# bathsoap

rajendra

12/5/2019

```
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(factoextra)

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at
https://goo.gl/13EFCZ

library(hrbrthemes)

## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use
these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto
Condensed and

##       if Arial Narrow is not on your system, please see
http://bit.ly/arialnarrow

library(GGally)

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

library(viridis)

## Loading required package: viridisLite

set.seed(123)

#Read Data
library(readr)

BathSoap <- read_csv("C:/Users/rajendra/Downloads/BathSoap.csv")

## Parsed with column specification:
## cols(
##   .default = col_double()
## )

## See spec(...) for full column specifications.
```

```r
str(BathSoap)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 600 obs. of  46
variables:
##  $ Member id            : num  1010010 1010020 1014020 1014030 1014190 ...
##  $ SEC                  : num  4 3 2 4 4 4 4 4 4 1 ...
##  $ FEH                  : num  3 2 3 0 1 3 2 3 3 3 ...
##  $ MT                   : num  10 10 10 0 10 10 10 10 10 5 ...
##  $ SEX                  : num  1 2 2 0 2 2 2 2 2 1 ...
##  $ AGE                  : num  4 2 4 4 3 3 4 2 4 4 ...
##  $ EDU                  : num  4 4 5 0 4 4 1 4 4 7 ...
##  $ HS                   : num  2 4 6 0 4 5 3 5 6 3 ...
##  $ CHILD                : num  4 2 4 5 3 2 2 3 4 4 ...
##  $ CS                   : num  1 1 1 0 1 1 1 0 1 1 ...
##  $ Affluence Index      : num  2 19 23 0 10 13 11 0 17 6 ...
##  $ No. of Brands        : num  3 5 5 2 3 3 4 3 2 4 ...
##  $ Brand Runs           : num  17 25 37 4 6 26 17 8 12 13 ...
##  $ Total Volume         : num  8025 13975 23100 1500 8300 ...
##  $ No. of  Trans        : num  24 40 63 4 13 41 26 25 27 18 ...
##  $ Value                : num  818 1682 1950 114 591 ...
##  $ Trans / Brand Runs   : num  1.41 1.6 1.7 1 2.17 1.58 1.53 3.13 2.25
1.38 ...
##  $ Vol/Tran             : num  334 349 367 375 638 ...
##  $ Avg. Price           : num  10.19 12.03 8.44 7.6 7.12 ...
##  $ Pur Vol No Promo - % : num  1 0.89 0.94 1 0.61 1 0.98 0.94 0.9 1 ...
##  $ Pur Vol Promo 6 %    : num  0 0.1 0.02 0 0.14 0 0.02 0 0.1 0 ...
##  $ Pur Vol Other Promo %: num  0 0.02 0.04 0 0.24 0 0 0.06 0 0 ...
##  $ Br. Cd. 57, 144      : num  0.38 0.02 0.03 0.4 0.05 0.08 0.45 0.04 0.39
0.07 ...
##  $ Br. Cd. 55           : num  0.13 0.08 0.55 0.6 0.14 0.07 0.05 0.79 0
0.12 ...
##  $ Br. Cd. 272          : num  0 0 0 0 0 0 0.01 0 0 0 ...
##  $ Br. Cd. 286          : num  0 0 0.03 0 0 0 0 0 0 0 ...
##  $ Br. Cd. 24           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Br. Cd. 481          : num  0 0.06 0 0 0 0 0 0 0 0 ...
##  $ Br. Cd. 352          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Br. Cd. 5            : num  0 0.14 0.02 0 0 0 0 0 0 0.4 ...
##  $ Others 999           : num  0.492 0.699 0.379 0 0.807 0.857 0.495 0.167
0.615 0.41 ...
##  $ Pr Cat 1             : num  0.23 0.29 0.12 0 0 0.22 0.07 0.04 0.11 0.61
...
##  $ Pr Cat 2             : num  0.56 0.55 0.32 0.4 0.05 0.45 0.66 0.04 0.89
0.1 ...
##  $ Pr Cat 3             : num  0.13 0.09 0.56 0.6 0.14 0.07 0.05 0.9 0
0.12 ...
##  $ Pr Cat 4             : num  0.07 0.06 0 0 0.81 0.27 0.23 0.02 0 0.17
...
##  $ PropCat 5            : num  0.5 0.46 0.24 0.4 0.81 0.49 0.82 0.06 0.7
0.24 ...
##  $ PropCat 6            : num  0 0.35 0.12 0 0 0.1 0 0 0.28 0.46 ...
```

```
##  $ PropCat 7           : num  0 0.03 0.03 0 0 0 0.02 0 0 0.15 ...
##  $ PropCat 8           : num  0 0.02 0.01 0 0.05 0.01 0.01 0 0 0 ...
##  $ PropCat 9           : num  0 0.01 0.01 0 0 0.07 0 0 0.02 0 ...
##  $ PropCat 10          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PropCat 11          : num  0 0.06 0 0 0 0 0 0 0 0 ...
##  $ PropCat 12          : num  0.03 0 0.02 0 0 0 0 0.01 0 0 ...
##  $ PropCat 13          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PropCat 14          : num  0.13 0.08 0.56 0.6 0.14 0.07 0.05 0.9 0
0.12 ...
##  $ PropCat 15          : num  0.34 0 0 0 0 0.27 0.1 0.03 0 0.03 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Member id` = col_double(),
##   ..   SEC = col_double(),
##   ..   FEH = col_double(),
##   ..   MT = col_double(),
##   ..   SEX = col_double(),
##   ..   AGE = col_double(),
##   ..   EDU = col_double(),
##   ..   HS = col_double(),
##   ..   CHILD = col_double(),
##   ..   CS = col_double(),
##   ..   `Affluence Index` = col_double(),
##   ..   `No. of Brands` = col_double(),
##   ..   `Brand Runs` = col_double(),
##   ..   `Total Volume` = col_double(),
##   ..   `No. of  Trans` = col_double(),
##   ..   Value = col_double(),
##   ..   `Trans / Brand Runs` = col_double(),
##   ..   `Vol/Tran` = col_double(),
##   ..   `Avg. Price` = col_double(),
##   ..   `Pur Vol No Promo - %` = col_double(),
##   ..   `Pur Vol Promo 6 %` = col_double(),
##   ..   `Pur Vol Other Promo %` = col_double(),
##   ..   `Br. Cd. 57, 144` = col_double(),
##   ..   `Br. Cd. 55` = col_double(),
##   ..   `Br. Cd. 272` = col_double(),
##   ..   `Br. Cd. 286` = col_double(),
##   ..   `Br. Cd. 24` = col_double(),
##   ..   `Br. Cd. 481` = col_double(),
##   ..   `Br. Cd. 352` = col_double(),
##   ..   `Br. Cd. 5` = col_double(),
##   ..   `Others 999` = col_double(),
##   ..   `Pr Cat 1` = col_double(),
##   ..   `Pr Cat 2` = col_double(),
##   ..   `Pr Cat 3` = col_double(),
##   ..   `Pr Cat 4` = col_double(),
##   ..   `PropCat 5` = col_double(),
##   ..   `PropCat 6` = col_double(),
##   ..   `PropCat 7` = col_double(),
```

