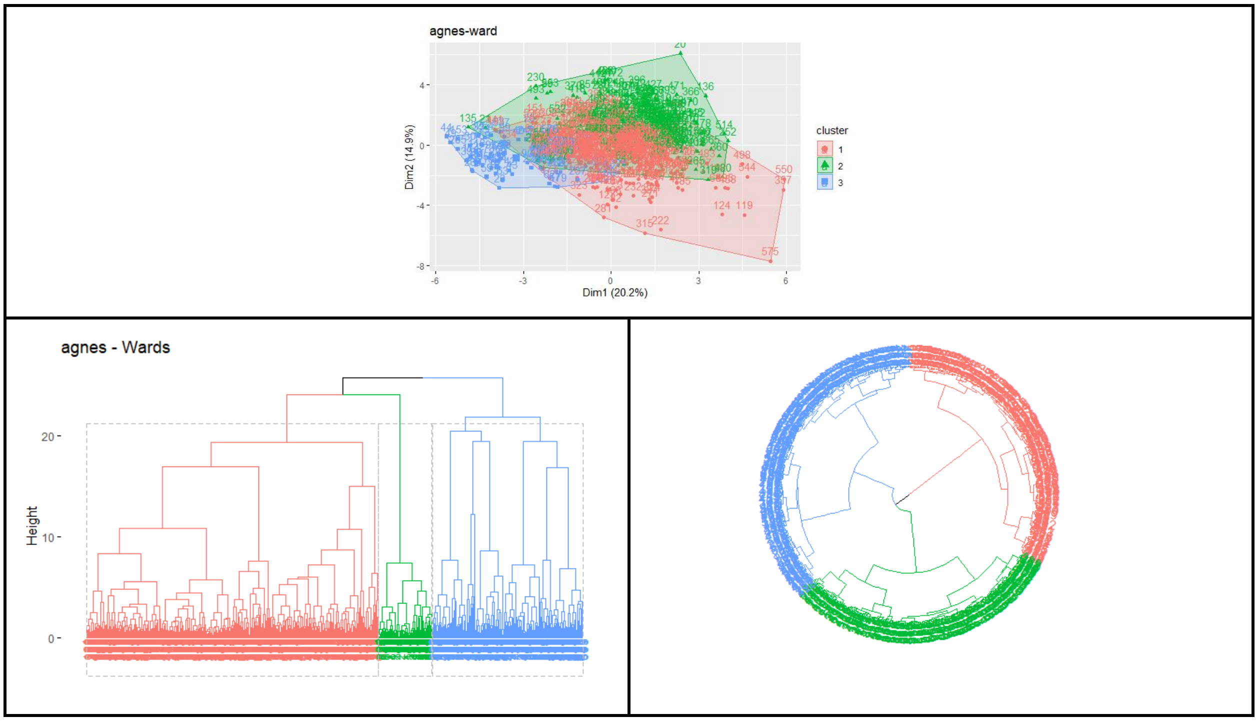
c) Purchase Behavior and Basis for Purchase combined variables

agglomerative/ agnes clustering		agglomerative coeff given by agnes
euclidean	complete	
	ward	
manhattan	complete	

<table>
<tr><td></td><td>ward</td><td>
</td></tr>
<tr><td>maximum</td><td>complete</td><td>
</td></tr>
<tr><td></td><td>ward</td><td>
</td></tr>
</table>

Davies-Bouldin’s index= 2.441306
For the combined variables our coefficient value is again best for ward method, i.e 0.97 for manhattan and 0.96 for euclidean and maximum.

We are cutting our dendrogram at cluster 3.



	Purchase Behaviour	Basis for Purchase	Both Purchase Behavior and Basis for Purchase
Davies-Bouldin's index	1.492116	2.973925	2.441306

Davies-Bouldin Index evaluates intra-cluster similarity and inter-cluster differences. It is the minimum for purchase behavior variable set. By this we can say that we got good clustering for these variable sets.

As shown in the above three tables(for three segmentation), we have tried multiple parameters for different distances and clustering methods for agglomerative clustering. AC describes the strength of the clustering structure. Agglomerative coefficient value close to 1 shows better clustering.

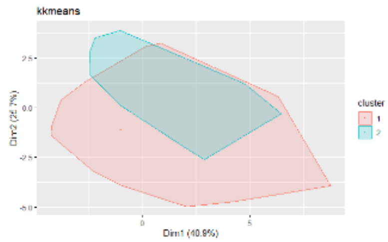
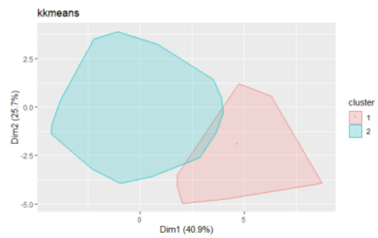
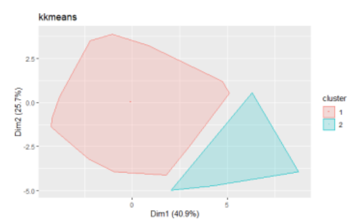
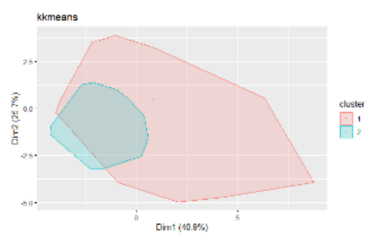
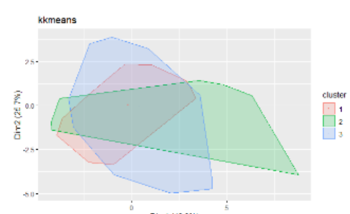
Here we see that Ward's method identifies the strongest clustering structure of all methods assessed. Ward's method aims to minimize the total within-cluster variance. For the euclidean method(ward clustering) **agglomerative coefficient is 0.98.**

