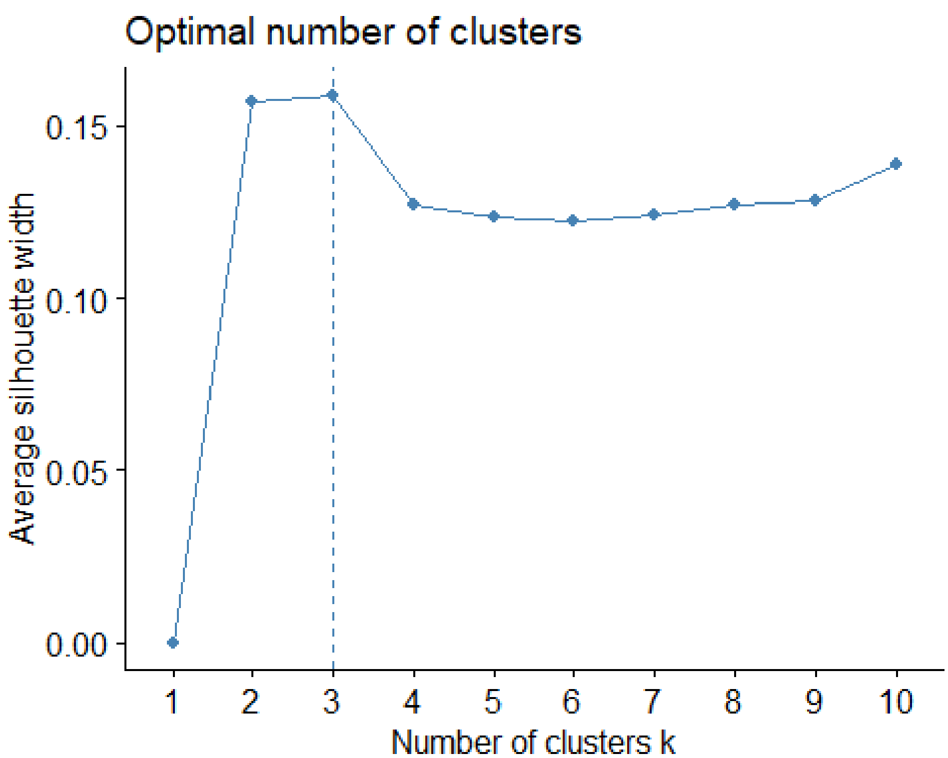
Final_Exam

Vijay

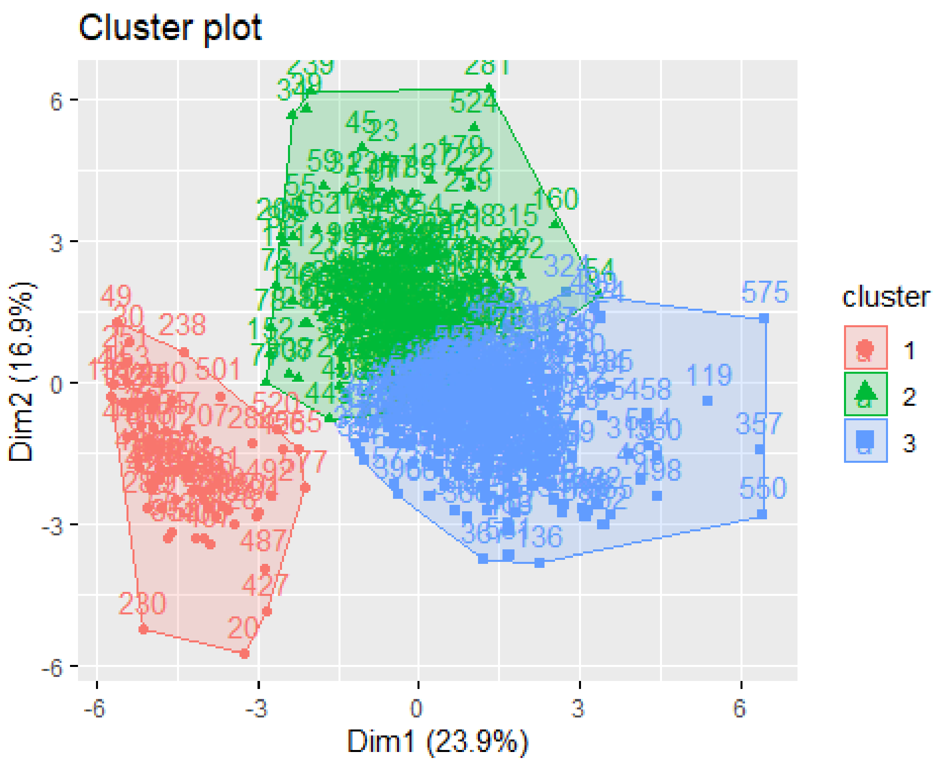
12/5/2019

```r
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(factoextra)

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

library(hrbrthemes)

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

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

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

library(GGally)

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

library(viridis)

## Loading required package: viridisLite

library(readr)
BathSoap<- read_csv("C:/Users/Vijay/Downloads/BathSoap (2).csv")

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

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

set.seed(123)

#Finding the brand Loyality
```

```r
# Here, I'm choosing the Bathsoap data columns from 23 to 31 which resembels the brand loyality varialbes. Loyality is measured by finding the maximum value amoung the rows, the highest value indicates the Brand and hence customer is loyal to that brand.
r1<-BathSoap[,23:31]
