Final Exam

Noorah

5/8/2021

Problem: 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.

Questions:

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.

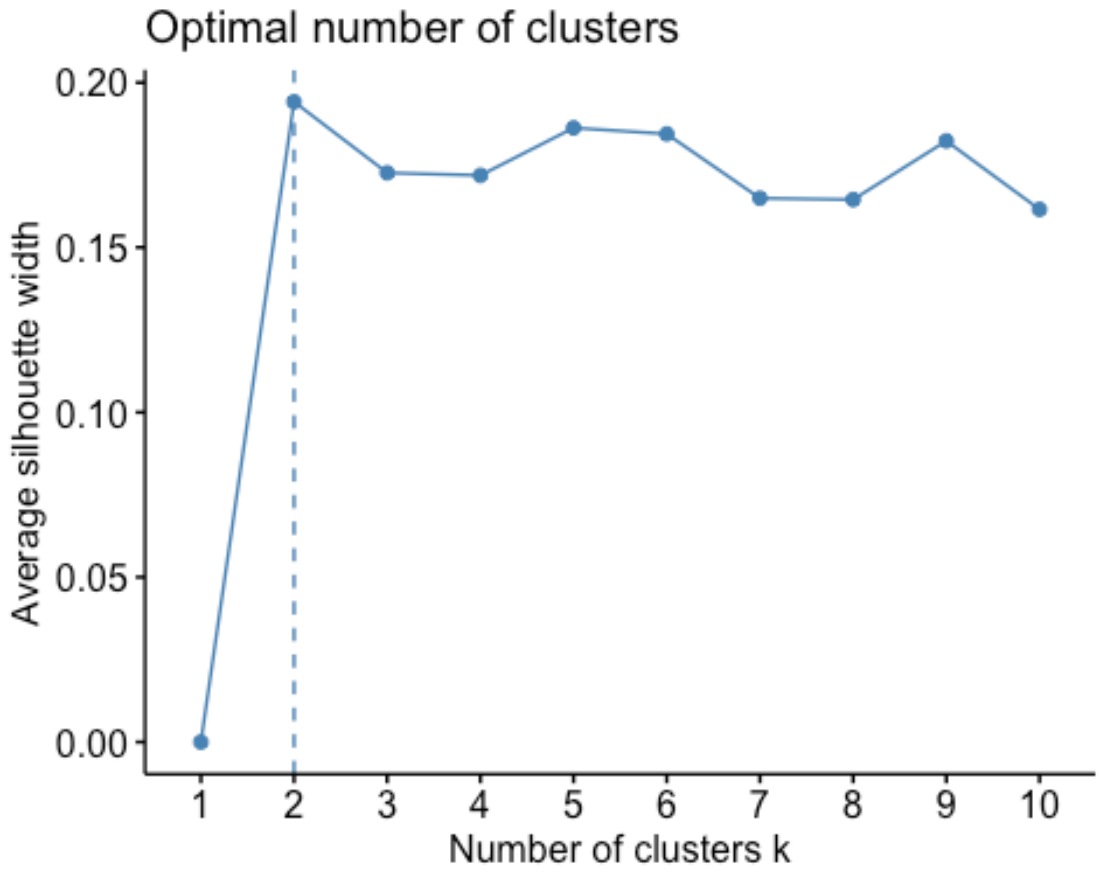
Methodology and Analysis:

```
###1st, I would import the data and clean the data
f<-read.csv("~/Desktop/MSBA-spring 2021/ML/Final/BathSoap.csv")
bs <- data.frame(sapply(f, function(x) as.numeric(gsub("%", "", x))))
```

```
###2nd, installing packages
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

library(readr)
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ tibble  3.1.1     ✓ stringr 1.4.0
## ✓ tidyr   1.1.3     ✓ forcats 0.5.1
## ✓ purrr   0.3.4

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

library(readxl)
library(FactoMineR)
set.seed(111)
```

When customer is loyal,the max precentage rate is higher in comparision with other brand purchase precentage, then we would have a loyal customer :)

```
###3rd, Changing binary variables to factor variables to find purchase behavior
bv <- bs[,23:31]
bv$MaxBrand <- apply(bv,1,max)
bsbinary <- cbind(bs[,c(19, 13, 15, 12, 31, 14, 16,20)], bsm = bv$MaxBrand)
bsbinary <- scale(bsbinary)
fviz_nbclust(bsbinary, kmeans, method = "silhouette")
```

