final exam

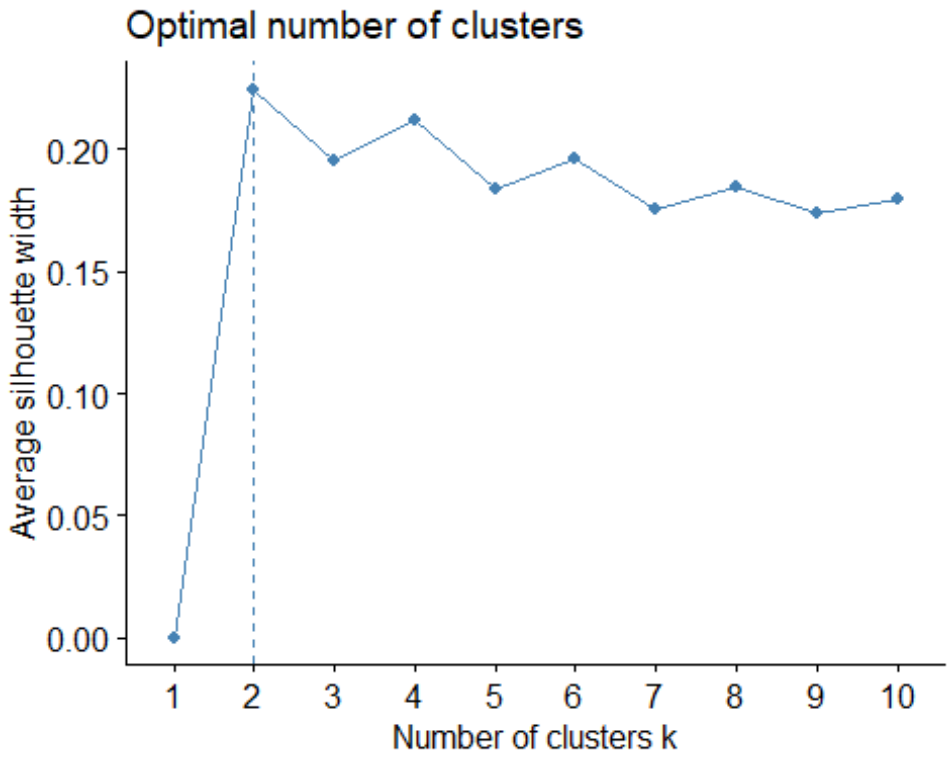
Rithin Gujja

04/12/2019

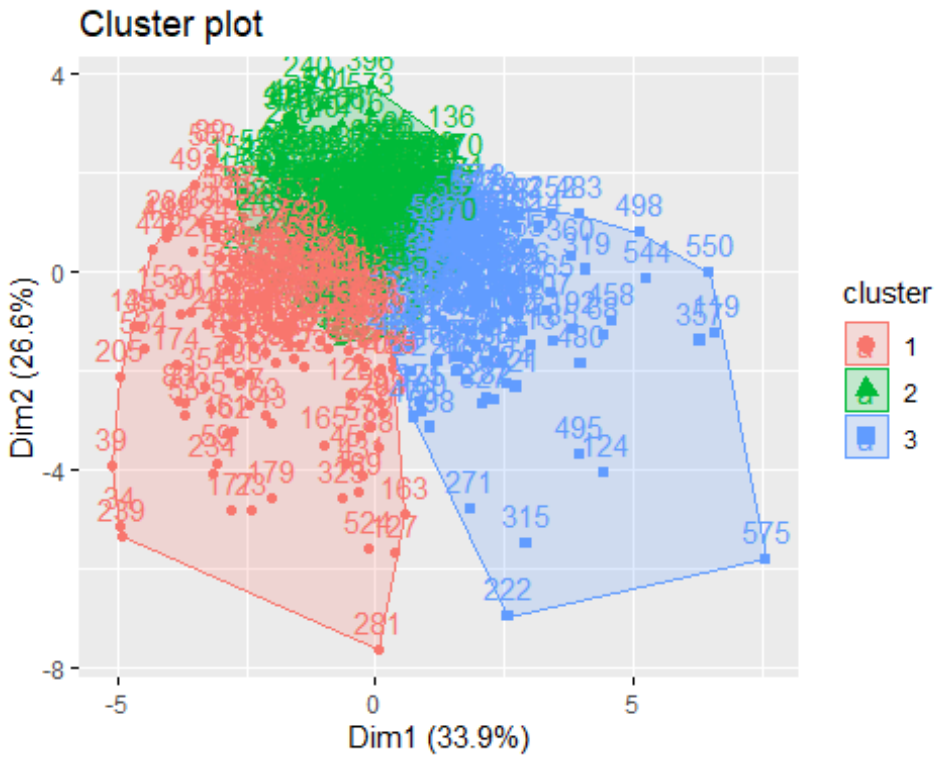
1Q)

```
## By taking variables that describe purchase behaviour.
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)
BathSoap<-read.csv("BathSoap.csv")
# To find the brand loyality, we are fininding the max value between the
brands. The max value in a particular brand corresponds the loyality of the
brand to the customer.
loy1<-BathSoap[,23:30]
BathSoap$Loyality_Brand<-as.numeric(apply(loy1,1,max))

data1<-BathSoap[,c(12,13,14,15,16,17,18,19,31,47)]
```

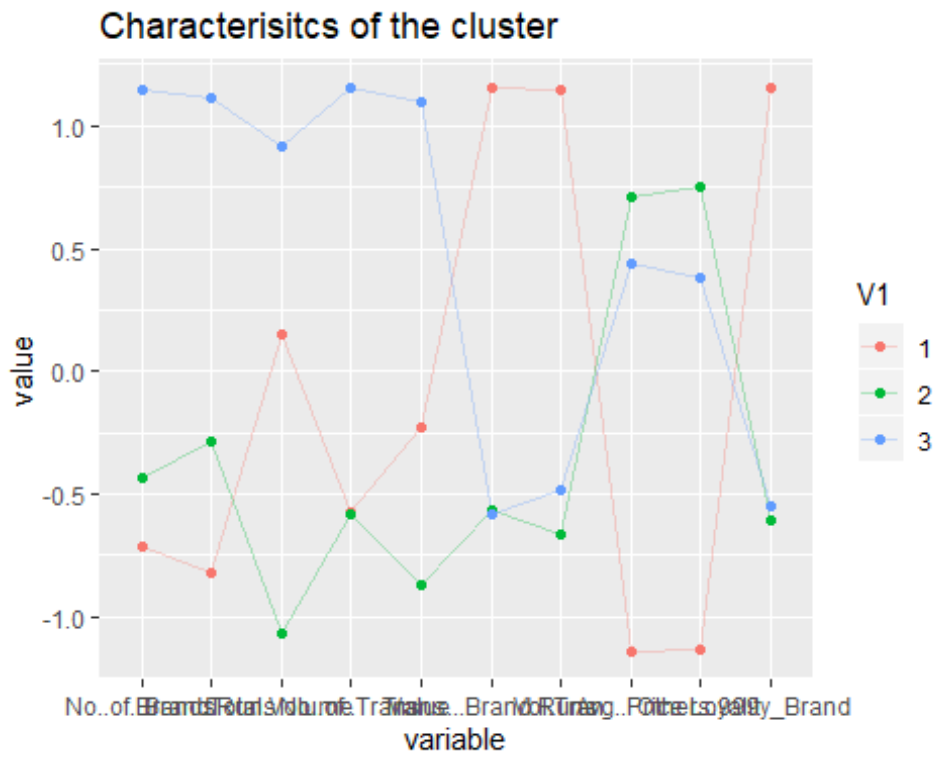
```r
data11<-as.data.frame(scale(data1))
fviz_nbclust(data11,kmeans,method = "silhouette")
```