BathSoap$Loyality<-as.numeric(apply(r1,1,which.max))

# 1) i. The data includes the demographics of the customer and the purchase behaviour.
# Demographics includes 'SEC'(socioeconomic class), 'SEX'(Male and Female), 'AGE','EDU'(Education level of the homemaker), 'HS'(Number of members in the household),'CS'(Television Availability) and 'Affluence Index'(Weighted value of durables)
# Purchase behaviour includes 'No.of brands','Brand Runs','Total Volume','No.of Trans','Trans/Brand runs','Vol/Trans,'Avg.Price',Promotions(20:22),'Loyality'.
data1<-BathSoap[,c(2,5,6,7,8,10:15,17:22,47)]
data1.s<-as.data.frame(scale(data1)) # scaling the data
# Elbow chart to estimate the optimal K
fviz_nbclust(data1.s,kmeans,method = "silhouette")
```

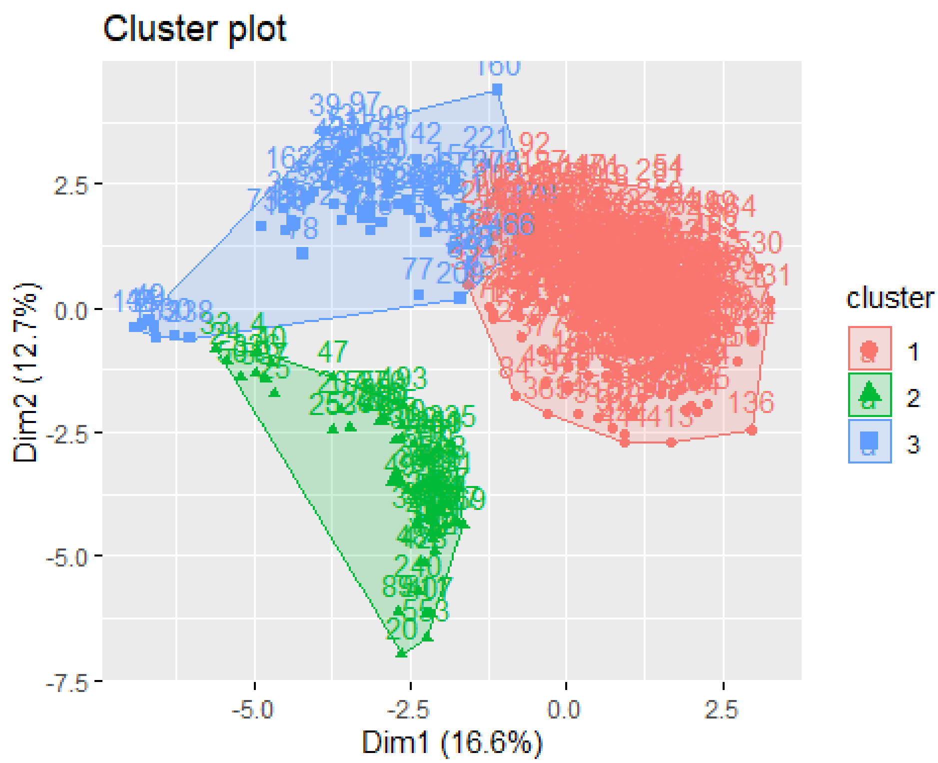



```r
# Choosing the optimal K as 3 and forming 3 clusters
model<-kmeans(data1.s,3,nstart=50)
# Visualizing the clusters
fviz_cluster(model,data1.s)
```




