BUA 466  
Segmenting Consumers of Bath Soap

short line

David Griffin   
3rd May 2023

**The Case**

CRISA is an Asian market research agency that tracks consumer purchase behavior in consumer goods. In this research project, CRISA tracks numerous consumer product categories and dozens of brands within each category. In order to track purchase behavior, CRISA sampled most of the Indian urban market. This resulted in a subset of 600 records to be analyzed. There is both transaction and household data. For the household data, the information includes:

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

The idea that CRISA is not assembling the following data for a specific brand, influences how *k* should be chosen. Therefore, it is likely that we would choose clusters that are distinct and separate from each other. By creating these distinct clusters CRISA can help its client identify and explore unique characteristics and behaviors of each segment and translate that information to create marketing actions specific for needs and preferences of that segment. This further leads to better targeting, cost-effective use of resources as well as higher customer satisfaction and higher sales and profit.

**The Problem**

CRISA has traditionally segmented markets on the basis of purchase demographics. They would now like to segment the market based on two key sets of variables more directly related to the purchase process and brand loyalty:

* Purchase Behavior
  + Volume
  + Frequency
  + Susceptibility to discounts
  + Brand Loyalty
* Basis of Purchase
  + Selling Proposition
  + Price

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Variable Type** | **Variable Name** | **Description** |
| Member ID | Member id | Unique identifier for each household class (1 =high, 5= low) |
| 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 |
|  | 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 (4 categories) |
|  | 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, 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 |

**Narrative**

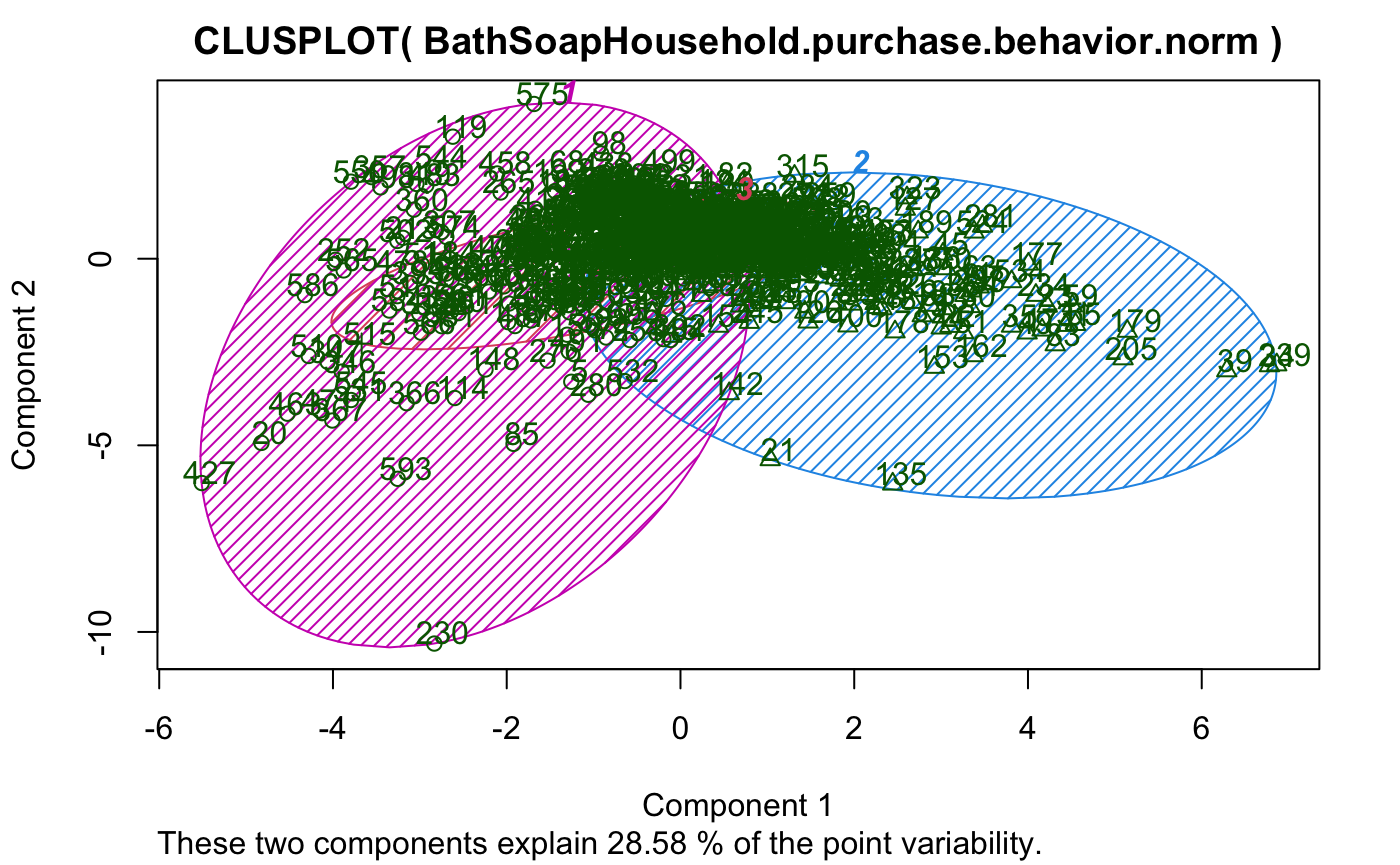
To start, we prepared the data by importing the data set “BathSoapHousehold.csv'' into an RMD file in RStudio. The data set consisted of 600 observations (rows) and 46 variables, which are described in the data dictionary included above. Afterwards, we needed to pre-process the data to have a better dataset usable for our analysis. We started by removing the first observation “Member.id'' because it is a unique identifier that does not describe any variable. Next, we made the data set into a data frame and omitted any missing values. After omitting any missing values, the next step we took was normalizing the data to put all of the observations on the same scale, ensuring that the amount of outliers accounted for by the model are kept to a minimum.

