ADS-505 Applied Data Science for Business

Module 5 Assignment 5.1: Use Case - Segmenting Consumers of Bath Soap (Using Python) (90 points)

BathSoapHousehold.csv is the dataset for this case study (linked in the assignment prompt in Blackboard).

For this assignment, you will use a Jupyter notebook to answer the questions in the "Assignment" section of this document. Be sure to include the number and question that you are answering in your notebook.

BUSINESS SITUATION

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and nondurable). In one major research project, CRISA tracks numerous consumer product categories (e.g., "detergents"), and, within each category, perhaps dozens of brands. To track purchase behavior, CRISA constituted household panels in over 100 cities and towns in India, covering most of the Indian urban market. The households were carefully selected using stratified sampling to ensure a representative sample; a subset of 600 records is analyzed here. The strata were defined on the basis of socioeconomic 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 it maintains the following information:

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

CRISA has two categories of clients: (1) advertising agencies that subscribe to the database services, obtain updated data every month, and use the data to advise their clients on advertising and promotion strategies; (2) consumer goods manufacturers, which 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 thus deploy promotion budgets more effectively. More effective market segmentation would enable CRISA's clients (in this case, a firm called IMRB) 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 the year. This would result in a more cost-effective allocation of the promotion budget to different market segments. It would also enable IMRB to design more effective customer reward systems and thereby increase brand loyalty.

DATA

The data in Table 21.8 profile each household, each row containing the data for one household.

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 of loyalty. However, a consumer who purchases one or two brands in quick succession, then settles on a third for a long streak, is different from a consumer who constantly switches back and forth among three brands. Therefore, 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.

ASSIGNMENT

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

   1.1. The variables that describe purchase behavior (including brand loyalty) **(10 points)**

1.2. The variables that describe the basis for purchase **(10 points)**

1.3. The variables that describe both purchase behavior and basis of purchase **(14 points)**

Note 1: How should *k* be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support two to five different promotional approaches.

Note 2: How should the percentages of total purchases 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? Consider using a single derived variable.

Table 21.8 Description of variables for each household

Variable type	**Variable name**	**Description**
Member ID	Member id	Unique identifier for each household
Demographics	SEC	Socioeconomic class (1 = high, 5 = low)
	FEH	Eating habits(1 = vegetarian, 2 = vegetarian but eat eggs, 3 = nonvegetarian, 0 = not specified)
	MT	Native language (see table in worksheet)
	SEX	Gender of homemaker (1 = male, 2 = female)
	AGE	Age of homemaker
	EDU	Education of homemaker (1 = minimum, 9 = maximum)
	HS	Number of members in household
	CHILD	Presence of children in household (4 categories)
	CS	Television availability (1 = available, 2 = unavailable)
	Affluence Index	Weighted value of durables possessed

Purchase summary 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	Average transactions per brand run
	Vol/Trans	Average volume per transaction
	Avg. Price	Average price of purchase
Purchase within promotion	Pur Vol	Percent of volume purchased
	No Promo - %	Percent of volume purchased under no promotion
	\rlapPur Vol Promo 6%	Percent of volume purchased under promotion code 6
	Pur Vol Other Promo %	Percent of volume purchased under other promotions
Brandwise purchase	Br. Cd. \rlap(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	Percent 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

(*Source*: © Cytel, Inc. and Datastats, LLC 2019)

2. Select what you think is the best segmentation and 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.) **(14 points)**
3. Develop a model that classifies the data into these segments. Since this information would most likely be used in targeting direct-mail promotions, it would be useful to select a market segment that would be defined as a success in the classification model. **(14 points)**
4. Briefly explain, in two to three paragraphs, the business objective, the data mining models used, why they were used, the model results, and your recommendations to your non-technical stakeholder team. **(18 points)**