Dimention Reduction. Analysis of Cardio taraining.

Intruducation

In this project we will analys, the information & dataset is a result of the medical examination, based on real patiant information.

Also, in this project we will use: PCA & MCA & CA

Dataset contains the following information:

Age | Objective Feature | age | int (days)
Height | Objective Feature | height | int (cm) |
Weight | Objective Feature | weight | float (kg) |
Gender | Objective Feature | gender | categorical code |
Systolic blood pressure | Examination Feature | ap_hi | int |
Diastolic blood pressure | Examination Feature | ap_lo | int |
Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
Smoking | Subjective Feature | smoke | binary |
Alcohol intake | Subjective Feature | alco | binary |
Physical activity | Subjective Feature | active | binary |
Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

Data preparation

install.packages("factoextra") install.packages("gridExtra") install.packages("tidyverse") install.packages("ggplot2")	library(factoextra) library(FactoMineR) library(ggplot2) library(dplyr) library(reshape2) library(corrplot) library(gridExtra) library(grid)

Loading data

setwd("D:\\R and R Studio\\Dimension Reduction\\DimensionReduction")

getwd()

data1 <- read.csv("cardio_train.csv", sep = ";")

View(data1)

```
Global Environment
Data
data1                    70000 obs. of 13 variables
  id : int 0 1 2 3 4 8 9 12 13 14 ...
  age : int 18393 20228 18857 17623 17474 21914 22113 22584 17668 …
  gender : int 2 1 1 2 1 1 1 2 1 1 ...
  height : int 168 156 165 169 156 151 157 178 158 164 ...
  weight : num 62 85 64 82 56 67 93 95 71 68 ...
  ap_hi : int 110 140 130 150 100 120 130 130 110 110 ...
  ap_lo : int 80 90 70 100 60 80 80 90 70 60 ...
  cholesterol: int 1 3 3 1 1 2 3 3 1 1 ...
  gluc : int 1 1 1 1 1 2 1 3 1 1 ...
  smoke : int 0 0 0 0 0 0 0 0 0 0 ...
  alco : int 0 0 0 0 0 0 0 0 0 0 ...
  active : int 1 1 0 1 0 0 1 1 1 0 ...
  cardio : int 0 1 1 1 0 0 0 1 0 0 ...
```

Data Analysis

head(data1)

