**CRISA CASE STUDY**

**Q1) a)**

**Variables used to describe purchase behavior:**

• Average Price • Brand Runs • Number of transactions • Number of brands • Others999 • Total volume • Value • Maximum brand loyalty

Maximum brand loyalty is obtained by taking maximum values out of the variables - Br. Cd. 57,144; Br. Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5. Others999 gives the share of transactions towards other brands which indicates that a customer is not brand loyal.

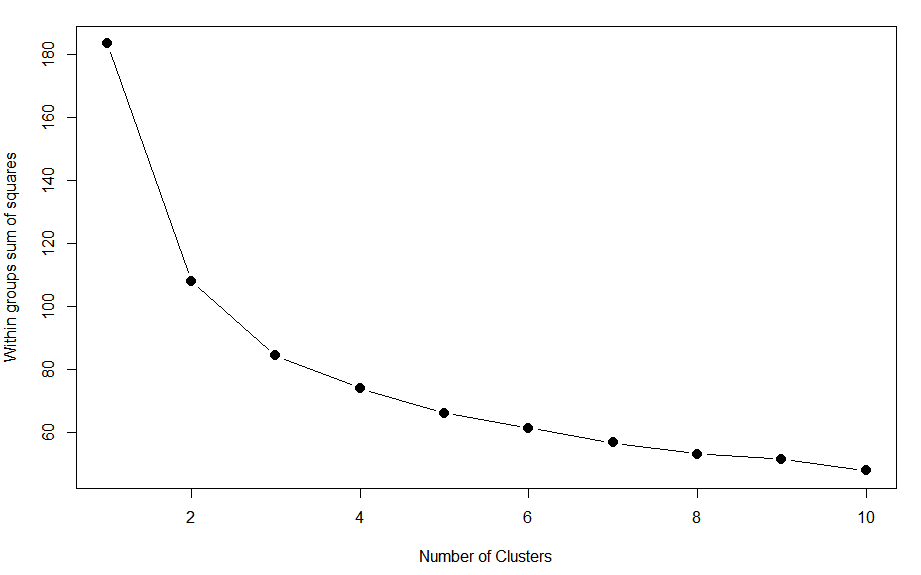
K-means algorithm is implemented on these variables and results are summarized in the table below:

**Step-1:** All variables were normalized from 0 to 1, except for max brand loyalty and others 999 which are already between 0 and 1. Also, any missing values in the data are removed

**Normalize <- function(x) {(x-min(x))/ (max(x)-min(x))}**

**data\_purchase1 <- na.omit(data\_purchase)**

**Step-2:** **Determining the optimal K using elbow graph**



From the above graph, there is no clear elbow curve visible, hence we will try 3,4 and 5 K values which could be the ideal values

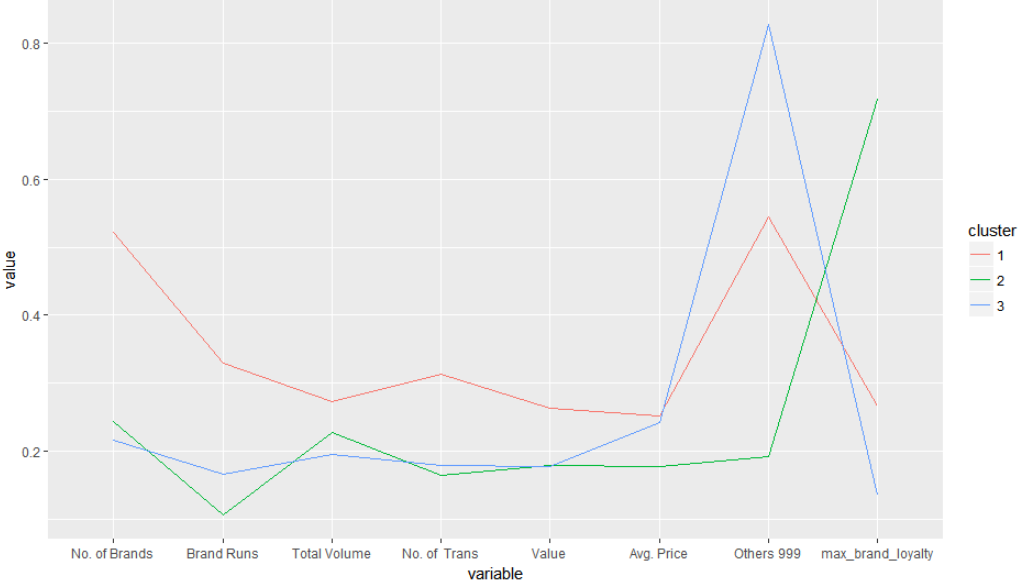
**Step-3: Summarizing the results for each K (from 3 to 5)**

**Distances and sizes summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of clusters** | **Size of each cluster** | **Total**  **SS** | **Avg distance between clusters(betweenss)** | **Avg distance to centroid within each cluster** |
| K=3 | 1: 205  2: 197  3: 198 | 183 | 99.2 | Avg (total withinss):84.2  Avg\_withinss (clus 1):32.7  Avg\_withinss (clus 2):27.9  Avg\_withinss (clus 3):23.6 |
| K=4 | 1: 152  2: 105  3: 186  4: 157 | 183 | 110.02 | Avg (total withinss):73.42  Avg\_withinss (clus 1):17.9  Avg\_withinss (clus 2):10.8  Avg\_withinss (clus 3):21.2  Avg\_withinss (clus 4):23.4 |
| K=5 | 1: 145  2: 100  3: 180  4: 135  5: 40 | 183 | 117.43 | Avg (total withinss):66  Avg\_withinss (clus 1):16.8  Avg\_withinss (clus 2):9.2  Avg\_withinss (clus 3):20.4  Avg\_withinss (clus 4):10.7  Avg\_withinss (clus 5):8.6 |

