PCA, CA and Clustering

Fujie Mei Sergio Delgado Mario Wang

April 29, 2023

Contents

1 Loading data and deleting columns

We will delete the columns we said that won't contained too many errors to be analyzeable.

```
df<-read.csv2("clean_data.csv")
df$X<-NULL
df$pdays<-NULL
df$previous<-NULL
df$errVar<-NULL
names(df)
```

```
##  [1] "age"               "job"               "marital"          
##  [4] "education"         "housing"           "loan"             
##  [7] "contact"           "month"             "day_of_week"      
## [10] "duration"          "campaign"          "poutcome"         
## [13] "emp.var.rate"      "cons.price.idx"    "cons.conf.idx"    
## [16] "euribor3m"         "nr.employed"       "y"                
## [19] "Age_group"         "Campaign_contacts" "mout"
```

```
vars_con = c("age","campaign","emp.var.rate","cons.price.idx","cons.conf.idx","euribor3m","nr.employed")
vars_dis = c("job","marital","education","housing","loan","contact","month","day_of_week")
vars_res= c("y","duration")
```

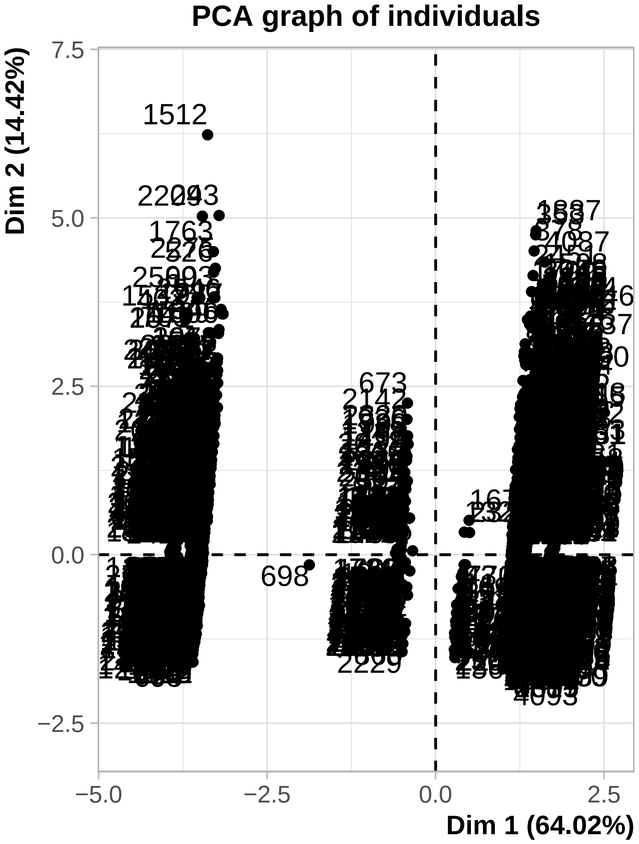
2 Principal Component Analysis (PCA)

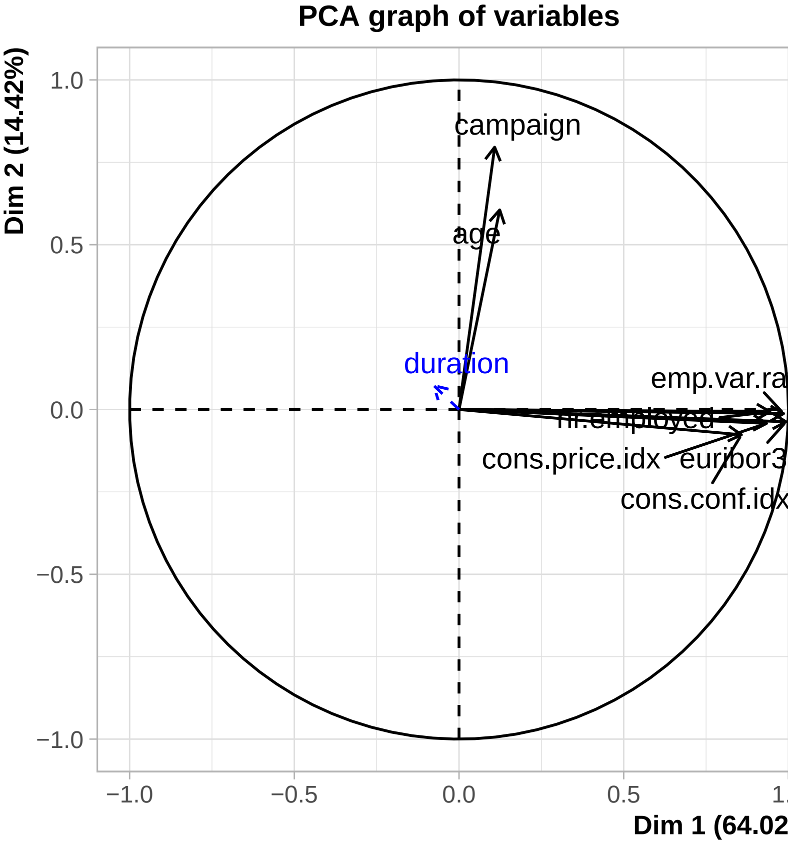
We are going to do a PCA analysis in our numerical variables from our dataset, from the PCA graph we can see that the target variable, duration, has little effect and the rest of the variable are very contributive to their respective axes. As we can see, they are very near to the axes and their length are very long.

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

From the Kayser's rule, all eigenvalues >1, we should consider 2 dimensions, on the other hand, with the Elbow's rule 4 dimensions is the most suitable. In our case we will take Kayser's rule into consideration because it's least number of components and the cummulative variation is almost 80%.

```
res.pca <- PCA(df[,c("duration",vars_con)],quanti.sup=c(1))
```

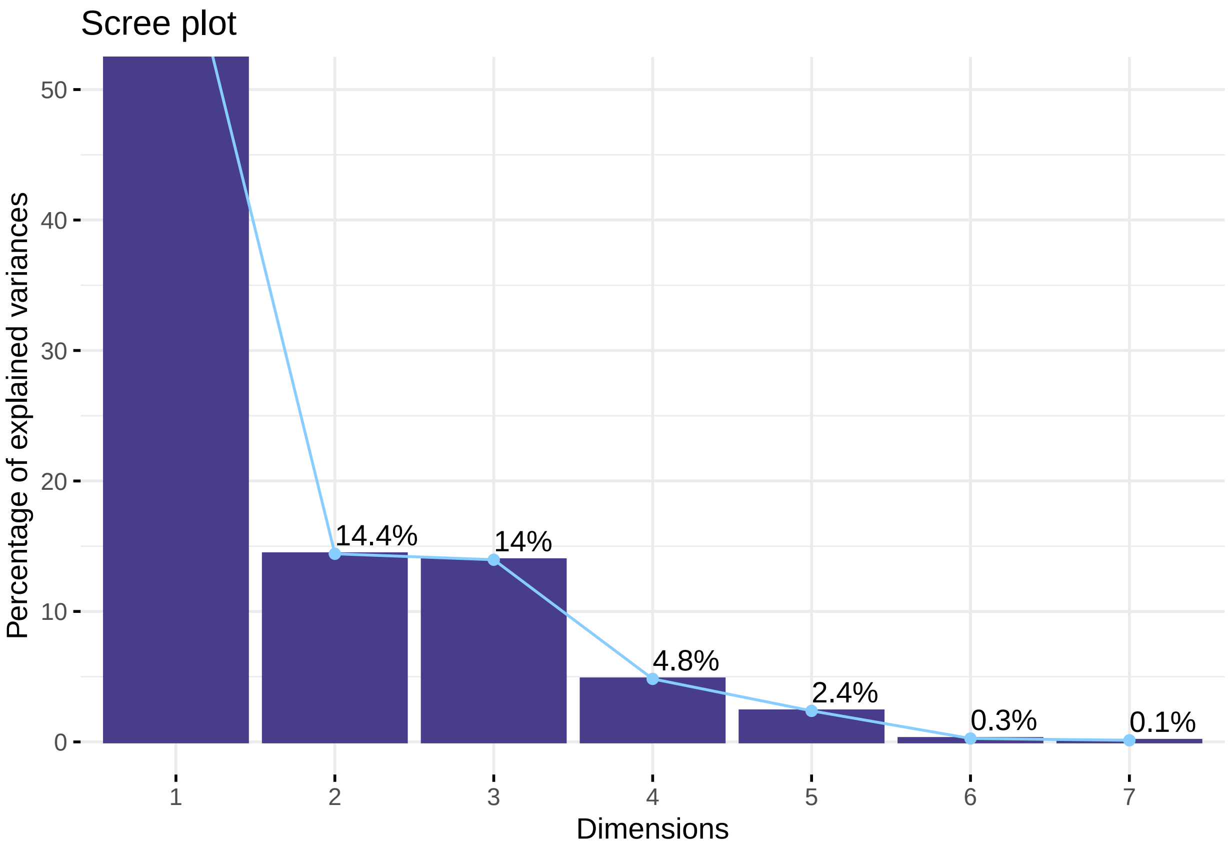




```
summary(res.pca)
```

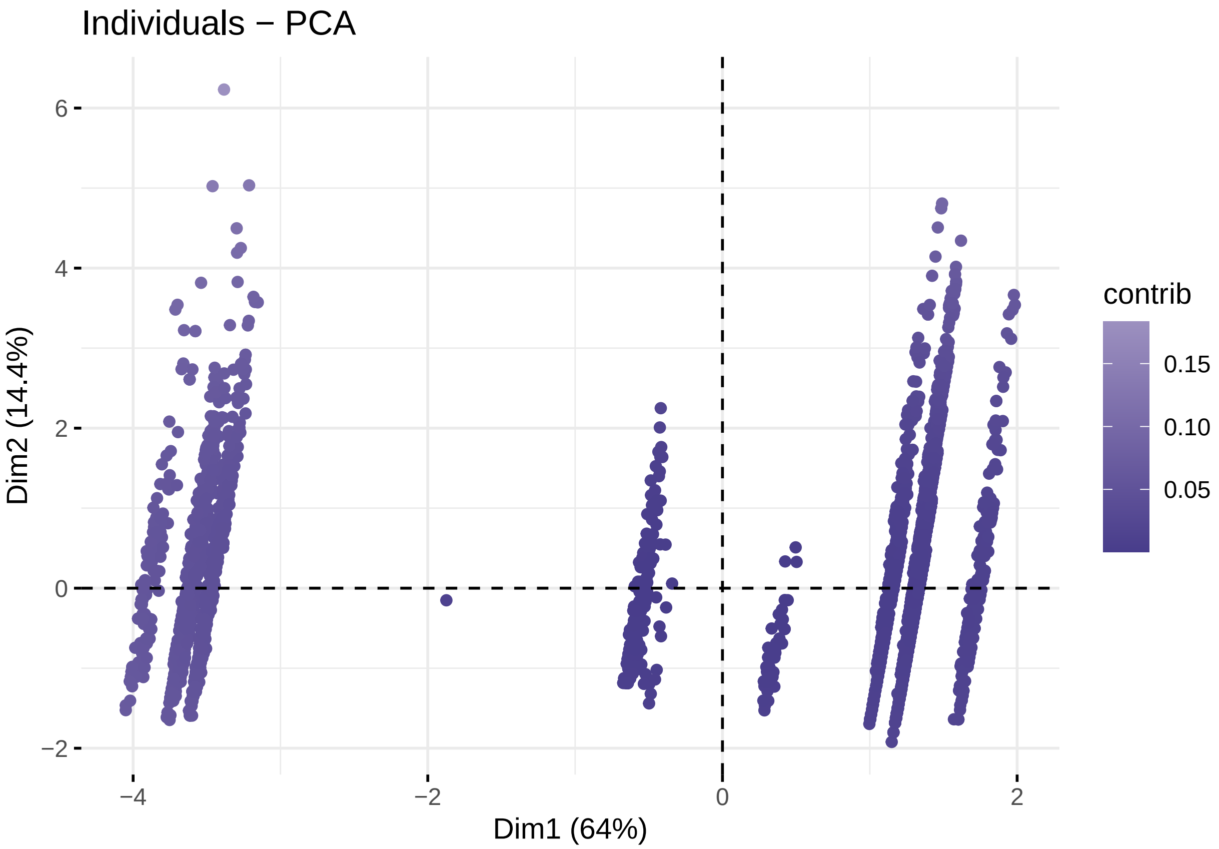
```
##
## Call:
## PCA(X = df[, c("duration", vars_con)], quanti.sup = c(1))
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               4.481   1.009   0.978   0.338   0.167   0.018   0.009
## % of var.             64.016  14.419  13.965   4.831   2.381   0.264   0.124
## Cumulative % of var.  64.016  78.435  92.400  97.231  99.612  99.876 100.000
##
## Individuals (the 10 first)
##                    Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1              |  1.905 |  1.185  0.006  0.387 |  0.650  0.008  0.116 | -0.511
## 2              |  3.706 | -3.569  0.057  0.928 |  0.579  0.007  0.024 | -0.792
## 3              |  2.215 |  1.208  0.007  0.298 | -1.289  0.033  0.339 | -0.359
## 4              |  3.640 | -3.508  0.055  0.929 | -0.389  0.003  0.011 | -0.826
## 5              |  2.007 |  1.083  0.005  0.291 | -0.613  0.007  0.093 | -0.923
## 6              |  3.682 | -3.625  0.059  0.969 | -0.328  0.002  0.008 |  0.517
## 7              |  2.790 |  1.810  0.015  0.421 |  1.432  0.041  0.263 | -0.642
## 8              |  2.328 | -0.474  0.001  0.042 |  0.551  0.006  0.056 |  1.837
## 9              |  1.968 |  1.072  0.005  0.296 | -0.917  0.017  0.217 | -0.071
## 10             |  1.609 | -0.611  0.002  0.144 | -0.522  0.005  0.105 | -0.833
##                   ctr   cos2
## 1               0.005  0.072 |
## 2               0.013  0.046 |
## 3               0.003  0.026 |
## 4               0.014  0.051 |
## 5               0.017  0.211 |
## 6               0.005  0.020 |
## 7               0.008  0.053 |
## 8               0.069  0.623 |
## 9               0.000  0.001 |
## 10              0.014  0.268 |
##
## Variables
##                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
## age             |  0.124  0.341  0.015 |  0.606 36.341  0.367 |  0.786 63.161
## campaign        |  0.108  0.261  0.012 |  0.796 62.742  0.633 | -0.593 35.973
## emp.var.rate    |  0.985 21.671  0.971 | -0.012  0.015  0.000 | -0.027  0.073
## cons.price.idx |  0.934 19.453  0.872 | -0.042  0.176  0.002 | -0.004  0.002
## cons.conf.idx   |  0.857 16.403  0.735 | -0.077  0.583  0.006 |  0.068  0.472
## euribor3m       |  0.993 21.988  0.985 | -0.037  0.138  0.001 | -0.009  0.009
## nr.employed     |  0.944 19.882  0.891 | -0.007  0.005  0.000 | -0.055  0.311
##                  cos2
## age             0.617 |
## campaign        0.352 |
## emp.var.rate    0.001 |
## cons.price.idx  0.000 |
## cons.conf.idx   0.005 |
## euribor3m       0.000 |
## nr.employed     0.003 |
##
## Supplementary continuous variable
##                   Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
## duration        | -0.076  0.006 |  0.072  0.005 | -0.116  0.013 |
```

```
fviz_screeplot(
  res.pca,
  addlabels=TRUE,
  ylim=c(0,50),
  barfill="darkslateblue",
  barcolor="darkslateblue",
  linecolor = "skyblue1"
)
```



## Individuals point of view: Are they any individuals “too contributive”? From what we can see in the graph of individuals, none is “too contributive”, as we can see contributions values that ranges from 0 to 0.20 more or less. So we can say in this part that almost all individuals contribute the same.

```
# head(res.pca$ind$contrib) # contribition of individuals to the princial components
fviz_pca_ind(res.pca, col.ind="contrib", geom = "point") +
scale_color_gradient2(low="darkslateblue", mid="white",
                      high="red", midpoint=0.40)
```

