Deliverable 2

PCA, CA and Clustering

Jordi Catafal, Lluís Cerdà, Judit Serna, Tomàs Serra

May 5, 2024

Contents

Principal Components Analysis (PCA)

```
#names(df)
vars_con
```

```
## [1] "age"      "trestbps" "chol"     "thalach"  "oldpeak"  "ca"
"target"
```

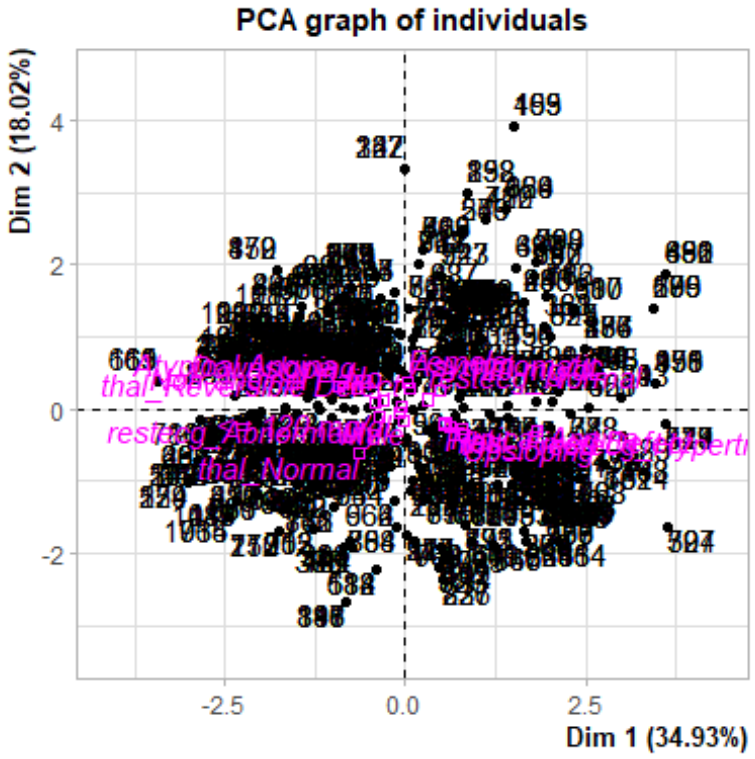
```
vars_dis
```

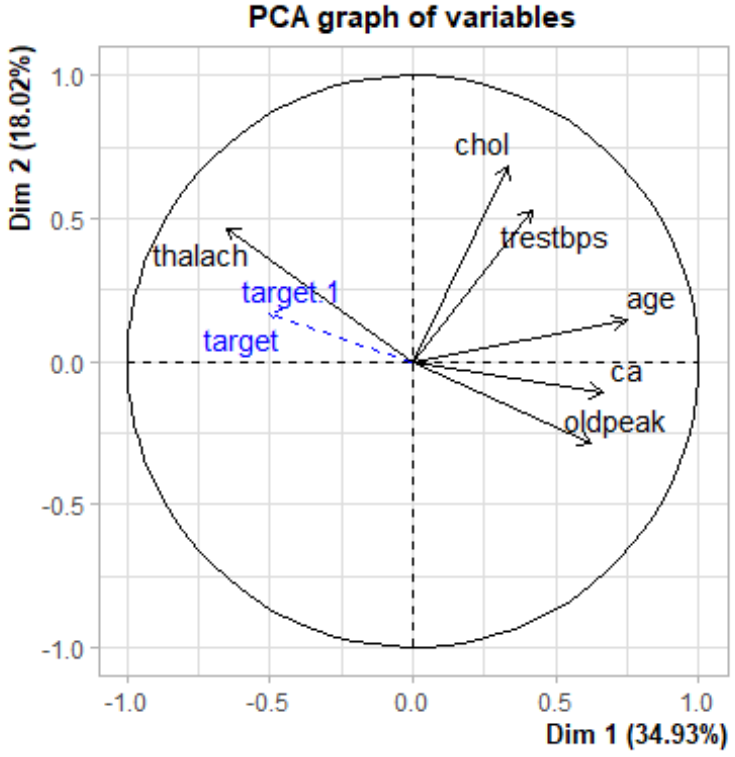
```
## [1] "sex"     "cp"      "fbs"     "restecg" "exang"   "slope"   "thal"
```

```
vars_res = c("target")


c(vars_res, vars_dis,vars_con, "mout") #All variables
```

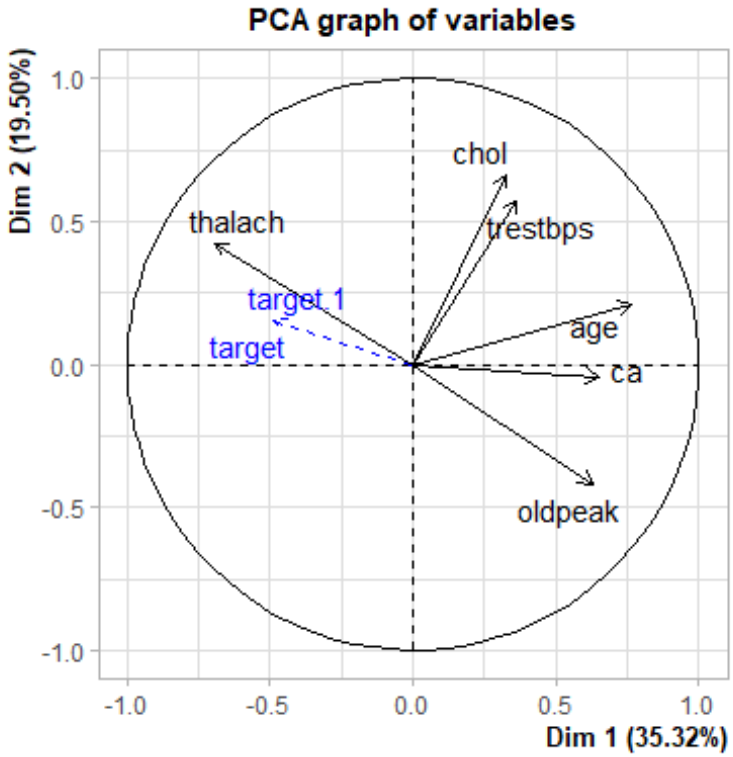
```
##  [1] "target"   "sex"      "cp"       "fbs"      "restecg"  "exang"
##  [7] "slope"    "thal"     "age"      "trestbps" "chol"     "thalach"
## [13] "oldpeak"  "ca"       "target"   "mout"
```

```
#summary( df[ , c(vars_res, vars_dis,vars_con, "mout") ] )
res.pca<-PCA(df[,c(vars_res, vars_dis, vars_con)], quali.sup=c(2:8),
quanti.sup= c(1,15))
```

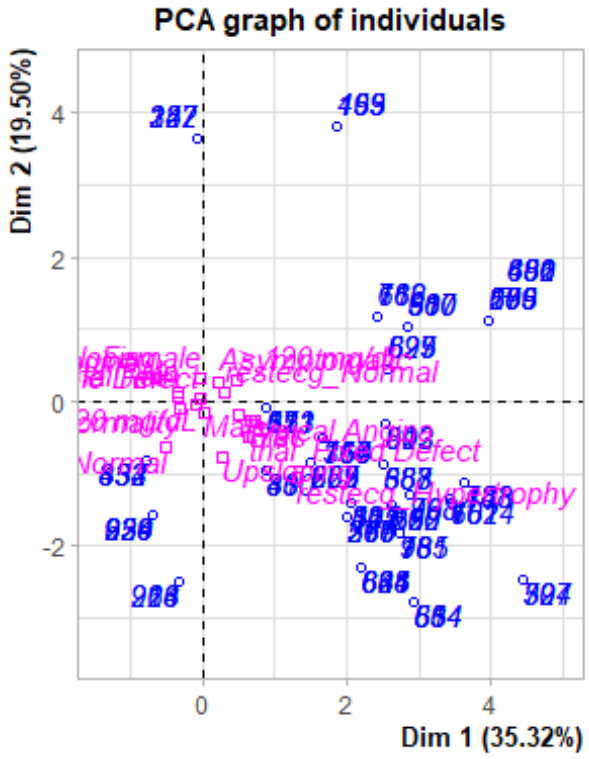





```
# Multivariant outliers should be included as supplementary observations
ll <- which( df$mout == "YesMOut")
res.pca<-PCA(df[,c(vars_res, vars_dis,
vars_con)],quali.sup=c(2:8),quanti.sup= c(1,15), ind.sup = ll )
```

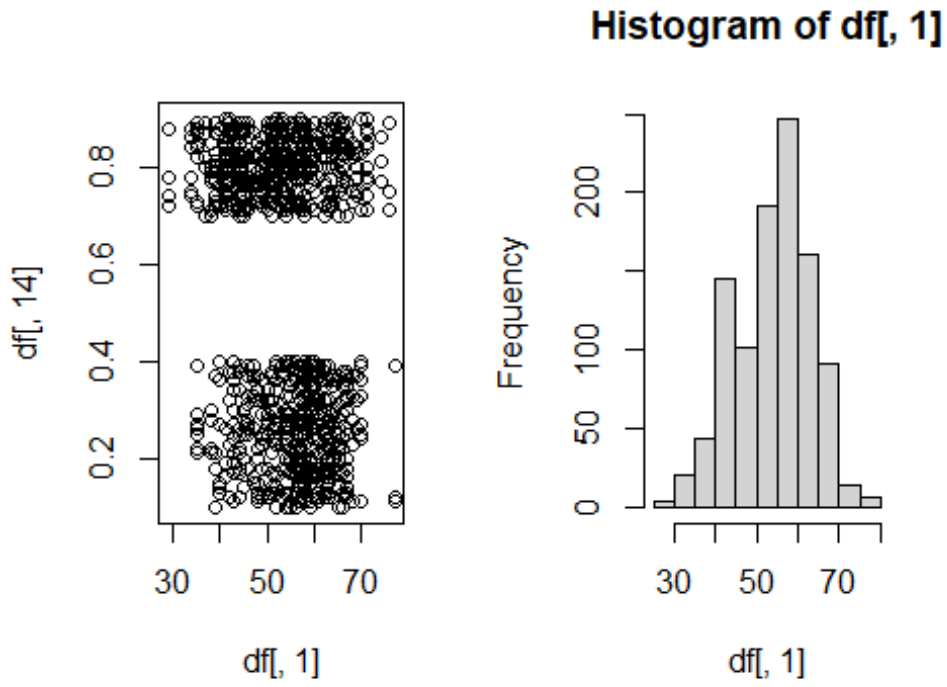



```
#plot.PCA(res.pca,choix=c("var"),axes = c(1, 2))
#plot.PCA(res.pca,choix=c("var"),invisible=c("var"))
#plot.PCA(res.pca,choix=c("var"),invisible=c("quanti.sup","var"))
plot.PCA(res.pca,choix=c("ind"),invisible=c("ind"))
```

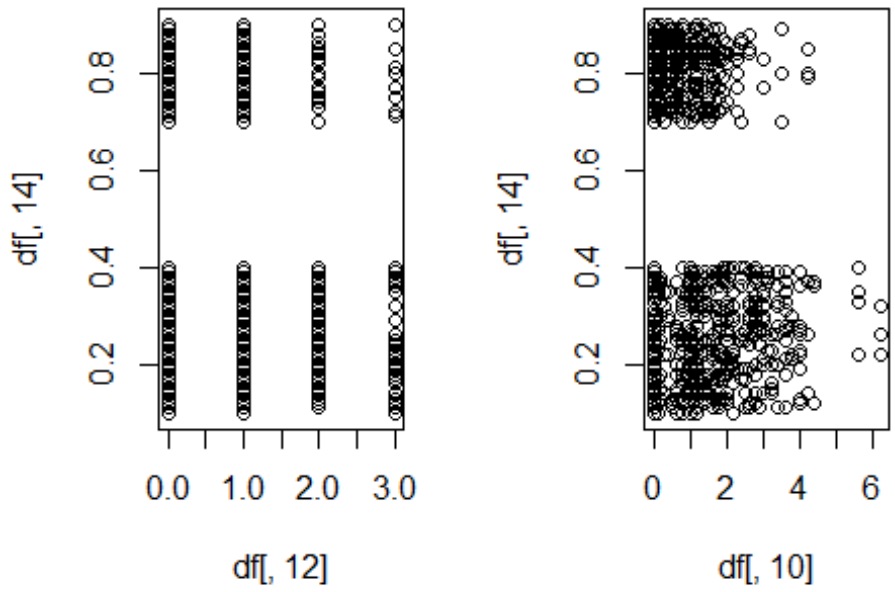


Aquest extrany PCA pot ser resultat de que la mostra que s'ha utilitzat per estudiar es gent jove amb problemes de cor principalment i en canvi la gent d'una major edat que ha format part de l'estudi no té aquests problemes. Això fa que quan més jove, doncs més probable de que tinguis problemes de cor. Però el colesterol, si que està correlacionat positivament en l'eix de les Y, que representa el conjunt de variables relacionades amb la sang. Ara bé, el "ca" o "oldpeak", són també causes d'atac de cor, al ser problemes comuns en gent gran, es correlacionen negativament amb el target pel que s'ha esmentat anteriorment.

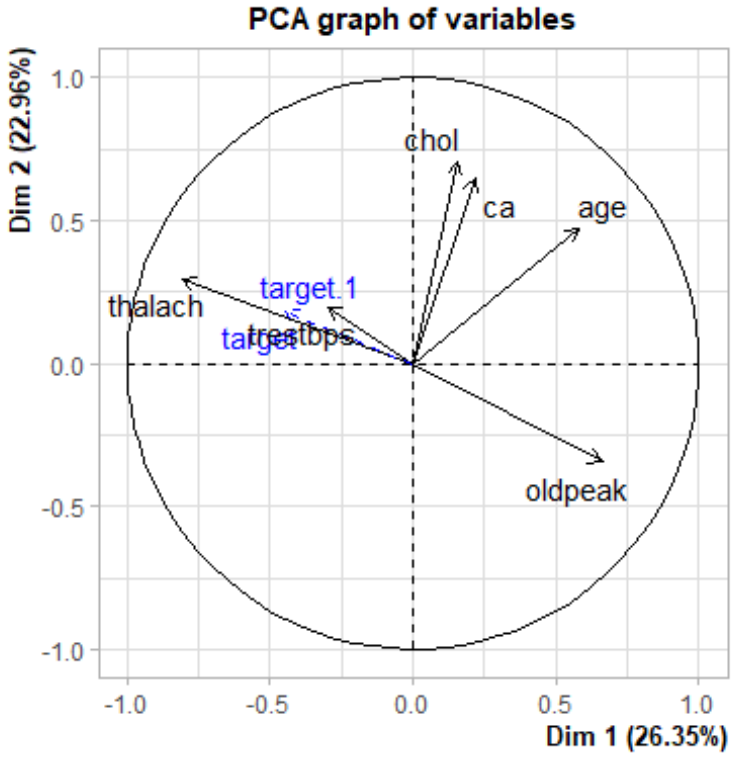
```
par(mfrow = c(1, 2))

plot(df[,1], df[,14]) #Here we can verify that there is only people with
problems or people with good heart health.
hist(df[,1]) #Here we verify that the most part of our population has
around 60 years old, but the majority of those individuals does not have
heart problems. And the most part of our population that has problems is
under those 50 years old. In addition, the youngest inviduals, younger
than 35, the majority has problems.
```




```
plot(df[,12], df[,14])

plot(df[,10], df[,14])
```



```r
plot(df[,1], df[,10]) #Here we conclude that not matter how bad a high
oldpeak is, that in our individuals with heart problems they have it
correct. So in reality is a factor that the bigger, the more probability
to have a problem. But for our individuals in this study we cannot
conclude that. And this affects the PCA.

#PCA with at least 45 years
df2 <- df[df$age < 45,]

res.pca45<-PCA(df2[,c(vars_res, vars_dis, vars_con)], quali.sup=c(2:8),
quanti.sup= c(1,15))

# Multivariant outliers should be included as supplementary observations
ll <- which( df2$mout == "YesMOut")
res.pca45<-PCA(df2[,c(vars_res, vars_dis,
vars_con)],quali.sup=c(2:8),quanti.sup= c(1,15), ind.sup = ll )
```



Eigenvalues and dominant axes analysis. How many axes we have to interpret according to Kaiser and Elbow's rule?

Individuals point of view: Are they any individuals "too contributive"? To better understand the axes meaning use the extreme individuals. Detection of multivariate outliers and influent data.

Interpreting the axes: Variables point of view coordinates, quality of representation, contribution of the variables

Perform a PCA taking into account also supplementary variables the supplementary variables can be quantitative and/or categorical

```
res.pca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of
variance
## comp 1  2.1191045              35.318409
35.31841
## comp 2  1.1699218              19.498697
54.81711
## comp 3  0.9050676              15.084460
69.90157
## comp 4  0.7087219              11.812031
81.71360
## comp 5  0.6339388              10.565647
92.27924
## comp 6  0.4632454               7.720756
100.00000
```

According to the Kaiser rule we have to take into account those components which have the 80% of the whole variables. So in this case we should consider comp 1 (35.38%), comp 2 (19.65%), comp 3 (14.99%) and comp 4 (11.71%)

```
res.pca$eig[,1]
```

```
##    comp 1    comp 2    comp 3    comp 4    comp 5    comp 6
## 2.1191045 1.1699218 0.9050676 0.7087219 0.6339388 0.4632454
```

```
# Eigenvalues are the square of the singular values (sdev)
eigenvalues <- res.pca$eig[,1]

# Calculate the proportion of variance explained
total_variance <- sum(eigenvalues)
explained_variance <- eigenvalues / total_variance

print(explained_variance)
```

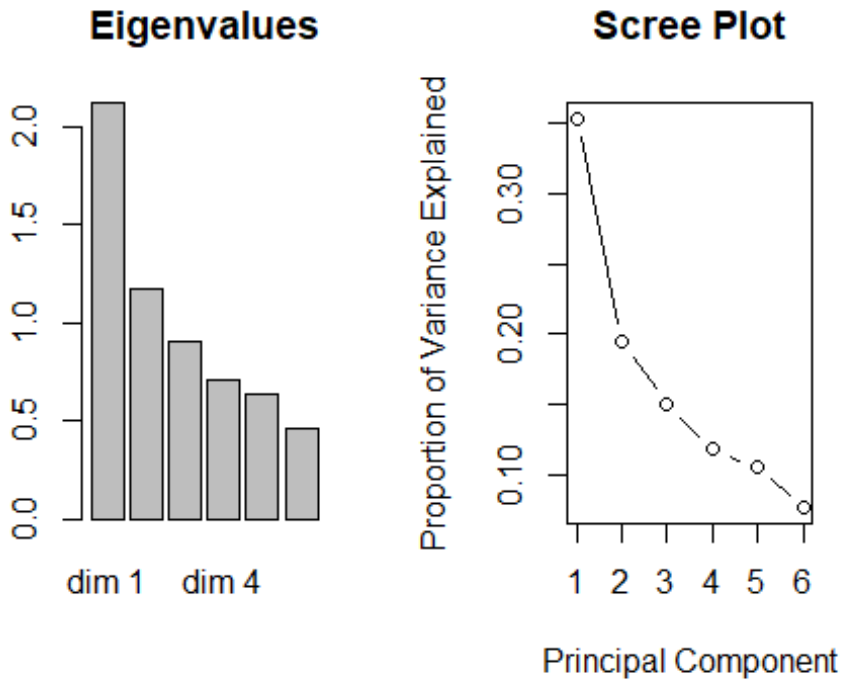
```
##     comp 1     comp 2     comp 3     comp 4     comp 5     comp 6
## 0.35318409 0.19498697 0.15084460 0.11812031 0.10565647 0.07720756
```

```
summary(res.pca, nb.dec = 2, ncp = 3, nbelements = 1)
```

```
##
## Call:
## PCA(X = df[, c(vars_res, vars_dis, vars_con)], ind.sup = ll,
##      quanti.sup = c(1, 15), quali.sup = c(2:8))
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6
## Variance                2.12   1.17   0.91   0.71   0.63   0.46
```

```
## % of var.             35.32  19.50  15.08  11.81  10.57   7.72
## Cumulative % of var.  35.32  54.82  69.90  81.71  92.28 100.00
##
## Individuals (the 1 first)
##             Dist   Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr
cos2
## 1        |  1.89 |  0.00  0.00  0.00 | -0.46  0.02  0.06 | -0.40  0.02
0.04 |
##
## Supplementary individuals (the 1 first)
##             Dist   Dim.1  cos2   Dim.2  cos2   Dim.3  cos2
## 7        |  5.22 |  3.44  0.43 | -1.34  0.07 | -1.52  0.08 |
##
## Variables (the 1 first)
##            Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr  cos2
## age      |  0.76 27.61  0.59 |  0.21  3.79  0.04 |  0.02  0.06  0.00 |
##
## Supplementary continuous variables (the 1 first)
##            Dim.1  cos2   Dim.2  cos2   Dim.3  cos2
## target   | -0.49  0.24 |  0.15  0.02 |  0.07  0.00 |
##
## Supplementary categories (the 1 first)
##             Dist   Dim.1  cos2 v.test   Dim.2  cos2 v.test   Dim.3
cos2 v.test
## Female   |  0.39 | -0.03  0.01  -0.43 |  0.32  0.69   6.06 | -0.06
0.02  -1.28
##
## Female   |
```

```
par(mfrow = c(1, 2))
barplot(res.pca$eig[, 1], main = "Eigenvalues", names.arg = paste("dim",
1:nrow(res.pca$eig)))

plot(explained_variance, xlab = "Principal Component", ylab = "Proportion
of Variance Explained", type = 'b', main = "Scree Plot")
```



Here we can confirm that we need the first 4 components to at least represent the 80% of the variables. We also have added a barplot were we also can see that the diference in the values represented between 4th and 5th are minimum.

