ML_FinalExam

Sindhu Vegiraju

05/05/2021

Problem Statement: 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 todesign more effective customer reward systems and thereby increase brand loyalty

Loading Required Packages

```
library(dplyr)
library(ISLR)
library(caret)
library(factoextra)
library(GGally)
library(hrbrthemes)
library(viridis)
set.seed(123)
```

Import BathSoap dataset and Cleaning the Data

```
BathSoap<- read.csv("D:/MACHINE LEARNING/Final exam/BathSoap.csv")
BSoap <- data.frame(sapply(BathSoap, function(x) as.numeric(gsub("%", "",
x))))
```

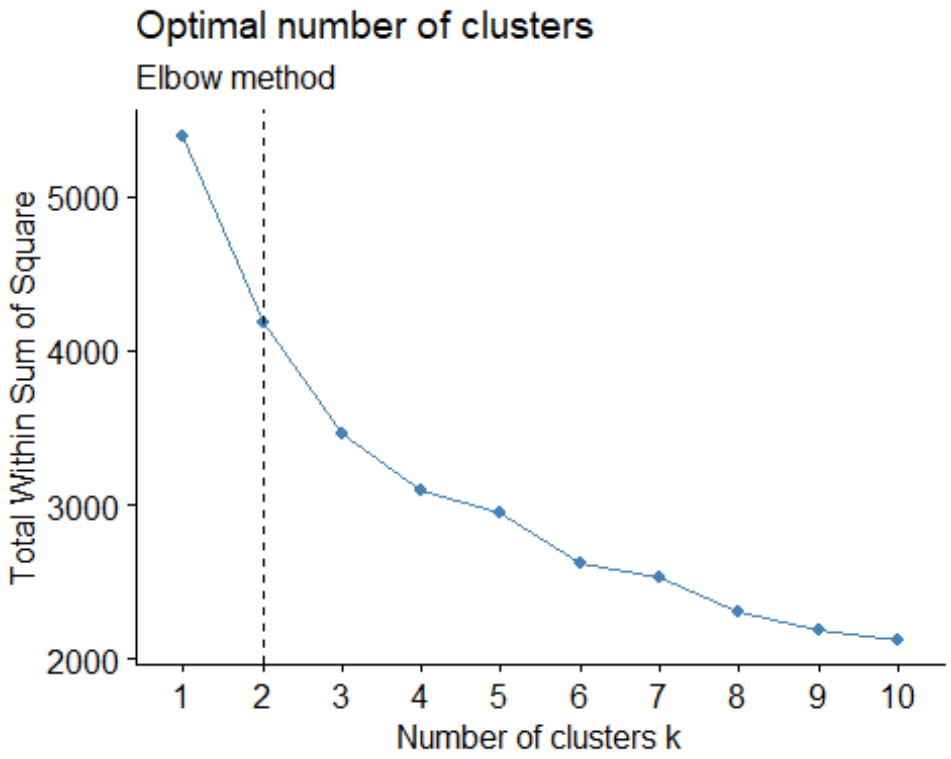
a)The variables that describe purchase behavior (including brand loyalty).

Variables used for this process are: Average Price, Brand Runs, Number of transactions, Number of brands, Others999,Total volume, Value, Maximum brand loyalty

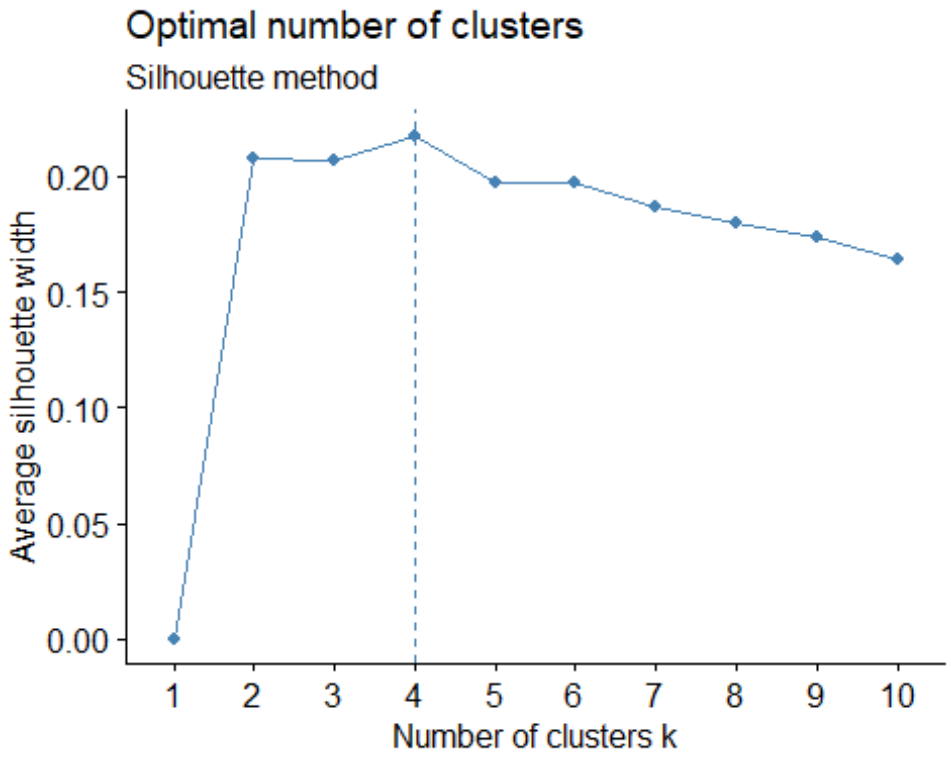
Maximum brand loyalty is obtained by taking maximum values out of the variables - Br. Cd. 57,144; Br. Cd. 55; Br. Cd. 272Cd.286; Br. Cd.24; Br. Cd.481; Br. Cd.352, Br. Cd.5. Others999

