IDS 572 - Group Assignment 3

CJ All, Eric Reitz, Vanessa Rodriguez

Due Date: 11/4/2020

Questions:

Before the Data:

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

   The goal is to identify advertising markets based on the purchasing process and brand loyalty for their advertising clients. Clearer market segmentation would allow their clients to create and implement more cost-efficient promotions targeted at the ideal customer base. An effect of this will allow them to use the savings to market at even more segments, giving them the best value for their marketing budget.

   From here, the company could start segmenting the market based on the purchase process and brand loyalty to save costs in promotions and target the right segments to increase revenue. By analyzing patterns in consumer behavior they will be able to predict the consumer journey and increase their sales by targeting the right audience. In the context of customer segmentation clustering will help us identify people in groups and put them together based on purchasing patterns to target the right audience and create customized advertising depending on what consumers want and like.

   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?

   We will use this data in a number of ways to assist the company with achieving its goal:

   1. Explore & mine the data to get a better understanding of what we're looking at
   2. Apply different machine learning & clustering techniques to our data to create the segments and purchasing groups we need to analyze
      a. These can include KNN, weighted KNN, etc.
   3. Draw conclusions and produces our solutions based on our models that we have drawn out

   The variables describing purchase behavior and basis-in-purchase are listed below:

1. Purchase behavior: volume of products, number of runs, number of transactions, Number of brands purchased average price, number of transactions, share to other brands, and brand loyalty

2. Basis of purchase: Percent of volume purchased not on promotion, Percent of volume purchased not on promotion code 6, Percent of volume purchased not on promotion on promo code other than 6, Price categories 1-4, and Selling Propositions 5-15

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.

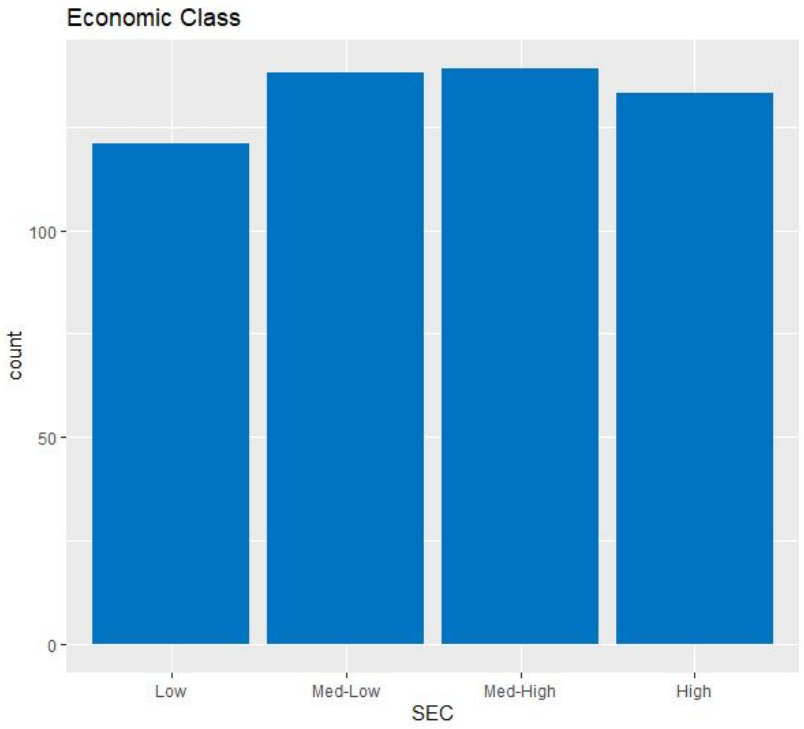
First we will cluster data based on the purchase behavior variables. After that we will cluster the data based on our basis of purchase variables, and then finally we will cluster data considering both sets of variables. For the first set the number of clusters will match the number variables, and after doing some initial clustering will tweak based on what we are seeing. If some clusters are very small or too closely joined with one another we will adjust the amounts from this. We'll apply this same approach in the other sets as well.

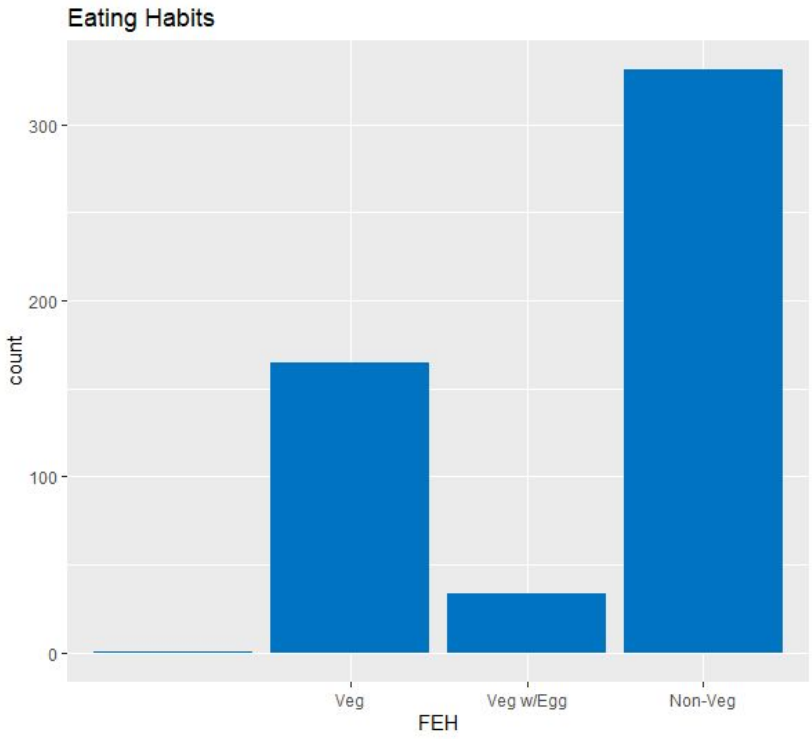
<u>Data Exploration:</u>

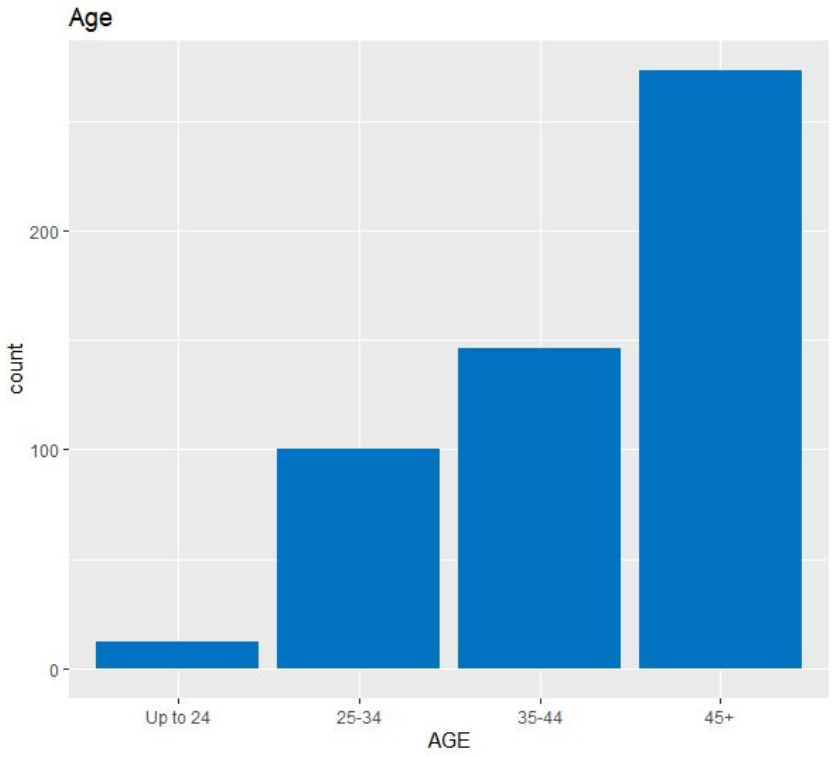
2. **Are there any missing values – how do you handle these?**

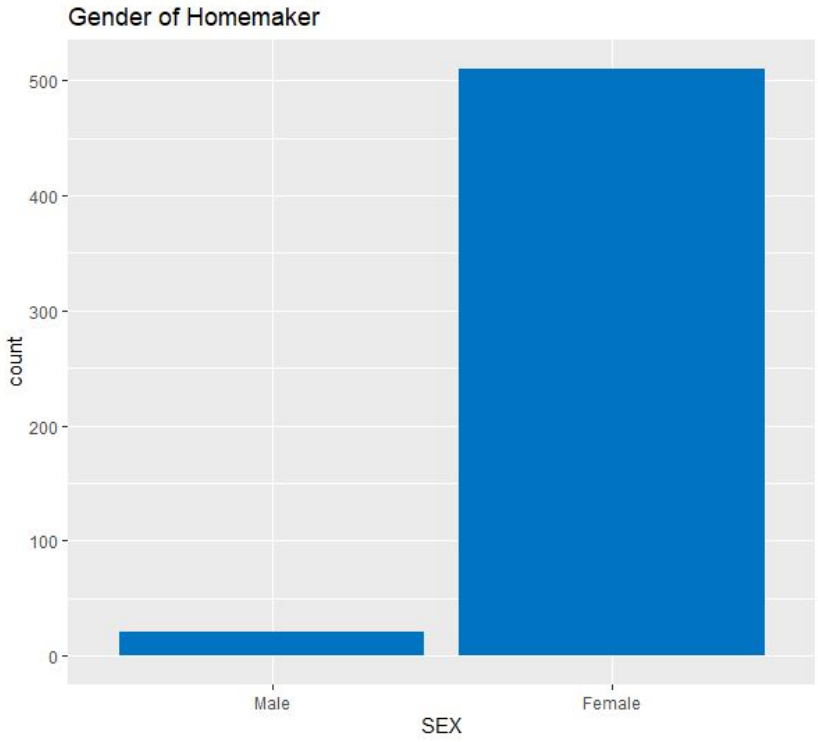
   Yes there are. Demographics have plenty of empty fields, we replaced these with Not Specified or Unavailable where applicable. Also we removed households that have a zero on the Affluence Index. If they are categorized as with zero they would not be ideal to base marketing on.

