ExamFIB-ADEI-Quiz1_19-20Q2

Júlia Gasull

November 10, 2020

Problem 1: All questions account for 1 point

25 personality self-report items taken from the International Personality Item Pool (ipip.ori.org) were included as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project (SAPA https://sapa-project.org). The data from 2800 subjects are included here. Three additional demographic variables (sex, education, and age) are also included. The first 25 items are organized by five putative factors: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. The item data were collected using a 6 point response scale: 1 Very Inaccurate 2 Moderately Inaccurate 3 Slightly Inaccurate 4 Slightly Accurate 5 Moderately Accurate 6 Very Accurate. The items given were sampled from the International Personality Item Pool of Lewis Goldberg using the sampling technique of SAPA. This is a sample data set taken from the much larger SAPA data bank.

Name Description * A1 I Am indifferent to the feelings of others. (q_146) * A2 Inquire about others' well-being. (q_1162) * A3 Know how to comfort others. (q_1206) * A4 Love children. (q_1364) * A5 Make people feel at ease. (q_1419) * C1 Am exacting in my work. (q_124) * C2 Continue until everything is perfect. (q_530) * C3 Do things according to a plan. (q_619) * C4 Do things in a half-way manner. (q_626) * C5 Waste my time. (q_1949) * E1 Don't talk a lot. (q_712) * E2 Find it difficult to approach others. (q_901) * E3 Know how to captivate people. (q_1205) * E4 Make friends easily. (q_1410) * E5 Take charge. (q_1768) * N1 Get angry easily. (q_952) * N2 Get irritated easily. (q_974) * N3 Have frequent mood swings. (q_1099 * N4 Often feel blue. (q_1479) * N5 Panic easily. (q_1505) * O1 Am full of ideas. (q_128) * O2 Avoid difficult reading material.(q_316) * O3 Carry the conversation to a higher level. (q_492) * O4 Spend time reflecting on things. (q_1738) * O5 Will not probe deeply into a subject. (q_1964) * gender gender Males = 1, Females =2 * education 1 = HS, 2 = finished HS, 3 = some college, 4 = college graduate 5 = graduate degree * age age in years

Firstly, load dataset and check available variables.

```
rm(list=ls())
setwd("~/Documents/uni/FIB-ADEI/exams")
load("~/Documents/uni/FIB-ADEI/exams/bfi_Raw.RData")
summary(df)
```

```
##        A1              A2              A3              A4           A5
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.0   1st Qu.:4.00
##  Median :2.000   Median :5.000   Median :5.000   Median :5.0   Median :5.00
##  Mean   :2.413   Mean   :4.802   Mean   :4.604   Mean   :4.7   Mean   :4.56
##  3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.0   3rd Qu.:5.00
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.0   Max.   :6.00
##  NA's   :16      NA's   :27      NA's   :26      NA's   :19    NA's   :16
##        C1              C2              C3              C4             C5
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:4.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :5.000   Median :5.00   Median :5.000   Median :2.000   Median :3.000
```

```
##  Mean   :4.502   Mean   :4.37   Mean   :4.304   Mean   :2.553   Mean   :3.297
##  3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.00   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :21      NA's   :24     NA's   :20      NA's   :26      NA's   :16
##        E1              E2              E3              E4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:4.000
##  Median :3.000   Median :3.000   Median :4.000   Median :5.000
##  Mean   :2.974   Mean   :3.142   Mean   :4.001   Mean   :4.422
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :23      NA's   :16      NA's   :25      NA's   :9
##        E5              N1              N2              N3
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :5.000   Median :3.000   Median :4.000   Median :3.000
##  Mean   :4.416   Mean   :2.929   Mean   :3.508   Mean   :3.217
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :21      NA's   :22      NA's   :21      NA's   :11
##        N4              N5              O1              O2              O3
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:4.000
##  Median :3.000   Median :3.00   Median :5.000   Median :2.000   Median :5.000
##  Mean   :3.186   Mean   :2.97   Mean   :4.816   Mean   :2.713   Mean   :4.438
##  3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:6.000   3rd Qu.:4.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.00   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :36      NA's   :29     NA's   :22                      NA's   :28
##        O4              O5          gender        education         age
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.00   Min.   : 3.00
##  1st Qu.:4.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:3.00   1st Qu.:20.00
##  Median :5.000   Median :2.00   Median :2.000   Median :3.00   Median :26.00
##  Mean   :4.892   Mean   :2.49   Mean   :1.672   Mean   :3.19   Mean   :28.78
##  3rd Qu.:6.000   3rd Qu.:3.00   3rd Qu.:2.000   3rd Qu.:4.00   3rd Qu.:35.00
##  Max.   :6.000   Max.   :6.00   Max.   :2.000   Max.   :5.00   Max.   :86.00
##  NA's   :14      NA's   :20                     NA's   :223
```

Load Required Packages for this deliverable

We load the necessary packages and set working directory

```
# Load Required Packages
options(contrasts=c("contr.treatment","contr.treatment"))

requiredPackages <- c("missMDA","chemometrics","mvoutlier","effects","FactoMineR","car", "factoextra","
missingPackages <- requiredPackages[!(requiredPackages %in% installed.packages()[,"Package"])]

if(length(missingPackages)) install.packages(missingPackages)
lapply(requiredPackages, require, character.only = TRUE)
```

Some useful functions

```
calcQ <- function(x) { # Function to calculate the different quartiles
  s.x <- summary(x)
  iqr<-s.x[5]-s.x[2]
  list(souti=s.x[2]-3*iqr, mouti=s.x[2]-1.5*iqr, min=s.x[1], q1=s.x[2], q2=s.x[3],
       q3=s.x[5], max=s.x[6], mouts=s.x[5]+1.5*iqr, souts=s.x[5]+3*iqr )
}

countNA <- function(x) { # Function to count the NA values
  mis_x <- NULL
  for (j in 1:ncol(x)) {mis_x[j] <- sum(is.na(x[,j])) }
  mis_x <- as.data.frame(mis_x)
  rownames(mis_x) <- names(x)
  mis_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {mis_i <- mis_i + as.numeric(is.na(x[,j])) }
  list(mis_col=mis_x,mis_ind=mis_i)
}

countX <- function(x,X) { # Function to count a specific number of appearences
  n_x <- NULL
  for (j in 1:ncol(x)) {n_x[j] <- sum(x[,j]==X) }
  n_x <- as.data.frame(n_x)
  rownames(n_x) <- names(x)
  nx_i <- rep(0,nrow(x))
  for (j in 1:ncol(x)) {nx_i <- nx_i + as.numeric(x[,j]==X) }
  list(nx_col=n_x,nx_ind=nx_i)
}
```

1. Define a binary factor for gender f.gender and a polytomic factor for education f.educ. Justify with R commands for the procedure and your answer. Calculate thresholds to identify severe outliers for the age variable (age).

- El gènere i l'educació es defineixen com a factors.

Binary factor for gender f.gender

```
df$f.gender <- factor(df$gender, labels=(c("sex.male","sex.famale")))
summary(df$gender)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.672   2.000   2.000
```

```
summary(df$f.gender)
```

```
##   sex.male sex.famale
##        919       1881
```

Polytomic factor for education f.educ

```
ll<-which(is.na(df$education))
df$education[ll]<-6
df$f.educ <- factor(df$education, labels=c("hs", "finished hs", "some college", "college graduate", "gra
summary(df$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.000   3.414   4.000   6.000
```

```
summary(df$f.educ) # --> we have NAs
```

```
##               hs      finished hs     some college college graduate
##              224              292             1249              394
##  graduate degree          unknown
##              418              223
```

- Es pot veure que a l'educació amb 223 missings. La imputació no és una solució raonable i cal definir un nivell específic desconegut.

Outliers for age

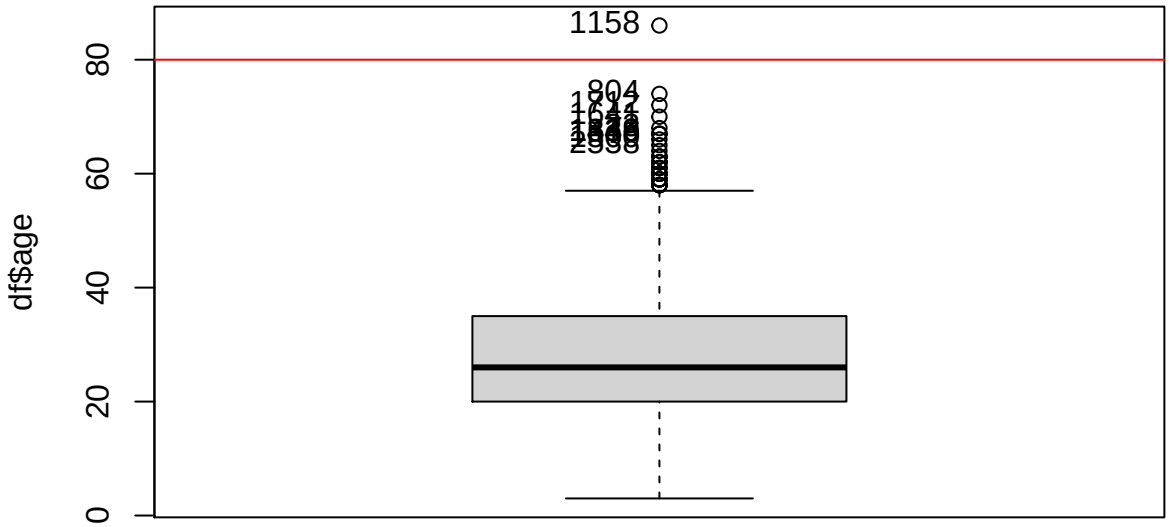
```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   20.00   26.00   28.78   35.00   86.00
```

```
Boxplot(df$age)
```

```
##  [1] 1158  804 1712 1641   23 1349 1436 1518 1500 2538
```

```
var_out<-calcQ(df$age)
abline(h=var_out$souts,col="red")
abline(h=var_out$souti,col="red")
```




```
llout<-which((df$age>80))
llout
```

```
## [1] 1158
```

```
# sumres<-summary(df$age)
# iqr<-as.numeric(sumres[5]-sumres[2]);iqr
# mildlow<-as.numeric(sumres[2]-1.5*iqr)
# mildup<-as.numeric(sumres[5]+1.5*iqr)
# sevlow<-as.numeric(sumres[2]-3*iqr)
# sevup<-as.numeric(sumres[5]+3*iqr)
# mildlow;mildup
# sevlow;sevup
# ll<-which(df$age>sevup);length(ll);ll
```

- L'edat és una variable numèrica sense missings.
- El càlcul dels llindars atípics greus determina que aquelles observacions superiors a 80 són atípics greus: només 1 persona compleix aquesta condició (obs. 1158).

- El llindar greu inferior no té sentit (ja que és -25).

