Machlearn_final

Abhishek_Sau

01/05/2021

```r
setwd("D:/Study/Assignments/MachLearn/MachLearn_final")
soap_df<- read.csv("BathSoap.csv")
```

Data Cleaning

```r
library(tidycomm)
```

```
## Warning: package 'tidycomm' was built under R version 4.0.5
```

```r
library(imputeMissings)
```

```
## Warning: package 'imputeMissings' was built under R version 4.0.5
```

```r
describe(soap_df)
```

```
## Warning: The `add` argument of `group_by()` is deprecated as of dplyr 1.0.0.
## Please use the `.add` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
## # A tibble: 19 x 13
##    Variable     N Missing       M      SD    Min    Q25    Mdn    Q75    Max
##    <chr>    <int>   <int>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Member.~   600       0 1.10e+6 4.56e+4 1.01e6 1.07e6 1.11e6 1.15e6 1.17e6
##  2 SEC        600       0 2.50e+0 1.12e+0 1.00e0 1.75e0 2.50e0 3.25e0 4.00e0
##  3 FEH        600       0 2.05e+0 1.13e+0 0.     1.00e0 3.00e0 3.00e0 3.00e0
##  4 MT         600       0 8.18e+0 4.29e+0 0.     4.00e0 1.00e1 1.00e1 1.90e1
##  5 SEX        600       0 1.74e+0 6.49e-1 0.     2.00e0 2.00e0 2.00e0 2.00e0
##  6 AGE        600       0 3.21e+0 8.65e-1 1.00e0 3.00e0 3.00e0 4.00e0 4.00e0
##  7 EDU        600       0 4.04e+0 2.19e+0 0.     3.00e0 4.50e0 5.00e0 9.00e0
##  8 HS         600       0 4.19e+0 2.30e+0 0.     3.00e0 4.00e0 5.00e0 1.50e1
##  9 CHILD      600       0 3.23e+0 1.22e+0 1.00e0 2.00e0 4.00e0 4.00e0 5.00e0
## 10 CS         600       0 9.32e-1 5.07e-1 0.     1.00e0 1.00e0 1.00e0 2.00e0
## 11 Affluen~   600       0 1.70e+1 1.14e+1 0.     1.00e1 1.50e1 2.40e1 5.30e1
## 12 No..of.~   600       0 3.64e+0 1.58e+0 1.00e0 2.00e0 3.00e0 5.00e0 9.00e0
## 13 Brand.R~   600       0 1.58e+1 1.04e+1 1.00e0 8.00e0 1.50e1 2.10e1 7.40e1
## 14 Total.V~   600       0 1.19e+4 7.77e+3 1.50e2 6.82e3 1.04e4 1.53e4 5.09e4
## 15 No..of.~   600       0 3.12e+1 1.74e+1 1.00e0 2.20e1 2.80e1 4.00e1 1.38e2
## 16 Value      600       0 1.34e+3 8.83e+2 2.00e1 7.90e2 1.22e3 1.68e3 6.37e3
## 17 Trans..~   600       0 2.62e+0 2.60e+0 1.00e0 1.42e0 1.85e0 2.69e0 2.30e1
## 18 Vol.Tran   600       0 4.15e+2 2.49e+2 9.44e1 2.51e2 3.62e2 4.91e2 2.52e3
## 19 Avg..Pr~   600       0 1.18e+1 3.74e+0 5.62e0 9.76e0 1.12e1 1.34e1 3.33e1
## # ... with 3 more variables: Range <dbl>, Skewness <dbl>, Kurtosis <dbl>
```

There are no missing values within the dataset. However, there are Zeros which can be treated as NAs as they don't hold meaning with respect to certain columns.

Data Transformation

```r
#Changing datatypes

soap_df$Member.id <- as.factor(soap_df$Member.id)
soap_df$SEC <- as.factor(soap_df$SEC)
soap_df$FEH <- as.factor(soap_df$FEH)
soap_df$MT <- as.factor(soap_df$MT)
soap_df$SEX <-as.factor(soap_df$SEX)
soap_df$AGE <- as.factor(soap_df$AGE)
soap_df$EDU <- as.factor(soap_df$EDU)
soap_df$HS <- as.factor(soap_df$HS)
soap_df$CHILD <- as.factor(soap_df$CHILD)
soap_df$CS <- as.factor(soap_df$CS)

#Changing datatype of columns related to purchase within promotions
soap_df$Pur.Vol.No.Promo....<- as.numeric(gsub("\\%", "", soap_df$Pur.Vol.No.Promo....))
soap_df$Pur.Vol.Other.Promo..<- as.numeric(gsub("\\%", "", soap_df$Pur.Vol.Other.Promo..))
soap_df$Pur.Vol.Promo.6.. <- as.numeric(gsub("\\%", "", soap_df$Pur.Vol.Promo.6..))

#Changing datatype of columns related to Brandwise Purchase
soap_df$Br..Cd..57..144 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..57..144))
soap_df$Br..Cd..55 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..55))
soap_df$Br..Cd..272 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..272))
soap_df$Br..Cd..286 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..286))
soap_df$Br..Cd..24 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..24))
soap_df$Br..Cd..481 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..481))
soap_df$Br..Cd..352 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..352))
soap_df$Br..Cd..5 <- as.numeric(gsub("\\%", "", soap_df$Br..Cd..5))
soap_df$Others.999 <- as.numeric(gsub("\\%", "", soap_df$Others.999))

#Changing datatype of columns related to Price category
soap_df$Pr.Cat.1 <- as.numeric(gsub("\\%", "", soap_df$Pr.Cat.1))
soap_df$Pr.Cat.2 <- as.numeric(gsub("\\%", "", soap_df$Pr.Cat.2))
soap_df$Pr.Cat.3 <- as.numeric(gsub("\\%", "", soap_df$Pr.Cat.3))
soap_df$Pr.Cat.4 <- as.numeric(gsub("\\%", "", soap_df$Pr.Cat.4))

#Changing datatype of columns related to Selling proposition wise purchase

soap_df$PropCat.5 <- as.numeric(gsub("\\%", "", soap_df$PropCat.5))
soap_df$PropCat.6 <- as.numeric(gsub("\\%", "", soap_df$PropCat.6))
soap_df$PropCat.7 <- as.numeric(gsub("\\%", "", soap_df$PropCat.7))
soap_df$PropCat.8 <- as.numeric(gsub("\\%", "", soap_df$PropCat.8))
soap_df$PropCat.9 <- as.numeric(gsub("\\%", "", soap_df$PropCat.9))
soap_df$PropCat.10 <- as.numeric(gsub("\\%", "", soap_df$PropCat.10))
soap_df$PropCat.11 <- as.numeric(gsub("\\%", "", soap_df$PropCat.11))
soap_df$PropCat.12 <- as.numeric(gsub("\\%", "", soap_df$PropCat.12))
soap_df$PropCat.13 <- as.numeric(gsub("\\%", "", soap_df$PropCat.13))
soap_df$PropCat.14 <- as.numeric(gsub("\\%", "", soap_df$PropCat.14))
soap_df$PropCat.15 <- as.numeric(gsub("\\%", "", soap_df$PropCat.15))
```

