Danielle Strejc, Olga Berezina, and Leticia Garcia
IDS 572- Fall 2020

Key Problems

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:

1. Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)
2. Basis of purchase (price, selling proposition)

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.

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.

Measuring Brand Loyalty

Several variables in this case measure aspects of brand loyalty. The number of different brands purchased by the customer is one measure. However, a consumer who purchases one or two brands in quick succession, and then settles on a third for a long streak is different from a consumer who constantly switches back and forth among three brands. So, how often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands – a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands. All three of these components can be measured with the data in the purchase summary worksheet.

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

Clustering approach

We will consider clustering based, first, on variables that describe purchase behavior, and then, based on variables that describe basis-for-purchase. A third clustering will then consider both sets of variables.
A key question is the number of clusters to consider – this can be based on how the clusters will

be used. It is likely that the marketing efforts would support 3-7 different promotional approaches. For clusters based on purchase behavior variables alone, or on basis-for-purchase variables alone, the fewer variables may support only 2-4 clusters. Clustering on the combined variables may allow for a higher number of useful clusters.

Remember – clusters are useful only so far as they carry a useful interpretation. And remember the business goal. Given the business goal, it is useful to consider demographic variables in addition to the variables used in clustering, for effective interpretation.

<u>Questions</u>

1. What is the business goal of clustering in this case study? Describe how you will use the data provided - household demographics, purchase behavior, basis-for-purchase. Which are the variables that describe purchase behavior, and those that describe basis-for-purchase? Describe your overall approach for clustering -- you do not need to talk about different clustering methods now; write about your approach for determining number of clusters, how you will evaluate alternate clustering, etc.

The business goal is clustering to determine which demographic variables affect purchase behavior. By summarizing demographics against purchasing, we will be able to see customer patterns and understand where most of the purchases are being done. We look at the data to try to determine the characteristics of households who are more loyal to a certain brand. Customer loyalty is measured by how often a customer purchases from a certain brand. If they are constantly switching from brands, then that is not considered loyalty. The implications of this analysis are that companies may be able to target their ads to certain demographics to promote their products. The variables that describe purchase behavior are: number of brands purchased, number of runs where they purchase the same brand, volume of product purchased, number of transactions, value in paise (local currency), transactions per brand run, volume per transaction, maximum of purchase by different major brands, and average price. Then for basis for purchase we use: purchase volume when there was a promotion versus when there was not, brand codelist for certain brands, promotion types, and proposition types. Using the demographics given to us, we can analyze which household, gender or even education status buy more when things are in promotion compared to when it's not. Also we can determine within the demographics which promotions are not working based on the number of return or percentage we received. This allows the company to set their promotion budget effectively.

For k-means, we use the elbow and silhouette methods to determine the optimal number of clusters. Here, we evaluate the models with average within- cluster sum of squares. This tells us the distance between each point and the center of the cluster. We want a smaller number, since that would indicate that the cluster has low variance. K means models tried vary k values to see how many clustering was best in terms of visuals and values it was giving back. Plotted variables against each other to look at the clustering for them.

For DBSCAN, we will look at how many clusters are created and the number of noise points(outliers) there are. This will show us how far apart the points are from each other and how many points are within each cluster. We use eps and minPts with DBSCAN. Eps is used to tell us how close the points need to be near each other in order to form a cluster. They should be neighbors. MinPts is used to tell us the minimum number of points that need to be used so that a dense region is formed. When trying to determine what the best parameters are, we need to look at how many outliers there are and how many points are there in each cluster.

For kernel k means, we compare cluster sizes and determine the optimal by looking at the average within-cluster sum of squares. To determine the optimal number of clusters for kernel k means, we look mainly at the average WCSS and to a lesser extent the plots that are made for each model.

2. Explore the data.

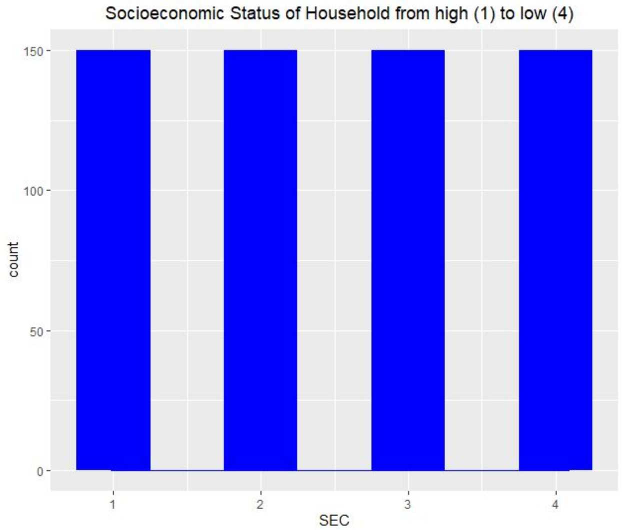
We utilize three individual datasets where we:

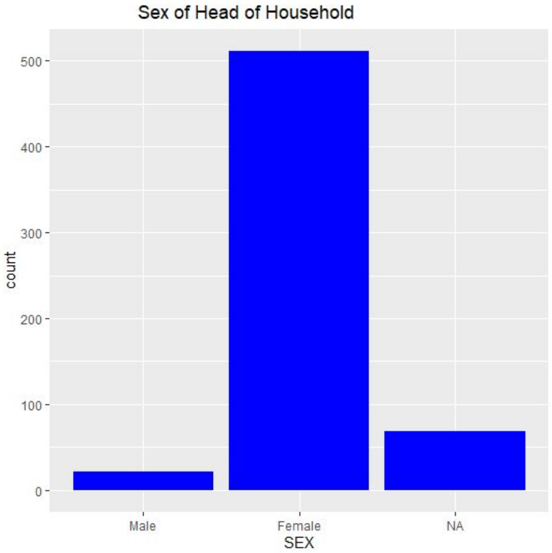
1. Put zeros in for missing values
2. Put median values for missing values

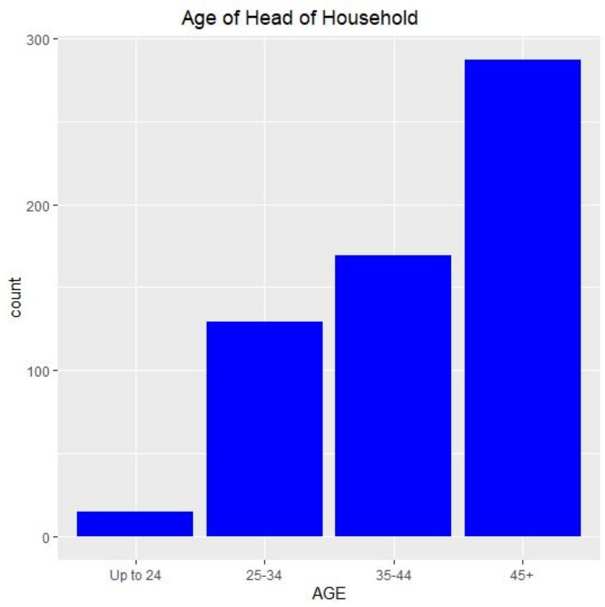
The reason we have different datasets is because we wanted to evaluate and see if having different values for missing values had an effect on the models.

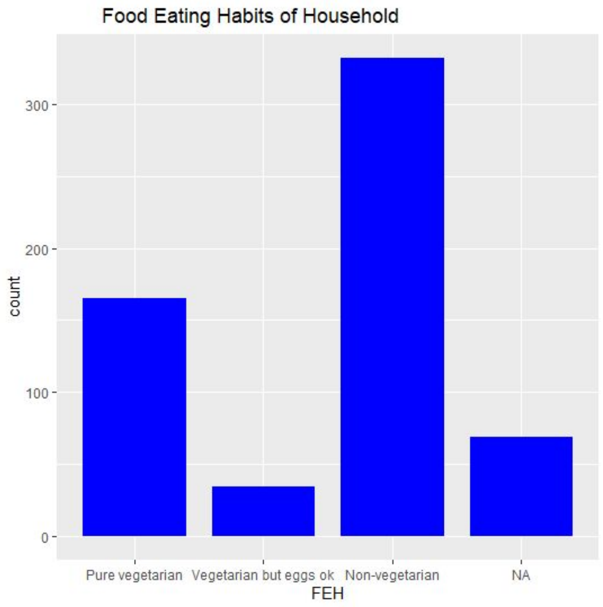
As we can see from the summary histograms, there are a significant number of NA's in six columns: food eating habits, mother tongue, sex, education level, household size, and television availability. To address these, we try three different methods- replace these categorical variables with 0, their median, or omit them altogether.

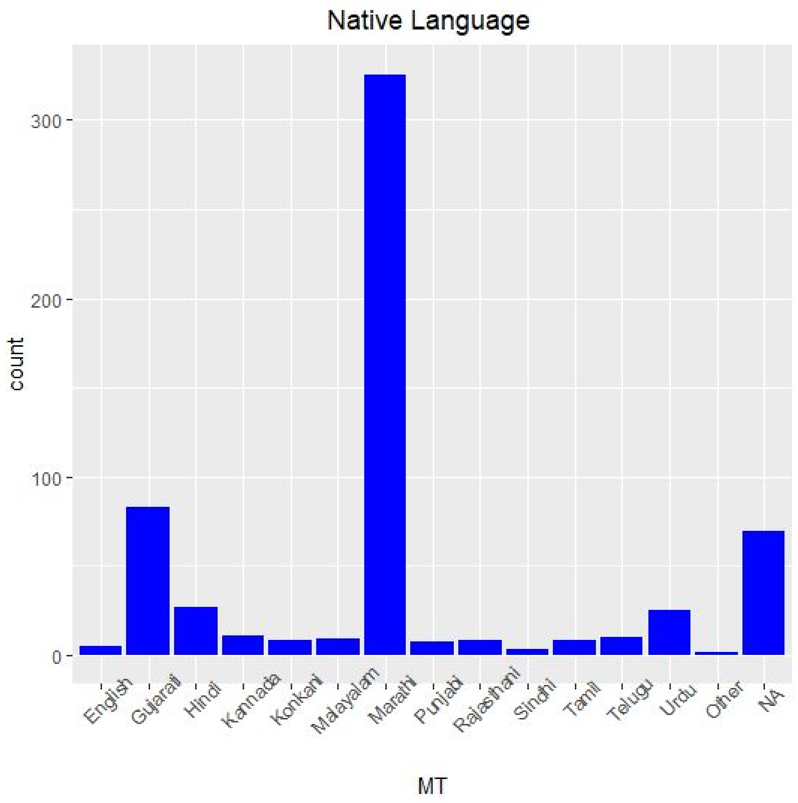
The households are overwhelmingly led by women, and that socioeconomic status is represented evenly across all four levels in the dataset.







