MVA Final Project

Javier Ferrando Monsonis
Marcel Porta Valles
Mehmet Fatih ??agil

February 20, 2018

Libraries

```
library(chemometrics)
library(DMwR)
library(mice)
library(missForest)
library(ggplot2)
library(graphics)
library(gridExtra)
library(Hmisc)
library(knitr)
library(FactoMineR)
library(DataExplorer)
```

```
## Warning: package 'DataExplorer' was built under R version 3.5.2
```

```
library(factoextra)
library(expm)
library(fpc)
library(cluster)

theme_set(theme_bw())
setwd("/Users/JaviFerrando/Desktop/MVA-Project")
```

```
heart_disease = read.csv("data/heart.csv")

# Find missing variables
which(is.na(heart_disease))
```

```
## integer(0)
```

```
head(heart_disease)
```

```
##   X...age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca
## 1      63   1  3      145  233   1       0     150     0     2.3     0  0
## 2      37   1  2      130  250   0       1     187     0     3.5     0  0
## 3      41   0  1      130  204   0       0     172     0     1.4     2  0
## 4      56   1  1      120  236   0       1     178     0     0.8     2  0
## 5      57   0  0      120  354   0       1     163     1     0.6     2  0
## 6      57   1  0      140  192   0       1     148     0     0.4     1  0
##   thal target
## 1    1      1
## 2    2      1
## 3    2      1
## 4    2      1
## 5    2      1
```

```
## 6    1      1
```

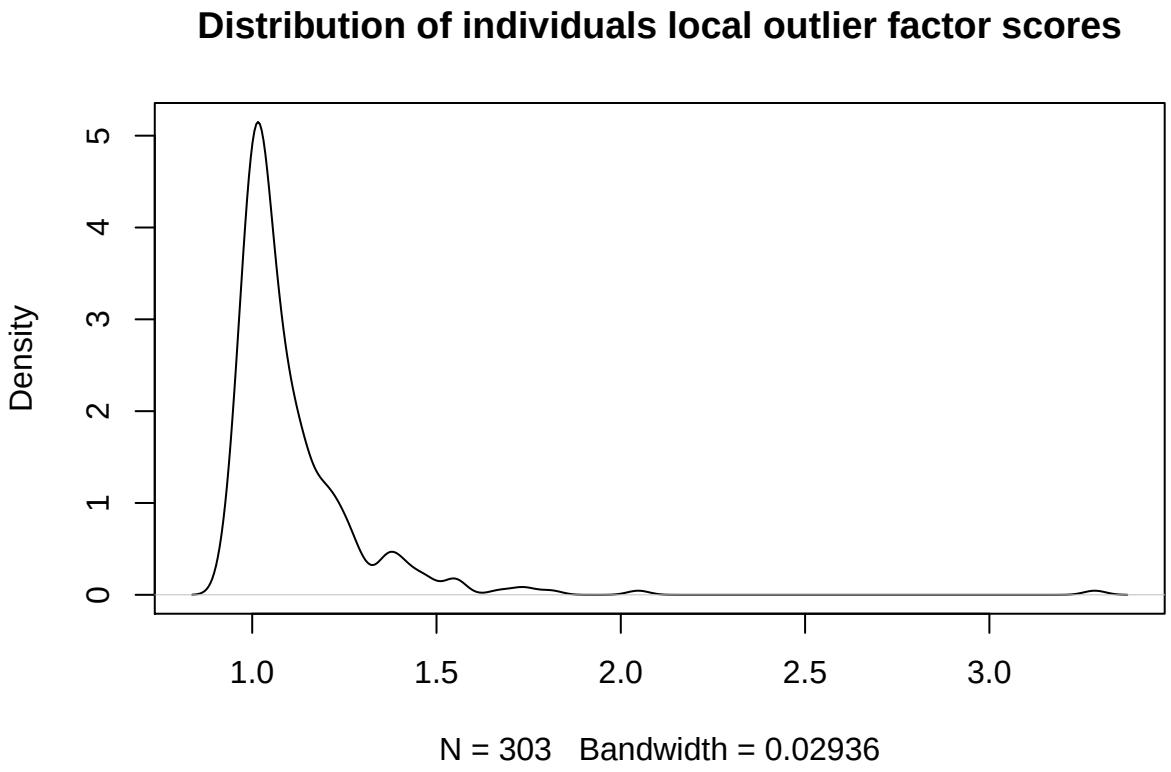
```
describe(heart_disease)
```

```
## heart_disease
##
##  14  Variables      303  Observations
## ---------------------------------------------------------------------------
## X...age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      303        0       41    0.999    54.37    10.36     39.1     42.0
##      .25      .50      .75      .90      .95
##     47.5     55.0     61.0     66.0     68.0
##
## lowest : 29 34 35 37 38, highest: 70 71 74 76 77
## ---------------------------------------------------------------------------
## sex
##        n  missing distinct     Info      Sum     Mean      Gmd
##      303        0        2    0.649      207   0.6832   0.4343
##
## ---------------------------------------------------------------------------
## cp
##        n  missing distinct     Info     Mean      Gmd
##      303        0        4    0.866    0.967    1.105
##
## Value          0     1     2     3
## Frequency    143    50    87    23
## Proportion 0.472 0.165 0.287 0.076
## ---------------------------------------------------------------------------
## trestbps
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      303        0       49    0.995    131.6    19.32      108      110
##      .25      .50      .75      .90      .95
##      120      130      140      152      160
##
## lowest :  94 100 101 102 104, highest: 174 178 180 192 200
## ---------------------------------------------------------------------------
## chol
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      303        0      152        1    246.3    55.95    175.0    188.0
##      .25      .50      .75      .90      .95
##    211.0    240.0    274.5    308.8    326.9
##
## lowest : 126 131 141 149 157, highest: 394 407 409 417 564
## ---------------------------------------------------------------------------
## fbs
##        n  missing distinct     Info      Sum     Mean      Gmd
##      303        0        2    0.379       45   0.1485   0.2538
##
## ---------------------------------------------------------------------------
## restecg
##        n  missing distinct     Info     Mean      Gmd
##      303        0        3     0.76   0.5281   0.5274
##
## Value          0     1     2
```

```
## Frequency    147   152     4
## Proportion 0.485 0.502 0.013
## ---------------------------------------------------------------------------
## thalach
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      303        0       91        1    149.6    25.77    108.1    116.0
##      .25      .50      .75      .90      .95
##    133.5    153.0    166.0    176.6    181.9
##
## lowest :  71  88  90  95  96, highest: 190 192 194 195 202
## ---------------------------------------------------------------------------
## exang
##        n  missing distinct     Info      Sum     Mean      Gmd
##      303        0        2     0.66       99   0.3267   0.4414
##
## ---------------------------------------------------------------------------
## oldpeak
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      303        0       40    0.964     1.04    1.225      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.8      1.6      2.8      3.4
##
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 4.0 4.2 4.4 5.6 6.2
## ---------------------------------------------------------------------------
## slope
##        n  missing distinct     Info     Mean      Gmd
##      303        0        3    0.798    1.399   0.6291
##
## Value          0     1     2
## Frequency     21   140   142
## Proportion 0.069 0.462 0.469
## ---------------------------------------------------------------------------
## ca
##        n  missing distinct     Info     Mean      Gmd
##      303        0        5    0.795   0.7294    1.005
##
## Value          0     1     2     3     4
## Frequency    175    65    38    20     5
## Proportion 0.578 0.215 0.125 0.066 0.017
## ---------------------------------------------------------------------------
## thal
##        n  missing distinct     Info     Mean      Gmd
##      303        0        4    0.778    2.314   0.6125
##
## Value          0     1     2     3
## Frequency      2    18   166   117
## Proportion 0.007 0.059 0.548 0.386
## ---------------------------------------------------------------------------
## target
##        n  missing distinct     Info      Sum     Mean      Gmd
##      303        0        2    0.744      165   0.5446   0.4977
##
## ---------------------------------------------------------------------------
```