Imputing mode in place of zeros

```r
soap_df[,c(5,7,8,10)][soap_df[,c(5,7,8,10)] == 0] <- NA
colSums(is.na(soap_df))
```

```
##             Member.id                    SEC                    FEH
##                     0                      0                      0
##                    MT                    SEX                    AGE
##                     0                     68                      0
##                   EDU                     HS                  CHILD
##                    73                     68                      0
##                    CS        Affluence.Index          No..of.Brands
##                    99                      0                      0
##            Brand.Runs           Total.Volume          No..of..Trans
##                     0                      0                      0
##                 Value    Trans...Brand.Runs               Vol.Tran
##                     0                      0                      0
##            Avg..Price  Pur.Vol.No.Promo....     Pur.Vol.Promo.6..
##                     0                      0                      0
## Pur.Vol.Other.Promo..        Br..Cd..57..144            Br..Cd..55
##                     0                      0                      0
##           Br..Cd..272            Br..Cd..286            Br..Cd..24
##                     0                      0                      0
##           Br..Cd..481            Br..Cd..352             Br..Cd..5
##                     0                      0                      0
##            Others.999               Pr.Cat.1               Pr.Cat.2
##                     0                      0                      0
##              Pr.Cat.3               Pr.Cat.4              PropCat.5
##                     0                      0                      0
##             PropCat.6              PropCat.7              PropCat.8
##                     0                      0                      0
##             PropCat.9             PropCat.10             PropCat.11
##                     0                      0                      0
##            PropCat.12             PropCat.13             PropCat.14
##                     0                      0                      0
##            PropCat.15
##                     0
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 4.0.5
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following object is masked from 'package:imputeMissings':
##
##     impute
## The following object is masked from 'package:tidycomm':
##
##     describe
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
soap_df$SEX <- impute(soap_df$SEX, mode)
soap_df$EDU <- impute(soap_df$EDU, mode)
soap_df$HS <- impute(soap_df$HS, mode)
soap_df$CS <- impute(soap_df$CS, mode)
```

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

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

The variables that define purchase behavior are - Variables used for this process are: Transaction/BrandRun Number of brands Volume/Transaction Average Price Value Purchase witin promotions Others999 Maximum brand loyalty (a derived value column)

NOTE: The intricacies of marketing to 5 segments would probably not be supported by clustering just based on purchase behavior, or clustering just based on basis for purchase, so I will stick to 2-3 clusters for the initial variables variables.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::compute()   masks imputeMissings::compute()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::src()       masks Hmisc::src()
## x dplyr::summarize() masks Hmisc::summarize()
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.0.4

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

```
library(ISLR)
library(parcoords)
```

```
## Warning: package 'parcoords' was built under R version 4.0.5
```

```
library(GGally)
```

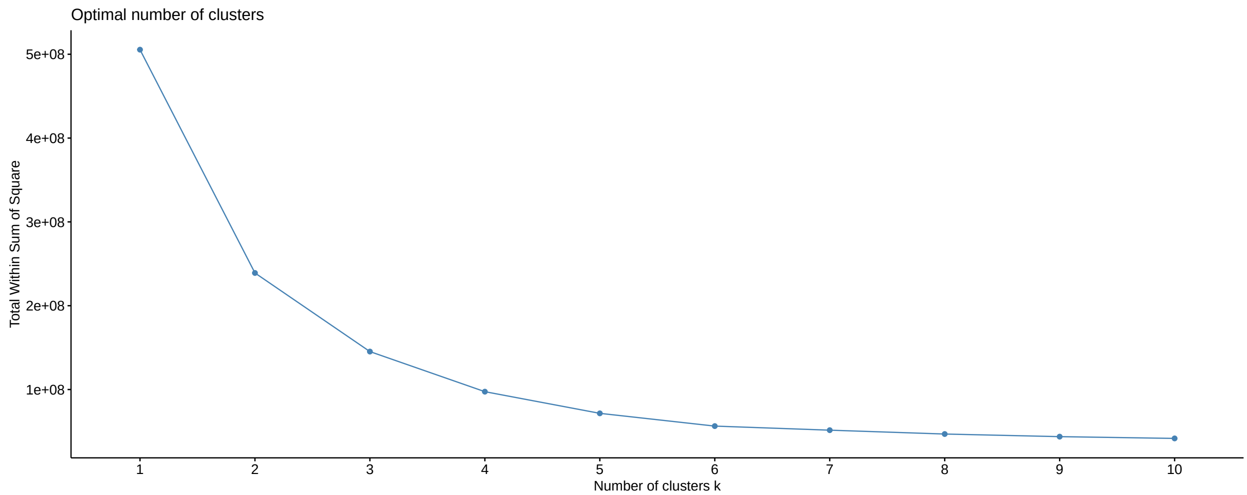
```
## Warning: package 'GGally' was built under R version 4.0.5

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

```
soap_df$max_brand_loy <- pmax(soap_df$Br..Cd..57..144, soap_df$Br..Cd..24,soap_df$Br..Cd..272, soap_df$B

library(caret)
```

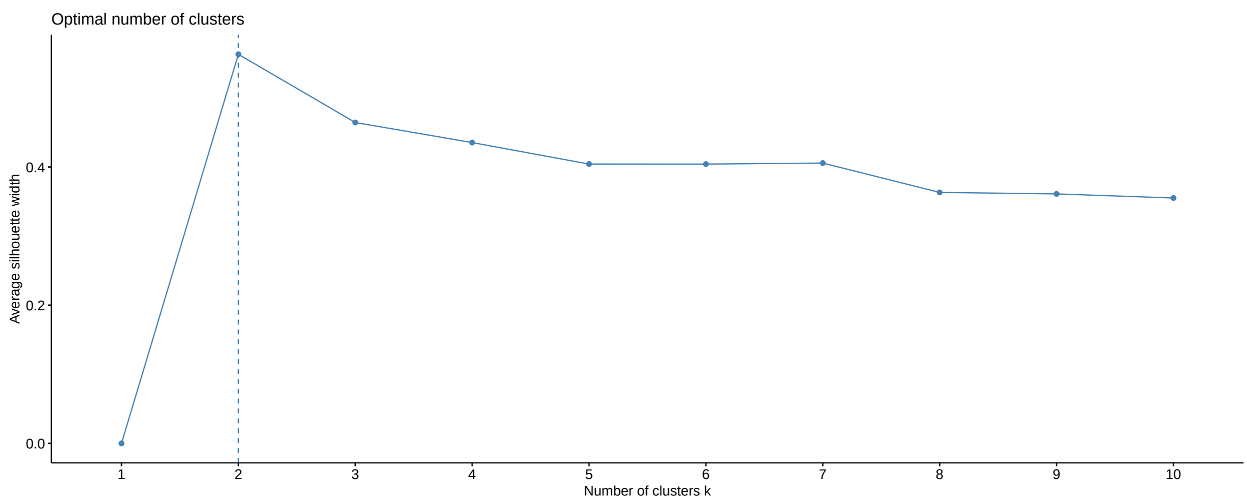
```
##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
```

```
##     lift
## The following object is masked from 'package:survival':
##
##     cluster
```

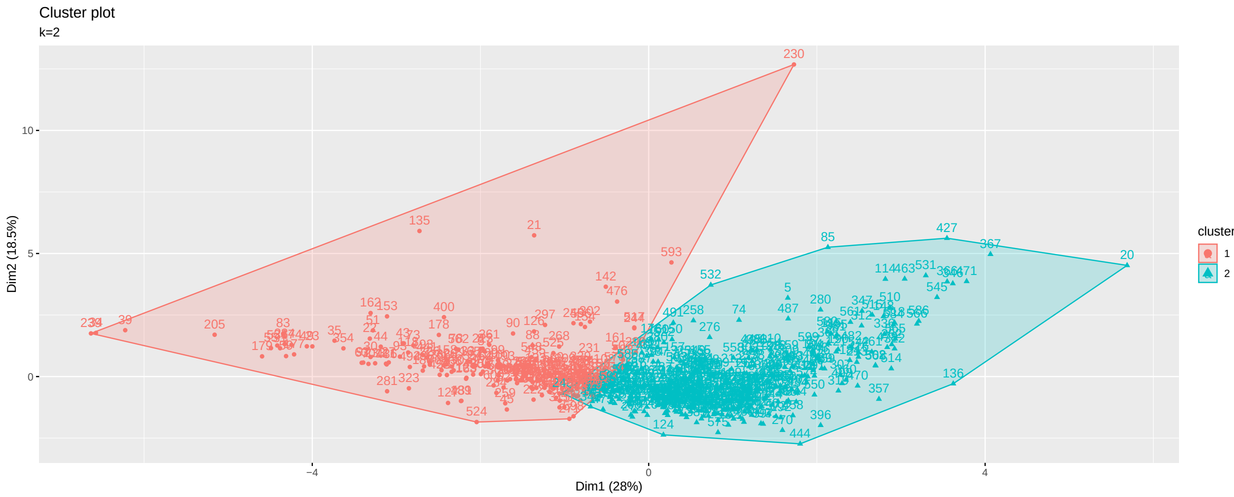
```r
prcs_bhvr <- soap_df[,c(12,16:22, 31, 47)]