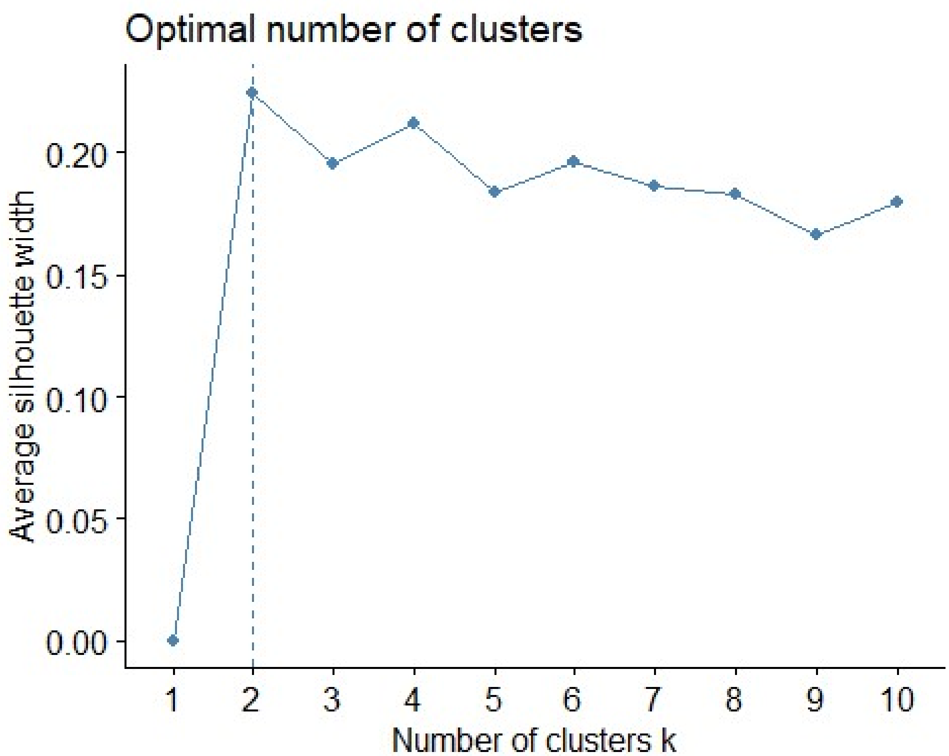
```
##   ..   `PropCat 8` = col_double(),
##   ..   `PropCat 9` = col_double(),
##   ..   `PropCat 10` = col_double(),
##   ..   `PropCat 11` = col_double(),
##   ..   `PropCat 12` = col_double(),
##   ..   `PropCat 13` = col_double(),
##   ..   `PropCat 14` = col_double(),
##   ..   `PropCat 15` = col_double()
##   .. )
```

```
#Customer Brand loyality
#Brand loyality is defined as the  customer buys spends a maximum amount of money in 8 brands.

r1<-BathSoap[,23:30]# Incldng 8 brands
BathSoap$Loyality_Brand<-as.numeric(apply(r1,1,max)) # Maximum value of the brand
table(BathSoap$Loyality_Brand)
```

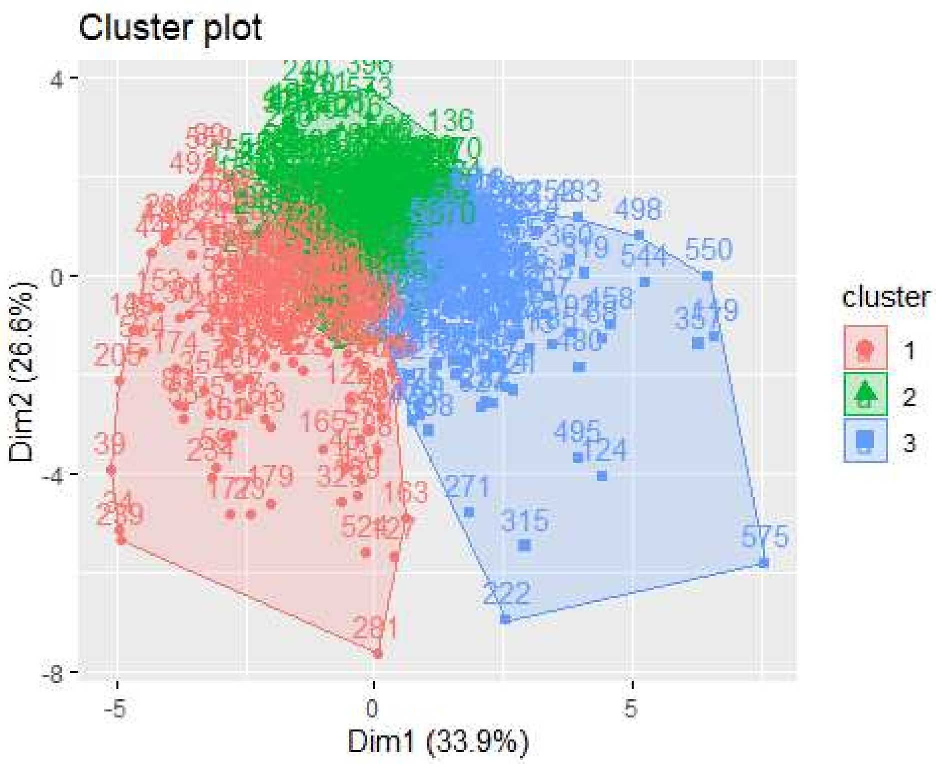
```
## 
##    0 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09  0.1 0.11 0.12 0.13 0.14 
##   24    4    9    8    8   13    8    6    9   16   10   15   10   13   18 
## 0.15 0.16 0.17 0.18 0.19  0.2 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 
##    7   10    5    6    6   11    9   10    4   10    9   11    7    9   11 
##  0.3 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39  0.4 0.41 0.42 0.43 0.44 
##   10   10    8    6    7    8    9    4    7    3    9    5    4    4    5 
## 0.45 0.46 0.47 0.48 0.49  0.5 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59 
##    3    4    2    9    2    4    1    3    6    6    3    8    5    3    5 
##  0.6 0.61 0.62 0.63 0.64 0.66 0.67 0.68 0.69  0.7 0.71 0.72 0.73 0.74 0.75 
##    6    3    4    4    3    5    8    1    5    3    4    3    4    1    5 
## 0.76 0.77 0.78 0.79  0.8 0.83 0.84 0.85 0.86 0.87 0.88 0.89  0.9 0.91 0.93 
##    4    3    2    5    4    3    5    1    4    3    2    7    3    2    2 
## 0.94 0.95 0.96 0.97 0.98 0.99    1 
##    4    2    3    3    3    2   15
```