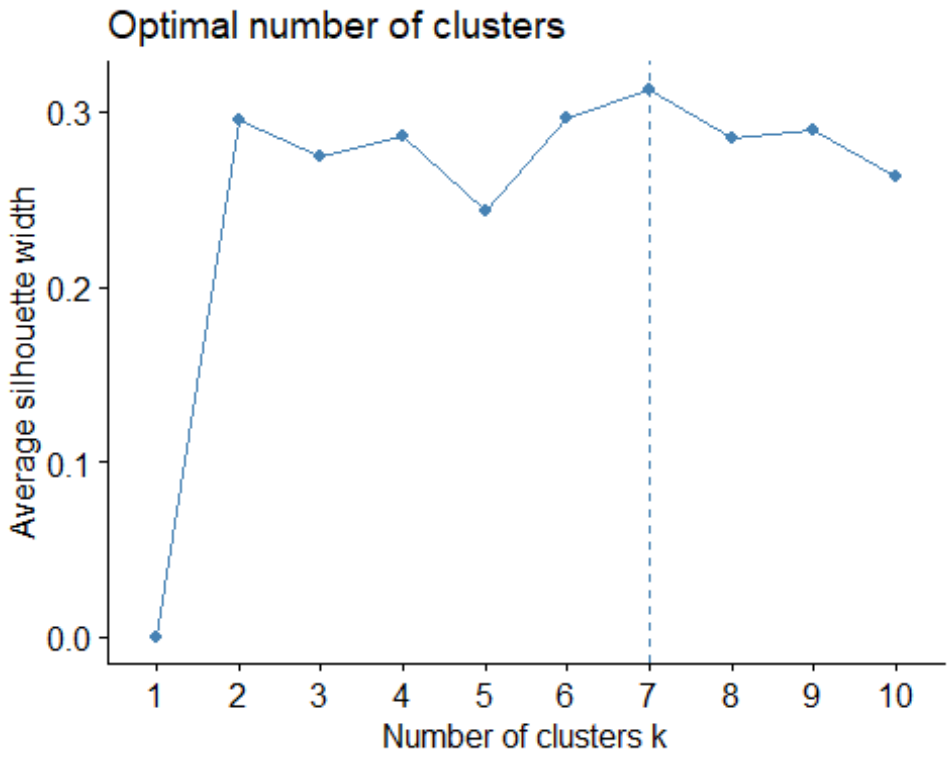

```r
# By taking thr optimal k value as 3
ms<-kmeans(data11,3,nstart=50)
fviz_cluster(ms,data11)
```



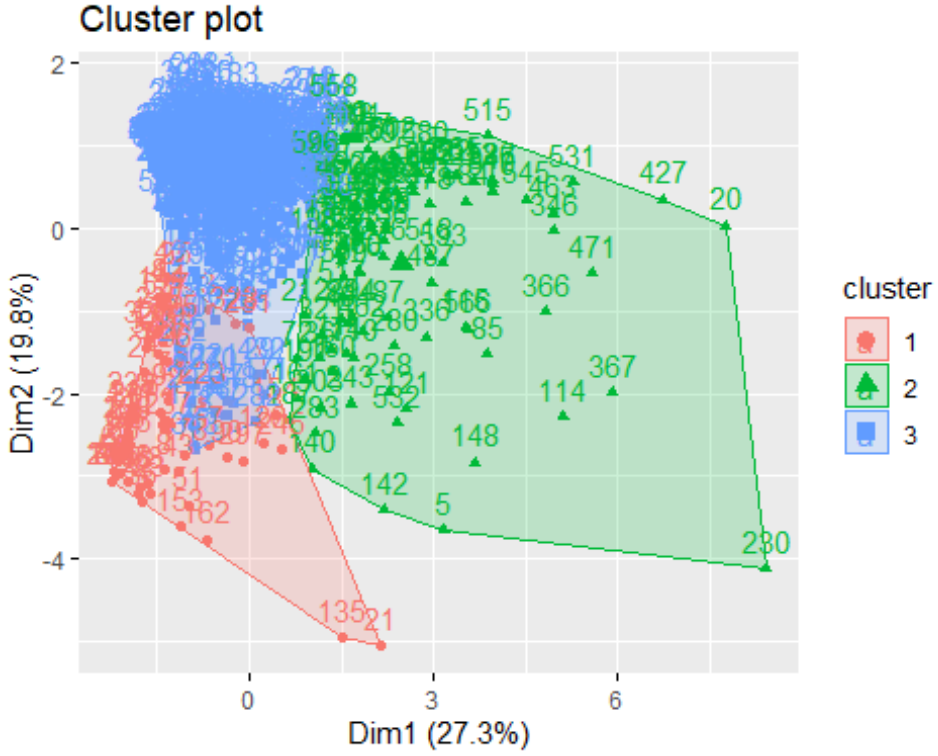

```
result<-as.data.frame(cbind(1:nrow(ms$centers),ms$centers))
result$V1<-as.factor(result$V1)
ms$size

## [1] 175 259 166

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




```
## by taking variables that describe basis of purchase

loy2<-BathSoap[,36:46]
BathSoap$max_prop_no<-as.numeric(apply(loy2,1,which.max))
BathSoap$max_prop<-as.numeric(apply(loy2,1,max))
data2<-BathSoap[,c(20:22,32:35,49)]
data22<-as.data.frame(scale(data2))
fviz_nbclust(data22,kmeans,method = "silhouette")
```



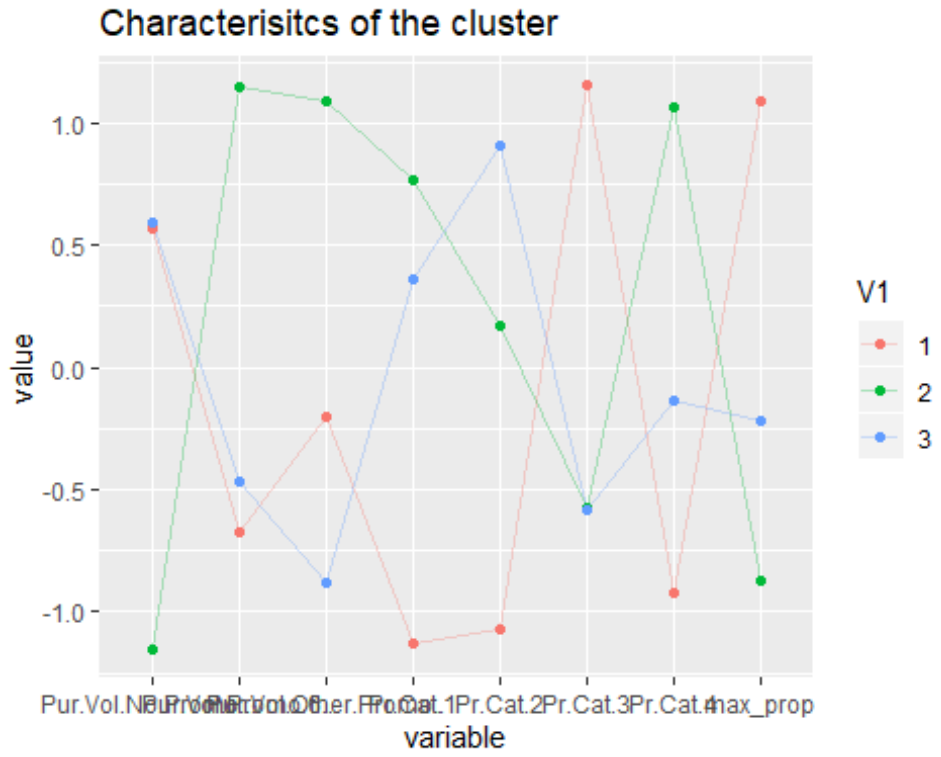

