title: “Machine Learning Final Exam”

author: “Sumit Dutt Mishra”

date: “05/02/2021”

output:

word_document: default

html_document: default

pdf_document: default

###Business Situation ####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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ISLR)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl
/ve3WBa
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
set.seed(123)
```

###Using k-means clustering for identifying 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 #####Reading And Cleaning the Data

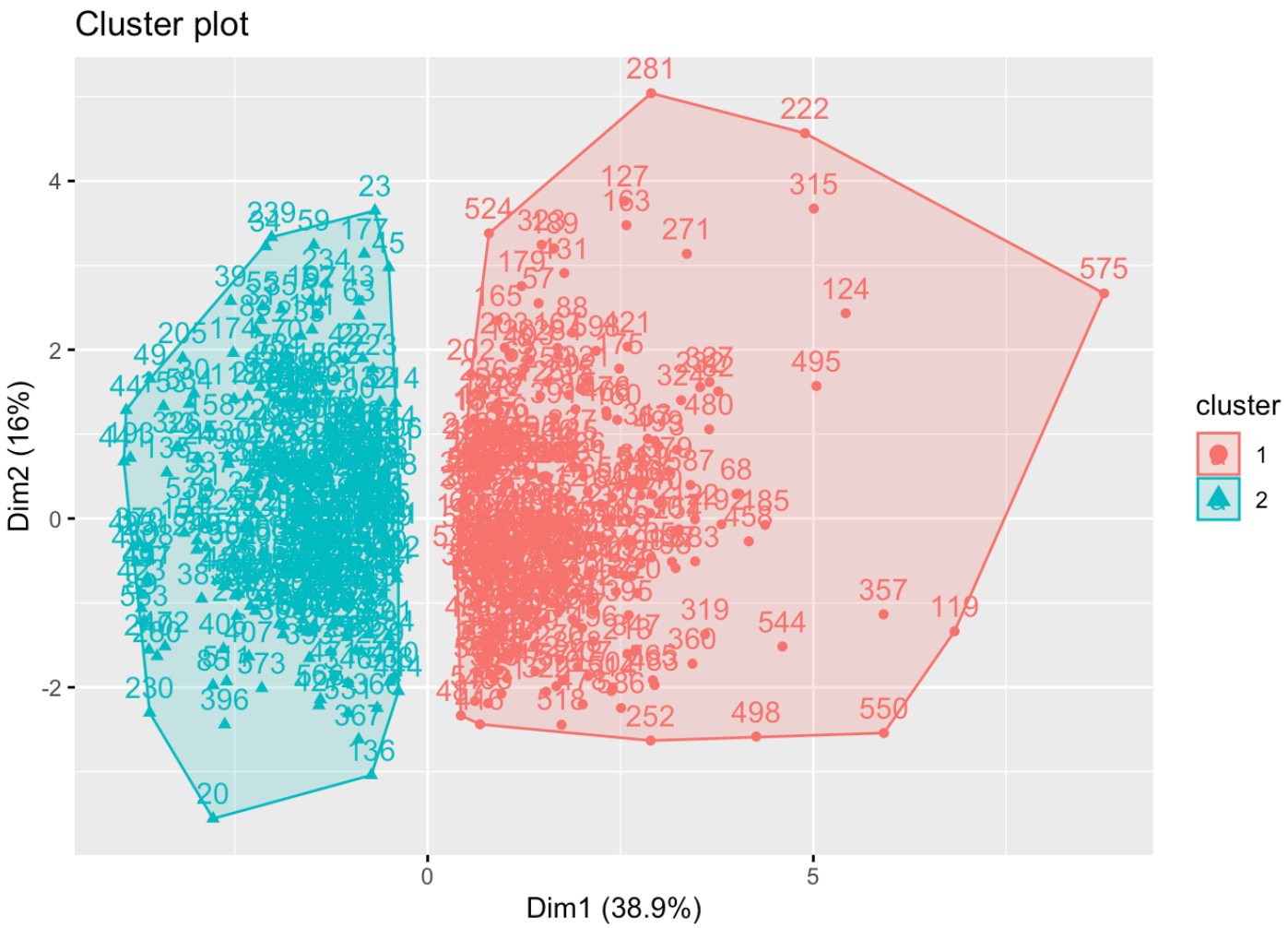
```
BathSoap <- read.csv("~/Downloads/BathSoap.csv")
BathsoapData <- data.frame(sapply(BathSoap, function(x) as.numeric(gsub("%", "", x
))))
```

####We gathered 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 to determine brand loyalty.

####The Max Brand purchase percentage is higher than the Other Brand purchase percentage when a customer is loyal to a company. As a result, the customer develops a sense of brand loyalty.

