QUIZ1 ADEI_1920Q2: Solutions to questions

L. Montero

March 30th, 2020

List of Questions

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. Available variables:

- 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

Source: The items are from the ipip (Goldberg, 1999). The data are from the SAPA project (Revelle, Wilt and Rosenthal, 2010) , collected Spring, 2010 (https://sapa-project.org).

References: 1. Goldberg, L.R. (1999) A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I. and Deary, I. and De Fruyt, F. and Ostendorf, F. (eds) Personality psychology in Europe. 7. Tilburg University Press. Tilburg, The Netherlands. 2. Revelle, W., Wilt, J., and Rosenthal, A. (2010) Individual Differences in Cognition: New Methods

for examining the Personality-Cognition Link In Gruszka, A. and Matthews, G. and Szymura, B. (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer. 3. Revelle, W, Condon, D.M., Wilt, J., French, J.A., Brown, A., and Elleman, L.G. (2016) Web and phone based data collection using planned missing designs. In Fielding, N.G., Lee, R.M. and Blank, G. (Eds). SAGE Handbook of Online Research Methods (2nd Ed), Sage Publications.

Firstly, load dataset and check available variables.

```r
rm(list=ls())
setwd("C:/Users/lidia/Dropbox/DOCENCIA/FIB-ADEI/EXAMENS/1920Q2")
load("C:/Users/lidia/Dropbox/DOCENCIA/FIB-ADEI/EXAMENS/1920Q2/bfi_Raw.RData")
summary(df)
```

```
##        A1              A2              A3              A4
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0
##  1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.0
##  Median :2.000   Median :5.000   Median :5.000   Median :5.0
##  Mean   :2.413   Mean   :4.802   Mean   :4.604   Mean   :4.7
##  3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.0
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.0
##  NA's   :16      NA's   :27      NA's   :26      NA's   :19
##        A5             C1              C2             C3
##  Min.   :1.00   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:4.00   1st Qu.:4.000   1st Qu.:4.00   1st Qu.:4.000
##  Median :5.00   Median :5.000   Median :5.00   Median :5.000
##  Mean   :4.56   Mean   :4.502   Mean   :4.37   Mean   :4.304
##  3rd Qu.:5.00   3rd Qu.:5.000   3rd Qu.:5.00   3rd Qu.:5.000
##  Max.   :6.00   Max.   :6.000   Max.   :6.00   Max.   :6.000
##  NA's   :16     NA's   :21      NA's   :24     NA's   :20
##        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   :3.297   Mean   :2.974   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
##  NA's   :26      NA's   :16      NA's   :23      NA's   :16
##        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   :4.422   Mean   :4.416   Mean   :2.929
##  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
##  NA's   :25      NA's   :9       NA's   :21      NA's   :22
##        N2              N3              N4              N5
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00
##  Median :4.000   Median :3.000   Median :3.000   Median :3.00
##  Mean   :3.508   Mean   :3.217   Mean   :3.186   Mean   :2.97
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :6.000   Max.   :6.000   Max.   :6.000   Max.   :6.00
##  NA's   :21      NA's   :11      NA's   :36      NA's   :29
##        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   :4.438   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.   :6.000
##  NA's   :22                      NA's   :28      NA's   :14
##       O5           gender        education          age
##  Min.   :1.00   Min.   :1.000   Min.   :1.00   Min.   : 3.00
##  1st Qu.:1.00   1st Qu.:1.000   1st Qu.:3.00   1st Qu.:20.00
##  Median :2.00   Median :2.000   Median :3.00   Median :26.00
##  Mean   :2.49   Mean   :1.672   Mean   :3.19   Mean   :28.78
##  3rd Qu.:3.00   3rd Qu.:2.000   3rd Qu.:4.00   3rd Qu.:35.00
##  Max.   :6.00   Max.   :2.000   Max.   :5.00   Max.   :86.00
##  NA's   :20                     NA's   :223
```

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).

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

```
##   sex.male sex.female
##        919       1881
```

```
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", "gradua
levels(df$educ)
```

```
## NULL
```

```
summary(df$educ)
```

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

```
sumres<-summary(df$age)
iqr<-as.numeric(sumres[5]-sumres[2]);iqr
```

```
## [1] 15
```

```
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
```

```
## [1] -2.5
```

```
## [1] 57.5
```

```
sevlow;sevup
```

```
## [1] -25
```

```
## [1] 80
```

```
ll<-which(df$age>sevup);length(ll);ll
```

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

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

Education can be seen to have 223 missing values. Imputation is not a reasonable solution and an specific level unknown has to be defined. Gender and education are defined as factors. Age is a numeric variable without missing data. Computation of severe outliers thresholds determines that those observations greater than 80 are severe outliers: only 1 person satisfy this condition (obs. 1158). Lower severe threshold does not make sense (since is -25). Follow R commands to figure out the calculus of these thresholds, based on 1.5/3 times Inter Quartilar Range from Q1/Q3. Or check theory slide notes.

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

```r
library(missMDA)
```

```
## Warning: package 'missMDA' was built under R version 3.6.2
```

```r
names(df)
```

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

```r
#summary(df[,c(1:25)])
res.impu<-imputePCA(df[,c(1:25,28)])
dfimpu<-as.data.frame(res.impu$completeObs)

#library(psych)
#describe(df[,1:25])
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.2
```

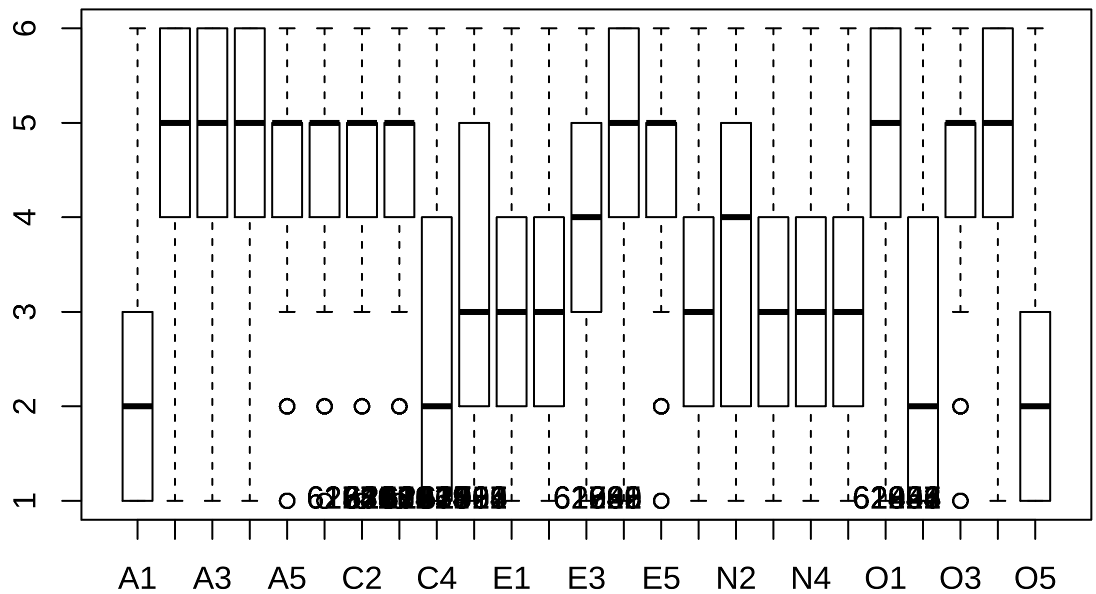
```
## Loading required package: carData
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```r
Boxplot(df[,1:25])
```

```
##  [1] "61629" "61640" "61788" "61840" "61873" "61926" "61932" "62282"
##  [9] "62551" "62552" "61654" "61682" "61761" "61921" "61979" "62038"
## [17] "62060" "62102" "62111" "62498" "61654" "61825" "61839" "61865"
## [25] "61918" "61921" "61969" "61979" "62029" "62079" "61654" "61701"
## [33] "61716" "62022" "62029" "62092" "62526" "62716" "62787" "62795"
## [41] "61629" "61682" "61761" "61788" "61825" "61840" "61865" "61989"
## [49] "62092" "62266" "61856" "61926" "62022" "62054" "62064" "62246"
## [57] "62327" "62328" "62443" "62491"
```

```r
Boxplot(dfimpu[,1:25])
```




```
##  [1] "61629" "61640" "61788" "61840" "61873" "61926" "61932" "62282"
##  [9] "62551" "62552" "61654" "61682" "61761" "61921" "61979" "62038"
## [17] "62060" "62102" "62111" "62498" "61654" "61825" "61839" "61865"
## [25] "61918" "61921" "61969" "61979" "62029" "62079" "61654" "61701"
## [33] "61716" "62022" "62029" "62092" "62526" "62716" "62787" "62795"
## [41] "61629" "61682" "61761" "61788" "61825" "61840" "61865" "61989"
## [49] "62092" "62266" "61856" "61926" "62022" "62054" "62064" "62246"
## [57] "62327" "62328" "62443" "62491"
```

```
lapply(df[,1:25],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
```

```r
lapply(dfimpu[,1:25],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% 
## 1.000000 3.000000 4.000000 4.000000 4.318546 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%  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
```

```r
df[,1:25]<-res.impu$completeObs[,1:25]
#summary(df)
```

All 25 first variables have missing values, between 9 and 36, except variable O2. Method imputePCA() from missMDA package has to be used for imputation of numeric variables. Check for reasonable imputation values has to be done using either graphics or quantiles. No problems seems to be present.

