Case 6 (CRISA) Bath Soap Solution Narrative

We look first at clusters based on purchase behavior, then clusters based on the basis for purchase, then clusters based on both. The complexity of marketing to 5 segments would probably not be supported by clustering just based on purchase behavior, or clustering just based on basis for purchase, so we will look at 2-3 clusters for those variables, and more when we cluster using both sets of variables.

In choosing k, we would seek a k that produces clusters that are distinct and separate from one another, in ways (variables) that are translatable into marketing actions. The variables we have been asked to consider are those that relate to purchase behavior (volume and frequency of purchase, brand loyalty), and a separate set that relate to the basis for purchase (response to promotions, pricing, and selling proposition).

Finally, we look at predictive models that classify customers into segments based on demographic data.

Clusters based on "purchase behavior"

Note: Some thought is needed about brand loyalty. For brand loyalty indicators, we have data on (1) percent of purchases devoted to major brands (i.e. is a customer a total devotee of brand A?), (2) a catch-all variable for percent of purchases devoted to other smaller brands (to reduce complexity of analysis), and (3) a derived variable that indicates the maximum share devoted to any one brand. Since CRISA is compiling this data for general marketing use, and not on behalf of one particular brand, we can say a customer who is fully devoted to brand A is similar to a customer fully devoted to brand B - both are fully loyal customers in their behavior. But if we include all the brand shares in the clustering, the analysis will treat those two customers as very different. So we will use only the derived variable for maximum purchase share for a brand, any brand, plus "max.brand.ind" and the "other.brand.ind," along with the purchase.ind (for volume, frequency, etc.). We will not use the individual values - "brand.ind."

output - 2 clusters

```
  No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value Trans...Brand.Runs
1     0.5657615  0.6822432    0.4080880     0.6179379  0.5331253         -0.2332518
2    -0.5508731 -0.6642894   -0.3973489    -0.6016764 -0.5190957          0.2271136
     Vol.Tran Avg..Price Others.999 max.brand.ind
1 -0.06467289  0.2205665  0.3799110    -0.4925259
2  0.06297097 -0.2147621 -0.3699134     0.4795647
> km$size
[1] 297 303
```

Comment: The two clusters are well-separated on everything, except transaction volume. Cluster 1 (n=297) is high activity & value, with low loyalty. Cluster 2 (n=303) is the reverse. ("Value" here is the meaning attached to the variable - total dollar value of purchases, not some broader meaning.)

Note: Due to the randomization element in the k-means process, different runs can produce different cluster results.

output - 3 clusters

```
  No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value Trans...Brand.Runs
1    -0.4869230 -0.7537170   0.06357029    -0.4467635 -0.1807197          0.6304400
2    -0.2789046 -0.2152595  -0.54334795    -0.4145185 -0.4639380         -0.2452216
3     0.8488821  1.0056324   0.71508972     1.0048322  0.8285120         -0.2353158
    Vol.Tran Avg..Price Others.999 max.brand.ind
1  0.4686479 -0.4790225 -1.1879685     1.3052831
2 -0.2847935  0.2360734  0.5998187    -0.5351342
3 -0.0289007  0.1079274  0.2472012    -0.4481440
```

Comment: Cluster 1 (n=282) is highly loyal, favoring main brands and bigger individual purchases, with middling overall value. Cluster 3 (n=163) is not at all loyal, favoring many brands, and of high value. Cluster 2 (n=155) is also not very loyal, but may be of the least interest since its customers have the lowest value.

Clusters based on "basis for purchase"

The variables used are: Pur_vol_no_promo, Pur_vol_promo_6, Pur_vol_other, all price categories, selling propositions 5 and 14 (most people seemed to be responding to one or the other of these promotions/propositions).

output - 2 clusters

```
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
1           -0.6618157         0.6133454             0.3062659  0.9065161
2            0.3859137        -0.3576499            -0.1785878 -0.5286017
    Pr.Cat.2   Pr.Cat.3    Pr.Cat.4   PropCat.5 PropCat.14
1 -0.4897589 -0.4120616  0.04392662 -0.16479110 -0.4095294
2  0.2855850  0.2402787 -0.02561420  0.09609191  0.2388021
```

Comment: The two clusters are well separated across most variables. Cluster 1 (n=77) responds to promotional offers and pricing category 1, and not to the two selling propositions chosen. Cluster 2 (n=523) purchases without needing promotional offers, likes pricing categories 2 and 3, and is somewhat responsive to the two selling propositions.