####We are using k-means clustering to group attributes into 2 sections:(i) Loyal Customers to Brand (ii) Disloyal Customers to Brand . For this we are taking k=2

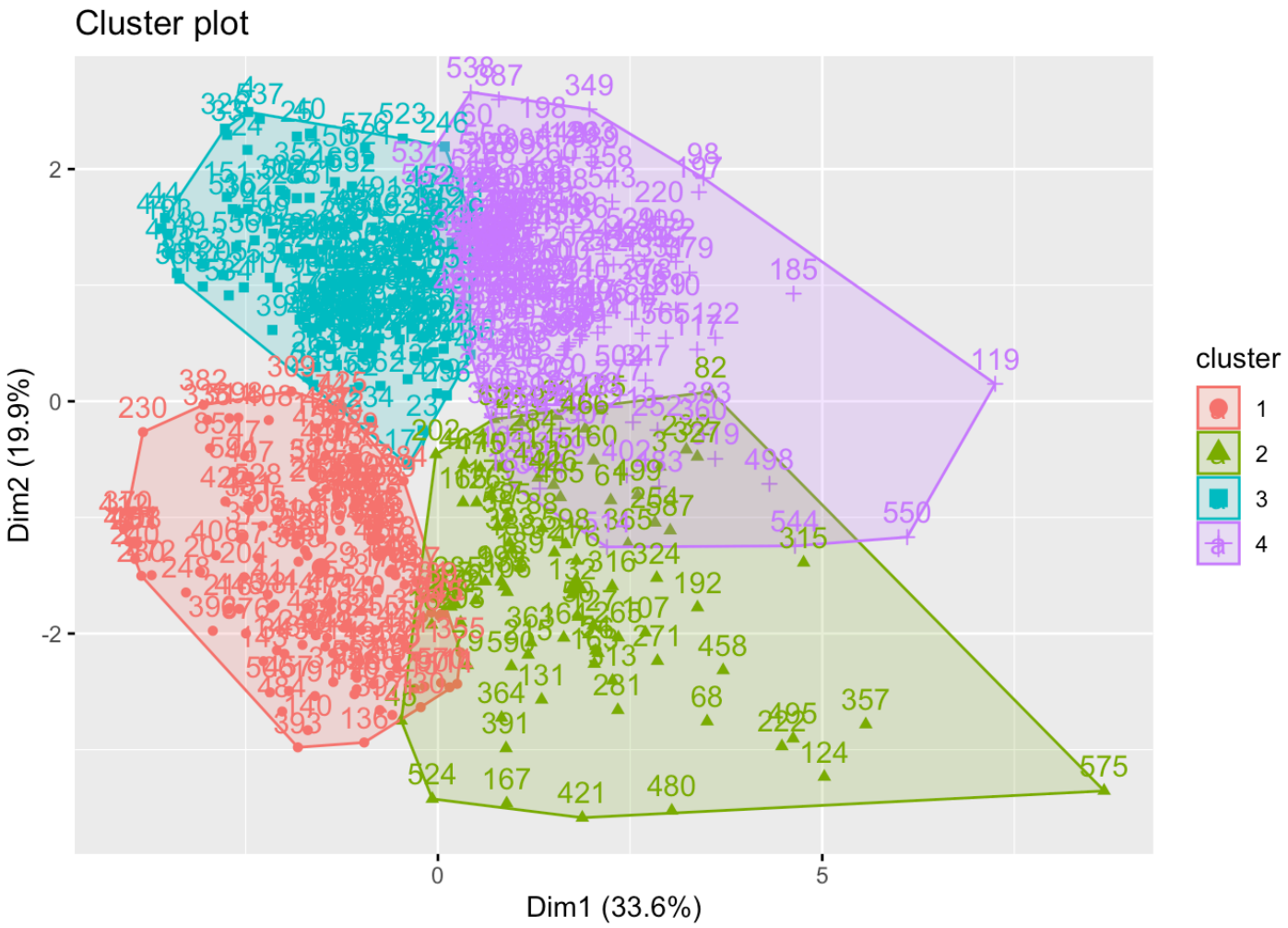
```
Loyalcustomers <- BathsoapData[,23:31]
Loyalcustomers$MaxBrand <- apply(Loyalcustomers,1,max)
BathSoapBrandLoyalty <- cbind(BathsoapData[,c(19, 13, 15, 12, 31, 14, 16,20)], Max
imumLoyal = Loyalcustomers$MaxBrand)
BathSoapBrandLoyalty <- scale(BathSoapBrandLoyalty)
Kmeans_model_1 <- kmeans(BathSoapBrandLoyalty, centers = 2, nstart = 25)
BathSoapBrandLoyalty <- cbind(BathSoapBrandLoyalty, Cluster = Kmeans_model_1$clust
er)
fviz_cluster(Kmeans_model_1, data = BathSoapBrandLoyalty)
```



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

####If we even cluster them further and take k=4

```
Kmeans_model_1 <- kmeans(BathSoapBrandLoyalty, centers = 4, nstart = 25)
BathSoapBrandLoyalty_4 <- cbind(BathSoapBrandLoyalty[,-10], Cluster = Kmeans_model
_1$cluster)
fviz_cluster(Kmeans_model_1, data = BathSoapBrandLoyalty_4)
```



#####Looking at the data for consumer purchase conduct.

####We've taken into account all of the selling propositions, chosen the best, and compared them to determine which are the most effective selling propositions to consider.

```
BathSoap_sellprep <- BathsoapData[,36:46]
BathSoap_sellprep$Max <- apply(BathSoap_sellprep,1,max)
BathSoap_sellprep$MaxBrand <- colnames(BathSoap_sellprep)[apply(BathSoap_sellprep,
1,which.max)]
```

#####Categories that are close to the Price Catagories. The same can be said for Promotions..