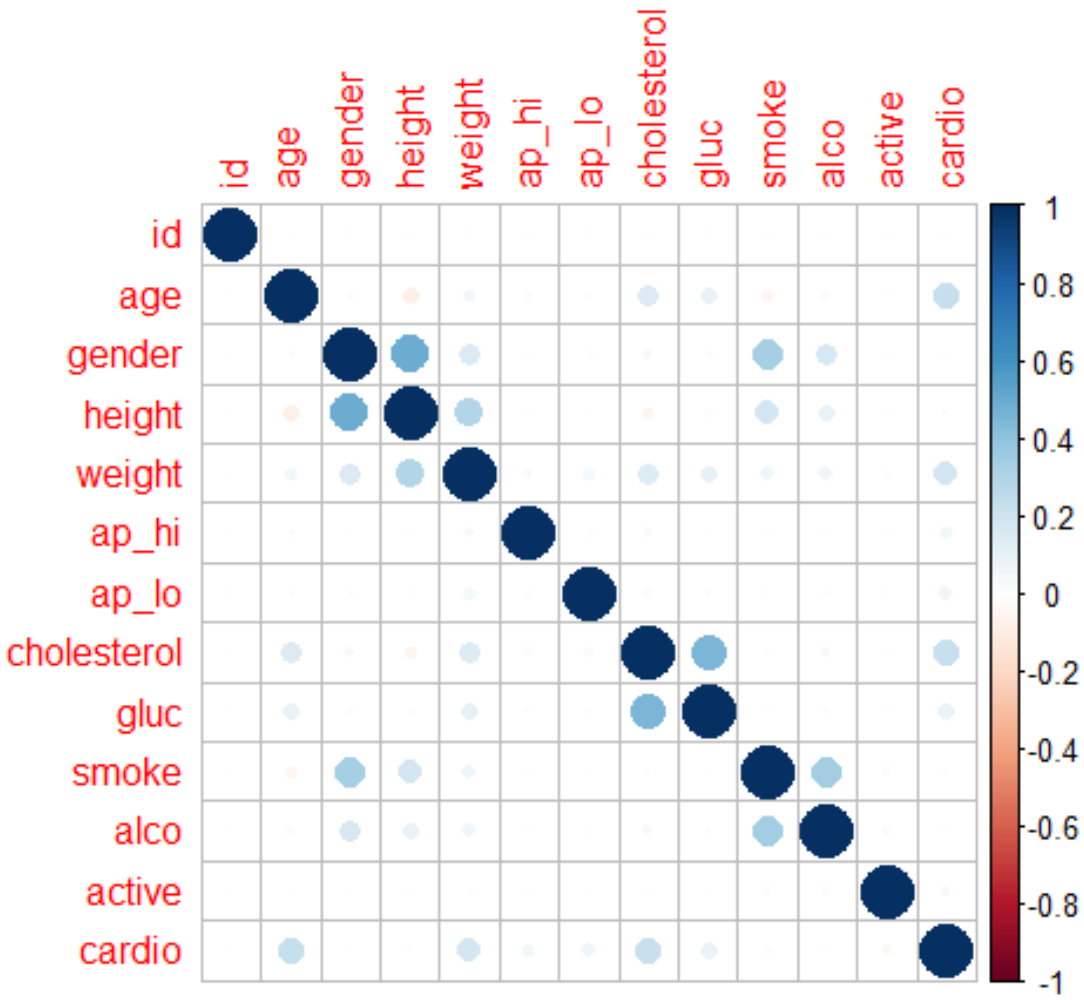
```
> head(data1)
  id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active cardio
1  0 18393      2    168     62   110    80           1    1     0    0      1      0
2  1 20228      1    156     85   140    90           3    1     0    0      1      1
3  2 18857      1    165     64   130    70           3    1     0    0      0      1
4  3 17623      2    169     82   150   100           1    1     0    0      1      1
5  4 17474      1    156     56   100    60           1    1     0    0      0      0
6  8 21914      1    151     67   120    80           2    2     0    0      0      0
```

summary(data1)

```
Console   Terminal   Jobs
D:/R and R Studio/Dimension Reduction/DimensionReduction/
> summary(data1)
       id             age            gender         height          weight
 Min.   :    0   Min.   :10798   Min.   :1.00   Min.   : 55.0   Min.   : 10.00
 1st Qu.:25007   1st Qu.:17664   1st Qu.:1.00   1st Qu.:159.0   1st Qu.: 65.00
 Median :50002   Median :19703   Median :1.00   Median :165.0   Median : 72.00
 Mean   :49972   Mean   :19469   Mean   :1.35   Mean   :164.4   Mean   : 74.21
 3rd Qu.:74889   3rd Qu.:21327   3rd Qu.:2.00   3rd Qu.:170.0   3rd Qu.: 82.00
 Max.   :99999   Max.   :23713   Max.   :2.00   Max.   :250.0   Max.   :200.00
      ap_hi            ap_lo           cholesterol         gluc           smoke
 Min.   : -150.0   Min.   :  -70.00   Min.   :1.000   Min.   :1.000   Min.   :0.00000
 1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.00000
 Median :  120.0   Median :   80.00   Median :1.000   Median :1.000   Median :0.00000
 Mean   :  128.8   Mean   :   96.63   Mean   :1.367   Mean   :1.226   Mean   :0.08813
 3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:0.00000
 Max.   :16020.0   Max.   :11000.00   Max.   :3.000   Max.   :3.000   Max.   :1.00000
      alco             active           cardio
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
 Median :0.00000   Median :1.0000   Median :0.0000
 Mean   :0.05377   Mean   :0.8037   Mean   :0.4997
 3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
```

Undersatnding data viusaly for it we need to show it in correlation.

