Segmenting Consumers of Bath Soap

Gordon Wall (gwall2)

12/1/2019

PROBLEM:

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 thus deploy promotion budgets more effectively. More effective market segmentation would enable CRISA's clients (in this case, a firm called IMRB) 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 the year. This would result in a more cost-effective allocation of the promotion budget to different market segments. It would also enable IMRB to design more effective customer reward systems and thereby increase brand loyalty.

PURCHASE BEHAVIOR CLUSTERING:

```
bath.soap = mutate(bath.soap, max.brand.loyal = apply(bath.soap[,23:30], 1,
max))
bath.soap = mutate(bath.soap, vol.br = Trans...Brand.Runs * Vol.Tran)
str(bath.soap)
## variable selection
bs.behavior = bath.soap[,c(12:16, 19, 31, 47:48)]
```

Variables selected were based on the parameters CRISA set for what's determined as "Purchase Behavior" (volume, frequency, and brand loyalty). **Volume** is covered 4 variables; Total.Volume, Value, Avg..Price, and a derived variable for AVG Volume per Brand Run, or vol.br (Avg Trans/Brand Run * Avg Vol/Trans). **Frequency** is covered by 2 variables; Brand.Runs and No..of..Trans. Most importantly, **Brand Loyalty** is covered by 3 variables; No..of.Brands, Others.999, and a derived variable for Maximum Brand Loyalty, or max.brand.loyalty (the max value for each household across all major brand categories). Considering Note 2 of the assignment, loyalty in this case is in a general sense. As stated, a consumer who buys all Brand A is equivalent in loyalty to a consumer who buys all Brand B; the scope of this problem is for general use and not to determine the loyalty of consumers to any one, specific brand. Therefore, clustering with the presence of a

maximum brand loyalty variable and the Others variable will show each households affinity to any major brand, as well as their likelihood to buy across many, minor brands. If a future cluster is centered around high max values, this cluster will be considered loyal. If a cluster is centered around high other.999 values, this cluster will be considered unloyal. Finally, a cluster with a center around low values of No..of.Brands will also be considered more loyal. A total of 9 variables are selected for analysis; We are ready to process this dataset now.

NOTE: Susceptibility to discounts was determined to be more fitting with the Basis of Purchase analysis and was moved accordingly. That way, clustering on purchase basis will be more robust and each clustering analysis will possess three factors to cluster on (Behavior: loyalty, vol, freq; Basis: price, sellprop, discount susc).

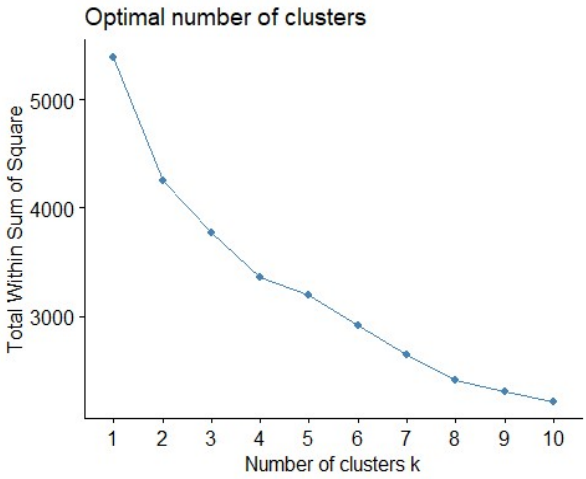
DATA PROCESSING

```
## normalize data (z-score)
scaled.behavior = scale(bs.behavior)
```

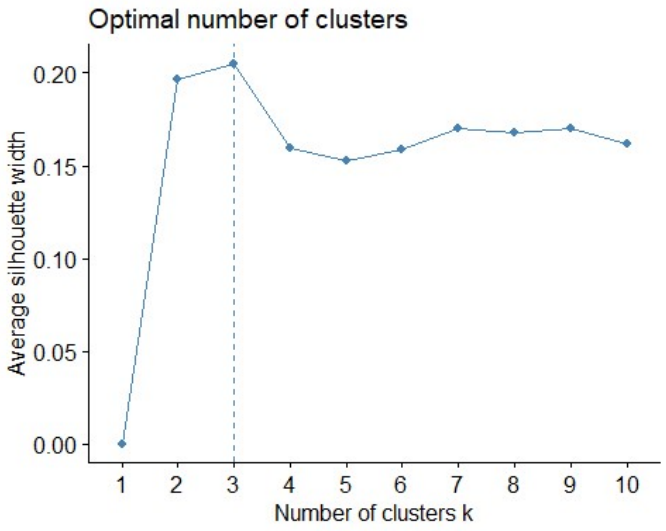
OPTIMAL CLUSTERS

Clusters will be important to the effectiveness of this study. CRISA has stated that they would implement anywhere between 2-5 marketing approaches when the study is complete, targeting consumers by both purchase behavior AND basis. The value of K should be chosen within those parameters and, thus, a maximum of 5 target groups will result from our final analysis. What this means is that each clustering analysis by either Behavior OR Basis should only possess a partial value of our total maximum clusters (5). We will consider K < 5 for each separate analysis before implementing the total in our final, combined analysis.

```
## optimal clusters
fviz_nbclust(scaled.behavior, kmeans, method = "wss")
```



```
fviz_nbclust(scaled.behavior, kmeans, method = "silhouette")
```



Elbow method reveals that an optimal number of clusters for the demographic data subset could be either **k = 2 or 3**. Silhouetting confirmation suggests that the optimal number of clusters is **k = 3**, but could also permit **k = 2** if needed. These results are congruent with K < 5 and, with that said, proceeding with 3 clusters for analysis seems reasonable.