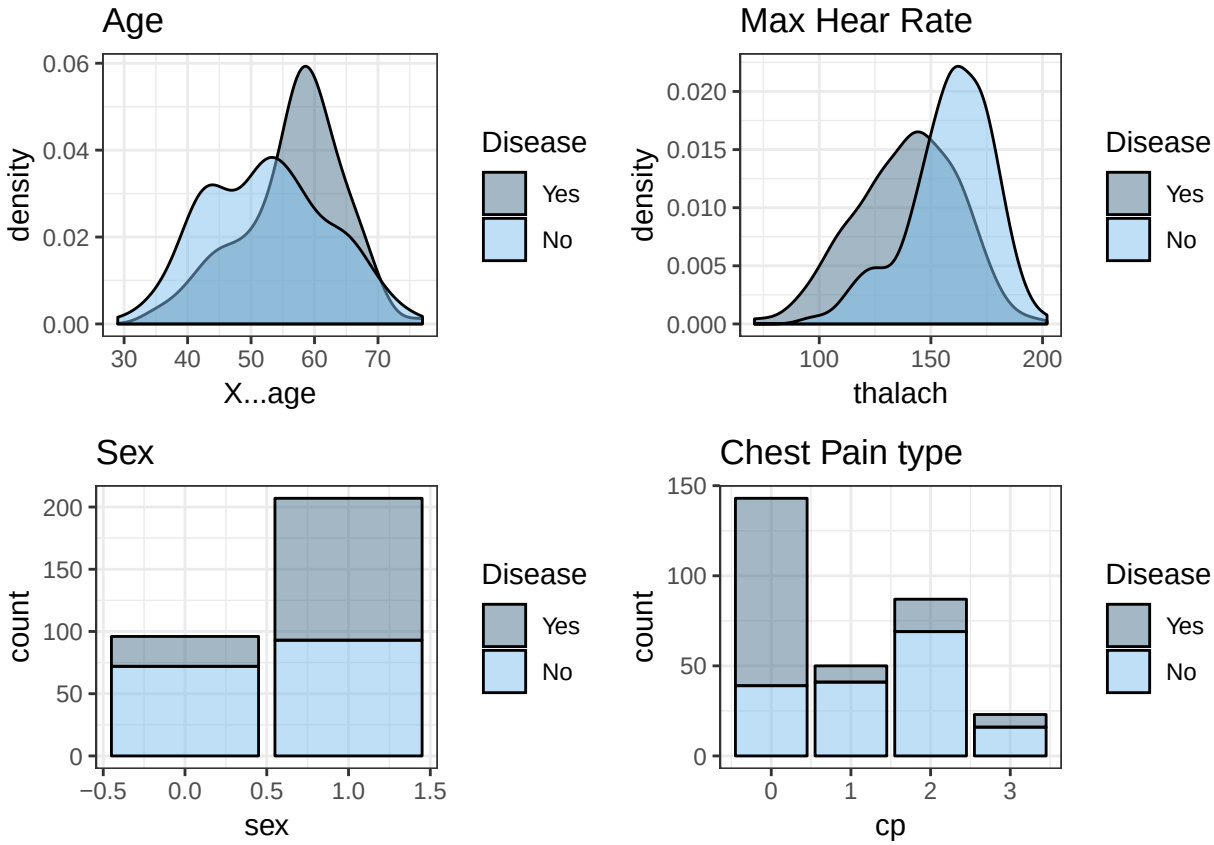
```r
classVar <- lapply(heart_disease,class)   # class of each variable
factor_heart <- heart_disease
factor_heart$target <- as.factor(heart_disease$target)
factor_heart$sex <- as.factor(heart_disease$sex)
factor_heart$fbs <- as.factor(heart_disease$fbs)
factor_heart$exang <- as.factor(heart_disease$exang)
factor_heart$restecg <- as.factor(heart_disease$restecg)
factor_heart$thal <- as.factor(heart_disease$thal)
factor_heart$slope <- as.factor(heart_disease$slope)
factor_heart$cp <- as.factor(heart_disease$cp)
factor_heart$ca <- as.factor(heart_disease$ca)
```

```r
#Outlier detection
############################
#Moutlier(heart_disease[,-14], quantile = 0.975, plot = TRUE, tol=1e-36) #Doesn't work


#Local Outlier Factor
outlier.scores <- lofactor(heart_disease[,-14], k=5)
plot(density(outlier.scores),main='Distribution of individuals local outlier factor scores')
```

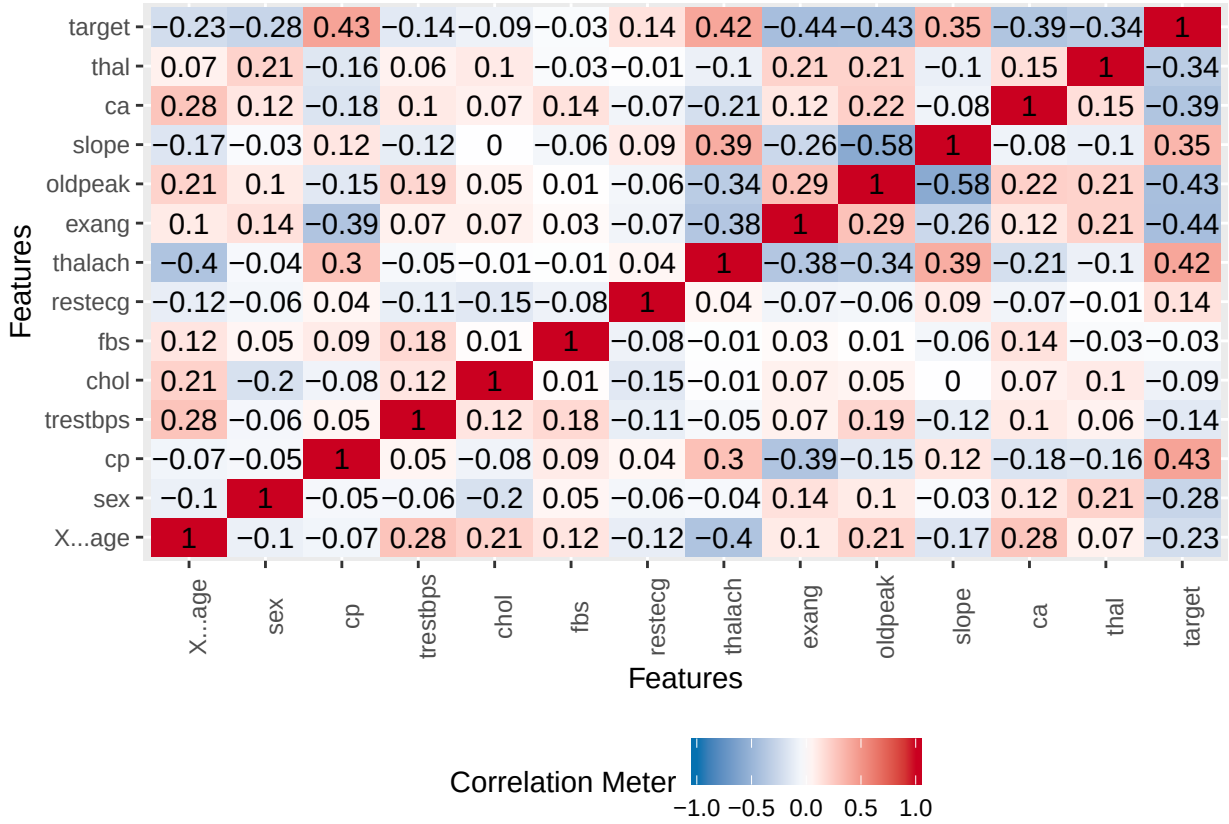



```r
#Exploratory Data Analysis
#Density of heart presence/absence disease by age
g1 <- ggplot(data=heart_disease, aes(x=X...age, fill=as.factor(target)))+
  geom_density(alpha=.5)+
  ggtitle("Age") +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("Yes", "No"))

#Density of heart presence/absence disease by Max heart rate
g2 <- ggplot(data=heart_disease, aes(x=thalach, fill=as.factor(target)))+
```