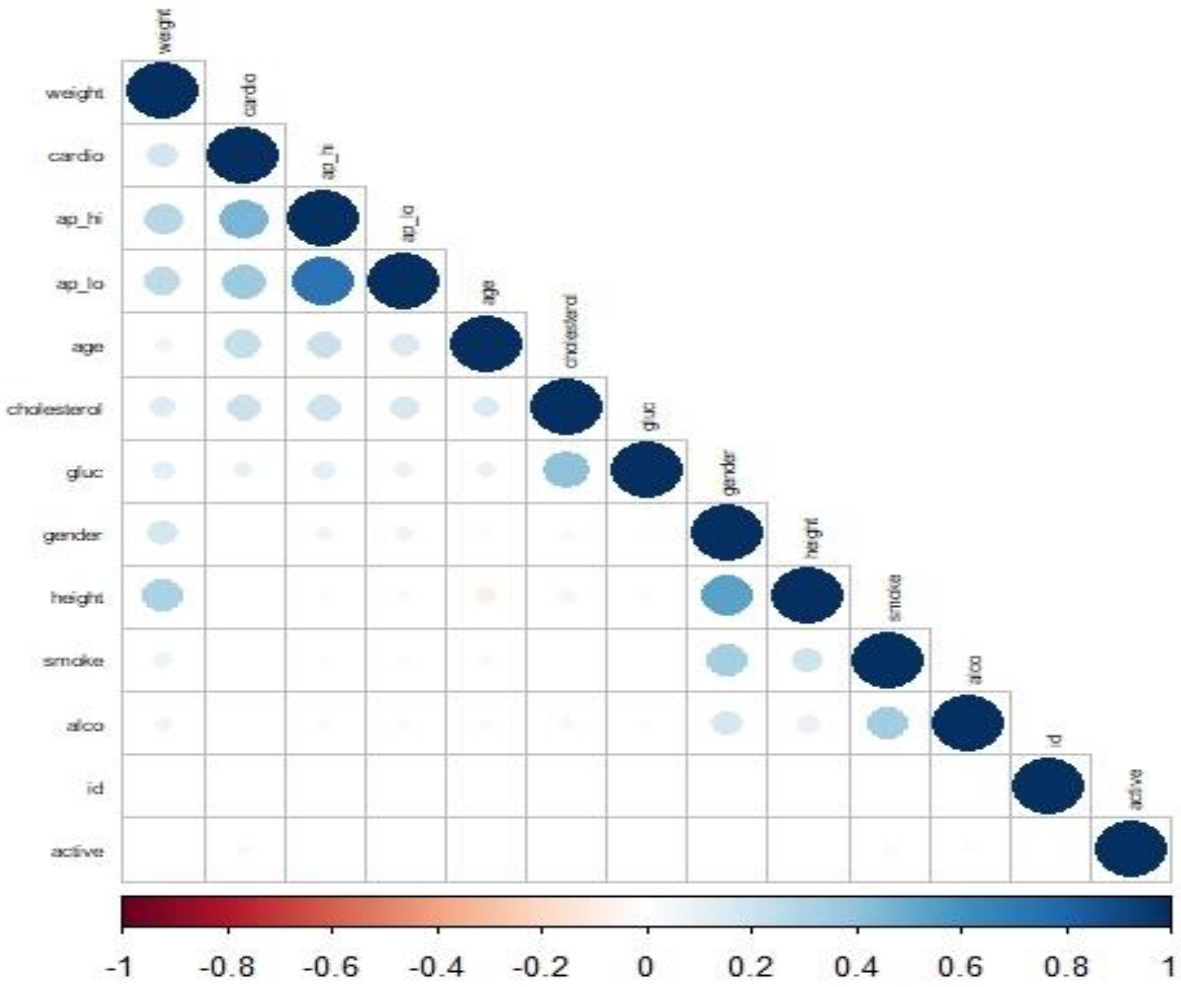
```
corr_data = cor(data1,method='pearson')

corrplot(corr_data)
```



Here, correlation suing matrix Cholesterol.

cor.matrix <- cor(data1, method = c("spearman"))