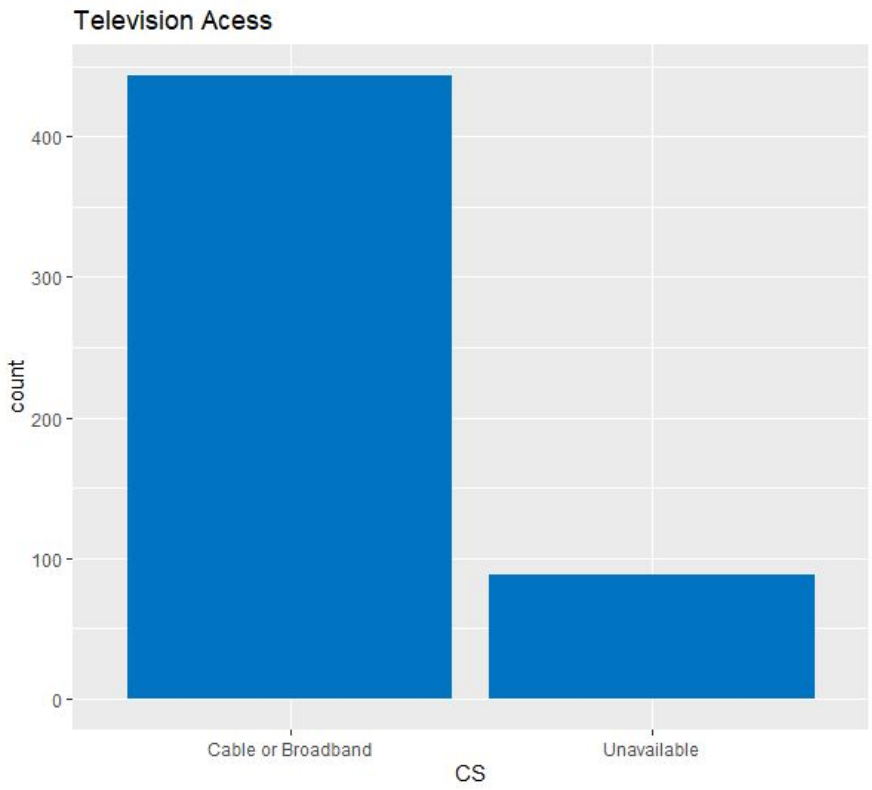
   Summarize the households in the data based on demographic variables – use plots, tables to help your description. (Bar graphs, circle graphs, line graphs, etc. could help for all of these variables)

