MVA Final Project

Javier Ferrando Monsonis
Marcel Porta Valles
Mehmet Fatih ??agil

February 20, 2018

Libraries

```
library(chemometrics)
```

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

```
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)
library(caret)
library(ROCR)
library(dplyr)
library(randomForest)
library(expm)
library(adegraphics)
library(fpc)
theme_set(theme_bw())
setwd("/Users/JaviFerrando/Desktop/MVA-Project")
```

```
heart_disease = read.csv("data/heart.csv")
columns <- colnames(heart_disease)
columns[1] <- "age"
colnames(heart_disease) <- columns

insert_nas <- function(x) {
  len <- length(x)
  n <- sample(1:floor(0.05*len), 1)
  i <- sample(1:len, n)
  x[i] <- NA
  x
}
```

```
heart_disease_missing <- sapply(heart_disease, insert_nas)
kable(head(heart_disease_missing))
```

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

```
knn_data <- knnImputation(heart_disease_missing, k = 1, scale = T)
kable(head(knn_data))
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
51	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	1	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

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

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

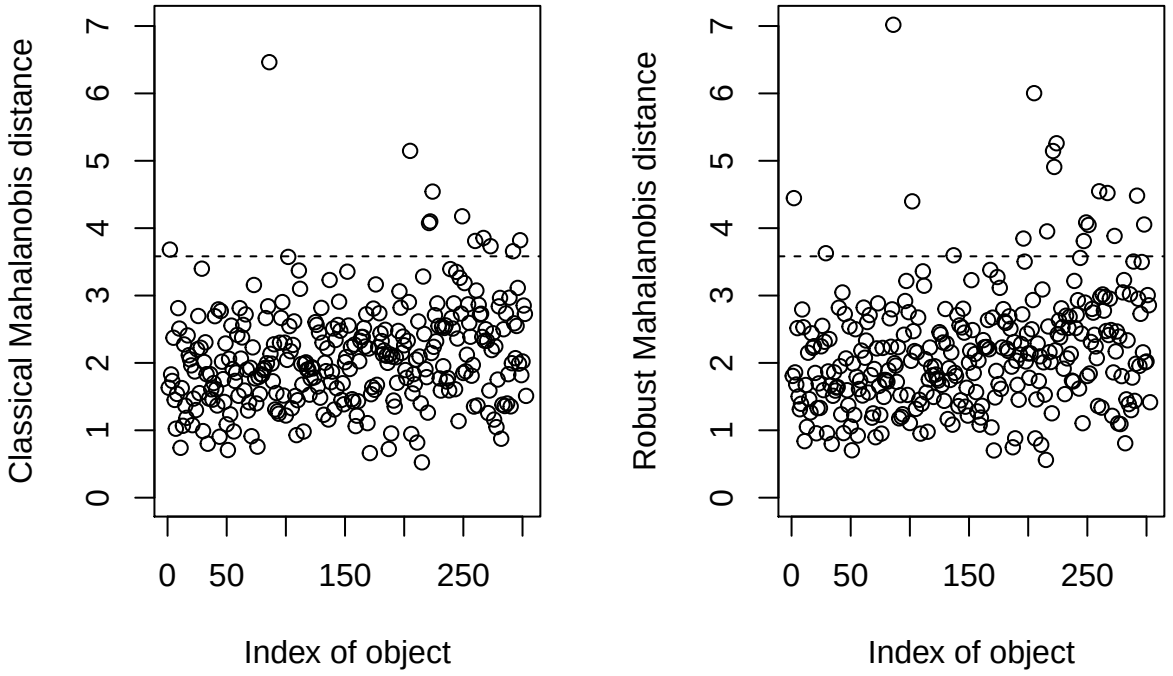
```
#kable(head(heart_disease))
#describe(heart_disease)
```

```
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)
```

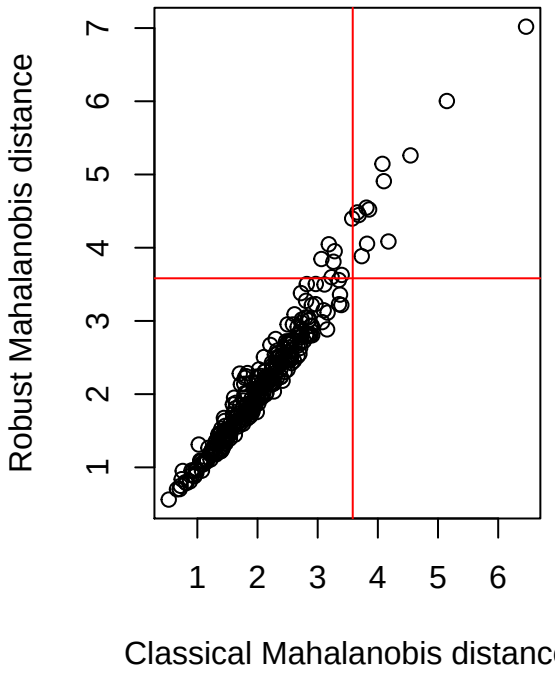
```
#Outlier detection
############################
cont_heart <- factor_heart[, sapply(factor_heart, class) != "factor"]
mout <- Moutlier(cont_heart, quantile = 0.975, plot = TRUE, tol=1e-36) #Doesn't work
```

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

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

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

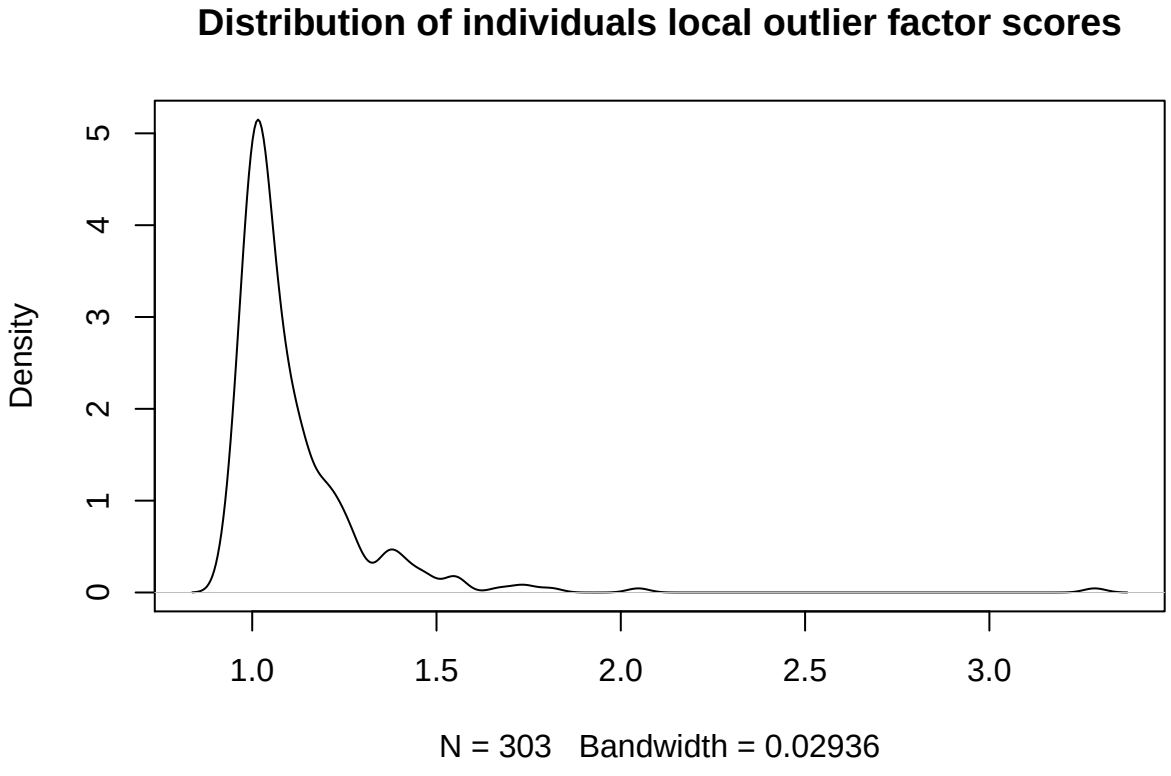
```
## Warning in axis(side = side, at = at, labels = labels, ...): "tol" is not a
## graphical parameter

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

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

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

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

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

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

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

## Warning in title(...): "tol" is not a graphical parameter
```




```
plot(mout$md,mout$rd,xlab='Classical Mahalanobis distance',ylab='Robust Mahalanobis distance')
abline(h = mout$cutoff, col="red")  # add cutoff line
abline(v = mout$cutoff, col="red")  # add cutoff line
```

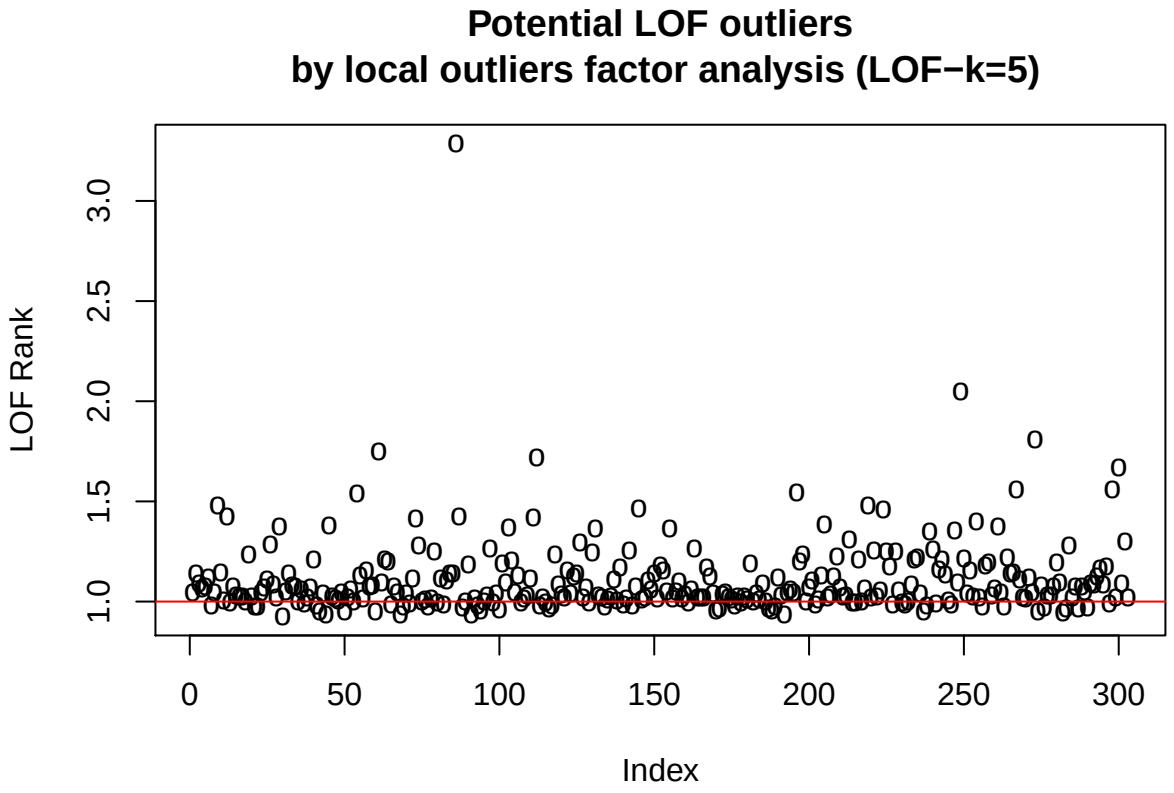



```
#Local Outlier Factor
outlier.scores <- lofactor(heart_disease[,-14], k=5)
plot(density(outlier.scores),main='Distribution of individuals local outlier factor scores')
```



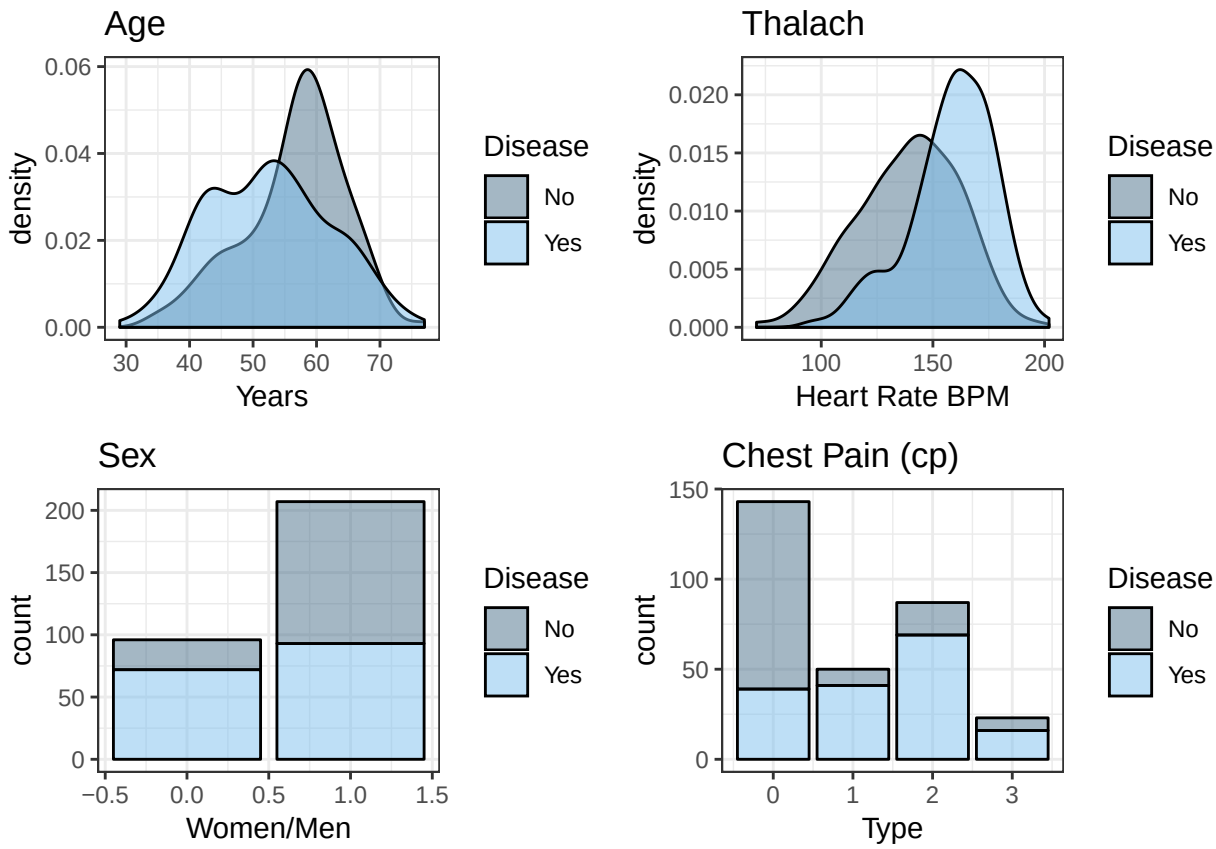

```
LOF_plot <- plot(outlier.scores,
                 pch="o",
                 cex=1,
                 main="Potential LOF outliers\n by local outliers factor analysis (LOF-k=5)",
```

```
                ylab="LOF Rank")
#LOF_plot_cutoff <- 0.5*(LOF_df[LOF_index_ordered[4],]$LOF_rank + LOF_df[LOF_index_ordered[5],]$LOF_ran
abline(h = 1, col="red")  # add cutoff line
```

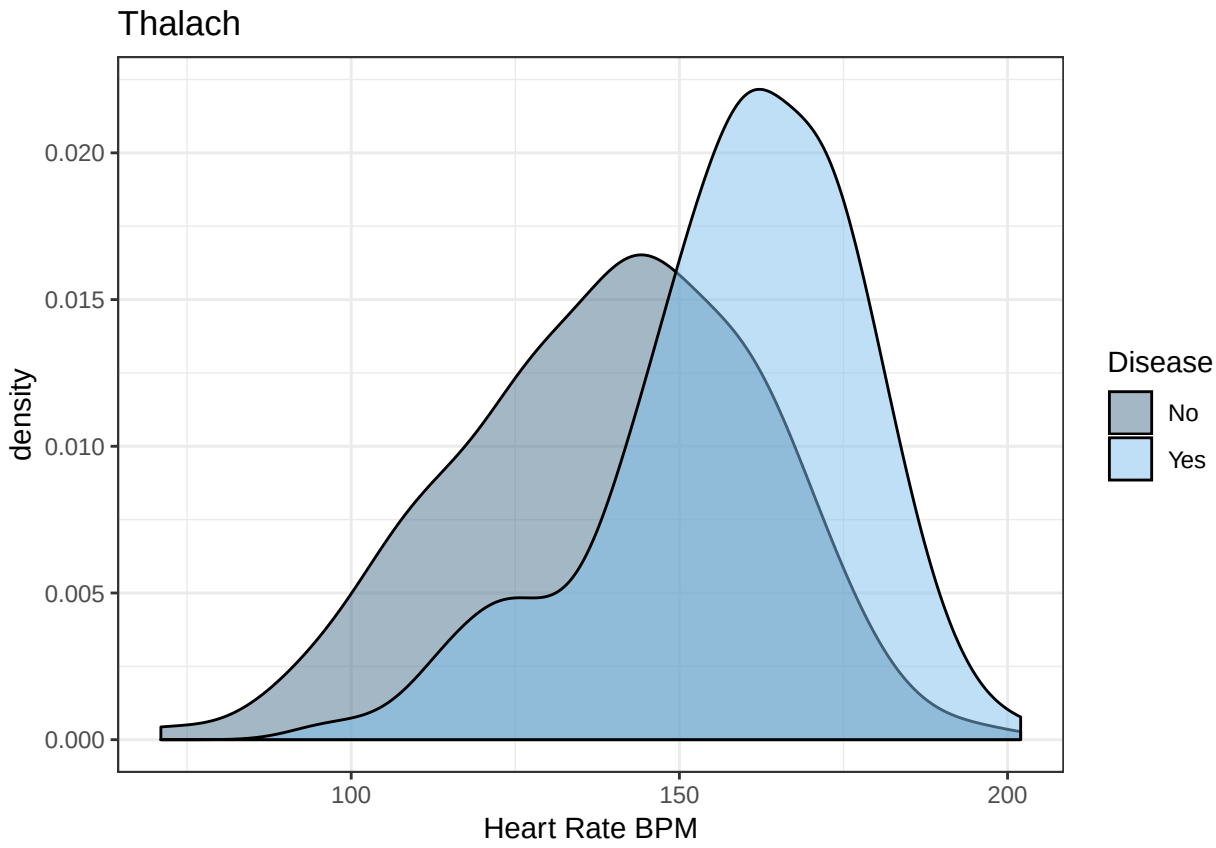