gives the share of transactions towards other brands which indicates that a customer is not brand loyal.For a customer to be considered as loyal, the Max Brand purchase percentage is expected to be higher than the Other Brand purchase percentage. K-means algorithm is implemented on these variables and results are summarized below.

```
Loyalty <- BSoap[,23:30]
Loyalty$Max_Brand <- apply(Loyalty,1,max)
BSoapLoyal <- cbind(BSoap[,c(12,13,14,15,16,19,20,31)], MaxLoyal =
Loyalty$Max_Brand)
BSoapLoyal <- scale(BSoapLoyal)
library(NbClust)
fviz_nbclust(BSoapLoyal, kmeans, method = 'wss') +geom_vline(xintercept = 2,
linetype = 2)+
 labs(subtitle = 'Elbow method')
```



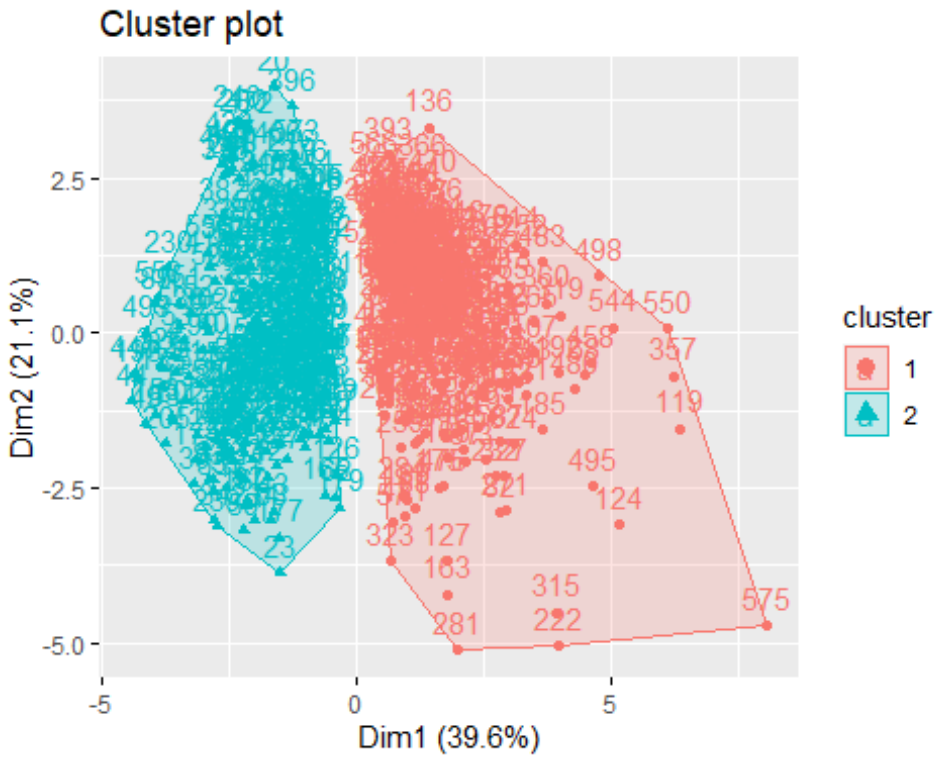

```
fviz_nbclust(BSoapLoyal, kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```



silhouette = 2 Elbow = 4 we would seek a k that produces clusters that are distinct and separate from one another, in ways (variables) that are translatable into marketing actions.The complexity of marketing to 5 segments would probably not be supported by clustering just based on purchase behavior so we will look at 2 clusters for those variables

Now we will run Kmeans with k=2 and nstart = 25 and plot the clusters using fviz_cluster which gives "Brand Loyal Customers" and "Not Brand Loyal Customers."

```
model1 <- kmeans(BSoapLoyal, centers = 2, nstart = 25)
BSoapLoyal <- cbind(BSoapLoyal, Cluster = model1$cluster)
fviz_cluster(model1, data = BSoapLoyal)
```



we will store the centers of the model in Output and print the size of the 2 clusters.