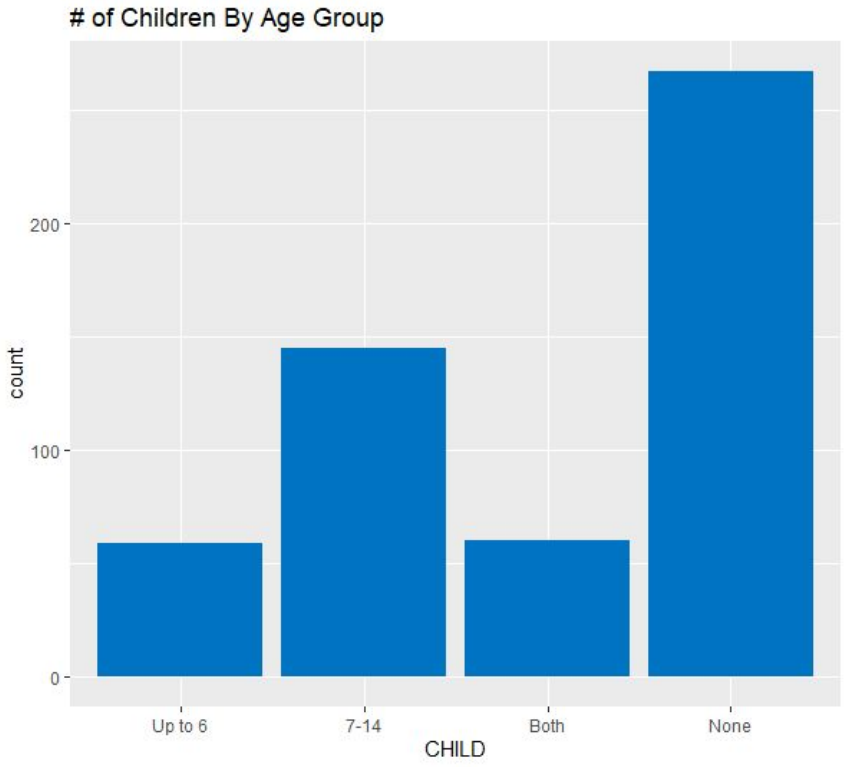

Economic Class
100
50
0
count
Low
Med-Low
Med-High
High
SEC

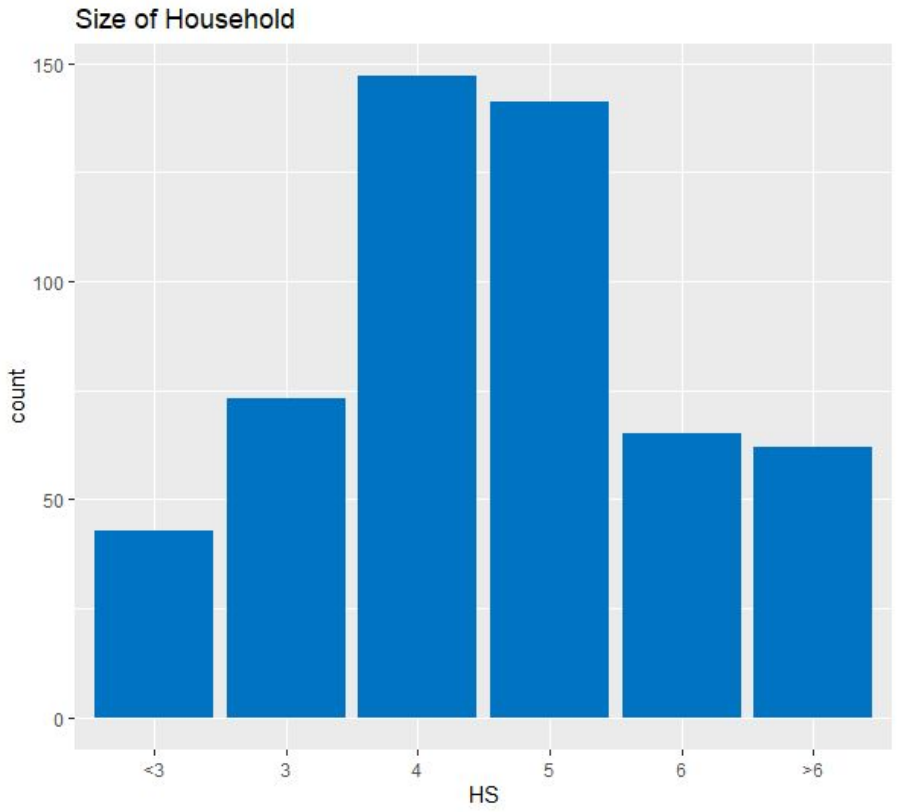

Eating Habits
300
200
100
0
count
Veg
Veg w/Egg
Non-Veg
FEH

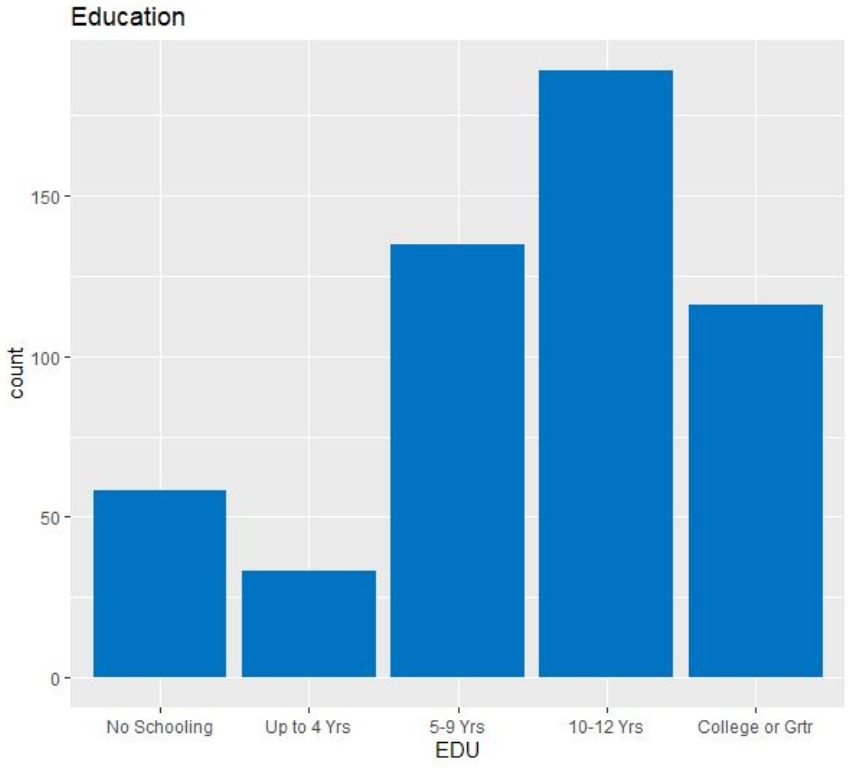

Age
count
200
100
0
Up to 24
25-34
35-44
45+
AGE

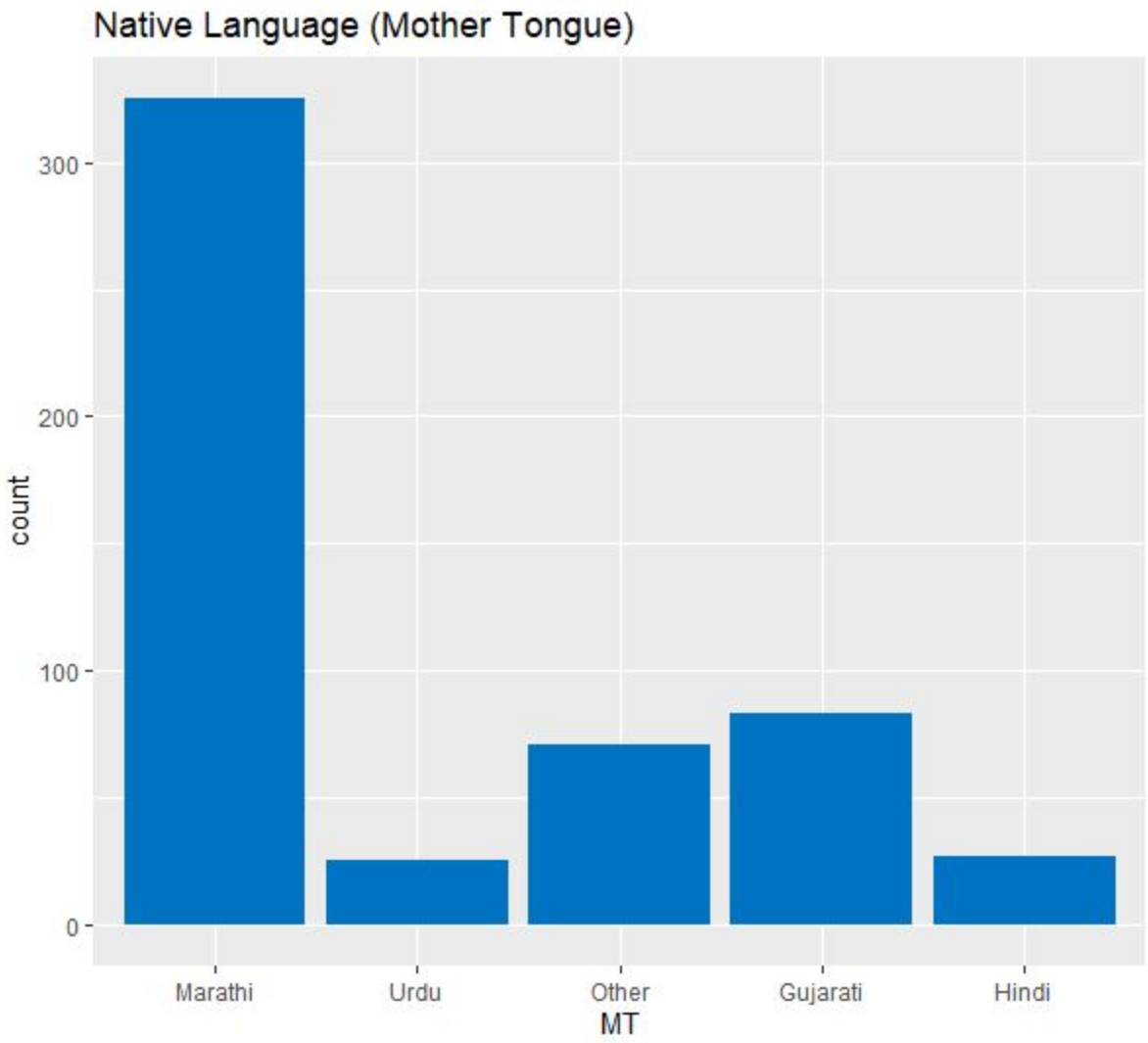

Gender of Homemaker
count
500
400
300
200
100
0
Male
Female
SEX


Television Acess
400
300
200
100
0
count
Cable or Broadband
Unavailable
CS


# of Children By Age Group
200
100
0
count
Up to 6
7-14
Both
None
CHILD


Size of Household
150
100
50
0
count
<3
3
4
5
6
>6
HS


Education
150
100
50
0
count
No Schooling
Up to 4 Yrs
5-9 Yrs
10-12 Yrs
College or Grtr
EDU



Demographically, there are several variables worth highlighting. The homemaker gender is almost all female making up over 96% of the total. At 83% households responded that they have Cable or Broadcast TV. While the native language of Marathi and the non-vegetarian food eating habits both make up more than 60% of households. About half of homemakers are aged 45 or over, with the other age categories making up the rest of it in a steady decline the younger the age range lowers. Households with children versus no children is a nearly even 50-50 split. Finally, The 4 to 5 person household make up about 54% of the households with a normal distribution on either higher or lower sized households.

Explore the purchase behavior variables, and those which describe basis-for-purchase. Will you use all these variables directly, or a subset of these, and/or use any data transformations?

Purchase behavior: *Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)].*

Basis of purchase: Percent of volume purchased not on promotion, Percent of volume purchased not on promotion code 6, Percent of volume purchased not on promotion on promo code other than 6.

How will you evaluate brand loyalty? Describe the variables you create and use to capture different aspects of brand loyalty.

We'll evaluate brand loyalty in three different ways, through the number of brands they purchase (the smaller amount the more loyal), the percentage of brand jumping (how often proportionally they switch brands), and the proportion of purchases of different brands (transactions per brand). The Number of Brands purchased from already exists, for Brand Jumping we will divide the number of brands by the number of brand runs, and for Proportion of Purchases Between Different Brands we will divide the number of brands by the number of overall transactions.

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

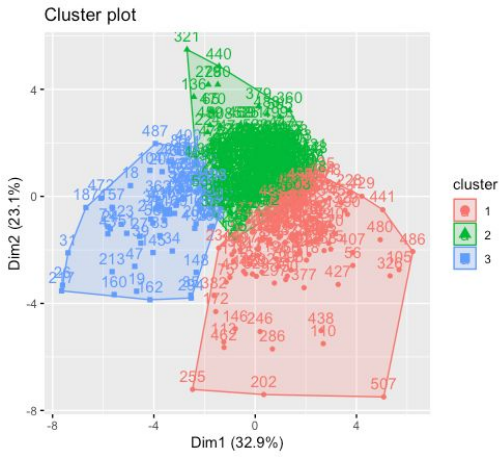
A.The variables that describe purchase behavior (including brand loyalty). Variables used: 'No__of_Brands', 'Brand_Runs', 'Total_Volume', 'No__of__Trans', 'Value', 'Trans___Brand_Runs', 'Vol_Tran', 'Avg__Price', 'maxBr', 'Others_999', 'Brand_Jumping', 'ProDifBrands')

We used k means with values of 3, 2 and 4. We determined that k=3 was our best choice; we based our decision on the distribution between the clusters groups to find which ones were close or similar to each other and we compared it to the other k means and k=3 had the largest distance between each cluster which helps us determine how to segment our market to determine their purchase behavior based on their preferences. Below are the details of the different k-means we ran.

K-means clustering with 3 clusters of sizes 150, 305, 76

```
Cluster means:
  No__of_Brands Brand_Runs Total_Volume No__of__Trans      Value Trans___Brand_Runs   Vol_Tran  Avg__Price      maxBr Others_999 Brand_Jumping
1     0.6627338  0.9096497    0.9091437     1.0131340  0.9928132         -0.2276797  0.1859528  0.07276688 -0.3552738  0.2314148    -0.4886633
2    -0.1398772 -0.1574144   -0.4790612    -0.3511893 -0.4148102         -0.2613891 -0.2541768  0.13823753 -0.2373391  0.2165347    -0.1981545
3    -0.7466781 -1.1636323    0.1281857    -0.5902285 -0.2948009          1.4983636  0.6530395 -0.69838790  1.6536777 -1.3257277     1.7596922
  ProDifBrands
1   -0.3647047
2    0.2227405
3   -0.1740810
```

Cluster 1: Seems to have the maximum number of soap brands and it also has the highest amount of repeat purchases of the same brand hence it also has the highest number of transactions.
Cluster 2: The average price on this cluster seems like it has the highest value meaning that the products on this cluster are the most expensive ones therefore it seems like the brand runs are low as well as the number of transactions.
Cluster 3: Purchases have high volume and as well as brand jumping that shows how often consumers switch brands. It also shows the lowest values on consumers considering other brands.

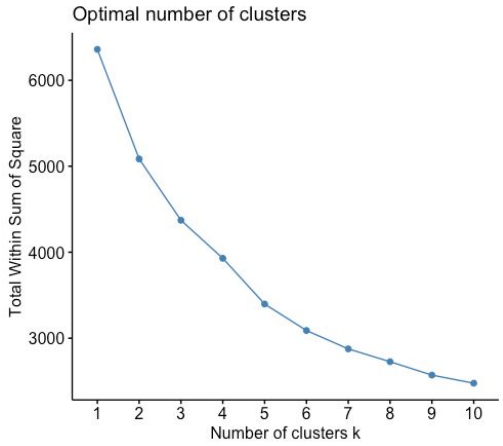


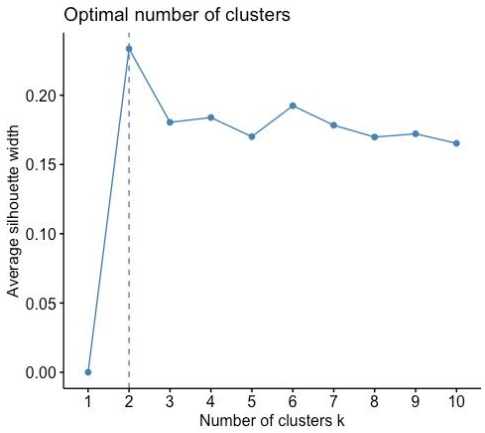
Clusters 1 and 3 are separated from each other but there are a few similarities between cluster 2,1 and 3 and 2.

[Q – how do you measure brand loyalty?]

From the graph we can see that the cluster with the highest major brand loyalty is cluster 3 with 81.84% maxbr and it also has the highest number of transactions with 6.76%. Cluster 3 also has the lowest number of different brand purchases further suggesting brand loyalty. Although when you consider brand jumping (the measure of how often they switch brands) it appears that over 60% of the time cluster 3 households are swapping between their few brands they support.