```
#Exploratory Data Analysis
#Density of heart presence/absence disease by age
g1 <- ggplot(data=heart_disease, aes(x=age, fill=as.factor(target)))+
  geom_density(alpha=.5)+
  labs(x = 'Years', title = 'Age') +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("No",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

#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)+
  labs(x = 'Heart Rate BPM', title = 'Thalach') +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("No", "Yes"))

#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")+
      labs(x = 'Women/Men', title = 'Sex') +
      #scale_x_discrete(breaks=c("0","1"),labels=c("Women", "Men")) +
      scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("No", "Yes"))

#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")+
  labs(x = 'Type', title = 'Chest Pain (cp)') +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("No", "Yes"))
```

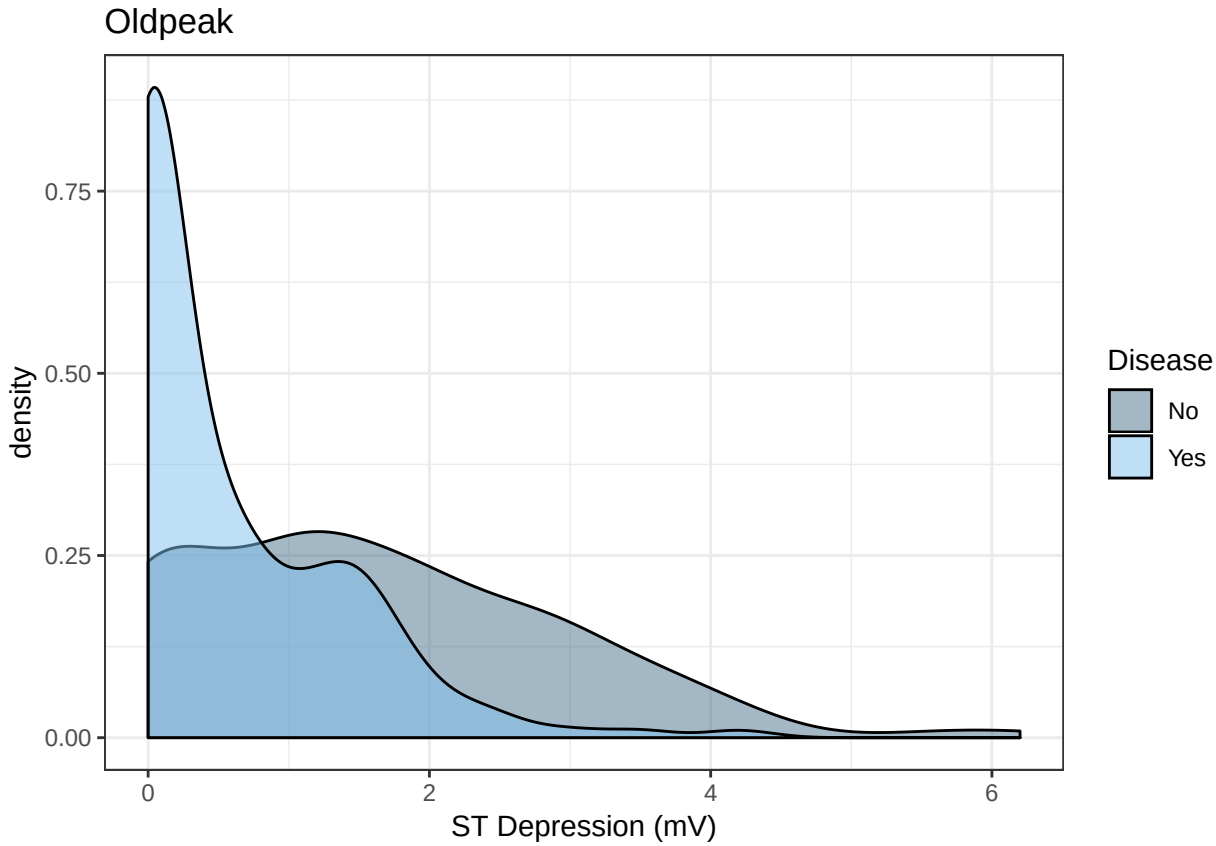
```
grid.arrange(g1, g2, g3, g4, ncol = 2)
```



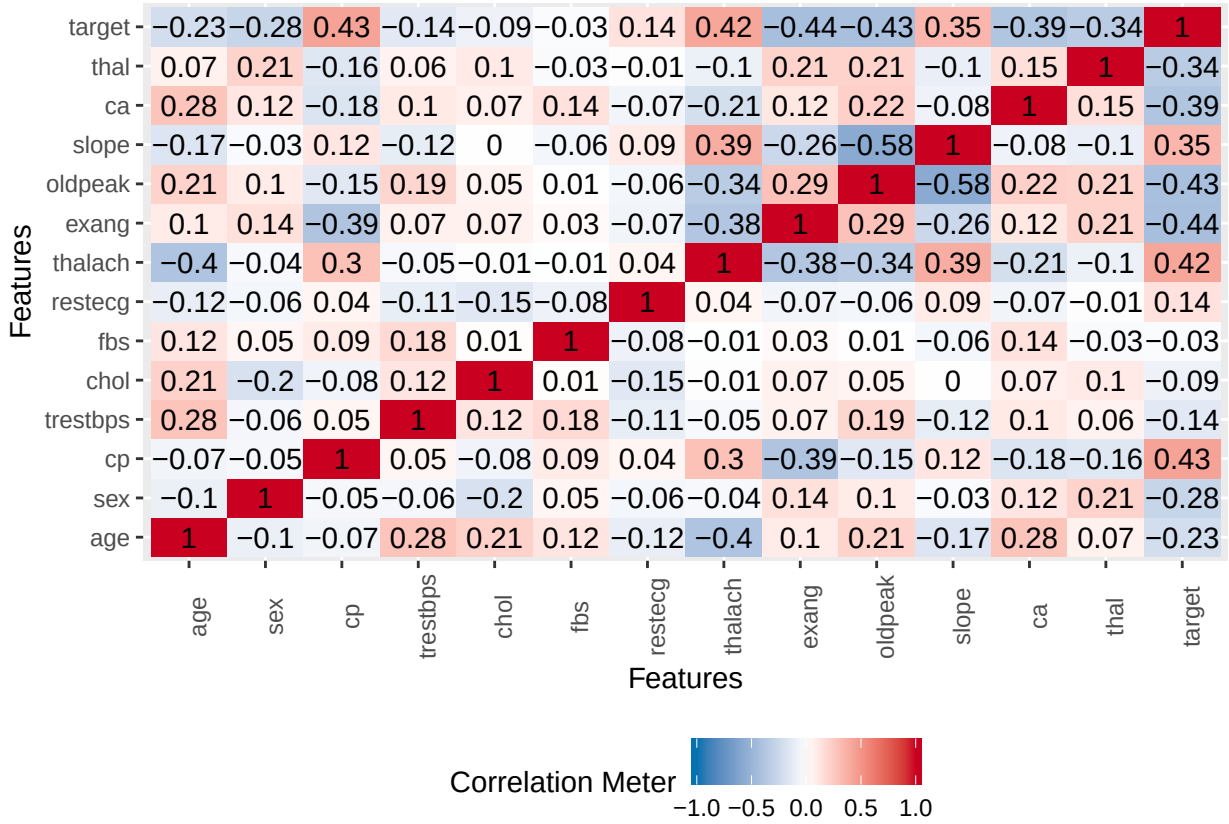

```
grid.arrange(g2)
```



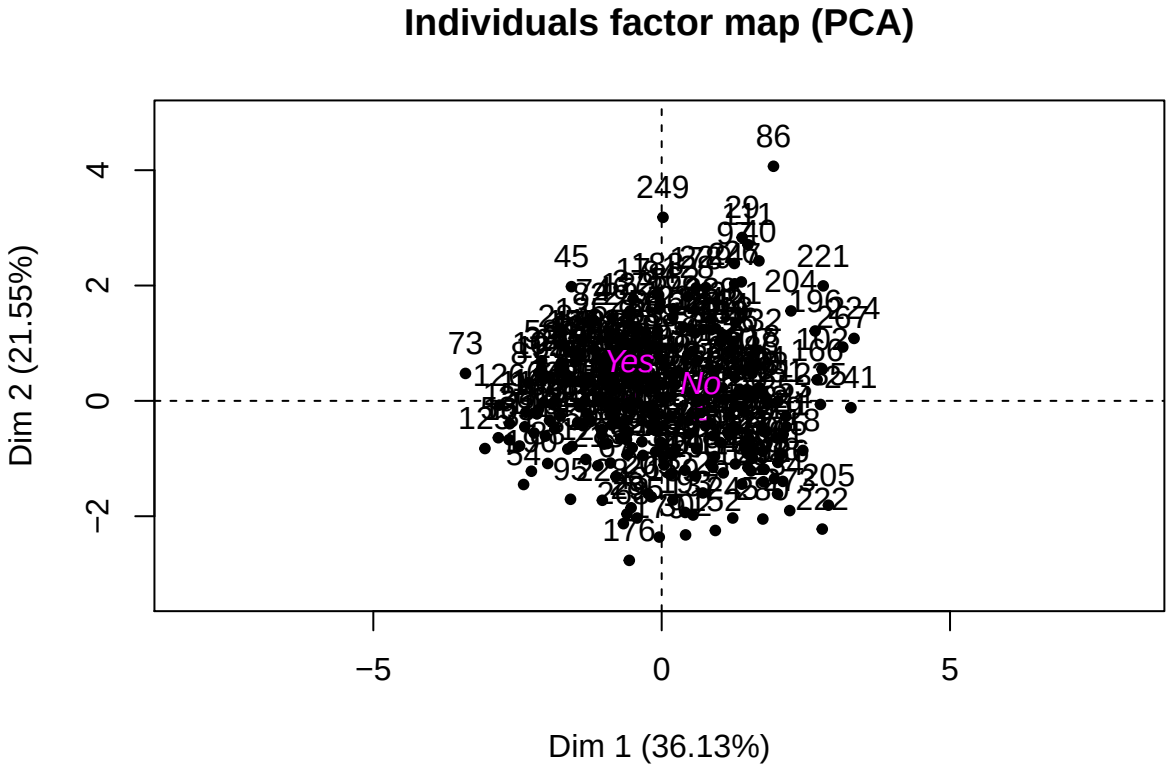

```
g5 <- ggplot(data=heart_disease, aes(x=oldpeak, fill=as.factor(target)))+
  geom_density(alpha=.5) +
  labs(x = 'ST Depression (mV)', title = 'Oldpeak') +
  scale_fill_manual(values = c('skyblue4', 'skyblue2'),name = "Disease", labels = c("No", "Yes"))
grid.arrange(g5)
```

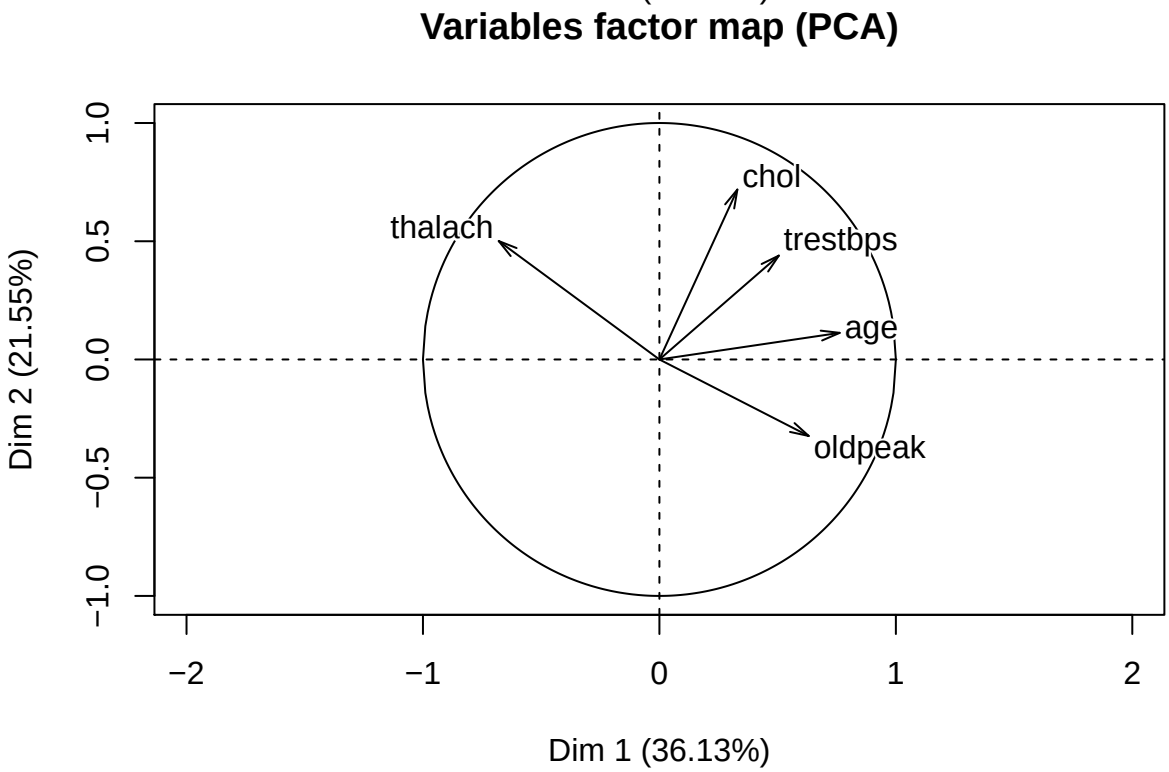



```
plot_correlation(heart_disease)
```

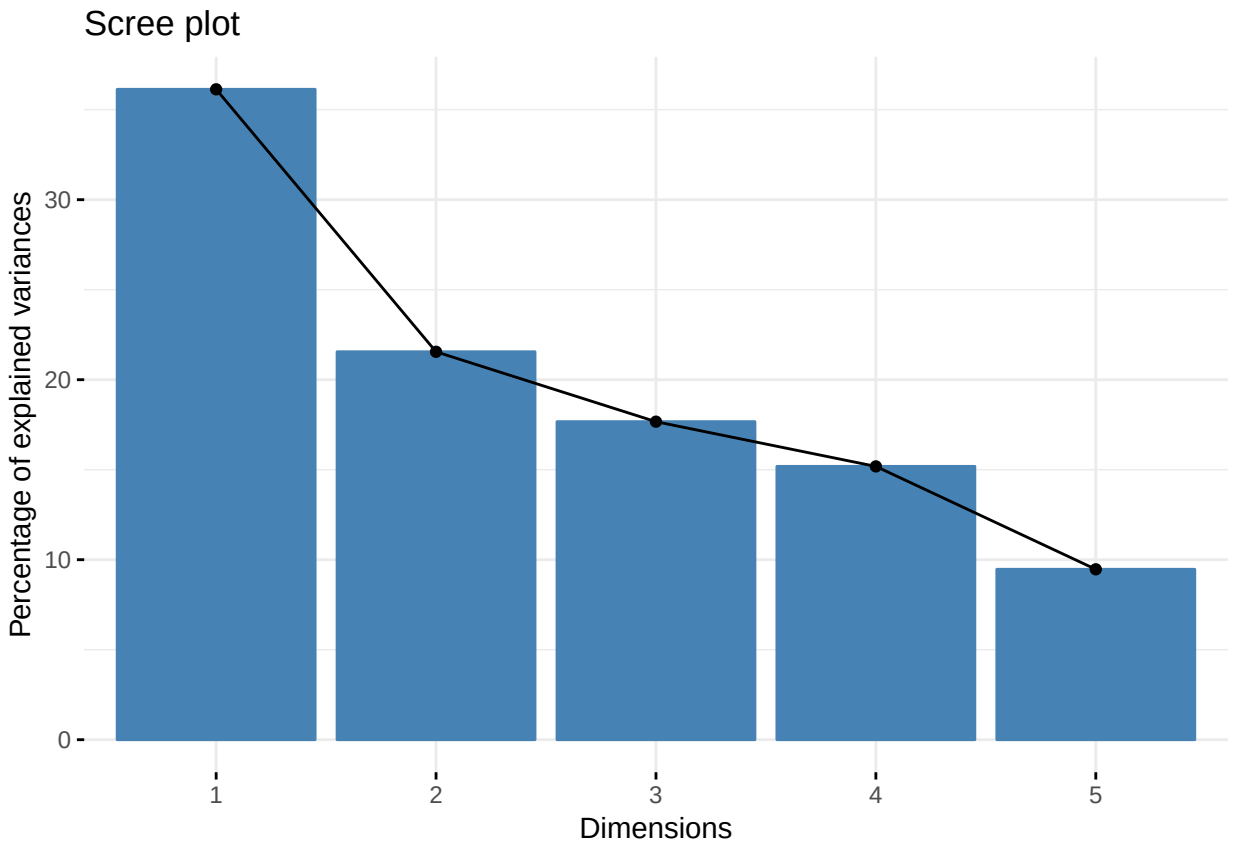