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

##   V1        SEC        SEX         AGE        EDU          HS        CS
## 1  1 -0.2628475 -2.6805048 -0.58631729 -1.8462679 -1.82239236 -1.8362598
## 2  2  0.5818049  0.3405506  0.03175431 -0.1427948  0.44193773  0.2955741
## 3  3 -0.4343853  0.3443883  0.11181105  0.5593413  0.05452243  0.1827527
##   Affluence Index No. of Brands Brand Runs Total Volume No. of  Trans
## 1       -1.491664    -0.7567786 -0.8757394   -1.0564151    -1.2079004
## 2       -0.273979    -0.3332655 -0.4268157    0.4400394    -0.1692321
## 3        0.587310     0.4638014  0.5718471   -0.1253429     0.4306589
##   Trans / Brand Runs   Vol/Tran Avg. Price Pur Vol No Promo - %
## 1         -0.3473341 -0.1165205  0.1847084         -0.01935313
## 2          0.3972486  0.5874040 -0.5934840          0.28061215
## 3         -0.2568194 -0.4738348  0.4628690         -0.23496155
##   Pur Vol Promo 6 % Pur Vol Other Promo %   Loyality
## 1        -0.1901672            0.27950215 -0.1585186
## 2        -0.2792348           -0.10590377 -0.4575637
## 3         0.2834282            0.02418215  0.4281616

# ii) The data includes the demographics of the customer and the purchase bas
is.
# Purchase basis includes the 'Price categorrywise purchase' and 'selling pro
positionwise purchase'.
data3<-BathSoap[,c(2,5,6,7,8,10,11,32:46)]
```

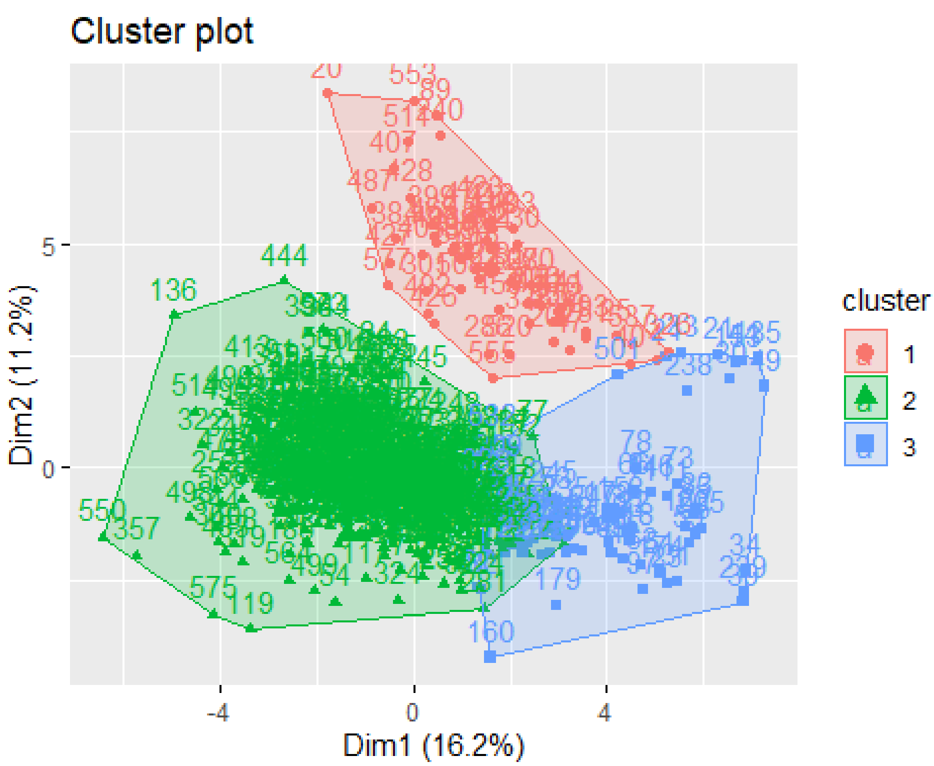
```
data3.s<-as.data.frame(scale(data3))
# Choosing K=3 as the optimal K
model1<-kmeans(data3.s,3,nstart=50)
fviz_cluster(model1,data3.s)
```




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