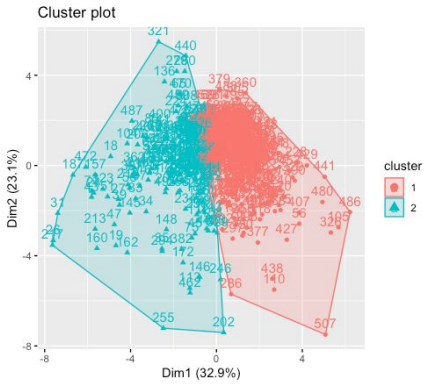
	cluster	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs	Brand_Jumping	ProDifBrands
1	1	2.560000	5.620000	1.973333	4.553333	21.30667	3.353333	0.2631133	4.820000	50.38000	26.26000	19797.433	2301.491	2.120133	0.2009333	0.1021333
2	2	2.432787	4.229508	1.950820	4.777049	19.50164	3.206557	0.2957148	3.573770	28.13443	15.31803	9373.761	1099.621	2.029443	0.2560328	0.1446230
3	3	2.894737	4.960526	1.973684	3.710526	14.05263	3.434211	0.8184605	2.631579	24.23684	5.00000	13933.421	1202.088	6.763816	0.6273684	0.1159211





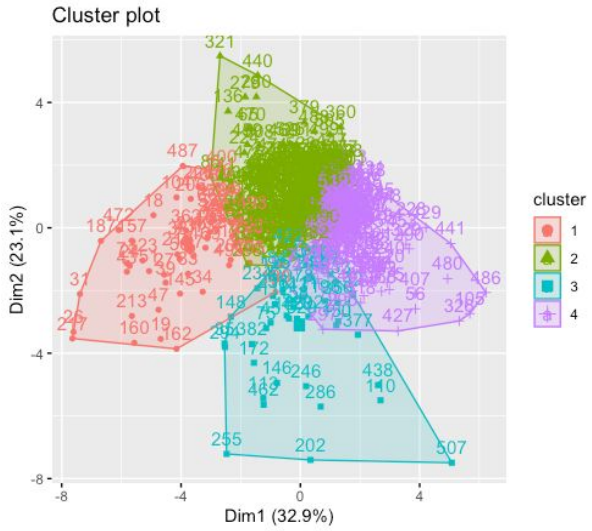
Silhouette plot measures the quality of a clustering and shows that the optimal k value for this data is k=2. The elbow method shows that the appropriate number of clusters is between 3 and 4.(bends)

K-means clustering with 2 clusters of sizes 347, 184



	cluster	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs	Brand_Jumping	ProDifBrands
1	1	2.423631	4.573487	1.962536	4.809798	20.97983	3.291066	0.2195159	4.170029	38.86744	21.291066	12312.76	1487.649	1.982104	0.2110086	0.1183285
2	2	2.744565	5.016304	1.956522	4.092391	15.93478	3.260870	0.6287554	3.076087	24.41848	8.711957	14212.09	1389.958	4.148152	0.4494022	0.1477174

K-means clustering with 4 clusters of sizes 63, 256, 47, 165



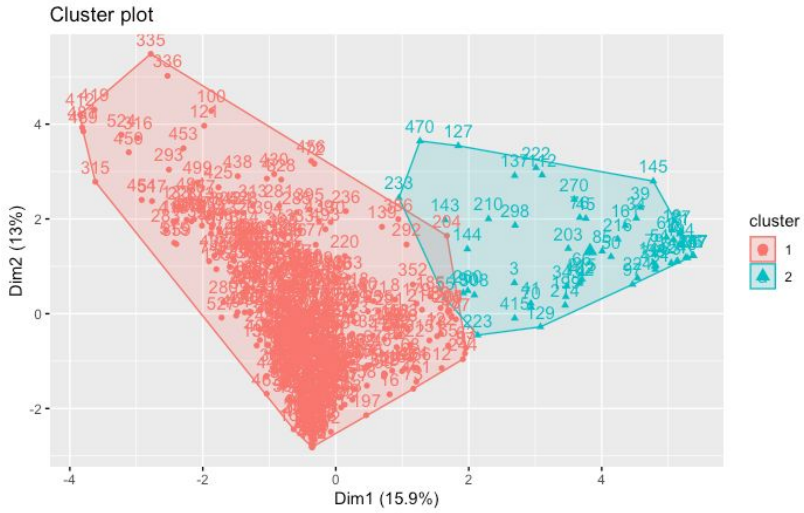
	cluster	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs	Brand_Jumping	ProDifBrands
1	1	2.793651	4.857143	1.968254	3.809524	14.30159	3.428571	0.8361270	2.571429	25.19048	4.380952	13795.238	1204.675	7.611587	0.6782540	0.1101587
2	2	2.546875	4.265625	1.949219	4.574219	18.44922	3.187500	0.3386367	3.382812	25.29297	13.363281	9245.516	1045.856	2.089258	0.2742188	0.1517969
3	3	2.851064	6.680851	2.000000	4.191489	18.74468	3.489362	0.3531702	3.531915	37.08511	16.425532	29452.234	3133.667	2.541064	0.2389362	0.1031915
4	4	2.327273	4.836364	1.963636	4.933333	22.46667	3.309091	0.2175576	4.963636	49.54545	27.406061	13741.485	1703.338	1.922667	0.1924242	0.1066061

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.]

'PropCat_5', 'PropCat_6', 'PropCat_7', 'PropCat_8','PropCat_9','PropCat_10','PropCat_11', 'PropCat_12', 'PropCat_13', 'PropCat_14', 'PropCat_15', 'Pur_Vol_No_Promo____', 'Pur_Vol_Promo_6__', 'Pur_Vol_Other_Promo__', 'Pr_Cat_1', 'Pr_Cat_2', 'Pr_Cat_3', 'Pr_Cat_4')

We decided to use all selling propositions to figure out what promotion is the most commonly used. We used k means with values of 2,3,4,5, and 9. We determined that k=2 was our best choice because based on the cluster distance and separation from each other as well as the points within the cluster it seems like they have enough space to divide them into different segments and target specific consumers. We noticed that clusters 5 and 9 were on top of each other and did not give us a good reason to believe it was the correct segmentation because they were all overlapping and when doing clustering we need to have separation within the clusters to determine how we want to analyze data. Below are the details of the different k-means we ran.

K-means clustering with 2 clusters of sizes 62, 469

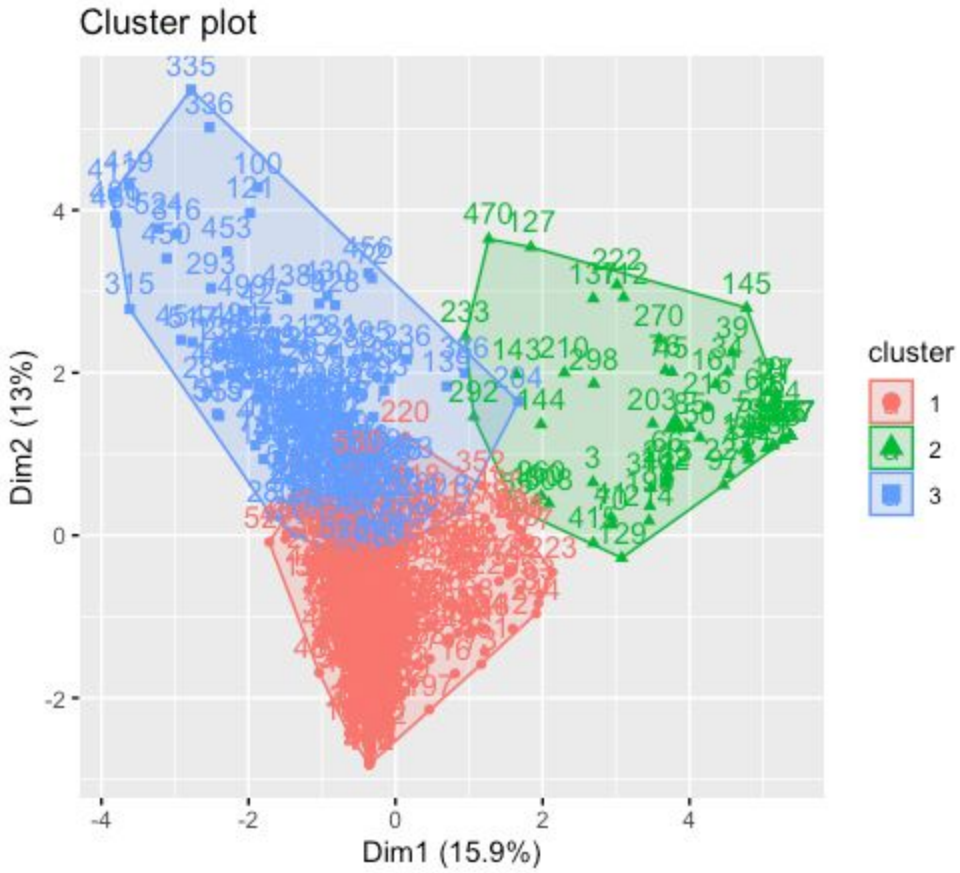



```
Cluster means:
  PropCat_5   PropCat_6   PropCat_7   PropCat_8   PropCat_9  PropCat_10  PropCat_11  PropCat_12  PropCat_13 PropCat_14  PropCat_15
1 -1.160079 -0.10836668 -0.45826040 -0.48868163 -0.17875851 -0.27054506 -0.22632924 -0.15739322 -0.24105792   2.502932 -0.22005762
2  0.153358  0.01432566  0.06058027  0.06460184  0.02363119  0.03576502  0.02991986  0.02080678  0.03186693  -0.330878  0.02909077
  Pur_Vol_No_Promo____ Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__   Pr_Cat_1   Pr_Cat_2   Pr_Cat_3    Pr_Cat_4
1            0.2218287        -0.4006236            0.22995012 -0.7731496 -1.2113391  2.4974357 -0.32927090
2           -0.0293249         0.0529609           -0.03039852  0.1022074  0.1601344 -0.3301514  0.04352835
```

	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	3.258065	5.258065	1.951613	3.080645	11.51613	3.145161	3.080645	0.7332581	3.209677	30.06452	10.01613	15480.08	1084.922	5.488065
2	2	2.439232	4.656716	1.961620	4.756930	20.25160	3.298507	2.997868	0.3121557	3.867804	34.36247	17.84648	12639.20	1502.562	2.368422

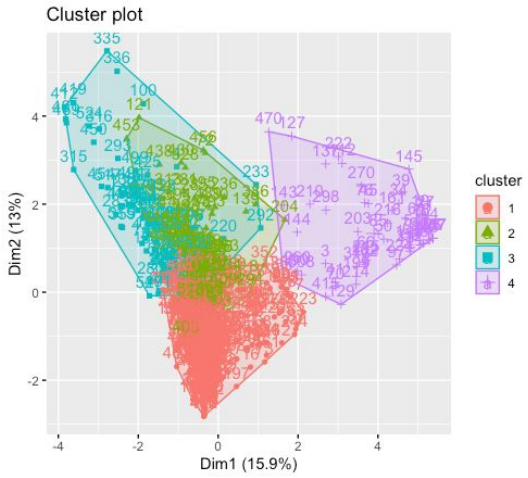
K-means clustering with 3 clusters of sizes 172, 62, 297

```
Cluster means:
   PropCat_5   PropCat_6   PropCat_7  PropCat_8   PropCat_9 PropCat_10  PropCat_11  PropCat_12 PropCat_13 PropCat_14  PropCat_15 Pur_Vol_No_Promo____
1 -0.3264437  0.10137844  0.21173737  0.5382685  0.12879035  0.3885625 -0.01373486  0.21173796  0.4453501 -0.4314532  0.06324250          -0.7022591
2 -1.1509785 -0.10817284 -0.46572279 -0.4850523 -0.17875851 -0.2705451 -0.22632924 -0.15739322 -0.2410579  2.5016593 -0.23741514           0.1933627
3  0.4293232 -0.03612921 -0.02540072 -0.2104678 -0.03726907 -0.1685487  0.05520137 -0.08976616 -0.2075913 -0.2723668  0.01293612           0.3663302
  Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__   Pr_Cat_1   Pr_Cat_2   Pr_Cat_3   Pr_Cat_4
1         0.6229257            0.3876986  1.0486291 -0.4756484 -0.4351306 -0.1023039
2        -0.3733065            0.2423929 -0.7728443 -1.2250359  2.4986332 -0.3098288
3        -0.2828223           -0.2751263 -0.4459524  0.5311911 -0.2696053  0.1239247
```