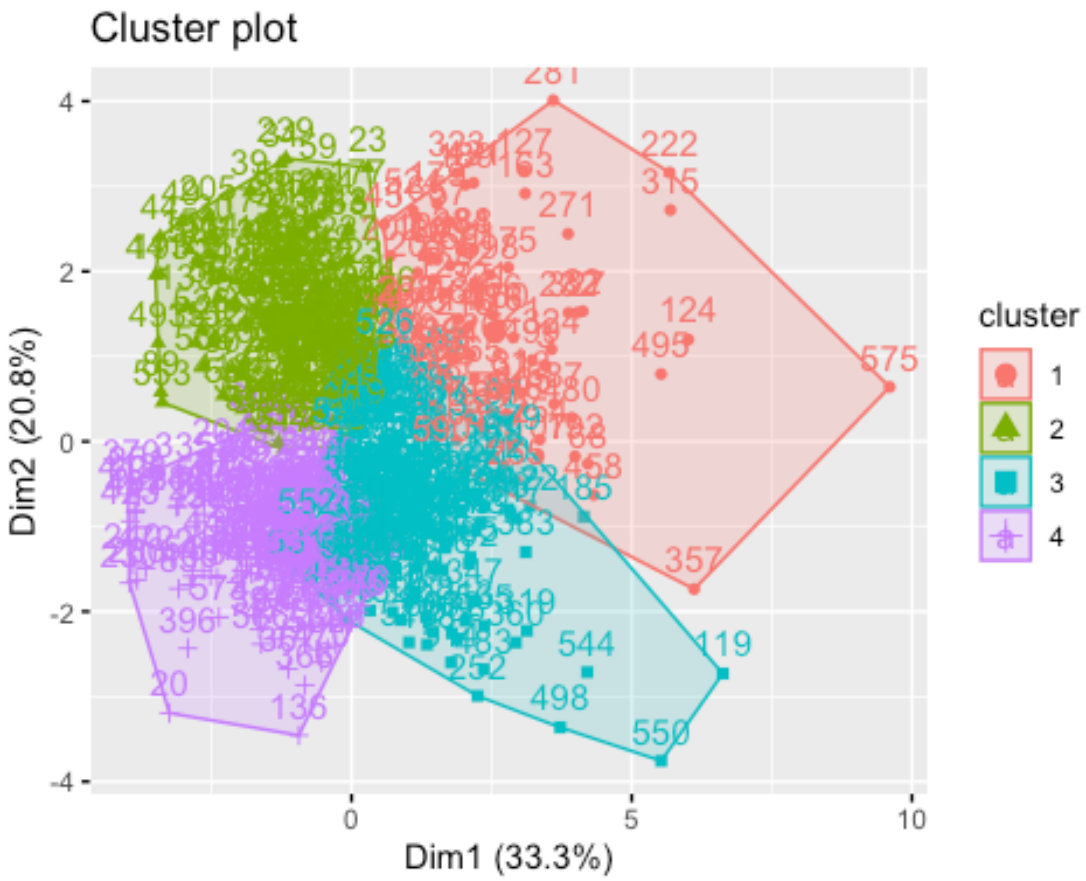


form the graph above we can see that k=2 will help to cluster our data into 2 clusters. cluster 1 = not loyal and cluster 2= loyal.

```
###4th, using k=2 to cluster the data
k2 <- kmeans(bsbinary, centers = 2, nstart = 25)
bsbinary <- cbind(bsbinary, Cluster = k2$cluster)
fviz_cluster(k2, data = bsbinary)
```

