Final Project

Machine Learning

Consumer Market Segmentation of Bath Soap

Prerak Kalpeshkumar Patel

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.

Key Problems

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.

Measuring Brand Loyalty

Several variables in this case measure aspects of brand loyalty. The number of different brands purchased by the customer is one measure of loyalty. However, a consumer who purchases one or two brands in quick succession, then settles on a third for a long streak, is different from a consumer who constantly switches back and forth among three brands. Therefore, how often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands—a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands.

Assignment

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.

Importing Libraries and Data

- library(readr)
- library(tidyverse)
- library(dplyr)
- library(readxl)
- library(FactoMineR)
- library(factoextra)
- library(Hmisc)

bs.data <- read_csv("BathSoap.csv")

Data PreProcessing

```
# Converting Binary variables from numeric to factor(i.e. Binary variables)
bs.data$SEC <- factor(bs.data$SEC)
bs.data$FEH <- factor(bs.data$FEH)
bs.data$MT <- factor(bs.data$MT)
bs.data$SEX <- factor(bs.data$SEX)
bs.data$AGE <- factor(bs.data$AGE)
bs.data$EDU <- factor(bs.data$EDU)
bs.data$HS <- factor(bs.data$HS)
bs.data$CHILD <- factor(bs.data$CHILD)
bs.data$CS <- factor(bs.data$CS)
bs.data$`Affluence Index` <- factor(bs.data$`Affluence Index`)

# Converting distinct number variables from numeric to integer
bs.data$`No. of Brands`<- as.integer(bs.data$`No. of Brands`)
bs.data$`Brand Runs`<- as.integer(bs.data$`Brand Runs`)
bs.data$`Total Volume`<- as.integer(bs.data$`Total Volume`)
bs.data$`No. of  Trans`<- as.integer(bs.data$`No. of  Trans`)

# Converting percentages in character to floating numericals
bs.data$`Pur Vol No Promo - %`<-  as.numeric(str_replace(bs.data$`Pur Vol No
Promo - %`,"%",""))/100
bs.data$`Pur Vol Promo 6 %`<-  as.numeric(str_replace(bs.data$`Pur Vol Promo
6 %`,"%",""))/100
bs.data$`Pur Vol Other Promo %`<-  as.numeric(str_replace(bs.data$`Pur Vol Ot
her Promo %`,"%",""))/100

bs.data$`Br. Cd. 24` <-  as.numeric(str_replace(bs.data$`Br. Cd. 24`,"%",""))
/100
bs.data$`Br. Cd. 57, 144`<-  as.numeric(str_replace(bs.data$`Br. Cd. 57, 144`
,"%",""))/100
bs.data$`Br. Cd. 55` <-  as.numeric(str_replace(bs.data$`Br. Cd. 55`,"%",""))
/100
```

```
bs.data$`Br. Cd. 272` <-  as.numeric(str_replace(bs.data$`Br. Cd. 272`,"%",""
))/100
bs.data$`Br. Cd. 286` <-  as.numeric(str_replace(bs.data$`Br. Cd. 286`,"%",""
))/100
bs.data$`Br. Cd. 481` <-  as.numeric(str_replace(bs.data$`Br. Cd. 481`,"%",""
))/100
bs.data$`Br. Cd. 352` <-  as.numeric(str_replace(bs.data$`Br. Cd. 352`,"%",""
))/100
bs.data$`Br. Cd. 5` <-  as.numeric(str_replace(bs.data$`Br. Cd. 5`,"%",""))/1
00

bs.data$`Others 999` <-  as.numeric(str_replace(bs.data$`Others 999`,"%",""))
/100

bs.data$`Pr Cat 1` <-  as.numeric(str_replace(bs.data$`Pr Cat 1`,"%",""))/100
bs.data$`Pr Cat 2` <-  as.numeric(str_replace(bs.data$`Pr Cat 2`,"%",""))/100
bs.data$`Pr Cat 3` <-  as.numeric(str_replace(bs.data$`Pr Cat 3`,"%",""))/100
bs.data$`Pr Cat 4` <-  as.numeric(str_replace(bs.data$`Pr Cat 4`,"%",""))/100

bs.data$`PropCat 5` <-  as.numeric(str_replace(bs.data$`PropCat 5`,"%",""))/1
00
bs.data$`PropCat 6` <-  as.numeric(str_replace(bs.data$`PropCat 6`,"%",""))/1
00
bs.data$`PropCat 7` <-  as.numeric(str_replace(bs.data$`PropCat 7`,"%",""))/1
00
bs.data$`PropCat 8` <-  as.numeric(str_replace(bs.data$`PropCat 8`,"%",""))/1
00
bs.data$`PropCat 9` <-  as.numeric(str_replace(bs.data$`PropCat 9`,"%",""))/1
00
bs.data$`PropCat 10`<-  as.numeric(str_replace(bs.data$`PropCat 10`,"%",""))/
100
bs.data$`PropCat 11`<-  as.numeric(str_replace(bs.data$`PropCat 11`,"%",""))/
100
bs.data$`PropCat 12`<-  as.numeric(str_replace(bs.data$`PropCat 12`,"%",""))/
100
bs.data$`PropCat 13`<-  as.numeric(str_replace(bs.data$`PropCat 13`,"%",""))/
100
bs.data$`PropCat 14`<-  as.numeric(str_replace(bs.data$`PropCat 14`,"%",""))/
100
bs.data$`PropCat 15`<-  as.numeric(str_replace(bs.data$`PropCat 15`,"%",""))/
100
```

Feature Engineering.

Ans. Variables used for evaluating brand loyalty:

- Average Price
- Brand Runs
- Number of transactions
- Number of brands
- Others999
- Total Volume
- Value
- Maximum Brand loyalty

Evaluating brand loyalty; max value among the brands shall correspond to be the values of brand loyalty as the max value in a particular brand implicitly links to loyalty of the brands that customer customer most often tends to spend the highest portion of the expenditure.

