FinalProjectML

Kunal Sharma

07/05/2021

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 to design more effective customer reward systems and thereby increase brand loyalty.

Question 1. Use k-means clustering to identify 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: 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 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? Consider using a single derived variable.

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.)

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.

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

Reading And Cleaning the Data

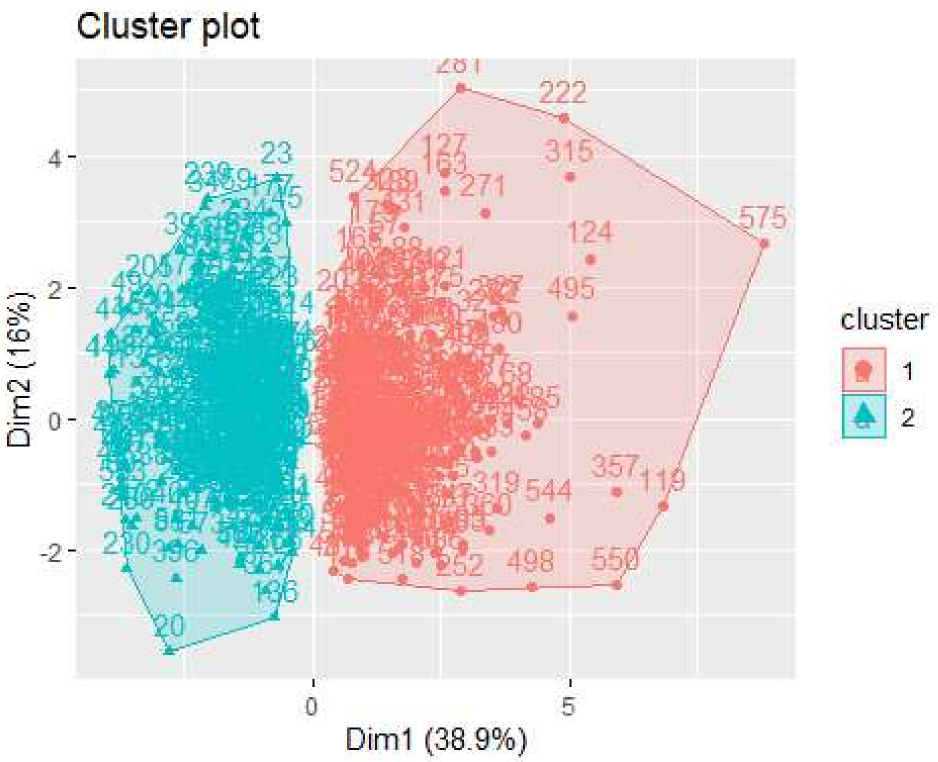
```
BathSoap<- read.csv("C://Users//admin//Downloads//BathSoap.csv")

BSData <- data.frame(sapply(BathSoap, function(x) as.numeric(gsub("%", "",
x))))
```

For computing brand loyalty, we used data from branded purchases based on the customer's purchase percentage on the Brand code, then found the highest brand loyal percentage and compared it to the other 999 brand purchases.

When a customer is loyal to a business, the Max Brand purchase percentage is higher than the Other Brand purchase percentage. As a consequence, the Customer's brand loyalty is created.

With k = 2, we use the kmeans clustering model to group the attributes that define brand loyalty into "Brand Loyal Customers" and "Not Brand Loyal Customers."