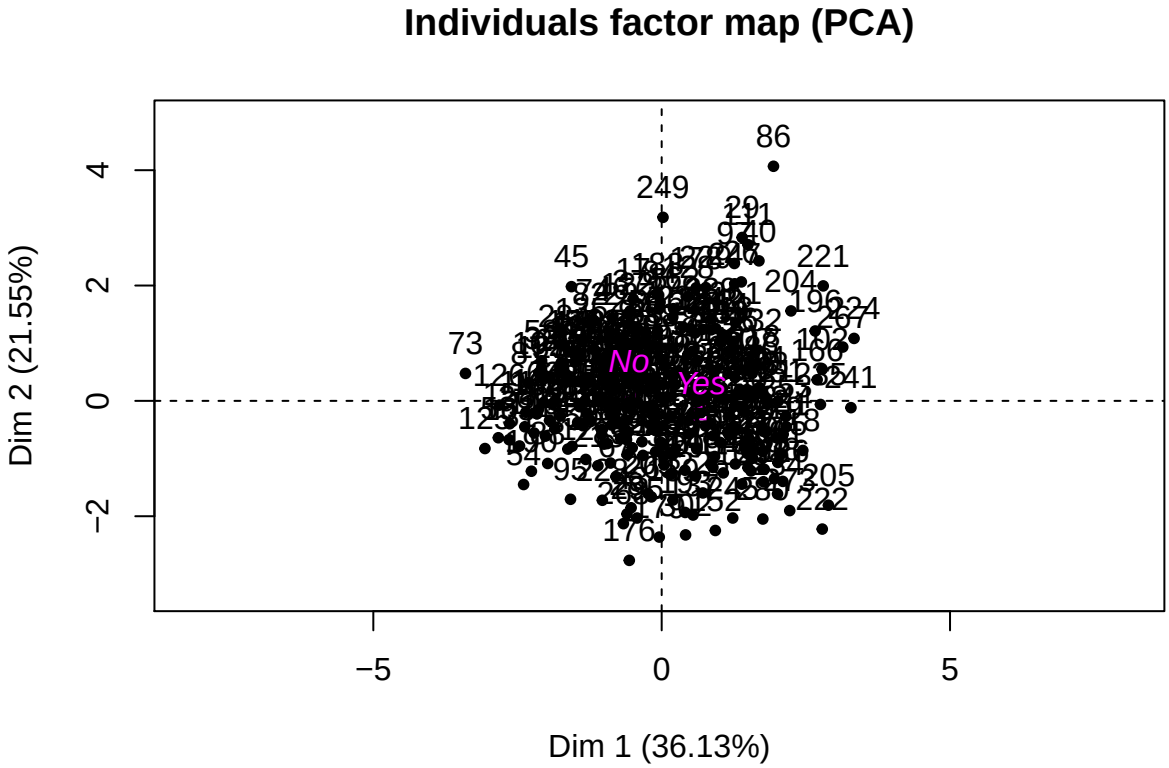
```
  geom_density(alpha=.5)+
  ggtitle("Max Hear Rate") +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("Yes", "No"))

#Density of heart presence/absence disease by sex
g3 <- ggplot(data=heart_disease, aes(x=sex, fill=as.factor(target)))+
      geom_bar(alpha=.5, color="black")+
      ggtitle("Sex") +
      scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("Yes", "No"))

#Density of heart presence/absence disease by chest type
g4 <- ggplot(data=heart_disease, aes(x=cp, fill=as.factor(target)))+
  geom_bar(alpha=.5, color="black")+
  ggtitle("Chest Pain type") +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("Yes", "No"))

grid.arrange(g1, g2, g3, g4, ncol = 2)
```

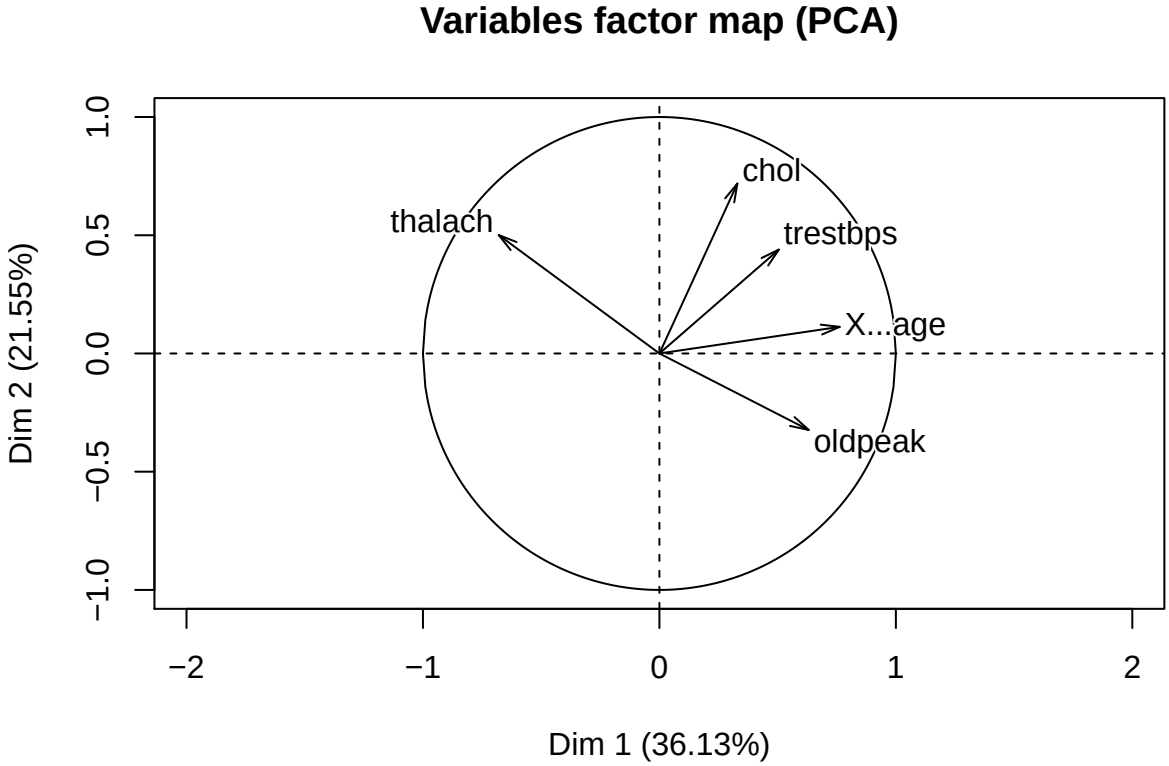



```
plot_correlation(heart_disease)
```