2. Conduct a suitable data imputation procedure to remove missing data included in dataset for numeric variables. Check imputation consistency for numeric variables.

```
library(missMDA)
summary(df)
```

```
##        A1              A2              A3              A4            A5
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0   Min.   :1.00
##  1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.0   1st Qu.:4.00
##  Median :2.000   Median :5.000   Median :5.000   Median :5.0   Median :5.00
##  Mean   :2.413   Mean   :4.802   Mean   :4.604   Mean   :4.7   Mean   :4.56
##  3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.0   3rd Qu.:5.00
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.0   Max.   :6.00
##  NA's   :16      NA's   :27      NA's   :26      NA's   :19    NA's   :16
##        C1              C2              C3              C4             C5
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:4.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :5.000   Median :5.00   Median :5.000   Median :2.000   Median :3.000
##  Mean   :4.502   Mean   :4.37   Mean   :4.304   Mean   :2.553   Mean   :3.297
##  3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.00   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :21      NA's   :24     NA's   :20      NA's   :26      NA's   :16
##        E1              E2              E3              E4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:4.000
##  Median :3.000   Median :3.000   Median :4.000   Median :5.000
##  Mean   :2.974   Mean   :3.142   Mean   :4.001   Mean   :4.422
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :23      NA's   :16      NA's   :25      NA's   :9
##        E5              N1              N2              N3
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :5.000   Median :3.000   Median :4.000   Median :3.000
##  Mean   :4.416   Mean   :2.929   Mean   :3.508   Mean   :3.217
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :21      NA's   :22      NA's   :21      NA's   :11
##        N4              N5              O1              O2             O3
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.00   1st Qu.:4.000   1st Qu.:1.000   1st Qu.:4.000
##  Median :3.000   Median :3.00   Median :5.000   Median :2.000   Median :5.000
##  Mean   :3.186   Mean   :2.97   Mean   :4.816   Mean   :2.713   Mean   :4.438
##  3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:6.000   3rd Qu.:4.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.00   Max.   :6.000   Max.   :6.000   Max.   :6.000
##  NA's   :36      NA's   :29     NA's   :22                      NA's   :28
##        O4              O5           gender        education          age
##  Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   : 3.00
##  1st Qu.:4.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:20.00
##  Median :5.000   Median :2.00   Median :2.000   Median :3.000   Median :26.00
##  Mean   :4.892   Mean   :2.49   Mean   :1.672   Mean   :3.414   Mean   :28.78
##  3rd Qu.:6.000   3rd Qu.:3.00   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:35.00
##  Max.   :6.000   Max.   :6.00   Max.   :2.000   Max.   :6.000   Max.   :86.00
```

```
##  NA's   :14      NA's   :20
##        f.gender                 f.educ
##  sex.male  : 919   hs              : 224
##  sex.famale:1881   finished hs     : 292
##                    some college    :1249
##                    college graduate: 394
##                    graduate degree : 418
##                    unknown         : 223
##
```

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

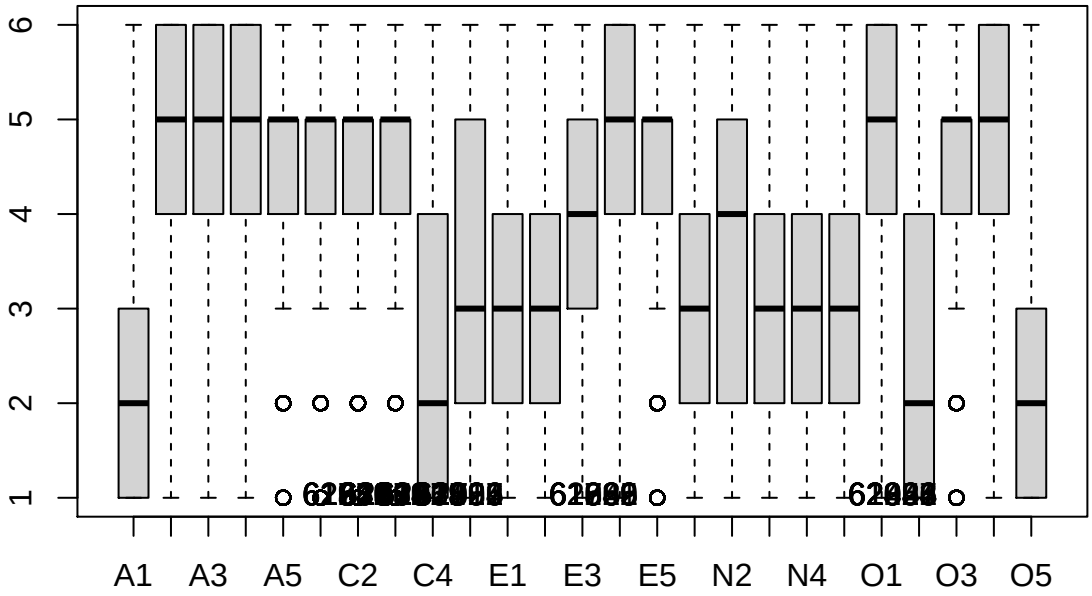
```
vars_quantitatives<-names(df)[c(1:25)]
res.imputation<-imputePCA(df[,vars_quantitatives],ncp=5)
dfimpu<-as.data.frame(res.imputation$completeObs)
summary(res.imputation$completeObs)
```

```
##        A1              A2              A3              A4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :2.000   Median :5.000   Median :5.000   Median :5.000
##  Mean   :2.414   Mean   :4.804   Mean   :4.605   Mean   :4.701
##  3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.072
##        A5              C1              C2              C3
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :5.000   Median :5.000   Median :5.000   Median :5.000
##  Mean   :4.561   Mean   :4.503   Mean   :4.372   Mean   :4.303
##  3rd Qu.:5.034   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        C4              C5              E1              E2
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :3.000   Median :3.000
##  Mean   :2.553   Mean   :3.296   Mean   :2.975   Mean   :3.142
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        E3              E4              E5              N1
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:2.000
##  Median :4.000   Median :5.000   Median :5.000   Median :3.000
##  Mean   :4.001   Mean   :4.421   Mean   :4.417   Mean   :2.932
##  3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        N2              N3              N4              N5
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :4.000   Median :3.000   Median :3.000   Median :3.000
```

```
##  Mean   :3.508   Mean   :3.216   Mean   :3.183   Mean   :2.969
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        O1              O2              O3              O4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :5.000   Median :2.000   Median :5.000   Median :5.000
##  Mean   :4.816   Mean   :2.713   Mean   :4.435   Mean   :4.892
##  3rd Qu.:6.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.146
##        O5
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.492
##  3rd Qu.:3.000
##  Max.   :6.000
```

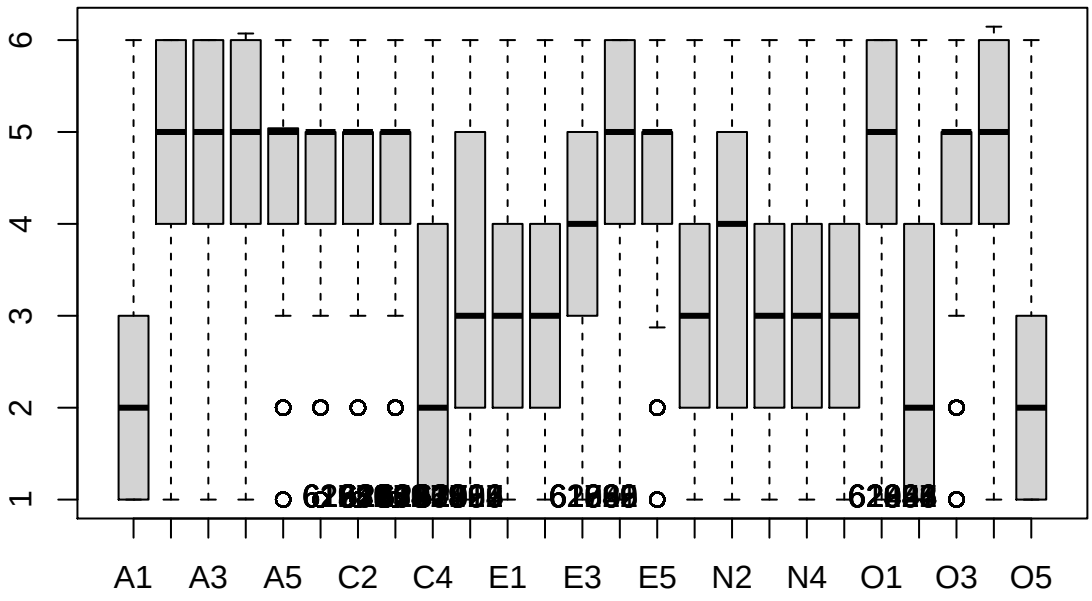
Checking imputation consistency

```
library(car)
Boxplot(df[,vars_quantitatives])
```




```
##  [1] "61629" "61640" "61788" "61840" "61873" "61926" "61932" "62282" "62551"
## [10] "62552" "61654" "61682" "61761" "61921" "61979" "62038" "62060" "62102"
## [19] "62111" "62498" "61654" "61825" "61839" "61865" "61918" "61921" "61969"
## [28] "61979" "62029" "62079" "61654" "61701" "61716" "62022" "62029" "62092"
## [37] "62526" "62716" "62787" "62795" "61629" "61682" "61761" "61788" "61825"
## [46] "61840" "61865" "61989" "62092" "62266" "61856" "61926" "62022" "62054"
## [55] "62064" "62246" "62327" "62328" "62443" "62491"
```

```
Boxplot(dfimpu[,vars_quantitatives])
```




```
##  [1] "61629" "61640" "61788" "61840" "61873" "61926" "61932" "62282" "62551"
## [10] "62552" "61654" "61682" "61761" "61921" "61979" "62038" "62060" "62102"
## [19] "62111" "62498" "61654" "61825" "61839" "61865" "61918" "61921" "61969"
## [28] "61979" "62029" "62079" "61654" "61701" "61716" "62022" "62029" "62092"
## [37] "62526" "62716" "62787" "62795" "61629" "61682" "61761" "61788" "61825"
## [46] "61840" "61865" "61989" "62092" "62266" "61856" "61926" "62022" "62054"
## [55] "62064" "62246" "62327" "62328" "62443" "62491"
```

```
lapply(df[,vars_quantitatives],quantile, probs=seq(0,1,0.1),na.rm=T)
```

```
## $A1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    1    2    2    2    3    4    5    6
##
## $A2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    6    6    6    6
##
## $A3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    5    6    6    6
##
## $A4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    4    4    5    5    6    6    6    6    6
##
## $A5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    4    5    5    5    6    6    6
##
## $C1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    4    5    5    5    6    6    6
##
## $C2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
```

```
##    1    2    3    4    4    5    5    5    5    6    6
##
## $C3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    5    6    6
##
## $C4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    3    4    5    6
##
## $C5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    3    3    4    4    5    6    6
##
## $E1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    5    5    6
##
## $E2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $E3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    3    4    4    4    5    5    6    6
##
## $E4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    6    6    6
##
## $E5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    6    6    6
##
## $N1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    4    5    6
##
## $N2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    3    4    4    4    5    6    6
##
## $N3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $N4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $N5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    5    5    6
##
```

```
## $O1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    6    6    6    6
##
## $O2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    4    4    5    6
##
## $O3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    4    5    5    5    5    6    6
##
## $O4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    5    5    5    5    6    6    6    6
##
## $O5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    3    4    4    6
```

```
lapply(dfimpu[,vars_quantitatives],quantile, probs=seq(0,1,0.1),na.rm=T)
```

```
## $A1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    1    2    2    2    3    4    5    6
##
## $A2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    6    6    6    6
##
## $A3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    5    6    6    6
##
## $A4
##       0%      10%      20%      30%      40%      50%      60%      70%
## 1.000000 2.000000 4.000000 4.000000 5.000000 5.000000 6.000000 6.000000
##      80%      90%     100%
## 6.000000 6.000000 6.071715
##
## $A5
##       0%      10%      20%      30%      40%      50%      60%      70%
## 1.000000 3.000000 4.000000 4.000000 4.261158 5.000000 5.000000 5.000000
##      80%      90%     100%
## 6.000000 6.000000 6.000000
##
## $C1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    4    5    5    5    6    6    6
##
## $C2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    5    6    6
##
## $C3
```

```
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    5    6    6
##
## $C4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    3    4    5    6
##
## $C5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    3    3    4    4    5    6    6
##
## $E1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    5    5    6
##
## $E2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $E3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    3    4    4    4    5    5    6    6
##
## $E4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    6    6    6
##
## $E5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    2    3    4    4    5    5    5    6    6    6
##
## $N1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    4    5    6
##
## $N2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    3    4    4    4    5    6    6
##
## $N3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $N4
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    2    2    2    3    4    4    5    5    6
##
## $N5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    3    3    4    5    5    6
##
## $O1
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    5    5    5    6    6    6    6
```

```
##
## $O2
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    4    4    5    6
##
## $O3
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    3    4    4    4    5    5    5    5    6    6
##
## $O4
##       0%      10%      20%      30%      40%      50%      60%      70%
## 1.000000 3.000000 4.000000 5.000000 5.000000 5.000000 5.000000 6.000000
##      80%      90%     100%
## 6.000000 6.000000 6.145828
##
## $O5
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    1    1    1    2    2    2    3    3    4    4    6
```

```
df[,vars_quantitatives]<-res.imputation$completeObs[,vars_quantitatives]
```