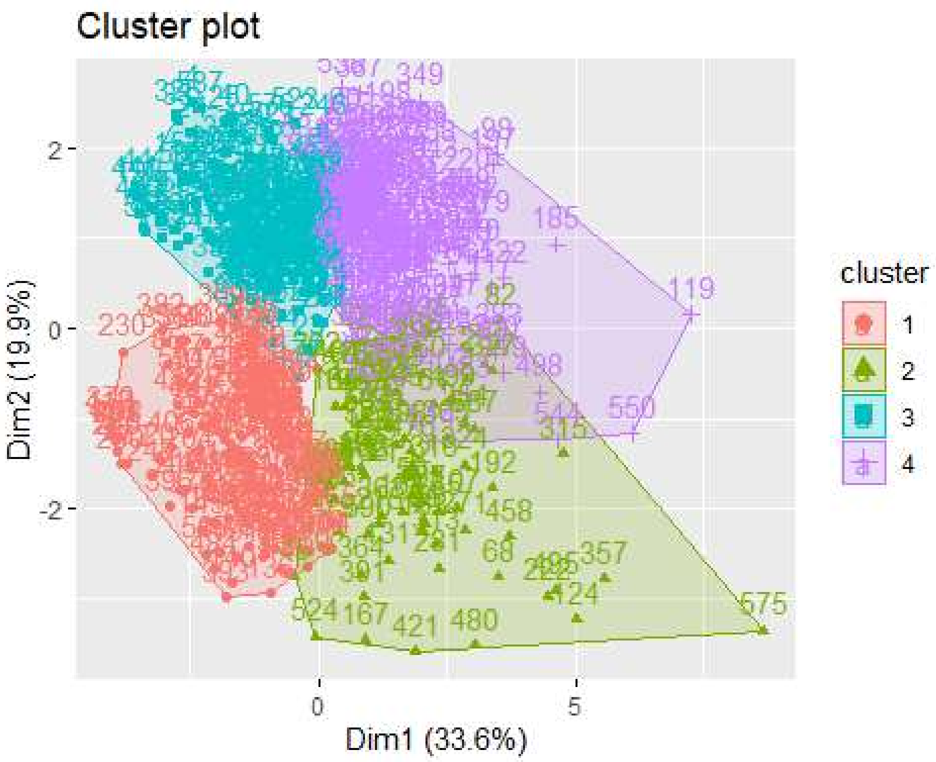
```
Loyal <- BSData[,23:31]

Loyal$MaxBrand <- apply(Loyal,1,max)

BathSoapBrandLoyalty <- cbind(BSData[,c(19, 13, 15, 12, 31, 14, 16,20)],
MaxLoyal = Loyal$MaxBrand)

BathSoapBrandLoyalty <- scale(BathSoapBrandLoyalty)

K_model_2 <- kmeans(BathSoapBrandLoyalty, centers = 2, nstart = 25)

BathSoapBrandLoyalty <- cbind(BathSoapBrandLoyalty, Cluster =
K_model_2$cluster)

fviz_cluster(K_model_2, data = BathSoapBrandLoyalty)
```



Customers in Cluster 1 are Brand Loyal, while customers in Cluster 2 are Brand Disloyal because they are unconcerned regarding products.

However, if we cluster them based on k = 4, we get the clusters shown below.

```
K_model_2 <- kmeans(BathSoapBrandLoyalty, centers = 4, nstart = 25)

BathSoapBrandLoyalty_4 <- cbind(BathSoapBrandLoyalty[,-10], Cluster =
K_model_2$cluster)

fviz_cluster(K_model_2, data = BathSoapBrandLoyalty_4)
```



Let’s take a look at the data for consumer purchase conduct.

Here, for Selling Proposition we’ve considered all of the selling propositions, selected the best of them, and compared them to show which are the most successful selling propositions to consider for the Model.

```
BathSoap_SP <- BSData[,36:46]

BathSoap_SP$Max <- apply(BathSoap_SP,1,max)
BathSoap_SP$MaxBrand <- colnames(BathSoap_SP)[apply(BathSoap_SP,1,which.max)]
```