fviz_nbclust(prcs_bhvr,kmeans,method = "wss")
```



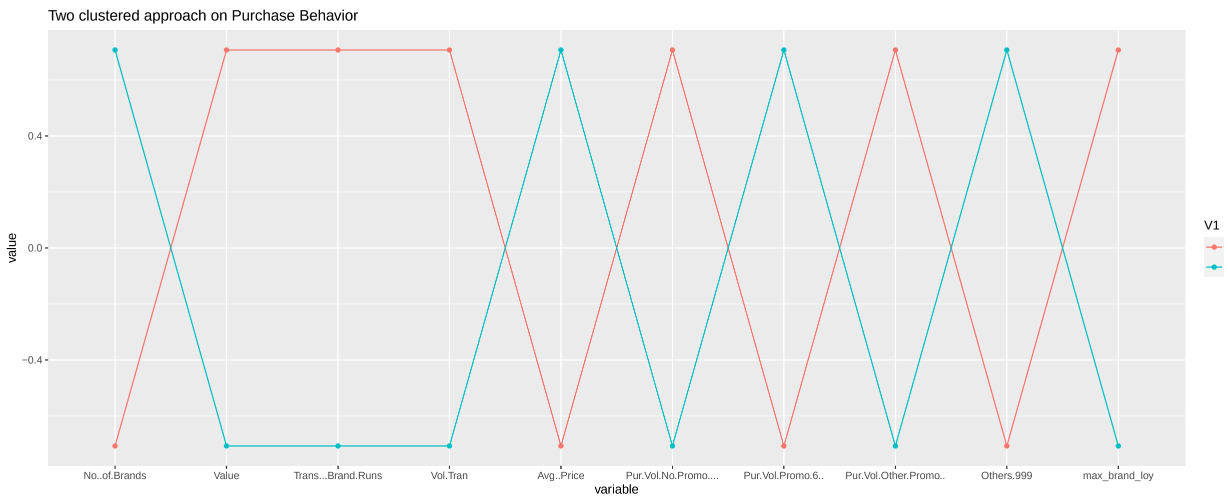

```r
fviz_nbclust(prcs_bhvr,kmeans,method = "silhouette")
```



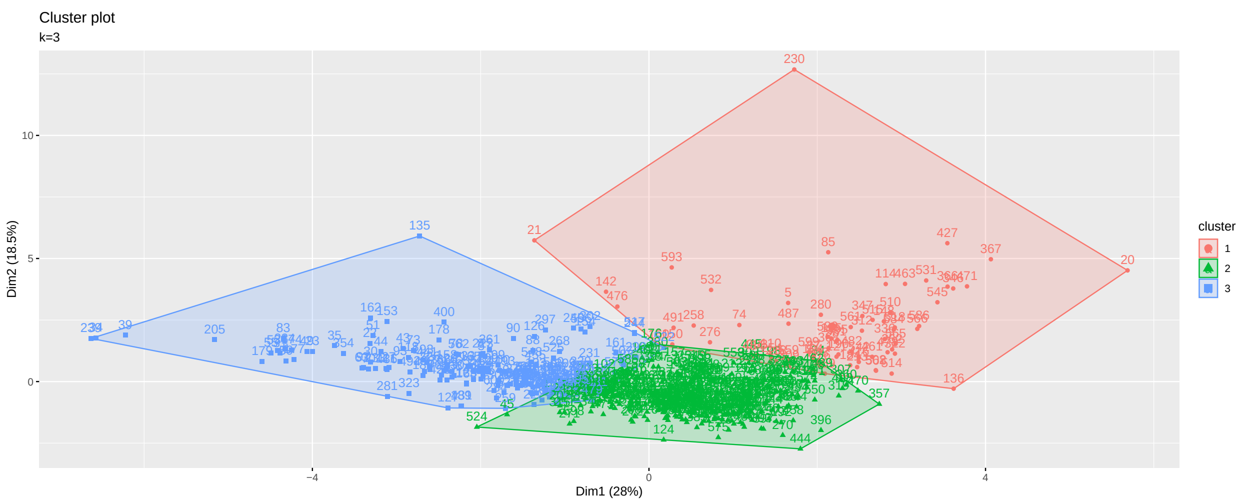

```r
set.seed(123)
norm_prcs_bhvr <- scale(prcs_bhvr)

prcs_means1 <- kmeans(norm_prcs_bhvr, centers = 2, nstart = 25)
fviz_cluster(prcs_means1, norm_prcs_bhvr)+ labs(subtitle = "k=2")
```



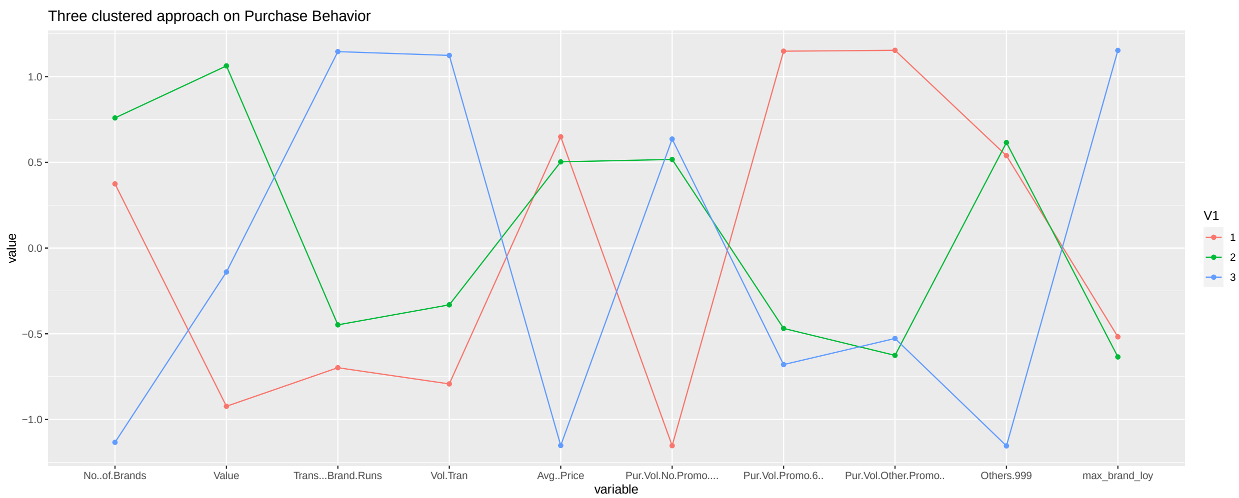

```
beh1 <- as.data.frame(cbind(1:nrow(prcs_means1$centers),prcs_means1$centers))
beh1$V1 <-as.factor(beh1$V1)

ggparcoord(data = beh1,
           columns = 2:11,groupColumn = 1,
           alphaLines = 1.0, showPoints = TRUE, title = "Two clustered approach on Purchase Behavior")
```




```
prcs_means2 <- kmeans(norm_prcs_bhvr, centers = 3, nstart = 25)
fviz_cluster(prcs_means2, norm_prcs_bhvr)+ labs(subtitle = "k=3")
```




