ML Final Exam

Yanxi Li

12/16/2020

I. Introduction

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods.

CRISA has traditionally segmented markets on the basis of purchaser demographics.

But they would now like to segment the market based on 2 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);

In this project, models will be built to select a market segmentation that would be defined as a success in the classification model.

II. Data Preparation

There are many ways to define the brand loyalty. Since the customer who buys all brand A just as loyal as a customer who buys all brand B, we use a derived variable called "Max_of_brand" which choose the maximum value for different brands except "Others.999" to define the brand loyalty.

For the proposition category from 5 to 15, I also select the maximum value named "Max_of_proposition" to reduce the unnecessary variables.

I noticed there are 1/6 missing data shows "0" in demographic columns and it should be removed when analyze the demographic part.

Since we first use variables in others parts (except the demography) and there is no missing value in other parts, I plan to use all data to normalize. Because the larger amount of data, the better estimate of the data mean & SD.

Data removing and data visualization for the demographic variables is under the IV's subtitle "Deal with demographic".

I change some character variables to numeric in order to build the k-means model.

The column 16 "Trans/Brand Runs" and column 17 "Vol/Trans" can be calculated from the other purchase behavior variables which is duplicated, so I did not put them in the model.

```
# load the libraries
library(tidyverse)   # data manipulation
library(factoextra)  # clustering algorithm & visualization
library(ggplot2)     # plot figure
library(FactoMineR)  # run PCA
library(cluster)     # agglomerative hierarchical clustering
library(dbscan)      # density based spatial clustering
library(gridExtra)   # arrange figures together

# load the data
data <- read.csv("BathSoap.csv")

# change variables of character to numeric
data[,20:46] <- sapply(data[,20:46], function(x) as.numeric(gsub("%", "", x)) / 100)

# Define Brand Loyalty
data$Max_of_brand <- apply(data[,23:30], 1, max)

# Max for the selling proposition 5 to 15
data$Max_of_proposition <- apply(data[,36:46], 1, max)

# change first column(Member.id) to row-names
rownames(data)<- data[,1]
# delete the first column
data <- data[,-1]

# copy the data
data_norm <- data
# normalize the numerical data
data_norm[,c(11:21, 30:34, 46, 47)] <- scale(data[,c(11:21, 30:34, 46, 47)])
```

III. K-means Clustering

We use 3 ways to build the k-means models and since the marketing efforts would support two to five different promotional approaches, I choose k values from 2 to 5.

a. The variables that describe purchase behavior (including brand loyalty) Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)
b. The variables that describe the basis for purchase Basis of purchase (price, selling proposition)
c. The variables that describe both purchase behavior and basis of purchase (both a & b)

Although the cluster visual assessment tells us where delineations occur between clusters, it does not tell us what the optimal number of clusters is.

For each of the 3 models, I run the k-means first with the k value from 2 to 5, and then use Elbow, Silhouette and Sum of Squares methods to identify the appropriate k value.

Sum of Squares method is to choose the optimal number of cluster by minimizing the within-cluster sum of squares (a measure of how tight each cluster is) and maximizing the between-cluster sum of squares (a measure of how separated each cluster is from the others).

a Purchase behavior

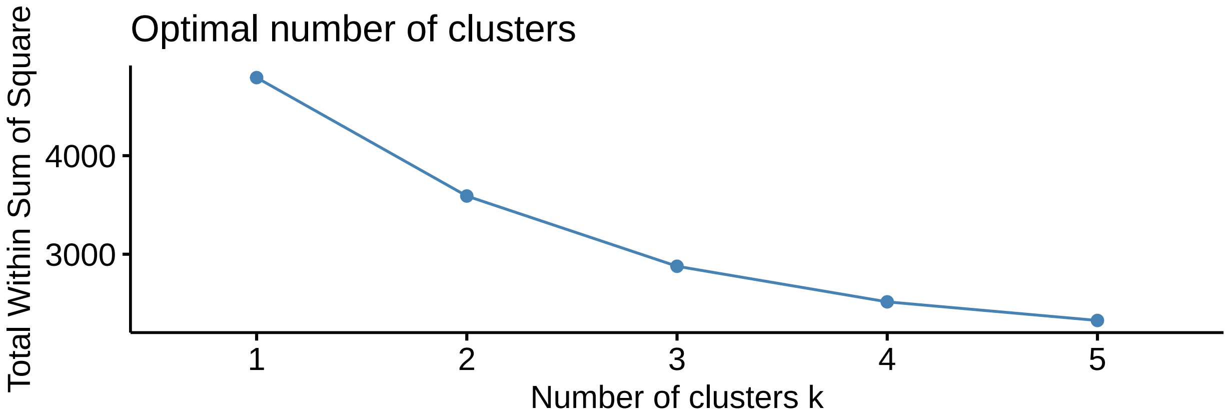
I use No..of.Brands, Brand.Runs, Total.Volume, No..of..Trans, Value, Avg..Price, Others.999, and Max_of_brand these 8 variables.

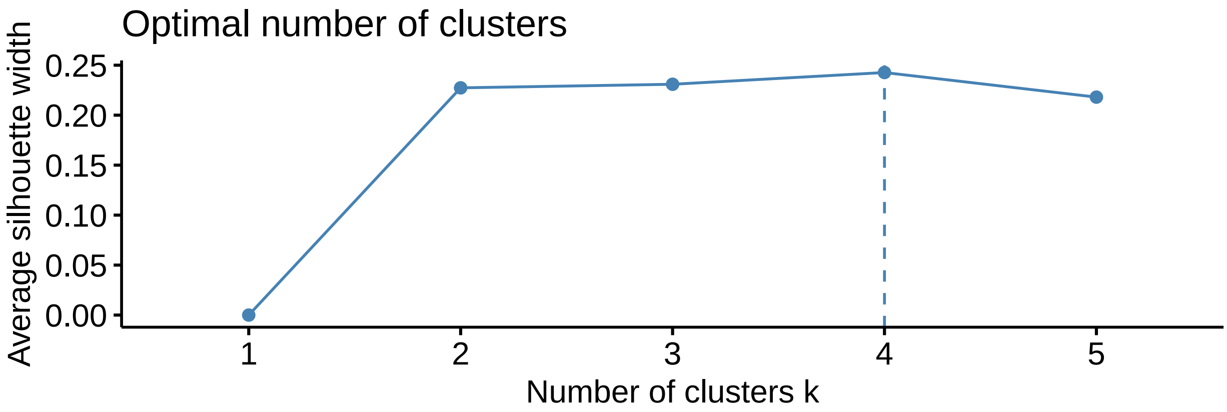
Columns number 11:15, 18, 30, 46 represents for these variables.

```
# generate the same random numbers
set.seed(123)

# build k-means with k value 2,3,4,5
a_k2 <- kmeans(data_norm[,c(11:15, 18, 30, 46)], centers = 2, nstart = 25) #iterate 25 times
a_k3 <- kmeans(data_norm[,c(11:15, 18, 30, 46)], centers = 3, nstart = 25)
a_k4 <- kmeans(data_norm[,c(11:15, 18, 30, 46)], centers = 4, nstart = 25)
a_k5 <- kmeans(data_norm[,c(11:15, 18, 30, 46)], centers = 5, nstart = 25)
```

Elbow & Silhouette method

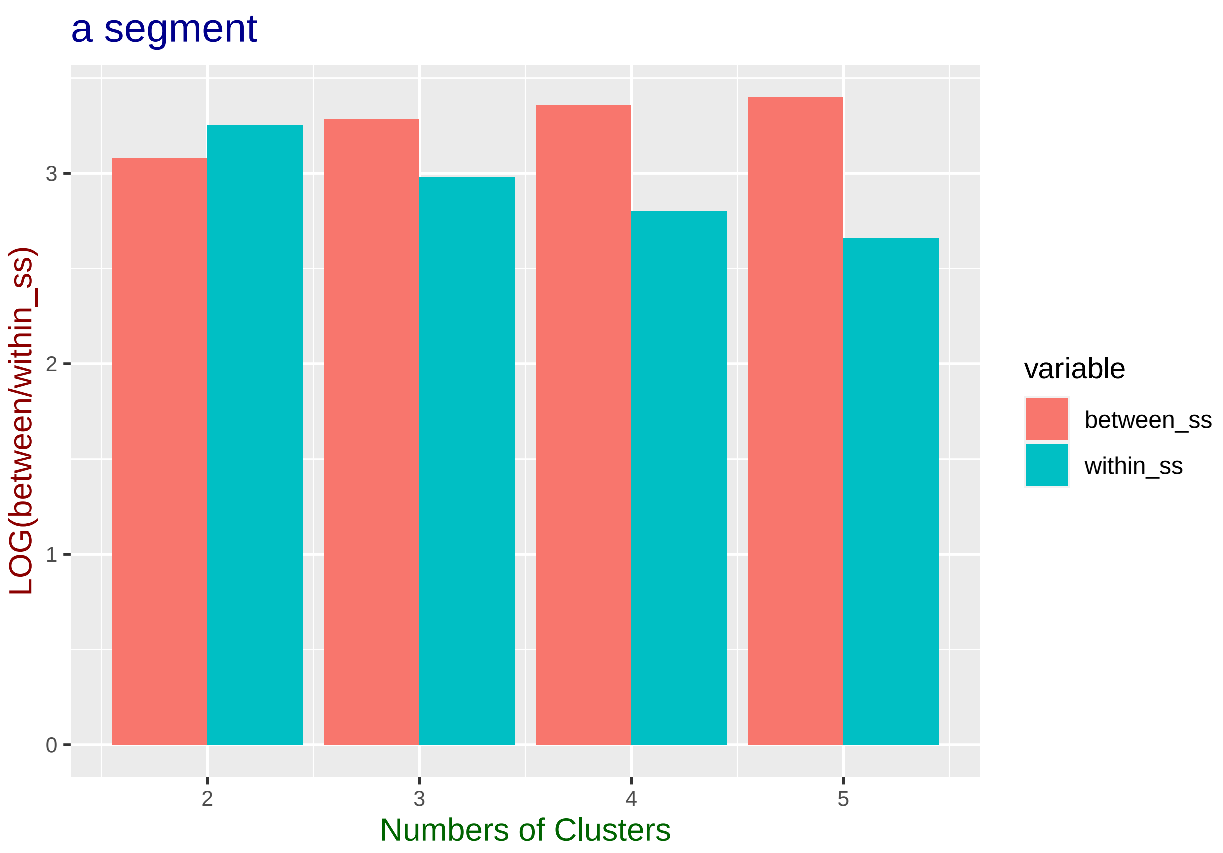
```
# use the Elbow Method to find the cluster numbers
Pa_elbow <- fviz_nbclust(data_norm[,c(11:15, 18, 30, 46)], kmeans,
                                                          method = "wss", k.max = 5)
# use the Silhouette Method to find the cluster numbers
Pa_silho <- fviz_nbclust(data_norm[,c(11:15, 18, 30, 46)], kmeans,
                                                   method = "silhouette", k.max = 5)

# display plots together
gridExtra::grid.arrange(Pa_elbow, Pa_silho, nrow = 2)
```