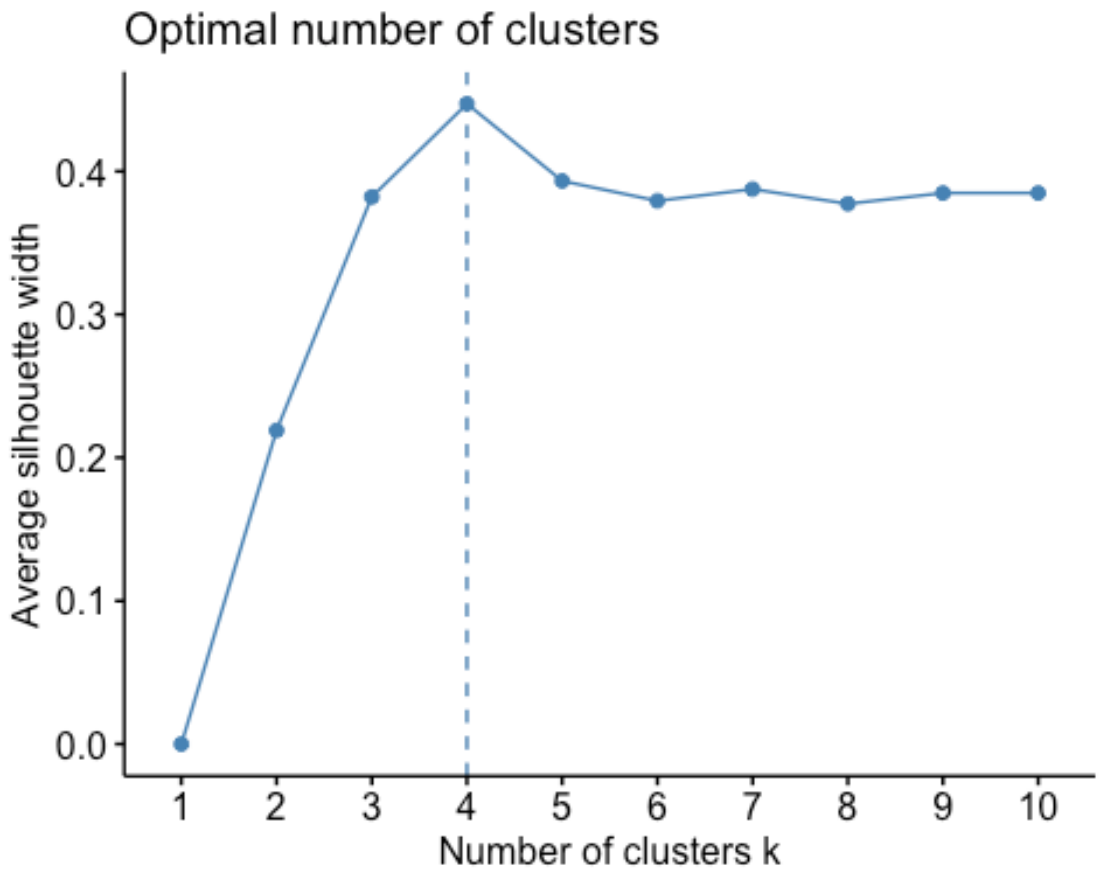
As shown from the graph, we have two clusters 1 and 2. cluster 1= not a loyal customer or we can say not intersted customer cluster 2= loyal customer

```
###using 4 clusters
k4 <- kmeans(bsbinary, centers = 4, nstart = 25)
bsbinary4 <- cbind(bsbinary[,-10], Cluster = k4$cluster)
fviz_cluster(k4, data = bsbinary4)
```