Next, we computed the normalized data on a euclidean scale to find the distance between data points using every observation and saving the result as a string. To get an idea of what our data looked like, we then created a dendrogram for a simple visual. However, the visual output did not provide much useful information because of the amount of data in the data set. We then created another dendrogram based off of the “average” method to see if that would make a difference, however it was not much more useful than the first dendrogram, so we decided to omit the output from this report for the sake of being concise. Due to this, we decided that utilizing dendrograms in a visual capacity would not be very efficient in our analysis, and decided to focus on using the dendrograms to make clusters. We cut out six clusters utilizing the dendrograms and plotted the six clusters in a 2-D space to have a better representation of how the data points were dispersed. After evaluating the six cluster plots we noticed the six clusters were not a good representation of the data due to it not maximizing between-cluster sum of squares. We decided to then cut out three clusters to find a better representation, however it still did not minimize the within-cluster sum of squares. We created a heatmap for an extra visualization of the data. There was not much information that could be extracted from the heat map. Our group switched to the k-means method to assign the amount of clusters to each graph. We assigned six clusters for the first k-means cluster plot. This plot was much better in minimizing within-cluster sum of squares but was not as good in maximizing between-cluster sum of squares. There was a lot of overlap so we decided to cut it to three clusters. Again, we got the same results with different amounts of clusters. To understand the difference between the hierarchical dendrogram and the k-means, we made a table including both 3-cluster and 6-cluster plots. We were still unsure why the clustering for k-means was overlapping, so to ensure we had the correct amount of clusters we used the elbow method. This confirmed that three clusters was the correct amount.

After conducting analysis of all of the observations we used the same method of understanding the data through hierarchical clustering and k-means clustering. Along with this, we tried to ensure that the amount of clusters and how they were applied were correct with respect to the segmentation of purchase behavior and basis of purchase for CRISA. The variables we conducted analysis on to describe purchase behavior included “Total.Volume”, “Vol.Tran”, “Pur.Vol.No.Promo....”, “Pur.Vol.Promo.6..”, “Pur.Vol.Other.Promo..”, “Brand.Runs”, “Trans...Brand.Runs”, “No..of.Brands”, “Br..Cd..57..144”, “Br..Cd..55”, “Br..Cd..272”, “Br..Cd..286”, “Br..Cd..24”, “Br..Cd..481”, “Br..Cd..352”, “Br..Cd..5”, and “Others.999”. The variables we conducted analysis on to describe basis of purchase included “Avg..Price”, “Pr.Cat.1”, “Pr.Cat.2”, “Pr.Cat.3”, “Pr.Cat.4”, “PropCat.5”, “PropCat.6”, “PropCat.7”, “PropCat.8”, “PropCat.9”, “PropCat.10”, “PropCat.11”, “PropCat.12”, “PropCat.13”, “PropCat.14”, and “PropCat.15”.