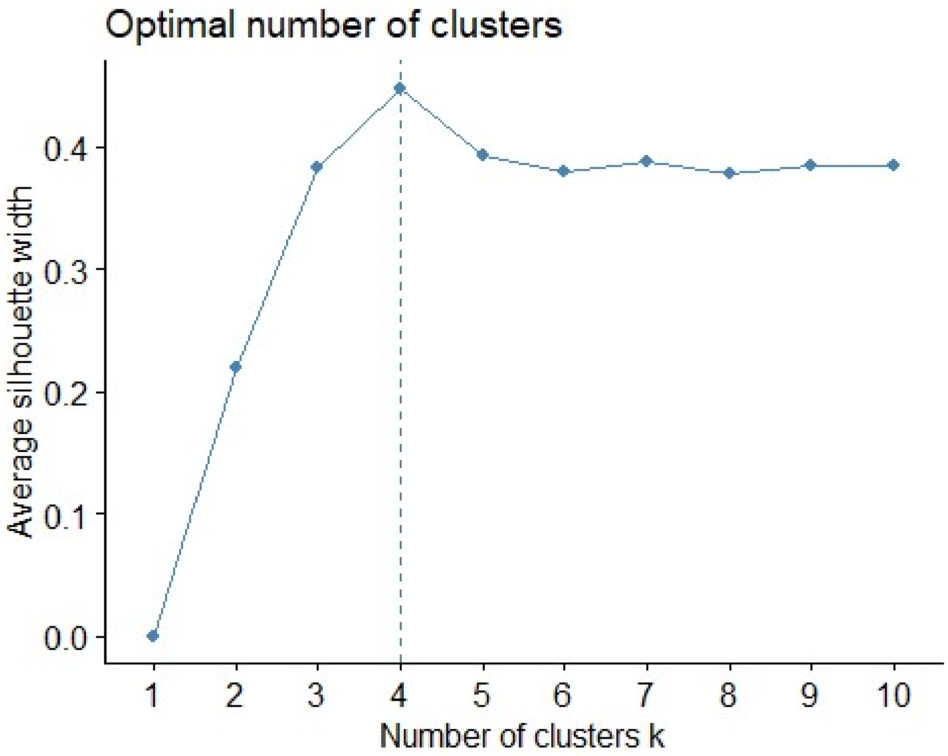
For the Price Catagories,catogories that are similar. Also the same can be said for promotions.

```
PriceCategory <- BSData[,32:35]
PriceCategory$Max <- apply(PriceCategory,1,max)
PriceCategory$MaxBrand <-
colnames(PriceCategory)[apply(PriceCategory,1,which.max)]

table(PriceCategory$MaxBrand)

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

Promo <- BSData[,20:22]
Promo$Max <- apply(Promo,1,max)
```

```
Promo$MaxBrand <- colnames(Promo)[apply(Promo,1,which.max)]

table(Promo$MaxBrand)

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

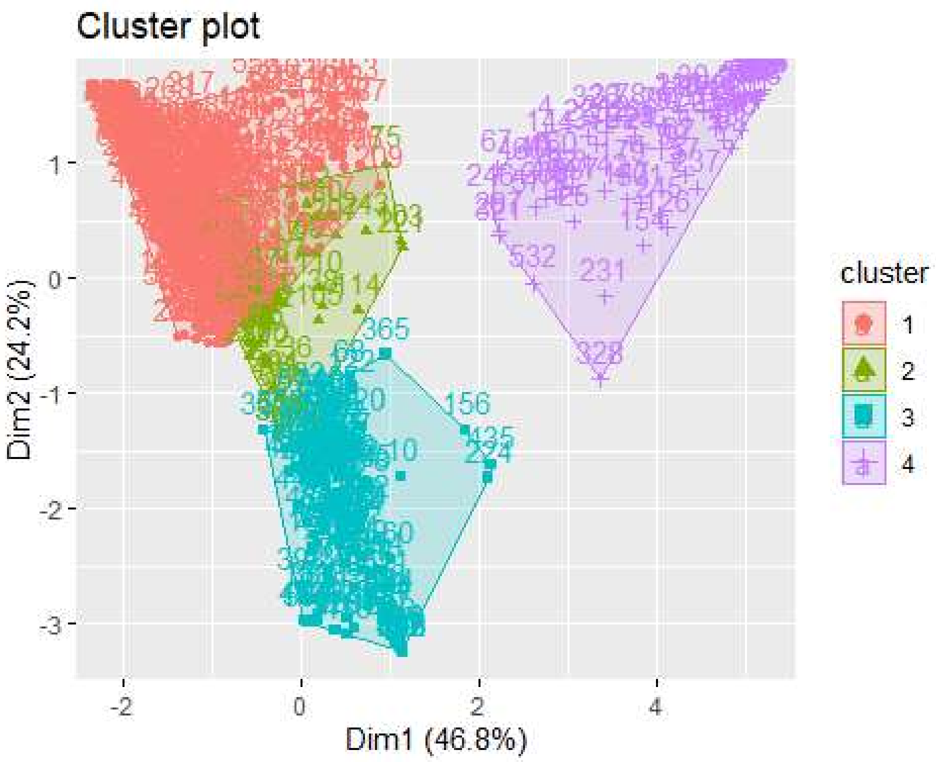
As a result, we've only considered the more powerful Selling Propositions when evaluating their effect. The same can be said for promotions and price categories.