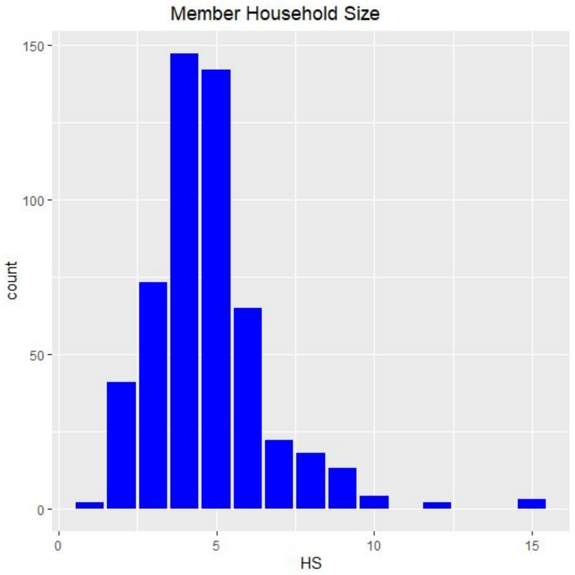


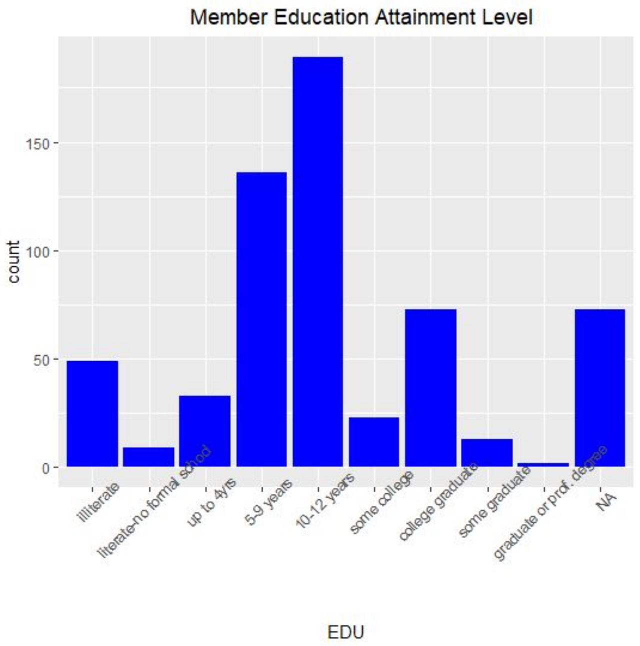


Households tend to be led by someone over 45 years, and most families were vegetarian or not vegetarian.

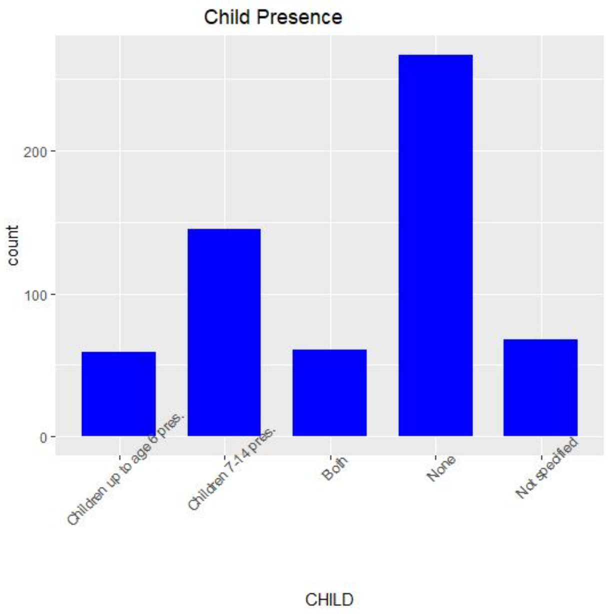
We see that there are quite a few languages that are not a significant proportion of the dataset, so we choose to only focus on a few of the languages and make them dummy variables so that the clusters could be formed more easily.

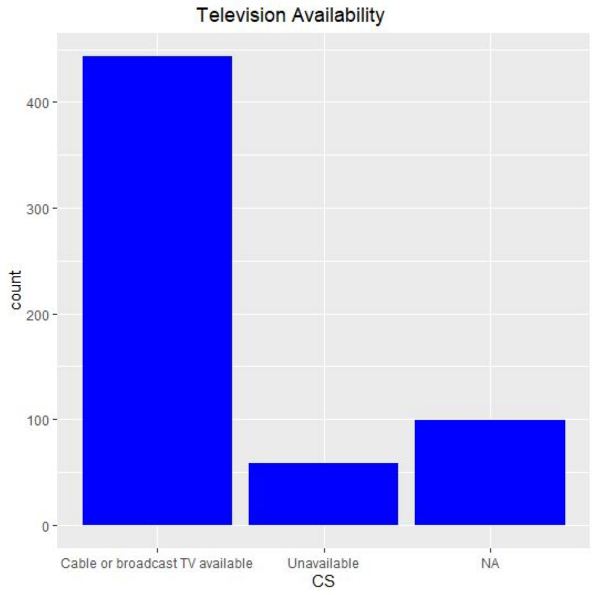
Below we see that household sizes tend to range 3-6 individuals. Education attainment is frequently somewhere between 5 and 12 years of schooling.

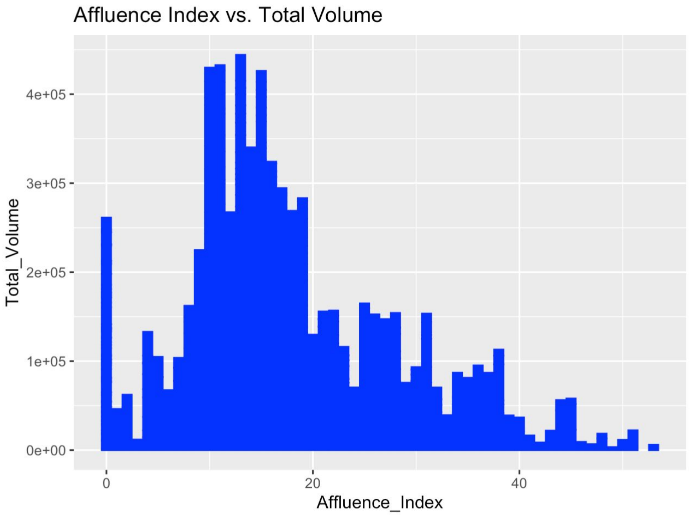




Below we consider whether or not there is a child present, and their age if so. Most commonly, there were no children at home or they were between the ages of 7 and 14. A majority of the households also have television availability where they live.

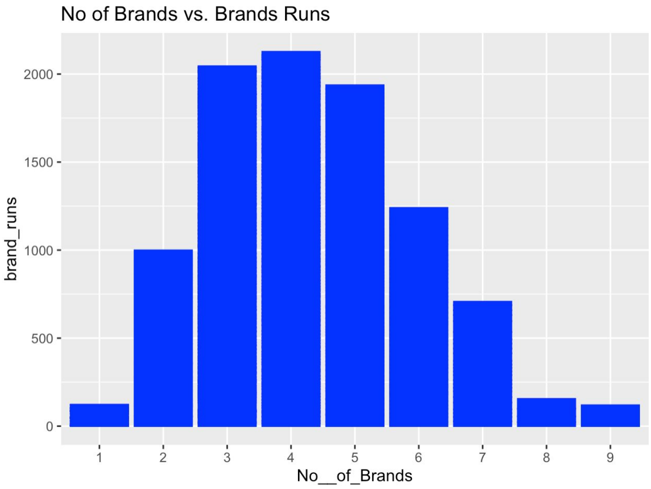


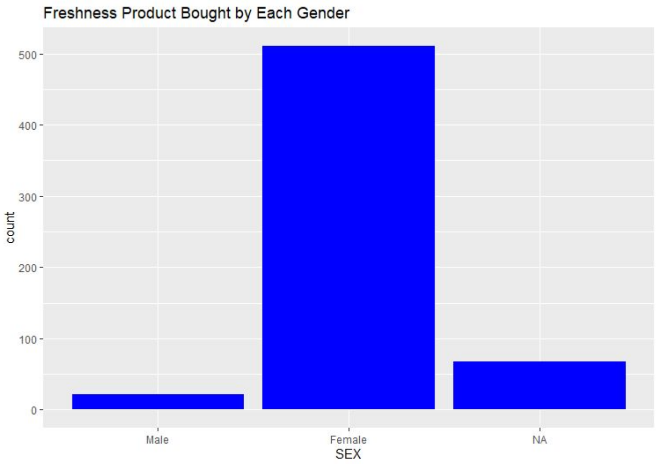




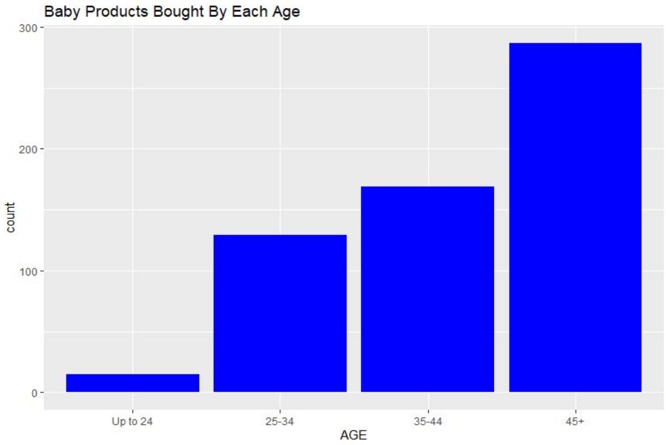
In the graph above, we are looking at the affluence index versus the total volume of purchases. Affluence index is the weight value of durables possessed, in other words individuals or household's economic and financial advantage in comparison to others. We see from the graph above that the most financial advantage doesn't seem to be buying a lot of products compared to those in the lower to middle section where we see that there's a high number of volume being purchased. This indicates that the lower to middle families index have a higher brand loyalty compared to higher financial families. When we reference the appendix of graphs we see that we have graphs that have a count of every family and their affluence index and we do see that we have a low count of higher income families which may affect the graph above.

When looking at the data we thought it was important to look at the purchasing history against demographic variables and see if we found any similar trends when looking at how many products were purchased or how frequently they brought products. Looking at the Socio Economic Class against purchasing and promotional sales there's a constant across all them, like the graph below dealing with health products purchased there's isn't much difference between each class. Looking at the number of products being purchased graphs that we see in the appendix pages, there's a huge amount of people purchasing low numbers of products and also when looking at how many frequently brought the same numbers of products we see that it's the same numbers of low products reappearing again. Higher in age you are the more likely to purchase compared to younger people and people with middle education also bought the most. When looking at the summary of demographic variables against purchasing and promotional sales, there's a clear indication that there's high correlation between few of them.