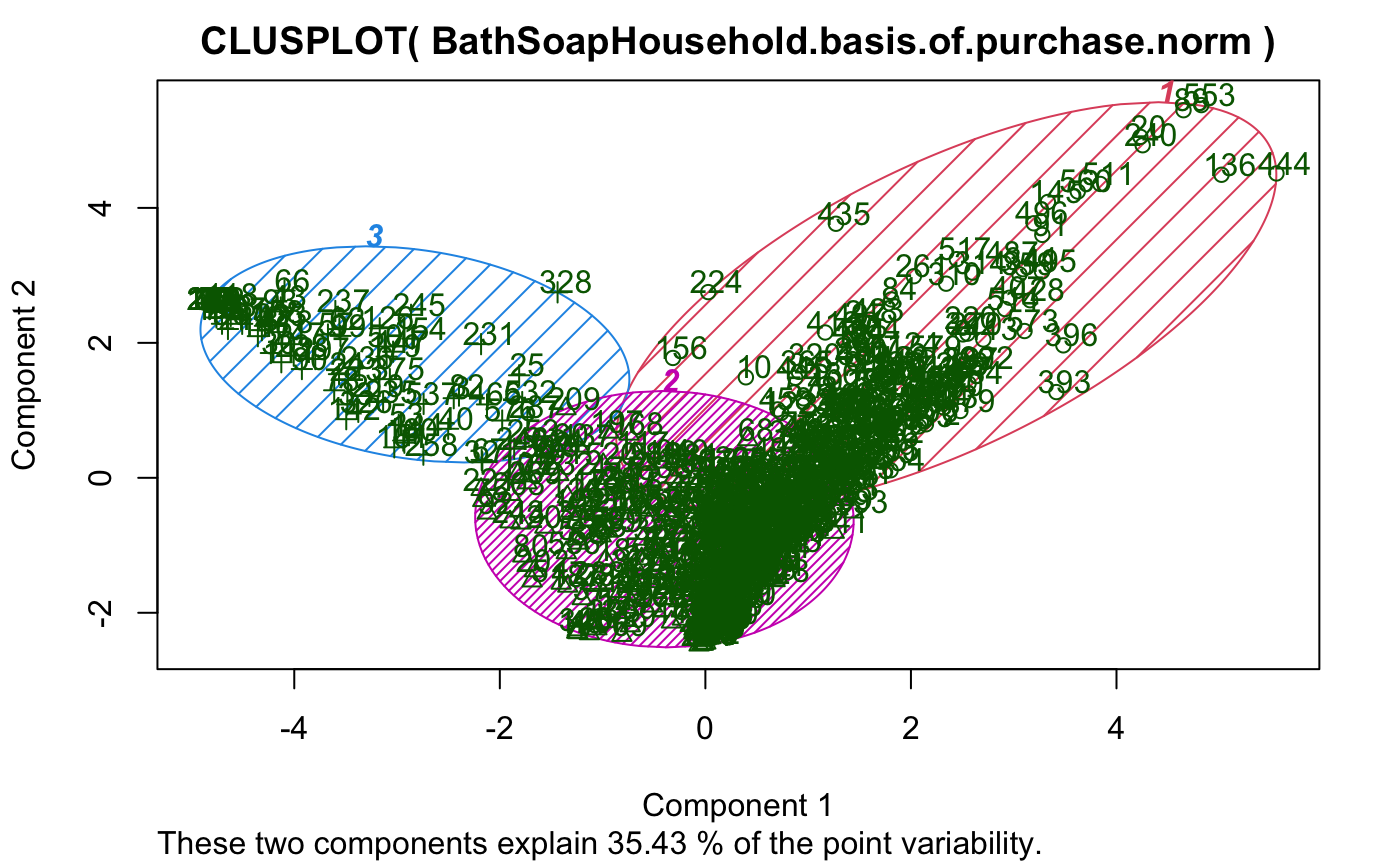
**Results**

Following all the steps mentioned above, we decided to interpret the cluster plots generated with the K-Means method. Since we already generated the code, we decided we would interpret the cluster plots that contained 3 clusters.