	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.696970	4.885522	1.959596	4.501684	18.16162	3.265993	2.909091	0.3630539	3.868687	33.10101	16.39731	13896.19	1492.002	2.491448
2	2	3.258065	5.274194	1.951613	3.048387	11.50000	3.145161	3.080645	0.7329355	3.161290	30.38710	10.16129	15871.61	1114.341	5.491290
3	3	1.994186	4.255814	1.965116	5.209302	23.86628	3.354651	3.151163	0.2243837	3.883721	36.42442	20.29651	10327.59	1510.191	2.154826

K-means clustering with 4 clusters of sizes 280, 101, 90, 60



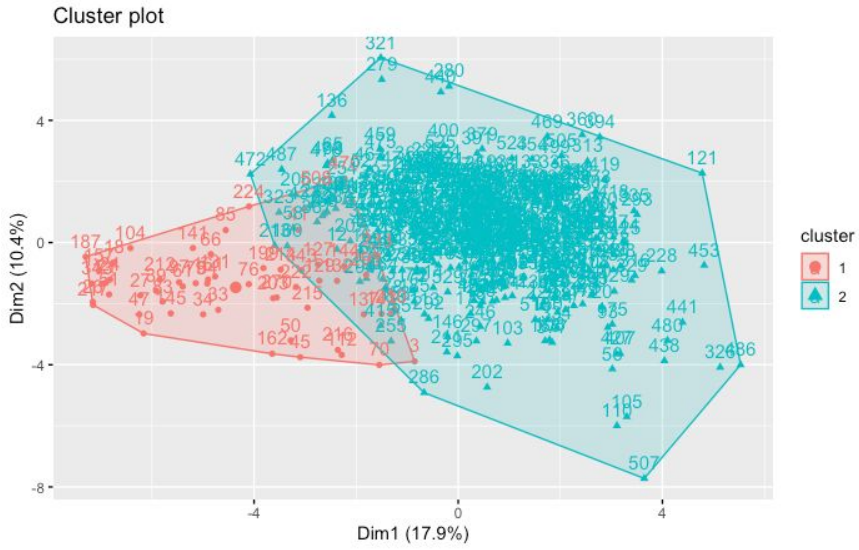
	clusKM	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.657143	4.853571	1.957143	4.514286	18.49286	3.267857	2.910714	0.3764536	3.867857	33.33214	16.45357	14045.01	1517.850	2.504714
2	2	1.910891	4.336634	1.960396	5.544554	24.67327	3.237624	2.960396	0.2197228	3.712871	36.70297	18.83168	10267.18	1663.171	2.447525
3	3	2.377778	4.388889	1.977778	4.644444	20.57778	3.455556	3.311111	0.2186333	4.055556	34.72222	20.94444	10904.72	1262.523	1.848333
4	4	3.250000	5.300000	1.950000	3.000000	11.50000	3.150000	3.083333	0.7431167	3.166667	30.25000	9.95000	15609.00	1089.352	5.603000

```
Cluster means:
    PropCat_5   PropCat_6    PropCat_7  PropCat_8   PropCat_9  PropCat_10  PropCat_11  PropCat_12  PropCat_13 PropCat_14  PropCat_15
1  0.41028250 -0.05774576  0.005538764 -0.1904269  0.01702277 -0.17717053  0.02291121 -0.12790254 -0.22068213 -0.2711653  0.03841402
2 -0.48484393  0.46939313  0.392766108  0.2419025 -0.18348023  0.58402560 -0.17605392  0.50519951  0.83649985 -0.4311629 -0.08666048
3  0.05459876 -0.28104252 -0.148056469  0.6504638  0.26493952  0.07591223  0.29283726 -0.06668124 -0.09257517 -0.3670315  0.13453044
4 -1.18039586 -0.09910108 -0.464919143 -0.4942392 -0.16799049 -0.27018232 -0.24981741 -0.15351881 -0.23939540  2.5417763 -0.23518262
  Pur_Vol_No_Promo____ Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__   Pr_Cat_1   Pr_Cat_2   Pr_Cat_3    Pr_Cat_4
1            0.4071692        -0.3420625           -0.25784028 -0.4568693  0.6022647 -0.2717095  0.02988809
2            0.2059379        -0.2693244            0.03571336  1.6350605 -0.7845524 -0.4385430 -0.40706978
3           -1.6640740         1.6618900            0.59795924  0.1004468 -0.1719472 -0.3538226  0.59827829
4            0.2493259        -0.4431809            0.24619832 -0.7709653 -1.2319845  2.5369258 -0.35166108
```

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

PURCHASE_BEHAVIOR_BASIS_FOR_PURCHASE<- c('No__of_Brands', 'Brand_Runs', 'Total_Volume', 'No__of__Trans', 'Value','Avg__Price', 'maxBr', 'Others_999', 'Brand_Jumping', 'ProDifBrands','PropCat_5', 'PropCat_6', 'PropCat_7','PropCat_8','PropCat_9','PropCat_10','PropCat_11', 'PropCat_12', 'PropCat_13', 'PropCat_14', 'PropCat_15', 'Pur_Vol_No_Promo____', 'Pur_Vol_Promo_6__', 'Pur_Vol_Other_Promo__', 'Pr_Cat_1', 'Pr_Cat_2', 'Pr_Cat_3', 'Pr_Cat_4')

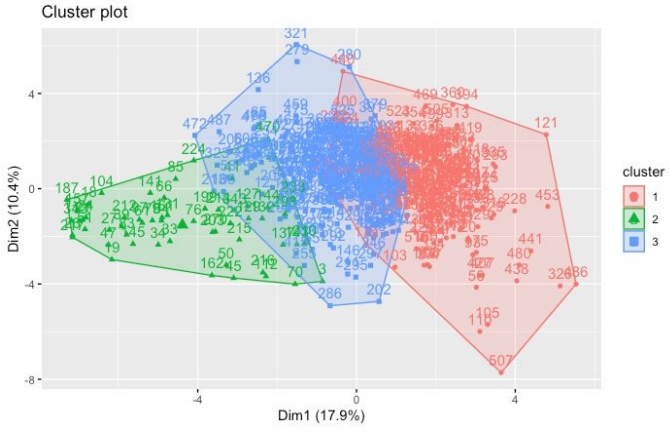
K-means clustering with 2 clusters of sizes 57, 474




```
Cluster means:
  No__of_Brands  Brand_Runs Total_Volume No__of__Trans       Value Avg__Price      maxBr Others_999 Brand_Jumping
1   -0.49810460 -0.73762250   0.34199494   -0.28088809 -0.45390096 -1.3444415  1.4474661 -1.2099313     1.0770493
2    0.05989865  0.08870144  -0.04112597    0.03377768  0.05458303  0.1616733 -0.1740624  0.1454981    -0.1295186
  ProDifBrands  PropCat_5  PropCat_6   PropCat_7  PropCat_8   PropCat_9  PropCat_10  PropCat_11  PropCat_12 PropCat_13
1  -0.14921911 -1.1909896 -0.1971017 -0.47092214 -0.5028283 -0.15486943 -0.26959048 -0.21902092 -0.16589450 -0.2399480
2   0.01794407  0.1432203  0.0237021  0.05662988  0.0604667  0.01862354  0.03241911  0.02633796  0.01994934  0.0288545
  PropCat_14  PropCat_15 Pur_Vol_No_Promo____ Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__    Pr_Cat_1   Pr_Cat_2   Pr_Cat_3
1  2.6197238 -0.25118303          0.20940637       -0.40715338           0.26746402 -0.78510473 -1.3066395  2.6157870
2 -0.3150301  0.03020555         -0.02518178        0.04896148          -0.03216339  0.09441133  0.1571275 -0.3145567
     Pr_Cat_4
1 -0.31872033
2  0.03832713
```

	kmClus_pbbp	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.438819	4.666667	1.959916	4.751055	20.17511	3.299578	2.995781	0.3132068	3.883966	34.41139	17.841772	12662.10	1500.402	2.369008
2	2	3.333333	5.228070	1.964912	2.982456	11.38596	3.122807	3.105263	0.7614561	3.017544	29.28070	9.368421	15538.86	1066.244	5.756842

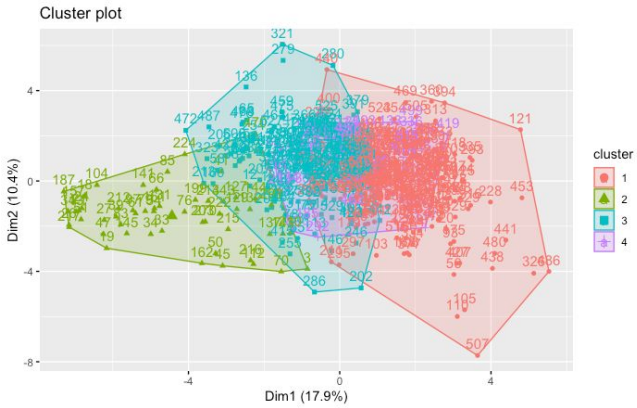
K-means clustering with 3 clusters of sizes 199, 56, 276