```
result1<-as.data.frame(cbind(1:nrow(model1$centers),model1$centers))
result1$V1<-as.factor(result1$V1)
result1
```

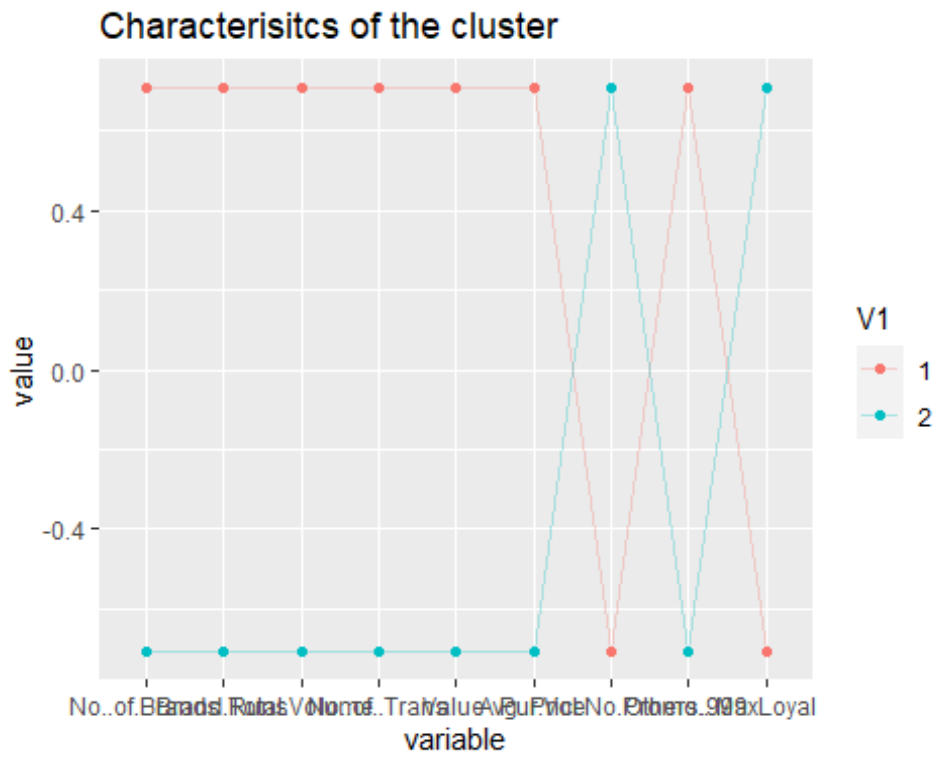
```
##   V1 No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
Avg..Price
## 1  1     0.5127116  0.6439617    0.3861580     0.5741146  0.4986122
0.1990075
## 2  2    -0.5444257 -0.6837944   -0.4100441    -0.6096268 -0.5294542 -
0.2113172
##   Pur.Vol.No.Promo.... Others.999   MaxLoyal
## 1           -0.1894729  0.3998053 -0.5092718
## 2            0.2011928 -0.4245356  0.5407731
```

```
model1$size
```

```
## [1] 309 291
```

Parallel plot to visualize the cluster.

```
ggparcoord(result1,columns = 2:ncol(result1), groupColumn = 1, showPoints =
TRUE, title = "Characterisitcs of the cluster",alphaLines = 0.3
)
```



Customers in Cluster 2 are Brand Loyal as they buy least number of brands. They have maximum brand loyalty

customers in Cluster 1 appear to buy from Others999 brands which indicate they are not brand loyal customers.They buy the highest number of brands.

b)The variables that describe the basis for purchase

Variables used for this process are:

1)All price categories

2)Selling propositions

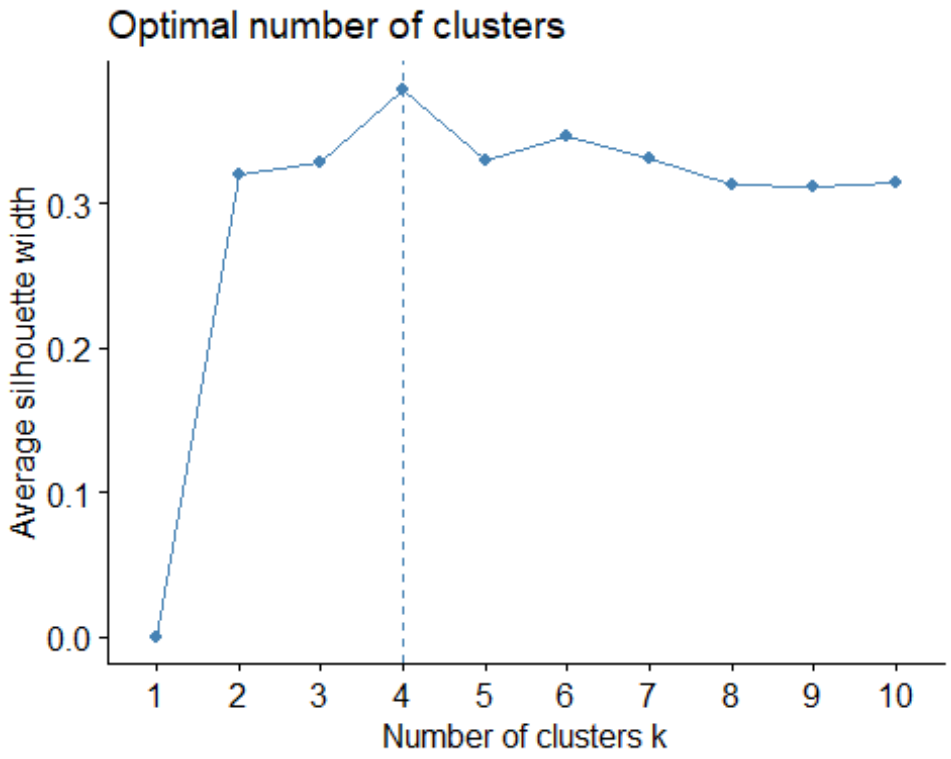
3)Purchase volume with no promotion, promotion 6 and other promotion

```
BSoap_SP <- BSoap[,36:46]
BSoap_SP$Max <- apply(BSoap_SP,1,max)
BSoap_SP$Max_Brand <- colnames(BSoap_SP)[apply(BSoap_SP,1,which.max)]
PCategory <- BSoap[,32:35]
PCategory$Max <- apply(PCategory,1,max)
PCategory$Max_Brand <- colnames(PCategory)[apply(PCategory,1,which.max)]
table(PCategory$Max_Brand)

