Segmenting Consumers of Bath Soap

Sandeep More

Kent State University

05/07/2021

Problem Statement

Help CRISA in segmenting 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)

Methodology

There are four key problems that need attention here

1. Segmenting the market based on purchase behavior.
2. Segmenting the market based on basis of purchase.
3. Segmenting the market based on both purchase behavior & basis of purchase.
4. Segmenting the market based on demography and purchase habits.

The data supplied contains all the variable needed to approach the problem. But the broad nature of problem means that we need to look at multiple angles and solutions.

Here, we start with identifying clusters based on purchase behavior (brand loyalty) then we identify clusters based on basis of purchase and then we take both purchase behavior and basis of purchase to identify clusters that depend on these variables. Later on, we include demographic information to understand how brand loyalty and basis of purchase are segmented based on demographic information. Finally, we develop a model that classifies data into these segments.

In choosing k we need to consider that k value needs to be mapped to marketing actions. We follow couple of methods to guide us choose the optimal value of K (using Elbow charts and Gap Statistic Analysis) and choose k based on the value of K that results in actionable marketing decisions. The decision of K is also made based on cluster topology i.e., more distinct clusters with minimum overlap as possible.

Analysis

Following is the analysis of various problems we were tasked to address

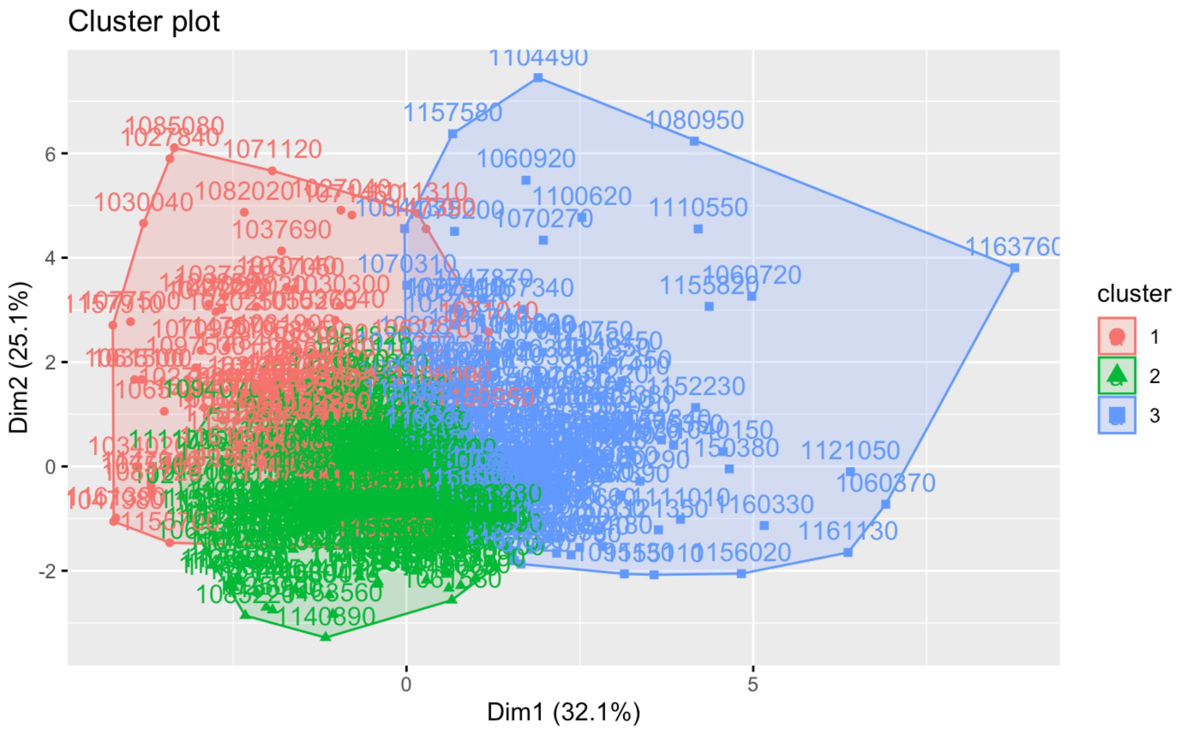
Segmentation based on Purchase Behavior – Brand Loyalty