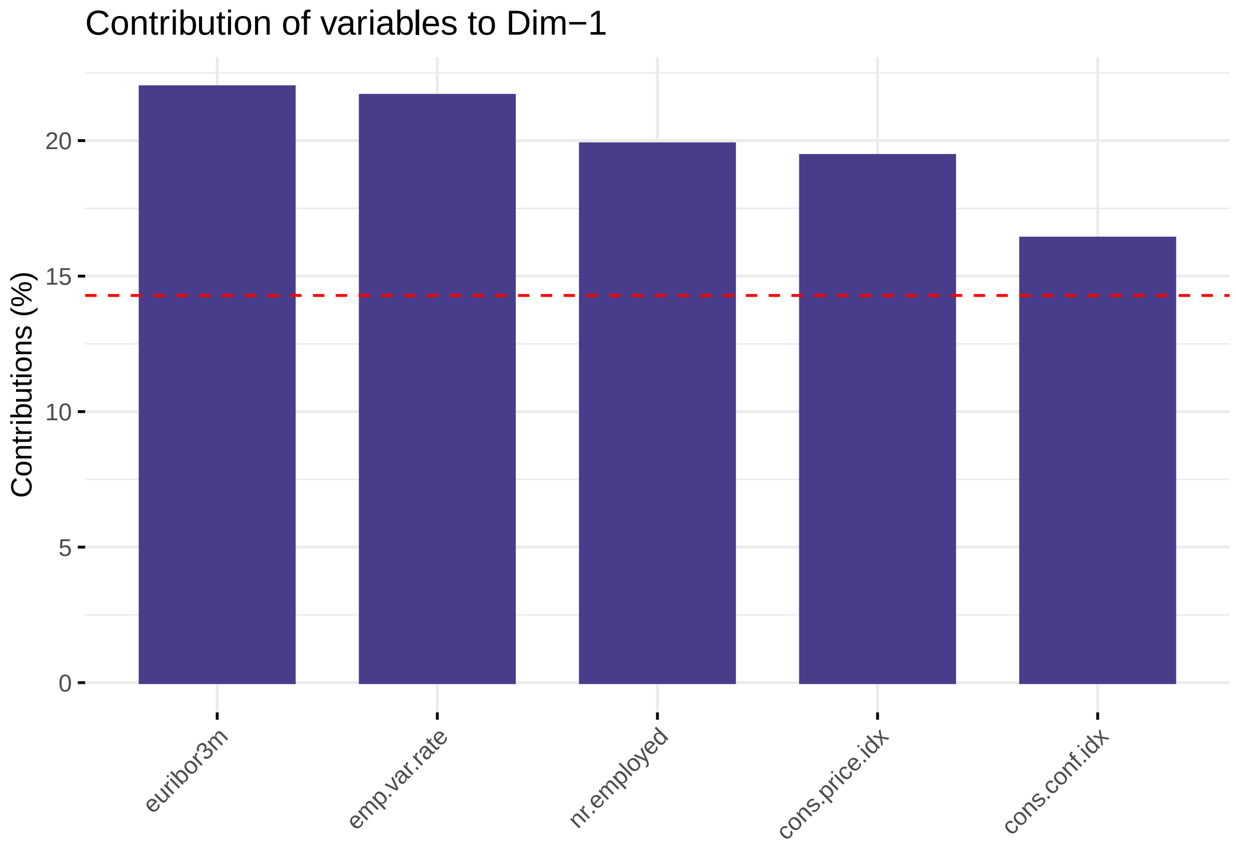


## Interpreting the axes

2.1.1 Dim1

We see that in the first dimension, the variables: euribor, emp.var.rate,nr.employed,cons.price.idx and cons.conf.idx are very contributive to the dimension and we can see that all relates to the economy, so we should name the ax as economic status. We can see that all of them are positively correlated such that as all of their values grow, the other variables will follow.

```
res.des<-dimdesc(res.pca)
fviz_contrib(  # contributions of variables to PC1
  res.pca,
  fill = "darkslateblue",
  color = "darkslateblue",
  choice = "var",
  axes = 1,
  top = 5)
```



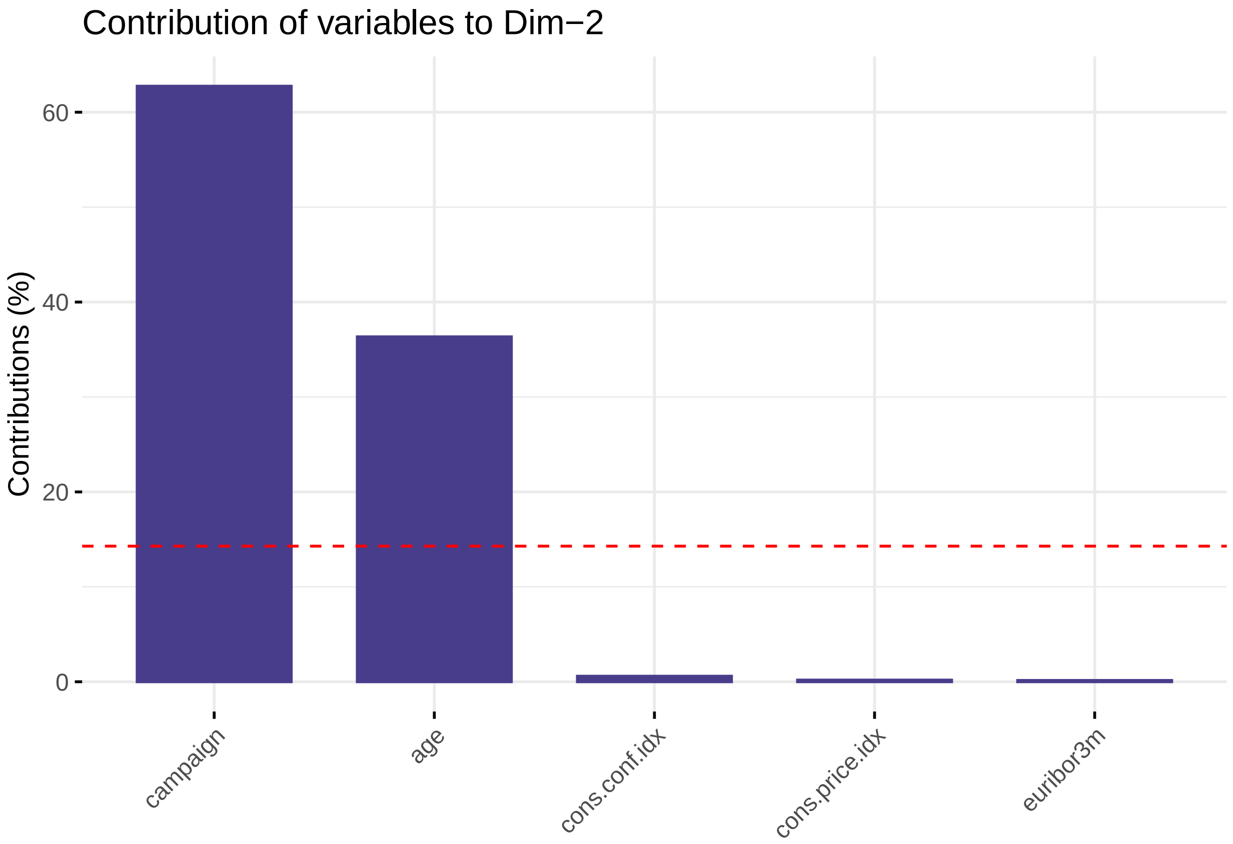
```r
res.des$Dim.1
```

```
##
## Link between the variable and the continuous variables (R-square)
## =================================================================================
##                correlation      p.value
## euribor3m       0.99263021 0.000000e+00
## emp.var.rate    0.98545330 0.000000e+00
## nr.employed     0.94389664 0.000000e+00
## cons.price.idx  0.93366222 0.000000e+00
## cons.conf.idx   0.85735341 0.000000e+00
## age             0.12353703 1.838123e-18
## campaign        0.10813127 1.766000e-14
## duration       -0.07640653 6.326626e-08
```

2.1.2 Dim2

In this dimension we see that the only variables that contribute significantly are campaign and age, since we think that campaign is the most relevant feature, we should name this as campaign calls, and it tells us that the older the person the more calls the person will receive.

```r
res.des<-dimdesc(res.pca)
fviz_contrib(  # contributions of variables to PC1
  res.pca,
  fill = "darkslateblue",
  color = "darkslateblue",
  choice = "var",
  axes = 2,
  top = 5)
```



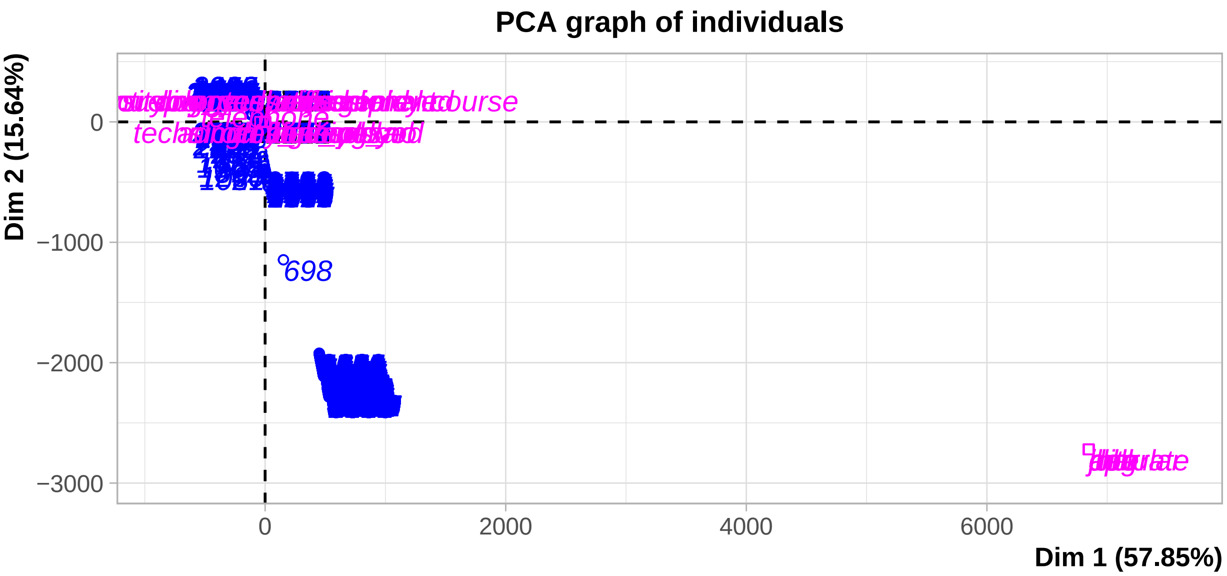
```
res.des$Dim.2
```

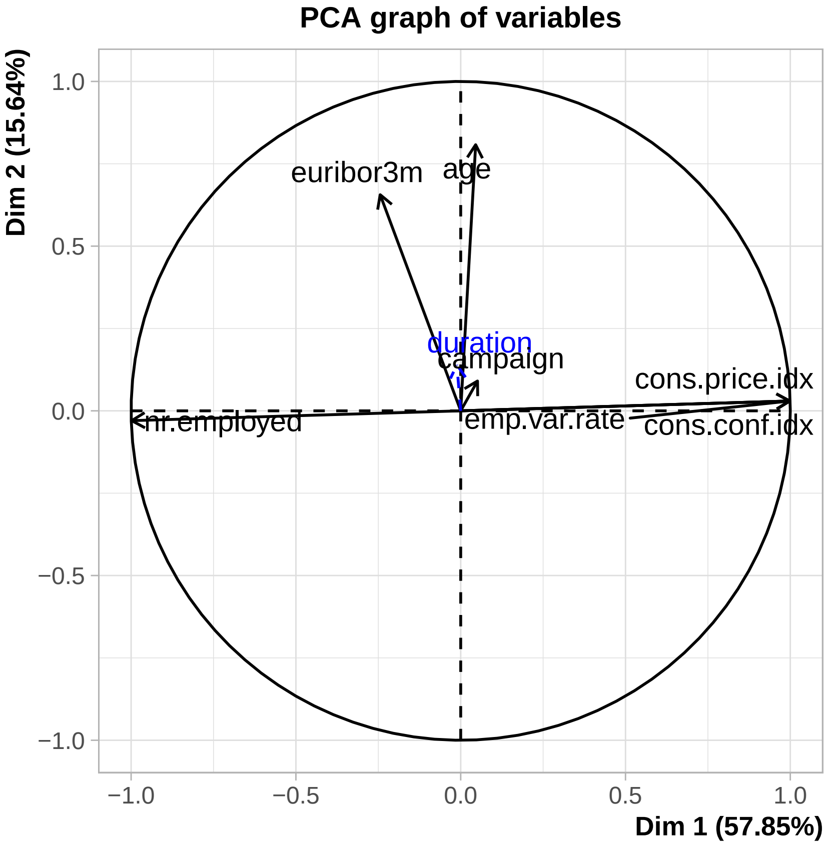
```
##
## Link between the variable and the continuous variables (R-square)
## =================================================================================
##                 correlation      p.value
## campaign         0.79578843 0.000000e+00
## age              0.60564701 0.000000e+00
## duration         0.07248939 2.878338e-07
## euribor3m       -0.03729857 8.347880e-03
## cons.price.idx -0.04214843 2.873808e-03
## cons.conf.idx  -0.07673524 5.552160e-08
```

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

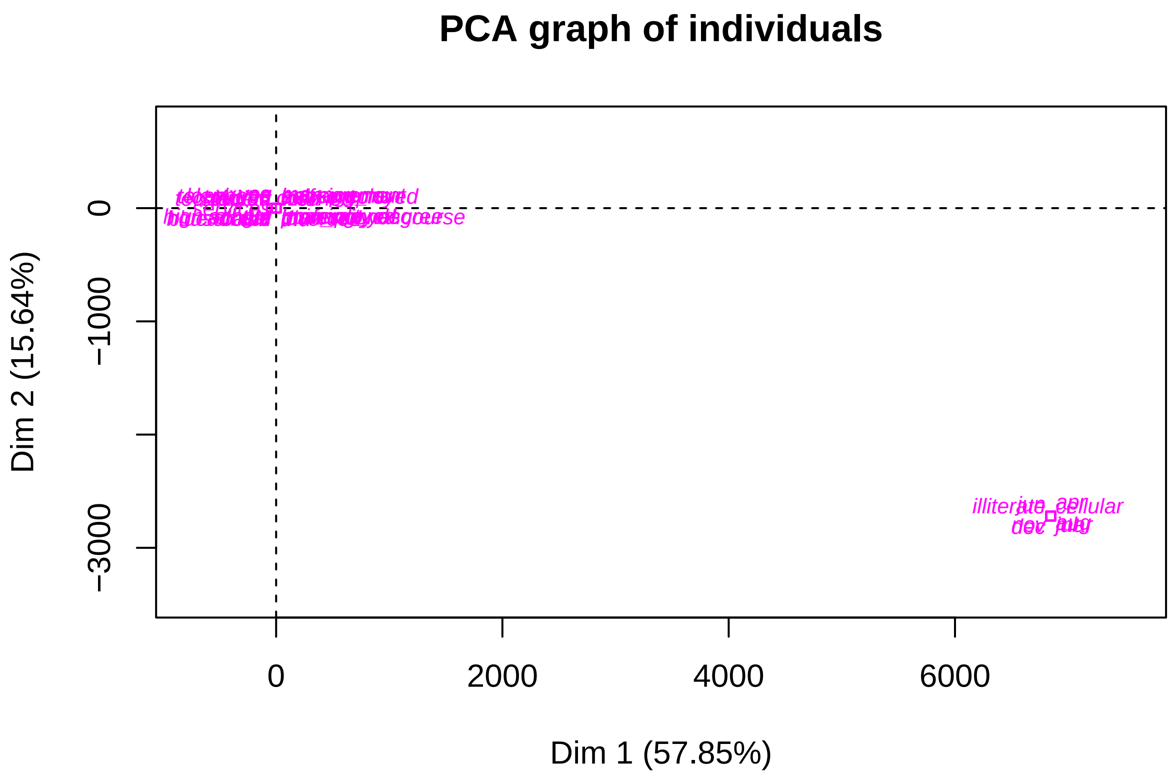
We will perform the PCA taking into account all the supplementary variables. The first we've gotten doesn't really make any sense, if we take a look at the PCA graph, we see that age is strongly correlated to euribor but we know for a fact that the euribor rates are totally independent from individuals, we can also see that the nr.employed are completely negatevely correlated to all the other economic variables, which doesn't make a lot of sense either, for example, it is counterintuitive to think that the higher the number of employed people, the lower the employment variation rate.

```
ll <- which( df$mout == "YesMOut")
res.pca_sup<-PCA(df[,c(vars_res, vars_con,vars_dis)],quali.sup=c(1,10:17),quanti.sup= c(2), ind.sup = l
```





```
plot(res.pca_sup, choix="ind",invisible=c("ind","ind.sup"), cex=0.7, graph.type = "classic")
```



