Assignment 1

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.

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)

Variable type	**Variable name**	**Description**
	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
	Pur 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. (57, 144), 55, 272, 286, 24, 481,	Percent of volume purchased of the brand

Variable type	Variable name	Description
	352, 5, and 999 (others)	
Price categorywise purchase	Price Cat 1 to 4	Percent of volume purchased under the price category
Selling propositionwise purchase	Proposition Cat 5 to 15	Percent of volume purchased under the product proposition category

Refer to page 518, 519, 520 for further background about the case.

Use “BathSoap.csv” to solve the below questions in Python and submit only the Python script file for your group thru BB. Give me your answers with # as comments in the code.

1. Use k-means clustering to run 2 and 3 clusters of households based on:
    a. The variables that describe purchase behavior (including brand loyalty).
    b. The variables that describe the basis for purchase.
    c. The variables that describe both purchase behavior and basis of purchase

Note 1: *For this analysis, we use all* ***purchaseIndicator****,* ***maxBrandIndicator*** *and* ***otherBrandIndicator*** *as a description of the customers purchase behavior*

- *Brand loyalty: bathSoap_df['maxBrandIndicator'] = bathSoap_df[brandIndicator].max(axis=1)*

Note 2: *for "basis for purchase", the variables used are: Pur Vol No Promo - %, Pur Vol Promo 6 %, Pur Vol Other Promo %, all price categories, selling propositions 5 and 14.*

2. Select what you think is the best clustering 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.)
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. (use logistic regression)

Good Luck!