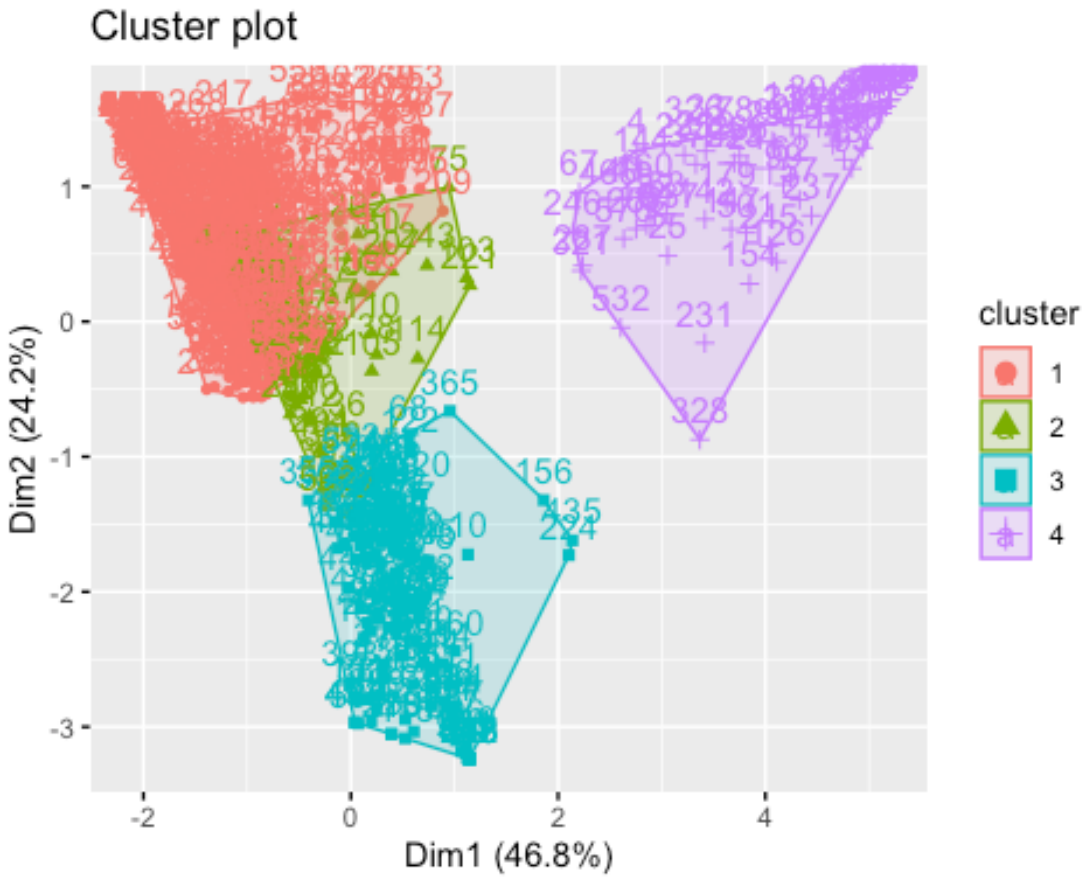


From the graph we see, I will use k=4

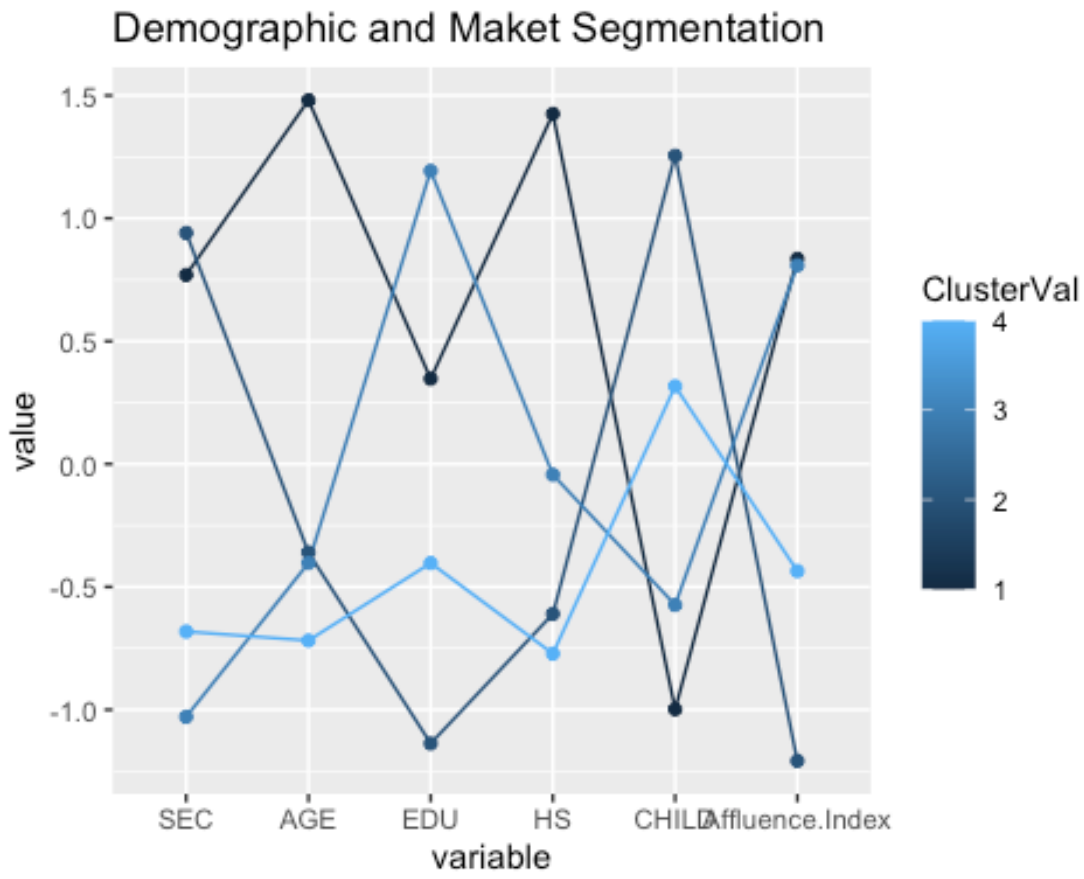
```
customerpb <- bs[,c(32,33,34,35,36,45)]
customerpb <- scale(customerpb)
fviz_nbclust(customerpb, kmeans, method = "silhouette")
```




```
###5th, finding customer behavior and basis purchase
cbp <- kmeans(customerpb, centers = 4, nstart = 25)
customerpb <- cbind(customerpb, Cluster = cbp$cluster)
fviz_cluster(cbp, data = customerpb)
```