##
## Pr.Cat.1 Pr.Cat.2 Pr.Cat.3 Pr.Cat.4
##      132      343       78       47
```

```
Promotion <- BSoap[,20:22]
Promotion$Max <- apply(Promotion,1,max)
Promotion$Max_Brand <- colnames(Promotion)[apply(Promotion,1,which.max)]
table(Promotion$Max_Brand)

##
##  Pur.Vol.No.Promo.... Pur.Vol.Other.Promo..     Pur.Vol.Promo.6..
##                   595                      1                     4
```

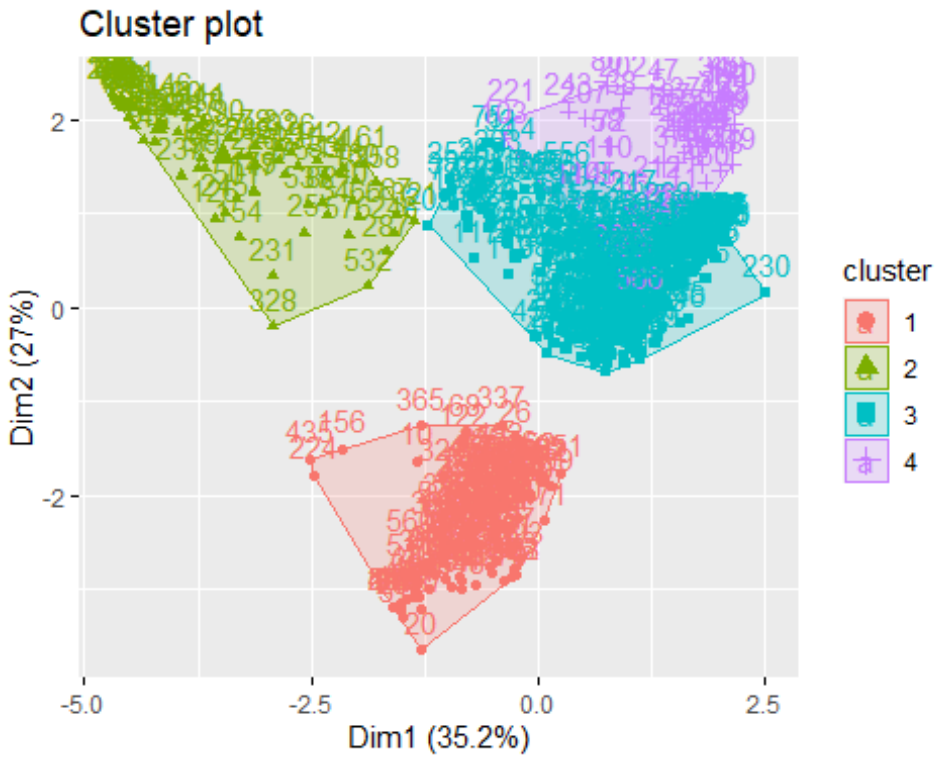
Most people seemed to be responding to only few of the propositions.Hence, we have only considered the more powerful Selling Propositions.The same goes for Promotions and price categories.

```
PurchaseBehaviour <- BSoap[,c(20,32,33,34,35,36,45)]
PurchaseBehaviour <- scale(PurchaseBehaviour)
fviz_nbclust(PurchaseBehaviour, kmeans, method = "silhouette")
```



The K means Clustering model is computed in order to measure basis for purchase. By considering the above silhouette method, k value is 4.

```
model2 <- kmeans(PurchaseBehaviour, centers = 4, nstart = 25)
PurchaseBehaviour <- cbind(PurchaseBehaviour, Cluster = model2$cluster)
fviz_cluster(model2, data = PurchaseBehaviour)
```