To measure brand loyalty, we need to consider:

- Number of different brands purchased by a consumer, number of consecutive instances of brand purchase, number of transactions etc.
- How much percent of a brand (any brand since we are only measuring loyalty) does a consumer buy?

We need to create a derived variable that looks at all the brand wise purchases and gets the max volume value for *any* brand purchased. We call this variable as `*brand.vol.max*` - the presumption here would be that if a customer buys more of brand "A" then they are loyal to that brand (which might or might not be true). We also consider other brands which indicates consumer's likelihood to choose other brands.

Following is a diagram of cluster with k=3



Here we only analyze cluster with k=3. We think clusters with k=3 are segmented more meaningfully as shown above then k=5 for purposes of this study (marketing to consumers)

```
K-means clustering with 3 clusters of sizes 102, 315, 183

Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value Trans...Brand.Runs   Vol.Tran  Avg..Price Others.999 brand.vol.max
1    -0.6140370 -0.9228896  -0.02107873    -0.4937273 -0.2381902          1.0843822  0.4463582 -0.35797842 -1.3384059      4.808923
2    -0.2502972 -0.2350536  -0.48909843    -0.3795004 -0.4523625         -0.2222636 -0.2464976  0.09119428  0.3144380      1.074950
3     0.7730896  0.9189979   0.85363953     0.9284306  0.9114185         -0.2218248  0.1755094  0.04255520  0.2047511      1.218621
```

Here we can see that Cluster 1 has following properties

- Lowest number of brands purchased
- Lowest Brand runs
- Highest average transaction/brand
- Lowest another brand purchase

- Highest volume purchase per brand

Cluster 3 has following properties

- Highest number of brands purchased
- Highest brand runs
- Highest sum of volumes purchased
- Highest number of transactions
- Low average transaction/brand
- Lowest volume purchase per brand

Cluster 2 lies somewhere in between Cluster 1 and Cluster 3.

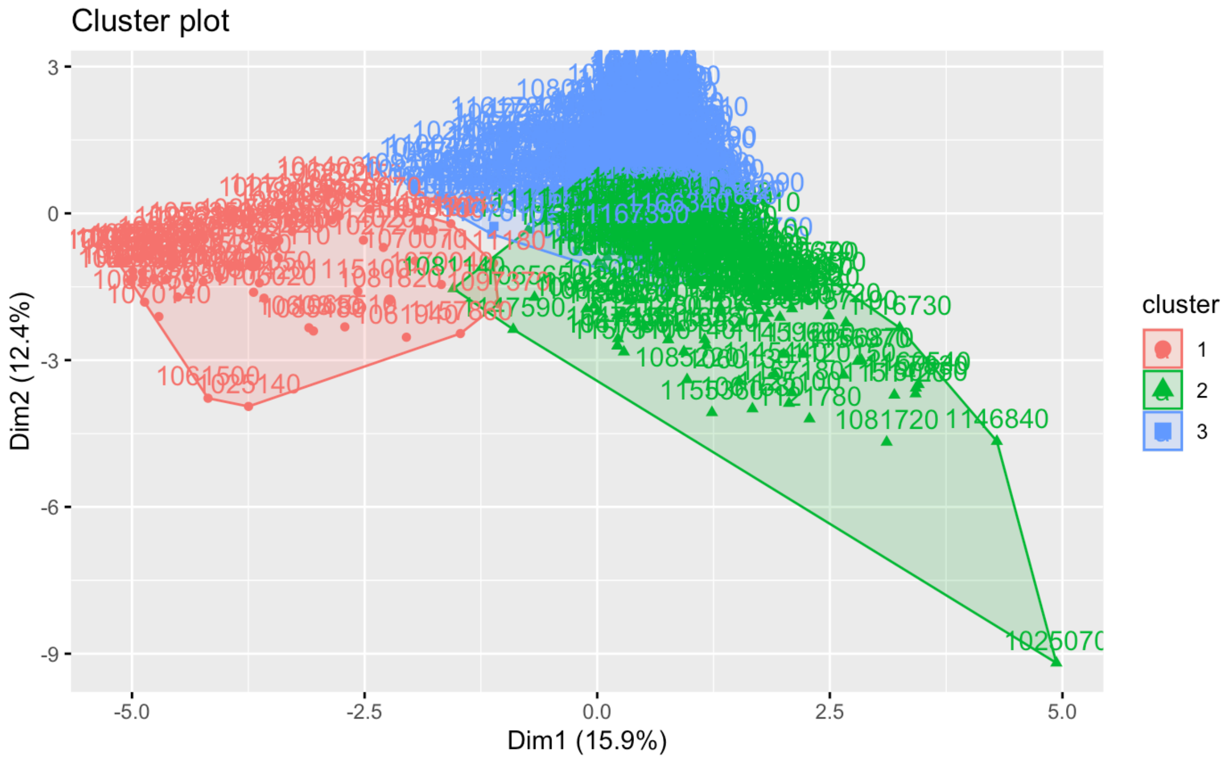
We can see that cluster 2 also has the highest number (n=315) than cluster 1 (n=102) and cluster 3 (n=183)