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

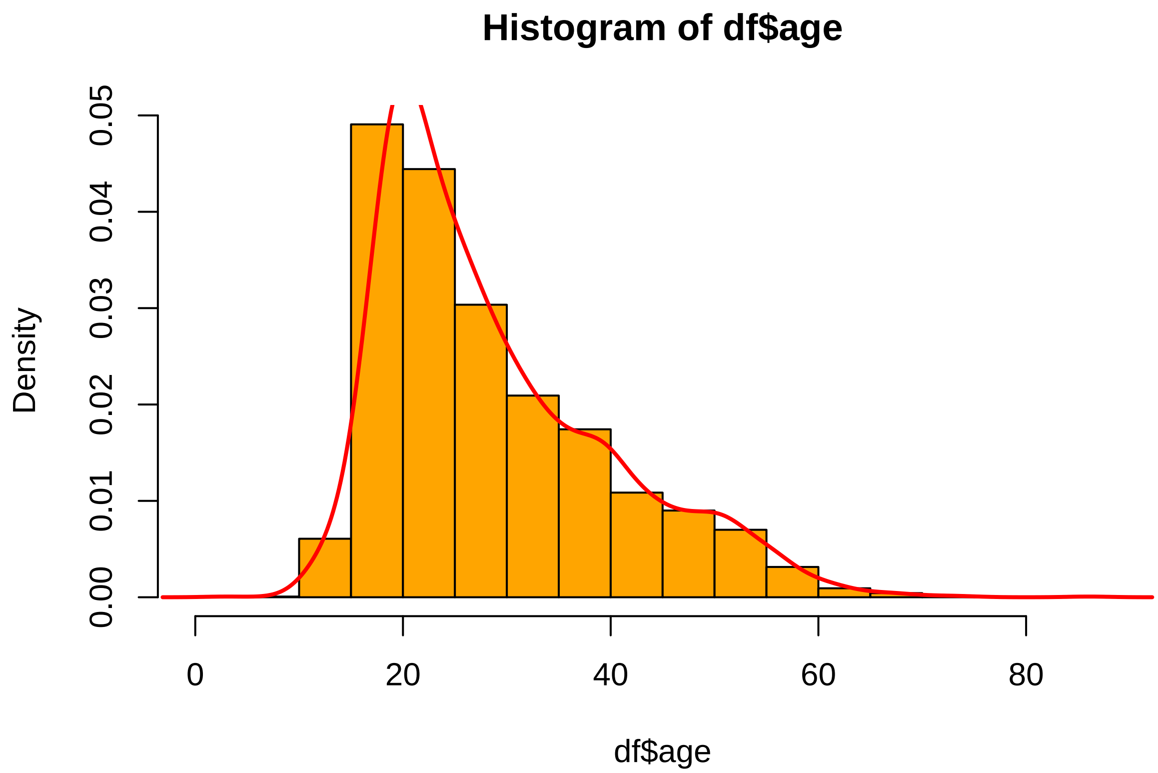
```r
summary(df$f.educ)
```

```
##                HS      finished HS     some college college graduate
##               224              292             1249              394
##   graduate degree          Unknown
##               418              223
```

Imputation for factors would have to use f.educ and f.gender data, so it is not likely that they contain enough information for a suitable imputation. If a set of factors had been included in dataset, then imputeMCA() in missMDA package would have to be used for imputation purposes. Actually, missing values of variable education have to be selected to define a new level in factor f.educ labelled as "Unknown".There are 223 observations with unknown education level.

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.

```r
hist(df$age,15,freq=F,col="orange")
lines(density(df$age),col="red",lwd=2)
```



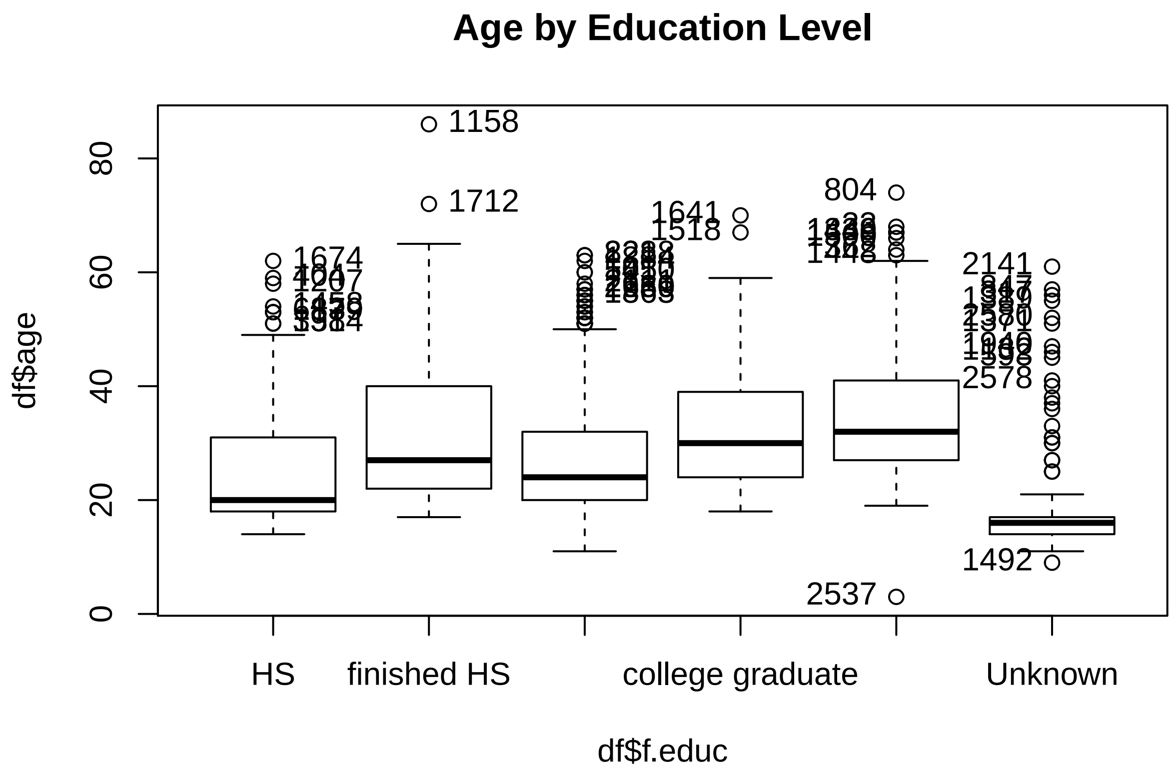

```
shapiro.test(df$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$age
## W = 0.91212, p-value < 2.2e-16
```

```
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"
```

```
tapply(df$age,df$f.gender,summary)
```

```
## $sex.male
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00   20.00   25.00   28.02   34.00   74.00
##
## $sex.female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   20.00   26.00   29.15   36.00   86.00
```

```
tapply(df$age,df$f.educ,summary)
```

```
## $HS
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   18.00   20.00   25.13   31.00   62.00
##
## $`finished HS`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   22.00   27.00   31.51   40.00   86.00
##
## $`some college`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   11.00   20.00   24.00   27.23   32.00   63.00
##
## $`college graduate`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   24.00   30.00   32.98   39.00   70.00
##
## $`graduate degree`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0    27.0    32.0    35.3    41.0    74.0
##
## $Unknown
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   16.00   17.95   17.00   61.00
```

```r
kruskal.test(df$age,df$f.gender)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$age and df$f.gender
## Kruskal-Wallis chi-squared = 7.8848, df = 1, p-value = 0.004985
```

```r
kruskal.test(df$age,df$f.educ)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$age and df$f.educ
## Kruskal-Wallis chi-squared = 733.88, df = 5, p-value < 2.2e-16
```

```r
pairwise.wilcox.test(df$age,df$f.educ,alternative="greater")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  df$age and df$f.educ
##
##                  HS      finished HS some college college graduate
## finished HS      2.4e-14 -           -            -
## some college     3.0e-09 1.0000      -            -
## college graduate < 2e-16 0.0036      < 2e-16      -
## graduate degree  < 2e-16 2.6e-08     < 2e-16      0.0019
## Unknown          1.0000  1.0000      1.0000       1.0000
##                  graduate degree
## finished HS      -
## some college     -
## college graduate -
## graduate degree  -
## Unknown          1.0000
##
## P value adjustment method: holm
```

```r
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 3.6.2
```

```r
res.condes<-condes(df[,c(1:25,28:30)],num.var=26,proba=0.01)
res.condes$quanti
```

```
##     correlation      p.value
## A4  0.14447666 1.567367e-14
## A5  0.12880606 7.850631e-12
## E5  0.11507222 1.017596e-09
## A2  0.11454173 1.214740e-09
## C1  0.07945618 2.562716e-05
## A3  0.06912295 2.518337e-04
## C3  0.06772844 3.353129e-04
## C5 -0.08699401 4.027149e-06
## N1 -0.08780808 3.266923e-06
## O5 -0.09968771 1.251611e-07
## N2 -0.10334852 4.230972e-08
## N5 -0.10391457 3.565737e-08
## E2 -0.10596202 1.906176e-08
## N3 -0.11174301 3.051191e-09
## C4 -0.14765217 4.074830e-15
## A1 -0.16115729 9.489506e-18
```

```r
res.condes$quali
```

```
##                R2      p.value
## f.educ 0.1703955 1.27964e-110
```

```r
res.condes$category
```

```
##                             Estimate      p.value
## f.educ=graduate degree    6.950567 1.064868e-39
## f.educ=college graduate   4.628827 4.548861e-16
## f.educ=finished HS        3.162830 9.046486e-06
## f.educ=HS                -3.216940 2.953164e-07
## f.educ=some college      -1.125088 2.627052e-11
## f.educ=Unknown          -10.400196 5.288444e-54
```

Numeric, graphic and inferential tools have to use to answer this question. First of all a rough assessment of age normality is performed: clearly shape is not symmetric and Shapiro-Wilk test rejects the null hypothesis of normality. Without normality, non-parametric methods have to be used. Summary statistics for groups of age defined according to f.gender/f.educ are not conclusive for f.gender, but differences appear for f.educ levels. Graphics: boxplot of age for f.gender is difficult to assess, but boxplot for each f.educ level show a clear different profile for age depending on the levels. Inferential tools: Null hypothesis for equal means age according to f.gender/f.educ are both rejected (pvalue=5e-03 for gender and pvalue=0 for education factor). Pairwise mean tests for f.educ can be computed and null hypothesis can be rejected (some of them). Null hypothesis can be defined as mean in group i greater (less) than mean in j.

A condes() method can be used for a fast answer: $quali shows global significance of f.educ and f.gender. It also shows graduate degree, college graduate, finished HS and female mean ages are over the mean and males, HS, some college and unknown are significantly under the global mean of age.