**Centroid plots: (plotting centers for each of the clusters)**

1. **K=3**



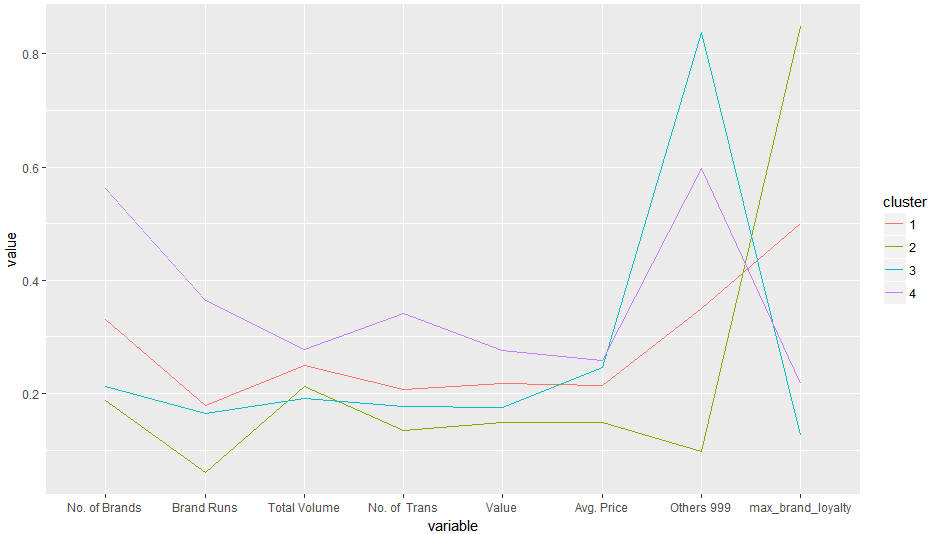
**Differences between each cluster:**

Clus 1: Large number of brands with highest brand runs. Highest number and volume of transactions whereas their brand loyalty is between cluster 1 and cluster 3 customers

Clus 2: Highest brand loyalty

Clus 3: Least brand loyalty because they tend to buy from Others999 brands. Least number of brands and total volume of transactions

1. **K=4**



**Differences between each cluster:**

Clus 1: Highest total volume and number of transactions. They have a significant peak for brand loyalty but also has considerable purchases from Others999 brands

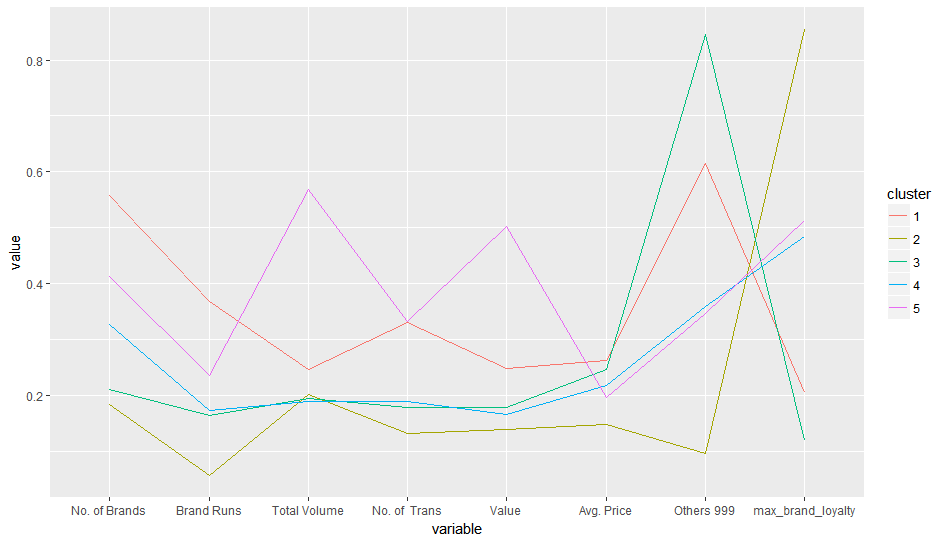
Clus 2: Least number of brands and high brand loyalty

Clus 3: moderate number of brands, brand runs, average price of transactions. They have a significant peak for Others999 which show that they are not brand loyal.

Clus 4: Not brand loyal and has high number of brands and volume

­­­­

1. **K=5**



**Differences between each cluster:**

Clus 1: Brand loyalty on lower side, but high number of brands

Clus 2: Highest brand loyalty with low number of brands and total volume

Clus 3: Least brand loyalty with peak in the graph observed

Clus 4: Least brand loyal customers after cluster 3

Clus 5: Highest total volume and number of transactions, value of goods purchased. Moderate brand loyalty

**After going through the distances and centroid plots, best K can be chosen using following criteria**

1. Distance within cluster (total\_withinss) should be minimum
2. Distance between clusters should be maximized (betweenss)
3. Information gained from the centroid plots (convergence)

Hence, we can say **K=5** is the best for this current segmentation on purchase behavior.

**Q1) b)**

**Variables used to describe basis of purchase**

All price categories, selling propositions, Purchase volume with no promotion, promotion 6 and other promotions

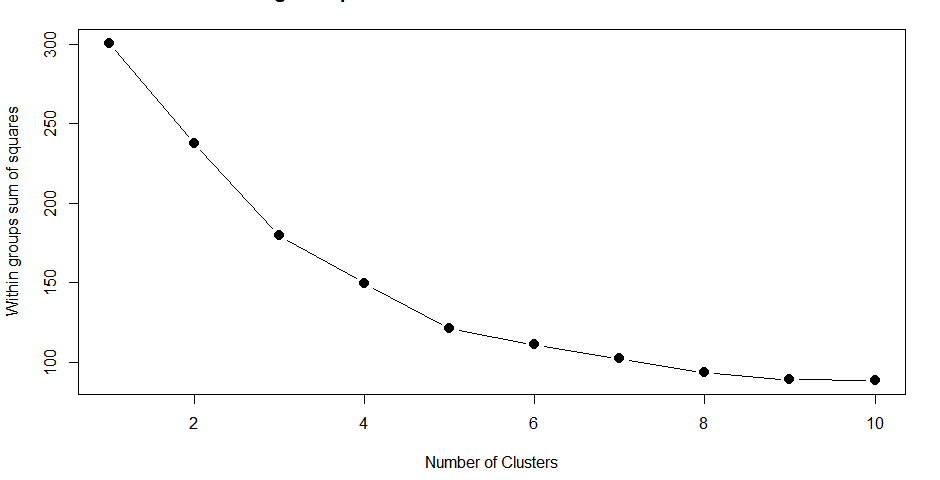
**Data exploration on selling propositions:**