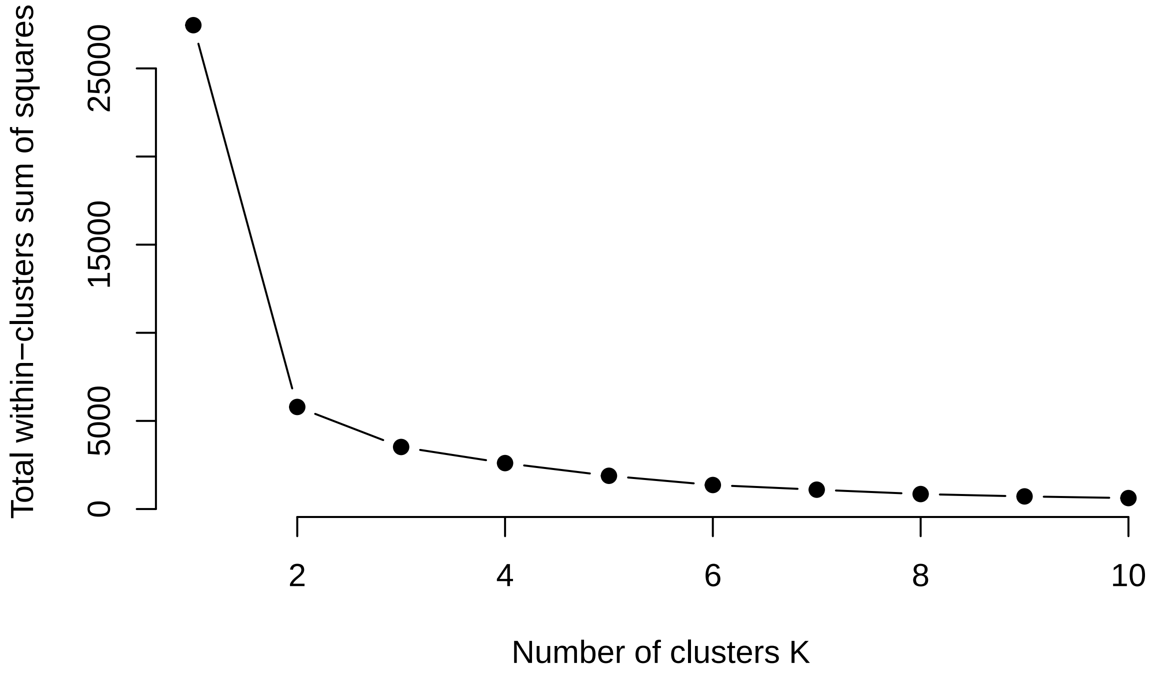
3 KMEANS

From this graph, we apply Elbow's method which reveals that the optimal number of clusters is 4. So we will proceed to model and interpret the kMEANs with 4 clusters.

```
dclu<- res.pca$ind$coord[,1:2]; # los dos ejes
k.max <- 10
wss <- sapply(1:k.max,
              function(k){kmeans(dclu, k, nstart=50,iter.max = 15 )$tot.withinss})
wss
```

```
##  [1] 27452.2639  5791.5810  3520.8784  2606.4789  1886.5163  1362.8296
##  [7]  1097.9665   849.8887   717.2235   622.5091
```

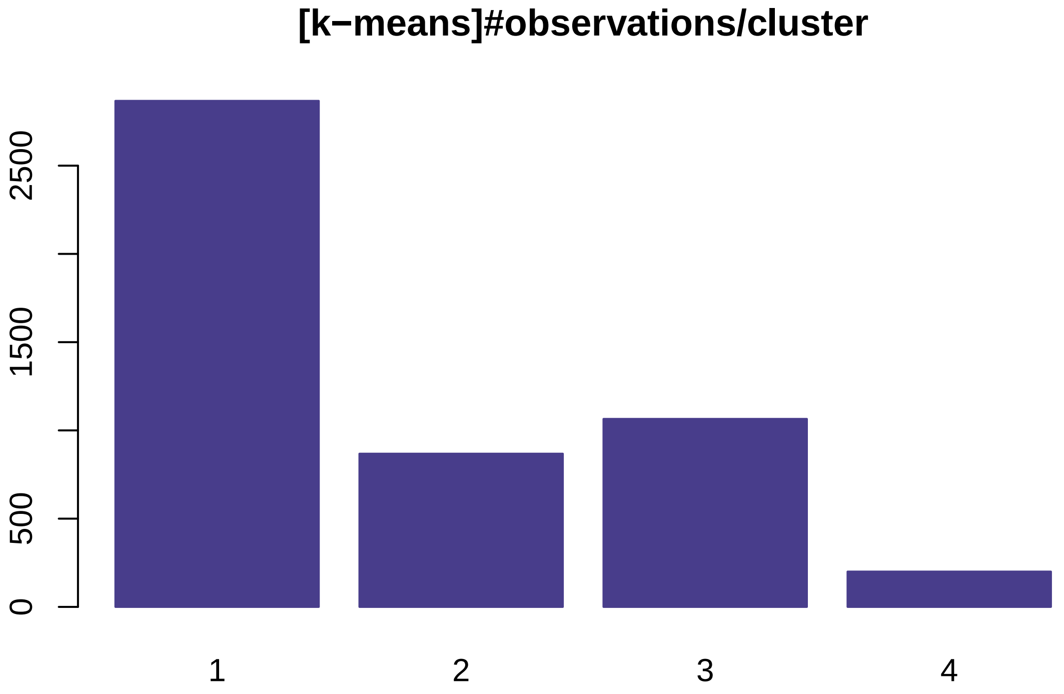
```
plot(1:k.max, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
```



```
  # coordenates are real - Euclidean metric
dist<-dist(dclu)
kc<-kmeans(dist, 4) #caclulate the distances, it turns into a matrix
```

We can see in this graph the distribution of individuals within each cluster.

```
df$claKM<-0
df$claKM<-kc$cluster
df$claKM<-factor(df$claKM)
barplot(table(df$claKM),col="darkslateblue",border="darkslateblue",main="[k-means]#observations/cluster
```



3.1 Interpret the results of the classification

3.1.1 The description of the clusters by the variables

- Cluster 1:
  - These are the people who will say yes to the campaign and being contacted by cellular, these people are young around the ages of 20-30 and are still students.
- Cluster 2:
  - The are people whho are more likely to say yes when they are being contacted on november by cellular.
- Cluster 3:
  - These are people who are being contacted by telephone and their response will be a no for the campaign, these people are more towards adults profiles from ages 30-50 and don't have higher degrees.
- Cluster 4:
  - These are people who are frequently contacted by campaign and are around the ages of 40-60 who will say no to a campaign.

We can see that there are two groups within the clusters, the cluster of people who will say yes and a cluster of people who will say no, in clusters 1-2 and 3-4 respectevely. So the campaign will be more succesful if it focuses on poeple of cluster's 1-2.

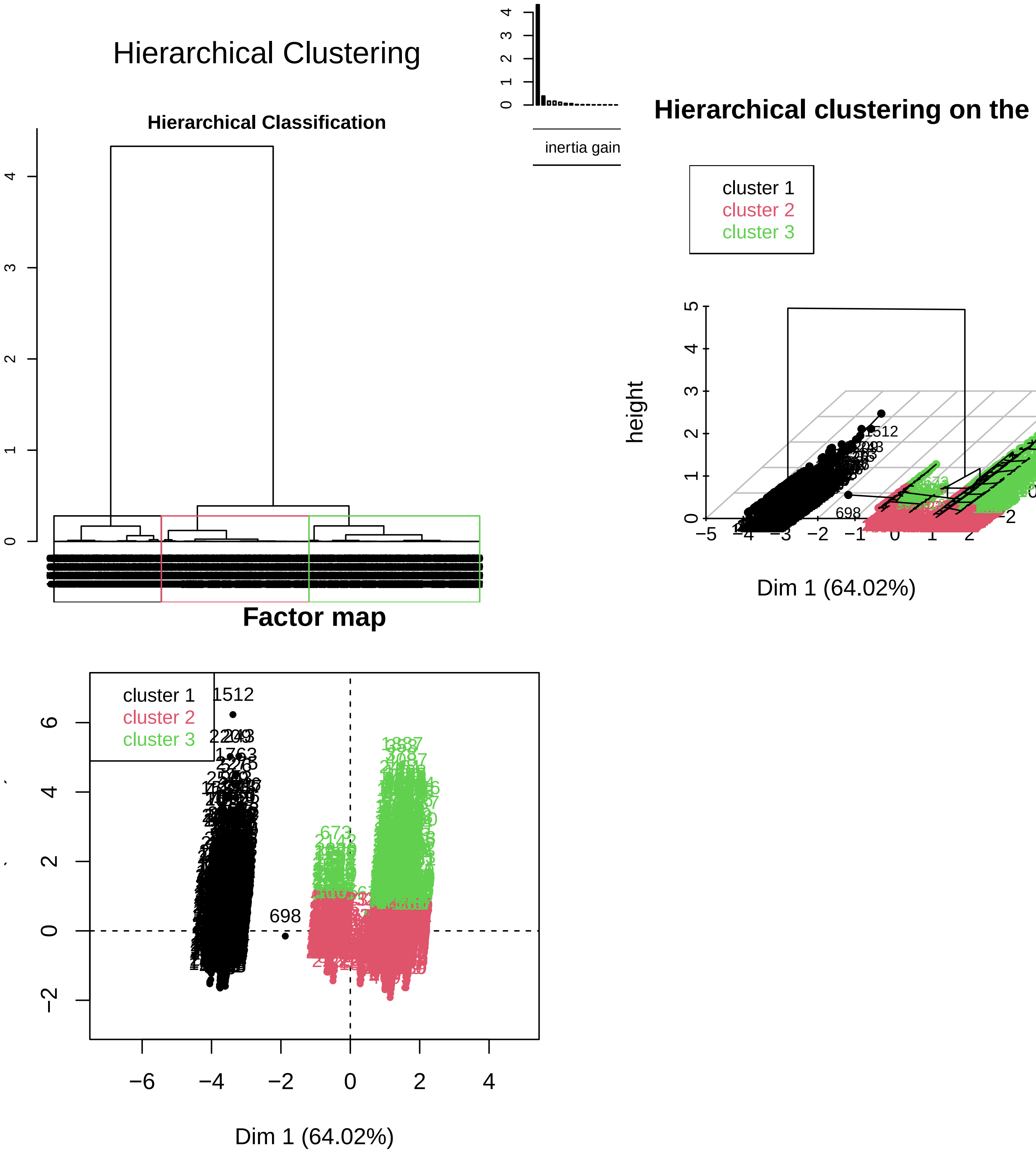
4 Hierarchical clustering

We've decided that numbers of cluster is the one that the algorithm gives us, with nb.clust=-1. ## Description of Clusters

- Cluster 1:
  - These are the people who will say yes to the campaign and being contacted by cellulars, and mostly single university graduates. Also, they are being called during the months of april and may, which are nearing summer seasons and these kind of people tend to have money saved.
- Cluster 2:
  - The are people who are more likely to say no when they are being contacted on november by cellular, cluster similar to the one in KMEANS. These people are divorced and don't have any housing loan.
- Cluster 3:
  - These are people who are married and retired which will most likely say no, and are most usually contacted by telephone. We see that these are people who have their life together already and aren't interested in these kind of campaigns anymore.

Following these trends, the company should focus specilly on people with chharacteristics similir to that of cluster 1. We can see something in common with these two methods which is that younger educated people are more likely to say yes.

```
res.pca<-PCA(df[,c('duration',vars_con,vars_dis,"y")],quanti.sup=1,quali.sup = c(9:17), ncp=2, graph=FA
res.hcpc<-HCPC(res.pca,order=TRUE, nb.clust = -1)
```



```
attributes(res.hcpc)
```

```
## $names
## [1] "data.clust" "desc.var"   "desc.axes"  "desc.ind"   "call"
##
## $class
## [1] "HCPC"
```