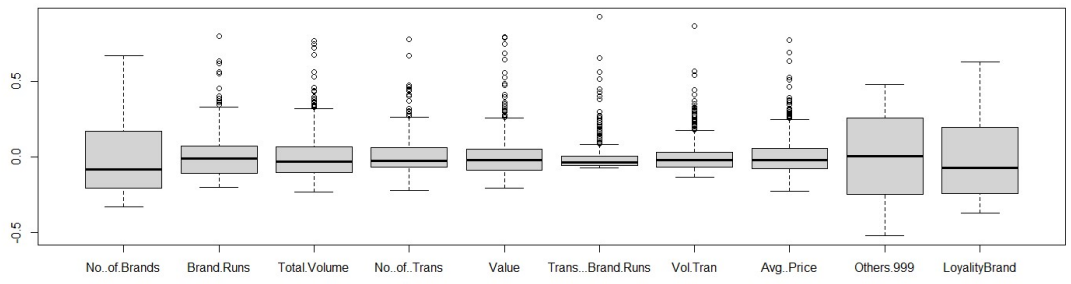
```
#loyal <- bs.data[,23:30]
bs.data$LoyalityBrand <- as.numeric(apply(bs.data[,23:30],1,max))
```

Q1. Use k-means clustering to identify clusters of households based on:

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

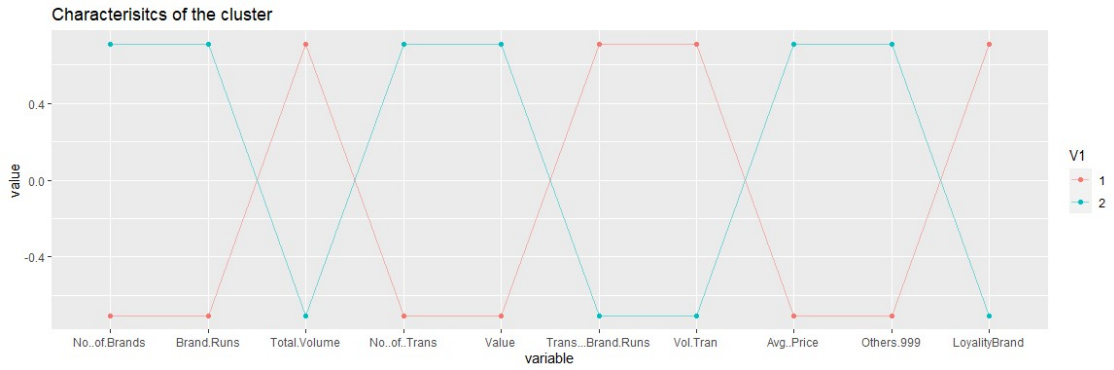
Creating a normalization and outlier removal function, which shall be used for every kmeans model in this project.

```
bs.data1.normalized <- data.frame(lapply(bs.data1, mean_norm_minmax))
boxplot(bs.data1.normalized)
```




```
# Vizual Scatterplot for the ph.cluster3 clusters
fviz_cluster(clusts1_2, bs.data1.normalized, palette = "Set2", ggtheme = them
e_minimal(), geom = "point" )
```


```
ggparcoord(out1, columns = 2:11, groupColumn = 1, showPoints = TRUE,  title =
"Characterisitcs of the cluster", alphaLines = 0.5 )
```

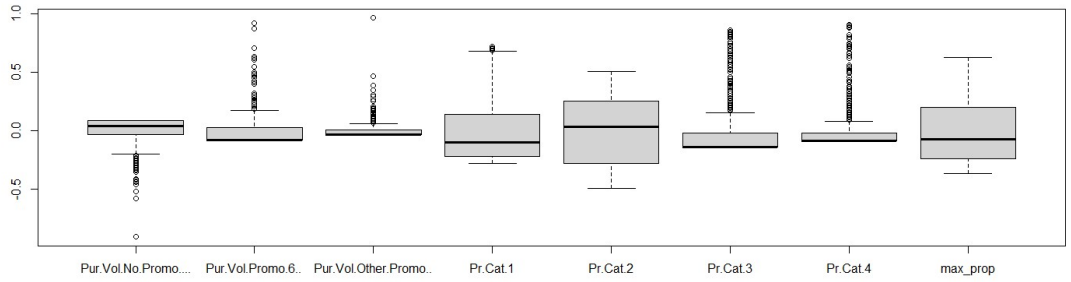


- cluster 2; Customers in this cluster appear to buy from others999 brands which indicate they are not brand loyal customers. They buy the heighst number of brands but the total volume of transaction is the least.
- cluster 1; Customers in this cluster has an opposite behavior compared to cluster 2. Customers in this cluster have maximum brand loyalty; they buy the least number of brands but have higher volumes of transaction in the limited transaction they do.

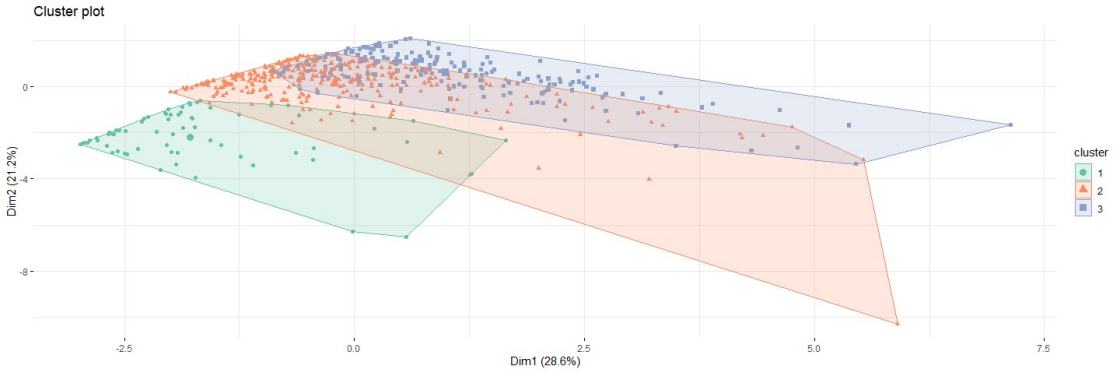
Q1 Use k-means clustering to identify clusters of households based on:

- (b). The variables that describe the basis for purchase
- variables used for clustering based on basis of purchase
  - All precise categories
  - Selling positions
  - Purchasing volume with no promotion, promotion 6 and other promotion