```
result2<-as.data.frame(cbind(1:nrow(model2$centers),model2$centers))
result2$V1<-as.factor(result2$V1)
result2
```

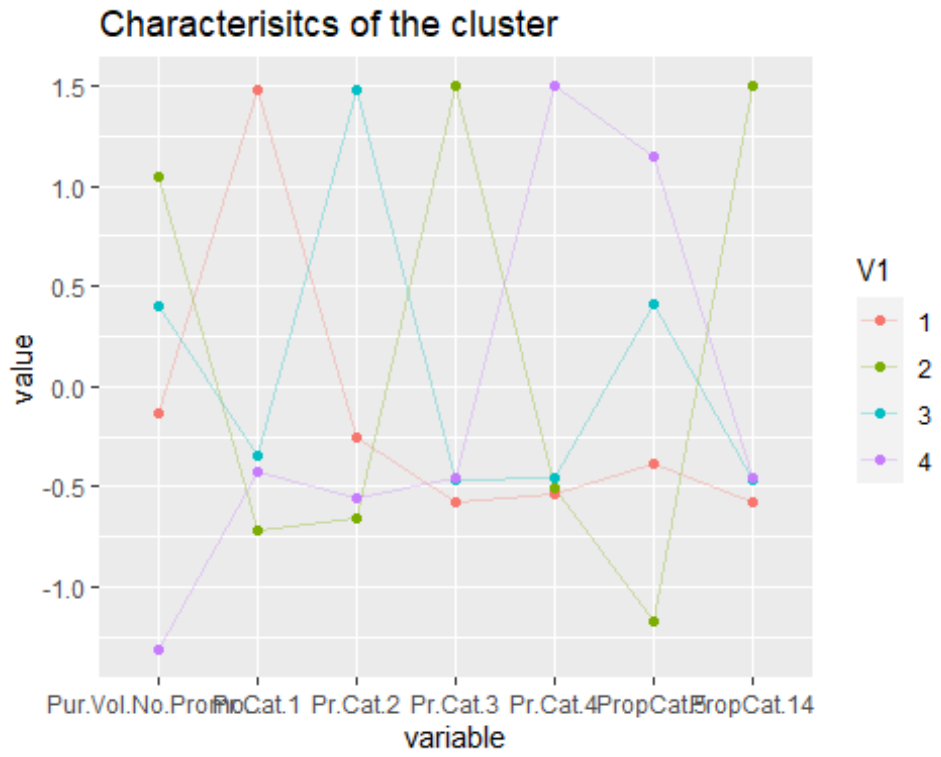
```
##   V1 Pur.Vol.No.Promo....   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3   Pr.Cat.4
## 1  1          -0.07783200  1.5565897 -0.7739911 -0.4690108 -0.3680463
## 2  2           0.19209812 -0.7884554 -1.1293188  2.3715353 -0.3204763
## 3  3           0.04536324 -0.3925317  0.7715511 -0.3159766 -0.2369244
## 4  4          -0.34625722 -0.4748374 -1.0435839 -0.2935805  2.7975778
##    PropCat.5 PropCat.14
## 1 -0.4103233 -0.4655254
## 2 -1.0922709  2.3739067
## 3  0.2789543 -0.3167251
## 4  0.9224559 -0.3012884
```

```
model2$size
```

```
## [1] 139  78 328  55
```

Parallel plot to visualize the cluster.

```
ggparcoord(result2,
           columns = 2:ncol(result2), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



Customers in clusters 4 do not purchase products without promotion offers,though availing promotional offers their maximum proportion of purchase is so low that they wont easily converted to loyal customers by offering more price offs.

Customers in clusters 1 are mainly those who purchase more volume of items even when not on promotion. These Customers evidently purchase products from a single price category. They purchase almost similarly both during price offs and no price offers.

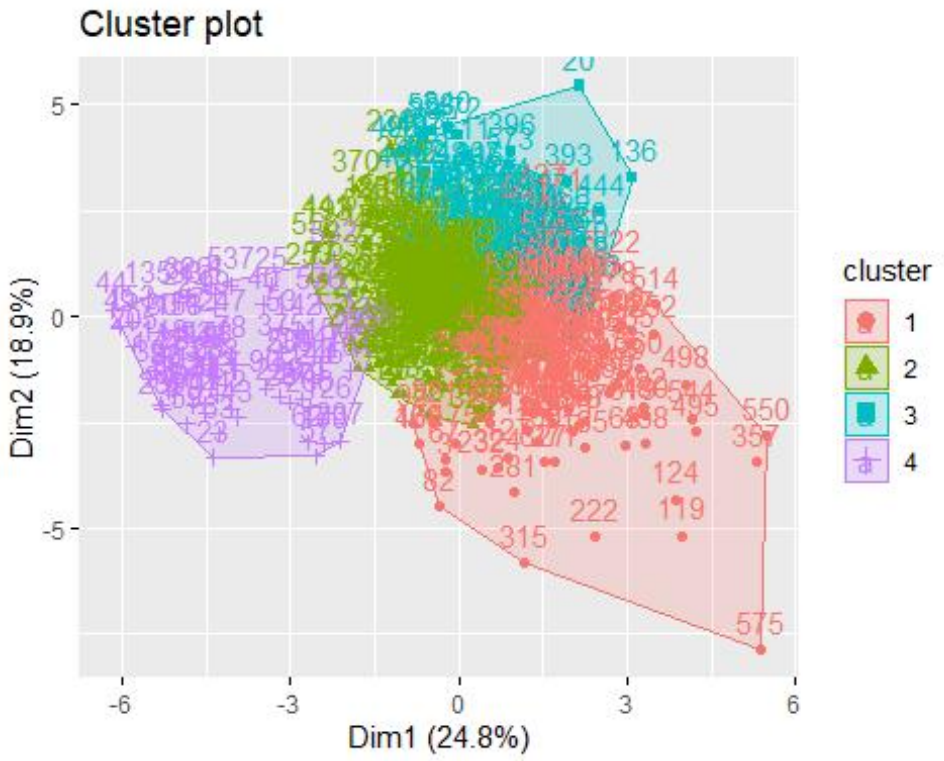
Customers in clusters 2,3 have a moderate behavior, they purchase products of a specific price category mostly. Their purchases are not affected with promotional offers.