- Les 25 primeres variables tenen missings, entre 9 i 36, excepte la variable O2.
- El mètode imputePCA () del paquet missMDA s'ha d'utilitzar per a la imputació de variables numèriques.
- Cal comprovar si hi ha valors d'imputació raonables mitjançant gràfics o quantils.
- Sembla que no hi ha cap problema.

3. Conduct a suitable data imputation procedure for factors. Summarize imputation results for f.education factor.

```
library(missMDA)
summary(df)
```

```
##        A1              A2              A3              A4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :2.000   Median :5.000   Median :5.000   Median :5.000
##  Mean   :2.414   Mean   :4.804   Mean   :4.605   Mean   :4.701
##  3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.072
##        A5              C1              C2              C3
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :5.000   Median :5.000   Median :5.000   Median :5.000
##  Mean   :4.561   Mean   :4.503   Mean   :4.372   Mean   :4.303
##  3rd Qu.:5.034   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        C4              C5              E1              E2
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :3.000   Median :3.000
##  Mean   :2.553   Mean   :3.296   Mean   :2.975   Mean   :3.142
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        E3              E4              E5              N1
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:2.000
```

```
##  Median :4.000   Median :5.000   Median :5.000   Median :3.000
##  Mean   :4.001   Mean   :4.421   Mean   :4.417   Mean   :2.932
##  3rd Qu.:5.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        N2              N3              N4              N5
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :4.000   Median :3.000   Median :3.000   Median :3.000
##  Mean   :3.508   Mean   :3.216   Mean   :3.183   Mean   :2.969
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##        O1              O2              O3              O4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :5.000   Median :2.000   Median :5.000   Median :5.000
##  Mean   :4.816   Mean   :2.713   Mean   :4.435   Mean   :4.892
##  3rd Qu.:6.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:6.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.146
##        O5            gender        education          age
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 3.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:20.00
##  Median :2.000   Median :2.000   Median :3.000   Median :26.00
##  Mean   :2.492   Mean   :1.672   Mean   :3.414   Mean   :28.78
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:35.00
##  Max.   :6.000   Max.   :2.000   Max.   :6.000   Max.   :86.00
##        f.gender                 f.educ
##  sex.male  : 919   hs              : 224
##  sex.famale:1881   finished hs     : 292
##                    some college    :1249
##                    college graduate: 394
##                    graduate degree : 418
##                    unknown         : 223
```

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
vars_categorical<-names(df)[c(29:30)]
summary(df[,vars_categorical])
```

```
##        f.gender                 f.educ
##  sex.male  : 919   hs              : 224
##  sex.famale:1881   finished hs     : 292
##                    some college    :1249
##                    college graduate: 394
##                    graduate degree : 418
##                    unknown         : 223
```

```
# res.input<-imputeMCA(df[,vars_categorical],ncp=2)
# summary(res.input$completeObs)
```

- La imputació de factors hauria d'utilitzar les dades f.educ i f.gender, de manera que no és probable que

continguin prou informació per a una imputació adequada.

- Si s'hagués inclòs un conjunt de factors al conjunt de dades, aleshores s'hauria d'utilitzar imputeMCA() al paquet missMDA amb finalitats d'imputació.
- De fet, els valors que falten de l'educació variable s'han de seleccionar per definir un nou nivell en el factor f.educ etiquetat com a "Desconegut".
- Hi ha 223 observacions amb un nivell d'educació desconegut.

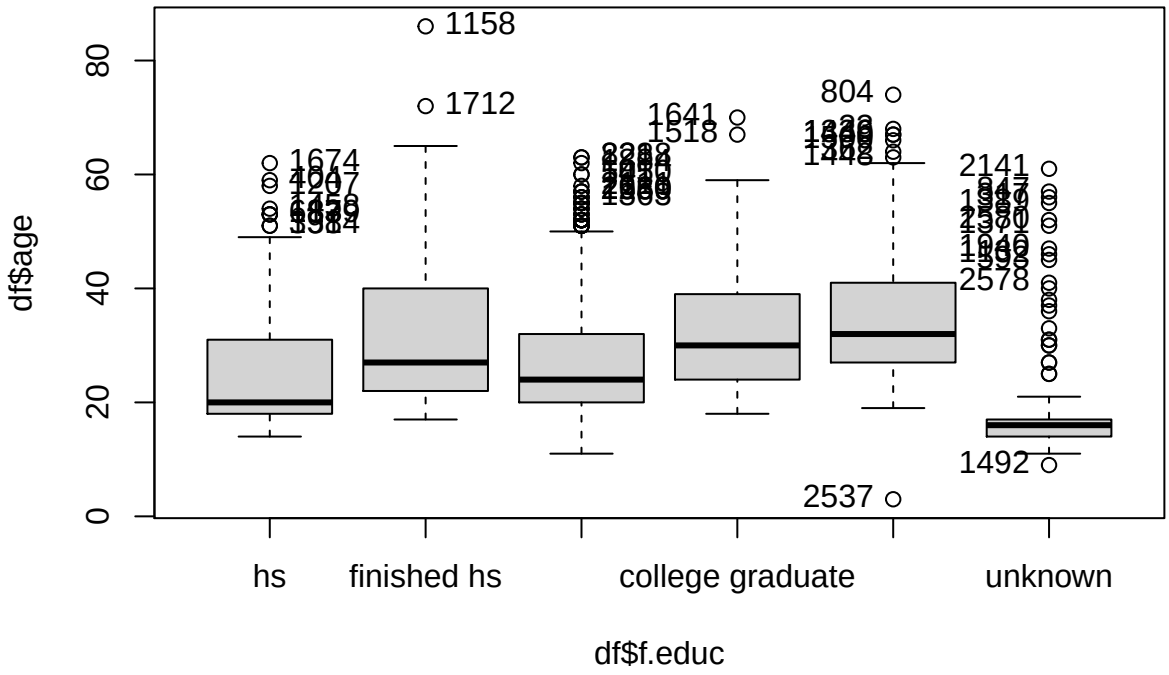
4. Can the average of age can be argued to be the same for all education levels (f.educ) and gender (f.gender)? Which are the groups that show significant greater values than the others? Use graphic, numeric and inferential tools.

Profiling of age -> f.educ and f.gender

```
Boxplot(df$age~df$f.gender,main="Age by Gender")
```


```
##  [1] "804"  "1641" "23"   "1349" "1436" "1518" "2538" "1448" "1100" "1242"
## [11] "1158" "1712" "1500" "362"  "821"  "2288" "1674" "1884" "2450" "2141"
```

```
Boxplot(df$age~df$f.educ,main="Age by Education Level")
```




```
##  [1] "183"  "338"  "404"  "613"  "1207" "1458" "1674" "1879" "1914" "1158"
## [11] "1712" "821"  "2288" "1884" "545"  "1010" "2111" "1969" "2280" "2630"
## [21] "1363" "1518" "1641" "2537" "23"   "362"  "804"  "1349" "1436" "1448"
## [31] "1500" "1492" "2141" "847"  "317"  "1389" "2580" "1371" "1940" "1132"
## [41] "593"  "2578"
```

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
vars_res<-names(df)[c(28)]
res.condes <- condes(df[, c(vars_res,vars_quantitatives, vars_categorical)],1)

res.condes$quali
```

```
##                   R2       p.value
## f.educ    0.170395545 1.279640e-110
## f.gender 0.002275621  1.158497e-02
```

```
res.condes$category
```

```
##                             Estimate      p.value
## f.educ=graduate degree     6.9505669 1.064868e-39
## f.educ=college graduate    4.6288269 4.548861e-16
## f.educ=finished hs         3.1628301 9.046486e-06
## f.gender=sex.famale        0.5651296 1.158497e-02
## f.gender=sex.male         -0.5651296 1.158497e-02
## f.educ=hs                 -3.2169400 2.953164e-07
```

```
## f.educ=some college      -1.1250879 2.627052e-11
## f.educ=unknown          -10.4001959 5.288444e-54
```

- Gràfics: boxplot d'edat per a f.gender és difícil d'avaluar, però boxplot per a cada nivell de f.educ mostra un perfil diferent clar per edat en funció dels nivells.
- Es pot utilitzar un mètode condes() per a una resposta ràpida:
  - $quali mostra la importància global de f.educ i f.gender.
  - També mostra que els estudis de postgrau, universitaris, HS acabats i les edats mitjanes femenines superen la mitjana i els homes, HS, algunes universitats i desconegudes estan significativament per sota de la mitjana global d'edat.

5. Let us assume that education (f.educ) is the target variable. Use a suitable feature selection and profiling tool to discuss global association between target and numerical variables/factors in dataset.