DATA CLUSTERING

```
## clustering with 3 centroids
k3.behavior = kmeans(scaled.behavior, centers = 3, nstart = 20)
```

ANALYSIS & VISUALIZATION

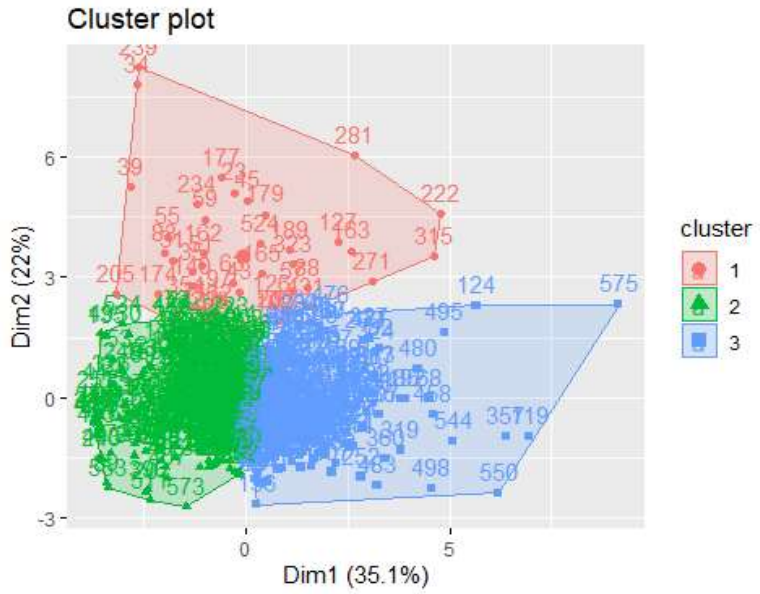
```
## analysis
k3.behavior$centers

##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1    -0.6974593 -0.8373175    1.7746590    -0.1369057  1.0483085
## 2    -0.4963588 -0.5590896   -0.5047393    -0.5816926 -0.5313059
## 3     0.7545089  0.8589412    0.3312993     0.7647961  0.4926129
##    Avg..Price Others.999 max.brand.loyal     vol.br
## 1 -0.95740979 -0.6255406     0.567867863  2.4428956
## 2 -0.05484281 -0.2043089    -0.008120829 -0.1145086
## 3  0.23787583  0.3699699    -0.089324977 -0.2829299

k3.behavior$size

## [1]  43 312 245
```

```
##visualization
fviz_cluster(k3.behavior, data = scaled.behavior)
```



Cluster 1 (43 obsv) is very loyal, indicated by high max.brand.loyalty, low others.999, and low no.of.brands values. While not frequent purchasers, they buy high volume with high value. Cluster 2 (312 obsv) is a middle-ground cluster. Their loyalty is ambiguous, with low volume and low value (this would be a hard segment to market to and, perhaps 2 clusters would be more encompassing of the data. Will re-visit later). Cluster 3 (245 obsv) is considerably unloyal and likes to shop around. They have mid-range volume and value.

Cluster 1 & 3 both prove promising in light of purchase behavior. Cluster 1 could be useful to major brands who want to secure and maintain their market share by catering to loyal consumers who by in high volume. Cluster 3 could be useful to newer, minor brands looking to take over market share from consumers who buy/spend much but aren't loyal to one brand (yet).

Graphing shows somewhat strong clustering, with only a small amount of overlap and dense populations around each of the three centroids.

PURCHASE BASIS CLUSTERING

DATA PRE-PROCESSING

```
## variable selection
colMeans(bath.soap[,c(36:46)])

##  PropCat.5  PropCat.6  PropCat.7  PropCat.8  PropCat.9 PropCat.10
## 0.46641667 0.15065000 0.14806667 0.13165000 0.07208333 0.04271667
## PropCat.11 PropCat.12 PropCat.13 PropCat.14 PropCat.15
## 0.05296667 0.01740000 0.04771667 0.14656667 0.04840000
bs.basis = bath.soap[,c(20:22, 32:36)]
```

Variables selected were based on the parameters CRISA set for what's determined as "Purchase Basis" (price, discount susceptibility, and selling proposition). **Price** is covered by 4 variables; every price catergory variable (Pr.Cat. 1-4). **Discount Suscetibility** is covered by 3 variables; all promotion catergory variables (pur.vol.no.promo, pur.vol.promo.6, and pur.vol.promo.other). **Selling Proposition** will be represented by a make-up of proposition category variables, however, there are 10 of these and not all of them are significant as some were very under-utilized by the consumer base. With an output of the variable means, we can see that Proposition Category 5 (prop.cat.5) has an overwhelming amount of purchase volume connected to it (avg: ~47%). We will use this variable in analysis. Price categories will represent which households are willing to purchase from each pricing level, actively showing us which consumer can pay for what. Promotion categories will represent which households bought high volumes of product under either promotion 6, no promotion, or all other promotions, actively showing which consumers only buy with promotion and which do not. And, finally, Selling Proposition categories will represent which consumers bought high volumes of product under the top 2 propositions. To remain consistent with using about 9 variables in analysis like we did for Purchase Behavior, a total of 8 variables are selected for analysis; We are ready to process this dataset now.