Sum of Squares mothod

```r
# combine within and between sum of squares in one data.frame
a_ssc <- data.frame(
  kmeans = c(2,3,4,5),
  within_ss = c(mean(a_k2$withinss), mean(a_k3$withinss), mean(a_k4$withinss), mean(a_k5$withinss)),
  between_ss = c(a_k2$betweenss, a_k3$betweenss, a_k4$betweenss, a_k5$betweenss))

# use pivot_longer to make variables together
a_ssc_pivot <- pivot_longer(a_ssc, cols = c("within_ss", "between_ss"),
                            names_to="variable", values_to = "value")

# ggplot to get bar-plot
ggplot(a_ssc_pivot, aes(x = kmeans, y = log10(value), fill = variable)) +
    geom_bar(stat='identity', position='dodge') +
    ggtitle("a segment") +      # name the title
    xlab("Numbers of Clusters") +        # name x axis
    ylab("LOG(between/within_ss)") +     # name y axis
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12), # y axis name's color & size
          axis.text.x = element_text(size=8),   # change x axis number's size
          axis.text.y = element_text(size=8),   # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size
```



Based on each group size and these 3 methods, k = 4 is better.

b Basis of purchase

I use Pur.Vol.No.Promo...., Pur.Vol.Promo.6.., Pur.Vol.Other.Promo.., Pr.Cat.1, Pr.Cat.2, Pr.Cat.3, Pr.Cat.4, Max_of_proposition these 8 variables.

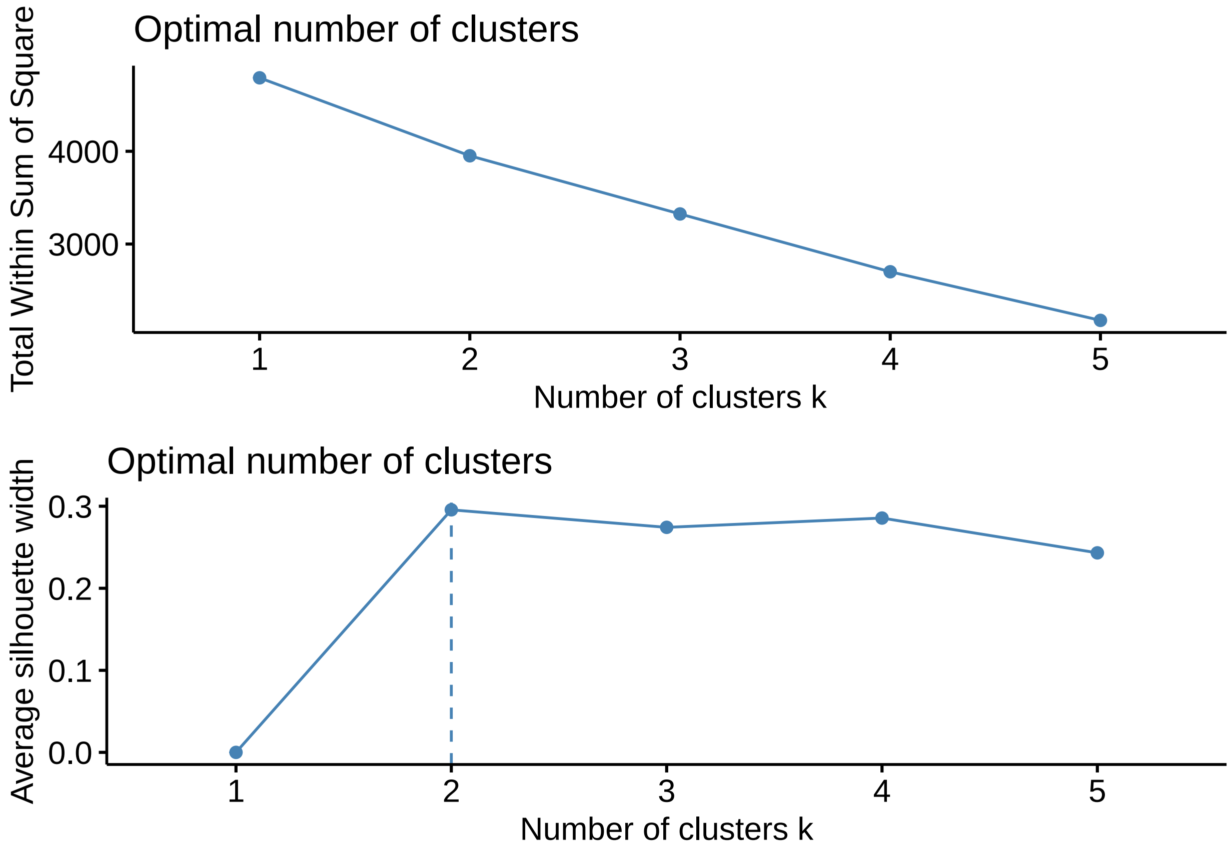
Columns number 19:21, 31:34, 47 represents these variables.

```
# generate the same random numbers
set.seed(123)

# build k-means with k value 2,3,4,5
b_k2 <- kmeans(data_norm[,c(19:21, 31:34, 47)], centers = 2, nstart = 25) #iterate 25 times
b_k3 <- kmeans(data_norm[,c(19:21, 31:34, 47)], centers = 3, nstart = 25)
b_k4 <- kmeans(data_norm[,c(19:21, 31:34, 47)], centers = 4, nstart = 25)
b_k5 <- kmeans(data_norm[,c(19:21, 31:34, 47)], centers = 5, nstart = 25)
```

Elbow & Silhouette mothod

```
# use the Elbow Method to find the cluster numbers
Pb_elbow <- fviz_nbclust(data_norm[,c(19:21, 31:34, 47)], kmeans,
                                                        method = "wss", k.max = 5)
# use the Silhouette Method to find the cluster numbers
Pb_silho <- fviz_nbclust(data_norm[,c(19:21, 31:34, 47)], kmeans,
                                                 method = "silhouette", k.max = 5)
```