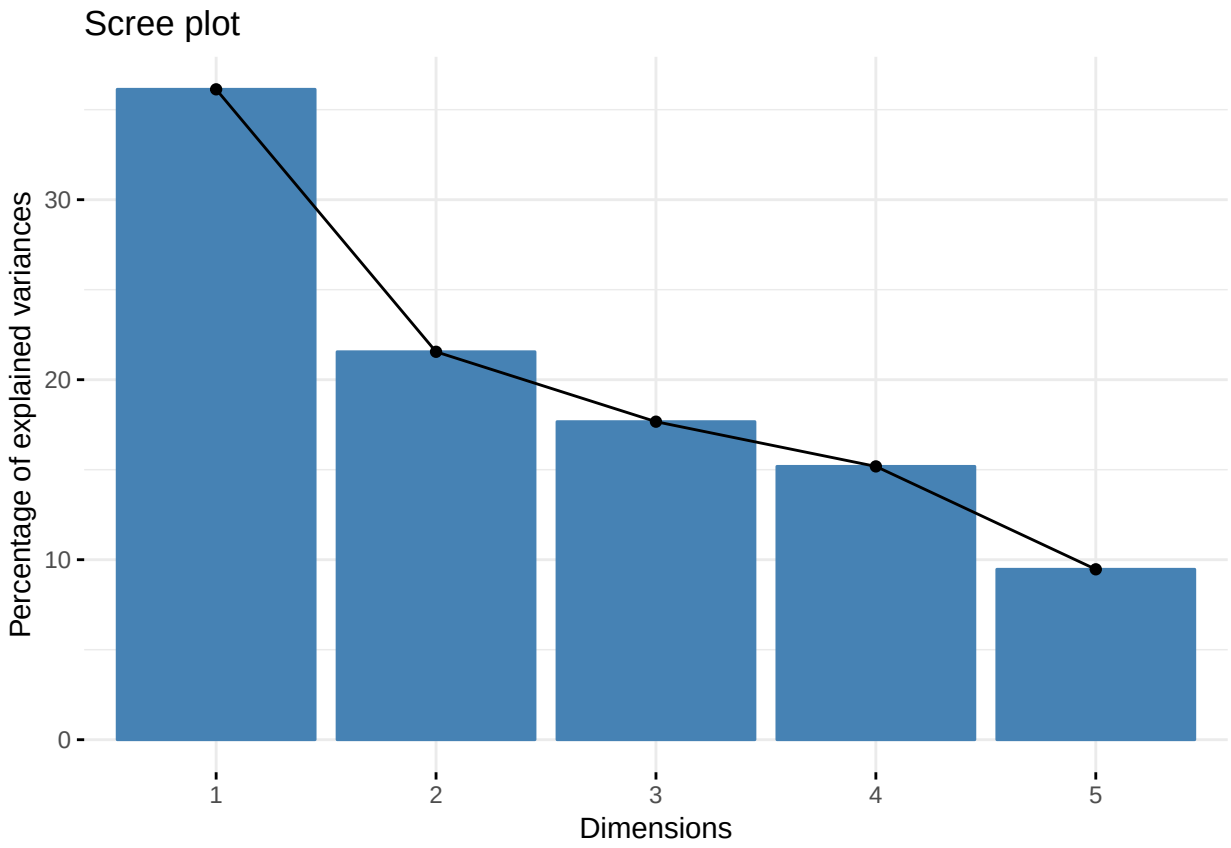



```
#PCA with categoriacal values
pca_facto <- factor_heart[, sapply(factor_heart, class) != "factor"]
#Some categorical values can be added as supplementary
#pca_facto$sex <- factor_heart$sex
#pca_facto$ca <- factor_heart$ca
pca_facto$disease <- heart_disease$target
pca_facto$disease[pca_facto$disease==0] <- "Yes"
pca_facto$disease[pca_facto$disease==1] <- "No"

pca_facto_heart <- PCA(pca_facto, quali.sup = 6, scale.unit = TRUE,  graph = TRUE)
```