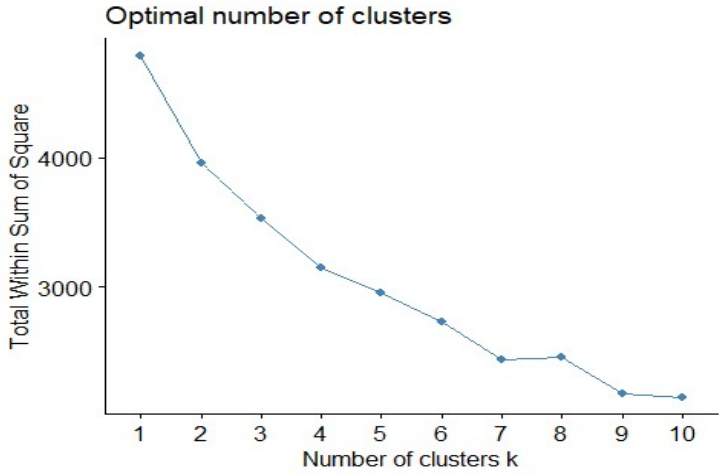
DATA PROCESSING

```
## normalize data (z-score)
scaled.basis = scale(bs.basis)
```

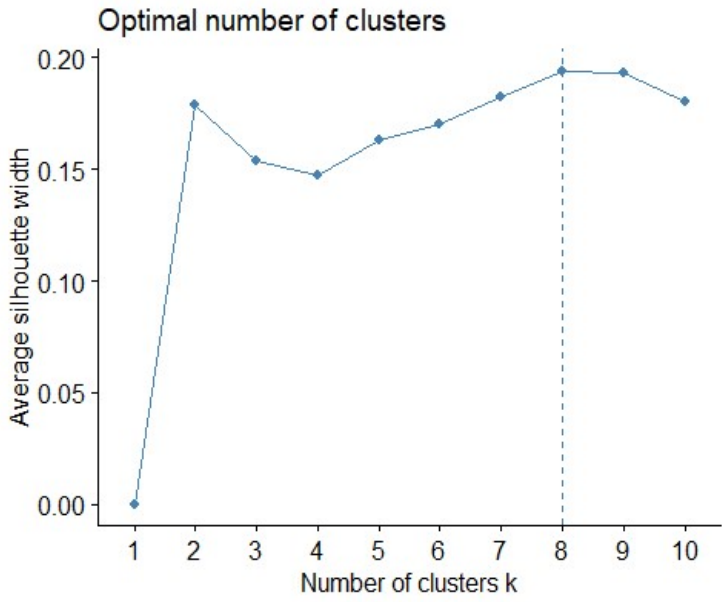
OPTIMAL CLUSTERS

Again, we will consider k < 5 to fit within the parameters of our overall goal of 2-5 marketing approaches.

```
## optimal clusters
fviz_nbclust(scaled.basis, kmeans, method = "wss")
```



```
fviz_nbclust(scaled.basis, kmeans, method = "silhouette")
```



The Elbow method from minimizing the total WSS indicates that an optimal number of clusters could be **k = 2**. Silhouetting further indicates this with the highest width spiking above **k = 2** as well. We will proceed with 2 clusters for analysis.

NOTE: Silhouetting indicates that 9 clusters is optimal, but this doesn't fit within our real-world parameters. CRISA has specified 2-5 total segments and 9 would be far too complicated to market to anyway (overlap, inefficient cost allocation).

DATA CLUSTERING

```
## clustering with 2 centroids
k2.basis = kmeans(scaled.basis, centers = 2, nstart = 20)
```

ANALYSIS & VISUALIZATION

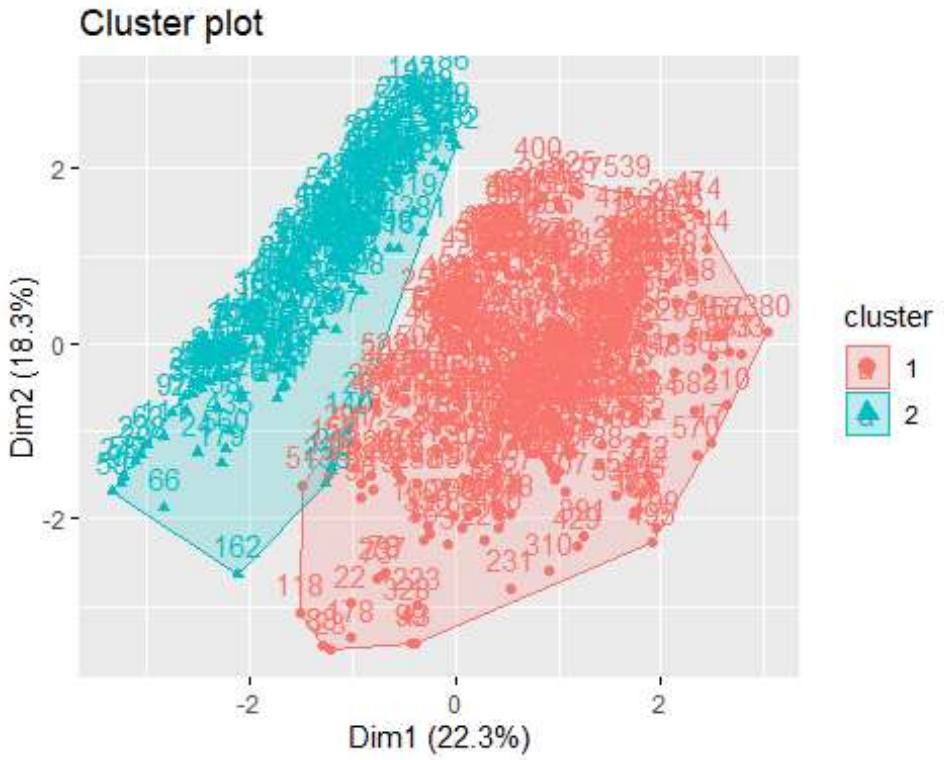
```
## analysis
k2.basis$centers
```

```
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
## 1            0.6531426         0.3735116             0.3664149  0.1056717
## 2           -1.2400244        -0.7091307            -0.6956573 -0.2006232
##      Pr.Cat.2    Pr.Cat.3    Pr.Cat.4   PropCat.5
## 1  0.09221621 -0.04906502  0.05604104  0.04877122
## 2 -0.17507715  0.09315244 -0.10639676 -0.09259463
```

```
k2.basis$size

## [1] 393 207

##visualization
fviz_cluster(k2.basis, data = scaled.basis)
```