Using the res.pca45, the PCA done by age greater than 45 years.

```
res.pca45$eig

##        eigenvalue percentage of variance cumulative percentage of
variance
## comp 1  1.5812967              26.354944
26.35494
## comp 2  1.3777648              22.962746
49.31769
## comp 3  1.0721064              17.868440
67.18613
## comp 4  0.8281359              13.802266
80.98840
## comp 5  0.6784422              11.307369
92.29577
## comp 6  0.4622541               7.704235
100.00000
```

According to the Kaiser rule we have to take into account those components which have the 80% of the whole variables. So in this case we should consider comp 1 (26.35%), comp 2 (22.96%), comp 3 (17.87%) and comp 4 (13.8%)

```r
res.pca45$eig[,1]
```

```
##    comp 1    comp 2    comp 3    comp 4    comp 5    comp 6
## 1.5812967 1.3777648 1.0721064 0.8281359 0.6784422 0.4622541
```

```r
# Eigenvalues are the square of the singular values (sdev)
eigenvalues45 <- res.pca45$eig[,1]

# Calculate the proportion of variance explained
total_variance45 <- sum(eigenvalues45)
explained_variance45 <- eigenvalues45 / total_variance45

print(explained_variance45)
```

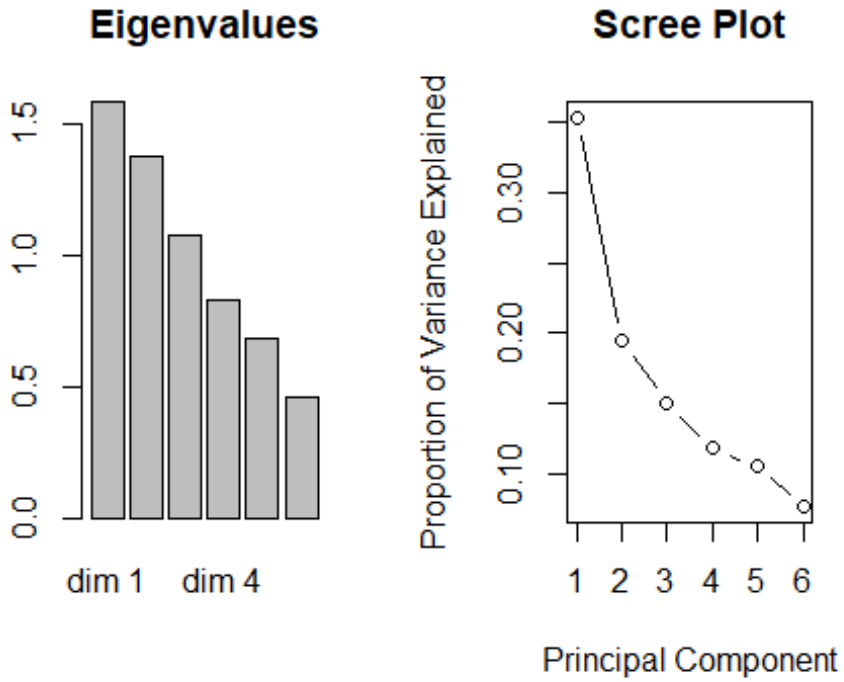
```
##     comp 1     comp 2     comp 3     comp 4     comp 5     comp 6
## 0.26354944 0.22962746 0.17868440 0.13802266 0.11307369 0.07704235
```

```r
summary(res.pca45, nb.dec = 2, ncp = 3, nbelements = 1)
```

```
##
## Call:
## PCA(X = df2[, c(vars_res, vars_dis, vars_con)], ind.sup = ll,
##      quanti.sup = c(1, 15), quali.sup = c(2:8))
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6
## Variance                1.58   1.38   1.07   0.83   0.68   0.46
## % of var.              26.35  22.96  17.87  13.80  11.31   7.70
## Cumulative % of var.   26.35  49.32  67.19  80.99  92.30 100.00
##
## Individuals (the 1 first)
##             Dist   Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr
cos2
## 13       |  2.49 | -1.64  0.97  0.44 | -0.93  0.35  0.14 | -0.06  0.00
0.00 |
##
## Supplementary individuals (the 1 first)
##            Dist   Dim.1 cos2   Dim.2 cos2   Dim.3 cos2
## 12       | 4.48 |  2.99 0.44 |  0.62 0.02 |  2.45 0.30 |
##
## Variables (the 1 first)
##            Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr  cos2
## age      |  0.58 21.29  0.34 |  0.47 16.08  0.22 | -0.20  3.78  0.04 |
##
## Supplementary continuous variables (the 1 first)
##            Dim.1  cos2   Dim.2  cos2   Dim.3  cos2
## target   | -0.44  0.20 |  0.18  0.03 | -0.10  0.01 |
##
## Supplementary categories (the 1 first)
##             Dist   Dim.1  cos2 v.test   Dim.2  cos2 v.test   Dim.3
```

```
cos2 v.test
## Female   |  0.54 | -0.15  0.07  -0.89 | -0.07  0.01  -0.43 | -0.46
0.74  -3.41
##
## Female   |
```

```
par(mfrow = c(1, 2))
barplot(res.pca45$eig[, 1], main = "Eigenvalues", names.arg =
paste("dim", 1:nrow(res.pca45$eig)))

plot(explained_variance, xlab = "Principal Component", ylab = "Proportion
of Variance Explained", type = 'b', main = "Scree Plot")
```



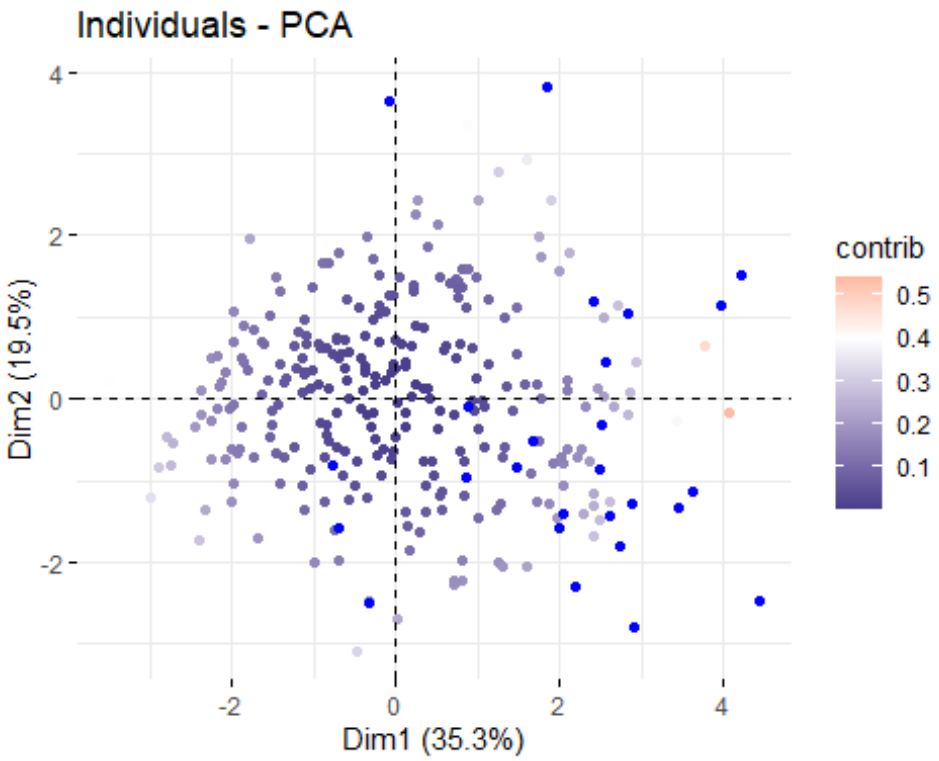
Here we can confirm that we need the first 4 components to at least represent the 80% of the variables. We also have added a barplot were we also can see that the diference in the values represented between 4th and 5th are minimum.

Individuals point of view

Contribution

```
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
fviz_pca_ind(res.pca, col.ind="contrib", geom = "point") +
scale_color_gradient2(low="darkslateblue", mid='white', high='red',
midpoint=0.40)
```



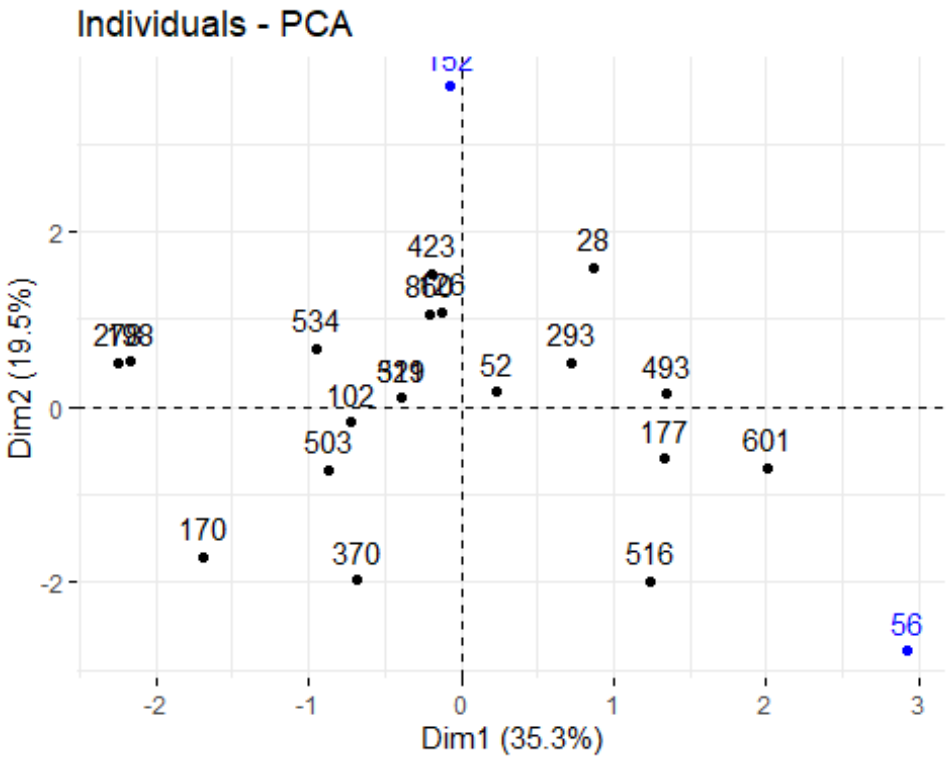
We can see that there are only two individuals who are slightly more contributive. There are rather more individuals who have a low contribution.

Extreme individuals

In dimension 1

We now look for the extreme individuals to better understand the axes meaning. First we plot them.

```
rang<-order(res.pca$ind$coord[,1])

contrib.extremes<-c(row.names(df)[rang[1:10]], row.names(df)
[rang[(length(rang)-10):length(rang)]])
fviz_pca_ind(res.pca, select.ind = list(names=contrib.extremes))
```



Here we have a closer look at the extrem individuals.

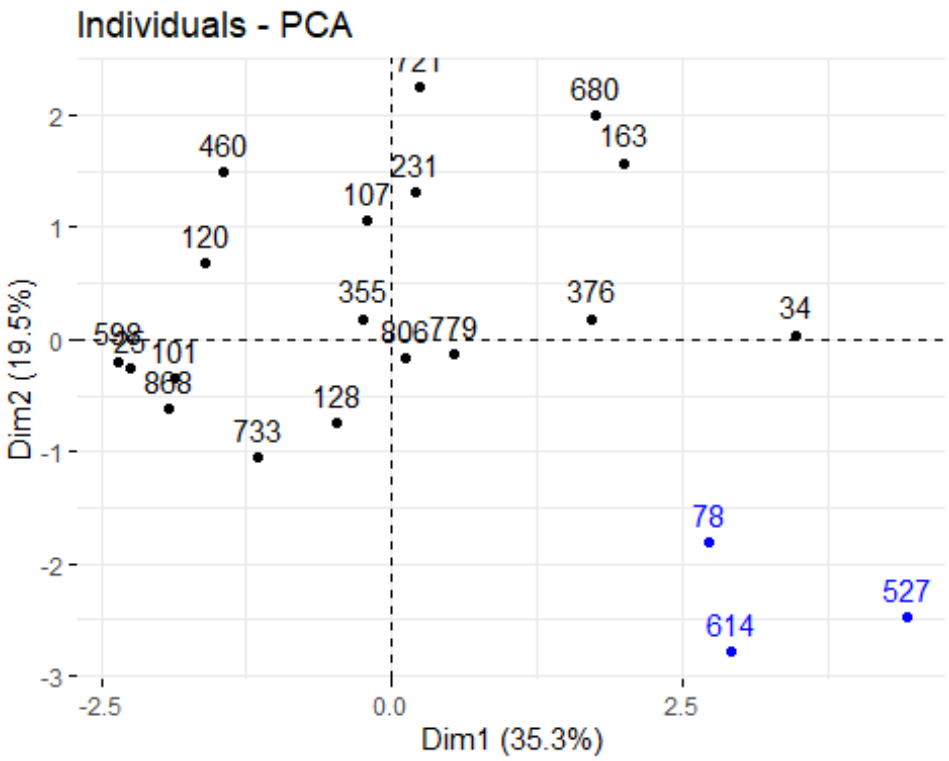
```
df[which(row.names(df) %in% row.names(df)[rang[length(rang)]]), 1:15]

##     age  sex             cp trestbps chol          fbs     restecg
thalach
## 521  59 Male Typical Angina      140  177 <= 120 mg/dL Abnormality
162
##     exang oldpeak       slope ca              thal target   mout
## 521   Yes       0 Downsloping  1 Reversible Defect    0.1 NoMOut

df[which(row.names(df) %in% row.names(df)[rang[1]]),1:15]

##    age    sex             cp trestbps chol          fbs     restecg
thalach
## 52  57 Female Typical Angina      140  241 <= 120 mg/dL Abnormality
123
##    exang oldpeak slope ca              thal target   mout
## 52   Yes     0.2  Flat  0 Reversible Defect    0.3 NoMOut
```

In dimension 2

We replicate the procedure conducted for dimension one, now focusing on dimension two.

```
rang<-order(res.pca$ind$coord[,2])

contrib.extremes<-c(row.names(df)[rang[1:10]], row.names(df)
```

```
[rang[(length(rang)-10):length(rang)]])
fviz_pca_ind(res.pca, select.ind = list(names=contrib.extremes))
```




```
df[which(row.names(df) %in% row.names(df)[rang[length(rang)]]), 1:15]

##     age  sex             cp trestbps chol          fbs restecg thalach
exang
## 806  53 Male Typical Angina      142  226 <= 120 mg/dL  Normal     111
Yes
##     oldpeak       slope ca              thal target   mout
## 806       0 Downsloping  0 Reversible Defect   0.85 NoMOut

df[which(row.names(df) %in% row.names(df)[rang[1]]),1:15]

##     age  sex               cp trestbps chol         fbs restecg
thalach exang
## 128  53 Male Non-anginal Pain      130  197 > 120 mg/dL  Normal
152    No
##     oldpeak     slope ca              thal target   mout
## 128     1.2 Upsloping  0 Reversible Defect   0.83 NoMOut
```

Interpreting the axes

Variables point of view coordinates

Initially, we examine the contribution of each variable to the dimensions. Age appears as the most significant contributor to the first dimension, while cholesterol (chol)

takes precedence in the second dimension. Notably, variables such as thalac, oldpeak, and ca only contribute significantly to one of the two dimensions.

```
res.pca$var$coord
```

```
##               Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
## age       0.7648891  0.21045293  0.02297982  0.1054981 -0.42541140
## trestbps  0.3635033  0.57344859  0.68589159  0.0954852  0.08742109
## chol      0.3231596  0.66386747 -0.49347469 -0.4274852  0.15854108
## thalach  -0.6938352  0.42237630  0.03948209  0.2434468  0.34222492
## oldpeak   0.6280475 -0.41936944  0.24434801 -0.3079318  0.47329539
## ca        0.6493316 -0.04237476 -0.35959724  0.5929946  0.28117582
```

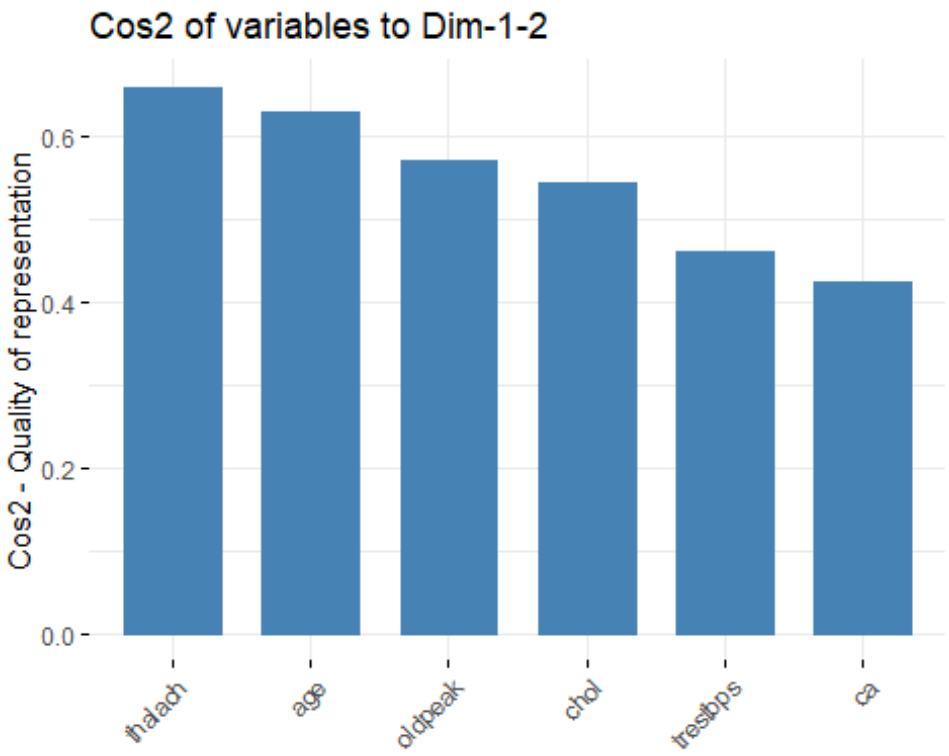
Quality of representation

Here, we observe the degree of representation for each category across the first four dimensions. In the plane of the first and second dimensions, all categories are well represented. However, in the plane of the third and fourth dimensions, categories such as oldpeak, thalac, and age show comparatively lower levels of representation.

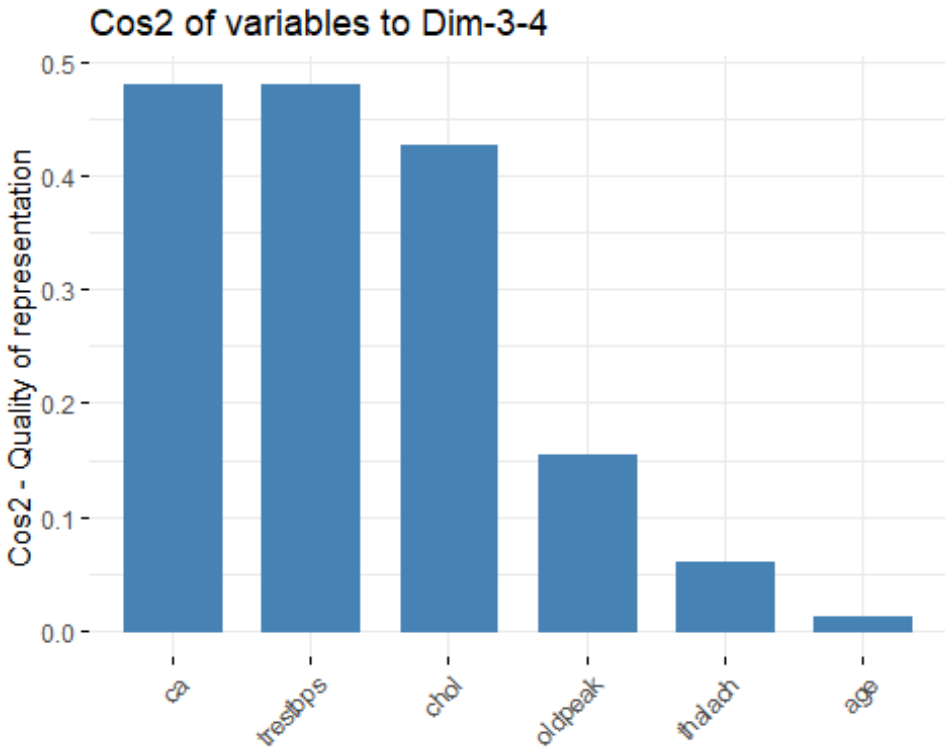
```
res.pca$var$cos2
```

```
##              Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
## age      0.5850554 0.044290436 0.000528072 0.011129844 0.180974859
## trestbps 0.1321346 0.328843288 0.470447271 0.009117423 0.007642447
## chol     0.1044321 0.440720018 0.243517267 0.182743629 0.025135273
## thalach  0.4814073 0.178401737 0.001558835 0.059266339 0.117117894
## oldpeak  0.3944436 0.175870727 0.059705952 0.094821966 0.224008525
## ca       0.4216315 0.001795621 0.129310176 0.351642654 0.079059841
```

```
fviz_cos2(res.pca, choice = "var", axes = 1:2)
```




```
fviz_cos2(res.pca, choice = "var", axes = 3:4)
```



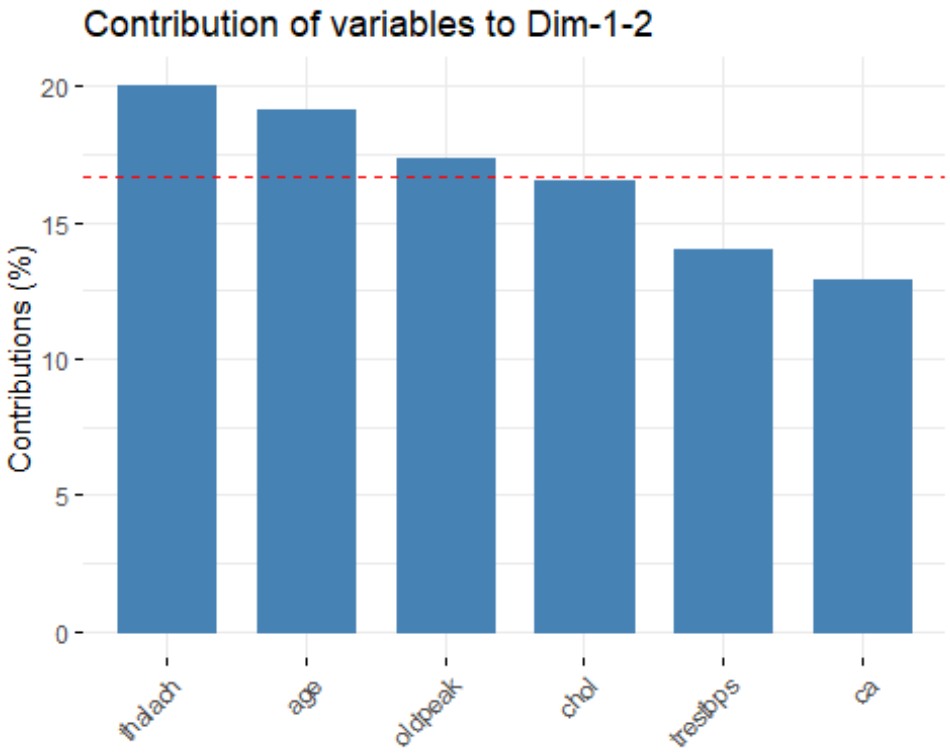
Contribution of the variables

Here, we seek to understand the contribution of categories across planes of different dimensions. We observe a similar pattern to that seen in representation of the categories.