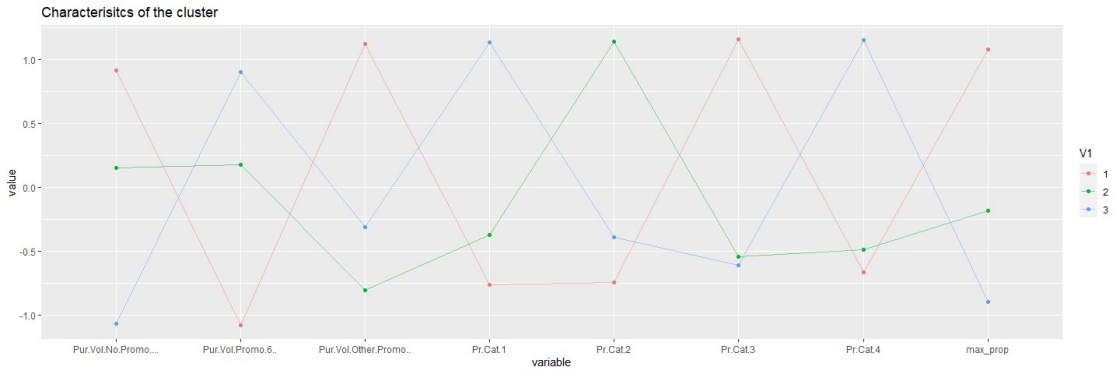
```
bs.data2.normalized <- data.frame(lapply(bs.data2, mean_norm_minmax))
boxplot(data.frame(lapply(bs.data2, mean_norm_minmax)))
```




```
# Vizual Scatterplot for the clusts2_3 clusters
fviz_cluster(clusts2_3, bs.data2.normalized,
             palette = "Set2", ggtheme = theme_minimal(), geom = "point" )
```




```
ggparcoord(out2_3, columns = 2:9, groupColumn = 1, showPoints = TRUE,  title
= "Characterisitcs of the cluster", alphaLines = 0.5 )
```



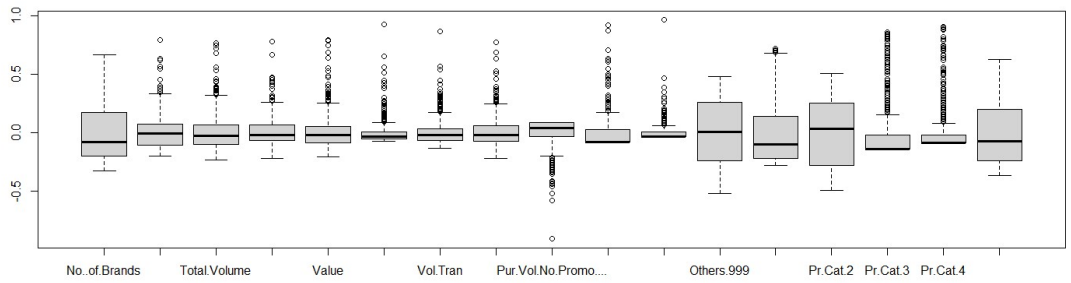
- cluster 3; Customers in this cluster don't purchase products without promotion offers, despite availing promotional offers their maximum proportion of purchase is so low that they won't efficaciously converted to loyal customers with offering more price offs.

- cluster 2; Customers in this cluster have a moderate behavior, they purchase products of a specific price category mostly. Their purchases aren't affected whether promotional offers.
- cluster 1; The behavior of Customers in this cluster evidently purchase products from a single price category. They purchase almost similarly both during price offs and no price offers. Customers in this cluster have a high brand loyalty.

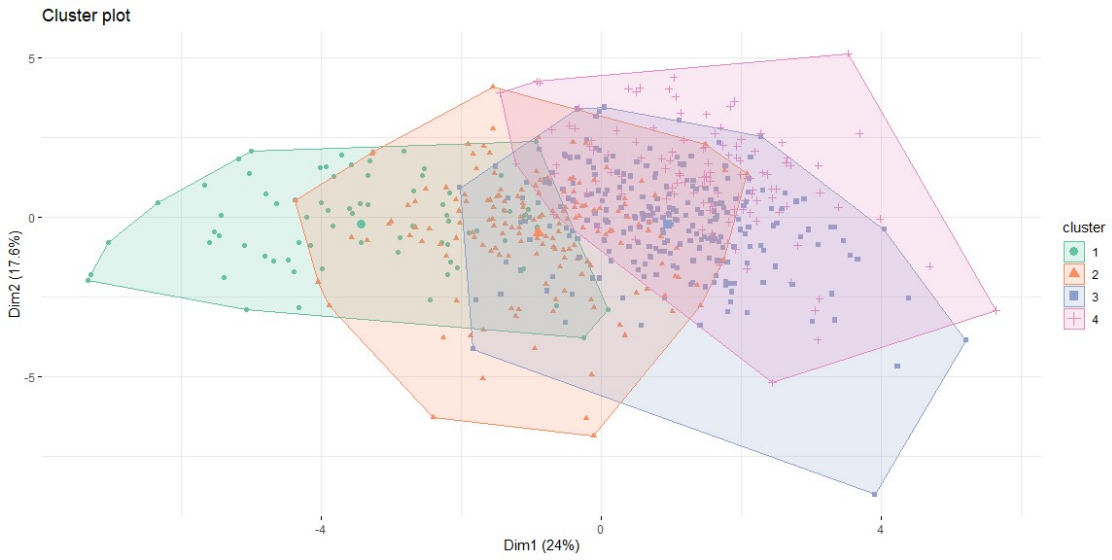
Q1. Use k-means clustering to identify clusters of households based on:

- (c). By taking the variables that describe the purchase behavior and basis of purchase and forming the cluster.

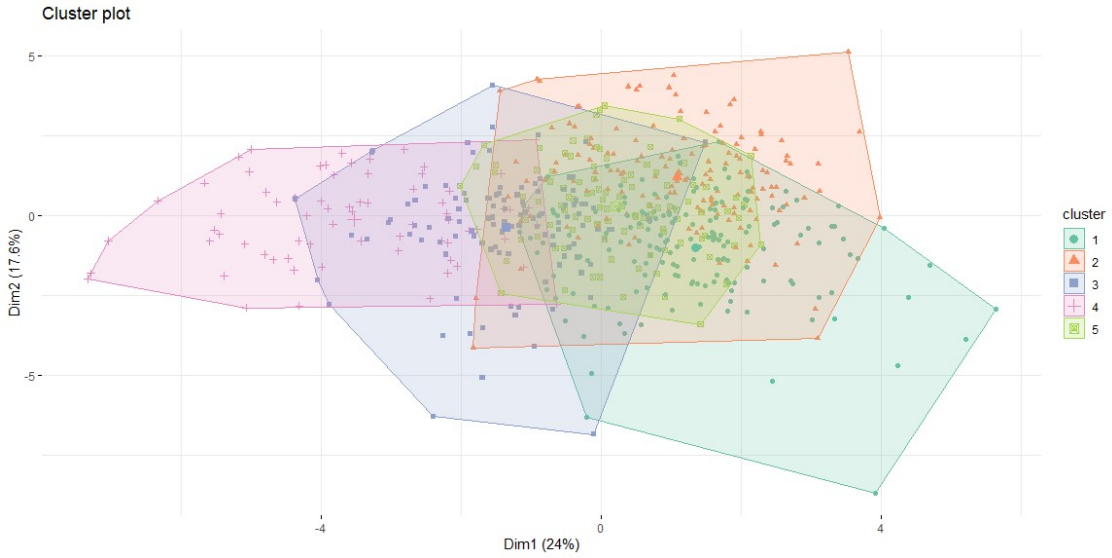
```
bs.data3 <- bs.data[,c(12:22,31:35,49)]
bs.data3.normalized <- data.frame(lapply(bs.data3, mean_norm_minmax))
boxplot(data.frame(lapply(bs.data3, mean_norm_minmax)))
```