```
#PCA with continuous values
pca_facto <- factor_heart[, sapply(factor_heart, class) != "factor"]
#Some categorical values can be added as supplementary
#pca_facto$sex <- factor_heart$sex
#pca_facto$ca <- factor_heart$ca
pca_facto$disease <- heart_disease$target
pca_facto$disease[pca_facto$disease==0]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

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

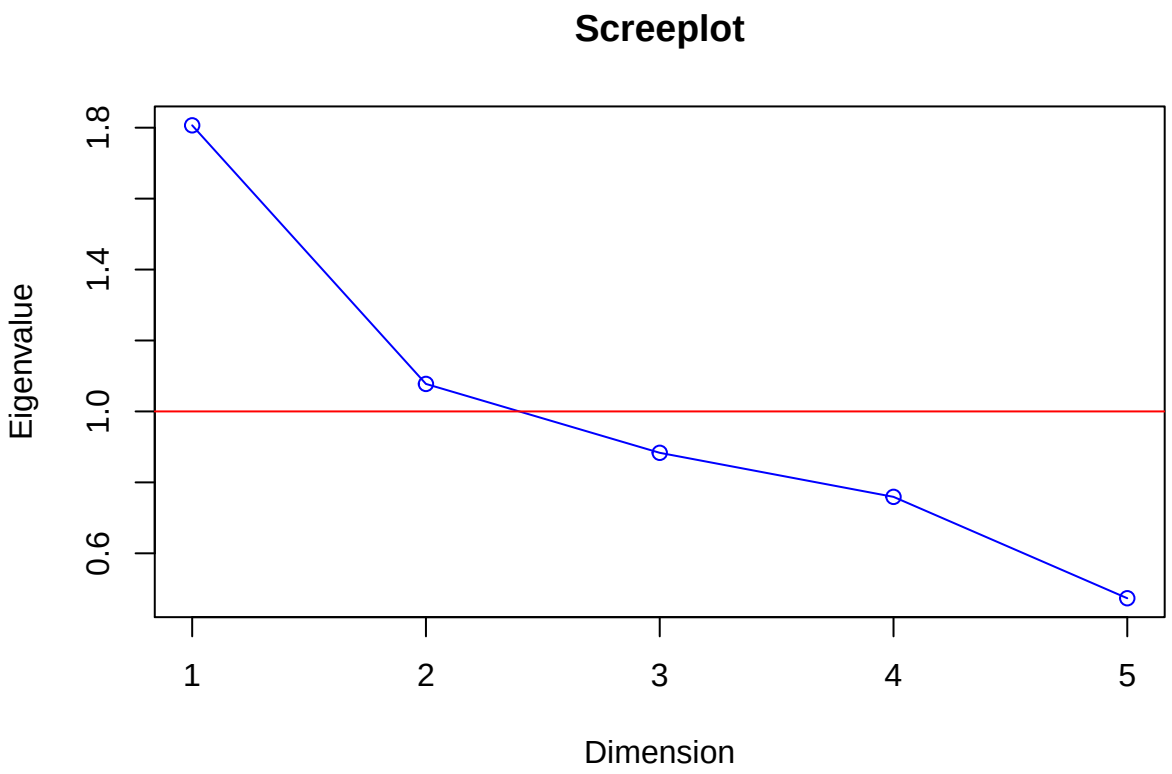




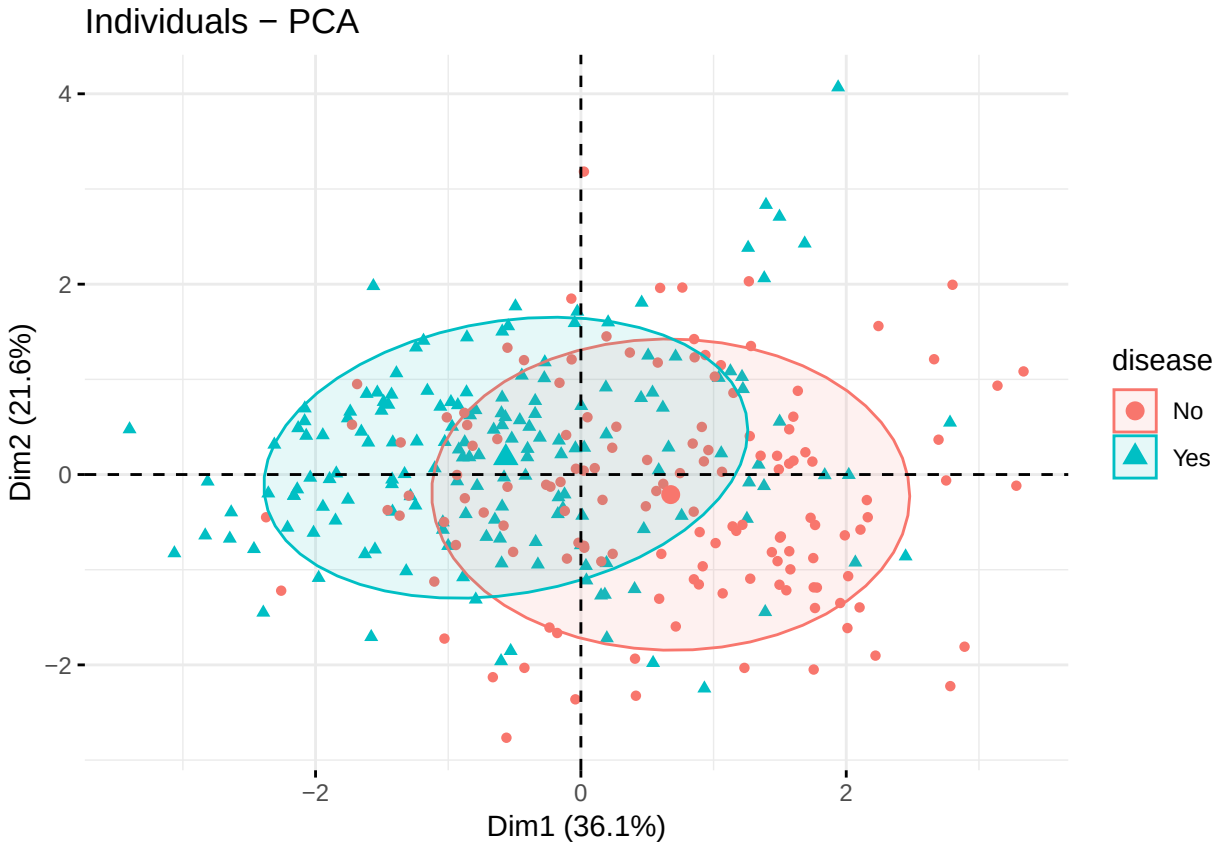

```
#Screeplots
fviz_screeplot(pca_facto_heart, addlabels = FALSE)
```




```
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
```

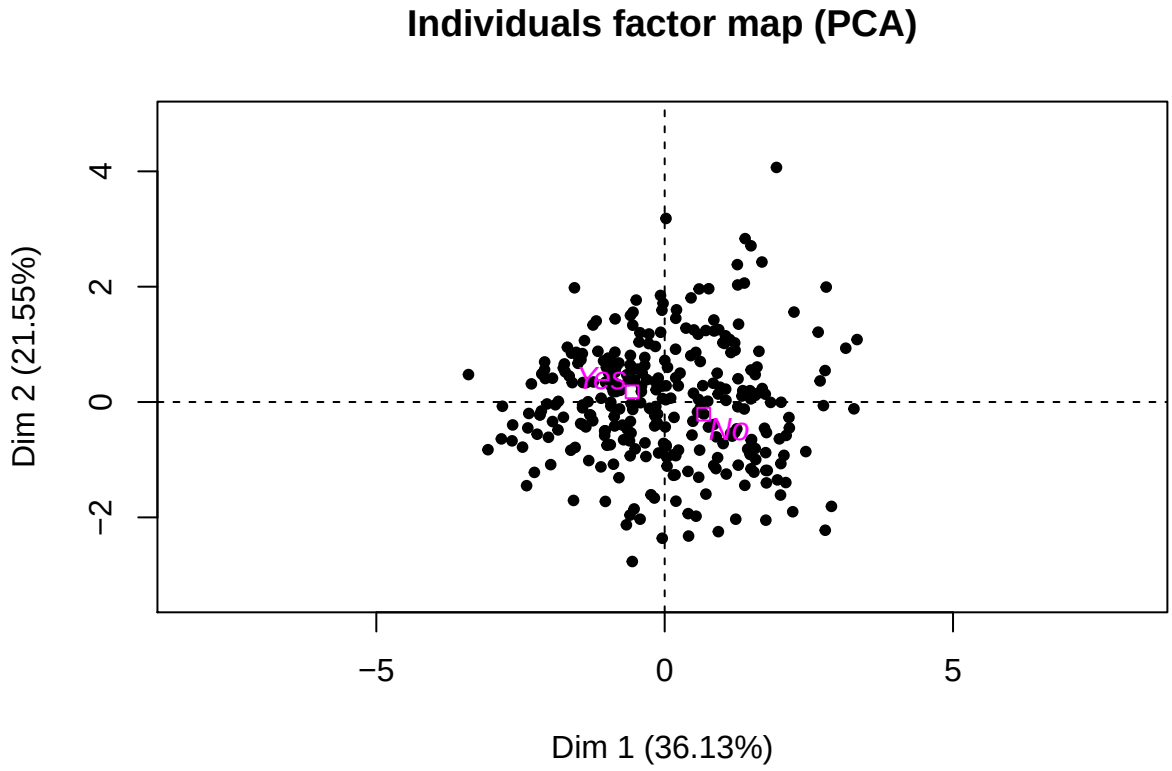
```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...):
## "scale.unit" is not a graphical parameter
```

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

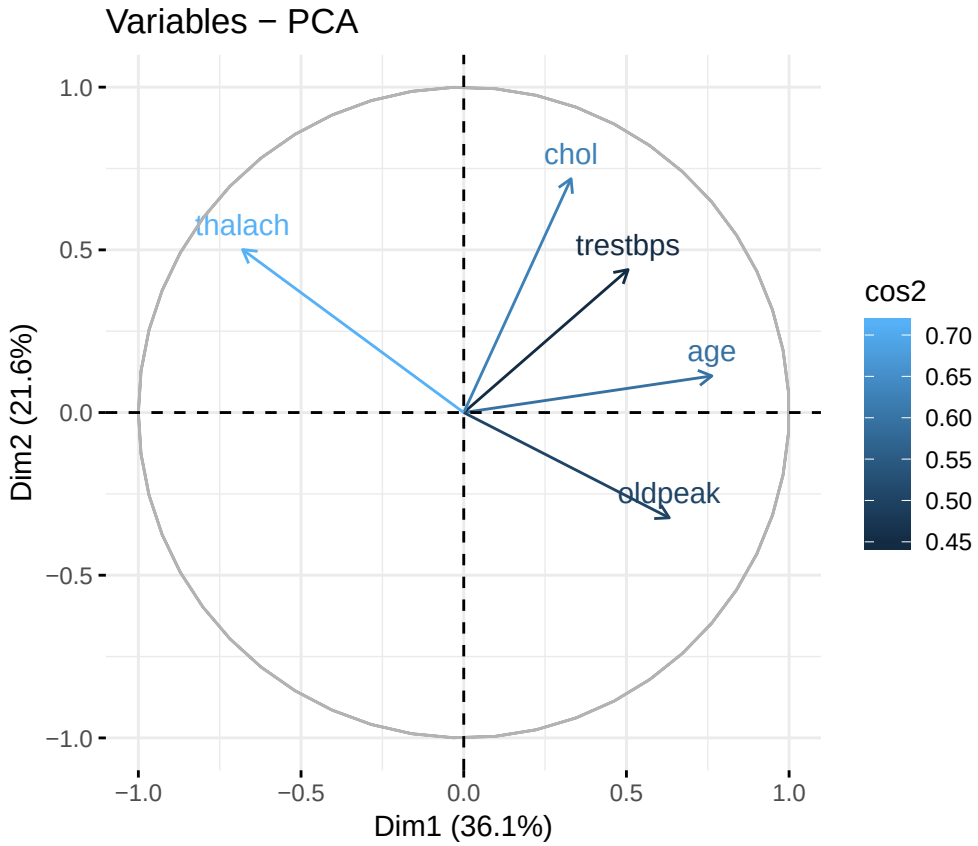
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "scale.unit" is not
## a graphical parameter
```

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

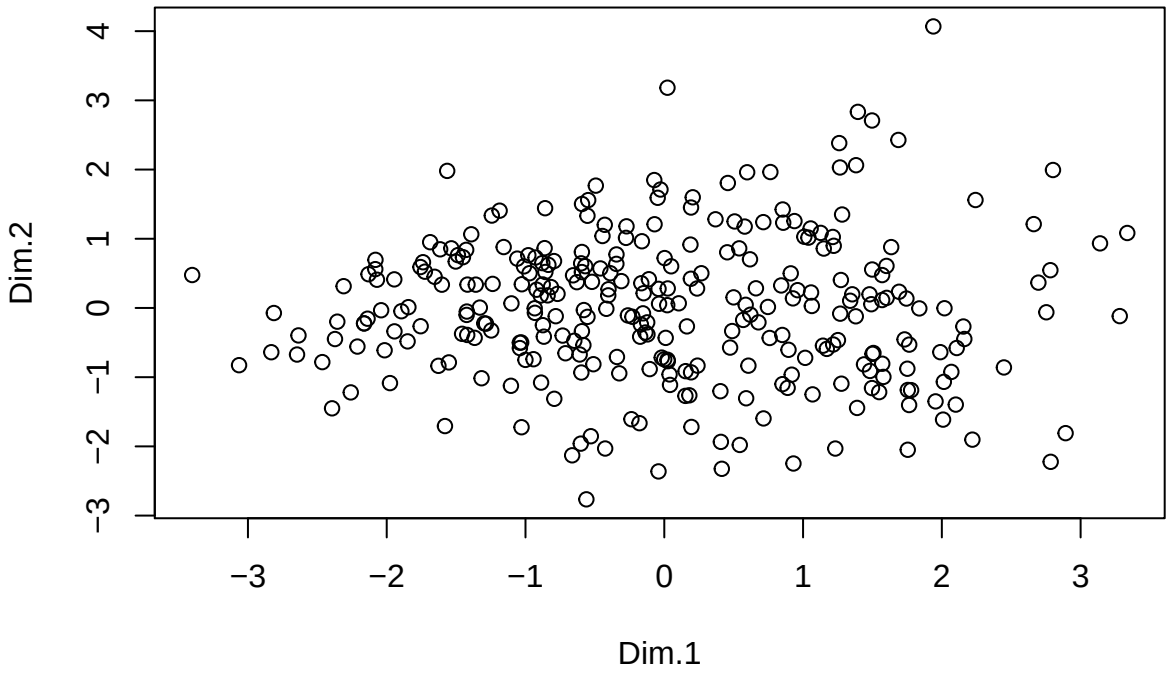
```
## Warning in text.default(xy, labels, cex = cex, ...): "scale.unit" is not a
## graphical parameter
```




```
#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
```

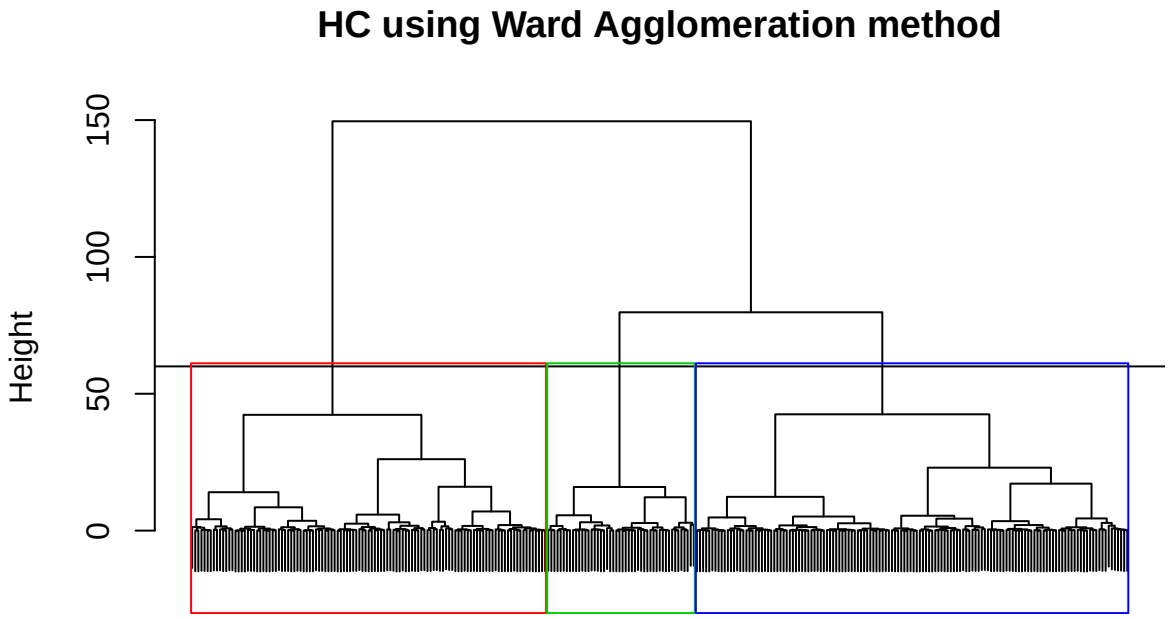



