Entrega2

PCA, CA and Clustering

Pol Renau Miguel Angel Merino

November 24, 2019

Contents

1 Carregar les dades

Carreguem les dades a analitzar, que ja han sigut processades a la Entrega 1 per a poder fer ara un anàlisi consistent. Separem també les variables continues de les discretes per facilitar-ne l'ús posteriorment.

```
load("mostra2.RData")

df$f.hpw<- factor(df$f.hpw)
df$f.educationNum <- factor(df$f.educationNum)
```

```
vars_con<-names(df)[c(3,5,11:13,24)];vars_con
```

```
## [1] "fnlwgt"        "education.num" "capital.gain"  "capital.loss"
## [5] "hr.per.week"   "i.rank"
```

```
vars_dis<-names(df)[c(7,9,10,15:23)];vars_dis
```

```
##  [1] "occupation"     "race"           "sex"            "y.bin"
##  [5] "f.type"         "f.marital"      "f.education"    "f.continent"
##  [9] "f.benefici"     "f.age"          "f.hpw"          "f.educationNum"
```

2 Carregar els paquets

Carregarem tots els paquets necessaris per utilitzar al llarg de la pràctica.

```
options(contrasts=c("contr.treatment","contr.treatment"))

requiredPackages <- c("FactoMineR", "car", "factoextra", "NbClust", "knitr")
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package
if(length(missingPackages)) install.packages(missingPackages)

lapply(requiredPackages, require, character.only = TRUE)
```

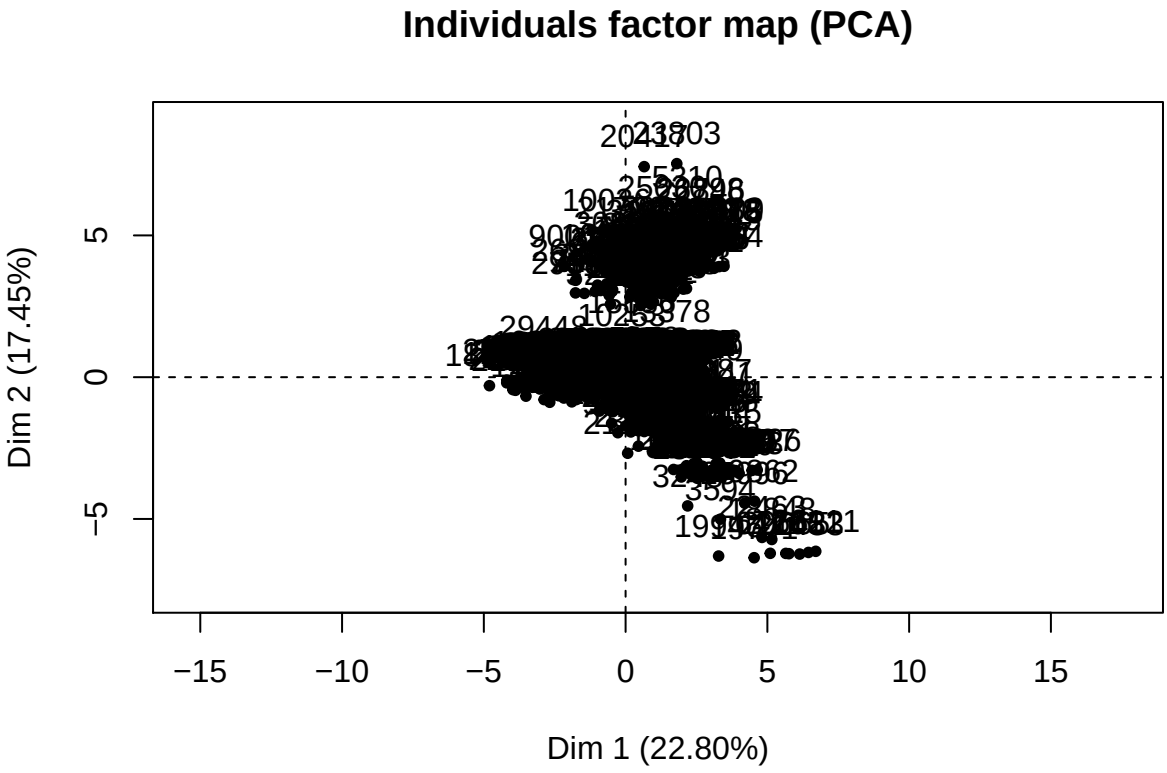
```
## Loading required package: FactoMineR

## Warning: package 'FactoMineR' was built under R version 3.6.1

## Loading required package: car

## Warning: package 'car' was built under R version 3.6.1

## Loading required package: carData

## Loading required package: factoextra

## Warning: package 'factoextra' was built under R version 3.6.1

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.1

## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13E

## Loading required package: NbClust

## Loading required package: knitr

## Warning: package 'knitr' was built under R version 3.6.1

## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
```

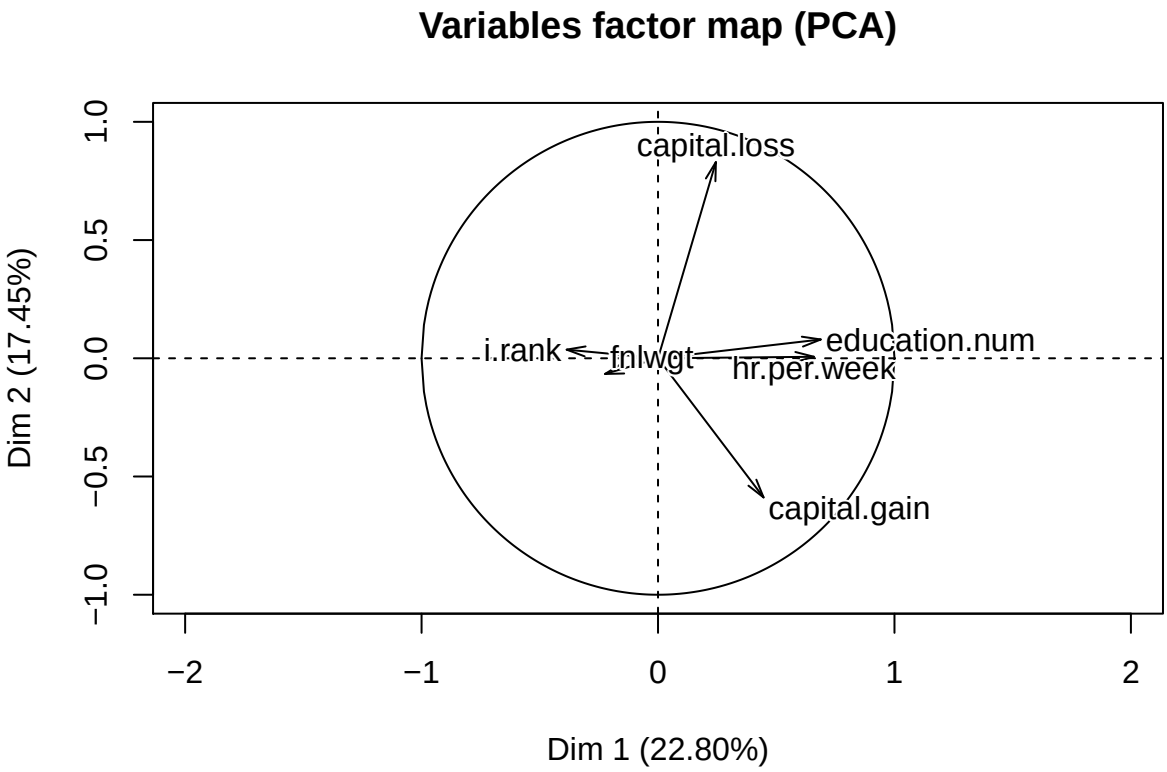
```
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
```

3 Anàlisi PCA

Inicialment comencem amb l'anàlisi de components principals. En primer lloc farem l'ànalisi sobre un PCA simple sense tenir en compte variables suplementàries i posteriorment sí que les tindrem en compte. Comencem:

```
res.pca <- PCA(df[,vars_con])
```





Com sabem, aquelles variables que tenen un angle de 90 graus o 270 aproximadament, depenent de com tinguem en compte la direcció, no estan relacionades.

Donant una primera ullada al PCA obtingut, veiem que per exemple hr.per.week i el capital loss estan molt poc relacionades. Mentre que podem veure que el education num i el hr.per.week estan bastant relacionada, ja que tenen direccions molt semblants. Es a dir que podriem dir que aquestes variables, a simple vista són grans candidates a estar relacionades positivament entre elles.

D'altre banda veiem que el i.rank esta inversament relacionat amb les hr.per.week, amb això podriem deduïr, que les persones que treballen més hores, tendeixen a no tenir errors en les enquestes ni a deixar preguntes en blanc.

3.1 Anàlisi dels *eigenvalues* i eixos dominants

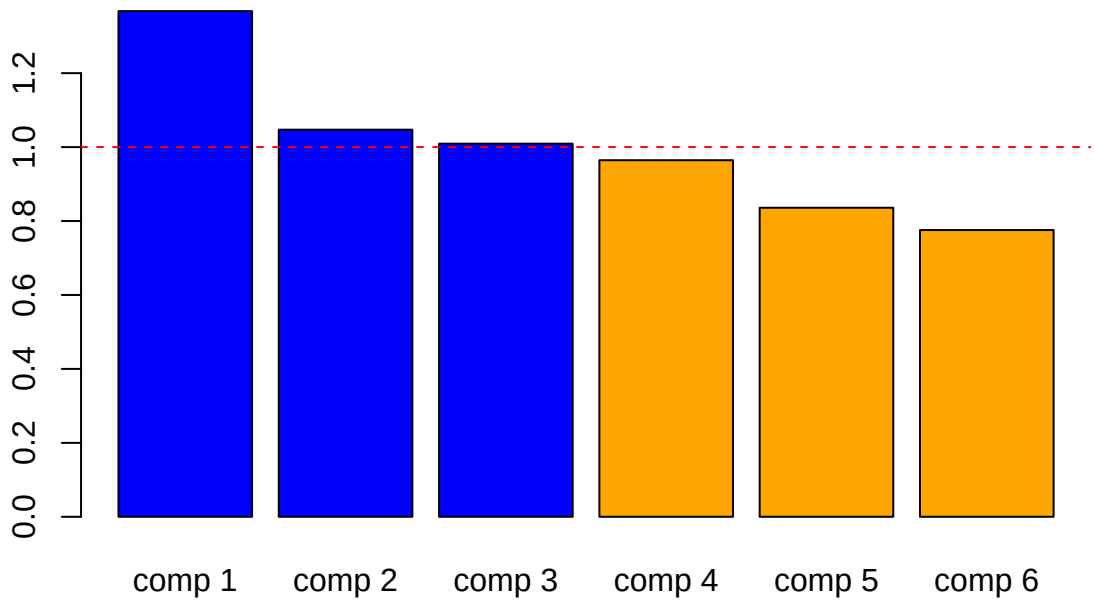
Dins els resultats del PCA, recollim els *eigenvalues* i el percentatge explicat per cada dimensió.

```
summary(res.pca,nb.dec=2,nbelements = Inf,nbind = 0)
```

```
##
## Call:
## PCA(X = df[, vars_con])
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6
## Variance                1.37   1.05   1.01   0.96   0.84   0.78
## % of var.              22.80  17.45  16.82  16.07  13.93  12.93
## Cumulative % of var.   22.80  40.25  57.07  73.14  87.07 100.00
```

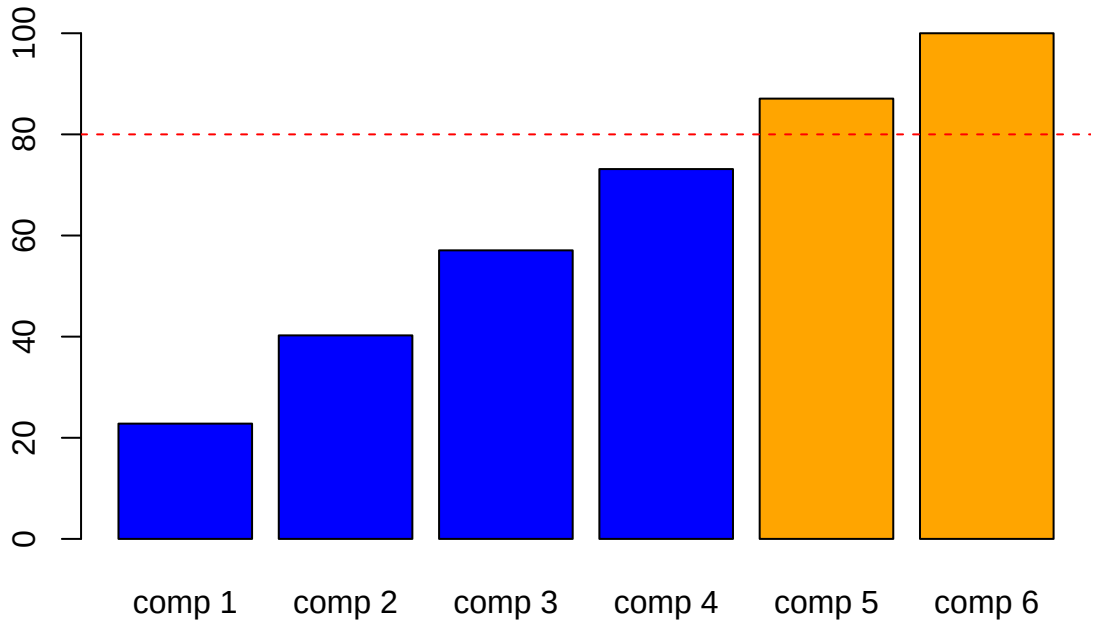
```
##
## Variables
##                  Dim.1   ctr  cos2   Dim.2   ctr  cos2   Dim.3   ctr  cos2
## fnlwgt         | -0.22  3.68  0.05 | -0.07  0.42  0.00 | -0.69 47.58  0.48
## education.num |  0.69 34.65  0.47 |  0.08  0.60  0.01 |  0.12  1.52  0.02
## capital.gain  |  0.45 14.58  0.20 | -0.59 33.08  0.35 |  0.27  7.48  0.08
## capital.loss  |  0.24  4.38  0.06 |  0.83 65.76  0.69 |  0.10  1.04  0.01
## hr.per.week   |  0.66 31.79  0.43 |  0.01  0.00  0.00 | -0.23  5.27  0.05
## i.rank         | -0.39 10.92  0.15 |  0.04  0.13  0.00 |  0.61 37.10  0.37
##
## fnlwgt         |
## education.num |
## capital.gain  |
## capital.loss  |
## hr.per.week   |
## i.rank         |
```

```r
colors<-c("Blue", "orange")
barplot(res.pca$eig[,1], col = colors[ifelse(res.pca$eig[,1] >= 1 , 1,2)])
abline(h=1, col = "red", lty=2)
```



Recordem que el **criteri de Kaiser** ens establia les dimensións rellevants com aquelles que tenien una variança major a 1.0. En aquest cas, podem observar que les dimensions que ens interessen son les **tres primeres**. És interessant veure que aquestes tres dimensions juntes expliquen un 57.07% de la *intertia*.

```r
barplot(res.pca$eig[,3],col= colors[ifelse(res.pca$eig[,3] < 80, 1, 2)])
abline(h=80,col="red",lty=2)
```

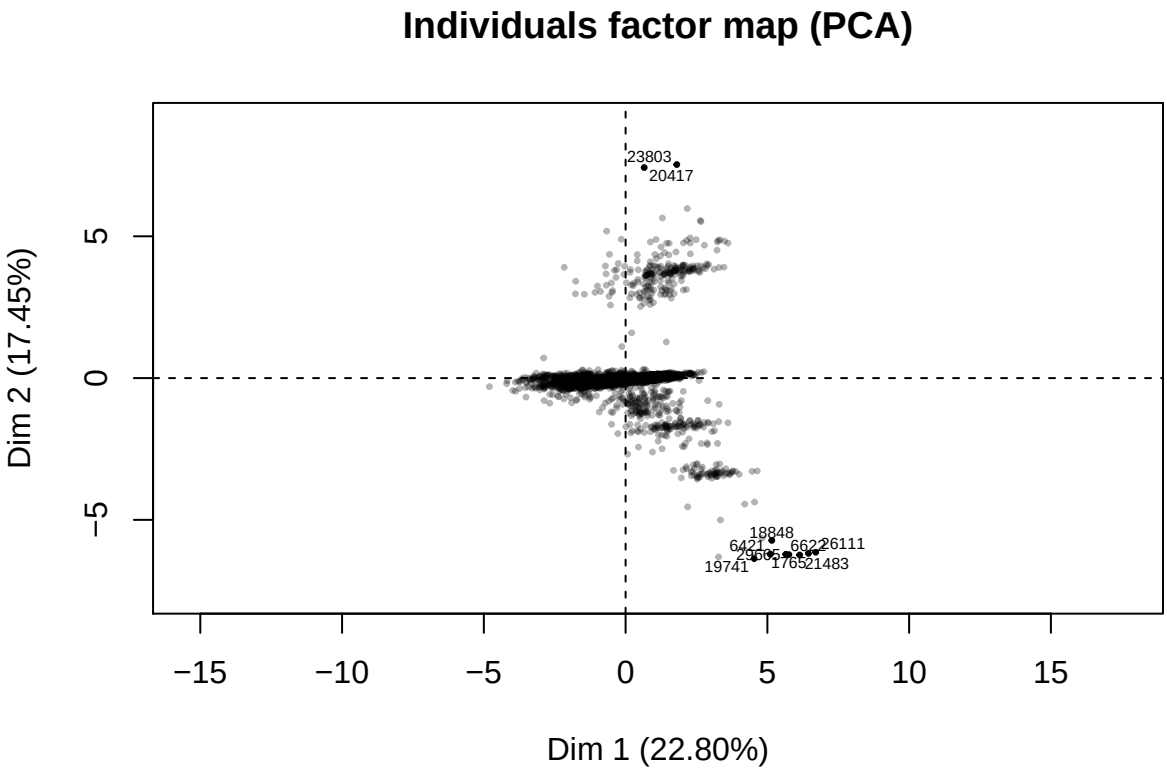


D'altra banda, veiem que en relació al **criteri d'Elbow** que el que ens diu és que recollim com a dimensións rellevants aquelles que arribin a explicar el 80% de la variança. En aquest cas, veiem que hauriem d'agafar les primeres 5 dimensións que expliquen el 87.07%, ja que amb 4 dimensións només arribariem al 73.14%.

3.2 Anàlisi dels individus

A continuació realitzarem un anàlisi des del punt de vista dels individus. Voldrem veure quins d'aquest són els més contributius, és a dir, quins es situarán més als extrems del plot. Concretament destacarem els 10 més contributius, i ressaltarem també en quin valor ho fan.