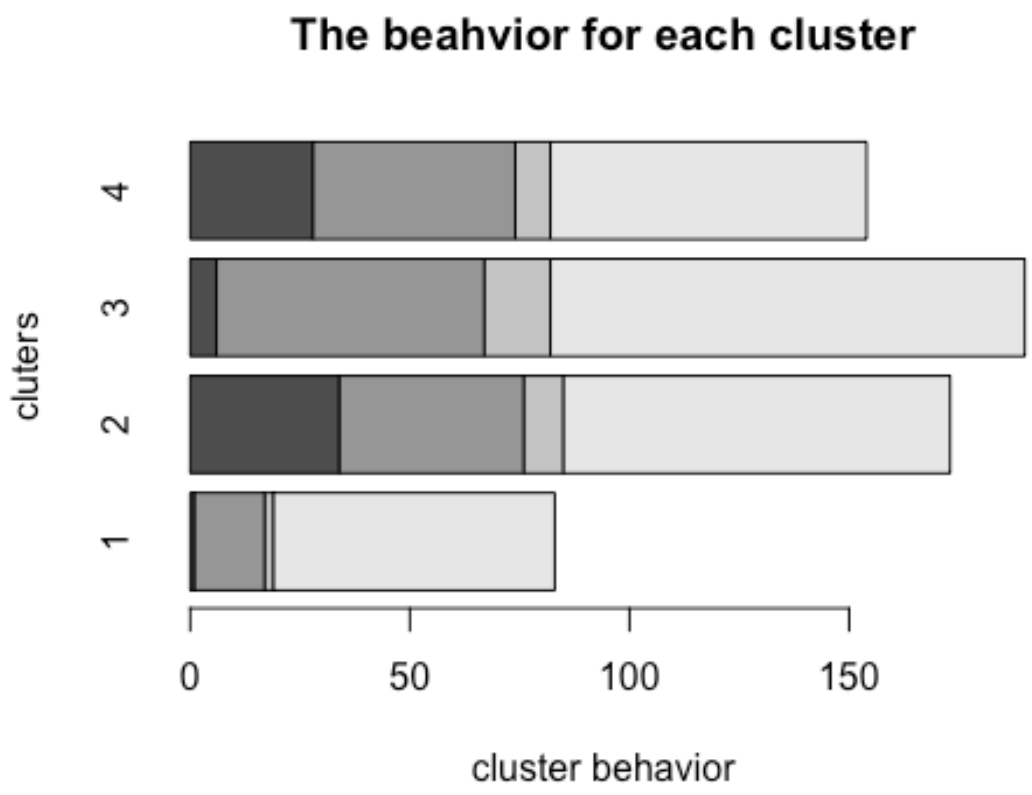



```
##2. Select what you think is the best segmentation and comment on the
characteristics (demographic, brand loyalty, and basis for purchase) of these
clusters. (This information would be used to guide the development of
advertising and promotional campaigns.)
Demographics <-cbind(bs[,2:11], ClusterVal = k4$cluster)
Centre_1 <- colMeans(Demographics[Demographics$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(1,5,6,7,8,10), groupColumn = 11,
           showPoints = TRUE,
           title = "Demographic and Maket Segmentation",
           alphaLines = 1)
```