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.

```r
names(df)
```

```
##  [1] "A1"        "A2"        "A3"        "A4"        "A5"
```

```
##  [6] "C1"        "C2"        "C3"        "C4"        "C5"
## [11] "E1"        "E2"        "E3"        "E4"        "E5"
## [16] "N1"        "N2"        "N3"        "N4"        "N5"
## [21] "O1"        "O2"        "O3"        "O4"        "O5"
## [26] "gender"    "education" "age"       "f.gender"  "f.educ"
```

```
res.catdes<-catdes(df[,c(1:25,28:30)],num.var=28,proba=0.01)
res.catdes$test.chi2
```

```
##                p.value df
## f.gender 0.0003503056  5
```

```
res.catdes$quanti.var
```

```
##            Eta2       P-value
## age 0.170395545 1.279640e-110
## A4  0.038149120  7.589005e-22
## A1  0.027952813  1.201449e-15
## C4  0.022730429  1.583479e-12
## C5  0.019288964  1.694486e-10
## O3  0.018711574  3.690319e-10
## E4  0.016014108  1.365382e-08
## A2  0.015117854  4.485212e-08
## C1  0.014955221  5.562131e-08
## O2  0.013667891  3.032756e-07
## A5  0.012544809  1.316374e-06
## A3  0.011749053  3.697040e-06
## C2  0.010532117  1.768863e-05
## E5  0.010161184  2.839841e-05
## N1  0.009664280  5.338181e-05
## O5  0.009540961  6.239791e-05
## N4  0.009281960  8.653265e-05
## O4  0.008702755  1.790611e-04
## N2  0.008668381  1.869235e-04
## C3  0.008279506  3.034758e-04
## E2  0.007757499  5.788560e-04
## E3  0.007101257  1.292346e-03
## N3  0.006607187  2.348837e-03
## O1  0.006073490  4.442871e-03
```

Globally associated to f.educ is f.gender factor. f.educ is globally associated to numeric variables age and 24 items more, being the most significance A4, A1, C4, C5 and O3 (all of them showing pvalues less than 1e-09).

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

```
library(FactoMineR)
res.catdes$category
```

```
## $HS
##                       Cla/Mod  Mod/Cla   Global     p.value    v.test
## f.gender=sex.male   10.119695 41.51786 32.82143 0.004525328  2.839013
## f.gender=sex.female  6.964381 58.48214 67.17857 0.004525328 -2.839013
##
## $`finished HS`
## NULL
##
## $`some college`
```

```
##                       Cla/Mod Mod/Cla   Global      p.value    v.test
## f.gender=sex.female 47.47475 71.4972 67.17857 1.205871e-05  4.376524
## f.gender=sex.male   38.73776 28.5028 32.82143 1.205871e-05 -4.376524
##
## $`college graduate`
## NULL
##
## $`graduate degree`
## NULL
##
## $Unknown
## NULL
```

```
res.catdes$quanti
```

```
## $HS
##        v.test Mean in category Overall mean sd in category Overall sd
## C4   2.693481          2.78952     2.553145        1.377601   1.369119
## E5  -3.099918          4.15247     4.417017        1.299913   1.331387
## age -5.115766         25.13393    28.782143       10.375964  11.125568
##          p.value
## C4  7.071015e-03
## E5  1.935742e-03
## age 3.124698e-07
##
## $`finished HS`
##       v.test Mean in category Overall mean sd in category Overall sd
## age 4.432176        31.513699    28.782143      12.227712  11.125568
## A1  4.026571         2.726282     2.413185       1.498345   1.403699
##          p.value
## age 9.328678e-06
## A1  5.659618e-05
##
## $`some college`
##        v.test Mean in category Overall mean sd in category Overall sd
## A4   8.334781         4.959001     4.700084       1.359491   1.474833
## E4   6.135781         4.609132     4.421009       1.349902   1.455627
## O2   4.887124         2.874299     2.713214       1.596449   1.564872
## A2   4.465375         4.914508     4.804686       1.126793   1.167638
## C2   4.055532         4.482939     4.370795       1.269912   1.312823
## A3   4.006821         4.714207     4.604769       1.234469   1.296718
## O5   3.361655         2.583046     2.489364       1.312532   1.323070
## A5   2.704708         4.632075     4.560564       1.234426   1.255248
## E2  -3.467854         3.024703     3.141687       1.564404   1.601558
## N4  -4.715235         3.029375     3.184460       1.539037   1.561502
## C4  -5.751950         2.387270     2.553145       1.297967   1.369119
## C5  -6.200094         3.083966     3.296112       1.582564   1.624476
## age -6.641486        27.225781    28.782143       9.445233  11.125568
##          p.value
## A4  7.764173e-17
## E4  8.474173e-10
## O2  1.023195e-06
## A2  7.992877e-06
## C2  5.002026e-05
## A3  6.154139e-05
```

```
## O5  7.747691e-04
## A5  6.836441e-03
## E2  5.246320e-04
## N4  2.414322e-06
## C4  8.822014e-09
## C5  5.642931e-10
## age 3.105357e-11
##
## $`college graduate`
##        v.test Mean in category Overall mean sd in category Overall sd
## age  8.077477        32.979695    28.782143      10.319741  11.125568
## O2  -2.743399         2.512690     2.713214       1.486348   1.564872
## A4  -3.563921         4.454574     4.700084       1.517742   1.474833
## A1  -3.728781         2.168707     2.413185       1.241342   1.403699
##          p.value
## age 6.612060e-16
## O2  6.080667e-03
## A4  3.653562e-04
## A1  1.924086e-04
##
## $`graduate degree`
##        v.test Mean in category Overall mean sd in category Overall sd
## age 12.986646        35.301435    28.782143      10.963622  11.125568
## O3   5.498533         4.739414     4.437910       1.119506   1.215245
## O1   3.327781         4.985034     4.816070       1.071881   1.125274
## O4   3.281653         5.072769     4.892402       1.101463   1.218096
## E4  -3.207547         4.210338     4.421009       1.458926   1.455627
## O5  -3.771794         2.264193     2.489364       1.320190   1.323070
## O2  -4.544317         2.392344     2.713214       1.412229   1.564872
## A1  -5.742488         2.049475     2.413185       1.240550   1.403699
##          p.value
## age 1.456720e-38
## O3  3.829639e-08
## O1  8.754079e-04
## O4  1.032006e-03
## E4  1.338723e-03
## O5  1.620777e-04
## O2  5.511363e-06
## A1  9.329557e-09
##
## $Unknown
##         v.test Mean in category Overall mean sd in category Overall sd
## C4    6.222562         3.100556     2.553145       1.390042   1.369119
## C5    5.325793         3.852018     3.296112       1.556293   1.624476
## A1    4.575801         2.825894     2.413185       1.469612   1.403699
## N1    4.337670         3.367629     2.931033       1.613464   1.566461
## N2    4.162202         3.914401     3.507424       1.456272   1.521751
## E2    3.503690         3.502242     3.141687       1.658987   1.601558
## N3    3.344357         3.559741     3.215912       1.563547   1.600022
## N4    2.945597         3.480002     3.184460       1.613780   1.561502
## O4   -3.199153         4.642011     4.892402       1.384490   1.218096
## E5   -3.508482         4.116874     4.417017       1.435163   1.331387
## C3   -3.538053         4.012189     4.304074       1.302358   1.283936
## E3   -3.667843         3.682383     4.000051       1.413264   1.347906
```

```
## C2   -4.077042         4.026877     4.370795       1.364893   1.312823
## A3   -4.152948         4.258746     4.604769       1.470946   1.296718
## O3   -4.407056         4.093786     4.437910       1.292218   1.215245
## A2   -5.500296         4.392021     4.804686       1.330553   1.167638
## A5   -5.726959         4.098655     4.560564       1.322163   1.255248
## C1   -6.098328         4.017937     4.502660       1.349187   1.237027
## A4   -7.510122         3.988390     4.700084       1.686754   1.474833
## age -15.151719        17.950673    28.782143       8.501769  11.125568
##          p.value
## C4  4.891010e-10
## C5  1.005137e-07
## A1  4.744012e-06
## N1  1.440008e-05
## N2  3.151928e-05
## E2  4.588591e-04
## N3  8.247357e-04
## N4  3.223324e-03
## O4  1.378321e-03
## E5  4.506720e-04
## C3  4.030887e-04
## E3  2.446052e-04
## C2  4.561232e-05
## A3  3.282193e-05
## O3  1.047851e-05
## A2  3.791547e-08
## A5  1.022466e-08
## C1  1.071836e-09
## A4  5.907243e-14
## age 7.381048e-52
```

Men are overrepresented in HS level (41.5% of HS group vs 32.82% globally, more than 10% of men included in the sample belong to HS group), while they are underrepresented in 'some college' groups. Specifically, numeric variables whose means are significantly different to overall mean for each f.educ level are:

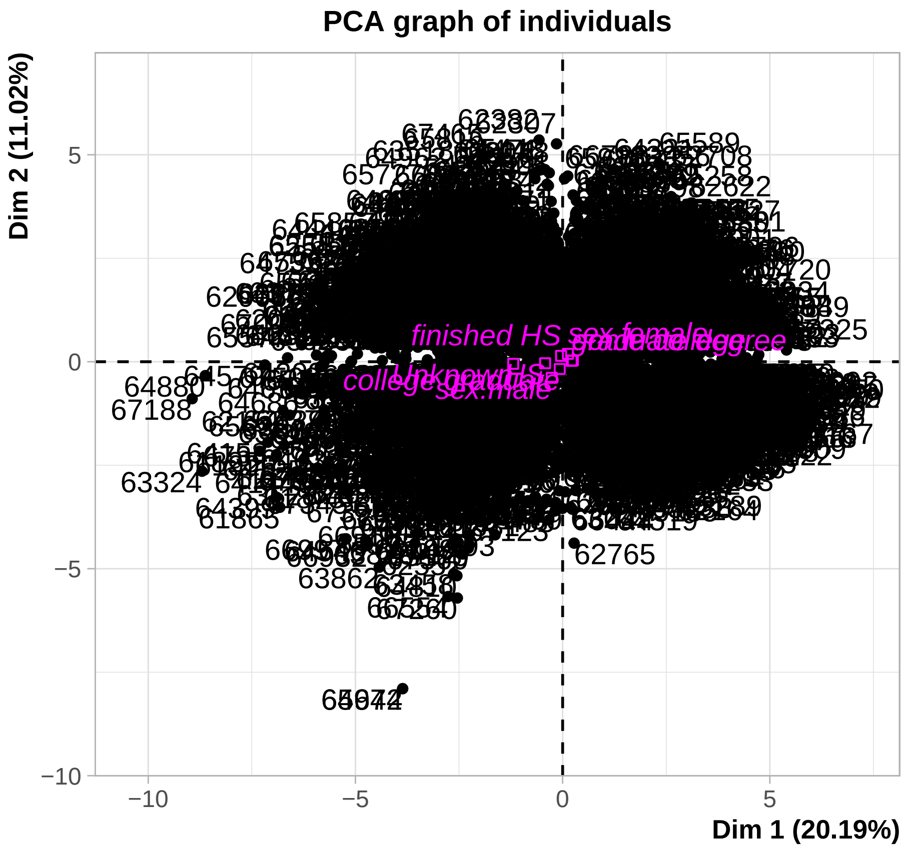
-For 'HS' level C4 and A1 are over the global mean, while age E5 and C3 are under the global mean. This is the direct answer to the question. -For 'finished HS' level age and A1 are over the global mean. -For 'some college' level A4, E4 and O2 are over the global mean, while age, C5 and C4 are under the global mean. -For 'college graduate' level age, O4 and C2 are over the global mean, while A1, A4 and O2 are under the global mean.