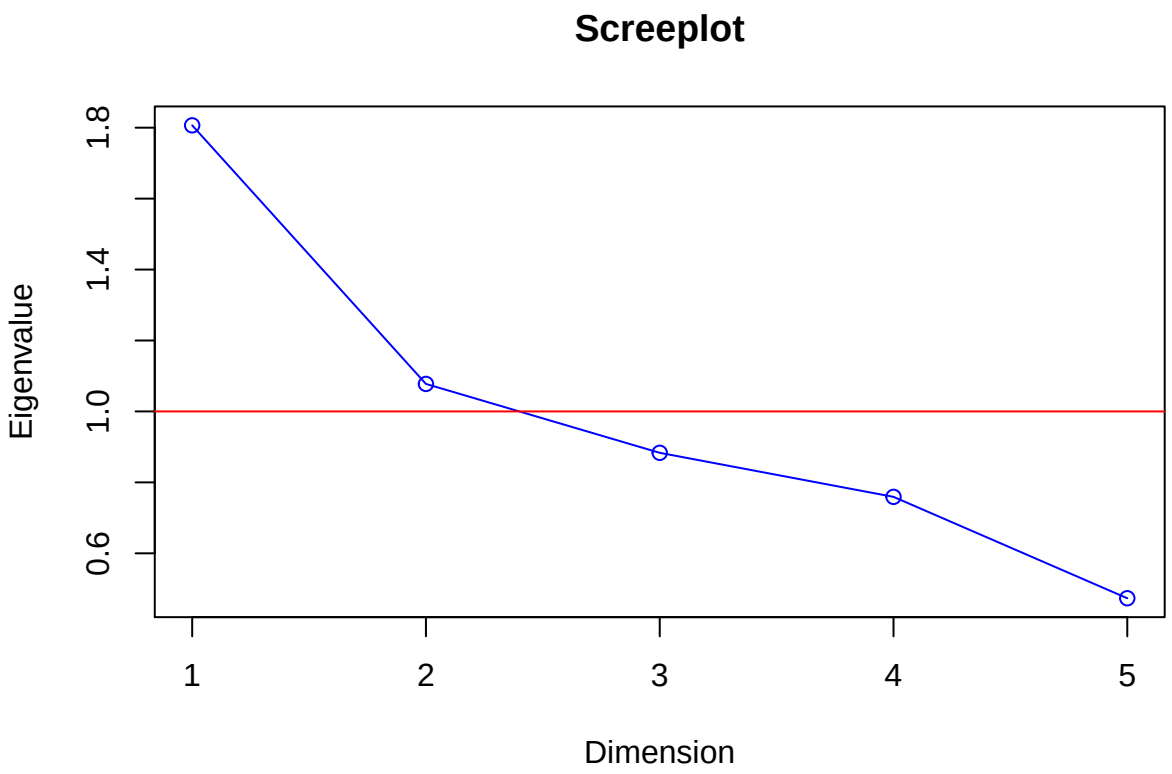




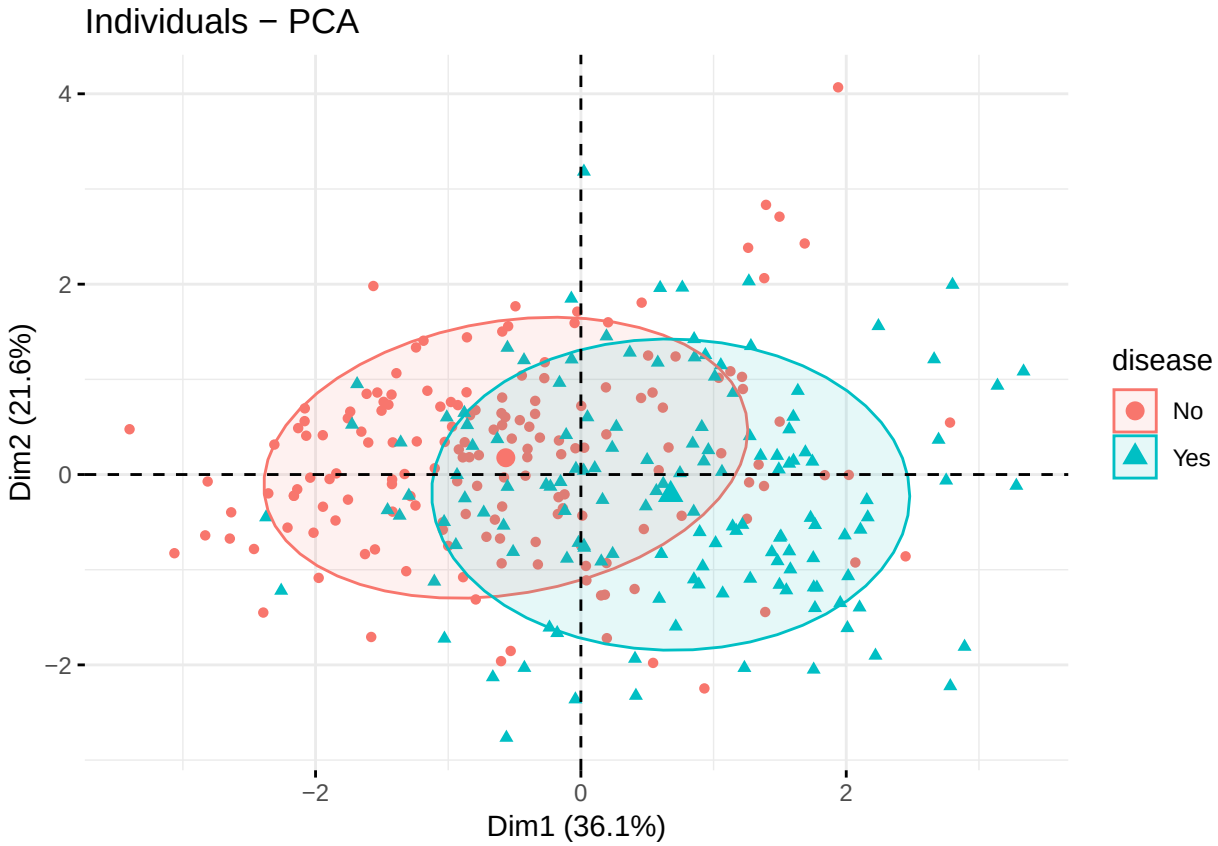

```
#Screeplots
fviz_screeplot(pca_facto_heart, addlabels = FALSE)
```




```
eigen_values <- pca_facto_heart$eig[,1]
plot(eigen_values, type="o", main="Screeplot",
     xlab='Dimension', ylab='Eigenvalue', col='blue')
abline(h=1,col="red")
```



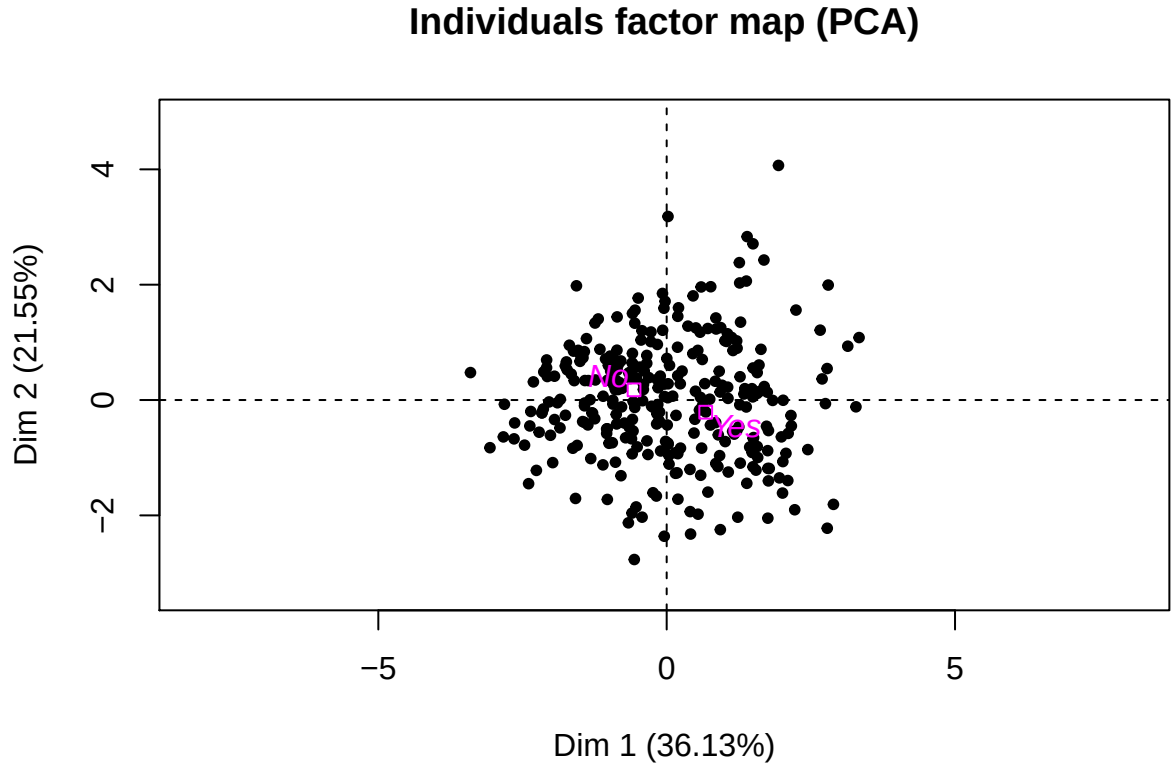

```
#Represented in Rp
#quali.sup -> Every modality is the centroide of the respective individuals having chosen that modality
fviz_pca_ind(pca_facto_heart, habillage = 6, geom = "point", label="quali",addEllipses =TRUE, ellipse.l
```