```r
beh2 <- as.data.frame(cbind(1:nrow(prcs_means2$centers),prcs_means2$centers))
beh2$V1 <-as.factor(beh2$V1)

ggparcoord(data = beh2,
           columns = 2:11,
           groupColumn = 1,
           alphaLines = 1.0,showPoints = TRUE, title = "Three clustered approach on Purchase Behavior")
```



For k=2, Cluster size - 391,209

Cluster 1 Characteristics - Low brand loyalty, Low average volume per transaction, Low activity i.e less no.of transaction and Less brand runs

Cluster 2 Characteristics - High brand loyalty, High Average volume per transaction, High activty i.e more number of purchases or transactions made

For k=3

Cluster 1 Characteristics - High Brand loyalty, Purchasing from only a few main brands, Purchasing in big volume per transaction.

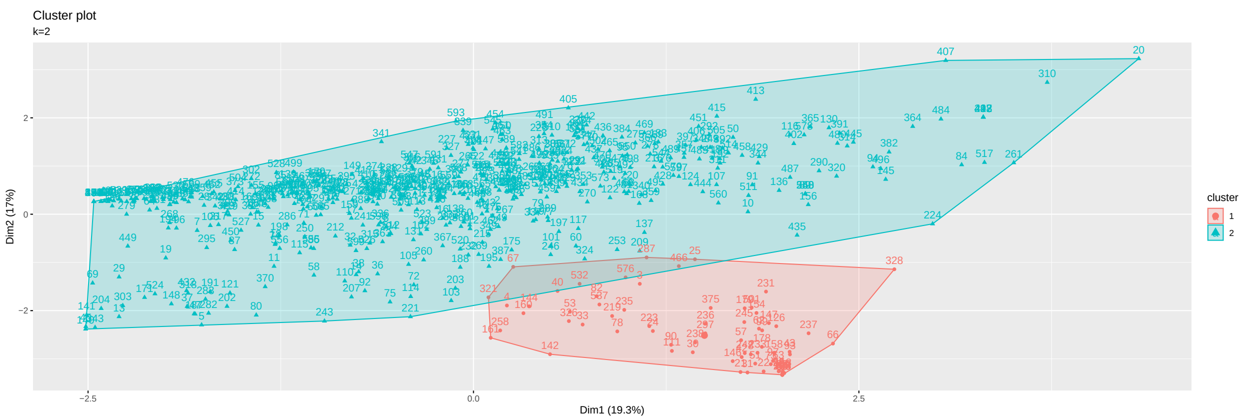
Cluster 2 characteristics - Lowest valued customers with no Brand loyalty and Purchases made from multiple brands

Cluster 3 characteristics - Customers in this cluster aren't very loyal, have the highest value but purchases are very small in volume per transaction

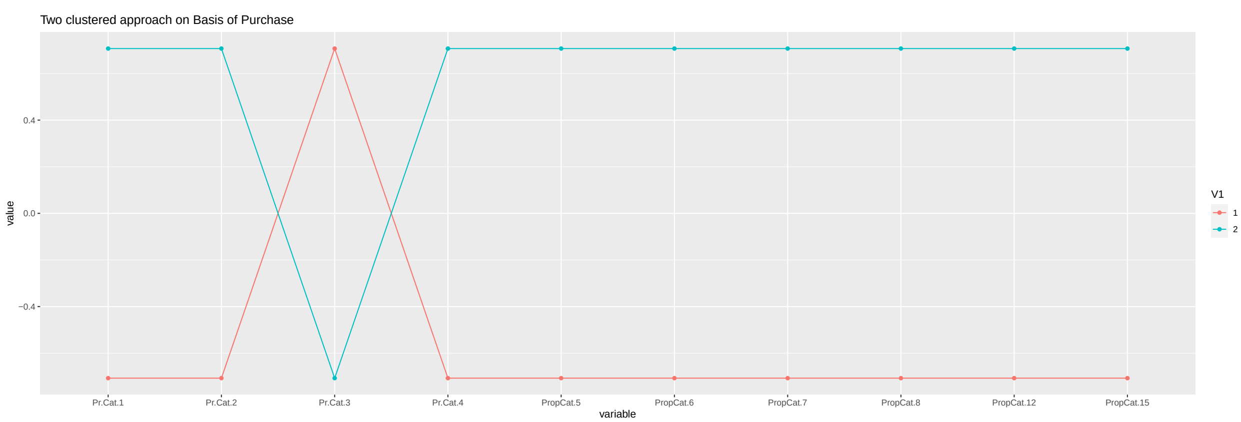
b.The variables that describe the basis for purchase

I observed that PropCat 9-11 & PropCat 13-14 have minimum distribution patterns for this variables. So I have considered only PropCat 5 – 8, PropCat 12, PropCat 15.

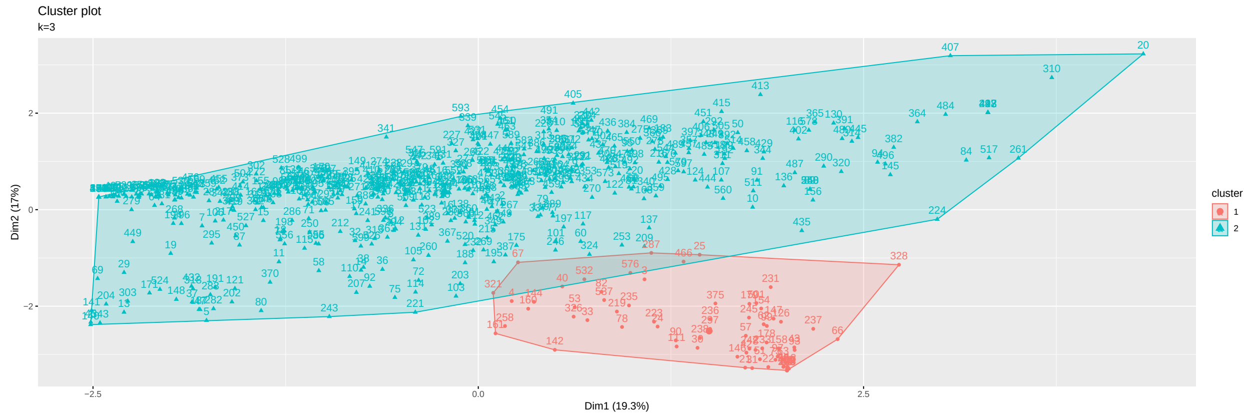
```
bas_prcs <- soap_df[, c(32:39,43,46)]

norm_bas_prcs <- as.data.frame(scale(bas_prcs))

bop_means1 <- kmeans(norm_bas_prcs, centers = 2, nstart = 25)
fviz_cluster(bop_means1, norm_bas_prcs)+ labs(subtitle = "k=2")
```



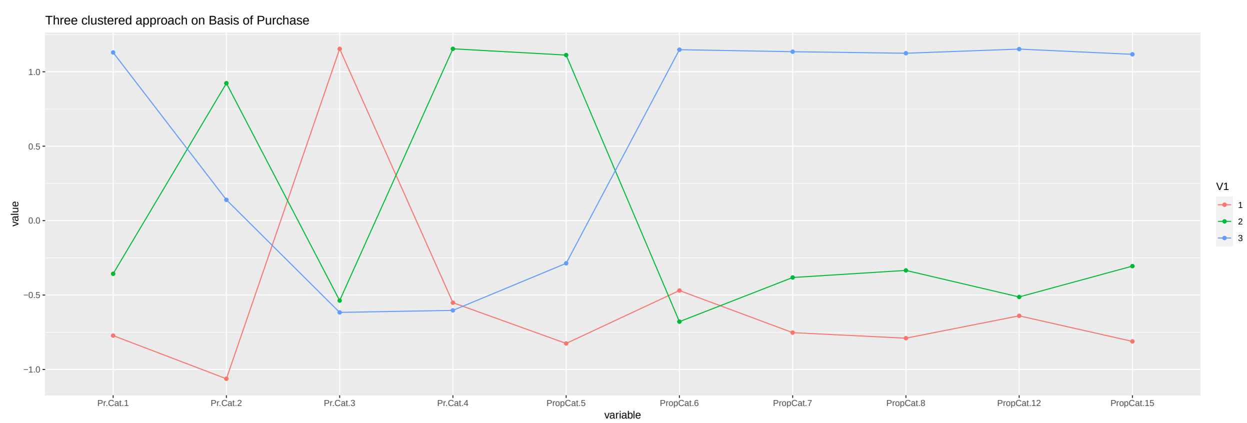