From 9 to 15, selling proposition variables have greater than 50% values to be 0, which are not contributing anything. Hence, we use variables from 5 to 8.

|  |  |
| --- | --- |
| **Variable** | **% of rows with 0 value** |
| PropCat 5 | 9% |
| PropCat 6 | 42% |
| PropCat 7 | 44% |
| PropCat 8 | 48% |
| PropCat 9 | 59% |
| PropCat 10 | 76% |
| PropCat 11 | 72% |
| PropCat 12 | 82% |
| PropCat 13 | 73% |
| PropCat 14 | 57% |
| PropCat 15 | 75% |

**Step-1:** All variables were normalized from 0 to 1, so no problem to implement K means

**Step-2:** **Determining the optimal K using elbow graph**



Unlike the graph in 1a), we can see after K=5, graph is flattening out and it indicates the convergence at K=5. But, we’ll be re affirming the K=5 by comparing it with K=4.

**Step-3: Summarizing the results for K=5 and K=4**

**Distances and sizes summary**

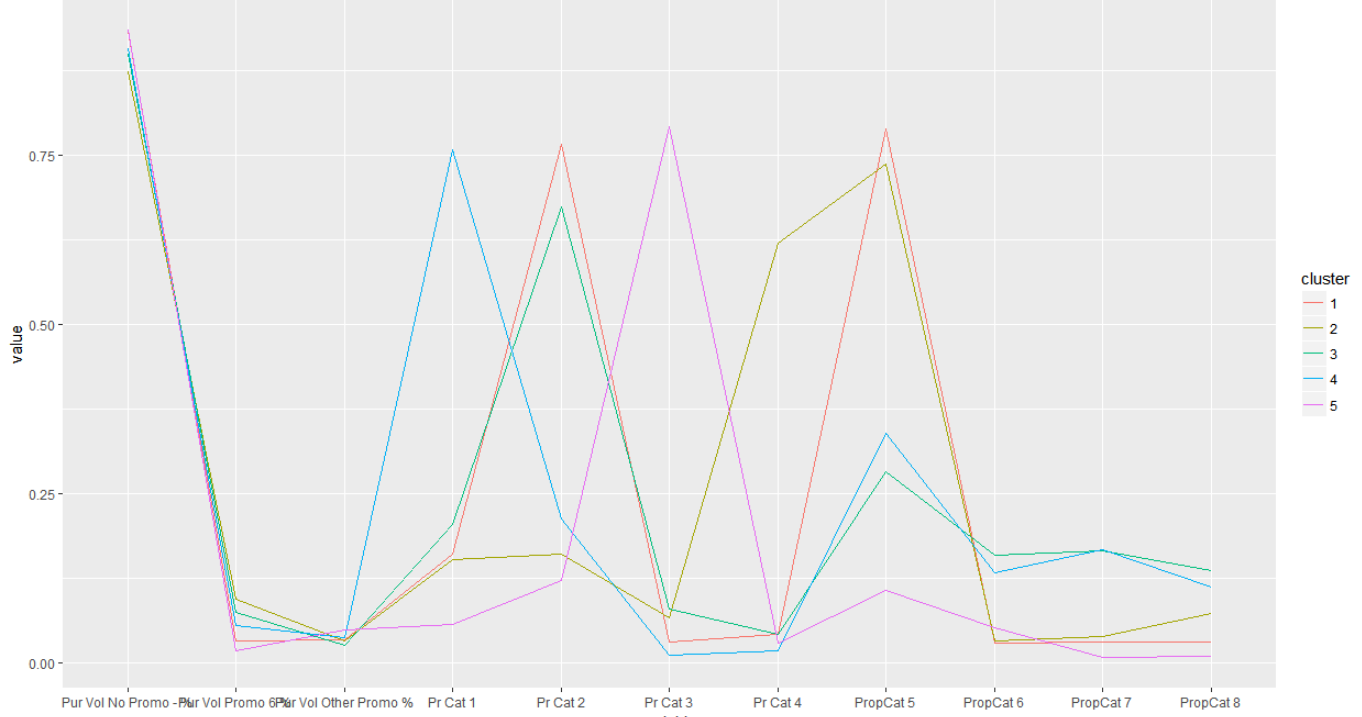
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of clusters** | **Size of each cluster** | **Total**  **SS** | **Avg distance between clusters** | **Avg distance to centroid within each cluster** |
| K=4 | 1: 57  2: 141  3: 323  4: 79 | 300 | 153.85 | Avg (total withinss):146.8  Avg\_withinss (clus 1):10.1  Avg\_withinss (clus 2):44.8  Avg\_withinss (clus 3):82.5  Avg\_withinss (clus 4):9.3 |
| K=5 | 1: 170  2: 55  3: 182  4: 119  5: 74 | 300 | 179.61 | Avg (total withinss):121.1  Avg\_withinss (clus 1):18.2  Avg\_withinss (clus 2):9.5  Avg\_withinss (clus 3):49  Avg\_withinss (clus 4):36  Avg\_withinss (clus 5):87 |

We can clearly see marked differences in distance within cluster being minimum and also distance between clusters is considerably higher when k=5. Also the size distribution among clusters is much better when k=5

**Hence, K=5 is the best K for this segmentation on basis of purchase**

**Centroid plots:**

**K=5**



**Differences between each cluster:**

Clus 1: Customers in this cluster buy popular soap which is beauty related and don’t use high number of promos

Clus 2: Customers in this cluster buy sub-popular soap which is beauty related and used promotion code 6 relatively higher than other clusters

Clus 3: Customers in this cluster buy popular soaps which are beauty and health related