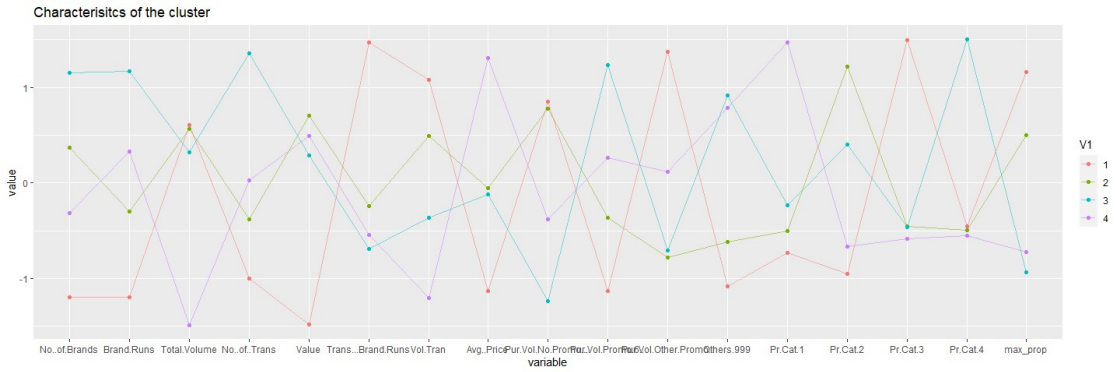

```
clusts3_4 <- kmeans(bs.data3.normalized, 4, 30)
clusts3_5 <- kmeans(bs.data3.normalized, 5, 30)

# Vizual Scatterplot for the clusters for k values 4 and 5
fviz_cluster(clusts3_4, bs.data3.normalized,
             palette = "Set2", ggtheme = theme_minimal(), geom = "point" )
```



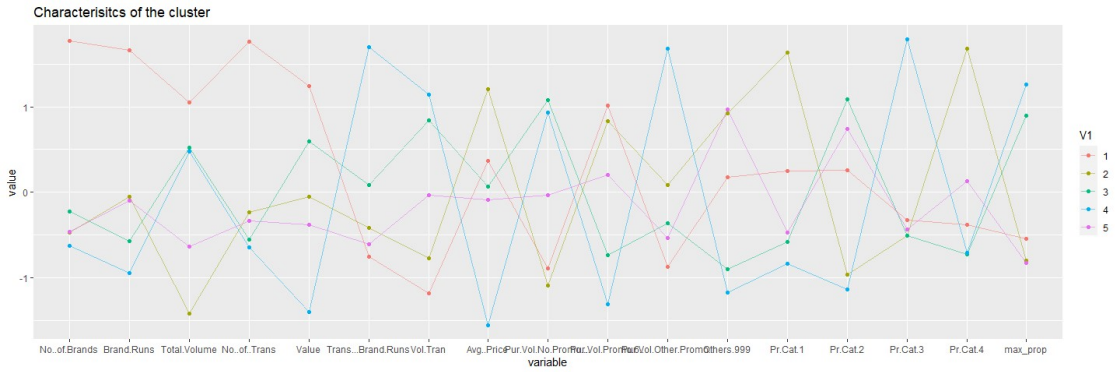

```
fviz_cluster(clusts3_5, bs.data3.normalized,
             palette = "Set2", ggtheme = theme_minimal(), geom = "point" )
```




```
ggparcoord(out3_4, columns = 2:18, groupColumn = 1, showPoints = TRUE,  title
= "Characterisitcs of the cluster", alphaLines = 0.5 )
```




```
ggparcoord(out3_5, columns = 2:18, groupColumn = 1, showPoints = TRUE,  title
= "Characterisitcs of the cluster", alphaLines = 0.5 )
```



How should K be chosen?

Ans.

- 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.

- In all 3 segmentation, we observe that for k= 4, distance within clusters is minimum and distance between clusters is maximum. But when we look at the centroid plots in Question 1. We notice that algorithm gives similar information for both k=2 and.3 Since we are getting similar information at 4 with minimum distance between clusters, we conclude that K-means algorithm with K=4 is the best model.

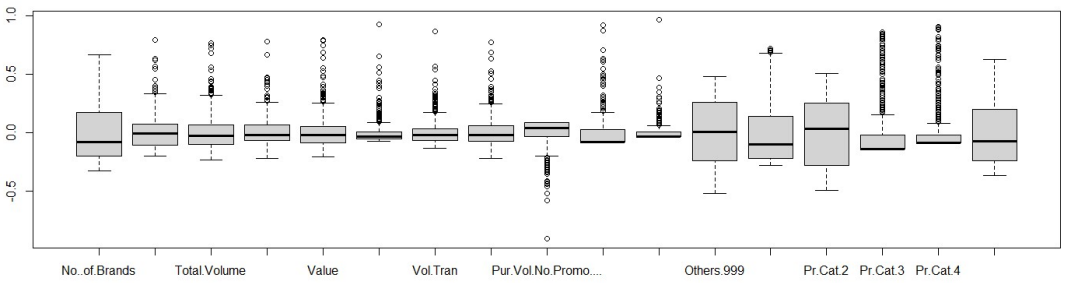
How should the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variable as is?

Ans.

- 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. As the variable was a string type initially using it as is in kmeans would give a garbage output. if you were trying to cluster customer profiles to do segmentation, you could count up words representing their interests in their profiles, and then have one column per interest, and count the number of times that word or n-gram appeared in the profile, then cluster on that matrix of numbers. But as kmeans calculated the euclidean distances, the string type variables is supposed to be transformed into a metric form.