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

```
#The variables used by us are No.Of Brands,Brand runs, Total Volume,Number of Transctions,value,trans/brandruns,vol/trans,avg price,others999,Loyality_Brand.

BS<-BathSoap[,c(12,13,14,15,16,17,18,19,31,47)]

data1.s<-as.data.frame(scale(BS)) # scaling the data
# Elbow chart to estimate the optimal K
fviz_nbclust(data1.s,kmeans,method = "silhouette")
```




```
# Choosing the optimal K as 3 and forming 3 clusters
model<-kmeans(data1.s,3,nstart=50)

# Visualizing the clusters
fviz_cluster(model,data1.s)
```



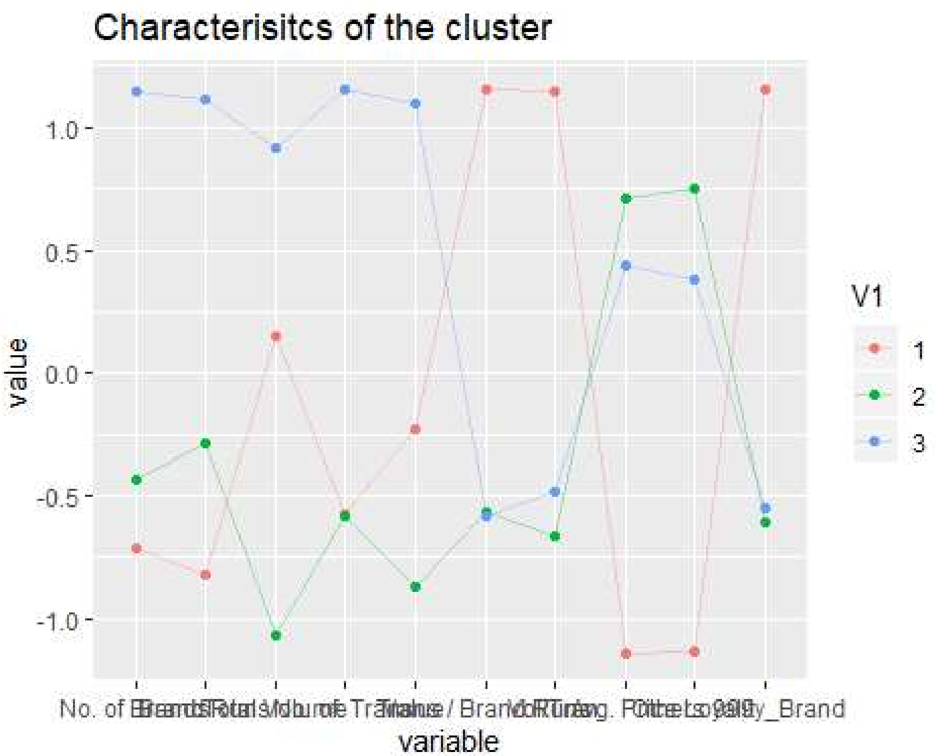

```
result<-as.data.frame(cbind(1:nrow(model$centers),model$centers))
result$V1<-as.factor(result$V1)
# Characteristics of the cluster
result
```

```
##   V1 No. of Brands Brand Runs Total Volume No. of  Trans       Value
## 1  1    -0.4934603 -0.7236194    0.1846318    -0.4114518 -0.05446008
## 2  2    -0.2759333 -0.2156513   -0.5323647    -0.4155637 -0.45499157
## 3  3     0.9507367  1.0993197    0.6359753     1.0821389  0.76730922
##   Trans / Brand Runs   Vol/Tran Avg. Price Others 999 Loyality_Brand
## 1          0.6077360  0.5853534 -0.4587115 -1.1282912      1.2402467
## 2         -0.2473209 -0.2735992  0.2283229  0.5990322     -0.5323700
## 3         -0.2548053 -0.1902088  0.1273427  0.2548289     -0.4768636
```

```
model$size
```

```
## [1] 175 259 166
```

```
# Parallel plot to visualize the cluster.
ggparcoord(result,
           columns = 2:ncol(result), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




Characterstics of cluster based on purchase behaviour

Loyality brand for cluster1 is very high and cluster 2 is very low because the number of brand runs is very low and high respctively.

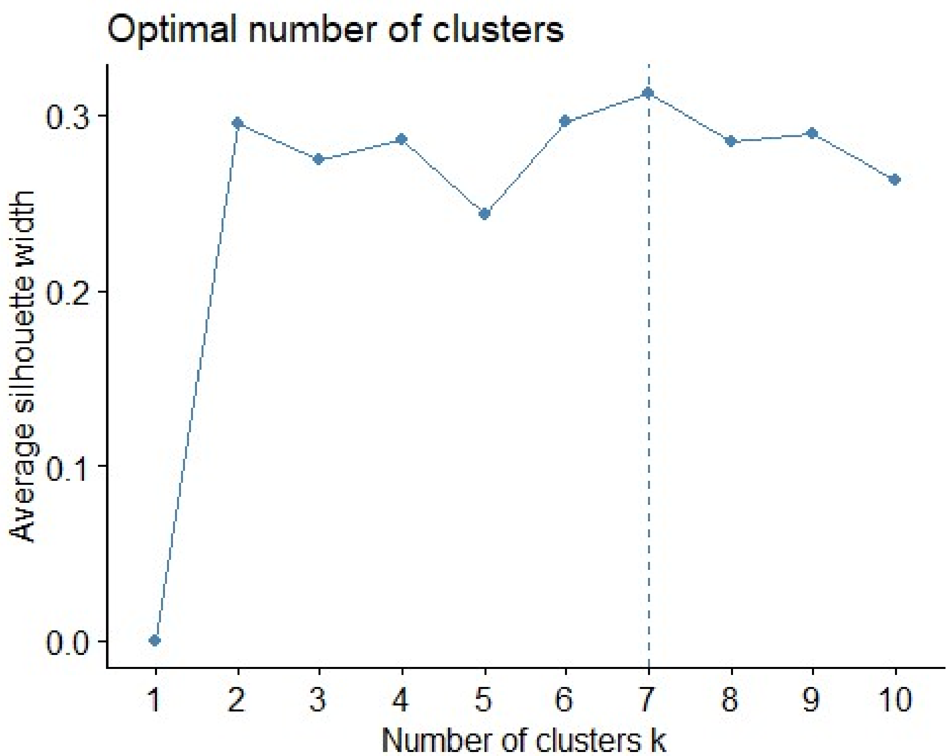
Average price is low for cluster 1 and high for cluster 2 and moreover the people are purchasing very low volume of brands from others 999 in cluster 1 when compared to cluster 2

High volume of transctions is very high in clster1 when compared to cluster 2.

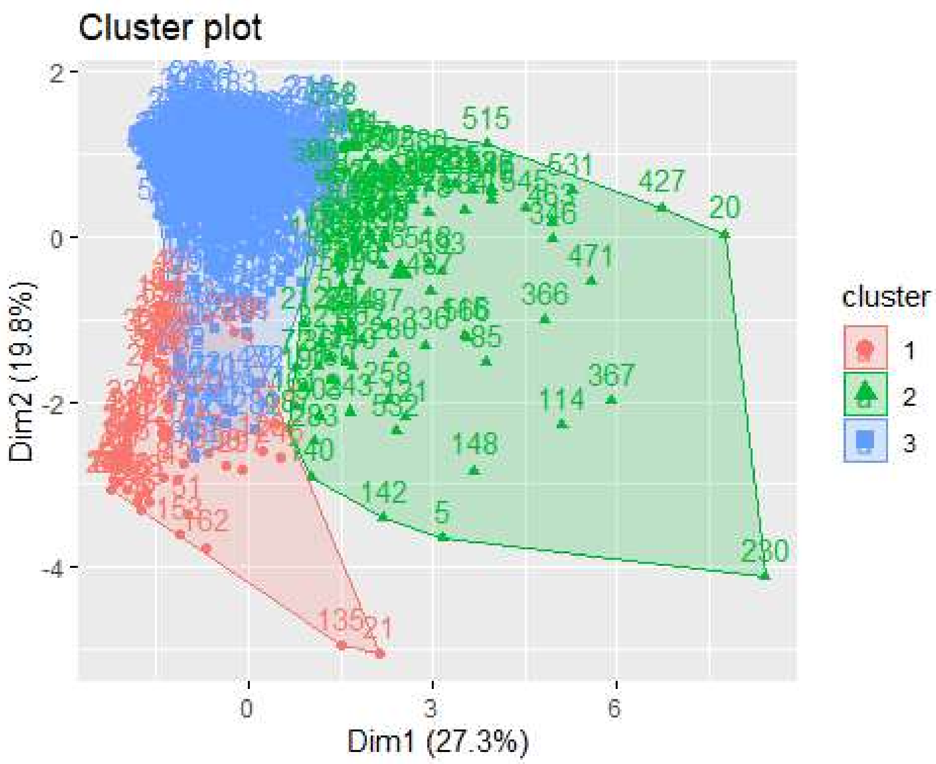
Cluster 3 is unremarkable from any measure

b. The variables that describe the basis for purchase

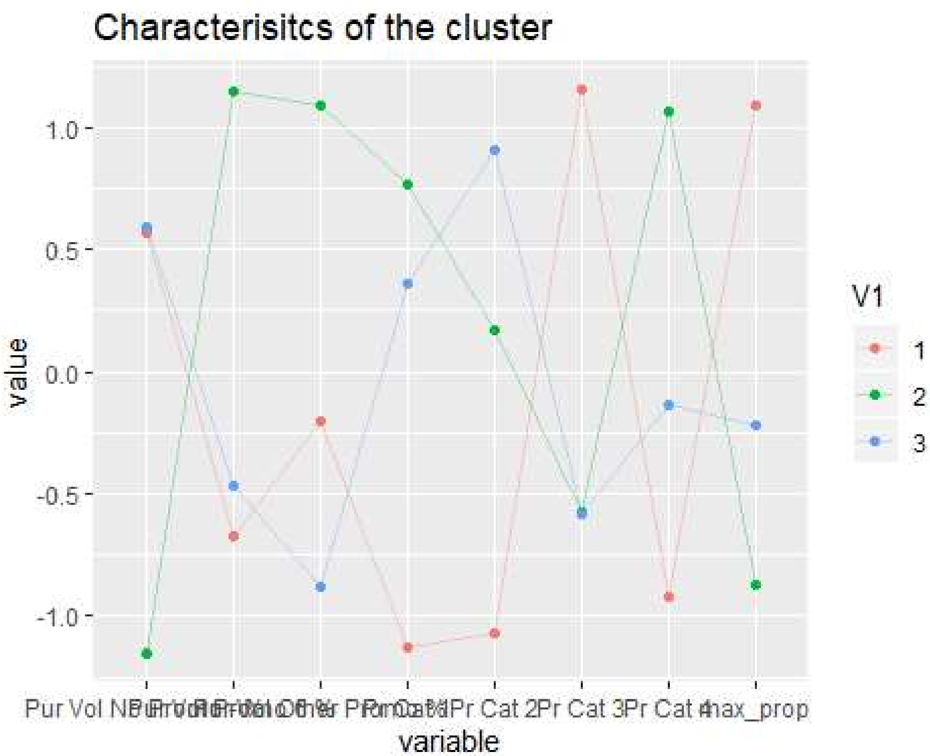
```
#Finding the maximum proptional category from the 10 categories
r2<-BathSoap[,36:46]
BathSoap$max_prop_no<-as.numeric(apply(r2,1,which.max))
BathSoap$max_prop<-as.numeric(apply(r2,1,max))