We can safely conclude that Cluster 1 has the most loyal customers and Cluster 3 has the least loyal customers

and Cluster 2 is where most of the customers are, who like to try out different brands and experiment with brands.

Basis of Purchase

For the basis of purchase, we will try to see if promotions affect the consumer purchase habits. Considering the elbow chart, Gap Stats Analysis and considering the marketing requirements we see that K=3 is perhaps a good number.



K-means clustering with 3 clusters of sizes 78, 193, 329

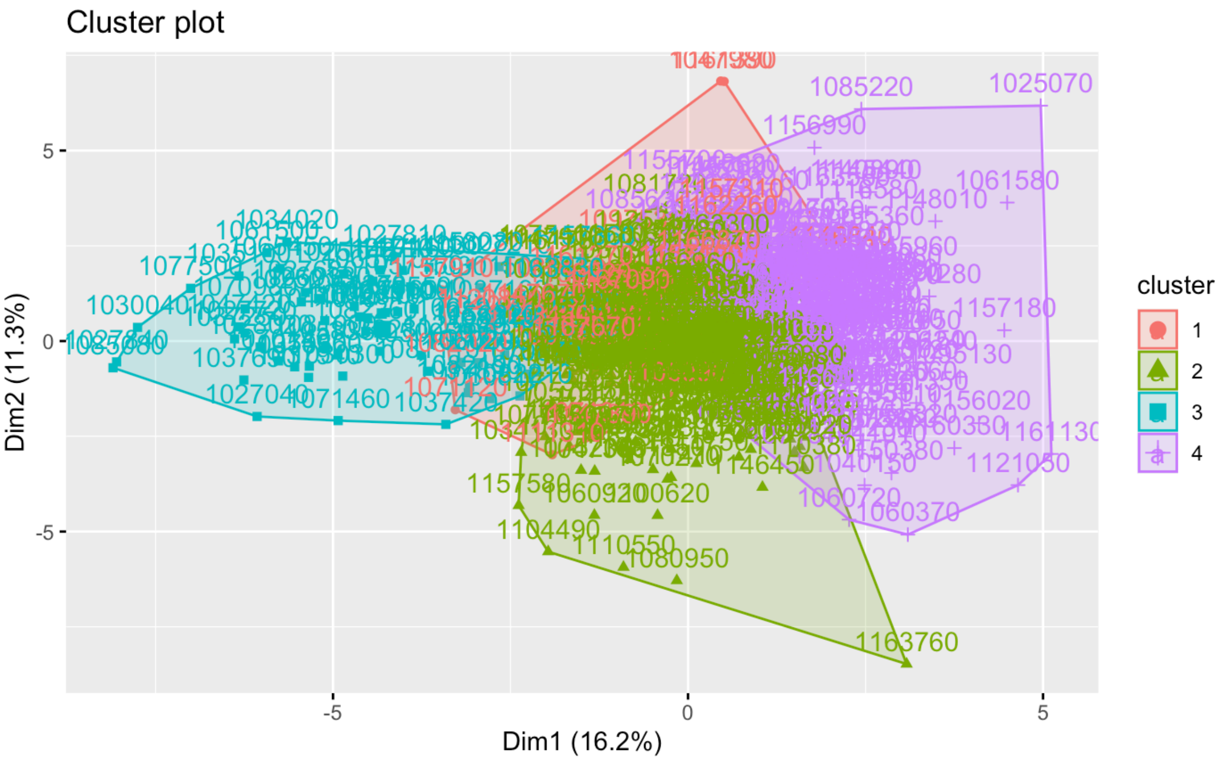
```
Cluster means:
  Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5   PropCat.6   PropCat.7  PropCat.8
1            0.1856666        -0.3842112             0.1912587 -0.7825205 -1.1334328  2.3701003 -0.3204763 -1.0914607 -0.17089192 -0.44919415 -0.4629703
2           -0.5626809         0.5576736             0.2131738  1.1091649 -0.4708722 -0.4653448 -0.2106562 -0.3516447  0.11719213  0.24917427  0.5131099
3            0.2860651        -0.2360563            -0.1703973 -0.4651435  0.5449425 -0.2889248  0.1995556  0.4650497 -0.02823256 -0.03967626 -0.1912417

    PropCat.9 PropCat.10  PropCat.11  PropCat.12 PropCat.13 PropCat.14  PropCat.15
