CAT 3 PROJECT:

ANALYSIS OF PERSONALITY AND SOCIAL BEHAVIOUR

Problem statement:

To classify general public of different age groups based on the analysis of 'Personality and Social Behaviour'.

Methods:

1. Elbow method of clustering.
2. Kmeans clustering.
3. Hierarchial clustering
4. PCA

Tools and libraries:

1. R studio.
2. library(dplyr)
3. library(factoextra)
4. library(ggplot2)
5. library(ggpubr)
6. library(cluster)

Code:

```
library(dplyr)

library(factoextra)

library(ggplot2)

library(ggpubr)

library(cluster)

dataset1=read.csv("C:\\Users\\Navika M S\\OneDrive\\Documents\\Semester 4\\PA lab\\Revision material\\Personality and Social Behavior Dataset For Analysis.csv")

print(dataset1)

summary(dataset1)

datasetm=read.csv("C:\\Users\\Navika M S\\OneDrive\\Documents\\Semester 4\\PA lab\\male.csv")

print(datasetm)

summary(datasetm)

datasetf=read.csv("C:\\Users\\Navika M S\\OneDrive\\Documents\\Semester 4\\PA lab\\female.csv")

print(datasetf)

summary(datasetf)
```

```
fviz_nbclust(dataset, kmeans, method="wss")
labs(subtitle="Elbow Method")
results11<-kmeans(dataset1,5)
results11
results11$size
results11$cluster
dataset1[,1:7] <-scale(dataset1[,1:7])
plot(dataset1[c("AGE","MIND")],col=results11$cluster)
points(results11$centers,pch=2,col="red")
plot(dataset1[c("AGE","ENERGY")],col=results11$cluster)
points(results11$centers,pch=2,col="red")
plot(dataset1[c("AGE","NATURE")],col=results11$cluster)
points(results11$centers,pch=2,col="red")
plot(dataset1[c("AGE","TACTICS")],col=results11$cluster)
points(results11$centers,pch=2,col="red")
plot(dataset1[c("AGE","IDENTITY")],col=results11$cluster)
points(results11$centers,pch=2,col="red")
clusplot(dataset1,results11$cluster)
sil <- silhouette(results11$cluster, dist(dataset1))
fviz_silhouette(sil)
#EUCLIDEAN
data.exc1<-dist(dataset1,method="euclidean")
round(as.matrix(data.exc1)[1:7,1:7],1)
# Use hcut() which compute hclust and cut the tree
hc.cut <- hcut(dataset1, k = 5, hc_method = "complete")
# Visualize dendrogram
fviz_dend(hc.cut, show_labels = TRUE, rect = TRUE)
res.pca <- prcomp(dataset1, scale = TRUE)
fviz_eig(res.pca)
```

```
fviz_pca_ind(res.pca,
             col.ind = "cos2", # Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
# Dimension reduction using PCA
res.pca <- prcomp(dataset1, scale = TRUE)
res.pca
# Coordinates of individuals
ind.coord <- as.data.frame(get_pca_ind(res.pca)$coord)
# Add clusters obtained using the K-means algorithm
ind.coord$cluster <- factor(results11$cluster)
# Add Species groups from the original data sett
ind.coord$Species <- df$Species
# Data inspection
head(ind.coord)
# Percentage of variance explained by dimensions
eigenvalue <- round(get_eigenvalue(res.pca), 1)
variance.percent <- eigenvalue$variance.percent
head(eigenvalue)
ggscatter(
  ind.coord, x = "Dim.1", y = "Dim.2",
  color = "cluster", palette = "npg", ellipse = TRUE, ellipse.type = "convex",
  shape = "circle", size = 1.5,  legend = "right", ggtheme = theme_bw(),
  xlab = paste0("Dim 1 (", variance.percent[1], "% )" ),
  ylab = paste0("Dim 2 (", variance.percent[2], "% )" )
) +
  stat_mean(aes(color = cluster), size = 4)
```

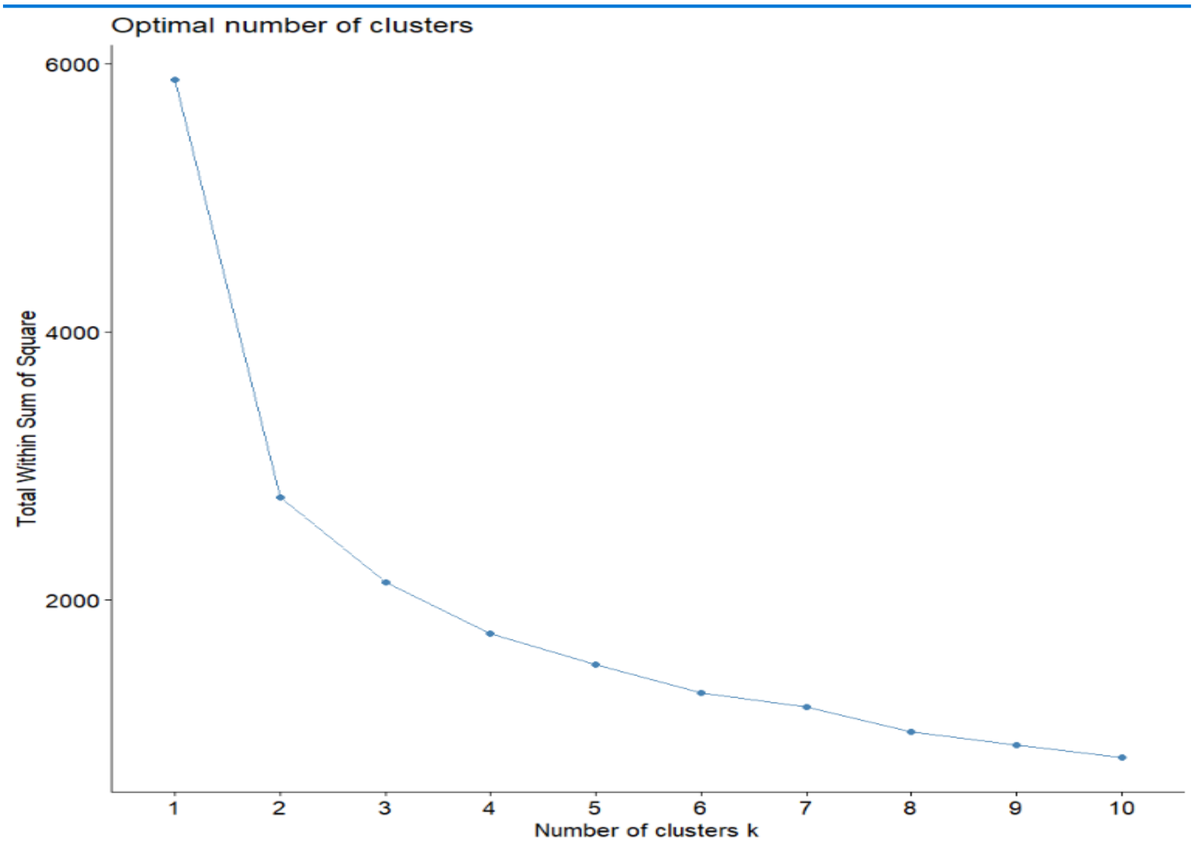
Output:

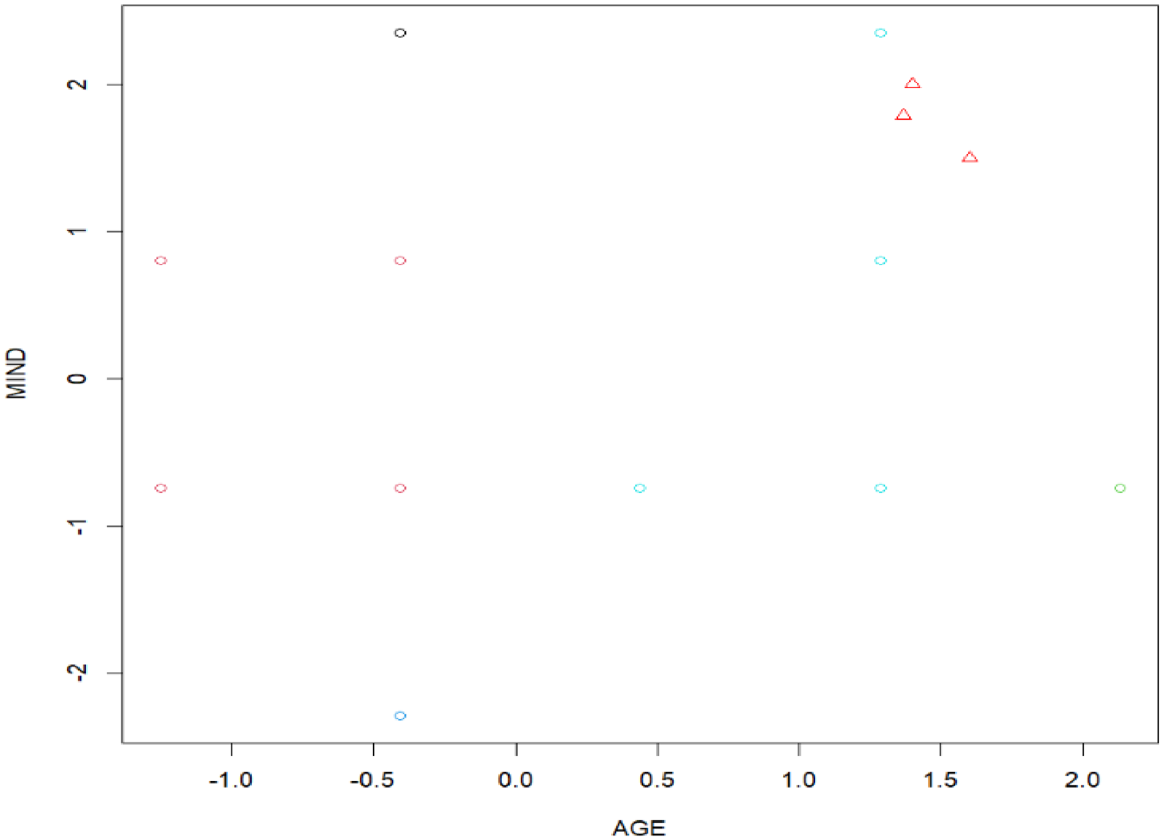
```
   GENDER AGE MIND ENERGY NATURE TACTICS IDENTITY
1       1   2    3      3      2       3        2
2       2   1    3      3      2       3        3
3       2   2    2      1      1       2        3
4       2   2    3      3      3       3        3
5       2   2    3      2      3       3        2
6       1   5    2      3      2       3        3
7       1   3    2      3      2       3        3
8       1   2    3      3      2       3        2
9       1   2    3      3      2       3        3
10      2   2    2      2      2       2        2
11      1   2    2      2      1       3        2
12      1   2    4      2      3       3        3
13      1   2    2      3      2       2        3
14      2   4    2      3      2       2        3
15      2   2    2      2      1       2        1
16      1   2    3      3      2       2        2
17      2   1    2      2      3       2        1
18      1   2    3      3      2       2        3
19      1   2    3      3      2       3        2
20      2   1    3      2      2       2        2
21      1   2    2      3      2       3        3
22      1   1    2      3      3       3        3
23      2   2    2      3      2       3        3
24      1   2    2      3      2       3        3
25      1   4    2      3      2       3        3
26      2   1    3      3      2       2        3
27      2   2    3      3      2       2        3
28      1   2    3      3      2       2        3
29      2   1    3      3      2       3        3
30      2   1    3      2      2       2        2
31      1   2    3      2      2       2        2
32      1   2    1      1      1       1        2
33      1   2    1      1      1       1        2
34      2   1    2      2      3       2        3
35      1   1    2      2      2       3        2
36      2   2    3      3      2       3        2
37      1   4    3      3      2       3        3
38      1   4    2      1      1       2        3
39      1   4    3      3      3       3        3
40      1   2    3      2      3       3        2
41      1   4    2      3      2       3        3
42      1   4    2      3      2       3        3
```