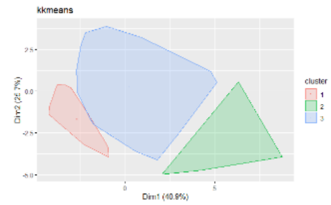
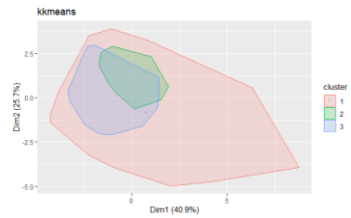
Next, we cut the dendrogram in order to create the desired number of clusters. We experimented with k=2,3 and 4 to cut the dendrogram for getting the desired number of clusters. We choose the number of clusters to be k=3 in all sets of variables , or as we can see in the dendrogram h=3 we get three clusters.

3. Kernel k-means:

In kernel k-means, we are mapping the data to higher dimensional space but not to linear dimensional space. Here we have experimented with different types of kernel like ploydot kernel and radial basis kernel.

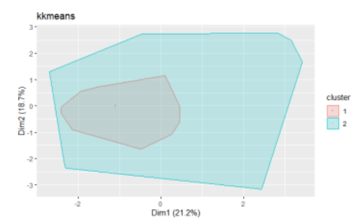
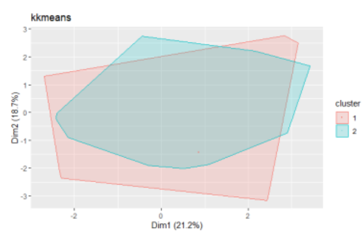
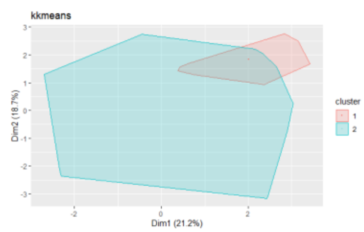
a) Purchase Behavior

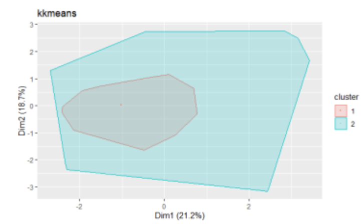
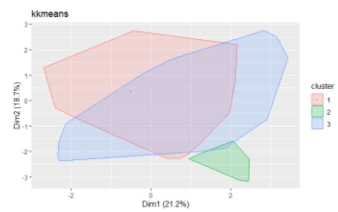
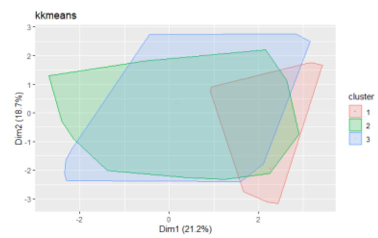
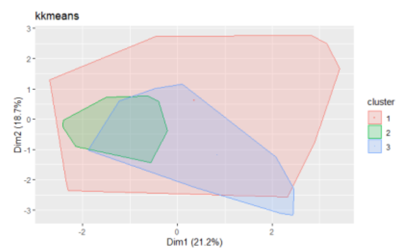
K Value			
For k=2	rbf kernel	Cluster size: [1] 238 362 Within-cluster sum of squares: [1] 2848.558 2720.512	
	polynomial kernel with degree 2	Cluster size: [1] 17 583 Within-cluster sum of squares: [1] 578.2558 4045.6776	
	polynomial kernel with degree 3	Cluster size: [1] 591 9 Within-cluster sum of squares: [1] 4276.7977 409.4864	
	rbf kernel with sigma=0.2	Cluster size: [1] 380 220 Within-cluster sum of squares: [1] 3552.508 1638.678	
For k=3	polynomial kernel with degree 2	Cluster size: [1] 441 21 138 Within-cluster sum of squares: [1] 2338.5355 685.4177 1776.7842	

	polynomial kernel with degree 3	Cluster size: [1] 61 9 530 Within-cluster sum of squares: [1] 1193.5060 409.4864 3445.8330	
	rbf kernel with sigma=0.2	Cluster size: [1] 203 194 203 Within-cluster sum of squares: [1] 2965.324 1183.911 1013.799	

By looking at the within-cluster sum of squares, we can say that the polynomial kernel with k=2 and degree =3 gives us better clusters as within-cluster sum of squares values are lower with these combinations. It means that data points are closely packed and are more relevant to the cluster.