Cluster 1 (393 obsv) shows a higher susceptibility to discounts, yet also buys much without them. This is useful to know because cluster 1 will provide a strong base level of profit for a brand, but will respond positively and more frequently to discounts offered. Further, cluster 1 responds well to the various pricing categories with an emphasis on category 1. They also responded positively to selling proposition five, and will make an all-around easy segment to target with promotions and discounts. Cluster 2 (207 obsv) is logically the opposite. They do not respond well to discounts, nor promotions, and do not contribute to nearly as many purchases in the various pricing categories. This is good information to know so brands can tailor current products to the customers who don't buy as much, or potentially research new products to fit the demographics of cluster 2 so they become more frequent, willing, and susceptible buyers.

Graphing shows somewhat strong clustering, with virtually no overlap and dense populations near each of the two centroids.

PURCHASE BEHAVIOR AND BASIS CLUSTERING

DATA PRE-PROCESSING

```
bs.bb = bath.soap[,c(12:16, 19:22, 31:36, 47:48)]
```

This is a combination dataframe of both the selected variables from behavior and basis analysis.

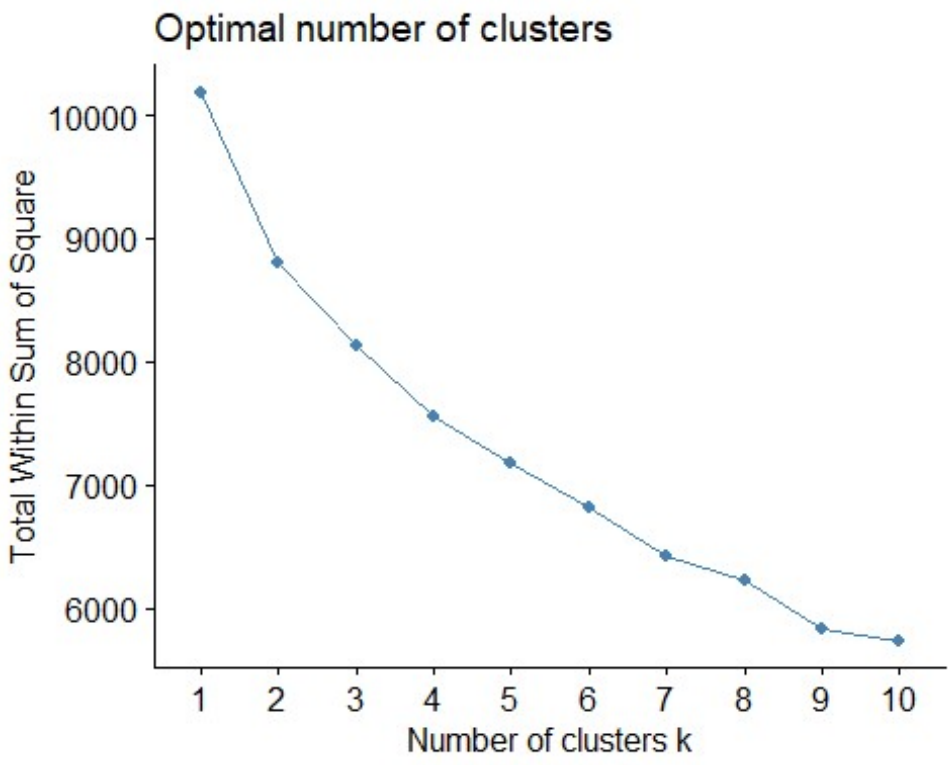
DATA PROCESSING

```
## normalize data (z-score)
scaled.bb = scale(bs.bb)
```