1 -0.16226455 -0.2570818 -0.22953559 -0.16301187 -0.2325107  2.3724613 -0.22967026
2  0.13143273  0.3787795 -0.01931633  0.23567662  0.4408922 -0.4620933  0.04781956
3 -0.03863186 -0.1612525  0.06575023 -0.09960688 -0.2035148 -0.2913920  0.02639850
```

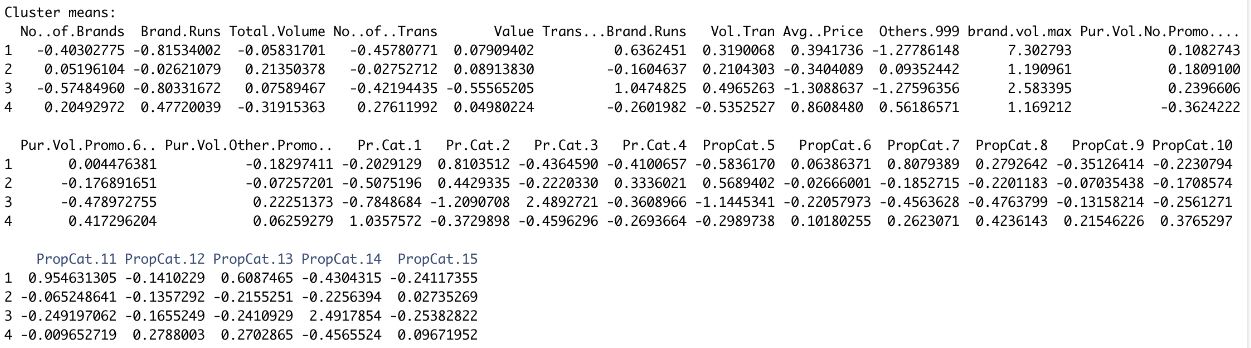
- Cluster 1 is responding nicely to price category 3 and proposition category 14. It does not respond well to price category 1, 5 and any proposition category other than 14 which is interesting. Cluster 1 also has least number of observations (n=78)
- Cluster 2 responds well to promotions and as expected does not respond well to no promotions so we can assume that the customers in this cluster are highly motivated by discounts and promotions. They also respond well to price category 1 and promotion categories 6-10,12,13,15 and do not respond well to other price category and purchase categories.
- Cluster 3 responds well to no promotions, looks like consumers in this cluster are not motivated by promotions. They respond well to price category 2, 4, 5 and promotion category 11. Interestingly this cluster has the largest number of observations (n=329)