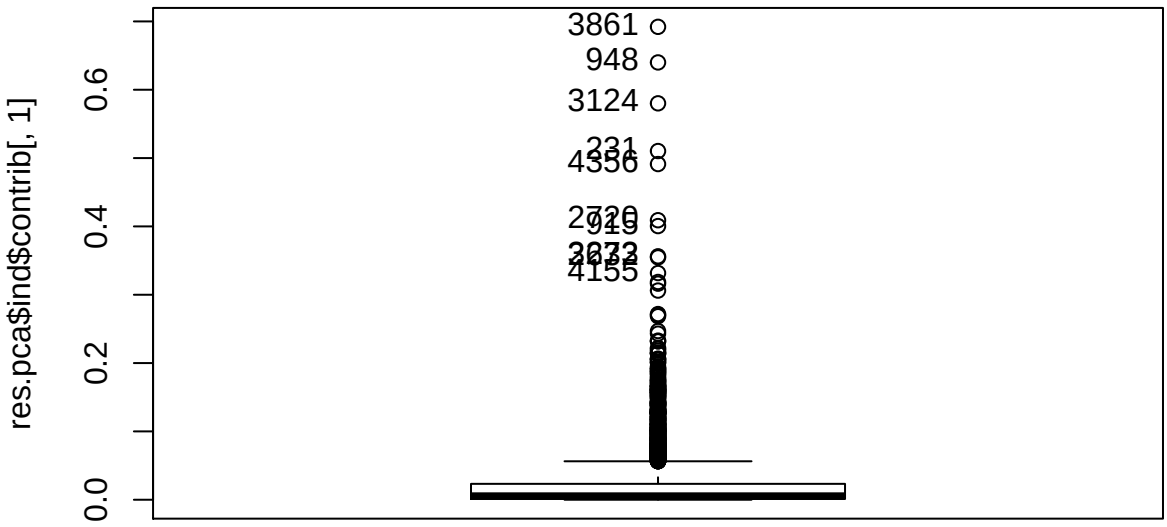
```r
plot(res.pca,choix="ind",select="contrib 10",cex=0.5)
```



3.3 Interpretació dels eixos

Seguidament, farem un anàlisi dels eixos. Per a decidir quins analitzarem, ens decantarem pel criteri de Kaiser i per tan realitzarem aquest anàlisi sobre els 3 primers.

```
#Dim 1
Boxplot(res.pca$ind$contrib[,1])
```




```
##  [1] 3861  948 3124  231 4356 2720  915 3273 2632 4155
```

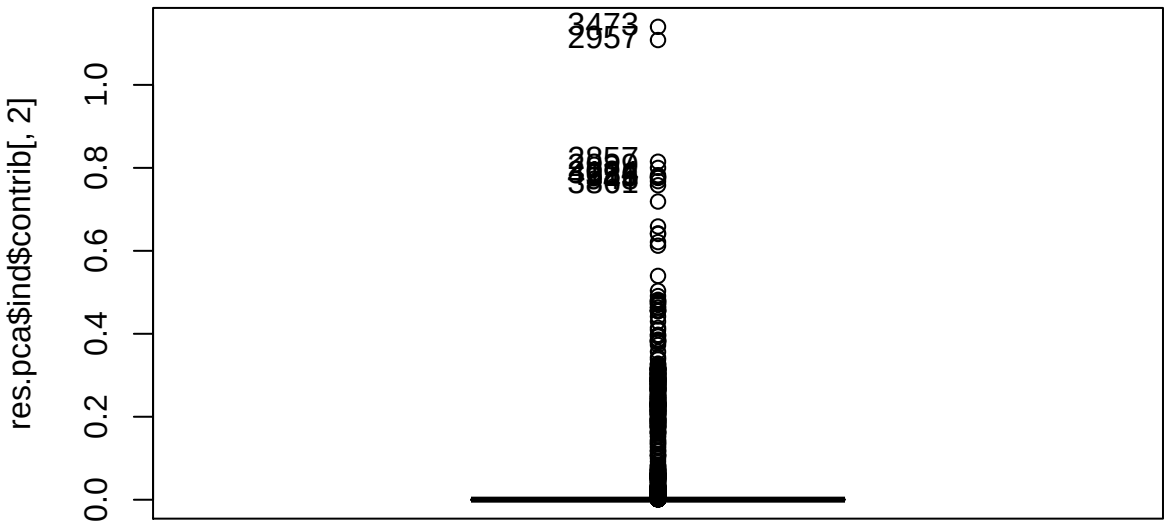
```
rang1<-order(res.pca$ind$contrib[,1],decreasing = T); rang1[1:10]
```

```
##  [1] 3861  948 3124  231 4356 2720  915 3273 2632 4155
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "26111" "6622"  "21483" "1765"  "29665" "18848" "6421"  "22463"
##  [9] "18195" "28326"
```

```
#Dim 2
Boxplot(res.pca$ind$contrib[,2])
```




```
##  [1] 3473 2957 2857 2889 3124  231 4356  915  948 3861
```

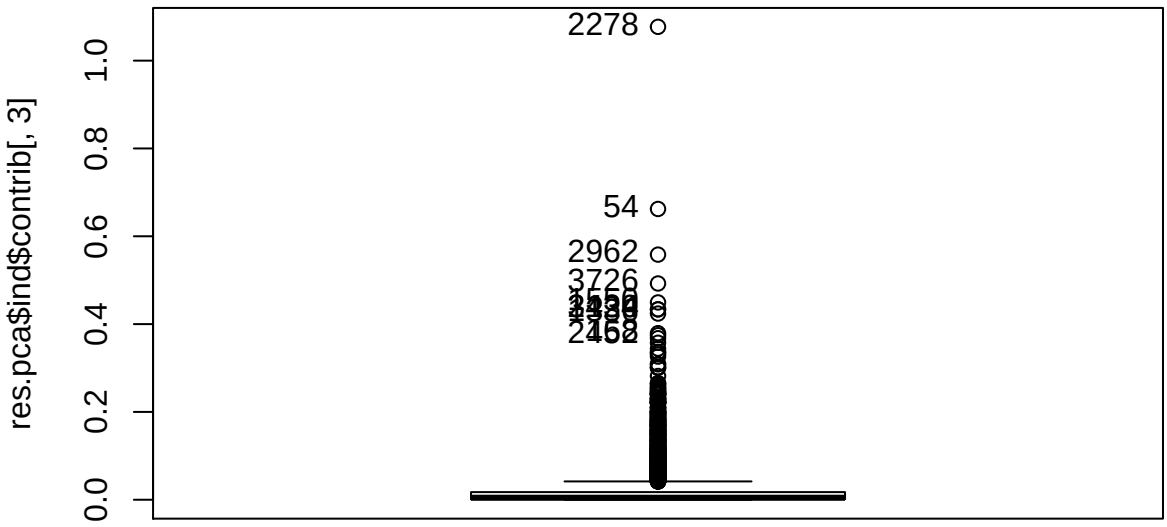
```
rang1<-order(res.pca$ind$contrib[,2],decreasing = T); rang1[1:10]
```

```
##  [1] 3473 2957 2857 2889 3124  231 4356  915  948 3861
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "23803" "20417" "19741" "19946" "21483" "1765"  "29665" "6421"
##  [9] "6622"  "26111"
```

```
#Dim 3
Boxplot(res.pca$ind$contrib[,3])
```




```
##  [1] 2278   54 2962 3726 1550 3434 2120 1336  168 2452
```

```
rang1<-order(res.pca$ind$contrib[,3],decreasing = T); rang1[1:10]
```

```
##  [1] 2278   54 2962 3726 1550 3434 2120 1336  168 2452
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "15570" "415"   "20489" "25360" "10674" "23524" "14415" "9246"
##  [9] "1292"  "16839"
```

3.4 Anàlisi PCA amb variables suplementaries

Finalment repetirem el PCA realitzat amb anterioritat però aquest cop utilitzant variables suplementaries. Aquestes variables són aquelles que no tenen influéncia en l'anàlisi de components principals i ens ajudaràn a poder interpretar millor les dimensións de la variablitat.

En el nostre cas, com a variables suplementaries qualitatives hem agafat:

- hr.per.week

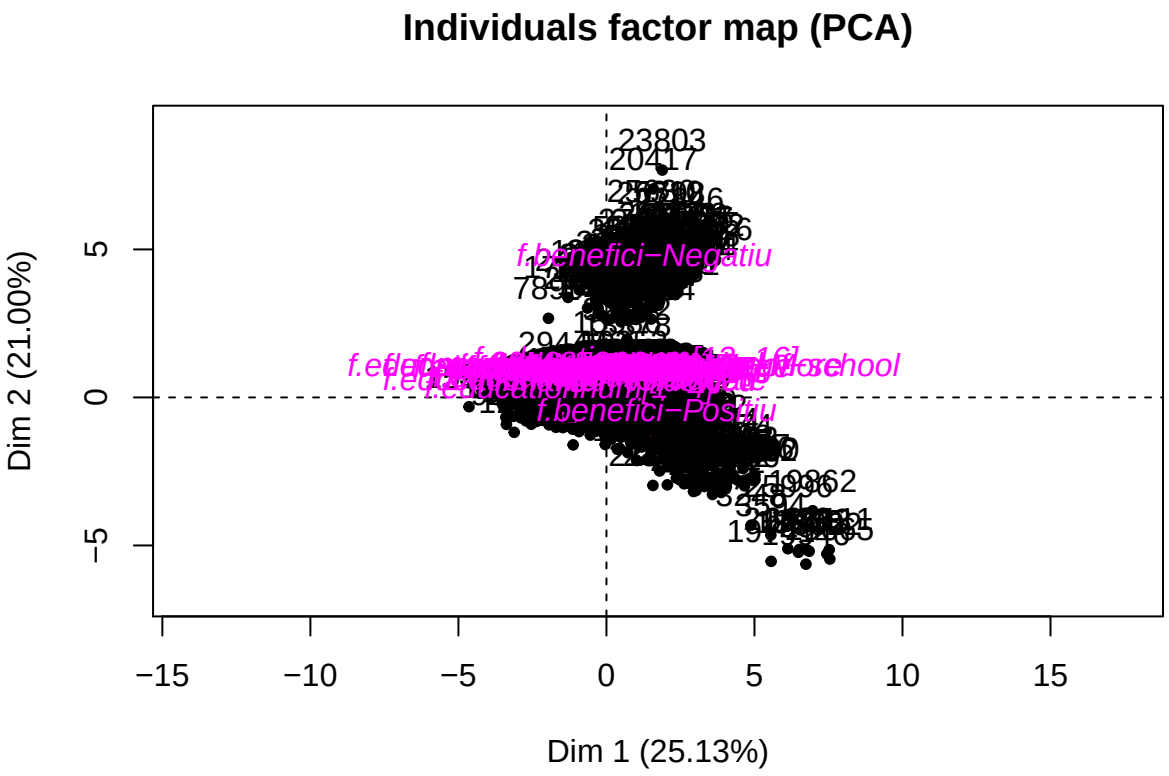
I com a suplementaries quantitatives:

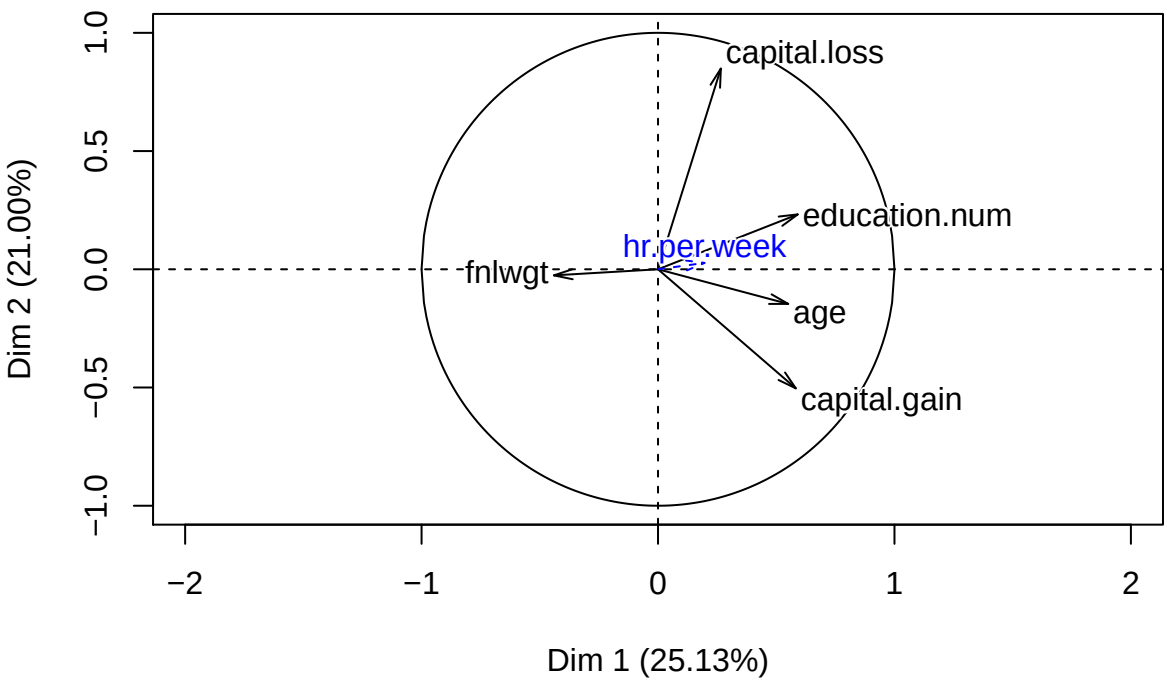
- y.bin
- f.type
- f.marital
- sex
- f.education

- f.continent
- f.benefici
- f.age
- f.hpw
- f.educationNum

La selecció ha estat aquesta perque considerem que compleixen el rol de variables complementaries i així aconseguirem fer un anàlisi millor del PCA.

```
res.pca <- PCA(df[,c(1,3,5,11:13,15:23)], quanti.sup=6,quali.sup =  7:15)
```






```
names(df[,c(1,3,5,11:13,15:24)])
```

```
##  [1] "age"            "fnlwgt"         "education.num"  "capital.gain"
##  [5] "capital.loss"   "hr.per.week"    "y.bin"          "f.type"
##  [9] "f.marital"      "f.education"    "f.continent"    "f.benefici"
## [13] "f.age"          "f.hpw"          "f.educationNum" "i.rank"
```

```
summary(res.pca,nb.dec=2,nbelements = Inf,nbind = 0)
```

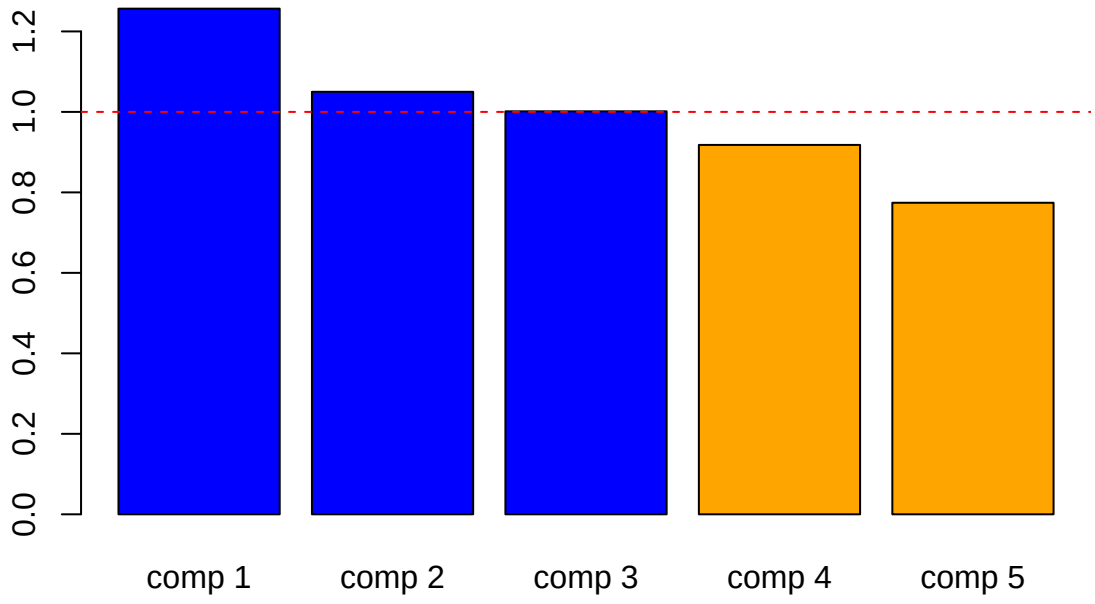
```
##
## Call:
## PCA(X = df[, c(1, 3, 5, 11:13, 15:23)], quanti.sup = 6, quali.sup = 7:15)
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5
## Variance                1.26   1.05   1.00   0.92   0.77
## % of var.              25.13  21.00  20.03  18.36  15.48
## Cumulative % of var.   25.13  46.13  66.16  84.52 100.00
##
## Variables
##                                  Dim.1   ctr  cos2   Dim.2   ctr  cos2
## age                            |  0.55 24.13  0.30 | -0.15  2.04  0.02
## fnlwgt                         | -0.44 15.43  0.19 | -0.03  0.06  0.00
## education.num                  |  0.59 27.79  0.35 |  0.23  5.13  0.05
## capital.gain                   |  0.58 27.03  0.34 | -0.50 24.15  0.25
## capital.loss                   |  0.27  5.62  0.07 |  0.85 68.62  0.72
```

```
##                                   Dim.3   ctr  cos2
## age                              | -0.45 20.27  0.20 |
## fnlwgt                           |  0.62 38.69  0.39 |
## education.num                    |  0.54 28.75  0.29 |
## capital.gain                     |  0.35 12.29  0.12 |
## capital.loss                     |  0.00  0.00  0.00 |
##
## Supplementary continuous variable
##                                   Dim.1 cos2   Dim.2 cos2   Dim.3 cos2
## hr.per.week                      |  0.20 0.04 |  0.03 0.00 |  0.08 0.01
##
## hr.per.week                      |
##
## Supplementary categories
##                                      Dist    Dim.1   cos2 v.test
## <=50K                            |   0.30 |  -0.28   0.86 -31.14 |
## >50K                             |   0.98 |   0.91   0.86  31.14 |
## f.typ-Civil                      |   0.22 |   0.15   0.44   4.49 |
## f.typ-Private                    |   0.14 |  -0.12   0.73 -11.47 |
## f.typ-SelfEm                     |   0.96 |   0.92   0.94  10.41 |
## f.typ-Other                      |   0.55 |   0.37   0.46   6.33 |
## f.marital-Married                |   0.37 |   0.29   0.62  16.75 |
## f.marital-No- Married            |   0.30 |   0.05   0.02   1.33 |
## f.marital-Never-married          |   0.79 |  -0.51   0.41 -21.64 |
## f.marital-Widowed                |   1.53 |   0.63   0.17   7.13 |
## f.education-Non-Graduate         |   1.74 |  -1.05   0.37 -25.46 |
## f.education-Some-college         |   0.22 |  -0.15   0.46  -4.80 |
## f.education-University-Or-More   |   0.31 |   0.22   0.49  14.70 |
## f.education-Assoc_AND_Proof-school |   0.80 |   0.58   0.53  11.54 |
## f.continent-America              |   0.01 |   0.00   0.18  -1.18 |
## f.continent-Asia                 |   0.39 |   0.08   0.05   0.70 |
## f.continent-Europa               |   0.21 |   0.13   0.39   0.98 |
## f.benefici-Neutre                |   0.32 |  -0.23   0.49 -36.24 |
## f.benefici-Positiu               |   2.67 |   1.70   0.41  30.70 |
## f.benefici-Negatiu               |   4.40 |   1.28   0.08  17.69 |
## f.age-[17,29]                    |   1.13 |  -0.77   0.46 -31.18 |
## f.age-(29,39]                    |   0.32 |  -0.08   0.06  -2.91 |
## f.age-(39,49]                    |   0.49 |   0.39   0.64  12.29 |
## f.age-(49,90]                    |   1.48 |   0.79   0.28  25.63 |
## f.hpw[10-20]                     |   0.50 |  -0.36   0.52  -4.52 |
## f.hpw[20-30]                     |   0.38 |  -0.35   0.83  -6.03 |
## f.hpw[30-40]                     |   0.16 |  -0.14   0.82  -3.19 |
## f.hpw[40-50]                     |   0.05 |  -0.04   0.64  -2.98 |
## f.hpw[50-60]                     |   0.50 |   0.47   0.88  13.10 |
## f.educationnum[1-4]              |   2.71 |  -1.25   0.21 -15.47 |
## f.educationnum[13-16]            |   1.40 |   1.00   0.51  34.46 |
## f.educationnum[5-8]              |   1.39 |  -0.97   0.48 -19.37 |
## f.educationnum[9-12]             |   0.18 |  -0.16   0.74 -12.49 |
##                                   Dim.2   cos2 v.test    Dim.3   cos2
```

```
## <=50K                                -0.03   0.01  -3.12 |  -0.10   0.10
## >50K                                  0.08   0.01   3.12 |   0.32   0.10
## f.typ-Civil                           0.04   0.03   1.36 |   0.02   0.01
## f.typ-Private                         0.00   0.00   0.11 |   0.02   0.02
## f.typ-SelfEm                         -0.12   0.02  -1.45 |   0.04   0.00
## f.typ-Other                          -0.07   0.02  -1.28 |  -0.29   0.28
## f.marital-Married                     0.00   0.00  -0.15 |  -0.10   0.07
## f.marital-No- Married                -0.10   0.10  -2.97 |  -0.19   0.43
## f.marital-Never-married               0.10   0.02   4.54 |   0.34   0.18
## f.marital-Widowed                    -0.42   0.07  -5.16 |  -0.93   0.37
## f.education-Non-Graduate             -0.39   0.05 -10.34 |  -0.92   0.28
## f.education-Some-college              0.04   0.03   1.39 |   0.05   0.05
## f.education-University-Or-More        0.06   0.04   4.80 |   0.13   0.19
## f.education-Assoc_AND_Proof-school    0.09   0.01   1.94 |   0.43   0.28
## f.continent-America                   0.00   0.05  -0.68 |   0.00   0.00
## f.continent-Asia                      0.09   0.05   0.83 |   0.13   0.11
## f.continent-Europa                    0.01   0.00   0.08 |  -0.14   0.41
## f.benefici-Neutre                    -0.08   0.06 -13.81 |  -0.08   0.07
## f.benefici-Positiu                   -1.43   0.29 -28.35 |   0.90   0.11
## f.benefici-Negatiu                    3.80   0.75  57.44 |   0.02   0.00
## f.age-[17,29]                         0.12   0.01   5.49 |   0.42   0.14
## f.age-(29,39]                         0.08   0.07   3.26 |   0.22   0.50
## f.age-(39,49]                        -0.01   0.00  -0.37 |  -0.09   0.04
## f.age-(49,90]                        -0.26   0.03  -9.21 |  -0.76   0.27
## f.hpw[10-20]                         -0.03   0.00  -0.44 |  -0.29   0.35
## f.hpw[20-30]                         -0.05   0.02  -1.03 |  -0.08   0.04
## f.hpw[30-40]                         -0.02   0.02  -0.48 |  -0.02   0.02
## f.hpw[40-50]                          0.00   0.00  -0.13 |  -0.01   0.03
## f.hpw[50-60]                          0.05   0.01   1.52 |   0.15   0.09
## f.educationnum[1-4]                  -0.65   0.06  -8.83 |  -1.63   0.36
## f.educationnum[13-16]                 0.33   0.06  12.45 |   0.69   0.24
## f.educationnum[5-8]                  -0.28   0.04  -6.19 |  -0.63   0.21
## f.educationnum[9-12]                 -0.04   0.05  -3.70 |  -0.07   0.14
##                                     v.test
## <=50K                               -12.20 |
## >50K                                 12.20 |
## f.typ-Civil                           0.73 |
## f.typ-Private                         2.26 |
## f.typ-SelfEm                          0.46 |
## f.typ-Other                          -5.47 |
## f.marital-Married                    -6.20 |
## f.marital-No- Married                -6.22 |
## f.marital-Never-married              16.17 |
## f.marital-Widowed                   -11.83 |
## f.education-Non-Graduate            -25.00 |
## f.education-Some-college              1.77 |
## f.education-University-Or-More       10.12 |
## f.education-Assoc_AND_Proof-school    9.48 |
## f.continent-America                  -0.16 |
```