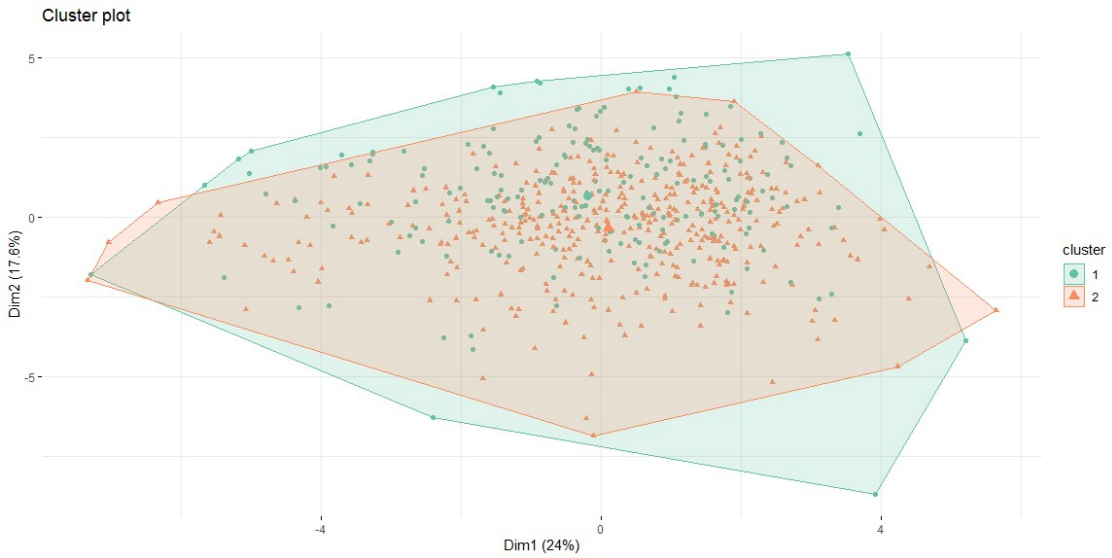
Adding the demographics to the basis of purchase variables

```
bs.data4 <- bs.data[,c(2:11,12:22,31:35,49)]
bs.data4.normalized <- data.frame(bs.data[,2:10],lapply(bs.data4[,11:27], mea
n_norm_minmax))
boxplot(data.frame(lapply(bs.data4[,c(11:27)], mean_norm_minmax)))
```

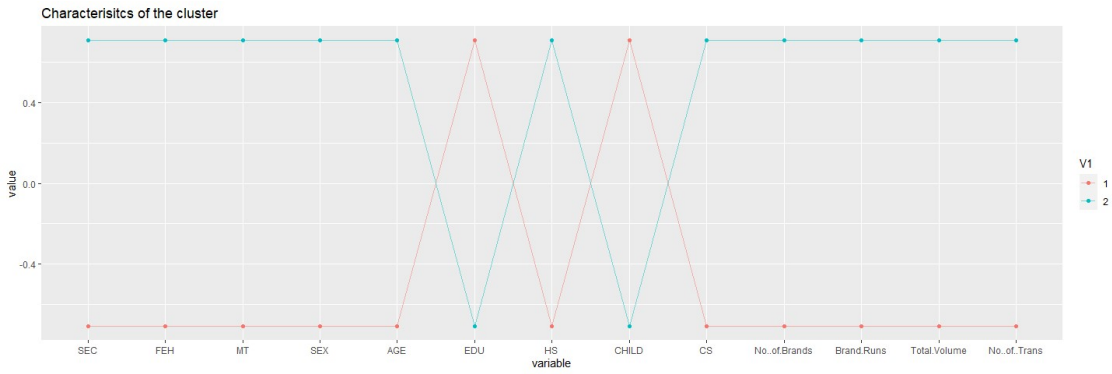


- We have considered 3 criteria to choose K:
  - Minimum distance within cluster
  - Maximum distance between clusters
  - Information from centroid plot of clusters

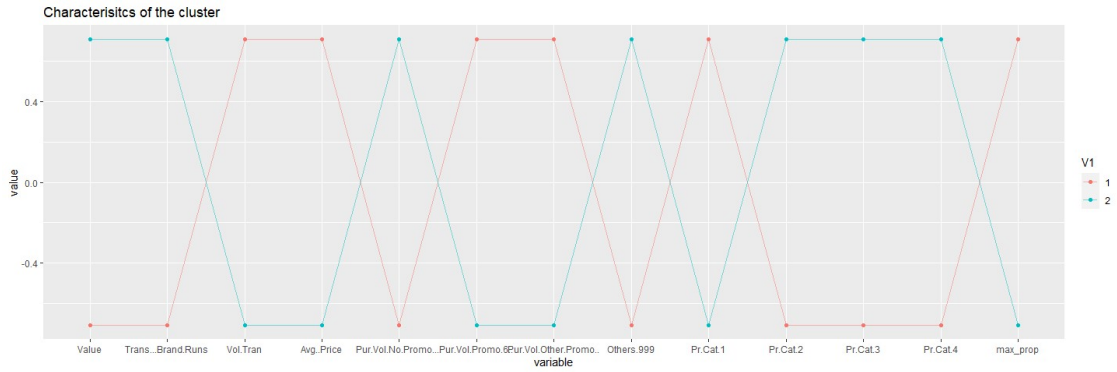
```
clusts4_4 <- kmeans(bs.data4.normalized, 2, 30)
fviz_cluster(clusts4_4, bs.data4.normalized[,10:26], palette = "Set2", ggthem
e = theme_minimal(), geom = "point" )
```




```
kmeans_basic_table4_4 <- data.frame(clusts4_4$size, clusts4_4$centers, cluste
r = factor(clusts4_4$cluster))
kmeans_basic_df4_4 <- data.frame(Cluster = factor(clusts4_4$cluster), bs.data
4.normalized)

out4 <- as.data.frame(cbind(1:nrow(clusts4_4$centers),clusts4_4$centers))
out4$V1 <- as.factor(out4$V1)

ggparcoord(out4, columns = 2:14, groupColumn = 1, showPoints = TRUE,  title =
"Characterisitcs of the cluster", alphaLines = 0.5 )
```

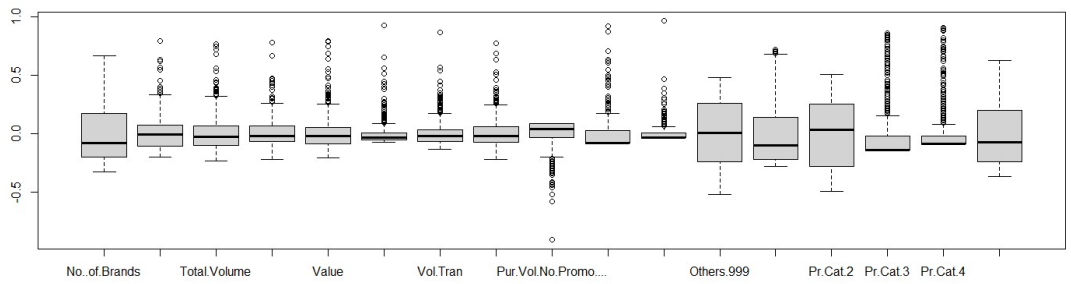



```
ggparcoord(out4, columns = 15:27, groupColumn = 1, showPoints = TRUE,  title
= "Characterisitcs of the cluster", alphaLines = 0.5 )
```