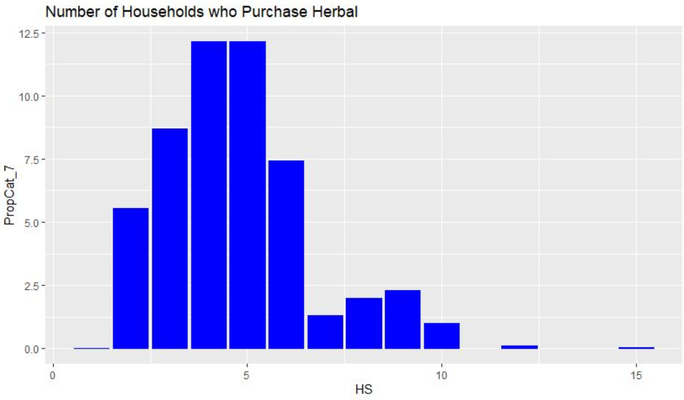

No of Brands vs. Brands Runs


Freshness Product Bought by Each Gender

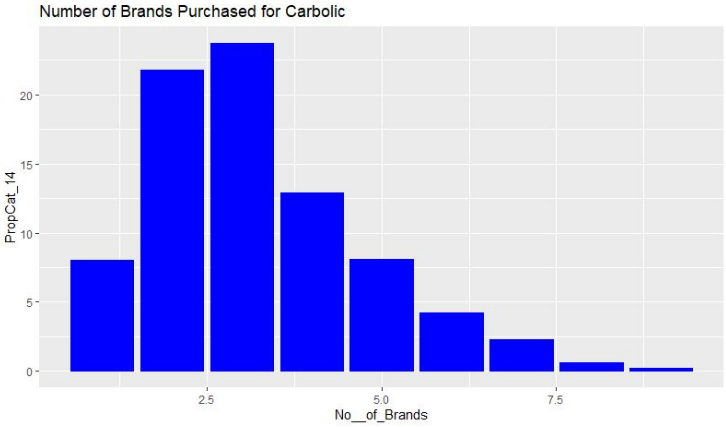
For this bar chart, we can see who is more responsible for purchasing household cleaning items and the amount being purchased. The majority of females seem to make the decision in the households compared to men.


Baby Products Bought By Each Age

This bar chart shows which age purchases baby products more. We can see that as age increases, so does the amount of baby products bought. This could either mean that people are having children later on in life or grandparents are the ones purchasing the products.



From these results, we can see that there is a medium amount of households that purchase herbal items. There are about 4-5 households who mostly purchase herbal items. They also have the highest count. There is not too much of an even distribution. However, it appears that as you increase households, the lesser number of herbal items are purchased.



Based on this chart, we can see that there are a few different brands that households can purchase from. There are mostly 2-3 different brands a household can pick from a company. This also tells us that households would be more loyal to a company who has more options to choose from their brands. In addition to this, CRISA would have a better understanding of which brands to focus on for better promotion deals.

3. Use k-means clustering to identify clusters of households based on:
a. The variables that describe purchase behavior (including brand loyalty).
[Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)].
[Q – how do you measure brand loyalty?]

Purchase Behavior

One way of measuring brand loyalty is looking at the demographic against purchasing history. In our summary page, we found trends within each demographic that was able to see which group brought the most and as well as which group brought it frequently afterwards. This allows CRISA to see which customers they need to target more frequently as it is important to keep old customers. Also, the company can perform a NPS(net promoter survey). In this survey, CRISA can ask different types of questions such as:

- How likely are you to recommend this brand to people you know?
- Would a loyalty program increase your chances of buying from this brand again?
- How satisfied are you with this product?
- Would you consider switching brands if you found a more affordable one?

After customers have completed the survey, the company can determine who is the most loyal based on their ratings. For this dataset in particular we use brand runs- where the same brands were purchased within a “run”. Brand loyalty is measured by the likelihood that a household stays with one brand and does not switch to others, so using brand_runs is significant. The other variables influence the ability to make clusters, but brand_runs is the one that indicates brand loyalty the most. Another variable that used to measure brand loyalty is total volume purchased. The more volume of purchases a person has the more likely that person is loyal to the products compared to someone who does not. Same can be said for transactions and avg price (which is part of the dataset) the more a person tends to have the more likely you can consider a person to be loyal to the products.

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

We would not want to sell all the promotions based on the results of the summary we had. We see that certain promotions gave zero returns on them compared to other promotions. We would exclude these in the future. CRISA should provide promotions to customers for a specific brand that they are loyal to. This would make the customer keep coming back to purchase it if they know they are receiving coupons for it. It was found that householders purchase cleaning items the most such as soap, and other cleaning products.These items were bought in a small bulk, which tells us there are multiple people living in a household. These are the types of items that CRISA needs to focus on for their promotions. For example, a company’s selling proposition for their soap could be “Provides the freshest smell, gets rid of all germs, and is organic” whereas their competitor would only say that their product gets rid of 90% of germs. Customers would obviously purchase the first brand as it looks like it provides the most benefits to them. The average price of products purchased is about 5 paise. There are a few products ranging between the 15-20 range. For general fresh products, households are willing to spend an average between 10-20. This is shown in the appendix document.

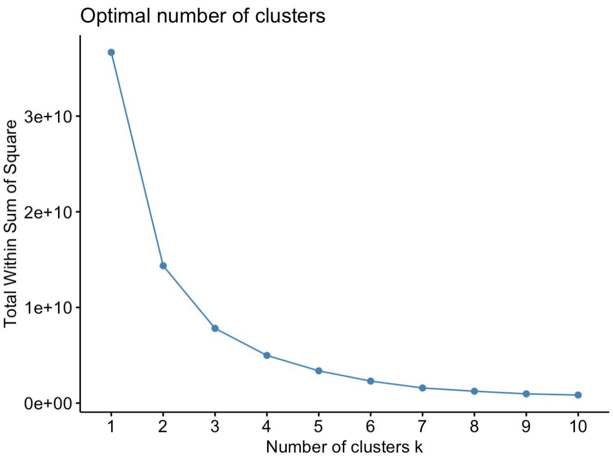
c. The variables that describe both purchase behavior and basis of purchase.

Socioeconomic status, affluence index, and household size are three variables that help describe purchase_behavior and basis_for purchase. A household’s economic standing can dictate what kind of products they purchase and the likelihood of buying when there’s a promotion. This leads to a better understanding for the company to see which brands customers are more loyal to. The company can also see how much each household is willing to spend according to their income. If customers only purchase the products when they’re on sale, that means the customer is not loyal to that brand. Instead, they are looking to purchase what is most affordable for them. In the affluence index charts we saw above, it was mentioned that households with more financial advantage did not purchase as much as those with less advantage. This conveys that CRISA should target middle income families more as they are their biggest customers. Household size is also another factor on how families purchase products. For example, multiple people can have a say in what to purchase as in what brand they prefer. Also, how much of a product needs to be bought. For example, a larger family will need to buy 5 packs of detergent whereas a smaller family will only need to buy 2 packs.

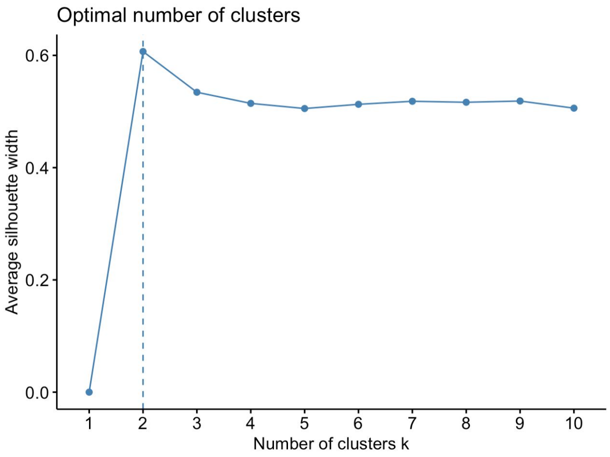
For each clustering in Q3 and in Q4 below:
(i) Describe your rationale for experimenting with different values of k.
(ii) Evaluate the clusters – based on generic performance measures for clustering.
(iii) Evaluate the clusters – based on the business problem and interpretation of clusters. Comment on the characteristics (demographic, brand loyalty and/or basis-for-purchase) of these clusters. This information will be used to guide the development of advertising and promotional campaigns.

(i)

For the k means models we determine the numbers of k values based on the elbow method and silhouette method to see the best optimized cluster. Below are the results of the graphs. We see the best cluster is at 2. Ultimately we decided to base the number of clusters relatively close to 2 clusters to see if we had a big difference between the models. The tables below are for both behavior purchase and basis purchase which row indicating the data frame we used as well the numbers of clusters we experimented with.



Optimal number of clusters



Optimal number of clusters

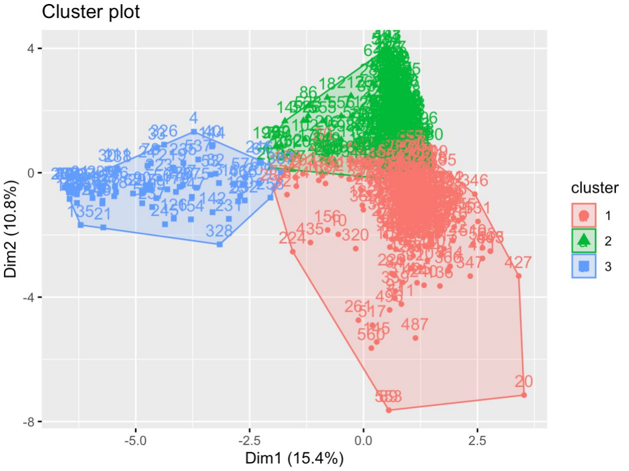
Purchase Behavior				
Model Setup / Evaluation				
Models	Data	Preprocess	K Values	between_SS / total_SS
Model 1	bsd1	No	3	33.7
Model 2	bsd1	No	5	49.3