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

```
##     duration           age           campaign      emp.var.rate
##  Min.   :   4.0   Min.   :18.00   Min.   :1.000   Min.   :-2.9000
##  1st Qu.: 175.0   1st Qu.:32.00   1st Qu.:1.000   1st Qu.:-1.8000
##  Median : 342.0   Median :38.00   Median :2.000   Median : 1.1000
##  Mean   : 479.8   Mean   :39.96   Mean   :2.006   Mean   : 0.3275
##  3rd Qu.: 686.0   3rd Qu.:47.00   3rd Qu.:2.033   3rd Qu.: 1.1000
##  Max.   :4199.0   Max.   :88.00   Max.   :8.000   Max.   : 1.4000
##
##  cons.price.idx  cons.conf.idx      euribor3m      nr.employed
##  Min.   :92.76   Min.   :-50.00   Min.   :1.244   Min.   :5076
##  1st Qu.:93.08   1st Qu.:-42.70   1st Qu.:1.811   1st Qu.:5099
##  Median :93.99   Median :-36.40   Median :4.856   Median :5191
##  Mean   :93.68   Mean   :-39.81   Mean   :3.965   Mean   :5174
##  3rd Qu.:93.99   3rd Qu.:-36.40   3rd Qu.:4.857   3rd Qu.:5191
##  Max.   :94.47   Max.   :-36.10   Max.   :5.045   Max.   :5228
##
##           job          marital                   education
##  blue-collar:1248   divorced: 530   basic              :1736
##  admin.     :1214   married :3103   high.school        :1169
##  technician : 757   single  :1367   illiterate         :   2
##  services   : 517                   professional.course: 631
##  management : 380                   university.degree  :1462
##  retired    : 192
##  (Other)    : 692
##         housing           loan           contact         month      day_of_week
##  housing_no :2442   loan_no :4278   cellular :2007   may    :3164   fri: 849
##  housing_yes:2558   loan_yes: 722   telephone:2993   apr    : 442   mon:1107
##                                                      jul    : 407   thu:1029
##                                                      jun    : 357   tue:1073
##                                                      aug    : 271   wed: 942
##                                                      nov    : 190
##                                                      (Other): 169
##      y        clust
##  y_no :2400   1:1262
##  y_yes:2600   2:2477
##               3:1261
##
##
##
##
```

```
attributes(res.hcpc$desc.var)
```

```
## $names
## [1] "test.chi2"  "category"   "quanti.var" "quanti"     "call"
##
## $class
## [1] "catdes" "list"
```

```r
# Factors globally related to clustering partition
res.hcpc$desc.var$test.chi2
```

```
##                  p.value df
## contact     0.000000e+00  2
## month       0.000000e+00 16
## y           0.000000e+00  2
## job         4.367799e-56 20
## marital     1.393485e-55  4
## education   5.162811e-17  8
## day_of_week 6.022360e-14  8
## housing     1.331316e-11  2
```

```r
# Categories over/under represented in each cluster
res.hcpc$desc.var$category
```

```
## $`1`
##                                Cla/Mod    Mod/Cla Global       p.value
## y=y_yes                      48.538462 100.000000  52.00  0.000000e+00
## contact=cellular             60.139512  95.641838  40.14  0.000000e+00
## month=apr                   100.000000  35.023772   8.84 1.337796e-294
## month=mar                   100.000000   9.984152   2.52  3.535437e-78
## marital=single               37.820044  40.966719  27.34  1.776329e-34
## month=jun                    47.338936  13.391442   7.14  6.958630e-21
## job=student                  67.647059   5.467512   2.04  1.214952e-19
## education=university.degree  32.831737  38.034865  29.24  5.437955e-15
## housing=housing_yes          29.124316  59.033281  51.16  8.970164e-11
## day_of_week=thu              32.555879  26.545166  20.58  2.862929e-09
## job=admin.                   31.136738  29.952456  24.28  8.689279e-08
## job=retired                  39.062500   5.942948   3.84  1.759644e-05
## job=management               21.052632   6.339144   7.60  4.795407e-02
## day_of_week=fri              22.497055  15.134707  16.98  4.206311e-02
## job=housemaid                15.384615   1.426307   2.34  9.955469e-03
## day_of_week=tue              21.528425  18.304279  21.46  1.411157e-03
## job=services                 19.535783   8.003170  10.34  1.282339e-03
## month=oct                     0.000000   0.000000   0.84  4.663341e-06
## housing=housing_no           21.171171  40.966719  48.84  8.970164e-11
## job=blue-collar              18.028846  17.828843  24.96  4.079496e-12
## education=basic              19.297235  26.545166  34.72  8.369367e-13
## marital=married              20.270706  49.841521  62.06  1.241676e-24
## month=nov                     0.000000   0.000000   3.80  2.848071e-25
## month=aug                     0.000000   0.000000   5.42  4.405850e-36
## month=jul                     0.000000   0.000000   8.14  9.888672e-55
## month=may                    16.561315  41.521395  63.28  9.308480e-75
## y=y_no                        0.000000   0.000000  48.00  0.000000e+00
## contact=telephone             1.837621   4.358162  59.86  0.000000e+00
##                                 v.test
## y=y_yes                            Inf
## contact=cellular                   Inf
## month=apr                    36.683515
## month=mar                    18.717943
## marital=single               12.245476
## month=jun                     9.374376
## job=student                   9.067754
## education=university.degree   7.816344
## housing=housing_yes           6.483361
## day_of_week=thu               5.939271
## job=admin.                    5.352197
## job=retired                   4.293391
## job=management               -1.977775
## day_of_week=fri              -2.032895
## job=housemaid                -2.577372
## day_of_week=tue              -3.192359
## job=services                 -3.219903
## month=oct                    -4.579390
## housing=housing_no           -6.483361
## job=blue-collar              -6.934400
## education=basic              -7.154966
## marital=married             -10.245355
## month=nov                   -10.386780
## month=aug                   -12.541851
## month=jul                   -15.580430
## month=may                   -18.293586
## y=y_no                            -Inf
## contact=telephone                 -Inf
##
## $`2`
##                              Cla/Mod    Mod/Cla Global       p.value     v.test
## contact=telephone           65.72001 79.4105773  59.86 1.675288e-177  28.406616
## y=y_no                      67.20833 65.1190957  48.00 8.887554e-130  24.237822
## month=may                   56.03666 71.5785224  63.28  1.075478e-33  12.098508
## month=nov                   82.10526  6.2979411   3.80  2.811309e-21   9.469521
## month=oct                   97.61905  1.6552281   0.84  5.879586e-12   6.882537
## housing=housing_no          53.93120 53.1691562  48.84  1.286627e-09   6.069058
## job=blue-collar             55.52885 27.9773920  24.96  1.031806e-06   4.885474
## month=jul                   60.19656  9.8909972   8.14  7.028006e-06   4.492839
## day_of_week=tue             54.98602 23.8191361  21.46  5.678910e-05   4.025770
## job=services                57.25338 11.9499394  10.34  2.111521e-04   3.705286
## education=high.school       53.63559 25.3128785  23.38  1.380871e-03   3.198620
## month=aug                   55.35055  6.0557126   5.42  4.947934e-02   1.964438
## marital=divorced            44.33962  9.4872830  10.60  1.133578e-02  -2.532174
## day_of_week=mon             44.98645 20.1049657  22.14  5.936734e-04  -3.434488
## education=university.degree 45.07524 26.6047638  29.24  4.914743e-05  -4.059645
## job=student                 29.41176  1.2111425   2.04  3.389517e-05  -4.145582
## housing=housing_yes         45.34793 46.8308438  51.16  1.286627e-09  -6.069058
## month=jun                   31.37255  4.5215987   7.14  6.194346e-13  -7.196135
## job=retired                 11.97917  0.9285426   3.84  2.957509e-29 -11.228411
## month=mar                    0.00000  0.0000000   2.52  7.630513e-39 -13.036054
## y=y_yes                     33.23077 34.8809043  52.00 8.887554e-130 -24.237822
## month=apr                    0.00000  0.0000000   8.84 3.785117e-141 -25.293301
## contact=cellular            25.41106 20.5894227  40.14 1.675288e-177 -28.406616
##
## $`3`
##                               Cla/Mod     Mod/Cla Global      p.value
## contact=telephone           32.442366 77.00237906  59.86 4.062364e-49
## y=y_no                      32.791667 62.41078509  48.00 1.642002e-32
## marital=married             29.745408 73.19587629  62.06 9.704644e-22
## job=retired                 48.958333  7.45440127   3.84 5.519471e-13
## month=aug                    44.649446  9.59555908   5.42 9.154379e-13
## month=jul                    39.803440 12.84694687   8.14 1.437399e-11
## day_of_week=mon              31.707317 27.83505155  22.14 3.140557e-08
## education=basic              29.550691 40.68199841  34.72 3.357623e-07
## month=may                    27.402023 68.75495638  63.28 2.638969e-06
## marital=divorced             33.773585 14.19508327  10.60 3.121665e-06
## job=management               32.368421  9.75416336   7.60 1.125170e-03
## job=housemaid                35.897436  3.33068993   2.34 9.519927e-03
## month=nov                    17.894737  2.69627280   3.80 1.500201e-02
## job=unemployed               15.702479  1.50674068   2.42 1.146364e-02
## education=high.school        21.813516 20.22204600  23.38 1.984436e-03
## education=university.degree 22.093023 25.61459159  29.24 9.774983e-04
## job=admin.                   21.087315 20.30134814  24.28 1.151868e-04
## month=oct                     2.380952  0.07930214   0.84 7.697880e-05
## day_of_week=thu              19.144801 15.62252181  20.58 2.716028e-07
## job=student                   2.941176  0.23790642   2.04 7.924325e-10
## month=mar                     0.000000  0.00000000   2.52 7.272218e-17
## y=y_yes                      18.230769 37.58921491  52.00 1.642002e-32
## marital=single               11.631309 12.60904044  27.34 1.547322e-46
## contact=cellular             14.449427 22.99762094  40.14 4.062364e-49
## month=apr                     0.000000  0.00000000   8.84 1.396587e-59
##                                 v.test
## contact=telephone             14.731231
## y=y_no                        11.872641
## marital=married                9.579998
## job=retired                    7.211855
## month=aug                      7.142657
## month=jul                      6.754085
## day_of_week=mon                5.533415
## education=basic                5.102180
## month=may                      4.697088
## marital=divorced               4.662648
## job=management                 3.257200
## job=housemaid                  2.592796
## month=nov                     -2.432330
## job=unemployed                -2.528239
## education=high.school         -3.092552
## education=university.degree  -3.296924
## job=admin.                    -3.856150
## month=oct                     -3.953616
## day_of_week=thu               -5.142156
## job=student                   -6.146436
## month=mar                     -8.342523
## y=y_yes                      -11.872641
## marital=single               -14.324094
## contact=cellular             -14.731231
## month=apr                    -16.278765
```

```
### desc.ind ###
### C. The description of the clusters by the individuals ###
```

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

```
## $`1`
##                                Cla/Mod    Mod/Cla Global       p.value
## y=y_yes                      48.538462 100.000000  52.00  0.000000e+00
## contact=cellular             60.139512  95.641838  40.14  0.000000e+00
## month=apr                   100.000000  35.023772   8.84 1.337796e-294
## month=mar                   100.000000   9.984152   2.52  3.535437e-78
## marital=single               37.820044  40.966719  27.34  1.776329e-34
## month=jun                    47.338936  13.391442   7.14  6.958630e-21
## job=student                  67.647059   5.467512   2.04  1.214952e-19
## education=university.degree  32.831737  38.034865  29.24  5.437955e-15
## housing=housing_yes          29.124316  59.033281  51.16  8.970164e-11
## day_of_week=thu              32.555879  26.545166  20.58  2.862929e-09
## job=admin.                   31.136738  29.952456  24.28  8.689279e-08
## job=retired                  39.062500   5.942948   3.84  1.759644e-05
## job=management               21.052632   6.339144   7.60  4.795407e-02
## day_of_week=fri              22.497055  15.134707  16.98  4.206311e-02
## job=housemaid                15.384615   1.426307   2.34  9.955469e-03
## day_of_week=tue              21.528425  18.304279  21.46  1.411157e-03
## job=services                 19.535783   8.003170  10.34  1.282339e-03
## month=oct                     0.000000   0.000000   0.84  4.663341e-06
## housing=housing_no           21.171171  40.966719  48.84  8.970164e-11
## job=blue-collar              18.028846  17.828843  24.96  4.079496e-12
## education=basic              19.297235  26.545166  34.72  8.369367e-13
## marital=married              20.270706  49.841521  62.06  1.241676e-24
## month=nov                     0.000000   0.000000   3.80  2.848071e-25
## month=aug                     0.000000   0.000000   5.42  4.405850e-36
## month=jul                     0.000000   0.000000   8.14  9.888672e-55
## month=may                    16.561315  41.521395  63.28  9.308480e-75
## y=y_no                        0.000000   0.000000  48.00  0.000000e+00
## contact=telephone             1.837621   4.358162  59.86  0.000000e+00
##                                 v.test
## y=y_yes                            Inf
## contact=cellular                   Inf
## month=apr                    36.683515
## month=mar                    18.717943
## marital=single               12.245476
## month=jun                     9.374376
## job=student                   9.067754
## education=university.degree   7.816344
## housing=housing_yes           6.483361
## day_of_week=thu               5.939271
## job=admin.                    5.352197
## job=retired                   4.293391
## job=management               -1.977775
## day_of_week=fri              -2.032895
## job=housemaid                -2.577372
## day_of_week=tue              -3.192359
## job=services                 -3.219903
## month=oct                    -4.579390
## housing=housing_no           -6.483361
## job=blue-collar              -6.934400
## education=basic              -7.154966
## marital=married             -10.245355
## month=nov                   -10.386780
## month=aug                   -12.541851
## month=jul                   -15.580430
## month=may                   -18.293586
## y=y_no                            -Inf
## contact=telephone                 -Inf
##
## $`2`
##                              Cla/Mod    Mod/Cla Global       p.value     v.test
## contact=telephone           65.72001 79.4105773  59.86 1.675288e-177  28.406616
## y=y_no                      67.20833 65.1190957  48.00 8.887554e-130  24.237822
## month=may                   56.03666 71.5785224  63.28  1.075478e-33  12.098508
## month=nov                   82.10526  6.2979411   3.80  2.811309e-21   9.469521
## month=oct                   97.61905  1.6552281   0.84  5.879586e-12   6.882537
## housing=housing_no          53.93120 53.1691562  48.84  1.286627e-09   6.069058
## job=blue-collar             55.52885 27.9773920  24.96  1.031806e-06   4.885474
## month=jul                   60.19656  9.8909972   8.14  7.028006e-06   4.492839
## day_of_week=tue             54.98602 23.8191361  21.46  5.678910e-05   4.025770
## job=services                57.25338 11.9499394  10.34  2.111521e-04   3.705286
## education=high.school       53.63559 25.3128785  23.38  1.380871e-03   3.198620
## month=aug                   55.35055  6.0557126   5.42  4.947934e-02   1.964438
## marital=divorced            44.33962  9.4872830  10.60  1.133578e-02  -2.532174
## day_of_week=mon             44.98645 20.1049657  22.14  5.936734e-04  -3.434488
## education=university.degree 45.07524 26.6047638  29.24  4.914743e-05  -4.059645
## job=student                 29.41176  1.2111425   2.04  3.389517e-05  -4.145582
## housing=housing_yes         45.34793 46.8308438  51.16  1.286627e-09  -6.069058
## month=jun                   31.37255  4.5215987   7.14  6.194346e-13  -7.196135
## job=retired                 11.97917  0.9285426   3.84  2.957509e-29 -11.228411
## month=mar                    0.00000  0.0000000   2.52  7.630513e-39 -13.036054
## y=y_yes                     33.23077 34.8809043  52.00 8.887554e-130 -24.237822
## month=apr                    0.00000  0.0000000   8.84 3.785117e-141 -25.293301
## contact=cellular            25.41106 20.5894227  40.14 1.675288e-177 -28.406616
##
## $`3`
##                               Cla/Mod     Mod/Cla Global      p.value
## contact=telephone           32.442366 77.00237906  59.86 4.062364e-49
## y=y_no                      32.791667 62.41078509  48.00 1.642002e-32
## marital=married             29.745408 73.19587629  62.06 9.704644e-22
## job=retired                 48.958333  7.45440127   3.84 5.519471e-13
## month=aug                   44.649446  9.59555908   5.42 9.154379e-13
## month=jul                   39.803440 12.84694687   8.14 1.437399e-11
## day_of_week=mon             31.707317 27.83505155  22.14 3.140557e-08
## education=basic             29.550691 40.68199841  34.72 3.357623e-07
## month=may                   27.402023 68.75495638  63.28 2.638969e-06
## marital=divorced            33.773585 14.19508327  10.60 3.121665e-06
## job=management              32.368421  9.75416336   7.60 1.125170e-03
## job=housemaid               35.897436  3.33068993   2.34 9.519927e-03
## month=nov                   17.894737  2.69627280   3.80 1.500201e-02
## job=unemployed              15.702479  1.50674068   2.42 1.146364e-02
## education=high.school       21.813516 20.22204600  23.38 1.984436e-03
## education=university.degree 22.093023 25.61459159  29.24 9.774983e-04
## job=admin.                  21.087315 20.30134814  24.28 1.151868e-04
## month=oct                    2.380952  0.07930214   0.84 7.697880e-05
## day_of_week=thu             19.144801 15.62252181  20.58 2.716028e-07
## job=student                  2.941176  0.23790642   2.04 7.924325e-10
## month=mar                    0.000000  0.00000000   2.52 7.272218e-17
## y=y_yes                     18.230769 37.58921491  52.00 1.642002e-32
## marital=single              11.631309 12.60904044  27.34 1.547322e-46
## contact=cellular            14.449427 22.99762094  40.14 4.062364e-49
## month=apr                    0.000000  0.00000000   8.84 1.396587e-59
##                                v.test
## contact=telephone            14.731231
## y=y_no                       11.872641
## marital=married               9.579998
## job=retired                   7.211855
## month=aug                     7.142657
## month=jul                     6.754085
## day_of_week=mon               5.533415
## education=basic               5.102180
## month=may                     4.697088
## marital=divorced              4.662648
## job=management                3.257200
## job=housemaid                 2.592796
## month=nov                    -2.432330
## job=unemployed               -2.528239
## education=high.school        -3.092552
## education=university.degree  -3.296924
## job=admin.                   -3.856150
## month=oct                    -3.953616
## day_of_week=thu              -5.142156
## job=student                  -6.146436
## month=mar                    -8.342523
## y=y_yes                     -11.872641
## marital=single              -14.324094
## contact=cellular            -14.731231
## month=apr                   -16.278765
```

5 CA

We will cut the duration, which is the numerical target into 8 levels. We will study the CA obtained from the Duration-Age_group and then Duration-education. We want to see this because in the clustering findings we discovered that young educated people are more likely to say yes, so we want to see if it affects the duration as well.

5.1 Eigenvalues and dominant axes(1)

We can see that independence test fails to refute H0 since the p-value= 0.3263>0.05, so there is no independence between duration and age. We can see that the farthest value is 10-20 from age which makes sense since teens aren't likely to be contacted.Since all the other values are around the center we can see that the duration is dependent on the age group(mostly).

```
aux2<-c(5,60,120,150,180,240,300,1200,2100)
duration_fact<-factor(cut(df$duration,breaks=aux2,include.lowest=T))
table(duration_fact)
```

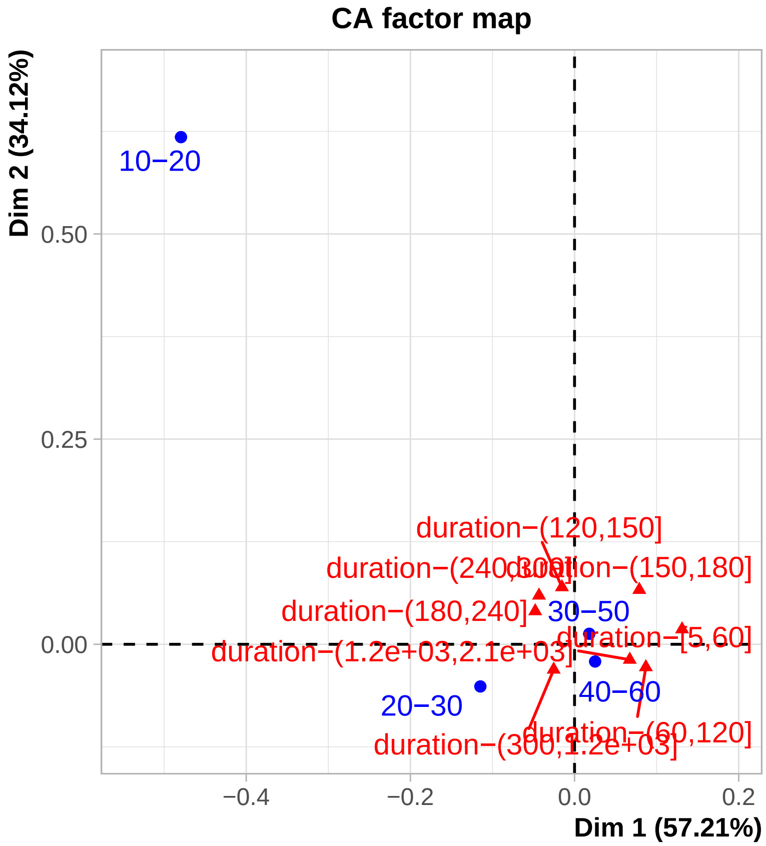
```
## duration_fact
##            [5,60]          (60,120]         (120,150]         (150,180]
##               175               493               321               330
##         (180,240]         (240,300]     (300,1.2e+03] (1.2e+03,2.1e+03]
##               543               422              2415               274
```

```
levels(duration_fact)<-paste0("duration-",levels(duration_fact))
df$duration_fact<-duration_fact

tt<-table(df[,c("Age_group","duration_fact")])
chisq.test(tt,  simulate.p.value = TRUE) #to see if the rows and columns are independents. H0: Rows and
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tt
## X-squared = 23.401, df = NA, p-value = 0.3048
```

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



The mean of eigenvalues = 0.001606341 making that only the first 2 dimensions satisfies Kaiser's criteria. So the dominant axes are 1 and 2 with a cummulative variance of 91.3%.