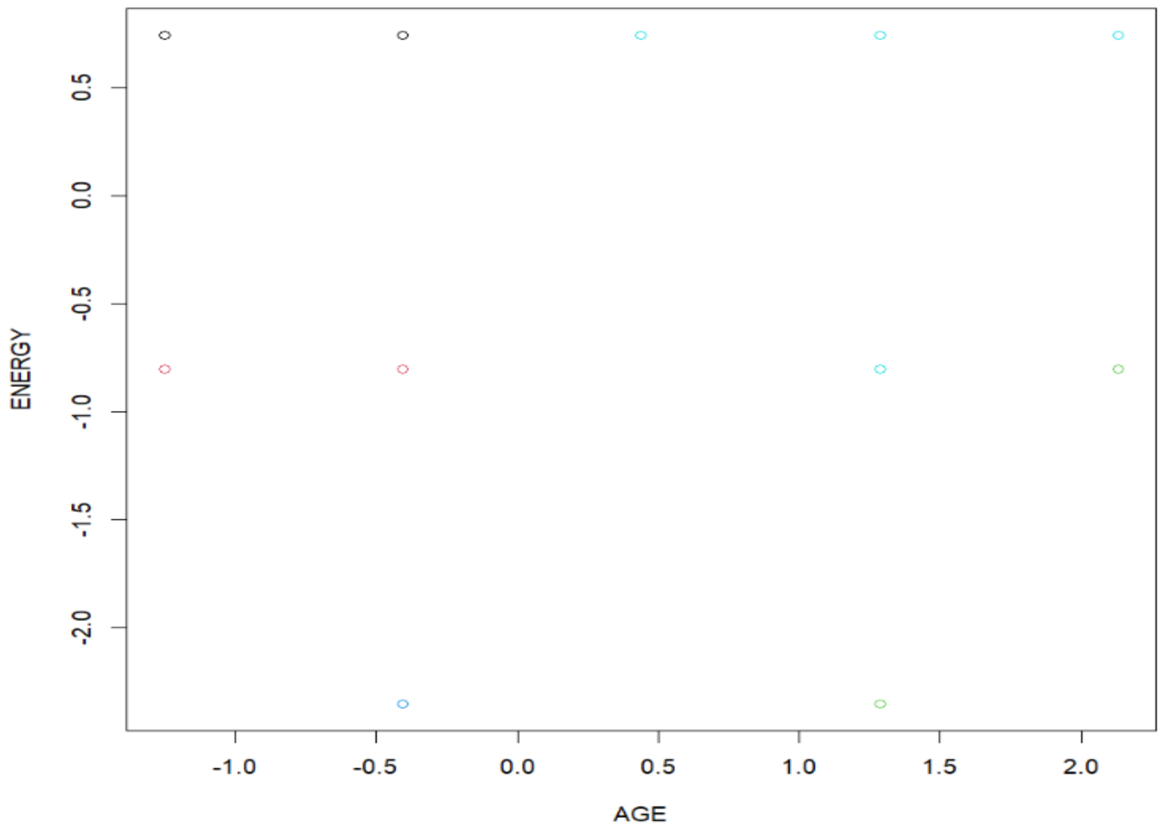
```
> summary(dataset1)
     GENDER          AGE            MIND          ENERGY          NATURE         TACTICS        IDENTITY   
 Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00   Min.   :1.00  
 1st Qu.:1.00   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:2.00  
 Median :1.00   Median :2.00   Median :2.00   Median :3.00   Median :2.00   Median :3.00   Median :3.00  
 Mean   :1.38   Mean   :2.48   Mean   :2.48   Mean   :2.52   Mean   :2.02   Mean   :2.52   Mean   :2.52  
 3rd Qu.:2.00   3rd Qu.:4.00   3rd Qu.:3.00   3rd Qu.:3.00   3rd Qu.:2.00   3rd Qu.:3.00   3rd Qu.:3.00  
 Max.   :2.00   Max.   :5.00   Max.   :4.00   Max.   :3.00   Max.   :3.00   Max.   :3.00   Max.   :3.00  
```

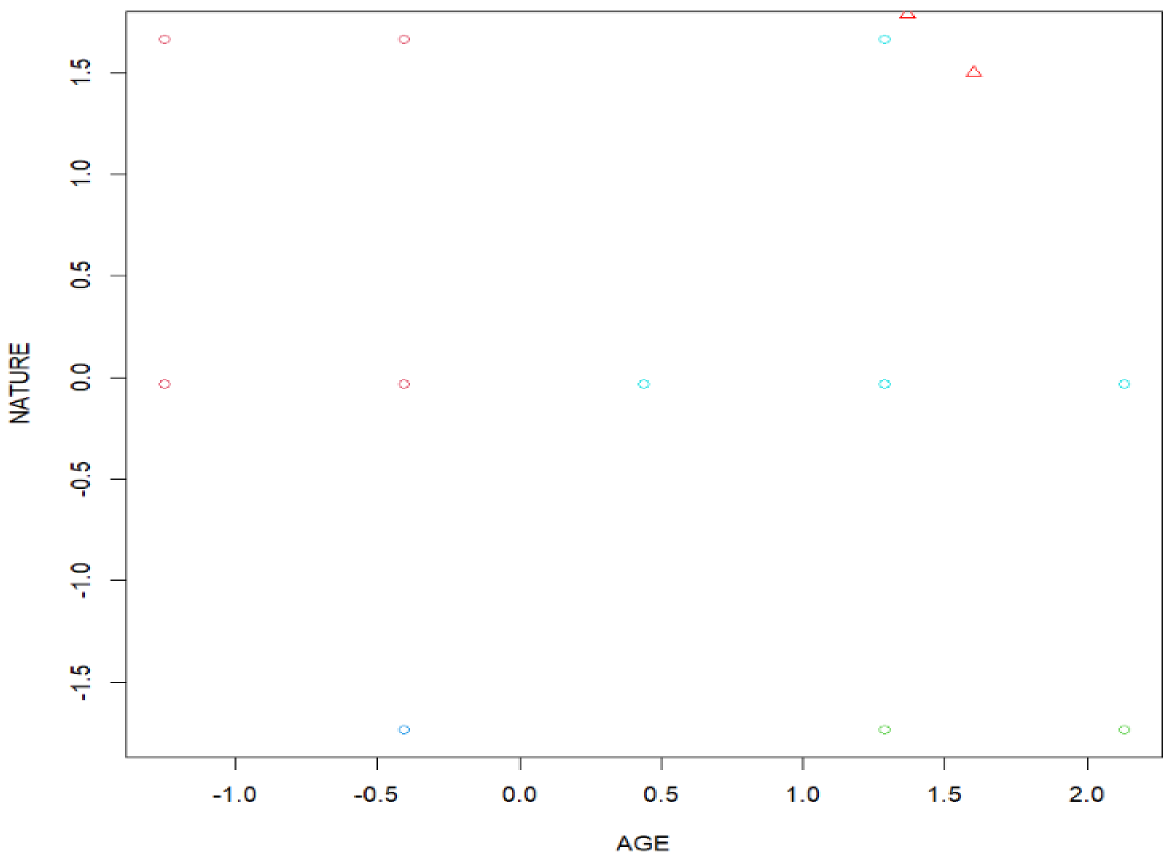
```
> fviz_nbclust(dataset, kmeans, method="wss")
> labs(subtitle="Elbow Method")
$subtitle
[1] "Elbow Method"

attr(,"class")
[1] "labels"
>
> results11<-kmeans(dataset1,5)
> results11
K-means clustering with 5 clusters of sizes 19, 10, 3, 5, 13

Cluster means:
    GENDER      AGE     MIND   ENERGY   NATURE  TACTICS IDENTITY
1 1.368421 1.789474 2.789474 2.947368 2.157895 2.684211 2.736842
2 1.600000 1.500000 2.500000 2.000000 2.400000 2.300000 2.000000
3 1.333333 4.333333 2.000000 1.666667 1.000000 2.333333 2.000000
4 1.400000 2.000000 1.600000 1.400000 1.000000 1.800000 2.000000
5 1.230769 4.000000 2.461538 2.923077 2.153846 2.769231 2.923077

Clustering vector:
 [1] 1 1 4 1 2 5 5 1 1 2 4 1 1 5 4 1 2 1 1 2 1 1 1 1 5 1 1 1 1 2 2 4 4 2 2 1 5 3 5 2 5 5 5 5 2 3 5 5 5 3