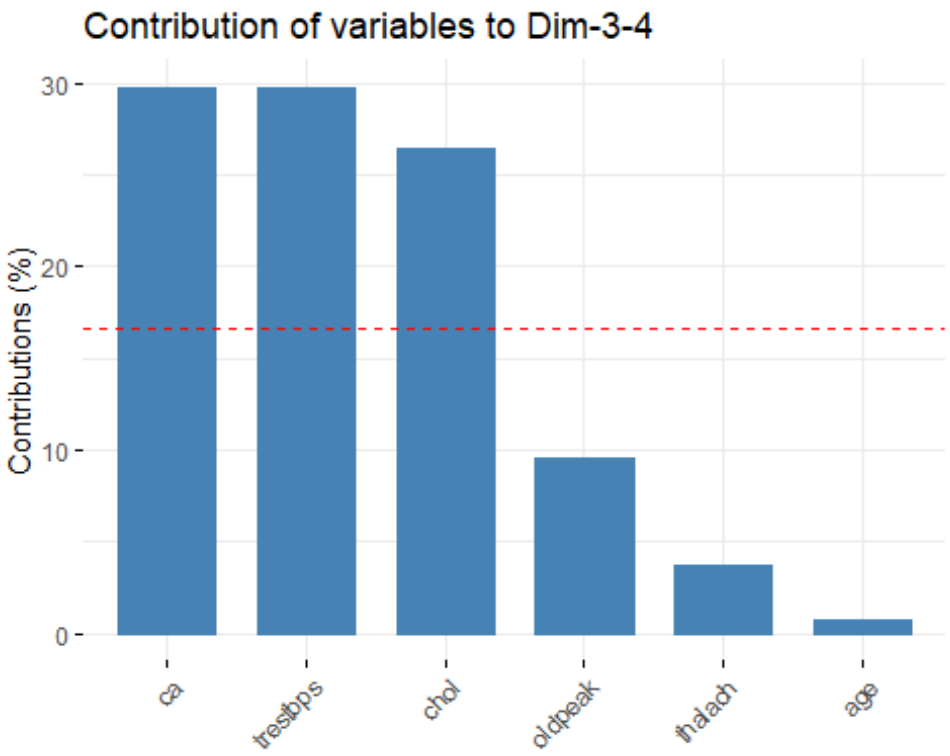
```
res.pca$var$contrib
```

```
##               Dim.1      Dim.2       Dim.3     Dim.4     Dim.5
## age      27.608613  3.7857603  0.05834614  1.570411 28.547684
## trestbps  6.235399 28.1081420 51.97924277  1.286460  1.205550
## chol      4.928125 37.6708946 26.90597631 25.784958  3.964937
## thalach  22.717485 15.2490306  0.17223414  8.362426 18.474636
## oldpeak  18.613694 15.0326905  6.59685023 13.379292 35.335984
## ca       19.896683  0.1534821 14.28735041 49.616454 12.471210
```

```
fviz_contrib(res.pca, choice = "var", axes = 1:2)
```

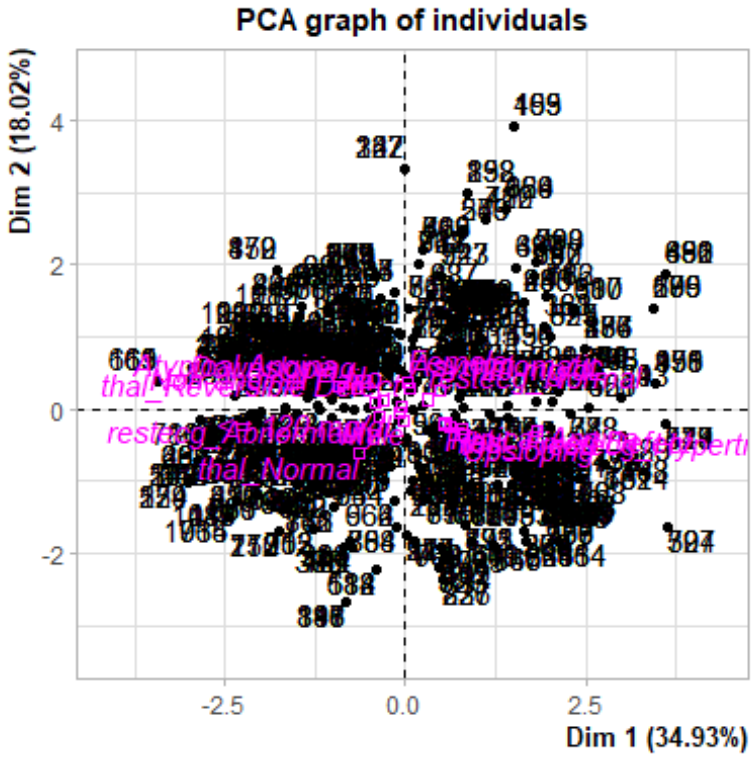


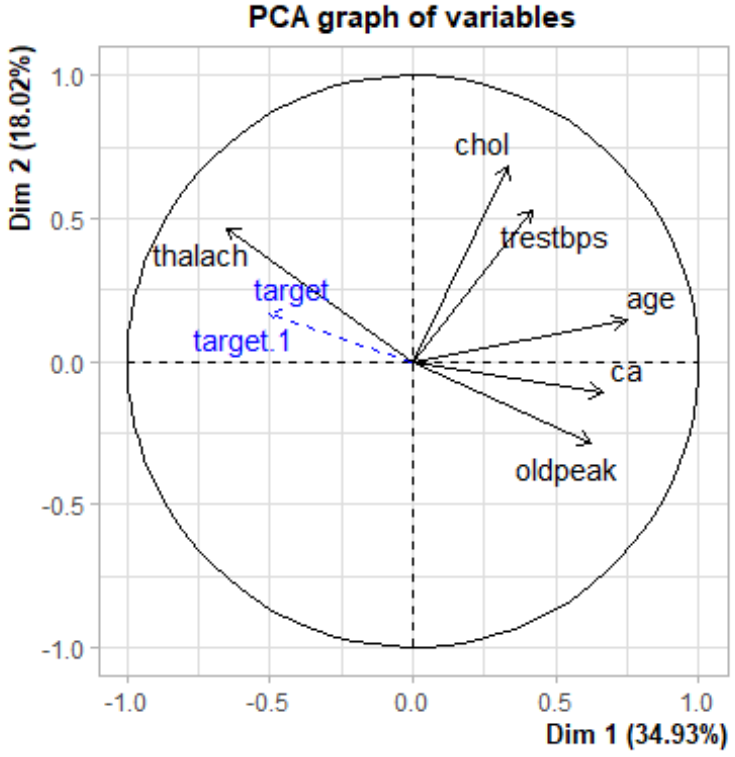

```
fviz_contrib(res.pca, choice = "var", axes = 3:4)
```



K-Means Clustering

```
dim(res.pca$ind$coord[, 1:2])

## [1] 932   2

#dist(res.pca$ind$coord[, 1:2])

res.pca<-PCA(df[,c(vars_res, vars_dis, vars_con)], quali.sup=c(2:8),
quanti.sup= c(1,15))
```






```
kc <- kmeans(res.pca$ind$coord[, 1:2], centers = 4)
```

Quality - Suma de quadrats total = Suma de quadrats intra + Suma de quadrats inter

```
100*kc$betweenss/kc$totss
```

```
## [1] 70.68443
```

```
df$kmeaclu <- factor(kc$cluster)

res.catdes <- catdes(df, 16)

res.catdes$test.chi2
```

```
##              p.value df
## cp      8.475467e-42  9
## slope   5.846107e-41  6
## mout    1.810566e-37  3
## exang   5.424431e-25  3
## restecg 2.488589e-14  6
## sex     1.046552e-08  3
## thal    1.898015e-07  6
## fbs     3.882112e-07  3
```

cp (p = 1.21e-42, df=9) and slope (p=3.62e-42, df=6) : A very small p-value suggests this variables have a significant association with the clusters. With 9 and 6 degrees of freedom, which suggests that the variables have several categories influencing how data is grouped into these clusters, since df > num. clusters.

exang (p=4.02e-26, df=3) and mout(p=6.29e-22, df=3): A significant p-value with fewer degrees of freedom than number of clusters, suggesting a strong association with the clusters.

restecg (p=1.08e-13, df=6) and thal (p=1.27e-07): Still a significant p-value, with 6 degrees of freedom.

sex (p=2.50e-07, df=3) and fbs (p=5.25e-06, df=3): Also significant p-value, since it is smaller than 0.05, with 3 degrees of freedom.

```
res.catdes$category[1]
```

```
## $`1`
##                           Cla/Mod   Mod/Cla    Global      p.value
v.test
## mout=YesMOut           72.043011 32.524272  9.073171 1.453745e-30
11.491611
## cp=Typical Angina      32.796781 79.126214 48.487805 9.759605e-24
10.044037
## slope=Flat             31.327801 73.300971 47.024390 1.514702e-17
8.526019
## exang=Yes              32.173913 53.883495 33.658537 2.073856e-11
6.700727
## slope=Upsloping        44.594595 16.019417  7.219512 7.118198e-07
4.958111
## restecg=Hypertrophy    73.333333  5.339806  1.463415 1.177767e-05
4.381663
## thal=Fixed Defect      35.365854 14.077670  8.000000 7.739897e-04
```

```
3.361933
## restecg=Normal         23.943662 57.766990 48.487805 2.933234e-03
2.974650
## thal=Reversible Defect 18.910256 85.922330 91.317073 3.649626e-03  -
2.906958
## restecg=Abnormality    14.814815 36.893204 50.048780 2.330790e-05  -
4.230590
## cp=Atypical Angina      5.988024  4.854369 16.292683 4.313136e-08  -
5.477526
## exang=No               13.970588 46.116505 66.341463 2.073856e-11  -
6.700727
## cp=Non-anginal Pain     7.042254  9.708738 27.707317 3.810613e-12  -
6.944033
## mout=NoMOut            14.914163 67.475728 90.926829 1.453745e-30 -
11.491611
## slope=Downsloping       4.690832 10.679612 45.756098 3.834205e-33 -
11.993693
```

This first cluster has high influence in individuals with sex=Female, cp=Asympomatic and fbs>=120 mg/dL. The mean of this variables in this cluster is significantly higher than the global one.

As less characteristic variables we can see sex=Male, cp=Typical Angina and fbs<=120 mg/dL. The mean in the cluster is significantly lower than the global one.

```
res.catdes$category[2]
```

```
## $`2`
##                          Cla/Mod   Mod/Cla    Global      p.value
v.test
## exang=Yes              32.463768 48.068670 33.658537 2.075698e-07
5.192427
## cp=Typical Angina      29.175050 62.231760 48.487805 1.785011e-06
4.776355
## mout=NoMOut            24.356223 97.424893 90.926829 1.468154e-05
4.333413
## slope=Flat             28.630705 59.227468 47.024390 2.274582e-05
4.236078
## sex=Male               26.086957 79.828326 69.560976 7.341938e-05
3.964924
## thal=Fixed Defect      39.024390 13.733906  8.000000 5.505501e-04
3.454874
## fbs=> 120 mg/dL        30.718954 20.171674 14.926829 1.305930e-02
2.482148
## fbs=<= 120 mg/dL       21.330275 79.828326 85.073171 1.305930e-02 -
2.482148
## thal=Reversible Defect 21.153846 84.978541 91.317073 2.342523e-04 -
3.678889
## sex=Female             15.064103 20.171674 30.439024 7.341938e-05 -
3.964924
## slope=Downsloping      17.057569 34.334764 45.756098 6.312769e-05 -
```

```
4.000804
## mout=YesMOut            6.451613  2.575107  9.073171 1.468154e-05 -
4.333413
## exang=No               17.794118 51.931330 66.341463 2.075698e-07 -
5.192427
## cp=Atypical Angina      8.383234  6.008584 16.292683 1.721886e-07 -
5.227101
```

The second cluster has high influence in individuals with cp=Typical Angina, slope=Flat and exang = Yes. We can see this is really contrary to the first cluster, the less characteristic variables are slope=Downsloping, cp=Non-anginal Pain and exang = No.

```
res.catdes$category[3]
```

```
## $`3`
##                          Cla/Mod   Mod/Cla    Global      p.value
v.test
## sex=Female             35.25641 45.454545 30.439024 1.465635e-08
5.665538
## cp=Asymptomatic        44.15584 14.049587  7.512195 3.754451e-05
4.122093
## fbs=> 120 mg/dL        35.94771 22.727273 14.926829 1.842168e-04
3.739732
## restecg=Normal         28.16901 57.851240 48.487805 8.724835e-04
3.328713
## slope=Downsloping      28.14499 54.545455 45.756098 1.761784e-03
3.127703
## thal=Reversible Defect 24.78632 95.867769 91.317073 2.390688e-03
3.036844
## mout=NoMOut            24.57082 94.628099 90.926829 1.769581e-02
2.371922
## mout=YesMOut           13.97849  5.371901  9.073171 1.769581e-02 -
2.371922
## restecg=Hypertrophy     0.00000  0.000000  1.463415 1.704774e-02 -
2.385676
## slope=Upsloping        12.16216  3.719008  7.219512 1.188846e-02 -
2.515438
## thal=Fixed Defect      12.19512  4.132231  8.000000 7.922484e-03 -
2.655356
## restecg=Abnormality    19.88304 42.148760 50.048780 4.976431e-03 -
2.808555
## cp=Typical Angina      18.51107 38.016529 48.487805 1.882105e-04 -
3.734337
## fbs=<= 120 mg/dL       21.44495 77.272727 85.073171 1.842168e-04 -
3.739732
## sex=Male               18.51332 54.545455 69.560976 1.465635e-08 -
5.665538
```

This cluster has high influence in individuals with sex=Female, cp=Asympomatic and fbs>=120 mg/dL. The mean of this variables in this cluster is significantly higher than the global one.

As less characteristic variables we can see sex=Male, cp=TypicalAngina and fbs<=120 mg/dL. The mean in the cluster is significantly lower than the global one.

```
res.catdes$category[4]
```

```
## $`4`
##                           Cla/Mod   Mod/Cla    Global      p.value
v.test
## slope=Downsloping      50.106610 68.313953 45.756098 4.178375e-25
10.350148
## exang=No               43.088235 85.174419 66.341463 4.399845e-21
9.422617
## cp=Atypical Angina     63.473054 30.813953 16.292683 4.430999e-18
8.667143
## restecg=Abnormality    43.469786 64.825581 50.048780 1.505856e-11
6.747335
## mout=NoMOut            36.158798 97.965116 90.926829 8.213448e-10
6.140747
## fbs=<= 120 mg/dL       36.467890 92.441860 85.073171 9.276942e-07
4.906387
## cp=Non-anginal Pain    44.718310 36.918605 27.707317 3.908961e-06
4.616165
## thal=Reversible Defect 35.149573 95.639535 91.317073 2.722179e-04
3.640397
## slope=Upsloping        22.972973  4.941860  7.219512 4.230924e-02 -
2.030466
## cp=Asymptomatic        18.181818  4.069767  7.512195 2.106107e-03 -
3.074847
## restecg=Hypertrophy     0.000000  0.000000  1.463415 2.058764e-03 -
3.081622
## thal=Fixed Defect      13.414634  3.197674  8.000000 1.972347e-05 -
4.267998
## fbs=> 120 mg/dL        16.993464  7.558140 14.926829 9.276942e-07 -
4.906387
## restecg=Normal         24.346076 35.174419 48.487805 1.177291e-09 -
6.083306
## mout=YesMOut            7.526882  2.034884  9.073171 8.213448e-10 -
6.140747
## cp=Typical Angina      19.517103 28.197674 48.487805 9.780999e-21 -
9.338390
## slope=Flat             19.087137 26.744186 47.024390 7.426371e-21 -
9.367510
## exang=Yes              14.782609 14.825581 33.658537 4.399845e-21 -
9.422617
```

Finally we have a cluster that has high influence in individuals with exang=Yes, cp=Typical Angina and slope = flat.

And less influence in individuals with sex=Female, cp=Atypical Angina and exang=No.

We can observe contrary characteristics between clusters, which makes sense.

```
res.catdes$quanti.var

##               Eta2       P-value
## thalach  0.5014750 8.442899e-154
## age      0.4612799 1.259363e-136
## oldpeak  0.4239113 8.928747e-122
## ca       0.3601399  1.556608e-98
## chol     0.2827808  2.778811e-73
## trestbps 0.2756022  4.429298e-71
## target   0.2372237  1.146794e-59
```

The values which seem to help the most to differentiate between clusters are age, talach(maximum heart rate), oldpeak (ST depression induced by exercise) and ca, with this same order, both thalach and oldpeak related to cardiovascular characteristics and also age as an important factor.

On the other hand both chol and trestbps seem to be the continuous variables that help less in differentiating clusters.

```
res.catdes$quanti[1]

## $`1`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak   18.882879        2.4533981    1.0715122      1.2333013
1.1744799
## ca        17.216087        1.6941748    0.6878049      0.9846714
0.9381339
## age       12.660154       61.5873786   54.4341463      5.5309926
9.0678636
## trestbps   7.390949      139.6747573  131.6117073     18.6739017
17.5081712
## chol       3.579025      257.5000000  246.0000000     47.1704974
51.5673370
## target   -11.436189        0.3328155    0.5363902      0.2074436
0.2856828
## thalach  -16.986062      124.7766990  149.1141463     20.1831094
22.9944987
##                p.value
## oldpeak  1.577475e-79
## ca       2.011359e-66
## age      9.830643e-37
## trestbps 1.457846e-13
## chol     3.448780e-04
```

```
## target   2.757329e-30
## thalach  1.041505e-64
```

We can see this first cluster has a higher influence in individuals with a higher maximum heart rate, a lower cholesterol and younger in comparison to the overall mean.

```
res.catdes$quanti[2]
```

```
## $`2`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak    2.177883         1.218884    1.0715122      0.9298043
1.1744799
## target    -4.961804         0.454721    0.5363902      0.2770647
0.2856828
## thalach   -8.852398       137.386266  149.1141463     17.3288363
22.9944987
## trestbps  -9.596588       121.931330  131.6117073     12.2577647
17.5081712
## chol     -10.098899       215.995708  246.0000000     32.8240854
51.5673370
##                 p.value
## oldpeak  2.941476e-02
## target   6.984151e-07
## thalach  8.565871e-19
## trestbps 8.263505e-22
## chol     5.586591e-24
```

The indiviudals in this second cluster are older, have higher cholesterol and a really low maximum heart rate, they also have a higher ST depression caused by activity in comparison to rest. The characteristics of this individuals look contrary to the individuals in the first cluster.

```
res.catdes$quanti[3]
```

```
## $`3`
##              v.test Mean in category Overall mean sd in category
Overall sd
## chol     14.703267      288.6198347   246.000000     56.0505610
51.567337
## trestbps 12.306414      143.7231405   131.611707     17.0904800
17.508171
## age       9.071141       59.0578512    54.434146      6.8740710
9.067864
## thalach   5.509569      156.2355372   149.114146     14.6859091
22.994499
## oldpeak  -4.350486        0.7842975     1.071512      0.7755039
1.174480
##                 p.value
## chol     6.142356e-49
```

```
## trestbps 8.365799e-35
## age      1.177774e-19
## thalach  3.597126e-08
## oldpeak  1.358359e-05
```

Individuals in this third clusters have a really high cholesterol, a high resting blood pressure(trest bps), and tend to be older than the global mean. Their overall higher heart rate and lower oldpeak suggest a better exercise condition in comparision to cluster 2.

```
res.catdes$quanti[4]
```