Segmentation based on Purchase Behavior and Basis of Purchase

Here we consider all the above variables. Looking at the elbow and gap stats results and considering that the marketing efforts would support two to five different promotional approaches, k=4 seems to be a good value.



K-means clustering with 4 clusters of sizes 40, 288, 70, 202


```
Cluster means:
  No..of.Brands  Brand.Runs Total.Volume No..of..Trans       Value Trans...Brand.Runs   Vol.Tran Avg..Price  Others.999 brand.vol.max Pur.Vol.No.Promo....
1   -0.40302775 -0.81534002  -0.05831701   -0.45780771  0.07909402          0.6362451  0.3190068  0.3941736 -1.27786148      7.302793            0.1082743
2    0.05196104 -0.02621079   0.21350378   -0.02752712  0.08913830         -0.1604637  0.2104303 -0.3404089  0.09352442      1.190961            0.1809100
3   -0.57484960 -0.80331672   0.07589467   -0.42194435 -0.55565205          1.0474825  0.4965263 -1.3088637 -1.27596356      2.583395            0.2396606
4    0.20492972  0.47720039  -0.31915363    0.27611992  0.04980224         -0.2601982 -0.5352527  0.8608480  0.56186571      1.169212           -0.3624222

  Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5   PropCat.6  PropCat.7  PropCat.8   PropCat.9 PropCat.10
1       0.004476381           -0.18297411 -0.2029129  0.8103512 -0.4364590 -0.4100657 -0.5836170  0.06386371  0.8079389  0.2792642 -0.35126414 -0.2230794
2      -0.176891651           -0.07257201 -0.5075196  0.4429335 -0.2220330  0.3336021  0.5689402 -0.02666001 -0.1852715 -0.2201183 -0.07035438 -0.1708574
3      -0.478972755            0.22251373 -0.7848684 -1.2090708  2.4892721 -0.3608966 -1.1445341 -0.22057973 -0.4563628 -0.4763799 -0.13158214 -0.2561271
4       0.417296204            0.06259279  1.0357572 -0.3729898 -0.4596296 -0.2693664 -0.2989738  0.10180255  0.2623071  0.4236143  0.21546226  0.3765297

    PropCat.11 PropCat.12 PropCat.13 PropCat.14  PropCat.15
1  0.954631305 -0.1410229  0.6087465 -0.4304315 -0.24117355
2 -0.065248641 -0.1357292 -0.2155251 -0.2256394  0.02735269
3 -0.249197062 -0.1655249 -0.2410929  2.4917854 -0.25382822
4 -0.009652719  0.2788003  0.2702865 -0.4565524  0.09671952
```

