Case Study

Myles Thomas

12/16/2020

Case Study - Segmenting Consumers of Bath Soap

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods. They would like to segment the soap buyers' market based on two sets of variables more directly related to the purchase process and to brand loyalty:

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

The reason for this segmentation is that once this market is segmented successfully with a new method, the agency can further segment those clusters using the traditional segmentation of markets on the basis of purchaser demographics.

Things to note before diving in:

K (K = The number of cluster) will be chosen by running a k-means algorithm in each part and seeing which number of k has the best spread of clusters.

Brand Loyalty is an important aspect of this Case Study, and there are a number of ways to quantify this idea. The transaction/brand runs ratio is a good start for seeing if a customer is loyal or not (A brand run is a string of consecutive times that a customer purchases the same brand). Brand-wise Volume % is more self explanatory: if a customer is brand loyal, the customer will purchase a make a large proportion of purchases to one brand. In order to account for the 9 Brandwise purchase columns, code will be written that checks to see if any of the first 8 columns have a % higher than 50%, since that would indicate the customer is making a majority of transactions for that brand.

The final derived predictor 'BrandLoyalYN' will give a customer a 1 if the customer has a Trans/Brand runs ratio of above 2.0 AND a 50% or higher in one of the first 8 brand columns.

1

Using Purchase behavior to Segment the Market

Starting off, here are quick descriptions of the relevant variables involved for Method 1, including whether they are categorical or quantitative:

Quantitative vars:

-Total.Volume, the total volume of products purchased (in grams)

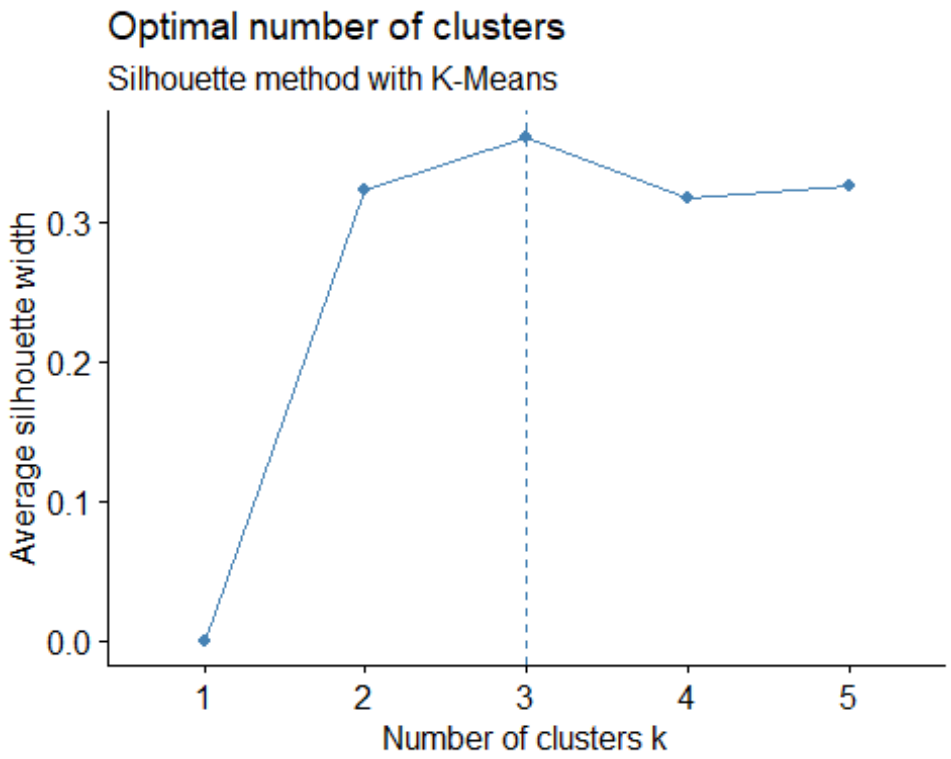
-No. of Trans, gives context for frequency of purchases