```
CustomerPurchaseBehaviour <- BSData[,c(32,33,34,35,36,45)]
CustomerPurchaseBehaviour <- scale(CustomerPurchaseBehaviour)
#View(CustomerPurchaseBehaviour)

fviz_nbclust(CustomerPurchaseBehaviour, kmeans, method = "silhouette")
```



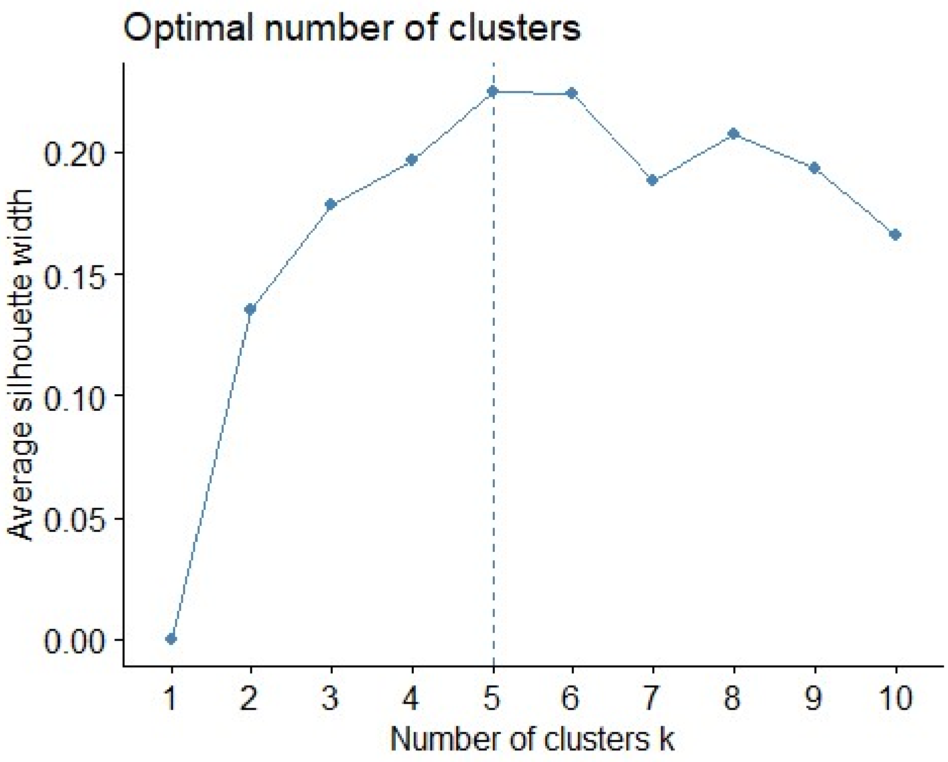
The K means Clustering model is computed in order to measure the customer's purchasing pattern. In this case, k = 4 will be used.