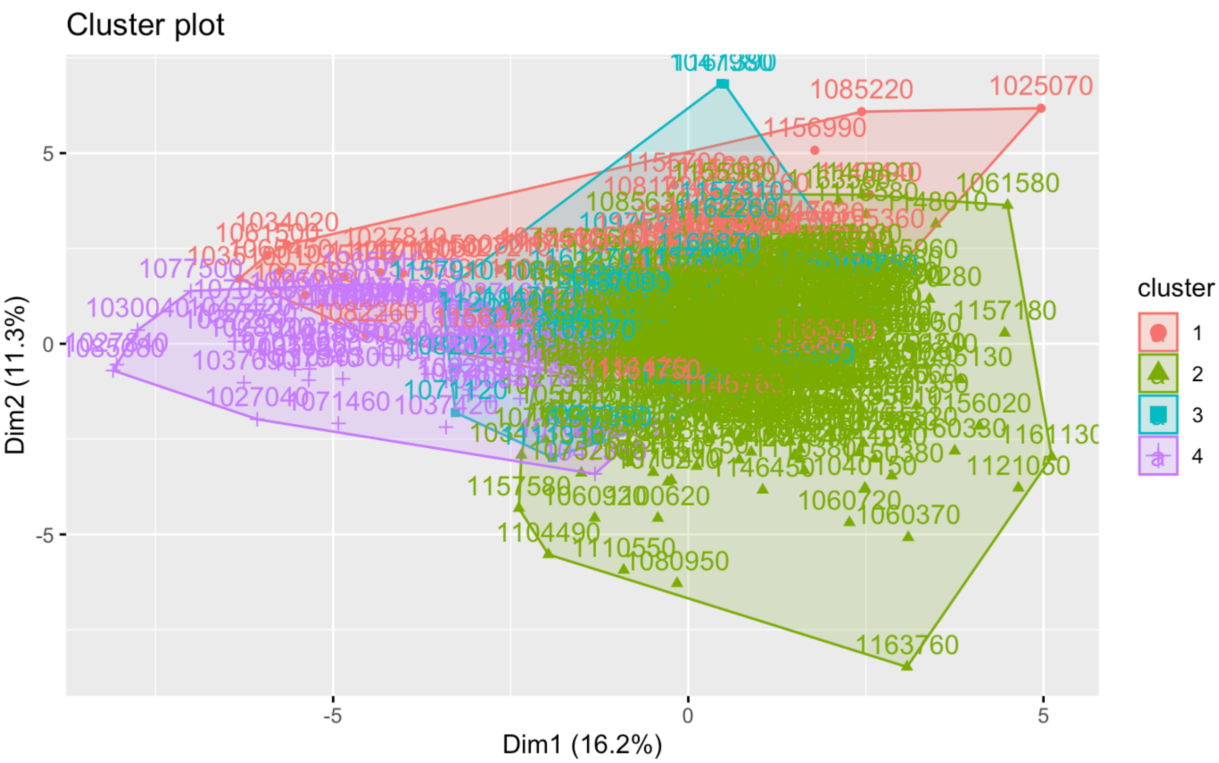
Here we can see that two clusters (Cluster 1 = 40, Cluster 3 = 70) have less candidates than other two (Cluster 2 = 288, Cluster 4 = 202).

- Cluster 1 consumers seems to be the most loyal customers who are not swayed by discounts and promotions. They do seem to care about promises that the products make (proposition). This cluster also has the least number of members (n=40). Looking at the data we see:
    - Brand volume purchased per brand is the highest (way higher than other clusters)
    - They have low numbers in brands purchased
    - Brand runs are low
    - Number of transactions of distinct brands are lowest
    - Average transaction/ brand run is highest
    - High Proposition cat 2, 7
- Cluster 2 can primarily be defined by the proposition category, they care a lot about product proposition, higher values in Proposition cat 2,4,5. They also seem to experiment with other brands (high no. of brands and high other brand purchases). It also has the highest number of members amongst all the clusters (n=288).
- Cluster 3 consumers can be classified as "frugal" given low "No. of brands", "No. of transactions" and "value". They have the most "Average transactions/brand run" and Volume per transaction. which indicates that they prefer to buy in bulk. They are inclined towards other promotions and discounts (high Purchase volume under other promotion). They also seemed to care about promises that the products make (proposition) indicated by high proposition category 14 and 2.