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

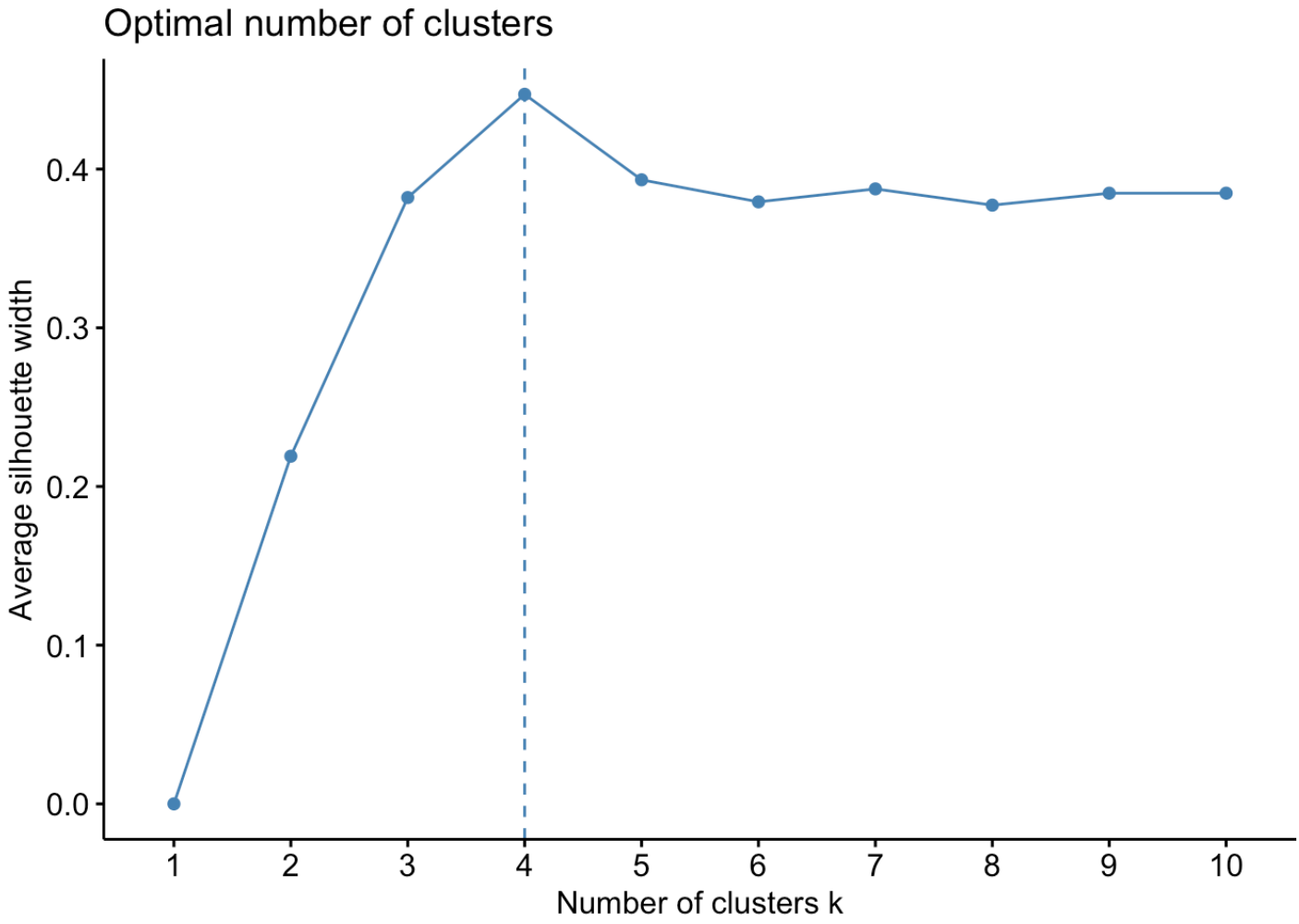
```
##
## Pr.Cat.1 Pr.Cat.2 Pr.Cat.3 Pr.Cat.4
##      132      343       78       47
```

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

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

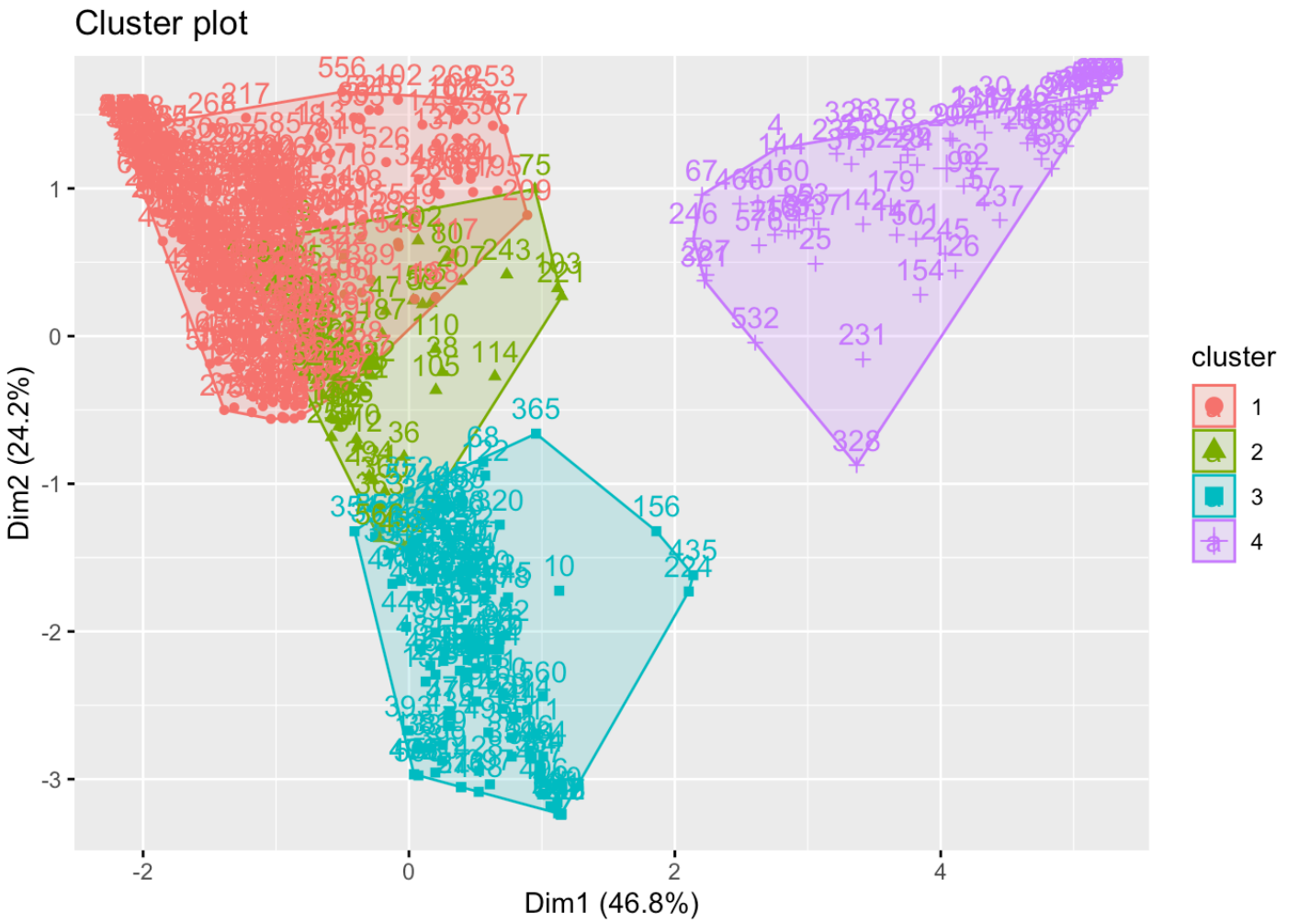
#####As a consequence, when assessing their effect, we've only looked at the more effective Selling Propositions. #####Promotions and price categories are in the same boat.