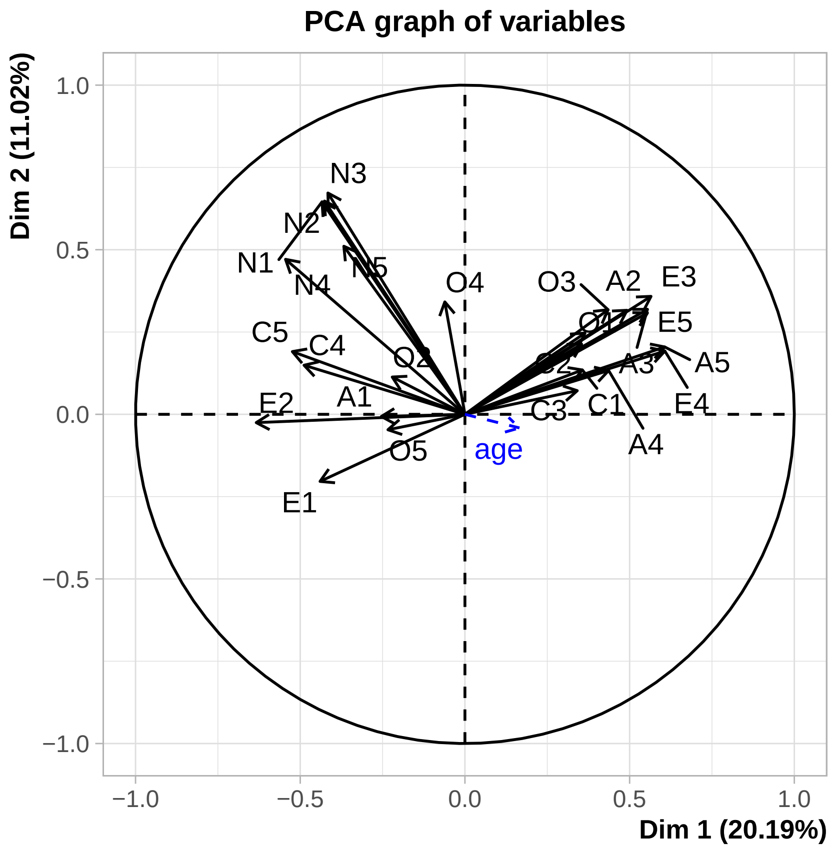
7. A Normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age. How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?.

```r
library(FactoMineR)
names(df)
```

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

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


PCA graph of individuals
Dim 2 (11.02%)
Dim 1 (20.19%)
5
0
−5
−10
−10
−5
0
5
finished HS
sex.female
college.grad
sex.male
62765
63324
67188
64880
61865
63862




```
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
## Variance               5.048   2.755   2.098   1.823   1.527   1.111
## % of var.             20.192  11.021   8.394   7.291   6.110   4.443
## Cumulative % of var.  20.192  31.214  39.608  46.898  53.008  57.451
##                        Dim.7   Dim.8   Dim.9  Dim.10  Dim.11  Dim.12
## Variance               0.847   0.811   0.733   0.697   0.682   0.659
## % of var.              3.387   3.244   2.932   2.788   2.728   2.636
## Cumulative % of var.  60.837  64.082  67.014  69.802  72.530  75.166
##                       Dim.13  Dim.14  Dim.15  Dim.16  Dim.17  Dim.18
## Variance               0.629   0.596   0.563   0.541   0.524   0.499
## % of var.              2.516   2.385   2.251   2.162   2.097   1.998
## Cumulative % of var.  77.683  80.068  82.319  84.481  86.578  88.576
##                       Dim.19  Dim.20  Dim.21  Dim.22  Dim.23  Dim.24
## Variance               0.490   0.457   0.433   0.409   0.406   0.386
## % of var.              1.960   1.828   1.733   1.635   1.622   1.542
## Cumulative % of var.  90.536  92.364  94.097  95.732  97.354  98.897
##                       Dim.25
## Variance               0.276
```

```
## % of var.                1.103
## Cumulative % of var. 100.000
##
## Variables
##                     Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## A1                | -0.254  1.276  0.064 | -0.007  0.002  0.000 |  0.141
## A2                |  0.492  4.800  0.242 |  0.316  3.623  0.100 | -0.179
## A3                |  0.554  6.079  0.307 |  0.319  3.691  0.102 | -0.266
## A4                |  0.436  3.773  0.190 |  0.133  0.642  0.018 | -0.147
## A5                |  0.607  7.289  0.368 |  0.205  1.519  0.042 | -0.291
## C1                |  0.357  2.520  0.127 |  0.135  0.659  0.018 |  0.529
## C2                |  0.354  2.488  0.126 |  0.213  1.652  0.046 |  0.515
## C3                |  0.341  2.306  0.116 |  0.072  0.187  0.005 |  0.408
## C4                | -0.488  4.726  0.239 |  0.149  0.808  0.022 | -0.470
## C5                | -0.524  5.438  0.275 |  0.191  1.318  0.036 | -0.290
## E1                | -0.440  3.834  0.194 | -0.204  1.508  0.042 |  0.340
## E2                | -0.634  7.953  0.401 | -0.025  0.023  0.001 |  0.285
## E3                |  0.565  6.324  0.319 |  0.358  4.660  0.128 | -0.155
## E4                |  0.607  7.299  0.368 |  0.192  1.341  0.037 | -0.381
## E5                |  0.554  6.084  0.307 |  0.308  3.447  0.095 |  0.084
## N1                | -0.434  3.728  0.188 |  0.645 15.117  0.417 |  0.020
## N2                | -0.426  3.602  0.182 |  0.648 15.234  0.420 |  0.072
## N3                | -0.416  3.429  0.173 |  0.672 16.408  0.452 |  0.042
## N4                | -0.545  5.877  0.297 |  0.471  8.046  0.222 |  0.111
## N5                | -0.368  2.677  0.135 |  0.511  9.462  0.261 | -0.032
## O1                |  0.365  2.634  0.133 |  0.247  2.211  0.061 |  0.257
## O2                | -0.221  0.965  0.049 |  0.113  0.466  0.013 | -0.390
## O3                |  0.435  3.742  0.189 |  0.318  3.663  0.101 |  0.190
## O4                | -0.061  0.074  0.004 |  0.342  4.234  0.117 |  0.282
## O5                | -0.234  1.083  0.055 | -0.047  0.080  0.002 | -0.353
##                     ctr   cos2
## A1                 0.949  0.020 |
## A2                 1.535  0.032 |
## A3                 3.377  0.071 |
## A4                 1.029  0.022 |
## A5                 4.026  0.084 |
## C1                13.349  0.280 |
## C2                12.649  0.265 |
## C3                 7.923  0.166 |
## C4                10.533  0.221 |
## C5                 4.005  0.084 |
## E1                 5.517  0.116 |
## E2                 3.875  0.081 |
## E3                 1.141  0.024 |
## E4                 6.930  0.145 |
## E5                 0.340  0.007 |
## N1                 0.019  0.000 |
## N2                 0.246  0.005 |
## N3                 0.084  0.002 |
## N4                 0.583  0.012 |
## N5                 0.050  0.001 |
## O1                 3.158  0.066 |
## O2                 7.239  0.152 |
## O3                 1.714  0.036 |
```

```
## O4                3.787  0.079 |
## O5                5.943  0.125 |
##
## Supplementary continuous variable
##                     Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
## age               |  0.166  0.027 | -0.041  0.002 |  0.044  0.002 |
##
## Supplementary categories
##                      Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test
## sex.male          |  0.701 | -0.263  0.140 -4.321 | -0.388  0.307 -8.650 |
## sex.female        |  0.342 |  0.128  0.140  4.321 |  0.190  0.307  8.650 |
## HS                |  0.505 | -0.423  0.702 -2.938 | -0.036  0.005 -0.340 |
## finished HS       |  0.340 | -0.037  0.012 -0.301 |  0.142  0.174  1.546 |
## some college      |  0.390 |  0.247  0.403  5.226 |  0.031  0.006  0.874 |
## college graduate  |  0.431 | -0.070  0.026 -0.668 | -0.176  0.167 -2.269 |
## graduate degree   |  0.586 |  0.216  0.136  2.131 |  0.022  0.001  0.290 |
## Unknown           |  1.338 | -1.192  0.795 -8.259 | -0.051  0.001 -0.478 |
##                   Dim.3   cos2 v.test
## sex.male          0.143  0.042  3.647 |
## sex.female       -0.070  0.042 -3.647 |
## HS               -0.045  0.008 -0.481 |
## finished HS      -0.019  0.003 -0.236 |
## some college     -0.058  0.022 -1.887 |
## college graduate  0.116  0.072  1.712 |
## graduate degree   0.224  0.147  3.434 |
## Unknown          -0.233  0.030 -2.506 |
```

Strictly following Kaiser criteria, we have to retain as many axes as eigenvalues greater than 1.0 (mean eigenvalue value). 6 axes satisfy the condition and explain 57.25% of the total inertia.