Within cluster sum of squares by cluster:
[1] 24.000000 13.900000  4.666667  8.400000 15.384615
 (between_SS / total_SS =  61.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
> results11$size
[1] 19 10  3  5 13
> results11$cluster
 [1] 1 1 4 1 2 5 5 1 1 2 4 1 1 5 4 1 2 1 1 2 1 1 1 1 5 1 1 1 1 2 2 4 4 2 2 1 5 3 5 2 5 5 5 5 2 3 5 5 5 3
>
> dataset1[,1:7] <-scale(dataset1[,1:7])
>
> plot(dataset1[c("AGE","MIND")],col=results11$cluster)
> points(results11$centers,pch=2,col="red")
>
> plot(dataset1[c("AGE","ENERGY")],col=results11$cluster)
> points(results11$centers,pch=2,col="red")
>
> plot(dataset1[c("AGE","NATURE")],col=results11$cluster)
> points(results11$centers,pch=2,col="red")
>
> plot(dataset1[c("AGE","TACTICS")],col=results11$cluster)
> points(results11$centers,pch=2,col="red")
>
> plot(dataset1[c("AGE","IDENTITY")],col=results11$cluster)
> points(results11$centers,pch=2,col="red")
>
> clusplot(dataset1,results11$cluster)
>
> sil <- silhouette(results11$cluster, dist(dataset1))
> fviz_silhouette(sil)
  cluster size ave.sil.width
1       1   19          0.16
2       2   10          0.17
3       3    3          0.05
4       4    5          0.09
5       5   13          0.25
>
> #EUCLIDEAN
> data.exc1<-dist(dataset1,method="euclidean")
> round(as.matrix(data.exc1)[1:7,1:7],1)
    1   2   3   4   5   6   7
1 0.0 2.7 5.0 3.1 3.1 3.4 2.4
2 2.7 0.0 4.3 1.9 2.9 4.2 3.1
3 5.0 4.3 0.0 5.1 4.7 5.1 4.5
4 3.1 1.9 5.1 0.0 2.2 4.0 3.2
5 3.1 2.9 4.7 2.2 0.0 4.6 3.9
6 3.4 4.2 5.1 4.0 4.6 0.0 1.7
7 2.4 3.1 4.5 3.2 3.9 1.7 0.0
```