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

```
## [1] 0.001606341
```

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

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 23.40117 (p-value =  0.3229672 
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3
## Variance               0.003   0.002   0.000
## % of var.             57.211  34.123   8.666
## Cumulative % of var.  57.211  91.334 100.000
##
## Rows
##                              Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## 10-20                       |     1.800 | -0.479 24.043  0.368 |  0.618 66.952
## 20-30                       |     2.121 | -0.115 63.923  0.831 | -0.051 21.491
## 30-50                       |     0.402 |  0.018  8.018  0.550 |  0.013  6.843
## 40-60                       |     0.496 |  0.025  4.017  0.223 | -0.021  4.713
##                              cos2    Dim.3    ctr   cos2
## 10-20                       0.612 |  0.112  8.717  0.020 |
## 20-30                       0.167 | -0.006  1.200  0.002 |
## 30-50                       0.280 | -0.010 16.440  0.171 |
## 40-60                       0.156 |  0.042 73.642  0.620 |
##
## Columns
##                              Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## duration-[5,60]             |     0.628 |  0.131 22.057  0.968 |  0.018  0.722
## duration-(60,120]           |     0.987 |  0.087 27.301  0.763 | -0.028  4.702
## duration-(120,150]          |     0.327 | -0.015  0.543  0.046 |  0.070 18.757
## duration-(150,180]          |     0.777 |  0.079 14.749  0.524 |  0.066 17.466
## duration-(180,240]          |     0.437 | -0.048  8.975  0.566 |  0.040 10.664
## duration-(240,300]          |     0.582 | -0.043  5.814  0.275 |  0.059 18.359
## duration-(300,1.2e+03]      |     0.790 | -0.025 11.396  0.398 | -0.031 28.135
## duration-(1.2e+03,2.1e+03] |     0.291 |  0.067  9.165  0.869 | -0.019  1.195
##                              cos2    Dim.3    ctr   cos2
## duration-[5,60]             0.019 | -0.015  1.906  0.013 |
## duration-(60,120]           0.078 |  0.040 37.519  0.159 |
## duration-(120,150]          0.942 |  0.008  0.974  0.012 |
## duration-(150,180]          0.370 | -0.036 19.833  0.107 |
## duration-(180,240]          0.401 | -0.011  3.388  0.032 |
## duration-(240,300]          0.518 |  0.037 28.764  0.206 |
## duration-(300,1.2e+03]      0.585 | -0.005  3.230  0.017 |
## duration-(1.2e+03,2.1e+03]  0.068 | -0.018  4.385  0.063 |
```

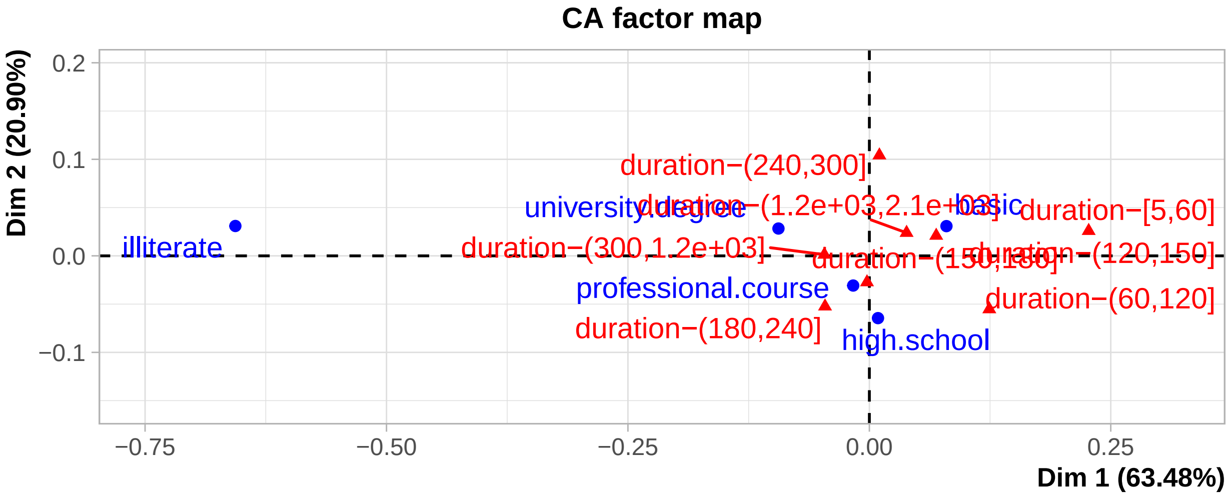
5.2 Eigenvalues and dominant axes(2)

We can see that independence test fails to refute H0 since the p-value=0.09445>0.05, so there is no independence between duration and education. From the factor map we can see that the farthest value is illiterate and the other values are really near from each other indicating that there is some dependence between them.

```
tt<-table(df[,c("education","duration_fact")])
chisq.test(tt,  simulate.p.value = TRUE) #to see if the rows and colum
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tt
## X-squared = 39.394, df = NA, p-value = 0.09345
```

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



The mean of eigenvalues = 0.001980396 making that only the first 3 dimensions satisfies Kaiser's criteria. So the dominant axes are 1 and 2 with a cummulative variance of 97.7%.

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

```
## [1] 0.001980396
```

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

```
##
## Call:
## CA(X = tt)
##
## The chi square of independence between the two variables is equal to 39.39403 (p-value =  0.0747902
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4
## Variance               0.005   0.002   0.001   0.000
## % of var.             63.476  20.902  13.305   2.317
## Cumulative % of var.  63.476  84.378  97.683 100.000
##
## Rows
##                             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## basic                     |     2.551 |  0.080 44.065  0.868 |  0.031 19.904
## high.school               |     1.165 |  0.009  0.372  0.016 | -0.065 58.748
## illiterate                |     0.426 | -0.657  3.447  0.407 |  0.031  0.023
## professional.course       |     0.938 | -0.017  0.705  0.038 | -0.031  7.232
## university.degree         |     2.841 | -0.094 51.412  0.910 |  0.028 14.092
##                             cos2    Dim.3    ctr   cos2
## basic                      0.129 | -0.004  0.437  0.002 |
## high.school                0.835 | -0.027 16.255  0.147 |
## illiterate                 0.001 | -0.476  8.636  0.214 |
## professional.course        0.128 |  0.078 73.242  0.823 |
## university.degree          0.082 | -0.007  1.430  0.005 |
##
## Columns
##                             Iner*1000    Dim.1    ctr   cos2    Dim.2    ctr
## duration-[5,60]           |     1.927 |  0.227 36.100  0.942 |  0.026  1.420
## duration-(60,120]          |     1.931 |  0.124 30.423  0.792 | -0.055 18.306
## duration-(120,150]         |     0.352 |  0.069  6.154  0.879 |  0.021  1.698
## duration-(150,180]         |     0.273 | -0.002  0.008  0.001 | -0.027  2.992
## duration-(180,240]         |     0.900 | -0.046  4.575  0.256 | -0.052 18.103
## duration-(240,300]         |     1.189 |  0.010  0.182  0.008 |  0.104 55.541
## duration-(300,1.2e+03]     |     1.205 | -0.047 20.930  0.874 |  0.001  0.046
## duration-(1.2e+03,2.1e+03] |     0.145 |  0.039  1.627  0.563 |  0.024  1.893
##                              cos2    Dim.3    ctr   cos2  
## duration-[5,60]             0.012 | -0.049  8.165  0.045 |
## duration-(60,120]           0.157 |  0.029  7.706  0.042 |
## duration-(120,150]          0.080 | -0.010  0.616  0.018 |
## duration-(150,180]          0.182 | -0.046 13.112  0.506 |
## duration-(180,240]          0.333 |  0.057 33.651  0.394 |
## duration-(240,300]          0.774 |  0.055 24.568  0.218 |
## duration-(300,1.2e+03]      0.001 | -0.015 10.991  0.096 |
## duration-(1.2e+03,2.1e+03]  0.216 | -0.015  1.191  0.086 |
```

5.3 Conclusions

All in all, we can see that the findings of CA relative to duration-age and duration-education are very linked to the findings of the clustering, so we can really say with a certain confidence that the age and education of an individual is really impactful on the target variables.

6 MCA

6.1 Eigenvalues and dominant axes analysis

We consider, according to the generalized Kaiser theorem, all those dimensions such that their eigenvalue is greater than the mean. We see that the average gives us 0.125. Therefore, we will take up to dimension 15, which represents the 60% of the sample.

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

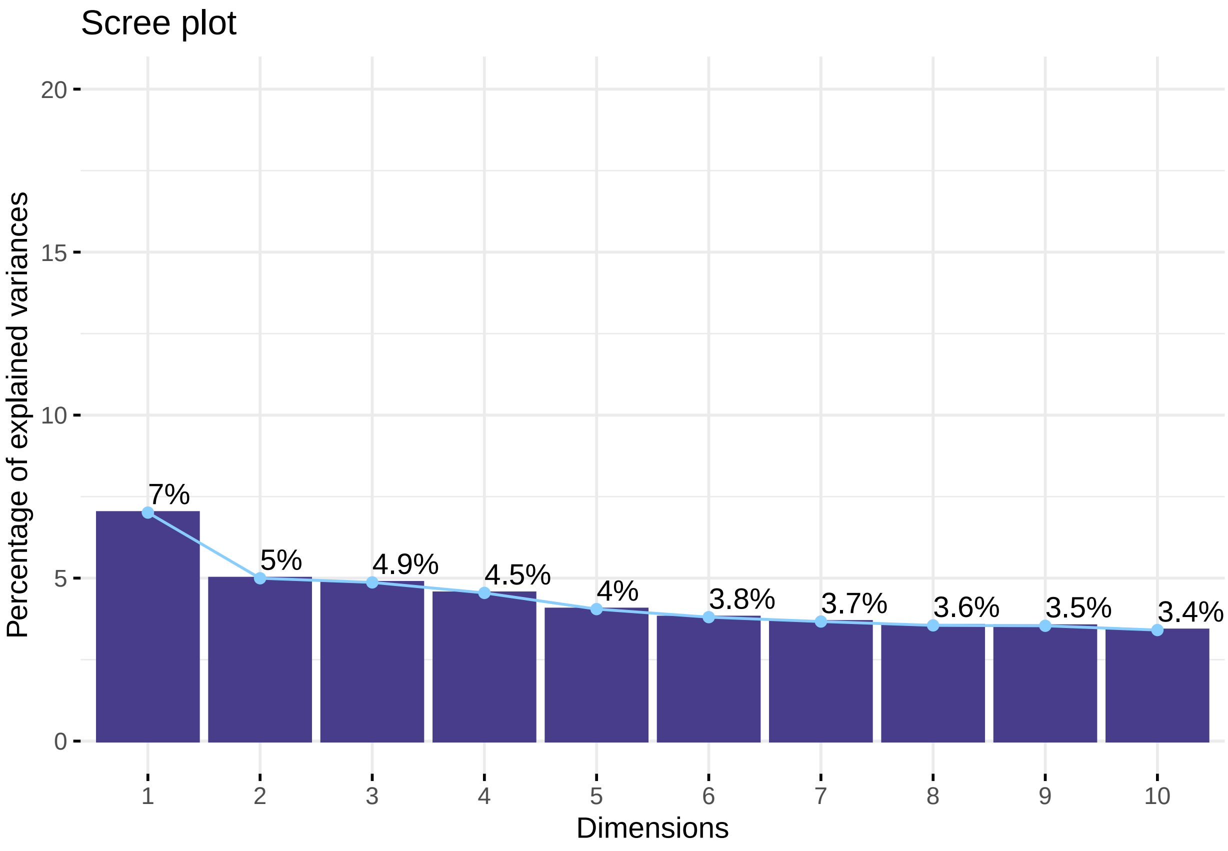
```
## [1] 0.125
```