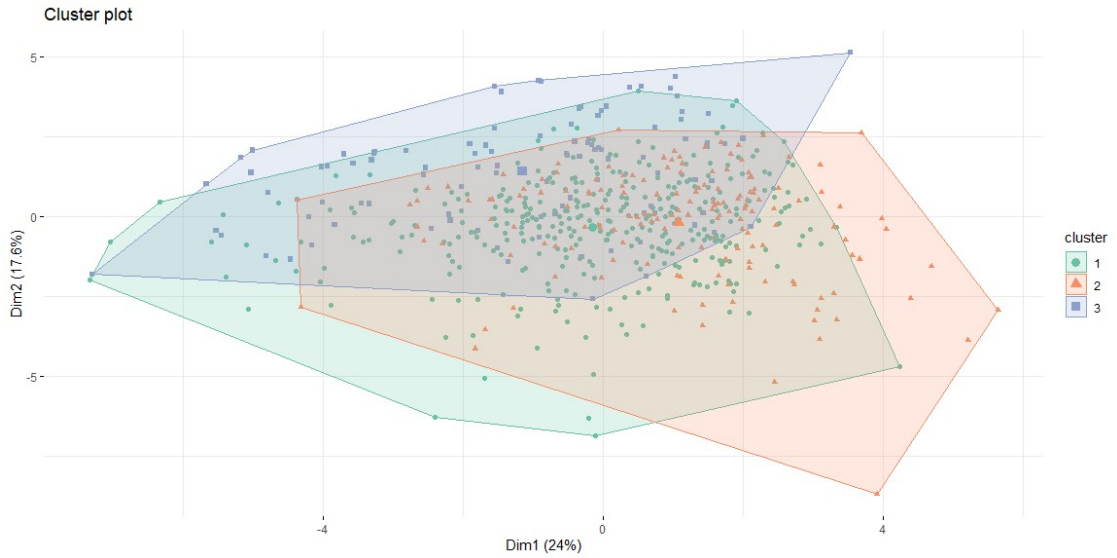


Forming clusters by using all variables

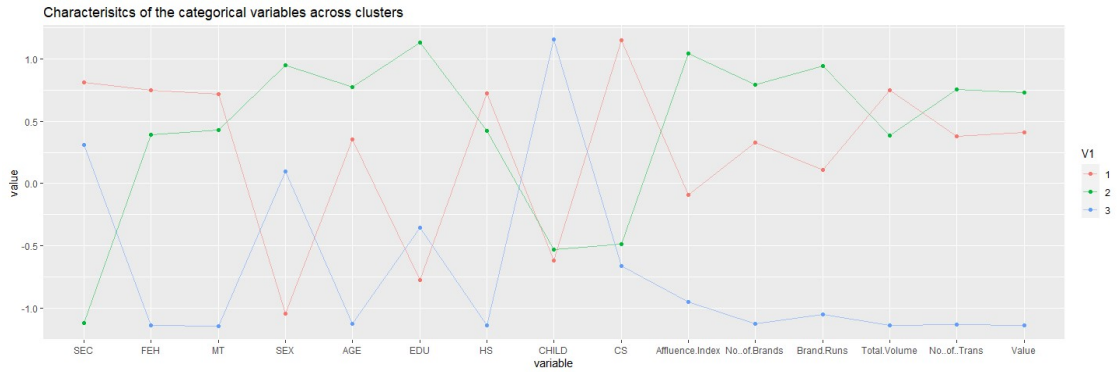
```
bs.data5 <- bs.data[,c(2:11,12:22,31:35,49)]
bs.data5.normalized <- data.frame(bs.data5[,1:10],lapply(bs.data5[,11:27], me
an_norm_minmax))
boxplot(data.frame(lapply(bs.data5[,c(11:27)], mean_norm_minmax)))
```



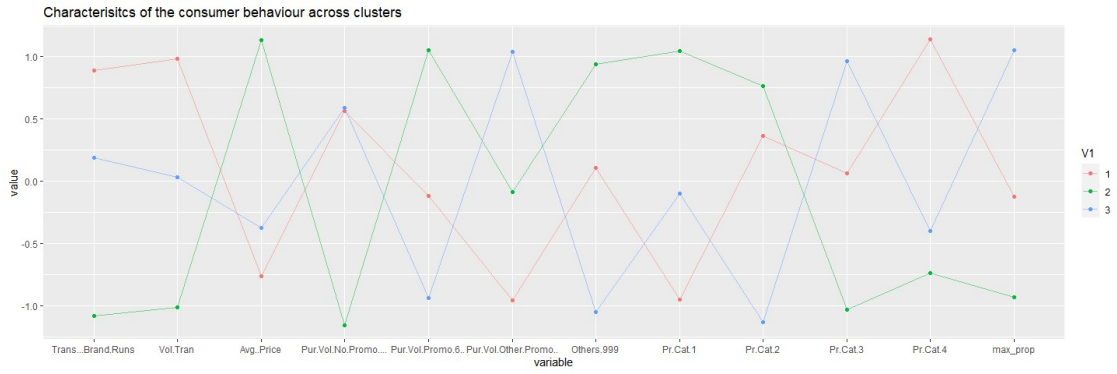

```
clusts5_4 <- kmeans(bs.data5.normalized, 3, 30)
fviz_cluster(clusts5_4, bs.data5.normalized[,11:27], palette = "Set2", ggthem
e = theme_minimal(), geom = "point" )
```



```
kmeans_basic_table5_4 <- data.frame(clusts5_4$size, clusts5_4$centers, cluste
r = factor(clusts5_4$cluster))
kmeans_basic_df5_4 <- data.frame(Cluster = factor(clusts5_4$cluster), bs.data
5.normalized)

out5 <- as.data.frame(cbind(1:nrow(clusts5_4$centers),clusts5_4$centers))
out5$V1 <- as.factor(out5$V1)

ggparcoord(out5, columns = 2:16, groupColumn = 1, showPoints = TRUE,  title =
"Characterisitcs of the categorical variables across clusters", alphaLines =
0.5 )
```




```
ggparcoord(out5, columns = 17:28, groupColumn = 1, showPoints = TRUE,  title
= "Characterisitcs of the consumer behaviour across clusters", alphaLines = 0
.5 )
```