```
PurchBehaviour <- BathsoapData[,c(32,33,34,35,36,45)]
PurchBehaviour <- scale(PurchBehaviour)
#View(PurchBehaviour)
fviz_nbclust(PurchBehaviour, kmeans, method = "silhouette")
```



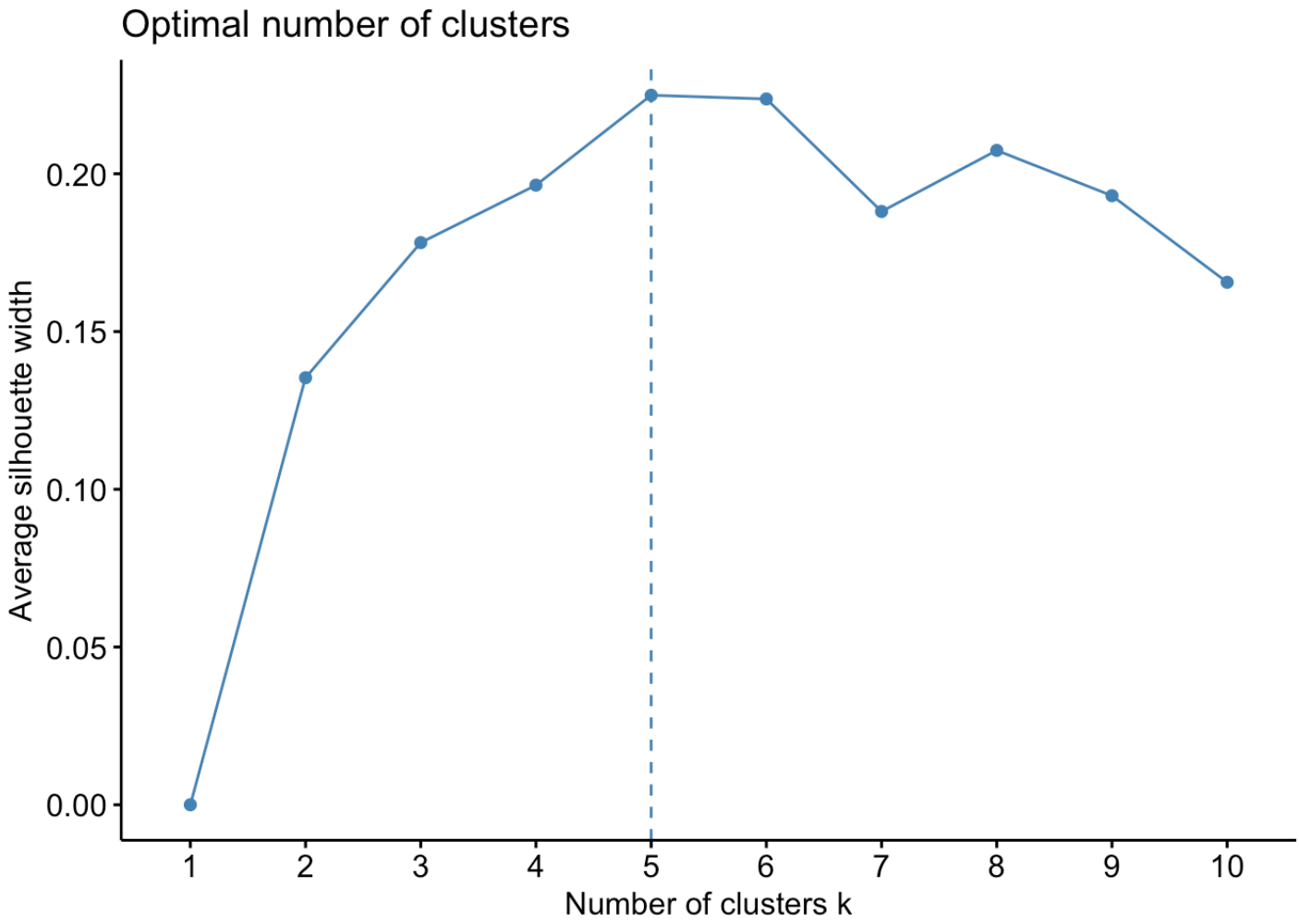
####To determine the customer's buying pattern, the K means Clustering model is used. k = 4 will be used in this situation.