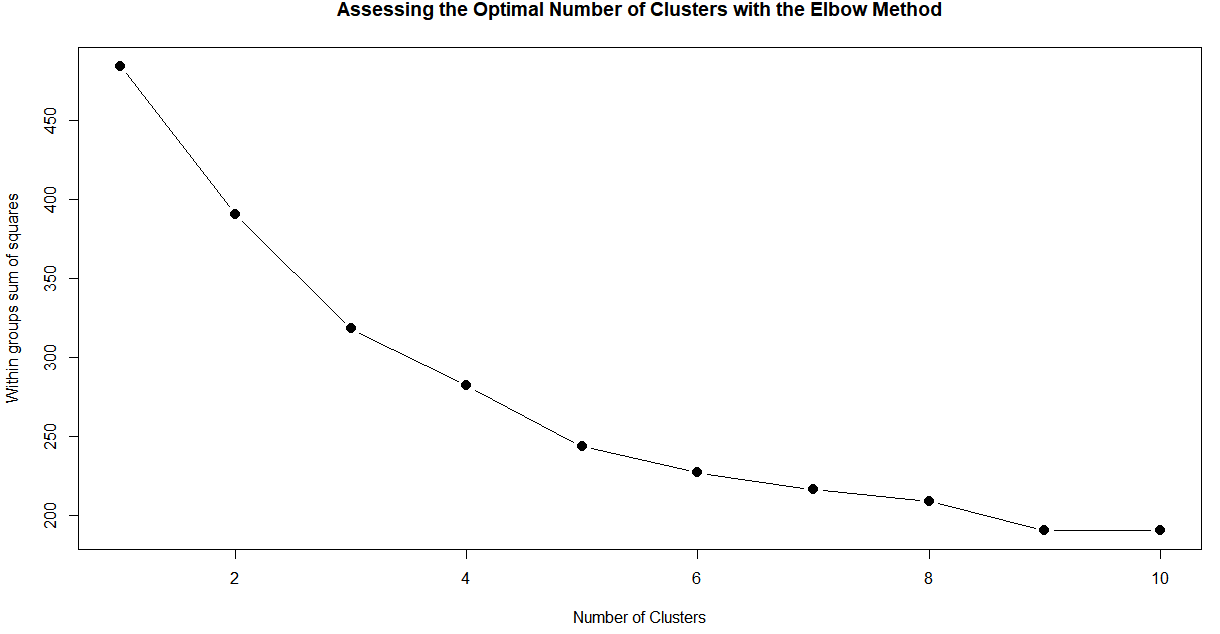
Clus 4: Same as cluster 3, but they buy high proportion of premium soaps

Clus 5: Customers in this cluster buy economic soap and don’t use high number of promos­­­

**Q1) c) The variables that describe both purchase behavior and basis of purchase:**

**Step-1:** All variables were normalized from 0 to 1, so no problem to implement K means

**Step-2:** **Determining the optimal K using elbow graph:**



Unlike the graph in 1a), we can see after K=5, graph is flattening out and it indicates the convergence at K=5. But, we’ll be re affirming the K=5 by comparing it with K=4.

**Step-3: Summarizing the results for K=5 and K=4**

**Distances and sizes summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number of clusters** | **Size of each cluster** | **Total**  **SS** | **Avg distance between clusters** | **Avg distance to centroid within each cluster** |
| K=4 | 1: 146  2: 74  3: 258  4: 122 | 484.16 | **205.80** | Avg (total withinss):**278.3**  Avg\_withinss (clus 1):61.7  Avg\_withinss (clus 2):17.7  Avg\_withinss (clus 3):128.7  Avg\_withinss (clus 4):70.1 |
| K=5 | 1: 73  2: 117  3: 222  4: 136  5: 52 | 484.16 | **240.8** | Avg (total withinss):**243.3**  Avg\_withinss (clus 1):17.7  Avg\_withinss (clus 2):66.6  Avg\_withinss (clus 3):89.5  Avg\_withinss (clus 4):56.06  Avg\_withinss (clus 5):13.93 |

We can clearly see marked differences in distance within cluster being minimum and also distance between clusters is considerably higher when k=5. Also the size distribution among clusters is much better when k=5

**Hence, K=5 is the best K for this segmentation on basis of purchase and purchase behaviour**

**Questions asked:**

**How should k be chosen?**

The value of ‘K’ should be chosen in such a way that:

1) The Intra cluster distances are minimum in all clusters

2) The clusters are well apart. That is, the inter cluster distances are maximum

Value of K is already chosen using these aspects in the above questions

**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?**

The percentages of total purchases should not be considered individually as they increase the inter cluster distances and the effectiveness of the clustering drops. Instead, consider MaxBrCode(Max proportion of purchase) which give the brand loyalty of the customer

**2. a)** Three types of segmentation tried out in the 1st question, where already best k is chosen each set of variables, now we have to choose best out of this three below models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type | Best K | Total SS | Between SS | BSS/TSS | Total Withins |
| Purchase behaviour | 5 | 183 | 117 | 0.631 | 66 |
| Basis of purchase | 5 | 300 | 179 | 0.596 | 121 |
| Both | 5 | 484 | 240 | 0.495 | 243 |

**BSS/TTS is high for the kmeans model on purchase behavior and also distance within clusters is minimal for that model. Hence we can use the purchase behavior model with K=5 as our best model**