****

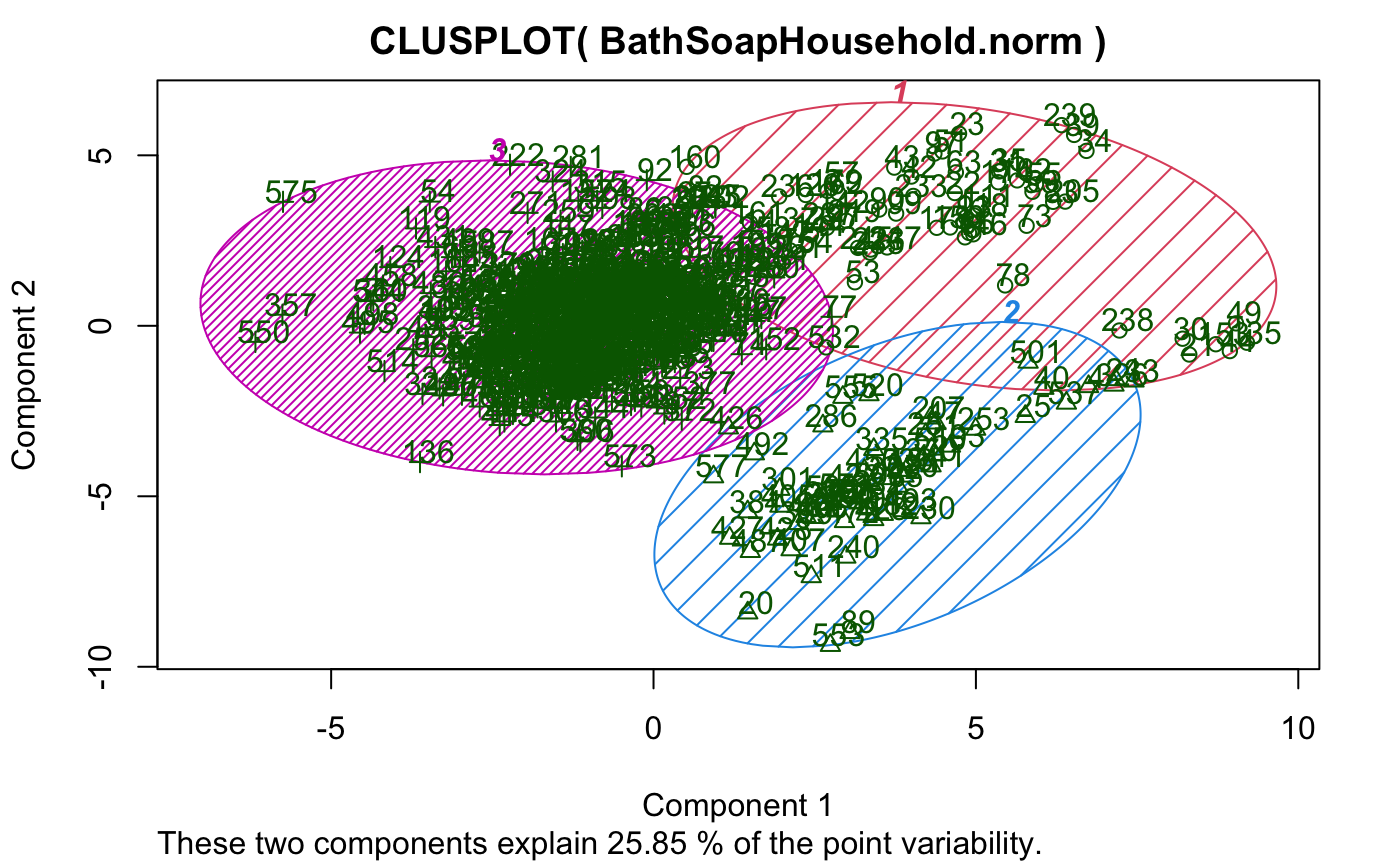
**[1.1]** K-Means method generated cluster plot with 3 clusters for “Purchase Behavior”