```
##   V1         SEC         SEX        AGE        EDU         HS         CS
## 1  1 -0.07479309  0.34723610  0.1055909  0.3476650  0.1965619  0.2192707
## 2  2 -0.43219025 -2.68050476 -0.5116665 -1.8462679 -1.8223924 -1.8362598
## 3  3  0.83859157  0.02327078 -0.2464889 -0.6765694  0.2680558  0.1346810
##   Affluence Index    Pr Cat 1    Pr Cat 2   Pr Cat 3    Pr Cat 4
## 1       0.3006678  0.06207737  0.2087226 -0.3535814  0.06352864
## 2      -1.4916636  0.46829067 -0.2291576 -0.1121263 -0.15786928
## 3      -0.6728730 -0.78758609 -1.1409055  2.3508034 -0.27362083
##      PropCat 5   PropCat 6   PropCat 7  PropCat 8   PropCat 9  PropCat 10
## 1  0.17036421  0.03523238  0.07809789  0.0525775  0.04003932  0.01395953
## 2 -0.01078491 -0.08632457 -0.04438654  0.1805590 -0.05520014  0.20032884
## 3 -1.07851836 -0.15277382 -0.46145256 -0.4865097 -0.20946734 -0.25650962
##     PropCat 11  PropCat 12  PropCat 13 PropCat 14  PropCat 15
## 1  0.06383757 -0.01302102 -0.02319284 -0.3552043  0.04897987
## 2 -0.17871700  0.28854389  0.45301047 -0.1027962 -0.22604282
## 3 -0.25817219 -0.15799155 -0.23049005  2.3533670 -0.12378092
```

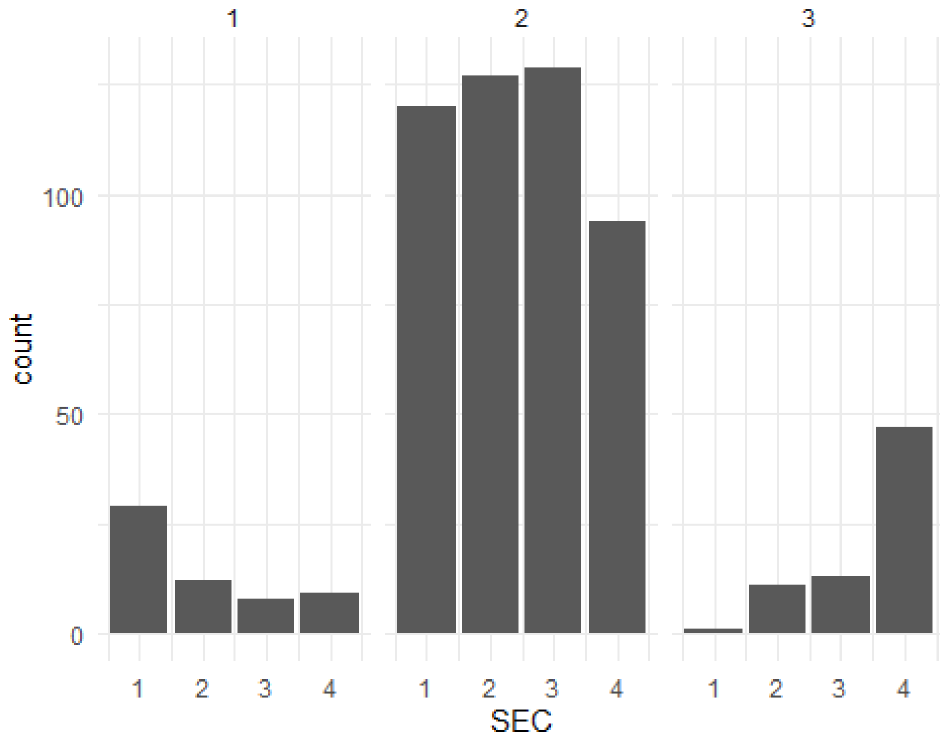
```r
# iii)
# Considering data includes demographics, purchase behaviour and purchase bas
is.
data4<-BathSoap[,c(2,5,6,7,8,10,11,12,13,15,17:22,32:47)]
data4.s<-as.data.frame(scale(data4))
model2<-kmeans(data4.s,3,nstart=50)
fviz_cluster(model2,data4.s)
```



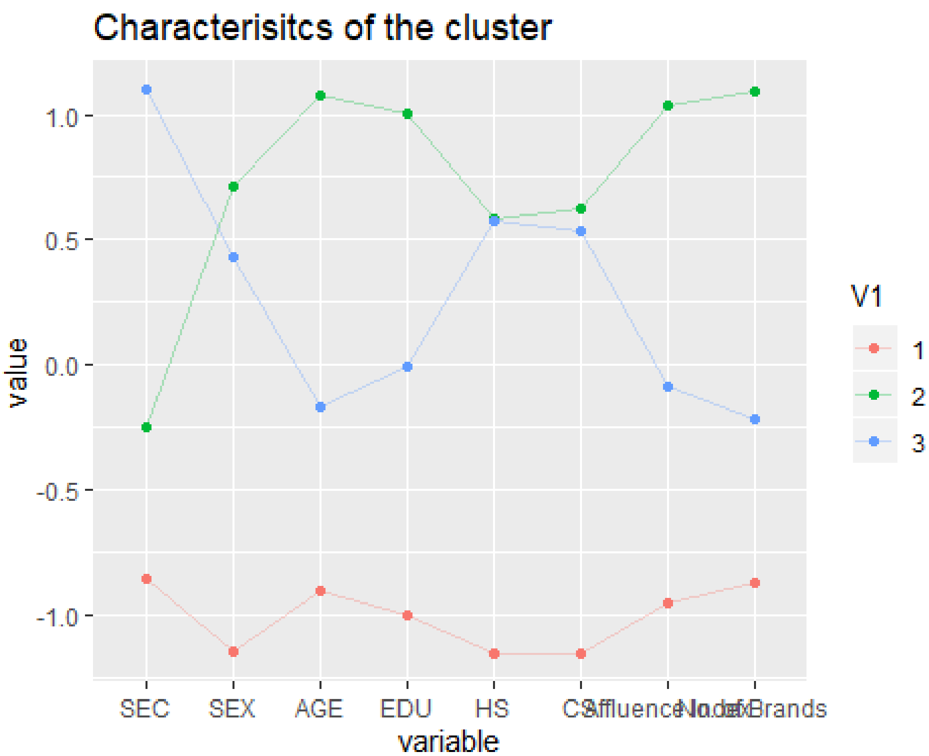