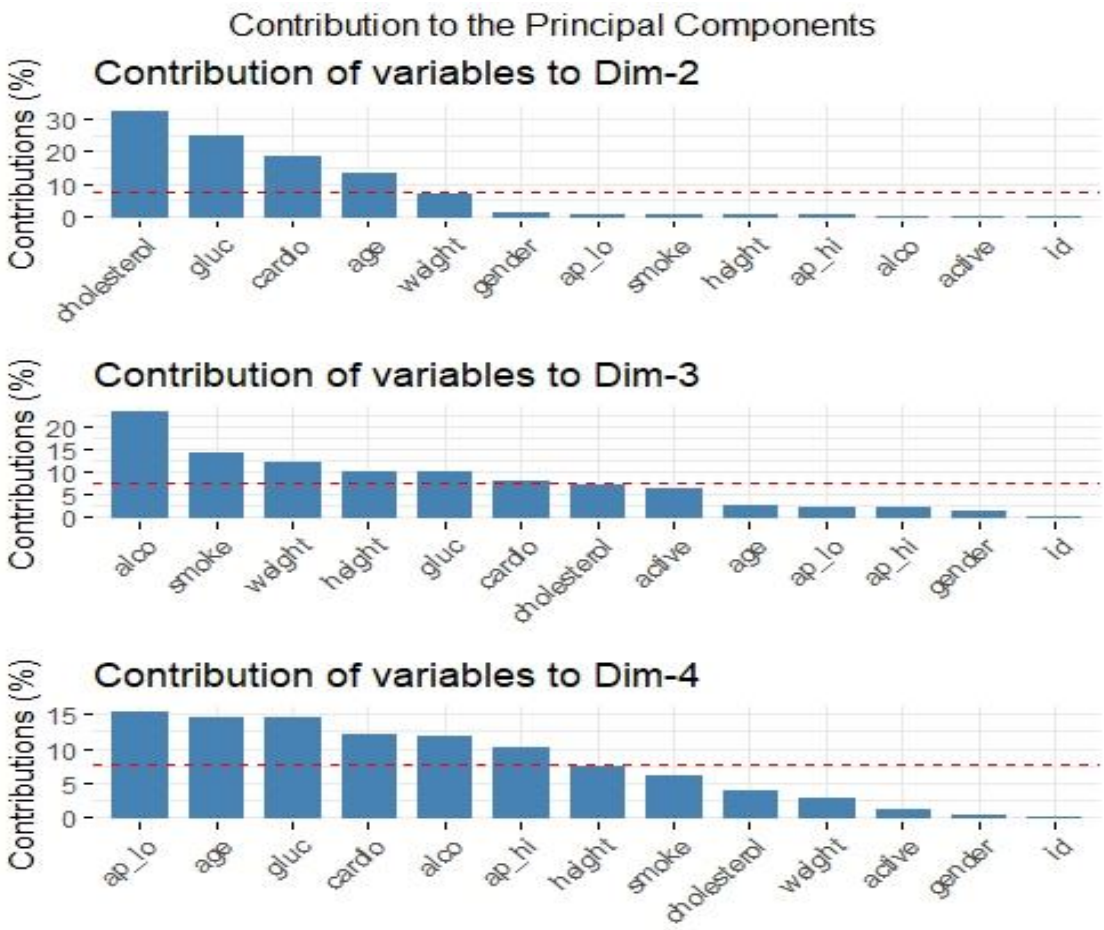
corrplot(cor.matrix, type = "lower", order = "hclust", tl.col = "black", tl.cex = 0.5)