-Pur.Vol.No.Promo.-.%, Percent of volume purchased not on promotion

-Pur.Vol.Promo.6.%, Percent of volume purchased on promo code 6

Categorical vars:

-BrandLoyalYN, the derived variable to decide if an observation can be deemed brand loyal, or not (binary)



Cluster sizes based on how many k chosen, Method 1

	C1	C2	C3	C4	C5
K = 2 clusters	120	480	-	-	-
K = 3 clusters	320	80	200	-	-
K = 4 clusters	242	72	116	170	-
K = 5 clusters	68	186	106	44	196

Cluster size k=3 at first glance seems like the best choice for K since each cluster size is above 50 and no cluster takes up too much of the data. By using R to decide the optimal

number of k, the function NbClust agrees that 3 clusters is the best number of k for Method 1. The resulting silhouette plot agrees by suggesting k=3.

Using Basis for Purchase to Segment the Market

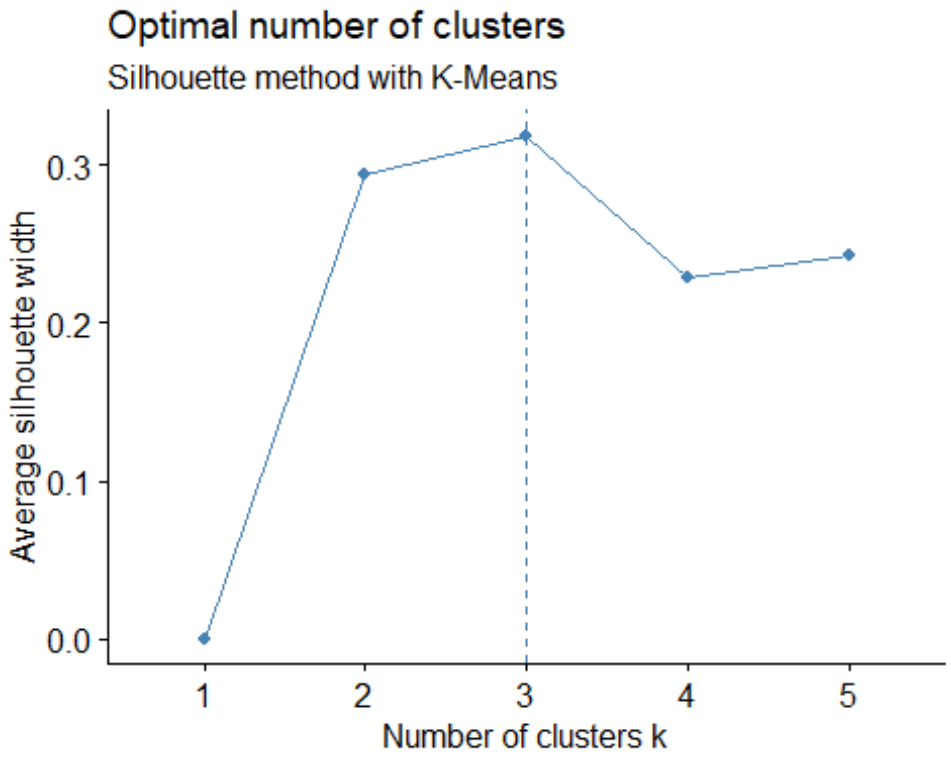
For this next method, all relevant variables involved are quantitative percentages, as they represent the % of Volume Purchased under the given categories. The 4 "Price Categories" are dependent on one another and add up to 100%, and in the same way the 11 "Proposition Categories" add up to 100%.