# variables considered for basis of purchase are purchase volume no
promo,6%,other promo,pric categories from 1 to 4 and bath soap maximumprop
BS1<-BathSoap[,c(20:22,32:35,49)]

data2.s<-as.data.frame(scale(BS1)) # scaling the data
# Elbow chart to estimate the optimal K
fviz_nbclust(data2.s,kmeans,method = "silhouette")
```




```
# Choosing the optimal K as 3 and forming 3 clusters
model1<-kmeans(data2.s,3,nstart=50)

# Visualizing the clusters
fviz_cluster(model1,data2.s)
```




```
result1<-as.data.frame(cbind(1:nrow(model1$centers),model1$centers))
result1$V1<-as.factor(result1$V1)
# Characteristics of the cluster
result1

##   V1 Pur Vol No Promo - % Pur Vol Promo 6 % Pur Vol Other Promo %
## 1  1            0.3313992        -0.5291617             0.1350441
## 2  2           -1.6753944         1.5351429             0.7942665
## 3  3            0.3591417        -0.2937761            -0.2159952
##       Pr Cat 1   Pr Cat 2   Pr Cat 3    Pr Cat 4    max_prop
## 1 -0.80050906 -1.2138944  2.5443612 -0.40695235  0.71615570
## 2  0.28747617 -0.2902592 -0.3105394  0.48441543 -0.38550796
## 3  0.05478764  0.2612340 -0.3221158 -0.05513508 -0.01753293

model1$size

## [1]  67 105 428

# Parallel plot to visualize the cluster.
ggparcoord(result1,
           columns = 2:ncol(result1), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