	kmClus_pbbp	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.075377	4.517588	1.974874	5.100503	24.12563	3.351759	3.00000	0.1871508	4.185930	41.16583	23.221106	11621.66	1603.086	2.080151
2	2	3.375000	5.250000	1.964286	2.910714	10.71429	3.142857	3.12500	0.7665179	3.000000	29.55357	9.339286	15754.73	1079.373	5.836964
3	3	2.695652	4.771739	1.949275	4.507246	17.43116	3.257246	2.98913	0.4046920	3.666667	29.46739	13.938406	13378.90	1422.129	2.573297

```
Cluster means:
  No__of_Brands Brand_Runs Total_Volume No__of__Trans       Value Avg__Price      maxBr Others_999 Brand_Jumping
1    0.25437258  0.6132958  -0.17969047     0.4480280  0.17484576  0.7236661 -0.6300661  0.5975695   -0.43132170
2   -0.50940340 -0.7404638   0.37074443    -0.2641530 -0.43852393 -1.3575967  1.4657767 -1.2122134    1.09003221
3   -0.08004911 -0.2919561   0.05433592    -0.2694384 -0.03709046 -0.2463194  0.1568828 -0.1848999    0.08982324
  ProDifBrands  PropCat_5   PropCat_6   PropCat_7  PropCat_8  PropCat_9 PropCat_10  PropCat_11 PropCat_12 PropCat_13
1   -0.1822992 -0.3042056  0.16082313  0.21366157  0.4054162  0.2497002  0.2882415 -0.03530744  0.2078548  0.3922856
2   -0.1917528 -1.1919488 -0.20434574 -0.47031210 -0.5075194 -0.1486743 -0.2693791 -0.25891994 -0.1639707 -0.2390375
3    0.1703467  0.4611813 -0.07449435 -0.05862745 -0.1893360 -0.1498717 -0.1531697  0.07799166 -0.1165969 -0.2343432
  PropCat_14  PropCat_15 Pur_Vol_No_Promo____ Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__   Pr_Cat_1   Pr_Cat_2   Pr_Cat_3
1 -0.4197922  0.11075889           -0.5257294         0.5079150            0.2186047  0.8909625 -0.3547940 -0.4189682
2  2.6418086 -0.25023289            0.2095598        -0.4162398            0.2828653 -0.7883740 -1.3244172  2.6382534
3 -0.2333429 -0.02908687            0.3365392        -0.2817596           -0.2150101 -0.4824369  0.5245339 -0.2332156
     Pr_Cat_4
1 -0.09959753
2 -0.31587119
3  0.13590107
```

K-means clustering with 4 clusters of sizes 229, 55, 190, 57




```
Cluster means:
  No__of_Brands Brand_Runs Total_Volume No__of__Trans       Value Avg__Price      maxBr Others_999 Brand_Jumping
1     0.3483677  0.5964179 -0.11703264     0.5032050  0.19822472  0.6340085 -0.5780871  0.4974625    -0.4033310
2    -0.5094034 -0.7451815  0.36730536    -0.2624206 -0.44472331 -1.3630202  1.4863317 -1.2251113     1.1140321
3    -0.0856985 -0.4137524  0.00888054    -0.4339981 -0.00868634 -0.1360113  0.5077300 -0.5460990     0.2512180
4    -0.6223914 -0.2979256  0.08616449    -0.3217748 -0.33830304 -0.7785911 -0.8041231  1.0038775    -0.2919363
  ProDifBrands  PropCat_5  PropCat_6   PropCat_7    PropCat_8   PropCat_9 PropCat_10  PropCat_11 PropCat_12 PropCat_13
1   -0.1871170 -0.3525259  0.3583968  0.22545074  0.27624258  0.23345257  0.2183264 -0.06636154  0.2329805  0.3145964
2   -0.1905857 -1.2052422 -0.1976499 -0.47023451 -0.50705004 -0.14225381 -0.2691601 -0.28665077 -0.1619770 -0.2380939
3    0.3662096  0.5266925 -0.2619154 -0.03786767 -0.18687275 -0.23126067 -0.1167381  0.21996334 -0.1587718 -0.2304290
4   -0.2850497  0.8235994 -0.3761089 -0.32579761  0.00235128 -0.02977461 -0.2282932 -0.19000789 -0.2504766 -0.2660687
  PropCat_14   PropCat_15 Pur_Vol_No_Promo____ Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__   Pr_Cat_1   Pr_Cat_2   Pr_Cat_3
1 -0.3651749  0.097527967          -0.1944181         0.1837736           0.08784702  0.7128352 -0.1315607 -0.3684363
2  2.6622576 -0.249248209           0.2498719        -0.4669854           0.28655347 -0.7846711 -1.3279358  2.6591213
3 -0.2664955  0.007932056           0.3788597        -0.3158629          -0.24419848 -0.4954670  0.8450710 -0.2692968
4 -0.2134207 -0.177760414          -0.7228866         0.7651577           0.18456671 -0.4551511 -1.0070109 -0.1879540
    Pr_Cat_4
1 -0.2752906
2 -0.3427623
3 -0.3024412
4  2.4448651
```

	kmClus_pbbp	SEC	HS	SEX	EDU	Affluence_Index	AGE	CHILD	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs
1	1	2.122271	4.524017	1.965066	5.205240	23.19651	3.314410	2.969432	0.2015197	4.331878	42.06550	23.048035	12092.14	1623.048	2.141616
2	2	3.363636	5.254545	1.963636	2.890909	10.70909	3.163636	3.145455	0.7722000	3.000000	29.58182	9.290909	15728.91	1074.080	5.900727
3	3	2.557895	4.794737	1.947368	4.531579	18.43158	3.252632	3.000000	0.5016789	3.657895	26.78421	12.689474	13037.59	1446.381	2.611053
4	4	3.315789	4.807018	1.982456	3.684211	14.19298	3.350877	3.052632	0.1390351	2.824561	28.61404	13.877193	13617.89	1164.945	2.455789

For each clustering in Q3 and in Q4 below:

(i) Describe your rationale for experimenting with different values of k.

When determining k, we wanted to figure out the best number of clusters to base our results on purchase behavior, the basis-for-purchase, and the both of them combined together. When deciding this number, we wanted to consider if generating more clusters per graph would either make our analysis either more informative, or more chaotic, since adding more clusters onto a graph for analysis can actually have negative effects when trying to draw comparisons between them. One example of a method we used to determine the optimal number k was the fviz_nbclust command to give us graphs of the optimal k number of clusters:

fviz_nbclust(xpb, kmeans, method = "wss")

fviz_nbclust(xpb, kmeans, method = "silhouette")

These two examples showcase the “Elbow Method” (looking at the “elbow/knee” of the graph line) and the average silhouette method.

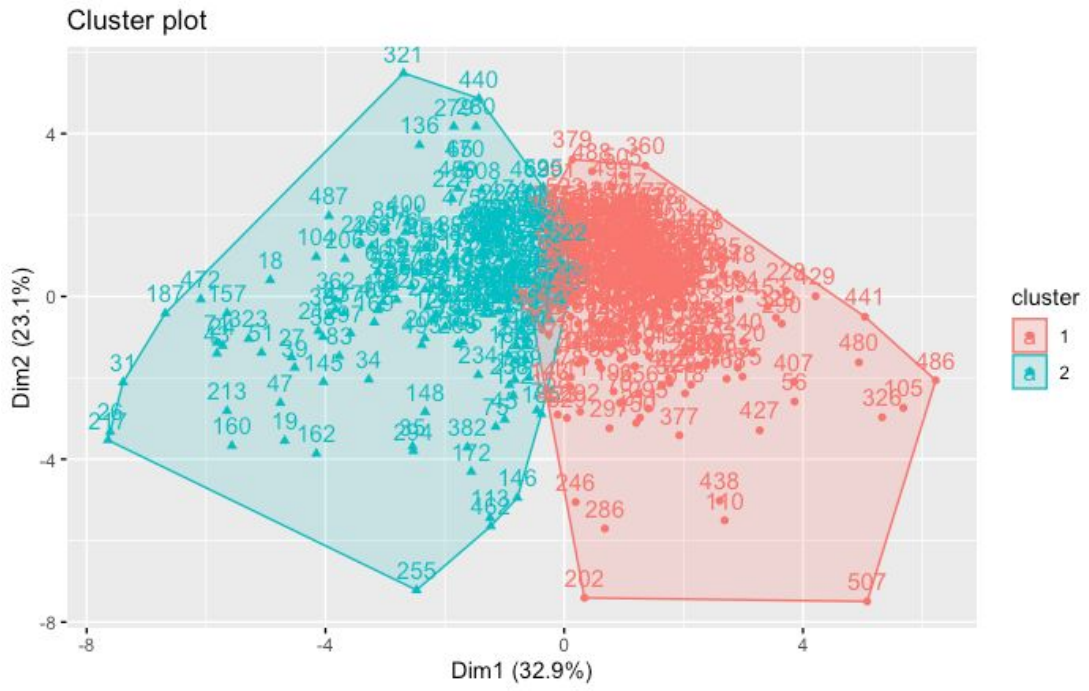
(ii) Evaluate the clusters – based on generic performance measures for clustering.

(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.

4.Try two other clustering methods for the questions above - from k-medoids, kernel-k-means, and DBSCAN clustering.

K-medoids

Pam partitioning
pam_pb<-pam(xpb,k=2,metric="euclidean")



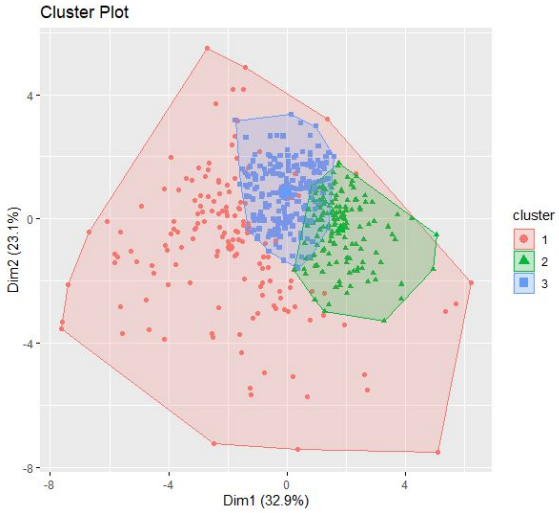
w

```
Medoids:
      ID No__of_Brands Brand_Runs Total_Volume No__of__Trans       Value Trans___Brand_Runs   Vol_Tran  Avg__Price
[1,] 424     0.7786595  0.5917335   0.01719226     0.3151974  0.09451801         -0.3838420 -0.3321342 -0.01316759
[2,] 243    -0.5094034 -0.8710714  -0.15593926    -0.5434240 -0.30603047          0.1476848  0.2128521 -0.46191161
          maxBr Others_999 Brand_Jumping ProDifBrands
[1,] -0.3665362  0.1978824    -0.3881353   0.02056919
[2,]  0.9936324 -0.9225628     0.4554558  -0.11768700
```