```
Purchase_K_model <- kmeans(CustomerPurchaseBehaviour, centers = 4, nstart =
25)

CustomerPurchaseBehaviour <- cbind(CustomerPurchaseBehaviour, Cluster =
Purchase_K_model$cluster)
#View(CustomerPurchaseBehaviour)

fviz_cluster(Purchase_K_model, data = CustomerPurchaseBehaviour)
```



Now we must understand the consumers’ brand loyalty as well as their purchasing behavior while developing a concept.

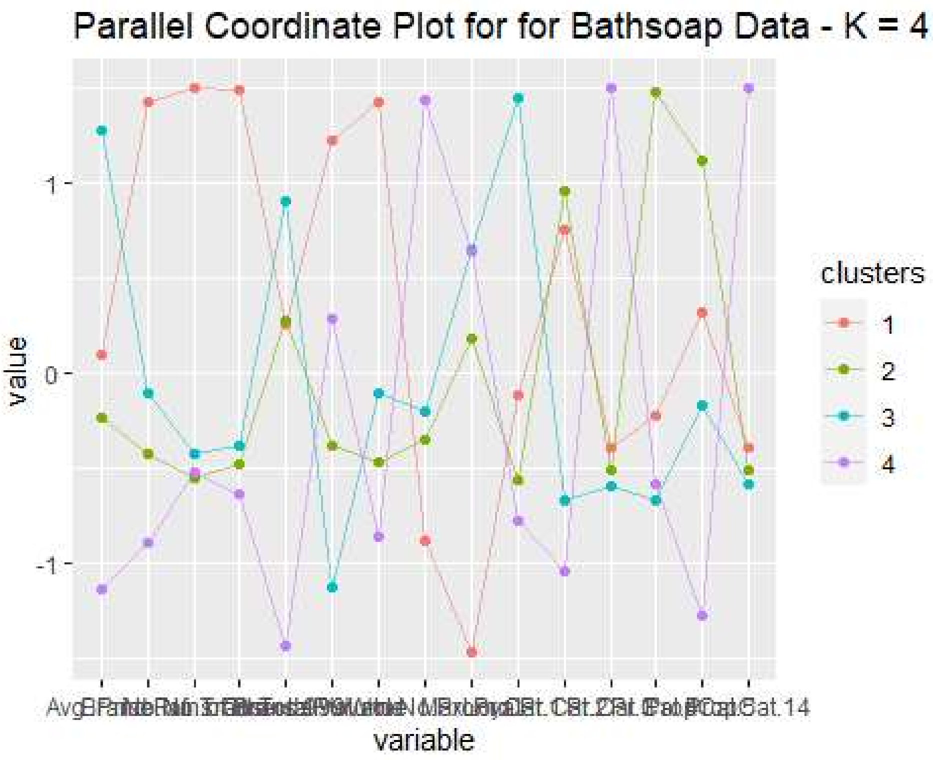
```
LoyalPurchase <- cbind(BathSoapBrandLoyalty[,-10],
CustomerPurchaseBehaviour[,-7])