```
# By choosing k=3
ms1<-kmeans(data22,3,nstart=50)
fviz_cluster(ms1,data22)
```



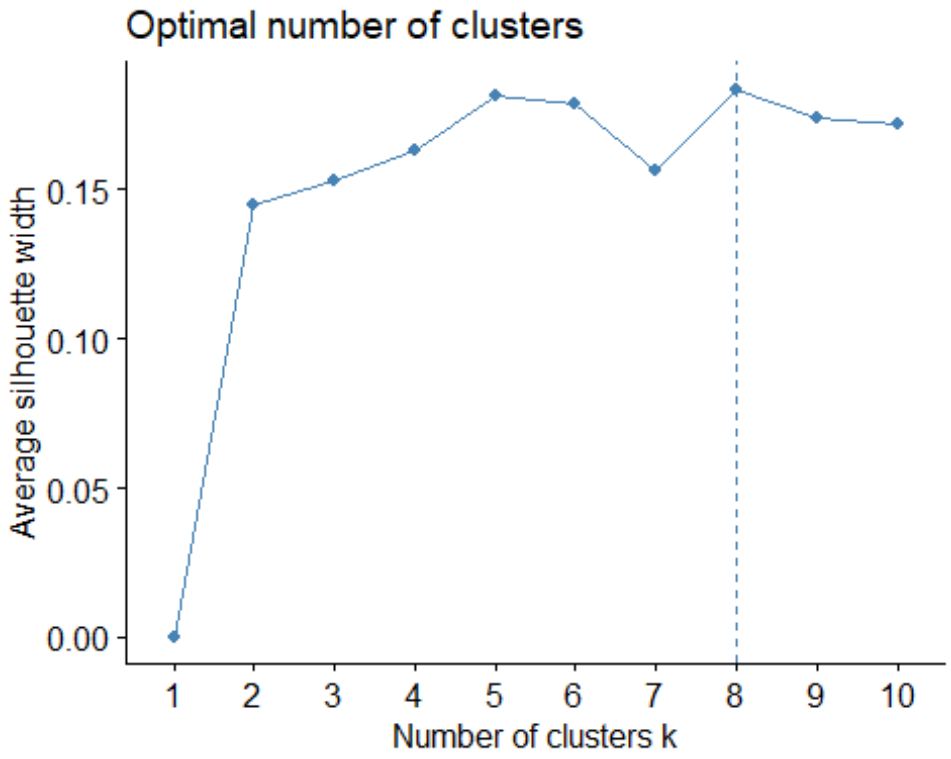
```
result1<-as.data.frame(cbind(1:nrow(ms1$centers),ms1$centers))
result1$V1<-as.factor(result1$V1)
ms1$size
```

```
## [1]  67 105 428
```

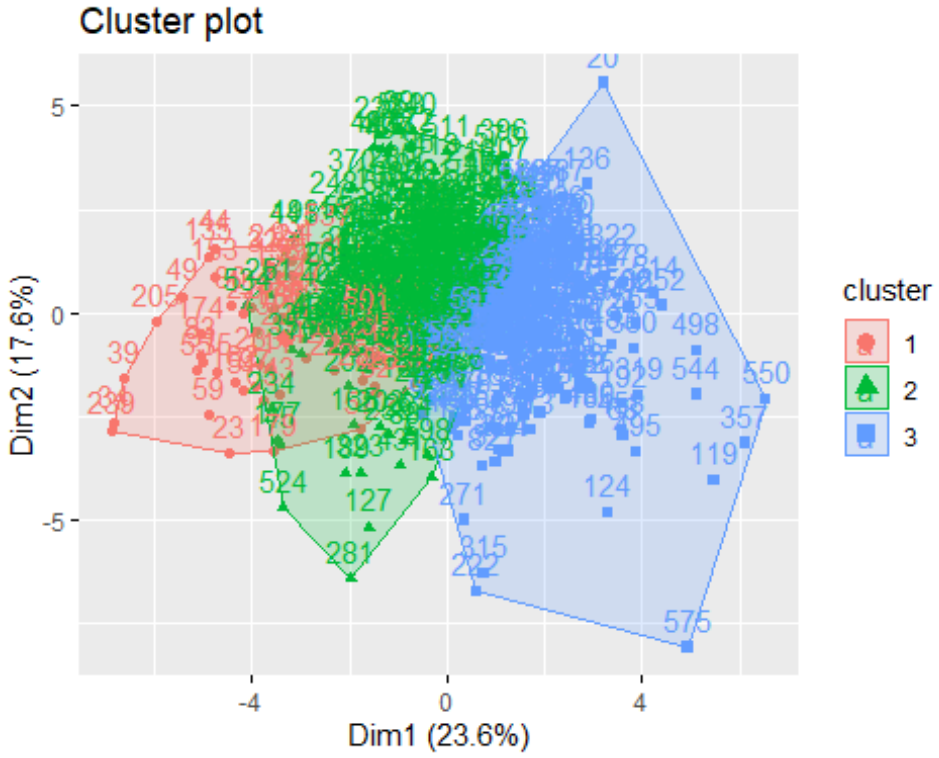
```
#visualize the cluster.
ggparcoord(result1,
           columns = 2:ncol(result1), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




```
## By taking the variables that describe the purchase behaviour and basis of purchase and forming the cluster.

data3<-BathSoap[,c(12:22,31:35,49)]
data33<-as.data.frame(scale(data3))
fviz_nbclust(data33,kmeans,method = "silhouette")
```