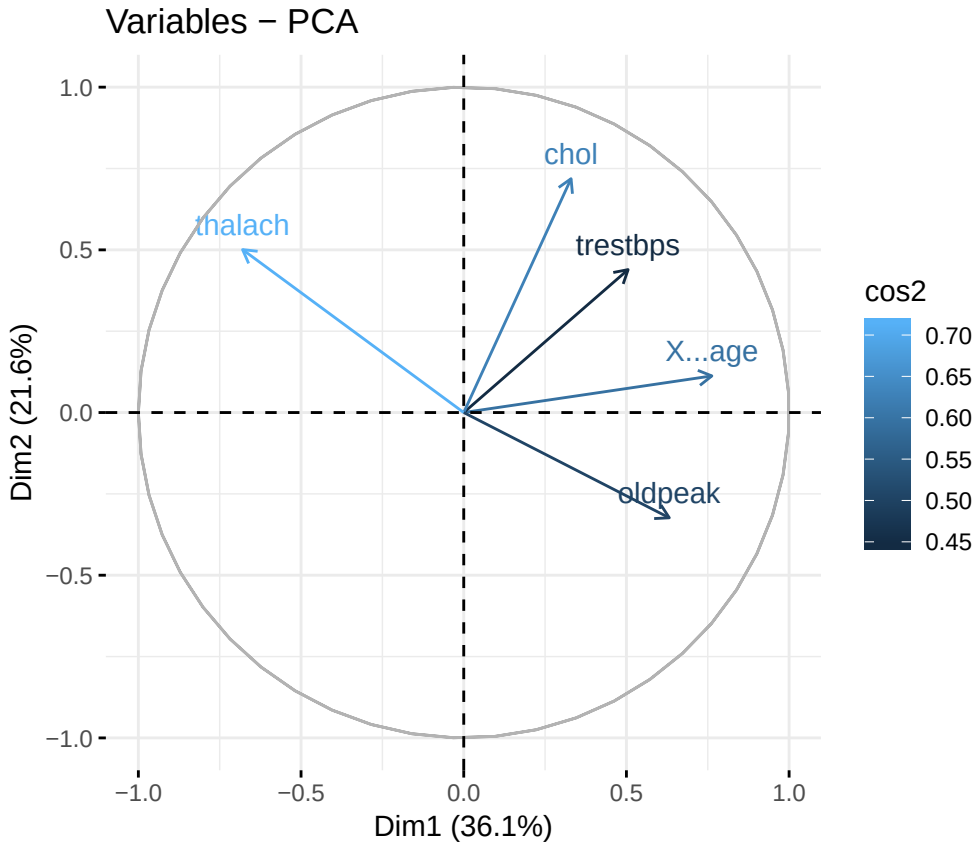

```
plot.PCA(pca_facto_heart, quali.sup = 6, scale.unit = TRUE,choix = 'ind',label="quali")
```

```
## Warning in plot.window(...): "quali.sup" is not a graphical parameter

## Warning in plot.window(...): "scale.unit" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "quali.sup" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "scale.unit" is not a graphical
## parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "quali.sup" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "scale.unit"
## is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "quali.sup" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "scale.unit"
## is not a graphical parameter

## Warning in box(...): "quali.sup" is not a graphical parameter

## Warning in box(...): "scale.unit" is not a graphical parameter

## Warning in title(...): "quali.sup" is not a graphical parameter

## Warning in title(...): "scale.unit" is not a graphical parameter

## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "quali.sup" is not a graphical parameter
```

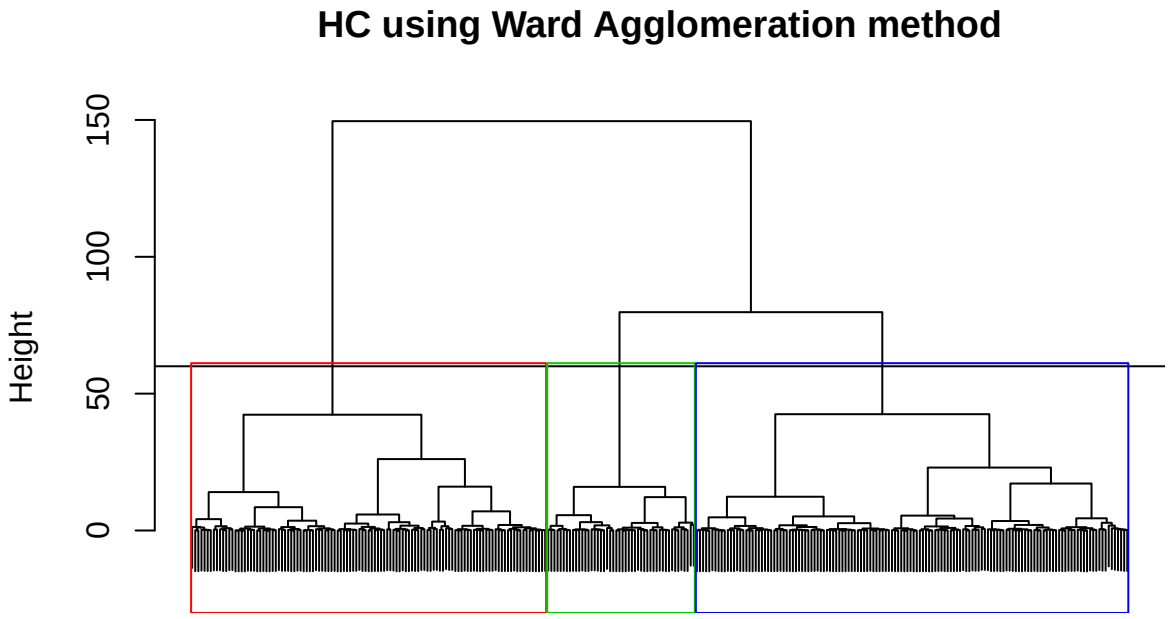
```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "scale.unit" is not a graphical parameter

## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "quali.sup" is not a graphical parameter

## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "scale.unit" is not a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "quali.sup" is not a
## graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "scale.unit" is not
## a graphical parameter

## Warning in text.default(xy, labels, cex = cex, ...): "quali.sup" is not a
## graphical parameter

## Warning in text.default(xy, labels, cex = cex, ...): "scale.unit" is not a
## graphical parameter
```