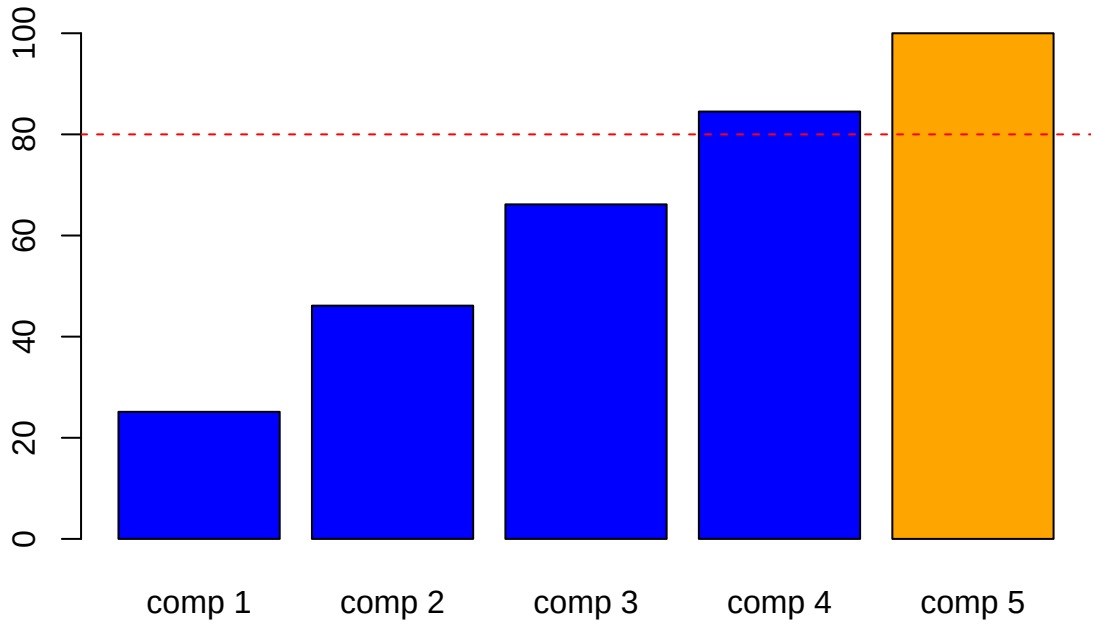
```
## f.continent-Asia                    1.20 |
## f.continent-Europa                  -1.13 |
## f.benefici-Neutre                  -15.02 |
## f.benefici-Positiu                  18.22 |
## f.benefici-Negatiu                   0.38 |
## f.age-[17,29]                       19.33 |
## f.age-(29,39]                        9.20 |
## f.age-(39,49]                       -3.37 |
## f.age-(49,90]                      -27.97 |
## f.hpw[10-20]                        -4.17 |
## f.hpw[20-30]                        -1.53 |
## f.hpw[30-40]                        -0.61 |
## f.hpw[40-50]                        -0.67 |
## f.hpw[50-60]                         4.62 |
## f.educationnum[1-4]                -22.56 |
## f.educationnum[13-16]               26.78 |
## f.educationnum[5-8]                -14.16 |
## f.educationnum[9-12]                -6.03 |
```

```r
barplot(res.pca$eig[,1], col = colors[ifelse(res.pca$eig[,1] >= 1 , 1,2)])
abline(h=1, col = "red", lty=2)
```



Podem observar que el nombre de dimensión a seleccionar amb el criteri de Kaiser és 3, el mateix que amb l'anàlisi de components principals sense variables suplementàries. En aquest cas s'explica un 66.16% de la variança.

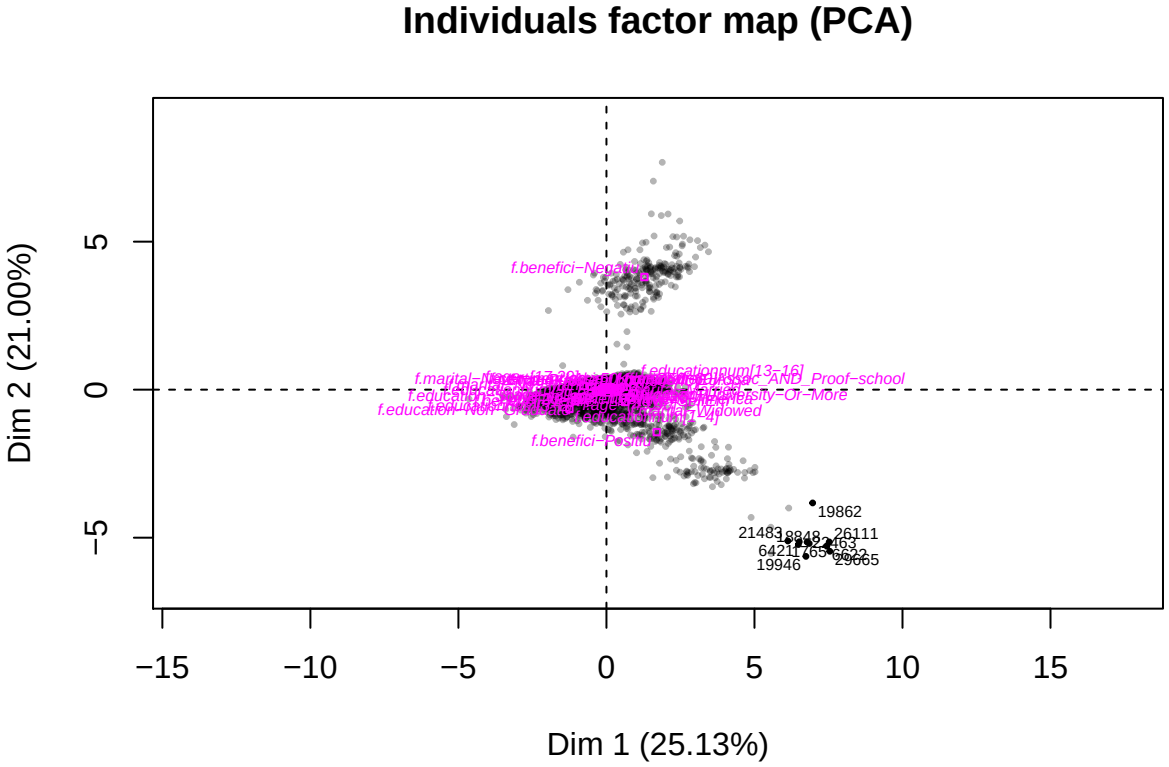
```
barplot(res.pca$eig[,3],col= colors[ifelse(res.pca$eig[,3] < 85, 1, 2)])
abline(h=80,col="red",lty=2)
```



En canvi, recordem que anteriorment amb el criteri d'Elbow seleccionavem les 5 primeres dimensións, i ara seleccionem les 4 primeres, que arriben a explicar un 84.52% de la variança.

Ara analitzem els individus més contributius.

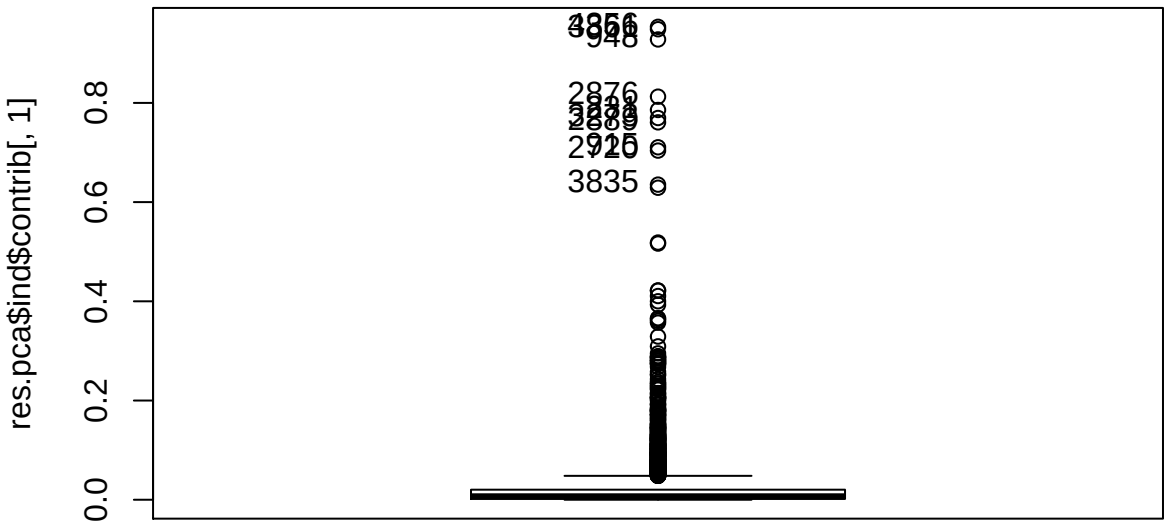
```
plot(res.pca,choix="ind",select="contrib 10",cex=0.5)
```



Veiem que com a diferència amb l'anàlisi anterior, tenim que ara tots els elemens més contributius els tenim situats a l'extrem del quadrant inferior dret, mentre que abans els teniem al quadrant dret pero distribuïts entre el superior i l'inferior.

Pel que fa a l'analisi de les dimensións, on seguim amb les tres primeres dimensións pel criteri de Kaiser, tenim:

```
#Dim 1
Boxplot(res.pca$ind$contrib[,1])
```




```
##  [1] 4356 3861  948 2876  231 3273 2889  915 2720 3835
```

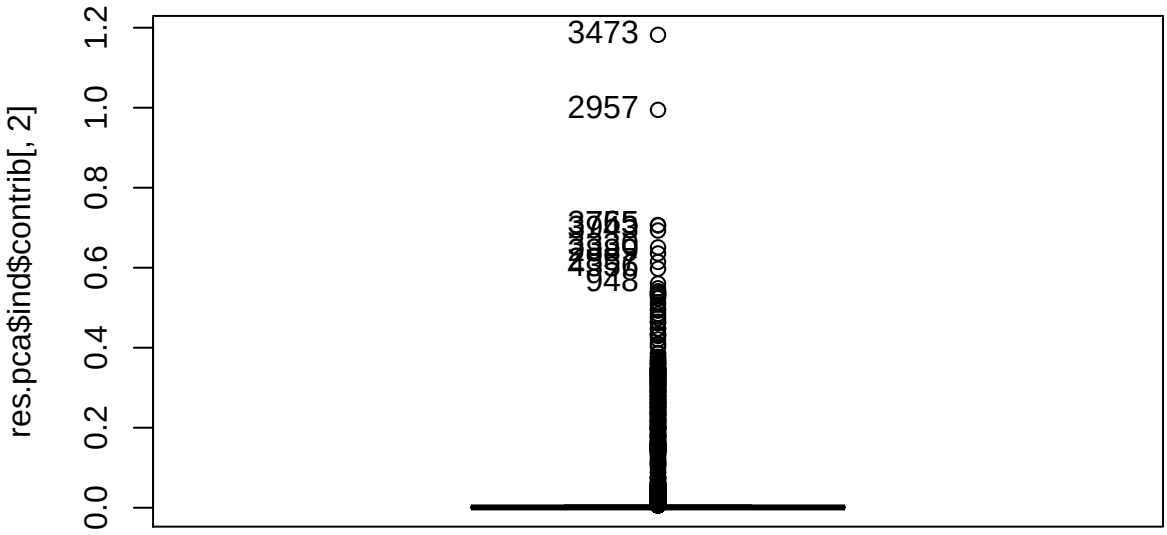
```
rang1<-order(res.pca$ind$contrib[,1],decreasing = T); rang1[1:10]
```

```
##  [1] 4356 3861  948 2876  231 3273 2889  915 2720 3835
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "29665" "26111" "6622"  "19862" "1765"  "22463" "19946" "6421"
##  [9] "18848" "25996"
```

```
#Dim 2
Boxplot(res.pca$ind$contrib[,2])
```




```
##  [1] 3473 2957 3765  755 3943 3330 2889 2857 4356  948
```

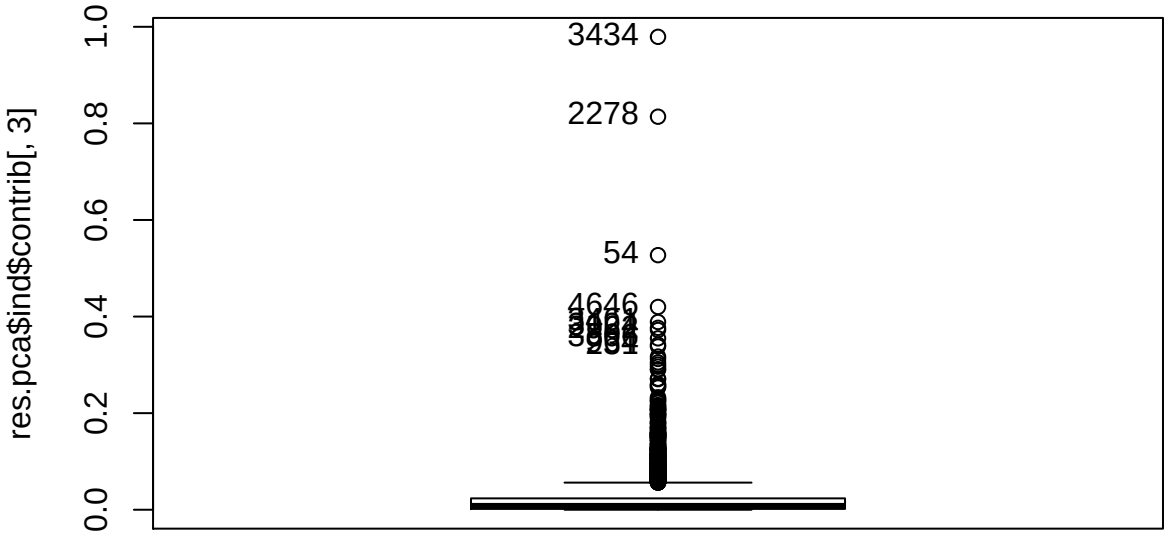
```
rang1<-order(res.pca$ind$contrib[,2],decreasing = T); rang1[1:10]
```

```
##  [1] 3473 2957 3765  755 3943 3330 2889 2857 4356  948
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "23803" "20417" "25630" "5310"  "26798" "22846" "19946" "19741"
##  [9] "29665" "6622"
```

```
#Dim 3
Boxplot(res.pca$ind$contrib[,3])
```




```
##  [1] 3434 2278   54 4646 3461 3124 2962 3806  964  231
```

```
rang1<-order(res.pca$ind$contrib[,3],decreasing = T); rang1[1:10]
```

```
##  [1] 3434 2278   54 4646 3461 3124 2962 3806  964  231
```

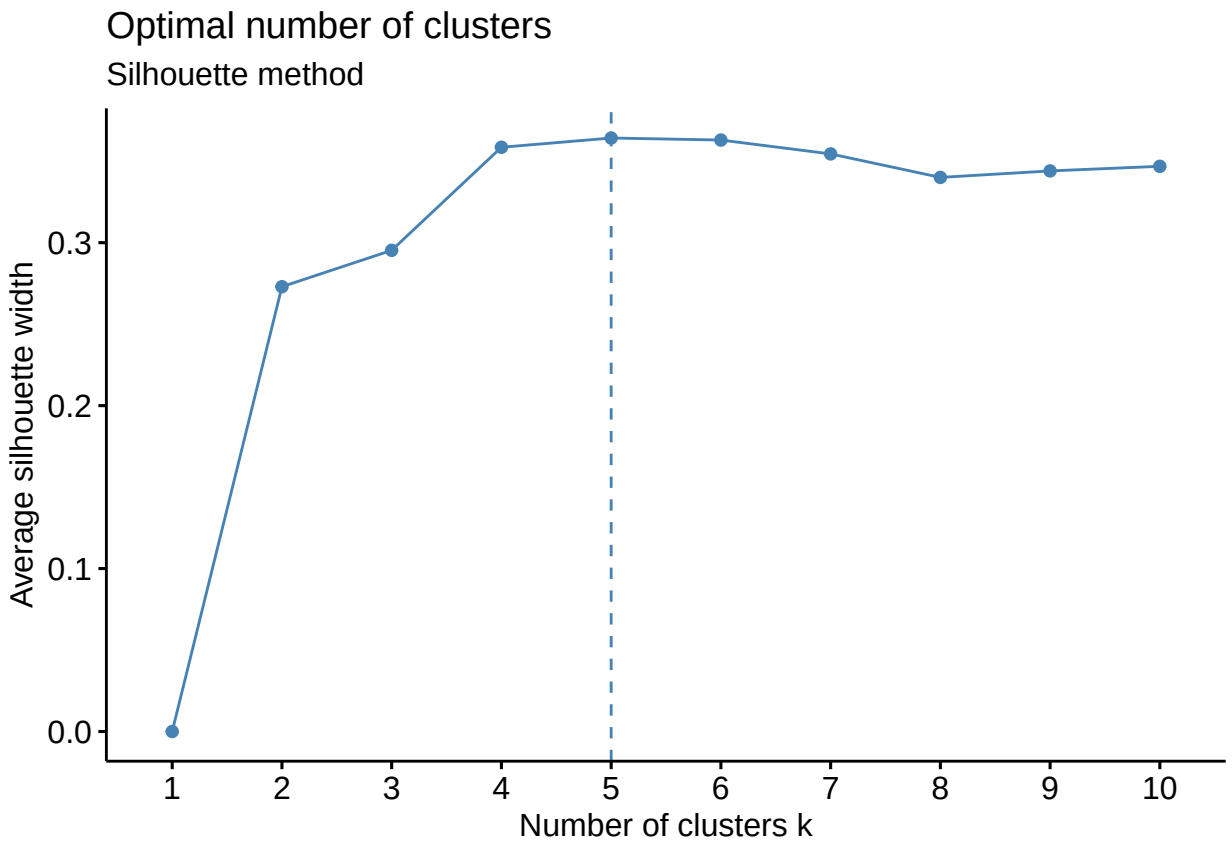
```
rownames(df[rang1[1:10],])
```

```
##  [1] "23524" "15570" "415"   "31793" "23702" "21483" "20489" "25852"
##  [9] "6728"  "1765"
```

4 Definir el nombre de Clusters

Abans de començar a executar Kmeans i Hierarchical Clustering, hem de definir quants clusters volem usar per aquests.

```
fviz_nbclust(res.pca$ind$coord[,1:3], kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```



Hem seguit el criteri de Silhouette, que ens escolleix aquells clusters que siguin millor per tenir menor distàncies dins del cluster, i maximitzar distàncies entre diferents clusters. Com podem veure a la gràfica, el valor òptim de clusters és 5, ja que és el que te un average silhouette width més elevat.