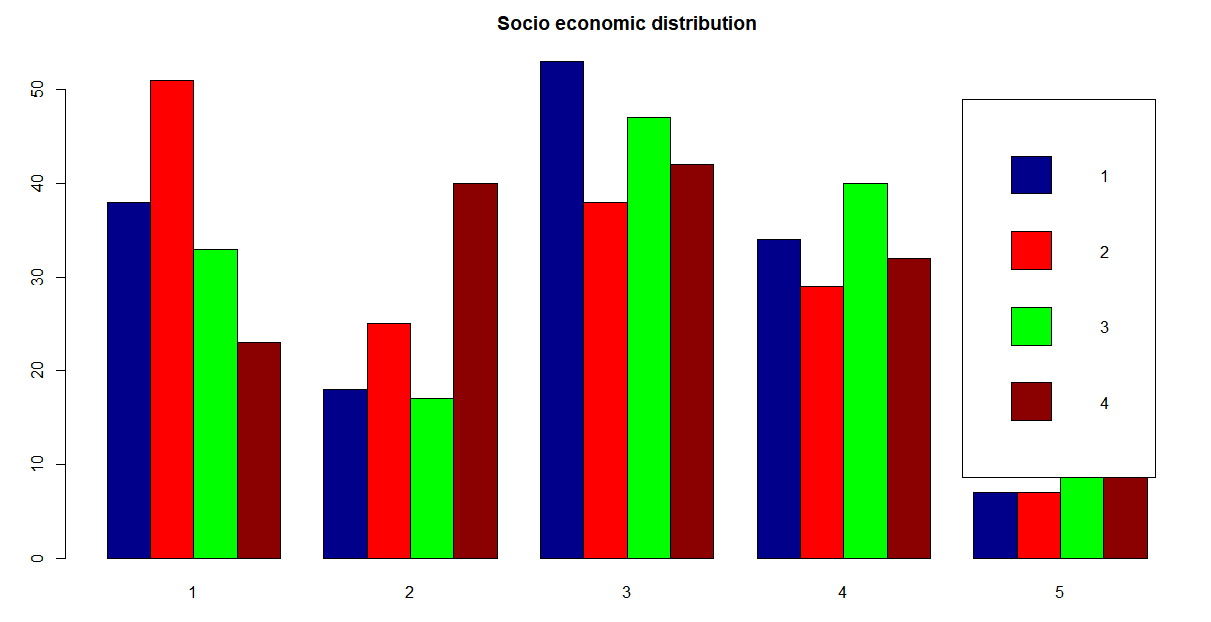
**2. B)**  **Comparison of demographic and other features f these clusters in best model:**

**i) SEC:**

Cluster 2 which has maximum brand loyaltyhave higher percentage of people who are high class belonging to lower economic classes

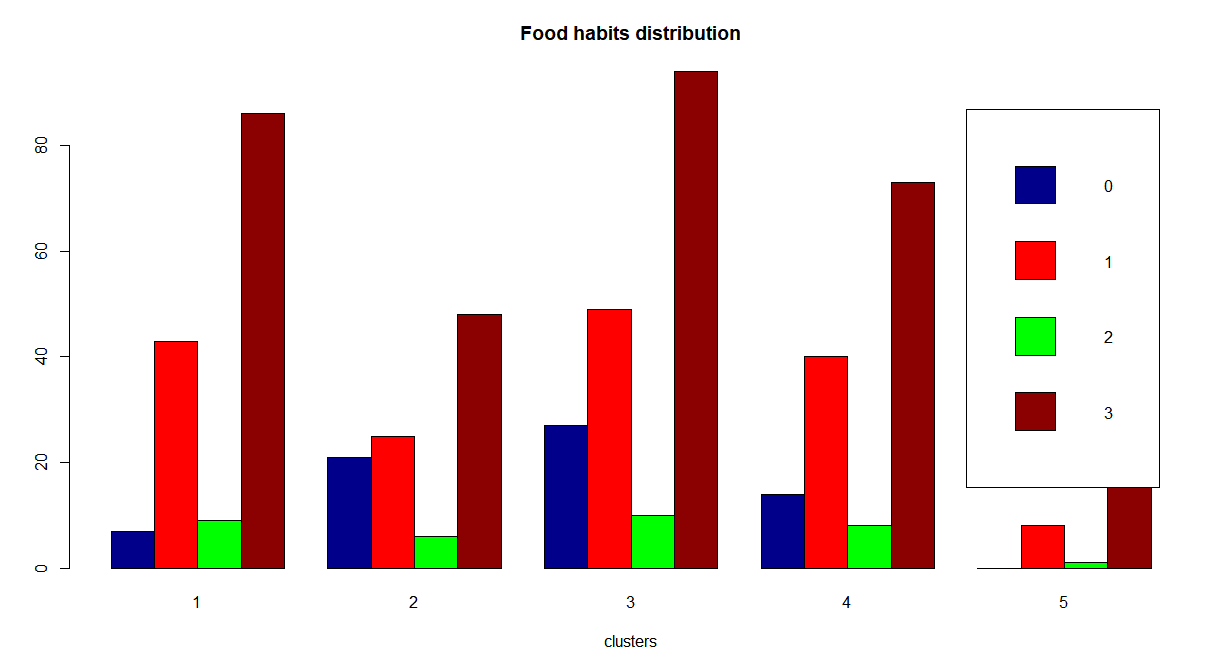
Cluster 1 which has least brand loyaltyhave higher percentage of people who belong to high class 1 and 2

This makes sense, because rich class have lot of options to choose from, whereas low class members buy something basic soap like lifebuoy over and over, hence you see this trend.