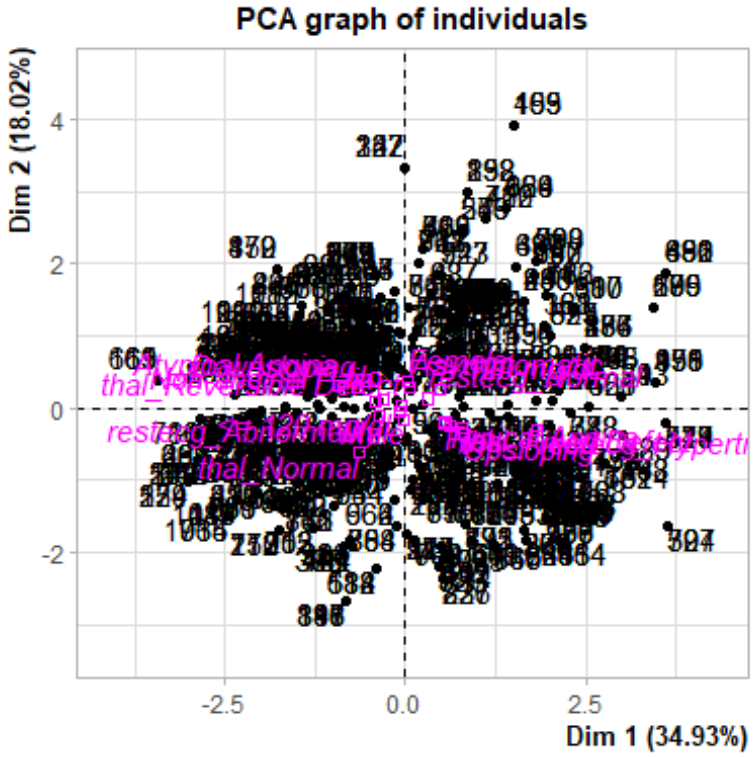
```
## $`4`
##              v.test Mean in category Overall mean sd in category
Overall sd
## thalach   17.316771      166.6220930  149.1141463     13.6512705
22.9944987
## target    13.376858        0.7044186    0.5363902      0.2223205
0.2856828
## chol      -7.297723      229.4534884  246.0000000     37.5569047
51.5673370
## trestbps  -8.822846      124.8197674  131.6117073     12.0653291
17.5081712
## ca       -13.996592        0.1104651    0.6878049      0.3401535
0.9381339
## oldpeak  -14.045024        0.3462209    1.0715122      0.6695518
1.1744799
## age      -20.490477       46.2645349   54.4341463      6.6910491
9.0678636
##               p.value
## thalach  3.515384e-67
## target   8.256497e-41
## chol     2.926779e-13
## trestbps 1.115869e-18
## ca       1.635273e-44
## oldpeak  8.264003e-45
## age      2.618166e-93
```

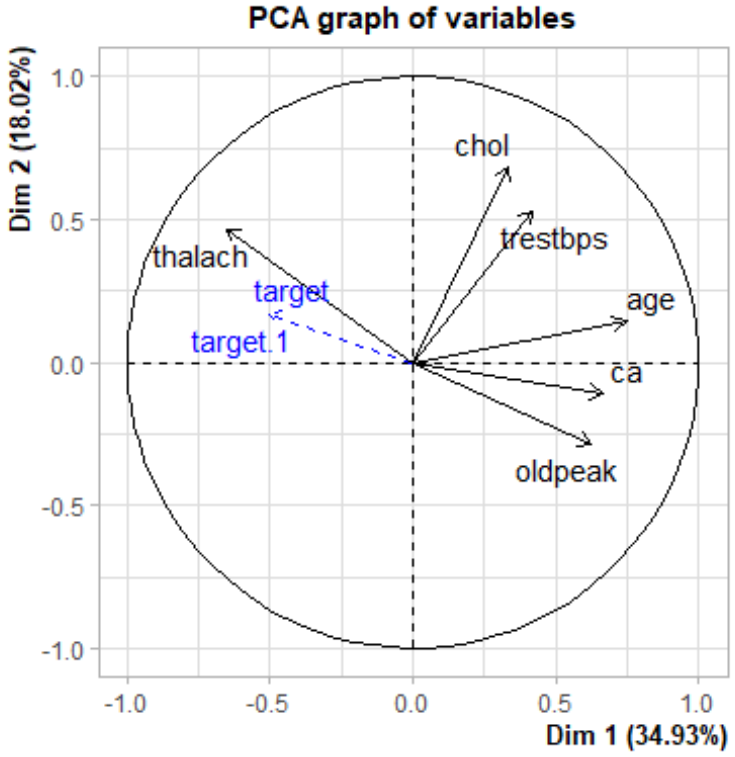
Finally, indivuals in the fourth cluster have a similar to the mean oldpeak, low max heart rate and low cholesterol. Age seems to not be relevant in individuals that ifluence this cluster.

Hierarchical Clustering (we will do it using the res.pca)

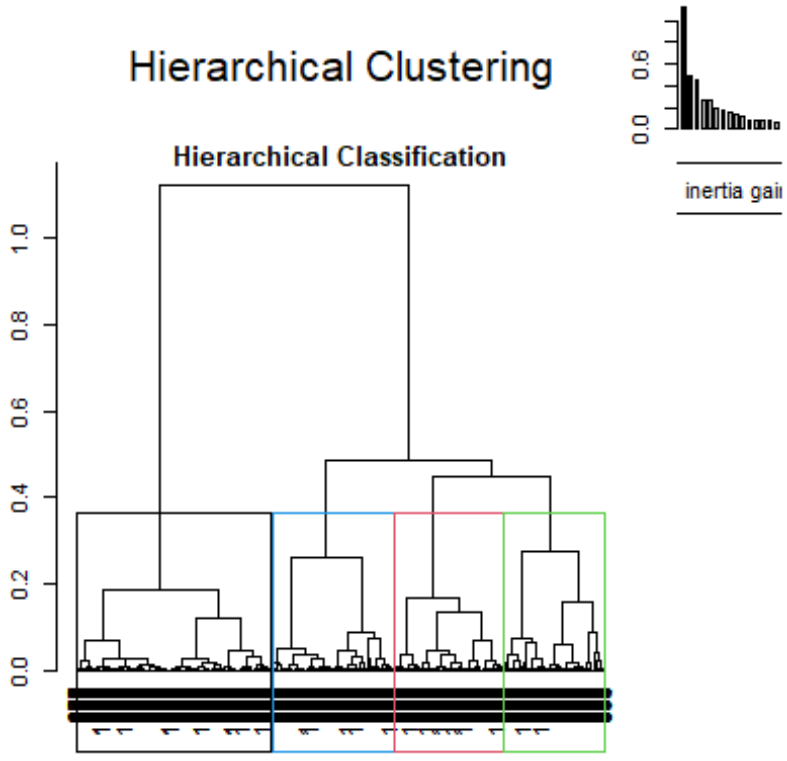
We continue now with hierarchical clustering. The function we will use is in the package FactoMineR. This function can work already with the result of the PCA function.

```
res.pca<-PCA(df[,c(vars_res, vars_dis, vars_con)], quali.sup=c(2:8),
quanti.sup= c(1,15))
```

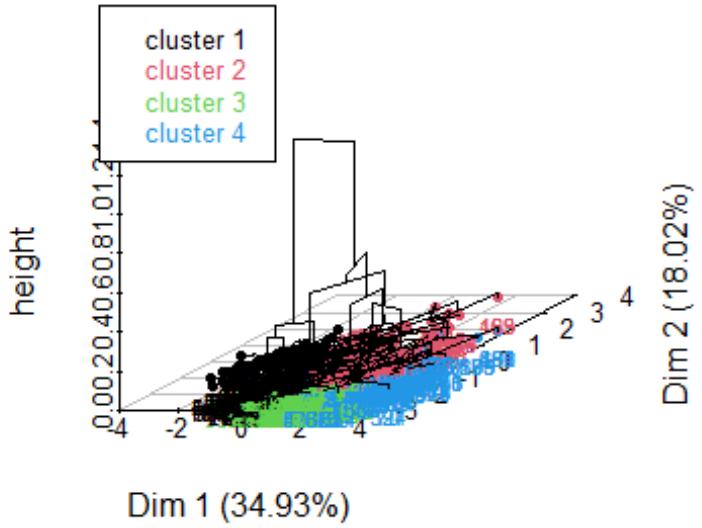


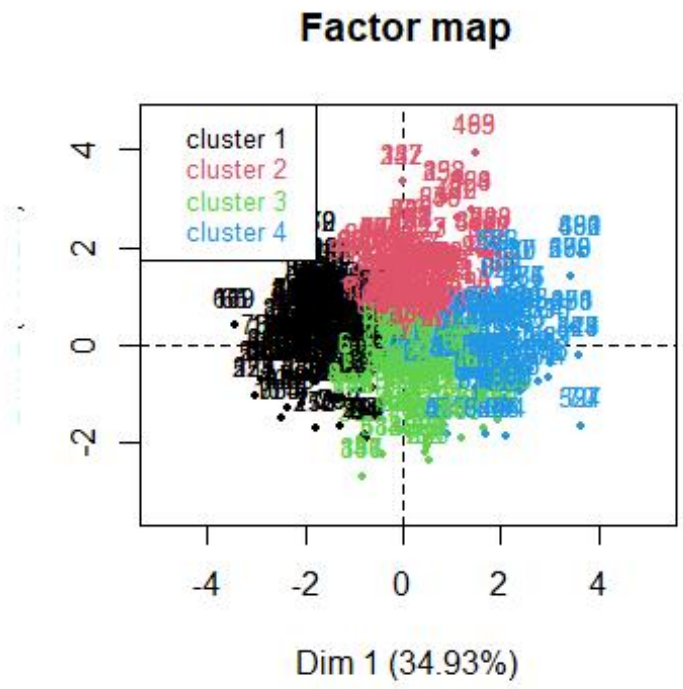



```
res.hcpc <- HCPC(res.pca, nb.clust=-1)
```



Hierarchical clustering on the factor map






```
res.hcpc

## **Results for the Hierarchical Clustering on Principal Components**
##    name
## 1  "$data.clust"
## 2  "$desc.var"
## 3  "$desc.var$quanti.var"
## 4  "$desc.var$quanti"
## 5  "$desc.var$test.chi2"
## 6  "$desc.axes$category"
## 7  "$desc.axes"
## 8  "$desc.axes$quanti.var"
## 9  "$desc.axes$quanti"
## 10 "$desc.ind"
## 11 "$desc.ind$para"
## 12 "$desc.ind$dist"
## 13 "$call"
## 14 "$call$t"
##    description
## 1  "dataset with the cluster of the individuals"
## 2  "description of the clusters by the variables"
## 3  "description of the cluster var. by the continuous var."
## 4  "description of the clusters by the continuous var."
## 5  "description of the cluster var. by the categorical var."
## 6  "description of the clusters by the categories."
## 7  "description of the clusters by the dimensions"
## 8  "description of the cluster var. by the axes"
```

```
## 9  "description of the clusters by the axes"
## 10 "description of the clusters by the individuals"
## 11 "parangons of each clusters"
## 12 "specific individuals"
## 13 "summary statistics"
## 14 "description of the tree"
```

Global View of the Results

```
head(res.hcpc$data.clust$clust, 100)
```

```
##   [1] 1 4 4 2 4 3 4 2 1 4 3 2 1 4 1 1 2 3 1 1 4 3 1 2 1 2 1 2 2 4 1 1
3 4 1 4 3
##  [38] 2 3 3 2 1 2 1 2 1 1 2 2 3 2 3 1 4 4 4 1 1 2 2 1 2 3 2 1 2 1 2 1
4 4 4 3 1
##  [75] 2 1 1 4 1 1 1 3 4 1 1 1 1 2 4 4 1 1 4 1 3 1 2 4 3 3
## Levels: 1 2 3 4
```

```
table(res.hcpc$data.clust$clust)
```

```
##
##   1   2   3   4
## 363 274 196 192
```

```
res.hcpc$desc.var
```

```
##
## Link between the cluster variable and the categorical variables (chi-
square test)
##
================================================================================
========
##              p.value df
## slope   8.187753e-42  6
## cp      4.362362e-33  9
## exang   2.980882e-23  3
## restecg 3.698370e-14  6
## sex     5.422185e-06  3
## thal    1.213876e-05  6
## fbs     1.765561e-05  3
##
## Description of each cluster by the categories
## ============================================
## $`1`
##                                Cla/Mod   Mod/Cla    Global      p.value
v.test
## slope=Downsloping           51.17271 66.115702 45.756098 2.464308e-22
9.720557
## exang=No                    44.55882 83.471074 66.341463 7.803736e-19
8.862789
## restecg=restecg_Abnormality 46.00390 65.013774 50.048780 1.070864e-12
7.121078
```

```
## cp=Atypical Angina          58.08383 26.721763 16.292683 6.521172e-11
6.531276
## cp=Non-anginal Pain         46.83099 36.639118 27.707317 2.987965e-06
4.671645
## fbs=<= 120 mg/dL            37.95872 91.184573 85.073171 2.857914e-05
4.184502
## thal=thal_Reversible Defect 36.43162 93.939394 91.317073 2.521883e-02
2.238034
## thal=thal_Fixed Defect      21.95122  4.958678  8.000000 6.571540e-03
-2.717811
## restecg=restecg_Hypertrophy  0.00000  0.000000  1.463415 1.340663e-03
-3.207130
## fbs=> 120 mg/dL             20.91503  8.815427 14.926829 2.857914e-05
-4.184502
## restecg=restecg_Normal      25.55332 34.986226 48.487805 1.289042e-10
-6.428459
## cp=Typical Angina           22.73642 31.129477 48.487805 1.088856e-16
-8.294672
## exang=Yes                   17.39130 16.528926 33.658537 7.803736e-19
-8.862789
## slope=Flat                  20.53942 27.272727 47.024390 2.179908e-21
-9.496056
##
## $`2`
##                               Cla/Mod   Mod/Cla    Global      p.value
## sex=Female                  37.179487 42.335766 30.439024 9.384319e-07
## fbs=> 120 mg/dL             41.176471 22.992701 14.926829 2.517890e-05
## cp=Asymptomatic             45.454545 12.773723  7.512195 2.508028e-04
## restecg=restecg_Normal      31.388330 56.934307 48.487805 1.104678e-03
## slope=Downsloping           31.130064 53.284672 45.756098 3.592801e-03
## thal=thal_Reversible Defect 27.670940 94.525547 91.317073 2.370529e-02
## restecg=restecg_Hypertrophy  0.000000  0.000000  1.463415 9.064131e-03
## restecg=restecg_Abnormality 23.001949 43.065693 50.048780 6.997787e-03
## cp=Typical Angina           21.529175 39.051095 48.487805 2.578903e-04
## slope=Upsloping              8.108108  2.189781  7.219512 4.170175e-05
## fbs=<= 120 mg/dL            24.197248 77.007299 85.073171 2.517890e-05
## sex=Male                    22.159888 57.664234 69.560976 9.384319e-07
##                                v.test
## sex=Female                   4.904128
## fbs=> 120 mg/dL              4.213190
## cp=Asymptomatic              3.661439
## restecg=restecg_Normal       3.262414
## slope=Downsloping            2.911863
## thal=thal_Reversible Defect  2.261872
## restecg=restecg_Hypertrophy -2.609626
## restecg=restecg_Abnormality -2.696950
## cp=Typical Angina           -3.654294
## slope=Upsloping             -4.097843
## fbs=<= 120 mg/dL            -4.213190
## sex=Male                    -4.904128
```

```
##
## $`3`
##                               Cla/Mod   Mod/Cla    Global      p.value
v.test
## exang=Yes                   29.85507 52.551020 33.658537 1.286263e-09
6.069103
## slope=Flat                  26.55602 65.306122 47.024390 1.165512e-08
5.704696
## cp=Typical Angina           24.94970 63.265306 48.487805 4.124002e-06
4.605033
## restecg=restecg_Hypertrophy 46.66667  3.571429  1.463415 1.645424e-02
2.398681
## cp=Atypical Angina          10.77844  9.183673 16.292683 1.746969e-03
-3.130184
## slope=Downsloping           12.79318 30.612245 45.756098 1.769884e-06
-4.778067
## exang=No                    13.67647 47.448980 66.341463 1.286263e-09
-6.069103
##
## $`4`
##                                 Cla/Mod   Mod/Cla    Global      p.value
## cp=Typical Angina           30.784708 79.687500 48.487805 1.100521e-22
## slope=Flat                  27.593361 69.270833 47.024390 6.109168e-12
## slope=Upsloping             48.648649 18.750000  7.219512 1.184655e-09
## exang=Yes                   28.695652 51.562500 33.658537 1.337124e-08
## thal=thal_Fixed Defect      39.024390 16.666667  8.000000 7.388695e-06
## restecg=restecg_Normal      23.138833 59.895833 48.487805 4.597031e-04
## restecg=restecg_Hypertrophy 53.333333  4.166667  1.463415 2.973958e-03
## sex=Male                    21.037868 78.125000 69.560976 3.599587e-03
## sex=Female                  13.461538 21.875000 30.439024 3.599587e-03
## thal=thal_Reversible Defect 17.094017 83.333333 91.317073 5.756389e-05
## restecg=restecg_Abnormality 13.450292 35.937500 50.048780 1.382248e-05
## exang=No                    13.676471 48.437500 66.341463 1.337124e-08
## cp=Non-anginal Pain          8.098592 11.979167 27.707317 1.010821e-08
## cp=Atypical Angina           1.796407  1.562500 16.292683 5.234794e-13
## slope=Downsloping            4.904051 11.979167 45.756098 3.571944e-28
##                                 v.test
## cp=Typical Angina             9.802299
## slope=Flat                    6.877080
## slope=Upsloping               6.082307
## exang=Yes                     5.681252
## thal=thal_Fixed Defect        4.482173
## restecg=restecg_Normal        3.503201
## restecg=restecg_Hypertrophy   2.970417
## sex=Male                      2.911274
## sex=Female                   -2.911274
## thal=thal_Reversible Defect  -4.022581
## restecg=restecg_Abnormality  -4.346662
## exang=No                     -5.681252
## cp=Non-anginal Pain          -5.728903
```

```
## cp=Atypical Angina           -7.219061
## slope=Downsloping           -11.006084
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                Eta2       P-value
## thalach  0.4997493 4.919041e-153
## oldpeak  0.4962024 1.806163e-151
## age      0.4703821 2.120391e-140
## ca       0.3378973  5.680376e-91
## trestbps 0.3282940  8.725125e-88
## chol     0.2347062  6.133439e-59
## target   0.2249908  3.762307e-56
## target.1 0.2249908  3.762307e-56
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##              v.test Mean in category Overall mean sd in category
Overall sd
## thalach   17.165341      165.7713499  149.1141463     14.0882717
22.9944987
## target.1  12.802854        0.6907438    0.5363902      0.2309836
0.2856828
## target    12.802854        0.6907438    0.5363902      0.2309836
0.2856828
## trestbps  -9.149772      124.8512397  131.6117073     12.1240111
17.5081712
## chol      -9.628293      225.0468320  246.0000000     36.7901578
51.5673370
## oldpeak  -12.064233        0.4735537    1.0715122      0.7936492
1.1744799
## ca       -13.476351        0.1542700    0.6878049      0.4307751
0.9381339
## age      -21.838374       46.0771350   54.4341463      6.5706251
9.0678636
##                p.value
## thalach   4.826379e-66
## target.1  1.580337e-37
## target    1.580337e-37
## trestbps  5.705407e-20
## chol      6.073004e-22
## oldpeak   1.631749e-33
## ca        2.155044e-41
## age      1.002594e-105
##
## $`2`
##             v.test Mean in category Overall mean sd in category
Overall sd
```

```
## chol     14.147658      283.7445255   246.000000     54.8747784
51.567337
## trestbps 11.347680      141.8905109   131.611707     16.4435237
17.508171
## age       9.421274       58.8540146    54.434146      6.4723332
9.067864
## thalach   5.389324      155.5255474   149.114146     14.9917643
22.994499
## oldpeak  -5.975958        0.7083942     1.071512      0.7017462
1.174480
##               p.value
## chol     1.930807e-45
## trestbps 7.615671e-30
## age      4.456475e-21
## thalach  7.072321e-08
## oldpeak  2.287424e-09
##
## $`3`
##              v.test Mean in category Overall mean sd in category
Overall sd
## age        6.381342       58.1530612   54.4341463      7.5232321
9.0678636
## target.1  -3.924386        0.4643367    0.5363902      0.2827995
0.2856828
## target    -3.924386        0.4643367    0.5363902      0.2827995
0.2856828
## chol      -6.159402      225.5867347  246.0000000     36.7595278
51.5673370
## trestbps -11.516476      118.6530612  131.6117073     10.7196197
17.5081712
## thalach  -15.102071      126.7959184  149.1141463     19.7204748
22.9944987
##               p.value
## age      1.755430e-10
## target.1 8.695111e-05
## target   8.695111e-05
## chol     7.302032e-10
## trestbps 1.089778e-30
## thalach  1.569257e-51
##
## $`4`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak   21.846866        2.7416667    1.0715122      1.0965064
1.1744799
## ca        16.712383        1.7083333    0.6878049      1.0744831
0.9381339
## trestbps   9.951869      142.9531250  131.6117073     17.8364354
17.5081712
## age        9.650456       60.1302083   54.4341463      5.7706806
```

```
9.0678636
## chol       1.962875      252.5885417  246.0000000     51.2695336
51.5673370
## target.1 -11.743181        0.3180208    0.5363902      0.1839419
0.2856828
## target   -11.743181        0.3180208    0.5363902      0.1839419
0.2856828
## thalach  -11.931905      131.2552083  149.1141463     17.8629760
22.9944987
##                 p.value
## oldpeak  8.325460e-106
## ca        1.064928e-62
## trestbps  2.474907e-23
## age       4.893781e-22
## chol      4.966068e-02
## target.1  7.655179e-32
## target    7.655179e-32
## thalach   8.070548e-33
```

Test of the association of categorical variables

```
res.hcpc$desc.var$test.chi2

##              p.value df
## slope   8.187753e-42  6
## cp      4.362362e-33  9
## exang   2.980882e-23  3
## restecg 3.698370e-14  6
## sex     5.422185e-06  3
## thal    1.213876e-05  6
## fbs     1.765561e-05  3
```

slope (p = 8.18e-42, df = 6): The extremely small p-value suggests that the variable 'slope' has a statistically significant association with the clusters. With 6 degrees of freedom, which suggests that the variable has several categories influencing how data is grouped into these clusters, since df > number of clusters.

cp (p= 4.36e-33, df = 9): Again, a very small p-value indicating strong statistical significance. With 9 degrees of freedom, which suggests that the variable has several categories influencing how data is grouped into these clusters.

exang (p = 2.98e-23, df = 3): A significant p-value with fewer degrees of freedom (3), suggesting a strong association between 'exang' and the clusters.

restecg (p = 3.69e-14, df = 6): This also shows a significant relationship but with a slightly higher p-value compared to the first three variables, indicating a somewhat less strong association with the clusters.

sex (p = 5.42e-06, df = 3): While still statistically significant, 'sex' has the least strong association among the variables listed,since it has the lower p-value. And with only 3 degrees of freedom.

thal (p = 1.21e-05, df = 6): 'thal' shows a significant but less strong association compared to the top variables, but with more degrees of freedom (6) than some of them, which suggests that the variable has several categories influencing how data is grouped into these clusters.

fbs (p = 1.76e-05, df = 3): Significant p-value, since it is smaller than 0.05, but less so than the top variables. And with 2 degrees of freedom.