C)The variables that describe both purchase behavior and basis of purchase

```
PurchaseBehaviour<-BSoap[,c(12:16,19:22,31:35,45)]
LP<-as.data.frame(scale(PurchaseBehaviour))
model3<-kmeans(LP,4,nstart=50)
```

When plotting the model for k = 4 and k = 5, we can see that the aspects can be resolved by simply using 4 clusters without drawing another 1. Hence, we'll use k = 4 here.

```
fviz_cluster(model3,LP)
```




```
result3<-as.data.frame(cbind(1:nrow(model3$centers),model3$centers))
result3$V1<-as.factor(result3$V1)
result3
```

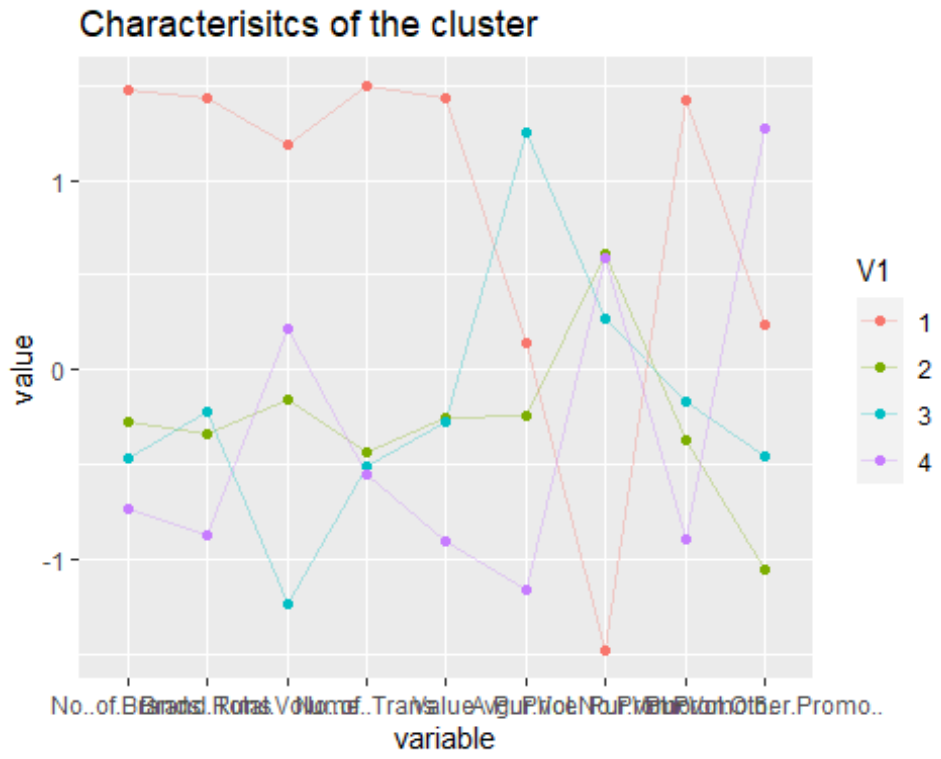
```
##   V1 No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
Avg..Price
## 1  1     0.9402265  1.0848692   0.57410652     0.9967751  0.7014988
0.1146557
## 2  2    -0.2698909 -0.3631494  -0.10394079    -0.3469387 -0.2044276 -
0.3126686
## 3  3    -0.3974749 -0.2663604  -0.64734600    -0.4009045 -0.2161136
1.3356727
## 4  4    -0.5838929 -0.8005685   0.08338649    -0.4334206 -0.5556157 -
1.3182534
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..  Others.999
## 1           -0.4970787         0.5851394            0.07623451  0.22631172
## 2            0.2186919        -0.2064677           -0.10097215 -0.03415958
## 3            0.1021136        -0.1167157           -0.01944329  0.52613844
## 4            0.2109945        -0.4375343            0.21655797 -1.26409557
##      Pr.Cat.1   Pr.Cat.2   Pr.Cat.3    Pr.Cat.4 PropCat.14
## 1  0.07462146  0.2231051 -0.2728659 -0.08890361 -0.2720400
## 2 -0.56612636  0.5716074 -0.2966520  0.31381339 -0.3008952
## 3  1.63258333 -0.8356254 -0.4841268 -0.35883152 -0.4775071
## 4 -0.79555111 -1.2196143  2.4956680 -0.33705699  2.4982280
```

```
model3$size
```

```
## [1] 164 252 114  70
```