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?.

```r
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
## Variance                5.05   2.76   2.10   1.82   1.53   1.11   0.85
## % of var.              20.19  11.02   8.39   7.29   6.11   4.44   3.39
## Cumulative % of var.   20.19  31.21  39.61  46.90  53.01  57.45  60.84
##                        Dim.8  Dim.9 Dim.10 Dim.11 Dim.12 Dim.13 Dim.14
## Variance                0.81   0.73   0.70   0.68   0.66   0.63   0.60
## % of var.               3.24   2.93   2.79   2.73   2.64   2.52   2.39
## Cumulative % of var.   64.08  67.01  69.80  72.53  75.17  77.68  80.07
##                       Dim.15 Dim.16 Dim.17 Dim.18 Dim.19 Dim.20 Dim.21
## Variance                0.56   0.54   0.52   0.50   0.49   0.46   0.43
## % of var.               2.25   2.16   2.10   2.00   1.96   1.83   1.73
## Cumulative % of var.   82.32  84.48  86.58  88.58  90.54  92.36  94.10
##                       Dim.22 Dim.23 Dim.24 Dim.25
## Variance                0.41   0.41   0.39   0.28
## % of var.               1.64   1.62   1.54   1.10
```

```
## Cumulative % of var.  95.73  97.35  98.90 100.00
##
## Variables
##                     Dim.1    ctr   cos2   Dim.2    ctr   cos2
## A1                | -0.25  1.28  0.06 | -0.01  0.00  0.00 |
## A2                |  0.49  4.80  0.24 |  0.32  3.62  0.10 |
## A3                |  0.55  6.08  0.31 |  0.32  3.69  0.10 |
## A4                |  0.44  3.77  0.19 |  0.13  0.64  0.02 |
## A5                |  0.61  7.29  0.37 |  0.20  1.52  0.04 |
## C1                |  0.36  2.52  0.13 |  0.13  0.66  0.02 |
## C2                |  0.35  2.49  0.13 |  0.21  1.65  0.05 |
## C3                |  0.34  2.31  0.12 |  0.07  0.19  0.01 |
## C4                | -0.49  4.73  0.24 |  0.15  0.81  0.02 |
## C5                | -0.52  5.44  0.27 |  0.19  1.32  0.04 |
## E1                | -0.44  3.83  0.19 | -0.20  1.51  0.04 |
## E2                | -0.63  7.95  0.40 | -0.03  0.02  0.00 |
## E3                |  0.57  6.32  0.32 |  0.36  4.66  0.13 |
## E4                |  0.61  7.30  0.37 |  0.19  1.34  0.04 |
## E5                |  0.55  6.08  0.31 |  0.31  3.45  0.09 |
## N1                | -0.43  3.73  0.19 |  0.65 15.12  0.42 |
## N2                | -0.43  3.60  0.18 |  0.65 15.23  0.42 |
## N3                | -0.42  3.43  0.17 |  0.67 16.41  0.45 |
## N4                | -0.54  5.88  0.30 |  0.47  8.05  0.22 |
## N5                | -0.37  2.68  0.14 |  0.51  9.46  0.26 |
## O1                |  0.36  2.63  0.13 |  0.25  2.21  0.06 |
## O2                | -0.22  0.96  0.05 |  0.11  0.47  0.01 |
## O3                |  0.43  3.74  0.19 |  0.32  3.66  0.10 |
## O4                | -0.06  0.07  0.00 |  0.34  4.23  0.12 |
## O5                | -0.23  1.08  0.05 | -0.05  0.08  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  -4.32 | -0.39  0.31  -8.65 |
## sex.female        |  0.34 |  0.13  0.14   4.32 |  0.19  0.31   8.65 |
## HS                |  0.50 | -0.42  0.70  -2.94 | -0.04  0.01  -0.34 |
## finished HS       |  0.34 | -0.04  0.01  -0.30 |  0.14  0.17   1.55 |
## some college      |  0.39 |  0.25  0.40   5.23 |  0.03  0.01   0.87 |
## college graduate  |  0.43 | -0.07  0.03  -0.67 | -0.18  0.17  -2.27 |
## graduate degree   |  0.59 |  0.22  0.14   2.13 |  0.02  0.00   0.29 |
## Unknown           |  1.34 | -1.19  0.79  -8.26 | -0.05  0.00  -0.48 |
```

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

```
## $quanti
##     correlation       p.value
## E4   0.60699466 1.488184e-281
## A5   0.60660985 4.188300e-281
## E3   0.56503469 5.795883e-236
## E5   0.55420354 3.134229e-225
## A3   0.55396728 5.319307e-225
```

```
## A2   0.49222293 8.238256e-171
## A4   0.43639863 1.498031e-130
## O3   0.43464926 2.088312e-129
## O1   0.36462129  8.715772e-89
## C1   0.35665465  9.185889e-85
## C2   0.35438275  1.228486e-83
## C3   0.34119245  2.802821e-77
## age  0.16564985  1.119648e-18
## O4  -0.06102942  1.233824e-03
## O2  -0.22069147  3.125168e-32
## O5  -0.23377252  4.585558e-36
## A1  -0.25382053  2.045460e-42
## N5  -0.36763735  2.438712e-90
## N3  -0.41603183 1.225228e-117
## N2  -0.42638872 4.294669e-124
## N1  -0.43383833 7.046188e-129
## E1  -0.43994580 6.831921e-133
## C4  -0.48846284 7.392605e-168
## C5  -0.52394638 2.978100e-197
## N4  -0.54467322 4.154837e-216
## E2  -0.63361021 3.371755e-314
##
## $quali
##                   R2      p.value
## f.educ   0.032215754 3.187109e-18
## f.gender 0.006669853 1.512737e-05
##
## $category
##                          Estimate      p.value
## f.educ=some college     0.4572505 1.627991e-07
## f.gender=sex.female     0.1953874 1.512737e-05
## f.educ=graduate degree  0.4259802 3.305420e-02
## f.educ=HS              -0.2132007 3.286679e-03
## f.gender=sex.male      -0.1953874 1.512737e-05
## f.educ=Unknown         -0.9823412 9.780837e-17
##
## attr(,"class")
## [1] "condes" "list "
```

```r
ddd$Dim.2
```

```
## $quanti
##     correlation       p.value
## N3   0.67238557  0.000000e+00
## N2   0.64788637  0.000000e+00
## N1   0.64539007  0.000000e+00
## N5   0.51058660 8.937682e-186
## N4   0.47084892 1.662478e-154
## E3   0.35831722  1.358530e-85
## O4   0.34157340  1.854528e-77
## A3   0.31888557  3.297121e-67
## O3   0.31769525  1.076535e-66
## A2   0.31595203  6.031240e-66
## E5   0.30817227  1.147153e-62
## O1   0.24679642  4.000931e-40
```

```
## C2   0.21336473  3.444210e-30
## A5   0.20458220  7.711571e-28
## E4   0.19221975  1.038573e-24
## C5   0.19057214  2.617404e-24
## C4   0.14916588  2.121687e-15
## C1   0.13470439  8.225381e-13
## A4   0.13300524  1.591790e-12
## O2   0.11336165  1.796037e-09
## C3   0.07183851  1.420592e-04
## age -0.04133694  2.872039e-02
## O5  -0.04681709  1.322768e-02
## E1  -0.20384340  1.202263e-27
##
## $quali
##                  R2      p.value
## f.gender 0.02672902 3.151797e-18
##
## $category
##                           Estimate      p.value
## f.gender=sex.female      0.2889712 3.151797e-18
## f.educ=college graduate -0.1644619 2.326792e-02
## f.gender=sex.male       -0.2889712 3.151797e-18
##
## attr(,"class")
## [1] "condes" "list "
```

It is difficult to summarize, but positive correlation to axis 1 are E4, A5 and negative correlated to E2 (Find it difficult to approach others. (q_901)) and N4 (Often feel blue. (q_1479)). It seems an axis of sociability. For axis 2, positive correlation appears for N3-Have frequent mood swings. (q_1099), N2 (Get irritated easily. (q_974)), N1 (Get angry easily. (q_952)) and inversely associated to E1 (not so intense): it seems an axis of psicological stability.

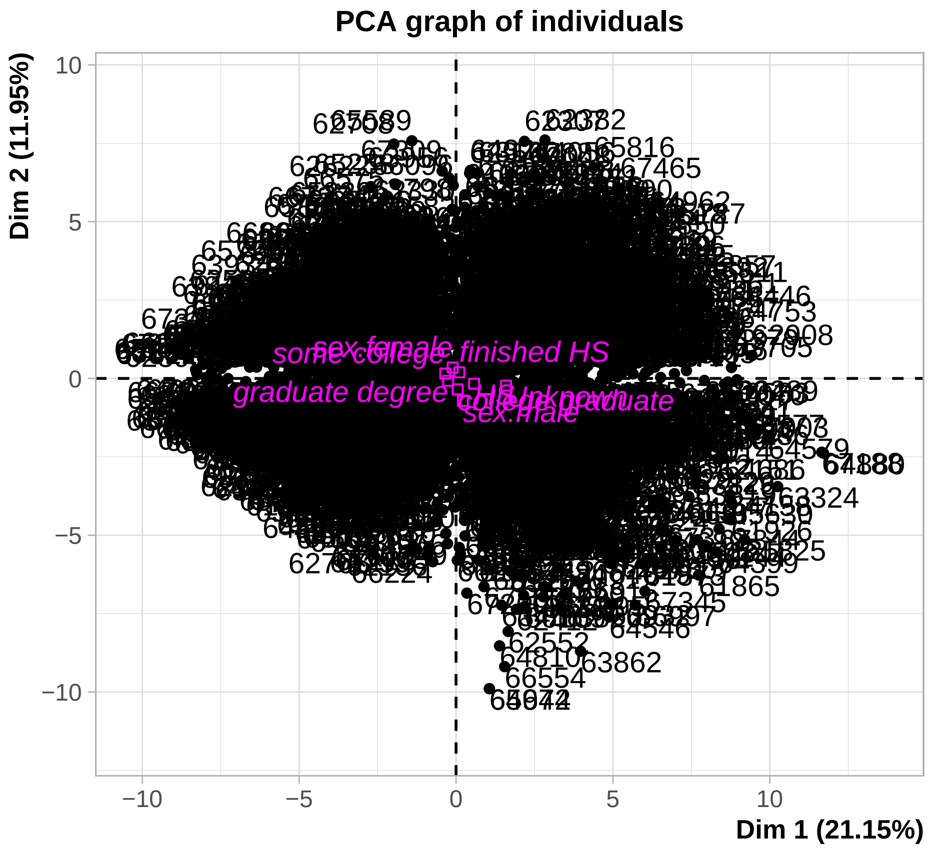
9. A Non-normalized Principal Component Analysis is addressed using as supplementary variables gender, education and age. How many axes do you have to retain according to Kaiser criteria? What's the inertia explained by retained Kaiser-based principal components?