Tests of categorical association

```
res.hcpc$desc.var$category[1]
```

```
## $`1`
##                                Cla/Mod   Mod/Cla    Global      p.value
v.test
## slope=Downsloping           51.17271 66.115702 45.756098 2.464308e-22
9.720557
## exang=No                    44.55882 83.471074 66.341463 7.803736e-19
8.862789
## restecg=restecg_Abnormality 46.00390 65.013774 50.048780 1.070864e-12
7.121078
## cp=Atypical Angina          58.08383 26.721763 16.292683 6.521172e-11
6.531276
## cp=Non-anginal Pain         46.83099 36.639118 27.707317 2.987965e-06
4.671645
## fbs=<= 120 mg/dL            37.95872 91.184573 85.073171 2.857914e-05
4.184502
## thal=thal_Reversible Defect 36.43162 93.939394 91.317073 2.521883e-02
2.238034
## thal=thal_Fixed Defect      21.95122  4.958678  8.000000 6.571540e-03
-2.717811
## restecg=restecg_Hypertrophy  0.00000  0.000000  1.463415 1.340663e-03
-3.207130
## fbs=> 120 mg/dL             20.91503  8.815427 14.926829 2.857914e-05
-4.184502
## restecg=restecg_Normal      25.55332 34.986226 48.487805 1.289042e-10
-6.428459
## cp=Typical Angina           22.73642 31.129477 48.487805 1.088856e-16
-8.294672
## exang=Yes                   17.39130 16.528926 33.658537 7.803736e-19
-8.862789
## slope=Flat                  20.53942 27.272727 47.024390 2.179908e-21
-9.496056
```

In cluster 1: We can detect several variables that have a strong influence in it. The first is the 'slope=Downsloping', since it has a very high and positive v-test value. And also we can see that the mean in the cluster is significantly higher than the global one. This same, occurs in 'exang=No' and 'res.tecg=restecg_Abnormality'.

And then the variable that are less characteristic, are the 'exang=Yes', 'slope=Flat' and 'cp=Typical Angina', in contrast of the variables with strong influence, have significant negative v.test values and the mean in the cluster is significantly lower than the global one, suggesting they are not typical of this cluster.

```
res.hcpc$desc.var$category[2]
```

```
## $`2`
##                                Cla/Mod   Mod/Cla    Global      p.value
## sex=Female                   37.179487 42.335766 30.439024 9.384319e-07
## fbs=> 120 mg/dL              41.176471 22.992701 14.926829 2.517890e-05
## cp=Asymptomatic              45.454545 12.773723  7.512195 2.508028e-04
## restecg=restecg_Normal       31.388330 56.934307 48.487805 1.104678e-03
## slope=Downsloping            31.130064 53.284672 45.756098 3.592801e-03
## thal=thal_Reversible Defect  27.670940 94.525547 91.317073 2.370529e-02
## restecg=restecg_Hypertrophy   0.000000  0.000000  1.463415 9.064131e-03
## restecg=restecg_Abnormality  23.001949 43.065693 50.048780 6.997787e-03
## cp=Typical Angina            21.529175 39.051095 48.487805 2.578903e-04
## slope=Upsloping               8.108108  2.189781  7.219512 4.170175e-05
## fbs=<= 120 mg/dL             24.197248 77.007299 85.073171 2.517890e-05
## sex=Male                     22.159888 57.664234 69.560976 9.384319e-07
##                                 v.test
## sex=Female                    4.904128
## fbs=> 120 mg/dL               4.213190
## cp=Asymptomatic               3.661439
## restecg=restecg_Normal        3.262414
## slope=Downsloping             2.911863
## thal=thal_Reversible Defect   2.261872
## restecg=restecg_Hypertrophy  -2.609626
## restecg=restecg_Abnormality  -2.696950
## cp=Typical Angina            -3.654294
## slope=Upsloping              -4.097843
## fbs=<= 120 mg/dL             -4.213190
## sex=Male                     -4.904128
```

In cluster 2: Here we can see a gender influence, since 'sex=Female' has a high positive v-test and a mean in the cluster is significantly higher than the global one. And the 'sex=Male' is a high negative v-test and a mean in the cluster is significantly lower than the global one.

Also a good point to mention is that 'cp=Asymptomatic' and 'restecg=restecg_Normal' are characteristic of this cluster, with positive v.test values and with a mean in the cluster is significantly higher than the global one.

And as well as the gender, the blood sugar levels are also significant, 'fbs=<= 120 mg/dL' and 'fbs=> 120 mg/dL' with corresponding positive and negative influences.

```
res.hcpc$desc.var$category[3]
```

```
## $`3`
##                                Cla/Mod   Mod/Cla    Global      p.value
```

```
v.test
## exang=Yes                    29.85507 52.551020 33.658537 1.286263e-09
6.069103
## slope=Flat                   26.55602 65.306122 47.024390 1.165512e-08
5.704696
## cp=Typical Angina            24.94970 63.265306 48.487805 4.124002e-06
4.605033
## restecg=restecg_Hypertrophy 46.66667  3.571429  1.463415 1.645424e-02
2.398681
## cp=Atypical Angina           10.77844  9.183673 16.292683 1.746969e-03
-3.130184
## slope=Downsloping            12.79318 30.612245 45.756098 1.769884e-06
-4.778067
## exang=No                     13.67647 47.448980 66.341463 1.286263e-09
-6.069103
```

In cluster 3: Catgories like exang=Yes, slope=Flat, and cp=Typical Angina are significantly more characteristic of this cluster, indicating specific profiles or conditions more prevalent within this cluster compared to the overall dataset. Since they have a mean in the cluster significantly higher than the global one. Also all of them have a very high positive v-test value suggesting, again that those categories are more prevalent in this cluster than globally.

Conversely, catgories like cp=Atypical Angina, slope=Downsloping, and exang=No are less characteristic of this cluster, occurring less frequently than expected globally.

```
res.hcpc$desc.var$category[4]

## $`4`
##                               Cla/Mod   Mod/Cla    Global      p.value
## cp=Typical Angina           30.784708 79.687500 48.487805 1.100521e-22
## slope=Flat                  27.593361 69.270833 47.024390 6.109168e-12
## slope=Upsloping             48.648649 18.750000  7.219512 1.184655e-09
## exang=Yes                   28.695652 51.562500 33.658537 1.337124e-08
## thal=thal_Fixed Defect      39.024390 16.666667  8.000000 7.388695e-06
## restecg=restecg_Normal      23.138833 59.895833 48.487805 4.597031e-04
## restecg=restecg_Hypertrophy 53.333333  4.166667  1.463415 2.973958e-03
## sex=Male                    21.037868 78.125000 69.560976 3.599587e-03
## sex=Female                  13.461538 21.875000 30.439024 3.599587e-03
## thal=thal_Reversible Defect 17.094017 83.333333 91.317073 5.756389e-05
## restecg=restecg_Abnormality 13.450292 35.937500 50.048780 1.382248e-05
## exang=No                    13.676471 48.437500 66.341463 1.337124e-08
## cp=Non-anginal Pain          8.098592 11.979167 27.707317 1.010821e-08
## cp=Atypical Angina           1.796407  1.562500 16.292683 5.234794e-13
## slope=Downsloping            4.904051 11.979167 45.756098 3.571944e-28
##                                 v.test
## cp=Typical Angina             9.802299
## slope=Flat                    6.877080
## slope=Upsloping               6.082307
## exang=Yes                     5.681252
```

```
## thal=thal_Fixed Defect        4.482173
## restecg=restecg_Normal        3.503201
## restecg=restecg_Hypertrophy   2.970417
## sex=Male                      2.911274
## sex=Female                   -2.911274
## thal=thal_Reversible Defect  -4.022581
## restecg=restecg_Abnormality  -4.346662
## exang=No                     -5.681252
## cp=Non-anginal Pain          -5.728903
## cp=Atypical Angina           -7.219061
## slope=Downsloping           -11.006084
```

In cluster 4: In this fourth cluster cp=Typical Angina: Dominates the cluster, indicating that typical angina is a primary symptom for many within this group. This suggests a specific cardiac issue or pain profile that is prevalent.

Slope Characteristics (Flat and Upsloping): Both flat and upsloping responses during exercise tests are significantly more common in this cluster compared to the global average, highlighting unique physiological responses to exercise among these patients.

Then, thal=thal_Fixed Defect: A higher prevalence of fixed defects in thallium stress tests characterizes this cluster, pointing to certain types of heart muscle damage or abnormalities.

The Electrocardiographic Findings: The cluster shows a higher occurrence of normal and hypertrophic findings in resting ECGs, indicating specific electrocardiographic profiles.

Finally, Gender (Male): Males are more prevalent in this cluster, suggesting possible gender-related susceptibility or risk factors in cardiac conditions featured within the group.

On the side of the less characteristic categories, we can see the thal=thal_Reversible Defect, Restecg Abnormalities, Absence of Exercise-Induced Angina (exang=No), Chest Pain Types (Non-anginal Pain and Atypical Angina) and Slope=Downsloping.

Global tests of numerical association.

```
res.hcpc$desc.var$quanti.var
```

```
##                Eta2       P-value
## thalach  0.4997493 4.919041e-153
## oldpeak  0.4962024 1.806163e-151
## age      0.4703821 2.120391e-140
## ca       0.3378973  5.680376e-91
## trestbps 0.3282940  8.725125e-88
## chol     0.2347062  6.133439e-59
## target   0.2249908  3.762307e-56
## target.1 0.2249908  3.762307e-56
```

To begin with, the thalach (Maximum heart rate achieved), has a very high eta2, suggesting that maximum heart rate significantly differentiates the clusters. This could imply that different clusters may be characterized by differing levels of physical fitness or cardiac function.

For the variable oldpeak, which represents the ST depression induced by exercise relative to rest, has as well as the thalach a high eta2 value, indicating a strong influence of exercise-induced ST depression on cluster differentiation, which might relate to underlying differences in cardiac ischemia among clusters.

The eta2 value for age is also high, suggesting significant variation in age across clusters, which could reflect differing disease prevalence or risk factors across age groups.

The eta2 value that we got for the Ca variable (Number of major vessels colored by fluoroscopy), it still being high indicating notable differences in the prevalence of observable coronary artery disease across clusters.

For the trestbps (Resting blood pressure) variable, we got a similar value as the Ca, which indicates a considerable influence of resting blood pressure on clustering, possibly reflecting different cardiovascular risk profiles.

And now,the following variable chol (Serum cholesterol), we find lower eta2 values than the others but still notable, suggesting variations in cholesterol levels influence cluster characteristics.

Specific tests of numerical association.

```
res.hcpc$desc.var$quanti[1]
```

```
## $`1`
##              v.test Mean in category Overall mean sd in category
Overall sd
## thalach   17.165341      165.7713499  149.1141463     14.0882717
22.9944987
## target.1  12.802854        0.6907438    0.5363902      0.2309836
0.2856828
## target    12.802854        0.6907438    0.5363902      0.2309836
0.2856828
## trestbps  -9.149772      124.8512397  131.6117073     12.1240111
17.5081712
## chol      -9.628293      225.0468320  246.0000000     36.7901578
51.5673370
## oldpeak  -12.064233        0.4735537    1.0715122      0.7936492
1.1744799
## ca       -13.476351        0.1542700    0.6878049      0.4307751
0.9381339
## age      -21.838374       46.0771350   54.4341463      6.5706251
9.0678636
##                p.value
```

```
## thalach   4.826379e-66
## target.1  1.580337e-37
## target    1.580337e-37
## trestbps  5.705407e-20
## chol      6.073004e-22
## oldpeak   1.631749e-33
## ca        2.155044e-41
## age      1.002594e-105
```

Cluster 1 is characterized by younger individuals with higher heart rates, a higher presence of the target condition, but lower blood pressure, cholesterol, ST depression, and fewer major vessels detected by fluoroscopy compared to the overall dataset.

This profile could suggest a subgroup with distinct cardiovascular health characteristics, possibly at different stages of cardiovascular disease or with different risk factors compared to the average population in the study.

```
res.hcpc$desc.var$quanti[2]
```

```
## $`2`
##             v.test Mean in category Overall mean sd in category
Overall sd
## chol     14.147658      283.7445255   246.000000     54.8747784
51.567337
## trestbps 11.347680      141.8905109   131.611707     16.4435237
17.508171
## age       9.421274       58.8540146    54.434146      6.4723332
9.067864
## thalach   5.389324      155.5255474   149.114146     14.9917643
22.994499
## oldpeak  -5.975958        0.7083942     1.071512      0.7017462
1.174480
##                 p.value
## chol     1.930807e-45
## trestbps 7.615671e-30
## age      4.456475e-21
## thalach  7.072321e-08
## oldpeak  2.287424e-09
```

In the case of cluster 2 is characterized by individuals who are generally older with significantly higher cholesterol and resting blood pressure, suggesting a profile potentially indicative of higher cardiovascular risk. However, this group shows somewhat higher maximum heart rates and notably lower ST depressions during exercise, which could indicate varying degrees of cardiovascular fitness or response to exercise despite other risk factors. These attributes underscore the cluster's distinct cardiovascular health characteristics, which could be pivotal for targeted interventions or further study into risk management for these individuals.

```
res.hcpc$desc.var$quanti[3]
```

```
## $`3`
##              v.test Mean in category Overall mean sd in category
Overall sd
## age        6.381342       58.1530612   54.4341463      7.5232321
9.0678636
## target.1  -3.924386        0.4643367    0.5363902      0.2827995
0.2856828
## target    -3.924386        0.4643367    0.5363902      0.2827995
0.2856828
## chol      -6.159402      225.5867347  246.0000000     36.7595278
51.5673370
## trestbps -11.516476      118.6530612  131.6117073     10.7196197
17.5081712
## thalach  -15.102071      126.7959184  149.1141463     19.7204748
22.9944987
##                p.value
## age      1.755430e-10
## target.1 8.695111e-05
## target   8.695111e-05
## chol     7.302032e-10
## trestbps 1.089778e-30
## thalach  1.569257e-51
```

For cluster 3 is characterized by older individuals with lower cholesterol, lower resting blood pressure, and lower maximum heart rates compared to the overall dataset.

Additionally, there is a lower presence of the target condition (likely indicating a lower prevalence of a specific health outcome) within this cluster. These findings suggest that Cluster 3 may represent a subgroup with relatively better cardiovascular health or a different risk profile compared to other clusters in the dataset.

```
res.hcpc$desc.var$quanti[4]
```

```
## $`4`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak   21.846866        2.7416667    1.0715122      1.0965064
1.1744799
## ca        16.712383        1.7083333    0.6878049      1.0744831
0.9381339
## trestbps   9.951869      142.9531250  131.6117073     17.8364354
17.5081712
## age        9.650456       60.1302083   54.4341463      5.7706806
9.0678636
## chol       1.962875      252.5885417  246.0000000     51.2695336
51.5673370
## target.1 -11.743181        0.3180208    0.5363902      0.1839419
0.2856828
## target   -11.743181        0.3180208    0.5363902      0.1839419
```

```
0.2856828
## thalach  -11.931905      131.2552083  149.1141463     17.8629760
22.9944987
##                 p.value
## oldpeak  8.325460e-106
## ca        1.064928e-62
## trestbps  2.474907e-23
## age       4.893781e-22
## chol      4.966068e-02
## target.1  7.655179e-32
## target    7.655179e-32
## thalach   8.070548e-33
```

Cluster 4 is characterized by individuals who tend to be older and exhibit significantly higher ST depressions during exercise, higher resting blood pressure, and a higher presence of major vessels colored by fluoroscopy compared to the overall dataset. However, despite these cardiovascular risk factors, individuals in this cluster demonstrate slightly higher cholesterol levels.

Additionally, there is a lower presence of the target condition (likely indicating a lower prevalence of a specific health outcome) and lower maximum heart rates in this cluster.

These findings suggest a subgroup with distinct cardiovascular health characteristics, possibly indicative of more advanced cardiovascular disease or different risk profiles compared to other clusters in the dataset.

Component Analysis (CA)

```
library(FactoMineR)
library(car)
head(df)
```

```
##   age    sex             cp trestbps chol          fbs     restecg
thalach
## 1  52   Male Typical Angina      125  212 <= 120 mg/dL Abnormality
168
## 2  53   Male Typical Angina      140  203  > 120 mg/dL      Normal
155
## 3  70   Male Typical Angina      145  174 <= 120 mg/dL Abnormality
125
## 4  61   Male Typical Angina      148  203 <= 120 mg/dL Abnormality
161
## 5  62 Female Typical Angina      138  294  > 120 mg/dL Abnormality
106
## 6  58 Female Typical Angina      100  248 <= 120 mg/dL      Normal
122
##   exang oldpeak       slope ca              thal target   mout kmeaclu
## 1    No     1.0 Downsloping  2 Reversible Defect   0.23 NoMOut       2
## 2   Yes     3.1   Upsloping  0      Fixed Defect   0.37 NoMOut       2
```

```
## 3   Yes     2.6   Upsloping  0 Reversible Defect   0.24 NoMOut       1
## 4    No     0.0 Downsloping  1 Reversible Defect   0.28 NoMOut       3
## 5    No     1.9        Flat  3 Reversible Defect   0.21 NoMOut       1
## 6    No     1.0        Flat  0 Reversible Defect   0.78 NoMOut       2
```

```
summary(df)
```

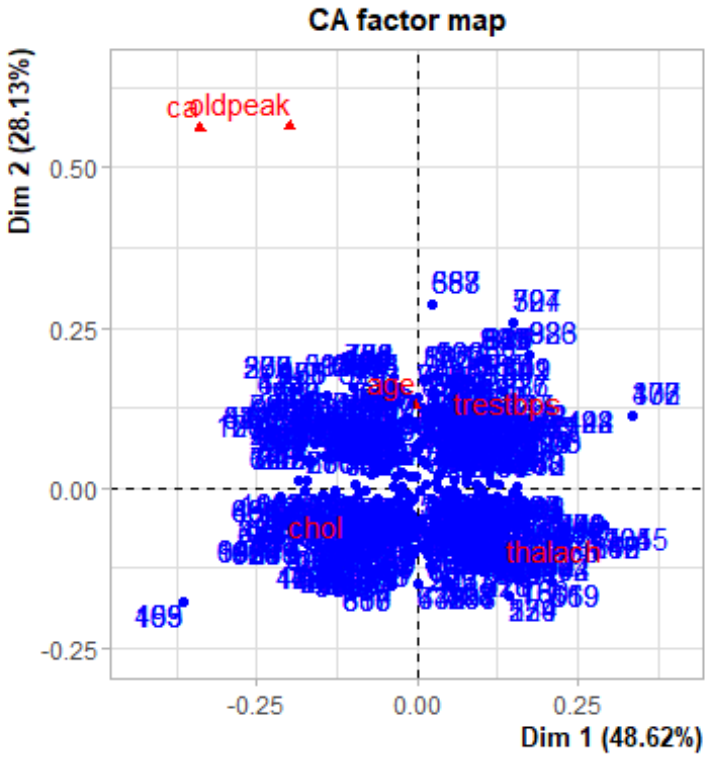
```
##       age            sex                     cp         trestbps
##  Min.   :29.00   Female:312   Asymptomatic    : 77   Min.   : 94.0
##  1st Qu.:48.00   Male  :713   Atypical Angina :167   1st Qu.:120.0
##  Median :56.00                Non-anginal Pain:284   Median :130.0
##  Mean   :54.43                Typical Angina  :497   Mean   :131.6
##  3rd Qu.:61.00                                       3rd Qu.:140.0
##  Max.   :77.00                                       Max.   :200.0
##       chol               fbs             restecg       thalach
exang
##  Min.   :126   <= 120 mg/dL:872   Abnormality:513   Min.   : 71.0   No
:680
##  1st Qu.:211   > 120 mg/dL :153   Hypertrophy: 15   1st Qu.:132.0
Yes:345
##  Median :240                      Normal     :497   Median :152.0
##  Mean   :246                                        Mean   :149.1
##  3rd Qu.:275                                        3rd Qu.:166.0
##  Max.   :564                                        Max.   :202.0
##     oldpeak              slope           ca
thal
##  Min.   :0.000   Downsloping:469   Min.   :0.0000   Fixed Defect     :
82
##  1st Qu.:0.000   Flat       :482   1st Qu.:0.0000   Normal           :
7
##  Median :0.800   Upsloping  : 74   Median :0.0000   Reversible
Defect:936
##  Mean   :1.072                     Mean   :0.6878
##  3rd Qu.:1.800                     3rd Qu.:1.0000
##  Max.   :6.200                     Max.   :3.0000
##      target            mout     kmeaclu
##  Min.   :0.1000   NoMOut :932   1:206
##  1st Qu.:0.2600   YesMOut: 93   2:233
##  Median :0.7100                 3:242
##  Mean   :0.5364                 4:344
##  3rd Qu.:0.8100
##  Max.   :0.9000
```

```
dim(df)
```

```
## [1] 1025   16
```

```
dfc<-df[,c(1,4,5,8,10,12)]
con<-c(1,4,5,8,10,12)
row.sum <- apply(dfc,1, sum)
head(row.sum, 100)
```

```
##   [1] 560.0 554.1 516.6 574.0 604.9 529.0 637.4 650.8 559.8 583.2
458.6 655.0
##  [13] 554.7 618.2 541.0 554.7 643.5 556.2 577.1 574.0 543.0 540.3
508.0 622.0
##  [25] 544.0 660.0 586.4 667.0 585.6 682.4 479.8 577.1 435.2 614.9
538.0 577.6
##  [37] 516.4 650.0 561.2 565.4 736.8 579.4 690.0 559.8 696.2 586.0
511.0 640.0
##  [49] 643.0 570.4 654.0 561.2 524.0 501.0 528.6 528.6 532.9 603.2
581.8 608.0
##  [61] 565.0 643.0 593.9 579.6 565.0 676.2 533.2 619.8 542.0 540.2
698.4 548.9
##  [73] 576.2 539.0 586.0 598.0 604.0 540.0 546.8 546.8 587.6 445.8
577.6 524.0
##  [85] 542.0 554.0 574.0 625.0 635.6 551.4 596.0 558.0 540.0 479.8
538.0 571.2
##  [97] 648.0 557.0 583.6 530.1

# Column margins
col.sum <- apply(dfc,2, sum)
col.sum

##      age trestbps     chol  thalach  oldpeak       ca
##  55795.0 134902.0 252150.0 152842.0   1098.3    705.0

# grand total
n <- sum(dfc)
n

## [1] 597492.3

res.ca <- CA(dfc)
```