```
res.mca$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## dim 1  0.27166749              7.0107741                          7.010774
## dim 2  0.19359439              4.9959843                         12.006758
## dim 3  0.18867610              4.8690607                         16.875819
## dim 4  0.17618335              4.5466671                         21.422486
## dim 5  0.15693106              4.0498338                         25.472320
## dim 6  0.14736533              3.8029761                         29.275296
## dim 7  0.14216283              3.6687181                         32.944014
## dim 8  0.13759710              3.5508930                         36.494907
## dim 9  0.13708679              3.5377236                         40.032631
## dim 10 0.13205455              3.4078594                         43.440490
## dim 11 0.13098810              3.3803382                         46.820828
## dim 12 0.12938254              3.3389042                         50.159733
## dim 13 0.12721618              3.2829981                         53.442731
## dim 14 0.12612378              3.2548073                         56.697538
## dim 15 0.12500640              3.2259716                         59.923509
## dim 16 0.12417242              3.2044496                         63.127959
## dim 17 0.12346130              3.1860981                         66.314057
## dim 18 0.12142900              3.1336516                        69.447709
## dim 19 0.12056091              3.1112492                        72.558958
## dim 20 0.11866219              3.0622501                        75.621208
## dim 21 0.11547654              2.9800397                        78.601248
## dim 22 0.11277008              2.9101955                        81.511443
## dim 23 0.11020115              2.8439007                        84.355344
## dim 24 0.10911690              2.8159200                        87.171264
## dim 25 0.10703430              2.7621755                        89.933439
## dim 26 0.09307584              2.4019571                        92.335397
## dim 27 0.08887914              2.2936552                        94.629052
## dim 28 0.07182627              1.8535813                        96.482633
## dim 29 0.05912249              1.5257418                        98.008375
## dim 30 0.04113547              1.0615606                        99.069935
## dim 31 0.03604000              0.9300646                       100.000000
```

We can also visualize the percentages of inertia explained by each MCA dimensions:

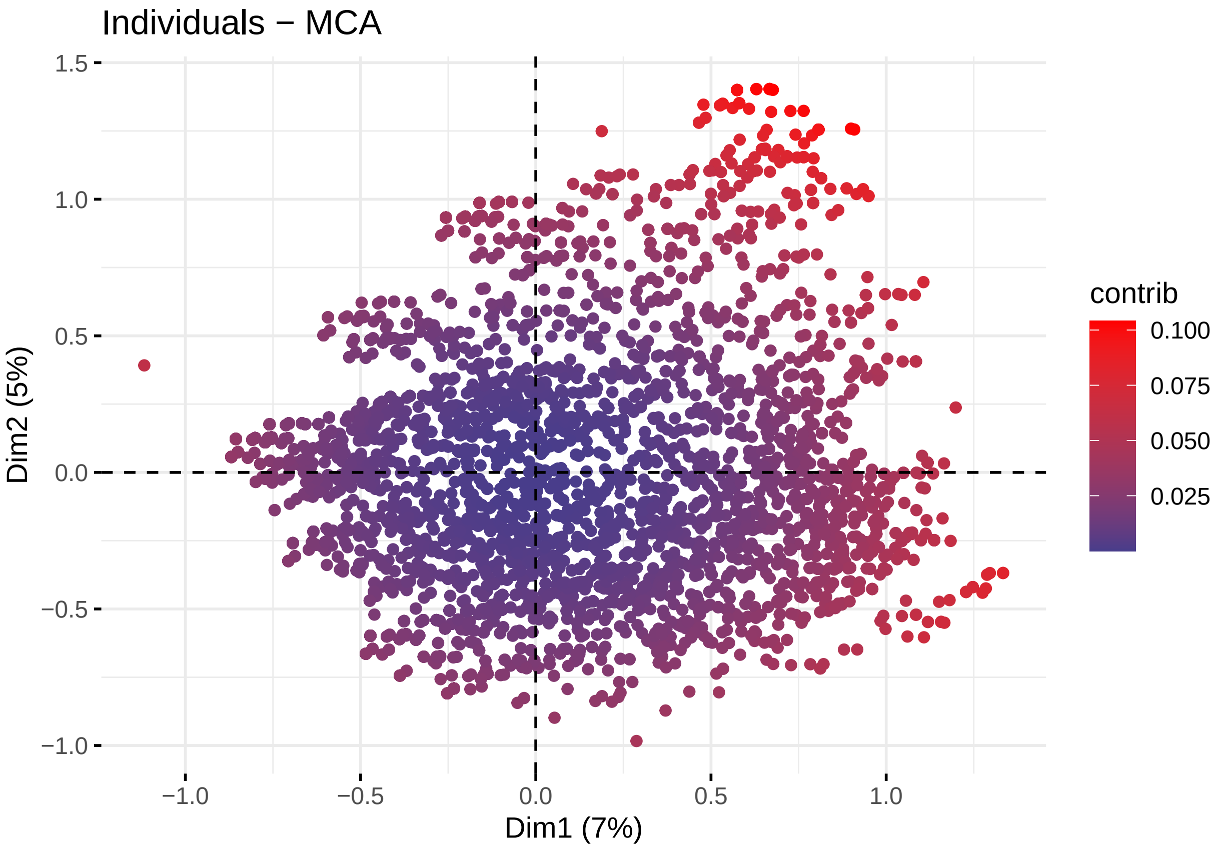
```
fviz_screeplot(
  res.mca,
  addlabels=TRUE,
  ylim=c(0,20),
  barfill="darkslateblue",
  barcolor="darkslateblue",
  linecolor="skyblue1"
)
```



6.2 Individuals point of view

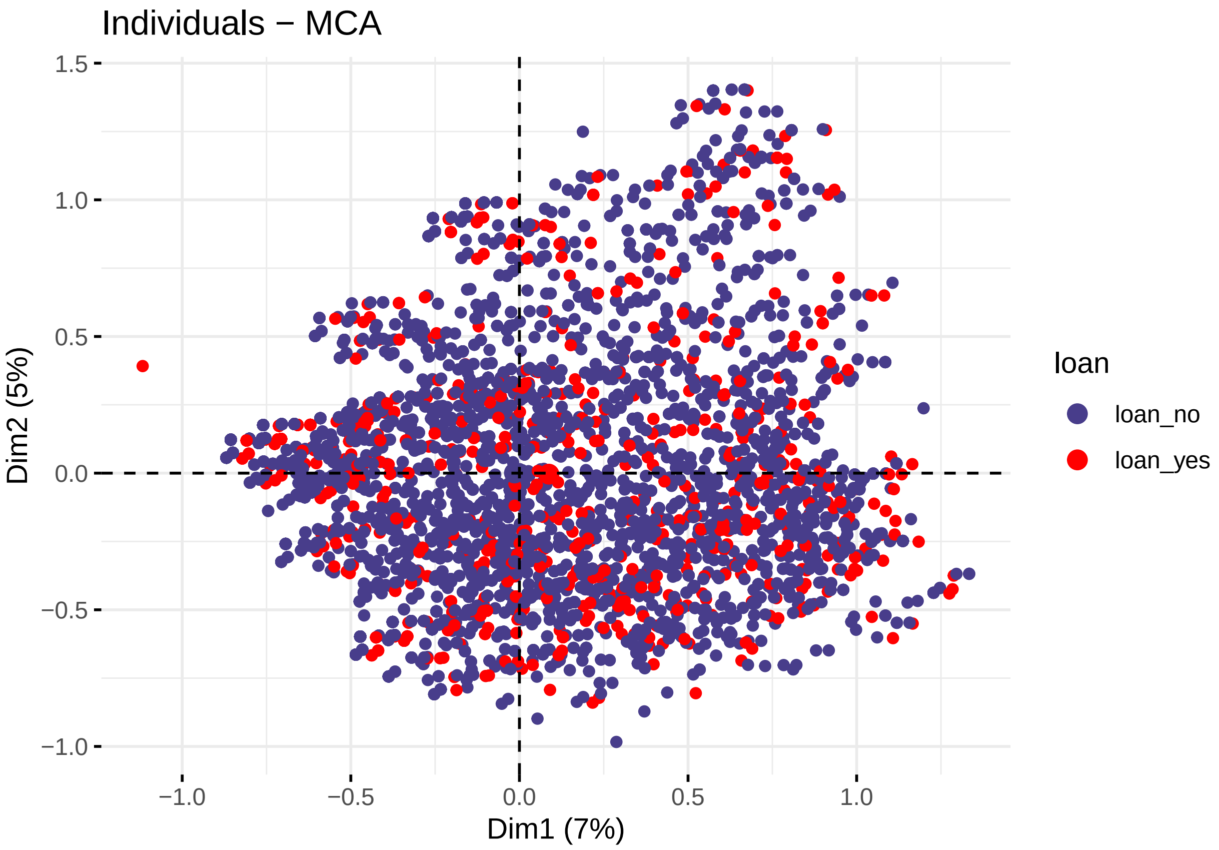
We can see in the legend that thhe contributions goes from 0.025 to 0.1 so we can't say that there are individuals who are too contributive.

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

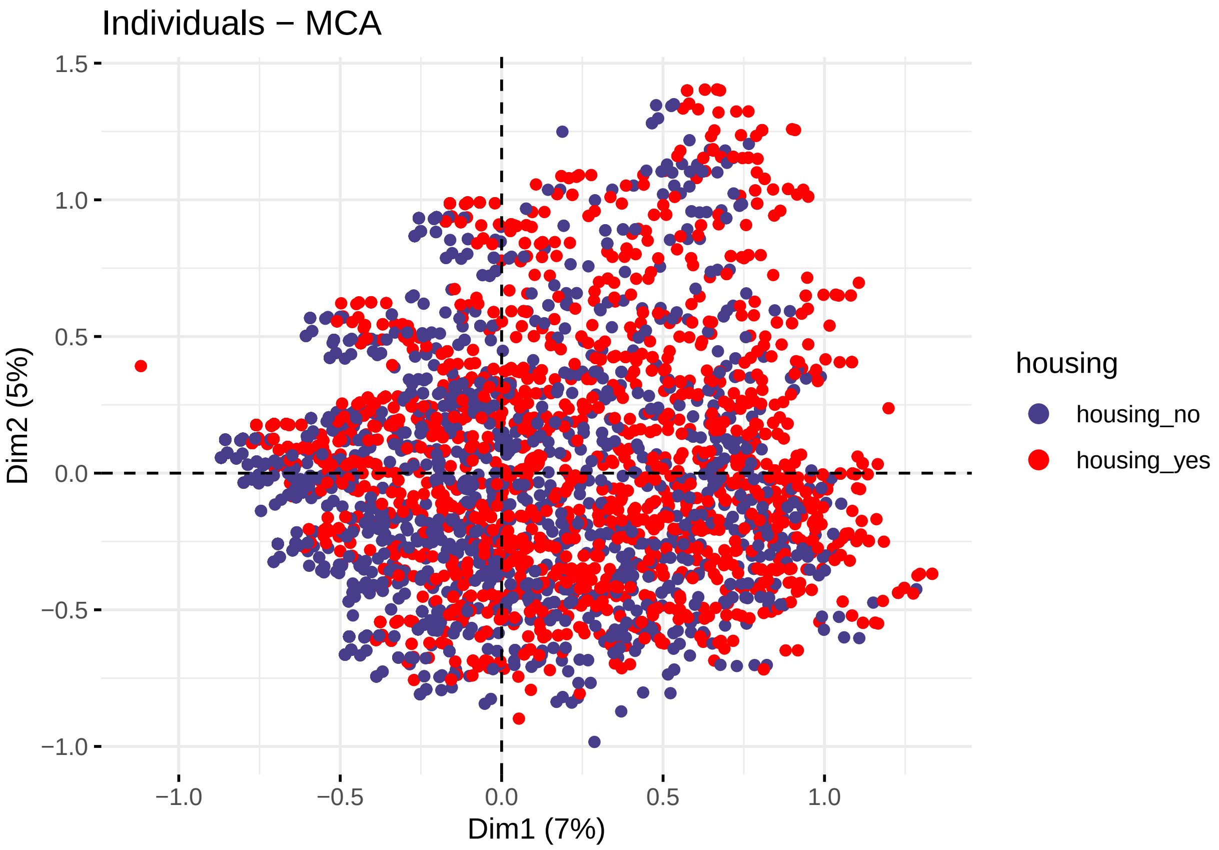


We've tried many variables but as we can see with these two they are mostly homogenous across the factorial map, that is, evenly distributed.

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



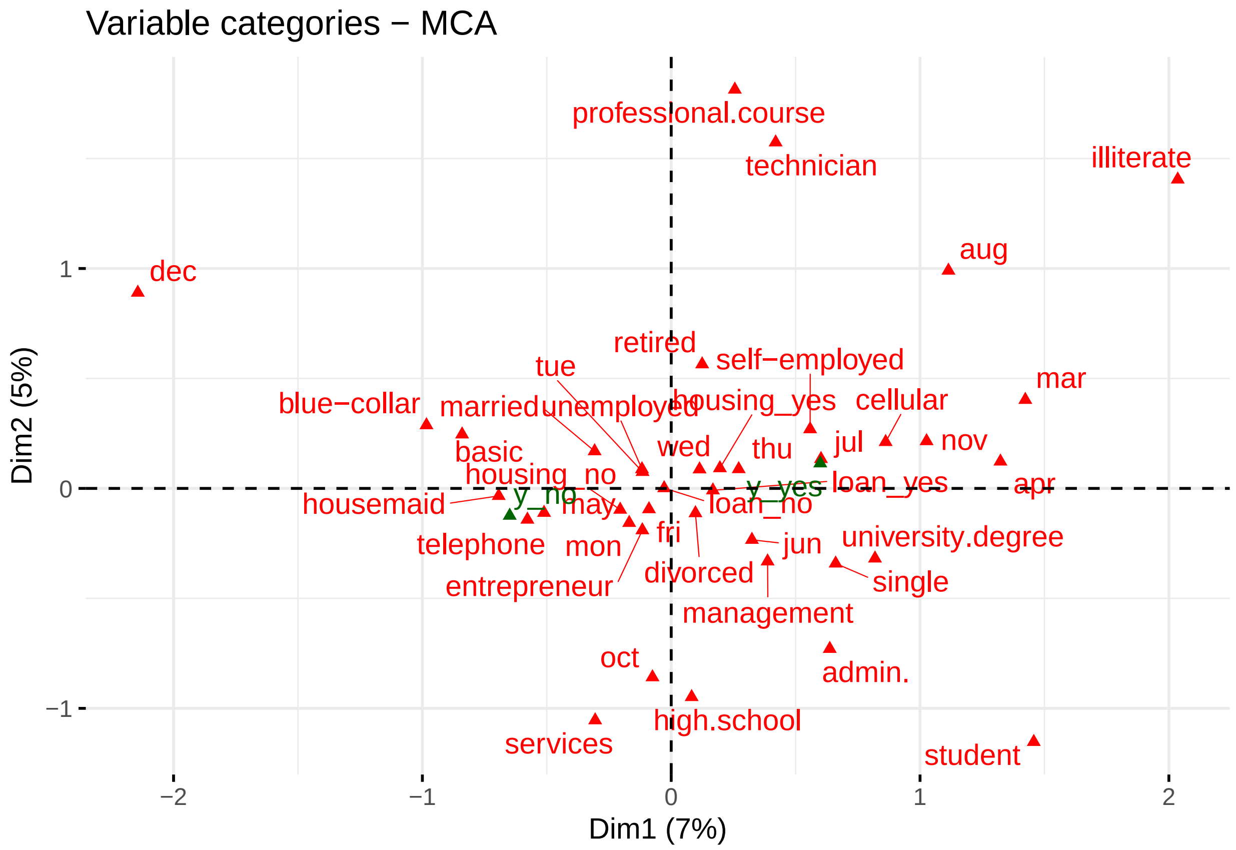
```r
fviz_mca_ind(res.mca, label="none", habillage="housing", palette=c("darkslateblue", "red"))
```



6.3 Interpreting map of categories: average profile versus extreme profiles (rare categories)

We can see that the month-december,education-illiterate are extreme profiles from the DIM1 and professional course and technician are etreme profiles from DIM2. All the remaining categories are all gravitating towards the center, we can clearly see the separation of categories respect to the variable "y", the ones near "yes" will make it more likely that the individual with those characteristic will says yes, and the same logics is applied for no.

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



6.4 Interpreting the axes association to factor map

```
res.desc <- dimdesc(res.mca, axes = c(1,2))
```

6.4.1 Description of dimension 1

The first dimension of the MCA plot is primarily driven by the contact type, education, and whether or not the client subscribed to a term deposit. Clients who were contacted via cellular communication, had a university degree, and subscribed to a term deposit are more likely to be positively associated with this dimension, while those who were contacted via telephone, had a lower level of education, and did not subscribe to a term deposit are more likely to be negatively associated with this dimension.

```
res.desc[[1]]
```

```
##
## Link between the variable and the categorical variable (1-way anova)
## =============================================
##                      R2       p.value
## y           0.388798633  0.000000e+00
## job         0.454019225  0.000000e+00
## education   0.452808059  0.000000e+00
## contact     0.498061874  0.000000e+00
## month       0.516277391  0.000000e+00
## marital     0.179201680 5.245561e-215
## housing     0.040126746  2.001070e-46
## day_of_week 0.028106884  8.505113e-30
## loan        0.004738099  1.105581e-06
```

```
##
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                    Estimate       p.value
## contact=cellular                0.375209177  0.000000e+00
## education=university.degree     0.181682749  0.000000e+00
## y=y_yes                         0.325258697  0.000000e+00
## month=apr                       0.511046411 3.415216e-204
## marital=single                  0.266172640 4.043475e-197
## job=admin.                      0.267302547 1.238292e-153
## month=aug                       0.402184394  3.366440e-82
## month=mar                       0.563181378  2.130686e-60
## job=student                     0.694645091  4.436658e-51
## month=nov                       0.356349157  4.177682e-48
## housing=housing_yes             0.104436599  2.001070e-46
## month=jul                       0.135213586  2.341728e-37
## job=technician                  0.153932743  1.385318e-36
## day_of_week=thu                 0.140160364  1.035432e-22
## job=management                  0.137101268  3.289527e-15
## job=self-employed               0.226136196  3.172369e-13
## loan=loan_yes                   0.051035445  1.105581e-06
## day_of_week=wed                 0.058158164  1.030385e-04
## education=illiterate            0.815840512  3.977185e-03
## month=dec                      -1.295868353  3.204833e-02
## marital=divorced               -0.027470225  1.761693e-02
## day_of_week=fri                -0.047788642  4.270685e-03
## education=high.school          -0.202385914  1.329667e-03
## day_of_week=tue                -0.061292801  2.020357e-05
## loan=loan_no                   -0.051035445  1.105581e-06
## month=jun                      -0.009469419  1.858941e-10
## day_of_week=mon                -0.089237085  1.787286e-10
## education=professional.course -0.111933787  5.757979e-12
## job=services                   -0.224117923  1.891021e-13
## job=housemaid                  -0.426172378  2.782137e-14
## housing=housing_no             -0.104436599  2.001070e-46
## marital=married                -0.238702415 5.327177e-185
## month=may                      -0.444877259  0.000000e+00
## contact=telephone              -0.375209177  0.000000e+00
## education=basic                -0.683203561  0.000000e+00
## job=blue-collar                -0.577497794  0.000000e+00
## y=y_no                         -0.325258697  0.000000e+00
```

6.4.2 Description of dimension 2

The dimension 2 appears to be strongly influenced by the type of job and level of education of the respondents, with some additional contribution from the month of last contact and marital status variables.