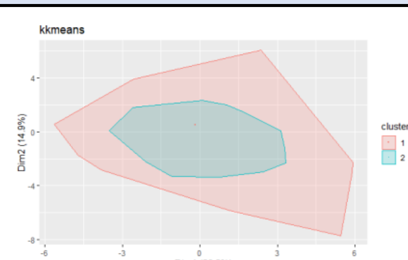
b) Basis for Purchase

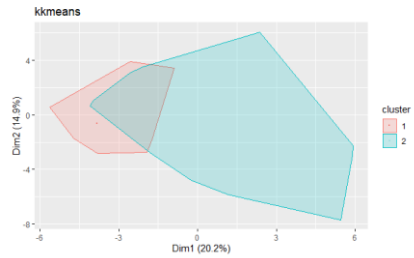
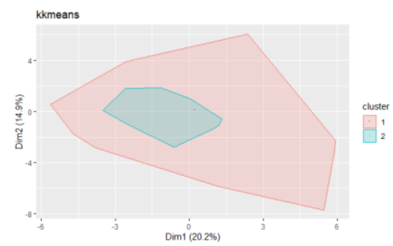
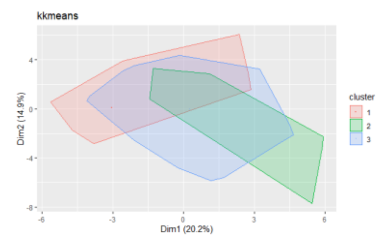
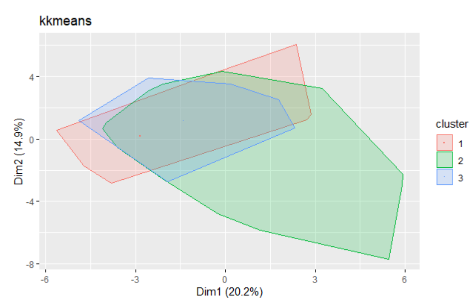
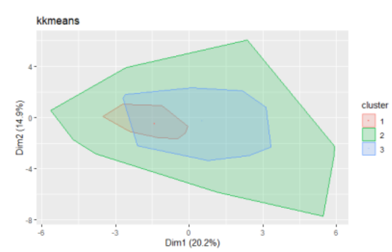
K Value			
For k=2	rbf kernel	Cluster size: [1] 224 376 Within-cluster sum of squares: [1] 1266.113 4742.680	
	polynomial kernel with degree 2	Cluster size: [1] 117 483 Within-cluster sum of squares: [1] 2443.901 3465.836	
	polynomial kernel with degree 3	Cluster size: [1] 25 575 Within-cluster sum of squares: [1] 914.9424 4738.6760	

	rbf kernel with sigma=0.2	Cluster size: [1] 245 355 Within-cluster sum of squares: [1] 1263.050 4690.019	
For k=3	polynomial kernel with degree 2	Cluster size: [1] 378 55 167 Within-cluster sum of squares: [1] 2637.112 1318.375 2416.851	
	polynomial kernel with degree 3	Cluster size: [1] 52 450 98 Within-cluster sum of squares: [1] 1343.289 2721.505 2006.459	
	rbf kernel with sigma=0.2	Cluster size: [1] 296 157 147 Within-cluster sum of squares: [1] 3919.070 1100.593 1345.355	

From the above data, we can say that the polynomial kernel with k=3 and degree =2 gives us better clusters as within-cluster sum of squares values are lower with these combinations. Cluster 2 is better than cluster 1 and cluster 3, which means that data points are closely packed and are more relevant to the cluster 2.

c) Purchase Behavior and Basis of Purchase variables:

K Value			
For k=2	rbf kernel	Cluster size: [1] 298 302 Within-cluster sum of squares: [1] 8622.073 3311.110	

	polynomial kernel with degree 2	Cluster size: [1] 51 549 Within-cluster sum of squares: [1] 2622.412 9723.458	
	rbf kernel with sigma=0.2	Cluster size: [1] 469 131 Within-cluster sum of squares: [1] 10420.749 1414.149	
For k=3	polynomial kernel with degree 2	Cluster size: [1] 53 36 511 Within-cluster sum of squares: [1] 2831.298 1941.312 8069.432	
	polynomial kernel with degree 3	Cluster size: [1] 49 539 12 Within-cluster sum of squares: [1] 2248.7512 9026.0559 950.9641	
	rbf kernel with sigma=0.2	Cluster size: [1] 59 328 213 Within-cluster sum of squares: [1] 886.2202 9094.9082 2174.9697	

For the combined variable set, polynomial kernels with k=3 and degree =3 could be considered as good clustering as within-cluster sum of squares values are lower with these combinations. Cluster 3 is better than cluster 1 and cluster 2 which means that data points are closely packed and are more relevant to the cluster 3.

<u>Davies-Bouldin's index for all sets of variables(Kernel k-means):</u>

Purchase of behavior		Basis for purchase		Combined variables	
k=2	1.987926	k=2	1.027648	k=2	1.174881
k=3	1.554175	k=3	2.42638	k=3	2.135912
k=4	1.727386	k=4	2.169107	k=4	2.335785
k=5	2.422906	k=5	1.491147	k=5	3.942959

We have calculated the Davies-Bouldin Index for all sets of variables. DB Index evaluates intra-cluster similarity and inter-cluster differences. Lower value of db index indicates good clustering. We see that the db index is lower for basis for purchase variable set (k=2) and purchase of behavior variables (k=4). Based on this we can say that these combinations will give us better clustering.