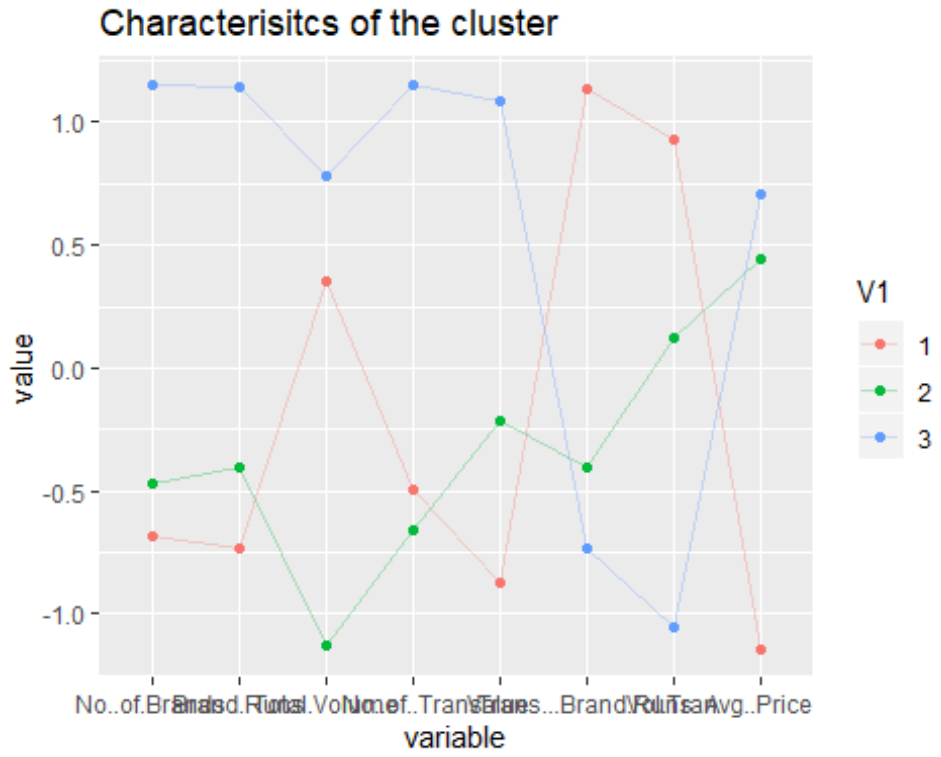

```r
# By taking k vcalue as 3 and taking 3 clusters.
ms2<-kmeans(data33,3,nstart=50)
fviz_cluster(ms2,data33)
```



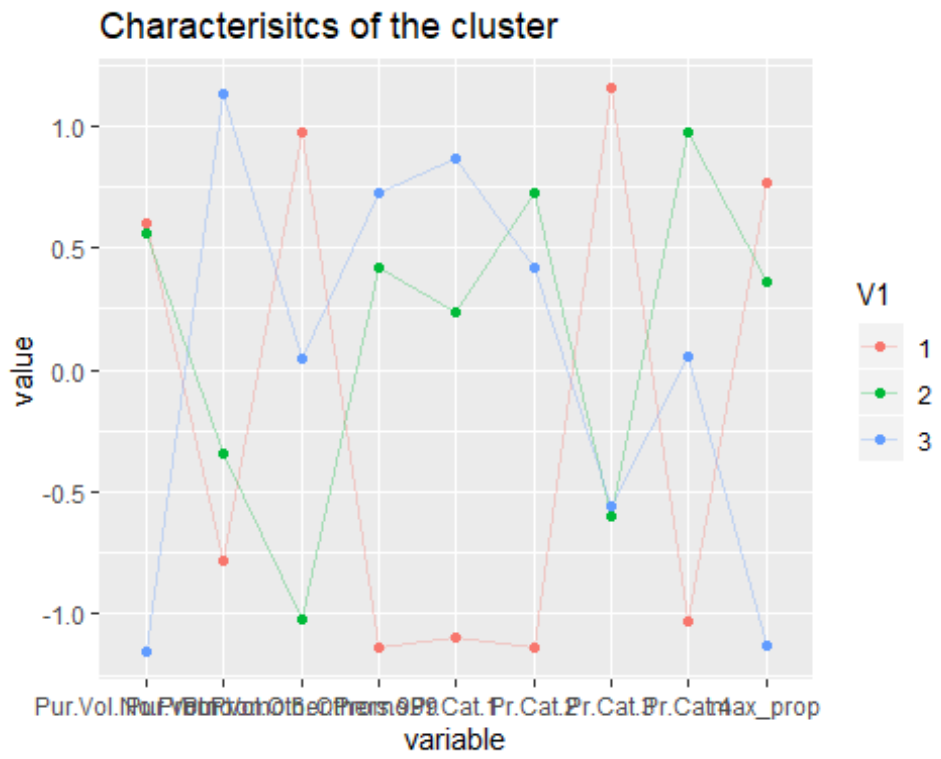
```
result2<-as.data.frame(cbind(1:nrow(ms2$centers),ms2$centers))
result2$V1<-as.factor(result2$V1)
ms2$size
```

```
## [1]  65 294 241
```

```
# visualizing the cluster.
ggparcoord(result2,
           columns = 2:9, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



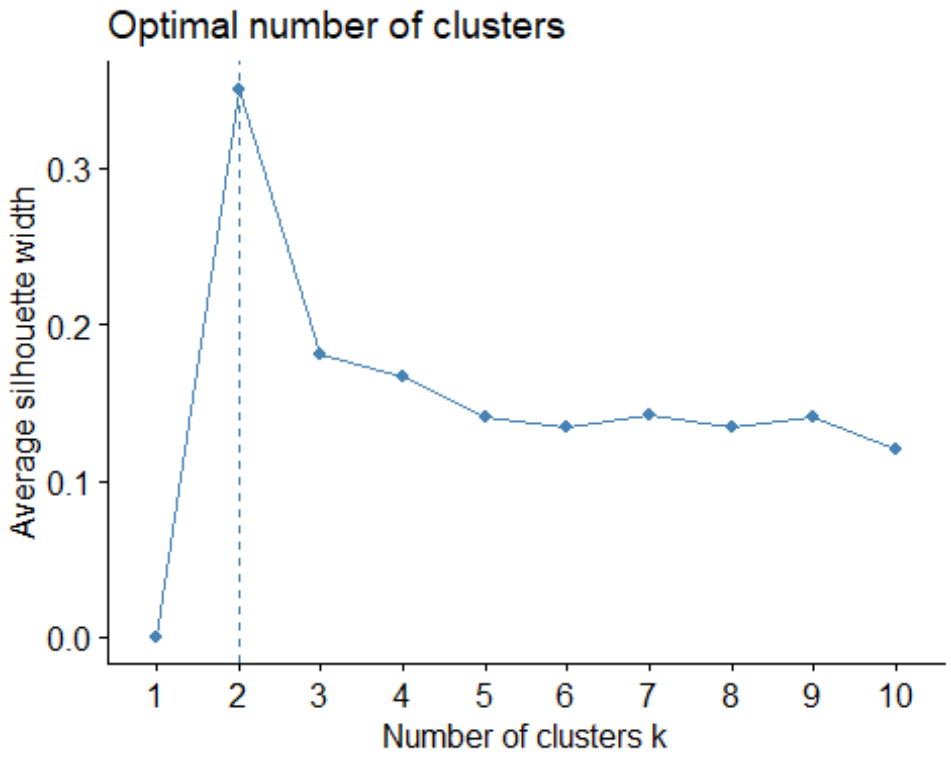

```
ggparcoord(result2,
           columns = 10:18, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