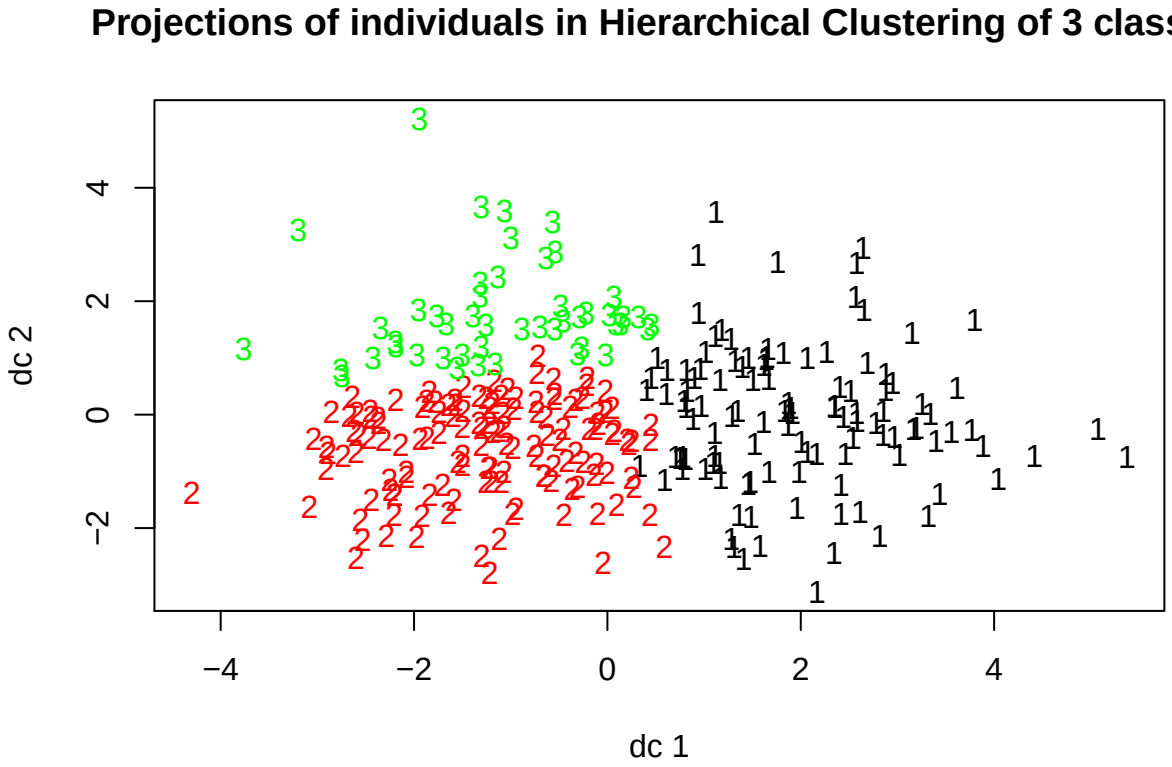

```
#Represented in Rn
#Projection of variables, show correlation between principal components
fviz_pca_var(pca_facto_heart, geom = c("arrow", "text"), col.var = "cos2")#By quality of representation
```




```
proj_indiv <- pca_facto_heart$ind$coord[,1:2] #individual projections on 1st factorial plane
#Clustering
hc_ward = hclust(dist(proj_indiv),method = "ward.D")
plot(hc_ward, main= "HC using Ward Agglomeration method", xlab="",sub="",cex=.9,  labels=FALSE)
abline(h=60)
rect.hclust(hc_ward, k = 3, border = 2:6)
```

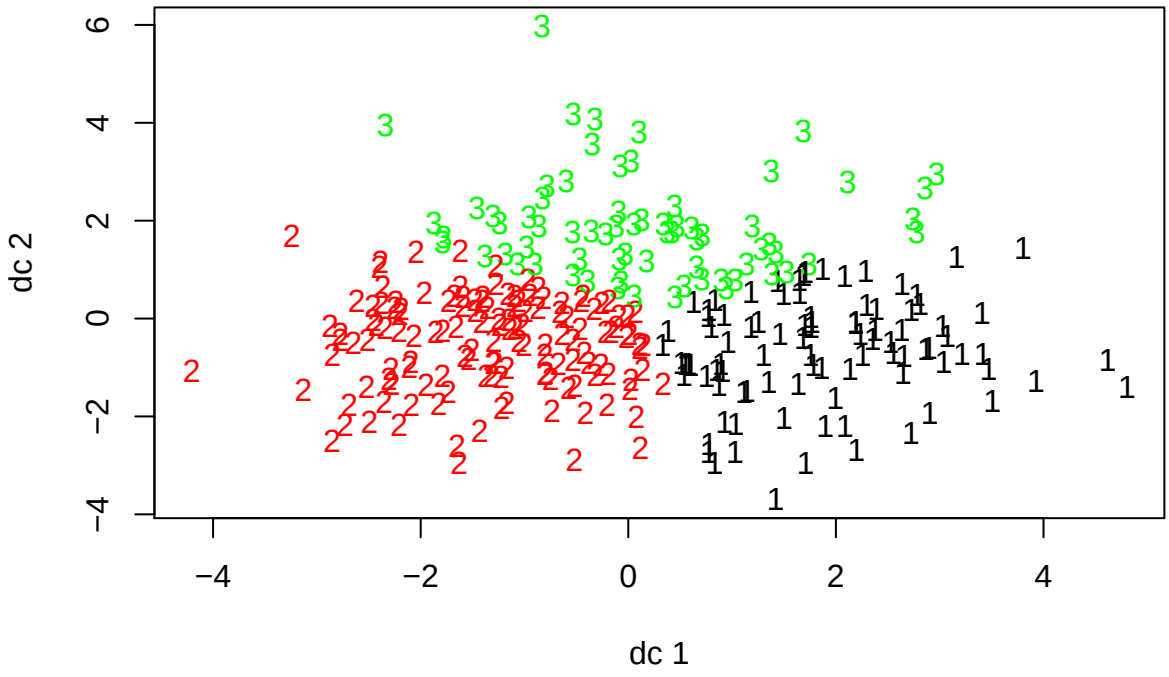


```r
#Association of individuals to clusters
classes <- cutree(hc_ward, h=50) #Depending on the height, number of clusters is chosen
plotcluster(proj_indiv, classes,main="Projections of individuals in Hierarchical Clustering of 3 classes
```




```r
get_centroids <- function(classes, n_classes){
  centroids <- NULL
  for(k in 1:n_classes){
    centroids <- rbind(centroids, colMeans(proj_indiv[classes == k, , drop = FALSE]))
  }
  return(centroids)
}
centroids <- get_centroids(classes, 3)
```

```r
#k_mean needs centroid of clusters
k_mean <- kmeans(proj_indiv, centroids)
plotcluster(proj_indiv, k_mean$cluster,main="Projections of individuals in K-means Clustering of 3 class
```




```
cal_idx_before <- calinhara(proj_indiv,classes,cn=max(classes))
cal_idx_after <- calinhara(proj_indiv,k_mean$cluster,cn=max(k_mean$cluster))

print(cal_idx_before)
```

```
## [1] 198.1154
```

```
print(cal_idx_after)
```