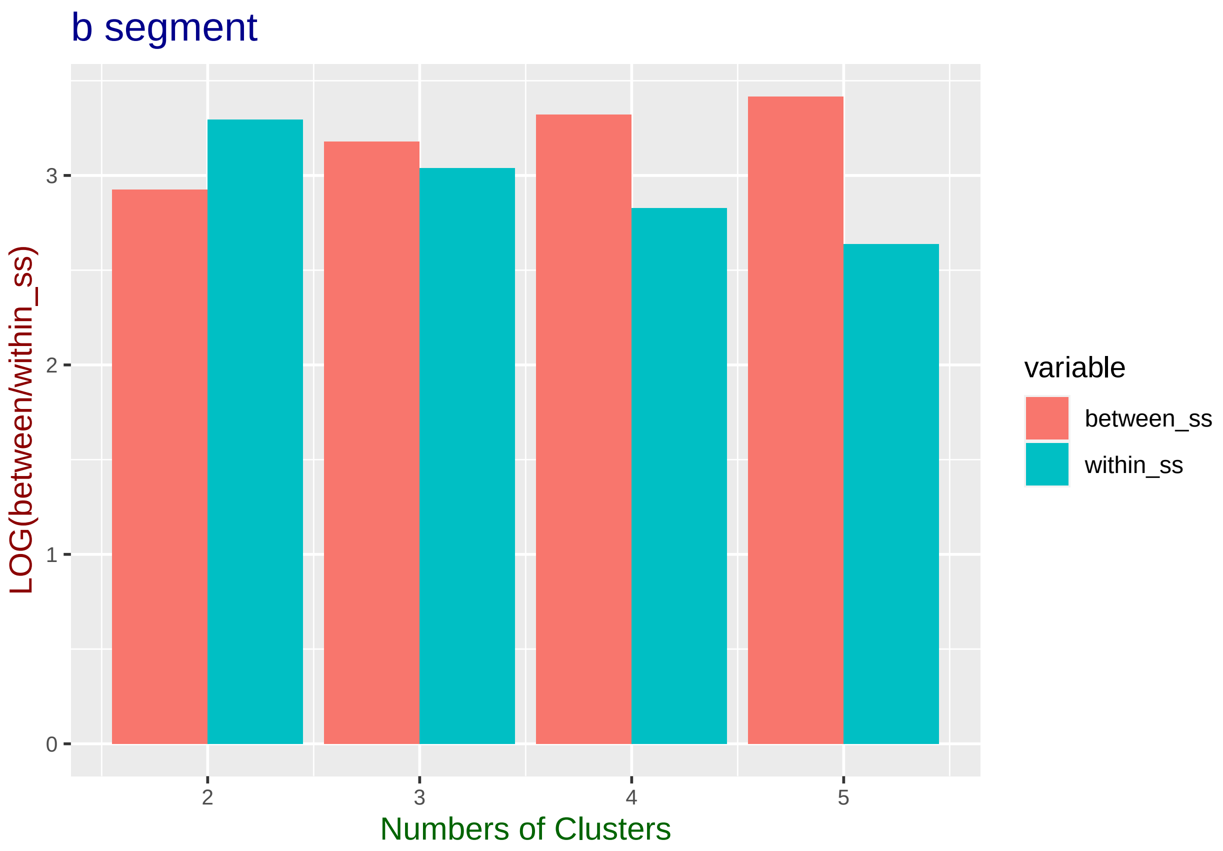
```
# display plots together
grid.arrange(Pb_elbow, Pb_silho, nrow = 2)
```



Sum of Squares method

```
# combine within and between sum of squares in one data.frame
b_ssc <- data.frame(
  kmeans = c(2,3,4,5),
  within_ss = c(mean(b_k2$withinss), mean(b_k3$withinss), mean(b_k4$withinss), mean(b_k5$withinss)),
  between_ss = c(b_k2$betweenss, b_k3$betweenss, b_k4$betweenss, b_k5$betweenss))

# use pivot_longer to make variables together
b_ssc_pivot <- pivot_longer(b_ssc, cols = c("within_ss", "between_ss"),
                            names_to="variable", values_to = "value")

# plot bar-plot
ggplot(b_ssc_pivot, aes(x = kmeans, y = log10(value), fill = variable)) +
    geom_bar(stat='identity', position='dodge') +
    ggtitle("b segment") +      # name the title
    xlab("Numbers of Clusters") +        # name x axis
    ylab("LOG(between/within_ss)") +     # name y axis
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),  # y axis name's color & size
          axis.text.x = element_text(size=8),    # change x axis number's size
          axis.text.y = element_text(size=8),    # change y axis number's size
```

```
        plot.title = element_text(color="DarkBlue", size=15)) # title's color & size
```



Based on cluster size and 3 methods, k = 3 is better.

c both a & b

Use the variables in a & b, total is 16 variables.

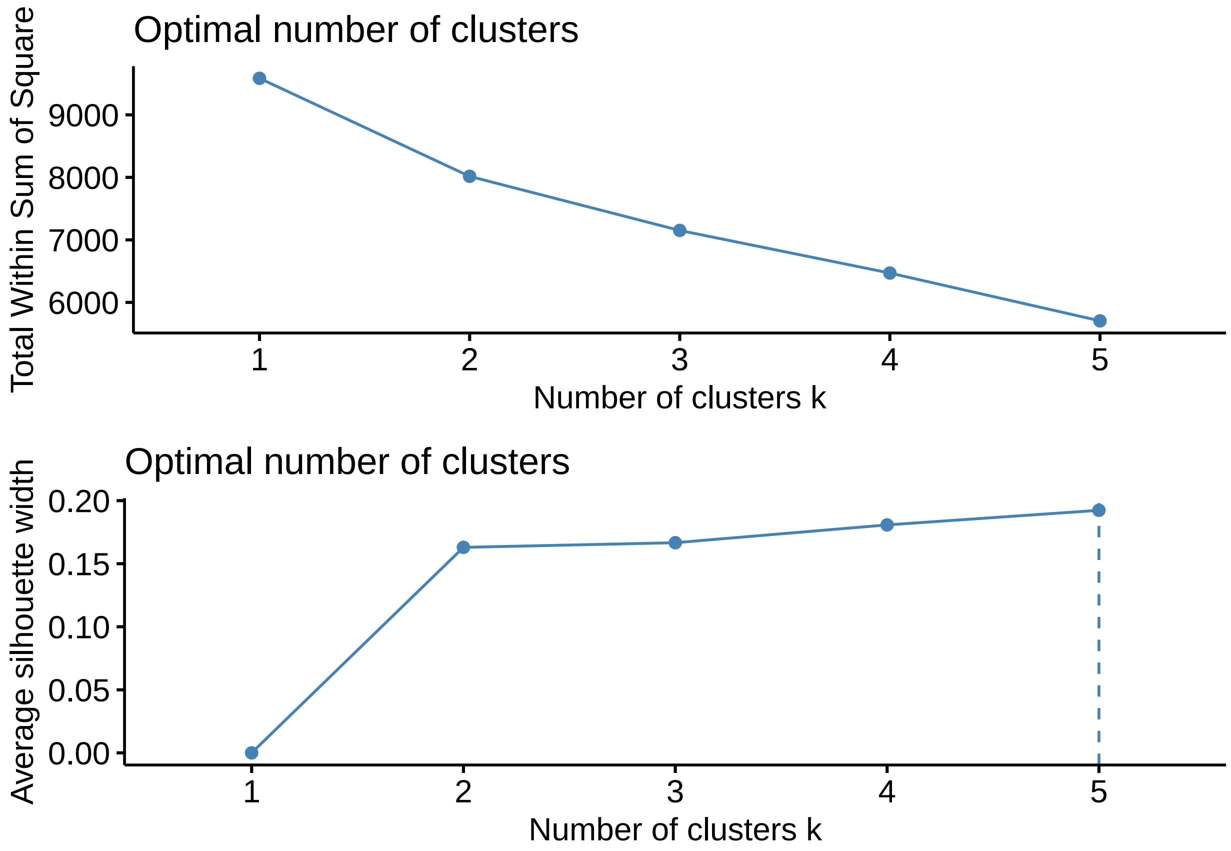
Columns number 11:15, 18:21, 30:34, 46, 47 are for these variables.

```
# generate the same random numbers
set.seed(123)

# build k-means with k value 2,3,4,5
c_k2 <- kmeans(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], centers = 2, nstart = 25)
c_k3 <- kmeans(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], centers = 3, nstart = 25)
c_k4 <- kmeans(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], centers = 4, nstart = 25)
c_k5 <- kmeans(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], centers = 5, nstart = 25)
```

Elbow & Silhouette method

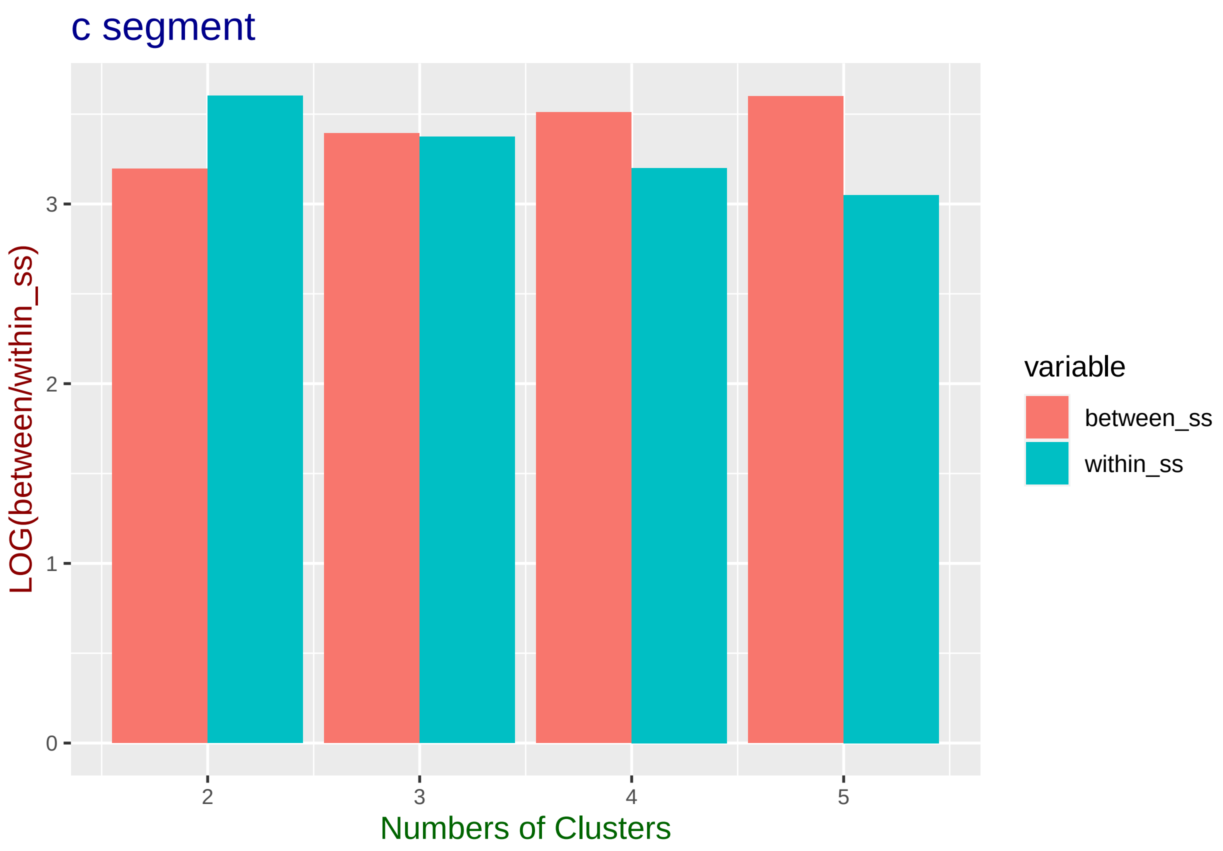
```
# use the Elbow Method to find the cluster numbers
Pc_elbow <- fviz_nbclust(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], kmeans,
                                                          method = "wss", k.max = 5)
# use the Silhouette Method to find the cluster numbers
```

```r
Pc_silho <- fviz_nbclust(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], kmeans,
                                                        method = "silhouette", k.max = 5)

# display plots together
grid.arrange(Pc_elbow, Pc_silho, nrow = 2)
```



