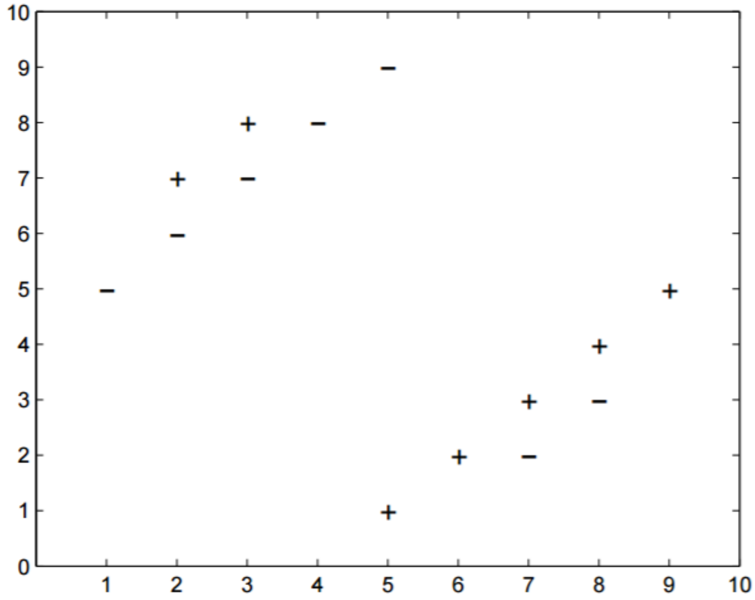
BUSINESS DATA MINING (IDS 572)

HOMEWORK 5
DUE DATE OF QUESTIONS 1 AND 2: WEDNESDAY NOVEMBER 29 AT 3:00 PM

- You can submit the answer of Question 3 with your HW6 submission. Due Date of Question 3: Wednesday December 06 at 3:00 PM
- Please provide succinct answers to the questions below.
- Please include all the R functions you use.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.

Problem 1. In the following questions you will consider a k-nearest neighbor classifier using Euclidean distance metric on a binary classification task. We assign the class of the test point to be the class of the majority of the k nearest neighbors. Note that a point can be its own neighbor.



(a) What value of k minimizes the training set error for this data set? What is the resulting training error? Explain.

(b) Why might using too large values k be bad in this dataset? Why might too small values of k also be bad?

(c) What value of k minimizes leave-one-out cross-validation error for this dataset? What is the resulting error?

Problem 2. Take the following points in two dimensional space:

$$(8,4),(3,3),(4,5),(0,1),(10,2),(3,7),(0,9),(8,1),(4,3),(9,4).$$

For this exercise, use the Manhattan distance metric: for instance, the distance from (3,3) to (8,1) is

$$|3-8|+|3-1|=7.$$

(a) Beginning with centroids at (1,1) and (8,8), do two iterations of the 2-means clustering algorithm, that is:

- allocate the points to centroids, then find the new centroids.

- again allocate the points to the centroids, and then get the new centroids.

If a point is equidistant between the centroids, assign it to the centroid that starts at (1,1). What are the resulting centroids and resulting clusters?

(b) Suppose we are interested in a binary (yes, no) output. Suppose outputs for the points above are yes, yes, no, no, yes, yes, no, no, yes, yes respectively. Consider the point (5,3).

i. What are the three closest points in our data set?

ii. Using the K-nearest neighbors approach, what would be the predicted output for (5,3) using K = 3 neighbors? (Use equal weights for each of three closest neighbors.)

Problem 3. Use R to answer this question.

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 CRISAs 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 marketsegments. 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 is BathSoap.xls. The data in the Table 1 below profiles each household each row contains the data for one household.