Price categories:

1. Premium soaps
2. Popular soaps
3. Economy/Carbolic soaps
4. Sub-popular soaps

Proposition categories:

5. Beauty
6. Health
7. Herbal
8. Freshness
9. Hair
10. Skin Care
11. Fairness
12. Baby
13. Glycerine
14. Garbolic
15. Others



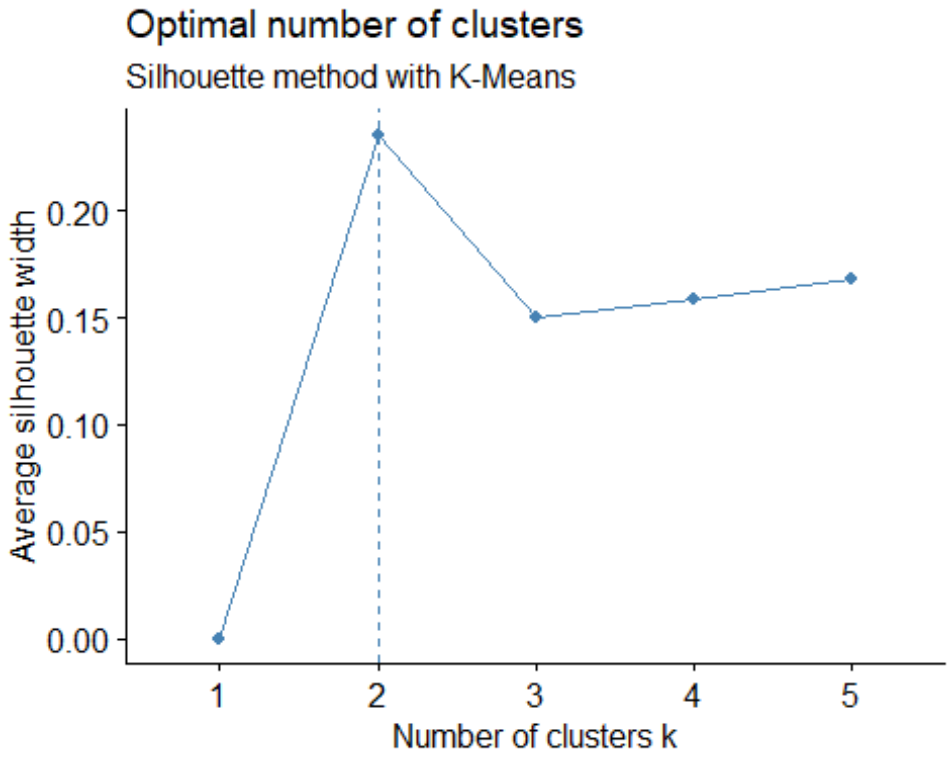
Cluster sizes based on how many k chosen, Method 2

	C1	C2	C3	C4	C5
K = 2 clusters	78	522	-	-	-
K = 3 clusters	78	76	446	-	-
K = 4 clusters	109	116	297	78	-
K = 5 clusters	114	299	53	74	60

Cluster size k=3 has once again selected by the NbClust algorithm. It should be noted that Cluster 3 is abnormally large (n=446). Once again, the silhouette plot agrees and suggests k=3, so the result k=3 from NbClust will be final.

Combining the prior 2 methods to Segment the Market

Running k-means clustering on all of the variables from both a) and b), which have already been described.



Cluster sizes based on how many k chosen, Method 3

	C1	C2	C3	C4	C5
K = 2 clusters	74	526	-	-	-
K = 3 clusters	326	73	201	-	-
K = 4 clusters	107	95	326	72	-
K = 5 clusters	163	63	61	73	240

Cluster size k=3 has once again selected by the NbClust algorithm. The resulting silhouette plot again disagrees and suggests k=2, but cluster size selection is subjective due to so many methods of decision. The NbClust method will remain superior here, so once again k=3.

2 - Selecting the Best Segmentation.