Question 4:

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

The quality of clusters depends on the clustering method, the number of clusters and distance measures used. We are using different sets of tuning parameters to obtain better clusters in each different method. Different distance methods used in our report are Euclidean, Manhattan, Maximum. We have also analysed several clustering methods to obtain better information.

Observing the best clusters from each Method for the three types of segmentations:

1) **Purchase_Behavior Results:**

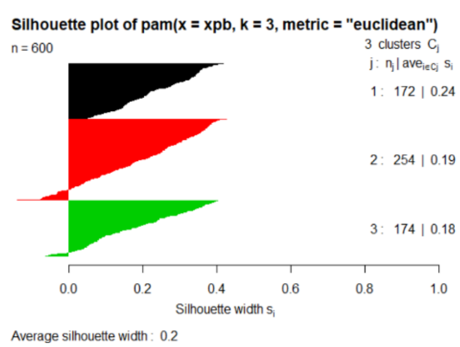
1) K-Means

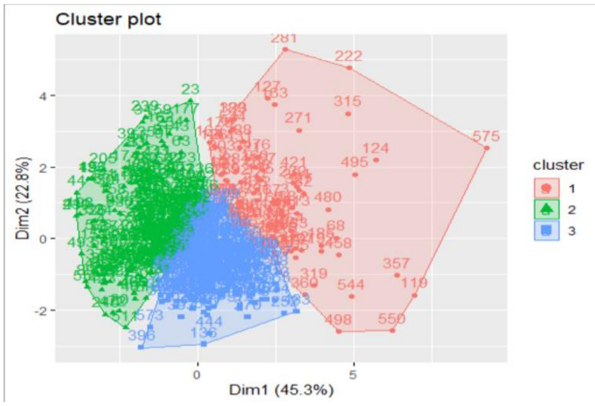
k=3, nstart= 60
3 clusters of sizes 259, 166, 175

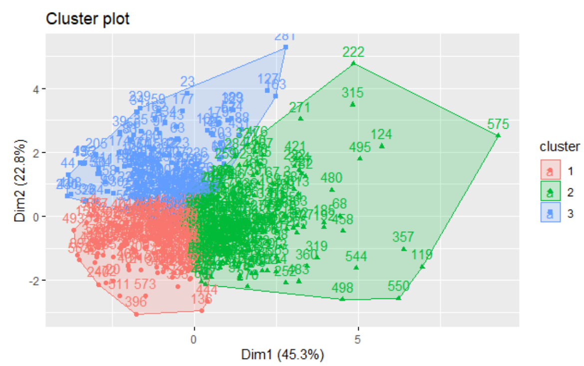
Within cluster sum of squares by cluster:
[1] 1242.739 1141.758 1585.147
(between_SS / total_SS = 33.7 %)

2) PAM

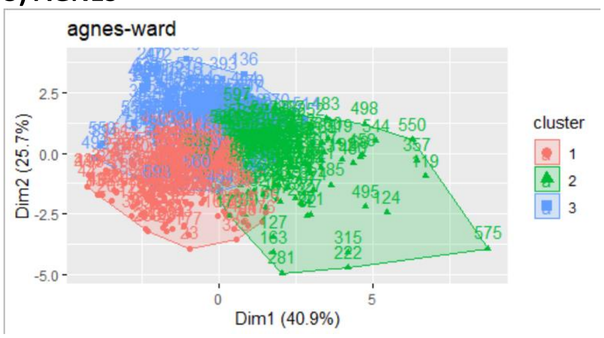
k=3
Distance Measure =euclidean
Average silhouette width of 0.2.





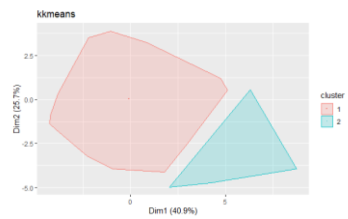


3) AGNES



k=3,
Agglomerative coefficient=0.98

4) Kernel K-Means



k=2, polynomial degree=3
Cluster size:
[1] 591 9

Within-cluster sum of squares:
[1] 4276.7977 409.4864

2) Basis of Purchase Results:

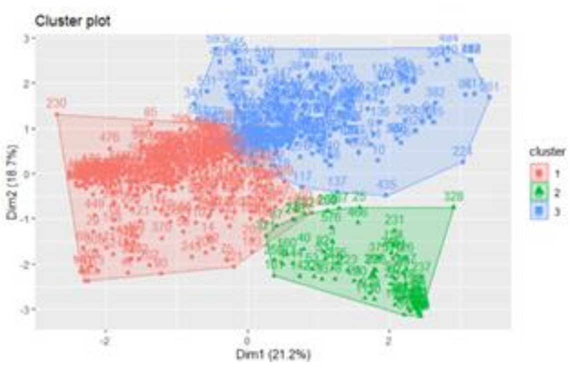
1) K-Means
k=3
nstart= 60
3 clusters of sizes 81, 215, 304

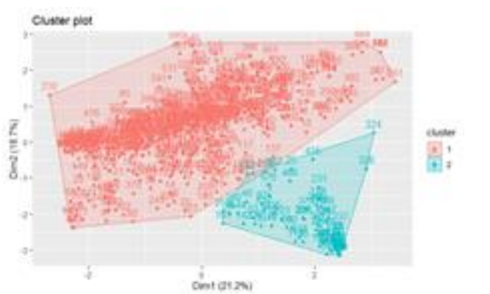
Within cluster sum of squares by cluster:
[1] 194.8455 1551.5754 2083.7473
(between_SS / total_SS = 29.0 %)