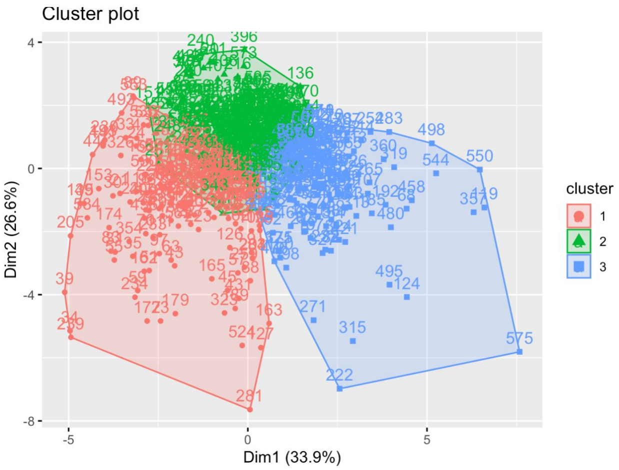
Model 3	bsd1	No	2	20.6
Model 4	bsd1PP	Yes	3	33.7
Model 5	bsd1PP	Yes	2	20.6
Model 6	df	No	3	33.7
Model 7	df	No	5	49.3
Model 8	df	No	2	20.6
Model 9	dfPP	Yes	2	20.6
Model 10	dfPP	Yes	3	33.7

Basis for Purchase				
Model Setup/Evaluation				
Models	**Data**	**Preprocess**	**K values**	**between_SS / total_SS**
Model 1	bsd1	No	3	19.8
Model 2	bsd1	No	2	12.8
Model 3	bsd1	No	5	30.4
Model 4	bsd1	No	4	25.2
Model 5	bsd1PP	Yes	3	19.8
Model 6	bsd1PP	Yes	2	12.8
Model 7	bsd1PP	Yes	4	25.2
Model 8	df	No	2	12.8
Model 9	df	No	3	19.8
Model 10	df	No	4	25.2
Model 11	dfPP	Yes	2	12.8
Model 12	dfPP	Yes	3	19.8
Model 13	dfPP	Yes	4	25.2

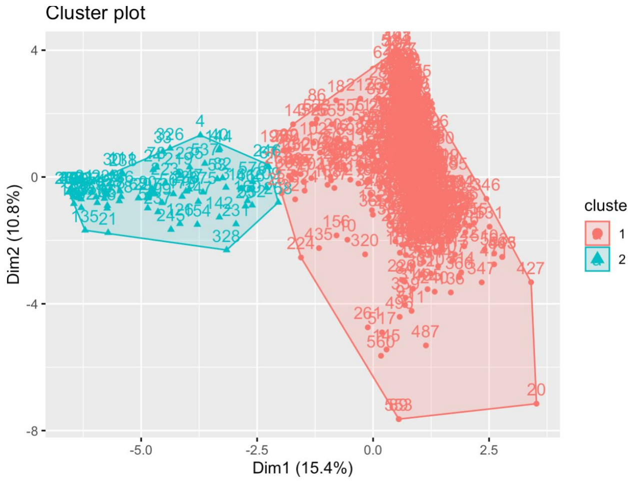
(ii)

In the above tables we show the different models perform as well the different values of k we used. We evaluate the model based on the WCSS values. We see that different data frames didn’t make a difference to the WCSS values nor did it make the plot of clustering any different. What we do see is if there's a difference in terms of k values and also we see a difference between purchase behavior or basis for purchase.





Results showed that WCSS value gets worse after 2 clusters and at the fifth clusters we start to see overlapping clusters which makes it difficult for analysis. Looking at the behavior purchase and basis for purchase there are different WSCC values for each number of clustering. We see the basis for purchase had lower WSCC compared to behavior purchase. We can see a difference in the plotting as well in the graphs above; we see the basis for purchase in the left have a more separation between each cluster compared to behavior purchase in the right where we see that the clusters are starting to overlap. The most optimal cluster was 2 and it had value 12.8 for basis for purchase which shown in the graph below.



(iii)

When looking at the clustering we are able to see the variables and the WSCC value for each clustering being processed in the tables below. We showed the clustering for 3 there’s one table for behavior purchase and basis for purchase. We show in the result below that demographic variables for tables are relatively the same and also have pretty same WSCC values as well. Which indicates if CRISCA wants to evaluate behavior purchase or basis for purchase they can use demographic variables for their decision making and marketing ideas. This doesn't come at a surprise based on the summary we found earlier.

Behavior Purchase

ClusKM	SEC	HS	SEX	EDU	Affluence Index	Age
1	2.339768	3.474903	1.644788	4.03861	16.06618	3.131274
2	2.822857	4.382857	1.702857	3.52000	13.86286	3.205714
3	2.409639	5.108434	1.921687	4.60241	20.99398	3.349398

CHILD	Maxbr	No of Brands	No of Trans	Brand runs	Total Volume	Trans Brands Runs
3.312741	0.5322054	3.200772	0.4155637	0.2156513	0.5323647	0.2473029
3.371429	1.2404272	2.857143	0.4114518	0.7236194	0.7236194	0.6076481
2.963855	0.4773106	5.138554	1.0821389	1.0993197	0.6359753	0.2547407

Basis for Purchase

ClusKM	SEC	HS	SEX	EDU	Affluence Index	Age
1	2.251701	3.989796	1.738095	4.309524	19.275510	3.289116
2	3.358974	4.230769	1.576923	2.461538	9.102564	3.012821
3	2.526316	4.438596	1.793860	4.241228	16.820175	3.184211

CHILD	Maxbr	No of Brands	No of Trans	Brand runs	Total Volume	Trans Brands Runs
3.261905	0.6635758	3.697279	0.1633456	0.2826735	0.1286162	0.1951789
3.435897	1.3089269	3.051282	0.3177751	0.6777811	0.4662258	0.9187686

3.127193	0.4078728	3.758772	0.1019174	0.1326275	0.1009382	0.0626374

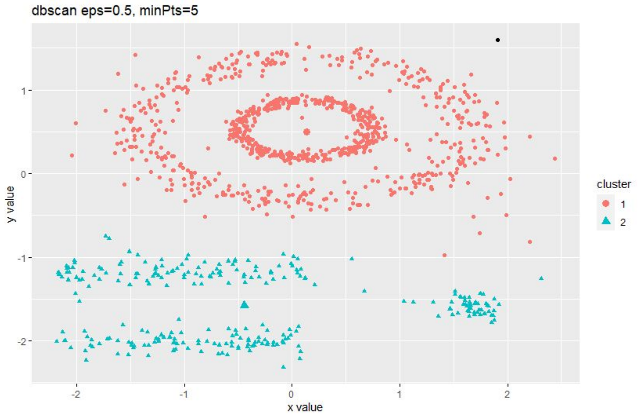
4. Try two other clustering methods (for a 2-person team, try one other method) for the questions above - from agglomerative clustering, k-medoids, kernel-k-means, and DBSCAN clustering. Show how you experiment with different parameter values for the different techniques, and how these affect the clusters obtained.

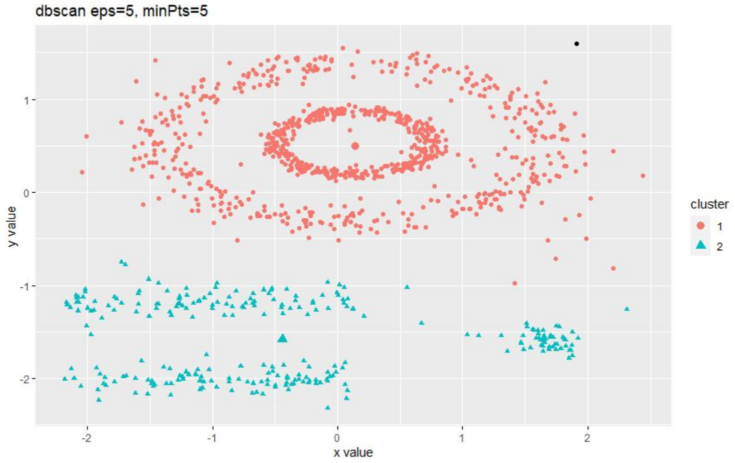
First clustering method: DBSCAN Clustering
Second clustering method: kernel k-means clustering

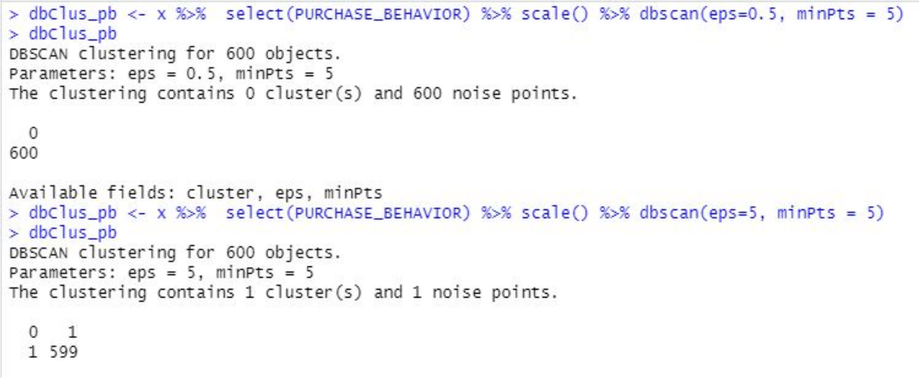
***The code below is to show how many points are specifically in each cluster with a certain eps and minPts

DBSCAN

Below are the DBSCAN graphs for Purchase_Behavior





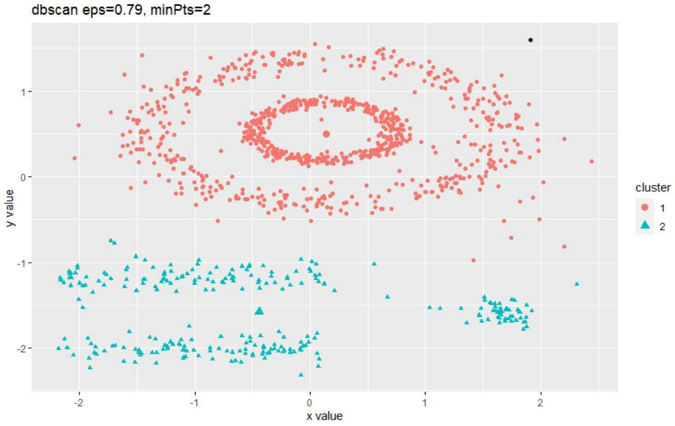


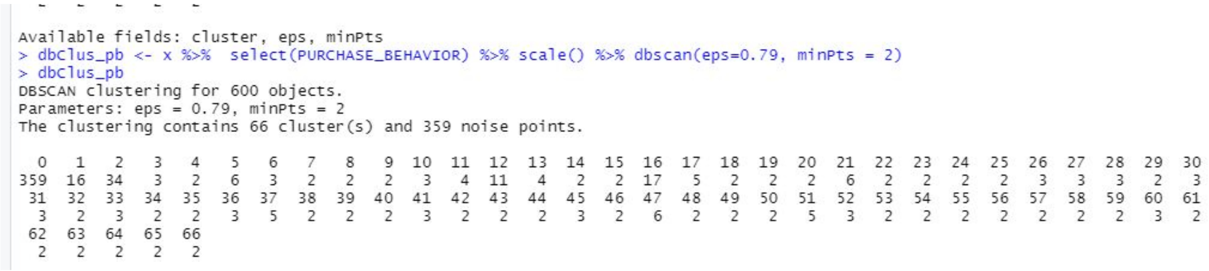
```
> dbClus_pb <- x %>%  select(PURCHASE_BEHAVIOR) %>% scale() %>% dbscan(eps=0.5, minPts = 5)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 0.5, minPts = 5
The clustering contains 0 cluster(s) and 600 noise points.

  0
600

Available fields: cluster, eps, minPts
> dbClus_pb <- x %>%  select(PURCHASE_BEHAVIOR) %>% scale() %>% dbscan(eps=5, minPts = 5)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 5, minPts = 5
The clustering contains 1 cluster(s) and 1 noise points.

  0   1
  1 599
```