```r
result2<-as.data.frame(cbind(1:nrow(model2$centers),model2$centers))
result2$V1<-as.factor(result2$V1)
#Characteristics of the clusters.
result2

##   V1        SEC        SEX        AGE        EDU         HS         CS
## 1  1 -0.4930657 -2.6805048 -0.5054616 -1.8462679 -1.8223924 -1.8362598
## 2  2 -0.0722551  0.3477149  0.1050528  0.3435663  0.1960299  0.2101638
## 3  3  0.8688570 -0.1105098 -0.2785838 -0.7554528  0.1883984  0.1073068
##   Affluence Index No. of Brands Brand Runs No. of  Trans
## 1      -1.4916636    -0.7195417 -0.8268658    -1.2068819
## 2       0.2930593     0.1545755  0.2076637     0.1880074
## 3      -0.7114080    -0.4294039 -0.6894961    -0.2550602
##   Trans / Brand Runs    Vol/Tran Avg. Price Pur Vol No Promo - %
## 1        -0.51413657 -0.20256792  0.4716573          -0.01829318
## 2        -0.09765404 -0.05769552  0.1423375          -0.02496721
## 3         1.05162946  0.53980323 -1.3090938           0.17771657
##   Pur Vol Promo 6 % Pur Vol Other Promo %    Pr Cat 1   Pr Cat 2
## 1       -0.12370247            0.18972283  0.53039031 -0.1940031
```