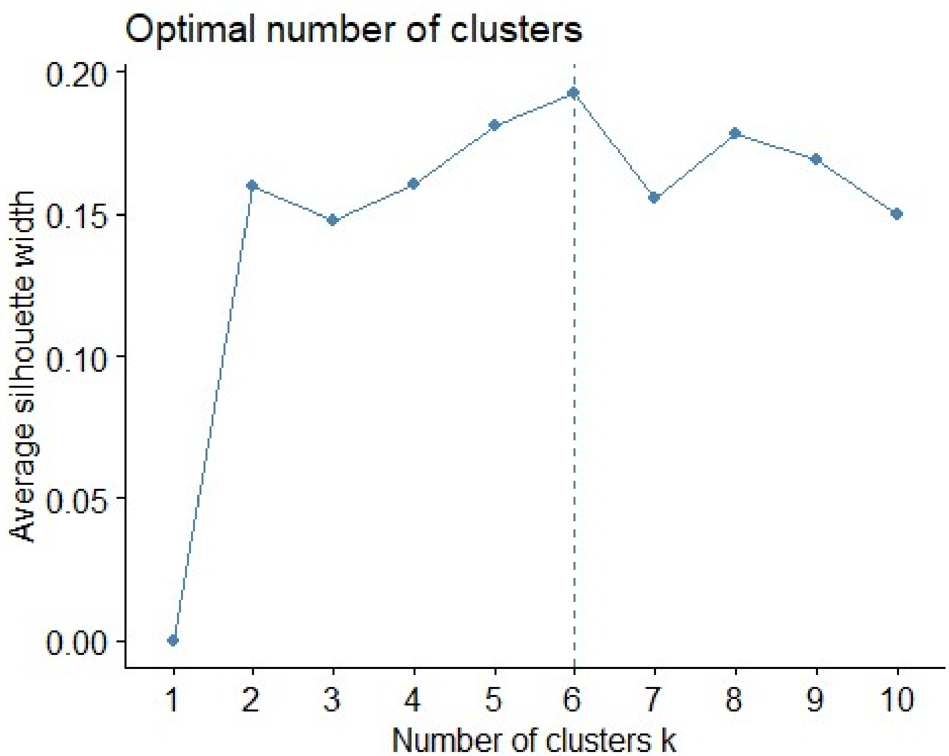

Characterstics of cluster

In cluster 1 we can see that maximum proption of purchase is coming when there is no promos and exactly reversible for cluster 2

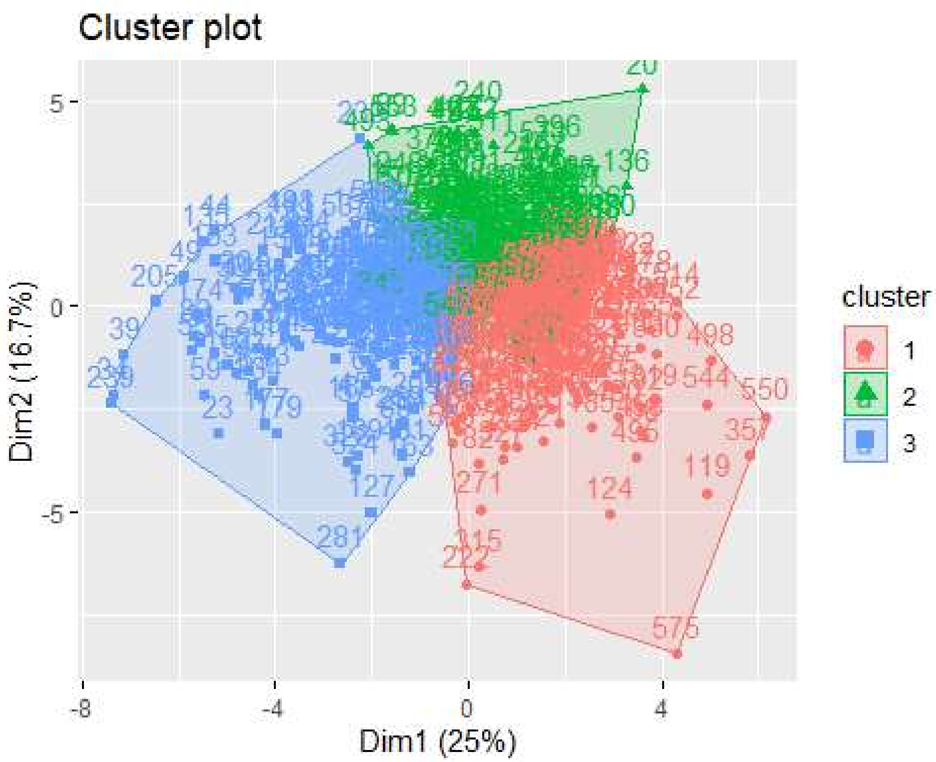
There is low for value for price category 1,category2,category 4 and high value of price category 3 then high chance of maximum proption of purchase in cluster1

c:

```
BS2<-BathSoap[,c(12:22,31:35,47,49)]

data3.s<-as.data.frame(scale(BS2)) # scaling the data
# Elbow chart to estimate the optimal K
fviz_nbclust(data3.s,kmeans,method = "silhouette")
```




```r
# Choosing the optimal K as 3 and forming 3 clusters
model2<-kmeans(data3.s,3,nstart=50)

# Visualizing the clusters
fviz_cluster(model2,data3.s)
```



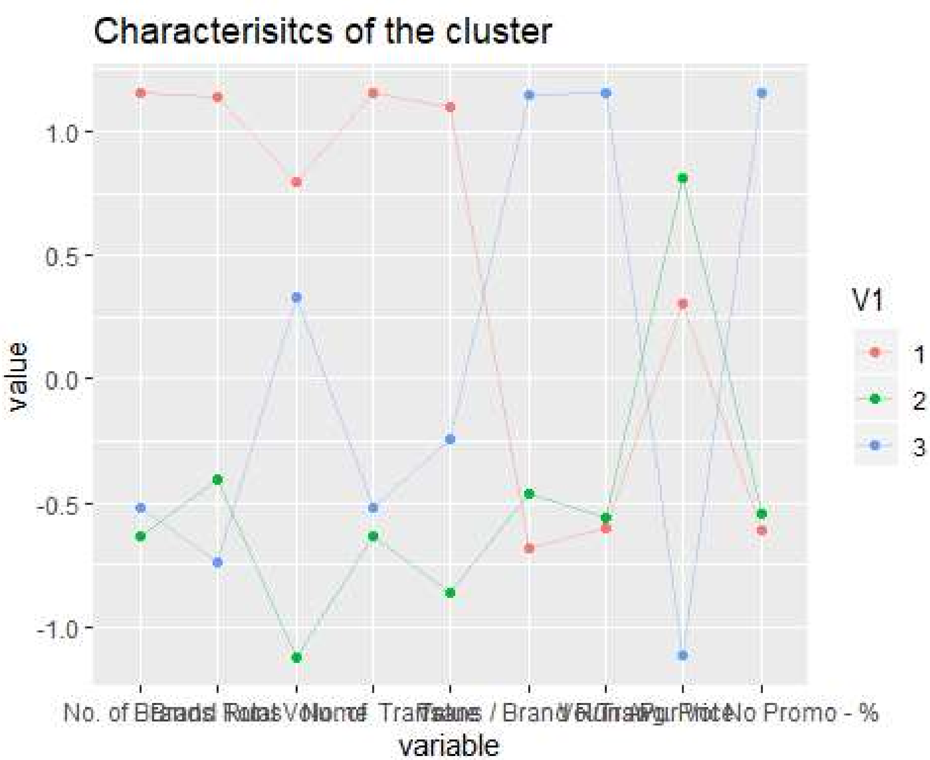

```
result2<-as.data.frame(cbind(1:nrow(model2$centers),model2$centers))
result2$V1<-as.factor(result2$V1)
# Characteristics of the cluster
result2
```

```
##   V1 No. of Brands Brand Runs Total Volume No. of  Trans      Value
## 1  1     0.8928878  0.9385645    0.3545583     0.8259381  0.4985096
## 2  2    -0.5355990 -0.3679096   -0.5358290    -0.4942869 -0.4202521
## 3  3    -0.4415878 -0.6533245    0.1379542    -0.4095943 -0.1290146
##   Trans / Brand Runs   Vol/Tran Avg. Price Pur Vol No Promo - %
## 1         -0.2783812 -0.2862997  0.1664992           -0.1572252
## 2         -0.1864074 -0.2675783  0.4436238           -0.1393148
## 3          0.4803077  0.5675279 -0.6092892            0.3042684
##   Pur Vol Promo 6 % Pur Vol Other Promo % Others 999    Pr Cat 1
## 1         0.2498496           -0.05534672  0.2112735  0.07752608
## 2         0.1431833            0.04128407  0.7468417  0.62463663
## 3        -0.4076960            0.01953425 -0.9514556 -0.68904124
##     Pr Cat 2   Pr Cat 3   Pr Cat 4 Loyality_Brand   max_prop
## 1  0.2494164 -0.2440511 -0.1761328     -0.4531279 -0.7869507
## 2 -0.4730066 -0.4261012  0.4469815     -0.5871606  0.1451538
## 3  0.1901929  0.6757571 -0.2438239      1.0569076  0.7061380
```