The data has been clustered into three groups, with the sizes of the groups being 345, 236, and 19 respectively. Looking at the cluster means, we can see that the first cluster has a negative mean for Total Volume, Vol.Tran, Pur.Vol.No.Promo, and Brand Runs, while the second cluster has a positive mean for these variables. The third cluster has a positive mean for Total Volume and Vol.Tran, but a negative mean for Pur.Vol.No.Promo and Brand Runs. The variable "Trans...Brand.Runs" has a negative mean for the first and third clusters, but a positive mean for the second cluster. The variable "No..of.Brands" has a positive mean for the first and third clusters, but a negative mean for the second cluster. The variables related to brand codes (“Br..Cd..57..144”, “Br..Cd..55, Br..Cd..272”, “Br..Cd..286”, “Br..Cd..24”, “Br..Cd..481”, “Br..Cd..352”, “Br..Cd..5”) have different mean values for each cluster, indicating that each group may prefer different brands. Overall, the clustering suggests that there are three distinct groups of households with different consumption patterns and brand preferences, meaning low brand loyalty.



**[1.2]** K-Means method generated cluster plot with 3 clusters for “Basis of Purchase”

The K-means clustering algorithm was applied to the dataset and resulted in 3 clusters with sizes 83, 23, and 494 respectively. The first cluster had the smallest size and showed a significantly higher average value for Pr.Cat.3, while having negative values for most of the other features. The second cluster had the smallest variance and had the highest average value for PropCat.15. It also had relatively low average values for most other features. The largest cluster, with size 494, showed the highest average values for Pr.Cat.1, Pr.Cat.2, and PropCat.5. It also had positive values for all features, except for Pr.Cat.3 and PropCat.14, which were negative.



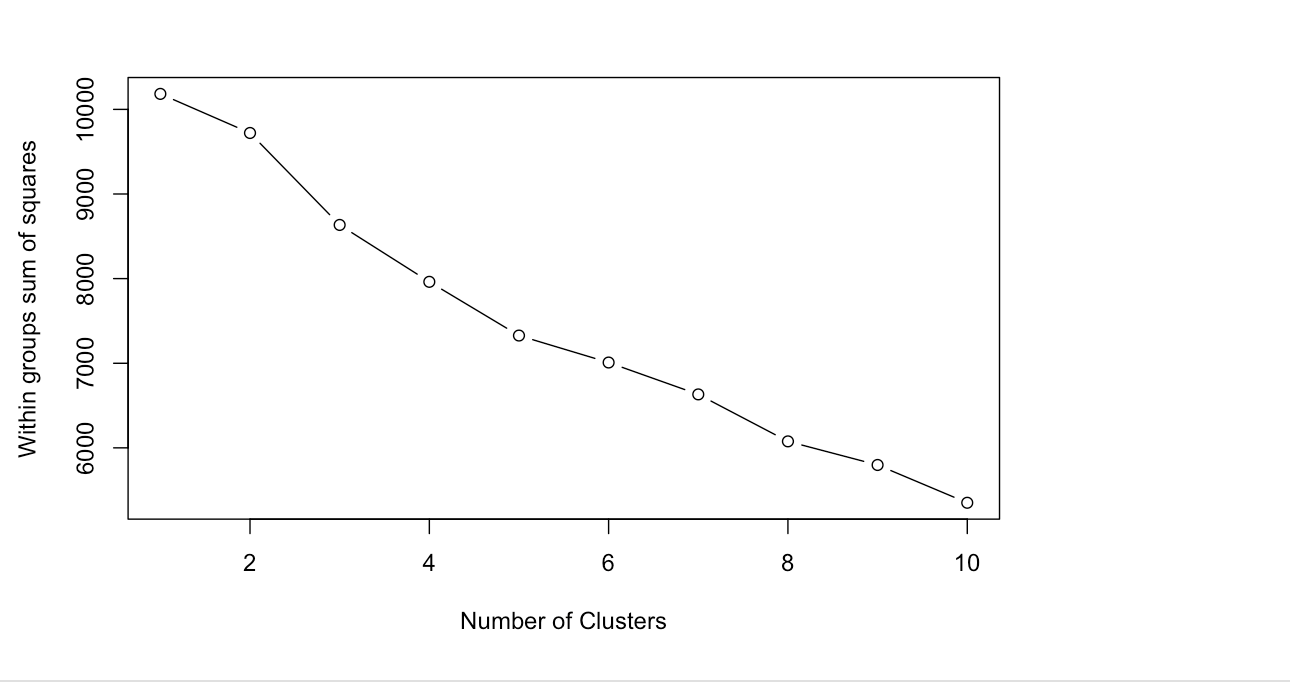
**[1.3]** K-Means method generated cluster plot with 3 clusters with all variables in data