```
bop1 <- as.data.frame(cbind(1:nrow(bop_means1$centers),bop_means1$centers))
bop1$V1 <-as.factor(bop1$V1)

ggparcoord(data = bop1,
           columns = 2:11,
           groupColumn = 1,
           alphaLines = 1,
           showPoints = TRUE, title = "Two clustered approach on Basis of Purchase")
```




```
bop_means2 <- kmeans(norm_bas_prcs, centers = 3, nstart = 25)
fviz_cluster(bop_means1, norm_bas_prcs)+ labs(subtitle = "k=3")
```




```r
bop2 <- as.data.frame(cbind(1:nrow(bop_means2$centers),bop_means2$centers))
bop2$V1 <-as.factor(bop2$V1)

ggparcoord(data = bop2,
           columns = 2:11,
           groupColumn = 1,
           alphaLines = 1.0,
           showPoints = TRUE, , title = "Three clustered approach on Basis of Purchase")
```



k=2

Cluster 1 Characteristics- High no. of Purchases made in No promotion followed by purchases made under other promotions, Purchases made by customers in this cluster fall under the price category 2 & 3, Customers in this cluster favor prop category 5,6 & 7

Cluster 2 Characteristics- High no. of Purchases made in Promotion code 6 followed by other promotions, likes price category 1 & 4 and Proposition categories 8, 12 & 15

k=3

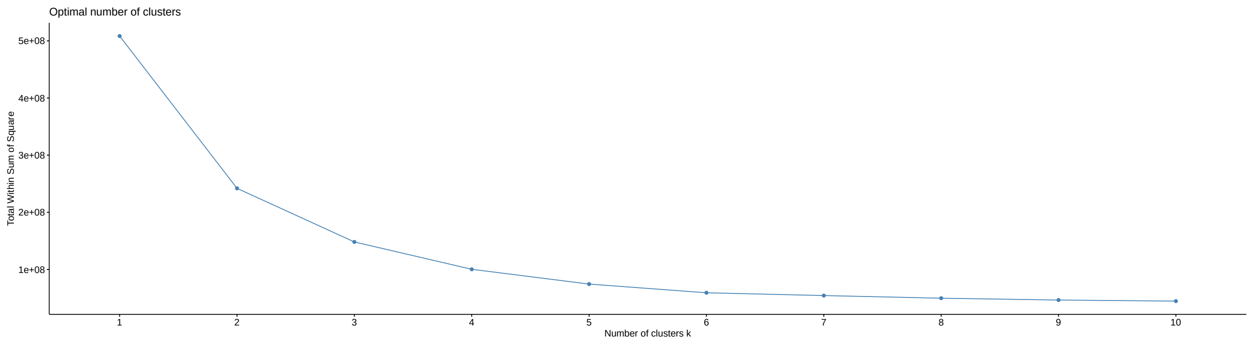
Cluster 1 Characteristics - High no. of Purchases made in No promotion, Most purchases were made on price category 2, Favor proposition category 5 more than others

Cluster 2 Characteristics - Most purchase are made under promotion code 6, likes to shop in the price category 1, and prefers propostion catergory 8 &12

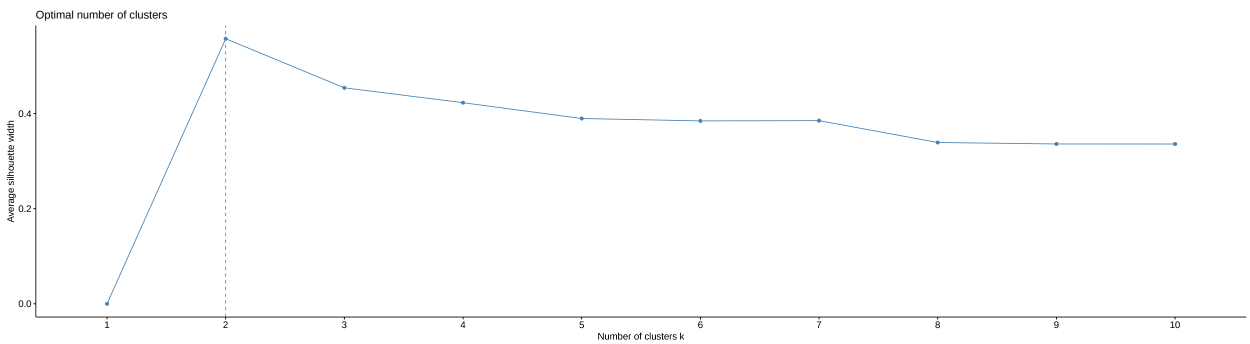
Cluster 3 Characteristics - Purchases are made under no promotion codes and other promotion codes, buys under the price category 2 & 4 mostly, and prefers proposition category 12

c.The variables that describe both purchase behavior and basis of purchase

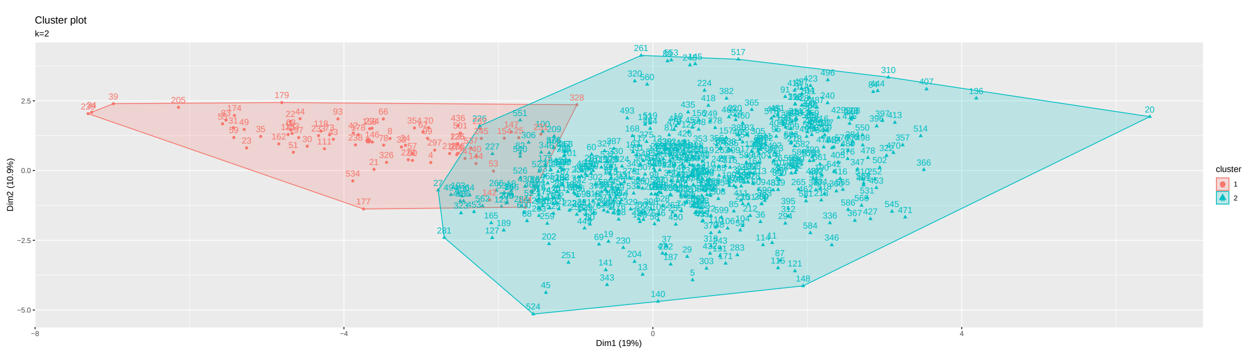
```r
pbbp <- cbind(prcs_bhvr, bas_prcs)

norm_pbbp <- scale(pbbp)

fviz_nbclust(pbbp,kmeans,method = "wss")
```



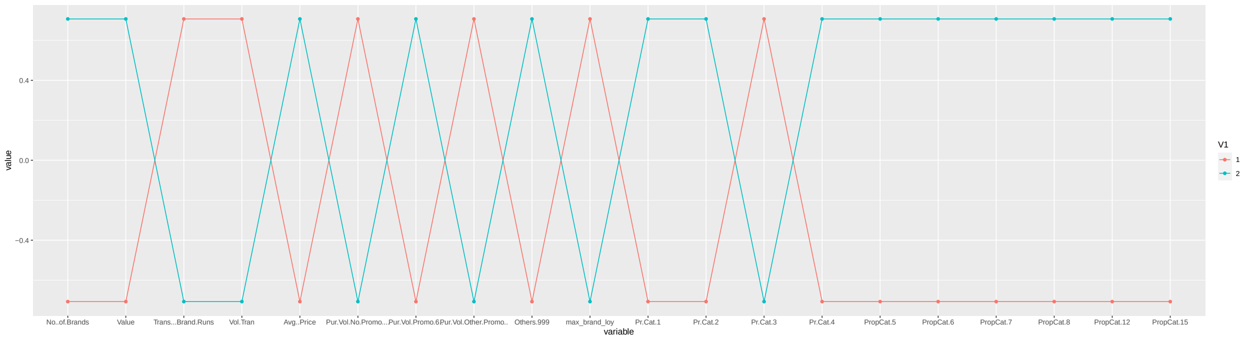

```r
fviz_nbclust(pbbp,kmeans,method = "silhouette")
```




```r
pbbp_means1 <- kmeans(norm_pbbp,2,25)
fviz_cluster(pbbp_means1, norm_pbbp)+ labs(subtitle = "k=2")
```