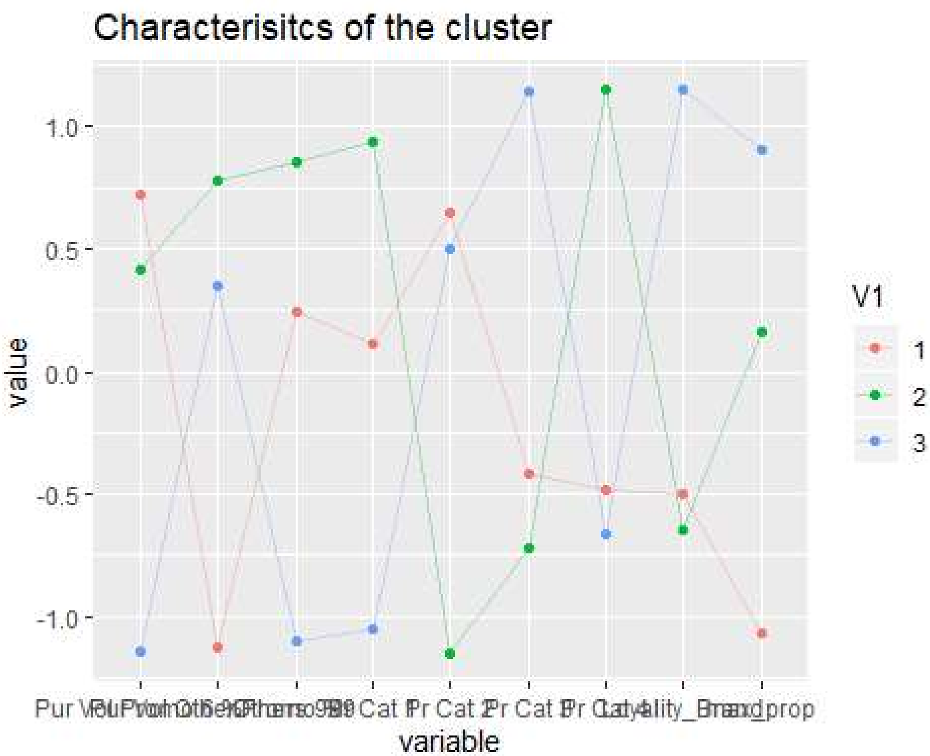
```
model2$size
```

```
## [1] 212 191 197
```

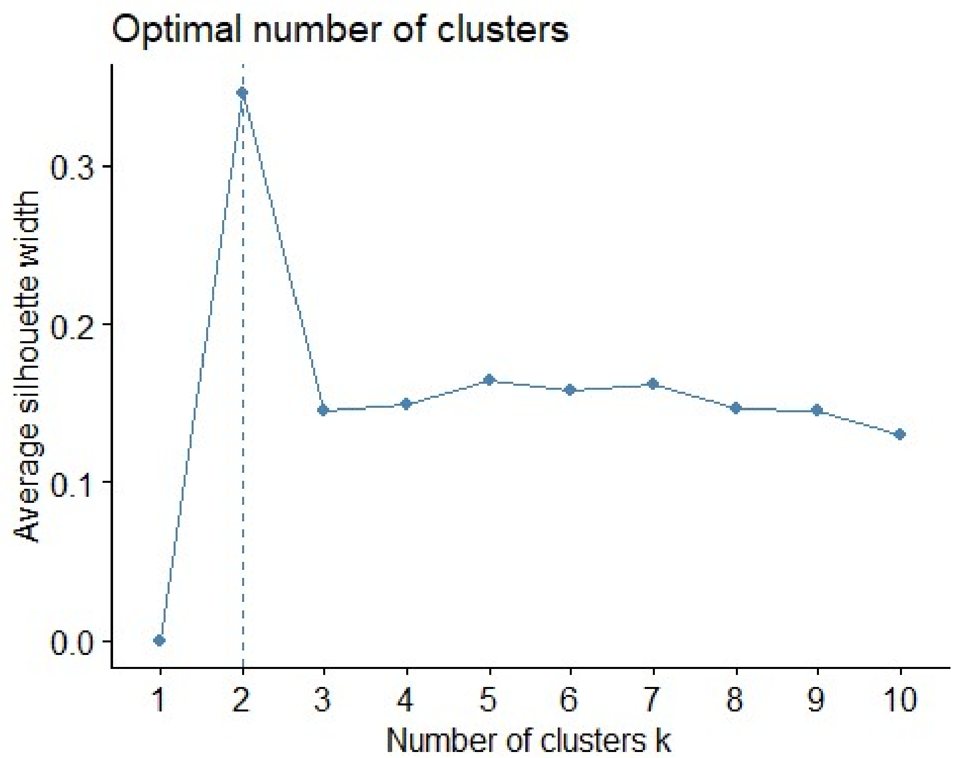
```r
# Parallel plot to visualize the cluster.
ggparcoord(result2,
           columns = 2:10, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



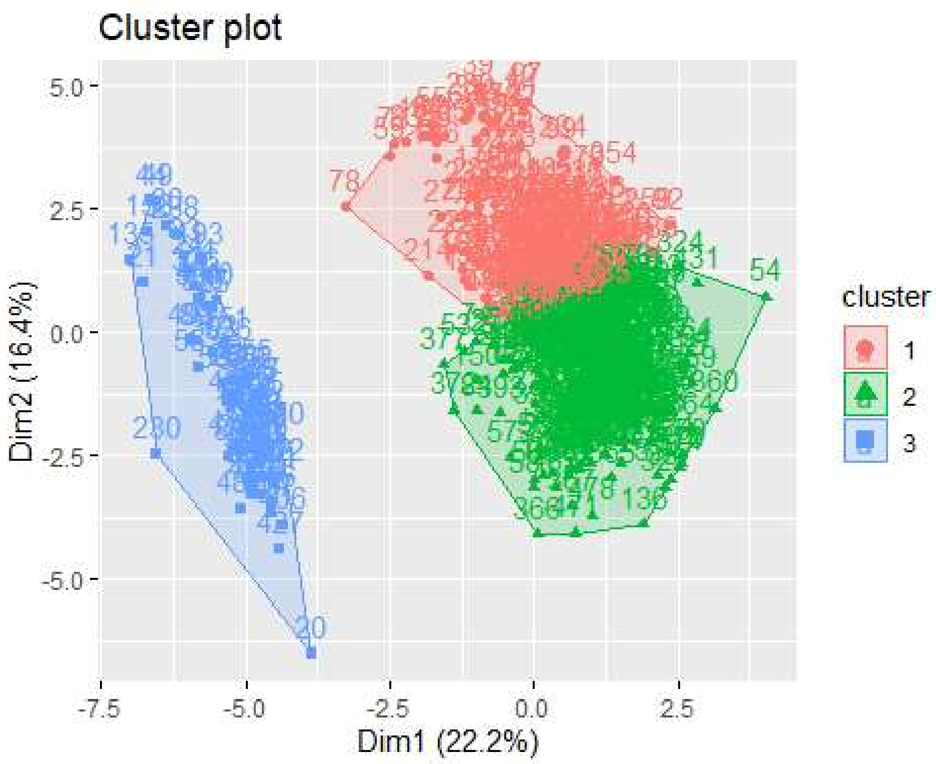