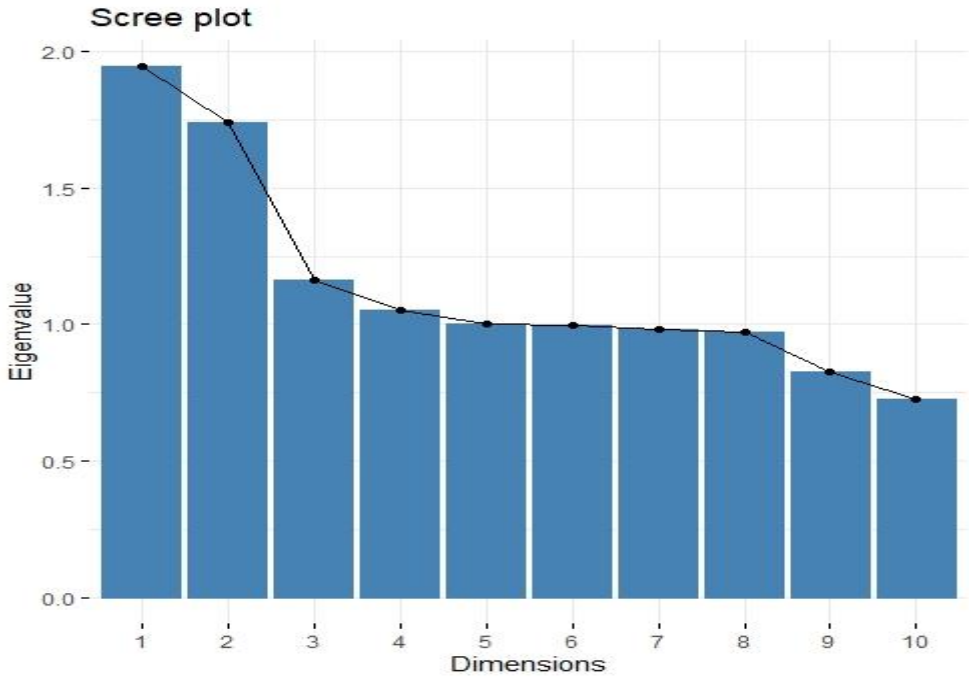
PCA

Choosing number in Component

```
> data.pca <- prcomp(data1, center=TRUE, scale=TRUE)

> eigen(cor(data1))$values

 [1] 1.9430876 1.7384125 1.1616289 1.0544779 1.0021538 0.9977607 0.9841246 0.9731207 0.8270677 0.7235359

[11] 0.6215278 0.5243303 0.4487716

var <- get_pca_var(data.pca)

a<-fviz_contrib(data.pca, "var",axes = 2)

b<-fviz_contrib(data.pca, "var",axes = 3)

c<-fviz_contrib(data.pca, "var",axes = 4)

grid.arrange(a,b,c,top='Contribution to the Principal Components')
```

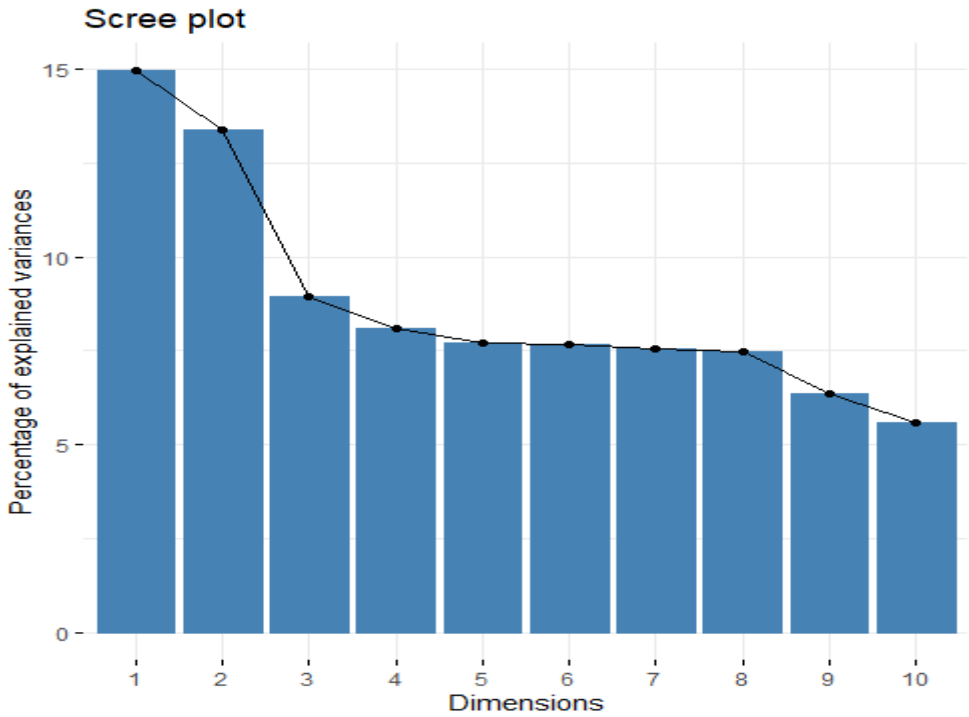


fviz_eig(data.pca, choice='eigenvalue')



Two diffrent contrubutions to see dimensioanl space result.

fviz_eig(data.pca)



summary(data.pca)