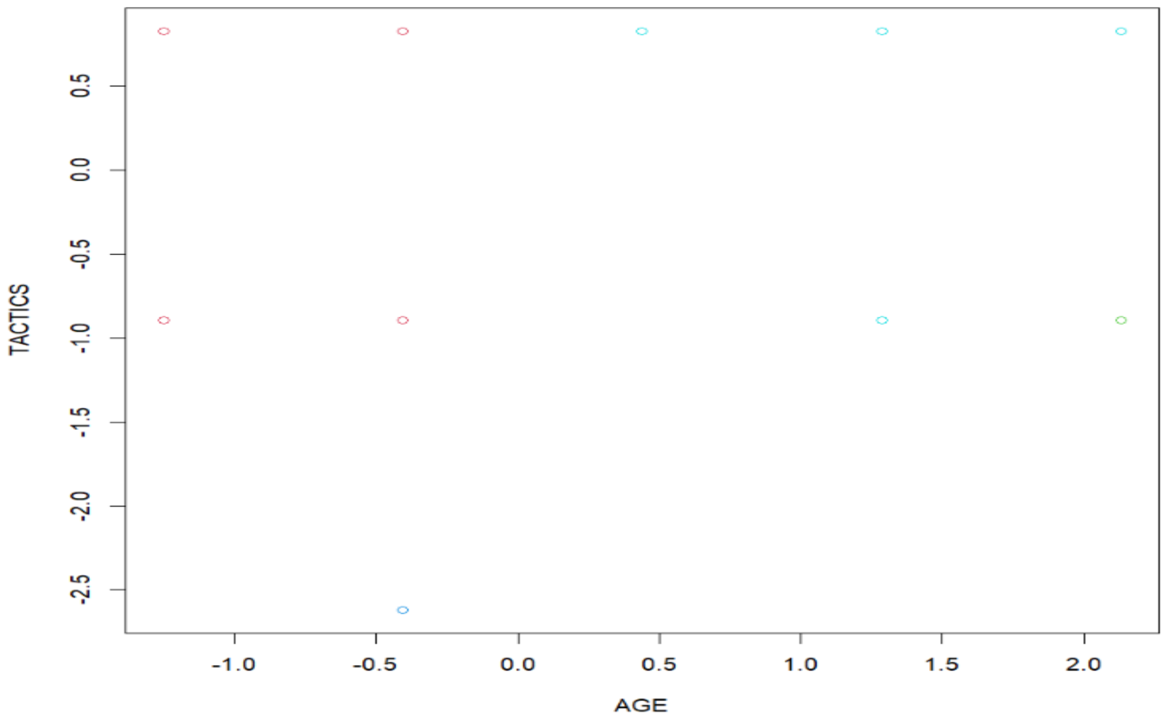
```
>
> # Use hcut() which compute hclust and cut the tree
> hc.cut <- hcut(dataset1, k = 5, hc_method = "complete")
> # Visualize dendrogram
> fviz_dend(hc.cut, show_labels = TRUE, rect = TRUE)
>
> res.pca <- prcomp(dataset1, scale = TRUE)
> fviz_eig(res.pca)
>
> fviz_pca_ind(res.pca,
+              col.ind = "cos2", # Color by the quality of representation
+              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
+              repel = TRUE     # Avoid text overlapping
+ )
> # Dimension reduction using PCA
> res.pca <- prcomp(dataset1, scale = TRUE)
> res.pca
Standard deviations (1, .., p=7):
[1] 1.5682067 1.2390121 0.8990680 0.8837855 0.7506537 0.6708209 0.6345826

Rotation (n x k) = (7 x 7):
                 PC1         PC2        PC3        PC4         PC5        PC6         PC7
GENDER    0.07723116 -0.53503227  0.6798017 -0.4357959  0.05765008  0.1463238 -0.17604903
AGE      -0.07109846  0.60733950  0.5411088  0.1614479  0.48867591  0.2138506  0.15060994
MIND     -0.42026964 -0.35851515  0.0765270  0.3055440  0.44899894 -0.6234309  0.07321646
ENERGY   -0.49033796  0.12342216  0.2164341 -0.2543599 -0.50662508 -0.1330876  0.59867498
NATURE   -0.44553406 -0.32181719 -0.2968844  0.0360027  0.28128147  0.6946220  0.21651221
TACTICS  -0.48913138  0.07826014  0.2038592  0.3845004 -0.39088833  0.1431537 -0.62605118
IDENTITY -0.36629546  0.30236089 -0.2503436 -0.6905242  0.26017368 -0.1532518 -0.37912186
> # Coordinates of individuals
> ind.coord <- as.data.frame(get_pca_ind(res.pca)$coord)
> # Add clusters obtained using the K-means algorithm
> ind.coord$cluster <- factor(results11$cluster)
> # Add Species groups from the original data sett
> ind.coord$Species <- df$Species
> # Data inspection
> head(ind.coord)
       Dim.1      Dim.2      Dim.3       Dim.4      Dim.5      Dim.6       Dim.7 cluster
1 -0.8126299 -0.2089837 -0.1335216  1.23081540 -0.8114864 -0.5758419  0.37416620       1
2 -1.1914308 -1.3215828  0.3875720 -0.91897494 -0.6836237 -0.7078555 -0.72962939       1
3  2.5156771 -0.2233651  0.2101749 -1.19229918  0.7987185 -0.5777534 -1.85581985       4
4 -2.0084231 -1.3545243  0.3409609 -0.72124750  0.2075711  0.6530279 -0.23442940       1
5 -0.6535140 -2.0377818  0.4138279  0.79660870  0.5675637  1.1084344 -0.54311543       2
6 -0.9394401  2.3791499  0.7136231  0.04349154  0.1577616  0.6816165  0.02577351       5
>
> # Percentage of variance explained by dimensions
> eigenvalue <- round(get_eigenvalue(res.pca), 1)
> variance.percent <- eigenvalue$variance.percent
> head(eigenvalue)
      eigenvalue variance.percent cumulative.variance.percent
Dim.1        2.5             35.1                        35.1
Dim.2        1.5             21.9                        57.1
Dim.3        0.8             11.5                        68.6
Dim.4        0.8             11.2                        79.8
Dim.5        0.6              8.0                        87.8
Dim.6        0.5              6.4                        94.2
>
> ggscatter(
+   ind.coord, x = "Dim.1", y = "Dim.2",
+   color = "cluster", palette = "npg", ellipse = TRUE, ellipse.type = "convex",
+   shape = "circle", size = 1.5,  legend = "right", ggtheme = theme_bw(),
+   xlab = paste0("Dim 1 (", variance.percent[1], "% )" ),
+   ylab = paste0("Dim 2 (", variance.percent[2], "% )" )
+ ) +
+   stat_mean(aes(color = cluster), size = 4)
>
```