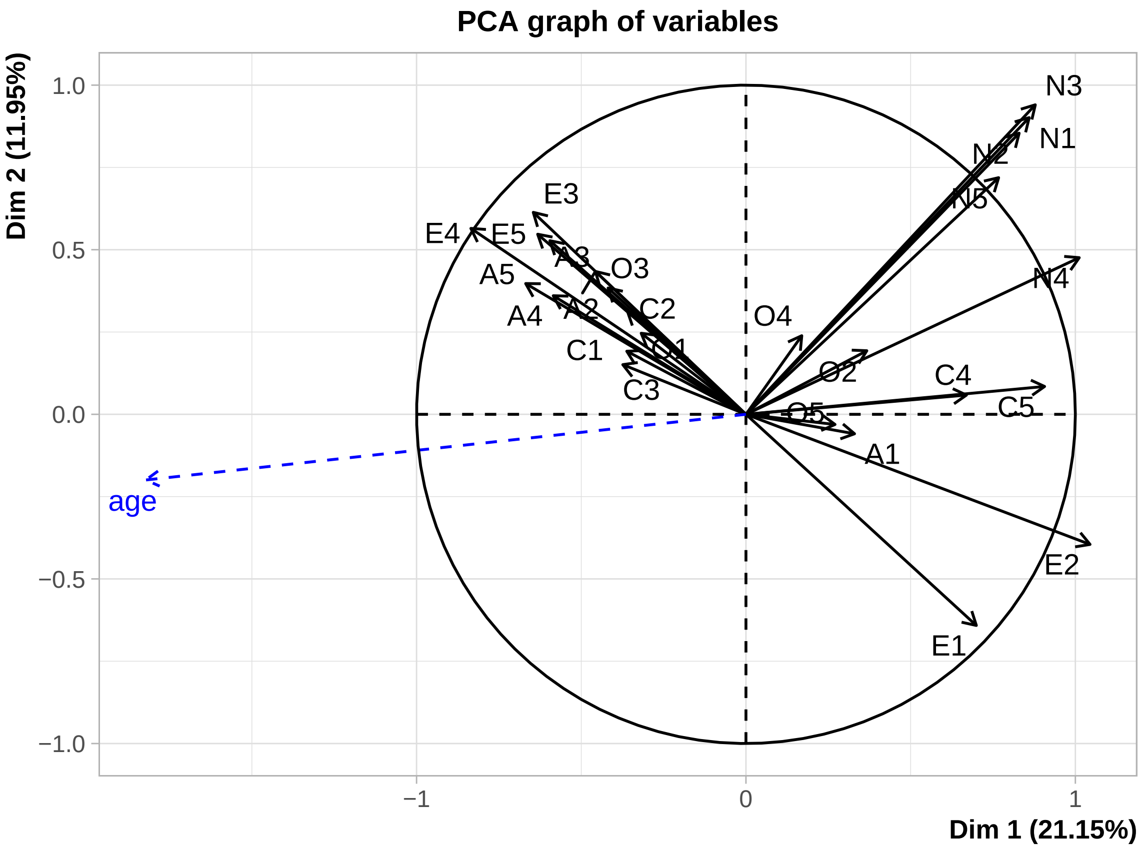
Strictly following Kaiser criteria, we have to retain as many axes as eigenvalues greater than 1.995566 (mean eigenvalue value). 6 axes satisfy the condition and explain 58.50% of the total inertia.

```r
names(df)
```

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

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


PCA graph of individuals
Dim 2 (11.95%)
Dim 1 (21.15%)
10
5
0
−5
−10
−10
−5
0
5
10
sex.female
some college
finished HS
graduate degree
college graduate
sex.male




```
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
## Variance               10.55   5.96   4.04   3.49   2.99   2.16   1.80
## % of var.              21.15  11.95   8.09   6.99   6.00   4.33   3.61
## Cumulative % of var.   21.15  33.10  41.19  48.17  54.17  58.50  62.11
##                        Dim.8  Dim.9 Dim.10 Dim.11 Dim.12 Dim.13 Dim.14
## Variance                1.75   1.50   1.43   1.34   1.22   1.17   1.14
## % of var.               3.51   3.01   2.87   2.69   2.44   2.35   2.29
## Cumulative % of var.   65.61  68.63  71.50  74.18  76.62  78.97  81.27
##                       Dim.15 Dim.16 Dim.17 Dim.18 Dim.19 Dim.20 Dim.21
## Variance                1.09   0.99   0.98   0.93   0.89   0.84   0.80
## % of var.               2.19   1.98   1.97   1.86   1.79   1.68   1.61
## Cumulative % of var.   83.46  85.44  87.41  89.26  91.06  92.74  94.35
##                       Dim.22 Dim.23 Dim.24 Dim.25
## Variance                0.76   0.74   0.69   0.63
## % of var.               1.53   1.48   1.38   1.27
## Cumulative % of var.   95.88  97.35  98.73 100.00
##
```

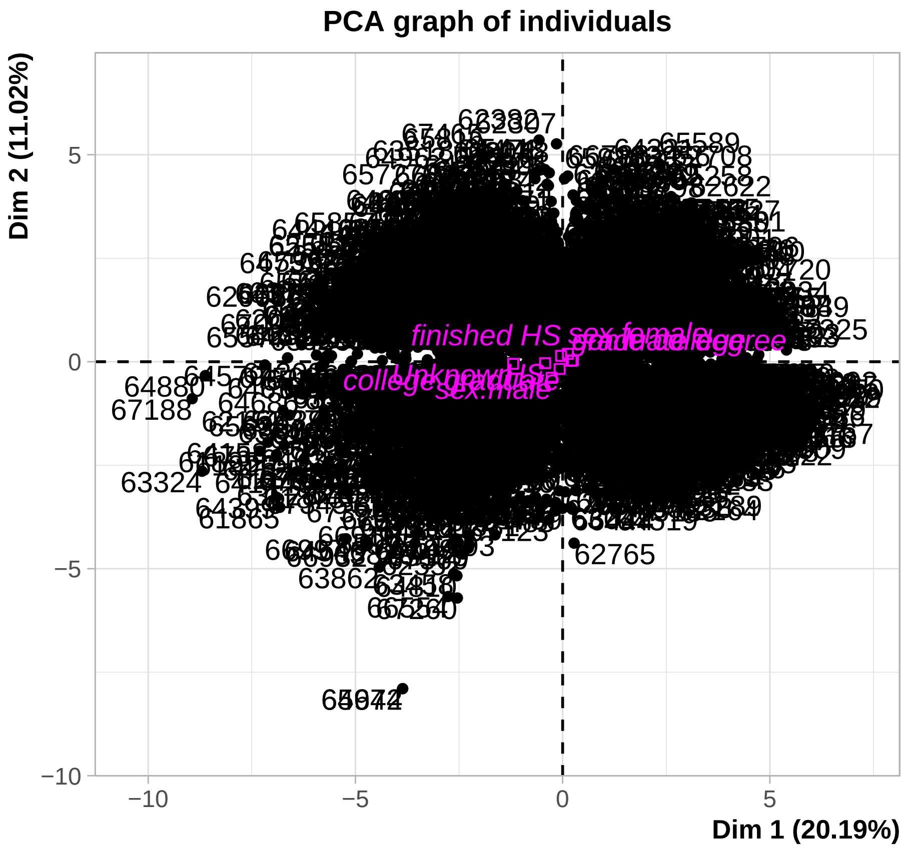
```
## Variables
##                     Dim.1    ctr   cos2    Dim.2    ctr   cos2
## A1                 |  0.33  1.03  0.06 | -0.06  0.06  0.00 |
## A2                 | -0.46  1.98  0.15 |  0.43  3.16  0.14 |
## A3                 | -0.60  3.36  0.21 |  0.53  4.67  0.17 |
## A4                 | -0.59  3.24  0.16 |  0.36  2.18  0.06 |
## A5                 | -0.67  4.24  0.28 |  0.40  2.65  0.10 |
## C1                 | -0.36  1.24  0.09 |  0.19  0.62  0.02 |
## C2                 | -0.36  1.23  0.08 |  0.31  1.64  0.06 |
## C3                 | -0.37  1.33  0.08 |  0.15  0.38  0.01 |
## C4                 |  0.67  4.25  0.24 |  0.06  0.06  0.00 |
## C5                 |  0.91  7.79  0.31 |  0.08  0.12  0.00 |
## E1                 |  0.70  4.64  0.19 | -0.64  6.90  0.16 |
## E2                 |  1.04 10.34  0.43 | -0.40  2.62  0.06 |
## E3                 | -0.65  3.95  0.23 |  0.61  6.32  0.21 |
## E4                 | -0.84  6.61  0.33 |  0.56  5.35  0.15 |
## E5                 | -0.63  3.79  0.23 |  0.55  5.02  0.17 |
## N1                 |  0.86  7.01  0.30 |  0.90 13.63  0.33 |
## N2                 |  0.83  6.52  0.30 |  0.85 12.24  0.31 |
## N3                 |  0.88  7.31  0.30 |  0.94 14.83  0.35 |
## N4                 |  1.01  9.70  0.42 |  0.48  3.80  0.09 |
## N5                 |  0.77  5.58  0.23 |  0.72  8.67  0.20 |
## O1                 | -0.32  0.96  0.08 |  0.25  1.02  0.05 |
## O2                 |  0.37  1.28  0.06 |  0.19  0.63  0.02 |
## O3                 | -0.42  1.66  0.12 |  0.38  2.47  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           |   1.01 |   0.22   0.05   2.48 |  -0.72   0.51 -10.87 |
## sex.female         |   0.49 |  -0.11   0.05  -2.48 |   0.35   0.51  10.87 |
## HS                 |   0.70 |   0.57   0.67   2.76 |  -0.17   0.06  -1.06 |
## finished HS        |   0.48 |   0.10   0.05   0.57 |   0.20   0.17   1.45 |
## some college       |   0.56 |  -0.34   0.38  -5.03 |   0.16   0.08   3.08 |
## college graduate   |   0.62 |   0.05   0.01   0.34 |  -0.35   0.31  -3.04 |
## graduate degree    |   0.81 |  -0.25   0.10  -1.72 |  -0.07   0.01  -0.67 |
## Unknown            |   1.86 |   1.60   0.74   7.65 |  -0.23   0.01  -1.45 |
```