```r
pbbp_centers1 <- as.data.frame(cbind(1:nrow(pbbp_means1$centers),pbbp_means1$centers))
pbbp_centers1$V1 <- as.factor(pbbp_centers1$V1)
ggparcoord(data = pbbp_centers1,
           columns = 2:21,
           groupColumn = 1,
           alphaLines = 1.0,
           showPoints = TRUE)
```



k=2 gives pretty distinguised clusters

Cluster 1 Characteristics - They have very less brand loyalty and buy products from any brand, which offers promotion. They prefer to buy under the price category 1.

Cluster 2 characteristics - Customers in this cluster appear to have higher volume per transaction and particularly consider no promotions while purchasing items. They are highly brand loyal and do not buy products of other brands, prefer to buy products under the price category 3

Q. How should k be chosen?

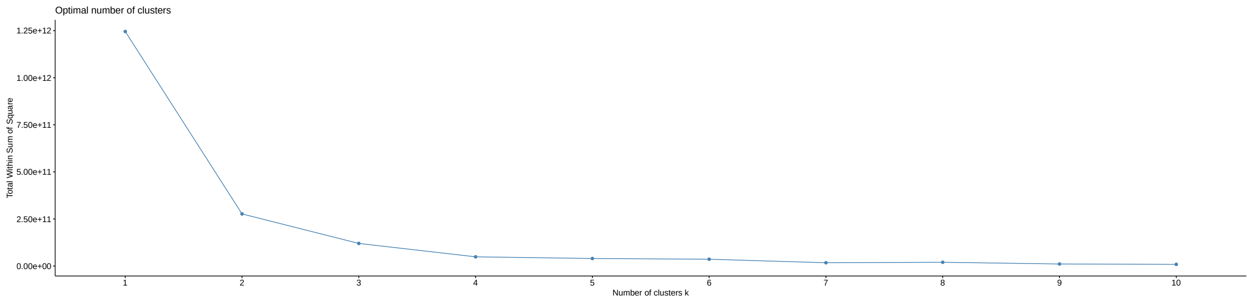
A - The value of 'K' should be chosen in such a way that: 1) The within cluster distances are minimum in all clusters 2) The inter cluster distances are maximum.

Q. 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.

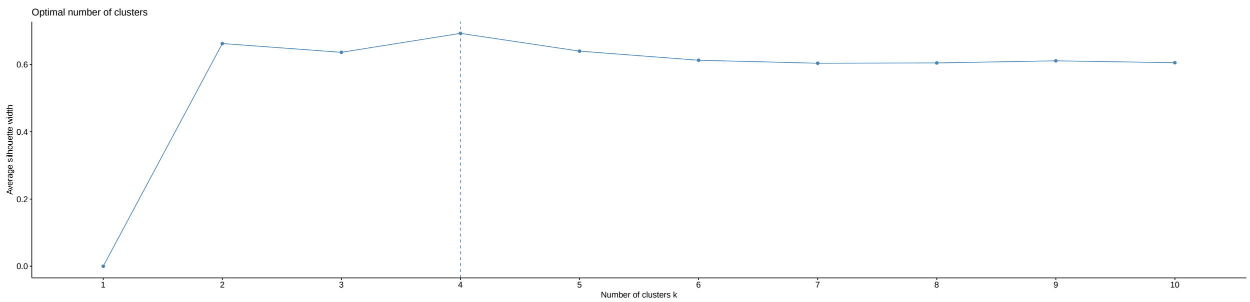
A - Since this data is being compiled for general use, and not to market and analyze one particular brand, we can say a customer who is fully devoted to brand A is similar to a customer fully devoted to brand B - both are fully loyal customers in their behavior. But if we include all the brand shares in the clustering, the analysis will treat those two customers as very different. The percentages of total purchases should not be considered individually as they increase the inter cluster distances and the effectiveness of the clustering drops. Hence I created a derived variable "max_brand_loy" which has the maximum proportion of brand purchase, which is used to determine brand loyalty.

Q2- Select what you think is the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters.

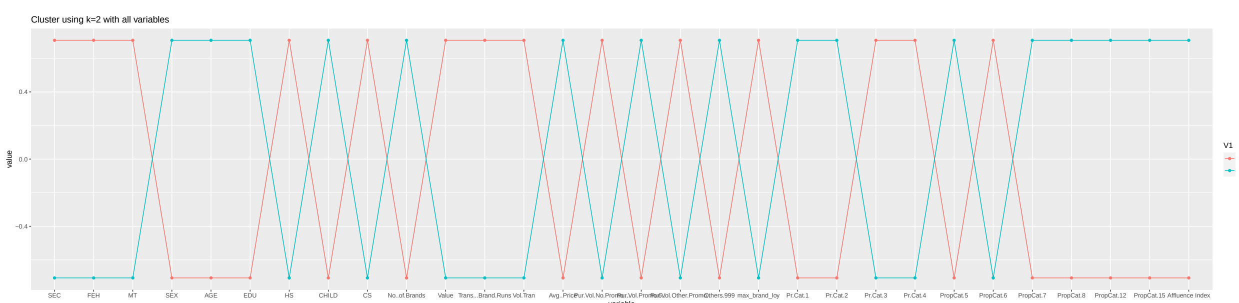
```
cat_var<-data.frame(apply((soap_df[which(colnames(soap_df) %in% c("SEC","FEH","MT","SEX","AGE","EDU","HS

#library(fastDummies)
#dummy_vars <- fastDummies::dummy_columns(cat_var) %>% select(-c("SEC","FEH","MT","SEX","AGE","EDU","HS
#all_vars <- cbind(dummy_vars, norm_pbbp) %>% mutate("Household Number"=soap_df$Member.id)

all_vars <- cbind(cat_var, norm_pbbp, "Affluence Index"= scale(soap_df[,11])) %>% mutate("Household Num

fviz_nbclust(all_vars,kmeans,method = "wss")
```



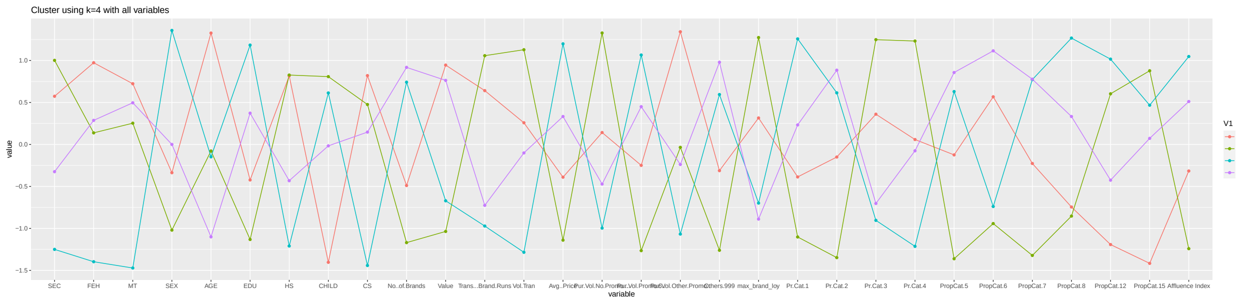

```
fviz_nbclust(all_vars,kmeans,method = "silhouette")
```



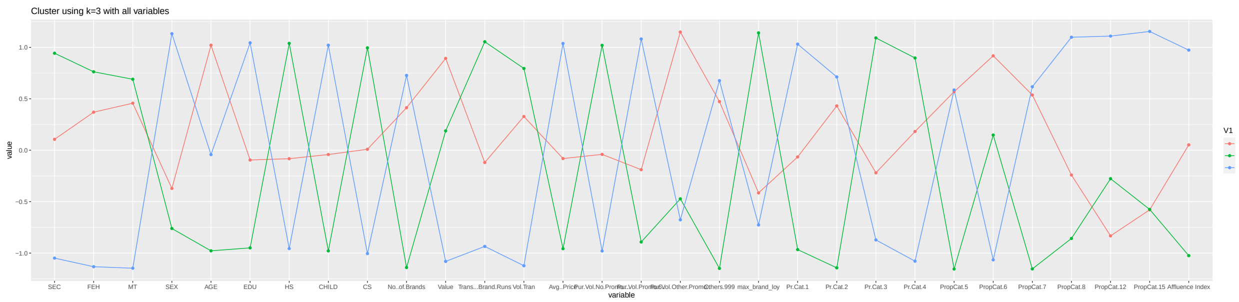