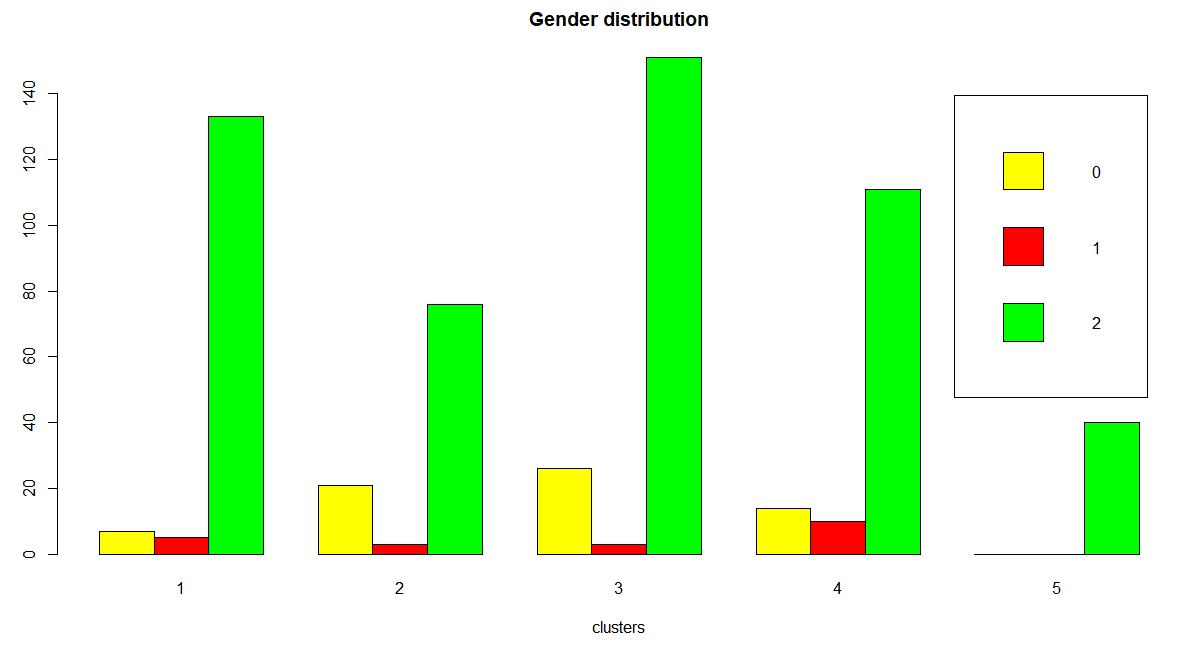
**ii) Food eating habits**

No Particular trend seen among the food eating habits and clusters



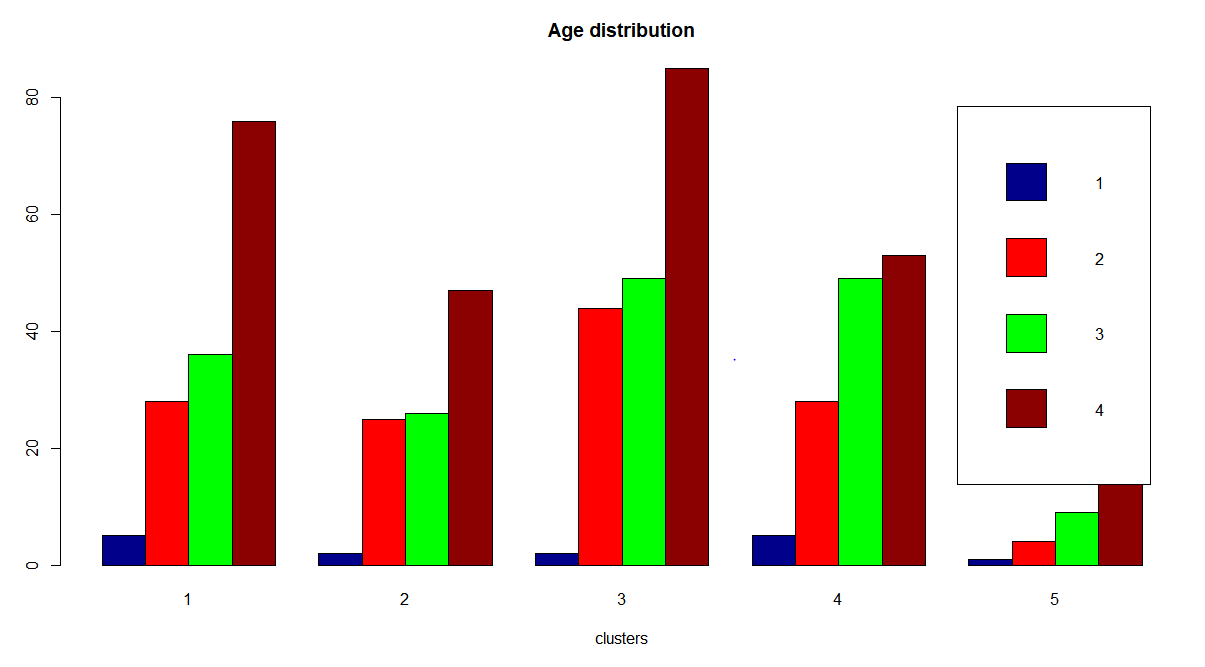
**iii) Gender**

Customers in cluster are majorly women



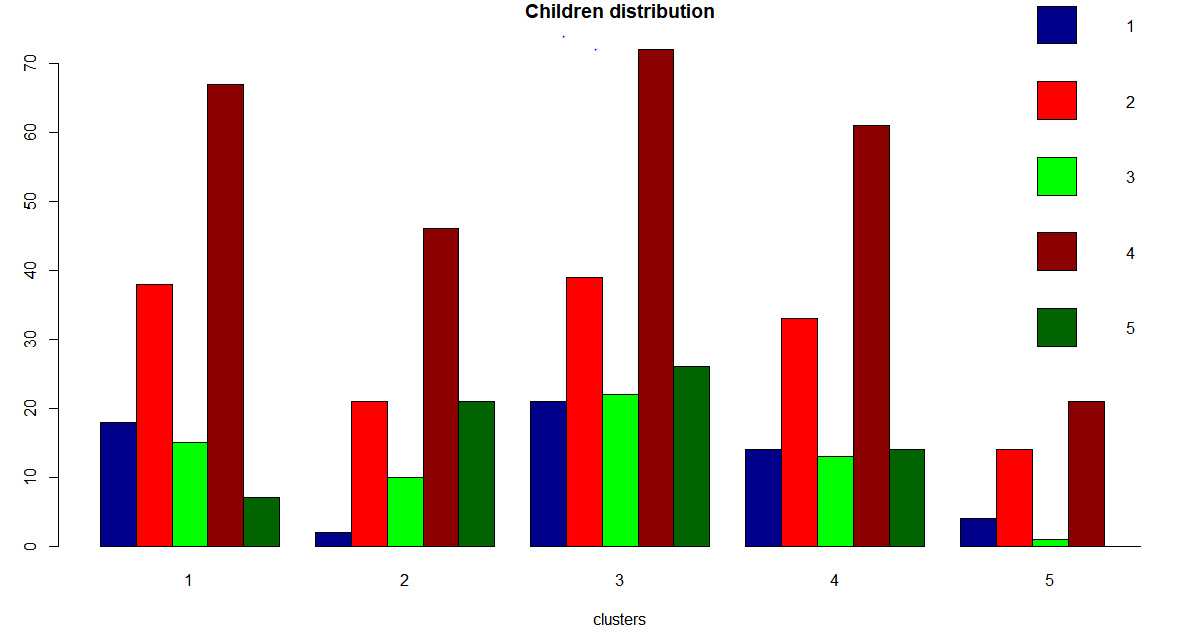
**iv) Age**

No significant trend, al clusters have people have with age greater than 45 as majority