Parallel plot to visualize the cluster.

```
ggparcoord(result3,
           columns = 2:10, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



Cluster1: Consumers in this cluster are least preferring the other brands though their no of brands, brand.runs values are least.

Cluster2: Consumers in this cluster are preferring other other brands though their total volume is least.

Cluster3: Consumers in this cluster are less interested in buying the products when they are on promotional offers like discounts,coupons etc.

Cluster4: Consumers in this cluster are moderate in buying the products when there are other promotional offers.

2)Choosing the best segmentation

We have considered 3 criteria to choose K: 1)Minimum distance within cluster 2)Maximum distance between clusters 3)Information from centroid plot of clusters

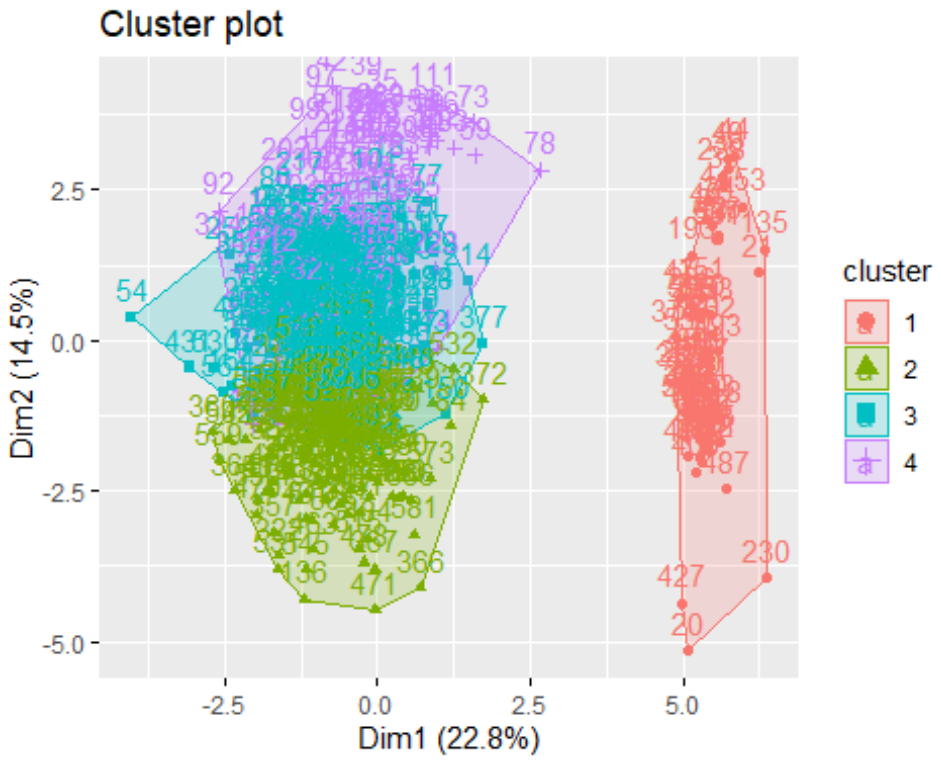
We observed that for K = 4 or 5, distance within cluster is minimum and distance between clusters is maximum. Since we are getting similar information at 4 with minimum distance within cluster and maximum distance between clusters, we conclude that K-means algorithm with K = 4 is the best model.

The Demographic result is computed for each cluster.We're simply attempting to interpret the demographic values of each cluster.

```
br<-BSoap[,23:30]
BSoap$Loyality<-as.numeric(apply(br,1,which.max))
Demo<-BSoap[,c(2:11,20:22,31:35,47)]
data<-as.data.frame(scale(Demo))
model2a<-kmeans(data,4,nstart=50)
```

Visualizing the clusters

```
fviz_cluster(model2a,data)
```



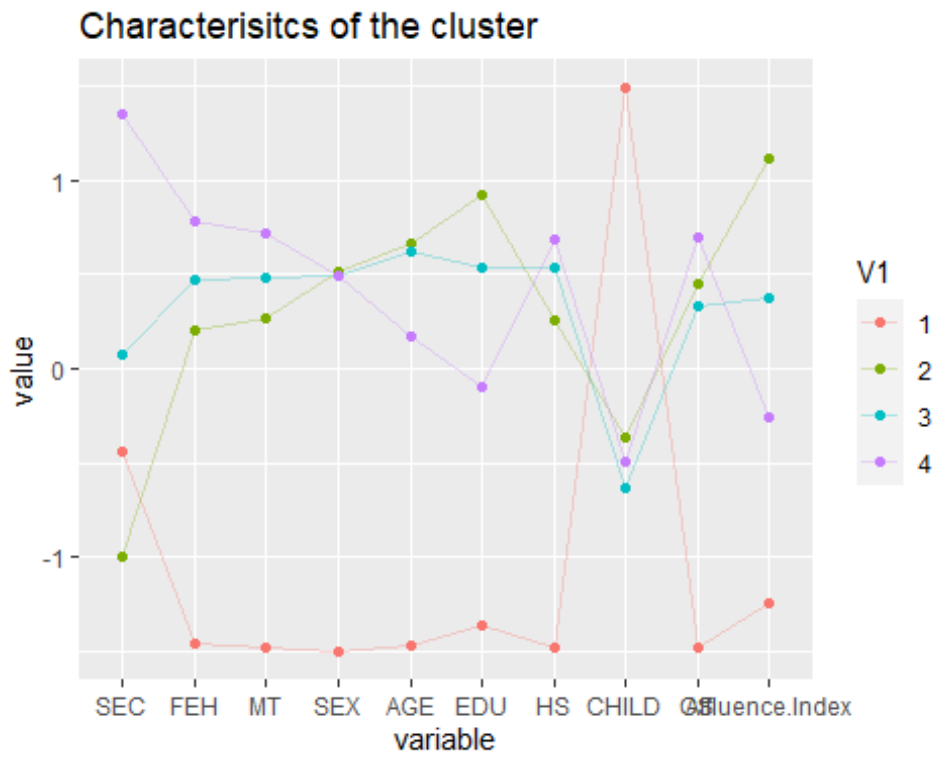

```
result2a<-as.data.frame(cbind(1:nrow(model2a$centers),model2a$centers))
result2a$V1<-as.factor(result2a$V1)
result2a
```

```
##   V1         SEC         FEH          MT        SEX         AGE         EDU
## 1  1 -0.26284751 -1.80475562 -1.90431150 -2.6805048 -0.58631729 -1.8462679
## 2  2 -0.64947785 -0.02721801  0.02887139  0.3586637  0.11625820  0.6704530
## 3  3  0.09294288  0.25345876  0.26304103  0.3356413  0.10013611  0.2395734
## 4  4  0.96680092  0.58218354  0.53424587  0.3333984 -0.04691692 -0.4515006
```