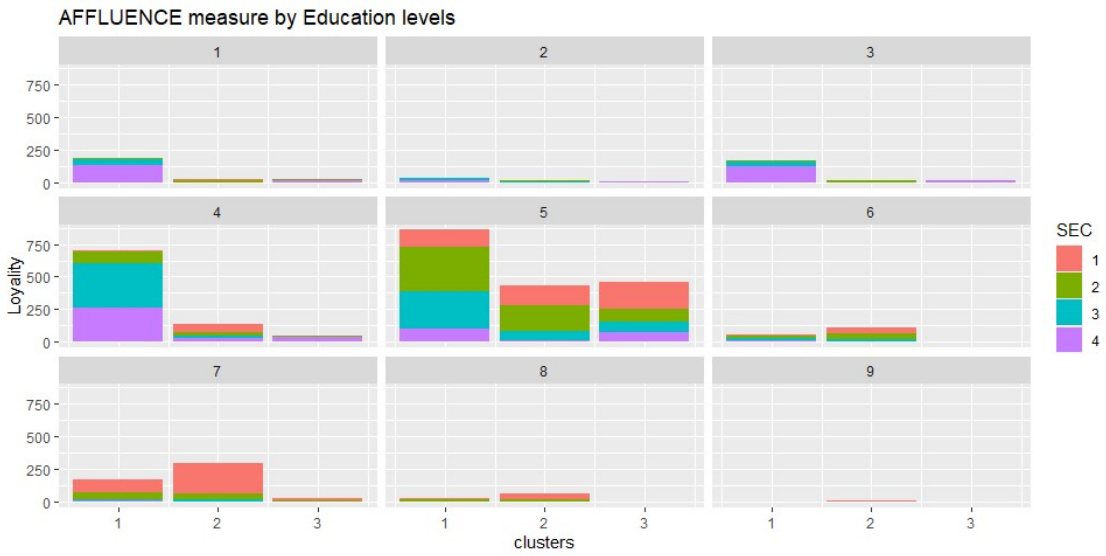
Q2. Select what you think is the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

- Customers buying more “other999” products are least loyal. Most customers fall in the segment who are not particularly brand loyal but prefer to buy value added packs and premium soaps.
- As most customers from cluster 1 have access to TV/cable, advertisements can be broadcast on television as an effective means of promoting products.

- Customers in cluster 2 are having a lower SEC and are buying products only when promotional offers are available which is consequently affecting the loyalty to degrade. Cluster 2 customers have a higher degree of House hold members but less of own children.
- From cluster 1; we can say that customers who do not care about price offers, are those who will buy products at any rate irrespective of the brand and they are generally attain good education or do not have any formal schooling.
- Customers with low SEC are categorized into cluster 2 Customers in this cluster group, they are buying products when there are only promo offers available and not maintaining the loyalty to the specific product.

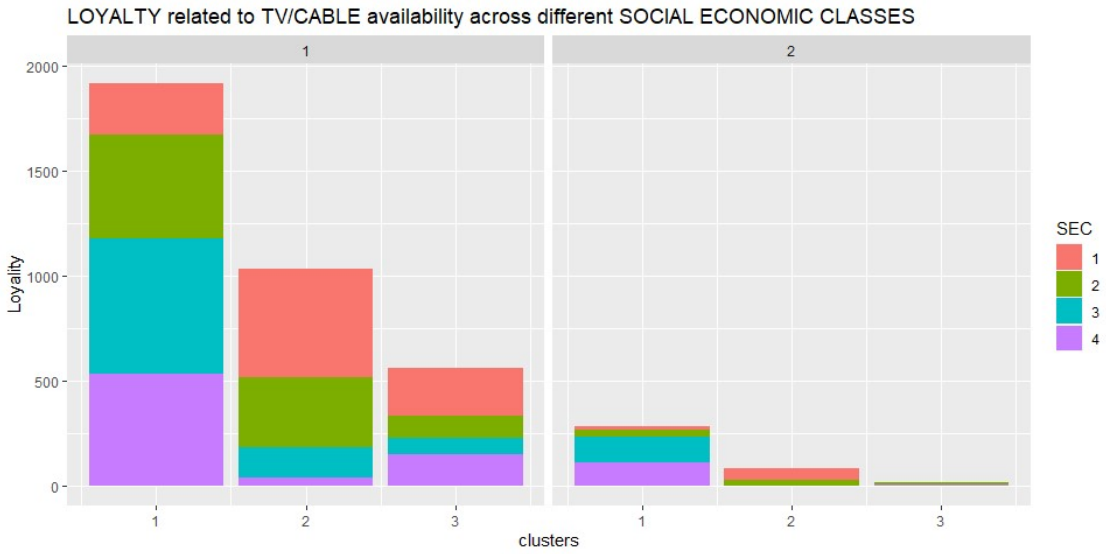
Q3. Develop a model that classifies the data into these segments. Since this information would most likely be used in targeting direct-mail promotions, it would be useful to select a market segment that would be defined as a success in the classification model

```
ggplot(datax, aes(x =clusters , y= Loyality, fill= SEC)) +  geom_bar(stat = '
identity') + facet_wrap(~EDU) + ggtitle("AFFLUENCE measure by Education level
s")
```



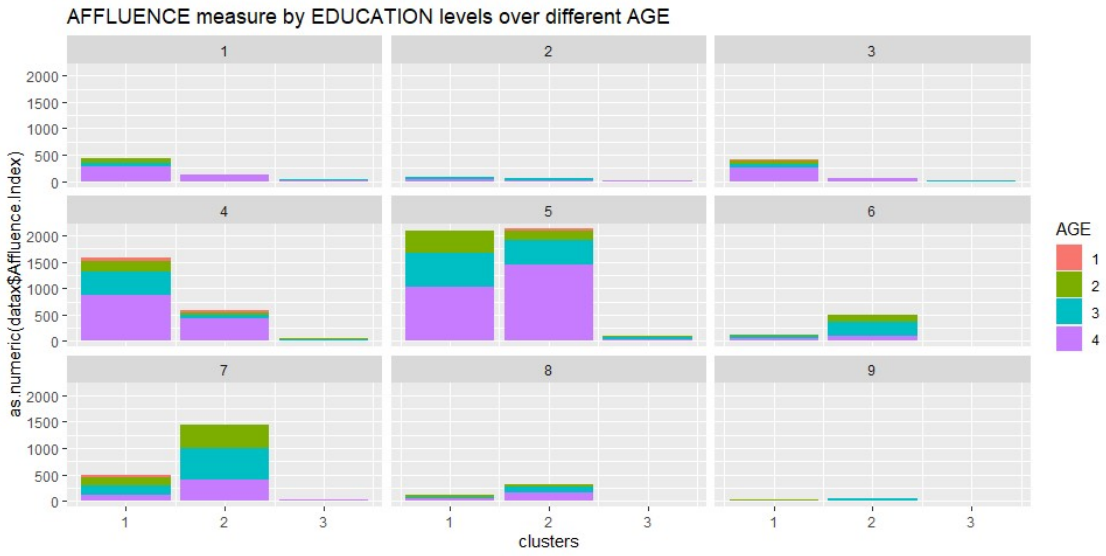
- Considering education as demographics, it seems most of customers have had education up to 4th and 5th level. There is a high proportion of college graduates in cluster 1 which buys value added packs and premium soaps which shows high brand royalty.

```
ggplot(datax, aes(x =clusters, y =Loyality, fill= SEC)) + geom_bar(stat = 'id
entity') + facet_wrap(~CS) + ggtitle("LOYALTY related to TV/CABLE availabilit
y across different SOCIAL ECONOMIC CLASSES")
```