Sum of Squares method

```r
# combine within and between sum of squares in one data.frame
c_ssc <- data.frame(
  kmeans = c(2,3,4,5),
  within_ss = c(mean(c_k2$withinss), mean(c_k3$withinss), mean(c_k4$withinss), mean(c_k5$withinss)),
  between_ss = c(c_k2$betweenss, c_k3$betweenss, c_k4$betweenss, c_k5$betweenss))

# use pivot_longer to make values together
c_ssc_pivot <- pivot_longer(c_ssc, cols = c("within_ss", "between_ss"),
                            names_to="variable", values_to = "value")

# plot bar-plot
ggplot(c_ssc_pivot, aes(x = kmeans, y = log10(value), fill = variable)) +
    geom_bar(stat='identity', position='dodge') +
    ggtitle("c segment") +      # name the title
    xlab("Numbers of Clusters") +        # name x axis
    ylab("LOG(between/within_ss)") +     # name y axis
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),  # y axis name's color & size
          axis.text.x = element_text(size=8),    # change x axis number's size
          axis.text.y = element_text(size=8),    # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size
```

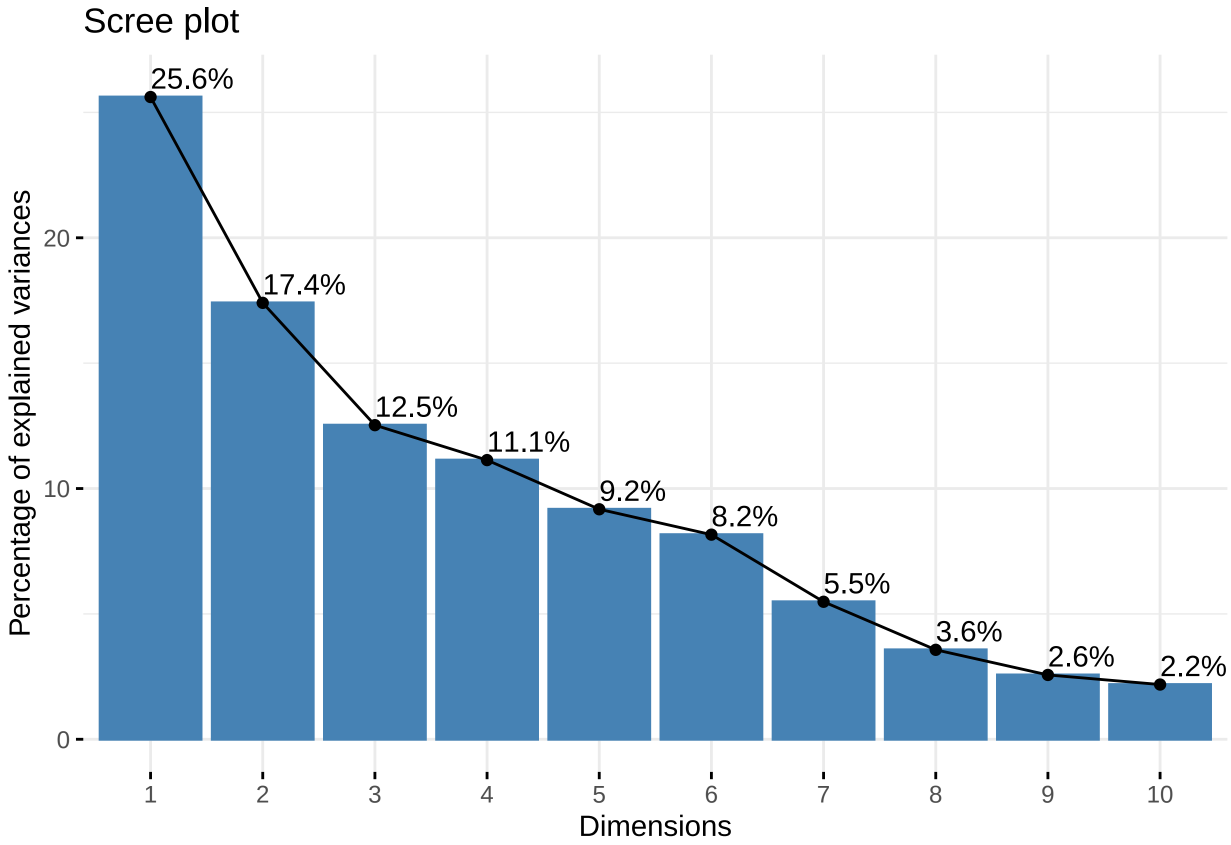


Based on cluster size and 3 methods, I choose k = 4.