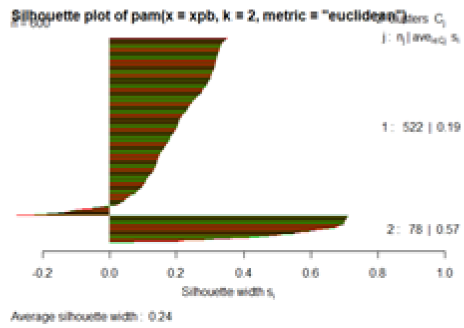
2) PAM
k=2

distance measure= euclidean

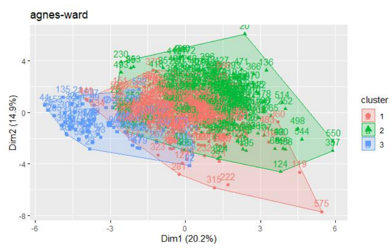
average silhouette width of 0.24





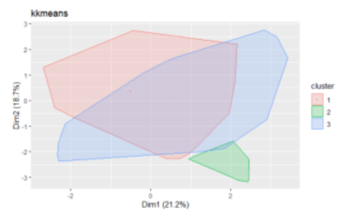


3) AGNES



k=3,
Agglomerative coefficient=0.98

4) Kernel K-Means



k=3, polynomial degree=2
Cluster size:
[1] 378 55 167

Within-cluster sum of squares:
[1] 2637.112 1318.375 2416.851

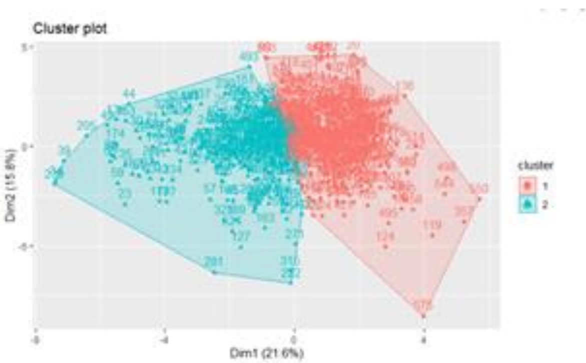
3) Purchase Behavior and Basis of Purchase results:

1) K-Means

k=2
nstart= 60

2 clusters of sizes 360, 240

Within cluster sum of squares by cluster:
[1] 5628.967 4165.342
(between_SS / total_SS = 13.9 %)

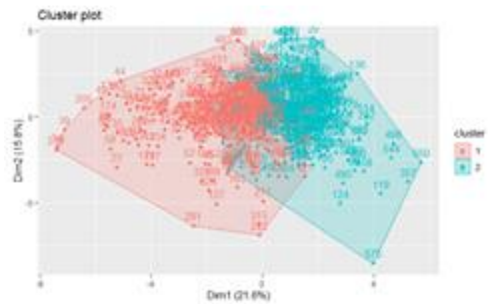


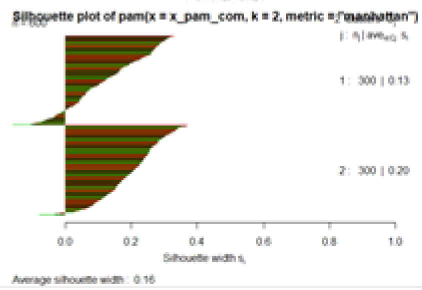
2) PAM

k=2

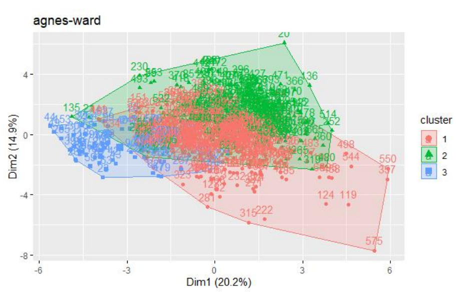
distance measure= manhattan

average silhouette width of 0.23



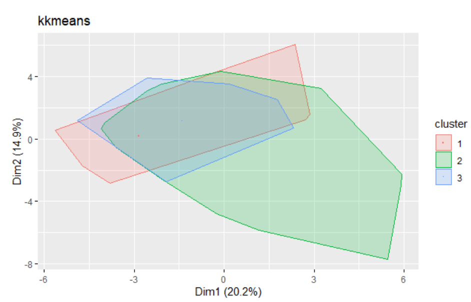


3) AGNES



k=3,
Agglomerative coefficient=0.97

4) Kernel K-Means



k=3, polynomial degree=3
Cluster size:
[1] 49 539 12

Within-cluster sum of squares:
[1] 2248.7512 9026.0559 950.9641

K means and k medoids (PAM) are different in the clusters and in the approach to build clusters. 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.

(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).