```
allvars_means1 <- kmeans(all_vars,2,25)
allvars_centers1 <- as.data.frame(cbind(1:nrow(allvars_means1$centers),allvars_means1$centers))
allvars_centers1$V1 <- as.factor(allvars_centers1$V1)
ggparcoord(data = allvars_centers1,
           columns = 2:31,
           groupColumn = 1,
           alphaLines = 1.0,
           showPoints = TRUE, title = "Cluster using k=2 with all variables")
```




```
allvars_means2 <- kmeans(all_vars,4,25)
allvars_centers2 <- as.data.frame(cbind(1:nrow(allvars_means2$centers),allvars_means2$centers))
allvars_centers2$V1 <- as.factor(allvars_centers2$V1)
ggparcoord(data = allvars_centers2,
           columns = 2:31,
           groupColumn = 1,
           alphaLines = 1.0,
           showPoints = TRUE, title = "Cluster using k=4 with all variables")
```




```
allvars_means3 <- kmeans(all_vars,3,25)
allvars_centers3 <- as.data.frame(cbind(1:nrow(allvars_means3$centers),allvars_means3$centers))
allvars_centers3$V1 <- as.factor(allvars_centers3$V1)
ggparcoord(data = allvars_centers3,
           columns = 2:31,
           groupColumn = 1,
           alphaLines = 1.0,
           showPoints = TRUE, title = "Cluster using k=3 with all variables")
```



K=2 (As per the elbow plot)

Cluster 1 Characteristics - Customers in this cluster, demographically, are low in socio-economic status, not as highly educated. In terms of purchase behavior, they are brand loyal and prefer buying in the price category 3, tend to purchase under no promotion. They are highly active which can be seen as they have high transaction per brand run.

Cluster 2 Characteristics - Highly educated customer base who are well to do in terms of socio economic level, no loyal towards any particular brand, tend to buy from different brands but have low activity and prefer to buy under the price category 1.

K=4 (As per the Silhouette plot)

Cluster 1 Characteristics - Not very responsive to promotions, pricing or selling propositions,

Cluster 2 Characteristics - Socio-economically High, Tend to buy under promotions, Mostly purchase under price category 1, Highly affluent

Cluster 3 Characteristics - Least brand loyalty, Doesn't have demographically distinct characteristics

Cluster 4 Characteristics - Price category 3 & 4, Socio-economically lowest, Have high loyalty towards brands, high activity i.e transaction/ brand run and volume/ transaction

This is the best approach k=3 (As per clusters analysis)

Cluster 1 Characteristics - Customers in this cluster are socio-economically high, Highly educated, Low brand loyalty, favours price category 1, influenced by promotions, least active customers

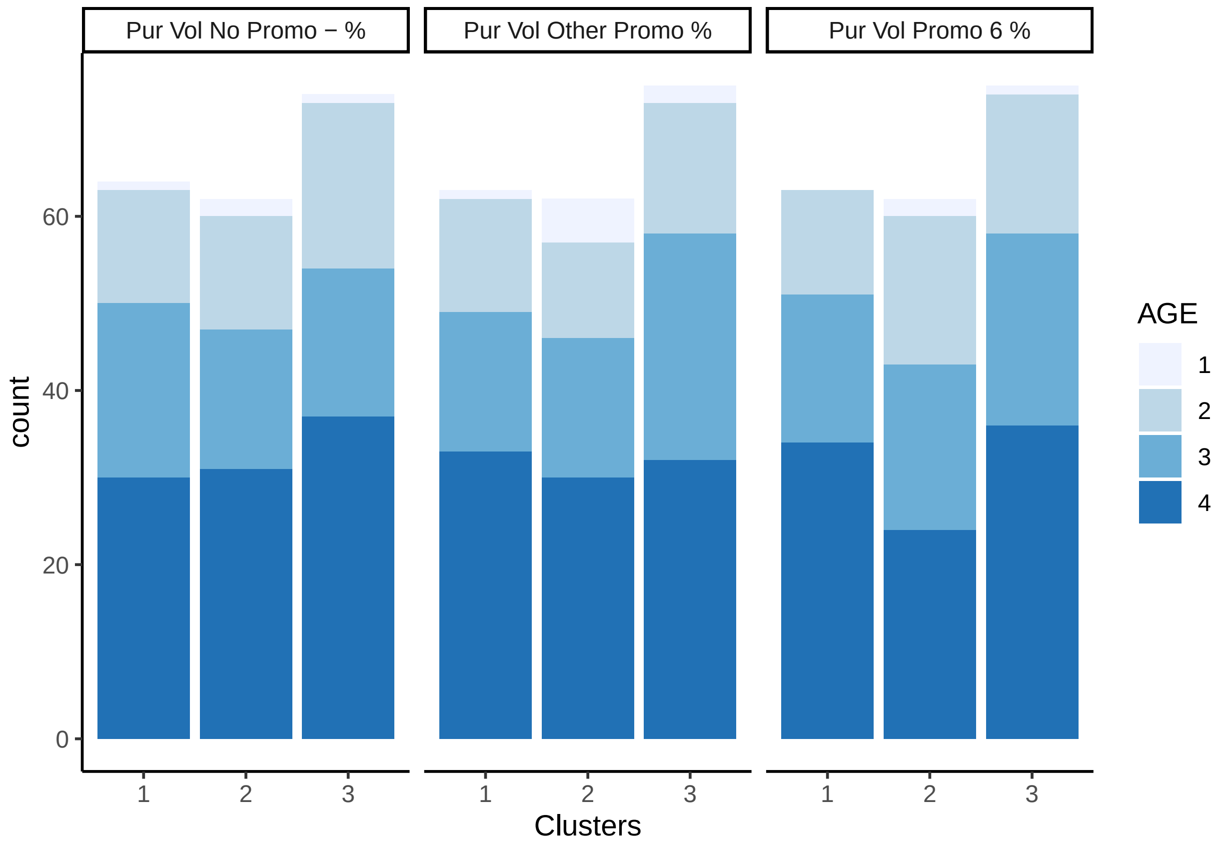
Cluster 2 Characteristics - Socioeconomically low, not as highly educated, Most purchases are not under promotion, highly active customer in terms of no. of transactions and volume of transactions, loyal towards brand

Cluster 3 Characteristics - Mid Socioeconomic group, moderately educated, moderately active in terms of transaction per brand run and volume in each transaction, not much influenced by promotions, moderatelyto low brand loyal

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.

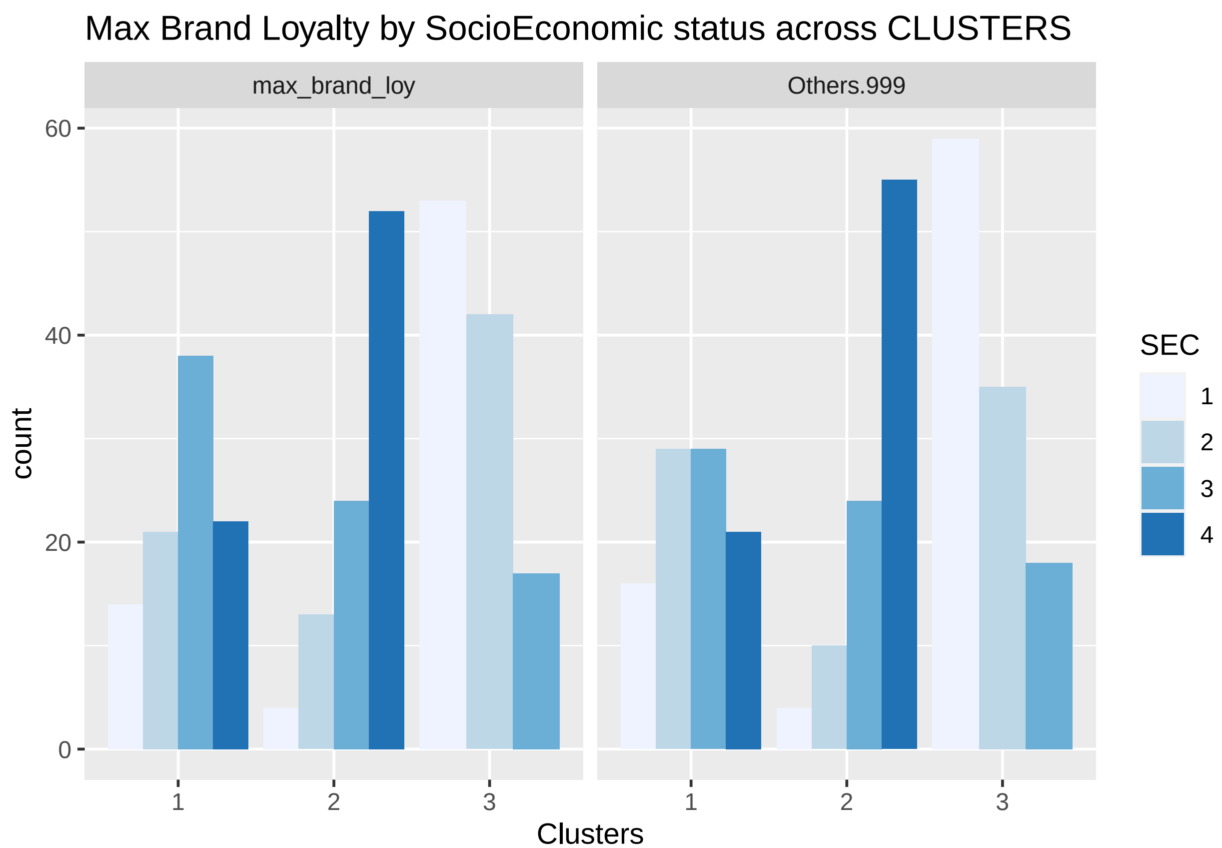
```
library(dplyr)
all_vars_clusters <- cbind("Household ID"=soap_df$Member.id, cat_var,"Affluence Index"=soap_df$Affluenc