```
## 2        0.07785093           -0.05930766  0.05513991  0.2082238
## 3       -0.40854436            0.23431495 -0.78719992 -1.2029587
##     Pr Cat 3    Pr Cat 4   PropCat 5   PropCat 6   PropCat 7   PropCat 8
## 1 -0.2293591 -0.14212727  0.03450861 -0.06826811 -0.03341022  0.21710786
## 2 -0.3410221  0.05705703  0.15863569  0.04453757  0.07267643  0.05016556
## 3  2.4108777 -0.25796419 -1.06333716 -0.23573764 -0.44750181 -0.50236207
##     PropCat 9 PropCat 10  PropCat 11  PropCat 12 PropCat 13 PropCat 14
## 1 -0.06557771  0.2244201 -0.17252232  0.31569143  0.4899944 -0.2208832
## 2  0.03620182  0.0115817  0.06075401 -0.01329945 -0.0250028 -0.3424752
## 3 -0.18349092 -0.2563856 -0.25761234 -0.16749111 -0.2315050  2.4135357
##    PropCat 15   Loyality
## 1 -0.22274756  0.0383162
## 2  0.06117496  0.1820726
## 3 -0.21990100 -1.2193953
```

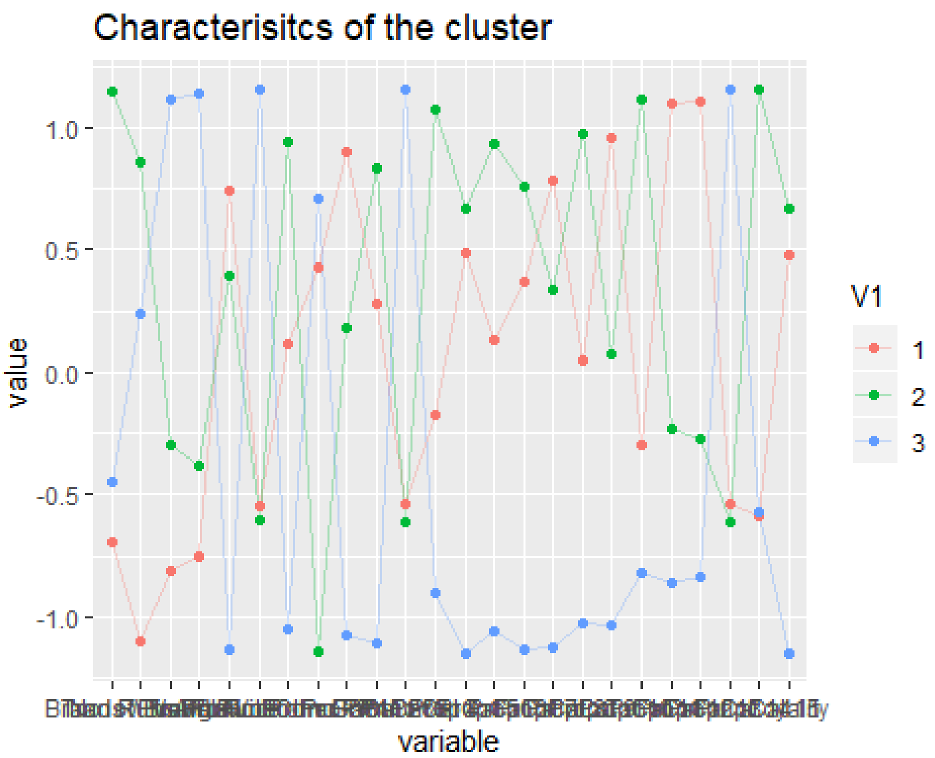
```
#2 : Visualizing the characteristics of the cluster
# From the chracteristics of the cluster above, it is infered that (ii) and (
iii) data forms the similar clusters. Hence visualising the characteristics o
f data with demographics and purchase basis since it has less varialbes compa
red to the (iii) case.
data4$clusters<-model2$cluster
# Formation of clusters, i.e size of the clusters.
ggplot(data4) +
  aes(x = SEC) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(clusters))
```




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

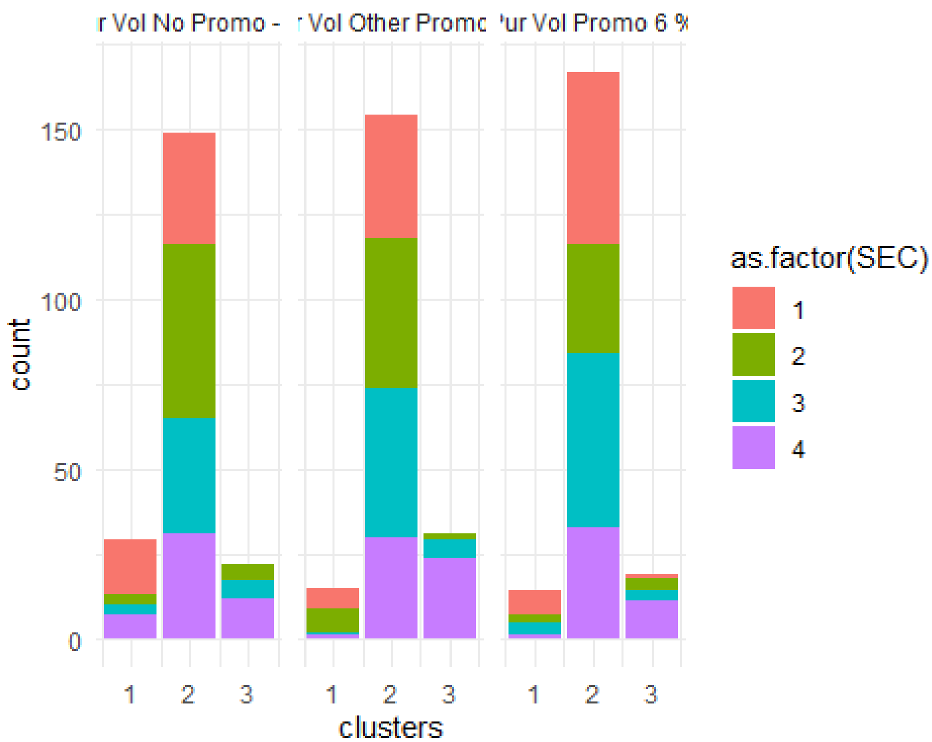