```
Purch_model <- kmeans(PurchBehaviour, centers = 4, nstart = 25)
PurchBehaviour <- cbind(PurchBehaviour, Cluster = Purch_model$cluster)
#View(PurchBehaviour)
fviz_cluster(Purch_model, data = PurchBehaviour)
```



####When creating a definition, we must now consider the customers' brand loyalty as well as their buying behaviour.

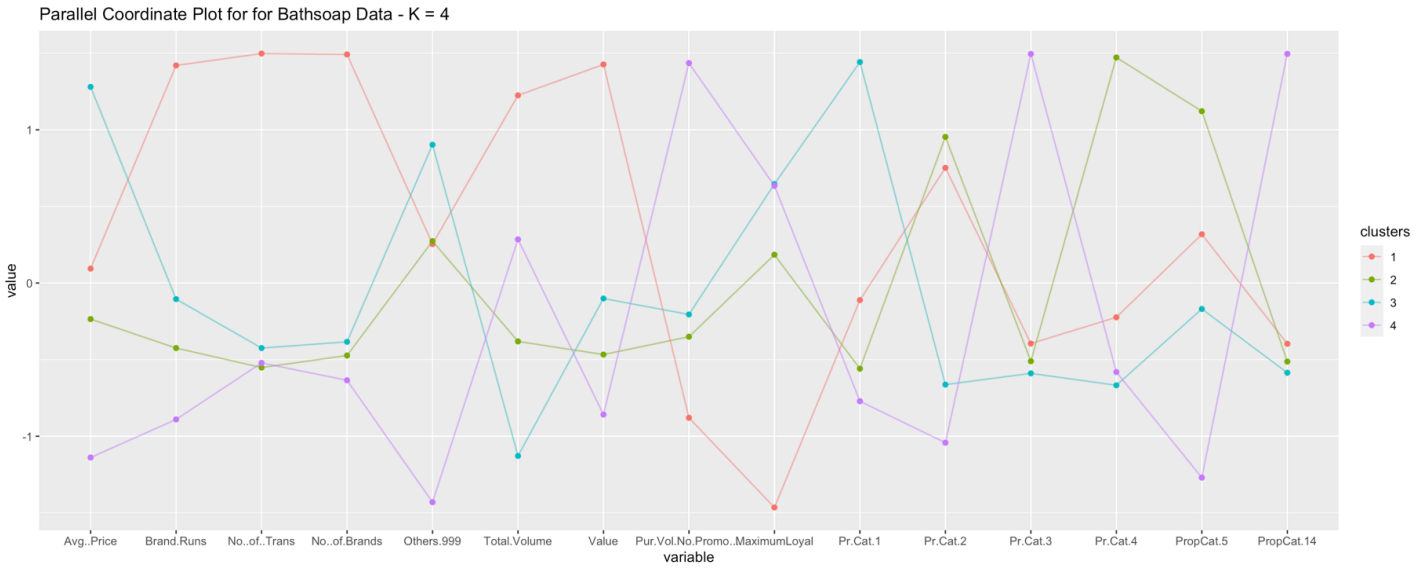
```
LoyalPurch <- cbind(BathSoapBrandLoyalty[,-10], PurchBehaviour[,-7])
fviz_nbclust(LoyalPurch, kmeans, method = "silhouette")
```




```
KMeans_All <- kmeans(LoyalPurch, centers = 4, nstart = 25)
```

####While plotting the model for k = 4 and k = 5, it is clearly visible that the aspects can be resolved by using 4 clusters without using 5. For this reason, we will use k=4 ###Selecting the best segmentation and commenting on the characteristics and Developing a model that classifies the data into these segments.

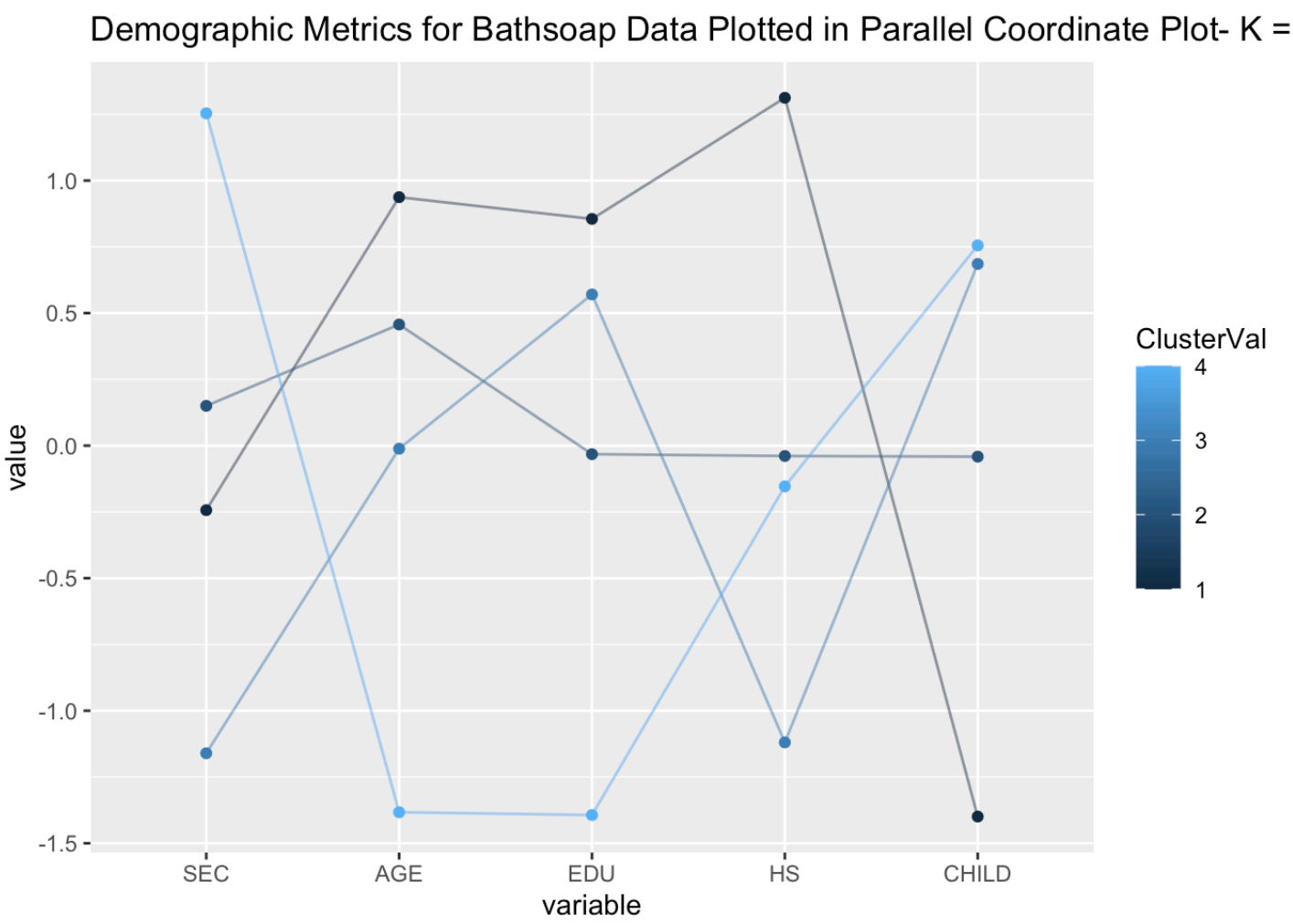
```
LoyalPurch<- cbind(LoyalPurch, Cluster = as.data.frame(KMeans_All$cluster))
clusters <- matrix(c("1","2","3","4"),nrow = 4)