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
vars_res<-names(df)[c(30)]
res.catdes <- catdes(df[, c(vars_res,vars_quantitatives, vars_categorical)], 1)
res.catdes$quanti.var
```

```
##           Eta2      P-value
## A4 0.038250591 6.574681e-22
## A1 0.027911554 1.272093e-15
## C4 0.022719413 1.607478e-12
## C5 0.019299303 1.671006e-10
## O3 0.019159725 2.017279e-10
## E4 0.016094231 1.227324e-08
## A2 0.015058808 4.849805e-08
## C1 0.014902578 5.963134e-08
## O2 0.013667891 3.032756e-07
## A5 0.012577693 1.261204e-06
## A3 0.011722108 3.828150e-06
## C2 0.010381110 2.145318e-05
## E5 0.010104322 3.053083e-05
## N1 0.009808988 4.443579e-05
## O5 0.009688231 5.178670e-05
## N4 0.009052973 1.154351e-04
## N2 0.008662120 1.883918e-04
## O4 0.008623240 1.977688e-04
## C3 0.008244104 3.171175e-04
## E2 0.007653338 6.580014e-04
## E3 0.006959286 1.535438e-03
## N3 0.006588420 2.402430e-03
## O1 0.005990587 4.901241e-03
## N5 0.005220952 1.204917e-02
```

Associat globalment a f.educ és el factor f.gender. f.educ s'associa globalment a l'edat de les variables numèriques i a 24 ítems més, sent les A4, A1, C4, C5 i O3 més significatives.

6. Profile HS education group according to available data in your dataset.

```
res.catdes$category["hs"]
```

```
## $hs
##                              Cla/Mod   Mod/Cla    Global      p.value
## f.educ.1=hs                100.000000 100.00000  8.000000 0.000000e+00
## f.gender=sex.male           10.119695  41.51786 32.821429 4.525328e-03
## f.gender=sex.famale          6.964381  58.48214 67.178571 4.525328e-03
## f.educ.1=unknown             0.000000   0.00000  7.964286 3.725170e-09
## f.educ.1=finished hs         0.000000   0.00000 10.428571 6.424107e-12
## f.educ.1=college graduate    0.000000   0.00000 14.071429 3.744301e-16
## f.educ.1=graduate degree     0.000000   0.00000 14.928571 3.545195e-17
## f.educ.1=some college        0.000000   0.00000 44.607143 1.442081e-61
##                                v.test
## f.educ.1=hs                       Inf
## f.gender=sex.male            2.839013
## f.gender=sex.famale         -2.839013
## f.educ.1=unknown            -5.895957
## f.educ.1=finished hs        -6.869913
## f.educ.1=college graduate   -8.146558
## f.educ.1=graduate degree    -8.427040
## f.educ.1=some college      -16.556294
```

```
res.catdes$quanti["hs"]
```

```
## $hs
##       v.test Mean in category Overall mean sd in category Overall sd
## C4  2.714228         2.790958     2.552636       1.378100   1.369844
## A1  2.138448         2.606140     2.413622       1.438861   1.404512
## O3 -2.177459         4.265698     4.435417       1.373329   1.215998
## A4 -2.355596         4.477871     4.700615       1.583869   1.475224
## E4 -2.410467         4.196429     4.421325       1.543068   1.455575
## C3 -2.499951         4.096930     4.302816       1.342935   1.284835
## E5 -3.088415         4.153491     4.417105       1.299218   1.331633
##        p.value
## C4 0.006643040
## A1 0.032480398
## O3 0.029446368
## A4 0.018493029
## E4 0.015932122
## C3 0.012421035
## E5 0.002012274
```

Els homes estan sobrerepresentats en el nivell HS (41,5% del grup HS vs 32,82% a nivell mundial, més del 10% dels homes inclosos a la mostra pertanyen al grup HS), mentre que estan subrepresentats en grups "d'alguns col · legis". En concret, les variables numèriques les mitjanes de les quals són significativament diferents de la mitjana global per a cada nivell f.educ són:

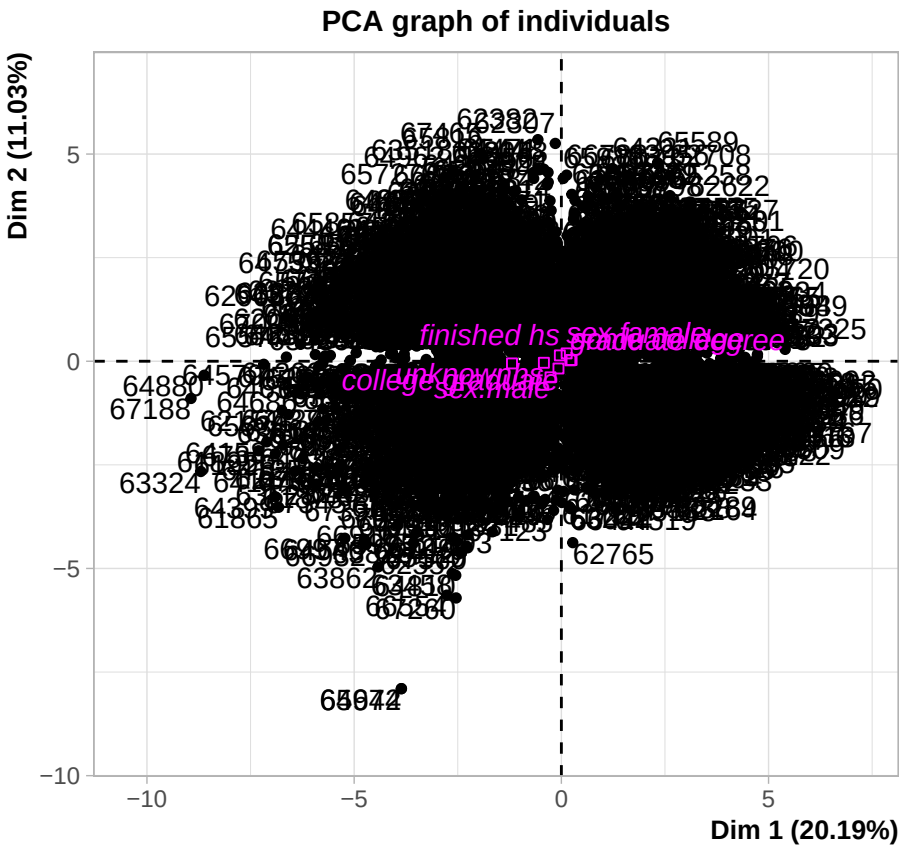
- Per als nivells "HS" C4 i A1 superen la mitjana global, mentre que l'edat E5 i C3 estan per sota de la mitjana global. Aquesta és la resposta directa a la pregunta.
- Per a l'edat del nivell "HS acabat" i A1 superen la mitjana global.

7. A Normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age. How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?.

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
res.pca <- PCA(df[, c(1:25, 28:30)], quanti.sup=26, quali.sup=27:28)
```



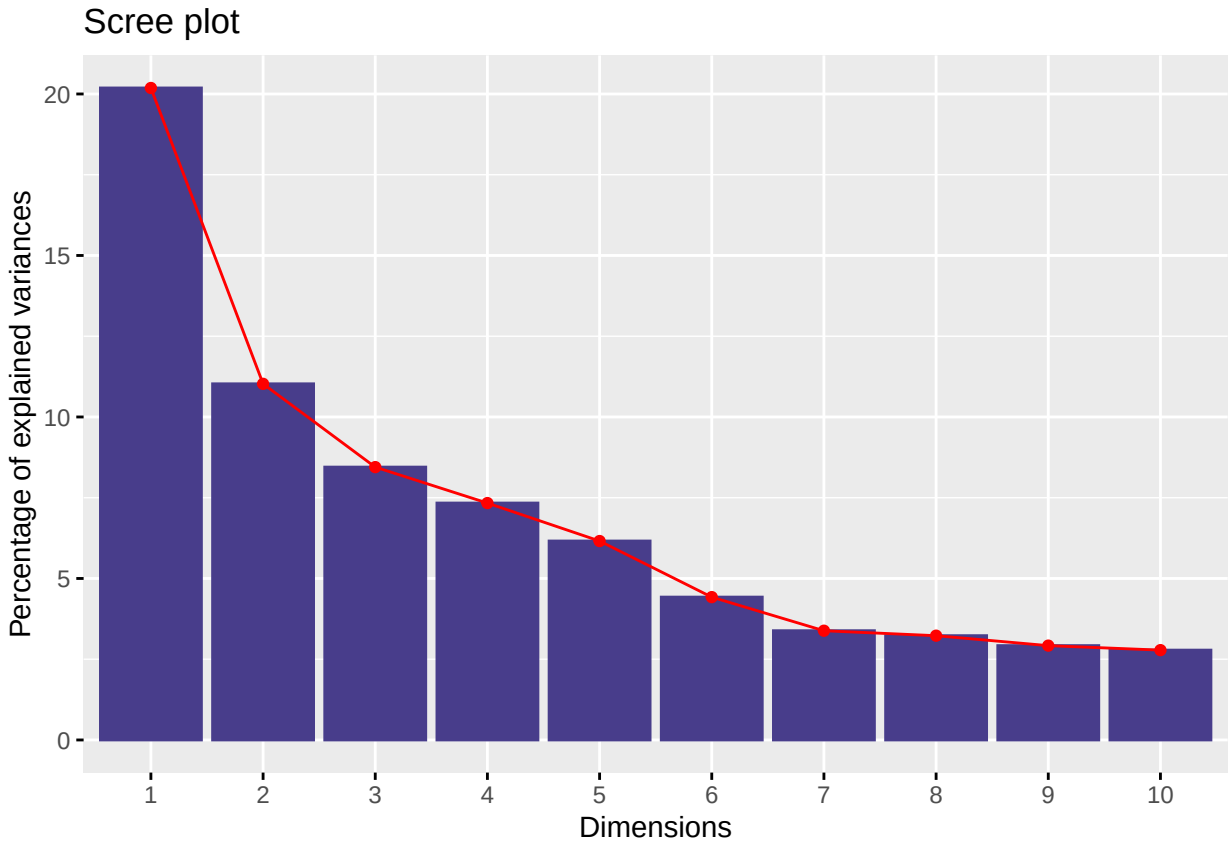

```
summary(res.pca,nbind=0,nbelements = 25)
```

```
##
## Call:
## PCA(X = df[, c(1:25, 28:30)], quanti.sup = 26, quali.sup = 27:28)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               5.046   2.756   2.112   1.834   1.540   1.106   0.846
## % of var.             20.185  11.026   8.448   7.334   6.159   4.423   3.382
## Cumulative % of var.  20.185  31.211  39.658  46.993  53.152  57.575  60.957
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.807   0.730   0.695   0.680   0.656   0.628   0.594
## % of var.              3.229   2.921   2.782   2.721   2.625   2.510   2.376
## Cumulative % of var.  64.186  67.107  69.888  72.609  75.234  77.745  80.121
##                       Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21
## Variance               0.560   0.540   0.523   0.498   0.488   0.455   0.432
```

```
## % of var.              2.242   2.159   2.093   1.993   1.953   1.821   1.730
## Cumulative % of var.  82.363  84.521  86.614  88.608  90.561  92.382  94.112
##                       Dim.22  Dim.23  Dim.24  Dim.25
## Variance               0.408   0.405   0.384   0.275
## % of var.              1.632   1.620   1.538   1.099
## Cumulative % of var.  95.743  97.364  98.901 100.000
##
## Variables
##                      Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## A1                | -0.253  1.267  0.064 | -0.006  0.001  0.000 |  0.140  0.923
## A2                |  0.492  4.800  0.242 |  0.316  3.621  0.100 | -0.179  1.520
## A3                |  0.554  6.088  0.307 |  0.317  3.649  0.101 | -0.268  3.395
## A4                |  0.436  3.767  0.190 |  0.134  0.649  0.018 | -0.146  1.011
## A5                |  0.606  7.281  0.367 |  0.204  1.505  0.041 | -0.293  4.054
## C1                |  0.357  2.523  0.127 |  0.137  0.685  0.019 |  0.531 13.351
## C2                |  0.354  2.483  0.125 |  0.217  1.708  0.047 |  0.517 12.660
## C3                |  0.341  2.303  0.116 |  0.074  0.201  0.006 |  0.410  7.965
## C4                | -0.488  4.728  0.239 |  0.147  0.780  0.022 | -0.472 10.545
## C5                | -0.523  5.425  0.274 |  0.189  1.300  0.036 | -0.292  4.026
## E1                | -0.441  3.847  0.194 | -0.202  1.481  0.041 |  0.343  5.578
## E2                | -0.634  7.957  0.402 | -0.024  0.021  0.001 |  0.286  3.881
## E3                |  0.565  6.323  0.319 |  0.357  4.624  0.127 | -0.159  1.203
## E4                |  0.607  7.296  0.368 |  0.190  1.310  0.036 | -0.383  6.949
## E5                |  0.554  6.091  0.307 |  0.309  3.456  0.095 |  0.084  0.332
## N1                | -0.434  3.728  0.188 |  0.646 15.122  0.417 |  0.018  0.015
## N2                | -0.426  3.597  0.182 |  0.649 15.259  0.421 |  0.068  0.221
## N3                | -0.416  3.429  0.173 |  0.673 16.437  0.453 |  0.039  0.072
## N4                | -0.545  5.876  0.297 |  0.472  8.075  0.223 |  0.111  0.584
## N5                | -0.368  2.681  0.135 |  0.510  9.451  0.261 | -0.034  0.055
## O1                |  0.365  2.635  0.133 |  0.247  2.212  0.061 |  0.256  3.107
## O2                | -0.221  0.965  0.049 |  0.113  0.461  0.013 | -0.389  7.154
## O3                |  0.435  3.752  0.189 |  0.316  3.623  0.100 |  0.189  1.688
## O4                | -0.060  0.072  0.004 |  0.343  4.280  0.118 |  0.282  3.768
## O5                | -0.234  1.086  0.055 | -0.049  0.088  0.002 | -0.354  5.944
##                     cos2
## A1                 0.019 |
## A2                 0.032 |
## A3                 0.072 |
## A4                 0.021 |
## A5                 0.086 |
## C1                 0.282 |
## C2                 0.267 |
## C3                 0.168 |
## C4                 0.223 |
## C5                 0.085 |
## E1                 0.118 |
## E2                 0.082 |
## E3                 0.025 |
## E4                 0.147 |
## E5                 0.007 |
## N1                 0.000 |
## N2                 0.005 |
## N3                 0.002 |
## N4                 0.012 |
```

```
## N5                0.001 |
## O1                0.066 |
## O2                0.151 |
## O3                0.036 |
## O4                0.080 |
## O5                0.126 |
##
## Supplementary continuous variable
##                     Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
## age              |  0.166  0.027 | -0.041  0.002 |  0.045  0.002 |
##
## Supplementary categories
##                      Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## sex.male          |  0.701 | -0.261  0.139 -4.303 | -0.388  0.306 -8.640 |
## sex.famale        |  0.342 |  0.128  0.139  4.303 |  0.190  0.306  8.640 |
## hs                |  0.505 | -0.422  0.699 -2.930 | -0.037  0.005 -0.349 |
## finished hs       |  0.338 | -0.037  0.012 -0.294 |  0.141  0.174  1.531 |
## some college      |  0.389 |  0.247  0.402  5.215 |  0.031  0.006  0.884 |
## college graduate  |  0.433 | -0.070  0.026 -0.667 | -0.175  0.164 -2.255 |
## graduate degree   |  0.586 |  0.216  0.136  2.131 |  0.022  0.001  0.296 |
## unknown           |  1.338 | -1.192  0.794 -8.255 | -0.053  0.002 -0.496 |
##                    Dim.3   cos2 v.test
## sex.male           0.143  0.042  3.651 |
## sex.famale        -0.070  0.042 -3.651 |
## hs                -0.046  0.008 -0.493 |
## finished hs       -0.020  0.003 -0.242 |
## some college      -0.057  0.021 -1.857 |
## college graduate   0.116  0.071  1.703 |
## graduate degree    0.224  0.147  3.423 |
## unknown           -0.235  0.031 -2.516 |
```

```
fviz_screeplot(
  res.pca,
  barfill = "darkslateblue",
  barcolor = "darkslateblue",
  linecolor = "red",
  ggtheme = theme_gray())
```