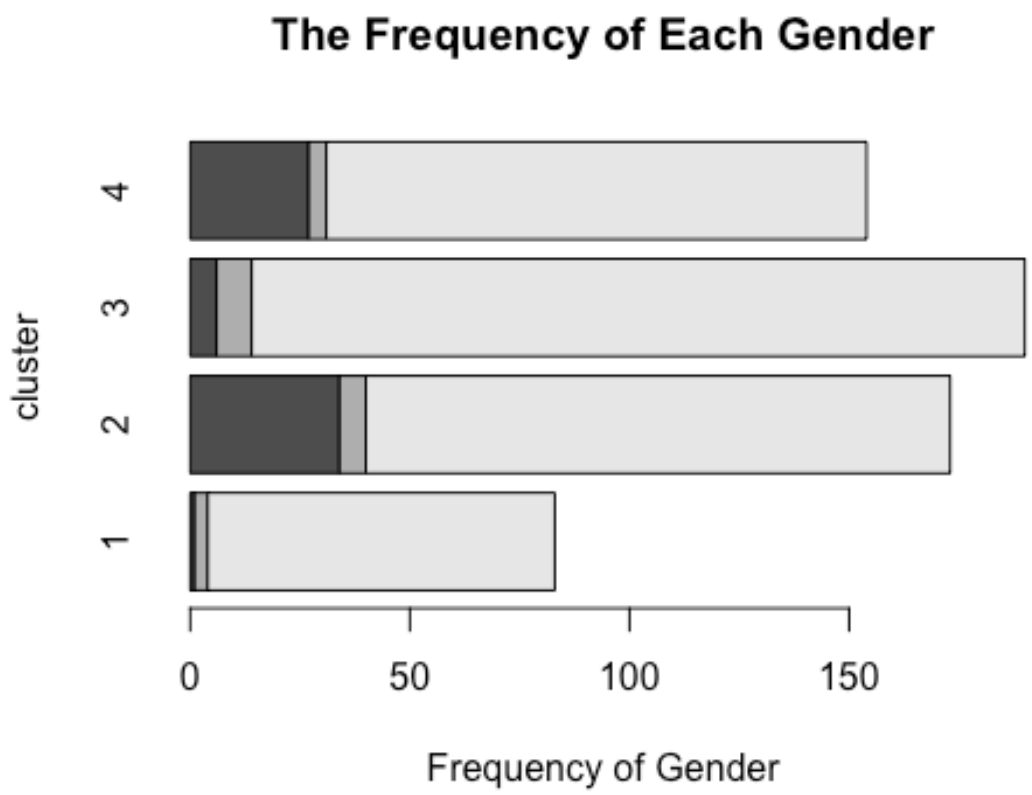


From the graph, we have find the demographic valuse for each cluster

```
###using bar pplot to find each cluster behavior
barplot(table(bs$FEH,k4$cluster), ylab = "cluters", xlab = "cluster
behavior", main = "The beahvior for each cluster",horiz=TRUE)
```




```
###using bar plot find each cluster gender frequency
barplot(table(bs$SEX,k4$cluster), ylab = "cluster", xlab = "Frequency of
Gender", main = "The Frequency of Each Gender",horiz=TRUE)
```



conclusion

As we can understand from the graphs created, we can see that most customers have access to television and this means marketing and attracting customers is more effective.

-Customers in cluster 1 have the highest education level in comparing with other clusters, that means they have more access to mail promotions

-Customers in cluster 3 have the higher profit and this means that they are loyal customers and they have more promotions

-Customers in cluster 4 have high consuming level but they don't care about promotions so that means cluster 3 are the most cluster to focus the promotions on