At first glipse, there seems to be some kind of dependence between thalac, tresbps, age and chol, due to the proximity of their values, also between ca and oldpeak.

```
res.ca
```

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  1025  categories; the column variable has 6
categories
## The chi square of independence between the two variables is equal to
14226.7 (p-value =  0 ).
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

```
summary(res.ca)
```

```
##
## Call:
## CA(X = dfc)
##
## The chi square of independence between the two variables is equal to
14226.7 (p-value =  0 ).
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
## Variance               0.012   0.007   0.002   0.002   0.001
## % of var.             48.617  28.126   9.520   7.975   5.761
## Cumulative % of var.  48.617  76.744  86.264  94.239 100.000
##
## Rows (the 10 first)
##            Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
Dim.3
## 1        |     0.016 |  0.095  0.073  0.532 | -0.008  0.001  0.004 |
0.072
## 2        |     0.020 |  0.102  0.083  0.479 |  0.058  0.046  0.153 | -
0.018
## 3        |     0.049 |  0.099  0.074  0.174 |  0.198  0.506  0.688 | -
0.006
## 4        |     0.021 |  0.124  0.127  0.701 |  0.059  0.050  0.159 |
0.008
## 5        |     0.050 | -0.187  0.304  0.710 |  0.096  0.138  0.187 |
0.033
## 6        |     0.016 | -0.100  0.076  0.562 | -0.025  0.008  0.035 |
0.030
## 7        |     0.057 | -0.175  0.283  0.574 | -0.003  0.000  0.000 |
0.101
## 8        |     0.009 | -0.055  0.029  0.367 |  0.028  0.013  0.094 | -
0.053
## 9        |     0.004 | -0.024  0.005  0.129 | -0.054  0.041  0.660 | -
0.028
## 10       |     0.036 | -0.170  0.244  0.792 |  0.040  0.023  0.043 |
0.031
##              ctr   cos2
## 1          0.213  0.304 |
## 2          0.014  0.016 |
## 3          0.001  0.001 |
## 4          0.003  0.003 |
## 5          0.049  0.022 |
## 6          0.035  0.051 |
## 7          0.479  0.190 |
## 8          0.134  0.334 |
## 9          0.033  0.180 |
## 10         0.042  0.027 |
##
## Columns
##            Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr   cos2
```

```
Dim.3
## age      |     2.830 | -0.003  0.005  0.000 |  0.128 22.866  0.541 |
0.076
## trestbps |     3.726 |  0.059  6.805  0.211 |  0.097 31.781  0.571 | -
0.056
## chol     |     6.172 | -0.114 47.730  0.895 | -0.039  9.453  0.103 | -
0.005
## thalach  |     6.699 |  0.141 43.665  0.754 | -0.075 21.597  0.216 |
0.025
## oldpeak  |     2.232 | -0.198  0.623  0.032 |  0.565  8.759  0.263 |
0.144
## ca       |     2.151 | -0.339  1.171  0.063 |  0.561  5.544  0.173 |
0.836
##             ctr   cos2
## age      23.788  0.191 |
## trestbps 30.692  0.187 |
## chol      0.499  0.002 |
## thalach   7.000  0.024 |
## oldpeak   1.683  0.017 |
## ca       36.338  0.383 |
```

- Null hypothesis (H0): the row and the column variables of the contingency table are independent.
- Alternative hypothesis (H1): row and column variables are dependent

```
chisq.test(dfc)

## Warning in chisq.test(dfc): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  dfc
## X-squared = 14227, df = 5120, p-value < 2.2e-16
```

We see that pvalue is really small, then we can reject independence. Let's find first the value for the innertia in our data:

```
sum(res.ca$eig[,1])

## [1] 0.02381069
```

We have an innertia of 0.0238 in our data. Our maximum inertia in our dataset would be:

```
max_inertia<-min(nrow(dfc)-1, ncol(dfc)-1)
max_inertia

## [1] 5
```

How much of this inertia are we explaining?

```
# How many components to choose?
res.ca$eig
```

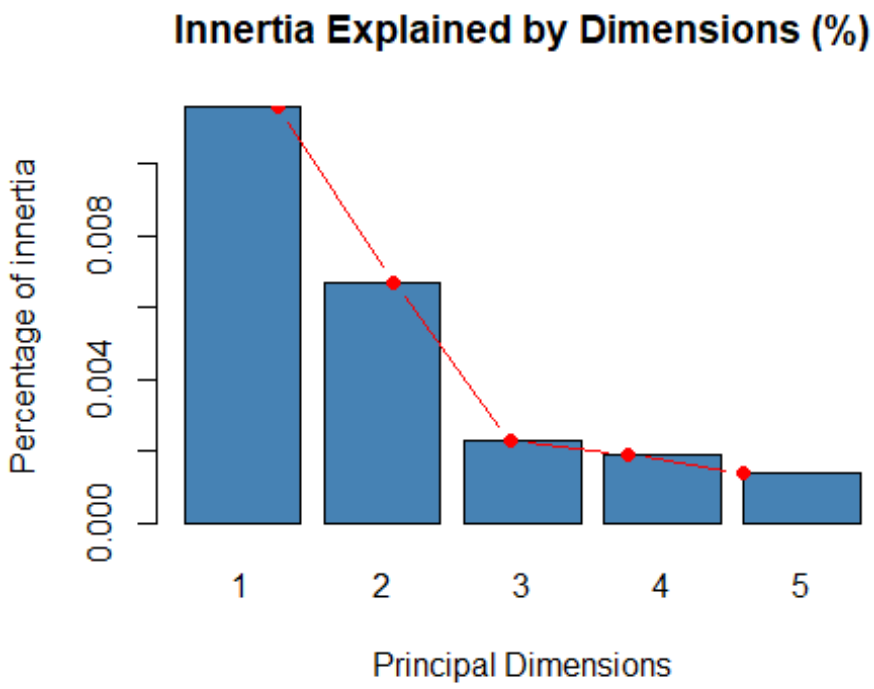
```
##        eigenvalue percentage of variance cumulative percentage of
variance
## dim 1 0.011576135              48.617389
48.61739
## dim 2 0.006697021              28.126114
76.74350
## dim 3 0.002266893               9.520486
86.26399
## dim 4 0.001898805               7.974590
94.23858
## dim 5 0.001371834               5.761422
100.00000
```

To determine the number of eigenvalues different from 0, you can simply count the number of eigenvalues greater than 0 in the output you provided. In this case, there are 5 eigenvalues provided, and all of them are greater than 0.

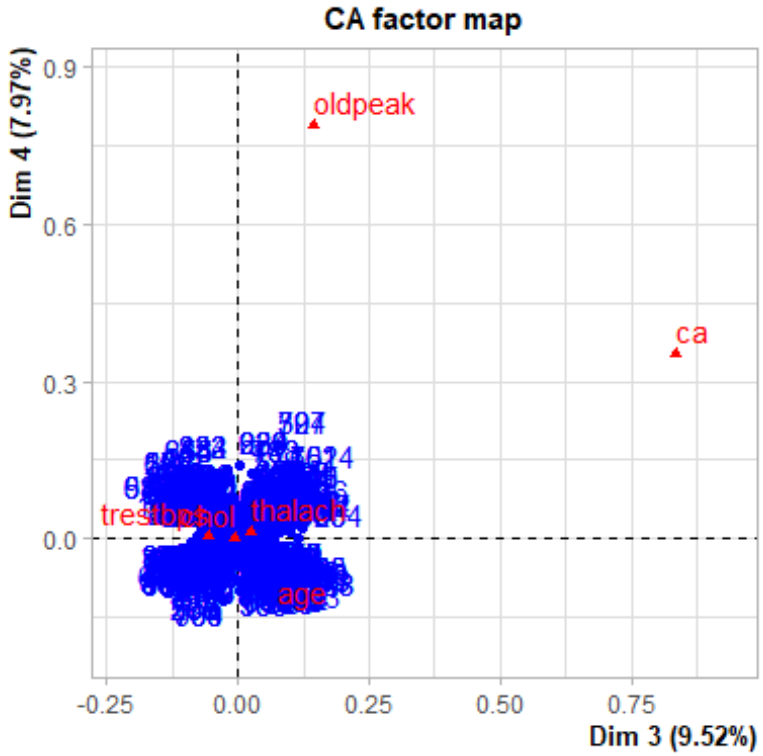
This is indeed related to the number of dimensions or factors retained in the correspondence analysis. Each non-zero eigenvalue corresponds to a dimension that explains a certain amount of variance in the data.

In a contingency table used for correspondence analysis, the number of non-zero eigenvalues typically corresponds to the smaller of the number of rows minus one and the number of columns minus one. This is because the maximum number of dimensions that can be extracted is limited by the minimum of the number of rows and the number of columns in the contingency table

```
eig.val <- res.ca$eig
barplot(eig.val[, 1],
        names.arg = 1:nrow(eig.val),
        main = "Innertia Explained by Dimensions (%)",
        xlab = "Principal Dimensions",
        ylab = "Percentage of innertia",
        col ="steelblue")
# Add connected line segments to the plot
lines(x = 1:nrow(eig.val), eig.val[, 1],
      type = "b", pch = 19, col = "red")
```




```
plot(res.ca, axes = 3:4)
```




```
# row margins
head(res.ca$call$marge.row, 20)
```

```
##  [1] 0.0009372506 0.0009273760 0.0008646137 0.0009606818 0.0010123980
##  [6] 0.0008853671 0.0010667920 0.0010892191 0.0009369158 0.0009760795
## [11] 0.0007675413 0.0010962484 0.0009283802 0.0010346577 0.0009054510
## [16] 0.0009283802 0.0010770013 0.0009308907 0.0009658702 0.0009606818

head(res.ca$call$marge.row*sum(dfc), 50)

##  [1] 560.0 554.1 516.6 574.0 604.9 529.0 637.4 650.8 559.8 583.2 458.6
655.0
## [13] 554.7 618.2 541.0 554.7 643.5 556.2 577.1 574.0 543.0 540.3 508.0
622.0
## [25] 544.0 660.0 586.4 667.0 585.6 682.4 479.8 577.1 435.2 614.9 538.0
577.6
## [37] 516.4 650.0 561.2 565.4 736.8 579.4 690.0 559.8 696.2 586.0 511.0
640.0
## [49] 643.0 570.4

# column margins
res.ca$call$marge.col

##         age    trestbps        chol     thalach     oldpeak
ca
## 0.093381957 0.225780315 0.422013807 0.255805807 0.001838183
0.001179932

res.ca$call$marge.col[rev(order(res.ca$call$marge.col))]*sum(dfc)

##     chol  thalach trestbps      age  oldpeak       ca
## 252150.0 152842.0 134902.0  55795.0   1098.3    705.0
```

Multiple Component Aanalysis (MCA)

Multiple Correspondence Analysis (MCA) it's an extension of Correspondence Analysis (CA), which can handle more than two categorical variables simultaneously. It helps to visualize patterns and associations between categories in a dataset by representing them in a lower-dimensional space.

We load the necessary libraries:

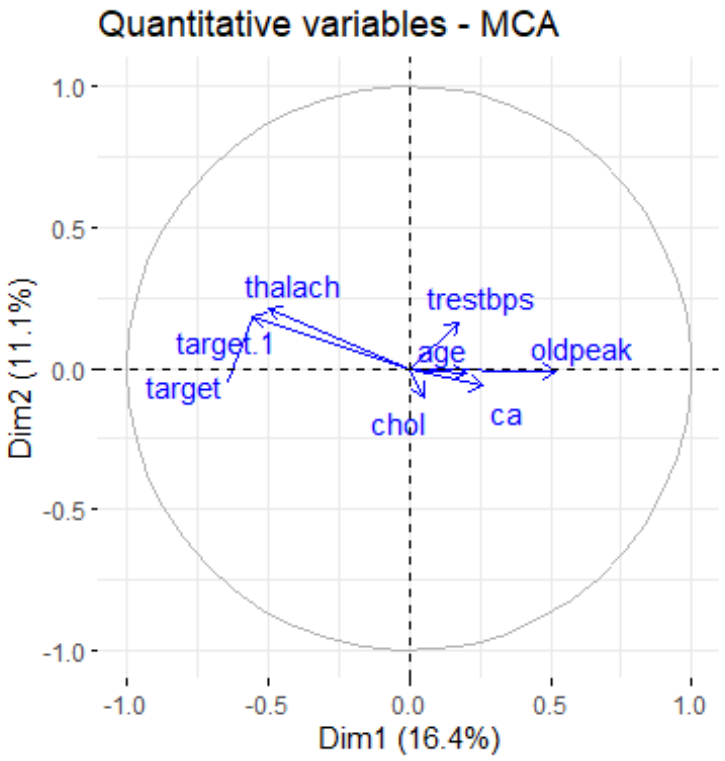
```
library(FactoMineR)
library(factoextra)
```

First, we compute MCA for our dataset, utilizing numerical variables as supplementary variables.

```
res.mca <- MCA(df[,c(vars_res, vars_dis, vars_con)], quanti.sup =
c(1,9:15), graph=FALSE)
```

When then look at the graph of the supplementary quantitative variables:

```
fviz_mca_var(res.mca, choice="quanti.sup", repel=TRUE)
```



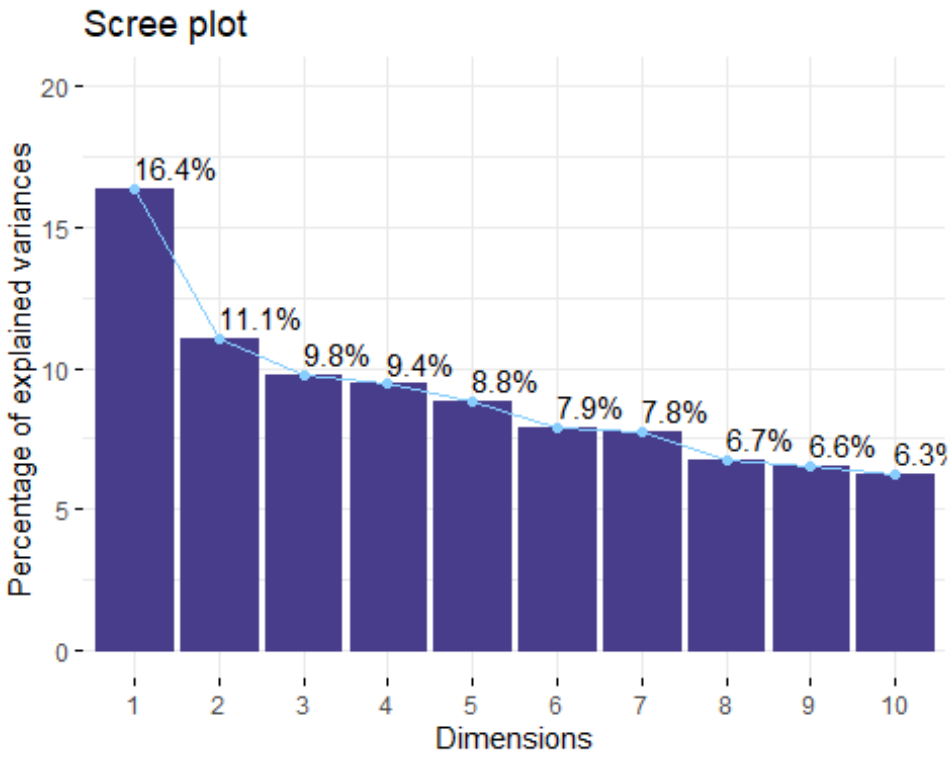

```
res.mca$quanti.sup

## $coord
##                Dim 1       Dim 2       Dim 3        Dim 4        Dim 5
## target   -0.55363343  0.18369741  0.14927747  0.037896342 -0.037690931
## age       0.20075549 -0.02051703  0.11593798 -0.061353432 -0.161468046
## trestbps  0.17355624  0.16358586  0.08935669 -0.009073522 -0.144135088
## chol      0.05118968 -0.10202456  0.10894730 -0.006131164 -0.110569593
## thalach  -0.49651950  0.21099025 -0.07883537  0.034648545 -0.005259775
## oldpeak   0.52044694 -0.01042545  0.09959675  0.128037999 -0.117473854
## ca        0.26080727 -0.06407901 -0.01933083 -0.058421180 -0.039286543
## target.1 -0.55363343  0.18369741  0.14927747  0.037896342 -0.037690931
```

We observe that thalach and target exhibit stronger correlations with dimension 2, whereas the remaining quantitative variables show higher correlations with dimension 1. This trend becomes clearer when considering specific values. Notably, trestbps displays a similar level of correlation across both dimensions. Additionally, age appears to be more closely associated with the third dimension, while oldpeak relates more strongly to the fourth. This observation might explain why we require four dimensions to capture a greater proportion of explained variances, as observed in the PCA.

Eigenvalues and dominant axes analysis

We want to know how many axes we need to consider for the next Hierarchical Classification stage. We can do this according to the generalized Kaiser theorem: all those dimensions such that their eigenvalue is greater than the mean.

```
mean(res.mca$eig[,1])

## [1] 0.1428571

head(get_eigenvalue(res.mca), 10)

##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1   0.2805925        16.367898                    16.36790
## Dim.2   0.1897640        11.069567                    27.43747
## Dim.3   0.1677624         9.786141                    37.22361
## Dim.4   0.1617123         9.433216                    46.65682
## Dim.5   0.1517120         8.849864                    55.50669
## Dim.6   0.1356620         7.913616                    63.42030
## Dim.7   0.1328645         7.750430                    71.17073
## Dim.8   0.1153906         6.731117                    77.90185
## Dim.9   0.1124421         6.559122                    84.46097
## Dim.10  0.1071469         6.250237                    90.71121

fviz_screeplot(res.mca, addlabels=TRUE, ylim=c(0,20),
barfill='darkslateblue',barcolor='darkslateblue', linecolor="skyblue1")
```

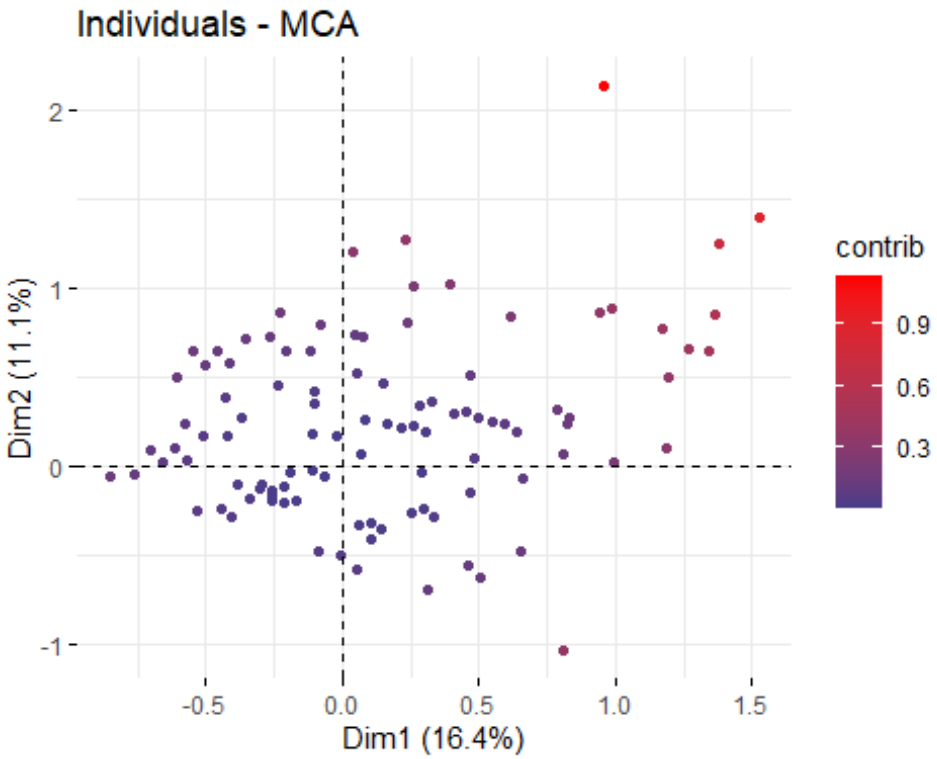


As we can see, the eigenvalue mean is 0.1428571. Therefore, we will take up to dimension 5, which represents the 55.51% of the sample.