```
num.clusters<-5
```

5 K-Means Classification

```
data.kmeans <- res.pca$ind$coord[,1:3]
kmeans.res <- kmeans(data.kmeans,num.clusters)
kmeans.res$centers
```

```
##         Dim.1       Dim.2       Dim.3
## 1  3.26391363 -2.34481583  1.73991469
## 2 -0.96092017 -0.04688232  0.59312008
## 3  1.32090535  3.86473700  0.03099065
## 4 -0.01238086 -0.27754030 -0.93034900
## 5  0.72502395  0.05001599  0.34294260
```

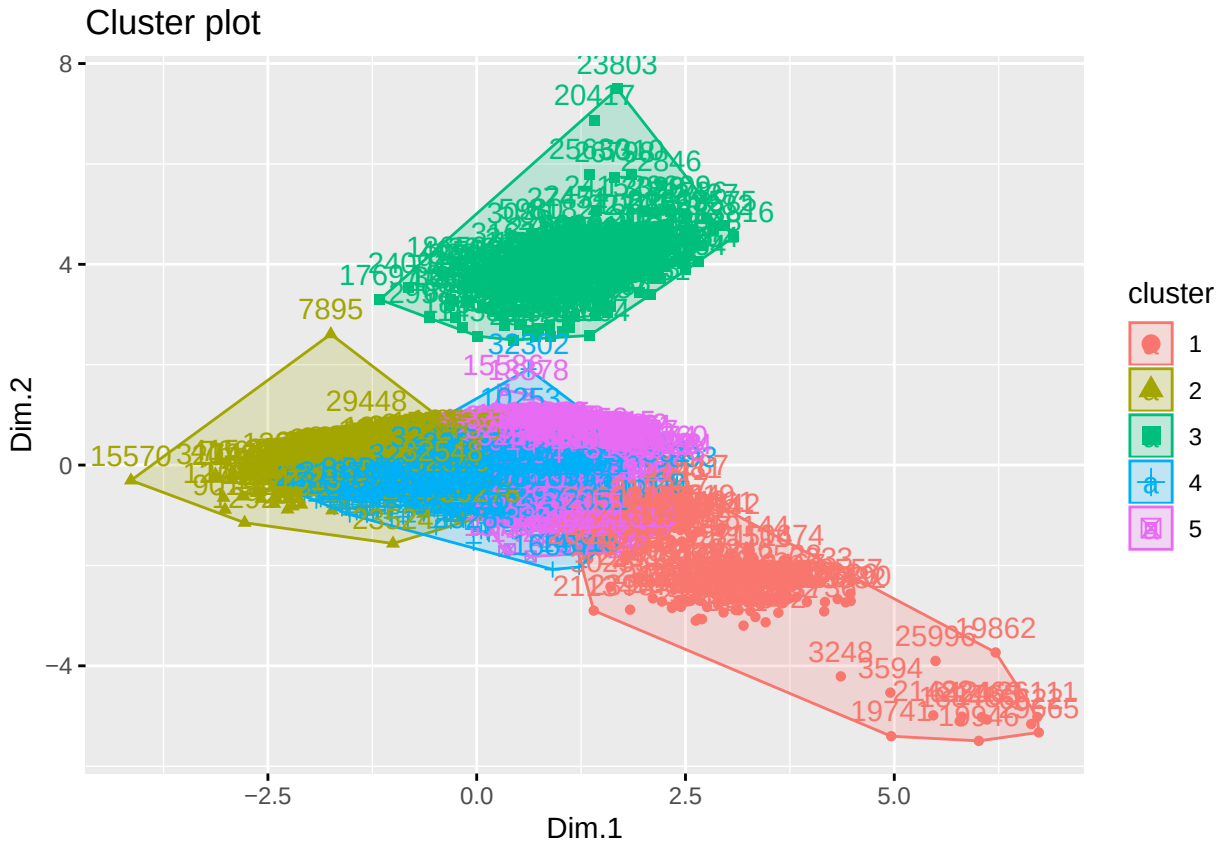
A continuació explicarem el resultat dels centres dels clusters obtinguts, ja que amb Kmeans, els centres són aquells que donen major explicailitat al cluster.

- Centre 1: Com veiem aquest centre te una gran relació positiva amb les dimensions 1 i 3, mentre que bastant negativa amb el que representa la dimensió 2. Per tant podem conclure que els individus que estiguin en aquest cluster, tindran molt en comú amb les variables que donen major explicabilitat positiva a les dimensions 1 i 3, mentre que tindràn gran representació amb aquelles que siguin negatives o inverses a la dimensió 2.

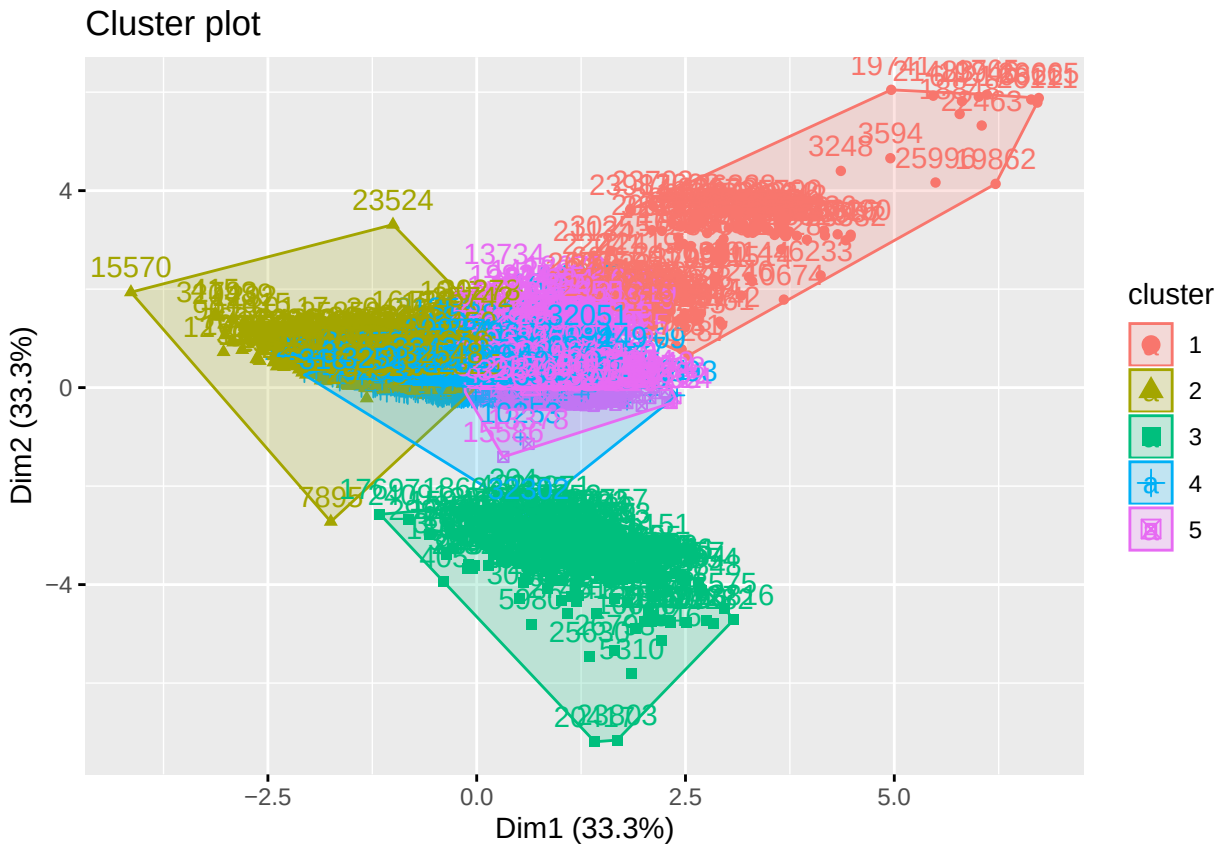
- Centre 2: Aquest centre te una relació negativa amb la dimensió 1 bastant significativa, no tant amb la tercera dimensió pero és positiva aquesta relació. Mentre que els integrants d'aquest grup, majoritariament no tenen gran relació amb la segona dimensió.
- Centre 3: Aquest grup esta fortament relacionat de forma postitiva amb les dimensions 1 i 2, com podem veure, no obstant veiem que la 3 dimensió no te una gran aportació en la descripció d'aquest grup.
- Centre 4: Aquest grup te una significativa relació negativa amb la 3ª dimensió, no obstant veiem que les altres dues no són tant rellevants, veiem que la 2ª potser te una mica de relació negativa, però sembla que no és tant significatiu com la dimensió 3.
- Centre 5: Aquest grup no sembla tenir una gran relació amb cap de les dimensions, pero veiem que la relació amb la primera dimensió és la més significativa.

A continuació mostrarem de manera gràfica el que hem explicat anteriorment.

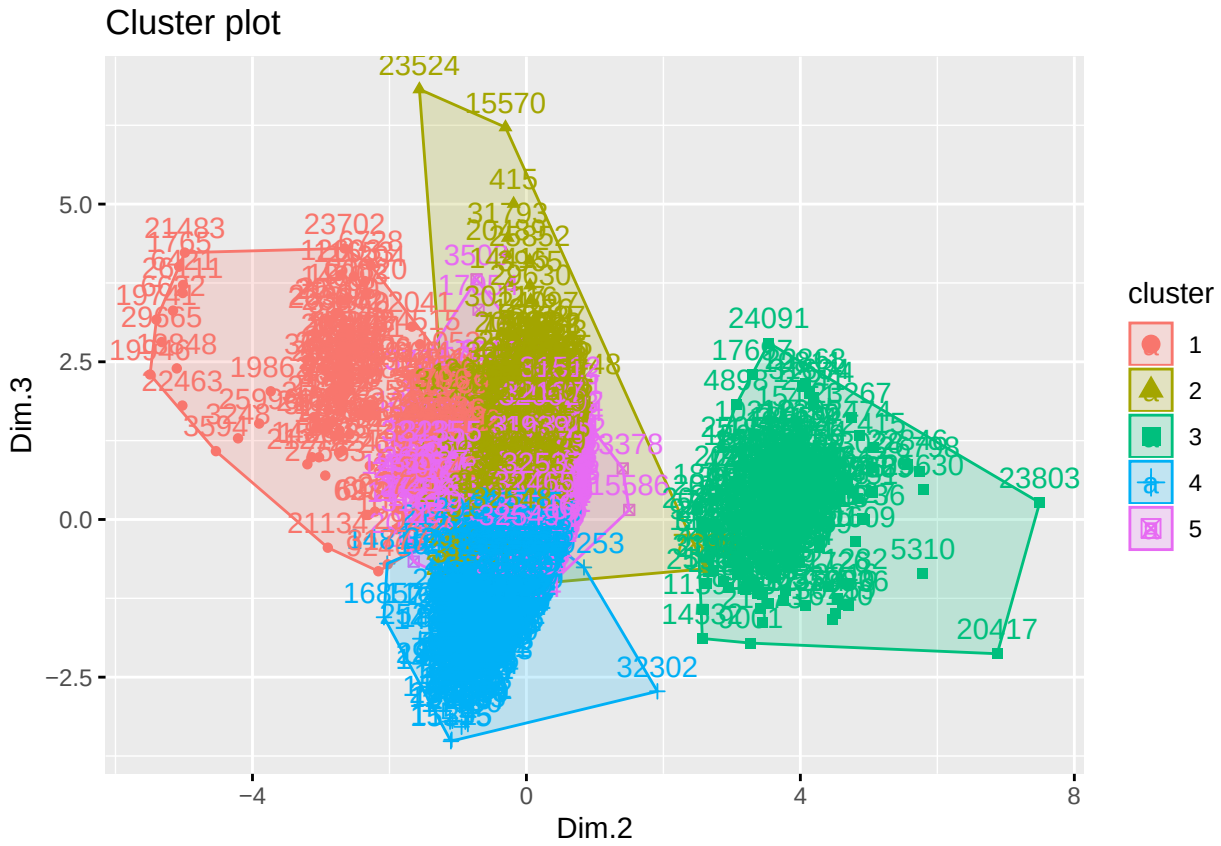
```
fviz_cluster(kmeans.res, data = data.kmeans[,1:2])
```




```
fviz_cluster(kmeans.res, data = data.kmeans[,1:3])
```

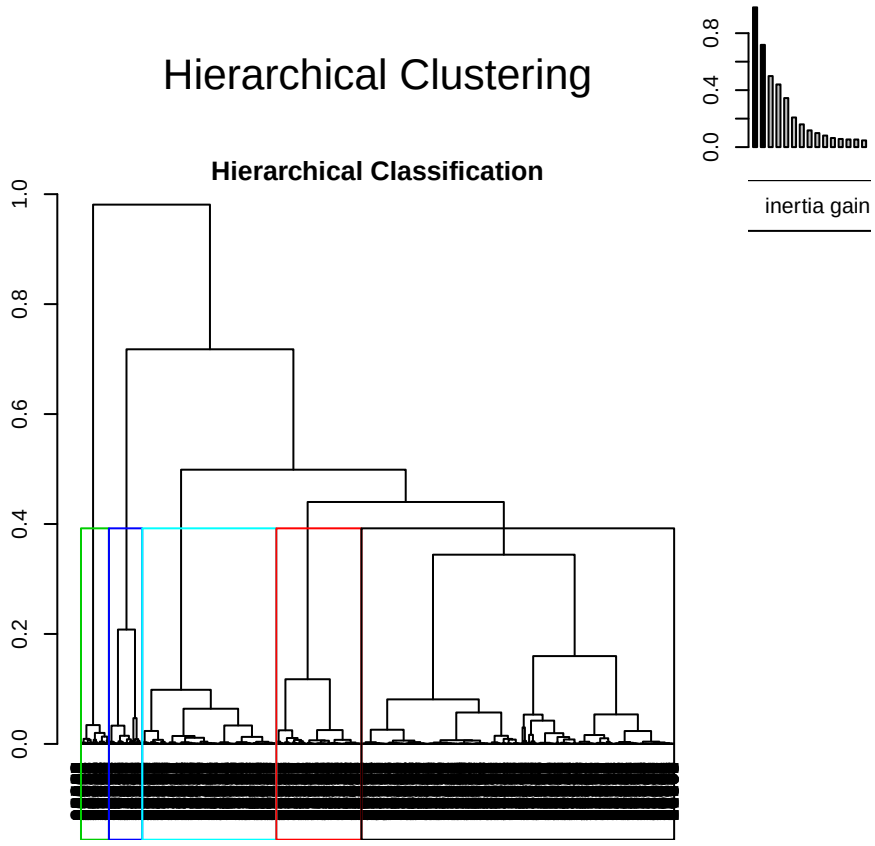


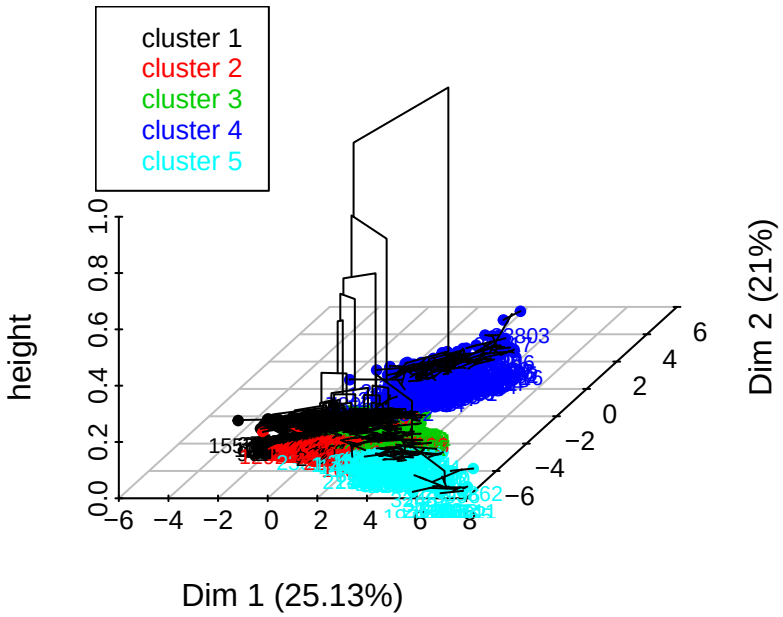

```
fviz_cluster(kmeans.res, data = data.kmeans[,2:3])
```

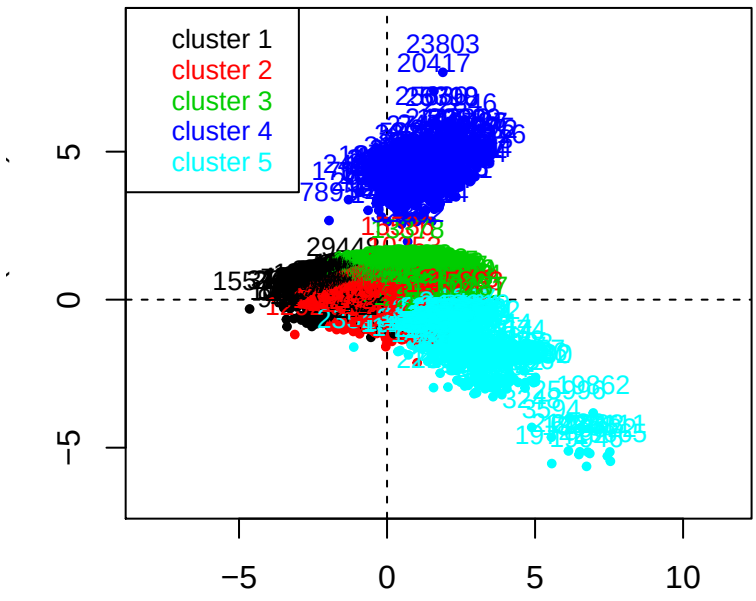


6 Hierarchical Clustering

```
res.hcpc<-HCPC(res.pca,nb.clust=num.clusters,graph=T)
```








```
output<-res.hcpc$desc.var;
output$quanti
```

```
## $`1`
##                   v.test Mean in category Overall mean sd in category
## fnlwgt          17.97953     2.268503e+05 192603.32029   1.205333e+05
## capital.gain   -10.07351     1.200517e+02    569.95991   6.634532e+02
## hr.per.week    -10.34382     3.803133e+01     39.81376   9.929309e+00
## capital.loss   -12.42875     2.187990e-01     91.13994   9.572298e+00
## education.num  -22.99849     8.987467e+00     10.03472   1.652864e+00
## age            -45.10583     2.767467e+01     38.43519   6.902466e+00
##                  Overall sd       p.value
## fnlwgt         1.078676e+05  2.818919e-72
## capital.gain   2.529242e+03  7.235097e-24
## hr.per.week    9.758405e+00  4.463617e-25
## capital.loss   4.142711e+02  1.824645e-35
## education.num  2.578697e+00 4.826412e-117
## age            1.350975e+01  0.000000e+00
##
## $`2`
##                    v.test Mean in category Overall mean sd in category
## age             41.054200     5.136247e+01     38.43519       9.723636
## capital.gain    -7.227739     1.438759e+02    569.95991     712.519239
## capital.loss    -9.321452     1.134137e+00     91.13994      29.608321
## fnlwgt         -12.373763     1.614937e+05 192603.32029   84355.478008
## education.num  -25.666137     8.492087e+00     10.03472       2.058977
```