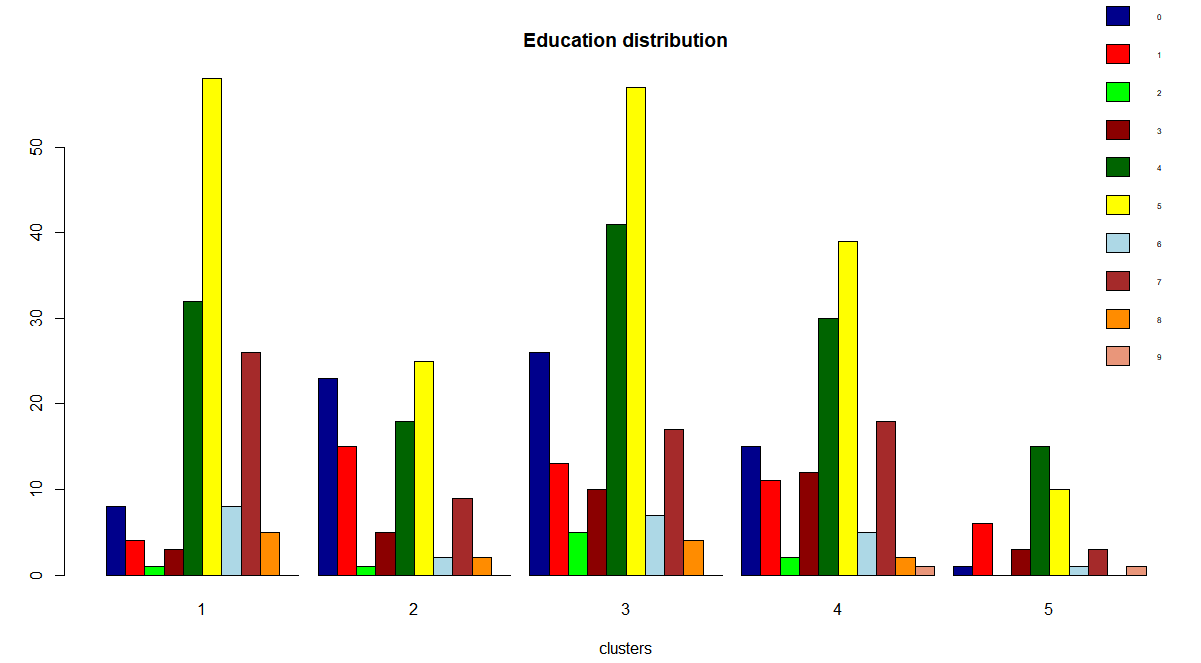
**V) Children**

Majority of clusters have customers with no children, no significant differences in that aspect



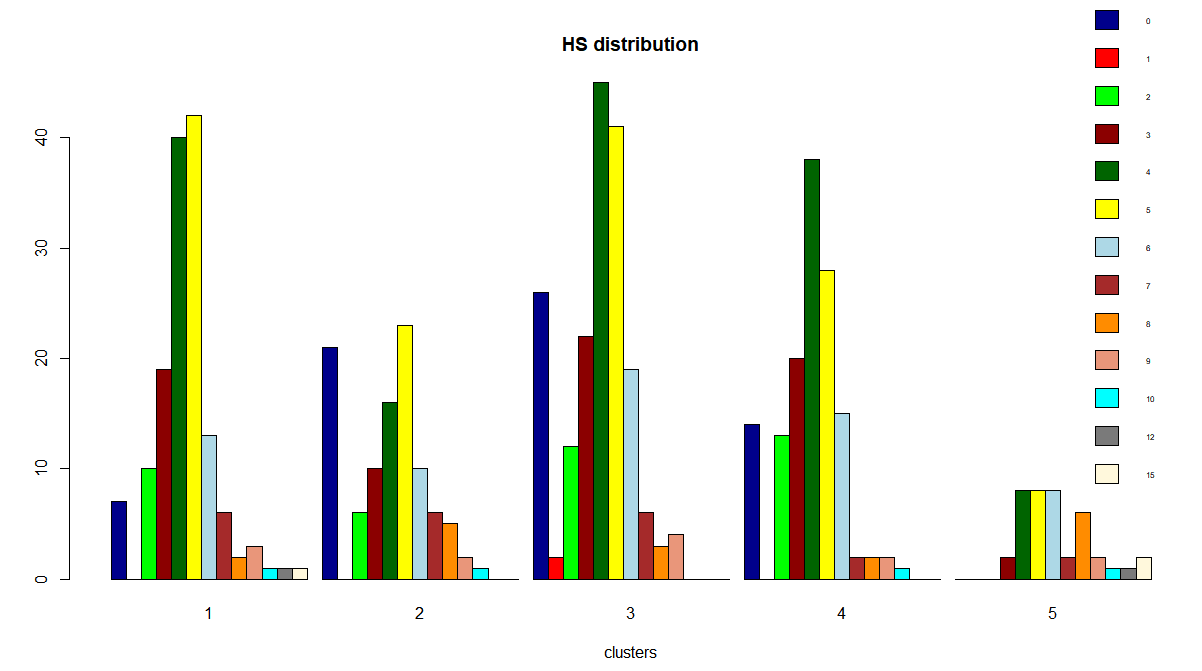
**Vi) Education**

Most of the customers in this sample have had 10-12 years of school. But, if you take a closer look at the clusters 1 and clusters3, we can see relatively higher college degree customers who buy premium soaps and economically well placed