Optimal number of clusters
6000
4000
2000
Total Within Sum of Square
1
2
3
4
5
6
7
8
9
10
Number of clusters k

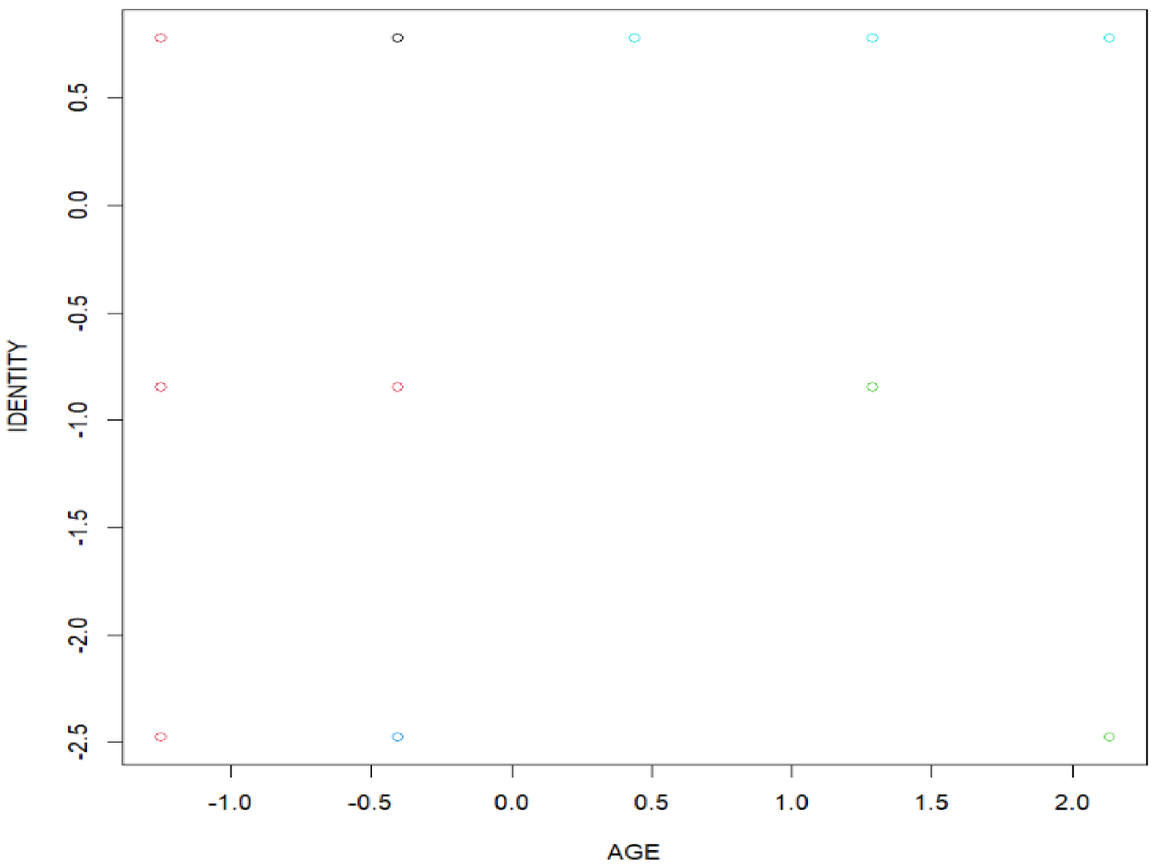

2
1
0
-1
-2
MIND
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
AGE


ENERGY
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