output - 3 clusters

```
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
1            0.2254203        -0.4365916             0.1897904 -0.7907928
2            0.3532652        -0.2908627            -0.2106668 -0.3833438
3           -0.6013139         0.5838886             0.2438889  0.8457362
```

```
    Pr.Cat.2   Pr.Cat.3   Pr.Cat.4   PropCat.5 PropCat.14
1 -1.1500950  2.4185104 -0.3535298 -1.12132288  2.4224950
2  0.7226866 -0.2959509 -0.1991685  0.30076127 -0.2955745
3 -0.6601121 -0.4186588  0.4191249 -0.04735407 -0.4206241
```

Comment: The clusters are well separated across most variables. Cluster 1 (n=74) is notable for its responsiveness to price category 3 and selling proposition 14 coupled with aversion to price categories 1 and 2, and selling proposition 5. Cluster 2 (n=97) is averse to promotions, likes pricing category 2, and is responsive to selling proposition 5. Cluster 3 (n=429) needs promotions, likes price categories 1 and 4, and is not responsive to the two selling propositions.

Clusters based on all of the above variables

output - 2 clusters

```
  No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
1    -0.3502754 -0.6333874    0.4567899    -0.3288100  0.04024226
2     0.1230697  0.2225415   -0.1604938     0.1155278 -0.01413917
  Trans...Brand.Runs   Vol.Tran Avg..Price Others.999 max.brand.ind
1          0.6311261  0.7877415 -0.8351173 -1.0478405      1.108272
2         -0.2217470 -0.2767740  0.2934196  0.3681602     -0.389393
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
1           0.26832815        -0.4055015            0.07840628 -0.7891886
2          -0.09427746         0.1424735           -0.02754815  0.2772825
    Pr.Cat.2   Pr.Cat.3    Pr.Cat.4  PropCat.5 PropCat.14
1 -0.1071528  1.1219292 -0.23817796 -0.3073079  1.1232341
2  0.0376483 -0.3941913  0.08368415  0.1079730 -0.3946498
> km$size
[1] 156 444
```

We can add demographic information:

```
             SEC      FEH       MT      SEX      AGE      EDU       HS    CHILD
meansc1 3.032051 2.128205 8.314103 1.717949 3.153846 3.288462 4.891026 3.288462
meansc2 2.313063 2.020270 8.130631 1.745495 3.234234 4.308559 3.945946 3.213964
               CS Affluence.Index
meansc1 0.9423077        12.97436
meansc2 0.9279279        18.44144
```

Comment: The two clusters are separated on almost all variables, Value being an important exception. Cluster 1 (n=156) is the more loyal, with lower socioeconomic status and affluence, and larger households.

output - 3 clusters

```
  No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
```

```
1     0.2661125  0.4647944   -0.2071309     0.2873510  0.02261697
2    -0.1901756 -0.3743002    0.2201507    -0.2591216  0.12901321
3    -0.4897439 -0.7232053    0.1549383    -0.3601611 -0.49703950
  Trans...Brand.Runs   Vol.Tran Avg..Price Others.999 max.brand.ind
1        -0.25041300 -0.4624500  0.5145414  0.5308713    -0.5830875
2         0.01151372  0.4389494 -0.2536159 -0.2962366     0.3169929
3         0.98611550  0.5108315 -1.3048669 -1.2378282     1.3858726
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
1           -0.3194874         0.3653116             0.0584290  0.6285869
2            0.3528103        -0.3437646            -0.1415754 -0.5648218
3            0.1974478        -0.4128870             0.2056018 -0.7941739
    Pr.Cat.2   Pr.Cat.3    Pr.Cat.4  PropCat.5 PropCat.14
1 -0.2883594 -0.3934908  0.09784716 -0.1550911 -0.3928812
2  0.7564480 -0.2715444 -0.02224660  0.5590708 -0.2734898
3 -1.1958286  2.4581358 -0.32964358 -1.1206861  2.4617501
> km$size
[1] 298 229  73
```

(demographic information is not added here, but could be)

Comment:

Cluster 1: (n=298) Low brand loyalty, responsive to price category 1

Cluster 2: (n=229) Responsive to price category 2 and selling proposition 5, otherwise somewhat middling.