Individuals point of view

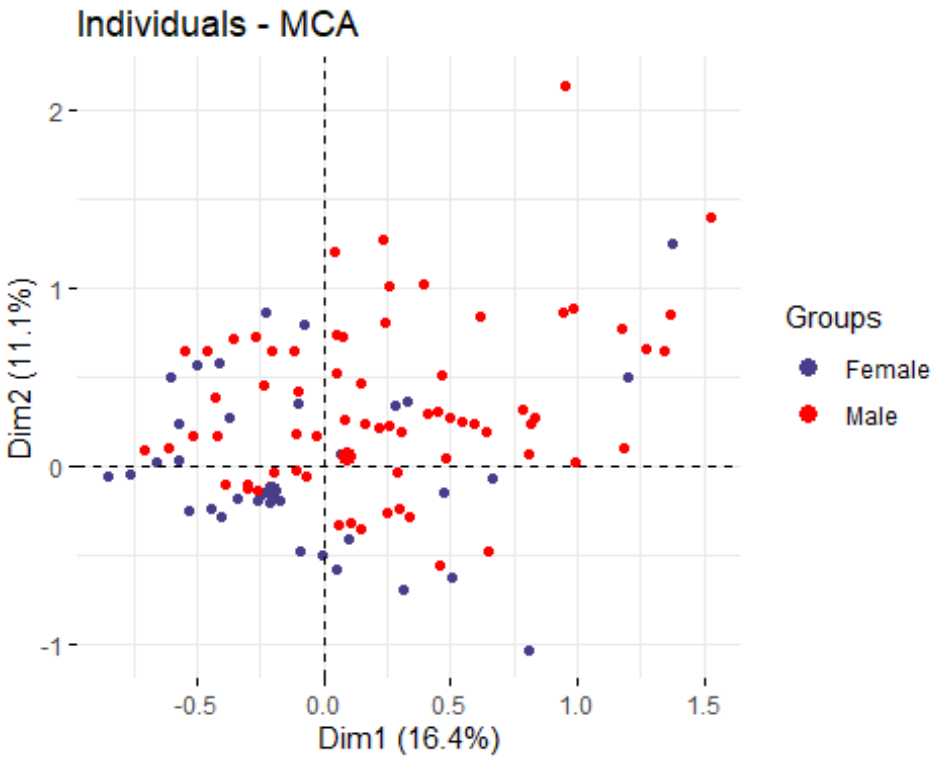
We now want to see if there are individuals "too contributive".

```
fviz_mca_ind(res.mca, geom=c('point'),col.ind="contrib", gradient.cols =
c("darkslateblue", "red"))
```



We observe that only three individuals appear to make significant contributions, with a small number showing lesser levels of contribution. However, the vast majority contribute minimally.

```
fviz_mca_ind(res.mca, label="none", habillage= df$sex,
palette=c("darkslateblue", "red"))
```

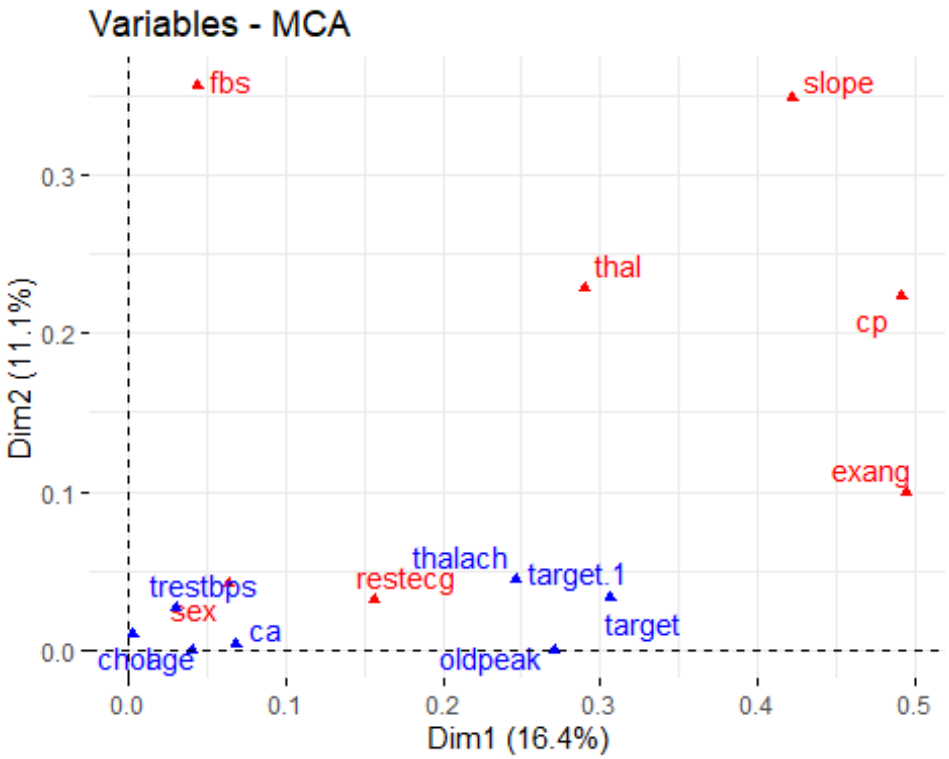


We see that there are quite a lot more males than females. With females more representend in the second dimension and males in the first.

Interpreting map of categories

In the subsequent plot, it's evident that quantitative variables display a predominant correlation along the first dimension, characterized by low values. Similarly, the qualitative variables sex and restecg demonstrate comparable correlations along this dimension. Conversely, the remaining qualitative variables showcase stronger correlations with either dimension 1 or 2. Notably, variables thal and slope exhibit similar correlations across both dimensions, whereas variables fbs, cp, and exang show stronger correlations with one dimension over the other—specifically, the second dimension for fbs and the first dimension for the latter two.

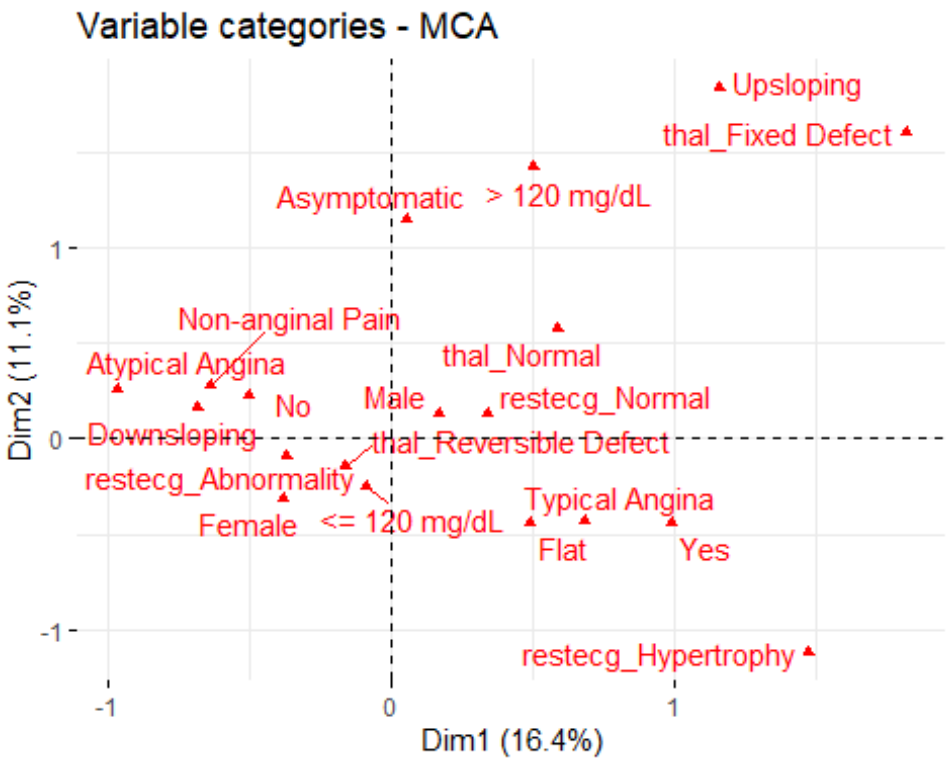
```
fviz_mca_var(res.mca, choice="mca.cor", repel=TRUE)
```



Now, we delve into the specific values of the variables. On the left side, we encounter values associated with a lower percentage of heart attacks, predominantly featuring female individuals. Conversely, in the upper right corner, we find values associated with a higher percentage of heart attacks.

As evident, the values corresponding to a moderate likelihood of experiencing a heart attack are centrally located in the gravity center, predominantly featuring male individuals.

```
fviz_mca_var(res.mca, repel=TRUE)
```



Interpreting the axes association to factor map

Variables point of view coordinates

First, we examine the contribution of each category value to each dimension, noting that higher values correspond to greater contributions. So for example, the value female for sex category has a high contribution to dimension 3, while male has a small contributioon to dimensions one and two.

```
res.mca$var$coord
```

```
##                              Dim 1       Dim 2       Dim 3        Dim
4
## Female                 -0.38190986 -0.30941315  0.98892143
0.041935330
## Male                    0.16711904  0.13539538 -0.43273981 -
0.018350383
## Asymptomatic            0.05497625  1.15155090 -0.99771217 -
0.149125221
## Atypical Angina        -0.96819838  0.26163919 -0.49829226
0.908634352
## Non-anginal Pain       -0.63963229  0.28529352  0.70316839 -
0.556813894
## Typical Angina          0.68231696 -0.42934914 -0.07980117
0.035967508
## <= 120 mg/dL           -0.08754261 -0.24982421 -0.07451881
0.163533899
```

```
## > 120 mg/dL              0.49893568  1.42383469  0.42470851 -
0.932036340
## restecg_Abnormality    -0.37104121 -0.09493797 -0.04363904
0.097299510
## restecg_Hypertrophy     1.47135299 -1.12128872  4.78221646
3.661127648
## restecg_Normal          0.33857917  0.13183604 -0.09928857 -
0.210928698
## No                     -0.50134209  0.22469405  0.07418328
0.059225308
## Yes                     0.98815253 -0.44287523 -0.14621632 -
0.116733940
## Downsloping            -0.68374036  0.16262550 -0.17228639
0.006775596
## Flat                    0.48755087 -0.44108763  0.06831892 -
0.229701012
## Upsloping               1.15776637  1.84233617  0.64692703
1.453218017
## thal_Fixed Defect       1.81598816  1.60844065  0.28536002
0.768147244
## thal_Normal             0.58458310  0.58281166  2.80496557 -
6.649019475
## thal_Reversible Defect -0.16346486 -0.14526903 -0.04597680 -
0.017569378
##                              Dim 5
## Female                 -0.28152335
## Male                    0.12319114
## Asymptomatic           -2.25878862
## Atypical Angina         0.61135890
## Non-anginal Pain       -0.11774620
## Typical Angina          0.21181028
## <= 120 mg/dL           -0.04793081
## > 120 mg/dL             0.27317430
## restecg_Abnormality     0.47964178
## restecg_Hypertrophy    -0.83427837
## restecg_Normal         -0.46990354
## No                     -0.16608601
## Yes                     0.32735793
## Downsloping             0.28426762
## Flat                   -0.33393454
## Upsloping               0.37344503
## thal_Fixed Defect       0.47676736
## thal_Normal             4.44401952
## thal_Reversible Defect -0.07500327
```

Quality of representation

Now, we assess the quality of representation for each category value by observing the graphical depiction of their representation levels. Notably, the category value "Typical Angina" emerges as the most well-represented in the first two dimensions, while in

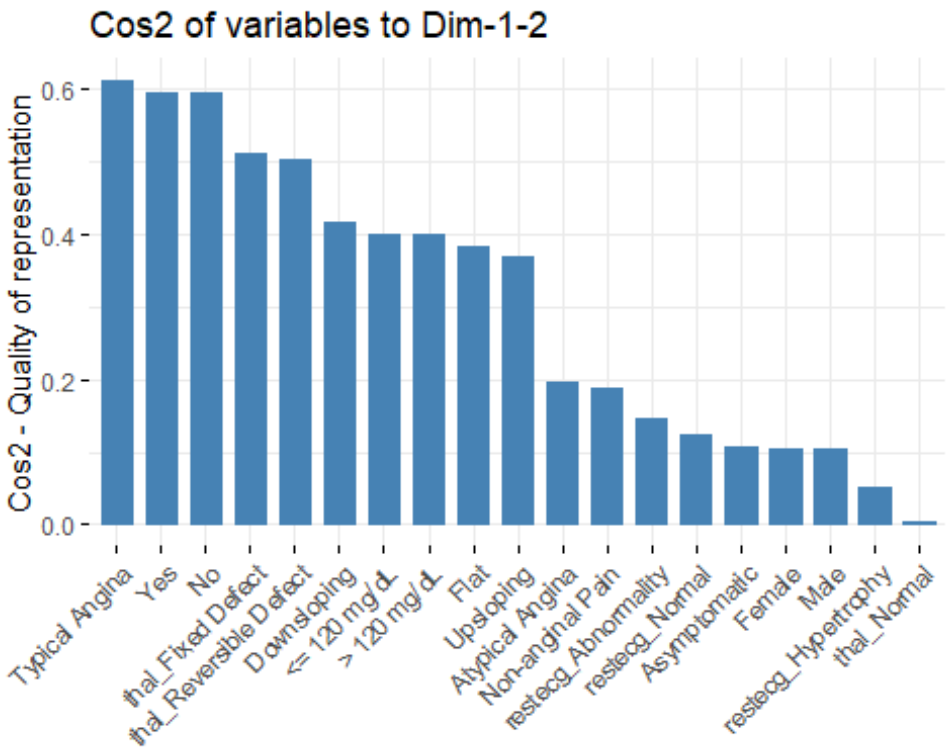
the third and fourth dimensions, “female” ranks as the second best representation. This aligns with the numerical values obtained earlier.

```
res.mca$var$cos2
```

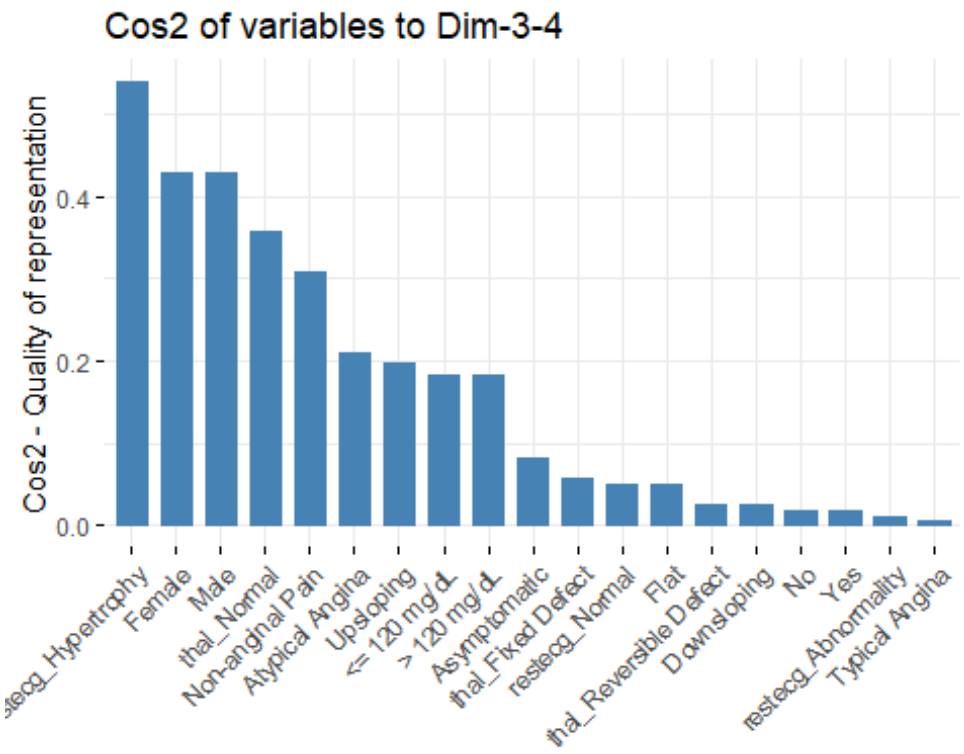
```
##                              Dim 1       Dim 2       Dim 3        Dim
4
## Female                 0.0638244084 0.041893111 0.427945670 7.695293e-
04
## Male                   0.0638244084 0.041893111 0.427945670 7.695293e-
04
## Asymptomatic           0.0002454893 0.107708174 0.080852403 1.806278e-
03
## Atypical Angina        0.1824558910 0.013324005 0.048327849 1.606969e-
01
## Non-anginal Pain       0.1568053543 0.031194925 0.189504190 1.188284e-
01
## Typical Angina         0.4382226265 0.173517654 0.005994334 1.217708e-
03
## <= 120 mg/dL           0.0436781342 0.355708370 0.031648773 1.524195e-
01
## > 120 mg/dL            0.0436781342 0.355708370 0.031648773 1.524195e-
01
## restecg_Abnormality    0.1379404687 0.009030822 0.001908086 9.485685e-
03
## restecg_Hypertrophy    0.0321516774 0.018672600 0.339647440 1.990672e-
01
## restecg_Normal         0.1079053376 0.016360280 0.009279423 4.187876e-
02
## No                     0.4954024523 0.099511430 0.010846807 6.913604e-
03
## Yes                    0.4954024523 0.099511430 0.010846807 6.913604e-
03
## Downsloping            0.3943487654 0.022308756 0.025038022 3.872514e-
05
## Flat                   0.2110022452 0.172701839 0.004143136 4.683527e-
02
## Upsloping              0.1043021021 0.264112502 0.032565804 1.643284e-
01
## thal_Fixed Defect      0.2867663471 0.224963593 0.007080899 5.130871e-
02
## thal_Normal            0.0023498643 0.002335644 0.054101005 3.039943e-
01
## thal_Reversible Defect 0.2810183411 0.221938135 0.022231217 3.246374e-
03
##                              Dim 5
## Female                 0.034681184
## Male                   0.034681184
## Asymptomatic           0.414413190
## Atypical Angina        0.072748101
```

```
## Non-anginal Pain       0.005313663
## Typical Angina         0.042229558
## <= 120 mg/dL           0.013093466
## > 120 mg/dL            0.013093466
## restecg_Abnormality    0.230505564
## restecg_Hypertrophy    0.010336937
## restecg_Normal         0.207845148
## No                     0.054369573
## Yes                    0.054369573
## Downsloping            0.068163651
## Flat                   0.098985113
## Upsloping              0.010851870
## thal_Fixed Defect      0.019765836
## thal_Normal            0.135800753
## thal_Reversible Defect 0.059162461
```

```
fviz_cos2(res.mca, choice = "var", axes = 1:2)
```




```
fviz_cos2(res.mca, choice = "var", axes = 3:4)
```



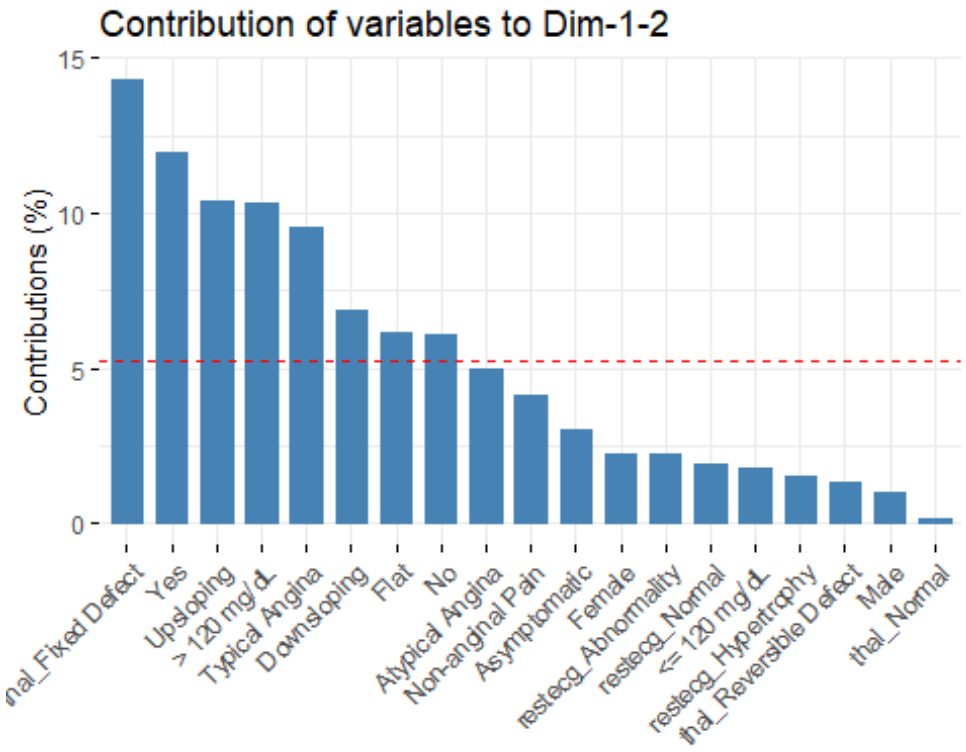
Contribution of the variables

Here, we observe the graphical representation of contributions for the variable categories. Once more, higher values indicate greater contributions. As expected, the variables contributing the most align with those that were best represented earlier.