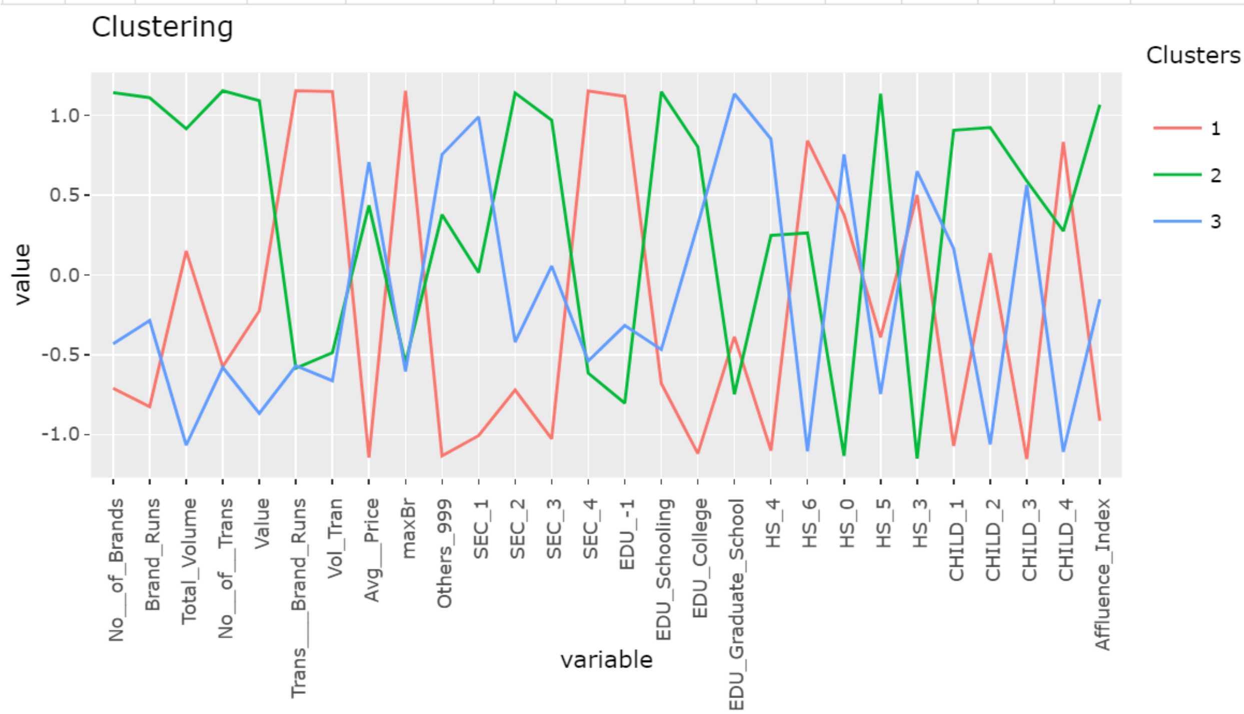
Based on the analysis above we would pick k-means clustering methods to choose the best segmentation.

The different segmentations by k-means are interpreted as follows:

1) Purchase_Behavior Results:

	clusKM	No__of_Brands	Brand_Runs	Total_Volume	No__of__Trans	Value	Trans___Brand_Runs	Vol_Tran	Avg__Price	maxBr	Others_999	SEC_1	SEC_2	SEC_3
1	1	5.138554	27.180723	16856.536	50.01205	2015.0398	1.954194	367.7342	12.31128	0.2350520	0.5977529	0.2409639	0.3012048	0.2650602
2	2	2.857143	8.228571	13349.429	23.98286	1289.2897	4.200463	560.6643	10.11735	0.7250765	0.1865303	0.1657143	0.2228571	0.2342857
3	3	3.200772	13.509653	7778.097	23.91120	935.5581	1.973567	346.9900	12.68932	0.2193920	0.7000982	0.3127413	0.2355212	0.2509653

SEC_4	EDU_-1	EDU_Schooling	EDU_College	EDU_Graduate_School	HS_4	HS_6	HS_0	HS_5	HS_3	CHILD_1	CHILD_2	CHILD_3	CHILD_4	Affluence_Index
0.1927711	0.09638554	0.6686747	0.1867470	0.01204819	0.2530120	0.12048193	0.02409639	0.3012048	0.09036145	0.13253012	0.2771084	0.10843373	0.4578313	20.99398
0.3771429	0.23428571	0.4800000	0.1200000	0.01714286	0.1714286	0.13714286	0.13142857	0.2228571	0.13142857	0.05714286	0.2514286	0.08571429	0.4742857	13.86286
0.2007722	0.13127413	0.5019305	0.1698842	0.03861004	0.2895753	0.08108108	0.15830116	0.2046332	0.13513514	0.10424710	0.2123552	0.10810811	0.4169884	16.60618



Cluster 1: This cluster shows a medium peak in Total volume of Bath Soap purchases. The average transaction per brand run and the Avg. volume per transaction is highest for this group with the highest Brand Loyalty compared to clusters 2 and 3. In terms of demographics, the population of this segment mainly belongs to the SEC_4 i.e the lowest socio-economic class, having minimum education (i.e from no education to Up to 4 years of school). This is also obvious by the Affluence Index which is the lowest among the three segments. The household size of this group is high along with an average of 2 to 4 children. It is also noted that television availability for this group has a higher number of non-availability than availability.