```r
ggparcoord(result2,
           columns = 11:19, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




```
BS1<-BathSoap[,c(2:11,20:22,31:35,47,49)]

data2.s<-as.data.frame(scale(BS1)) # scaling the data
# Elbow chart to estimate the optimal K
fviz_nbclust(data2.s,kmeans,method = "silhouette")
```




```r
# Choosing the optimal K as 3 and forming 3 clusters
model1<-kmeans(data2.s,3,nstart=50)

# Visualizing the clusters
fviz_cluster(model1,data2.s)
```



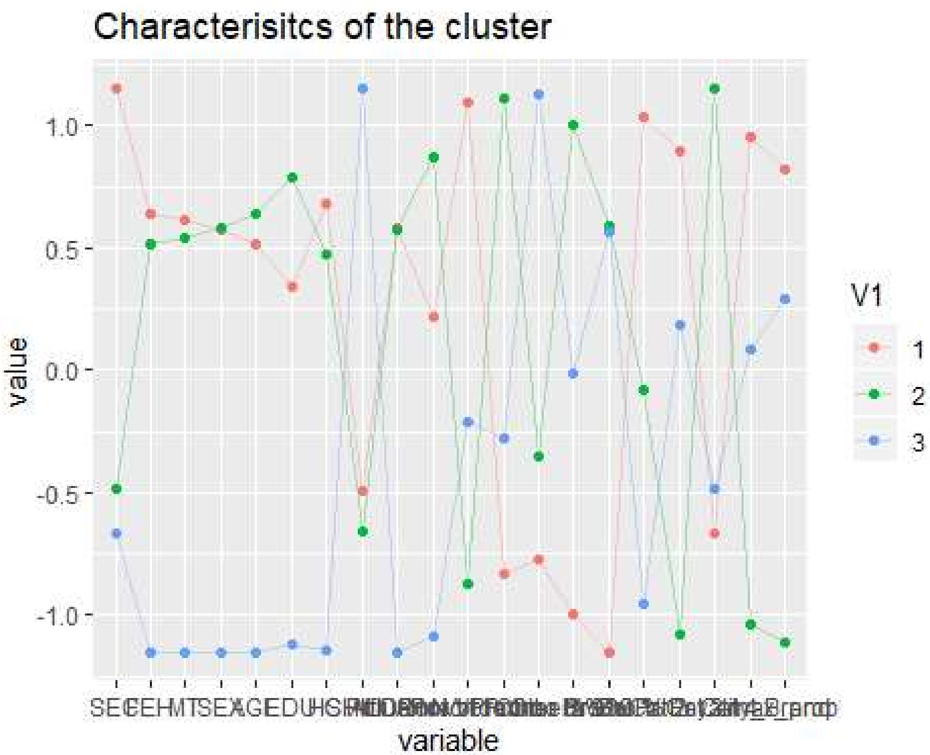

```
result1<-as.data.frame(cbind(1:nrow(model1$centers),model1$centers))
result1$V1<-as.factor(result1$V1)
# Characteristics of the cluster
result1
```

```
##   V1        SEC        FEH         MT        SEX         AGE         EDU
## 1  1  0.4112713  0.3256288  0.3025349  0.3267729  0.04667653 -0.09929787
## 2  2 -0.1957460  0.1730272  0.2075035  0.3522446  0.09210753  0.43959241
## 3  3 -0.2628475 -1.8047556 -1.9043115 -2.6805048 -0.58631729 -1.84626790
##          HS       CHILD         CS Affluence Index Pur Vol No Promo - %
## 1  0.3903697 -0.08543193  0.2425434      -0.2088666           0.34055052
## 2  0.1373364 -0.24631996  0.2299530       0.4332789          -0.20282369
## 3 -1.8223924  1.45152536 -1.8362598      -1.4916636          -0.01935313
##   Pur Vol Promo 6 % Pur Vol Other Promo % Others 999   Pr Cat 1
## 1        -0.3709517          -0.086197988 -0.8647333 -0.6509860
## 2         0.2643283          -0.005076589  0.5591275  0.3299121
## 3        -0.1901672           0.279502153 -0.1655857  0.3183424
##      Pr Cat 2   Pr Cat 3   Pr Cat 4 Loyality_Brand   max_prop
## 1  0.25868188  0.5417135 -0.2232630      0.9020746  0.5484161
## 2 -0.08409727 -0.3722635  0.1735000     -0.5988584 -0.3931422
## 3 -0.35527740  0.2108060 -0.1845976      0.2486052  0.2926239
```