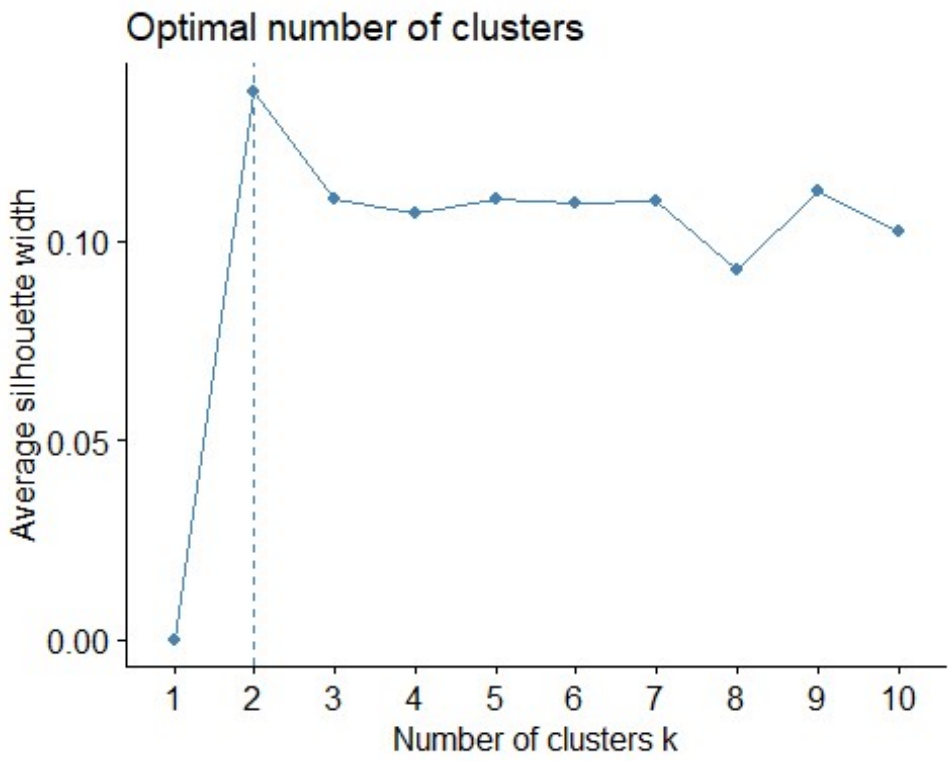
OPTIMAL CLUSTERS

Again, we will consider k <= 5 to fit within the parameters of our overall goal of 2-5 marketing approaches. Now we can consider up to and including 5 clusters. Let's see what WSS and Silhouetting have to say.

```
## optimal clusters
fviz_nbclust(scaled.bb, kmeans, method = "wss")
```



```
fviz_nbclust(scaled.bb, kmeans, method = "silhouette")
```



Both methods lean toward **k = 2** as a good cluster amount. However, with this many variables being considered, too few clusters might not capture some of the segmented differences we're looking for. Thus, we will try multiple configurations and determine the best.

DATA CLUSTERING

```
## clustering with 2 centroids
k2.bb = kmeans(scaled.bb, centers = 2, nstart = 20)

## clustering with 3 centroids
k3.bb = kmeans(scaled.bb, centers = 3, nstart = 20)

## clustering with 4 centroids
k4.bb = kmeans(scaled.bb, centers = 4, nstart = 20)

## clustering with 5 centroids
k5.bb = kmeans(scaled.bb, centers = 5, nstart = 20)
```

ANALYSIS & VISUALIZATION

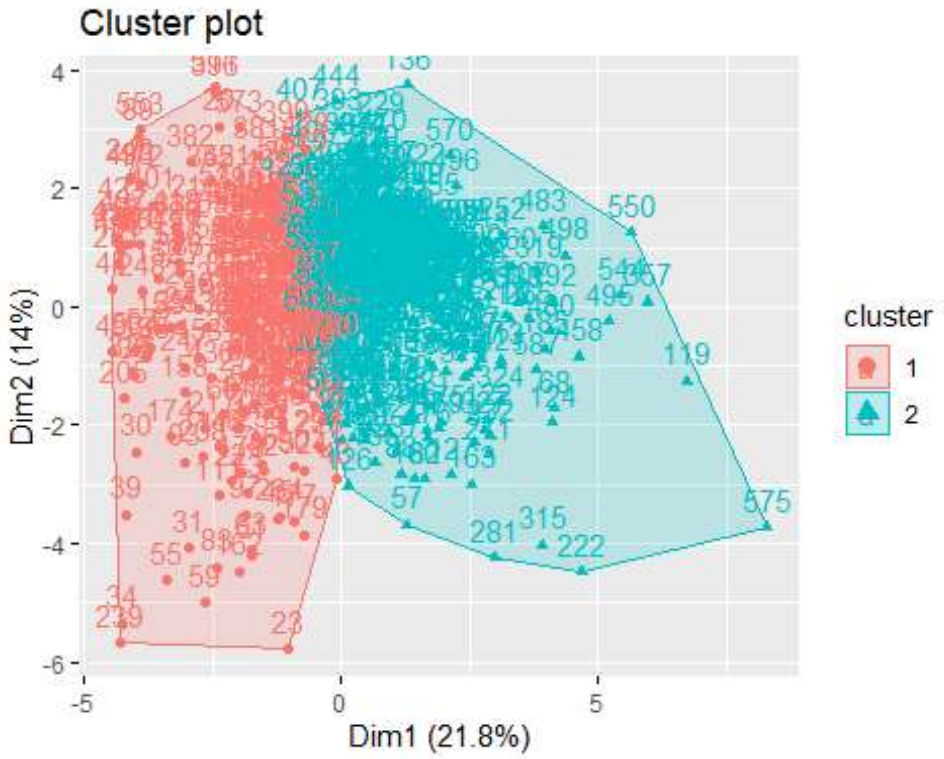
```
## analysis of 2 clusters
k2.bb$centers
```

```
##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1    -0.6689573 -0.8141212   -0.3597434    -0.6960649 -0.5503022
## 2     0.3845187  0.4679594    0.2067816     0.4001003  0.3163155
##   Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1 -0.3044237           -0.7632083        -0.5990353            -0.3848957
## 2  0.1749837            0.4386945         0.3443274             0.2212392
##   Others.999   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3    Pr.Cat.4  PropCat.5
## 1 -0.4915316 -0.3310407 -0.3946812  0.2587869 -0.09089833 -0.1943789
## 2  0.2825339  0.1902832  0.2268640 -0.1487515  0.05224865  0.1117296
##   max.brand.loyal     vol.br
## 1      0.03953568  0.3858398
## 2     -0.02272523 -0.2217819
```

```
k2.bb$size
```

```
## [1] 219 381
```

```
##visualization
fviz_cluster(k2.bb, data = scaled.bb)
```