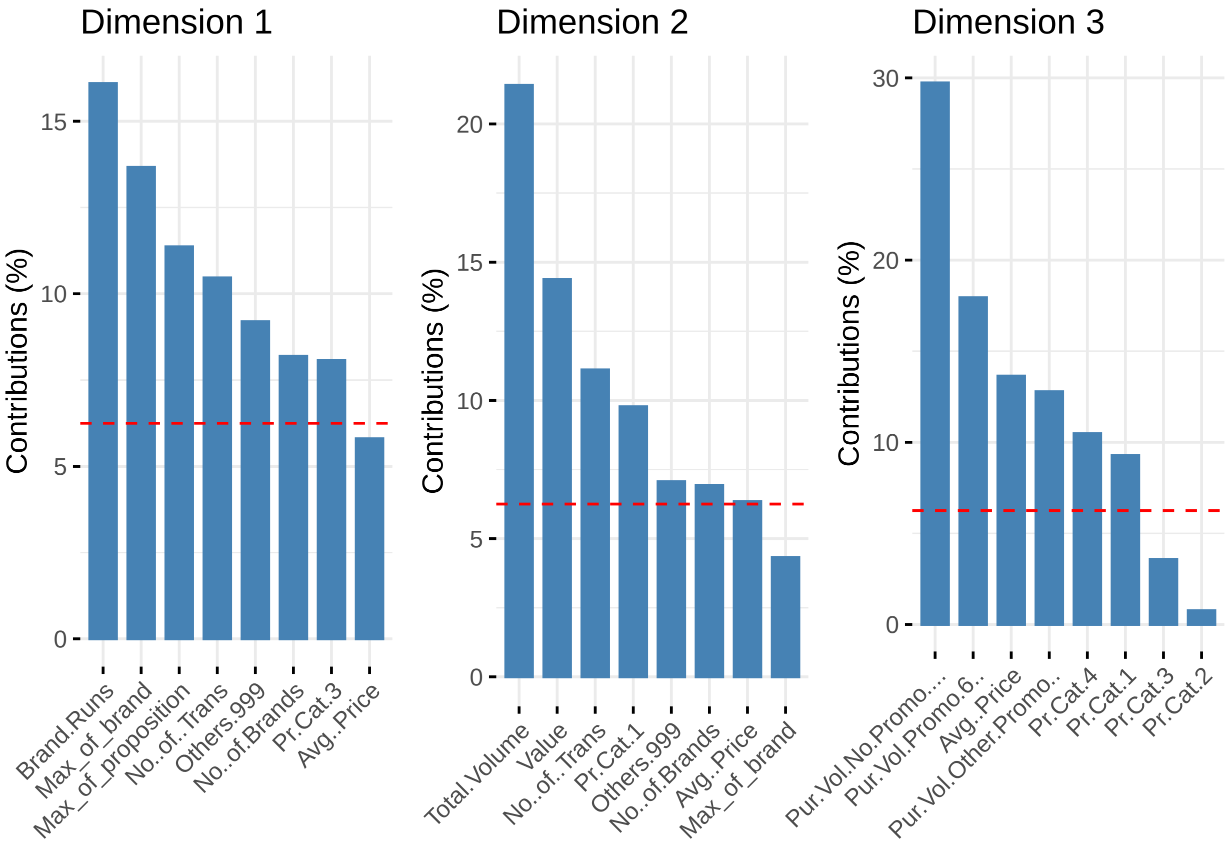
Best segmentation

Dimension reduction can help with choose variables like the PCA(principle components analysis).

```r
# build PCA
c_res.pca <- PCA(data_norm[,c(11:15, 18:21, 30:34, 46, 47)],  graph = FALSE)
# Visualize percentage of explained variables
fviz_screeplot(c_res.pca, addlabels = TRUE, ylim = c(0, 26))
```



```r
# Contributions of variables to PC1
Contri_1 <- fviz_contrib(c_res.pca, choice = "var", axes = 1, top = 8, title = "Dimension 1")
# Contributions of variables to PC2
Contri_2 <- fviz_contrib(c_res.pca, choice = "var", axes = 2, top = 8, title = "Dimension 2")
# Contributions of variables to PC3
Contri_3 <- fviz_contrib(c_res.pca, choice = "var", axes = 3, top = 8, title = "Dimension 3")

# display different clusters together
grid.arrange(Contri_1, Contri_2, Contri_3, nrow = 1)
```

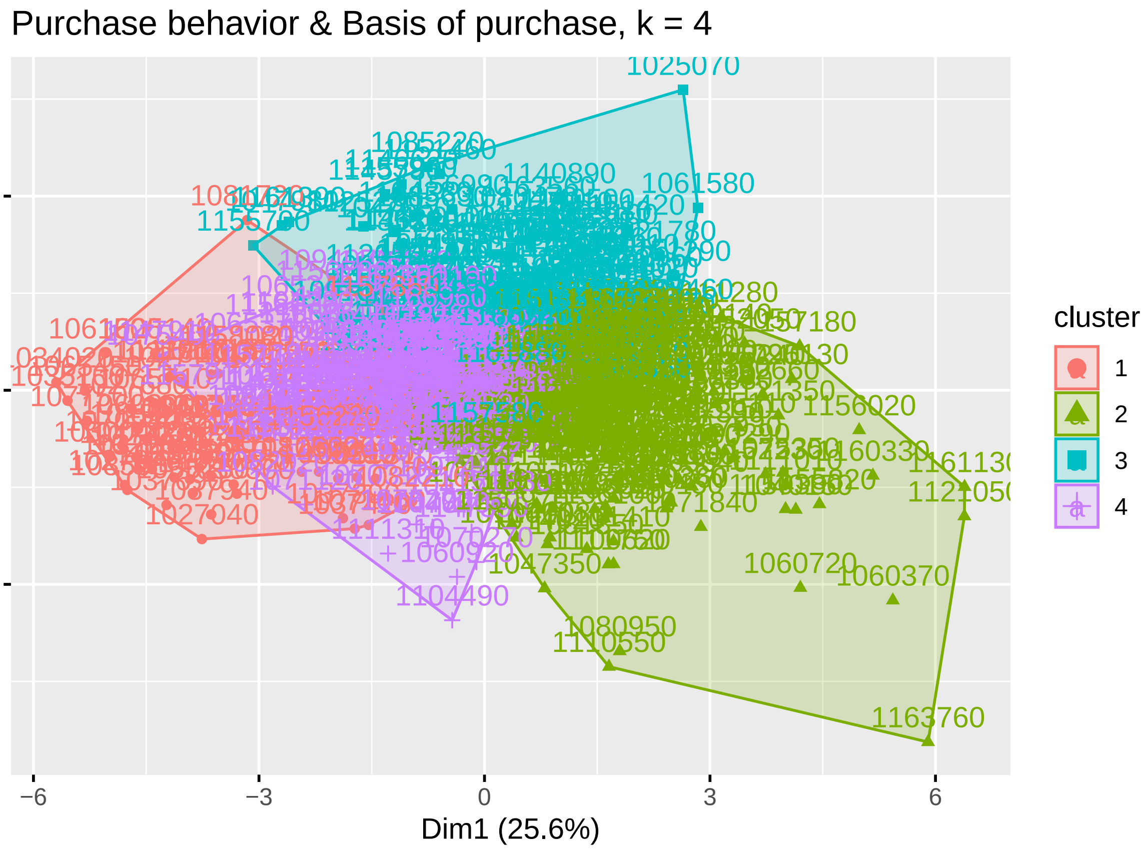


Dimension 1,2,3 contribute nearly 60% of all the variables. The main variables for these 3 dimensions include both a & b segments variables.

Since both a & b variables are important for the model, I choose the c segmentation with cluster numbers k = 4.

Visualize the best cluster model

```
fviz_cluster(c_k4, data = data_norm[,c(11:15, 18:21, 30:34, 46, 47)],
                        main = "Purchase behavior & Basis of purchase, k = 4")
```



```
# each cluster size
c_k4$size
```

```
[1]  69 202 154 175
```

IV. Comment clusters

(demographic, brand loyalty, basis for purchase)

As I mentioned above, demographic variables have 1/6 missing (which is “0”).

Let’s remove the missing data in demographic and visualize these variables.

Deal with demographic

Variables explanation:

SEC : Socioeconomic class(1 = high, 5 = low)

FEH : Eating habits(1 = vegetarian, 2 = vegetarian but eggs, 3 = non-vegetarian)

MT : Native language

SEX : Gender of homemaker(1 = male, 2 = female)

AGE : Age of homemaker

EDU : Education of homemaker(1 = min, 9 = max)

HS : Number of members in household

CHILD : Presence of children in household(4 categories)

CS : Television availability(1 = available, 2 = unavailable)

Affluence Index : Weighted value of durable possessed

Removing missing data in demographic

```r
# add cluster to data
data_norm$cluster <- c_k4$cluster
# transfer cluster to factor
# data_norm$cluster <- factor(data_norm$cluster)

# change missing data 0 to NA , 0 means NA
data_norm[,c(1:10)][data_norm[,c(1:10)] == 0] <- NA
# remove missing data
data_norm_remove <- data_norm[complete.cases(data_norm),]   # 495 rest
```

Visual demographic 10 variables :

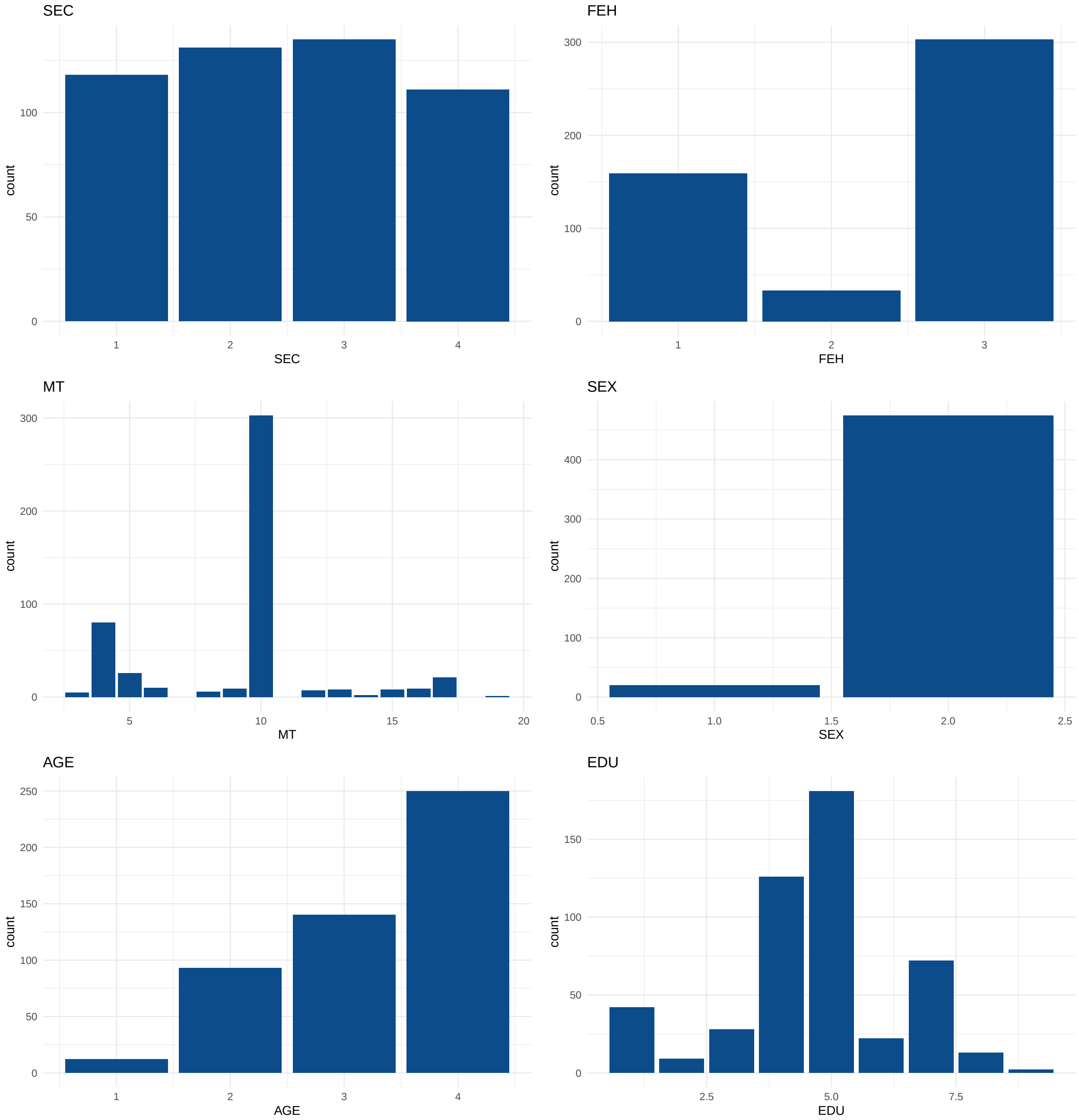
For SEX, almost all the homemakers are female.