Cluster 3: (n=73) Highly loyal, low value, highly responsive to price category 3 and selling proposition 14.

output - 4 clusters

```
  No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
1     1.0183723  1.0937387    0.5223769     1.0473004  0.62417542
2    -0.5956884 -0.8027591    0.1018284    -0.4229423 -0.54648845
3    -0.3602143 -0.4774911    0.0319774    -0.3916182 -0.07913369
4    -0.3125952 -0.1493687   -0.6136206    -0.3755403 -0.29597650
  Trans...Brand.Runs   Vol.Tran  Avg..Price Others.999 max.brand.ind
1        -0.26749530 -0.2899399  0.03778485  0.2792907    -0.5041869
2         1.06003591  0.5313002 -1.33358415 -1.2706794     1.4348797
3         0.03362367  0.3906571 -0.32486388 -0.2041763     0.2729229
4        -0.22671331 -0.4364318  0.95096397  0.5243290    -0.4553992
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..     Pr.Cat.1
1          -0.03730309        0.08521843           -0.04816127  0.003979717
2           0.24351852       -0.43398564            0.15638262 -0.810295976
3           0.40274089       -0.32881894           -0.24376200 -0.609484479
4          -0.58440832        0.52248401            0.29514671  1.126863451
    Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5 PropCat.14
1  0.2300430 -0.2333055 -0.0535363 -0.1028166 -0.2368240
2 -1.2241619  2.5119045 -0.3351477 -1.1600056  2.5144713
3  0.6820653 -0.3058539  0.2120839  0.6366967 -0.3076327
4 -0.5852009 -0.4470861 -0.0748432 -0.2173710 -0.4423368
> km$size
[1] 163  69 207 161
```

We can add demographic information to these clusters:

```
             SEC      FEH       MT      SEX      AGE      EDU       HS    CHILD
meansc1 2.435583 2.343558 9.177914 1.901840 3.306748 4.558282 4.907975 2.987730
meansc2 3.420290 2.057971 7.710145 1.536232 3.028986 2.347826 3.913043 3.521739
meansc3 2.435583 2.343558 9.177914 1.901840 3.306748 4.558282 4.907975 2.987730
meansc4 1.919255 1.689441 7.074534 1.608696 3.211180 4.322981 3.248447 3.478261
              CS Affluence.Index
meansc1 1.0122699        20.34356
meansc2 0.8695652         8.42029
meansc3 1.0122699        20.34356
meansc4 0.8571429        19.00000
```

Comment:

Cluster 1 (n=163) is distinguished mostly by the purchase behavior variables - it has low brand loyalty together with high value, volume and frequency. The brand switching seems to be intrinsic - this group is not particularly responsive to promotions, pricing or selling propositions. Demographically it is relatively affluent and educated.

Cluster 2 (n=69) stands out in both groups of variables - it has high loyalty, low value and price per purchase, and very differential response to price (unresponsive to categories 1, 2 and 4, highly responsive to category 3), and selling proposition (unresponsive to #5, highly responsive to #14). Demographically it has low affluence and education.

Cluster 3 (n=207) is a "gray" cluster, it is not characterized by very extreme/distinctive values across all variables, but is responsive to price category 2 and selling proposition 5 (similar to cluster 2 in the 3-cluster analysis). Demographically it is relatively affluent and educated.

Cluster 4 (n=161) is characterized by low volume, low loyalty, and sensitivity to promotions and price (responsive to cat. 1, unresponsive to 2 and 3), and unmoved by selling proposition. Demographically, it is affluent, of high socio-economic status, and has relatively small family size.

Best cluster approach

There is no single "right" approach to clustering; different approaches are feasible depending on different marketing purposes. CRISA is a marketing agency and owns the data, which it collected at considerable expense, so it will want to be able to use both the data and the segmentation analysis in different ways for different clients. Here are just a few possible marketing approaches:

1. Establishing named customer "personas," corresponding to the cluster segments, for use by a client's sales and marketing teams.

2. Establishing named customer "personas," corresponding to the cluster segments, for use by CRISA in providing marketing services to clients.

note: The difference between #1 and #2 is that #1, being confined to a single client, can use that client's customer data to refine and do more analysis. #2 would have to rely on the data collected by CRISA.

3. "Capture affluent market share" campaign for a client who wants to target more affluent consumers who are not wedded to their current brand, and secure more brand share.

4. "Down market" campaign for a data-poor client to build a "value" brand for less affluent consumers, much as Dollar General has done in the U.S.