The clusters have sizes 66, 61, and 473, respectively. Cluster 1 has a high affluence index, indicating that the customers within this cluster are relatively wealthy. They tend to purchase a moderate number of brands and have a relatively high number of transactions. However, they do not purchase products with promotions very frequently. They tend to purchase expensive products, which suggests that they have a relatively high income. Cluster 2, on the other hand, has a low affluence index and tends to purchase fewer brands and have fewer transactions. They are more likely to purchase products with promotions and tend to buy products with a lower average price. Cluster 3 is characterized by an average affluence index and an average number of brands and transactions. They tend to purchase products with promotions and buy products with a relatively low average price. In addition to these purchasing behaviors, the data also provides information about the brands that customers within each cluster tend to purchase. For example, customers in cluster 1 tend to purchase brand codes 55 and 144, while customers in cluster 2 tend to purchase brand code 286. Furthermore, the data provides information on the proportion of purchases that customers make in each product category.

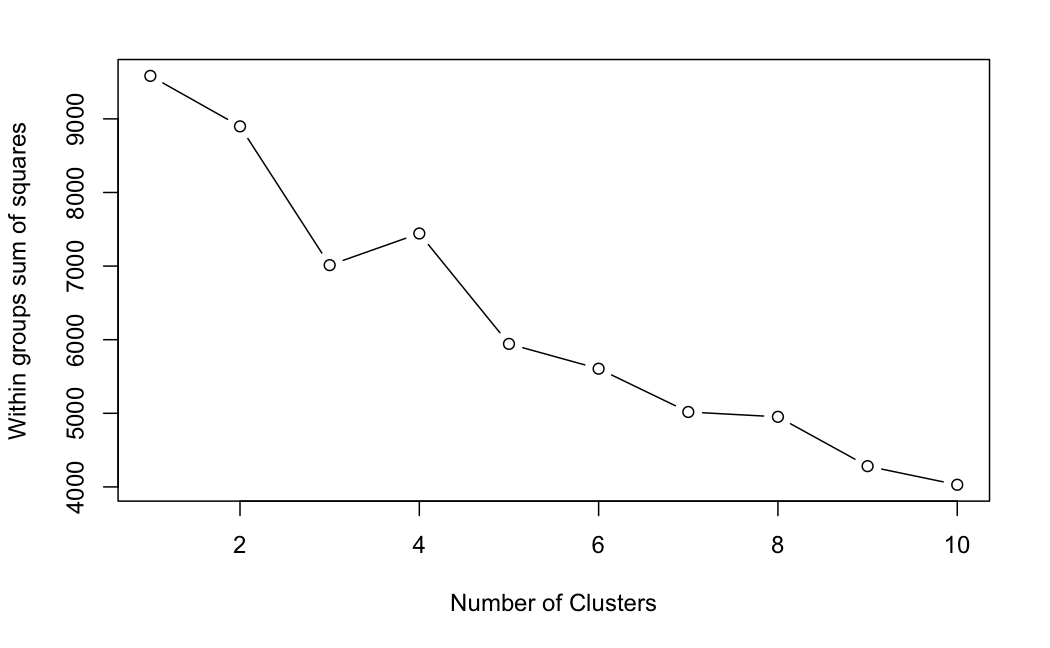
Overall, the data suggests that the clusters represent groups of customers with distinct purchasing behaviors and demographic characteristics. These insights can be used to inform marketing and sales strategies targeted towards each cluster.