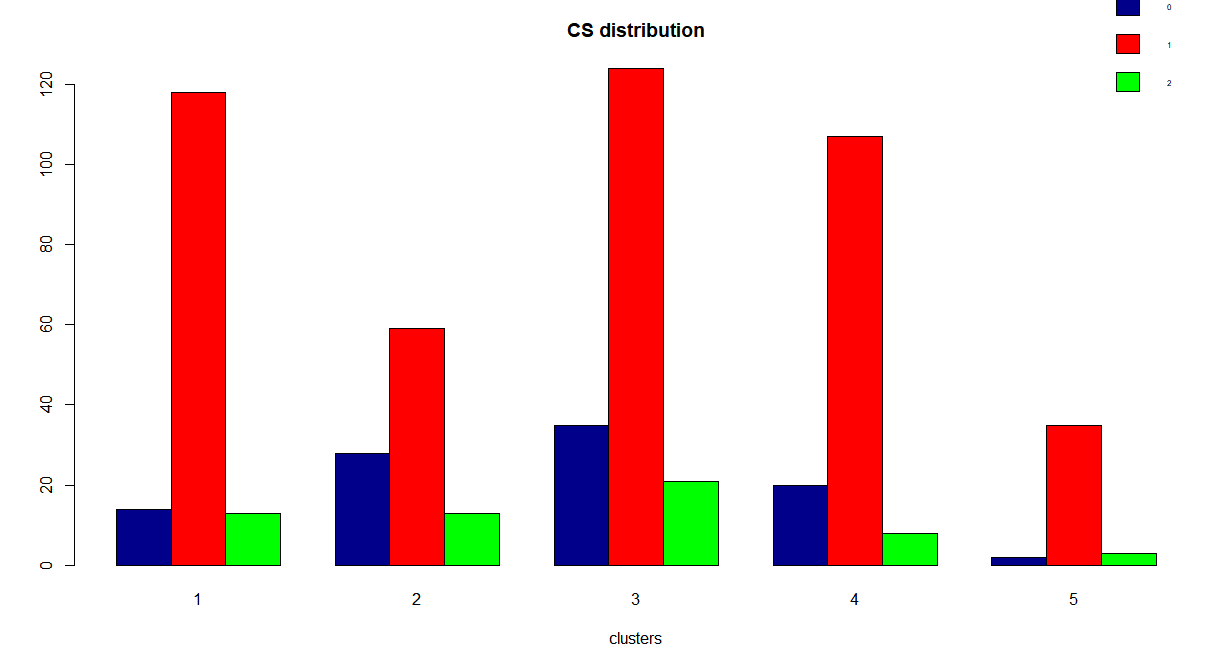


**Vii) Household Size**

Most customers seem to have 4-5 people in the household. It looks these people do not care about brands and that they prefer to buy value packs and premium soaps.



**Viii) CS**

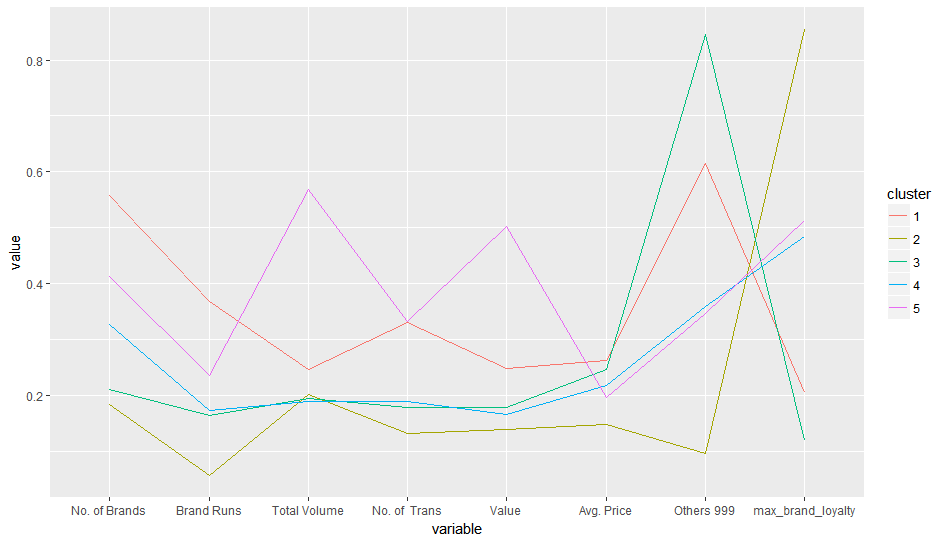
Most people across clusters have a cable TV available.

**iX) Mother tongue**

This demographic does not appear to be significant as all clusters are dominated by customer base whose mother tongue is Marathi. Looks like this sample data was gathered from a locality with a predominantly Marathi speaking population

**X) Brand Loyalty**

Cluster 2 has high brand loyalty, 4 and 5 have moderate brand loyalty values. 1 and 3 has least



**Xi)**

**Price categories**

Below table gives the share of members who had purchased atleast 50% in their respective price categories:

**Ex:-** In cluster 1, there are total of 145 people, of which only 31 people had a value greater than 0.50 in Pr Cat1, dividing 31/145 gives 21% which is an indication of share of that particular type of product in that cluster

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Pr Cat 1 | Pr Cat 2 | Pr Cat 3 | Pr Cat 4 |
| 1 | 21% | 60% | 0% | 1% |
| 2 | 9% | 44% | 47% | 0% |
| 3 | 35% | 38% | 0% | 21% |
| 4 | 13% | 66% | 14% | 0% |
| 5 | 8% | 68% | 15% | 0% |

**From this table, we can see cluster1 customers are majorly buying regular soaps followed by premium. Cluster 3 has equal share of regular soaps and premium soaps.**

**Cluster 2 and cluster 4 has majority of customers buying regular/economic soaps**

**This makes sense because clusters 1 and 3 have higher class people compared to 2,4 and 5 which is why the premium soap share is high**

**Promotions**

**Using similar criteria as above, we can see in the most of clusters people are not using any promo codes**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster | Pur Vol No Promo - % | | Pur Vol Promo 6 % | | Pur Vol Other Promo % |
| 1 | 92% | 0% | | 0% | |
| 2 | 96% | 0% | | 1% | |
| 3 | 92% | 0% | | 0% | |
| 4 | 95% | 0% | | 0% | |
| 5 | 98% | 0% | | 0% | |

**Conclusion:**

Most consumers are females, thus most of the ads should be targeted for women. Also, most customers fall in the segment who are not particularly brand loyal but prefer to buy value added packs and premium soaps. As most people have a TV/cable, advertisements can be broadcast on television as an effective means of promoting products.

Not many people care about price offs. People buying on basis of price offs are very small across all clusters. Thus, in order to promote brand loyalty, manufacturers should promote their brands by gifting coupons or exchange offers