- Cluster 4 consumers are high spenders, they have high number of brands and the highest brand runs. They seem to buy less volume and more frequently paying the most (highest Average price per transaction). They also seem to experiment a lot (high other brand purchases). They are also inclined towards discounts and promotions (especially promo code 6), and appear to be most influenced by produce proposition 1,6,8,9,10,12,13 and 15

Best Segmentation

Here we include demographic info in cluster creation. Looking at the charts and considering that the marketing efforts would support two to five different promotional approaches, k=4 seems to be a good value.



K-means clustering with 4 clusters of sizes 66, 431, 44, 59

```
Cluster means:
          SEC         FEH          MT        SEX         AGE        EDU         HS      CHILD          CS Affluence.Index No..of.Brands Brand.Runs
Total.Volume No..of..Trans      Value Trans...Brand.Runs    Vol.Tran  Avg..Price Others.999 brand.vol.max Pur.Vol.No.Promo.... Pur.Vol.Promo.6..
Pur.Vol.Other.Promo..
1 -0.25727195 -1.80475562 -1.90431150 -2.6805048 -0.63162774 -1.8462679 -1.8223924  1.4515254 -1.83625981      -1.4916636    -0.7291330 -0.8592797
-1.04269343   -1.19207192 -1.0030708         -0.3390380 -0.09233208  0.08100038 -0.1174025      1.804981          -0.04196942       -0.17848551
 0.30204415
2 -0.04250689  0.19250649  0.23562435  0.3498234  0.07788562  0.3255918  0.1970988 -0.2279307  0.21699402       0.2912677     0.1888753  0.2892358
0.09722397    0.22804946  0.1798034         -0.1695007 -0.08946189  0.12093198  0.2993729      1.109575          -0.03578233        0.07641421
-0.03941985
3 -0.32497510  0.09758683  0.02197947  0.1932169  0.38373836  0.3434338  0.1340527  0.2004250  0.08988684       0.1834901    -0.2879318 -0.7150001
-0.06291343   -0.34526198  0.1058201          0.5722869  0.22458798  0.44214547 -1.2173356      7.002966           0.09611289        0.01766136
-0.18328995
4  0.84066652  0.53982630  0.39259944  0.2989470 -0.14857221 -0.5692790  0.4988137 -0.1081565  0.40192716      -0.5959346    -0.3493813 -0.6184432
0.50309197   -0.07493294 -0.2703163          1.1906898  0.58932409 -1.30376457 -1.1477692      2.485438           0.23666440       -0.37172170
0.08677629
     Pr.Cat.1   Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5   PropCat.6    PropCat.7   PropCat.8   PropCat.9  PropCat.10  PropCat.11   PropCat.12
PropCat.13 PropCat.14  PropCat.15
1  0.25018675 -0.3180835  0.2329356 -0.1761849 -0.1036176 -0.12097375 -0.077753582  0.12702773 -0.08816523  0.16504361 -0.18779002  0.254557229
0.08188569  0.2447797 -0.23086921
2  0.08450512  0.1370934 -0.3226239  0.1039235  0.2298877  0.03855987 -0.003162087  0.01112694  0.06577698  0.03090294 -0.03339647 -0.001078873
-0.04796716 -0.3246806  0.08682974
3 -0.15176518  0.7532476 -0.4423098 -0.3851764 -0.6015721  0.03898175  0.761486517  0.36708717 -0.31333455 -0.20915792  0.94217859 -0.158347532
0.64844527 -0.4362395 -0.22221010
4 -0.78400616 -1.2074004  2.4260808 -0.2748316 -1.1148077 -0.17542767 -0.457810355 -0.49714199 -0.14820766 -0.25439144 -0.24860743 -0.158787990
-0.22478307  2.4233292 -0.21032217
```