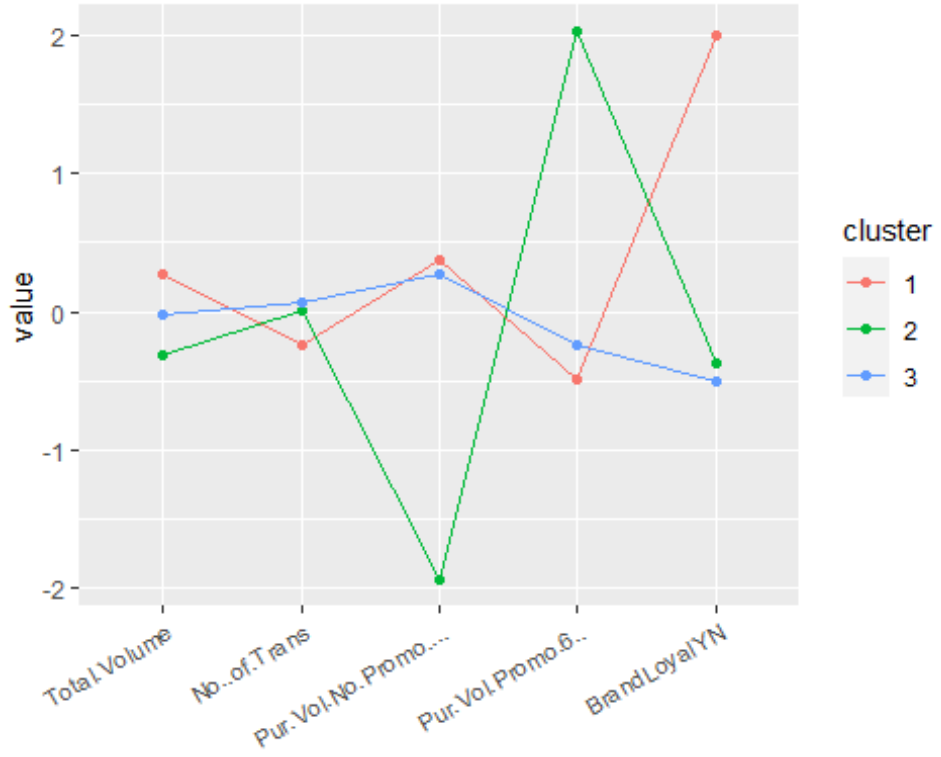
First, observe the cluster sizes for all 3 approaches.

Cluster size for all 3 approaches

	C1	C2	C3
Method 1 -	320	80	200
Method 2 -	78	76	446
Method 3 -	326	73	201

Now, to aggregate the means of each variable to look at the discrepancies/differences.

Method 1:



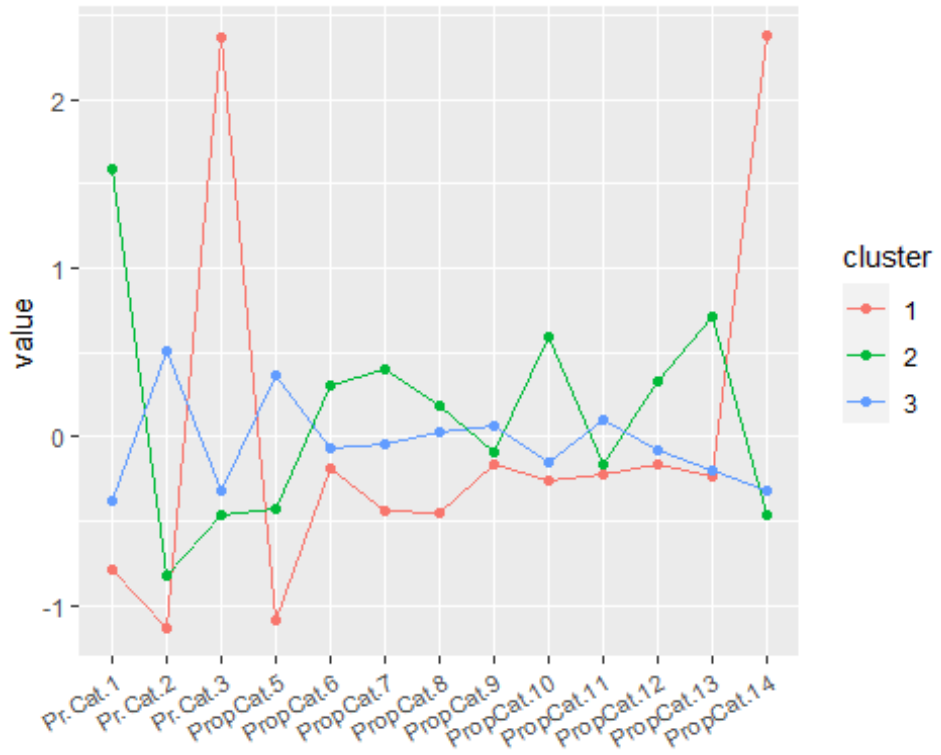
Cluster names to summarize:

1 - Brand Loyal

2 - Frugal Buyer

3 - Frequent Shopper, Not Loyal

Method 2:



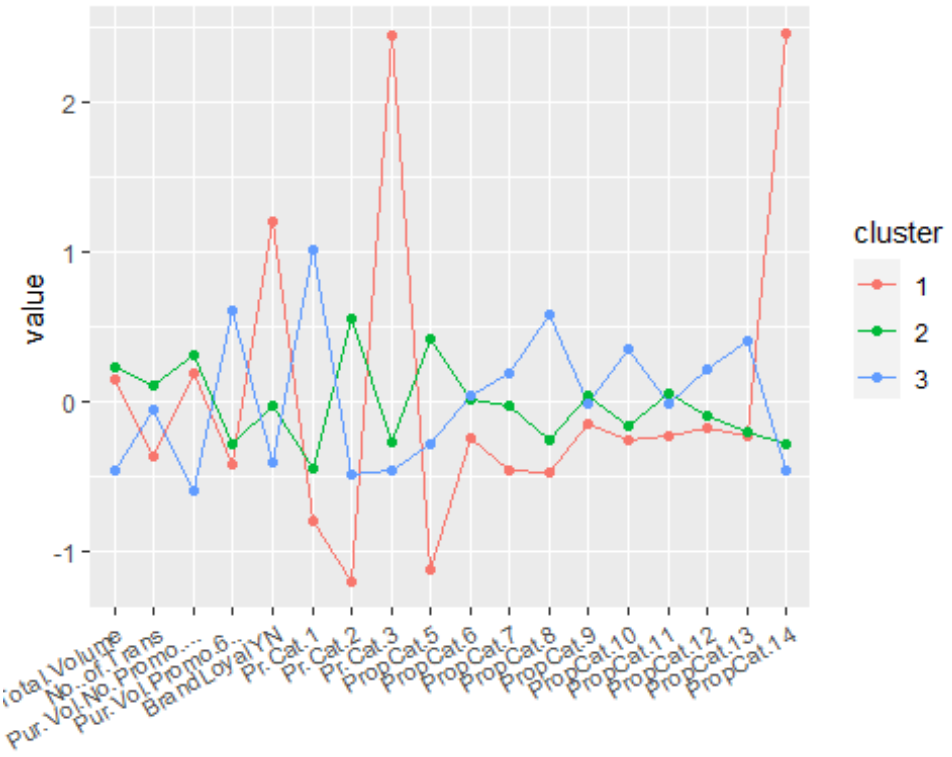
Cluster names to summarize:

1 - Garbolic

2 - High end

3 - Middle Class

Method 3:



Cluster names to summarize:

1 - Loyal to Garbolic Soap

2 - Joe Shmoe (Average person) buys the first soap in sight

3 - Outward beauty, uses promo code 6

Looking at the means, here is what stood out for METHOD 1 (Purchase Behavior):

- Cluster 1 has middling results, except for a very high proportion (100%) of 'Yes' for Brand Loyalty. This cluster represents a customer who is Brand Loyal.
- Cluster 2 is the most variant, with a very low proportion for Purchase Volume w/ no promo code and a very high proportion (24%) of Purchase Volume w/ promo code 6. This cluster represents a customer who is frugal, and very succeptable to using promo codes.
- Cluster 3 is very average, as it has middling proportions for all 5 predictors. Due to the highest value for # of transactions and lowest proportion for brand loyalty, this cluster represents a customer who shops often but does not care about promos or brand loyalty.