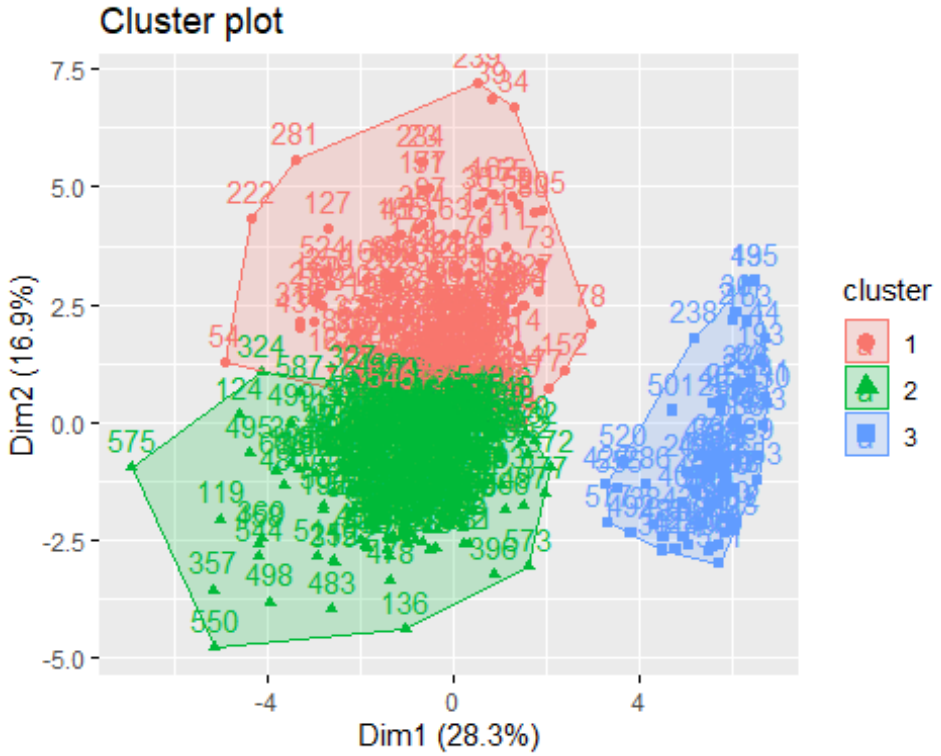

```
## comparing the cluster sizes
ms$size

## [1] 175 259 166

ms1$size

## [1]  67 105 428

ms2$size

## [1]  65 294 241
```

2Q)

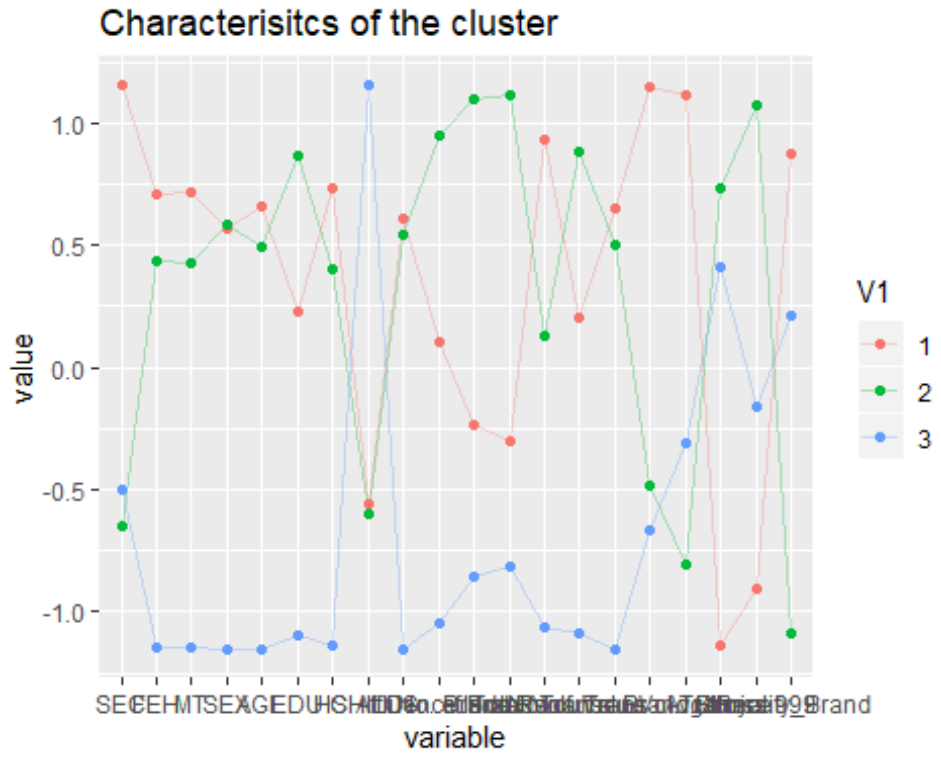
```
## adding demographics to the describe purchase behaviour variables
data4<-BathSoap[,c(2:11,12,13,14,15,16,17,18,19,31,47)]
data44<-as.data.frame(scale(data4))
fviz_nbclust(data44,kmeans,method = "silhouette")
```



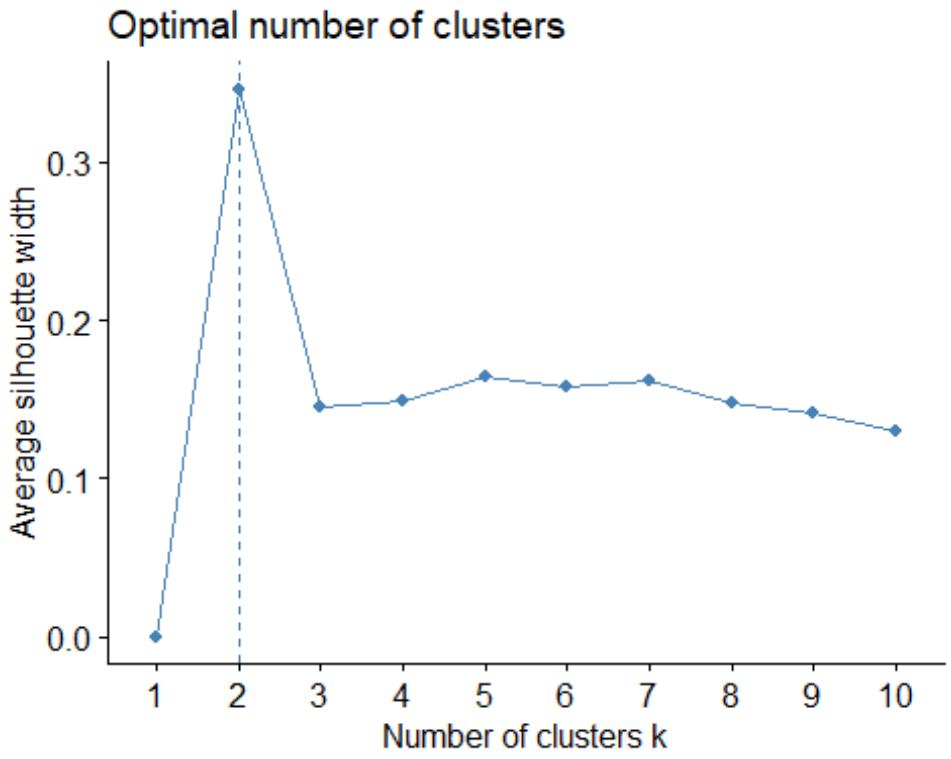

```
# Taking k value as 3
mp<-kmeans(data44,3,nstart=50)
fviz_cluster(mp,data44)
```