```
proj_indiv <- pca_facto_heart$ind$coord[,1:2] #individual projections on 1st factorial plane
plot(proj_indiv)
```

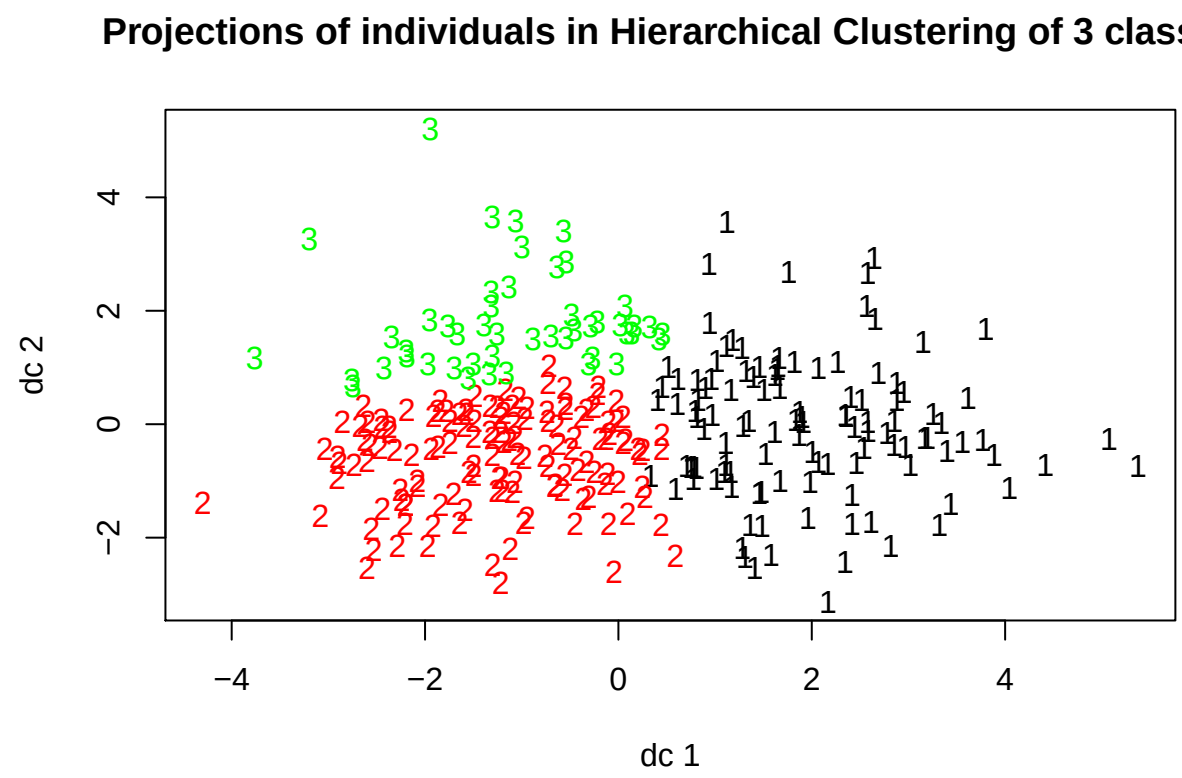



```
#Clustering
hc_ward = hclust(dist(proj_indiv),method = "ward.D")
plot(hc_ward, main= "HC using Ward Agglomeration method", xlab="",sub="",cex=.9,  labels=FALSE)
```

```r
abline(h=60)
rect.hclust(hc_ward, k = 3, border = 2:6)
```



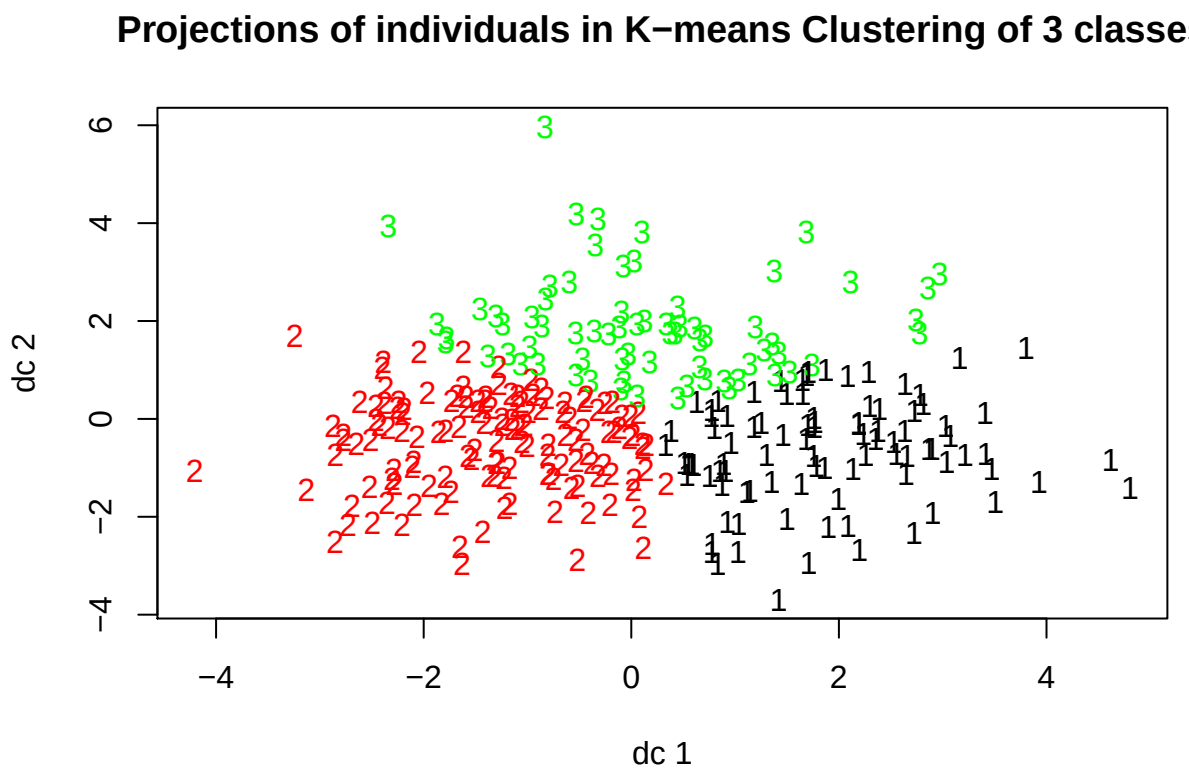

```r
#Association of individuals to clusters
classes <- cutree(hc_ward, h=50) #Depending on the height, number of clusters is chosen
plotcluster(proj_indiv, classes,main="Projections of individuals in Hierarchical Clustering of 3 classes
```




```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)
```

```
#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 clas
```




```
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)
```

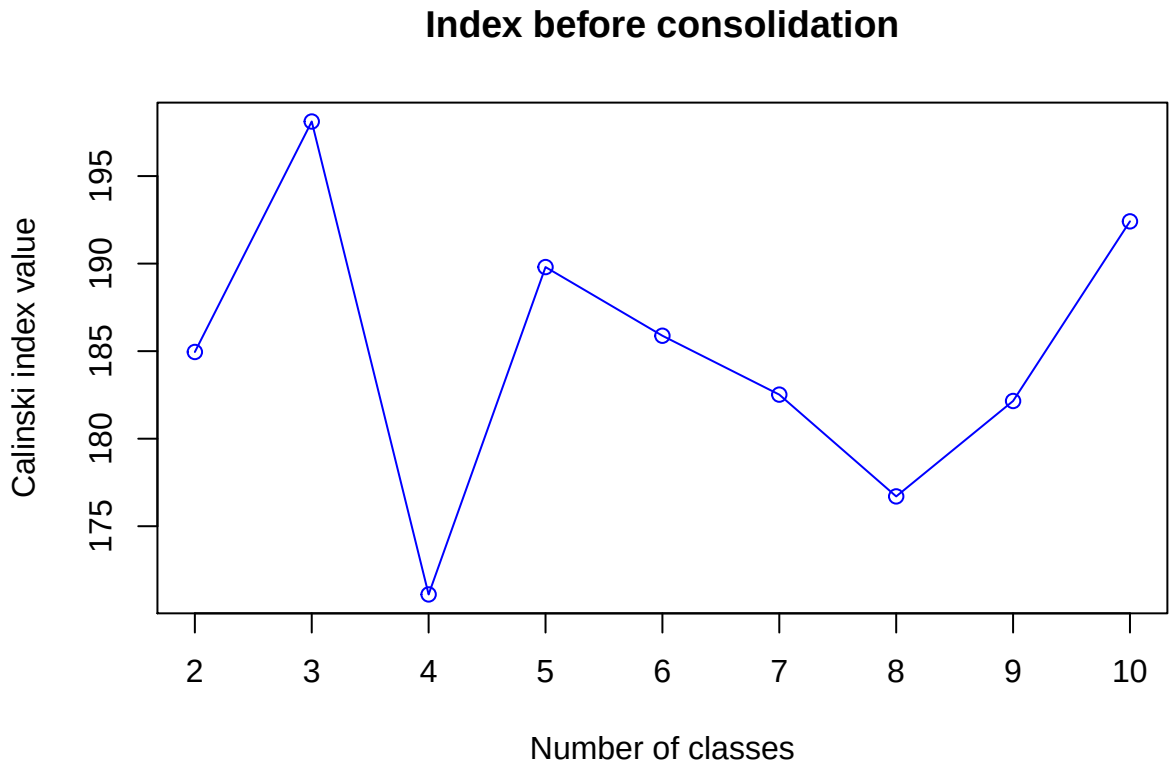
```
## [1] 226.1952
```

```
#Improvement
```

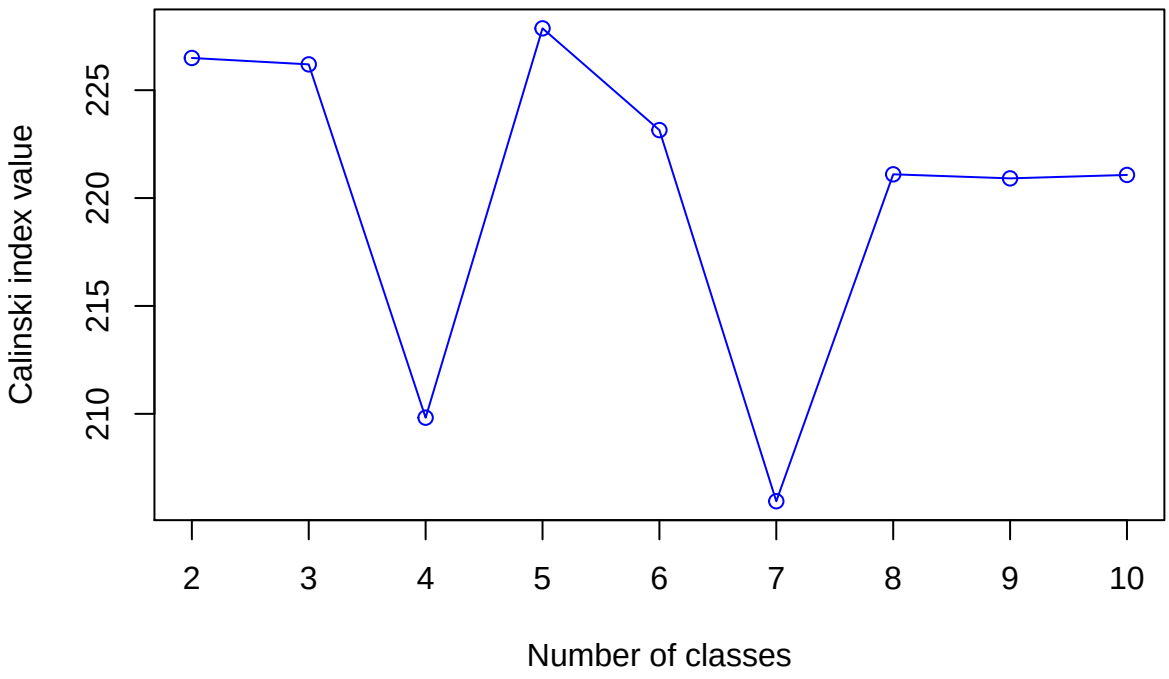
```
Calinski_Harabassza <- function (projections, hc, kind, n_classes){
  classes <- cutree(hc, k=n_classes)
  centroids <- get_centroids(classes, n_classes)
  if(kind=='hc'){
    index <- calinhara(proj_indiv,classes,cn=max(classes))
  }
  if(kind=='kmeans'){
    kmeans_classes <- kmeans(proj_indiv, 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))
```




```
first_factorial <- proj_indiv
df <- data.frame(first_factorial, Class = as.factor(k_mean$cluster))
df2 <- cbind(as.factor(k_mean$cluster),heart_disease[,1:13])
catdes_k_means <- catdes(df2, num.var = 1, proba = 0.05, row.w = NULL)

catdes_k_means$quanti$`1`[1:6,4]  #  p-values for cluster 1
```

```
##    oldpeak        age      exang         ca        sex       chol
##  1.2264232  7.0074040  0.4997398  1.0078889  0.4249693 41.3369221
```

```
catdes_k_means$quanti$`2`[1:6,4]  #  p-values for cluster 2
```

```
##    thalach      slope    restecg         cp       thal         ca
## 14.0625861  0.5818251  0.4831867  0.9355780  0.5354285  0.9521994
```

```
catdes_k_means$quanti$`3`[1:6,4]  #  p-values for cluster 3
```

```
##       chol   trestbps        age        fbs    restecg        sex
## 54.9322793 18.6108154  6.6042040  0.4199125  0.5182388  0.4991830
```

```
factor_heart$disease[factor_heart$target==0] <- "No"
factor_heart$disease[factor_heart$target==1] <- "Yes"
factor_heart$target <- NULL

factor_heart2 <- factor_heart
factor_heart2$age<-cut(factor_heart2$age, seq(0,80,10), right=FALSE)
factor_heart2$age <- paste("Age", factor_heart2$age, sep="_")
min(factor_heart2$oldpeak)
```