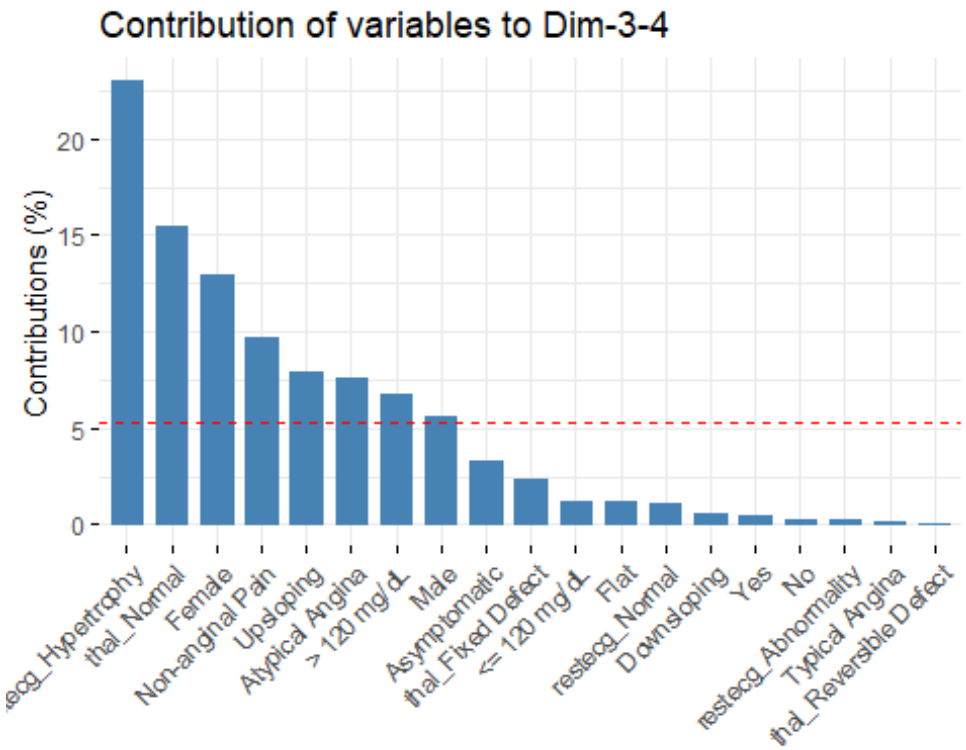
```
res.pca$var$contrib
```

```
##              Dim.1     Dim.2     Dim.3     Dim.4       Dim.5
## age      26.573814  1.880155  1.063934 30.934313  0.01433326
## trestbps  8.257378 26.078341 54.413145  1.069699  0.15174479
## chol      5.282763 43.197025 28.711497 11.310311  9.85587951
## thalach  20.120457 20.193673  0.860683  6.209484 26.07973258
## oldpeak  18.588577  7.537726  8.326780 49.392631  2.27566859
## ca       21.177011  1.113080  6.623961  1.083562 61.62264127
```

```
fviz_contrib(res.mca, choice = "var", axes = 1:2)
```




```r
fviz_contrib(res.mca, choice = "var", axes = 3:4)
```



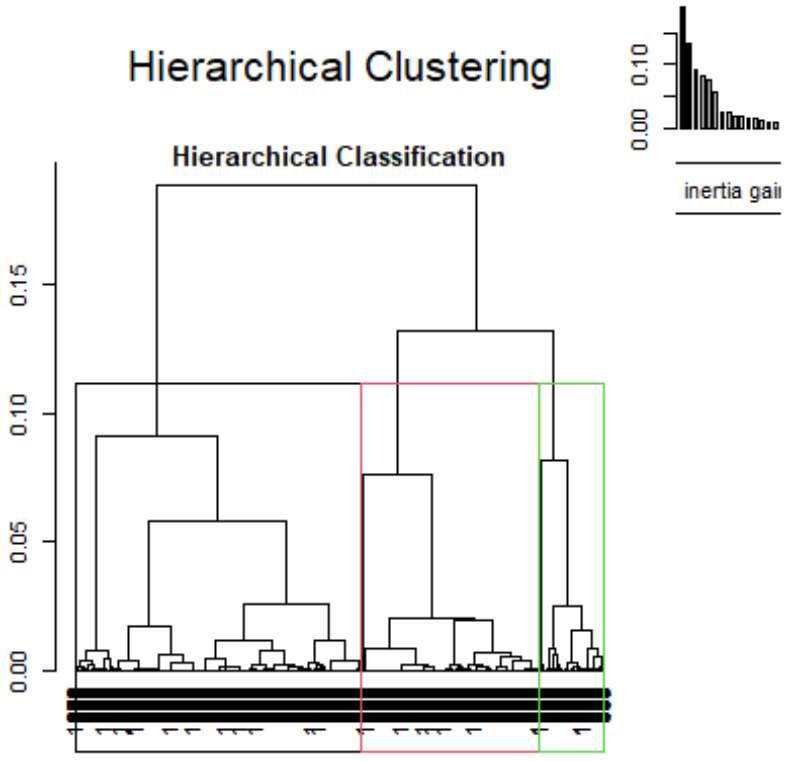
MCA with suplementary variables

We now perform the MCA again, but putting as suplemenatry the individuals found as multivariate outliers.

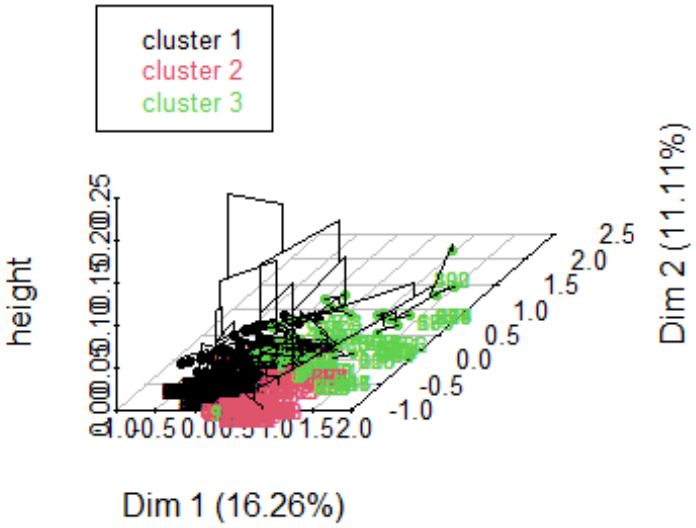
```
res.mca <- MCA(df[,c(vars_res, vars_dis, vars_con)], quanti.sup =
c(1,9:15), ind.sup = ll, graph = FALSE)
```

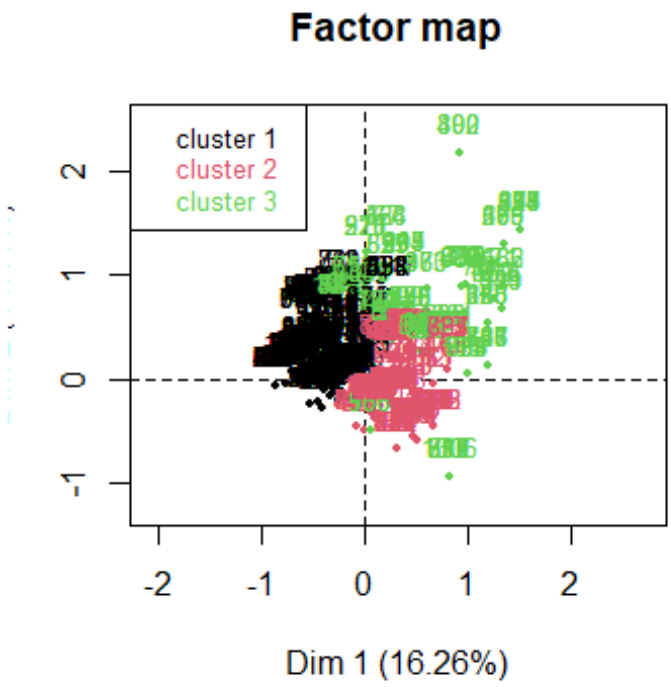
Hierarchical Clustering (from MCA)

```
res.hcpc.mca <- HCPC(res.mca, nb.clust=-1)
```



Hierarchical clustering on the factor map





The hierarchical clustering using the mca results ends in three different clusters, one less than using pca results.

Test of the association of categorical variables

```
res.hcpc.mca$desc.var$test.chi2
```

```
##                p.value df
## slope   2.088192e-168  4
## thal    4.783440e-117  4
## exang   2.876502e-103  2
## cp       5.085997e-98  6
## restecg  1.896935e-24  4
## sex      1.323096e-05  2
## fbs      3.237123e-05  2
```

Variables like “slope,” “thal,” and “exang” show extremely low p-values, suggesting strong relationships. However, “sex” and “fbs” also exhibit significance, with slightly higher p-values.

```
res.hcpc.mca$desc.var$category[1]
```

```
## $`1`
##                                Cla/Mod   Mod/Cla    Global
p.value
## exang=No                     76.331361 96.089385 66.666667 3.043409e-
111
## slope=Downsloping            84.881210 73.184358 45.660750  2.284374e-
```

```
83
## cp=Non-anginal Pain         83.802817 44.320298 28.007890  4.991151e-
37
## cp=Atypical Angina          90.909091 27.932961 16.272189  4.295908e-
30
## thal=thal_Reversible Defect 57.481163 99.441341 91.617357  4.913993e-
25
## restecg=restecg_Abnormality 60.433071 57.169460 50.098619  1.776532e-
06
## sex=Female                  64.102564 37.243948 30.769231  1.987260e-
06
## cp=Asymptomatic             70.666667  9.869646  7.396450  1.291218e-
03
## restecg=restecg_Normal      46.843177 42.830540 48.422091  1.590317e-
04
## restecg=restecg_Hypertrophy  0.000000  0.000000  1.479290  1.085952e-
05
## sex=Male                    48.005698 62.756052 69.230769  1.987260e-
06
## slope=Upsloping              0.000000  0.000000  7.001972  2.837654e-
25
## thal=thal_Fixed Defect       0.000000  0.000000  7.692308  7.553867e-
28
## slope=Flat                  30.000000 26.815642 47.337278  4.621146e-
45
## cp=Typical Angina           19.591837 17.877095 48.323471 1.672384e-
101
## exang=Yes                    6.213018  3.910615 33.333333 3.043409e-
111
##                                 v.test
## exang=No                     22.411453
## slope=Downsloping            19.344128
## cp=Non-anginal Pain          12.713257
## cp=Atypical Angina           11.397643
## thal=thal_Reversible Defect  10.334611
## restecg=restecg_Abnormality   4.777313
## sex=Female                    4.754716
## cp=Asymptomatic               3.217925
## restecg=restecg_Normal       -3.776525
## restecg=restecg_Hypertrophy  -4.399311
## sex=Male                     -4.754716
## slope=Upsloping             -10.387129
## thal=thal_Fixed Defect      -10.938379
## slope=Flat                  -14.086144
## cp=Typical Angina           -21.389529
## exang=Yes                   -22.411453
```

This first cluster has a strong influence with variables like exang=No, slope=Downsloping or ca=Non anginal pain, some less characteristic variables could

be slope=Flat, cp=Typical Angina and exang=Yes, suggesting individuals with this characteristics are not typical of this cluster.

```
res.hcpc.mca$desc.var$category[2]
```

```
## $`2`
##                                Cla/Mod   Mod/Cla    Global      p.value
## cp=Typical Angina           63.469388 91.202346 48.323471 2.859491e-93
## exang=Yes                   74.260355 73.607038 33.333333 4.622191e-84
## slope=Flat                  57.916667 81.524927 47.337278 6.054078e-57
## thal=thal_Reversible Defect 36.275565 98.826979 91.617357 2.832638e-11
## sex=Male                    37.179487 76.539589 69.230769 2.887510e-04
## restecg=restecg_Normal      38.289206 55.131965 48.422091 2.383105e-03
## restecg=restecg_Abnormality 30.118110 44.868035 50.098619 1.792091e-02
## restecg=restecg_Hypertrophy  0.000000  0.000000  1.479290 2.025789e-03
## sex=Female                  25.641026 23.460411 30.769231 2.887510e-04
## cp=Asymptomatic             13.333333  2.932551  7.396450 4.215514e-05
## slope=Upsloping              0.000000  0.000000  7.001972 6.131022e-14
## thal=thal_Fixed Defect       0.000000  0.000000  7.692308 2.615867e-15
## cp=Atypical Angina           0.000000  0.000000 16.272189 1.652019e-33
## cp=Non-anginal Pain          7.042254  5.865103 28.007890 5.745548e-34
## slope=Downsloping           13.606911 18.475073 45.660750 4.114031e-37
## exang=No                    13.313609 26.392962 66.666667 4.622191e-84
##                                v.test
## cp=Typical Angina            20.486184
## exang=Yes                    19.426334
## slope=Flat                   15.902856
## thal=thal_Reversible Defect   6.655019
## sex=Male                      3.625188
## restecg=restecg_Normal        3.037802
## restecg=restecg_Abnormality  -2.367248
## restecg=restecg_Hypertrophy  -3.086425
## sex=Female                   -3.625188
## cp=Asymptomatic              -4.095339
## slope=Upsloping              -7.505253
## thal=thal_Fixed Defect       -7.907986
## cp=Atypical Angina          -12.063217
## cp=Non-anginal Pain         -12.149872
## slope=Downsloping           -12.728357
## exang=No                    -19.426334
```

Contrary to the first cluster, indivuduals with slope=Flat, cp=Typical Angina and exang=Yes are common in this second cluster, other non typical characteristics of this group could be cp=Non-anginal Pain, slope=Downsloping and exang=No.

```
res.hcpc.mca$desc.var$category[3]
```

```
## $`3`
##                                  Cla/Mod   Mod/Cla    Global
p.value
## thal=thal_Fixed Defect      100.000000 57.352941  7.692308 1.100260e-
```

```
79
## slope=Upsloping             100.000000 52.205882  7.001972 2.033410e-
71
## restecg=restecg_Hypertrophy 100.000000 11.029412  1.479290 4.071985e-
14
## fbs=> 120 mg/dL              25.000000 27.941176 14.990138 2.530947e-
05
## exang=Yes                    19.526627 48.529412 33.333333 8.286107e-
05
## cp=Typical Angina            16.938776 61.029412 48.323471 1.474465e-
03
## sex=Male                     14.814815 76.470588 69.230769 4.691337e-
02
## sex=Female                   10.256410 23.529412 30.769231 4.691337e-
02
## cp=Non-anginal Pain           9.154930 19.117647 28.007890 1.119552e-
02
## restecg=restecg_Abnormality   9.448819 35.294118 50.098619 2.018391e-
04
## exang=No                     10.355030 51.470588 66.666667 8.286107e-
05
## fbs=<= 120 mg/dL             11.368910 72.058824 85.009862 2.530947e-
05
## slope=Downsloping             1.511879  5.147059 45.660750 4.692625e-
29
## thal=thal_Reversible Defect   6.243272 42.647059 91.617357 3.506349e-
70
##                                 v.test
## thal=thal_Fixed Defect       18.901896
## slope=Upsloping              17.869629
## restecg=restecg_Hypertrophy   7.558674
## fbs=> 120 mg/dL               4.212022
## exang=Yes                     3.935971
## cp=Typical Angina             3.179663
## sex=Male                      1.987081
## sex=Female                   -1.987081
## cp=Non-anginal Pain          -2.536536
## restecg=restecg_Abnormality  -3.716704
## exang=No                     -3.935971
## fbs=<= 120 mg/dL             -4.212022
## slope=Downsloping           -11.187542
## thal=thal_Reversible Defect -17.710070
```

Finally the third cluster is composed with individuals with thal=Fixed Defect, slope=Upsloping and restecg=Hypertrophy. On the other hand, fbs =< 120 mg/dL, slopw=Downsloping and thal=Reversible Defect are some of the less common characteristics in individuals in this third cluster.

```
res.hcpc.mca$desc.var$quanti.var
```

```
##                Eta2      P-value
## oldpeak  0.25258535 1.217284e-64
## thalach  0.23997059 5.751309e-61
## target   0.21809791 9.739262e-55
## target.1 0.21809791 9.739262e-55
## ca       0.06050462 1.987218e-14
## age      0.04979871 6.106504e-12
## trestbps 0.03738909 4.308935e-09
## chol     0.02037771 3.021001e-05
```

Variables like “oldpeak,” “thalach,” and “ca” exhibit notably high Eta2 values and extremely low p-values, indicating substantial contributions and strong associations. On the other hand, “age,” “trestbps,” and “chol” also show statistical significance but with relatively lower Eta2 values, suggesting comparatively lesser impact on the dataset’s variability.

```
res.hcpc.mca$desc.var$quanti[1]
```

```
## $`1`
##              v.test Mean in category Overall mean sd in category
Overall sd
## thalach   15.567596      159.6424581  149.1015779     17.9693175
22.8658841
## target.1  14.574760        0.6609870    0.5376923      0.2495858
0.2856772
## target    14.574760        0.6609870    0.5376923      0.2495858
0.2856772
## chol      -2.253698      242.9888268  246.4319527     51.8568064
51.5928799
## trestbps  -3.520578      129.7914339  131.6242604     16.1411742
17.5808598
## age       -7.072221       52.5847300   54.4802761      9.2037150
9.0513060
## ca        -7.240683        0.4878957    0.6893491      0.8121188
0.9395669
## oldpeak  -15.300455        0.5351955    1.0630178      0.7602372
1.1649733
##                p.value
## thalach  1.208649e-54
## target.1 4.065488e-48
## target   4.065488e-48
## chol     2.421515e-02
## trestbps 4.306081e-04
## age      1.524739e-12
## ca       4.464295e-13
## oldpeak  7.592347e-53
```

We can see individuals with a high maximum heart rate (thalach) in this cluster, with a little below the average cholesterol and trestbps, and with a low number of major vessels detected by fluoroscopy (ca) and a low ST depression (oldpeak).

```
res.hcpc.mca$desc.var$quanti[2]
```

```
## $`2`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak    8.498107        1.5000000    1.0630178       1.094802
1.1649733
## ca         7.419934        0.9970674    0.6893491       1.023187
0.9395669
## age        5.771066       56.7859238   54.4802761       8.430390
9.0513060
## chol       4.287813      256.1964809  246.4319527      52.402174
51.5928799
## thalach  -12.321489      136.6656891  149.1015779      21.380500
22.8658841
## target.1 -12.919705        0.3747801    0.5376923       0.249966
0.2856772
## target   -12.919705        0.3747801    0.5376923       0.249966
0.2856772
##                p.value
## oldpeak  1.927077e-17
## ca       1.171787e-13
## age      7.877145e-09
## chol     1.804406e-05
## thalach  6.940049e-35
## target.1 3.484747e-38
## target   3.484747e-38
```

In this second cluster we can observe a high influence in individuals with a high ST depression, a high numer of vessels in the fluoroscopy test and generally above the average in terms of age. These individuals have also a high cholesterol comparing it to the mean.

```
res.hcpc.mca$desc.var$quanti[3]
```

```
## $`3`
##              v.test Mean in category Overall mean sd in category
Overall sd
## oldpeak  10.628375        2.0514706    1.0630178      1.5085046
1.1649733
## trestbps  6.056804      140.1250000  131.6242604     21.1254786
17.5808598
## age       2.357599       56.1838235   54.4802761      8.2303261
9.0513060
## chol     -2.643496      235.5441176  246.4319527     44.0292128
51.5928799
## target.1 -3.435682        0.4593382    0.5376923      0.2704495
0.2856772
## target   -3.435682        0.4593382    0.5376923      0.2704495
0.2856772
## thalach  -5.719152      138.6617647  149.1015779     23.2740137
```

```
22.8658841
##               p.value
## oldpeak  2.199099e-26
## trestbps 1.388526e-09
## age      1.839356e-02
## chol     8.205470e-03
## target.1 5.910641e-04
## target   5.910641e-04
## thalach  1.070569e-08
```

Finally in this last group we can find extremely high ST depression and a high blood pressure. The typical individuals of this clusters have an above the mean age, and they tend to have a low max heart rate.

Parangons

```
res.hcpc.mca$desc.ind$para

## Cluster: 1
##        76        82       104       139       150
## 0.2486547 0.2486547 0.2486547 0.2486547 0.2486547
## ------------------------------------------------------------
## Cluster: 2
##        93       112       161       163       181
## 0.2933109 0.2933109 0.2933109 0.2933109 0.2933109
## ------------------------------------------------------------
## Cluster: 3
##       178       389       757       906       595
## 0.4762593 0.4762593 0.4762593 0.4762593 0.5237542
```

This are the Parangons for cluster 1, 2 and 3.

Class-specific variables

```
res.hcpc.mca$desc.ind$dist

## Cluster: 1
##      320      330      360       51      764
## 3.277887 3.277887 3.277887 1.715935 1.715935
## ------------------------------------------------------------
## Cluster: 2
##      687      735      894       15      182
## 3.735682 3.735682 3.735682 3.735682 1.465070
## ------------------------------------------------------------
## Cluster: 3
##        7      151      662     1014      100
## 3.321113 3.321113 3.321113 3.321113 2.709853
```

Class-specific individuals for clusters 1, 2 and 3.

Comparison of clusters obtained after K-Means (based on PCA) and/or Hierarchical Clustering (based on PCA) focusing on the target.

thalach 15.567596 159.6424581 149.1015779 17.9693175 22.8658841 target.1 14.574760 0.6609870 0.5376923 0.2495858 0.2856772 target 14.574760 0.6609870 0.5376923 0.2495858 0.2856772 chol -2.253698 242.9888268 246.4319527 51.8568064 51.5928799 …

In the first cluster of hierarchical clustering using mca we can find a high influence of individuals with a high probability of getting a heart attack, the cluster mean is significantly higher than the global mean. We can see is also related to individuals with a high maximum heart rate (thalach), with a little below the average cholesterol and trestbps, and with a low number of major vessels detected by fluoroscopy (ca) and a low ST depression (oldpeak).

thalach 17.165341 165.7713499 149.1141463 14.0882717 22.9944987 target.1 12.802854 0.6907438 0.5363902 0.2309836 0.2856828 target 12.802854 0.6907438 0.5363902 0.2309836 0.2856828 trestbps -9.149772 124.8512397 131.6117073 12.1240111 17.5081712 chol -9.628293 225.0468320 246.0000000 36.7901578 51.5673370 oldpeak -12.064233 0.4735537 1.0715122 0.7936492 1.1744799 ca -13.476351 0.1542700 0.6878049 0.4307751 0.9381339 age -21.838374 46.0771350 54.4341463 6.5706251 9.0678636

A similar conclusion can be done by looking at the first cluster of the herarchical clustering using PCA, seeing the same individuals with a high heart rate, low cholesterol, young…

thalach 10.046405 159.9523810 149.1141463 14.4981976 22.9944987 target 7.550224 0.6375873 0.5363902 0.2623760 0.2856828 chol 3.636029 254.7968254 246.0000000 43.6305193 51.5673370 trestbps 2.602203 133.7492063 131.6117073 11.5521162 17.5081712 age -2.229345 53.4857143 54.4341463 6.2168960 9.0678636 ca -7.620883 0.3523810 0.6878049 0.6470586 0.9381339 oldpeak -10.487070 0.4936508 1.0715122 0.6486795 1.1744799

Finally all said before resonates with the first cluster of kmeans, all variables have a similar means and go on the same line.