fviz_nbclust(LoyalPurchase, kmeans, method = "silhouette")
```



```
K_Means_All <- kmeans(LoyalPurchase, centers = 4, nstart = 25)
```

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. As a result, we'll use k = 4 here.

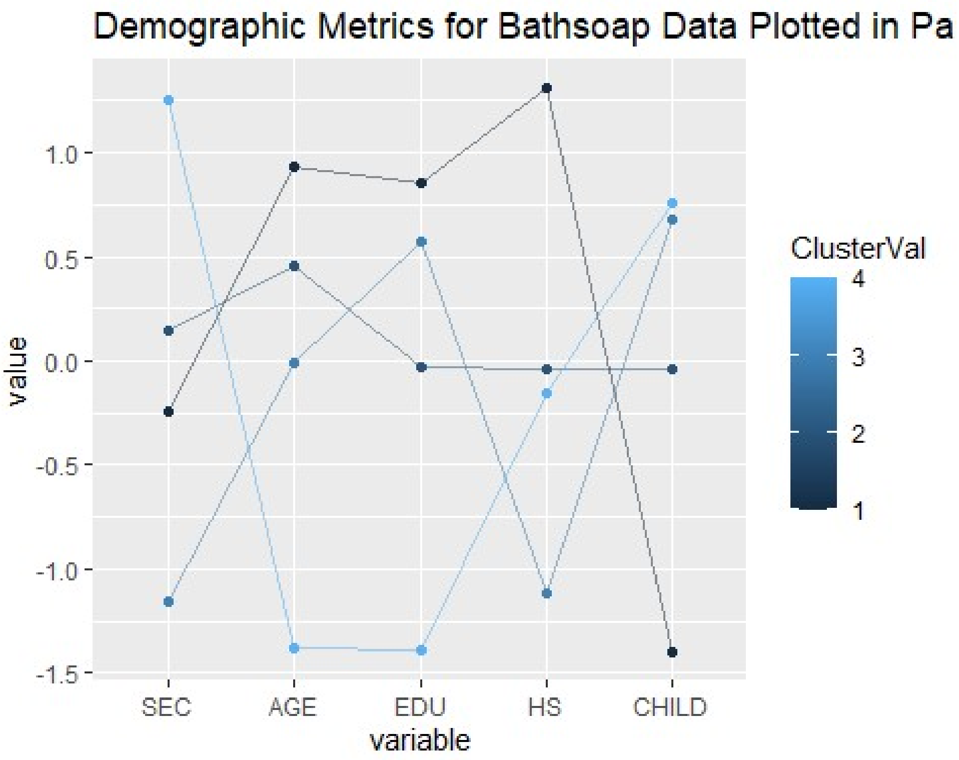
```
LoyalPurchase <- cbind(LoyalPurchase, Cluster =
as.data.frame(K_Means_All$cluster))
clusters <- matrix(c("1","2","3","4"),nrow = 4)
LoyalPurchase_Centroids <- cbind(clusters,as.data.frame(K_Means_All$centers))

ggparcoord(LoyalPurchase_Centroids,
           columns = 2:16, groupColumn = 1,
           showPoints = TRUE,
           title = "Parallel Coordinate Plot for for Bathsoap Data - K = 4",
           alphaLines = 0.5)
```



For each cluster, the Demographic result is computed.

We’re simply attempting to decipher the demographic values of each cluster.

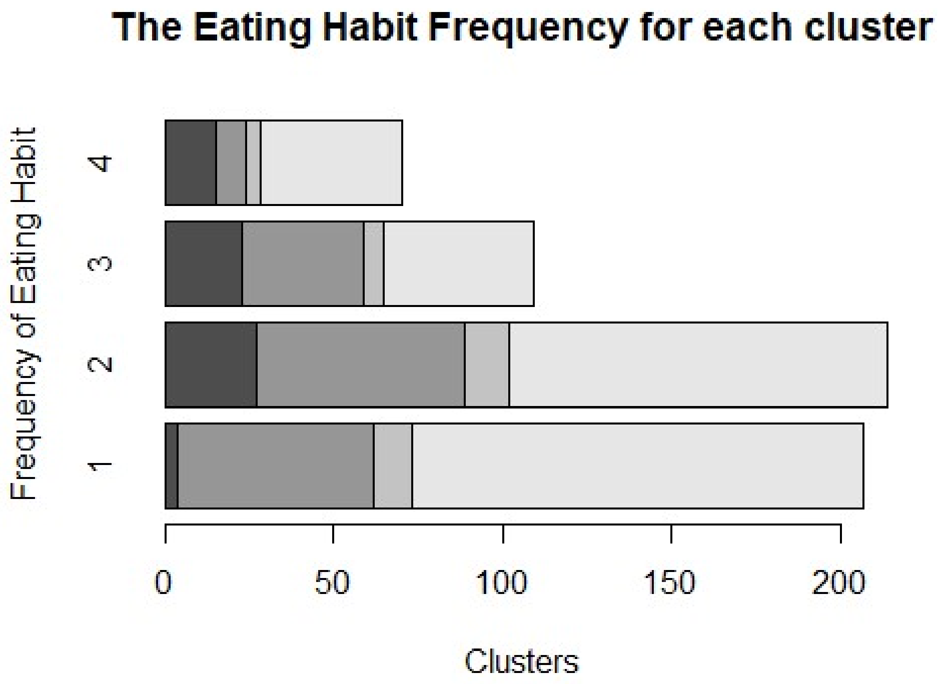
```
Demographics <-cbind(BSData[,2:11], ClusterVal = K_Means_All$cluster)