Seguint estrictament els criteris de Kaiser, hem de conservar tants eixos com valors propis superiors a 1,0 (valor mitjà del valor propi). 6 eixos compleixen la condició i expliquen el 57,25% de la inèrcia total.

8. Try to explain the meaning of the axes in the first factorial plane. Which 3 variables have the greatest correlation with each factor in the first factorial plane?.

```
summary(res.pca,nb.dec=2,nbind=0,nbelements = 25,ncp=2)
```

```
##
## Call:
## PCA(X = df[, c(1:25, 28:30)], quanti.sup = 26, quali.sup = 27:28)
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8
## Variance                5.05   2.76   2.11   1.83   1.54   1.11   0.85   0.81
## % of var.              20.19  11.03   8.45   7.33   6.16   4.42   3.38   3.23
## Cumulative % of var.   20.19  31.21  39.66  46.99  53.15  57.57  60.96  64.19
##                        Dim.9 Dim.10 Dim.11 Dim.12 Dim.13 Dim.14 Dim.15 Dim.16
## Variance                0.73   0.70   0.68   0.66   0.63   0.59   0.56   0.54
## % of var.               2.92   2.78   2.72   2.63   2.51   2.38   2.24   2.16
## Cumulative % of var.   67.11  69.89  72.61  75.23  77.74  80.12  82.36  84.52
##                       Dim.17 Dim.18 Dim.19 Dim.20 Dim.21 Dim.22 Dim.23 Dim.24
## Variance                0.52   0.50   0.49   0.46   0.43   0.41   0.41   0.38
## % of var.               2.09   1.99   1.95   1.82   1.73   1.63   1.62   1.54
## Cumulative % of var.   86.61  88.61  90.56  92.38  94.11  95.74  97.36  98.90
##                       Dim.25
## Variance                0.27
## % of var.               1.10
```

```
## Cumulative % of var. 100.00
##
## Variables
##                    Dim.1   ctr  cos2   Dim.2   ctr  cos2
## A1                | -0.25  1.27  0.06 | -0.01  0.00  0.00 |
## A2                |  0.49  4.80  0.24 |  0.32  3.62  0.10 |
## A3                |  0.55  6.09  0.31 |  0.32  3.65  0.10 |
## A4                |  0.44  3.77  0.19 |  0.13  0.65  0.02 |
## A5                |  0.61  7.28  0.37 |  0.20  1.50  0.04 |
## C1                |  0.36  2.52  0.13 |  0.14  0.68  0.02 |
## C2                |  0.35  2.48  0.13 |  0.22  1.71  0.05 |
## C3                |  0.34  2.30  0.12 |  0.07  0.20  0.01 |
## C4                | -0.49  4.73  0.24 |  0.15  0.78  0.02 |
## C5                | -0.52  5.42  0.27 |  0.19  1.30  0.04 |
## E1                | -0.44  3.85  0.19 | -0.20  1.48  0.04 |
## E2                | -0.63  7.96  0.40 | -0.02  0.02  0.00 |
## E3                |  0.56  6.32  0.32 |  0.36  4.62  0.13 |
## E4                |  0.61  7.30  0.37 |  0.19  1.31  0.04 |
## E5                |  0.55  6.09  0.31 |  0.31  3.46  0.10 |
## N1                | -0.43  3.73  0.19 |  0.65 15.12  0.42 |
## N2                | -0.43  3.60  0.18 |  0.65 15.26  0.42 |
## N3                | -0.42  3.43  0.17 |  0.67 16.44  0.45 |
## N4                | -0.54  5.88  0.30 |  0.47  8.08  0.22 |
## N5                | -0.37  2.68  0.14 |  0.51  9.45  0.26 |
## O1                |  0.36  2.64  0.13 |  0.25  2.21  0.06 |
## O2                | -0.22  0.97  0.05 |  0.11  0.46  0.01 |
## O3                |  0.44  3.75  0.19 |  0.32  3.62  0.10 |
## O4                | -0.06  0.07  0.00 |  0.34  4.28  0.12 |
## O5                | -0.23  1.09  0.05 | -0.05  0.09  0.00 |
##
## Supplementary continuous variable
##                    Dim.1  cos2   Dim.2  cos2
## age               |  0.17  0.03 | -0.04  0.00 |
##
## Supplementary categories
##                     Dist   Dim.1  cos2 v.test   Dim.2  cos2 v.test
## sex.male          |  0.70 | -0.26  0.14  -4.30 | -0.39  0.31  -8.64 |
## sex.famale        |  0.34 |  0.13  0.14   4.30 |  0.19  0.31   8.64 |
## hs                |  0.50 | -0.42  0.70  -2.93 | -0.04  0.01  -0.35 |
## finished hs       |  0.34 | -0.04  0.01  -0.29 |  0.14  0.17   1.53 |
## some college      |  0.39 |  0.25  0.40   5.21 |  0.03  0.01   0.88 |
## college graduate  |  0.43 | -0.07  0.03  -0.67 | -0.17  0.16  -2.26 |
## graduate degree   |  0.59 |  0.22  0.14   2.13 |  0.02  0.00   0.30 |
## unknown           |  1.34 | -1.19  0.79  -8.26 | -0.05  0.00  -0.50 |
```

```
ddd<-dimdesc(res.pca,axes=1:2)
ddd$Dim.1
```

```
## $quanti
##     correlation       p.value
## E4   0.60678995 2.581025e-281
## A5   0.60614875 1.444464e-280
## E3   0.56488388 8.228260e-236
## E5   0.55440753 1.984370e-225
## A3   0.55427553 2.667432e-225
```

```
## A2   0.49216762 9.110530e-171
## A4   0.43596883 2.866036e-130
## O3   0.43515011 9.837259e-130
## O1   0.36465680  8.358251e-89
## C1   0.35678688  7.893779e-85
## C2   0.35395304  2.001486e-83
## C3   0.34087539  3.950740e-77
## age  0.16553197  1.185144e-18
## O4  -0.06048571  1.364268e-03
## O2  -0.22069220  3.123673e-32
## O5  -0.23405023  3.779005e-36
## A1  -0.25289736  4.131111e-42
## N5  -0.36782900  1.940469e-90
## N3  -0.41595283 1.369594e-117
## N2  -0.42605611 6.978760e-124
## N1  -0.43371160 8.518441e-129
## E1  -0.44057379 2.613257e-133
## C4  -0.48845621 7.481154e-168
## C5  -0.52321855 1.295696e-196
## N4  -0.54453581 5.596826e-216
## E2  -0.63367986 2.742826e-314
##
## $quali
##                   R2      p.value
## f.educ   0.032157174 3.459080e-18
## f.gender 0.006613973 1.642898e-05
##
## $category
##                           Estimate      p.value
## f.educ=some college      0.4562950 1.729604e-07
## f.gender=sex.famale      0.1945324 1.642898e-05
## f.educ=graduate degree   0.4255240 3.308024e-02
## f.educ=hs               -0.2123329 3.374662e-03
## f.gender=sex.male       -0.1945324 1.642898e-05
## f.educ=unknown          -0.9820438 1.005788e-16
##
## attr(,"class")
## [1] "condes" "list "
```

```
ddd$Dim.2
```

```
## $quanti
##     correlation       p.value
## N3   0.67310604  0.000000e+00
## N2   0.64853273  0.000000e+00
## N1   0.64561243  0.000000e+00
## N5   0.51041261 1.251271e-185
## N4   0.47178460 3.396083e-155
## E3   0.35701147  6.100910e-85
## O4   0.34346948  2.353798e-78
## A3   0.31715546  1.837812e-66
## O3   0.31599938  5.756314e-66
## A2   0.31591431  6.259562e-66
## E5   0.30865970  7.195507e-63
## O1   0.24690220  3.699901e-40
```

```
## C2   0.21698706  3.441627e-31
## A5   0.20367439  1.330497e-27
## E4   0.19001736  3.566309e-24
## C5   0.18928569  5.355159e-24
## C4   0.14666526  6.213252e-15
## C1   0.13738225  2.856760e-13
## A4   0.13377746  1.180392e-12
## O2   0.11271951  2.218237e-09
## C3   0.07440298  8.123050e-05
## age -0.04056719  3.183001e-02
## O5  -0.04936074  8.992190e-03
## E1  -0.20207827  3.449439e-27
##
## $quali
##                   R2      p.value
## f.gender 0.02667282 3.420504e-18
##
## $category
##                            Estimate      p.value
## f.gender=sex.famale       0.2887232 3.420504e-18
## f.educ=college graduate -0.1630700 2.408087e-02
## f.gender=sex.male       -0.2887232 3.420504e-18
##
## attr(,"class")
## [1] "condes" "list "
```