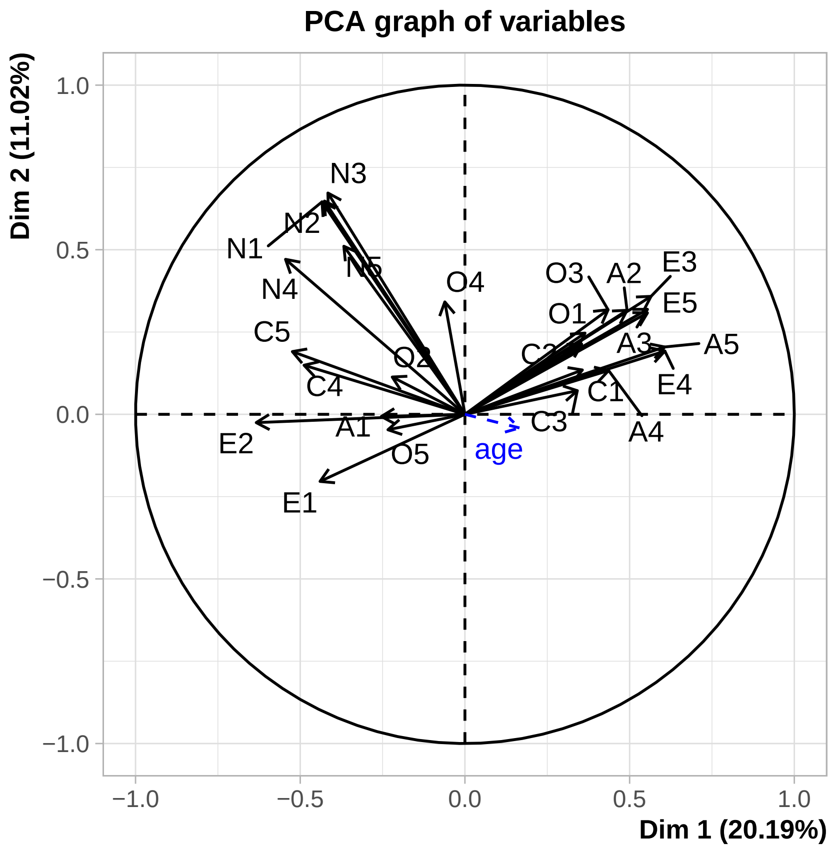
```r
mean(res.pcann$eig[,1])
```

```
## [1] 1.995566
```

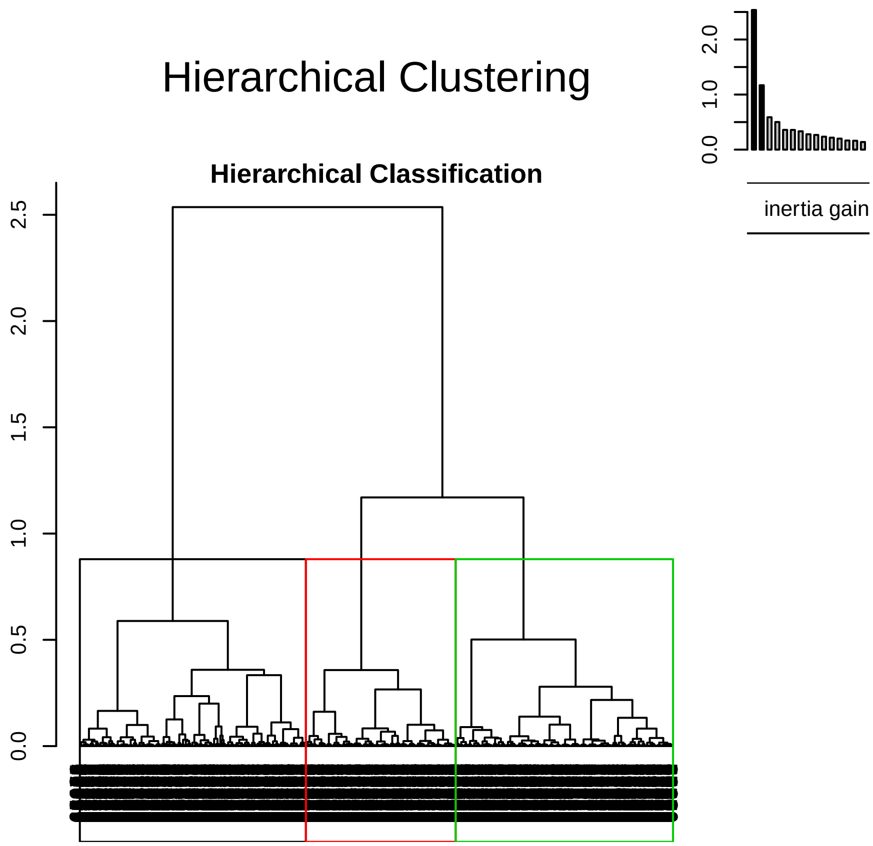
10. A Hierarchical Clustering is addressed. A non-default criteria for selecting the number of clusters to 3 has to be set. Explain the characteristics of cluster number 1.

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

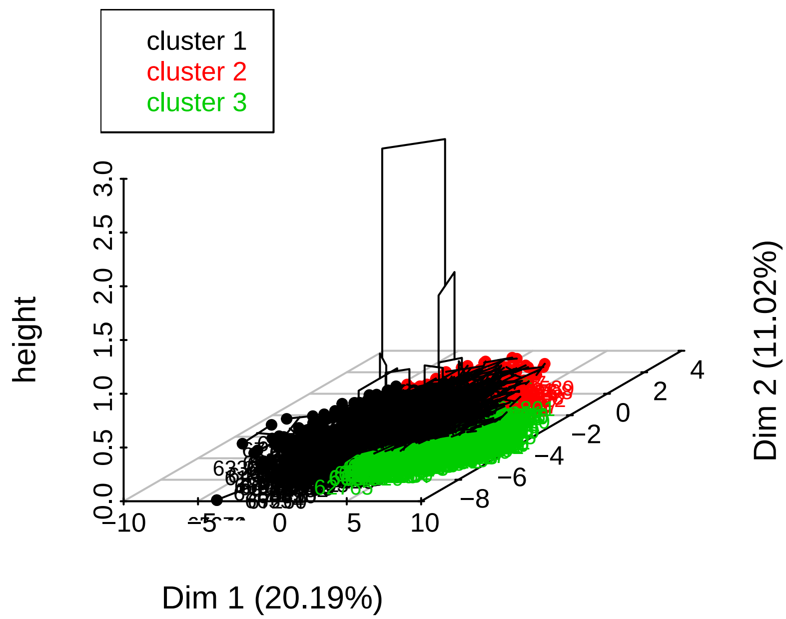

PCA graph of individuals
Dim 2 (11.02%)
Dim 1 (20.19%)
finished HS
sex.female
graduate degree
college degree
Unknown
sex.male
62765
63324
67188
61865
63862
−10
−5
0
5

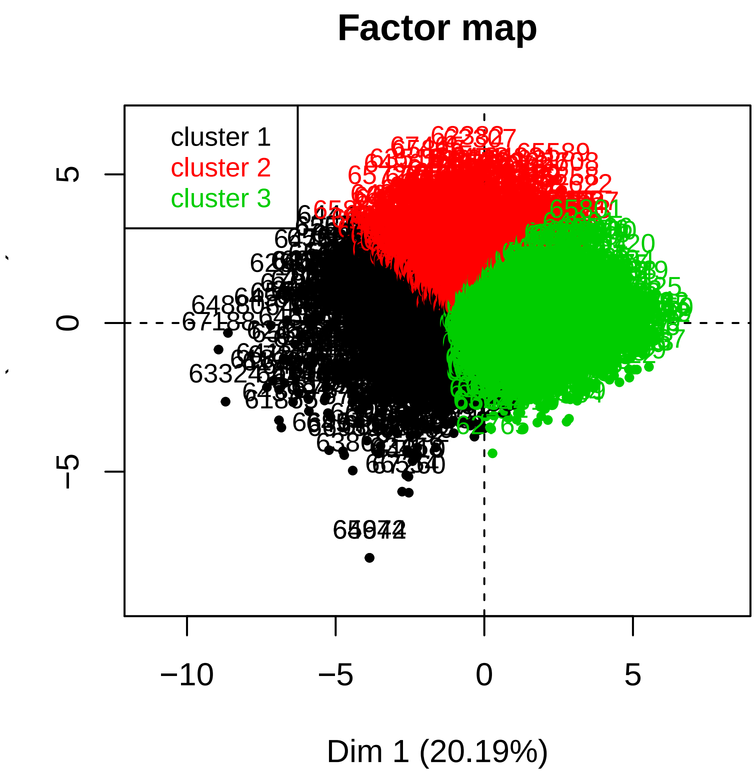



```
res.hcpc<-HCPC(res.pca,nb.clust=-1,graph=T)
```


Hierarchical Clustering
Hierarchical Classification
0.0
0.5
1.0
1.5
2.0
2.5
inertia gain
0.0
1.0
2.0

Hierarchical clustering on the factor map






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

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =================================================================================
##               p.value df
## f.gender 2.303901e-13  2
## f.educ   8.067390e-12 10
##
## Description of each cluster by the categories
## =============================================
## $`1`
##                          Cla/Mod  Mod/Cla    Global      p.value    v.test
## f.gender=sex.male       38.95539 42.61905 32.821429 9.031198e-13  7.144518
## f.educ=Unknown          47.08520 12.50000  7.964286 2.270547e-08  5.590011
## f.educ=college graduate 35.27919 16.54762 14.071429 1.481146e-02  2.436957
## f.educ=some college     25.70056 38.21429 44.607143 7.925025e-06 -4.467199
## f.gender=sex.female     25.62467 57.38095 67.178571 9.031198e-13 -7.144518
##
## $`2`
##                      Cla/Mod  Mod/Cla   Global      p.value    v.test
## f.gender=sex.female 33.70548 74.32591 67.17857 6.973169e-08  5.391861
## f.gender=sex.male   23.83025 25.67409 32.82143 6.973169e-08 -5.391861
##
## $`3`
##                      Cla/Mod   Mod/Cla    Global      p.value    v.test
```