Centre_1 <- colMeans(Demographics[Demographics$ClusterVal == "1",])
Centre_2 <- colMeans(Demographics[Demographics$ClusterVal == "2",])
Centre_3 <- colMeans(Demographics[Demographics$ClusterVal == "3",])
Centre_4 <- colMeans(Demographics[Demographics$ClusterVal == "4",])

Centroid <- rbind(Centre_1, Centre_2, Centre_3, Centre_4)

ggparcoord(Centroid,
           columns = c(1,5,6,7,8), groupColumn = 11,
           showPoints = TRUE,
           title = "Demographic Metrics for Bathsoap Data Plotted in Parallel
Coordinate Plot- K = 4",
           alphaLines = 0.5)
```



We are presenting it in a barplot since there are a few attributes that are categorical.

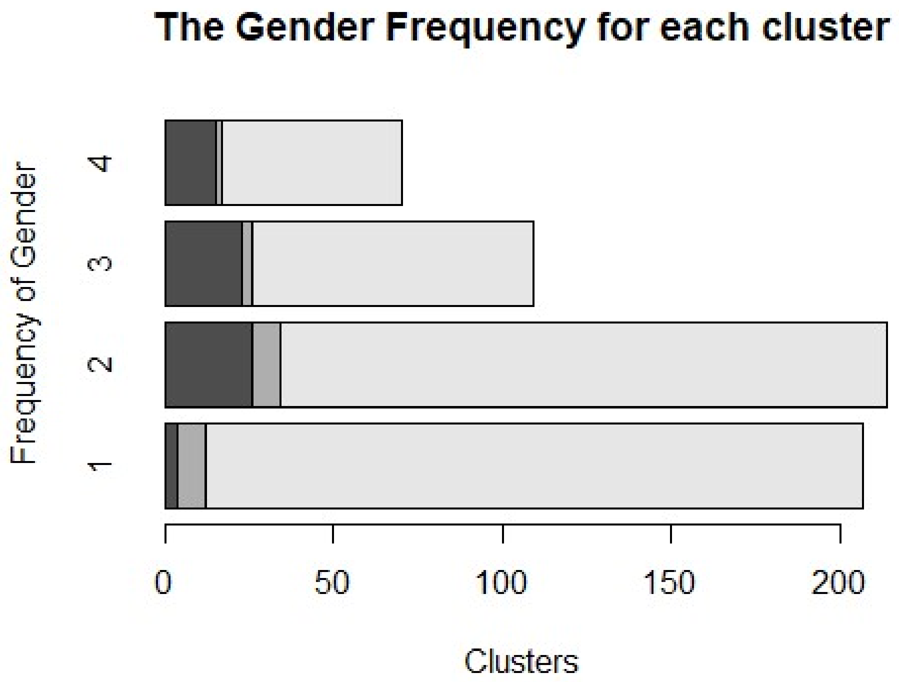
Plotting Eating Habit Frequency (Not Specified,Vegetarian Who Eats Eggs, Vegetarian,Non-Vegetarian):