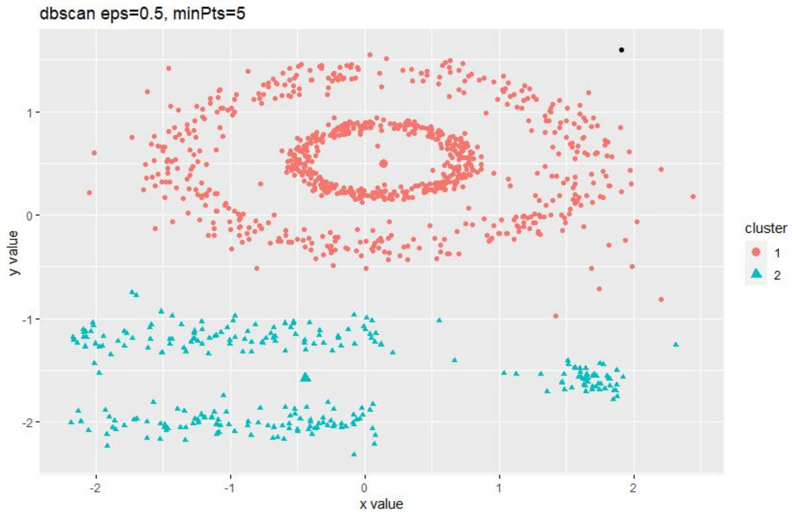


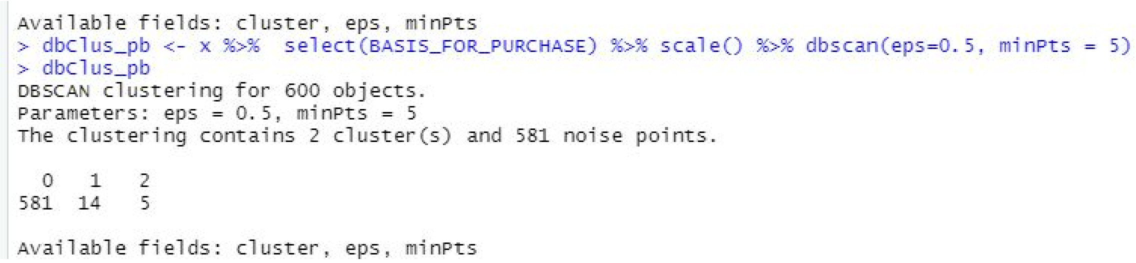


```
Available fields: cluster, eps, minPts
> dbClus_pb <- x %>%  select(PURCHASE_BEHAVIOR) %>% scale() %>% dbscan(eps=0.79, minPts = 2)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 0.79, minPts = 2
The clustering contains 66 cluster(s) and 359 noise points.

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30 
359  16  34   3   2   6   3   2   2   2   3   4  11   4   2   2  17   5   2   2   2   6   2   2   2   2   3   3   3   2   3 
 31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61 
  3   2   3   2   2   3   5   2   2   2   3   2   2   2   3   2   6   2   2   2   5   3   2   2   2   2   2   2   2   3   2 
 62  63  64  65  66 
  2   2   2   2   2 
```

Now, we will look at the DCSCAN graphs for Basis_for_Behavior:

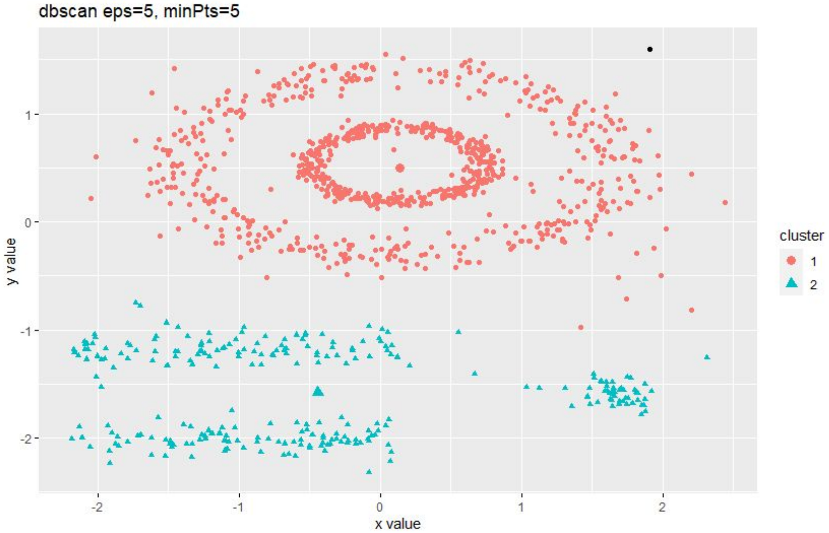


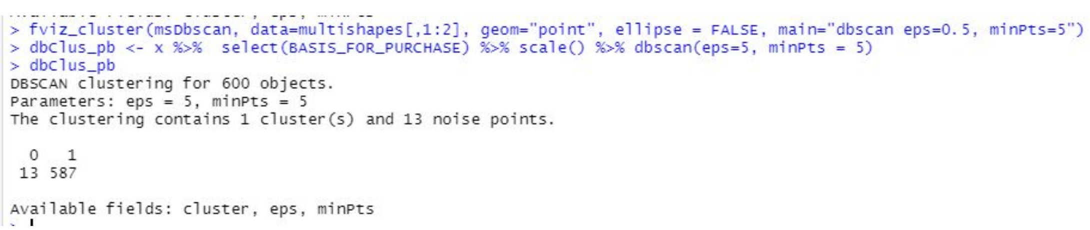


```
Available fields: cluster, eps, minPts
> dbClus_pb <- x %>%  select(BASIS_FOR_PURCHASE) %>% scale() %>% dbscan(eps=0.5, minPts = 5)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 0.5, minPts = 5
The clustering contains 2 cluster(s) and 581 noise points.

  0   1   2 
581  14   5 

Available fields: cluster, eps, minPts
```

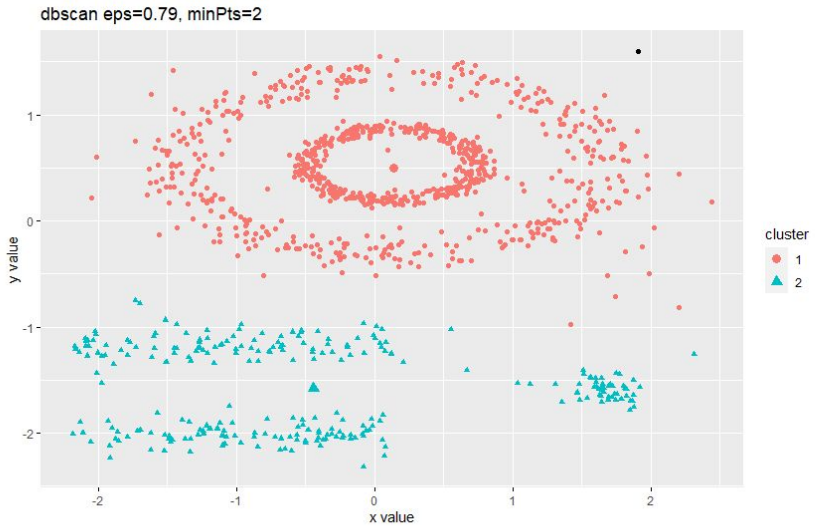


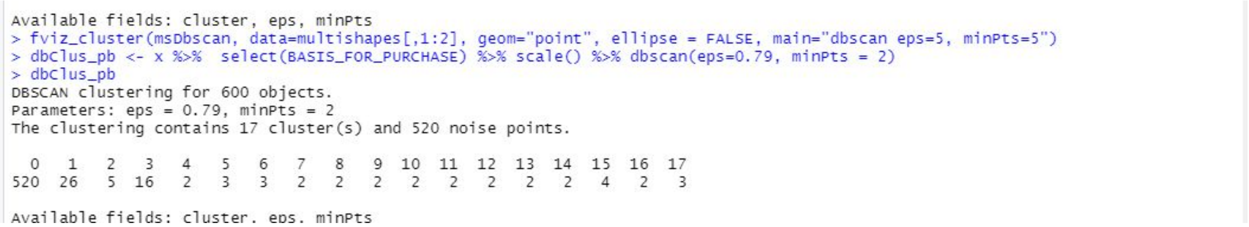


```
> fviz_cluster(msDbscan, data=multishapes[,1:2], geom="point", ellipse = FALSE, main="dbscan eps=0.5, minPts=5")
> dbClus_pb <- x %>%  select(BASIS_FOR_PURCHASE) %>% scale() %>% dbscan(eps=5, minPts = 5)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 5, minPts = 5
The clustering contains 1 cluster(s) and 13 noise points.

  0   1 
 13 587 

Available fields: cluster, eps, minPts
>
```




```
Available fields: cluster, eps, minPts
> fviz_cluster(msDbscan, data=multishapes[,1:2], geom="point", ellipse = FALSE, main="dbscan eps=5, minPts=5")
> dbClus_pb <- x %>%  select(BASIS_FOR_PURCHASE) %>% scale() %>% dbscan(eps=0.79, minPts = 2)
> dbClus_pb
DBSCAN clustering for 600 objects.
Parameters: eps = 0.79, minPts = 2
The clustering contains 17 cluster(s) and 520 noise points.

  0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
520  26   5  16   2   3   3   2   2   2   2   2   2   2   2   4   2   3 

Available fields: cluster, eps, minPts
```

Below are tables that show some of the eps and minPts values that were used to determine what values would provide the best solution. We could see that none of them do. Both Purchase_Behavior and Basis_for_Purchase provide a few clusters with very few points. The majority are noise points.