```
##                  Overall sd       p.value
## age           1.350975e+01  0.000000e+00
## capital.gain  2.529242e+03  4.911002e-13
## capital.loss  4.142711e+02  1.147593e-20
## fnlwgt        1.078676e+05  3.624359e-35
## education.num 2.578697e+00 2.792898e-145
##
## $`3`
##                   v.test Mean in category Overall mean sd in category
## education.num 45.962010     1.316228e+01     10.03472       1.045628
## hr.per.week    8.877367     4.209973e+01     39.81376       9.003224
## age            3.982535     3.985494e+01     38.43519      11.216924
## fnlwgt        -5.652381     1.765143e+05 192603.32029   92991.145555
## capital.gain  -6.393085     1.432747e+02    569.95991     712.747618
## capital.loss  -8.282990     5.920218e-01     91.13994      19.652982
##                  Overall sd      p.value
## education.num 2.578697e+00 0.000000e+00
## hr.per.week   9.758405e+00 6.846193e-19
## age           1.350975e+01 6.818422e-05
## fnlwgt        1.078676e+05 1.582404e-08
## capital.gain  2.529242e+03 1.625719e-10
## capital.loss  4.142711e+02 1.201201e-16
##
## $`4`
##                   v.test Mean in category Overall mean sd in category
## capital.loss  67.752502       1921.96429     91.13994     339.394614
## education.num  6.879981         11.19196     10.03472       2.713019
## hr.per.week    3.287360         41.90625     39.81376       9.246870
## age            2.728151         40.83929     38.43519      11.575107
## capital.gain  -3.454755          0.00000    569.95991       0.000000
##                 Overall sd      p.value
## capital.loss   414.271117 0.000000e+00
## education.num    2.578697 5.986072e-12
## hr.per.week      9.758405 1.011313e-03
## age             13.509755 6.369036e-03
## capital.gain  2529.242128 5.507936e-04
##
## $`5`
##                   v.test Mean in category Overall mean sd in category
## capital.gain  60.360622      11637.13115    569.95991   5.229969e+03
## education.num 10.045398         11.91257     10.03472   2.211002e+00
## age            7.517837         45.79781     38.43519   1.165851e+01
## hr.per.week    6.427519         44.36066     39.81376   8.250914e+00
## fnlwgt        -2.470810     173282.62295 192603.32029   1.137583e+05
## capital.loss  -3.034810          0.00000     91.13994   0.000000e+00
##                  Overall sd      p.value
## capital.gain  2.529242e+03 0.000000e+00
## education.num 2.578697e+00 9.625806e-24
## age           1.350975e+01 5.568968e-14
```

```
## hr.per.week   9.758405e+00 1.297032e-10
## fnlwgt        1.078676e+05 1.348073e-02
## capital.loss  4.142711e+02 2.406869e-03
```

6.1 Descripció dels clusters

6.1.1 Cluster 1:

Tenen un capital.gain bastant per sota de la mitjana global. Tenen un capital.loss molt petit o practicament inexistent -> Aquest grup la majoria no tenen capital.loss i tenen un capital gain positiu majoritariament però distant de la mitjana(no tenen inversions de gran Capital) Treballen menys hores de la mitjana, però poc distant unes 38 hores Aquest grup tenen una edad bastant més jove (mitjana de 27 anys) Cluster 2: L'edat d'aquest grup es bastant superior a la mitja global (51 anys) En aquest cluster majoritariament tindrem a les persones d'edat superior. Igual que al primer clusten en quan al capital gain i capital loss Aquest grup te menys hores d'estudi que la mitjana.

6.1.2 Cluster 3:

Aquest cluster són dels que tenen major hores invertides en educació, i estan per sobre de la mirjana En quan a les hores treballades també estan en mitjana significativament per sobre de la global, però tampoc molt No tenen practicament capital loss, i tenen un capital gain bastant reduït respecte la mitjana Les edats són bastant properes a la mitjana global. (Mitjana edat segurament)

6.1.3 Cluster 4:

Aquest grup te un capital loss bastant significant respecte a la mitjana (1921). No tenen capitalgain, es a dir que són un grup que les inversions que tenen són perdudes. En quan a les hores invertides en educació, són superiors a la mitjana però no molt significatiu. Les hores treballades properes a les 42 per setmana fins a 2 hores lluny de la mitjana

6.1.4 Cluster 5:

Aquest grup tenen inversions de capital gain bastant bones, i no en tenen de negatives Estan en quan a anys d'estudi semblants al cluster 4. I l'edat es també aproximadament mitjana edat. (molt proper a la mitjana 44)

7 Análisis CA

Seguidament, farem l'anàlisi de correspondències simples per poder analitzar les relacions entre dos factors de la nostra mostra. En el nostre cas, utilitzarem les variables qualitatives "f.education" i "f.hpw".

Inicialment estudiem els perfils marginals per fila:

```
prop.table(table(df$f.education,df$f.hpw),1)
```

```
##
##                                  f.hpw[10-20] f.hpw[20-30]
##   f.education-Non-Graduate          0.09702660   0.12206573
##   f.education-Some-college          0.05555556   0.11302682
##   f.education-University-Or-More    0.02477134   0.04992378
```

```
##   f.education-Assoc_AND_Proof-school   0.01797753   0.04269663
##
##                                     f.hpw[30-40] f.hpw[40-50]
##   f.education-Non-Graduate             0.14397496   0.55555556
##   f.education-Some-college             0.11206897   0.57279693
##   f.education-University-Or-More       0.11432927   0.61318598
##   f.education-Assoc_AND_Proof-school   0.11011236   0.61348315
##
##                                     f.hpw[50-60]
##   f.education-Non-Graduate             0.08137715
##   f.education-Some-college             0.14655172
##   f.education-University-Or-More       0.19778963
##   f.education-Assoc_AND_Proof-school   0.21573034
```

En base a la darrera columna, podriem dir que a major grau educatiu, major quantita d'hores es treballen. No obstant, veiem que més o menys son perfils homogenis i per tant no podem confirmar que hi hagi dependència entre ells.

Proseguim amb l'anàlisi dels perfils marginals per columna:

```
prop.table(table(df$f.education,df$f.hpw),2)
```

```
##
##                                     f.hpw[10-20] f.hpw[20-30]
##   f.education-Non-Graduate             0.32124352   0.22543353
##   f.education-Some-college             0.30051813   0.34104046
##   f.education-University-Or-More       0.33678756   0.37861272
##   f.education-Assoc_AND_Proof-school   0.04145078   0.05491329
##
##                                     f.hpw[30-40] f.hpw[40-50]
##   f.education-Non-Graduate             0.16487455   0.12522046
##   f.education-Some-college             0.20967742   0.21093474
##   f.education-University-Or-More       0.53763441   0.56754850
##   f.education-Assoc_AND_Proof-school   0.08781362   0.09629630
##
##                                     f.hpw[50-60]
##   f.education-Non-Graduate             0.06341463
##   f.education-Some-college             0.18658537
##   f.education-University-Or-More       0.63292683
##   f.education-Assoc_AND_Proof-school   0.11707317
```

Tot i que també podem veure com hi aparèixen algunes relacions, per a assegurar-nos el que farem es realitzar un test de la chi-quadrat en base a la hipòtesi nula H0 definida com que les files i les columnes no son dependents.

```
chisq.test(table(df$f.education,df$f.hpw))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(df$f.education, df$f.hpw)
## X-squared = 204.63, df = 12, p-value < 2.2e-16
```

Veiem que el p-value és menor a 2.2e-16, és a dir, és un valor tant ínfim que ens permet rebutjar la hipòtesi nula.

Aquest seria l'anàlisi per via descriptiva, a continuació hauria d'anar l'anàlisi de CA però per motius que desconeixem no ens és possible executar aquest anàlis, per tant, ens quedarem amb els resultats obtinguts mitjançant l'estudi descriptiu.

```
#res.ca1<-CA(table(df[,c("f.education","f.hpw")]))
```

A continuació, repetirem aquest anàlisi pero tractant de trobar relació amb la variable "f.marital".

```
prop.table(table(df$f.marital,df$f.hpw),1)
```

```
##
##                            f.hpw[10-20] f.hpw[20-30] f.hpw[30-40]
##   f.marital-Married           0.01671939   0.04383190   0.08178943
##   f.marital-No- Married       0.02375297   0.05344418   0.14845606
##   f.marital-Never-married     0.07198444   0.11543450   0.14526589
##   f.marital-Widowed           0.16129032   0.16774194   0.18064516
##
##                            f.hpw[40-50] f.hpw[50-60]
##   f.marital-Married           0.62720289   0.23045639
##   f.marital-No- Married       0.62470309   0.14964371
##   f.marital-Never-married     0.55512322   0.11219196
##   f.marital-Widowed           0.41935484   0.07096774
```

```
prop.table(table(df$f.marital,df$f.hpw),2)
```

```
##
##                            f.hpw[10-20] f.hpw[20-30] f.hpw[30-40]
##   f.marital-Married           0.19170984   0.28034682   0.32437276
##   f.marital-No- Married       0.10362694   0.13005780   0.22401434
##   f.marital-Never-married     0.57512953   0.51445087   0.40143369
##   f.marital-Widowed           0.12953368   0.07514451   0.05017921
##
##                            f.hpw[40-50] f.hpw[50-60]
##   f.marital-Married           0.48959436   0.62195122
##   f.marital-No- Married       0.18553792   0.15365854
##   f.marital-Never-married     0.30194004   0.21097561
##   f.marital-Widowed           0.02292769   0.01341463
```

Novament, podem veure algunes relacions pero res rellevant que ens faci pensar que existeix una relació. En tot cas, realitzem el test de la chi-quadrat i ho comprovem:

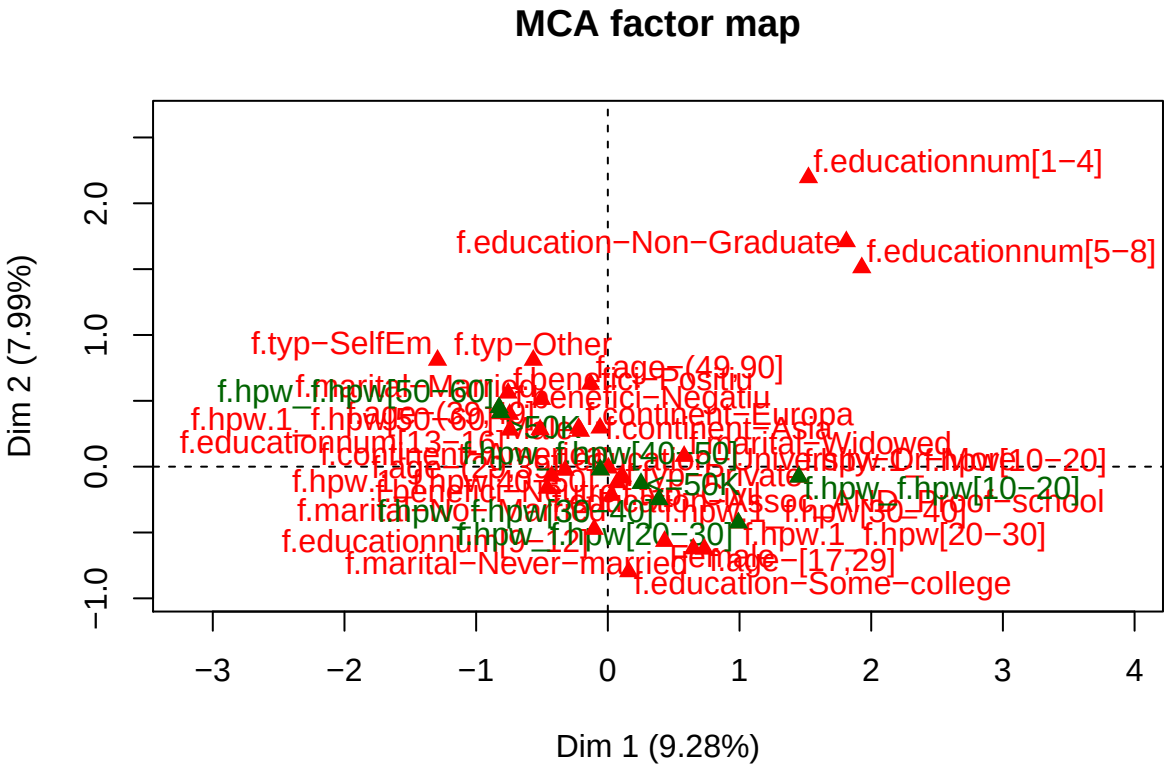
```
chisq.test(table(df$f.marital,df$f.hpw))
```

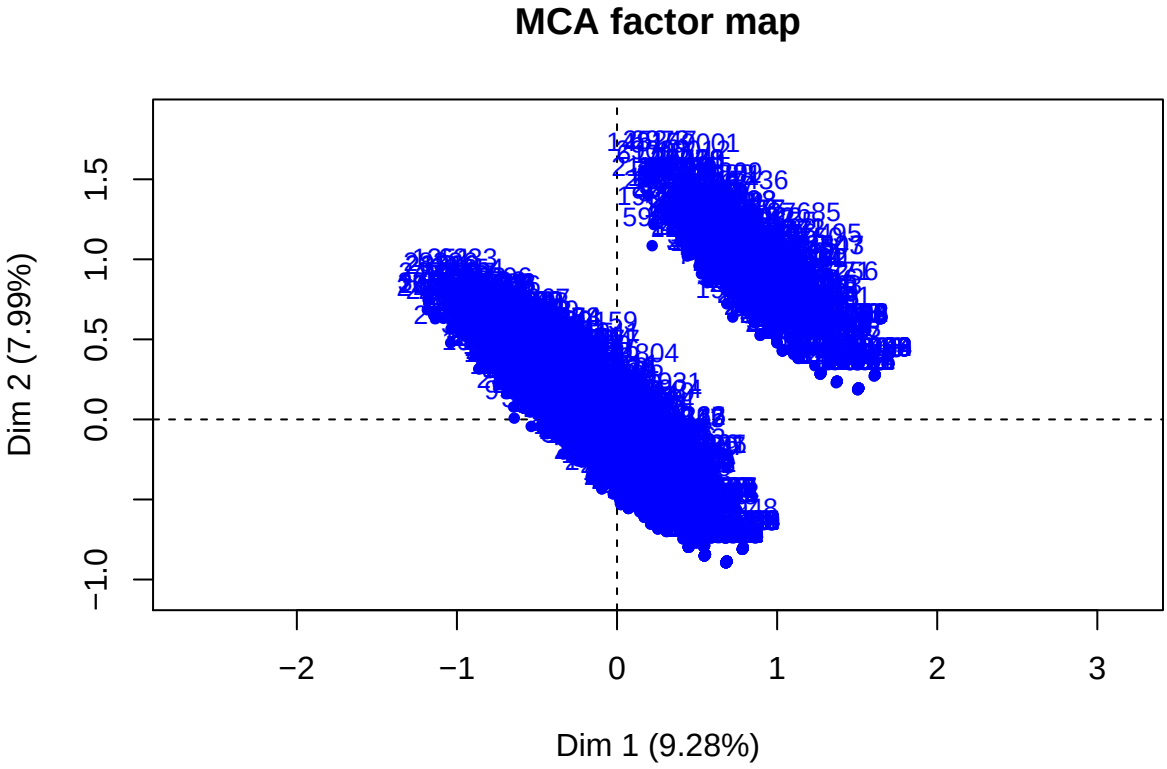
```
##
##  Pearson's Chi-squared test
##
## data:  table(df$f.marital, df$f.hpw)
## X-squared = 368.34, df = 12, p-value < 2.2e-16
```

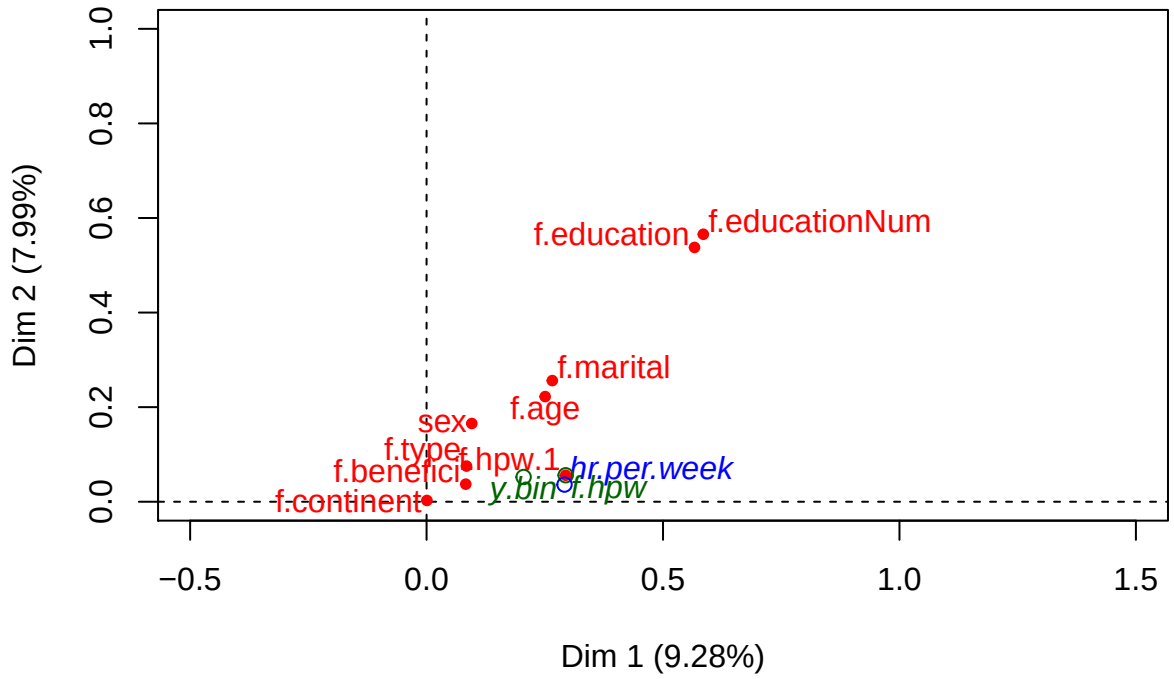
Efectivament, el p-value és ínfim i novament la hipòtesi nula queda rebutjada.

8 MCA

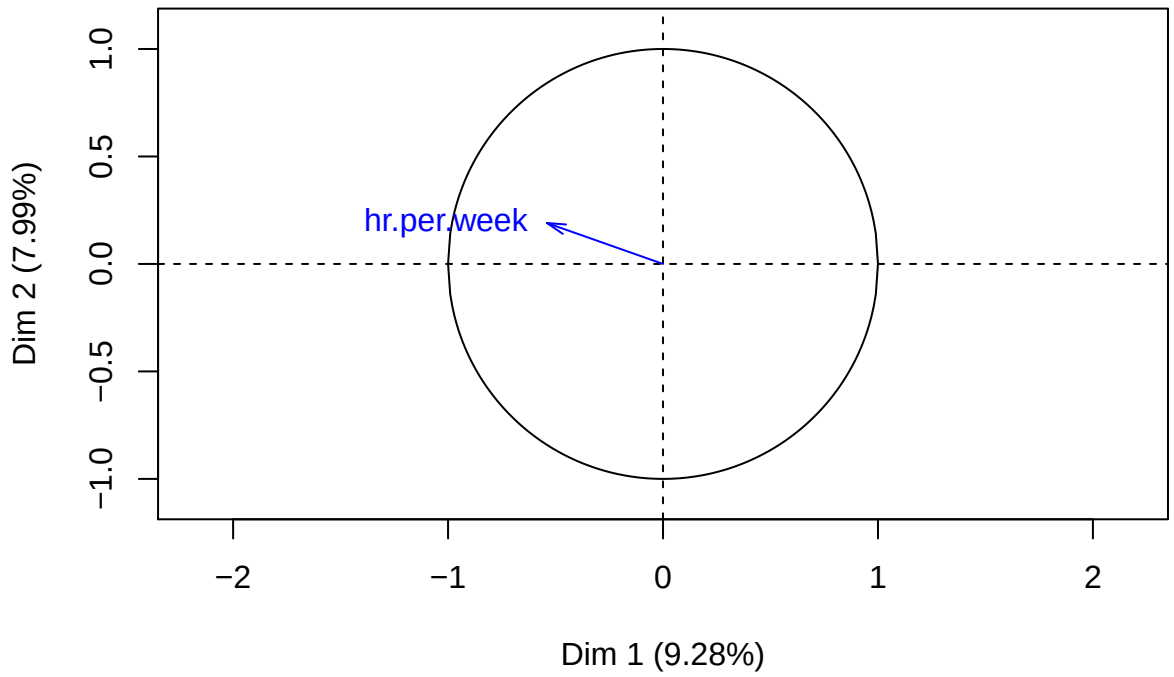
```
res.mca1<-MCA(df[,c(13,22,15:23,10)],quanti.sup=1,quali.sup = 2:3)
```







Supplementary variables on the MCA factor map

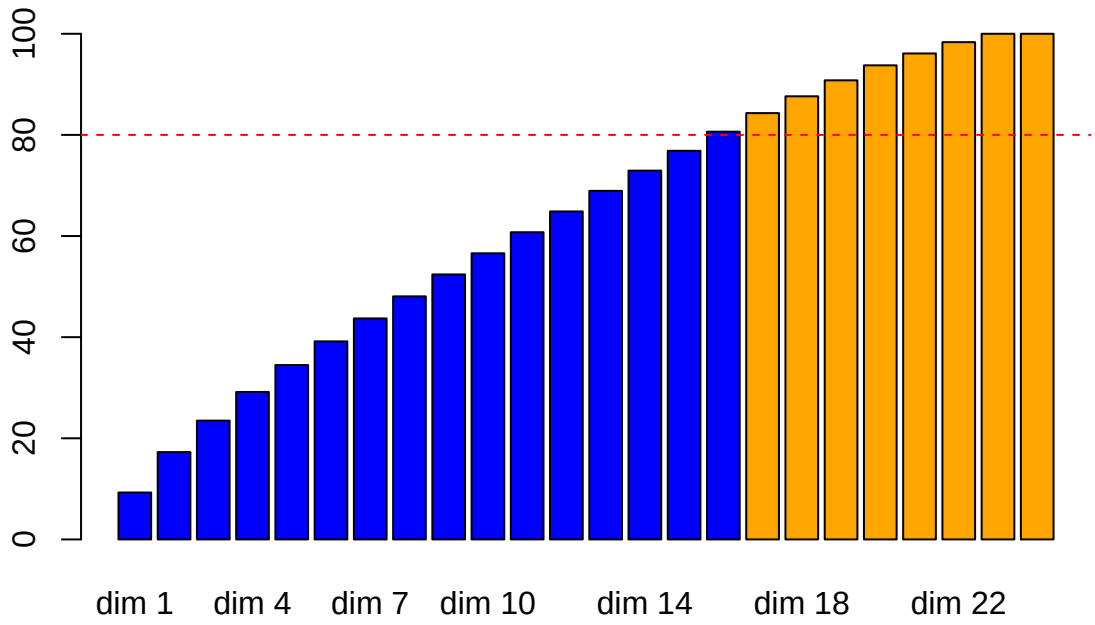