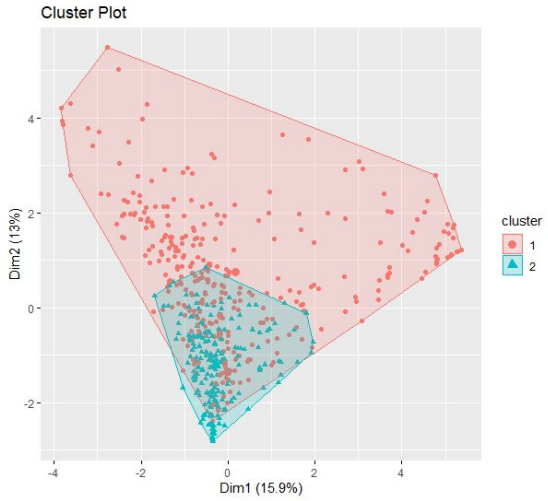
Kernel-k-Means

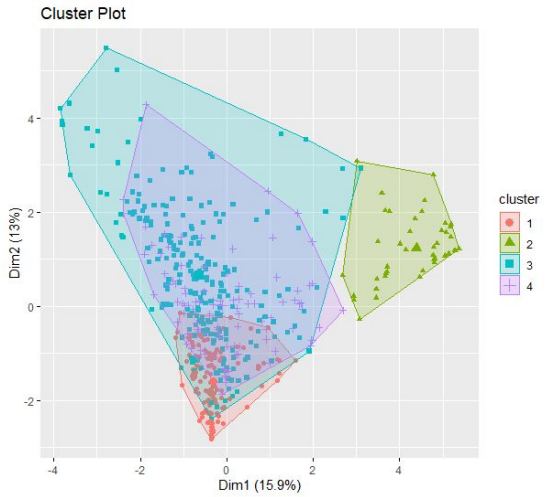
When observing Kernel-k-Mean clusters, the initial k that's set for each one is 3. The k is then changed for experimentation to 2 and 4.

Purchase Behavior Kernel-k-Means (k is set to 2, 3, and 4 respectively)

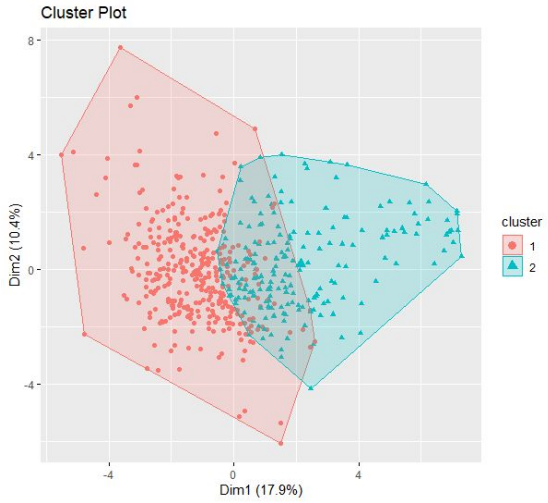


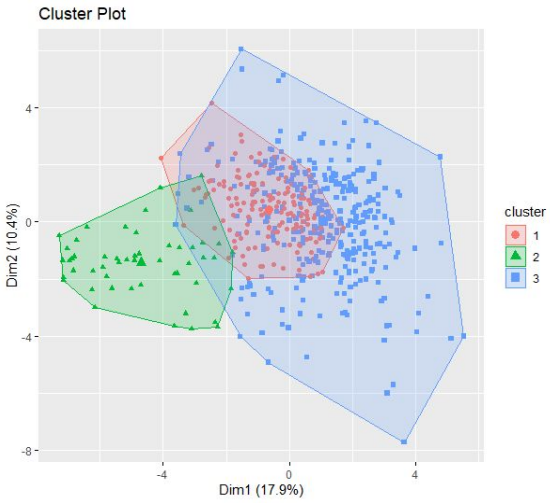
Basis of Purchase Kernel-k-Means

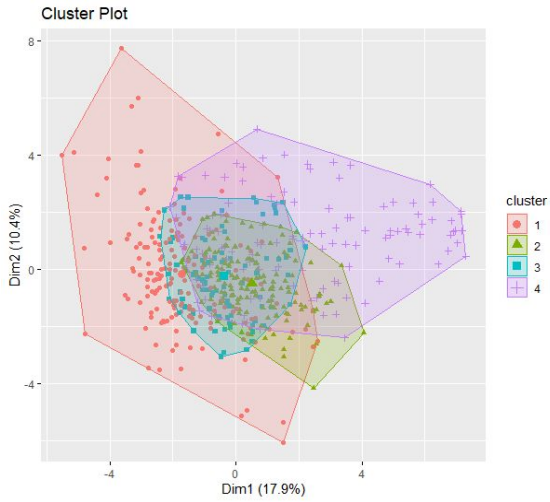




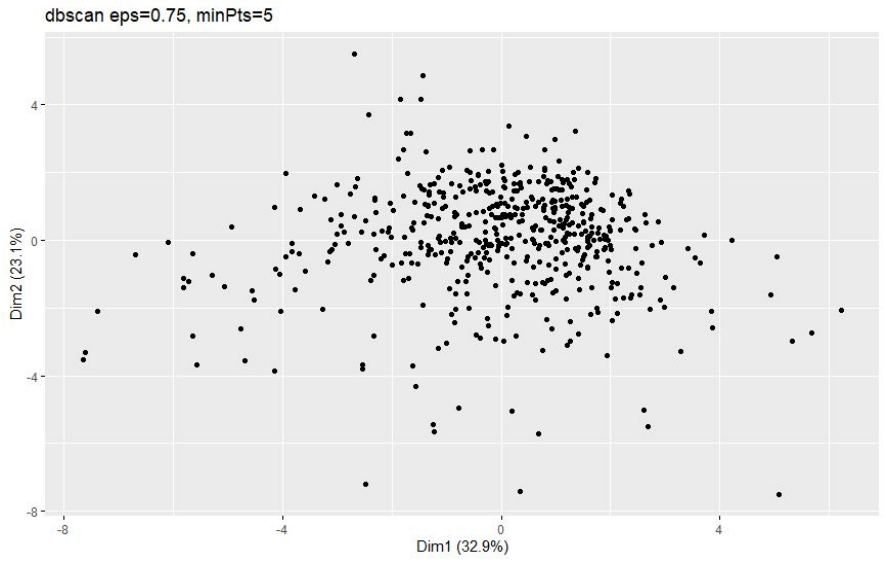
Purchase Behavior + Basis of Purchase Kernel-k-Means





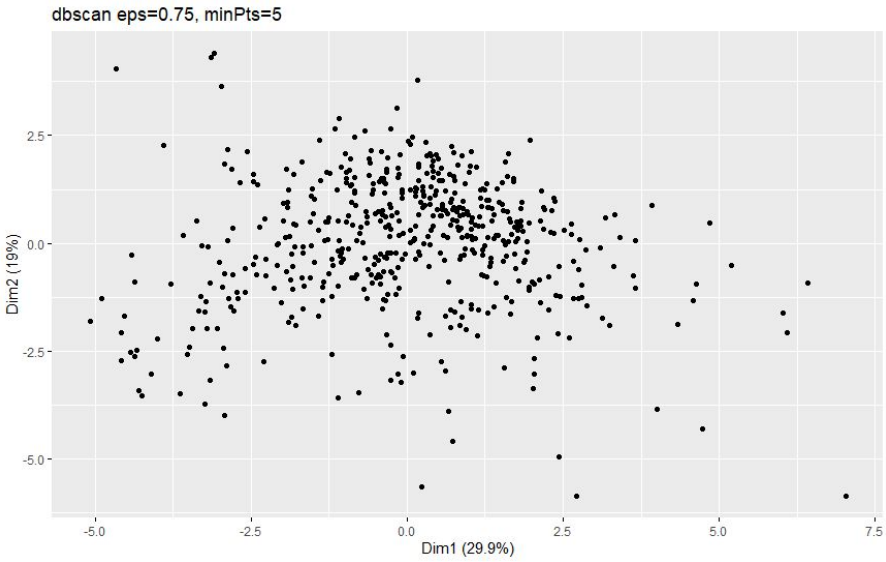


DBSCAN Purchase Behavior (we changed k for each graph and it returned the same result visually for us to analyze)



DBSCAN Basis for Purchase

DBSCAN Purchase Behavior & Basis for Purchase



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

After going through and evaluating each method with different parameters, we noticed that there was not much change within each visual analysis. When looking at the Purchase Behavior Kernel-k-Means, we observe that values 2-4 for k don't affect the fact that smaller clusters are overlapped by one bigger, making it seem pointless to pick an optimal value of k besides 1. As for the Basis of Purchase Kernel-k-Means, we notice this same trend continuing up until when k=4, as another separate cluster develops from the others, meaning that we can pick an optimal # for k that's higher than 4 to start clustering this model from. Oddly enough for the Purchase Behavior + Basis of Purchase Kernel-k-Means, we notice an opposite trend continue: the higher the k value is, the more congestion occurs with each cluster overlapping each other. We can probably attribute this to there being more variables within the clusters, and that maybe the optimal number for k is =2 since there are 2 clearly defined clusters.

In terms of determining the clusters for DBSCAN, we can tell that this method doesn't work for this set of data since the only clusters that emerge from the data occur for the Basis of Purchase. Even so, it's difficult to tell where they are because of the black points on each visualization.

After the Data:

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.

Note: running dbscan we noticed that this method of clustering did not produce useful results for analysis or comparison purposes so we will be ignoring these. Taking a look below we listed out the various clustering models between K-Means and Kernel K-Means procedures which do produce different sized clusters. The interpretability of the plots gave us a clear indication that the Kernel K-means method did not produce the best results due to a lack of separation between clusters which left us with focusing on the K-means clustering method to hone in on the ideal segmentations.

We used the k-medoids since it uses a good separation between clusters and closeness of points within the clusters which shows us that the segmentation is potentially relevant. If we were to apply this clustering for our prediction of sales of bar soaps to better target our consumers this could help improve our segmentation. We also noticed that the k means work better with k=2 and k=3. The clusters seem to have better separation and the clusters were better distributed from the center.