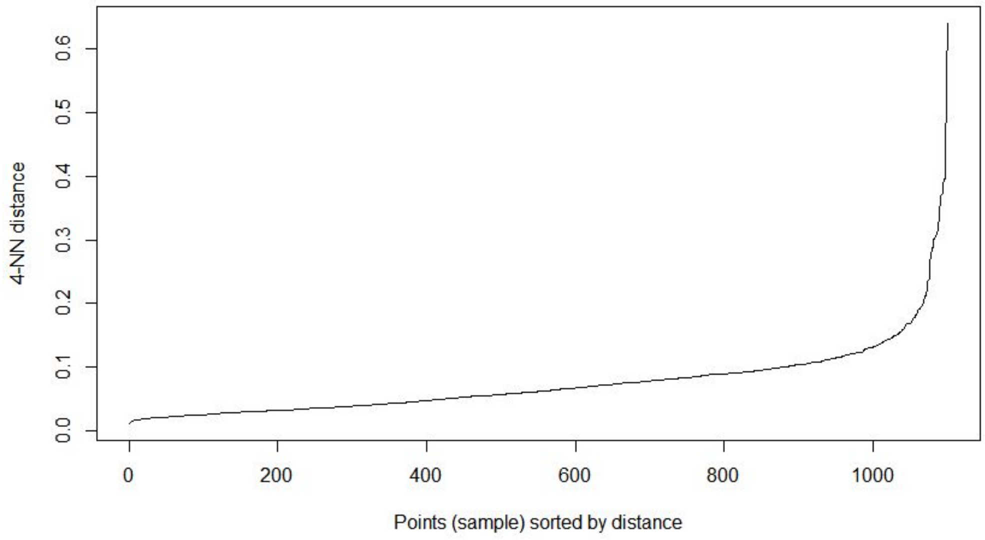
Purchase_Behavior:

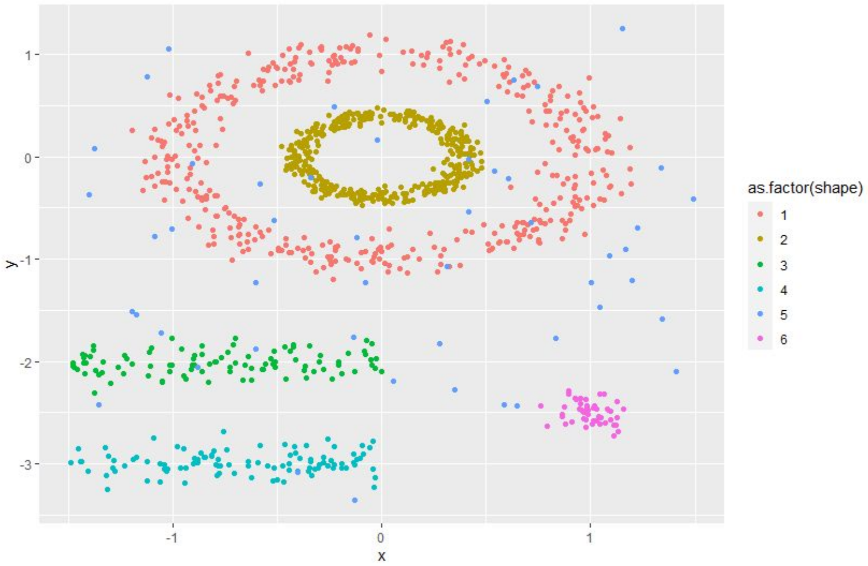
	Cluster 1	**Cluster 2**	**Cluster 3**
eps	0.5	5	0.79
minPts	5	5	2
Total clusters	0	1	66
Noise points	600	1	359

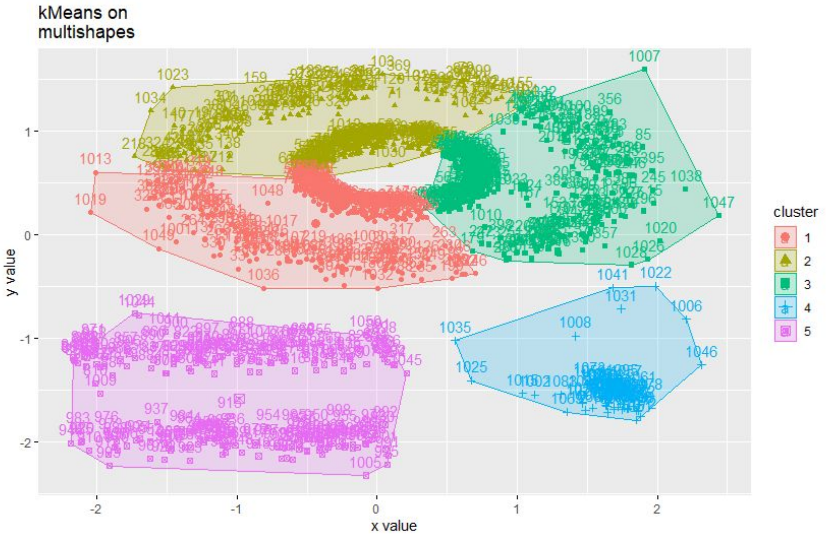
Basis_for_Purchase:

	Cluster 1	Cluster 2	Cluster 3
eps	0.5	5	0.79
minPts	5	5	2
Total clusters	2	1	17
Noise points	581	13	520

When using DBSCAN, we can see that it is not the best technique to use with this dataset. This is because DBSCAN doesn't work well with clusters of similar density. It only works effectively when you are trying to separate high and low density clusters. Also, DBSCAN requires you to carefully choose its parameters (eps, minPts). For example, if you look at the table and screenshots above for Purchase_Behavior and Basis_for_Purchase(located above), you can see that we tried to use different types of parameters for each cluster and were not able to reach a "nice" solution. There was a high number of outliers for all parameters that were tested. For example, in Purchase_Behavior, when we used eps=5 and minPts=5, there were 600 outliers and zero clusters. For Basis_for_Purchase, when we used the same parameters (eps=5 and minpts=5), there were only 2 clusters and 581 noise points(outliers). The same happens with the rest of the dataset.







Kernel K-means

We implement a kernel k-means clustering model to build on the k-means clustering. Kernel models project data into a non-linear feature space in order to potentially cluster those clusters that may not be linearly separable in the original space.

To get the optimal cluster number- we see the clusters corresponding to the lowest average within-cluster sum of squares. Two tends to be the best, but for some models it is closer than others. We decide to do models with 2 and 3 clusters to compare. 4 and 5 were almost never the optimal number.

Model Prediction	Sigma value	Best cluster (usually)
BFP	.005	2 and 3 very close
BFP	def	2
PB	.005	2
PB	def	2 and 3 close

Comparing average WCSS, we found that kernel = radial gives the best values consistently. We looked at using a linear kernel, but that did not improve the average WCSS, so we decided to stick to **kernel=radial and then compare between the different sigma values and 2 or 3 clusters**.

When we do kernel K means, the numbers are different every time, but we find that the differences between sigma = .005 and sigma as the default are so small that we just decide to go with the default since that seems to have more instances of better average WCSS. Here is a table showing how we came to that conclusion. At the end of this section is a table with the overall average WCSS that informed our decision.

Kernel KMeans with Purchase Behavior				
Model	**sigma**	**C1**	**C2**	**C3**
bsd1PP	.005	16.23	9.252	-
bsd1PP	.005	21.08	8.814	9.929
bsd1PP	default	15.06	6.828	-
bsd1PP	default	7.835	12.40	12.34

These are with the original dataset and the action is to omit any missing values. We find slightly better values with the default sigma, so we continue with that here. This is looking at 2 versus 3 clusters

Kernel KMeans with Purchase Behavior				
	sigma	**C1**	**C2**	**C3**
Bsd	def	7.430	17.57	-
bsd	def	18.13	9.465	6.999
bsdPP	def	6.457	16.44	-

bsdPP	def	8.952	23.06	5.153

Below we look at the dataframe where we set NAs=zero. We compare between no pre-process and pre-process with “scale”, and also clusters = 2 and clusters = 3.

Kernel KMeans With Purchase Behavior				
	sigma	**C1**	**C2**	**C3**
bsd1	def	7.029	16.48	-
Bsd1	def	11.84	16.07	7.843
bsd1PP	def	15.06	6.828	-
bsd1PP	def	7.835	12.40	12.34

This table is the one where we substitute the median value into the columns with blanks. Again we compare with and without pre-process and between 2 and 3 clusters.

Kernel KMeans With Purchase Behavior				
	sigma	**C1**	**C2**	**C3**
df	def	12.49	7.274	-
df	def	9.272	22.834	6.938
dfPP	def	15.32	6.406	-
dfPP	def	14.10	15.99	5.922

Finally, we compare the averages for all of the models illustrated above. In general, we find that the difference between linear and radial is quite small and likely insignificant since the averages are so close together. 2 clusters perform better than three, which agrees with what we assumed from earlier.

Kernel K Means- Purchase Behavior			
Models	**Number of clusters**	**Sigma value**	**Average WCSS**
Bsd1	2	def	11.52
-	3	def	11.88

bsd1PP	2	def	11.04
-	3	def	11.22
df	2	def	10.20
-	3	def	11.50
dfPP	2	def	10.98
-	3	def	11.16
bsd	2	def	11.72
-	3	def	11.94
bsdPP	2	def	11.17
-	3	def	11.49

Basis For Purchase

Below is a starting point for basis for purchase. We compare a dataset with 2 clusters vs 3. We already learned that the radial kernel was best for this problem also, so now we compare sigma values to see which is preferable.

<table>
<tr><th colspan="5">Kernel KMeans with Basis For Purchase</th></tr>
<tr><th>Model</th><th>sigma</th><th>C1</th><th>C2</th><th>C3</th></tr>
<tr><td>bsd1PP</td><td>.005</td><td>32.36</td><td>24.69</td><td>-</td></tr>
<tr><td>bsd1PP</td><td>.005</td><td>60.88</td><td>54.97</td><td>17.66</td></tr>
<tr><td>bsd1PP</td><td>def</td><td>18.07</td><td>36.92</td><td>-</td></tr>
<tr><td>bsd1PP</td><td>def</td><td>55.81</td><td>46.77</td><td>12.50</td></tr>
</table>

We find from average WCSS that default sigma once again has a slight improvement over sigma = .005, so we continue to use the default value for the rest of our models on the different datasets.

Here, we compare between no pre-process and pre-process with “scale”. These are with the original dataset and the action is to omit any missing values. We find slightly better values with the default sigma, so we continue with that here.

Kernel KMeans with Basis For Purchase				
	sigma	C1	C2	C3
Bsd	def	12.35	39.90	-
bsd	def	42.35	14.42	24.86
bsdPP	def	39.18	12.49	-
bsdPP	def	19.90	17.97	48.60

Below we look at the dataframe where we set NAs=zero. We compare between no pre-process and pre-process with “scale” between 2 and 3 clusters.