```
res.mca1$eig
```

```
##          eigenvalue percentage of variance
## dim 1  2.473667e-01           9.276251e+00
```

```
## dim 2  2.131668e-01           7.993755e+00
## dim 3  1.659487e-01           6.223077e+00
## dim 4  1.511913e-01           5.669673e+00
## dim 5  1.418411e-01           5.319040e+00
## dim 6  1.249312e-01           4.684919e+00
## dim 7  1.209554e-01           4.535828e+00
## dim 8  1.163890e-01           4.364589e+00
## dim 9  1.155793e-01           4.334223e+00
## dim 10 1.115897e-01           4.184614e+00
## dim 11 1.110344e-01           4.163788e+00
## dim 12 1.098109e-01           4.117907e+00
## dim 13 1.083024e-01           4.061339e+00
## dim 14 1.068169e-01           4.005634e+00
## dim 15 1.045242e-01           3.919659e+00
## dim 16 1.008600e-01           3.782250e+00
## dim 17 9.789245e-02           3.670967e+00
## dim 18 8.861292e-02           3.322984e+00
## dim 19 8.422891e-02           3.158584e+00
## dim 20 7.905225e-02           2.964459e+00
## dim 21 6.270639e-02           2.351490e+00
## dim 22 5.971316e-02           2.239243e+00
## dim 23 4.415271e-02           1.655727e+00
## dim 24 1.915193e-28           7.181973e-27
##        cumulative percentage of variance
## dim 1                              9.276251
## dim 2                             17.270007
## dim 3                             23.493084
## dim 4                             29.162756
## dim 5                             34.481796
## dim 6                             39.166715
## dim 7                             43.702543
## dim 8                             48.067132
## dim 9                             52.401355
## dim 10                            56.585968
## dim 11                            60.749757
## dim 12                            64.867664
## dim 13                            68.929003
## dim 14                            72.934637
## dim 15                            76.854296
## dim 16                            80.636546
## dim 17                            84.307513
## dim 18                            87.630497
## dim 19                            90.789081
## dim 20                            93.753540
## dim 21                            96.105030
## dim 22                            98.344273
## dim 23                           100.000000
## dim 24                           100.000000
```

```r
colors<-c("Blue","Orange")
barplot(res.mca1$eig[,3], col= colors[ifelse(res.mca1$eig[,3] < 81, 1, 2)])
abline(h=80,col="red",lty=2)
```




```r
summary(res.mca1,nbind=0,nbelements = 5, ncp=12 )
```

```
##
## Call:
## MCA(X = df[, c(13, 22, 15:23, 10)], quanti.sup = 1, quali.sup = 2:3)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
## Variance               0.247   0.213   0.166   0.151   0.142   0.125
## % of var.              9.276   7.994   6.223   5.670   5.319   4.685
## Cumulative % of var.   9.276  17.270  23.493  29.163  34.482  39.167
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.121   0.116   0.116   0.112   0.111   0.110
## % of var.              4.536   4.365   4.334   4.185   4.164   4.118
## Cumulative % of var.  43.703  48.067  52.401  56.586  60.750  64.868
##                       Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.108   0.107   0.105   0.101   0.098   0.089
## % of var.              4.061   4.006   3.920   3.782   3.671   3.323
## Cumulative % of var.  68.929  72.935  76.854  80.637  84.308  87.630
##                       Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24
## Variance               0.084   0.079   0.063   0.060   0.044   0.000
```

```
## % of var.              3.159   2.964   2.351   2.239   1.656   0.000
## Cumulative % of var.  90.789  93.754  96.105  98.344 100.000 100.000
##
## Categories (the 5 first)
##                         Dim.1     ctr    cos2  v.test     Dim.2     ctr
## f.typ-Civil         |   0.072   0.045   0.001   2.418 |  -0.121   0.147
## f.typ-Private       |   0.097   0.296   0.022  10.276 |  -0.085   0.267
## f.typ-SelfEm        |  -1.293   2.435   0.056 -16.314 |   0.809   1.105
## f.typ-Other         |  -0.566   1.018   0.024 -10.764 |   0.809   2.410
## f.marital-Married   |  -0.504   5.304   0.221 -32.404 |   0.511   6.342
##                        cos2  v.test     Dim.3     ctr    cos2  v.test
## f.typ-Civil           0.003  -4.075 |   0.484   3.028   0.056  16.321 |
## f.typ-Private         0.017  -9.061 |  -0.184   1.597   0.080 -19.550 |
## f.typ-SelfEm          0.022  10.202 |   0.135   0.039   0.001   1.701 |
## f.typ-Other           0.050  15.375 |   0.448   0.949   0.015   8.512 |
## f.marital-Married     0.228  32.892 |  -0.097   0.291   0.008  -6.222 |
##                       Dim.4     ctr    cos2  v.test     Dim.5     ctr
## f.typ-Civil           0.611   5.300   0.089  20.610 |  -0.045   0.031
## f.typ-Private        -0.157   1.277   0.059 -16.688 |  -0.131   0.945
## f.typ-SelfEm          0.178   0.075   0.001   2.243 |   1.066   2.884
## f.typ-Other          -0.188   0.183   0.003  -3.572 |   0.939   4.882
## f.marital-Married    -0.279   2.664   0.068 -17.952 |   0.222   1.791
##                        cos2  v.test
## f.typ-Civil           0.000  -1.533 |
## f.typ-Private         0.041 -13.906 |
## f.typ-SelfEm          0.038  13.445 |
## f.typ-Other           0.067  17.850 |
## f.marital-Married     0.043  14.260 |
##
## Categorical variables (eta2)
##                       Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## f.type              | 0.084 0.075 0.084 0.093 0.112 |
## f.marital           | 0.266 0.256 0.622 0.136 0.436 |
## f.education         | 0.567 0.538 0.033 0.390 0.078 |
## f.continent         | 0.001 0.003 0.009 0.023 0.011 |
## f.benefici          | 0.083 0.037 0.006 0.017 0.027 |
##
## Supplementary categories (the 5 first)
##                         Dim.1    cos2  v.test     Dim.2    cos2  v.test
## f.hpw_f.hpw[10-20]  |   1.450   0.089  20.558 |  -0.078   0.000  -1.111 |
## f.hpw_f.hpw[20-30]  |   0.991   0.077  19.138 |  -0.427   0.014  -8.242 |
## f.hpw_f.hpw[30-40]  |   0.388   0.020   9.758 |  -0.250   0.008  -6.279 |
## f.hpw_f.hpw[40-50]  |  -0.057   0.005  -4.791 |  -0.023   0.001  -1.963 |
## f.hpw_f.hpw[50-60]  |  -0.826   0.142 -25.992 |   0.449   0.042  14.145 |
##                       Dim.3    cos2  v.test     Dim.4    cos2  v.test
## f.hpw_f.hpw[10-20]    0.751   0.024  10.652 |   0.968   0.040  13.727 |
## f.hpw_f.hpw[20-30]    0.372   0.011   7.178 |   0.173   0.002   3.350 |
## f.hpw_f.hpw[30-40]    0.350   0.016   8.789 |   0.407   0.022  10.221 |
## f.hpw_f.hpw[40-50]   -0.096   0.014  -8.019 |  -0.200   0.059 -16.772 |
```

```
## f.hpw_f.hpw[50-60]  -0.241   0.012  -7.576 |   0.114   0.003   3.593 |
##                      Dim.5    cos2  v.test
## f.hpw_f.hpw[10-20]   1.344   0.076  19.054 |
## f.hpw_f.hpw[20-30]   0.893   0.063  17.255 |
## f.hpw_f.hpw[30-40]  -0.437   0.025 -10.991 |
## f.hpw_f.hpw[40-50]  -0.256   0.097 -21.467 |
## f.hpw_f.hpw[50-60]   0.490   0.050  15.416 |
##
## Supplementary categorical variables (eta2)
##                      Dim.1 Dim.2 Dim.3 Dim.4 Dim.5
## f.hpw              | 0.294 0.056 0.063 0.086 0.234 |
## y.bin              | 0.205 0.052 0.001 0.004 0.007 |
##
## Supplementary continuous variable
##                       Dim.1    Dim.2    Dim.3    Dim.4    Dim.5
## hr.per.week         | -0.540 |  0.190 | -0.220 | -0.149 | -0.118 |
```

```
dimdesc(res.mca1,prob=0.01)
```

```
## $`Dim 1`
## $`Dim 1`$quanti
##             correlation p.value
## hr.per.week  -0.5400644       0
##
## $`Dim 1`$quali
##                          R2       p.value
## f.hpw           0.29412178  0.000000e+00
## f.education     0.56666911  0.000000e+00
## f.hpw.1         0.29412178  0.000000e+00
## f.educationNum 0.58518482  0.000000e+00
## f.marital       0.26594042 5.000745e-318
## f.age           0.25075951 6.179964e-297
## y.bin           0.20516323 3.692360e-239
## sex             0.09560636 8.346913e-106
## f.type          0.08445217  1.719246e-90
## f.benefici      0.08287326  6.144800e-90
##
## $`Dim 1`$category
##                                                    Estimate       p.value
## f.educationNum=f.educationnum[5-8]            0.6351455926  0.000000e+00
## f.education=f.education-Non-Graduate          0.7660636523  0.000000e+00
## f.age=f.age-[17,29]                           0.3940074197 1.434139e-278
## y.bin=<=50K                                   0.2652265970 3.692360e-239
## f.marital=f.marital-Never-married             0.2292959985 1.126177e-235
## sex=Female                                    0.1623367806 8.346913e-106
## f.educationNum=f.educationnum[1-4]            0.4337173284 5.688190e-104
## f.hpw.1=f.hpw.1_f.hpw[10-20]                  0.5274386600  3.101053e-98
## f.hpw=f.hpw_f.hpw[10-20]                      0.5274386600  3.101053e-98
## f.benefici=f.benefici-Neutre                  0.2856672407  2.386940e-91
## f.hpw.1=f.hpw.1_f.hpw[20-30]                  0.2992570455  7.384299e-85
```

```
## f.hpw=f.hpw_f.hpw[20-30]                            0.2992570455  7.384299e-85
## f.type=f.typ-Private                                 0.2583972187  5.065487e-25
## f.marital=f.marital-Widowed                          0.1944965067  1.963997e-13
## f.age=f.age-(49,90]                                 -0.0357071360  1.674553e-06
## f.hpw.1=f.hpw.1_f.hpw[40-50]                        -0.2219659496  1.614442e-06
## f.hpw=f.hpw_f.hpw[40-50]                            -0.2219659496  1.614442e-06
## f.education=f.education-Some-college                -0.0558173683  6.262613e-09
## f.educationNum=f.educationnum[9-12]                 -0.3754062275  2.325308e-20
## f.hpw.1=f.hpw.1_f.hpw[30-40]                        -0.0005038467  1.063881e-22
## f.hpw=f.hpw_f.hpw[30-40]                            -0.0005038467  1.063881e-22
## f.education=f.education-Assoc_AND_Proof-school      -0.3636464822  1.135504e-24
## f.type=f.typ-Other                                  -0.0713253622  2.503774e-27
## f.benefici=f.benefici-Negatiu                       -0.1398792715  3.385367e-31
## f.age=f.age-(29,39]                                 -0.1318051240  6.692046e-42
## f.benefici=f.benefici-Positiu                       -0.1457879691  1.837004e-54
## f.type=f.typ-SelfEm                                 -0.4329721623  1.674415e-61
## f.age=f.age-(39,49]                                 -0.2264951596  2.110643e-77
## sex=Male                                            -0.1623367806 8.346913e-106
## f.hpw.1=f.hpw.1_f.hpw[50-60]                        -0.6042259093 1.918673e-160
## f.hpw=f.hpw_f.hpw[50-60]                            -0.6042259093 1.918673e-160
## f.educationNum=f.educationnum[13-16]                -0.6934566936 9.552551e-201
## y.bin=>50K                                          -0.2652265970 3.692360e-239
## f.marital=f.marital-Married                         -0.3440501409 6.058695e-260
## f.education=f.education-University-Or-More          -0.3465998018 1.045414e-263
##
##
## $`Dim 2`
## $`Dim 2`$quanti
##              correlation      p.value
## hr.per.week   0.1904179 4.867078e-40
##
## $`Dim 2`$quali
##                          R2       p.value
## f.education    0.53779302  0.000000e+00
## f.educationNum 0.56556399  0.000000e+00
## f.marital      0.25618147 2.029753e-304
## f.age          0.22231777 1.508136e-258
## sex            0.16542881 8.498594e-189
## f.type         0.07538323  2.361836e-80
## f.hpw          0.05600235  5.251911e-58
## f.hpw.1        0.05600235  5.251911e-58
## y.bin          0.05228491  2.053554e-57
## f.benefici     0.03713626  9.433883e-40
## f.continent    0.00269435  1.651008e-03
##
## $`Dim 2`$category
##                                                    Estimate       p.value
## f.education=f.education-Non-Graduate            0.71096349  0.000000e+00
## f.educationNum=f.educationnum[5-8]              0.29260966 4.083400e-287
```