```
## [1] 0
```

```
factor_heart2$oldpeak<-cut(factor_heart2$oldpeak, seq(0,7,1), right=FALSE)
factor_heart2$oldpeak <- paste("Oldp", factor_heart2$oldpeak, sep="_")
factor_heart2$thalach<-cut(factor_heart2$thalach, seq(70,220,20), right=FALSE)
factor_heart2$thalach <- paste("thalach", factor_heart2$thalach, sep="_")
factor_heart2$trestbps<-cut(factor_heart2$trestbps, seq(80,220,20), right=FALSE)
factor_heart2$trestbps <- paste("thres", factor_heart2$trestbps, sep="_")
factor_heart2$chol<-cut(factor_heart2$chol, seq(100,600,100), right=FALSE)
factor_heart2$chol <- paste("Col", factor_heart2$chol, sep="_")
#factor_heart2$age <- NULL
kable(head(factor_heart2))
```

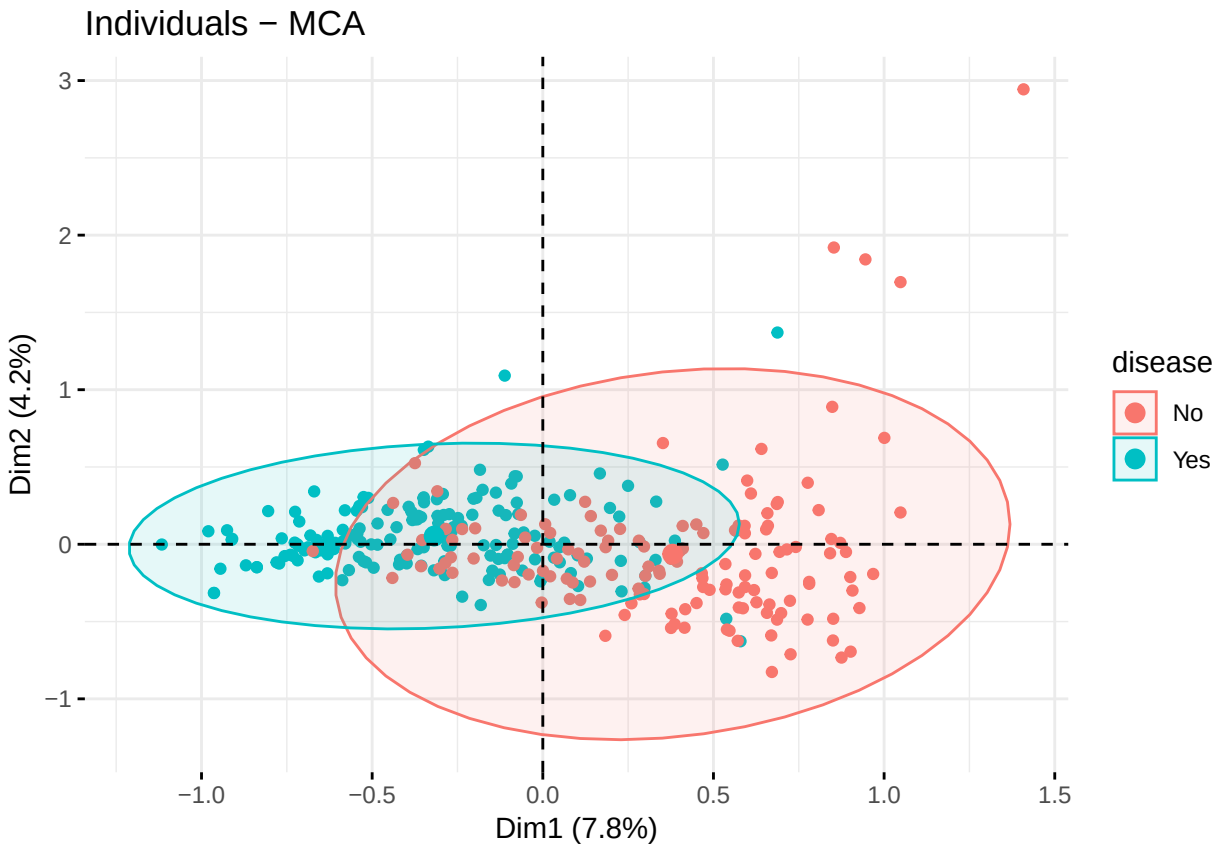
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slop
Age_[60,70)	1	3	thres_[140,160)	Col_[200,300)	1	0	thalach_[150,170)	0	Oldp_[2,3)	0
Age_[30,40)	1	2	thres_[120,140)	Col_[200,300)	0	1	thalach_[170,190)	0	Oldp_[3,4)	0
Age_[40,50)	0	1	thres_[120,140)	Col_[200,300)	0	0	thalach_[170,190)	0	Oldp_[1,2)	2
Age_[50,60)	1	1	thres_[120,140)	Col_[200,300)	0	1	thalach_[170,190)	0	Oldp_[0,1)	2
Age_[50,60)	0	0	thres_[120,140)	Col_[300,400)	0	1	thalach_[150,170)	1	Oldp_[0,1)	2
Age_[50,60)	1	0	thres_[140,160)	Col_[100,200)	0	1	thalach_[130,150)	0	Oldp_[0,1)	1

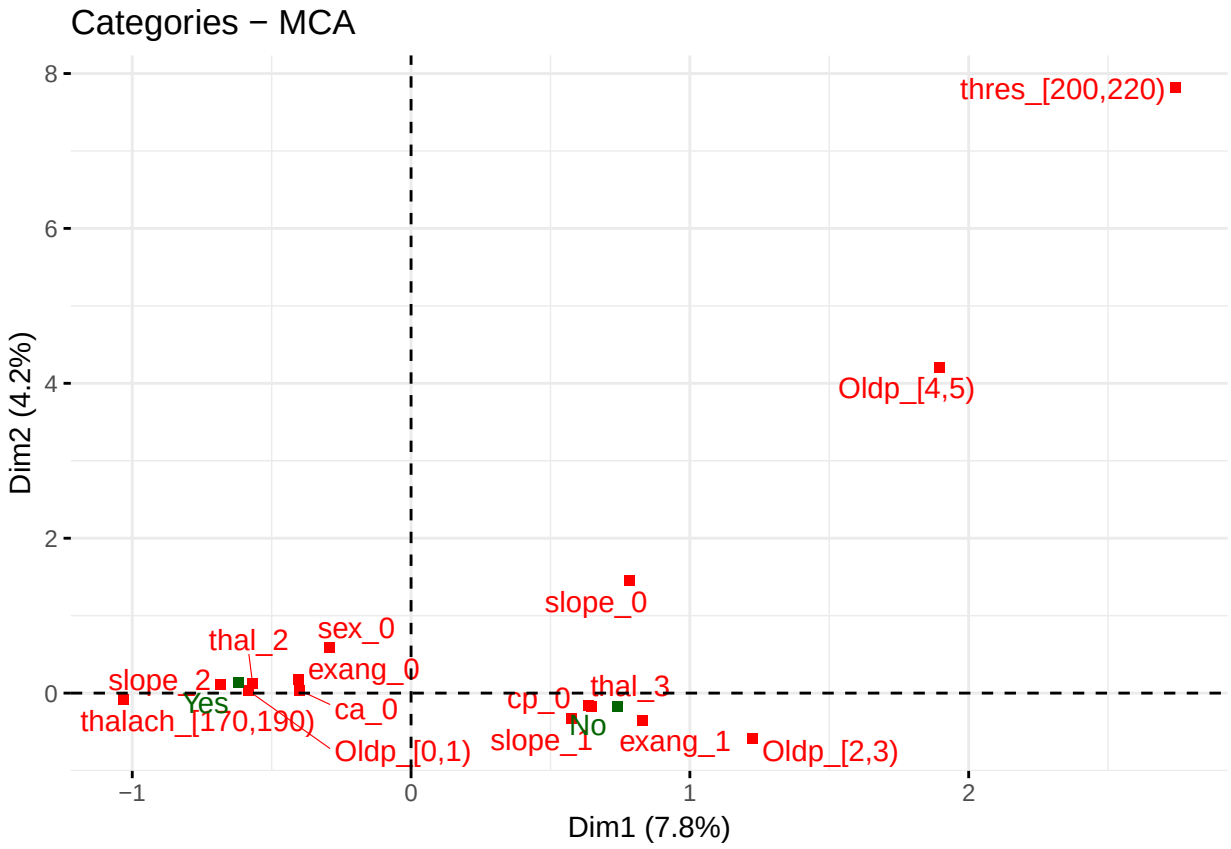
```
mcaHeart <- MCA(factor_heart2,ncp=7,
                #quanti.sup=c(10),
                quali.sup=c(14),
                excl=NULL,
                graph = FALSE,
                level.ventil = 0.00,
                axes = c(1,2),
                row.w = NULL,
                method="Indicator",
                na.method="NA",
                tab.disj=NULL)

# mcaHeart <- MCA(factor_heart,ncp=7,
#                 quanti.sup=c(1,4,5,8,10),
#                 quali.sup=c(14),
#                 excl=NULL,
#                 graph = FALSE,
#                 level.ventil = 0.00,
#                 axes = c(1,2),
#                 row.w = NULL,
#                 method="Indicator",
#                 na.method="NA",
#                 tab.disj=NULL)
summary(mcaHeart)
```

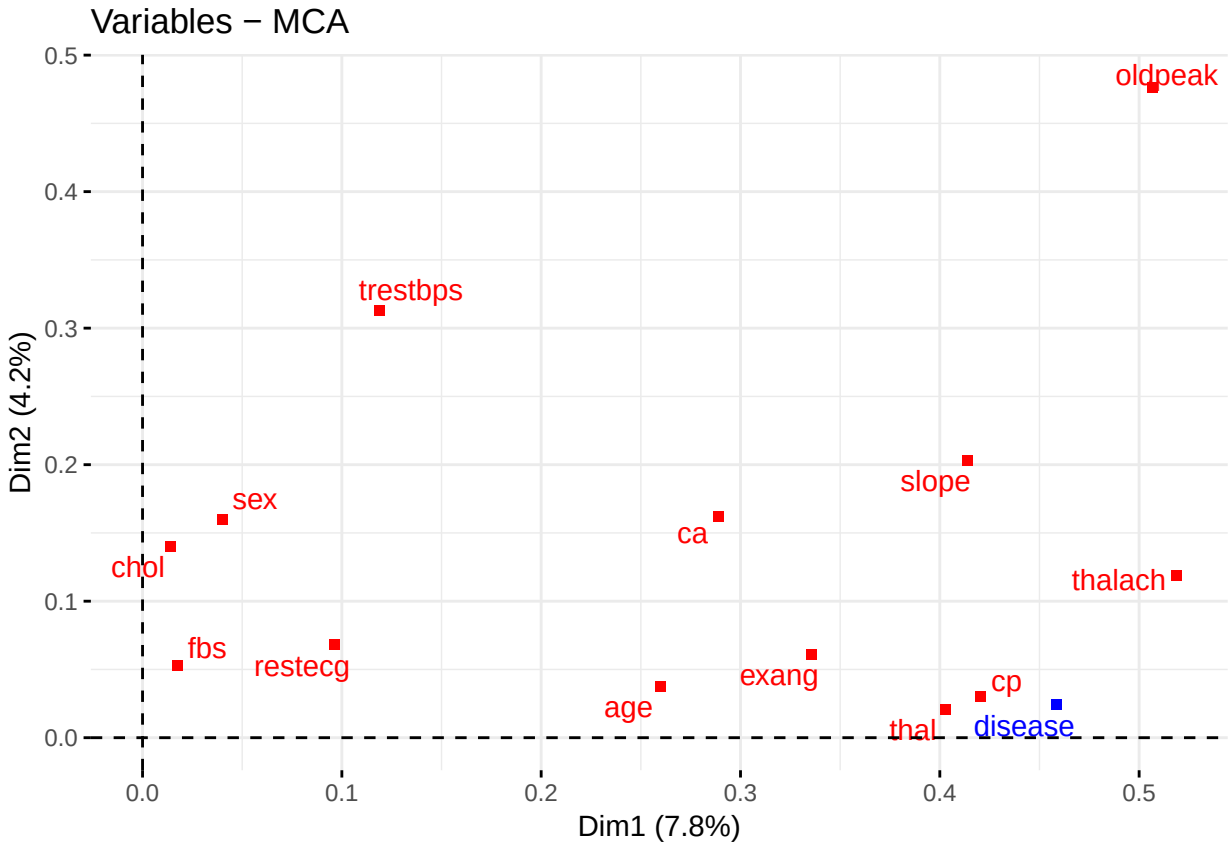
```
##
## Call:
## MCA(X = factor_heart2, ncp = 7, quali.sup = c(14), excl = NULL,
##       graph = FALSE, level.ventil = 0, axes = c(1, 2), row.w = NULL,
##       method = "Indicator", na.method = "NA", tab.disj = NULL)
##
##
## Eigenvalues
```