Kernel KMeans With Basis For Purchase				
	sigma	C1	C2	C3
bsd1	def	51.22	13.80	-
Bsd1	def	53.61	54.44	12.70
bsd1PP	def	18.07	36.92	-
bsd1PP	def	55.81	46.77	12.50

This table is the one where we substitute the median value into the columns with blanks. Again we compare with and without pre-process.

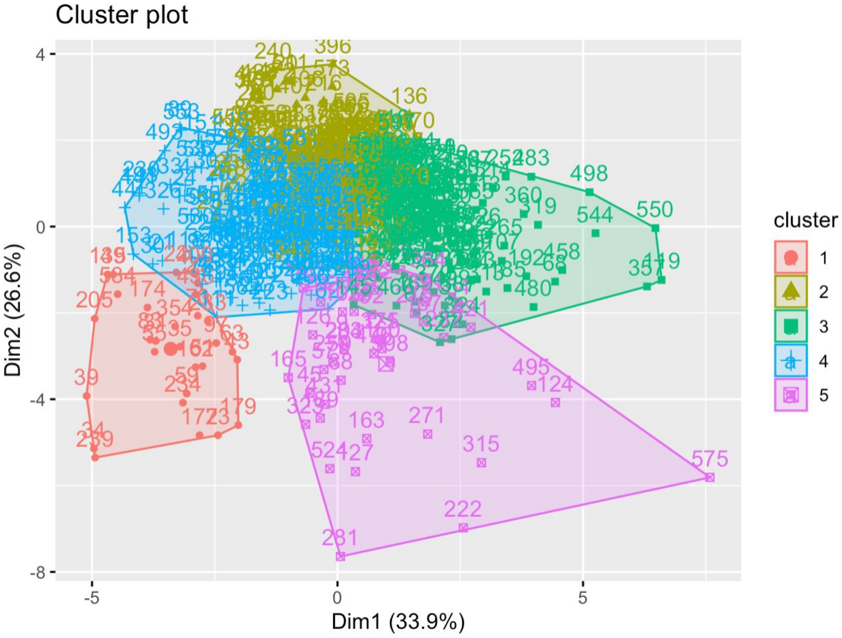
Kernel KMeans With Basis For Purchase				
	sigma	C1	C2	C3
df	def	13.70	52.67	-
df	def	39.86	15.89	26.37
dfPP	def	21.64	32.75	-
dfPP	def	56.95	54.44	13.28

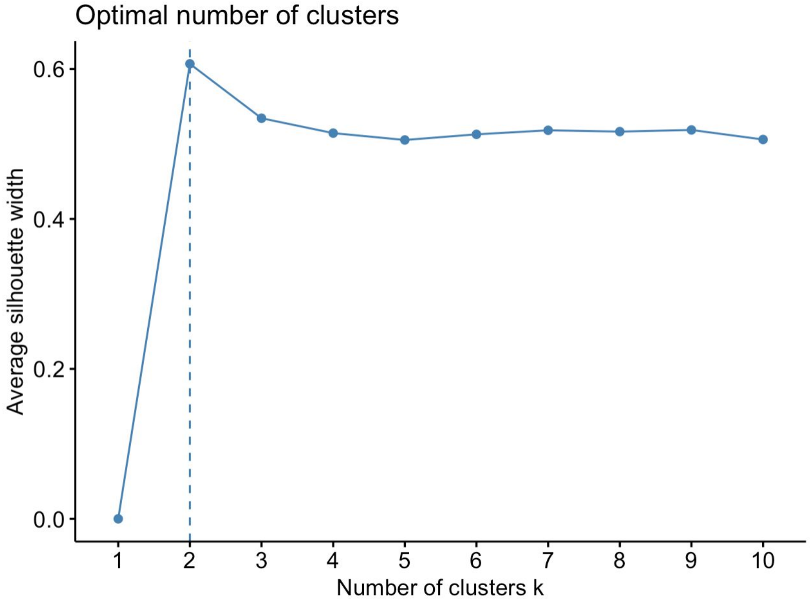
Finally, we compare the averages for all of the models illustrated above. In general, 2 clusters perform better than three, which agrees with our preliminary analysis. Across models that were pre-processed versus those that were not, the difference is not significant enough to draw conclusions from. What we see from this table compared to the purchase behavior, is that Basis For Purchase is tougher to make into accurate clusters. The values are more than double those of purchase behavior indicating that it may be easier to predict a households purchase behavior than their Basis For Purchase. This is relevant to advertisers, since they may want to focus ads that are for brand loyalty rather than for getting purchases from promotions or a certain category type.

Kernel K Means- Basis For Purchase			
Models	**Number of clusters**	**Sigma value**	**Average WCSS**
Bsd1	2	def	28.33
-	3	def	30.99
bsd1PP	2	def	28.81
-	3	def	31.20
-	2	.005	29.50
-	3	.005	31.64
df	2	def	28.45
-	3	def	30.59
dfPP	2	def	29.03
-	3	def	31.05
bsd	2	def	28.19
-	3	def	30.25
bsdPP	2	def	28.24
-	3	def	30.10

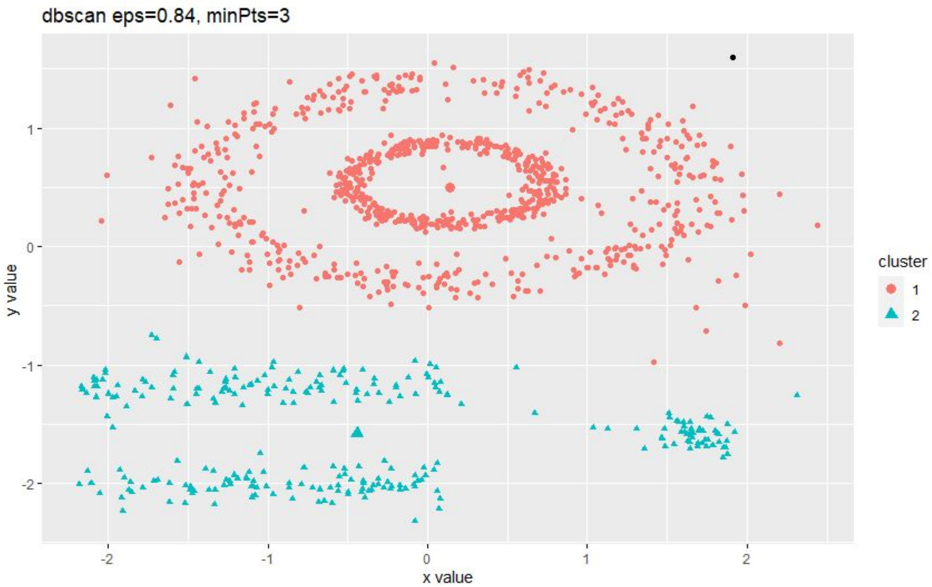
5. (a) Compare the clusters obtained in Q3 and Q4. Are the clusters obtained from the different procedures similar/different? Describe how they are similar/different – in terms of number and size of clusters, within cluster spread and separation between clusters; also, very importantly, interpretability.

DBScan models didn't give a good enough result compared to other clusters models we made. Kmean and Kernel K means models are very similar in the plotting of the model but gave different results in terms of the size and spread of the cluster. K means models we saw that in plotting clusters more than 5, resulted in difficult models to interpret as shown in the graph below. We have clusters that overlap each other. When plotting the model we also saw the most optimal was 2 clusters (using the silhouette method) but clustering for 3 and 4 were more better to use further break down variables we had into groups.





The DBScan model had a large disproportion of cluster sizes compared to other two models that we had. Kmean and Kernel K means models had relatively the same numbers for each model where they range in the hundreds. DBScan sizes went in the low tens to the hundreds and also had many outliers to it as shown below.



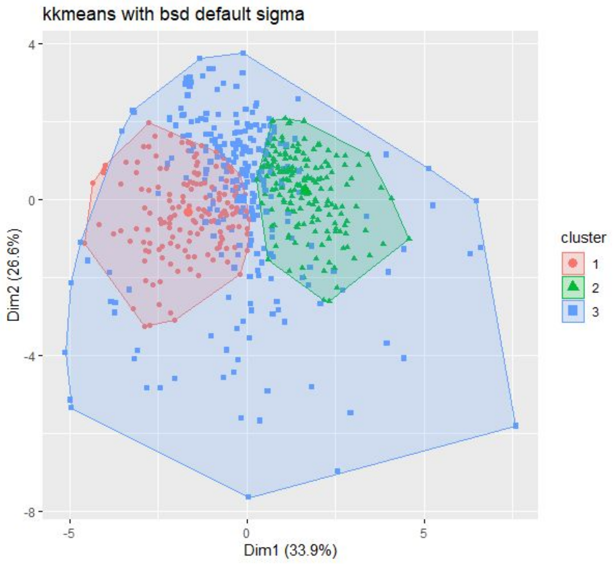
K means model overall gave the same performances within each dataset when doing behavior purchase and basis for purchase. Datasets didn't make any difference to the average within clusters sums of squares. Nor did they make any differences to the plotting of the models, all of them gave the same results. The only differences in the k means models were the number of clusters being made and then also looking at the variables tables afterwards which gave the distance measurement from the middle of the clusters. DBScan plotting of the models showed a lot of outliers.

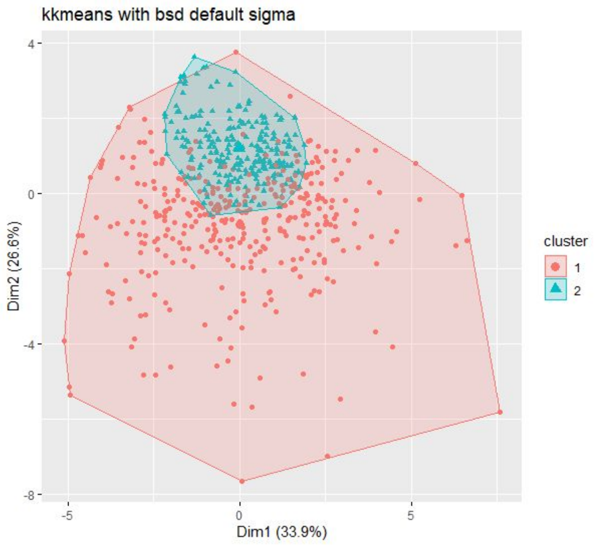
The kernel k means model gave better results for purchase behavior than basis for purchase. Our preliminary analysis indicated that 2 or 3 clusters would almost always be best for average WCSS, so that is what we use. From this, 2 clusters consistently performed better than 3. Between the dataframes, there was not a significant difference in average WCSS, which is what we used to evaluate the models. Pre-processing also had a minimal impact on the average WCSS, so we cannot conclude that that helps cluster formation.