LoyalPurch_Centroid <- cbind(clusters,as.data.frame(KMeans_All$centers))
ggparcoord(LoyalPurch_Centroid,
           columns = 2:16, groupColumn = 1,
           showPoints = TRUE,
           title = "Parallel Coordinate Plot for for Bathsoap Data - K = 4",
           alphaLines = 0.5)
```



####The Demographic result is computed for each cluster

Converting the demographic values of each cluster.

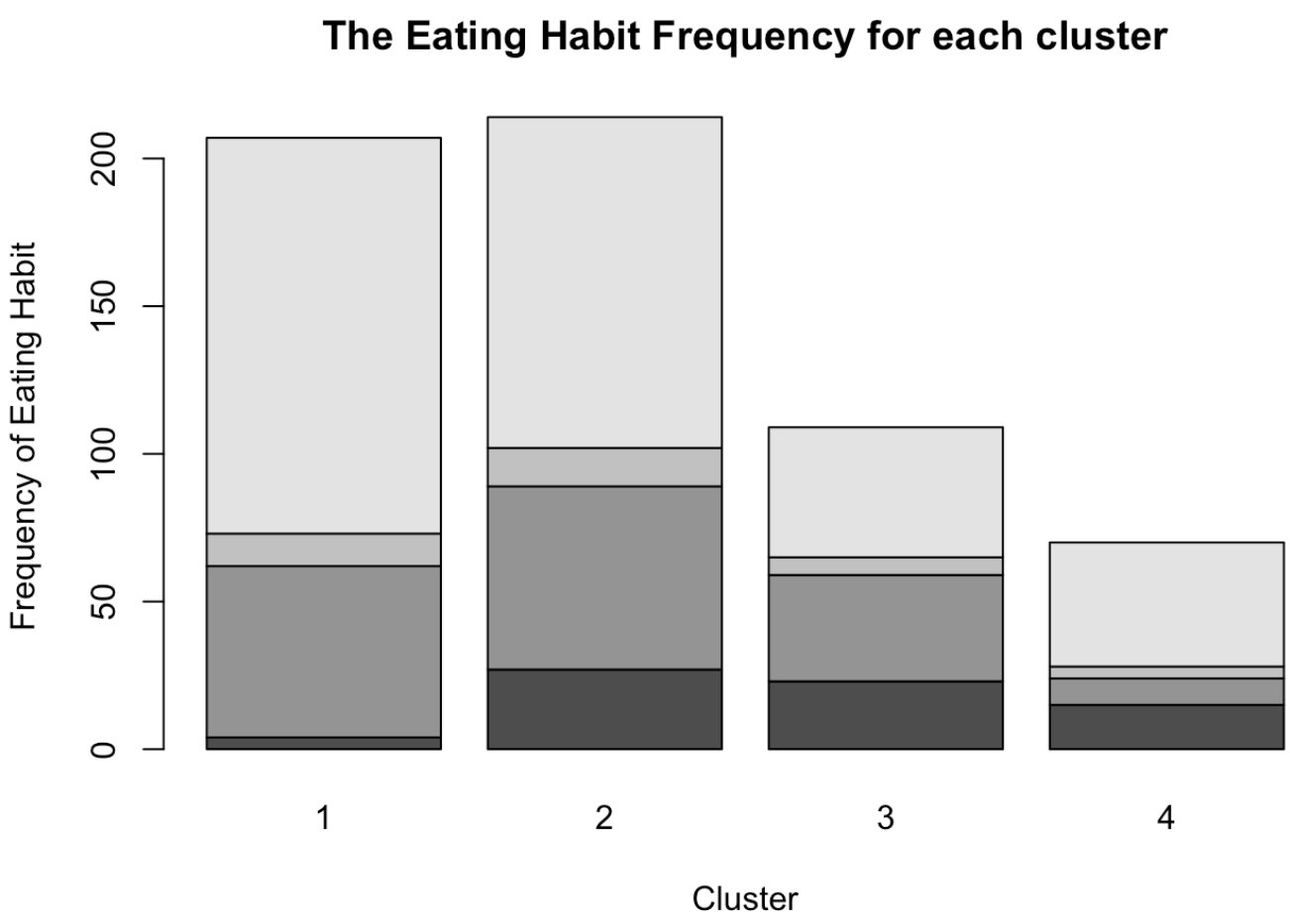
```
Demograph <-cbind(BathsoapData[,2:11], ClusterVal = KMeans_All$cluster)
Center_1 <- colMeans(Demograph[Demograph$ClusterVal == "1",])
Center_2 <- colMeans(Demograph[Demograph$ClusterVal == "2",])
Center_3 <- colMeans(Demograph[Demograph$ClusterVal == "3",])
Center_4 <- colMeans(Demograph[Demograph$ClusterVal == "4",])
Centroids <- rbind(Center_1, Center_2, Center_3, Center_4)
ggparcoord(Centroids,
           columns = c(1,5,6,7,8), groupColumn = 11,
           showPoints = TRUE,
           title = "Demographic Metrics for Bathsoap Data Plotted in Parallel Coor
dinate Plot- K = 4",
           alphaLines = 0.5)
```