all_vars_clusters$SEC <- as.factor(all_vars_clusters$SEC)
all_vars_clusters$EDU <- as.factor(all_vars_clusters$EDU)
all_vars_clusters$Clusters <- as.factor(all_vars_clusters$Clusters)

library(ggplot2)

ggplot(all_vars_clusters) +
  aes(x =Clusters,fill= AGE) +
  geom_bar() +
  scale_fill_hue() +
  theme_classic() +
        scale_fill_brewer(palette = "Blues")+
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other Promo %")))
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```

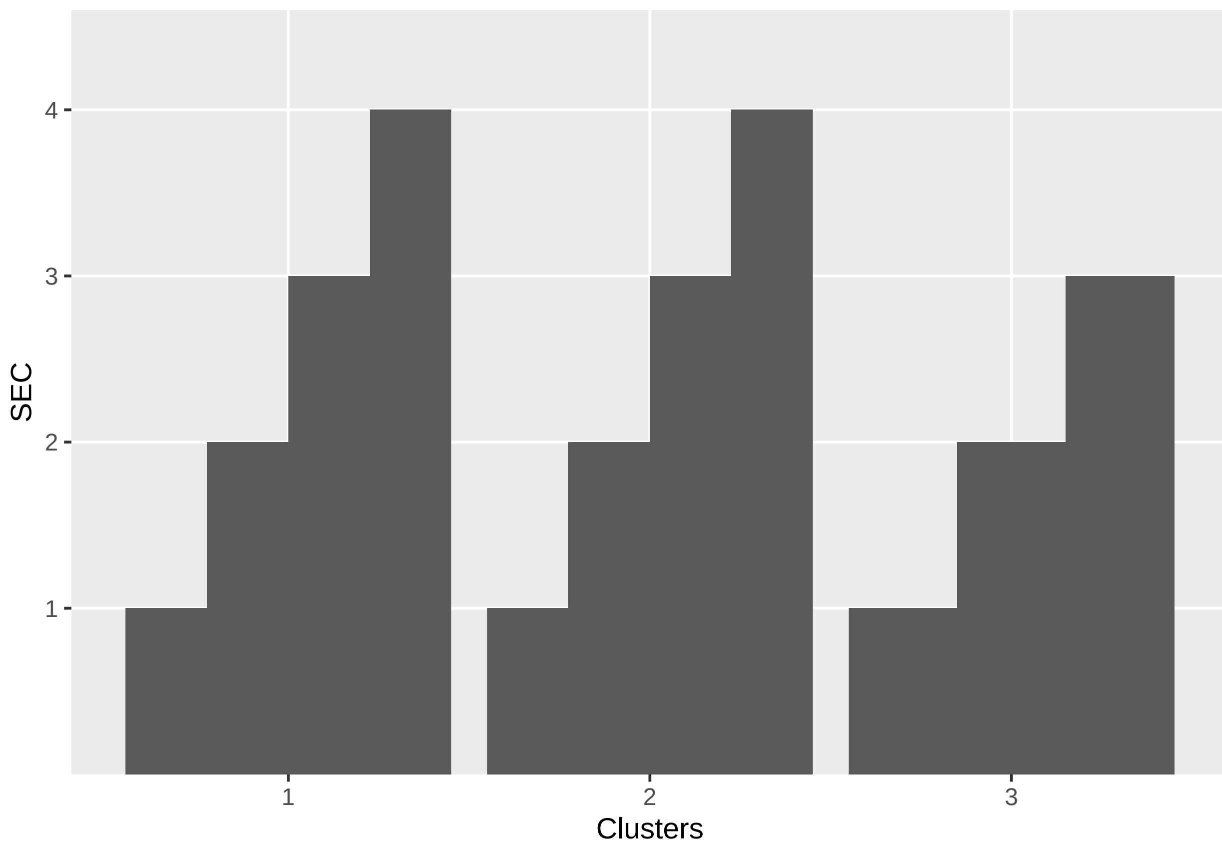


```r
ggplot(all_vars_clusters) +
  aes(x =Clusters,fill= SEC) +
  geom_bar(position = "dodge") +
  scale_fill_hue() +
  theme_grey() +
        scale_fill_brewer(palette = "Blues")+
  facet_wrap(vars(c("Others.999", "max_brand_loy")))+
        ggtitle("Max Brand Loyalty by SocioEconomic status across CLUSTERS")
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```




```r
ggplot(all_vars_clusters, aes(Clusters, SEC)) +geom_bar(stat = "identity", position = "dodge")
```



Conclusion : The customer base can be divided into three major segments:

Customers in cluster 1 are socio-economically well to do, Highly educated. But demonstrate Low brand loyalty, favors price category 1 and are influenced by promotions with very high average purchase price.

Suggestion: Since they are least active customers as they have low transaction/Brand run and low volume purchase per transaction, they should be given promotion codes to increase their activity. This could possibly increase their loyalty towards brand.

Customers in cluster 2 are Socioeconomically low, not as highly educated. These customers are very loyal towards brands and hardly make purchases from other brands, Most purchases are not under promotion, highly active customer in terms of no. of transactions and volume of transactions.

Suggestion: We should pay attention to these customers as they give more business than the rest of the two segments. However they are not enticed by Promotions, so we need to come up with schemes to retain and reward them.

Customers of Cluster 3 fall under Mid Socioeconomic group, moderately educated, moderately active in terms of transaction per brand run and volume in each transaction, not much influenced by promotions but have the potential, moderately to low brand loyal. These customers are of high value

Suggestion: This group needs to be targeted by promotions heavily as they have the potential to move incline towards the brand and become loyal.