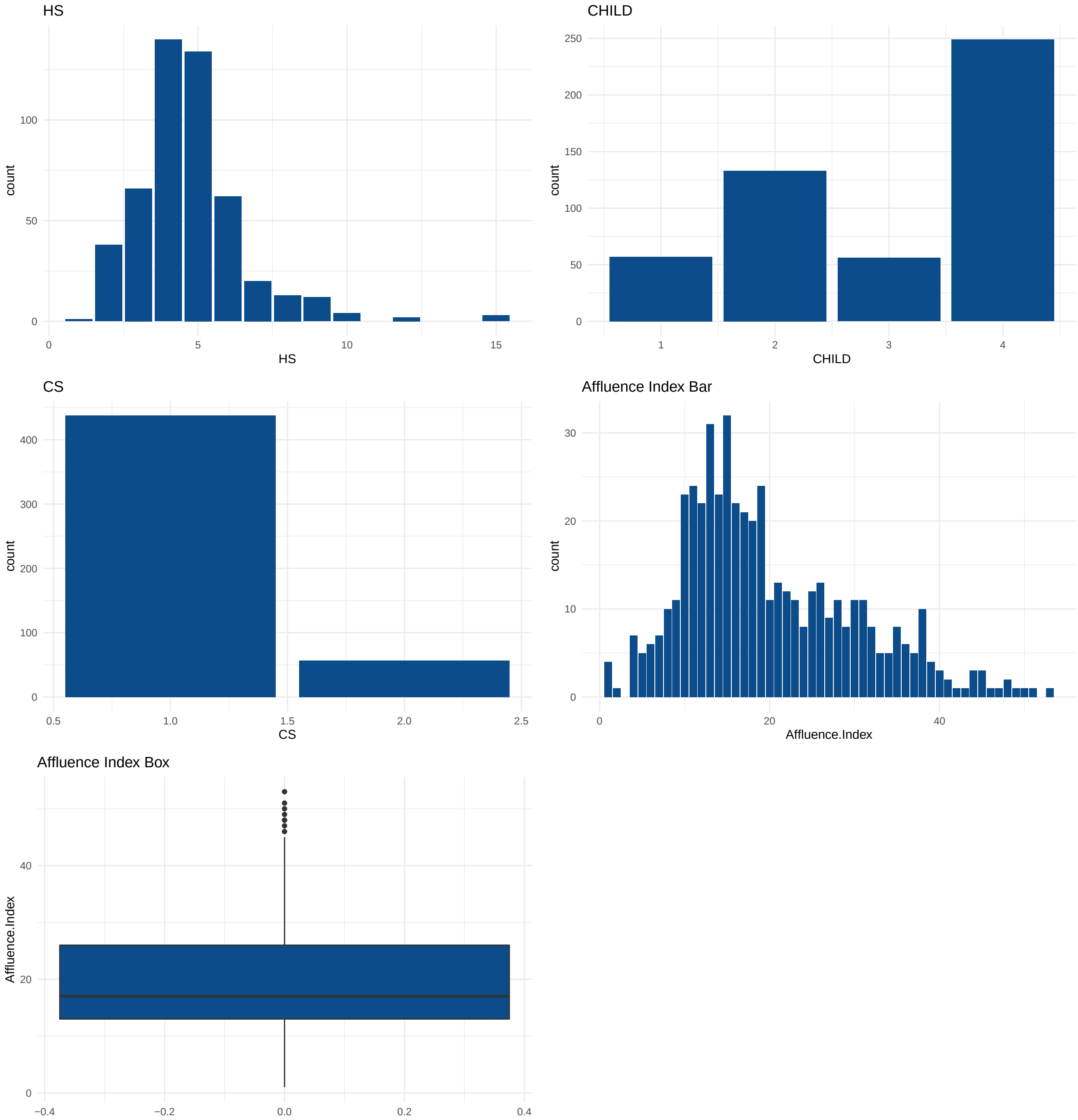
For CS, almost all the households with television available.

For Affluence.Index, many households is from 10 to 20 and there are several outliers.

For EDU, most homemaker have the average education level.

For MT, many households say the N0.10 language.


SEC
count
0
50
100
1
2
3
4
SEC
FEH
count
0
100
200
300
1
2
3
FEH
MT
count
0
100
200
300
5
10
15
20
MT
SEX
count
0
100
200
300
400
0.5
1.0
1.5
2.0
2.5
SEX
AGE
count
0
50
100
150
200
250
1
2
3
4
AGE
EDU
count
0
50
100
150
2.5
5.0
7.5
EDU



Since almost all the homemakers are female and almost all household are television available, it's unnecessary to analyze the SEX and CS.

I choose only 4/10 demographic variables to comment because the rest variables cannot tell any difference among the 4 clusters.

Demographic

separate the clusters

```
# divide into different clusters
cluster_1 <- data_norm_remove[which(data_norm_remove$cluster == 1),] # size 43
```

```
cluster_2 <- data_norm_remove[which(data_norm_remove$cluster == 2),] # size 185
cluster_3 <- data_norm_remove[which(data_norm_remove$cluster == 3),] # size 116
cluster_4 <- data_norm_remove[which(data_norm_remove$cluster == 4),] # size 151
```

Affluence Index

```
# Affluence Index
Afflu <- rbind(median(cluster_1$Affluence.Index), median(cluster_2$Affluence.Index),
                          median(cluster_3$Affluence.Index), median(cluster_4$Affluence.Index))

# change to data frame
Afflu_median <- data.frame(Afflu)
# add cluster
Afflu_median$cluster <- c(1,2,3,4)
# change cluster to factor
Afflu_median$cluster <- factor(Afflu_median$cluster)
```

SEC, EDU, HS

```
# change to factor
data_norm_remove$SEC <- factor(data_norm_remove$SEC)
data_norm_remove$EDU <- factor(data_norm_remove$EDU)
data_norm_remove$HS <- factor(data_norm_remove$HS)

# make the data and x axis
p <- ggplot(data = data_norm_remove, aes(x = cluster))

# SEC(Socioeconomics class)
h <- p + geom_histogram(binwidth = 0.2, aes(fill = SEC), color = "Black")
T1 <- h + xlab("Cluster") +     # name x axis
    ylab("count") +   # name y axis
    ggtitle("socioeconomics class") +    # name the figure title
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),# y axis name's color & size
          axis.text.x = element_text(size=8),    # change x axis number's size
          axis.text.y = element_text(size=8),    # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size

# EDU(education)
h <- p + geom_histogram(binwidth = 0.2, aes(fill = EDU), color = "Black")
T2 <- h + xlab("Cluster") +     # name x axis
    ylab("count") +   # name y axis
    ggtitle("education") +    # name the figure title
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),# y axis name's color & size
          axis.text.x = element_text(size=8),    # change x axis number's size
          axis.text.y = element_text(size=8),    # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size

# HS(numbers in house)
h <- p + geom_histogram(binwidth = 0.2, aes(fill = HS), color = "Black")
T3 <- h + xlab("Cluster") +     # name x axis
    ylab("count") +   # name y axis
    ggtitle("numbers in house") +    # name the figure title
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),# y axis name's color & size
          axis.text.x = element_text(size=8),    # change x axis number's size
          axis.text.y = element_text(size=8),    # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size
```

Brand loyalty & Basis purchase

```
# transfer centers to data frame
centers_c <- data.frame(c_k4$centers)

# add cluster
centers_c$cluster <- c(1,2,3,4)
# transfer cluster to factor
centers_c$cluster <- factor(centers_c$cluster)

# choose useful variables
centers_c_select <- centers_c[,c(7:16, 17)]

# use the pivot_longer to make variables together
centers_c_plot <- centers_c_select %>% select(cluster, Pur.Vol.No.Promo....,
        Pur.Vol.Promo.6.., Pur.Vol.Other.Promo.., Others.999, Pr.Cat.1, Pr.Cat.2,
                        Pr.Cat.3, Pr.Cat.4, Max_of_brand, Max_of_proposition) %>%
        pivot_longer(Pur.Vol.No.Promo.... : Max_of_proposition,
                                     names_to = "variables", values_to = "value")

# use ggplot to plot the picture
t <- ggplot(centers_c_plot, aes(variables, value, colour = cluster)) + geom_point()
final_plot <- t + xlab("Variables") +    # name x axis
    ylab("Normalized values") +   # name y axis
    ggtitle("Brand loyalty & Basis of purchase") +   # name the title
    theme(axis.title.x = element_text(color="DarkGreen", size=12),# x axis name's color & size
          axis.title.y = element_text(color="Darkred", size=12),  # y axis name's color & size
          axis.text.x = element_text(size=8, angle=45, vjust=1, hjust=1),   # x axis word's size
          axis.text.y = element_text(size=8),   # change y axis number's size
          plot.title = element_text(color="DarkBlue", size=15)) # title's color & size
```

