Assignment 3 – Market Segmentation (Segmenting Consumers of Bath Soap)

Due: Nov 1st 2020

Business Situation

CRISA is a leading market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and non-durable). In one major project, CRISA tracks about 30 product categories (e.g. detergents, etc.) and within each category, about 60 – 70 brands. To track purchase behavior, CRISA has constituted about 50,000 household panels in 105 cities and towns in India, covering about 80% of the Indian urban market. (In addition to this, there are 25,000 sample households selected in rural areas; however, we are working with only urban market data). The households are carefully selected using stratified sampling. The strata are defined on the basis of socio-economic status, and the market (a collection of cities).

CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and, for the household data, maintains the following information:

- Demographics of the households (updated annually)
- Possession of durable goods (car, washing machine, etc.; updated annually) and a computed "affluence index" on this basis
- Purchase data of product categories and brands (updated monthly).

CRISA has two categories of clients: (1) Advertising agencies who subscribe to the database services; they obtain updated data every month and use it to advise their clients on advertising and promotion strategies. (2) Consumer goods manufacturers who monitor their market share using the CRISA database.

Key Problems

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 more effectively deploy promotion budgets.

The better and more effective market segmentation would enable CRISA's clients 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 a year. This would result in a more cost-effective allocation of the promotion budget to different market-segments. It would also enable CRISA to design more effective customer reward systems and thereby increase brand loyalty.

Measuring Brand Loyalty

Several variables in this case measure aspects of brand loyalty. The number of different brands purchased by the customer is one measure. However, a consumer who purchases one or two brands in quick succession, and then settles on a third for a long streak is different from a consumer who constantly switches back and forth among three brands. So, how often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands – a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands.

All three of these components can be measured with the data in the purchase summary worksheet.

Note: How should the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is?

Clustering approach

We will consider clustering based, first, on variables that describe purchase behavior, and then, based on variables that describe basis-for-purchase. A third clustering will then consider both sets of variables.

A key question is the number of clusters to consider – this can be based on how the clusters will be used. It is likely that the marketing efforts would support 3-7 different promotional approaches. For clusters based on purchase behavior variables alone, or on basis-for-purchase variables alone, the fewer variables may support only 2-4 clusters. Clustering on the combined variables may allow for higher number of useful clusters.

Remember – clusters are useful only so far as they carry a useful interpretation. And remember the business goal. Given the business goal, *it is useful to consider demographic variables in addition to the variables used in clustering, for effective interpretation*.

Questions

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

2. Explore the data.
   Are there any missing values – how do you handle these?
   Summarize the households in the data based on demographic variables – use plots, tables to help your description.

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?

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

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

   a. The variables that describe purchase behavior (including brand loyalty).

      [*Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)].*
      [Q – how do you measure brand loyalty?]

   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.]

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

<u>For each clustering in Q3 and in Q4 below:</u>
(i) Describe your rationale for experimenting with different values of k.
(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 a 2-person team, try one other method*) for the questions above - from agglomerative clustering, k-medoids, kernel-k-means, and DBSCAN clustering.

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

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.

   (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.

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

Data

Data file – Assgt3_BathSoap.xls
The data in the Table 1 below profiles each household – each row contains the data for one household.

Though not used in the assignment, two additional datasets were used in the derivation of the summary data.

CRISAPurchaseData is a transaction database in which each row is a transaction. Multiple rows in this dataset corresponding to a single household were consolidated into a single row of household data in CRISASummaryData.

The Durables sheet in the data file contains information used to calculate the affluence index. Each row corresponds to a household, and each column represents a durable consumer good. A 1 in a column indicates that the durable is possessed by the household; a 0 indicates that it is not possessed. This value is multiplied by the weight assigned to the durable item. The sum of all the weighted values of the durables possessed gives the affluence index.

Table 1

Member Identification	**Member id**		Unique identifier for each household
Demographics	**SEC**	1 – 5 categories	Socio Economic Class (1=high, 5=low)
	FEH	1 – 3 categories	Food eating habits (1=vegetarian, 2=veg. but eat eggs, 3=non veg., 0=not specified)
	MT		Native language (see table in worksheet)
	SEX	1: male 2: Female	Sex of homemaker
	AGE		Age of homemaker
	EDU	1 – 9 categories	Education of homemaker (1=minimum, 9 = maximum)
Demographics	**HS**	1 - 9	Number of members in the household
	CHILD	1 – 4 categories	Presence of children in the household
	CS	1 - 2	Television available. 1: Available 2: Not Available
	Affluence Index		Weighted value of durables possessed

Summarized Purchase Data

Purchase summary of the house hold over the period	**No. of Brands**		Number of brands purchased
	Brand Runs		Number of instances of consecutive purchase of brands
	Total Volume		Sum of volume
	No. of Trans		Number of purchase transactions; Multiple brands purchased in a month are counted as separate transactions
	Value		Sum of value
	Trans / Brand Runs		Avg. transactions per brand run
	Vol/Tran		Avg. volume per transaction
	Avg. Price		Avg. price of purchase

Purchase within Promotion	**Pur Vol No Promo - %**		Percent of volume purchased under no-promotion
	Pur Vol Promo 6 %		Percent of volume purchased under Promotion Code 6
	Pur Vol Other Promo %		Percent of volume purchased under other promotions

Brand wise purchase	**Br. Cd. (57, 144), 55, 272, 286, 24, 481, 352, 5 and 999 (others)**		Percent of volume purchased of the brand
Price category wise purchase	**Price Cat 1 to 4**		Per cent of volume purchased under the price category

Selling proposition wise purchase	**Proposition Cat 5 to 15**		Percent of volume purchased under the product proposition category