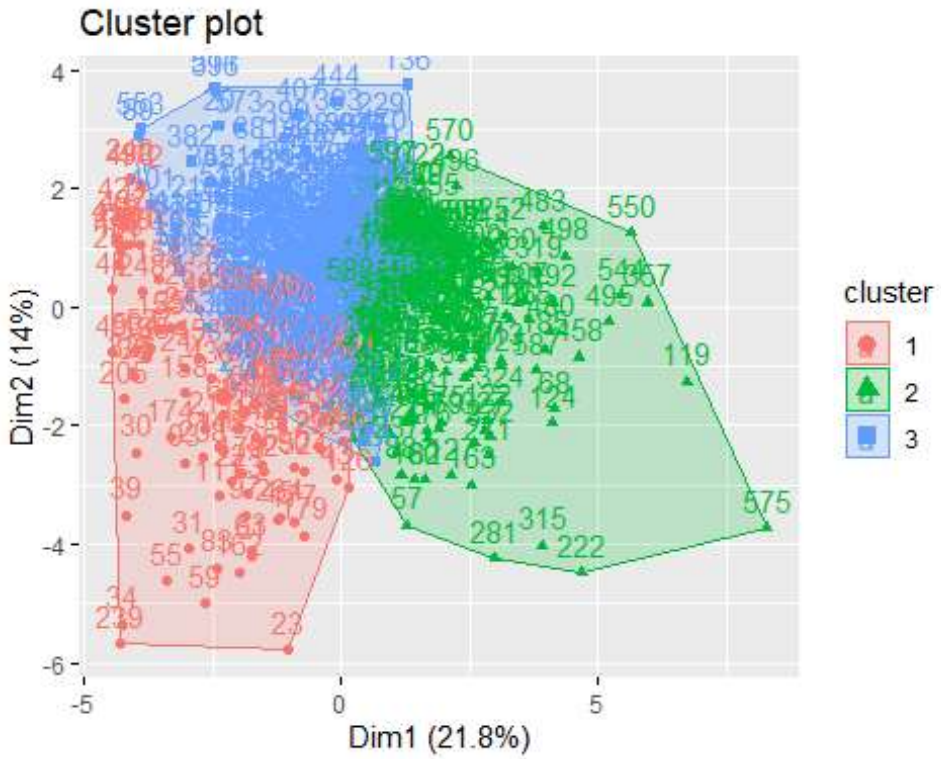
```
## analysis of 3 clusters
k3.bb$centers
```

```
##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1    -0.7934935 -0.9460681  -0.06721598    -0.6469651 -0.5530054
## 2     0.7846531  0.9243742   0.58836920     0.8945483  0.6974310
## 3    -0.2822380 -0.3282307  -0.41118524    -0.4191868 -0.3084415
##    Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1 -0.97798379           -0.5409478       -0.56474765            -0.2005059
## 2  0.09937947            0.4564311        0.40539737             0.3035853
## 3  0.29602324           -0.1341665       -0.08729841            -0.1494656
##    Others.999   Pr.Cat.1    Pr.Cat.2    Pr.Cat.3   Pr.Cat.4   PropCat.5
## 1 -0.76221508 -0.7338001 -0.96609032  1.00545404 -0.1037724 -0.85955722
## 2  0.33930291  0.1111678  0.07467922  0.01376655  0.2058386 -0.07806222
## 3  0.03640778  0.1949518  0.30985522 -0.39036946 -0.1135069  0.38291763
##   max.brand.loyal     vol.br
## 1      0.12956405  0.9441033
## 2     -0.05116297 -0.2225696
## 3     -0.01102166 -0.1918001
```

```
k3.bb$size
```

```
## [1] 107 210 283
```

```
##visualization
fviz_cluster(k3.bb, data = scaled.bb)
```