Comments each cluster

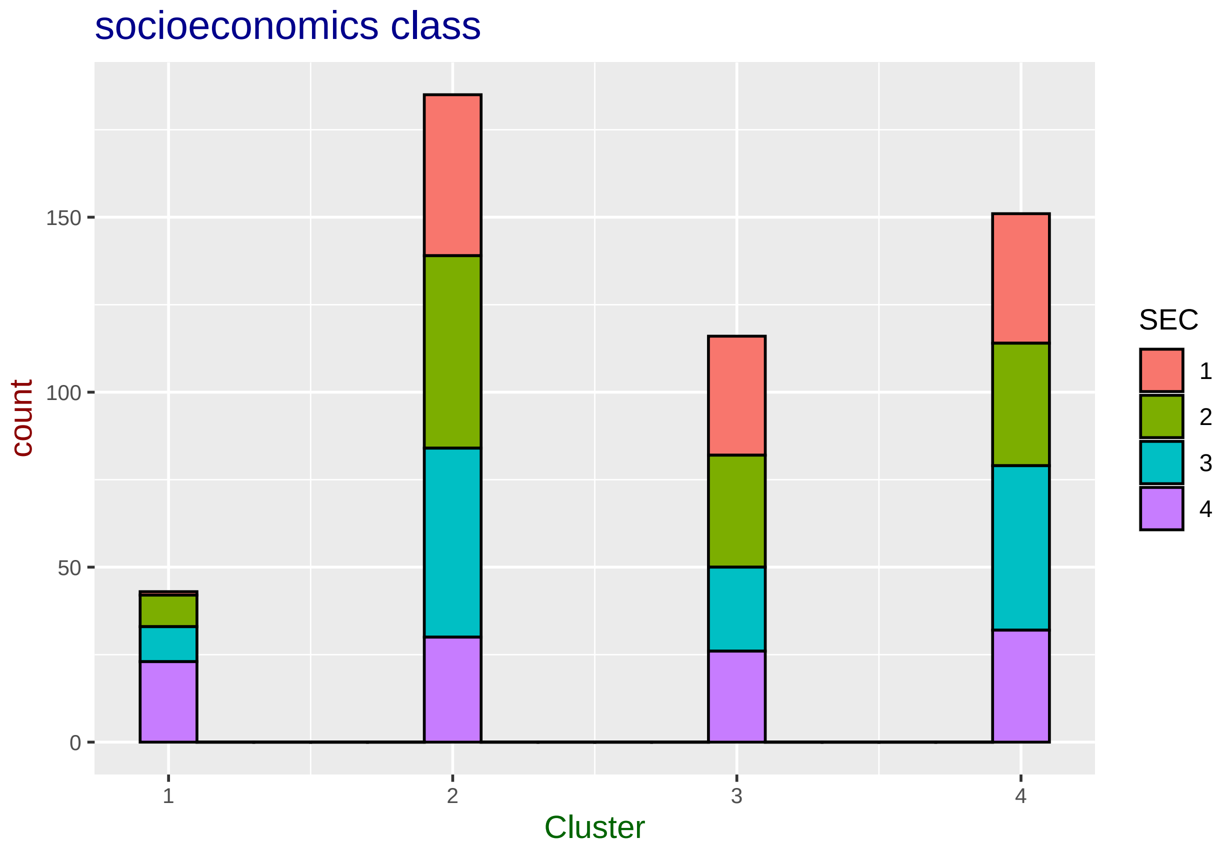
show results

```
# Affluence Index
Afflu_median
```

```
##   Afflu cluster
## 1    11       1
## 2    19       2
## 3    18       3
## 4    16       4
```

SEC

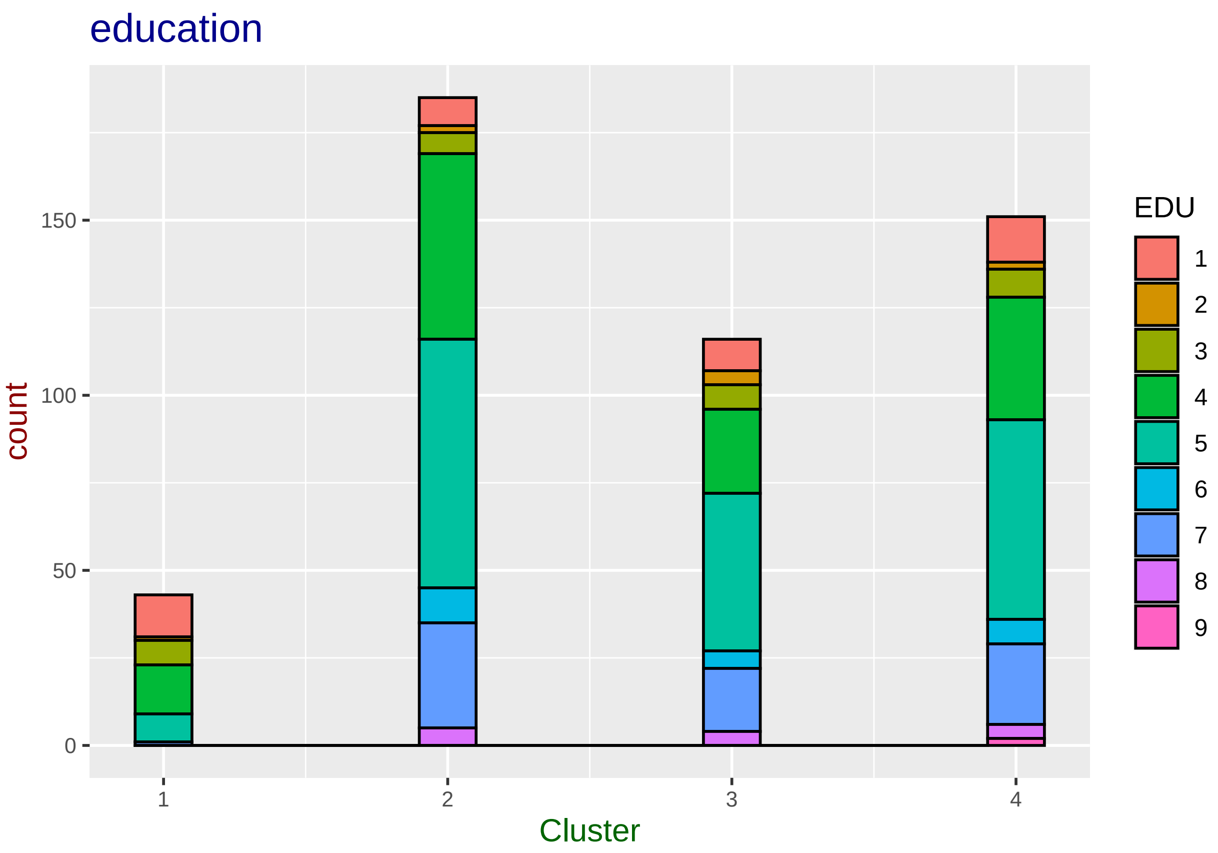
```
T1
```



Cluster 1 have low socioeconomics among the clusters.