- Cluster 1 consumers seems to be ranked low in socioeconomic class have more children and are less affluent and are younger. They tend to have least number of brands; low brand runs and lowest total volume. They also seem to spend less and buy less on brands. They prefer other promotions (not promo code 6) prefer proposition category 1,10 and 12. It also has the second lowest number of members (n=66)
- Cluster 2 has more females and well educated and lowest number of children and are most affluent. They tend to prefer more brands have more brand runs and more brand transactions with least volume, which makes us conclude that they experiment a lot. They also don't seem to be motivated by promotions. Cluster 2 has highest membership (n=431)
- Cluster 3 consumers seems to be ranked lowest in socioeconomic class. One thing that stands out is that they have the highest rate of volume purchase for any brand. They also seem to be motivated by some brand propositions (2,7,6,11,13)

- Cluster 4 consumers are ranked the highest in socioeconomic class and are the most educated. They can be classified as loyal consumers; they have low number of brands and low brand runs and buy bigger volumes. They also have highest brand runs which makes us think they experiment frequently. They appear to buy at the lower price and do not seem to take advantage of promotions. They have the most average transactions/brand run and volume per transaction. which indicates that they prefer to buy in bulk. One interesting thing that stands out is that they are driven by proposition category 3 and 14

Model

Target Audience

CRISA has two categories of clients:

- Advertising agencies - which use the data to advise their clients on advertising and promotion strategies.
- Consumer goods manufacturers - which monitor their market share using the CRISA database.

Multiple models are needed to be looked at depending on client needs and their marketing strategies.

Following are the scenarios where models can be used to segment the consumers

Advertising agencies

Create a model based on customer demography and usage (like the one described in "Best Segmentation" approach) for the client's sales and marketing team to use. They can further enhance the model by adding their own consumer data.

Manufacturers

This model will be for the manufacturers of the soap. They might be interested in promotions and brand propositions. So, they can use the "Basis of Purchase" and "Brand Loyalty" models described earlier to tailor their promotions and brand message.

Model

Here we develop a model that would be used in targeting direct-mail promotions. We define success criteria as “Engage maximum number of consumers with the brand”. We use this success criteria to develop a model that classifies consumers based on following categories (sub goals)

- More likely to engage
- Less likely to engage
- Can be swayed

We are looking for clusters with demographic data along with purchase data. For segments like “More likely to engage” we need to consider the education, purchase habits and responsiveness to promotions.

Looking at the previous models we analyzed the one described in “Best Segmentation” fits the bill.

Recalling our analysis for Cluster 2

> Cluster 2 has more females and well educated and lowest number of children and are most affluent. They tend to prefer more brands have more brand runs and more brand transactions with least volume, which makes us conclude that they experiment a lot. They also don't seem to be motivated by promotions. Cluster 2 has highest membership (n=431)

This sounds like a good match to our success criteria “Engage maximum number of consumers with the brand” and matches our goal “More likely to engage”. We see consumers here are well educated and prefer more brand runs which means they are willing to try and experiment. They do not seem to be motivated by promotion which indicates that they likely make decisions based on their own research and are not influenced by promotions and discounts. These customers could be engaged more by brand proposition.

Cluster 3 seems to be a good fit for our sub goal “Can be swayed”

> Cluster 3 consumers seems to be ranked lowest in socioeconomic class. One thing that stands out is that they have the highest rate of volume purchase for any brand. They also seem to be motivated by some few brand propositions (2,7,6,11,13)

We can also generalize this model to focus on affluent customers or customers with higher socioeconomic scores.

Conclusion

We hope that the analysis discussed in this report and the model presented above helps CRISA in segmenting the market based on their key criteria. Here we presented a model that fits the success criteria “Engage maximum number of consumers with the brand”. We can use the analysis presented above and the clusters described in the report to define other success criteria such as

- Target most affluent customers
- Target budget customers
- Target customers that care about brand messaging

Using these success criteria, they can better serve their client needs. These models can further be enriched by adding more purchase and demographic data by clients of CRISA to get more benefits out of the model.