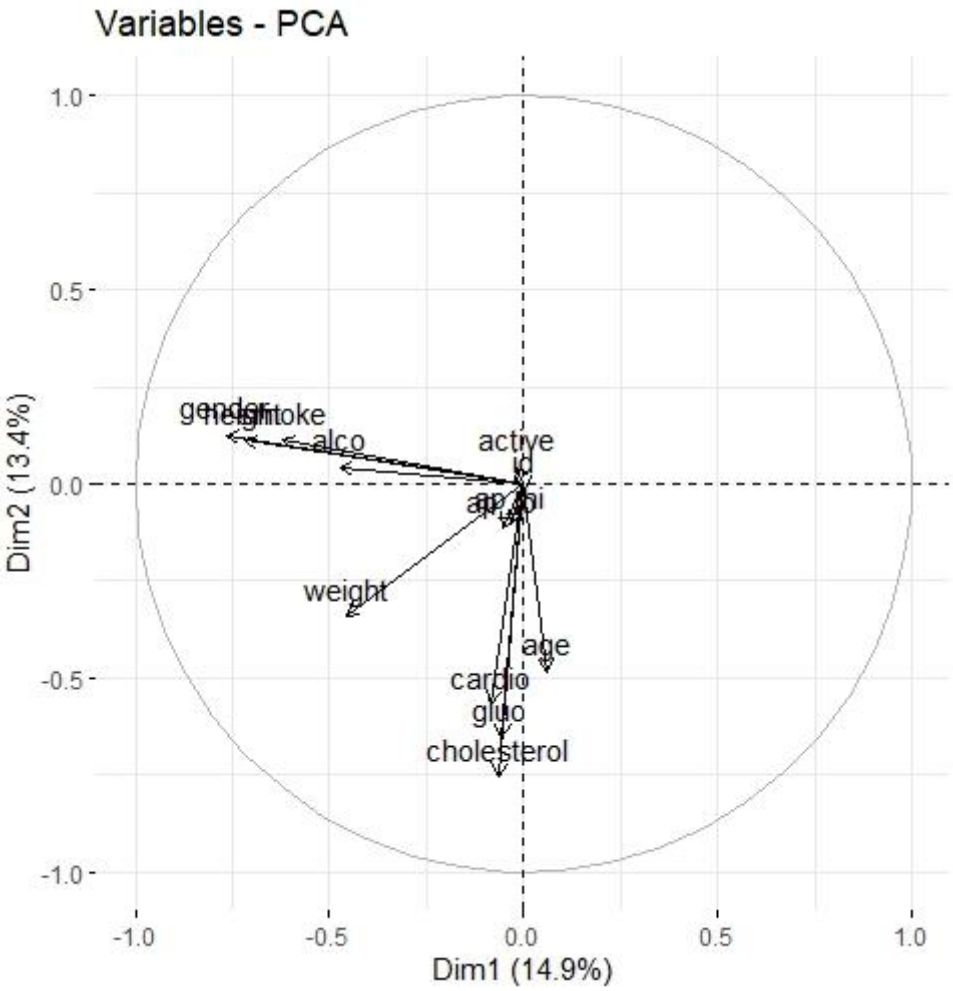
```
> summary(data.pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6    PC7     PC8     PC9    PC10    PC11
Standard deviation     1.3939 1.3185 1.07779 1.02688 1.00108 0.99888 0.9920 0.98647 0.90943 0.85061 0.78837
Proportion of Variance 0.1495 0.1337 0.08936 0.08111 0.07709 0.07675 0.0757 0.07486 0.06362 0.05566 0.04781
Cumulative Proportion  0.1495 0.2832 0.37255 0.45366 0.53075 0.60750 0.6832 0.75806 0.82168 0.87734 0.92515
                          PC12    PC13
Standard deviation     0.72411 0.66990
Proportion of Variance 0.04033 0.03452
Cumulative Proportion  0.96548 1.00000
```

If we look at the plot of components and variance that they explain.In the analysis, the all component will be taken into consideration. (Mainly, 3 component)

Component Analysis

Cleare visoin of issues

fviz_pca_var(data.pca, col.ind = "Age")



Conclution

To conculude the project, analysing current data, it gives us more understanding about most case in cardiovascular disease.

References

Source information; /kaggle/input/cardio_train.csv

Movite based on: RPubs - Dimension Reduction for nominal data and qualitative data