Purchase behavior

K-means clustering with 3 clusters of sizes 150, 305, 76(Best cluster)
K-means clustering with 2 clusters of sizes 347, 184
K-means clustering with 4 clusters of sizes 63, 256, 47, 165

Basis for purchase
K-means clustering with 2 clusters of sizes 62, 469
K-means clustering with 3 clusters of sizes 172, 62, 297
K-means clustering with 4 clusters of sizes 280, 101, 90, 60

Purchase and basis behaviour
K-means clustering with 2 clusters of sizes 57, 474
K-means clustering with 3 clusters of sizes 199, 56, 276
K-means clustering with 4 clusters of sizes 229, 55, 190, 57

Purchase behavior

Kernel K-means clustering with 2 clusters of sizes 287, 284
Kernel K-means clustering with 3 clusters of sizes 163, 199, 169
Kernel K-means clustering with 4 clusters of sizes 142, 131, 94, 164

Basis for purchase

Kernel K-means clustering with 2 clusters of sizes 329, 202
Kernel K-means clustering with 3 clusters of sizes 234, 153, 144
Kernel K-means clustering with 4 clusters of sizes 120, 90, 109, 212

Purchase and basis behavior

Kernel K-means clustering with 2 clusters of sizes 294, 237
Kernel K-means clustering with 3 clusters of sizes 206, 50, 275
Kernel K-means clustering with 4 clusters of sizes 132, 77, 190, 132

(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.

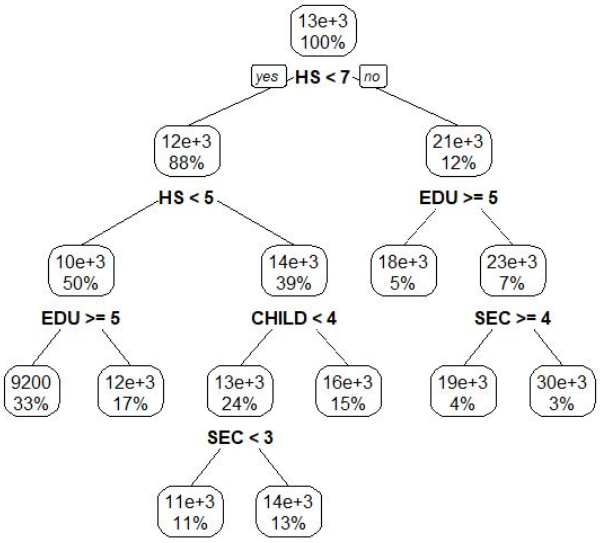
We believe the best segmentation is the clusters created from the K-means method producing 3 clusters based on the purchase behavior variables. The size of the clusters are 150,305 and 76.

	cluster	SEC	HS	SEX	EDU	Affluence_Index	AGE	maxBr	No__of_Brands	No__of__Trans	Brand_Runs	Total_Volume	Value	Trans___Brand_Runs	Brand_Jumping	ProDifBrands
1	1	2.560000	5.620000	1.973333	4.553333	21.30667	3.353333	0.2631133	4.820000	50.38000	26.26000	19797.433	2301.491	2.120133	0.2009333	0.1021333
2	2	2.432787	4.229508	1.950820	4.777049	19.50164	3.206557	0.2957148	3.573770	28.13443	15.31803	9373.761	1099.621	2.029443	0.2560328	0.1446230
3	3	2.894737	4.960526	1.973684	3.710526	14.05263	3.434211	0.8184605	2.631579	24.23684	5.00000	13933.421	1202.088	6.763816	0.6273684	0.1159211

(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: for the decision trees, volume was the variable that we focused our graphs around, because we wanted to analyze what information the decision trees can give us that can help us determine the clusters based on the amount of traffic for different customer segments.

Running a decision tree on different columns based on the entire data:

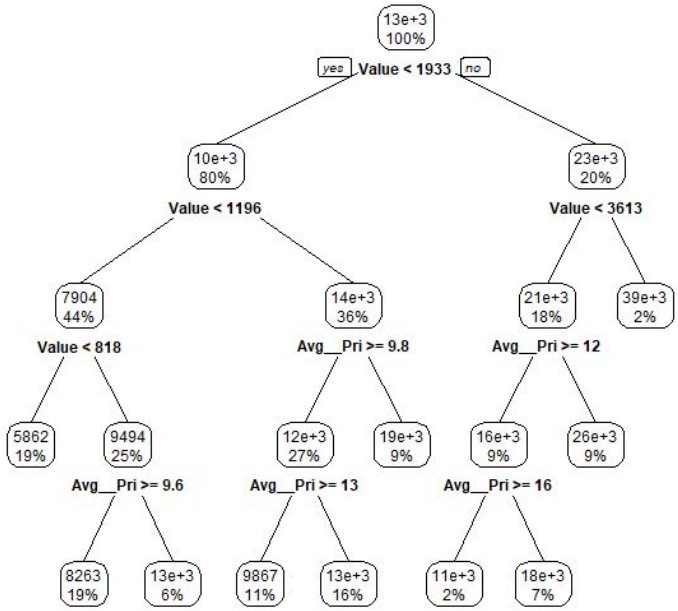


Variable Importance:

HS	SEC	EDU	CHILD	AGE	SEX
6324049179	1772474478	1412805562	1005473873	191825252	13717732

(These importance values don't make a lot of sense)

Trying the same concept for Purchase Behavior:

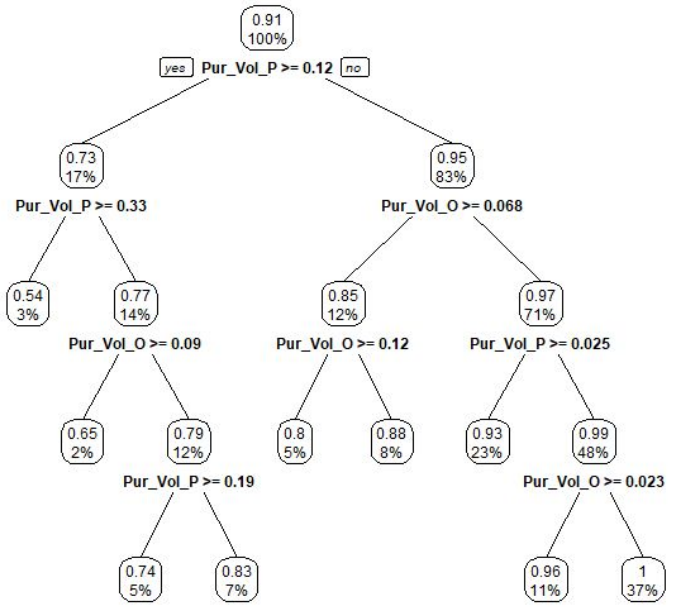


Variable Importance:

```
dt2$variable.importance
```

Value	Avg__Price	Vol_Tran	No__of__Trans	Brand_Runs	Others_999	maxBr Trans___Brand_Runs
20858744582	5343076791	4528517399	3839289811	1951273153	1049683024	947875154
ProDifBrands	Brand_Jumping					854379599
598094836	349330953					

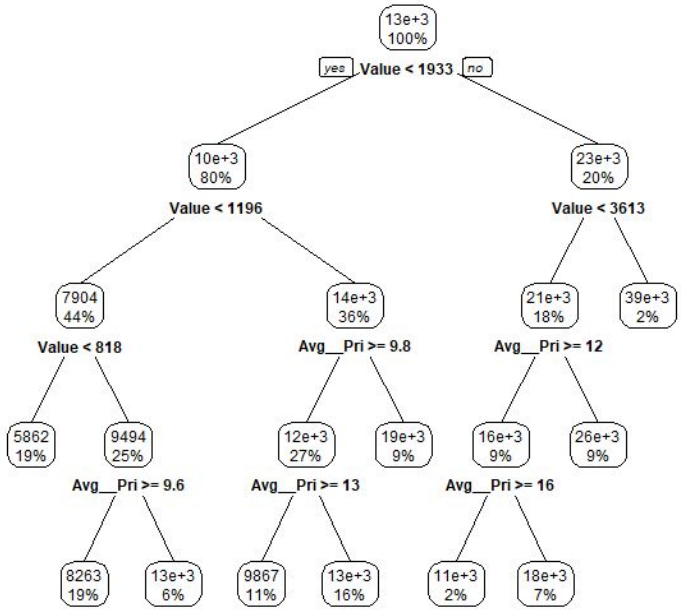
Basis for Purchase:



Variable Importance:

```
dt3$variable.importance
  Pur_Vol_Promo_6__ Pur_Vol_Other_Promo__          PropCat_8          PropCat_9          PropCat_6          Pr_Cat_3
        4.608031030          1.255750062        0.121774723        0.046083957        0.029302074        0.022874863
         PropCat_14          PropCat_11          PropCat_10          PropCat_13          PropCat_5          Pr_Cat_4
        0.022874863        0.015263876        0.011682867        0.009149945        0.007788578        0.002543979
         PropCat_15
        0.001313775
```

Purchase Behavior & Basis for Purchase:



Variable Importance:

```
dt4$variable.importance
               Value           Avg__Price      No__of__Trans          Pr_Cat_1         PropCat_13          Pr_Cat_3         PropCat_14
         20858744582          5556153716           3307646323        2769549345         1845570279        1783897150         1748235663
         Brand_Runs           Pr_Cat_2           PropCat_10          Pr_Cat_4         ProDifBrands      Brand_Jumping          PropCat_8
         1210168182          1111395491            839579566         710292686          598094836         476255783          114237032
Pr_Vol_No_Promo____           Others_999
          111254771           103908504
```

When looking at the decision trees, the two questions that we need to answer are as follows:

1. Do these trees give us a new way to look at the regular clusters based on the trees' leaves/clusters?
2. Do they allow us to interpret the clusters in a useful way, and give us new insights and information?

The answer to both of those questions is yes. We can answer #1 by looking at each leaf to describe different mini clusters, if you will, that can describe new areas of data to analyze. For example, in the first decision tree that we made, we see that the HS, EDU, SEC, and CHILD columns are set >= or <= to a certain number. These technically give us a brand new way to see different, more specific clusters within the data as a certain percentage and number of those particular groups belong to each leaf within the decision tree. For #2, viewing the clusters in this manner is useful because we can then use these groups in other real life examples to

make important decisions when segmenting different clusters and groups from each other. In this particular case of market segmentation, we can assist CRISA by looking at these different groups here like we did with the other clusters to determine which promotions to run on them to better understand how they can segment their market. The new insights and information come into play when we see our leaves and clusters on a more granular level; when we run clusters, we don't see specific information right away. We just see different points separated from one another in different groups and whether or not they overlap each other. However, with these decision trees, we can see now if certain nodes have an SEC (socio economic class) of 4 or higher for example, then see whether or not they have a household of more than 4 children, we can start to then run more promotions on this particular group and segment our data in similar fashion with different values to get a better understanding of our data, and later, a better understanding of how we can incur more revenue for our business.

(Note - you may develop decision trees for alternate clustering, and use these to help choose the 'best' clustering).