AGE

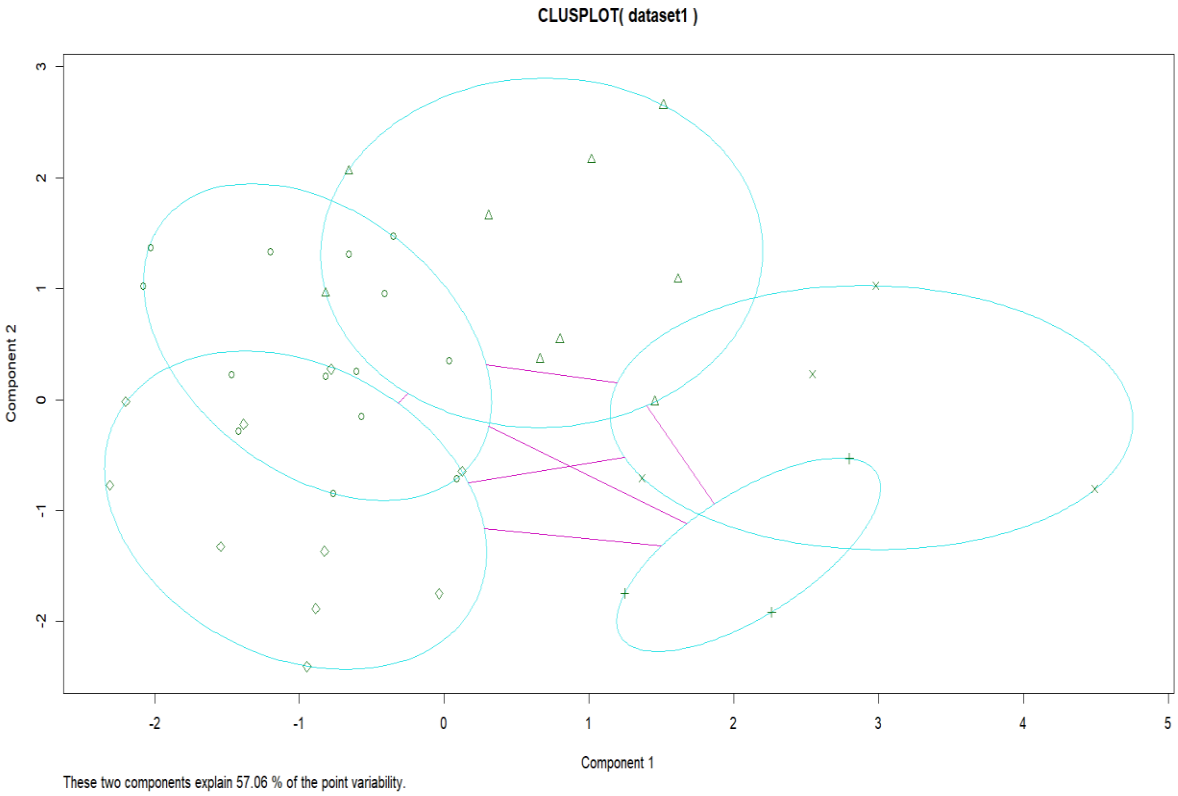

NATURE
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
AGE

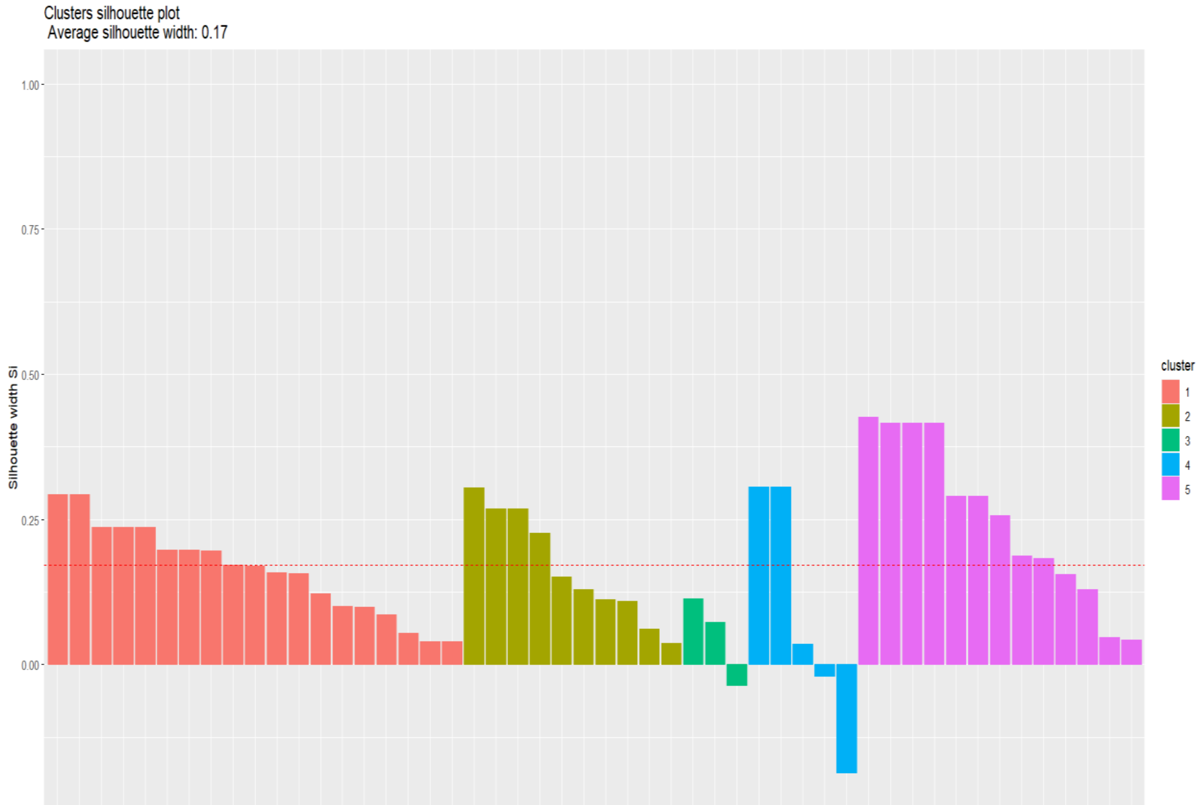


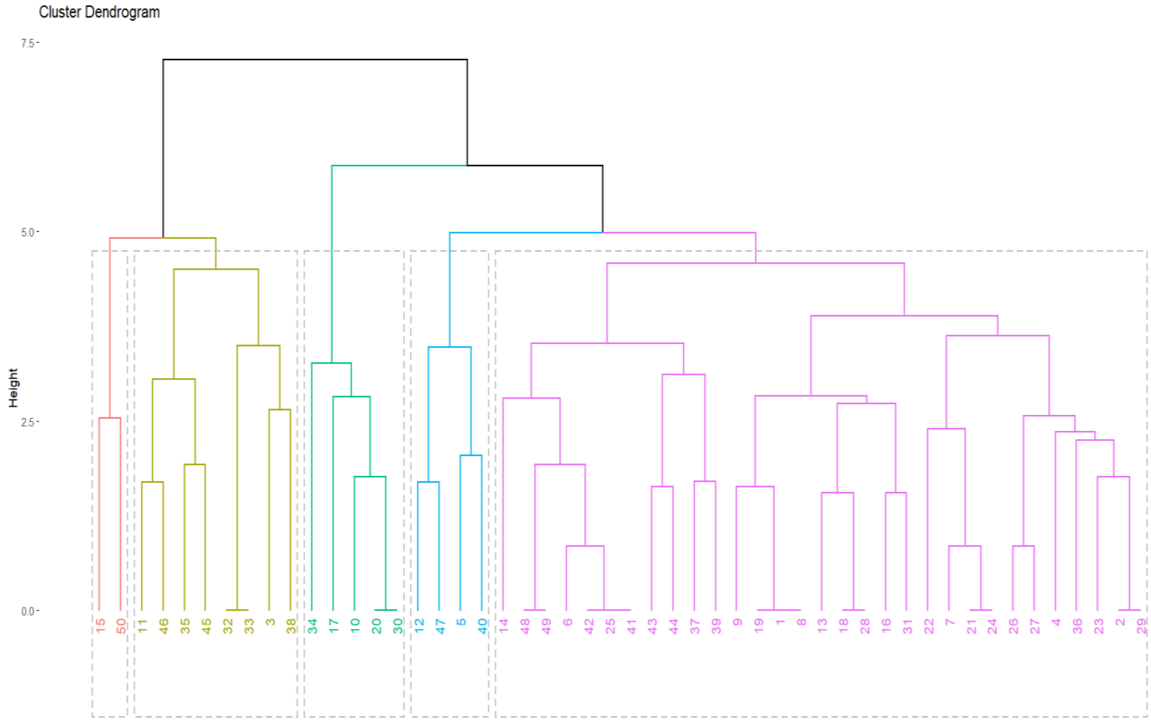
TACTICS
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
AGE