Cluster 2: This segment shows the highest in the No: of brands purchased, Number of instances of consecutive purchase of brands, Total volume and sum of value compared to the other clusters. We observe that the average Price of purchase is also high. This analysis is obvious as we see that the brand loyalty for this group is pretty low. In terms of demographics, the population of this segment mainly belongs to the SEC_2 and SEC_3 which are the middle class families, having average education of schooling as well as college. In terms of family size this segment has an average of 1-2 children per family. The Affluence index of this segment is the highest.

Cluster 3: This segment shows the lowest in Total Volume purchases, Number of purchase transactions, Value, Avg. transactions per brand run and Avg. volume per transaction. The number of brands purchased and the Brands runs are quite low for this group. The brand loyalty is low and is about the same level of cluster 2. In terms of demographics, the population of this segment belongs to SEC_1 which is the highest of all the socio-economic conditions but with a medium Affluence Index. This segment has the highest education level of graduate School and professional degrees.

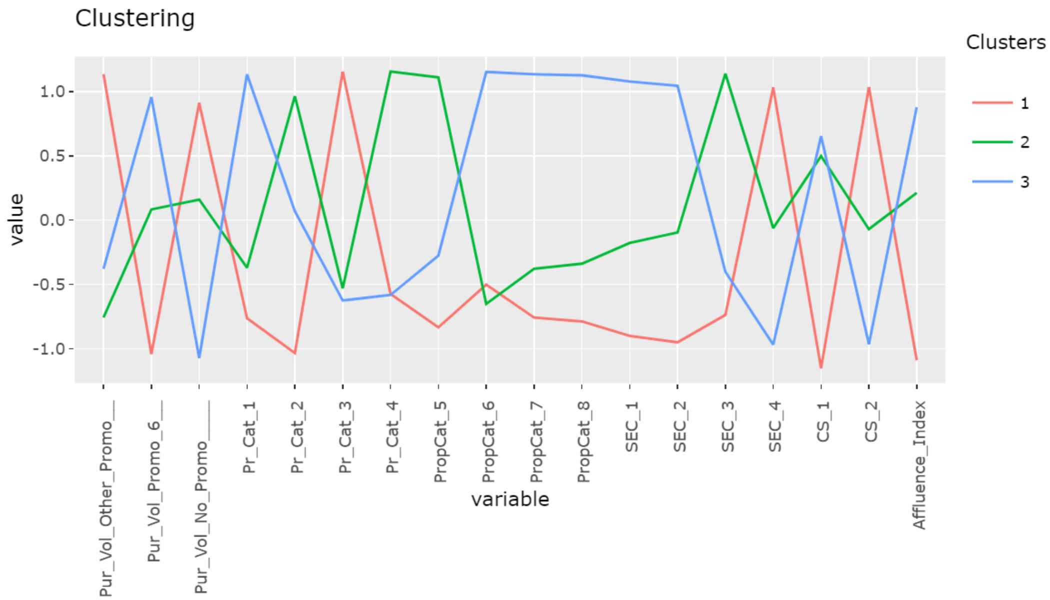
In conclusion, from the above analysis we see that cluster 1 which is of lower socio economic conditions has the maximum brand loyalty as well as highest volume as well as highest average transaction per a single brand. Whereas, cluster 2 and cluster 3 are from middle and higher socio economic conditions, having better educational status compared to cluster 1 but having lower brand loyalty.

Based on this different marketing strategies can be developed to target each of these segments separately.

2) Basis_for_Purchase results

	clusKMbp	Pur_Vol_Other_Promo__	Pur_Vol_Promo_6__	Pur_Vol_No_Promo___	Pr_Cat_1	Pr_Cat_2	Pr_Cat_3	Pr_Cat_4	PropCat_5	PropCat_6
1	1	0.03041806	0.04892563	0.9206563	0.15697596	0.6344386	0.05875599	0.14982946	0.6779938	0.06234805
2	2	0.04528756	0.01714400	0.9375684	0.05807375	0.1563918	0.75937397	0.02616046	0.1124928	0.06905901
3	3	0.03339350	0.07366496	0.8929415	0.53487133	0.4202229	0.01930871	0.02559711	0.2747735	0.14347237

PropCat_7	PropCat_8	SEC_1	SEC_2	SEC_3	SEC_4	CS_1	CS_2	Affluence_Index
0.047831385	0.04471701	0.18421053	0.2302632	0.2894737	0.29605263	0.7631579	0.10855263	16.648026
0.009857612	0.01043611	0.04938272	0.1728395	0.1975309	0.58024691	0.5185185	0.18518519	8.975309
0.199099130	0.15651031	0.41860465	0.3069767	0.2139535	0.06046512	0.7860465	0.04651163	20.576744



Cluster 1: This segment shows the highest volume of purchases for Other promotions and no promotions but with the lowest volume of purchases with promotion_6 (Branded Offers). In terms of price categories, this group is inclined the highest towards price category 3 and lower inclination towards price categories 1, 2 and 4. In terms of Proposition categories, we don't observe any high inclination towards any one category. It may seem this is not a basis of their purchase. In terms of socio-economic conditions, this segment majorly belongs to SEC_4 which is of the lower socio-economic class. This is obvious with the lowest Affluence Index seen in the graph and non availability of television. In other words we see this segment more inclined towards buying soaps with promotions with lower prices, extra grammage, free gift, Value added packs and not inclined towards a particular brand.