```
## f.marital=f.marital-Married                          0.26483162 7.036748e-269
## f.educationNum=f.educationnum[1-4]                   0.60850679 2.311503e-226
## sex=Male                                             0.19822950 8.498594e-189
## f.age=f.age-(49,90]                                  0.26007840 1.363331e-123
## y.bin=>50K                                           0.12429243  2.053554e-57
## f.type=f.typ-Other                                   0.21046029  1.171849e-54
## f.hpw.1=f.hpw.1_f.hpw[50-60]                         0.23783928  2.340611e-46
## f.hpw=f.hpw_f.hpw[50-60]                             0.23783928  2.340611e-46
## f.benefici=f.benefici-Positiu                        0.12227287  3.484522e-30
## f.type=f.typ-SelfEm                                  0.21051016  1.095925e-24
## f.age=f.age-(39,49]                                  0.10022317  8.441479e-24
## f.benefici=f.benefici-Negatiu                        0.04749246  7.056934e-10
## f.continent=f.continent-Asia                         0.05055267  5.611181e-03
## f.continent=f.continent-America                     -0.08964683  3.483425e-04
## f.education=f.education-Assoc_AND_Proof-school      -0.15346333  3.040568e-04
## f.type=f.typ-Civil                                  -0.21865681  4.541055e-05
## f.education=f.education-University-Or-More          -0.11124402  4.260895e-08
## f.hpw.1=f.hpw.1_f.hpw[30-40]                        -0.08494128  3.158323e-10
## f.hpw=f.hpw_f.hpw[30-40]                            -0.08494128  3.158323e-10
## f.marital=f.marital-No- Married                     -0.07132371  3.511322e-12
## f.hpw.1=f.hpw.1_f.hpw[20-30]                        -0.16663597  1.340925e-16
## f.hpw=f.hpw_f.hpw[20-30]                            -0.16663597  1.340925e-16
## f.type=f.typ-Private                                -0.20231364  9.094128e-20
## f.educationNum=f.educationnum[13-16]                -0.27660319  1.594847e-27
## f.benefici=f.benefici-Neutre                        -0.16976533  3.873396e-40
## y.bin=<=50K                                         -0.12429243  2.053554e-57
## sex=Female                                          -0.19822950 8.498594e-189
## f.age=f.age-[17,29]                                 -0.31835282 1.826945e-195
## f.education=f.education-Some-college                -0.44625614 1.386409e-205
## f.marital=f.marital-Never-married                   -0.25882634 3.587646e-215
## f.educationNum=f.educationnum[9-12]                 -0.62451326  0.000000e+00
##
##
## $`Dim 3`
## $`Dim 3`$quanti
##              correlation      p.value
## hr.per.week  -0.2204362 2.185325e-53
##
## $`Dim 3`$quali
##                           R2       p.value
## f.marital       0.621983433  0.000000e+00
## f.age           0.393250737  0.000000e+00
## sex             0.195910918 3.270083e-227
## f.educationNum  0.087886644  2.339152e-94
## f.type          0.083836313  8.453842e-90
## f.hpw           0.062772012  2.242640e-65
## f.hpw.1         0.062772012  2.242640e-65
## f.education     0.033039706  2.304348e-34
## f.continent     0.009028227  4.441393e-10
```

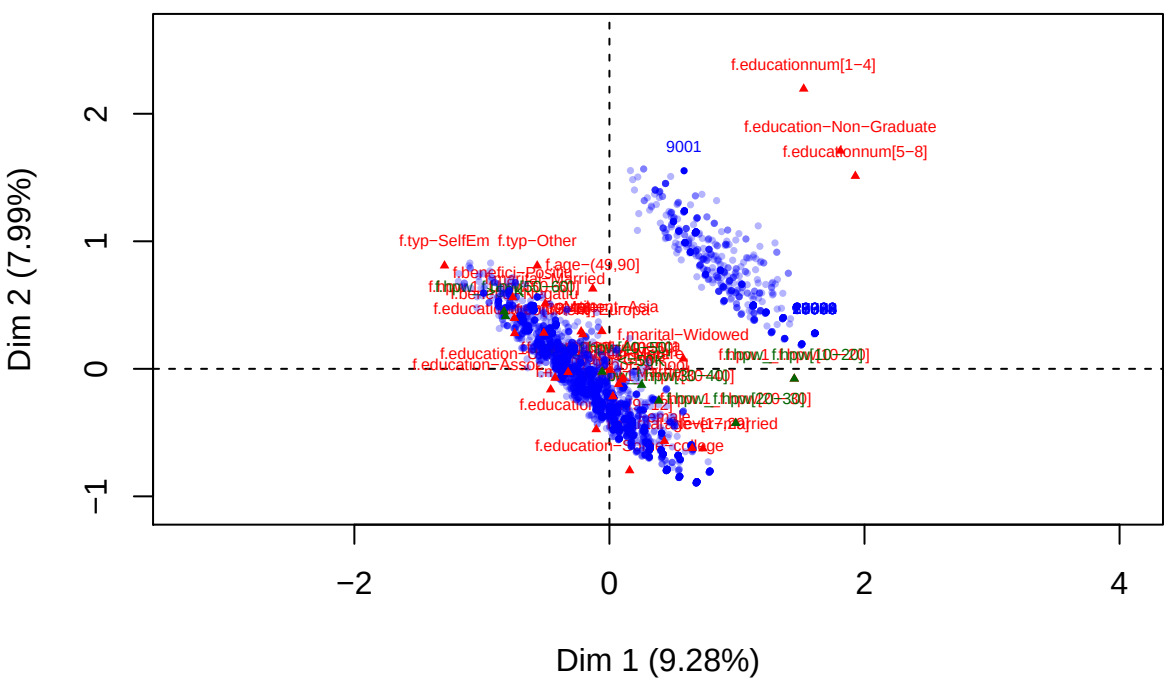
```
## f.benefici     0.005830482  9.328732e-07
##
## $`Dim 3`$category
##                                                  Estimate       p.value
## f.age=f.age-(49,90]                           0.402578631  0.000000e+00
## f.marital=f.marital-Widowed                   1.009453206  0.000000e+00
## sex=Female                                    0.190335232 3.270083e-227
## f.type=f.typ-Civil                            0.107173601  1.495334e-61
## f.educationNum=f.educationnum[9-12]           0.062575657  8.500403e-48
## f.education=f.education-Some-college          0.106228544  6.971252e-30
## f.hpw.1=f.hpw.1_f.hpw[10-20]                  0.213412395  8.727274e-27
## f.hpw=f.hpw_f.hpw[10-20]                      0.213412395  8.727274e-27
## f.educationNum=f.educationnum[1-4]            0.288326571  1.880551e-23
## f.hpw.1=f.hpw.1_f.hpw[30-40]                  0.049855408  1.111667e-18
## f.hpw=f.hpw_f.hpw[30-40]                      0.049855408  1.111667e-18
## f.type=f.typ-Other                            0.092551047  1.308426e-17
## f.hpw.1=f.hpw.1_f.hpw[20-30]                  0.058845491  6.167248e-13
## f.hpw=f.hpw_f.hpw[20-30]                      0.058845491  6.167248e-13
## f.benefici=f.benefici-Positiu                 0.085181434  1.731665e-04
## f.continent=f.continent-America               0.064059484  3.344322e-04
## f.education=f.education-Assoc_AND_Proof-school 0.041896907  8.496674e-04
## f.age=f.age-(39,49]                           0.007543772  9.325464e-04
## f.benefici=f.benefici-Negatiu                -0.093720199  8.421034e-05
## f.education=f.education-Non-Graduate         -0.084064812  1.595170e-05
## f.marital=f.marital-Married                  -0.386907762  4.556821e-10
## f.continent=f.continent-Asia                 -0.213502846  2.352920e-10
## f.hpw.1=f.hpw.1_f.hpw[50-60]                 -0.190594127  3.008086e-14
## f.hpw=f.hpw_f.hpw[50-60]                     -0.190594127  3.008086e-14
## f.hpw.1=f.hpw.1_f.hpw[40-50]                 -0.131519167  8.589029e-16
## f.hpw=f.hpw_f.hpw[40-50]                     -0.131519167  8.589029e-16
## f.age=f.age-(29,39]                          -0.110621743  2.431692e-16
## f.education=f.education-University-Or-More    -0.064060639  3.823645e-17
## f.educationNum=f.educationnum[5-8]           -0.213019963  3.364523e-31
## f.educationNum=f.educationnum[13-16]         -0.137882266  2.642902e-38
## f.type=f.typ-Private                         -0.164830223  1.263445e-88
## f.marital=f.marital-No- Married              -0.011872112 2.874291e-165
## f.age=f.age-[17,29]                          -0.299500660 1.387540e-220
## sex=Male                                     -0.190335232 3.270083e-227
## f.marital=f.marital-Never-married            -0.610673332 4.579902e-233
```

Per el MCA seguint el criteri estudiat a clase, hauriem d'agafar fins a 15 dimensions que són aquelles que donen una explicabilitat d'un 80% de les variables per a tenir un bon model.

També veiem que les hr.per.week esta negativament relaciont amb la dimensió 1.

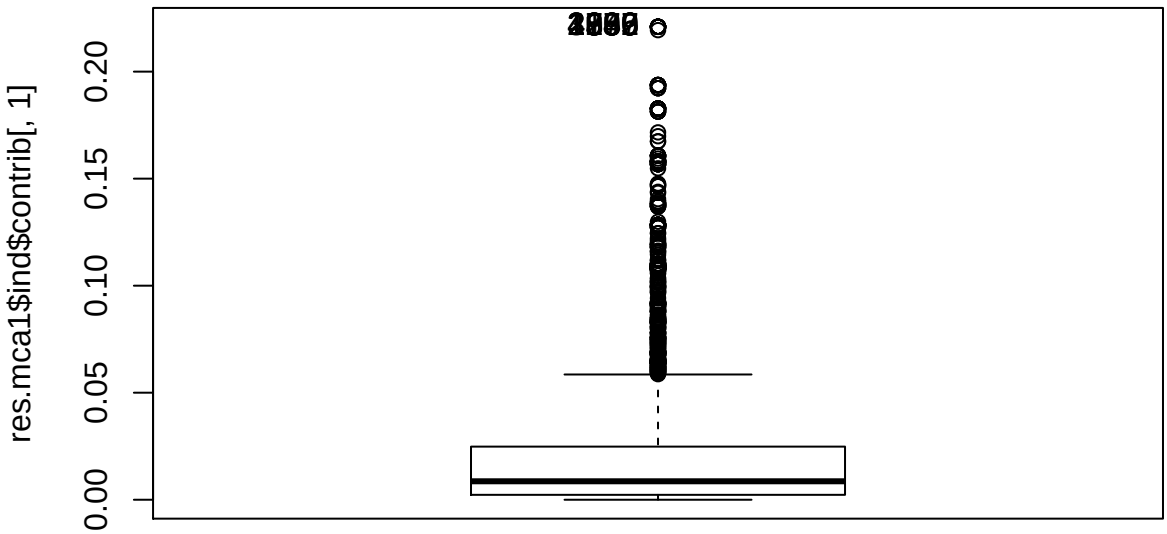
8.1 Individual point of view:

```
plot(res.mca1,choix="ind",select="contrib 10",cex=0.5)
```



No podem tenir una gran apreciació de les variables més contributibes així que farem un anàlisis per les diferents Dimensions a estudiar.

```
#Dim 1
Boxplot(res.mca1$ind$contrib[,1])
```




```
##  [1] 1266 1357 1879 2389 2595 2742 2900 3855 4086 4132
```

```
rang1<-order(res.mca1$ind$contrib[,1],decreasing = T); rang1[1:10]
```

```
##  [1] 1266 1357 1879 2389 2595 2742 2900 3855 4086 4132
```

```
rownames(df[rang1[1:10],])
```

```
##  [1] "8856"  "9395"  "12803" "16433" "17908" "19004" "20022" "26089"
##  [9] "27760" "28093"
```

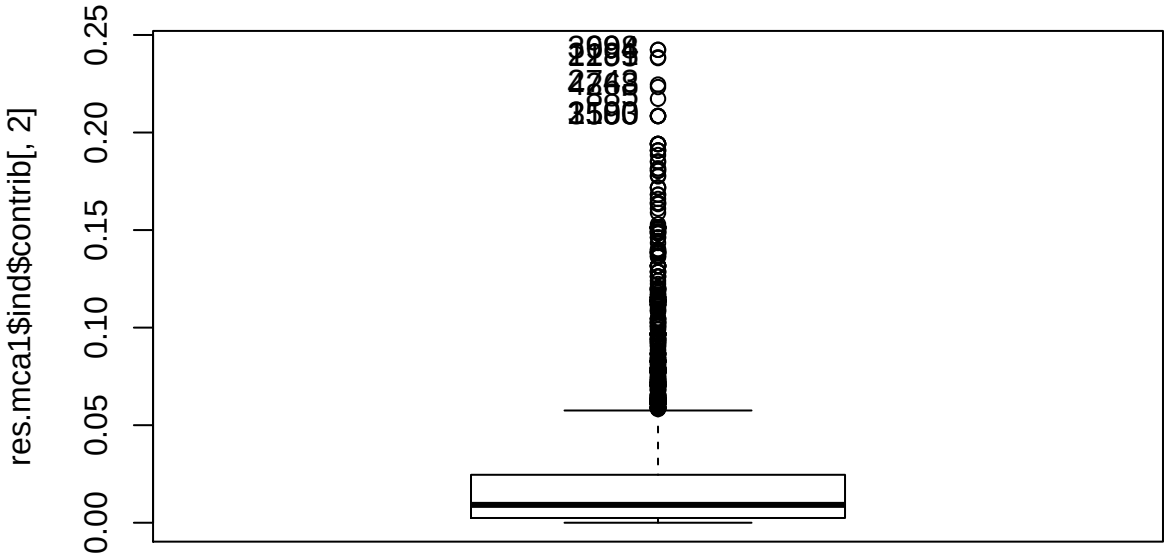
```
df[rang1[1:10],c(vars_con,vars_dis)]
```

```
##       fnlwgt education.num capital.gain capital.loss hr.per.week i.rank
## 8856  154908             6            0            0          10      0
## 9395  182042             7            0            0          19      0
## 12803 117549             6            0            0          12      0
## 16433 267965             7            0            0          15      0
## 17908 131180             7            0            0          16      0
## 19004 198830             7            0            0          10      0
## 20022 276540             8            0            0          15      0
## 26089 225507             7            0            0          15      0
## 27760 317702             6            0            0          15      0
## 28093  36877             6            0            0          10      0
##              occupation  race    sex y.bin        f.type
## 8856      Other-service White Female <=50K f.typ-Private
## 9395      Other-service White Female <=50K f.typ-Private
## 12803             Sales Black Female <=50K f.typ-Private
```

```
## 16433             Sales White Female <=50K f.typ-Private
## 17908    Prof-specialty White Female <=50K f.typ-Private
## 19004      Adm-clerical White Female <=50K f.typ-Private
## 20022             Sales Black Female <=50K f.typ-Private
## 26089 Handlers-cleaners Black Female <=50K f.typ-Private
## 27760             Sales Black Female <=50K f.typ-Private
## 28093             Sales White Female <=50K f.typ-Private
##                     f.marital              f.education         f.continent
## 8856  f.marital-Never-married f.education-Non-Graduate f.continent-America
## 9395  f.marital-Never-married f.education-Non-Graduate f.continent-America
## 12803 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 16433 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 17908 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 19004 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 20022 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 26089 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 27760 f.marital-Never-married f.education-Non-Graduate f.continent-America
## 28093 f.marital-Never-married f.education-Non-Graduate f.continent-America
##              f.benefici         f.age        f.hpw      f.educationNum
## 8856  f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 9395  f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 12803 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 16433 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 17908 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 19004 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 20022 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 26089 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 27760 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
## 28093 f.benefici-Neutre f.age-[17,29] f.hpw[10-20] f.educationnum[5-8]
```

En la primera dimensió veiem que tenim que els individus més contributius d'aquesta dimensió són bastant joves, i tendeixen a treballar poques hores a la setmana. També podem veure que són aquells que tenen meys anys dedicats a la educació, ja que no tenen tituls superiors o universitaris. També ens ha soptat veure que tots els individus de gran contribució en aquesta dimensió són dones.

```
#Dim 2
Boxplot(res.mca1$ind$contrib[,2])
```




```
##  [1]  998 3694 2181 1295 2743 4268  885 1593 2500 3180
```

```
rang2<-order(res.mca1$ind$contrib[,2],decreasing = T); rang1[1:10]
```

```
##  [1] 1266 1357 1879 2389 2595 2742 2900 3855 4086 4132
```

```
rownames(df[rang2[1:10],])
```

```
##  [1] "6973"  "25247" "14815" "9001"  "19012" "29093" "6171"  "10978"
##  [9] "17213" "21921"
```

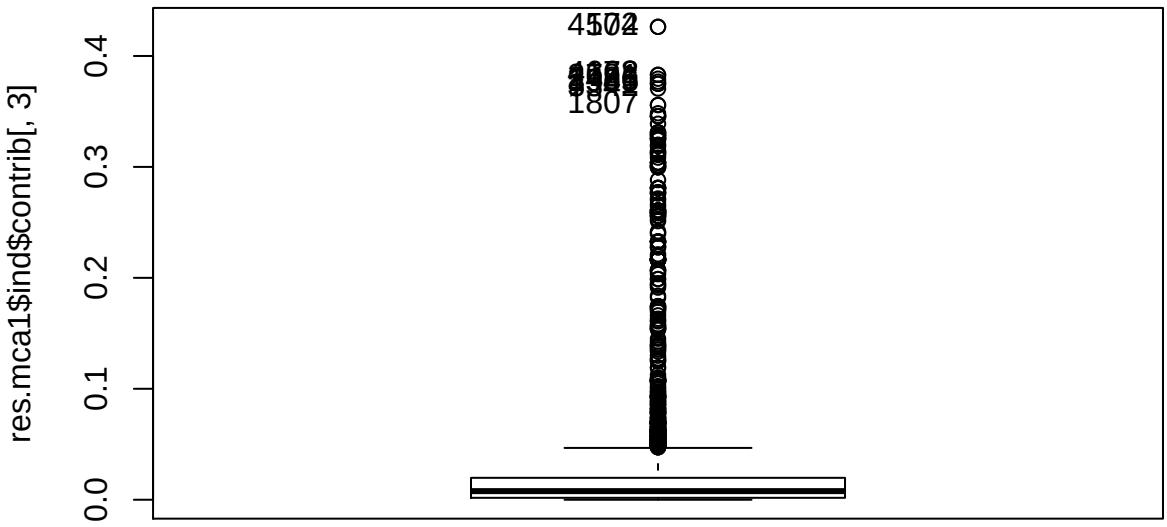
```
df[rang2[1:10],c(vars_con,vars_dis)]
```

```
##       fnlwgt education.num capital.gain capital.loss hr.per.week i.rank
## 6973  104003             4            0            0          55      0
## 25247  36327             4            0            0          50      0
## 14815 141409             6         7688            0          50      0
## 9001  266707             2            0         2179          18      0
## 19012 131417             3         1797            0          21      0
## 29093  75577             4         2580            0          50      0
## 6171  242184             4            0            0          55      0
## 10978 167380             4            0            0          40      0
## 17213 236470             4            0            0          40      0
## 21921  33487             4            0            0          40      0
##               occupation  race  sex y.bin        f.type         f.marital
## 6973               Sales White Male <=50K   f.typ-Other f.marital-Married
## 25247 Machine-op-inspct White Male <=50K   f.typ-Other f.marital-Married
## 14815              Sales White Male  >50K   f.typ-Other f.marital-Married
```

```
## 9001   Transport-moving White Male <=50K   f.typ-Other f.marital-Married
## 19012   Farming-fishing White Male <=50K   f.typ-Other f.marital-Married
## 29093      Adm-clerical White Male <=50K f.typ-Private f.marital-Married
## 6171    Exec-managerial White Male  >50K   f.typ-Other f.marital-Married
## 10978   Farming-fishing White Male <=50K   f.typ-Other f.marital-Married
## 17213   Farming-fishing White Male <=50K   f.typ-Other f.marital-Married
## 21921      Craft-repair White Male <=50K   f.typ-Other f.marital-Married
##                    f.education         f.continent         f.benefici
## 6973  f.education-Non-Graduate f.continent-America  f.benefici-Neutre
## 25247 f.education-Non-Graduate f.continent-America  f.benefici-Neutre
## 14815 f.education-Non-Graduate f.continent-America f.benefici-Positiu
## 9001  f.education-Non-Graduate f.continent-America f.benefici-Negatiu
## 19012 f.education-Non-Graduate f.continent-America f.benefici-Positiu
## 29093 f.education-Non-Graduate f.continent-America f.benefici-Positiu
## 6171  f.education-Non-Graduate f.continent-America  f.benefici-Neutre
## 10978 f.education-Non-Graduate f.continent-America  f.benefici-Neutre
## 17213 f.education-Non-Graduate f.continent-America  f.benefici-Neutre
## 21921 f.education-Non-Graduate f.continent-America  f.benefici-Neutre
##                f.age        f.hpw      f.educationNum
## 6973  f.age-(49,90] f.hpw[50-60] f.educationnum[1-4]
## 25247 f.age-(49,90] f.hpw[50-60] f.educationnum[1-4]
## 14815 f.age-(49,90] f.hpw[50-60] f.educationnum[5-8]
## 9001  f.age-(49,90] f.hpw[10-20] f.educationnum[1-4]
## 19012 f.age-(49,90] f.hpw[20-30] f.educationnum[1-4]
## 29093 f.age-(49,90] f.hpw[50-60] f.educationnum[1-4]
## 6171  f.age-(39,49] f.hpw[50-60] f.educationnum[1-4]
## 10978 f.age-(49,90] f.hpw[40-50] f.educationnum[1-4]
## 17213 f.age-(49,90] f.hpw[40-50] f.educationnum[1-4]
## 21921 f.age-(49,90] f.hpw[40-50] f.educationnum[1-4]
```

En Aquesta dimensió els que destaquen a major escala són homes blancs d'edat més gran, que estan casats, que han invertit pocs anys en la seva educació i que en conseqüéncia no tenen una titulació superior.