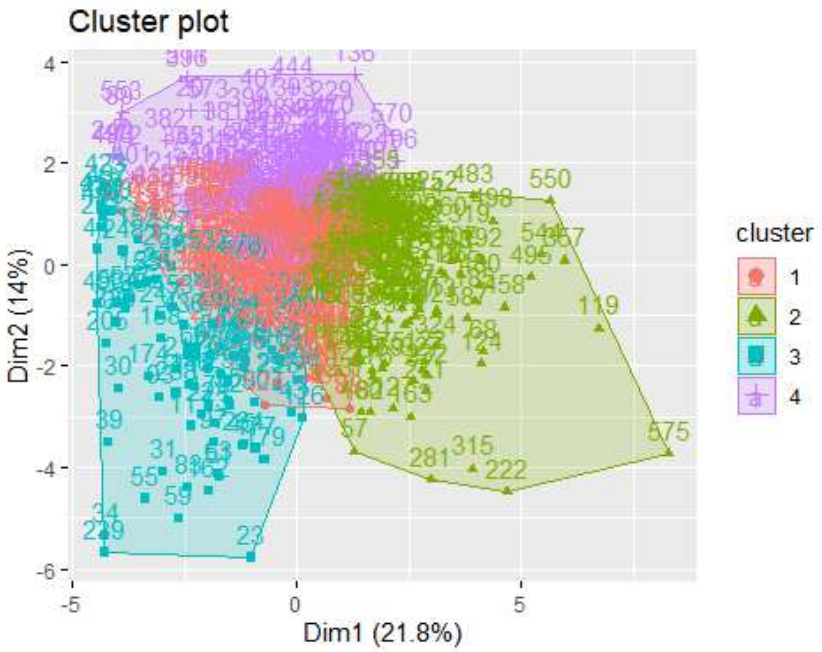
```
## analysis of 4 clusters
k4.bb$centers

##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1    -0.2633455 -0.2998961  -0.07423707    -0.3043260 -0.1691462
## 2     0.9529255  1.0614472   0.63663428     1.0322151  0.7339641
## 3    -0.7814683 -0.9724793  -0.07478773    -0.6499729 -0.5800643
## 4    -0.2623549 -0.2066557  -0.67379545    -0.3954396 -0.2749672
##    Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1 -0.29887674           -0.1654704      -0.129011272           -0.24062580
## 2  0.04005069            0.5565102       0.459180736            0.36629914
## 3 -1.08713501           -0.4483163      -0.538201755           -0.13358819
## 4  1.24530635           -0.1393646      -0.004822078            0.01300678
##    Others.999    Pr.Cat.1   Pr.Cat.2    Pr.Cat.3   Pr.Cat.4  PropCat.5
## 1  0.01941382 -0.29960753  0.5585710 -0.27014312  0.2800319  0.7414333
## 2  0.27496797  0.05860053  0.1802464  0.06552744  0.1061204 -0.1525075
## 3 -0.82852389 -0.72597038 -0.9863585  1.15608832 -0.2583602 -1.0008986
## 4  0.20332911  0.95795389 -0.4658118 -0.47534745 -0.4270793 -0.3180068
##   max.brand.loyal      vol.br
## 1      0.14437824 -0.09729109
## 2     -0.05686947 -0.23791702
## 3      0.29707592  1.08203620
## 4     -0.38470325 -0.30647898

k4.bb$size

## [1] 213 169  92 126

##visualization
fviz_cluster(k4.bb, data = scaled.bb)
```



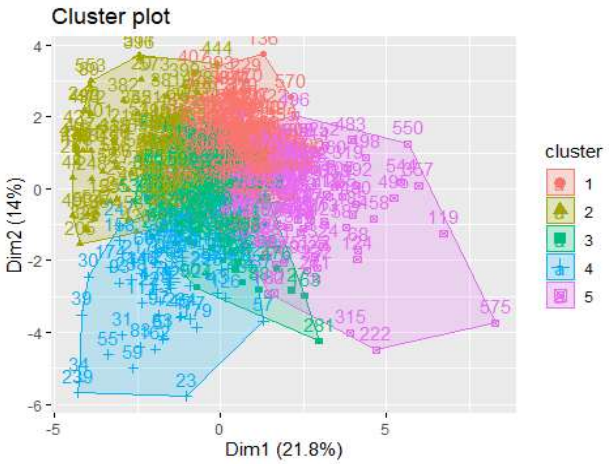
```
## analysis of 5 clusters
k5.bb$centers
```

```
##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1     0.0539875  0.1651161   -0.3465337   -0.02307384 -0.2338782
## 2    -0.8385974 -0.8832661   -0.8642682   -0.99533693 -0.8067789
## 3    -0.1510458 -0.3477854    0.3738130   -0.22383935  0.3269236
## 4    -0.5564891 -0.7456049    0.5086826   -0.20441679 -0.2296750
## 5     1.1335527  1.2603336    0.7963150    1.28007542  1.0031569
##   Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1  0.2256255           0.50979797         0.6079096             0.2616937
## 2  0.2234448          -0.98083724        -0.6516114            -0.5152960
## 3 -0.1529591          -0.46624274        -0.4829418            -0.3553937
## 4 -1.2493786          -0.01872256        -0.4603329             0.1587386
## 5  0.2359407           0.44155013         0.2267449             0.2448620
##   Others.999   Pr.Cat.1    Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5
## 1  0.4705526  0.2947515  0.08019933 -0.3643955  0.2914127  0.1274780
## 2 -0.3664026 -0.2103280 -0.57513878 -0.2440166 -0.1840191 -0.5178024
## 3 -0.4270261 -0.1368771  0.89765156 -0.2781831 -0.1614638  1.1598864
## 4 -0.7249199 -0.6085017 -0.89494931  1.7318455 -0.1551607 -0.7936049
## 5  0.3017350  0.1432577  0.10989580  0.1337603 -0.1094884 -0.3143838
##   max.brand.loyal      vol.br
## 1     -0.26816109 -0.31643106
## 2     -0.69347274 -0.06866315
## 3      0.78487833  0.09332442
## 4      0.81239217  1.37748354
## 5     -0.03332307 -0.24080066
```

```
k5.bb$size
```

```
## [1] 205 109 103  66 117
```

```
##visualization
fviz_cluster(k5.bb, data = scaled.bb)
```



BEST SEGMENTATION

It appears the best make-up, given the output of our 4 models, is **k = 3**. Graphing here shows 3 distinct clusters around 3 centroids, with little overlap. Further, the three clusters seem to be indicative of groups that are each targetable in their own way. Cluster 1 (107 obsv) is extremely loyal to their chosen brand, and low in susceptibility to promotion or discount. They represent a segment that will buy what they want no matter what, and won't change preference easily. This good for potential clients to learn about their base market share. Cluster 2 (210 obsv) is the opposite; they are very unloyal and responsive to promotion 6/others. Further, they represent a segment that is high in transaction volume and value, consistently switching up their preference in search of the best deal or newest product. They could make or break any given CRISA client if a portion of them was captured in the market by promoting new products well and offering ample discounts. Cluster 3 (283 obsv) is a middle of the road type cluster. They are neither too loyal or too unloyal, but their average transaction price is high and they demand a strong presence in pricing categories 1 & 2 and proposition category 5. They seem to be the careful shopper, not eager to switch preference too fast, but willing to pay top dollar for the right things. This could also be valuable to a CRISA client when determining who to make and market their higher-end products to under a carefully planned promotion.