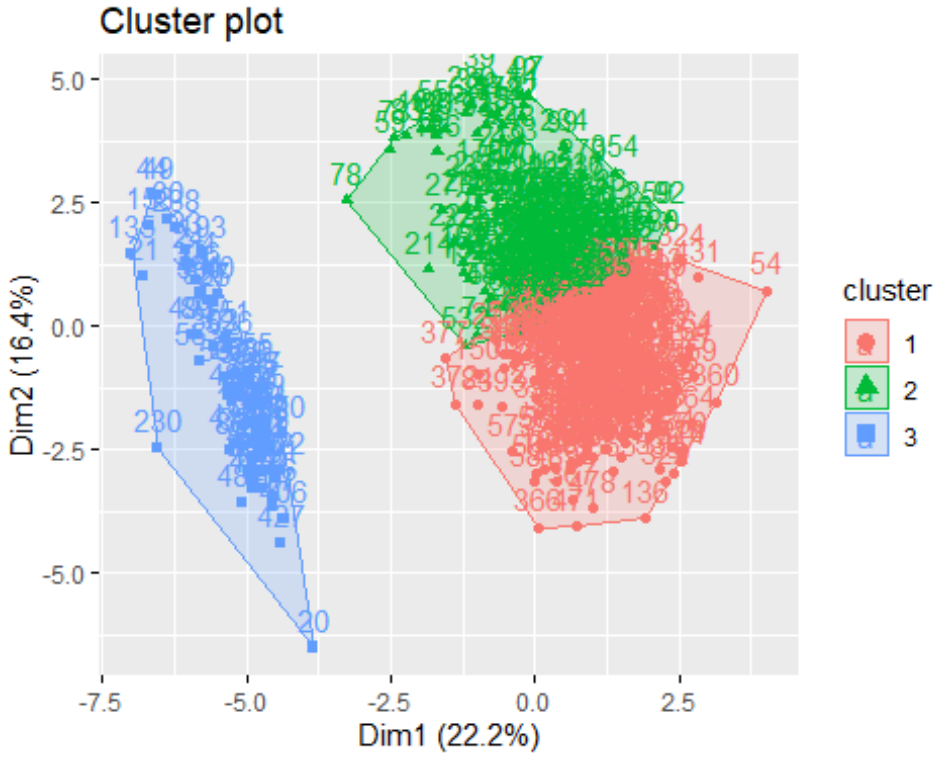
```
out<-as.data.frame(cbind(1:nrow(mp$centers),mp$centers))
out$V1<-as.factor(out$V1)

#visualizing the cluster the cluster.
ggparcoord(out,
           columns = 2:ncol(out), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



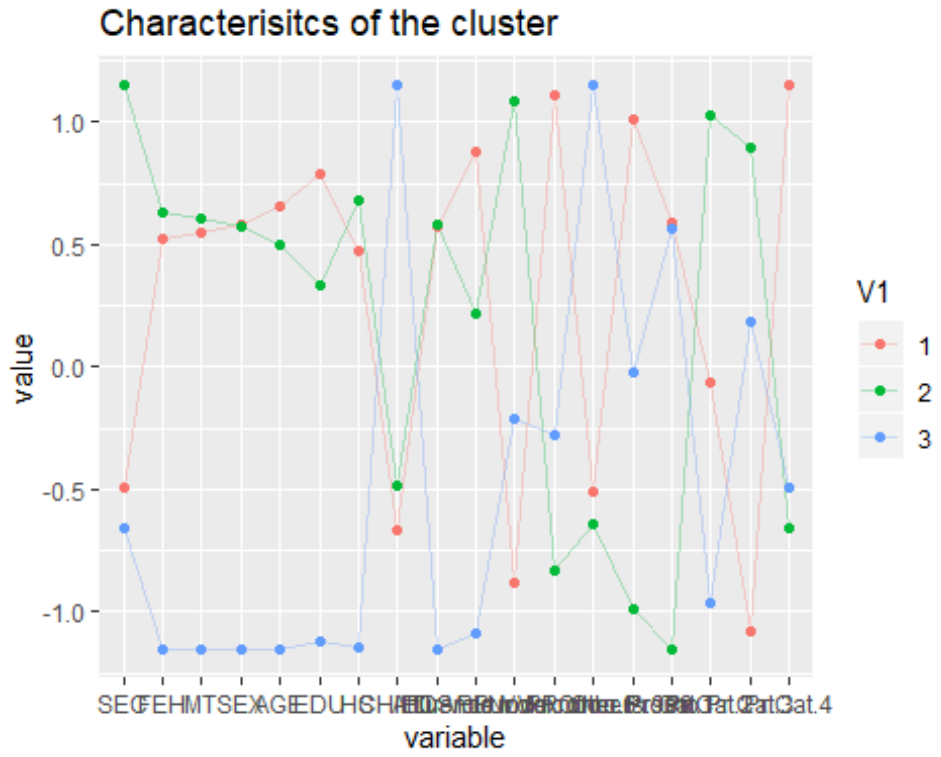

```
## adding the demographics to the basis of purchase variables

data5<-BathSoap[,c(2:11,20:22,31:35,47,49)]
data55<-as.data.frame(scale(data5))
fviz_nbclust(data55,kmeans,method = "silhouette")
```



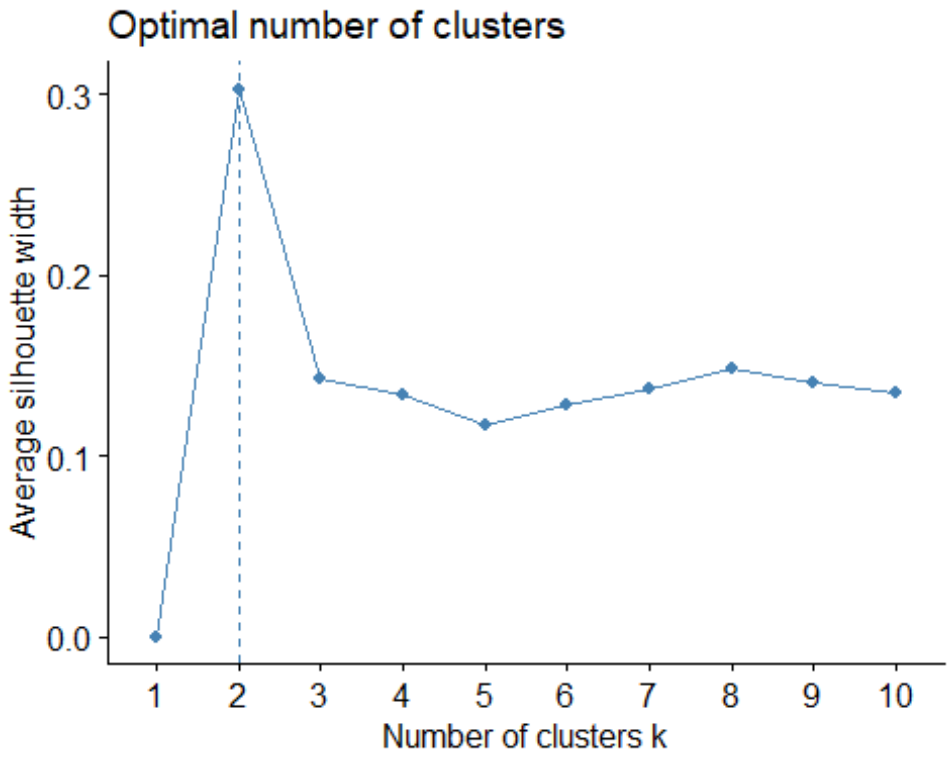