```
#Dim 3
Boxplot(res.mca1$ind$contrib[,3])
```




```
##  [1]  504 4172  168 4672 2781 2486 1986 3545 3342 1807
```

```
rang3<-order(res.mca1$ind$contrib[,3],decreasing = T); rang1[1:10]
```

```
##  [1] 1266 1357 1879 2389 2595 2742 2900 3855 4086 4132
```

```
rownames(df[rang3[1:10],])
```

```
##  [1] "3589"  "28477" "1292"  "31942" "19202" "17109" "13513" "24247"
##  [9] "22931" "12307"
```

```
df[rang3[1:10],c(vars_con,vars_dis)]
```

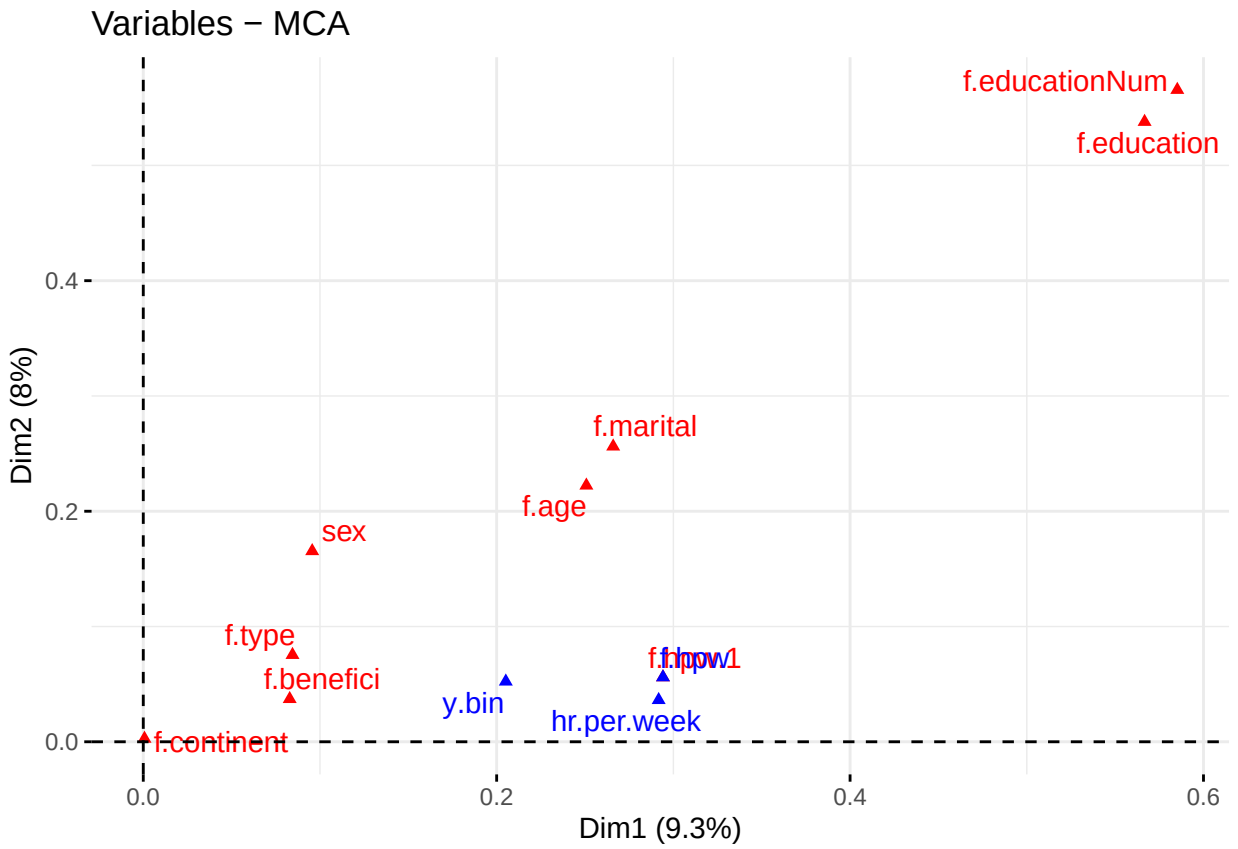
```
##       fnlwgt education.num capital.gain capital.loss hr.per.week i.rank
## 3589  104661            10            0            0          12      2
## 28477 116165            10            0            0          14      2
## 1292  795830             2            0            0          30      0
## 31942  26857             4            0            0          35      0
## 19202 255927            10            0            0          24      0
## 17109 180869            10            0            0          35      0
## 13513 197218             9            0            0          18      0
## 24247 334666             9            0            0          12      0
## 22931 392886             9            0            0          14      0
## 12307 124686             4            0            0          10      0
##              occupation  race    sex y.bin        f.type         f.marital
## 3589       Adm-clerical White Female <=50K   f.typ-Civil f.marital-Widowed
## 28477      Adm-clerical White Female <=50K   f.typ-Civil f.marital-Widowed
## 1292      Other-service White Female <=50K   f.typ-Other f.marital-Widowed
```

```
## 31942   Farming-fishing White Female <=50K   f.typ-Other f.marital-Widowed
## 19202      Adm-clerical White Female <=50K   f.typ-Civil f.marital-Widowed
## 17109      Adm-clerical White Female <=50K   f.typ-Civil f.marital-Widowed
## 13513     Other-service White Female <=50K   f.typ-Civil f.marital-Widowed
## 24247      Adm-clerical White Female <=50K   f.typ-Civil f.marital-Widowed
## 22931   Farming-fishing White Female <=50K   f.typ-Other f.marital-Widowed
## 12307 Machine-op-inspct White Female <=50K f.typ-Private f.marital-Widowed
##                          f.education         f.continent        f.benefici
## 3589        f.education-Some-college f.continent-America f.benefici-Neutre
## 28477       f.education-Some-college f.continent-America f.benefici-Neutre
## 1292        f.education-Non-Graduate f.continent-America f.benefici-Neutre
## 31942       f.education-Non-Graduate f.continent-America f.benefici-Neutre
## 19202       f.education-Some-college f.continent-America f.benefici-Neutre
## 17109       f.education-Some-college f.continent-America f.benefici-Neutre
## 13513 f.education-University-Or-More f.continent-America f.benefici-Neutre
## 24247 f.education-University-Or-More f.continent-America f.benefici-Neutre
## 22931 f.education-University-Or-More f.continent-America f.benefici-Neutre
## 12307       f.education-Non-Graduate f.continent-America f.benefici-Neutre
##                f.age        f.hpw       f.educationNum
## 3589  f.age-(49,90] f.hpw[10-20] f.educationnum[9-12]
## 28477 f.age-(49,90] f.hpw[10-20] f.educationnum[9-12]
## 1292  f.age-(49,90] f.hpw[30-40]  f.educationnum[1-4]
## 31942 f.age-(49,90] f.hpw[30-40]  f.educationnum[1-4]
## 19202 f.age-(49,90] f.hpw[20-30] f.educationnum[9-12]
## 17109 f.age-(49,90] f.hpw[30-40] f.educationnum[9-12]
## 13513 f.age-(49,90] f.hpw[10-20] f.educationnum[9-12]
## 24247 f.age-(49,90] f.hpw[10-20] f.educationnum[9-12]
## 22931 f.age-(49,90] f.hpw[10-20] f.educationnum[9-12]
## 12307 f.age-(49,90] f.hpw[10-20]  f.educationnum[1-4]
```

En aquesta dimensió no podem donar una gran explicabilitat de les variables. Ja que no tenen gaire relació entre elles.

Finalment podem conclure que els individus que contribueixen més a les dades són persones o bé molt joves, o bé persones d'abançada edat, que de retruc han invertit pocs anys en la seva educació.

```
fviz_mca_var(res.mca1, choice = "mca.cor",
             repel = TRUE, #
             ggtheme = theme_minimal())
```



Com podem veure tenim que les categories extranyes són la f.education, i la educationNum, es a dir que veiem que la educació es bastant contributiu alhora de decidir.

8.2 Interpreting the axes

```
out<-dimdesc(res.mca1)
out$`Dim 1`$quanti
```

```
##             correlation p.value
## hr.per.week  -0.5400644       0
```

```
out$`Dim 1`$quali
```

```
##                         R2       p.value
## f.hpw          0.29412178  0.000000e+00
## f.education    0.56666911  0.000000e+00
## f.hpw.1        0.29412178  0.000000e+00
## f.educationNum 0.58518482  0.000000e+00
## f.marital      0.26594042 5.000745e-318
## f.age          0.25075951 6.179964e-297
## y.bin          0.20516323 3.692360e-239
## sex            0.09560636 8.346913e-106
## f.type         0.08445217  1.719246e-90
## f.benefici     0.08287326  6.144800e-90
```

```
out$`Dim 2`$quanti
```

```
##             correlation      p.value
```

```
## hr.per.week   0.1904179 4.867078e-40
```

```
out$`Dim 2`$quali
```

```
##                         R2       p.value
## f.education    0.53779302  0.000000e+00
## f.educationNum 0.56556399  0.000000e+00
## f.marital      0.25618147 2.029753e-304
## f.age          0.22231777 1.508136e-258
## sex            0.16542881 8.498594e-189
## f.type         0.07538323  2.361836e-80
## f.hpw          0.05600235  5.251911e-58
## f.hpw.1        0.05600235  5.251911e-58
## y.bin          0.05228491  2.053554e-57
## f.benefici     0.03713626  9.433883e-40
## f.continent    0.00269435  1.651008e-03
```

```
out$`Dim 3`$quanti
```

```
##             correlation      p.value
## hr.per.week  -0.2204362 2.185325e-53
```

```
out$`Dim 3`$quali
```

```
##                         R2       p.value
## f.marital      0.621983433  0.000000e+00
## f.age          0.393250737  0.000000e+00
## sex            0.195910918 3.270083e-227
## f.educationNum 0.087886644  2.339152e-94
## f.type         0.083836313  8.453842e-90
## f.hpw          0.062772012  2.242640e-65
## f.hpw.1        0.062772012  2.242640e-65
## f.education    0.033039706  2.304348e-34
## f.continent    0.009028227  4.441393e-10
## f.benefici     0.005830482  9.328732e-07
## y.bin          0.001377660  1.050194e-02
```

Veiem que la dimensió 1 com hem dit anteirorment esta negativament relacionada amb les hr.per.week, també que amb els factors que te major relació són el education i educationNum, tot i que la edat i l'estat civil també semblen estar relacionats però no a tant gran escala.

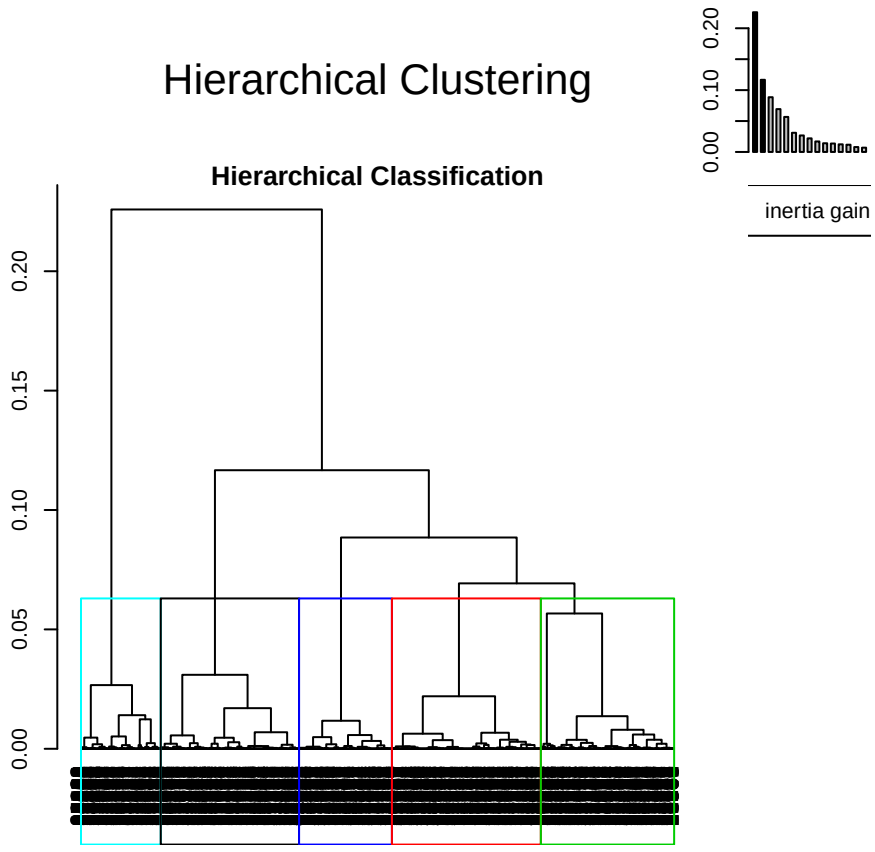
Per la dimensió 2 veiem que ten una certa relació positiva tot i que no molt significant amb la vairable hr.per.week. Un altre cop veiem que les variables més explicatives d'aquestes són els anys d'educació.

Finalment per la dimensió 3 que esta poc relacionada negativament amb el target, tenim que les variables més explicatives per ella, són l'estat civil i l'edat dels individus.

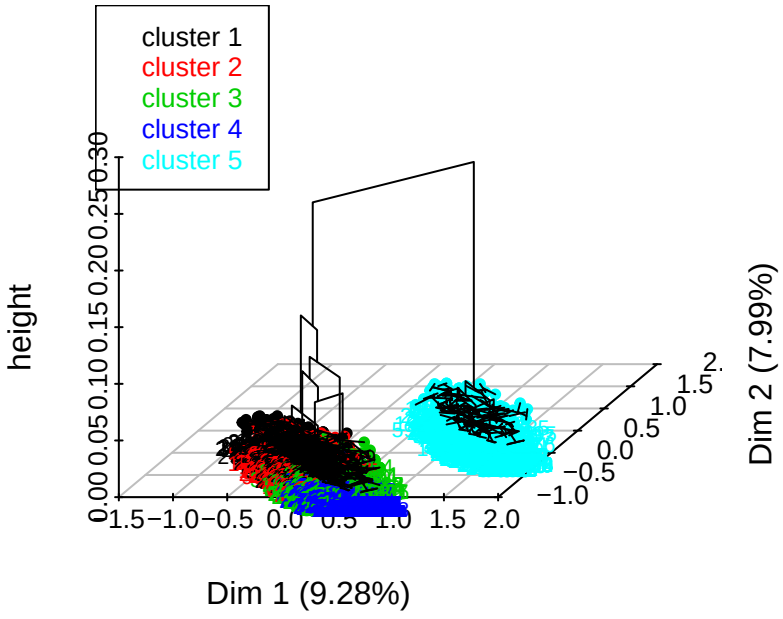
9 Hierarchical Clustering (MCA)

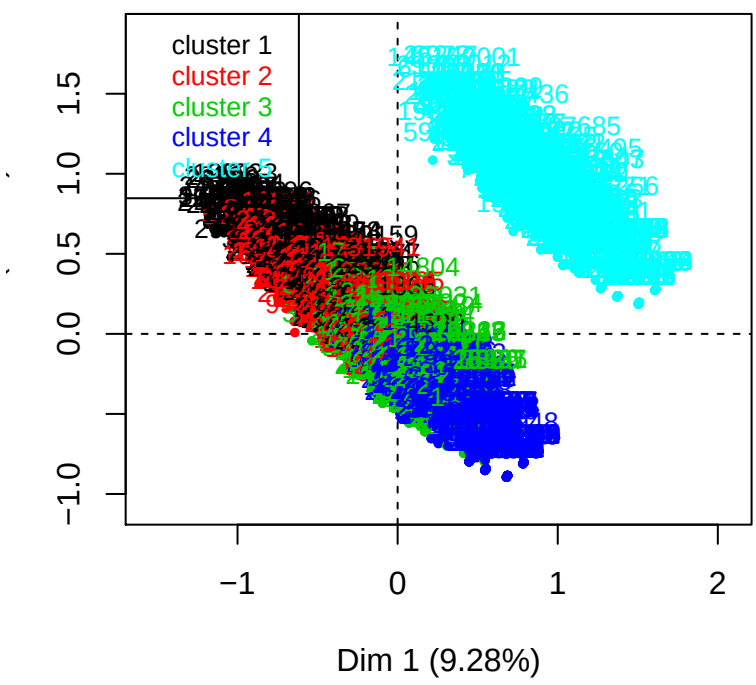
```
num.clusters <- 5
#obtenim que el nombre de clusters es de 5.
```

```
res.hc<-HCPC(res.mca1,nb.clust = num.clusters)
```



Hierarchical clustering on the factor map






```
res.hc$desc.var$quanti
```

```
## $`1`
##                v.test Mean in category Overall mean sd in category
## hr.per.week 14.68475         44.01713     39.81376       8.399916
##              Overall sd      p.value
## hr.per.week   9.758405 8.073113e-49
##
## $`2`
##                v.test Mean in category Overall mean sd in category
## hr.per.week 16.44532          43.5977     39.81376       8.048121
##              Overall sd      p.value
## hr.per.week   9.758405 9.060272e-61
##
## $`3`
##                 v.test Mean in category Overall mean sd in category
## hr.per.week -8.231952         37.25795     39.81376       8.182259
##              Overall sd      p.value
## hr.per.week   9.758405 1.841878e-16
##
## $`4`
##                 v.test Mean in category Overall mean sd in category
## hr.per.week -14.64573         35.93466     39.81376       9.963593
##              Overall sd      p.value
## hr.per.week   9.758405 1.434522e-48
##
```

```
## $`5`
##                v.test Mean in category Overall mean sd in category
## hr.per.week -11.66357         35.62441     39.81376       10.82322
##             Overall sd      p.value
## hr.per.week   9.758405 1.956728e-31
```

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

```
## $`1`
##                   v.test Mean in category Overall mean sd in category
## fnlwgt          17.97953     2.268503e+05 192603.32029   1.205333e+05
## capital.gain   -10.07351     1.200517e+02    569.95991   6.634532e+02
## hr.per.week    -10.34382     3.803133e+01     39.81376   9.929309e+00
## capital.loss   -12.42875     2.187990e-01     91.13994   9.572298e+00
## education.num -22.99849     8.987467e+00     10.03472   1.652864e+00
## age            -45.10583     2.767467e+01     38.43519   6.902466e+00
##                  Overall sd       p.value
## fnlwgt         1.078676e+05  2.818919e-72
## capital.gain   2.529242e+03  7.235097e-24
## hr.per.week    9.758405e+00  4.463617e-25
## capital.loss   4.142711e+02  1.824645e-35
## education.num 2.578697e+00 4.826412e-117
## age            1.350975e+01  0.000000e+00
##
## $`2`
##                    v.test Mean in category Overall mean sd in category
## age             41.054200     5.136247e+01     38.43519       9.723636
## capital.gain    -7.227739     1.438759e+02    569.95991     712.519239
## capital.loss    -9.321452     1.134137e+00     91.13994      29.608321
## fnlwgt        -12.373763     1.614937e+05 192603.32029   84355.478008
## education.num -25.666137     8.492087e+00     10.03472       2.058977
##                  Overall sd       p.value
## age            1.350975e+01  0.000000e+00
## capital.gain   2.529242e+03  4.911002e-13
## capital.loss   4.142711e+02  1.147593e-20
## fnlwgt         1.078676e+05  3.624359e-35
## education.num 2.578697e+00 2.792898e-145
##
## $`3`
##                   v.test Mean in category Overall mean sd in category
## education.num 45.962010     1.316228e+01     10.03472       1.045628
## hr.per.week    8.877367     4.209973e+01     39.81376       9.003224
## age            3.982535     3.985494e+01     38.43519      11.216924
## fnlwgt        -5.652381     1.765143e+05 192603.32029   92991.145555
## capital.gain  -6.393085     1.432747e+02    569.95991     712.747618
## capital.loss  -8.282990     5.920218e-01     91.13994      19.652982
##                  Overall sd      p.value
## education.num 2.578697e+00 0.000000e+00
## hr.per.week   9.758405e+00 6.846193e-19
## age           1.350975e+01 6.818422e-05
```

```
## fnlwgt        1.078676e+05 1.582404e-08
## capital.gain  2.529242e+03 1.625719e-10
## capital.loss  4.142711e+02 1.201201e-16
##
## $`4`
##                   v.test Mean in category Overall mean sd in category
## capital.loss  67.752502       1921.96429     91.13994     339.394614
## education.num  6.879981         11.19196     10.03472       2.713019
## hr.per.week    3.287360         41.90625     39.81376       9.246870
## age            2.728151         40.83929     38.43519      11.575107
## capital.gain  -3.454755          0.00000    569.95991       0.000000
##                 Overall sd      p.value
## capital.loss   414.271117 0.000000e+00
## education.num    2.578697 5.986072e-12
## hr.per.week      9.758405 1.011313e-03
## age             13.509755 6.369036e-03
## capital.gain  2529.242128 5.507936e-04
##
## $`5`
##                   v.test Mean in category Overall mean sd in category
## capital.gain  60.360622      11637.13115    569.95991   5.229969e+03
## education.num 10.045398         11.91257     10.03472   2.211002e+00
## age            7.517837         45.79781     38.43519   1.165851e+01
## hr.per.week    6.427519         44.36066     39.81376   8.250914e+00
## fnlwgt        -2.470810     173282.62295 192603.32029   1.137583e+05
## capital.loss  -3.034810          0.00000     91.13994   0.000000e+00
##                  Overall sd      p.value
## capital.gain  2.529242e+03 0.000000e+00
## education.num 2.578697e+00 9.625806e-24
## age           1.350975e+01 5.568968e-14
## hr.per.week   9.758405e+00 1.297032e-10
## fnlwgt        1.078676e+05 1.348073e-02
## capital.loss  4.142711e+02 2.406869e-03
```

Respecte als clusters fets i analitzats comparantlos amb els obtinguts en el PCA, si ens centrem en el hr.per.week veiem que els grups que ha fet al fer-ho amb el MCA, no són tant distants als que havia fet anteriorment. Peró si que veiem més clarament una separació horaria. per exemple els primers dos clusters la majoria de gent, són persones que tendeixen a tenir grans quantitats d'hores de feina, mentres que els altres no.