####We are using a barplot because there are a few attributes that are categorical.

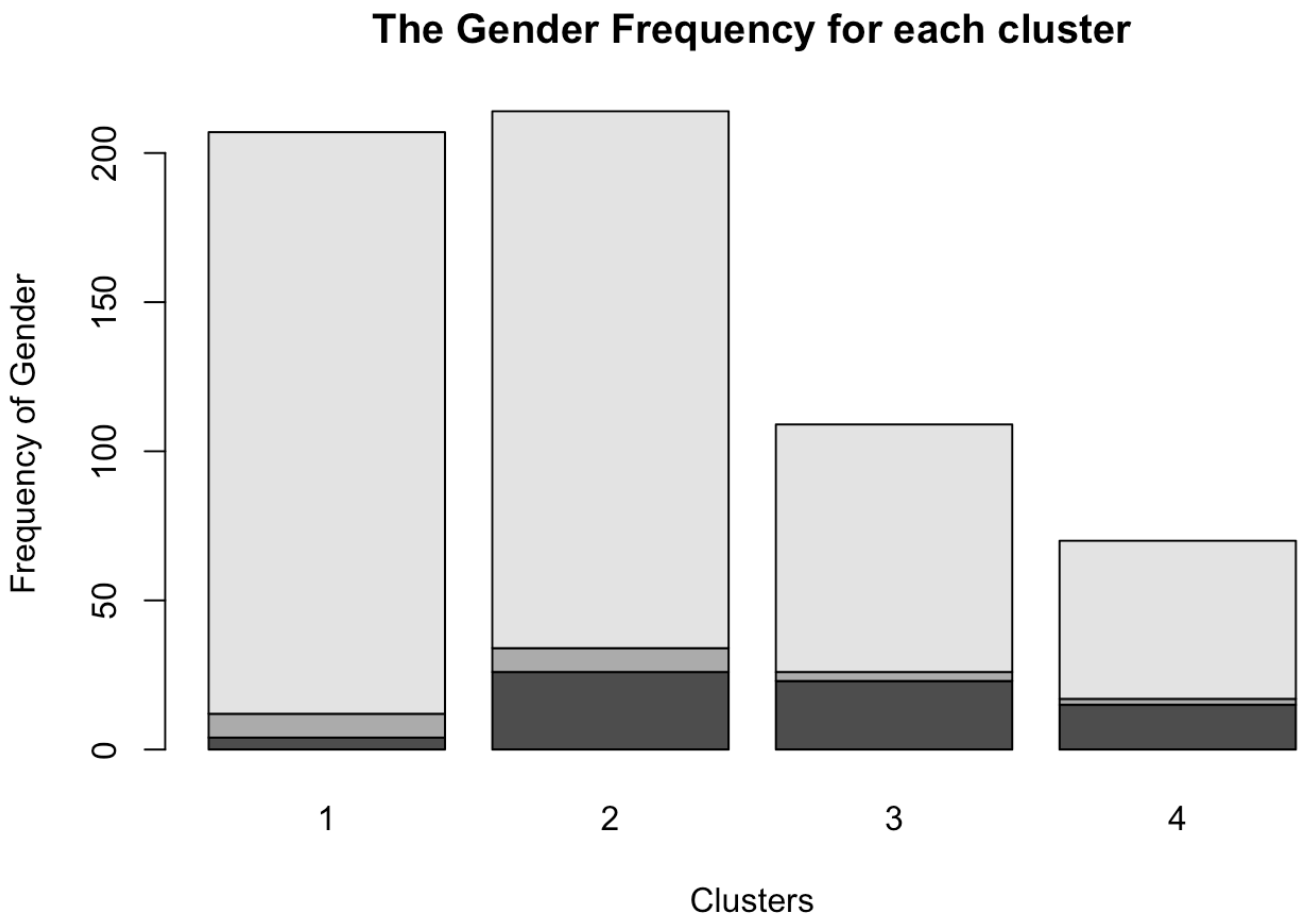
####Plotting Eating Habit Frequency (Not Specified,Vegetarian Who Eats Eggs, Vegetarian,Non-Vegetarian):

```
barplot(table(BathsoapData$FEH,KMeans_All$cluster), xlab = "Cluster", ylab = "Freq
uency of Eating Habit", main = "The Eating Habit Frequency for each cluster")
```



####Plotting the Frequency of Gender like Male, Female and NA:

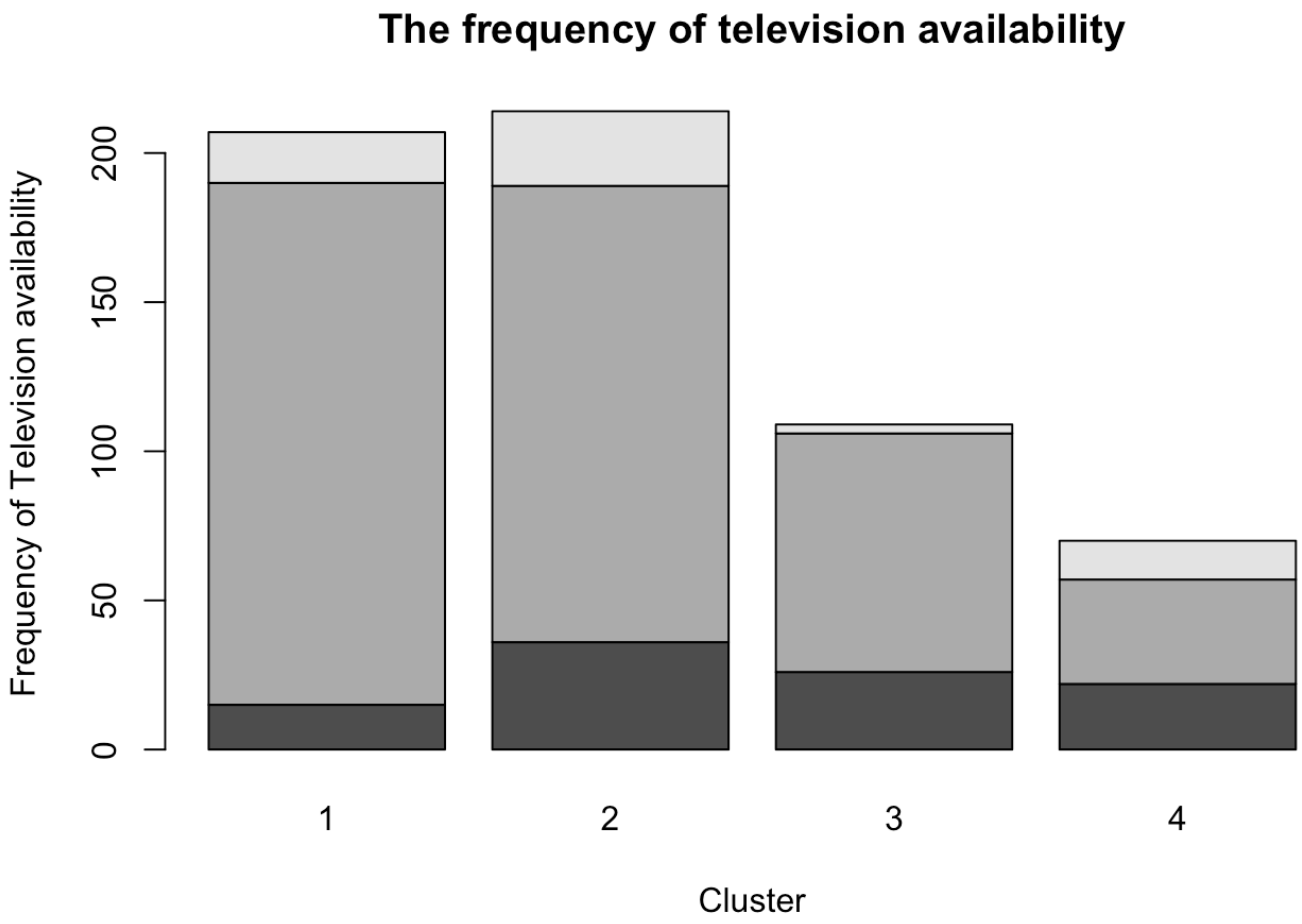
```
barplot(table(BathsoapData$SEX,KMeans_All$cluster), xlab = "Clusters", ylab = "Fre
quency of Gender", main = "The Gender Frequency for each cluster")
```



####The female population has a higher purchasing rate, as females in clusters 1 and 2 have the most females.

####Plotting the Television Availability Frequency as Availability, Unavailability and Unspecified:

```
barplot(table(BathsoapData$CS, KMeans_All$cluster), xlab = "Cluster", ylab = "Freq
uency of Television availability", main = "The frequency of television availabilit
y")
```



####Since almost everybody has access to television, a television advertising offer may be successful in attracting customers. ####The selling proposition is high also for those with codes 5 and 14 . Hence, these are good propositions that we can use in the future. ####Price Categoru 1 and 2 have both received psoitive reviews , so they can be used to draw customers interest in the future. ####We may assume that cluster 3 customers are brand loyal if we consider the remediation for a higher profit on the soaps to be sold . As a result , any advertising soap promotional offers can be sent to Cluster 3 customers. ####Cluster 4 customers are similarly unaware of promotional offers, but their purchases remain high , so sending them a promotional email would not help us get higher revenue. Instead, cluster 1 customers who purchase over promotions are the most frequent , and they should received priority mails. ####The benefit spectrum can be extended if the average price is higher. As a result, customers in Cluster 3 will focus on submitting high priced items to the mail for recommendations.