```
##             HS       CHILD         CS Affluence.Index Pur.Vol.No.Promo....
## 1 -1.822392359  1.45152536 -1.8362598     -1.49166356          -0.01935313
## 2  0.002611964 -0.06273596  0.2378116      0.80740525          -0.49313060
## 3  0.294046449 -0.28044565  0.1110297      0.08413614           0.33067546
## 4  0.454198276 -0.16183387  0.5109515     -0.53158740           0.03150557
##   Pur.Vol.Promo.6.. Pur.Vol.Other.Promo.. Others.999   Pr.Cat.1   Pr.Cat.2
## 1       -0.19016719            0.27950215 -0.1655857  0.3183424 -0.3552774
## 2        0.51365950            0.15612319  0.5142118  0.9950870 -0.4041570
## 3       -0.27613898           -0.19256288 -0.2235640 -0.4600110  0.8361828
## 4       -0.05803017            0.02073987 -0.1935783 -0.7072681 -1.0488348
##     Pr.Cat.3   Pr.Cat.4    Loyality
## 1  0.2108060 -0.1845976 -0.05529907
## 2 -0.4433352 -0.1820650  0.11271925
## 3 -0.2900938 -0.2792758 -0.02002257
## 4  1.2222027  1.0335161 -0.09656119

model2a$size

## [1]  68 172 250 110
```

Parallel plot to visualize the cluster.

```
ggparcoord(result2a,
           columns = 2:11, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



Customers in clusters 3 and 4 are having high SEC and are buying products irrespective of the promos and maintaining the loyalty to the product.

Customers with low SEC are falling in clusters 1 and 2 and are buying products when there is an promo offer and not maintaining the loyalty to the product.

Customers in Cluster 2 who are having high education are preferring other brands who are not the loyal ones to the brand.

Customers in Cluster 3 who are buying more products even when not on Promotional offers are the ones who are not preferring others brands who have to be considered as loyal customers

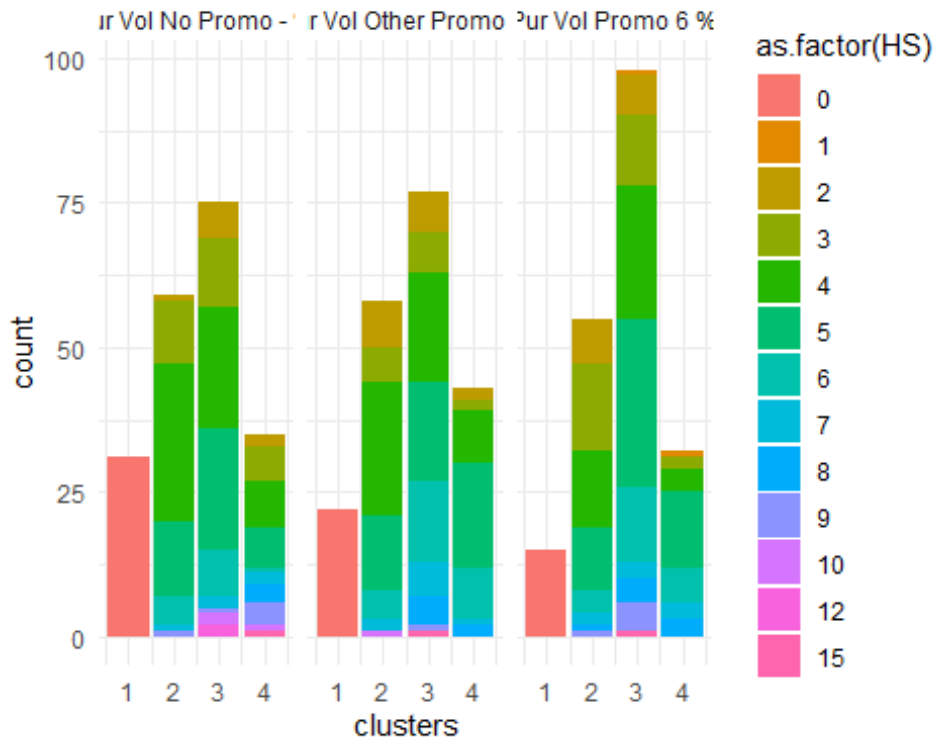
Customers in Cluster 1 who are having less education are buying more no of products when they are on various other promotions like coupons, discounts.

Customers in Cluster 4 with maximum household are less interested in buying the products when they are in a promotional offer of 6% and they tend to prefer less products belonging to other brands.

3)Develop a model that classifies the data into these segments.

```
ab<-BSoap[,c(2:11,20:22,31:35,47)]
ab$clusters <- model2a$cluster
max_price_cat <- as.factor(apply(ab[,c(15:18)],1,which.max))
ggplot(ab) +
  aes(x = clusters,fill= as.factor(HS)) +
```

```
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other
Promo %")))
```



In Cluster 3 the consumers buying in other promo and promo of 6% are higher than the ones purchasing in no promotional offers and also they are the ones who tending to buy more number of products from other brands and they are not brand loyal.

In Cluster 1 there are no people in house holds. In Cluster 2 the consumers are being loyal to the brand as they are buying less from other brands.

In Cluster4 as the customers are moderately brand loyal with High television availability, they can be targeted with advertisements.

By seeing the characteristics of the cluster line graph diagram, the people of high SEC must be given more promos to preserve their loyalty. People with low SEC must not be given promo codes because they are using their promo code and switching to other product who are providing promo code and not maintaining the loyalty.

Brand loyalty comes in the case when people have an option of exchange offers or coupons. Not many people care about price offs. Thus, in order to promote brand loyalty, manufacturers should promote their brands by gifting coupons or exchange offers.

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.