```
res.desc[[2]]
```

```
##
## Link between the variable and the categorical variable (1-way anova)
## =============================================
##                      R2      p.value
## job         0.691224421 0.000000e+00
## education   0.676742542 0.000000e+00
## month       0.079818846 9.388966e-85
## marital     0.050964218 1.739981e-57
## contact     0.029940676 6.615411e-35
## y           0.014137106 3.303438e-17
## day_of_week 0.011153159 1.972311e-11
## housing     0.008898955 2.330139e-11
##
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                  Estimate       p.value
## education=professional.course  0.604588052  0.000000e+00
## job=technician                 0.720719739  0.000000e+00
## month=aug                      0.359914771  7.414765e-65
## marital=married                0.116201309  3.211613e-54
## contact=cellular               0.077658706  6.615411e-35
## job=blue-collar                0.154955587  3.277243e-32
## y=y_yes                        0.052356952  3.303438e-17
## job=retired                    0.276720638  1.240984e-15
## housing=housing_yes            0.041517652  2.330139e-11
## month=mar                      0.101335181  4.567219e-06
## job=self-employed              0.146859347  4.499815e-04
## day_of_week=thu                0.038775635  1.526750e-03
## month=nov                      0.018736054  2.495384e-03
## day_of_week=wed                0.038303385  3.045035e-03
## day_of_week=tue                0.033092483  5.489247e-03
## education=illiterate           0.424733650  4.691648e-02
## job=entrepreneur              -0.055252227  7.917238e-03
## month=apr                     -0.022002410  6.965792e-03
## marital=divorced              -0.007768646  6.241464e-03
## month=jul                     -0.016880733  4.732601e-03
## day_of_week=fri               -0.041388290  2.597441e-03
## month=jun                     -0.178674277  4.587300e-06
## month=oct                     -0.453580636  2.224987e-08
## day_of_week=mon               -0.068783213  3.484534e-09
## housing=housing_no            -0.041517652  2.330139e-11
## job=management                -0.117563670  1.602679e-11
## y=y_no                        -0.052356952  3.303438e-17
## month=may                     -0.124631199  5.180098e-25
## job=student                   -0.478780854  2.410154e-32
## contact=telephone             -0.077658706  6.615411e-35
## education=basic               -0.085245402  2.245510e-37
## education=university.degree   -0.333427014  2.157554e-48
## marital=single                -0.108432663  1.404283e-50
## job=services                  -0.435419880 5.021304e-151
## job=admin.                    -0.292540341 3.198180e-205
## education=high.school         -0.610649285  0.000000e+00
```

6.5 Perform a MCA taking into account also supplementary variables

6.5.1 Description of dimensions

```
res.desc <- dimdesc(res.mca_sup, axes = c(1,2))
```

6.5.1.1 Description of dimension 1

The first dimension is positively correlated with the duration of the last contact, which means that clients who had longer contacts are more likely to be positioned towards the positive end of the first dimension. The first dimension is negatively correlated with the age and the economic indicators, such as the number of employees, employment variation rate, consumer price index, consumer confidence index, and the euribor 3 month rate. This means that older clients and clients with higher economic indicators are more likely to be positioned towards the negative end of the first dimension. The first dimension is negatively correlated with the binary variable that indicates whether the client subscribed to a term deposit or not. This means that clients who did not subscribe to a term deposit are more likely to be positioned towards the negative end of the first dimension.

```
res.desc[[1]]
```

```
##
## Link between the variable and the continuous variables (R-square)
## =================================================================================
##                 correlation       p.value
## duration         0.2326336  1.997449e-62
## age             -0.1860164  3.645494e-40
## nr.employed     -0.3244683 6.414285e-123
## emp.var.rate    -0.4649930 9.507227e-267
## euribor3m       -0.4682708 5.380749e-271
## cons.conf.idx   -0.5659700  0.000000e+00
## cons.price.idx  -0.5734473  0.000000e+00
##
## Link between the variable and the categorical variable (1-way anova)
## =============================================
##                      R2       p.value
## y           0.388798633  0.000000e+00
## job         0.454019225  0.000000e+00
## education   0.452808059  0.000000e+00
## contact     0.498061874  0.000000e+00
## month       0.516277391  0.000000e+00
## marital     0.179201680 5.245561e-215
## housing     0.040126746  2.001070e-46
## day_of_week 0.028106884  8.505113e-30
## loan        0.004738099  1.105581e-06
##
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                   Estimate       p.value
## contact=cellular               0.375209177  0.000000e+00
## education=university.degree    0.181682749  0.000000e+00
## y=y_yes                        0.325258697  0.000000e+00
## month=apr                      0.511046411 3.415216e-204
## marital=single                 0.266172640 4.043475e-197
## job=admin.                     0.267302547 1.238292e-153
## month=aug                      0.402184394  3.366440e-82
## month=mar                      0.563181378  2.130686e-60
## job=student                    0.694645091  4.436658e-51
## month=nov                      0.356349157  4.177682e-48
## housing=housing_yes            0.104436599  2.001070e-46
## month=jul                      0.135213586  2.341728e-37
## job=technician                 0.153932743  1.385318e-36
## day_of_week=thu                0.140160364  1.035432e-22
## job=management                 0.137101268  3.289527e-15
## job=self-employed              0.226136196  3.172369e-13
## loan=loan_yes                  0.051035445  1.105581e-06
## day_of_week=wed                0.058158164  1.030385e-04
## education=illiterate           0.815840512  3.977185e-03
## month=dec                     -1.295868353  3.204833e-02
## marital=divorced              -0.027470225  1.761693e-02
## day_of_week=fri               -0.047788642  4.270685e-03
## education=high.school         -0.202385914  1.329667e-03
## day_of_week=tue               -0.061292801  2.020357e-05
## loan=loan_no                  -0.051035445  1.105581e-06
## month=jun                     -0.009469419  1.858941e-10
## day_of_week=mon               -0.089237085  1.787286e-10
## education=professional.course -0.111933787  5.757979e-12
## job=services                  -0.224117923  1.891021e-13
## job=housemaid                 -0.426172378  2.782137e-14
## housing=housing_no            -0.104436599  2.001070e-46
## marital=married               -0.238702415 5.327177e-185
## month=may                     -0.444877259  0.000000e+00
## contact=telephone             -0.375209177  0.000000e+00
## education=basic               -0.683203561  0.000000e+00
## job=blue-collar               -0.577497794  0.000000e+00
## y=y_no                        -0.325258697  0.000000e+00
```

6.5.1.2 Description of dimension 2 Age is weakly positively correlated with the second dimension of the MCA, meaning that it has some association with the categorical variables being analyzed. Duration has a weak positive correlation with the second dimension of the MCA, indicating that it also has some relationship with the categorical variables being analyzed. The number of employees and consumer confidence index have a weak positive and negative correlation, respectively, with the second dimension of the MCA, suggesting that they have some association with the categorical variables being analyzed. Education and job have the strongest association, with an R-squared value of around 0.67-0.69, followed by month, marital, contact, and housing. The variable "y" (indicating whether or not the client subscribed to a term deposit) has a relatively weak association with the categorical variables, with an R-squared value of 0.014. Among the categories of the categorical variables, several have a relatively strong association with the dimension, either positively or negatively. For example, professional course education, technician job, and August month are positively associated with the dimension, while illiterate education, entrepreneur job, and October month are negatively associated with the dimension.

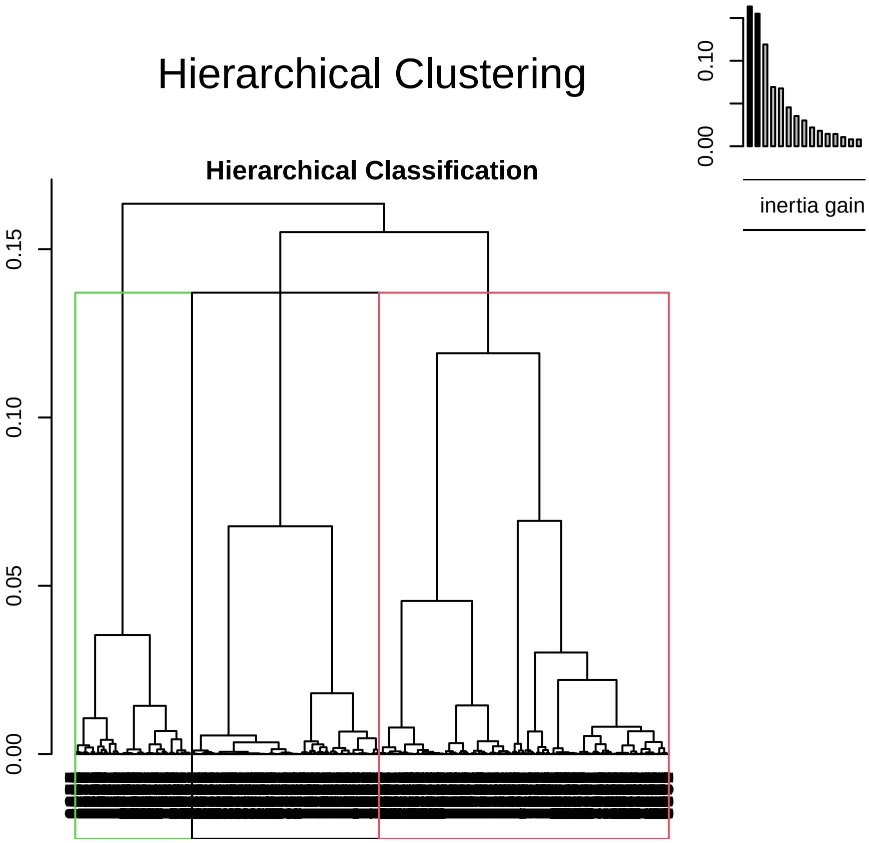
```
res.desc[[2]]
```

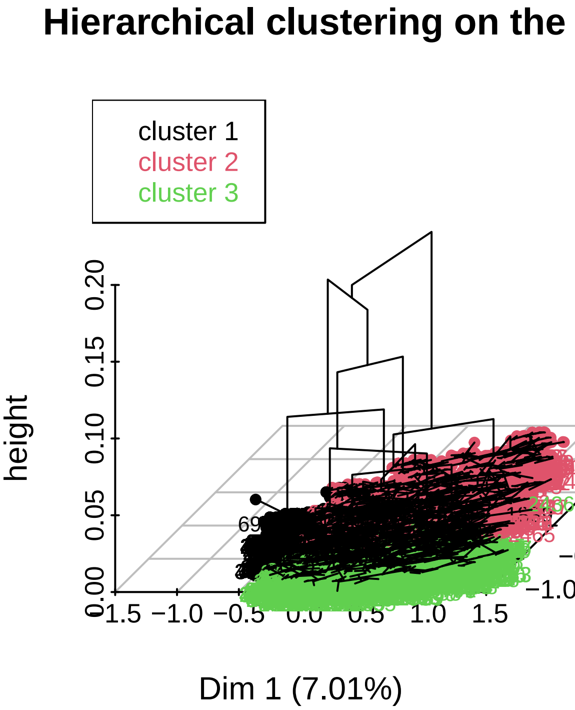
```
##
## Link between the variable and the continuous variables (R-square)
## =================================================================================
##                 correlation      p.value
## age              0.16111055 1.992409e-30
## duration         0.08057452 1.161164e-08
## nr.employed      0.02812660 4.672916e-02
## cons.conf.idx  -0.05189257 2.416841e-04
## cons.price.idx -0.10615595 5.241665e-14
##
## Link between the variable and the categorical variable (1-way anova)
## =============================================
##                       R2      p.value
## job          0.691224421 0.000000e+00
## education    0.676742542 0.000000e+00
## month        0.079818846 9.388966e-85
## marital     0.050964218 1.739981e-57
## contact     0.029940676 6.615411e-35
## y           0.014137106 3.303438e-17
## day_of_week 0.011153159 1.972311e-11
## housing     0.008898955 2.330139e-11
## 
## Link between variable abd the categories of the categorical variables
## ================================================================
##                                   Estimate       p.value
## education=professional.course  0.604588052  0.000000e+00
## job=technician                 0.720719739  0.000000e+00
## month=aug                      0.359914771  7.414765e-65
## marital=married                0.116201309  3.211613e-54
## contact=cellular               0.077658706  6.615411e-35
## job=blue-collar                0.154955587  3.277243e-32
## y=y_yes                        0.052356952  3.303438e-17
## job=retired                    0.276720638  1.240984e-15
## housing=housing_yes            0.041517652  2.330139e-11
## month=mar                      0.101335181  4.567219e-06
## job=self-employed              0.146859347  4.499815e-04
## day_of_week=thu                0.038775635  1.526750e-03
## month=nov                      0.018736054  2.495384e-03
## day_of_week=wed                0.038303385  3.045035e-03
## day_of_week=tue                0.033092483  5.489247e-03
## education=illiterate           0.424733650  4.691648e-02
## job=entrepreneur              -0.055252227  7.917238e-03
## month=apr                     -0.022002410  6.965792e-03
## marital=divorced              -0.007768646  6.241464e-03
## month=jul                     -0.016880733  4.732601e-03
## day_of_week=fri               -0.041388290  2.597441e-03
## month=jun                     -0.178674277  4.587300e-06
## month=oct                     -0.453580636  2.224987e-08
## day_of_week=mon               -0.068783213  3.484534e-09
## housing=housing_no            -0.041517652  2.330139e-11
## job=management                -0.117563670  1.602679e-11
## y=y_no                        -0.052356952  3.303438e-17
## month=may                     -0.124631199  5.180098e-25
## job=student                   -0.478780854  2.410154e-32
## contact=telephone             -0.077658706  6.615411e-35
## education=basic               -0.085245402  2.245510e-37
## education=university.degree   -0.333427014  2.157554e-48
## marital=single                -0.108432663  1.404283e-50
## job=services                  -0.435419880 5.021304e-151
## job=admin.                    -0.292540341 3.198180e-205
## education=high.school         -0.610649285  0.000000e+00
```

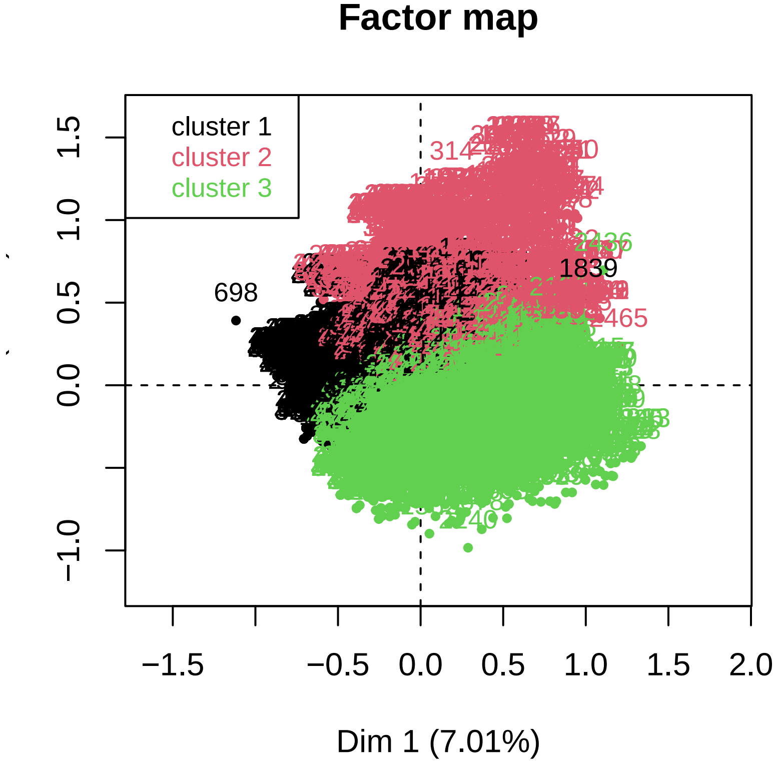
7 Hierarchical Clustering (from MCA)

We've decided that numbers of cluster is the one that the algorithm gives us, with nb.clust=-1.

```
res.hcpcMCA <- HCPC(res.mca,nb.clust = -1, order = TRUE)
```







7.1 Description of clusters

- Cluster 1:

    - The first cluster are people who are more likely to say no contacted via telephone and have a basic type of education and have a blue-collar kind of job and are married.
- Cluster 2:
    - The second cluster are people who are more likely to say yes being contacted by cellular and are educated from a professional course and are technicians.They are also married and young.
- Cluster 3:
    - The first cluster are people who are almost guaranteed to say say yes, they are university educated and are working on more technical jobs such as managment and adminsitration, they are young and most likely single as well.

From this clustering analysis, we can see that the clusters aren't very different than the previous ones, young university graduates are still the people who are more likely to say yes.