```
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               0.264   0.142   0.134   0.129   0.126   0.119
## % of var.              7.805   4.192   3.961   3.803   3.736   3.513
## Cumulative % of var.   7.805  11.997  15.958  19.761  23.496  27.009
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.110   0.108   0.106   0.101   0.097   0.096
## % of var.              3.263   3.188   3.131   2.981   2.869   2.844
## Cumulative % of var.  30.273  33.460  36.592  39.573  42.442  45.286
##                       Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.093   0.090   0.088   0.085   0.081   0.079
## % of var.              2.759   2.661   2.587   2.521   2.390   2.332
## Cumulative % of var.  48.045  50.706  53.293  55.814  58.204  60.536
##                       Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24
## Variance               0.079   0.077   0.075   0.072   0.070   0.068
## % of var.              2.322   2.283   2.213   2.140   2.079   1.997
## Cumulative % of var.  62.858  65.141  67.355  69.495  71.574  73.571
##                       Dim.25  Dim.26  Dim.27  Dim.28  Dim.29  Dim.30
## Variance               0.066   0.064   0.063   0.060   0.057   0.055
## % of var.              1.950   1.891   1.850   1.782   1.693   1.635
## Cumulative % of var.  75.521  77.412  79.262  81.043  82.736  84.371
##                       Dim.31  Dim.32  Dim.33  Dim.34  Dim.35  Dim.36
## Variance               0.054   0.052   0.050   0.046   0.044   0.040
## % of var.              1.596   1.543   1.491   1.364   1.291   1.192
## Cumulative % of var.  85.968  87.511  89.002  90.366  91.657  92.849
##                       Dim.37  Dim.38  Dim.39  Dim.40  Dim.41  Dim.42
## Variance               0.039   0.037   0.034   0.031   0.029   0.026
## % of var.              1.138   1.085   1.008   0.916   0.852   0.771
## Cumulative % of var.  93.987  95.072  96.080  96.996  97.848  98.619
##                       Dim.43  Dim.44
## Variance               0.024   0.023
## % of var.              0.712   0.669
## Cumulative % of var.  99.331 100.000
##
## Individuals (the 10 first)
##                 Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## 1             |  0.527  0.347  0.055 |  0.515  0.618  0.052 |  0.515  0.652
## 2             | -0.428  0.229  0.039 |  0.110  0.028  0.003 |  0.790  1.537
## 3             | -0.681  0.579  0.257 |  0.057  0.008  0.002 | -0.091  0.020
## 4             | -0.718  0.645  0.376 | -0.104  0.025  0.008 |  0.229  0.129
## 5             | -0.262  0.086  0.043 |  0.111  0.029  0.008 | -0.379  0.354
## 6             |  0.229  0.065  0.020 | -0.107  0.027  0.004 |  0.451  0.500
## 7             | -0.206  0.053  0.026 |  0.191  0.085  0.022 | -0.453  0.505
## 8             | -0.656  0.538  0.274 | -0.209  0.101  0.028 |  0.359  0.318
## 9             | -0.194  0.047  0.014 |  0.300  0.210  0.032 |  0.204  0.102
## 10            | -0.484  0.293  0.126 | -0.003  0.000  0.000 |  0.079  0.015
##                cos2
## 1             0.052 |
## 2             0.132 |
## 3             0.005 |
## 4             0.038 |
## 5             0.090 |
## 6             0.076 |
## 7             0.124 |
## 8             0.082 |
```

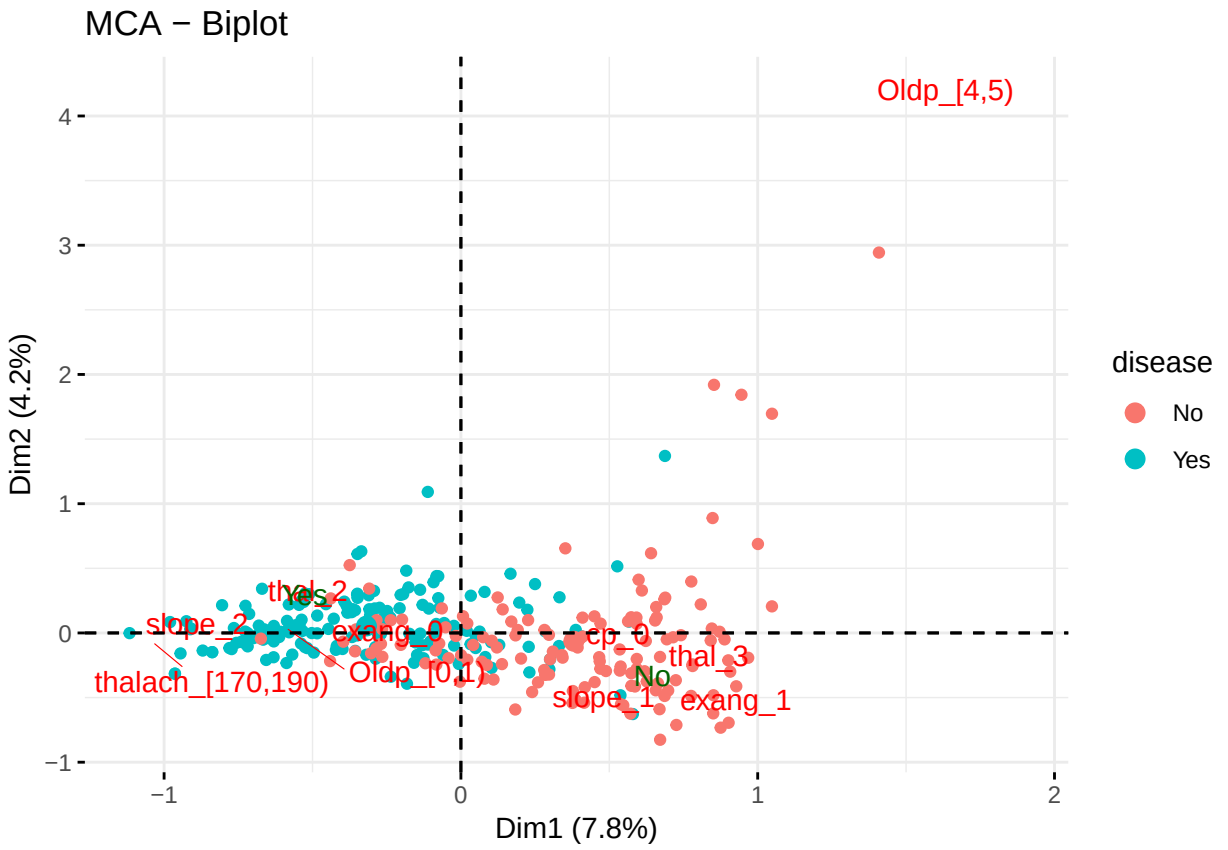
```
## 9            0.015 |
## 10           0.003 |
##
## Categories (the 10 first)
##                Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test
## Age_[20,30) | -1.800  0.311  0.011 -1.800 |  0.241  0.010  0.000  0.241 |
## Age_[30,40) | -1.140  1.873  0.068 -4.521 | -0.337  0.305  0.006 -1.338 |
## Age_[40,50) | -0.641  2.839  0.128 -6.214 | -0.230  0.679  0.016 -2.228 |
## Age_[50,60) |  0.158  0.299  0.017  2.298 | -0.011  0.003  0.000 -0.165 |
## Age_[60,70) |  0.532  2.178  0.102  5.540 |  0.246  0.865  0.022  2.559 |
## Age_[70,80) |  0.271  0.071  0.003  0.870 |  0.310  0.172  0.003  0.996 |
## sex_0       | -0.294  0.798  0.040 -3.481 |  0.588  5.932  0.160  6.955 |
## sex_1       |  0.136  0.370  0.040  3.481 | -0.273  2.751  0.160 -6.955 |
## cp_0        |  0.638  5.586  0.363 10.474 | -0.164  0.687  0.024 -2.692 |
## cp_1        | -0.957  4.396  0.181 -7.390 |  0.035  0.011  0.000  0.271 |
##              Dim.3    ctr   cos2 v.test
## Age_[20,30)  4.476  3.794  0.066  4.476 |
## Age_[30,40)  1.347  5.156  0.095  5.344 |
## Age_[40,50)  0.305  1.271  0.029  2.962 |
## Age_[50,60)  0.059  0.081  0.002  0.854 |
## Age_[60,70) -0.618  5.780  0.137 -6.429 |
## Age_[70,80) -0.459  0.398  0.007 -1.472 |
## sex_0       -0.721  9.462  0.241 -8.538 |
## sex_1        0.335  4.388  0.241  8.538 |
## cp_0        -0.031  0.026  0.001 -0.505 |
## cp_1         0.247  0.579  0.012  1.911 |
##
## Categorical variables (eta2)
##               Dim.1 Dim.2 Dim.3
## age         | 0.260 0.038 0.287 |
## sex         | 0.040 0.160 0.241 |
## cp          | 0.421 0.030 0.047 |
## trestbps    | 0.119 0.313 0.068 |
## chol        | 0.014 0.140 0.228 |
## fbs         | 0.018 0.053 0.002 |
## restecg     | 0.096 0.068 0.002 |
## thalach     | 0.519 0.119 0.314 |
## exang       | 0.336 0.061 0.001 |
## oldpeak     | 0.507 0.476 0.192 |
##
## Supplementary categories
##                 Dim.1    cos2  v.test     Dim.2    cos2  v.test     Dim.3
## No          |   0.740   0.458  11.767 |  -0.170   0.024  -2.694 |   0.048
## Yes         |  -0.619   0.458 -11.767 |   0.142   0.024   2.694 |  -0.040
##                cos2  v.test
## No            0.002   0.766 |
## Yes           0.002  -0.766 |
##
## Supplementary categorical variables (eta2)
##               Dim.1 Dim.2 Dim.3
## disease     | 0.458 0.024 0.002 |
```

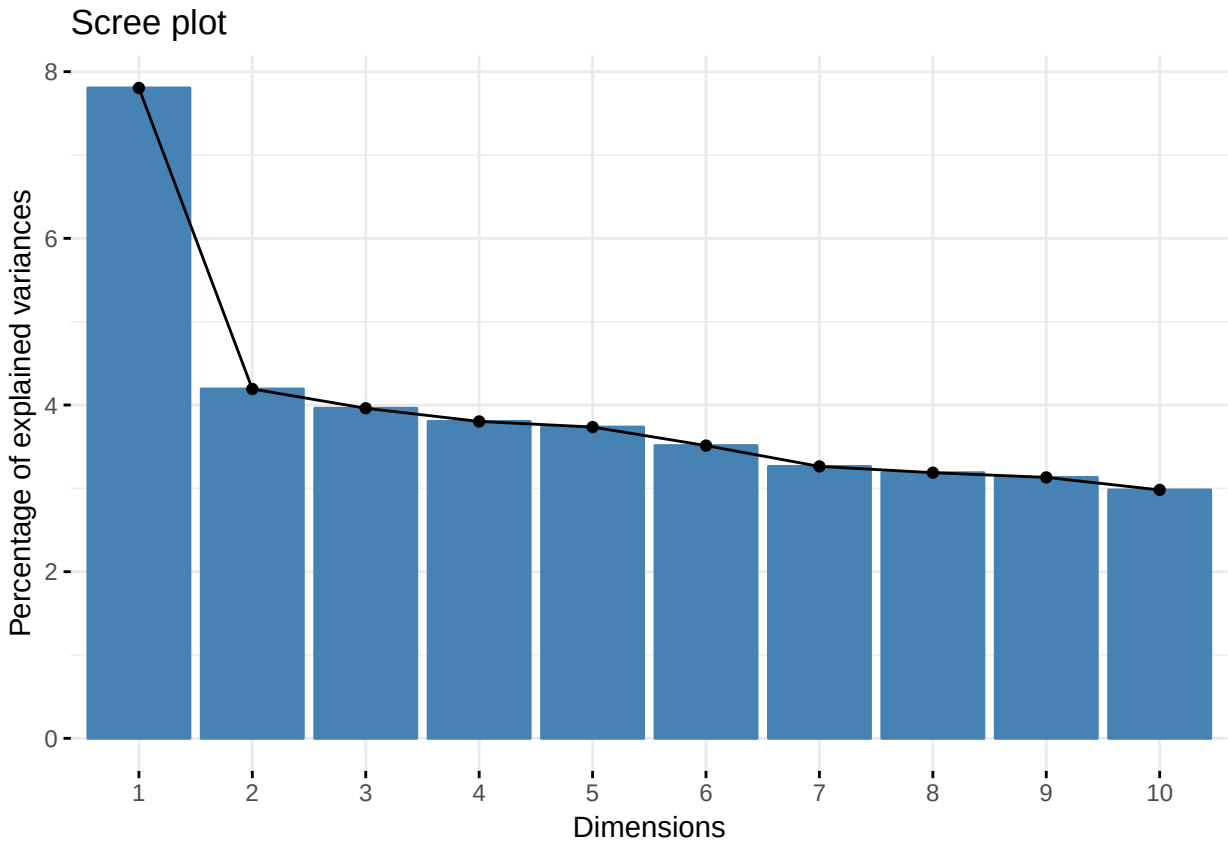

Individuals − MCA
Dim2 (4.2%)
Dim1 (7.8%)
disease
No
Yes

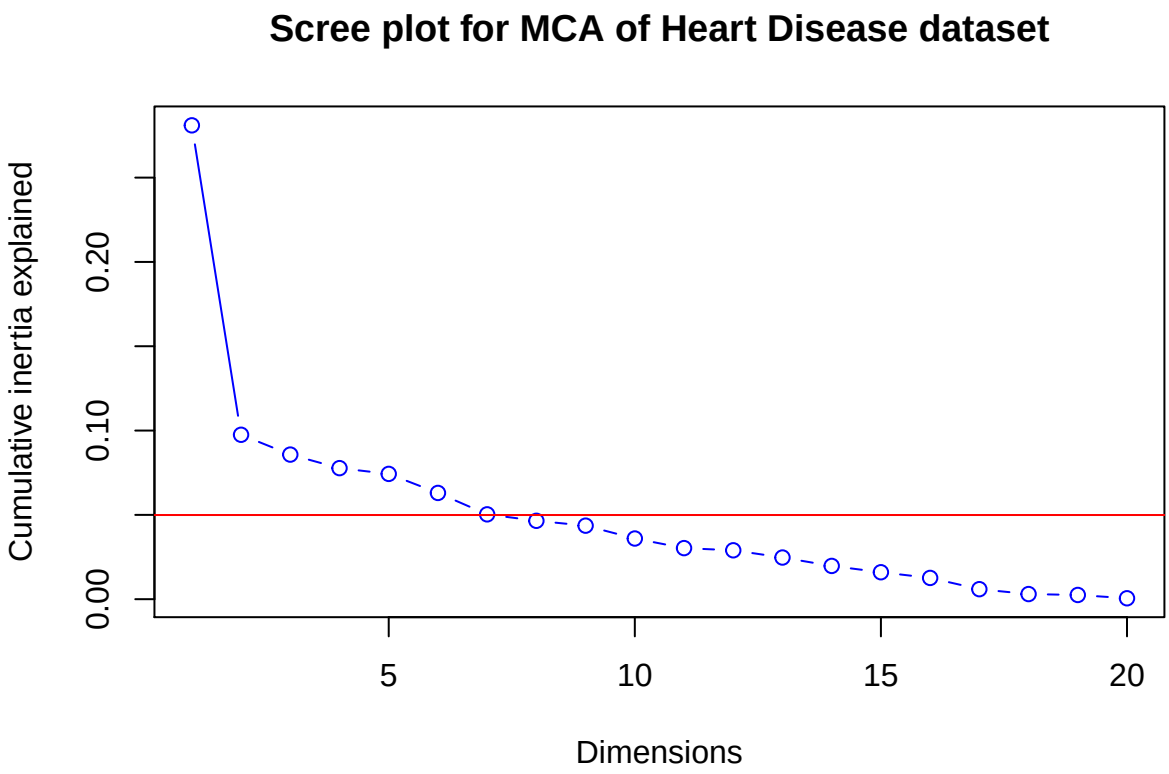

Categories − MCA
Dim2 (4.2%)
Dim1 (7.8%)
thres_[200,220)
Oldp_[4,5)
slope_0
thal_2
sex_0
slope_2
exang_0
Yes
ca_0
thalach_[170,190)
Oldp_[0,1)
cp_0
thal_3
No
slope_1
exang_1
Oldp_[2,3)

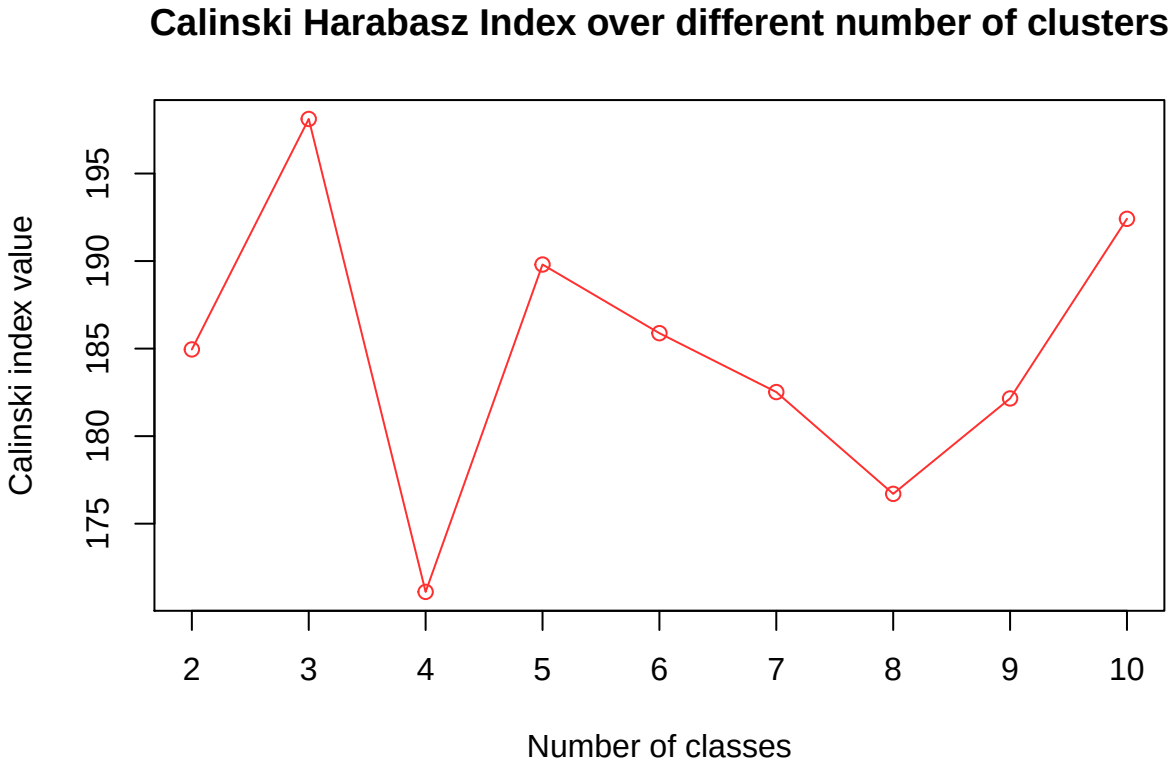


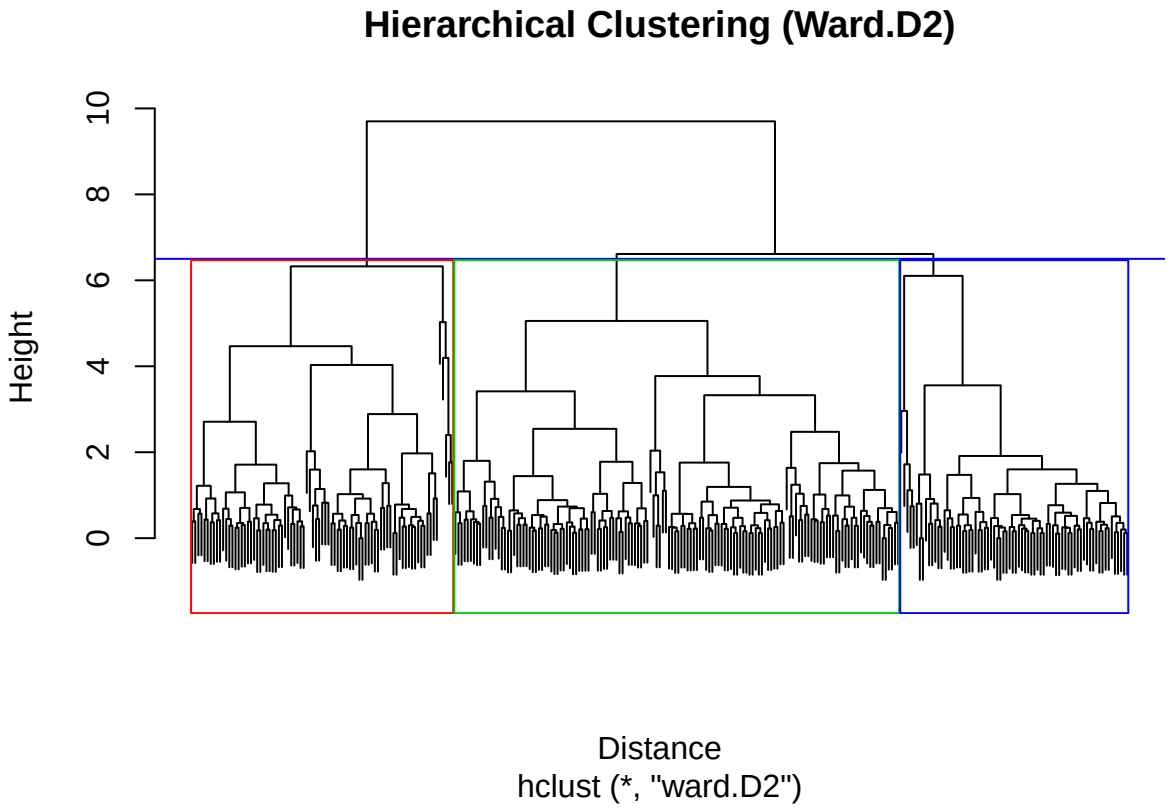

```
##
## Attaching package: 'data.table'
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```


MCA − Biplot
Oldp_[4,5)
thal_2
Yes
slope_2
exang_0
cp_0
thal_3
thalach_[170,190)
Oldp_[0,1)
No
slope_1
exang_1
Dim2 (4.2%)
Dim1 (7.8%)
disease
No
Yes


Scree plot
Percentage of explained variances
Dimensions


Scree plot for MCA of Heart Disease dataset
Cumulative inertia explained
0.00
0.10
0.20
5
10
15
20
Dimensions


Calinski Harabasz Index over different number of clusters
Calinski index value
175
180
185
190
195
2
3
4
5
6
7
8
9
10
Number of classes

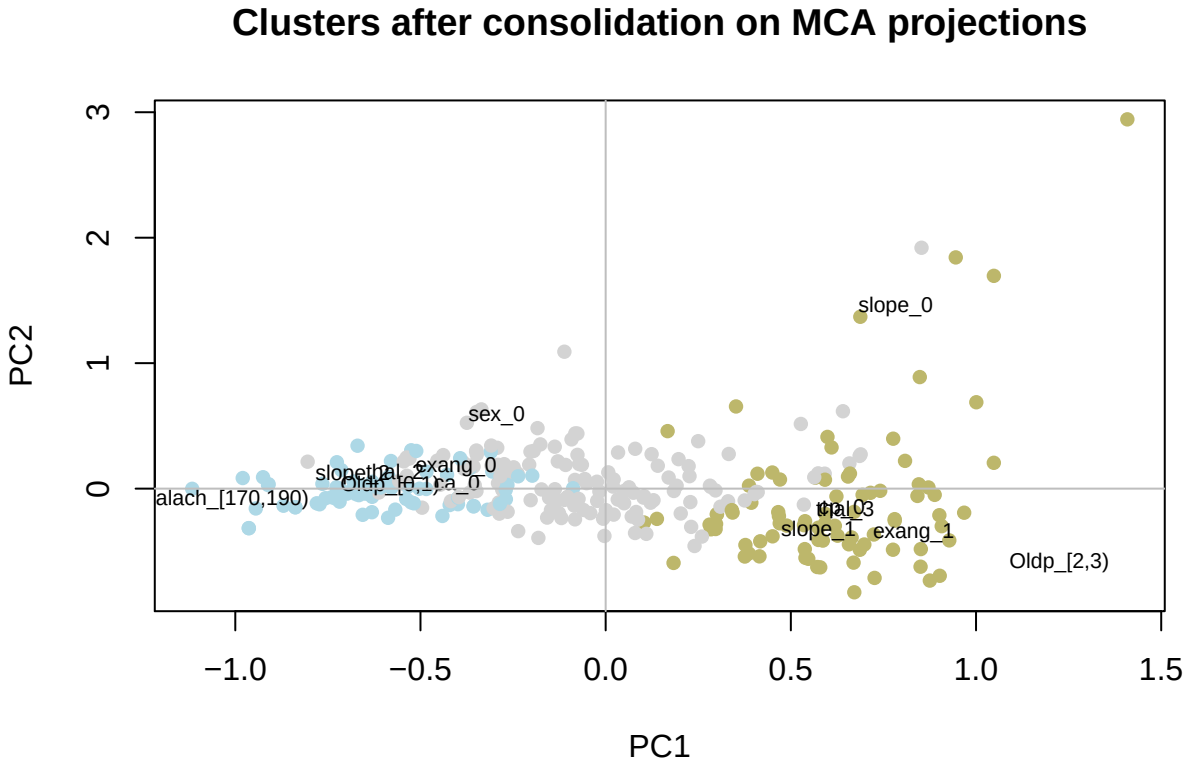



```
## Warning: package 'plotrix' was built under R version 3.5.2
```

```
##
## Attaching package: 'plotrix'
```

```
## The following object is masked from 'package:gplots':
##
##     plotCI
```