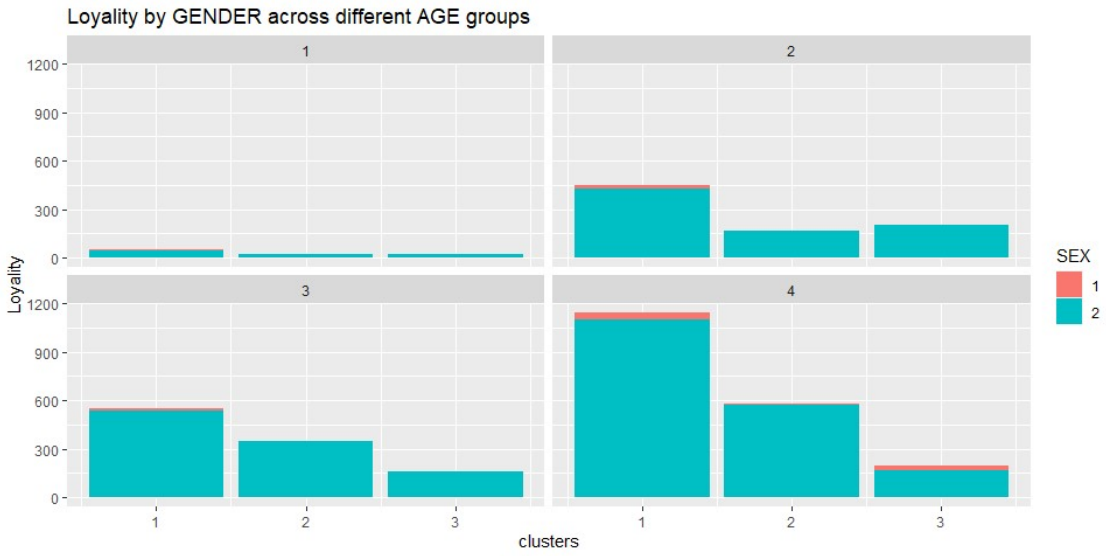
- Cluster 1 includes customers who show a high tendency to buy premium soaps. Another interesting fact is that there is a high percentage o customers from other SEC sections in cluster 1, indicating that they prefer to buy any kind of soap. Hence customers with high social economic status don't care about premium or popular soaps, but their brand royalty is high.
- As most customers from cluster 1 have access to TV/cable, advertisements can be broadcast on television as an effective means of promoting products.

```
ggplot(datax, aes(x =clusters, y=as.numeric(datax$Affluence.Index), fill= AGE
)) +  geom_bar(stat = 'identity') + facet_wrap(~EDU) + ggtitle("AFFLUENCE mea
sure by EDUCATION levels over different AGE")
```



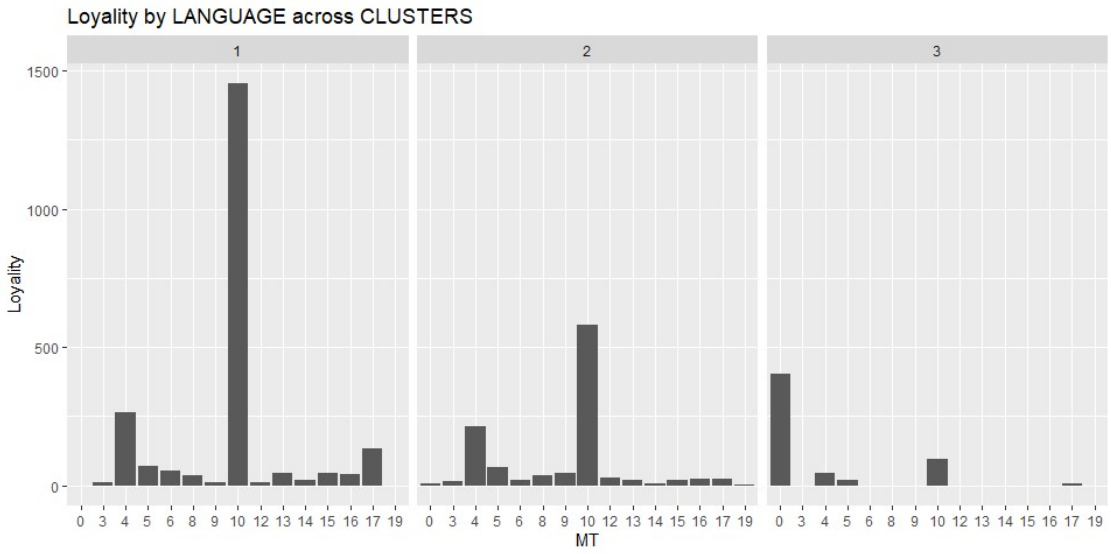
- Cluster 3 consists customers with highly affluent people across all education levels.
- Customers of Age group 4 are most affluent customer.
- Cluster 3 presents don't have a higher customer base loyal to the brand, but have a high scale of affluence. Customers Potential to be converted into brand loyal customers.

```
ggplot(datax, aes(x =clusters, y=Loyality, fill= SEX)) + geom_bar(stat = 'ide
ntity') + facet_wrap(~AGE) + ggtitle("Loyality by GENDER across different AGE
groups")
```



- Maximum number of customers in each cluster are women. All clusters also have the highest proportion of women across every cluster. No strong conlusion can be devised out of it except the brand is loved by women or it manufactures products for specific gender type.

```
ggplot(datax, aes(x =MT, y=Loyality)) + facet_wrap(~factor(clusters)) + geom_
bar(stat = 'identity')  + ggtitle("Loyality by LANGUAGE across CLUSTERS")
```



- This demographic does not appear to be significant as most clusters are dominated by a customer base who speak a common mother tongue. Looks like the sample data data was gathered from a locality speaking a specific language.

Conclusion:

- Most customers are female, thus most of the ads and future product development plans should be framed accordingly. Also most of the customers fall in the cluster where they are not brand loyal but prefer to buy value added packs and premium packs and premium soaps.
- As most of the customer base have TV/Cable; advertisements can be broadcasted on television as an effective means of promoting the products.