DEMOGRAPHICS

```
## map cluster numbers to observations
bs.clustered = cbind(bath.soap, Cluster =k3.bb$cluster)

bs.democlust = aggregate(cbind(SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, CS,
Affluence.Index) ~ Cluster, data = bs.clustered, mean, na.rm = FALSE)
bs.democlust
```

```
##   Cluster      SEC      FEH       MT      SEX      AGE      EDU       HS
## 1       1 3.093458 1.878505 7.504673 1.476636 3.112150 2.420561 3.962617
## 2       2 2.452381 2.380952 9.366667 1.923810 3.295238 4.638095 5.114286
## 3       3 2.310954 1.865724 7.551237 1.699647 3.190813 4.215548 3.593640
##      CHILD        CS Affluence.Index
## 1 3.485981 0.8504673        8.757009
## 2 2.952381 1.0523810       20.776190
## 3 3.346290 0.8727915       17.356890
```

Cluster 1 (107 obsv) seems to be comprised of lower socioeconomic class households with younger, less-educated homemakers and more average children. This makes sense given their high loyalty, low transaction volume, and lower presence in the price categories, as they have a lot of family-related bills to take care of and may not have the financial means to keep switching up their preference. They stick to what they know to be affordable and effective for their needs.

Cluster 2 (210 obsv) seems to be comprised of high socioeconomic class households with highly educated homemakers. They also have the highest affluence index average,

showing that this cluster contains people with a lot of valued possessions and spending potential. This is consistent with the results of the analysis, which show their high transaction volume and value, combined with their unloyalty and sampling of many brands. This cluster has money and likes to spend it with variable, newly-peaked interests.

Cluster 3 (283 obsv) seems to be comprised of the highest socioeconomic class households on average with high education and lower average age than cluster 2. They also possess decent affluence index. This cluster also has strong financial means supported by their analysis; they carry much weigh in pricing categories 1 & 2 and sport high-valued transactions. This customer is most likely a money-minded, successful mid-life man/woman who is conscious with their money, but will pay well for what they deem to be a quality product.

These results are necessary to determine the best segmentation of households. Because CRISA isn't working for any specific brand (they operate on general data analysis for this project to be used for any client), they can begin defining these consumer segments to sell to various clients based on those clients' individual goals. Does the client want to produce an affordable product that becomes the standard of the middle and lower class? The potential with marketing to Cluster 1 is staggering in this regard. They may not be as susceptible to promotion, but are highly loyal and could become impenetrable foundation of market share for a brand looking to control this market. Does the client want to appeal to affluent consumers with money to blow? Cluster 2 has low loyalty and high spending activity/power, making them a great target for the latest-and-greatest product on the market. Does the client have long-term branding goals? By making a high-quality and value-add product, they could easily target Cluster 3 which does their research and will buy in mass what they determine to be the best. They're also the largest cluster so the profit potential here is huge. Moving on to developing our classification model, we design this model to sort households into these three target segments with the most accuracy.

CLASSIFICATION MODEL

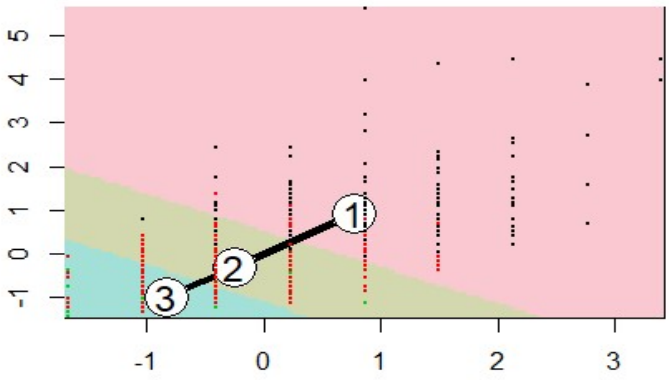
```
set.seed(123)

k3 = kcca(scaled.bb, k=3, kccaFamily("kmeans"))
k3
## kcca object of family 'kmeans'
##
## call:
## kcca(x = scaled.bb, k = 3, family = kccaFamily("kmeans"))
##
## cluster sizes:
##
##   1   2   3
## 207 290 103

clusters_index.train = predict(k3)
# clusters_index.test = predict(k3, newdata = bathsoap.test)

image(k3)
points(scaled.bb, col=clusters_index.train, pch=19, cex=0.3)

# points(bathsoap.test, col=clusters_index.test, pch=22, bg="red")
```



Here we can see a similar sized clustering to what our kmeans function originally produced for behavior and basis variables with 3 clusters. Using kcca with prediction methods, the above code shows a model that could be used to predict classifications of new data into the three target segments previously discussed. In the commented-out code "bathsoap.test" would be new data that CRISA provides. They sampled 600 observations for this project but are in possession of data from the entire Indian market to utilize for test

data. The commented-out code further shows our model predicting with the new data and then plotting the predicted observations with differently-sized and distinguishable points on the image of our original cluster to see where they match up to the training data clusters. Splitting our 600 observation into subsets for validation and test isn't as effective because kmeans clustering is unsupervised learning, and data partitioning is typically utilized in supervised learning. With new data present from CRISA, we could use nearest-neighbor classification on the centroids to predict what clusters the new data would belong to and section the Indian market into our 3 target segments for implementing marketing approaches.