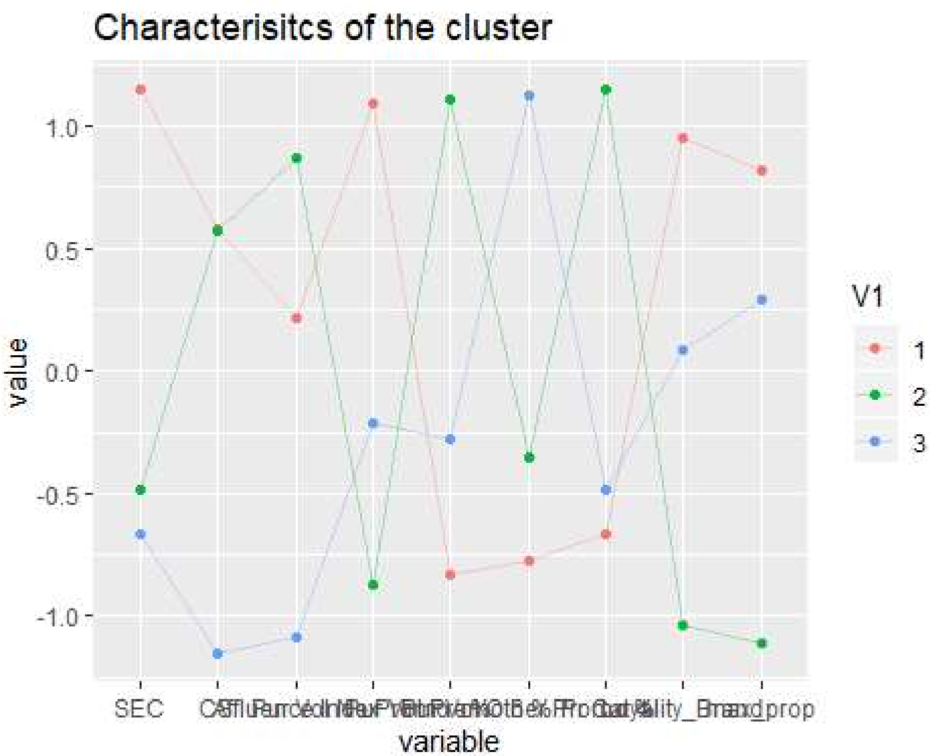
```
model1$size
```

```
## [1] 201 331  68
```

```r
# Parallel plot to visualize the cluster.
ggparcoord(result1,
           columns = 2:21, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```

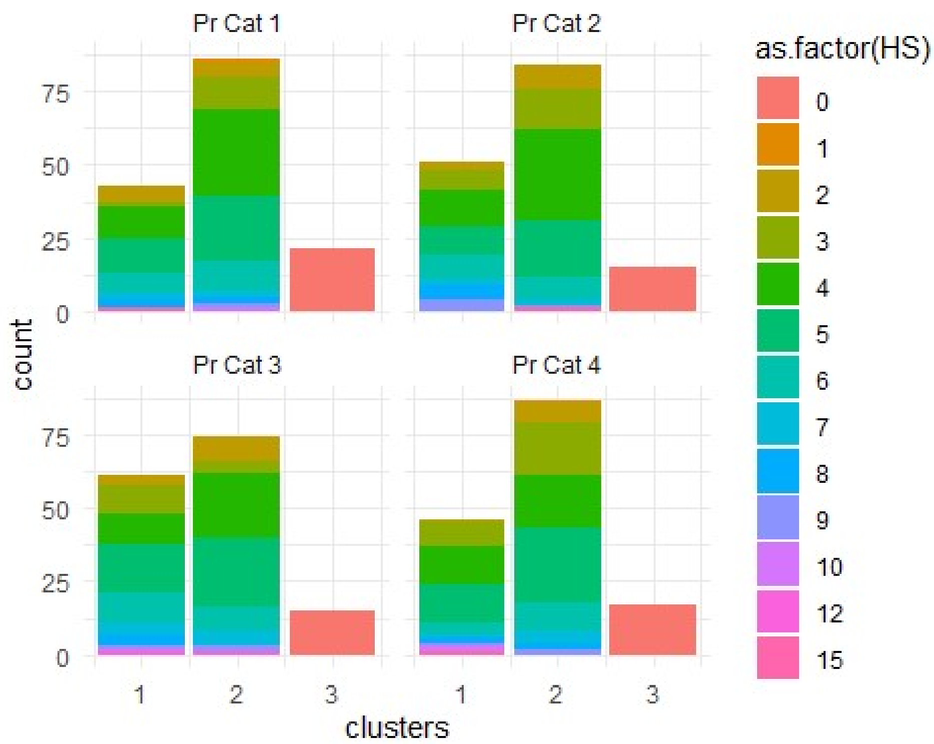



```r
ggparcoord(result1,
           columns = c(2,10:14,19:21), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```

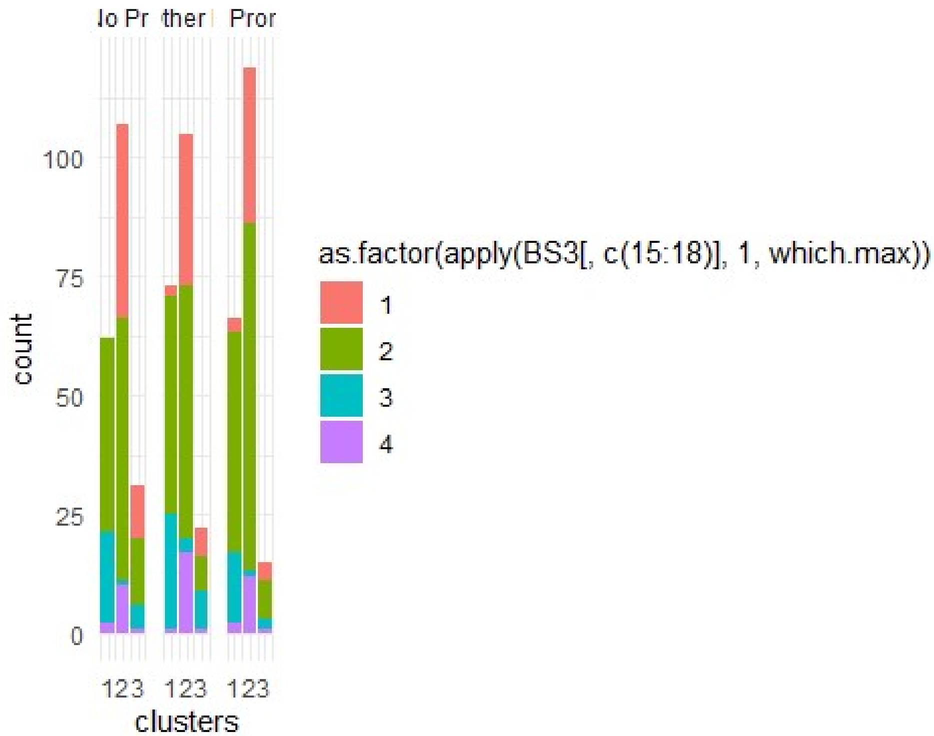



When we compare all the three models it is observed that variables with reasons of purchase can explain all the characterstics of data when it is compared with purchase bhaviour and combination of both. so the best segementation of the model is reasons of purchase.

```
r1<-BathSoap[,23:31]
BathSoap$Loyality<-as.numeric(apply(r1,1,which.max))
BS3 <- BathSoap[,c(2:4,6:11,19,20:22,31:35,47,48,50)]
BS3$clusters <- model1$cluster
ggplot(BS3) +
  aes(x = clusters,fill=as.factor(HS)) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pr Cat 1","Pr Cat 2","Pr Cat 3","Pr Cat 4")))
```




```r
ggplot(BS3) +
  aes(x = clusters,fill= as.factor(apply(BS3[,c(15:18)],1,which.max)))+
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other
Promo %")))
```




Suggested mail promotions when it comes in cluster 1 there is a minimum purchase of pricecategory 4 and price category 1 even though where there is availabilty of all promos or no promos.

when it comes in cluster 2 there is a minimum purchase of pricecategory 3 even though where there is availabilty of all promos or no promos.

when it comes in cluster 3 there is a minimum purchase of pricecategory 4 even though where there is availabilty of all promos or no promos.