```
## choosing k value as 3 and forming 3 clusters
mp1<-kmeans(data55,3,nstart=50)
fviz_cluster(mp1,data55)
```



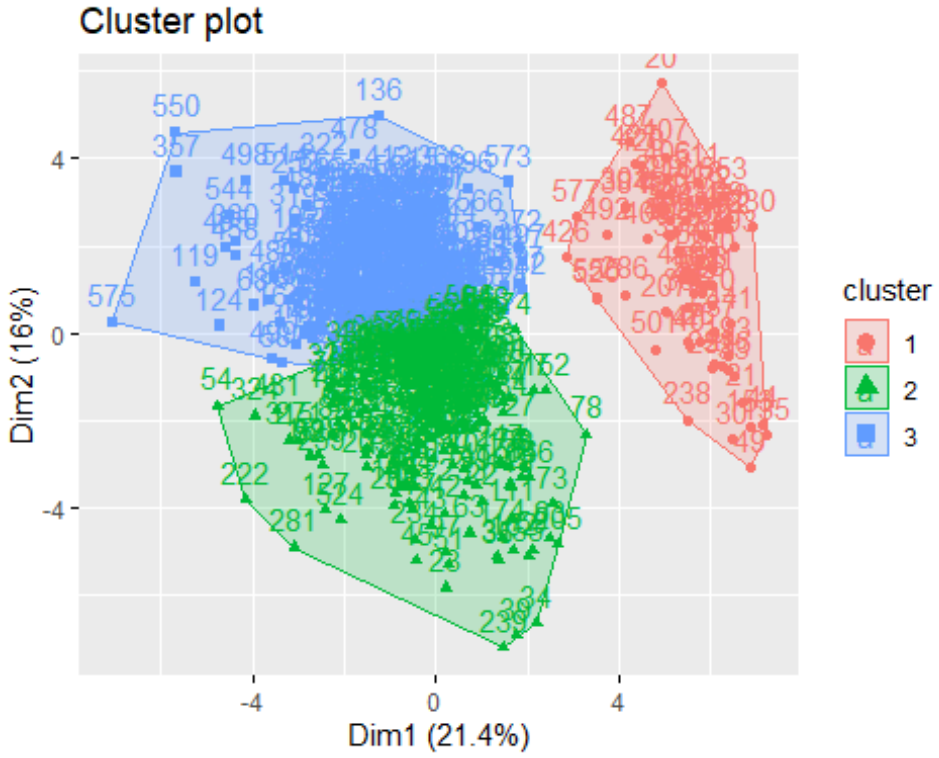
```r
out1<-as.data.frame(cbind(1:nrow(mp1$centers),mp1$centers))
out1$V1<-as.factor(out1$V1)

# visualizing the cluster.
ggparcoord(out1,
           columns = 2:19, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



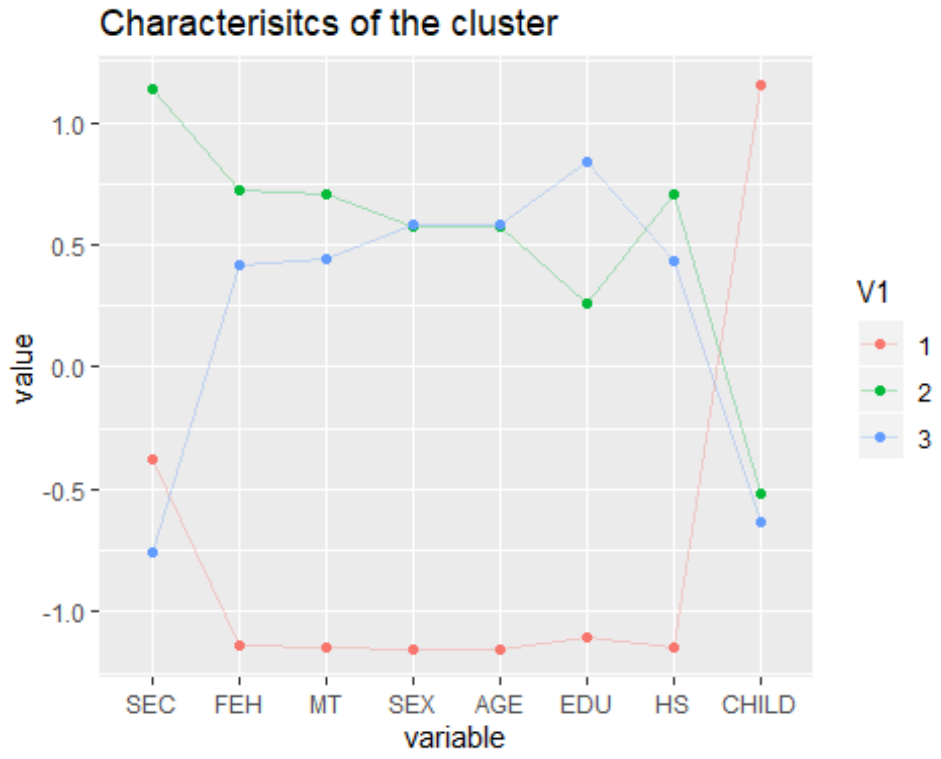

```r
## taking all the variables and forming the cluster

data6<-BathSoap[,c(2:11,12:22,31:35,49)]
data66<-as.data.frame(scale(data6))
fviz_nbclust(data66,kmeans,method = "silhouette")
```



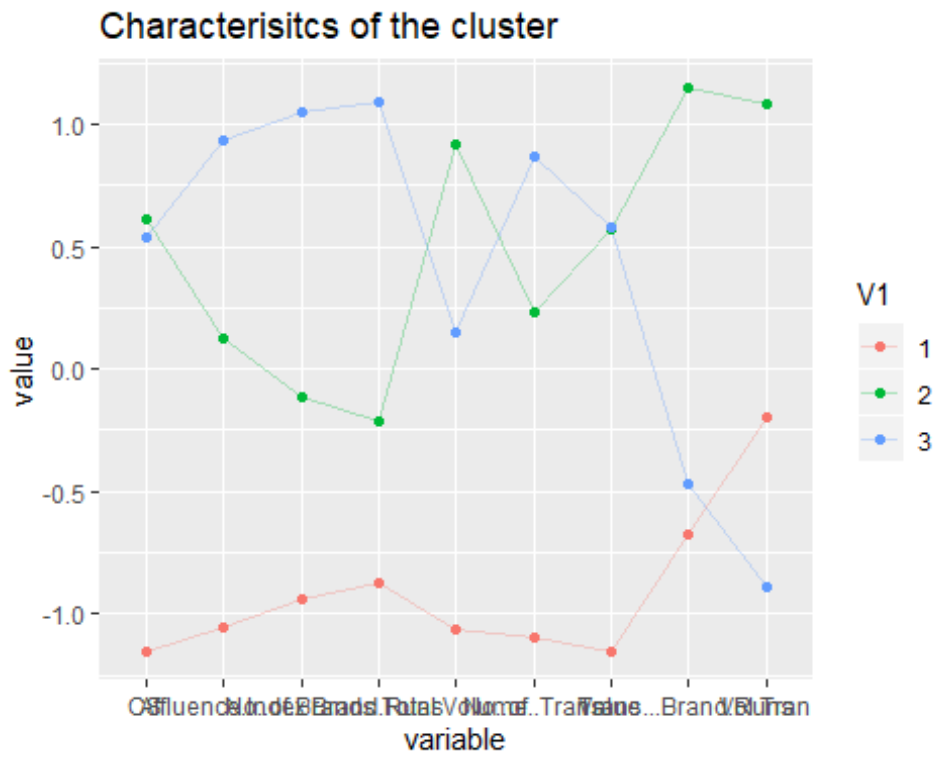

```r
# Choosing the optimal K as 3 and forming 3 clusters
mp2<-kmeans(data66,3,nstart=50)
fviz_cluster(mp2,data66)
```



```r
out2<-as.data.frame(cbind(1:nrow(mp2$centers),mp2$centers))
out2$V1<-as.factor(out2$V1)

# Parallel plot to visualize the cluster.
ggparcoord(out2,
           columns = 2:9, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