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




```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## ================================================================================
##               p.value df
## thalach  1.932712e-47 12
## oldpeak  7.962991e-23 12
## slope    8.435305e-21  4
## exang    6.027553e-20  2
## age      8.573237e-20 10
## cp       1.776157e-19  6
## thal     1.421244e-16  6
## ca       3.911596e-13  8
## restecg  1.111592e-06  4
## sex      4.784081e-06  2
## chol     1.204882e-03  8
## trestbps 5.918866e-03 12
##
## Description of each cluster by the categories
## =============================================
## $`1`
##                               Cla/Mod    Mod/Cla    Global      p.value
## thalach=thalach_[150,170) 73.148148 71.1711712 35.643564 7.033818e-23
## sex=0                     56.250000 48.6486486 31.683168 2.026298e-06
## thal=2                    45.783133 68.4684685 54.785479 2.700666e-04
## exang=0                   42.647059 78.3783784 67.326733 1.669393e-03
## cp=2                      50.574713 39.6396396 28.712871 1.677088e-03
## age=Age_[50,60)           46.400000 52.2522523 41.254125 3.389516e-03
## restecg=0                 44.897959 59.4594595 48.514851 3.934508e-03
## ca=3                      10.000000  1.8018018  6.600660 7.771682e-03
## thal=3                    27.350427 28.8288288 38.613861 7.760270e-03
```

```
## cp=0                       28.671329 36.9369369 47.194719 6.718614e-03
## thal=1                      5.555556  0.9009009  5.940594 2.677075e-03
## exang=1                    24.242424 21.6216216 32.673267 1.669393e-03
## thalach=thalach_[110,130) 14.634146  5.4054054 13.531353 1.107965e-03
## age=Age_[30,40)            0.000000  0.0000000  4.950495 8.652009e-04
## thalach=thalach_[130,150) 19.178082 12.6126126 24.092409 2.786118e-04
## chol=Col_[100,200)        12.000000  5.4054054 16.501650 3.168007e-05
## thalach=thalach_[170,190) 13.793103  7.2072072 19.141914 2.790836e-05
## oldpeak=Oldp_[2,3)         5.882353  1.8018018 11.221122 1.787026e-05
## age=Age_[40,50)           15.277778  9.9099099 23.762376 7.998732e-06
## sex=1                     27.536232 51.3513514 68.316832 2.026298e-06
##                              v.test
## thalach=thalach_[150,170)  9.847403
## sex=0                      4.750784
## thal=2                     3.642439
## exang=0                    3.143502
## cp=2                       3.142156
## age=Age_[50,60)            2.930010
## restecg=0                  2.883365
## ca=3                      -2.661831
## thal=3                    -2.662326
## cp=0                      -2.710479
## thal=1                    -3.002573
## exang=1                   -3.143502
## thalach=thalach_[110,130) -3.261571
## age=Age_[30,40)           -3.331047
## thalach=thalach_[130,150) -3.634415
## chol=Col_[100,200)        -4.161041
## thalach=thalach_[170,190) -4.189895
## oldpeak=Oldp_[2,3)        -4.289963
## age=Age_[40,50)           -4.465218
## sex=1                     -4.750784
##
## $`2`
##                              Cla/Mod   Mod/Cla    Global      p.value
## thalach=thalach_[170,190) 84.482759 55.056180 19.141914 7.495100e-23
## oldpeak=Oldp_[0,1)        47.590361 88.764045 54.785479 1.094766e-15
## age=Age_[40,50)           68.055556 55.056180 23.762376 2.364851e-15
## slope=2                   50.000000 79.775281 46.864686 6.186719e-14
## exang=0                   40.686275 93.258427 67.326733 2.355934e-11
## ca=0                      43.428571 85.393258 57.755776 7.602943e-11
## cp=1                      64.000000 35.955056 16.501650 2.619648e-08
## thal=2                    41.566265 77.528090 54.785479 1.913940e-07
## restecg=1                 42.763158 73.033708 50.165017 2.408284e-07
## age=Age_[30,40)           86.666667 14.606742  4.950495 3.950633e-06
## chol=Col_[100,200)        50.000000 28.089888 16.501650 8.135919e-04
## trestbps=thres_[160,180)   4.761905  1.123596  6.930693 5.776580e-03
## thalach=thalach_[90,110)   0.000000  0.000000  5.280528 3.227860e-03
## thalach=thalach_[130,150) 15.068493 12.359551 24.092409 1.519712e-03
## ca=2                       7.894737  3.370787 12.541254 8.782061e-04
## thalach=thalach_[110,130)  4.878049  2.247191 13.531353 4.626749e-05
## age=Age_[50,60)           16.800000 23.595506 41.254125 4.527240e-05
## thal=3                    15.384615 20.224719 38.613861 1.533432e-05
## oldpeak=Oldp_[2,3)         0.000000  0.000000 11.221122 3.143779e-06
```

```
## restecg=0                 16.326531 26.966292 48.514851 1.105417e-06
## oldpeak=Oldp_[1,2)         8.974359  7.865169 25.742574 1.064415e-06
## ca=1                       6.153846  4.494382 21.452145 4.485823e-07
## age=Age_[60,70)            5.000000  4.494382 26.402640 1.008709e-09
## cp=0                      11.888112 19.101124 47.194719 1.103577e-10
## exang=1                    6.060606  6.741573 32.673267 2.355934e-11
## slope=1                   10.000000 15.730337 46.204620 1.543010e-12
##                             v.test
## thalach=thalach_[170,190)  9.841015
## oldpeak=Oldp_[0,1)         8.015739
## age=Age_[40,50)            7.920538
## slope=2                    7.504068
## exang=0                    6.682068
## ca=0                       6.508254
## cp=1                       5.565125
## thal=2                     5.207508
## restecg=1                  5.164697
## age=Age_[30,40)            4.613963
## chol=Col_[100,200)         3.348129
## trestbps=thres_[160,180)  -2.760201
## thalach=thalach_[90,110)  -2.945162
## thalach=thalach_[130,150) -3.170893
## ca=2                      -3.326891
## thalach=thalach_[110,130) -4.073723
## age=Age_[50,60)           -4.078781
## thal=3                    -4.323830
## oldpeak=Oldp_[2,3)        -4.661195
## restecg=0                 -4.871879
## oldpeak=Oldp_[1,2)        -4.879340
## ca=1                      -5.047093
## age=Age_[60,70)           -6.108026
## cp=0                      -6.452036
## exang=1                   -6.682068
## slope=1                   -7.070568
##
## $`3`
##                             Cla/Mod    Mod/Cla    Global      p.value
## exang=1                   69.696970 66.9902913 32.673267 1.396795e-19
## cp=0                      59.440559 82.5242718 47.194719 1.744311e-19
## slope=1                   57.142857 77.6699029 46.204620 1.533863e-15
## oldpeak=Oldp_[2,3)        94.117647 31.0679612 11.221122 7.942302e-15
## thal=3                    57.264957 65.0485437 38.613861 1.696461e-11
## thalach=thalach_[110,130) 80.487805 32.0388350 13.531353 6.453248e-11
## thalach=thalach_[130,150) 65.753425 46.6019417 24.092409 1.604492e-10
## thalach=thalach_[90,110)  87.500000 13.5922330  5.280528 9.533717e-06
## thal=1                    83.333333 14.5631068  5.940594 1.457395e-05
## ca=3                      80.000000 15.5339806  6.600660 1.929255e-05
## sex=1                     40.579710 81.5533981 68.316832 3.031969e-04
## age=Age_[60,70)           47.500000 36.8932039 26.402640 3.615991e-03
## ca=1                      49.230769 31.0679612 21.452145 4.356482e-03
## slope=0                   61.904762 12.6213592  6.930693 7.929604e-03
## trestbps=thres_[160,180)  61.904762 12.6213592  6.930693 7.929604e-03
## oldpeak=Oldp_[3,4)        64.705882 10.6796117  5.610561 9.468838e-03
## sex=0                     19.791667 18.4466019 31.683168 3.031969e-04
```

```
## age=Age_[40,50)           16.666667 11.6504854 23.762376 2.527795e-04
## cp=1                       4.000000  1.9417476 16.501650 5.344978e-08
## cp=2                      10.344828  8.7378641 28.712871 5.914839e-09
## ca=0                      19.428571 33.0097087 57.755776 4.518008e-10
## thalach=thalach_[170,190)  1.724138  0.9708738 19.141914 5.236652e-11
## thalach=thalach_[150,170)  5.555556  5.8252427 35.643564 5.686563e-17
## thal=2                    12.650602 20.3883495 54.785479 2.036748e-18
## exang=0                   16.666667 33.0097087 67.326733 1.396795e-19
## oldpeak=Oldp_[0,1)        10.843373 17.4757282 54.785479 1.300306e-21
## slope=2                    7.042254  9.7087379 46.864686 1.814133e-22
##                              v.test
## exang=1                    9.052541
## cp=0                       9.028255
## slope=1                    7.974188
## oldpeak=Oldp_[2,3)         7.768497
## thal=3                     6.730015
## thalach=thalach_[110,130)  6.532844
## thalach=thalach_[130,150)  6.395093
## thalach=thalach_[90,110)   4.427487
## thal=1                     4.335031
## ca=3                       4.272925
## sex=1                      3.612553
## age=Age_[60,70)            2.909853
## ca=1                       2.851125
## slope=0                    2.655053
## trestbps=thres_[160,180)   2.655053
## oldpeak=Oldp_[3,4)         2.594646
## sex=0                     -3.612553
## age=Age_[40,50)           -3.659427
## cp=1                      -5.439435
## cp=2                      -5.819148
## ca=0                      -6.234992
## thalach=thalach_[170,190) -6.564047
## thalach=thalach_[150,170) -8.371549
## thal=2                    -8.755237
## exang=0                   -9.052541
## oldpeak=Oldp_[0,1)        -9.549733
## slope=2                   -9.751695
```

	Cla/Mod	Mod/Cla	Global	p.value	v.test
thalach=thalach_[150,170)	73.148148	71.1711712	35.643564	0.0000000	9.847403
sex=0	56.250000	48.6486486	31.683168	0.0000020	4.750784
thal=2	45.783133	68.4684685	54.785479	0.0002701	3.642439
exang=0	42.647059	78.3783784	67.326733	0.0016694	3.143502
cp=2	50.574713	39.6396396	28.712871	0.0016771	3.142156
age=Age_[50,60)	46.400000	52.2522523	41.254125	0.0033895	2.930010
restecg=0	44.897959	59.4594595	48.514851	0.0039345	2.883365
ca=3	10.000000	1.8018018	6.600660	0.0077717	-2.661831
thal=3	27.350427	28.8288288	38.613861	0.0077603	-2.662326
cp=0	28.671329	36.9369369	47.194719	0.0067186	-2.710479
thal=1	5.555556	0.9009009	5.940594	0.0026771	-3.002573
exang=1	24.242424	21.6216216	32.673267	0.0016694	-3.143502
thalach=thalach_[110,130)	14.634146	5.4054054	13.531353	0.0011080	-3.261571
age=Age_[30,40)	0.000000	0.0000000	4.950495	0.0008652	-3.331047

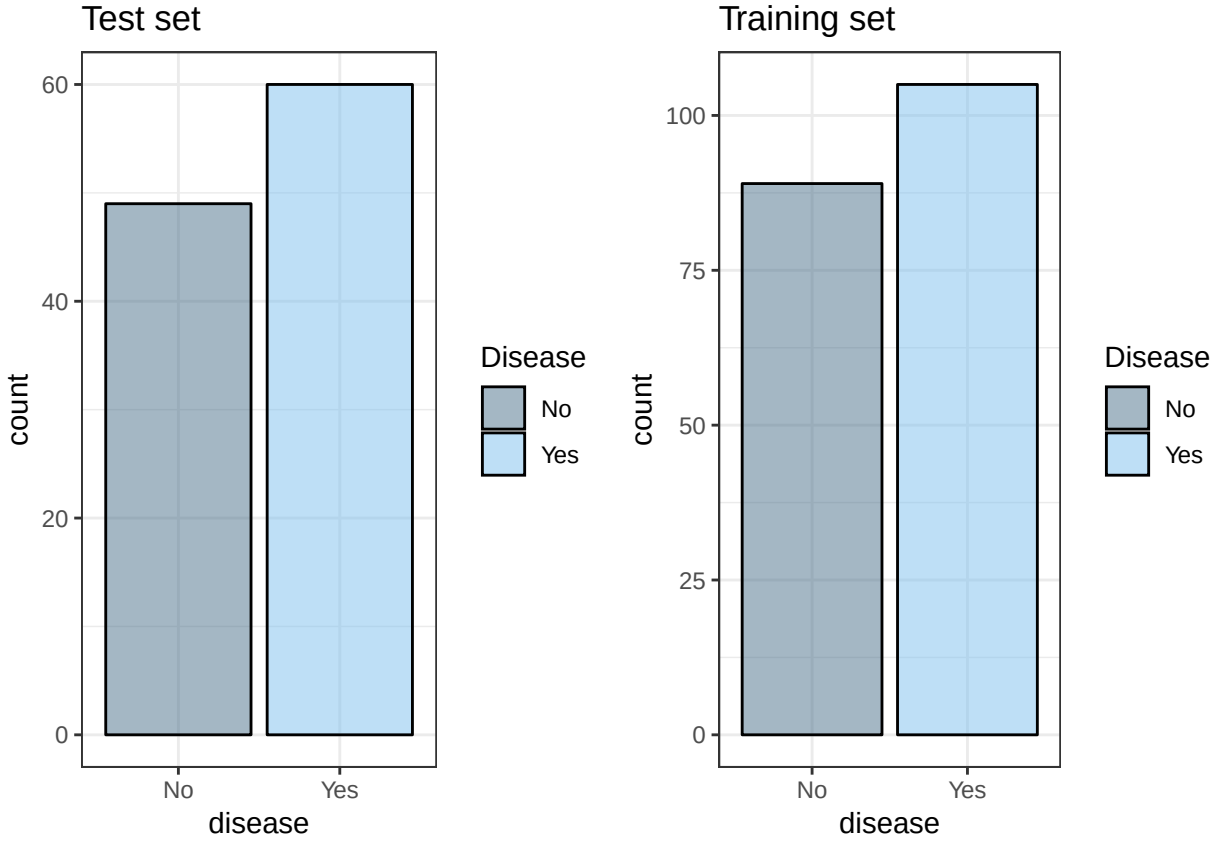
	Cla/Mod	Mod/Cla	Global	p.value	v.test
thalach=thalach_[130,150)	19.178082	12.6126126	24.092409	0.0002786	-3.634415
chol=Col_[100,200)	12.000000	5.4054054	16.501650	0.0000317	-4.161041
thalach=thalach_[170,190)	13.793103	7.2072072	19.141914	0.0000279	-4.189895
oldpeak=Oldp_[2,3)	5.882353	1.8018018	11.221122	0.0000179	-4.289963
age=Age_[40,50)	15.277778	9.9099099	23.762376	0.0000080	-4.465218
sex=1	27.536232	51.3513514	68.316832	0.0000020	-4.750784

	Cla/Mod	Mod/Cla	Global	p.value	v.test
thalach=thalach_[170,190)	84.482759	55.056180	19.141914	0.0000000	9.841015
oldpeak=Oldp_[0,1)	47.590361	88.764045	54.785479	0.0000000	8.015739
age=Age_[40,50)	68.055556	55.056180	23.762376	0.0000000	7.920538
slope=2	50.000000	79.775281	46.864686	0.0000000	7.504068
exang=0	40.686275	93.258427	67.326733	0.0000000	6.682068
ca=0	43.428571	85.393258	57.755776	0.0000000	6.508254
cp=1	64.000000	35.955056	16.501650	0.0000000	5.565125
thal=2	41.566265	77.528090	54.785479	0.0000002	5.207508
restecg=1	42.763158	73.033708	50.165017	0.0000002	5.164697
age=Age_[30,40)	86.666667	14.606742	4.950495	0.0000040	4.613963
chol=Col_[100,200)	50.000000	28.089888	16.501650	0.0008136	3.348129
trestbps=thres_[160,180)	4.761905	1.123595	6.930693	0.0057766	-2.760201
thalach=thalach_[90,110)	0.000000	0.000000	5.280528	0.0032279	-2.945162
thalach=thalach_[130,150)	15.068493	12.359551	24.092409	0.0015197	-3.170893
ca=2	7.894737	3.370786	12.541254	0.0008782	-3.326891
thalach=thalach_[110,130)	4.878049	2.247191	13.531353	0.0000463	-4.073723
age=Age_[50,60)	16.800000	23.595506	41.254125	0.0000453	-4.078781
thal=3	15.384615	20.224719	38.613861	0.0000153	-4.323830
oldpeak=Oldp_[2,3)	0.000000	0.000000	11.221122	0.0000031	-4.661195
restecg=0	16.326531	26.966292	48.514851	0.0000011	-4.871879
oldpeak=Oldp_[1,2)	8.974359	7.865169	25.742574	0.0000011	-4.879340
ca=1	6.153846	4.494382	21.452145	0.0000004	-5.047093
age=Age_[60,70)	5.000000	4.494382	26.402640	0.0000000	-6.108026
cp=0	11.888112	19.101124	47.194719	0.0000000	-6.452036
exang=1	6.060606	6.741573	32.673267	0.0000000	-6.682068
slope=1	10.000000	15.730337	46.204620	0.0000000	-7.070568

	Cla/Mod	Mod/Cla	Global	p.value	v.test
exang=1	69.696970	66.9902913	32.673267	0.0000000	9.052541
cp=0	59.440559	82.5242718	47.194719	0.0000000	9.028255
slope=1	57.142857	77.6699029	46.204620	0.0000000	7.974188
oldpeak=Oldp_[2,3)	94.117647	31.0679612	11.221122	0.0000000	7.768497
thal=3	57.264957	65.0485437	38.613861	0.0000000	6.730015
thalach=thalach_[110,130)	80.487805	32.0388350	13.531353	0.0000000	6.532844
thalach=thalach_[130,150)	65.753425	46.6019417	24.092409	0.0000000	6.395093
thalach=thalach_[90,110)	87.500000	13.5922330	5.280528	0.0000095	4.427487
thal=1	83.333333	14.5631068	5.940594	0.0000146	4.335031
ca=3	80.000000	15.5339806	6.600660	0.0000193	4.272925
sex=1	40.579710	81.5533981	68.316832	0.0003032	3.612553
age=Age_[60,70)	47.500000	36.8932039	26.402640	0.0036160	2.909853
ca=1	49.230769	31.0679612	21.452145	0.0043565	2.851125