```
barplot(table(BSData$FEH,K_Means_All$cluster), xlab = "Clusters", ylab =
"Frequency of Eating Habit", main = "The Eating Habit Frequency for each
cluster",horiz=TRUE)
```



Plotting the Frequency of Gender(NA, Male, Female):

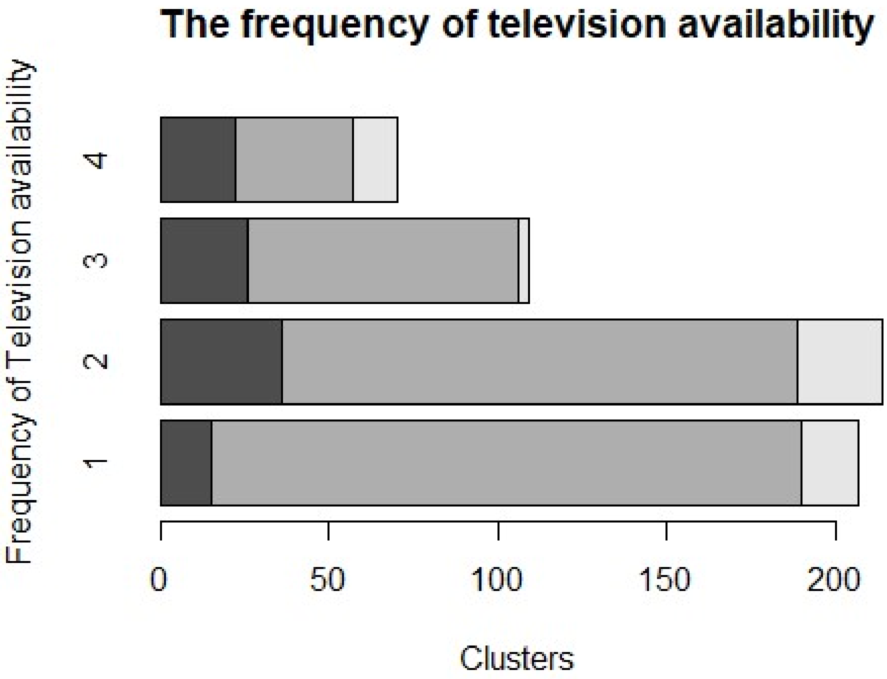
```
barplot(table(BSData$SEX,K_Means_All$cluster), xlab = "Clusters", ylab =
"Frequency of Gender", main = "The Gender Frequency for each
cluster",horiz=TRUE)
```



The female population has a higher purchasing rate, with females from clusters 1 and 2 having the most females.

Plotting the Television Availability Frequency (Unspecified, Availability, Not Available):

```
barplot(table(BSData$CS, K_Means_All$cluster), xlab = "Clusters", ylab =
"Frequency of Television availability", main = "The frequency of television
availability",horiz=TRUE)
```



Since almost all viewers have access to television, having a promotional offer on television can be effective in attracting customers.

Even, for those with codes 5 and 14, the Selling Proposition is strong. These are strong perpositions, Henec, and they can be used in the future.

Similarly, Price Catagory 1 and 2 have received positive feedback, so they can be used again in the future to attract customers' attention.

Customers in cluster 1 often have a higher educational level, implying that they have a decent job and that their email is reviewed often, allowing for the mailing of promotions.

Now, if we consider the remediation for a higher profit on the soaps to be sold, we can conclude that Cluster 3 customers are brand loyal. Thus, any promotional deals for branded soaps can be sent to Cluster 3 customers.

Similarly, Cluster 4 consumers are ignorant about promotional deals, but their sales are still high, so sending them a promotional email will not help us make a lot of money. Rather, the Cluster 1 customers who buy over the promotions are the most numerous, and they should be the mail priority.

Where the Average Price is higher, the profit range can be expanded. As a result, Cluster 3 consumers will concentrate on sending high-priced products to the mail for recommendations.