Eix 1: * És difícil resumir-lo, però la correlació positiva amb l'eix 1 és + E4 - Make friends easily. + A5 - Make people feel at ease. * i negativa correlacionada amb * E2 - Find it difficult to approach others * N4 - Often feel blue.

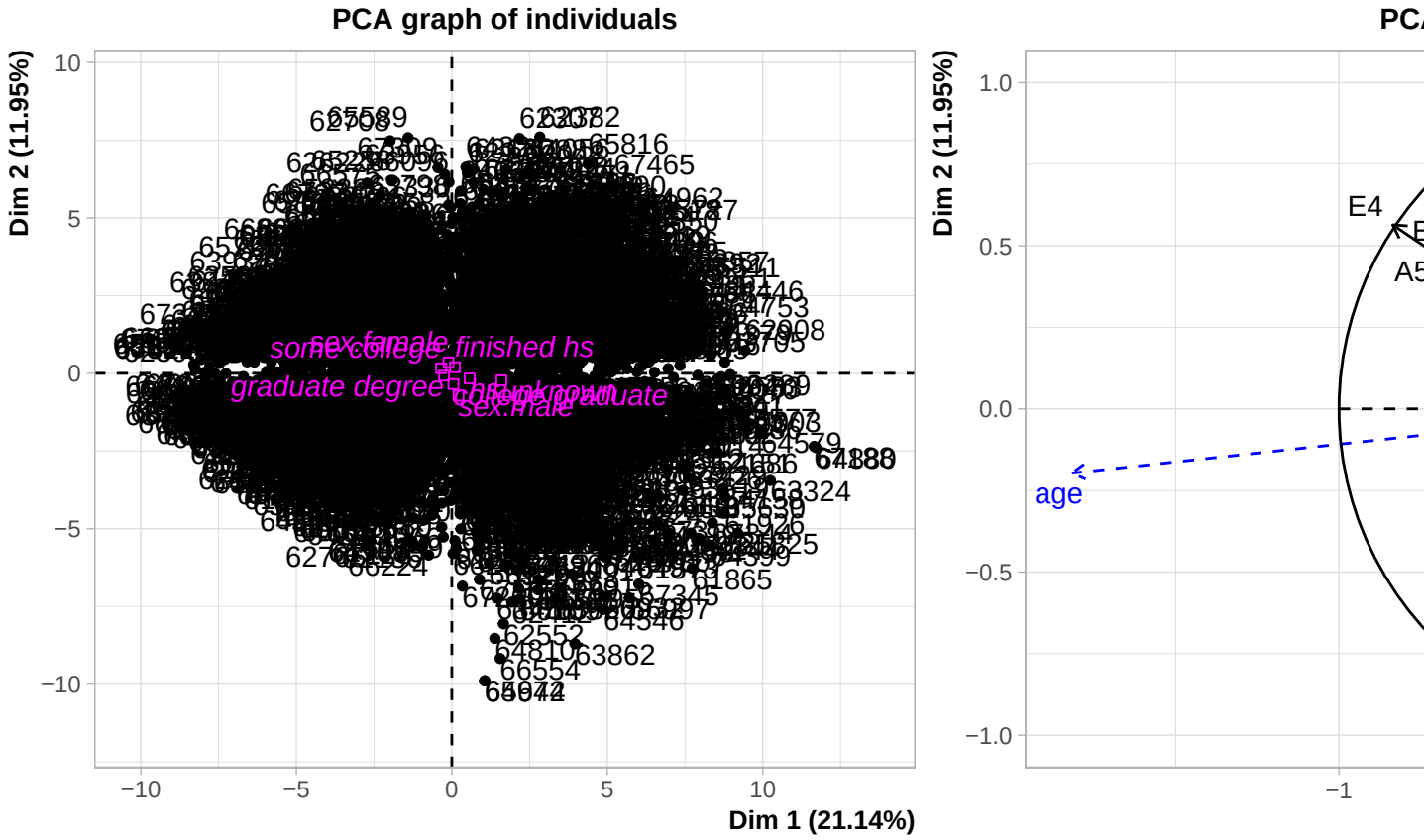
Eix 2: * Per a l'eix 2, apareix una correlació positiva per a + N3 - Have frequent mood swings. + N2 - Get irritated easily. + N1 - Get angry easily. * s'associa inversament a + E1 - Don't talk a lot.

9. A Non-normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age. How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?

```
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"        "C1"
##  [7] "C2"        "C3"        "C4"        "C5"        "E1"        "E2"
## [13] "E3"        "E4"        "E5"        "N1"        "N2"        "N3"
## [19] "N4"        "N5"        "O1"        "O2"        "O3"        "O4"
## [25] "O5"        "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
res.pcann <- PCA(df[, c(1:25, 28:30)], quanti.sup=26, quali.sup=27:28, scale.unit=FALSE)
```




```
summary(res.pcann,nb.dec=2,nbind=0,ncp=2,nbelements = 25)
```

```
##
## Call:
## PCA(X = df[, c(1:25, 28:30)], scale.unit = FALSE, quanti.sup = 26,
##      quali.sup = 27:28)
##
##
## Eigenvalues
##                        Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8
## Variance               10.55   5.97   4.06   3.50   3.02   2.15   1.80   1.75
## % of var.              21.14  11.95   8.14   7.02   6.04   4.31   3.60   3.50
## Cumulative % of var.   21.14  33.09  41.23  48.25  54.29  58.61  62.21  65.72
##                        Dim.9 Dim.10 Dim.11 Dim.12 Dim.13 Dim.14 Dim.15 Dim.16
## Variance               1.50   1.43   1.34   1.22   1.17   1.14   1.09   0.99
## % of var.              3.01   2.86   2.68   2.43   2.34   2.28   2.19   1.98
## Cumulative % of var.  68.72  71.58  74.26  76.70  79.04  81.32  83.51  85.48
##                       Dim.17 Dim.18 Dim.19 Dim.20 Dim.21 Dim.22 Dim.23 Dim.24
## Variance               0.98   0.93   0.89   0.84   0.80   0.76   0.74   0.69
## % of var.              1.96   1.85   1.78   1.68   1.60   1.52   1.48   1.38
## Cumulative % of var.  87.44  89.30  91.08  92.76  94.37  95.89  97.36  98.74
##                       Dim.25
## Variance               0.63
## % of var.              1.26
## Cumulative % of var. 100.00
##
## Variables
```

```
##                   Dim.1   ctr  cos2   Dim.2   ctr  cos2
## A1              |  0.33  1.02  0.05 | -0.06  0.05  0.00 |
## A2              | -0.46  1.98  0.15 |  0.43  3.16  0.14 |
## A3              | -0.60  3.37  0.21 |  0.53  4.64  0.16 |
## A4              | -0.58  3.23  0.16 |  0.36  2.20  0.06 |
## A5              | -0.67  4.23  0.28 |  0.40  2.64  0.10 |
## C1              | -0.36  1.24  0.09 |  0.19  0.62  0.02 |
## C2              | -0.36  1.22  0.07 |  0.32  1.67  0.06 |
## C3              | -0.37  1.32  0.08 |  0.15  0.39  0.01 |
## C4              |  0.67  4.24  0.24 |  0.06  0.06  0.00 |
## C5              |  0.91  7.76  0.31 |  0.08  0.12  0.00 |
## E1              |  0.70  4.66  0.19 | -0.64  6.89  0.16 |
## E2              |  1.05 10.36  0.43 | -0.40  2.62  0.06 |
## E3              | -0.65  3.95  0.23 |  0.61  6.31  0.21 |
## E4              | -0.84  6.61  0.33 |  0.56  5.33  0.15 |
## E5              | -0.63  3.79  0.23 |  0.55  5.03  0.17 |
## N1              |  0.86  7.01  0.30 |  0.90 13.67  0.33 |
## N2              |  0.83  6.51  0.30 |  0.86 12.27  0.32 |
## N3              |  0.88  7.31  0.30 |  0.94 14.84  0.35 |
## N4              |  1.01  9.72  0.42 |  0.48  3.79  0.09 |
## N5              |  0.77  5.59  0.23 |  0.72  8.65  0.20 |
## O1              | -0.32  0.96  0.08 |  0.25  1.01  0.05 |
## O2              |  0.37  1.27  0.05 |  0.19  0.63  0.02 |
## O3              | -0.42  1.67  0.12 |  0.38  2.43  0.10 |
## O4              |  0.17  0.27  0.02 |  0.24  0.96  0.04 |
## O5              |  0.27  0.69  0.04 | -0.03  0.02  0.00 |
##
## Supplementary continuous variable
##                   Dim.1  cos2   Dim.2  cos2
## age             | -1.82  0.03 | -0.20  0.00 |
##
## Supplementary categories
##                      Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## sex.male         |   1.00 |   0.22   0.05   2.46 |  -0.72   0.51 -10.86 |
## sex.famale       |   0.49 |  -0.11   0.05  -2.46 |   0.35   0.51  10.86 |
## hs               |   0.70 |   0.57   0.66   2.75 |  -0.17   0.06  -1.06 |
## finished hs      |   0.47 |   0.10   0.05   0.57 |   0.19   0.17   1.44 |
## some college     |   0.56 |  -0.34   0.37  -5.01 |   0.16   0.08   3.10 |
## college graduate |   0.62 |   0.05   0.01   0.34 |  -0.35   0.31  -3.04 |
## graduate degree  |   0.81 |  -0.25   0.10  -1.71 |  -0.08   0.01  -0.68 |
## unknown          |   1.86 |   1.59   0.74   7.63 |  -0.23   0.01  -1.44 |
```