```
res.hcpcMCA$desc.var$category    # description of each cluster by the categories
```

```
## $`1`
##                                 Cla/Mod    Mod/Cla Global       p.value
## education=basic               88.536866 93.4346505  34.72  0.000000e+00
## job=blue-collar               90.544872 68.6930091  24.96  0.000000e+00
## marital=married               43.022881 81.1550152  62.06  8.588297e-90
## month=may                     39.475348 75.9270517  63.28  6.253971e-40
## contact=telephone             39.859673 72.5227964  59.86  1.923064e-38
## y=y_no                        40.875000 59.6352584  48.00  7.647273e-31
## job=retired                   64.062500  7.4772036   3.84  2.478293e-19
## job=housemaid                 65.811966  4.6808511   2.34  2.179802e-13
## housing=housing_no            36.650287 54.4072948  48.84  3.509902e-08
## loan=loan_no                  33.543712 87.2340426  85.56  1.763566e-02
## loan=loan_yes                 29.085873 12.7659574  14.44  1.763566e-02
## month=jun                     26.330532  5.7142857   7.14  5.442846e-03
## month=oct                     11.904762  0.3039514   0.84  2.044557e-03
## job=self-employed             21.951220  2.1884498   3.28  1.829562e-03
## marital=divorced              24.339623  7.8419453  10.60  5.904849e-06
## month=mar                     12.698413  0.9726444   2.52  1.551889e-07
## housing=housing_yes           29.319781 45.5927052  51.16  3.509902e-08
## month=aug                     18.081181  2.9787234   5.42  2.320910e-08
## month=nov                     15.263158  1.7629179   3.80  2.243860e-08
## month=apr                     16.742081  4.4984802   8.84  1.588880e-15
## job=student                    0.000000  0.0000000   2.04  1.266359e-18
## job=services                  14.313346  4.4984802  10.34  4.919823e-24
## job=management                 9.210526  2.1276596   7.60  2.317436e-29
## y=y_yes                       25.538462 40.3647416  52.00  7.647273e-31
## contact=cellular              22.521176 27.4772036  40.14  1.923064e-38
## marital=single                13.240673 11.0030395  27.34  1.071892e-81
## education=professional.course  1.743265  0.6686930  12.62  8.741972e-99
## education=high.school          7.271172  5.1671733  23.38 9.106072e-121
## job=technician                 1.453104  0.6686930  15.14 5.026176e-124
## job=admin.                     4.036244  2.9787234  24.28 6.683809e-169
## education=university.degree    0.752394  0.6686930  29.24 1.822607e-288
##                                  v.test
## education=basic                     Inf
## job=blue-collar                     Inf
## marital=married               20.092469
## month=may                     13.225476
## contact=telephone             12.965368
## y=y_no                        11.546966
## job=retired                     8.989734
## job=housemaid                   7.337279
## housing=housing_no              5.513889
## loan=loan_no                    2.373180
## loan=loan_yes                  -2.373180
## month=jun                      -2.779585
## month=oct                      -3.083682
## job=self-employed              -3.116589
## marital=divorced               -4.529768
## month=mar                      -5.246294
## housing=housing_yes            -5.513889
## month=aug                      -5.586201
## month=nov                      -5.592064
## month=apr                      -7.969834
## job=student                    -8.808673
## job=services                  -10.111357
## job=management                -11.249942
## y=y_yes                       -11.546966
## contact=cellular              -12.965368
## marital=single                -19.144683
## education=professional.course -21.095530
## education=high.school         -23.367708
## job=technician                -23.686006
## job=admin.                    -27.701574
## education=university.degree   -36.296721
## 
## $`2`
##                                 Cla/Mod    Mod/Cla Global       p.value
## education=professional.course 98.256735 64.3153527  12.62  0.000000e+00
## job=technician                98.546896 77.3858921  15.14  0.000000e+00
## month=aug                     33.948339  9.5435685   5.42  4.370508e-09
## contact=cellular              21.674141 45.1244813  40.14  4.695792e-04
## housing=housing_yes           20.914777 55.4979253  51.16  2.701465e-03
## y=y_yes                       20.769231 56.0165975  52.00  5.442728e-03
## marital=single                21.433797 30.3941909  27.34  1.876250e-02
## day_of_week=tue               21.435228 23.8589212  21.46  4.520494e-02
## job=self-employed             13.414634  2.2821577   3.28  4.683164e-02
## marital=married               18.175959 58.5062241  62.06  1.172465e-02
## y=y_no                        17.666667 43.9834025  48.00  5.442728e-03
## housing=housing_no            17.567568 44.5020747  48.84  2.701465e-03
## month=may                     17.951960 58.9211618  63.28  1.890152e-03
## job=retired                   10.416667  2.0746888   3.84  7.639167e-04
## contact=telephone             17.674574 54.8755187  59.86  4.695792e-04
## job=housemaid                  5.982906  0.7261411   2.34  4.085151e-05
## job=student                    2.941176  0.3112033   2.04  8.064158e-07
## job=entrepreneur               5.319149  1.0373444   3.76  2.337388e-08
## education=university.degree   12.722298 19.2946058  29.24  6.788955e-15
## job=management                 4.473684  1.7634855   7.60  2.825998e-18
## job=services                   5.996132  3.2157676  10.34  1.657293e-19
## education=high.school          7.784431  9.4398340  23.38  1.357278e-34
## job=blue-collar                4.086538  5.2904564  24.96  1.033624e-69
## job=admin.                     3.377265  4.2531120  24.28  2.510399e-75
## education=basic                3.859447  6.9502075  34.72 9.461812e-110
##                                    v.test
## education=professional.course        Inf
## job=technician                       Inf
## month=aug                       5.869523
## contact=cellular                3.497535
## housing=housing_yes             2.999812
## y=y_yes                         2.779592
## marital=single                  2.350216
## day_of_week=tue                 2.002742
## job=self-employed              -1.987820
## marital=married                -2.520325
## y=y_no                         -2.779592
## housing=housing_no             -2.999812
## month=may                      -3.106971
## job=retired                    -3.365548
## contact=telephone              -3.497535
## job=housemaid                  -4.102610
## job=student                    -4.933808
## job=entrepreneur               -5.584971
## education=university.degree    -7.788351
## job=management                 -8.718224
## job=services                   -9.033854
## education=high.school         -12.267285
## job=blue-collar               -17.649115
## job=admin.                    -18.364872
## education=basic               -22.257880
##
## $`3`
##                                 Cla/Mod   Mod/Cla Global       p.value
## job=admin.                    92.586491 47.009619  24.28 1.130221e-317
## education=university.degree   86.525308 52.906734  29.24 5.568529e-293
## education=high.school         84.944397 41.530740  23.38 2.823746e-198
## job=management                86.315789 13.718110   7.60  9.214754e-60
## job=services                  79.690522 17.231284  10.34  1.601752e-55
## marital=single                65.325530 37.348390  27.34  1.298027e-52
## job=student                   97.058824  4.140527   2.04  1.769413e-28
## contact=cellular              55.804684 46.842325  40.14  2.043330e-20
## y=y_yes                       53.692308 58.385613  52.00  4.690501e-18
## month=apr                     64.027149 11.836052   8.84  7.929581e-13
## job=self-employed             64.634146  4.433292   3.28  1.144098e-05
## month=nov                     62.631579  4.976997   3.80  3.082353e-05
## job=entrepreneur              62.234043  4.893350   3.76  5.529699e-05
## marital=divorced              55.471698 12.296110  10.60  1.953103e-04
## month=mar                     63.492063  3.345880   2.52  3.648824e-04
## month=jun                     55.742297  8.322877   7.14  1.906125e-03
## housing=housing_yes           49.765442 53.241322  51.16  4.835333e-03
## month=oct                     66.666667  1.171058   0.84  1.472746e-02
## day_of_week=tue               44.920783 20.158929  21.46  3.192193e-02
## housing=housing_no            45.782146 46.758678  48.84  4.835333e-03
## job=housemaid                 28.205128  1.380176   2.34  1.320427e-05
## job=retired                   25.520833  2.049352   3.84  1.204909e-10
## y=y_no                        41.458333 41.614387  48.00  4.690501e-18
## contact=telephone             42.465753 53.157675  59.86  2.043330e-20
## month=may                     42.572693 56.336261  63.28  1.736625e-22
## marital=married               38.801160 50.355500  62.06  2.422605e-60
## education=professional.course  0.000000  0.000000  12.62 4.520998e-197
## job=technician                 0.000000  0.000000  15.14 2.066279e-241
## job=blue-collar                5.368590  2.802175  24.96 4.677485e-309
## education=basic                7.603687  5.520703  34.72  0.000000e+00
##                                   v.test
## job=admin.                     38.103288
## education=university.degree    36.581805
## education=high.school          30.041448
## job=management                 16.304193
## job=services                   15.696359
## marital=single                 15.265513
## job=student                    11.069216
## contact=cellular                9.260052
## y=y_yes                         8.660658
## month=apr                       7.162367
## job=self-employed               4.387977
## month=nov                       4.167295
## job=entrepreneur                4.032031
## marital=divorced                3.725006
## month=mar                       3.564261
## month=jun                       3.104482
## housing=housing_yes             2.817804
## month=oct                       2.439012
## day_of_week=tue                -2.145387
## housing=housing_no             -2.817804
## job=housemaid                  -4.356692
## job=retired                    -6.438713
## y=y_no                         -8.660658
## contact=telephone              -9.260052
## month=may                      -9.756126
## marital=married               -16.385626
## education=professional.course -29.949094
## job=technician                -33.180374
## job=blue-collar               -37.579332
## education=basic                     -Inf
```

```
res.hcpcMCA$desc.var$test.chi2   # categorical variables which characterizes the clusters
```

```
##                 p.value df
## job       0.000000e+00 20
## education 0.000000e+00  8
## marital   6.642756e-90  4
## month     1.440972e-42 16
## contact   1.954621e-36  2
## y         6.175640e-30  2
## housing   1.237023e-07  2
```

7.2 Parangons and class-specific individuals.

```
res.hcpcMCA$desc.ind$para  # representative individuals of each cluster
```

```
## Cluster: 1
##        234       2836       4098       1223       1121
## 0.09147034 0.09147034 0.09147034 0.14811486 0.16283767
## ------------------------------------------------------------
## Cluster: 2
##      2942       478      2056      1367       660
## 0.2780758 0.3440311 0.3548909 0.3577262 0.4071450
## ------------------------------------------------------------
## Cluster: 3
##      1938      1515      1389        31       566
## 0.1828545 0.2159257 0.2336711 0.2483546 0.2483755
```

What we obtain are the more representative individuals, paragons, for each cluster. We get the rownames of each paragon in every single cluster.

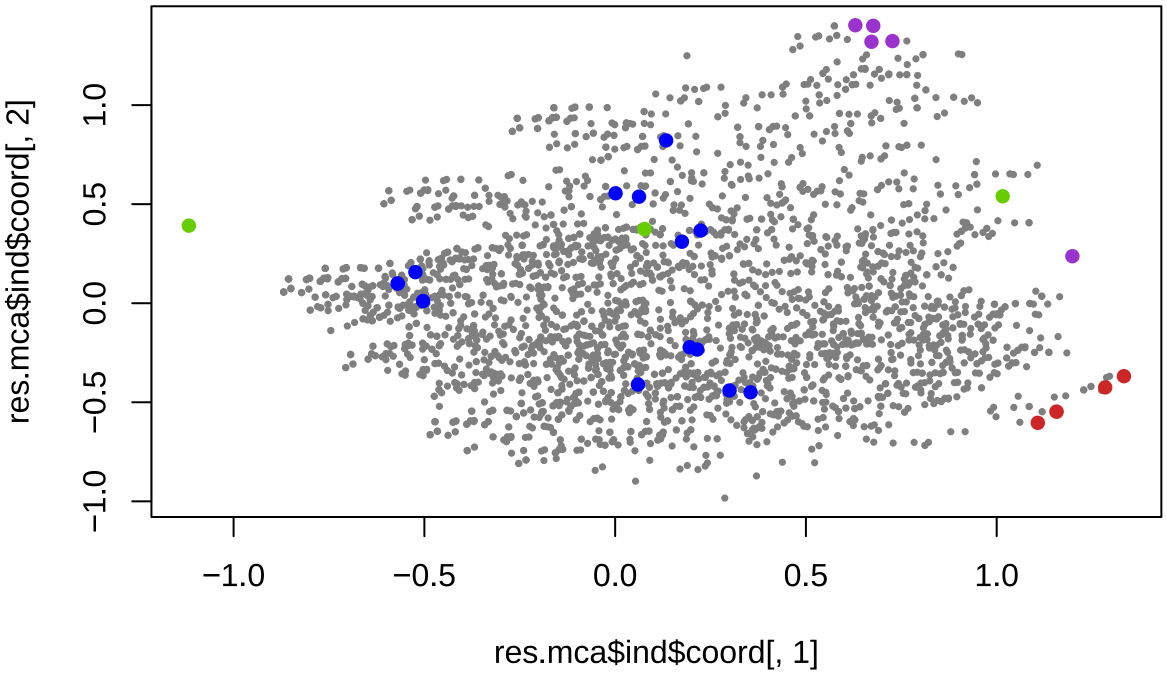
```
res.hcpcMCA$desc.ind$dist  # individuals distant from each cluster
```

```
## Cluster: 1
##      698     1839      269     1541     2063
## 2.251399 2.244362 1.675209 1.675209 1.675209
## ------------------------------------------------------------
## Cluster: 2
##     2465     1602     2162       23       94
## 2.296221 1.918623 1.891750 1.867893 1.866062
## ------------------------------------------------------------
## Cluster: 3
##       76      144      586     1978     1813
## 2.574272 2.527872 2.524225 2.524225 2.492180
```

We get the grpahical representation for the individuals that characterize classes (para and dist).

```
# characteristic individuals
para1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[1]]))
dist1<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[1]]))
para2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[2]]))
dist2<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[2]]))
para3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$para[[3]]))
dist3<-which(rownames(res.mca$ind$coord)%in%names(res.hcpcMCA$desc.ind$dist[[3]]))

plot(res.mca$ind$coord[,1],res.mca$ind$coord[,2],col="grey50",cex=0.5,pch=16)
points(res.mca$ind$coord[para1,1],res.mca$ind$coord[para1,2],col="blue",cex=1,pch=16)
points(res.mca$ind$coord[dist1,1],res.mca$ind$coord[dist1,2],col="chartreuse3",cex=1,pch=16)
points(res.mca$ind$coord[para2,1],res.mca$ind$coord[para2,2],col="blue",cex=1,pch=16)
points(res.mca$ind$coord[dist2,1],res.mca$ind$coord[dist2,2],col="darkorchid3",cex=1,pch=16)
points(res.mca$ind$coord[para3,1],res.mca$ind$coord[para3,2],col="blue",cex=1,pch=16)
points(res.mca$ind$coord[dist3,1],res.mca$ind$coord[dist3,2],col="firebrick3",cex=1,pch=16)
```



7.3 Comparison of clusters obtained after ihierachical clustering (based on PCA) on target duration and binary target.

Given the following description from clusters in MCA:

- Cluster 1:
  - The first cluster are people who are more likely to say no contacted via telephone and have a basic type of education and have a blue-collar kind of job and are married.
- Cluster 2:
  - The second cluster are people who are more likely to say yes being contacted by cellular and are educated from a professional course and are technicians.They are also married and young.
- Cluster 3:
  - The first cluster are people who are almost guaranteed to say say yes, they are university educated and are working on more technical jobs such as managment and adminsitration, they are young and most likely single as well.

and then PCA:

- Cluster 1:
  - These are the people who will say yes to the campaign and being contacted by cellulars, and mostly single university graduates. Also, they are being called during the months of april and may, which are nearing summer seasons and these kind of people tend to have money saved.
- Cluster 2:
  - The are people who are more likely to say no when they are being contacted on november by cellular, cluster similar to the one in KMEANS. These people are divorced and don't have any housing loan.
- Cluster 3:
  - These are people who are married and retired which will most likely say no, and are most usually contacted by telephone. We see that these are people who have their life together already and aren't interested in these kind of campaigns anymore.

We can comparethe clusters,but we can't say anything about the duration but we can clearly see some trends on the binary target: * In both methods we can see that the people who will say yes are young people, who are highly educated with most of them having university degrees and having good jobs and are contacted by cellular, a clear indication they are young. And the people who say no are tending towards older people who are married and have their life together already, the majority of them being retired and are contacted with a telephone.