**Optimizing Model Performance**

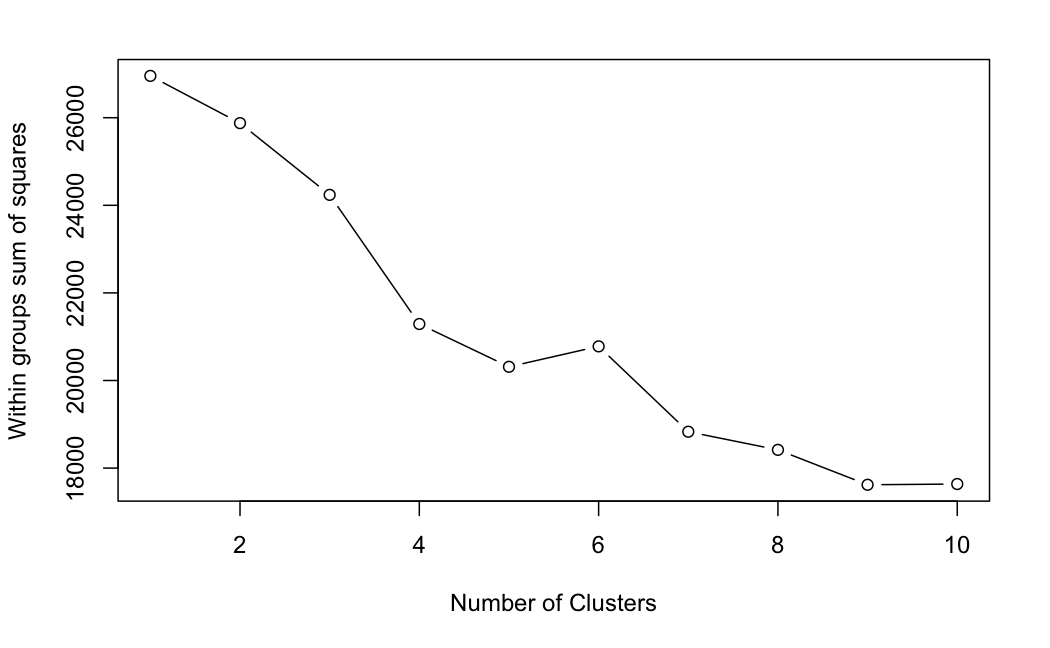
We utilized the elbow method in order to figure out the optimal cluster size. In Figure 1.4, the optimal cluster size is 5 because the within groups sum of squares begins to steadily decline. In Figure 1.5, the optimal cluster size is 5. In Figure 1.6, the optimal cluster size is 4.

****

**[1.4]** Elbow method generated for “Purchase Behavior”



**[1.5]** Elbow method generated for “Basis for Purchase”

****

**[1.6]** Elbow method generated with all variables

**Business Takeaways and Recommendations**

With this data, CRISA can identify customers that belong to specific cluster segments and implement appropriate marketing strategies. Moreover, CRISA can charge their clients for the data that can be used internally for their business needs.

CRISA initially segmented markets on the basis of purchase demographics, however, creating standardized promotional strategies is fruitless and not as effective. The following cluster segmentation based on purchase behavior and basis of purchase allows CRISA to determine unique characteristics and purchase patterns.

The following customer data takes into account that customers have different purchase behavior, therefore marketing strategies must be tailored for each individual group. Each cluster has different characteristics and responds differently to marketing stimuli. CRISA should use this information to tailor their marketing strategies to effectively target each segment.

When using clustering to segment customers, it is important to choose variables that can be translated into marketing actions in order to influence customer behavior in a meaningful way.