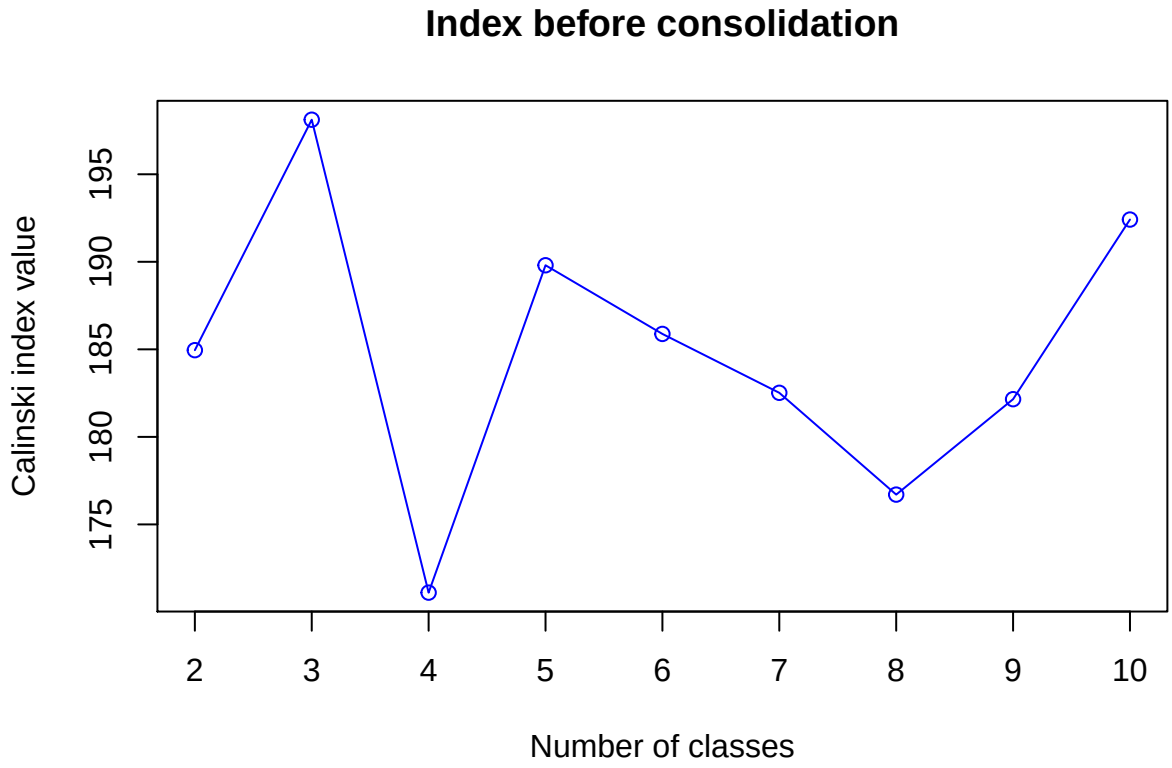
```
## [1] 226.1952
```

```
#Improvement
```

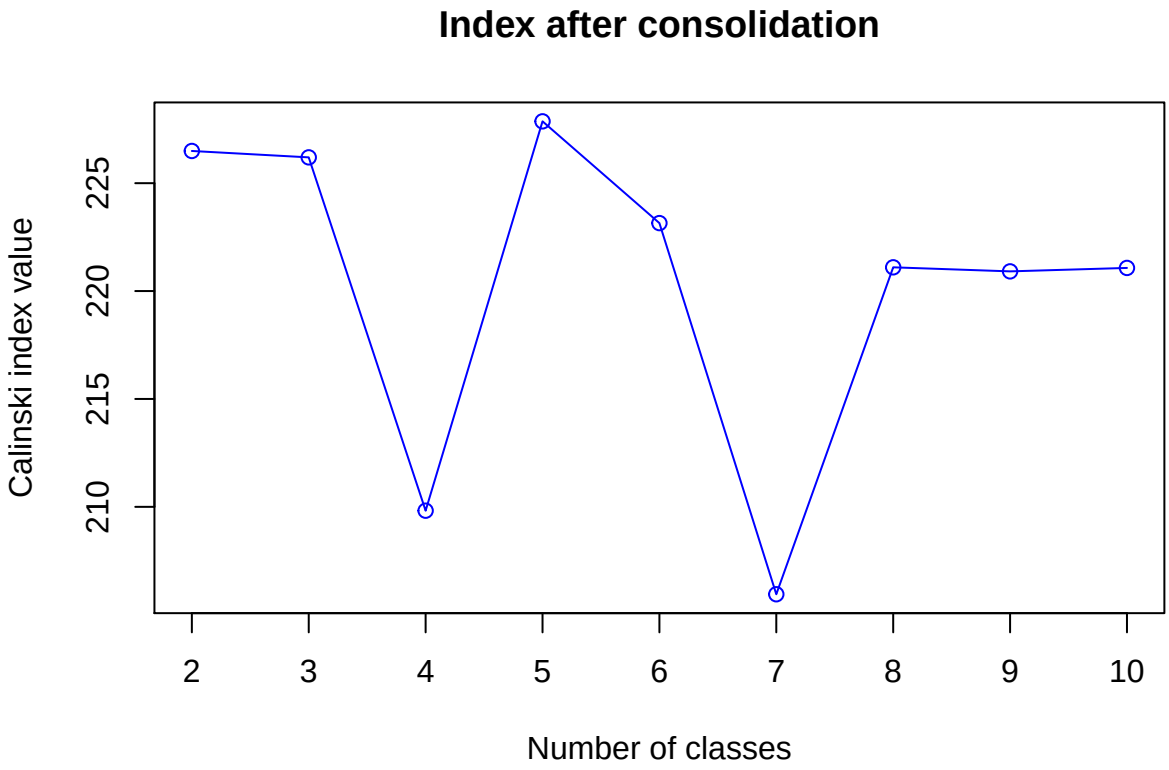
```
Calinski_Harabassza <- function (projections, hc, kind, n_classes){
  classes <- cutree(hc, k=n_classes)
  centroids <- get_centroids(classes, n_classes)
  if(kind=='hc'){
    index <- calinhara(proj_indiv,classes,cn=max(classes))
  }
  if(kind=='kmeans'){
    kmeans_classes <- kmeans(proj_indiv, centers = centroids)$cluster
    index <-calinhara(proj_indiv,kmeans_classes,cn=max(kmeans_classes))
  }
  return(index)
}
get_indexes <- function(until, kind){
  indexes <- c()
  for (n_classes in 2:until){
    indexes <- c(indexes, Calinski_Harabassza(proj_indiv, hc_ward, kind, n_classes))
  }
  return(indexes)
```

```
}
```

```
indexes_before <- get_indexes(10, 'hc')
plot(indexes_before, type = "o", xlab = 'Number of classes', ylab = 'Calinski index value'
, main = 'Index before consolidation', col = 'blue', xaxt
= "n")
axis(1, at=1:9, labels = c(2, 3, 4, 5, 6, 7,8,9,10))
```




```
indexes_after <- get_indexes(10, 'kmeans')
plot(indexes_after, type = "o", xlab = 'Number of classes', ylab = 'Calinski index value'
, main = 'Index after consolidation', col = 'blue', xaxt
= "n")
axis(1, at=1:9, labels = c(2, 3, 4, 5, 6, 7,8,9,10))
```


Index after consolidation
Calinski index value
210
215
220
225
2
3
4
5
6
7
8
9
10
Number of classes