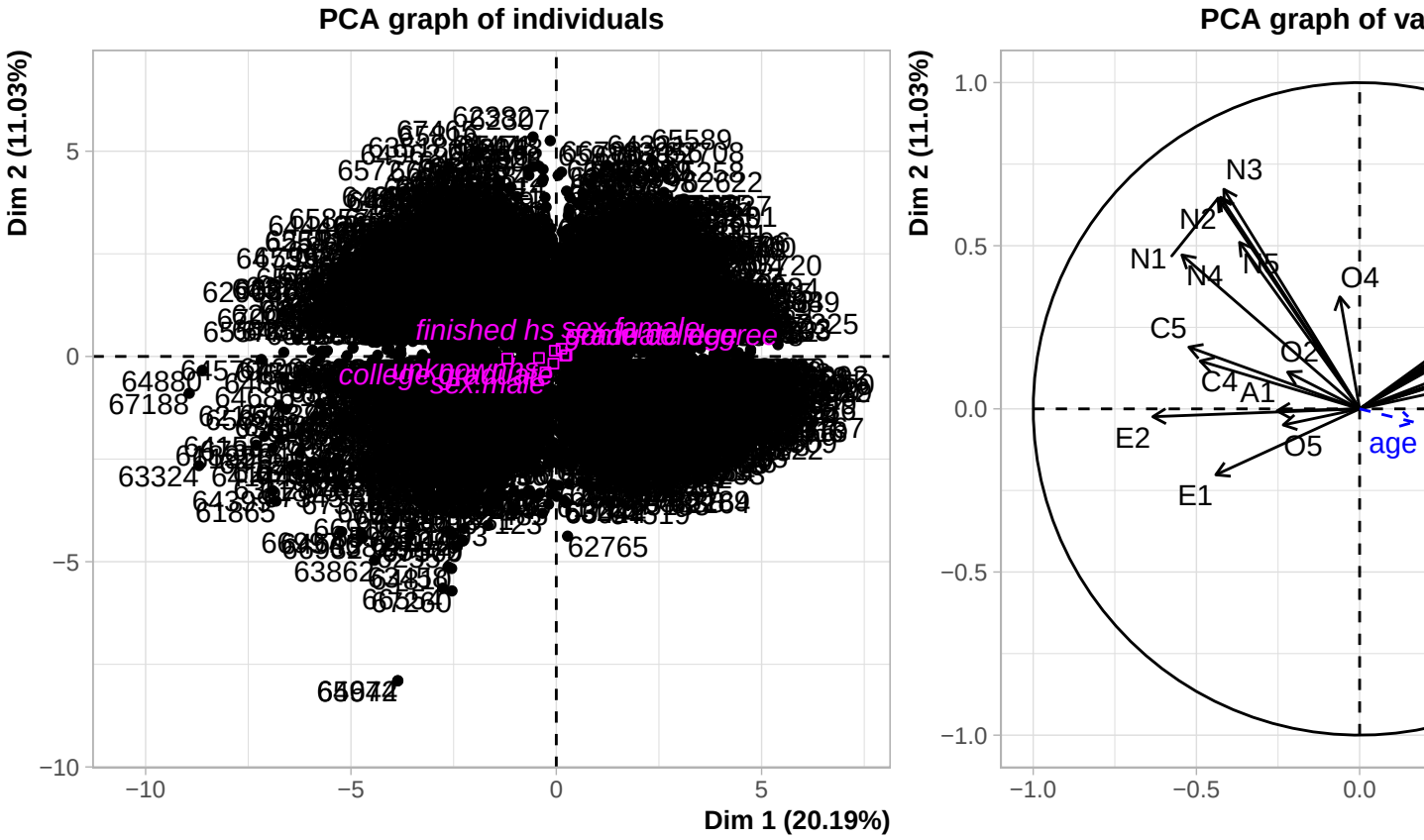
```
mean(res.pcann$eig[,1])
```

```
## [1] 1.996997
```

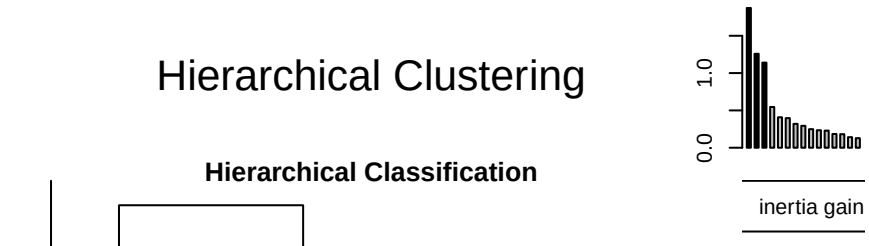
- Seguint estrictament els criteris de Kaiser, hem de conservar tants eixos com a valors propis superiors a la mitjana (> mean(res.pcann$eig[,1]) [1] 1.996997).
- 6 dimensions compleixen la condició i expliquen el 58.61% de la inèrcia total

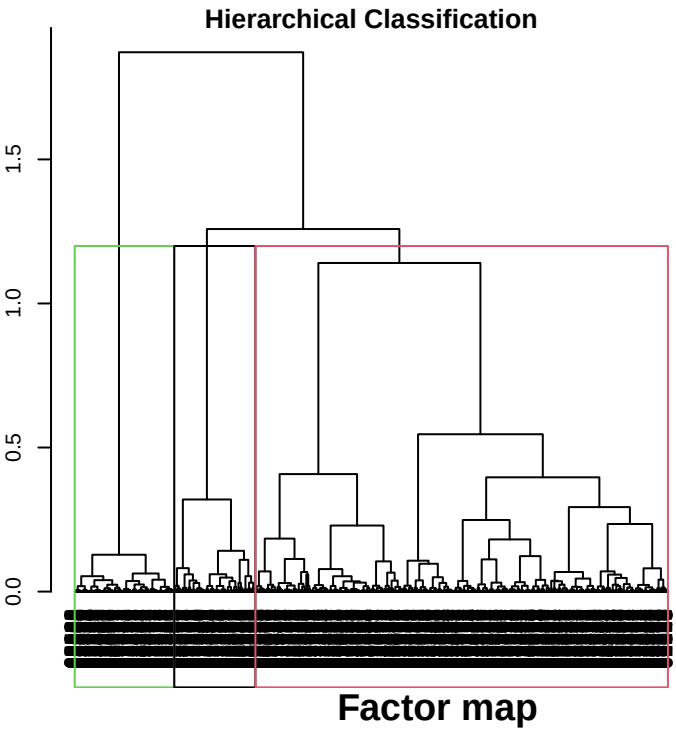
10. A Hierarchical Clustering is addressed. A non-default criteria for selecting the number of clusters to 3 has to be set. Explain the characteristics of cluster number 1.

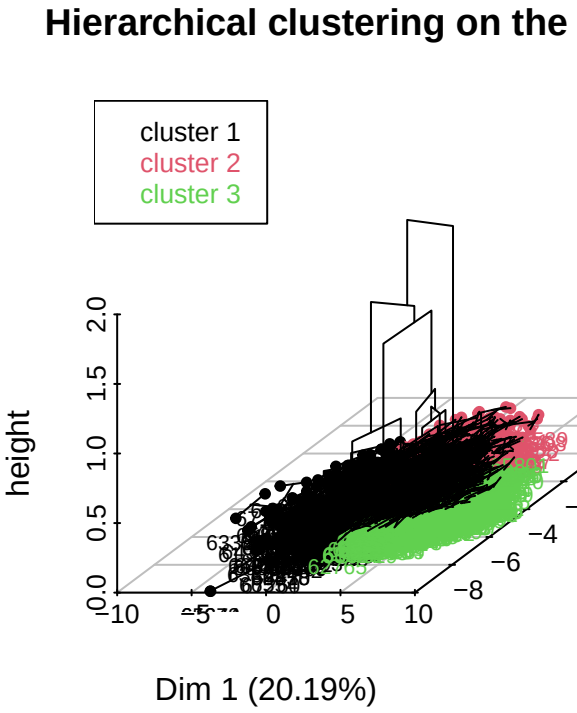
```
# 6 dimensions have to be selected according to Kaiser's criteria
res.pca <- PCA(df[, c(1:25, 28:30)], quanti.sup=26, quali.sup=27:28, ncp=6)
```

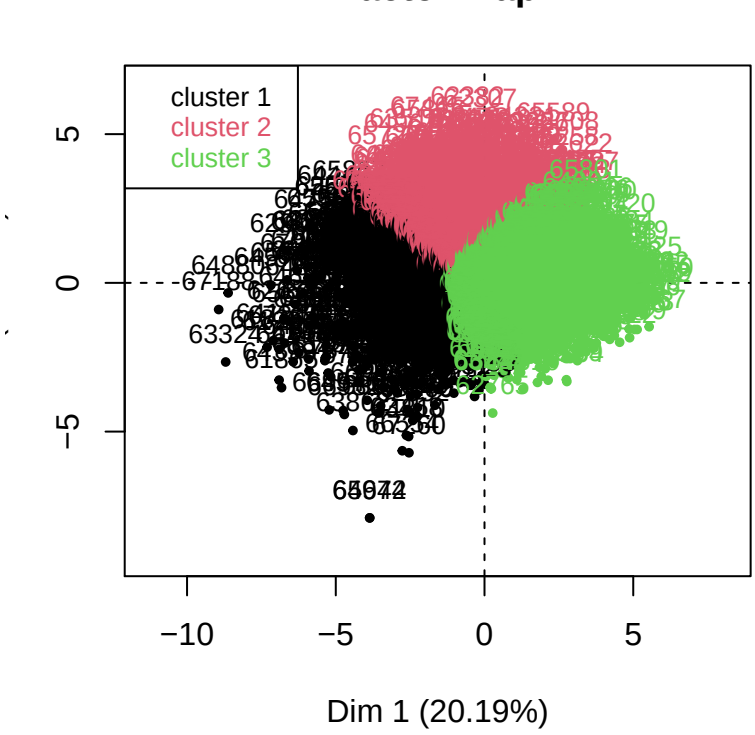



```
res.hcpc <- HCPC(res.pca,nb.clust=3, graph=T)
```