```r
ggparcoord(out2,
           columns = 10:18, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




```
mp$size
```

```
## [1] 202 330  68
```

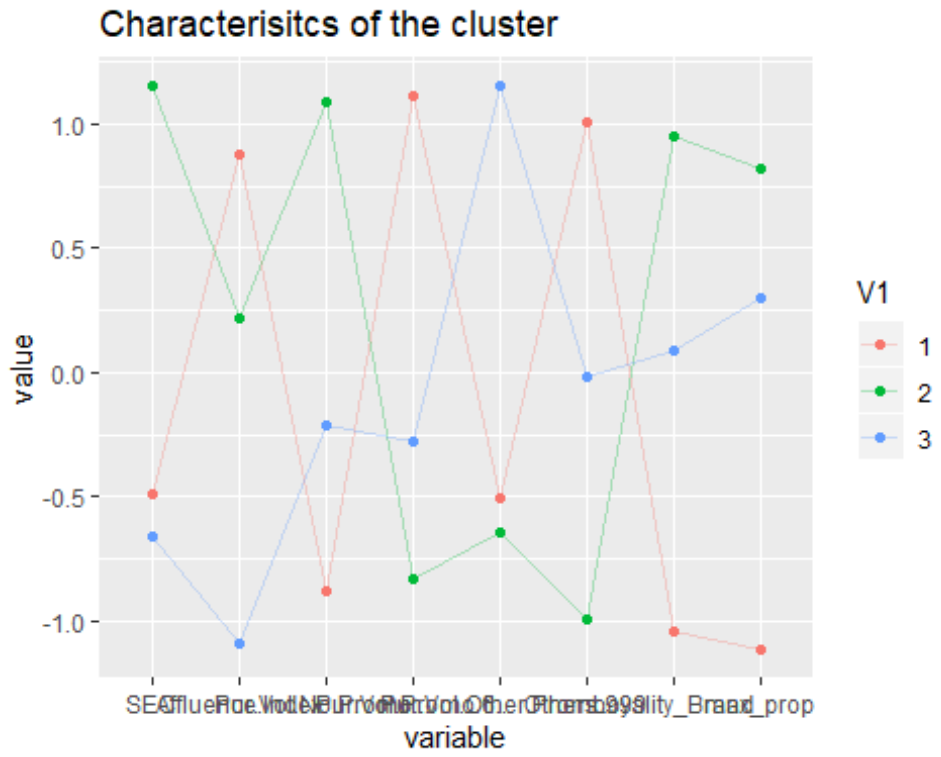
```
mp1$size
```

```
## [1] 329 203  68
```

```
mp2$size
```

```
## [1]  68 256 276
```

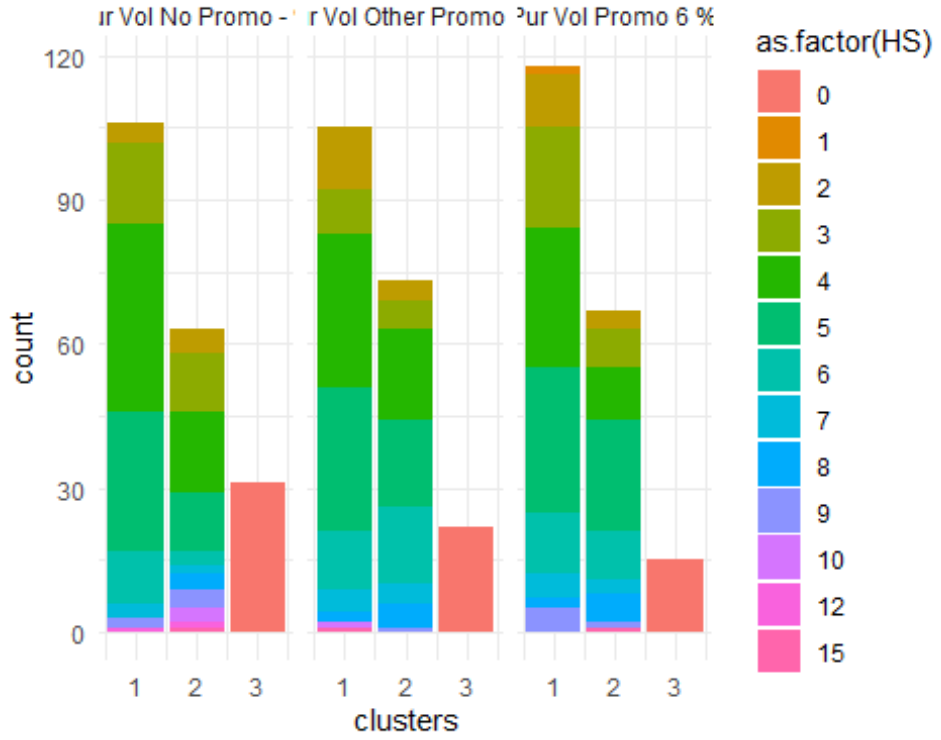
```
## By comparing the above clusters formed cluster mp and cluster mp1 are
having the same sizes, but cluster mp1 has less variables and forming the
similar cluster to size.
## By taking cluster with demographics basis of purchase is the optimal
segmentation.

ggparcoord(out1,
           columns = c(2,11:15,20,21), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```




```
## By seeing the visualization customers buying more other products are not
loyal.
## People with cluster 2 are having high SEC and are  buying products
irrespective of the promos and maintaining the loyality to the product.
## people with low SEC are fallin in cluster 1 and 3 and are buying products
when there is an promo offer and not maintaining the loyality to the product.
```

3Q)

```
ri<-BathSoap[,23:31]
BathSoap$Loyality<-as.numeric(apply(ri,1,which.max))
datax <- BathSoap[,c(2:4,6:11,19,20:22,31:35,47,48,50)]
datax$clusters <- mp1$cluster
ggplot(datax) +
  aes(x =clusters,fill= as.factor(HS)) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other
Promo %")))
```




```
# In cluster 1 and cluster 2 the people buying in other promo and promo of 6%
is larger then purchasing in no promo.

# In cluster 3 there are no people in house holds.

## By seeing the characterics of the cluster line graph diagram, the people
of high SEC must be given more promos to preserve their loyality.

## People with low SEC must not be given promo codes because they are using
their promo code and switching to other product who are providing promo code
and not maintaining the loyality.
```