What stood out for METHOD 2 (Basis of Purchase):

- Cluster 1 has very high proportions (77% each) for Economy/Carbolic price category and Garbolic product proposition category. This cluster represents a customer who is into Garbolic and Economy priced soaps.
- Cluster 2 has average proportions besides very high proportions in Premium soaps price category (72%) and Glycerine product proposition category (%). This cluster is for the expensive soap buyer, since Glycerin soap is more expensive than most other types of soap.
- Cluster 3 has middling proportions throughout except for a relatively high proportion of Popular soaps. This cluster is for the middle-upper middle class soap buyer.

What stood out for METHOD 3 (Combination Method):

By combining parts A and B, the importance for the brand loyalty variable is diminished (One predictor that seems to be very important). Instead of having a huge weight like it did in Method 1, it seems that now Cluster 1 has relatively high brand loyalty and Clusters 2/3 with below average brand loyalty. It appears that brand loyalty is a very useful part of the segmentation, but it is possible that with only a few predictors the weight was too much. Anyhow,

- Cluster 1 has abnormally high proportions for Brand Loyalty (68%), Economy Soaps (80%) and Garbolic Soaps (79%. This cluster represents the buyer who is loyal to middle-priced garbolic soap brands.
- Cluster 2 has no abnormalities, as each proportion is average throughout. This cluster represents the average soap buyer.
- Cluster 3 has relatively high proportions for the Beauty/Freshness/Skin care propositions, as well as the highest proportion for use of promo code 6 (11%). This cluster represents those who worry a lot about their skin and outward appearance.

Decision:

The Basis for Purchase Method alone is the worst. The Purchase Behavior clusters are good, especially with how brand loyalty is an important factor, but the fact that all 3 clusters have similar values for use of promo codes removes an important part of the analysis. The combination method not only takes into account brand loyalty, creates good sized clusters and has distinct features for all 3 of the clusters, but being the only method that properly takes into account the use of promo codes makes Method 3 the recommended method going forward.

3 - Build a Classification Model

Since the information observed this far is in an effort to find a group to be targeted by direct-mail promotions, developing a model that defines observations as "1" success if it is classified in the correct cluster (The cluster most susceptible to using promotions) and a "0" if not should do the trick.

Using The Combination Method's clusters, Cluster 3 becomes the "success" group while Clusters 1 and 2 come together and are the "0". This is because cluster 3 is most susceptible to using promo codes.

Logistic Regression:

```
##  Accuracy
## 0.9791667
```

Classification tree:

```
##  Accuracy
## 0.9041667
```

k-NN:

```
##  Accuracy
## 0.9666667
```

Interpretations and conclusions of the results:

The Logistic Regression Model and K-NN models both had great accuracy, as they predicted 5 and 8 observations incorrect, respectively.

LR:

```
##           Reference
## Prediction   0   1
##          0 158   1
##          1   4  77
```

k-NN:

```
##           Reference
## Prediction   0   1
##          0 160   6
##          1   2  72
```

Due to the ease of interpretations and 97.9% Accuracy of the Logistic Regression Model, that is the model that will be recommended.

Overall, when using brand loyalty and purchase behavior to group people into clusters, it is possible to find those more likely to using promo codes whom can be targeted in a marketing campaign. Despite the fact that the Basis of Purchase variables were unable to segment the market into relevant clusters, and that the Purchase Behavior clusters lacked interpretation of the use of promo codes, the goal was still accomplished.

By combining the data and methods, this allowed for segmenting into relevant clusters of a demographic of choice. This led to a model being fit that accurately predicted almost 100% of observations in the validation data set. Using this information going forward, the market research agency should be able to further segment the market using traditional demographics and improve performance/success with future marketing campaigns.