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

```
##
```

```
## Link between the cluster variable and the categorical variables (chi-square test)
## ================================================================================
##               p.value df
## f.gender 3.312144e-13  2
## f.educ   1.068090e-11 10
##
## Description of each cluster by the categories
## =============================================
## $`1`
##                            Cla/Mod  Mod/Cla    Global      p.value    v.test
## f.gender=sex.male       38.95539 42.56837 32.821429 1.127038e-12  7.114029
## f.educ=unknown          47.08520 12.48514  7.964286 2.443182e-08  5.577273
## f.educ=college graduate 35.27919 16.52794 14.071429 1.552311e-02  2.419939
## f.educ=some college     25.78062 38.28775 44.607143 9.899771e-06 -4.419351
## f.gender=sex.famale     25.67783 57.43163 67.178571 1.127038e-12 -7.114029
##
## $`2`
##                        Cla/Mod  Mod/Cla   Global      p.value    v.test
## f.gender=sex.famale 33.70548 74.23888 67.17857 9.841918e-08  5.329619
## f.gender=sex.male   23.93906 25.76112 32.82143 9.841918e-08 -5.329619
##
## $`3`
##                        Cla/Mod   Mod/Cla    Global      p.value    v.test
## f.educ=some college 43.23459 48.868778 44.607143 2.540649e-04  3.658127
## f.educ=unknown      18.38565  3.710407  7.964286 2.211532e-12 -7.020451
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##           Eta2       P-value
## N3  0.34794344 1.896487e-260
## N2  0.32690687 3.656861e-241
## N1  0.32517673 1.325205e-239
## N4  0.30524376 6.348819e-222
## E3  0.28986110 1.269430e-208
## E4  0.28984909 1.299822e-208
## A5  0.27027465 4.250732e-192
## A3  0.26709834 1.846805e-189
## E2  0.25355681 2.432030e-178
## E5  0.25004768 1.716174e-175
## N5  0.20985875 8.711098e-144
## A2  0.20968042 1.194351e-143
## C5  0.18780176 4.595871e-127
## E1  0.18239233 4.944495e-123
## O3  0.15285565 1.771950e-101
## C4  0.14367533  6.235599e-95
## A4  0.14228348  6.043242e-94
## O1  0.09693753  1.178511e-62
## C2  0.07587939  1.179038e-48
## C1  0.06591332  3.858949e-42
## C3  0.06002771  2.519995e-38
## A1  0.04145500  1.927913e-26
## O2  0.03388765  1.151185e-21
## O4  0.03301271  4.082666e-21
```

```
## O5  0.02860825  2.350089e-18
## age 0.01712806  3.213398e-11
##
## Description of each cluster by quantitative variables
## ======================================================
## $`1`
##         v.test Mean in category Overall mean sd in category Overall sd
## E2   23.559798         4.231451     3.142382       1.370509   1.602382
## E1   22.495063         4.029885     2.974726       1.493634   1.625971
## C5   14.556178         3.978066     3.295928       1.472660   1.624451
## C4   13.416473         3.082821     2.552636       1.346132   1.369844
## N4   11.908391         3.719900     3.183149       1.452022   1.562437
## A1    8.444417         2.755769     2.413622       1.337729   1.404512
## O5    7.395481         2.774659     2.492212       1.329898   1.323893
## O2    3.903164         2.889417     2.713214       1.575193   1.564872
## N2    3.553650         3.663794     3.507751       1.371753   1.522135
## N5    2.971998         3.107332     2.969069       1.526378   1.612638
## N1    2.943622         3.065574     2.932482       1.480823   1.567304
## N3    2.293638         3.322225     3.216355       1.507437   1.600029
## O4   -1.967554         4.822770     4.891937       1.255672   1.218579
## age  -4.474584        27.346017    28.782143      11.033637  11.125568
## C3  -12.062330         3.855724     4.302816       1.320750   1.284835
## C1  -13.293589         4.028014     4.502523       1.311110   1.237326
## C2  -14.565247         3.819361     4.371579       1.387430   1.314241
## O1  -16.300620         4.286250     4.815602       1.269402   1.125699
## A4  -19.254606         3.881188     4.700615       1.583338   1.475224
## O3  -20.670132         3.710323     4.435417       1.326453   1.215998
## A2  -24.187988         3.989373     4.804418       1.220805   1.168056
## E5  -26.394368         3.403162     4.417105       1.344304   1.331633
## A5  -26.445666         3.603445     4.561126       1.213214   1.255304
## A3  -27.258020         3.584852     4.604916       1.312151   1.297226
## E4  -27.827839         3.252817     4.421325       1.415591   1.455575
## E3  -28.348423         2.898245     4.000863       1.217943   1.348274
##           p.value
## E2  9.962274e-123
## E1  4.639204e-112
## C5   5.335860e-48
## C4   4.842147e-41
## N4   1.070227e-32
## A1   3.055658e-17
## O5   1.408974e-13
## O2   9.494331e-05
## N2   3.799248e-04
## N5   2.958691e-03
## N1   3.243963e-03
## N3   2.181129e-02
## O4   4.911937e-02
## age  7.656023e-06
## C3   1.669899e-33
## C1   2.521852e-40
## C2   4.672899e-48
## O1   9.769551e-60
## A4   1.291540e-82
## O3   6.433975e-95
```

```
## A2  2.976486e-129
## E5  1.590115e-153
## A5  4.092538e-154
## A3  1.335163e-163
## E4  1.997804e-170
## E3  8.753067e-177
##
## $`2`
##        v.test Mean in category Overall mean sd in category Overall sd
## N3  27.042714         4.450935     3.216355      1.2158923   1.600029
## N1  25.771594         4.084968     2.932482      1.3471741   1.567304
## N2  25.519998         4.616096     3.507751      1.1031618   1.522135
## N5  20.379386         3.906780     2.969069      1.5459410   1.612638
## N4  18.879488         4.024803     3.183149      1.3164928   1.562437
## A3   9.897311         4.971247     4.604916      1.0100649   1.297226
## E5   9.839652         4.790962     4.417105      1.0674726   1.331633
## E3   9.806655         4.378122     4.000863      1.1077085   1.348274
## C5   9.639989         3.742740     3.295928      1.5660191   1.624451
## O4   9.330579         5.216354     4.891937      0.9990437   1.218579
## A2   9.277820         5.113626     4.804418      0.9106669   1.168056
## O3   8.280681         4.722719     4.435417      1.0238368   1.215998
## C4   7.603949         2.849838     2.552636      1.4078531   1.369844
## E4   6.606221         4.695690     4.421325      1.2485755   1.455575
## O2   6.344319         2.996487     2.713214      1.6464387   1.564872
## C2   5.880706         4.592098     4.371579      1.2444415   1.314241
## O1   4.939957         4.974269     4.815602      1.0454416   1.125699
## A5   4.669653         4.728379     4.561126      1.1173231   1.255304
## A4   3.632132         4.853499     4.700615      1.4080868   1.475224
## C1   3.258556         4.617563     4.502523      1.1407284   1.237326
## A1   2.359527         2.508178     2.413622      1.4905917   1.404512
## age -2.818019        27.887588    28.782143     10.2487338  11.125568
## E1  -7.855368         2.610291     2.974726      1.5127232   1.625971
##           p.value
## N3  4.652628e-161
## N1  1.846560e-146
## N2  1.182615e-143
## N5   2.548333e-92
## N4   1.682076e-79
## A3   4.276160e-23
## E5   7.597302e-23
## E3   1.054045e-22
## C5   5.419379e-22
## O4   1.052937e-20
## A2   1.729822e-20
## O3   1.224733e-16
## C4   2.872291e-14
## E4   3.942542e-11
## O2   2.234121e-10
## C2   4.085209e-09
## O1   7.813977e-07
## A5   3.017092e-06
## A4   2.810897e-04
## C1   1.119809e-03
## A1   1.829825e-02
```

```
## age  4.832101e-03
## E1   3.986016e-15
##
## $`3`
##         v.test Mean in category Overall mean sd in category Overall sd
## A5   20.404327         5.160742     4.561126      0.9103214   1.255304
## E4   19.876467         5.098618     4.421325      1.0404725   1.455575
## E3   17.350021         4.548484     4.000863      1.0926233   1.348274
## A3   16.241955         5.098154     4.604916      1.0130562   1.297226
## E5   15.486262         4.899867     4.417105      1.0553174   1.331633
## A4   14.637253         5.206114     4.700615      1.1356020   1.475224
## A2   13.946159         5.185765     4.804418      0.9782905   1.168056
## O3   11.586087         4.765233     4.435417      1.0123441   1.215998
## O1   10.634832         5.095858     4.815602      0.9103020   1.125699
## C3   10.408164         4.615873     4.302816      1.1301221   1.284835
## C1    9.398405         4.774755     4.502523      1.1444196   1.237326
## C2    8.121102         4.621436     4.371579      1.1758689   1.314241
## age   6.851114        30.566516    28.782143     11.5948717  11.125568
## O4   -6.943733         4.693853     4.891937      1.2911718   1.218579
## O5   -8.187669         2.238457     2.492212      1.2027500   1.323893
## O2   -9.636836         2.360181     2.713214      1.4182546   1.564872
## A1  -10.142475         2.080140     2.413622      1.3085159   1.404512
## E1  -13.698290         2.453314     2.974726      1.4180326   1.625971
## C4  -19.745745         1.919427     2.552636      1.0773754   1.369844
## N5  -21.984040         2.139130     2.969069      1.2595931   1.612638
## C5  -22.732534         2.431445     3.295928      1.3729361   1.624451
## E2  -23.016574         2.278988     3.142382      1.2880222   1.602382
## N1  -27.036694         1.940488     2.932482      1.0624731   1.567304
## N2  -27.371835         2.532403     3.507751      1.2565698   1.522135
## N3  -27.624434         2.181634     3.216355      1.1652980   1.600029
## N4  -28.952492         2.124162     3.183149      1.1634226   1.562437
##           p.value
## A5   1.530530e-92
## E4   6.505630e-88
## E3   1.971709e-67
## A3   2.546782e-59
## E5   4.295498e-54
## A4   1.625020e-48
## A2   3.320103e-44
## O3   4.847792e-31
## O1   2.051979e-26
## C3   2.275693e-25
## C1   5.539657e-21
## C2   4.619706e-16
## age  7.327703e-12
## O4   3.818712e-12
## O5   2.663340e-16
## O2   5.588367e-22
## A1   3.579127e-24
## E1   1.039431e-42
## C4   8.726790e-87
## N5  4.093648e-107
## C5  2.136449e-114
## E2  3.181096e-117
```

```
## N1  5.476420e-161
## N2  5.937160e-165
## N3  5.662303e-168
## N4  2.610678e-184
```

```
(res.hcpc$call$t$within[1]-res.hcpc$call$t$within[3])/res.hcpc$call$t$within[1]
```

```
## [1] 0.217464
```

Es seleccionen tres clústers que són suficients per representar la complexitat d'aquest conjunt de dades.

- Clúster 1:
  - E2 - Find it difficult to approach others.
  - E1 - Don't talk a lot.
  - C5 - Waste my time
  - C4 - Do things in a half-way manner.
- Clúster 2:
  - N3 - Have frequent mood swings.
  - N1 - Get angry easily.
  - N2 - Get irritated easily.
  - N5 - Panic easily.
- Clúster 3:
  - A5 - Make people feel at ease.
  - E4 - Make friends easily.
  - A2 - Inquire about others' well-being.
  - O3 - Carry the conversation to a higher level.