In both models we saw the best WCSS values came from 2 clusters. Basis for purchase ended up having the best value for WCSS value for both models but when looking at the variable summarization of WCSS for both had revelantly the same numbers. In the end we decided that the best result came from 2 clustering based on both models interpretation as well using purchase behavior since it gave the lowest WCSS value for further analysis on demographic influences.

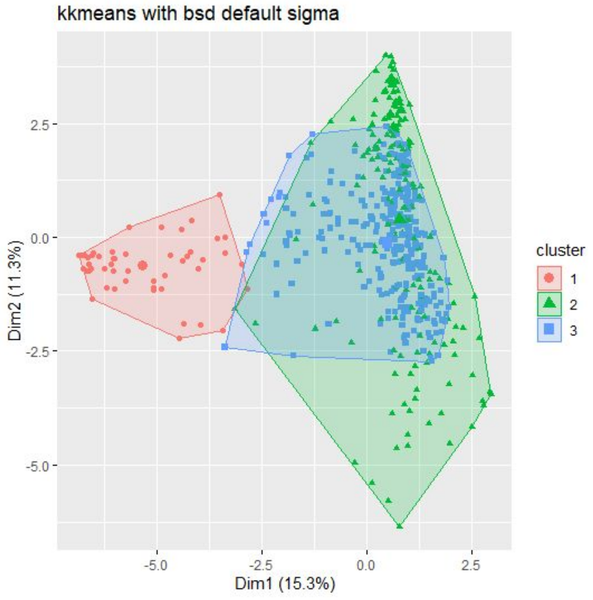
Kernel K-mean sample plots

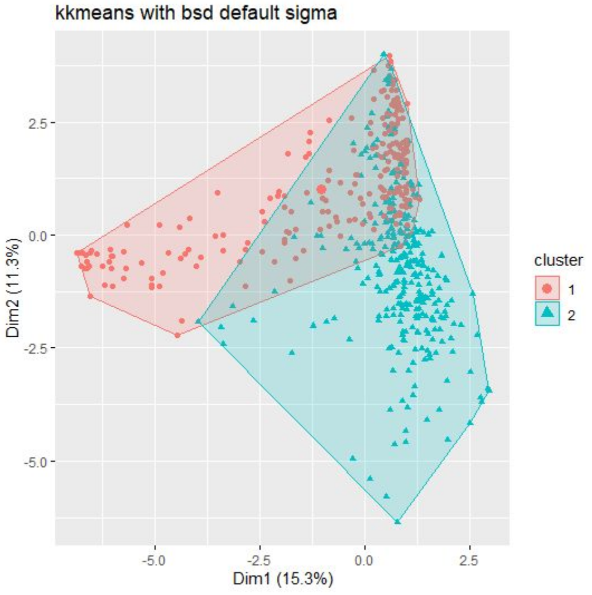
Below are two plots comparing 2 and 3 clusters with kernel k means for Purchase Behavior. See the appendix for plots comparing with sigma=.005.





Below are two plots comparing 2 and 3 clusters with kernel k means for Basis For Purchase





(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.

The best segmentation was determined using the average within clusters sums of squares and then also looking at the variables summarization table that k means model given. In the part above we concluded that the best clustering was from the 2 to 3 clusters. When looking at the kernel k model we saw that the preprocessing did very slightly better in WSCC value, but we are not sure if this difference is significant.

Below we compare the average WCSS for the original dataset with 2 and 3 clusters between kernel k-means and k-means. We look at both purchase behavior and basis for purchase since the models each found a different one to be easier to predict than the other. Overall, we go with purchase behavior to determine the best segmentation, using kernel k-means.

Comparing Kernel K means to K means for Purchase Behavior		
Model	**Cluster Number**	**Average WCSS**
K means	2	20.0
K means	3	33.7
Kernel K means	2	11.72
Kernel K means	3	11.94

Comparing Kernel K means to K means for Basis For Purchase		
Model	**Cluster Number**	**Average WCSS**
K means	2	12.8
K means	3	19.8
Kernel K means	2	28.19
Kernel K means	3	30.25

Based on that, we did variables summarization with clustering of 2 and 3 and also did a comparison between preprocess data and no preprocess values to see if there's a difference between them and there wasn't. What we do see is that for demographic variables we get lower values WSCC with clustering 2 and 3. This indicates that those are the variables that make up the inner clustering which helps interpret the behavior purchase. Similar results were given on the basis for purchase which shown in the appendix pages. With basis for purchase, there are a lot of variables involving promotion sales so we can use results variables tables to determine which specific demographic are in the cluster.

Purchase Behavior (Preprocessed Data)

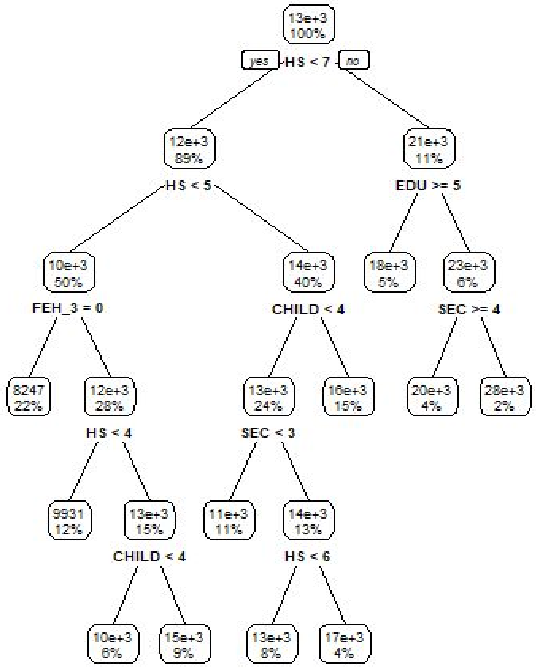
ClusKM	SEC	HS	SEX	EDU	Affluence Index	Age
1	2.339768	3.474903	1.644788	4.03861	16.06618	3.131274
2	2.822857	4.382857	1.702857	3.52000	13.86286	3.205714
3	2.409639	5.108434	1.921687	4.60241	20.99398	3.349398

CHILD	Maxbr	No of Brands	No of Trans	Brand runs	Total Volume	Trans Brands Runs
3.312741	0.5322054	3.200772	0.4155637	0.2156513	0.5323647	0.2473029
3.371429	1.2404272	2.857143	0.4114518	0.7236194	0.7236194	0.6076481
2.963855	0.4773106	5.138554	1.0821389	1.0993197	0.6359753	0.2547407

(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). Does the decision tree provide a similar interpretation to that you find from the description of cluster centers; does it provide alternate or additional information which will be useful in understanding the clusters.
(Note - you may develop decision trees for alternate clustering, and use these to help choose the ‘best’ clustering).

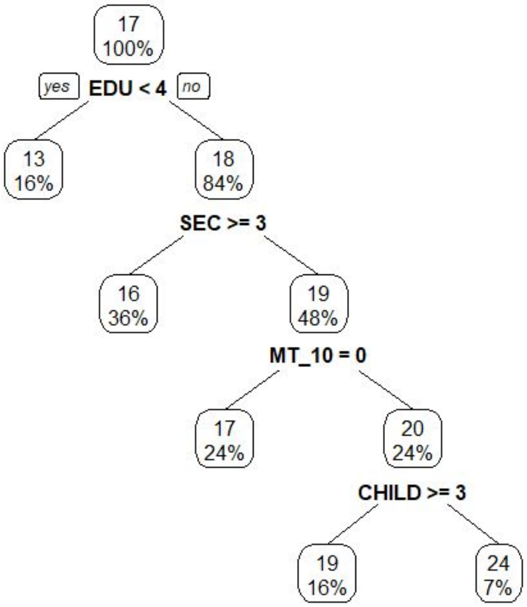
Total Volume

We created a simple decision tree model to predict “total volume” purchased by a household with the idea that some of the variables that predicted brand loyalty earlier would also be significant here. Household size, child presence, socioeconomic status, and food eating habits are the most important predictors. Surprisingly, when removing blank values, the affluence index is far less significant. The model looks the same with or without affluence index. By viewing this over-simplified model we can confirm that the variables found to be significant earlier make sense as contributors to brand loyalty. In the graph below we see household and education, which were found in our clustering models to have the lowest mean WSCC value so it's not surprising to find that they have strong influence here. From a company’s perspective, they may want to target “volume-based” ads to households that have fewer than 5 members, based on what we see from the decision tree below.

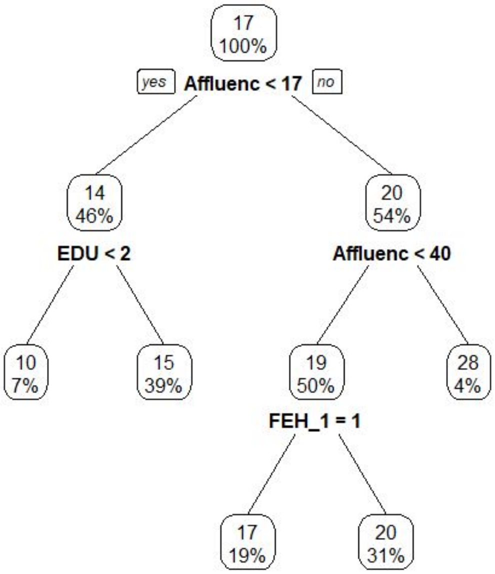


Brand Runs

We take a look at a decision tree predicting brand runs- which is an indicator of brand loyalty. When we remove the affluence index, we get this tree below. The significant splits in order of importance are: education of head of household, child presence, socioeconomic status, and language not being Marathi. From this decision tree, advertisers may target families with these criteria. We still see these trends that we predict and found in the clusters; where education, socio economic status, mother language, and the number of children made an impact on how often a family ended up buying products more frequently compared to others. A company may want to make advertisements targeted to households with medium educational levels and a lower-medium socioeconomic status if they are trying to improve their brand loyalty.



If we include the affluence index it overwhelms the other categories, and we do not learn as much about specific demographic information. But if companies wanted to consider a family's affluence, they would want to target families with a "medium level of "affluence" somewhere between 17 and 40 on the scale they use for the data, since those appear to be more loyal to a certain brand.



Based on the decision trees above our predictions were correct on the demographic variables that had an influence in deciding purchasing and promotion sales. Based on the results here, if we do other things involving sales and promotions, we will have the same results where demographics are going to have an impact on it. Higher income households have more purchasing power than lower/middle income households.