"Down market" scenario

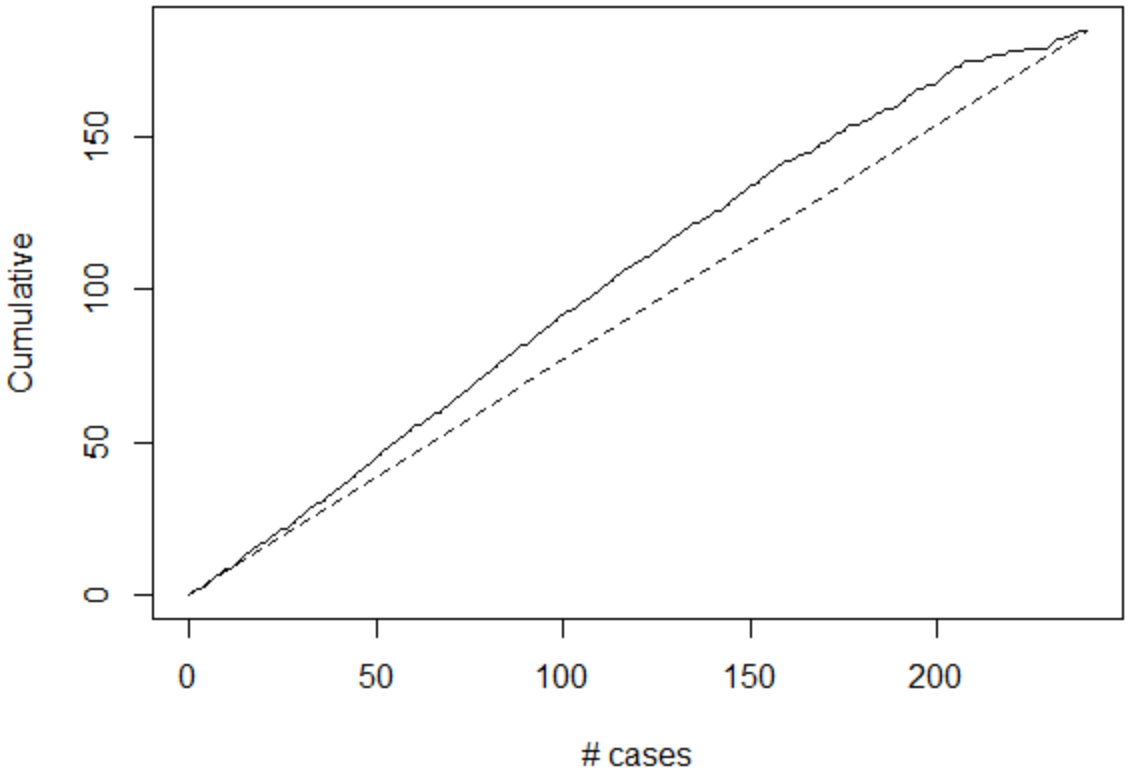
This fourth scenario is the one we will explore further to develop a predictive model, and classify people into either "value conscious" or not. "Data poor" means that the client has, or can get, demographic data on their customers, but not detailed purchase data (particularly involving other brands). So a predictive model is to be built using just demographic data. We will look at the results of clustering into two segments based on CRISA's own detailed purchase data, then classify people into those two segments.

Recall our characterization of the two segments:

Comment: The two clusters are separated on almost all variables, Value being an important exception. Cluster 1 (n=156) is the more loyal, with lower socioeconomic status and affluence, and larger households.

So our "success" category is cluster 1, the less affluent group, lower socioeconomic group, which also turns out to be highly loyal and, as it happens, spends roughly as much as the more affluent group. This is a promising group around which to build a down-market brand strategy.

Multiple models were tried and assessed; see code for details (you will need to re-run the confusion matrix and plotting section after each model.) Random forest performed best; its lift curve is shown below. Only demographic predictors are used - CRISA will not have the detailed purchase information for its client's customers.



You can see that the lift curve shows the model has only modest predictive power.

What's next?

Many data mining algorithms are iterative in an mathematical sense - iteration is used to find a good, if not best, solution. The modeling process itself is also iterative. In initial exploration, we do not seek the perfect model, merely something to get started. Results are assessed, and we typically continue with a modified approach.

Several steps can be explored next to improve predictive performance:

1. Some of the demographic categorical variables may not have much value being treated as is, as ordered categorical variables. They could be reviewed and turned into binary dummies.

2. Instead of using a two-cluster model, a multi-cluster model could be used in hopes of deriving more distinguishable clusters. The non-success clusters could then be consolidated. For example, cluster #2 in the 4-cluster model is similar to our cluster 1 ("success") in the 2-cluster model, only more sharply defined.

3. Demographic predictors could be added to the original clustering process.

4. The clustering process, which includes a randomization component that yields variability in resulting clusters, can be repeated, to ensure that the cluster labels reflect some degree of stability. Repetition should show some clustering results that are consistent across various runs. Choosing for your labels a clustering result that is very inconsistent with the others could mean that you are labeling your market segments according to a chance fluke.

5. In the real world, going beyond the parameters of this case study, CRISA would probably work with the client to add the client's own purchase data to the model to improve it over time.