Cluster 2: This segment shows medium volume purchases for with or without promotions with the highest inclination towards price categories 2 and 4. In terms of Proposition categories, we see highest inclination towards category 5 i.e Beauty. This group highly belongs to SEC_3, but we also see some presence in SEC_4 as well as SEC_2. This group is the middle class with a medium Affluence Index along with a decent exposure to television advertisements.

Cluster 3: This segment shows the highest volume of purchases with promotion_6 (Branded Offers), a medium volume of purchases for Other promotions and alternatively it is the lowest volume with no promotion. It has the highest average volume purchased for Price Category 1 and decreases with the Price Categories 2, 3 and 4. With respect to proposition wise purchase, this group is more inclined towards Proposition categories 6,7 and 8 which are respectively Health, Herbal and Freshness related soaps. This segment belongs to SEC_1 and SEC_2 which are higher socio-economic conditions. This is obvious with the highest Affluence Index seen in the graph. We also observe that television availability for this segment is the highest, hence it is the most exposed group with respect to marketing through televisions.

In conclusion, from the above analysis we see that cluster 1 which is of lower socio economic conditions and is inclined towards other promotions, price reductions, free gifts and not showing a particular inclination towards any particular proposition category.. Whereas, cluster 2 and cluster 3 are from middle and higher socio economic conditions, are more inclined towards branded offers and proposition categories of Beauty, Health, Herbal and Freshness.

Based on this different marketing strategies can be developed to target each of these segments separately.

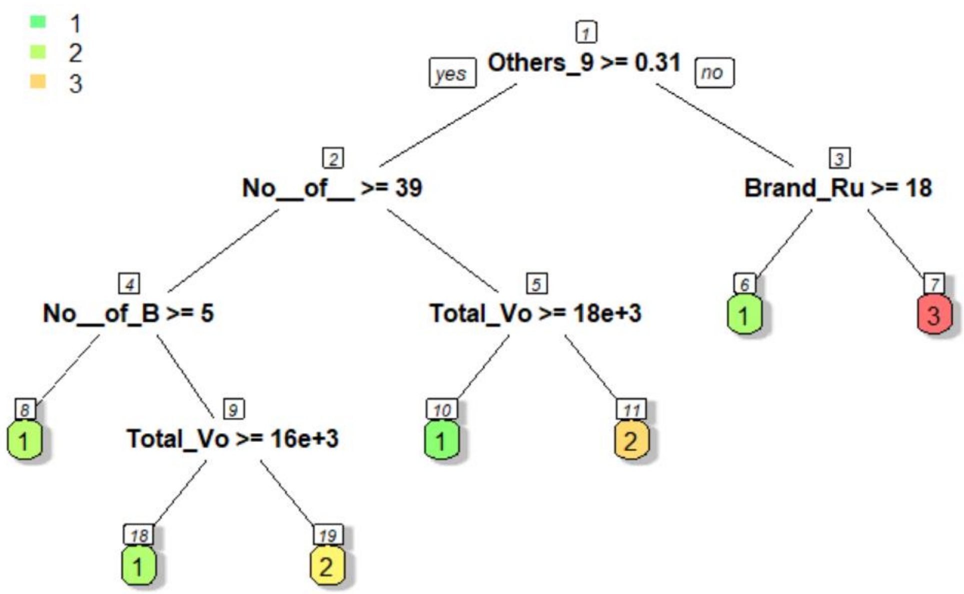
Conclusion:
Based on the above interpretations, for the purpose of marketing strategy any one of the above two basis of segmentation i.e purchase behavior or basis of purchase can be used.

Combining Purchase behavior and basis for purchase variables provided us with only 2 distinct clusters. As the value of k increased the segmentations were less distinct and a high overlap was developed, hence this has been disregarded.

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

Here we attempt to build a decision tree in order to interpret the clusters for one segmentation. The best segmentation obtained so far is for Purchase-Behavior.

<u>**Decision Tree:**</u>



Training Data Accuracy = 90%

```
          Reference
Prediction   1   2   3
         1  93   2  12
         2  16 177   6
         3   1   4 109
```

Test Data Accuracy = 84%

```
          Reference
Prediction  1  2  3
         1 39  2  3
         2 16 70 13
         3  1  4 32
```

The Decision tree was reasonably effective to classify the data into the three clusters. However the accuracy of classification is not a 100%, we obtain an accuracy of 84% on test data. By converting the tree to decision rules, a company can target the required market segment by following the tree rules.

We observe out of all the variables in the dataset, only the following were used to develop the decision tree:
1)Others_999
2)No__of__Trans
3)Total_Volume
4)No__of_Brands
5)Brand_Runs

We also observe that the above variables were used in the variable set for the purpose of clustering the market based on Purchase Behavior. However we also see that many variables have not been used such as Brand Loyalty. We also see costly variables related to demographics such as SEC, EDU, CHILD, CS are not used in building the decision Tree. Although these demographic variables can be used in the interpretation of the clustering, decision tree analysis provides a much cheaper way of classifying the groups. But there is a trade-off with lower accuracy hence a certain amount of cost is lost here.

References:

"Customer Segmentation via Cluster Analysis", Optimove, Mobius Solutions, 2020.
https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis

"Introduction to Segmentation and Clustering", Towards Data Science, Ifeoma Ojialor, December 2019.
https://towardsdatascience.com/introduction-to-segmentation-and-clustering-703b2ad2578a