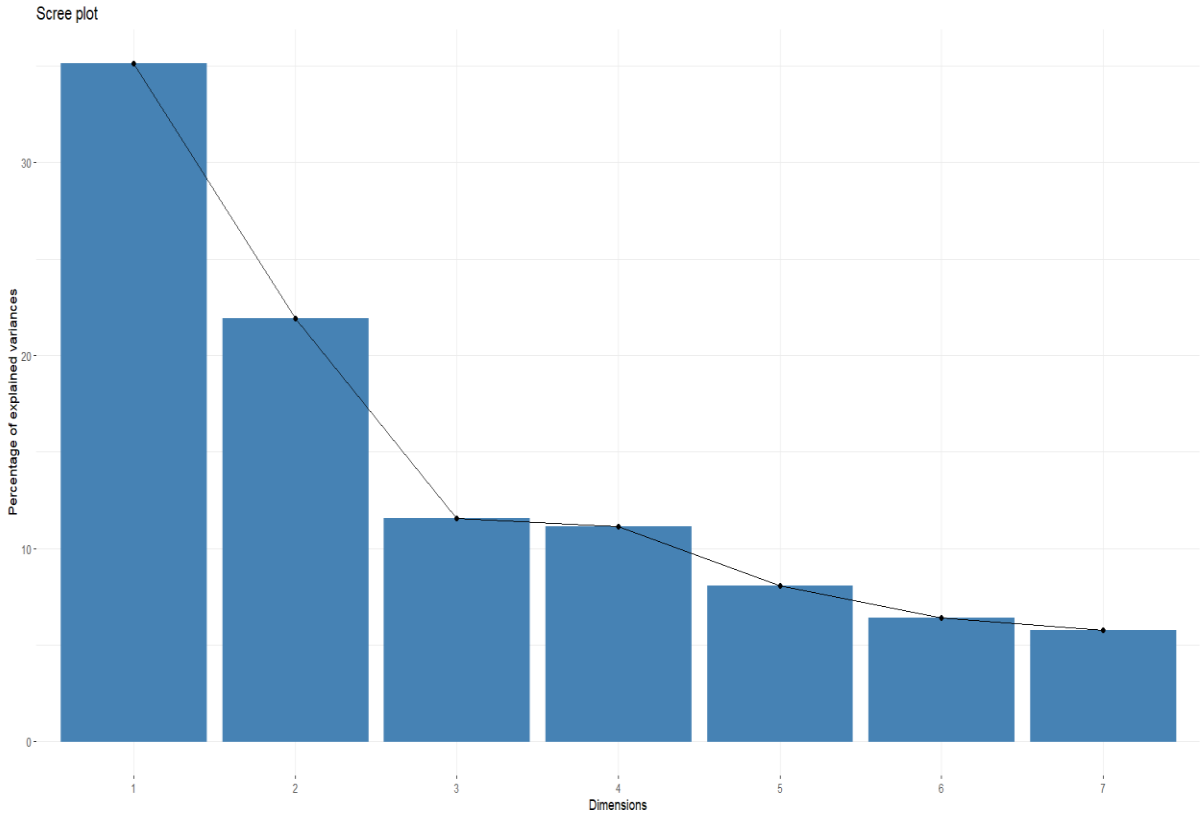


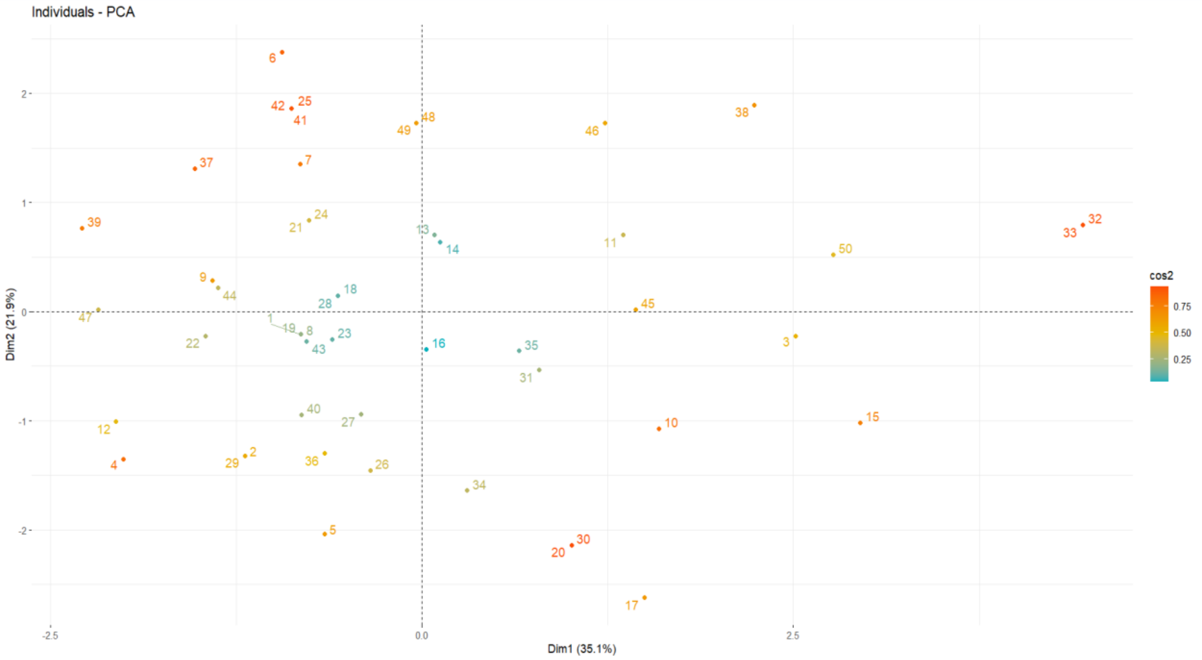
IDENTITY
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
AGE

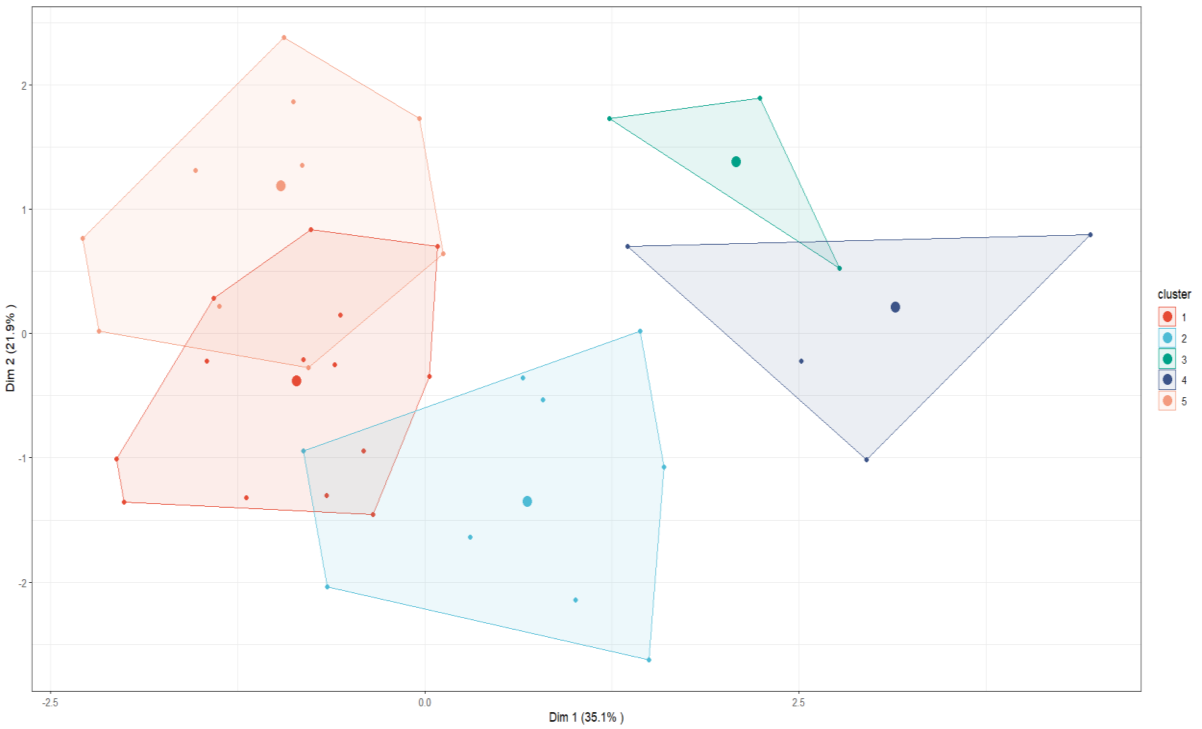

CLUSPLOT(dataset1)
Component 2
Component 1
These two components explain 57.06 % of the point variability.


Clusters silhouette plot
Average silhouette width: 0.17
Silhouette width Si
cluster
1
2
3
4
5


Cluster Dendrogram
Height
7.5
5.0
2.5
0.0
15 50 11 46 35 45 32 33 3 38 34 17 10 20 30 12 47 5 40 14 48 49 6 42 25 41 43 44 37 39 9 19 1 8 13 18 28 16 31 22 7 21 24 26 27 4 36 23 2 29


Scree plot
Percentage of explained variances
30
20
10
0
1 2 3 4 5 6 7
Dimensions





Interpretation:

- From the plot of the elbow method, we could observe that the line is bent at the range 5 making it the optimal number of clusters for the dataset.
- By using k means clustering the clusters have been formed.
- By using PCA we found that the variables whose value is greater than 0.5 has a great influence on once personality and social behaviour.
- The within-cluster sum of squares is a measure of the variability of the observations within each cluster. In general, a cluster that has a small sum of squares is more

compact than a cluster that has a large sum of squares. Clusters that have higher values exhibit greater variability of the observations within the cluster

- In Silhouhette method
    - - Observations with a large silhouhette Si (almost 1) are very well clustered.
    - - A small Si (around 0) means that the observation lies between two clusters.
    - - Observations with a negative Si are probably placed in the wrong cluster

- For the age category 13-18 it observed that people under this category are **extroverts and intuitive**. They can cope with emotions and are well planned. They are assertive.

- The people under age category 19-24 **have similar personality and behavioural pattern** of people below 19.

- People of age between 25-30 are **extroverts and strongly intuitive** people. Their capacity to cope with emotions are high. They are very well planned and assertive people.

- 31-59 highly **extroverted and highly intuitive**, can easily manage their emotions and are planned.

- Within sum of squares **less**, the cluster is **more compact**.

- Within sum of squares of MIND is the greatest, hence making it the most dissimilar cluster.