```
## f.educ=some college 43.31465 48.870822 44.607143 2.469734e-04  3.665378
## f.educ=Unknown      18.38565  3.703704  7.964286 1.896052e-12 -7.041923
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##           Eta2       P-value
## N3  0.34876858 3.227595e-261
## N2  0.32558352 5.702340e-240
## N1  0.32520323 1.254388e-239
## N4  0.30655043 4.563696e-223
## E4  0.29104254 1.236849e-209
## E3  0.28985213 1.292062e-208
## A5  0.26969800 1.283022e-191
## A3  0.26508472 8.567006e-188
## E2  0.25215023 3.383701e-177
## E5  0.24919838 8.355890e-175
## N5  0.21034708 3.669376e-144
## A2  0.20972241 1.108828e-143
## C5  0.18862254 1.117575e-127
## E1  0.18189876 1.149895e-122
## O3  0.15326770 8.973411e-102
## C4  0.14420090  2.642392e-95
## A4  0.14261871  3.498180e-94
## O1  0.09751128  4.845401e-63
## C2  0.07599473  9.901835e-49
## C1  0.06582140  4.428313e-42
## C3  0.05940283  6.383109e-38
## A1  0.04203449  8.275538e-27
## O2  0.03405374  9.051483e-22
## O4  0.03282860  5.328050e-21
## O5  0.02832719  3.522014e-18
## age 0.01728051  2.586750e-11
##
## Description of each cluster by quantitative variables
## =====================================================
## $`1`
##         v.test Mean in category Overall mean sd in category Overall sd
## E2   23.454477         4.226251     3.141687       1.369127   1.601558
## E1   22.454916         4.028386     2.974798       1.493125   1.625071
## C5   14.513925         3.976859     3.296112       1.472112   1.624476
## C4   13.338115         3.080402     2.553145       1.344438   1.369119
## N4   11.882418         3.720176     3.184460       1.450143   1.561502
## A1    8.424656         2.754623     2.413185       1.337596   1.403699
## O5    7.352829         2.770246     2.489364       1.329124   1.323070
## O2    3.817889         2.885714     2.713214       1.572467   1.564872
## N2    3.497848         3.661109     3.507424       1.369696   1.521751
## N5    2.890577         3.103213     2.968657       1.523331   1.612245
## N1    2.871027         3.060884     2.931033       1.477797   1.566461
## N3    2.230828         3.318970     3.215912       1.505534   1.600022
## O4   -2.004696         4.821898     4.892402       1.255727   1.218096
## age  -4.476912        27.344048    28.782143      11.040055  11.125568
## C3  -11.992404         3.859507     4.304074       1.319944   1.283936
## C1  -13.265351         4.028871     4.502660       1.311911   1.237027
```

```
## C2  -14.572952         3.818411     4.370795       1.387198   1.312823
## O1  -16.366670         4.284322     4.816070       1.268400   1.125274
## A4  -19.274909         3.879313     4.700084       1.582601   1.474833
## O3  -20.699502         3.711619     4.437910       1.326620   1.215245
## A2  -24.190360         3.989161     4.804686       1.221139   1.167638
## E5  -26.345503         3.404277     4.417017       1.344742   1.331387
## A5  -26.419397         3.603063     4.560564       1.213978   1.255248
## A3  -27.161159         3.587863     4.604769       1.311775   1.296718
## E4  -27.885006         3.249062     4.421009       1.414524   1.455627
## E3  -28.348570         2.896792     4.000051       1.217736   1.347906
##           p.value
## E2  1.189928e-121
## E1  1.145743e-111
## C5   9.889614e-48
## C4   1.389284e-40
## N4   1.460803e-32
## A1   3.618111e-17
## O5   1.940555e-13
## O2   1.345987e-04
## N2   4.690281e-04
## N5   3.845359e-03
## N1   4.091402e-03
## N3   2.569250e-02
## O4   4.499553e-02
## age  7.573054e-06
## C3   3.894368e-33
## C1   3.676956e-40
## C2   4.174562e-48
## O1   3.308245e-60
## A4   8.725377e-83
## O3   3.499698e-95
## A2  2.810242e-129
## E5  5.778956e-153
## A5  8.203233e-154
## A3  1.869332e-162
## E4  4.055636e-171
## E3  8.716657e-177
##
## $`2`
##        v.test Mean in category Overall mean sd in category Overall sd
## N3  27.120140         4.455065     3.215912      1.2166416   1.600022
## N1  25.823267         4.086181     2.931033      1.3490254   1.566461
## N2  25.507693         4.615888     3.507424      1.1050891   1.521751
## N5  20.459869         3.910635     2.968657      1.5470715   1.612245
## N4  18.980980         4.030846     3.184460      1.3136976   1.561502
## A3   9.912763         4.971837     4.604769      1.0059997   1.296718
## E3   9.790369         4.376899     4.000051      1.1085967   1.347906
## C5   9.764682         3.749092     3.296112      1.5662436   1.624476
## E5   9.749020         4.787674     4.417017      1.0688579   1.331387
## O4   9.316749         5.216483     4.892402      0.9983749   1.218096
## A2   9.259691         5.113439     4.804686      0.9109201   1.167638
## O3   8.315184         4.726475     4.437910      1.0212739   1.215245
## C4   7.760448         2.856558     2.553145      1.4081918   1.369119
## E4   6.596893         4.695227     4.421009      1.2477534   1.455627
```

```
## O2   6.443819         3.001172     2.713214      1.6488295   1.564872
## C2   5.790455         4.587878     4.370795      1.2424899   1.312823
## O1   5.063380         4.978777     4.816070      1.0451340   1.125274
## A5   4.649632         4.727233     4.560564      1.1168769   1.255248
## A4   3.611554         4.852189     4.700084      1.4083383   1.474833
## C1   3.157163         4.614188     4.502660      1.1402270   1.237027
## A1   2.506996         2.513677     2.413185      1.4883777   1.403699
## age -2.856649        27.874560    28.782143     10.2486998  11.125568
## E1  -7.732733         2.615949     2.974798      1.5132280   1.625071
##           p.value
## N3  5.699381e-162
## N1  4.859224e-147
## N2  1.619557e-143
## N5   4.907036e-93
## N4   2.449831e-80
## A3   3.663709e-23
## E3   1.238428e-22
## C5   1.596113e-22
## E5   1.862582e-22
## O4   1.199595e-20
## A2   2.050248e-20
## O3   9.160984e-17
## C4   8.463015e-15
## E4   4.198643e-11
## O2   1.165044e-10
## C2   7.019587e-09
## O1   4.118871e-07
## A5   3.325275e-06
## A4   3.043672e-04
## C1   1.593121e-03
## A1   1.217621e-02
## age  4.281387e-03
## E1   1.052622e-14
##
## $`3`
##         v.test Mean in category Overall mean sd in category Overall sd
## A5   20.385174         5.158696     4.560564      0.9116163   1.255248
## E4   19.925777         5.098991     4.421009      1.0399977   1.455627
## E3   17.354055         4.546833     4.000051      1.0914402   1.347906
## A3   16.125912         5.093560     4.604769      1.0198486   1.296718
## E5   15.515562         4.899881     4.417017      1.0553382   1.331387
## A4   14.666053         5.205686     4.700084      1.1349854   1.474833
## A2   13.956242         5.185602     4.804686      0.9769061   1.167638
## O3   11.573480         4.766672     4.437910      1.0118126   1.215245
## O1   10.573555         5.094190     4.816070      0.9099437   1.125274
## C3   10.368259         4.615247     4.304074      1.1303353   1.283936
## C1    9.461211         4.776237     4.502660      1.1436575   1.237027
## C2    8.207915         4.622674     4.370795      1.1740210   1.312823
## age   6.885214        30.572719    28.782143     11.5865580  11.125568
## O4   -6.891459         4.696181     4.892402      1.2904182   1.218096
## O5   -8.153247         2.237209     2.489364      1.2040214   1.323070
## O2   -9.644343         2.360434     2.713214      1.4184190   1.564872
## A1  -10.256173         2.076663     2.413185      1.3075449   1.403699
## E1  -13.767075         2.451840     2.974798      1.4172126   1.625071
```

```
## C4  -19.806800        1.919262     2.553145      1.0761019   1.369119
## N5  -21.969333        2.140713     2.968657      1.2595977   1.612245
## C5  -22.795557        2.430513     3.296112      1.3715291   1.624476
## E2  -23.000762        2.280618     3.141687      1.2882043   1.601558
## N1  -26.999893        1.942401     2.931033      1.0624035   1.566461
## N2  -27.290326        2.536678     3.507424      1.2592351   1.521751
## N3  -27.620675        2.182881     3.215912      1.1646548   1.600022
## N4  -29.004964        2.125771     3.184460      1.1624081   1.561502
##           p.value
## A5   2.264107e-92
## E4   2.432423e-88
## E3   1.837998e-67
## A3   1.677651e-58
## E5   2.722392e-54
## A4   1.063539e-48
## A2   2.882363e-44
## O3   5.615780e-31
## O1   3.952219e-26
## C3   3.457680e-25
## C1   3.043967e-21
## C2   2.250616e-16
## age  5.770091e-12
## O4   5.522286e-12
## O5   3.542807e-16
## O2   5.194253e-22
## A1   1.110203e-24
## E1   4.021685e-43
## C4   2.600975e-87
## N5  5.659355e-107
## C5  5.074829e-115
## E2  4.580121e-117
## N1  1.482190e-160
## N2  5.525391e-164
## N3  6.282566e-168
## N4  5.696362e-185
```

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

```
## [1] 0.2580477
```

Three clusters are selected which it is enough to represent the complexity of this dataset (it explains less than 26% of total inertia in data). Cluster 1 contains 39% of the male observations in the sample. On average 32.82% of the data units belong to male gender, but in Cluster 1 males are overrepresented (42.6%). 'Unknown' is also overrepresented being 12.5% in Cluster 1 and 8% globally and 'some graduate' educated people represents 44.61% of the sample, but only 38.21% of them are included in Cluster 1.

Cluster 1 shows mean values of E2, E1, C5 and C4 remarkably over the global mean, while the global mean in the sample, while E3, E4, A3 are clearly under the global mean in this cluster. It indicates difficult approach and not very social and communicative behavior.

Do not forget to Knit to .pdf before posting your answers in ATENEA.