```r
ggparcoord(result2,
           columns = 10:ncol(result2), groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster",
           alphaLines = 0.3
)
```



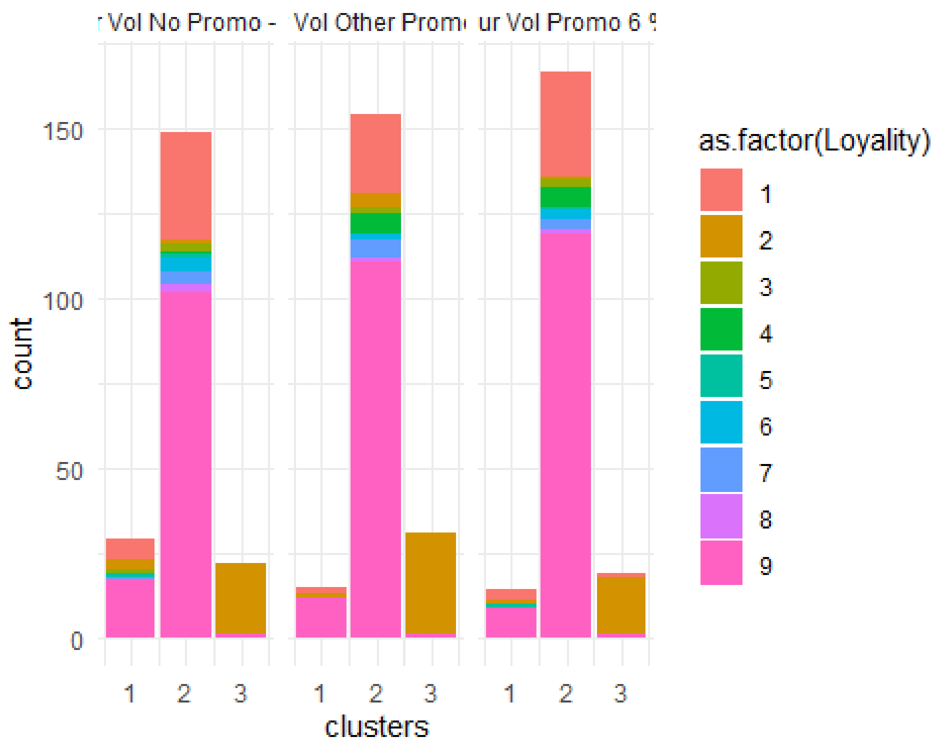

```
# Description of the clusters:
# Cluster 1 :
# There are high number of customers who belong to low economic level.
# The value of "Trans/Brands Runs","Vol/Tran" are low and high average price.
# The most of the customers in the cluster 1 not utilizes the promotions for purchase, some utilizes the promotion "other promo%" at high rate and few utilizes the promotion "Promo6%".
# The customers made least purchases from price catalog 1,2,4 and high purchases from catalog 3.
# The selling propoption is low except 14th catalog.

# Cluster 2 :
# The cluster consists of high economic status customers.
# There are low house hold members, availability of TV, Affluence Index, brand runs, number of transactions, volume of transactions.
# The average price of the purchases is low.
# The purchases fall in Price catalog 1 and selling propoption in catalog 8,10,12,13 & 15.

# Cluster 3 :
# There are more number of educated people, house hold members, tv availability, affluence index, number of brands and high frequency of purchase.
# The customers are less utilizing the promotion discount "Pur Vol No Promo - %","Pur Vol Other Promo %" and higly utilizing "Pur Vol Promo 6 %".
# The purchases fall under price catalogs 2,4&5. Less under catalog 3
# The selling proposition falls under catalog 5,6,7,9,11 and 15. Least in cat
```

```
alog 14

#Q 3:
data4$clusters<-model2$cluster

# Visualizing the output.
# For the targeted marketing, the below plots would help to infer the data re
lation between the clusters formed,promotions,loyality and socio economic sta
tus.
ggplot(data4) +
  aes(x = clusters,fill=as.factor(SEC)) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other
Promo %")))
```




```
# The above graph depicts that cluster 2 has mix of all stauts customers. Hen
ce the below graph tells about the target marketing.

ggplot(data4) +
  aes(x = clusters,fill=as.factor(Loyality)) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
```

```
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other Promo %")))
```




```
# The above graph depicts that  cluster3 is loyal to brand2 with avaling all types of promotions and cluster2 is loyal to brand9( which is 'other 999' i.e, not loyal to any brand). Hence marketing team would focus on cluster2 customers to increase thier loyality by pitching the promotion offers.
```