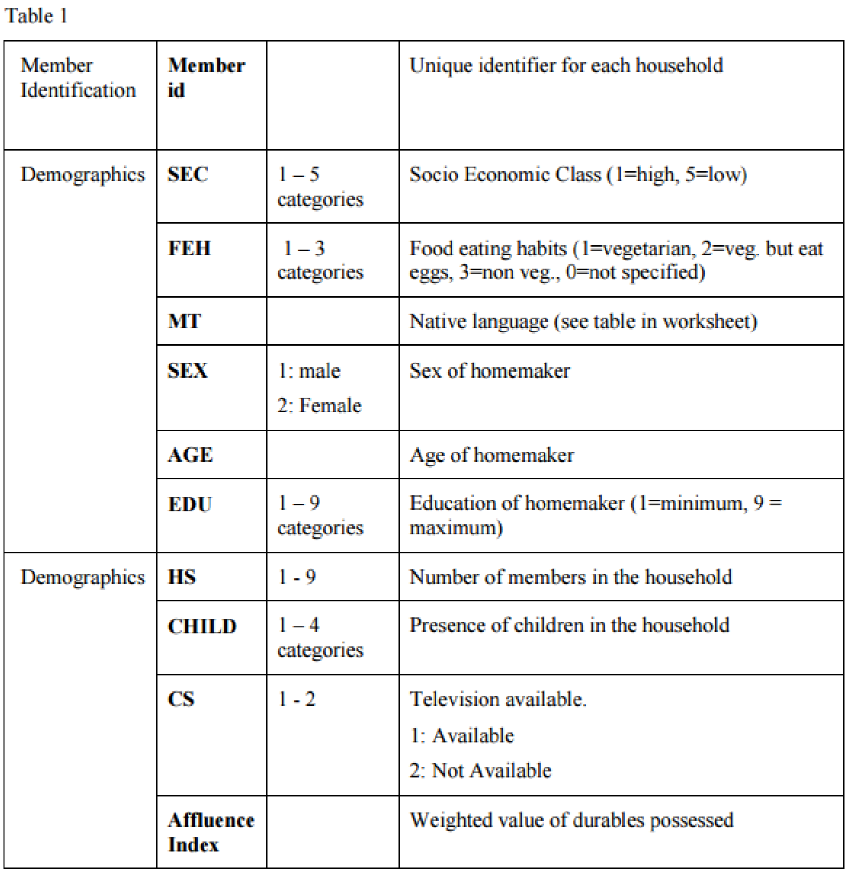


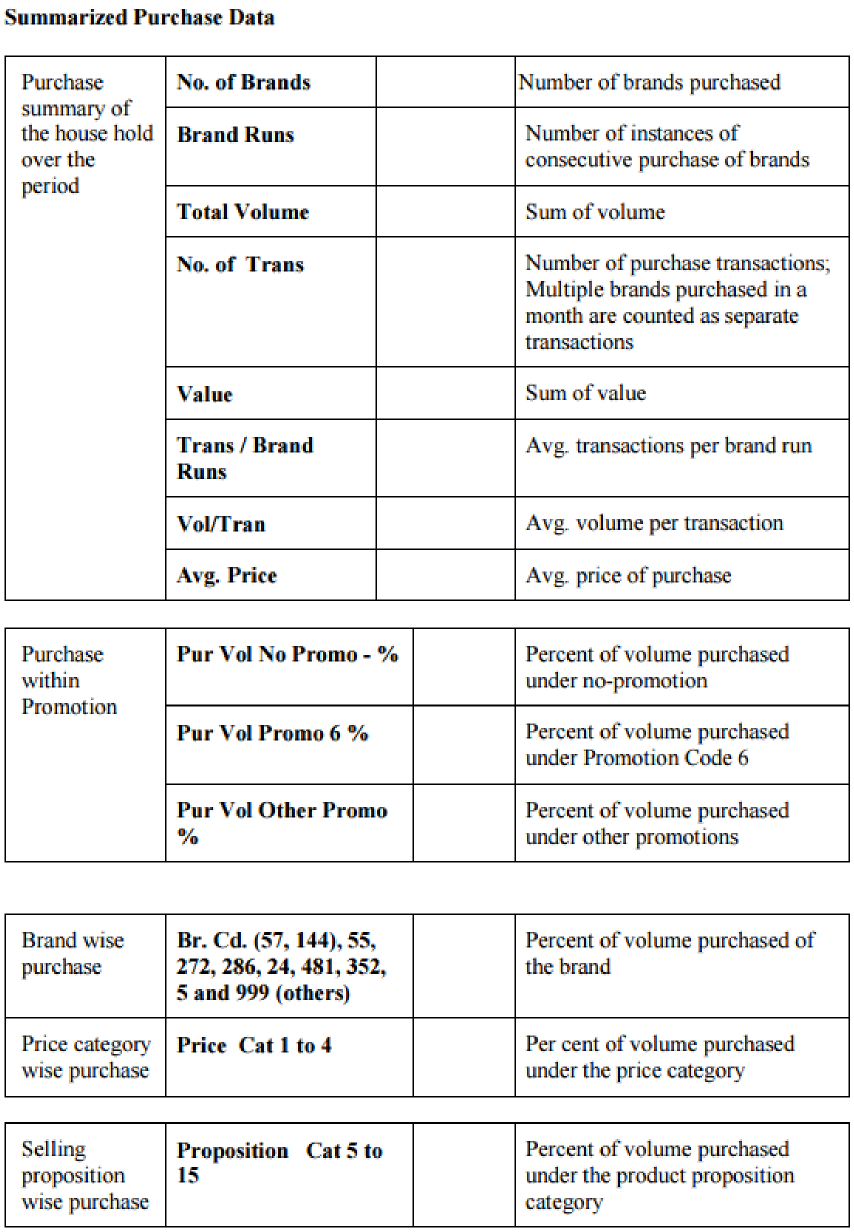
Table 1

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

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.



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).

   [Variables: # brands, brand runs, total volume, # transactions, value, Avg. price, share to other brands, max to one brand].

   (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?

2. (a) Select what you think is the best segmentation - explain why you think this is the "best".
(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.)