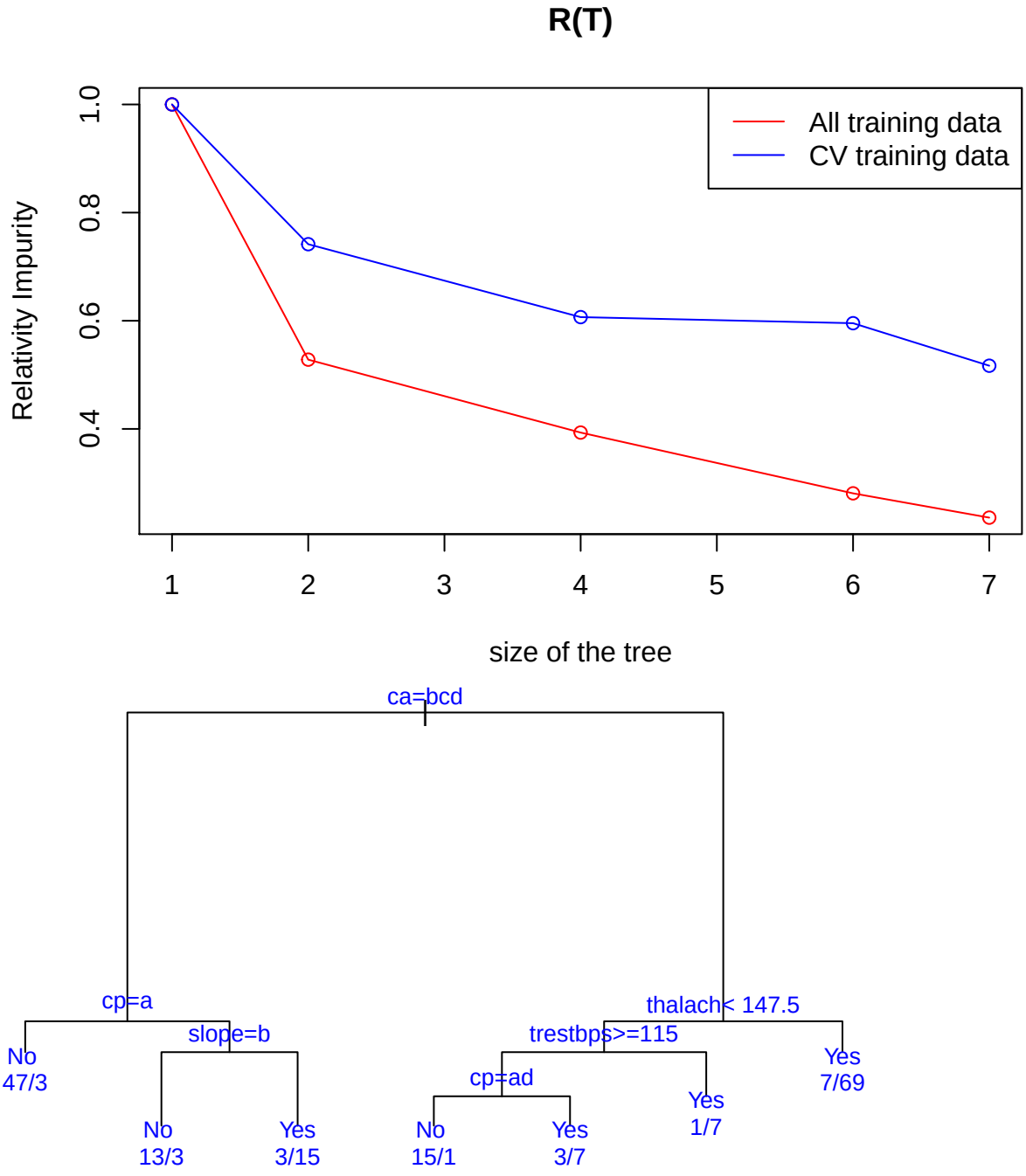
	Cla/Mod	Mod/Cla	Global	p.value	v.test
slope=0	61.904762	12.6213592	6.930693	0.0079296	2.655053
trestbps=thres_[160,180)	61.904762	12.6213592	6.930693	0.0079296	2.655053
oldpeak=Oldp_[3,4)	64.705882	10.6796117	5.610561	0.0094688	2.594646
sex=0	19.791667	18.4466019	31.683168	0.0003032	-3.612553
age=Age_[40,50)	16.666667	11.6504854	23.762376	0.0002528	-3.659427
cp=1	4.000000	1.9417476	16.501650	0.0000001	-5.439435
cp=2	10.344828	8.7378641	28.712871	0.0000000	-5.819148
ca=0	19.428571	33.0097087	57.755776	0.0000000	-6.234992
thalach=thalach_[170,190)	1.724138	0.9708738	19.141914	0.0000000	-6.564048
thalach=thalach_[150,170)	5.555556	5.8252427	35.643564	0.0000000	-8.371549
thal=2	12.650602	20.3883495	54.785479	0.0000000	-8.755237
exang=0	16.666667	33.0097087	67.326733	0.0000000	-9.052541
oldpeak=Oldp_[0,1)	10.843374	17.4757282	54.785479	0.0000000	-9.549733
slope=2	7.042254	9.7087379	46.864686	0.0000000	-9.751695

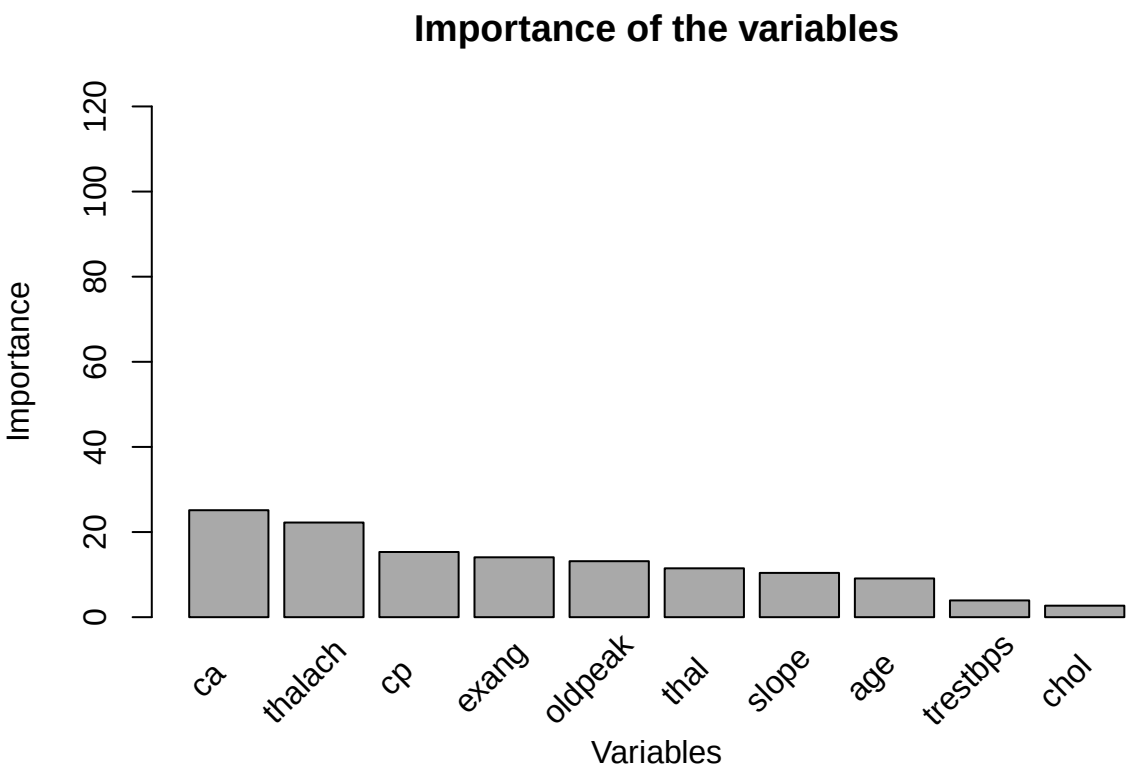
```
## Loading required package: caTools
```

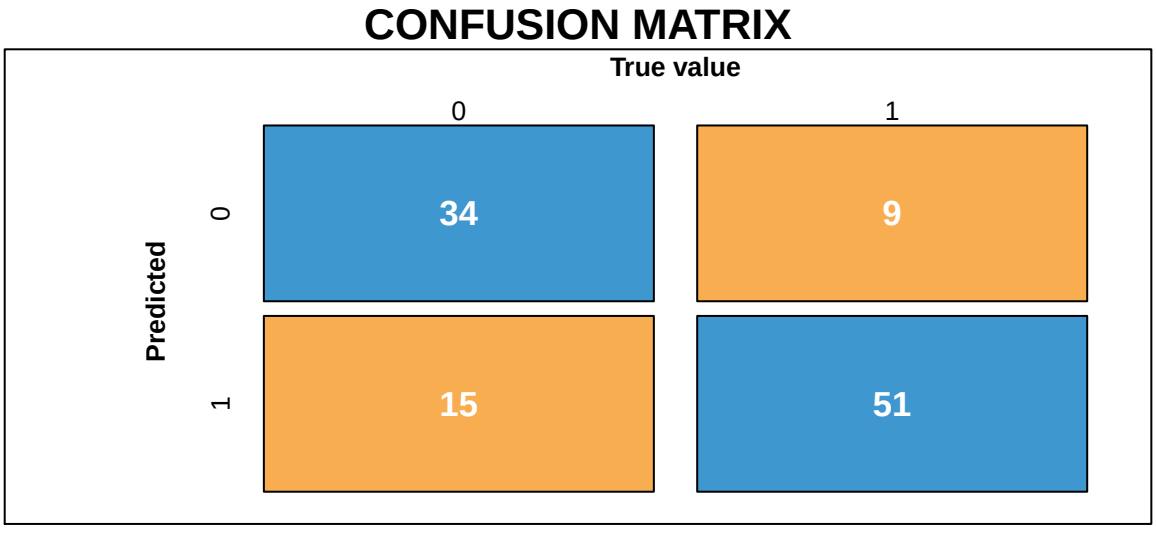



```
##
## Classification tree:
## rpart(formula = disease ~ ., data = train, control = rpart.control(cp = 0.001,
##     xval = 10))
##
## Variables actually used in tree construction:
## [1] ca       cp       slope    thalach  trestbps
##
## Root node error: 89/194 = 0.45876
```

```
##
## n= 194
##
##         CP nsplit rel error  xerror     xstd
## 1 0.471910      0   1.00000 1.00000 0.077983
## 2 0.067416      1   0.52809 0.74157 0.074146
## 3 0.056180      3   0.39326 0.60674 0.070141
## 4 0.044944      5   0.28090 0.59551 0.069736
## 5 0.001000      6   0.23596 0.51685 0.066561
```



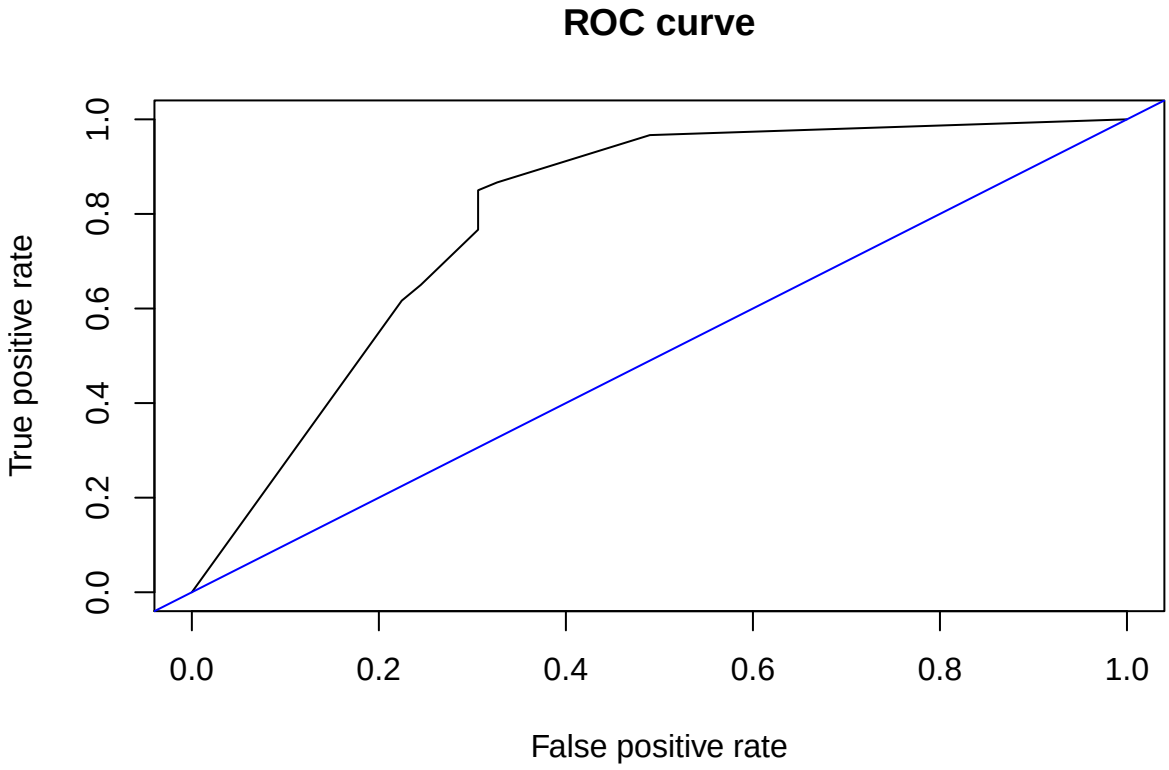




DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.85	0.694	0.773	0.85	0.81

Accuracy	Kappa
0.78	0.55



AUC.value
0.7943878

```
test$type <- 'test'
train$type <- 'train'
test_train <- rbind(test,train)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

```
g0 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = thalach, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'PIE', title = 'contexts')
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

```
g7 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = age, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'Years', title = 'Age') +
  theme(legend.position="none")
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

```
g8 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = thalach, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'Heart Rate BPM', title = 'Thalach') +
  theme(legend.position="none")
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

```
g9 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = oldpeak, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'ST Depression (mV)', title = 'Oldpeak') +
  theme(legend.position="none")
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

```
g10 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = chol, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'mg/dl', title = 'Chol') +
  theme(legend.position="none")
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

```
g11 <- test_train[test_train$type==c('test','train'),] %>%
  ggplot(aes(x = trestbps, fill=type)) +
  #geom_density(alpha = 0.5) +
  geom_histogram()+
  labs(x = 'mm/Hg ', title = 'Trestbps') +
  theme(legend.position="none")
```

```
## Warning in test_train$type == c("test", "train"): longer object length is
## not a multiple of shorter object length
```

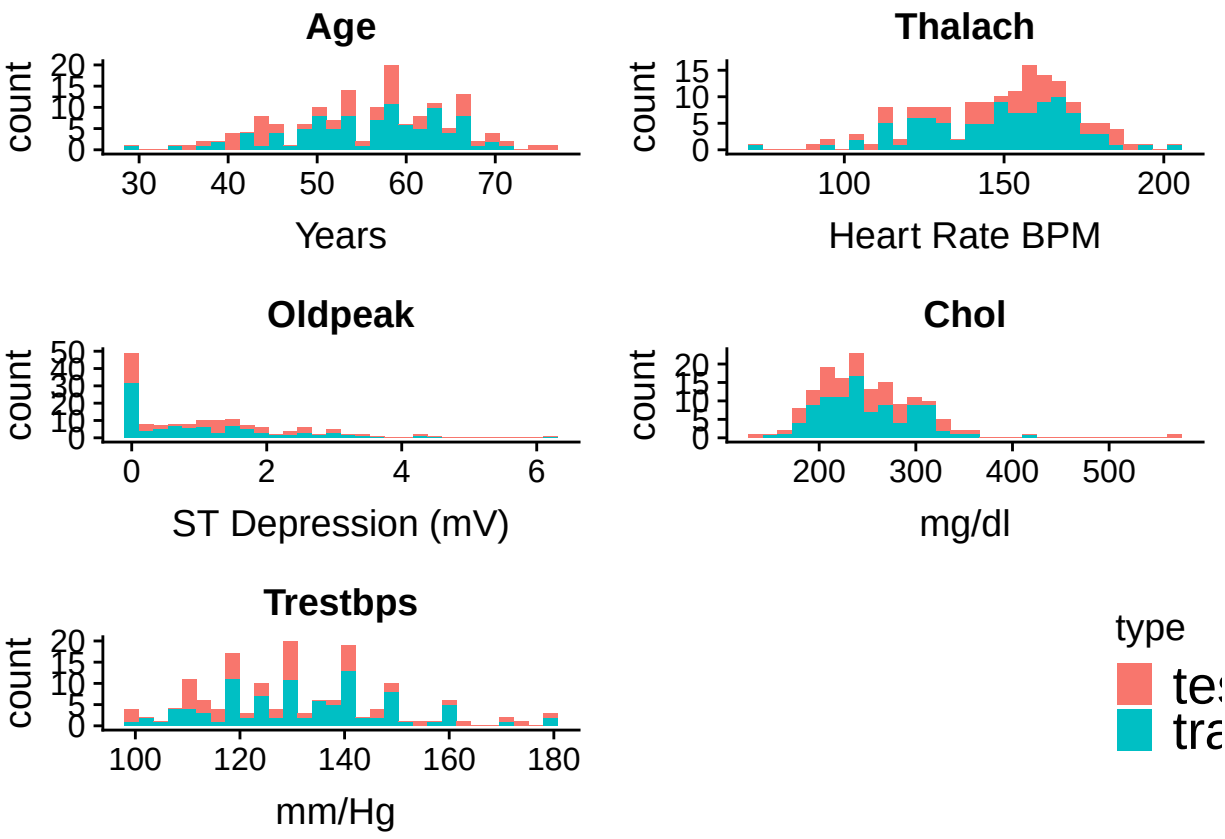
```
legend <- get_legend(g0 + theme(legend.position=c(0.8, 0.6)) + theme(legend.box = "horizontal") + theme
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
grid.arrange(g7, g8, g9, g10, g11, legend,ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Age
count
Years
Thalach
Heart Rate BPM
Oldpeak
ST Depression (mV)
Chol
mg/dl
Trestbps
mm/Hg
type