Assignment 3 – Market Segmentation
Segmenting Consumers of Bath Soap
Due: Nov 6th, 2016

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 like now 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.

Data

Data file – 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

Questions

1. Use k-means clustering to identify clusters of households based on
    a. The variables that describe purchase behavior (including brand loyalty). How do you evaluate brand loyalty?

        [*Variables: #brands, brand runs, total volume, #transactions, value, avg. price, share to other brands, (brand loyalty)]*.

    b. The variables that describe basis-for-purchase.

        [*Variables: pur_vol_no_promo, pur_vol_promo_6, pur_vol_other, all price categories, selling propositions]*

        [Note – would you use all selling_propositions? Explore the data.]

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

Note: How should k be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support 2-5 different promotional approaches.

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?

d. Try k-medoids, kernel k-means, agglomerative clustering, and DBSCAN clustering. You do not need to try all techniques on all combinations in (a)-(c) above; you may pick one set of variables in (a) thru (c) that you feel may be best suited for addressing the segmentation need.
How do different parameter values for the different techniques affect the clusters obtained? Are the clusters obtained from the different procedures similar? What might be some reasons for differences in clusters obtained using different procedures? Which would you pick as your 'best' and why?

2. (a) 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 useful0.

(b) Comment on the characteristics (demographic, brand loyalty and basis-for-purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

3. For the best segmentation, obtain a description of the clusters. You may base this on attributes describing the clusters (not restricted to attributes used for clustering). You should also build a decision tree to help describe the clusters – how effective is the tree in identifying the different clusters? Does the tree help in explaining/interpreting the different clusters? (explain why/why not).
(You may use a decision tree to help choose the 'best' clustering).