EDU

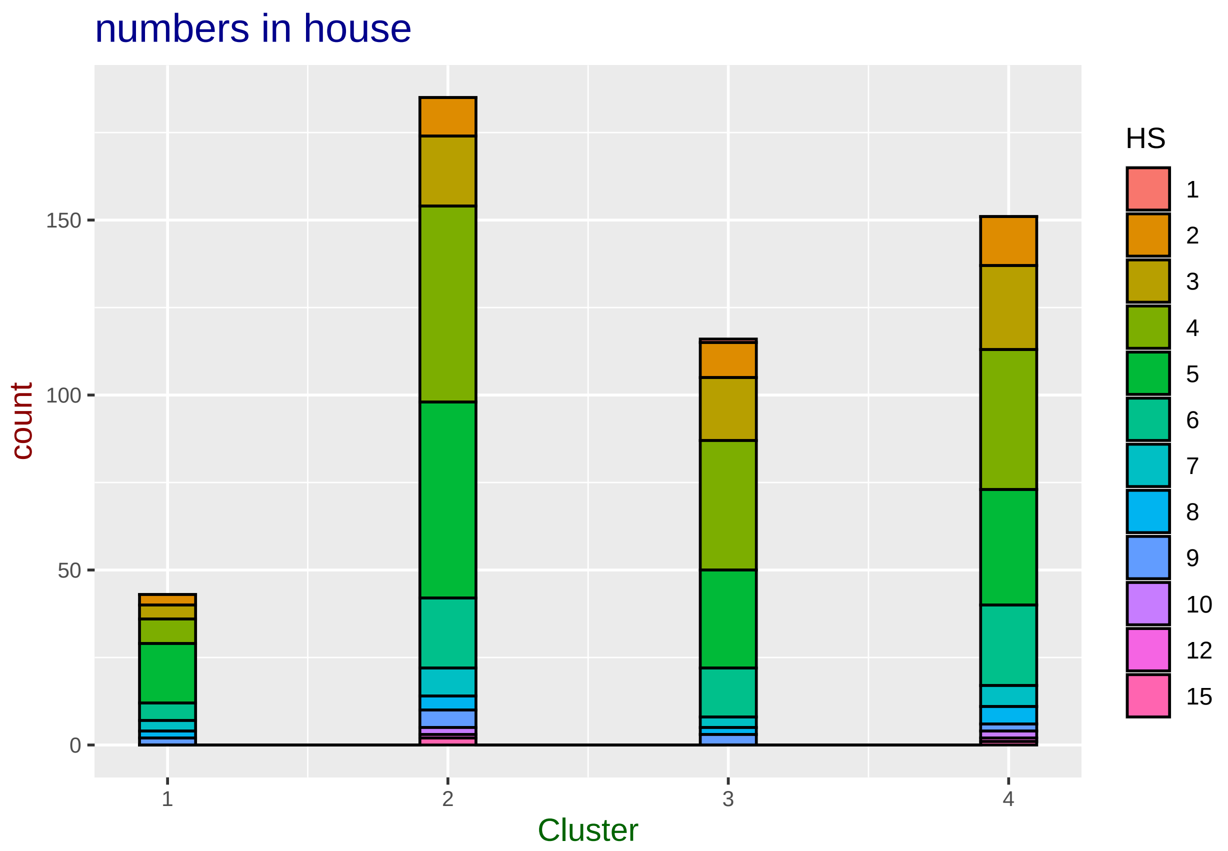
```
T2
```



Cluster 1 has low education level and cluster 4 has the highest education level 9 occurs.

HS

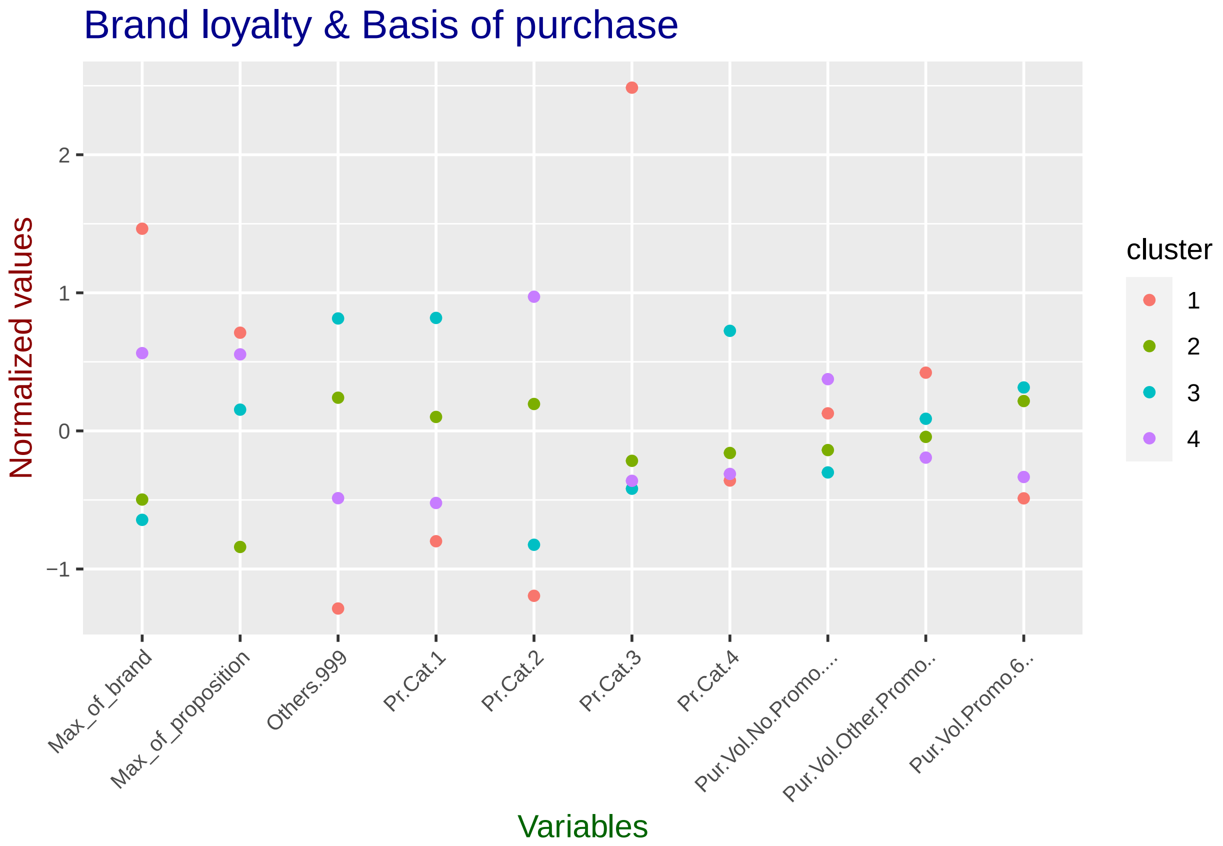
```
T3
```



Cluster 1 have the most average numbers of people in household. And cluster 2 & 4 occurs the categorical for 12 and 15.

Brand loyalty & Basis of purchase

```
final_plot
```



Comments

---

cluster 1 :

Socioeconomics is low.

Homemakers average education level is low.

The number of people in household is about 5.

The median of Affluence Index is 11.

They have the highest other promotion among other clusters but the least promotion 6 and relatively high purchase on no promotion.

They have super high purchase under price category 3 and also highest response to proposition.

They also have the highest brand loyalty for the selected brands and the least brand loyalty for others 999.

cluster 2 :

Socioeconomics is relatively higher.

Homemakers average education levels are relatively high.

The median of Affluence Index is 19.

The number of people in household is about 4 and 5.

They have lowest response to selling proposition.

Many of them seems to purchase on promotions 6.

They have a relatively low brand loyalty for the selected brands but have a relatively high brand loyalty for brands in others 999.

---

cluster 3 :

The median of Affluence Index is 18.

The number of people in household is about 4 and 5.

They purchase highest on promotion 6 and relatively high in other promotion. Cluster 3 also has the least on no promotion.

They have highest purchase under price category 1&4.

They have the lowest brand loyalty for the selected brand and have the highest brand loyalty for others 999.

---

cluster 4 :

The median of Affluence Index is 16.

The highest education level 9 occurs.

The number of people in household is about 4 and 5.

The purchases in this cluster have the highest purchase on no promotions. They purchase lowest on other promotion and relatively low on promotion 6. They have highest purchase under price category 2.

They have the high brand loyalty for the selected brands and the relatively low brand loyalty for others 999.

V. Develop a model

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.

The unsupervised machine learning methods we learned in class are k-means clustering, DBSCAN(density based spatial clustering of application with noise), and hierarchical clustering.

The K-means model we have built above, and I will build DBSCAN & Hierarchical clustering as follows.

Hierarchical

I plot dendragram to extract 4 clusters. The cluster performance is really bad, there is really few numbers(even cannot see it clearly) in the first cluster.

The hierarchical clustering is not suitable for this dataset.

```r
d <- dist(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], method = "euclidean")
# compute divisive hierarchical clustering
hc_complete <- hclust(d, method = "complete")
# plot dendrogram
#plot(hc_complete, cex = 0.05)
#rect.hclust(hc_complete, k = 4, border = 1:4)
```

Due to increasing commercialization, consumer data is increasing exponentially.

Clustering models need to possess the capability to process this enormous data effectively.

The paper shows K-Means clustering gives better performance for a large number of observations while hierarchical clustering has the ability to handle fewer data points.

DBSCAN

I choose many different value for epsilon and min-points, but the noise are more than 500 since we only have 600 households.

The bad performance happened maybe because of the high dimensionality since DBSCAN is not appropriate for high dimensional dataset.

```r
# make sure for the same random numbers
# set.seed(123)
# db_cluster <- dbscan :: dbscan(data_norm[,c(11:15, 18:21, 30:34, 46, 47)], eps = 0.3, minPts = 5)
# db_cluster
```

VI. Conclusion

Since DBSCAN and Hierarchical Clustering is not suitable for this dataset, finally K-means could be the best model!

Households have highest brand loyalty would be defined as a success category.

So our promising group is cluster 1 which have low socioeconomics, low education level , low affluence index, and more numbers of people in a household.

Since it has the highest brand loyalty, so the company should build more promotion strategies to them.

Reference

[1]. 10 Tips for Choosing the Optimal Number of Clusters. Matt.O(2019) Retrieved from : https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92

[2]. TRIPATHI, Shreya; BHARDWAJ, Aditya; E, Poovammal. Approaches to Clustering in Customer Segmentation. International Journal of Engineering & Technology, [S.l.], v. 7, n. 3.12, p. 802-807, 2018.