Chuvilina Anna, MBD171

**1 Assignment on Principal Component Analysis**

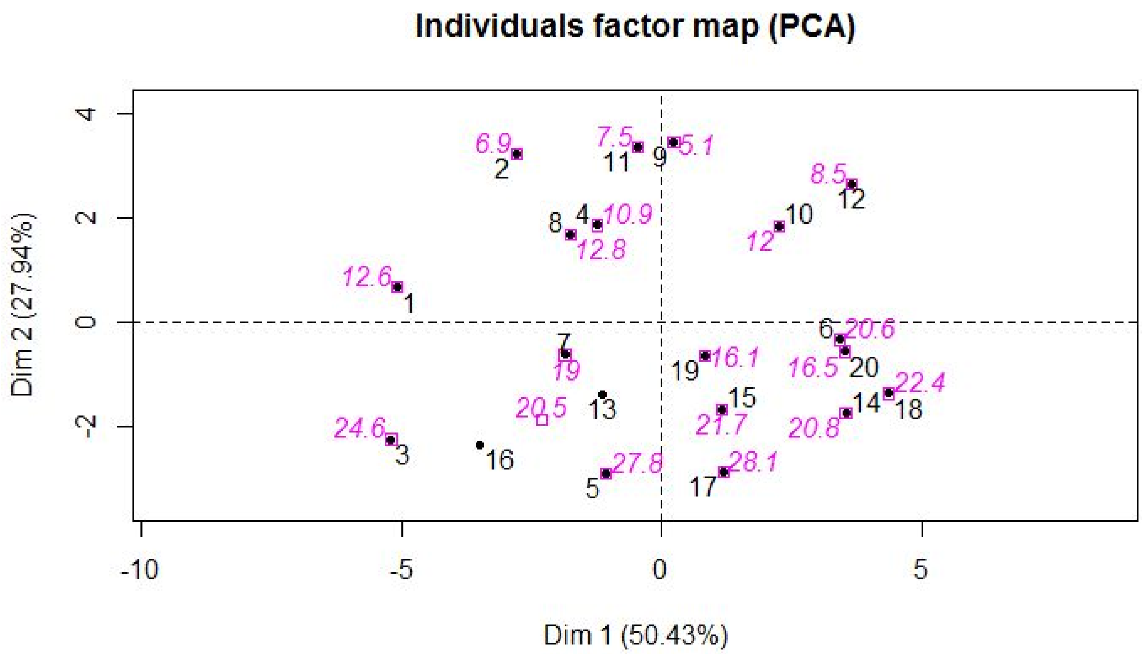
1. Get the multivariate data. Clearly specify the data you have chosen in your report.
I have chosen fat data frame with 252 observations on the following 19 variables. A data set containing many physical measurements of 252 males. Most of the variables can be measured with a scale or tape measure. Can they be used to predict the percentage of body fat? If so, this offers an easy alternative to an underwater weighing technique.

2. Use FactoMineR package to study individuals:
(a) Plot the individuals in the plane corresponding to the first two principal components (PCs), see [1], p.31. Comment on the resulting cloud.

---

```
da = fat[1:20, c('body.fat', 'body.fat.siri', 'density', 'weight', 'height', 'BMI', 'ffweight', 'neck',
'chest', 'abdomen', 'hip', 'thigh', 'knee', 'ankle', 'bicep', 'forearm', 'wrist')]
daPCA = PCA(da, quali.sup = 1)
```

---




I suppose that 3 and 12 observations have most distance locations.

---

```
 da[3,]
  body.fat body.fat.siri density weight height  BMI ffweight neck chest abdomen  hip thigh
knee
3     24.6          25.3  1.0414    154  66.25 24.7      116   34  95.8    87.9 99.2  59.6 38.9
  ankle bicep forearm wrist
3    24  28.8    25.2  16.6
> da[12,]
   body.fat body.fat.siri density weight height  BMI ffweight neck chest abdomen   hip thigh
12      8.5           7.8  1.0812    216     76 26.3    197.7 39.4 103.6    90.9 107.7  66.2
```

```
   knee ankle bicep forearm wrist
12 39.2  25.9  37.2    30.2    19
```

---

We can see that all values quite different.
I suppose that 6 and 20 are nearest observations.

---

```
> da[6,]
  body.fat body.fat.siri density weight height  BMI ffweight neck chest abdomen   hip thigh
6     20.6          20.9  1.0502 210.25  74.75 26.5      167   39 104.5    94.4 107.8    66
  knee ankle bicep forearm wrist
6   42  25.6  35.7    30.6  18.8
> da[20,]
   body.fat body.fat.siri density weight height  BMI ffweight neck chest abdomen hip thigh
knee
20     16.5          16.5   1.061 211.75   73.5 27.6    176.8   40 106.2   100.5 109  65.8 40.6
   ankle bicep forearm wrist
20    24  37.1    30.1  18.2
```

---

How we can see some variables have nearest value, but another variables have a liitle bit difference.

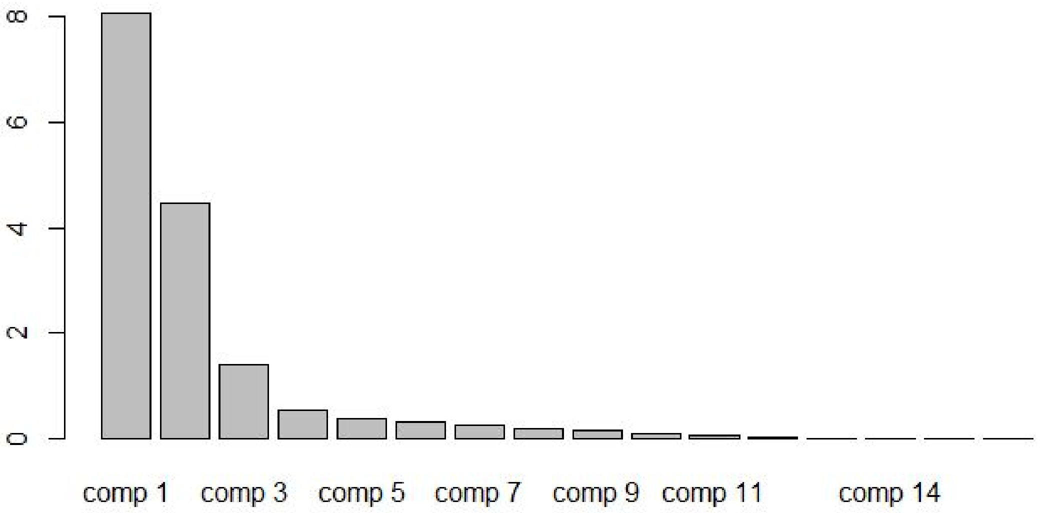
(b) Justify the choice of the PCs by plotting the eigenvalues, [1],p.32. Calculate how much of the total variability is explained by the first two PCs.

---

```
> daPCA$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1  8.069496e+00           5.043435e+01                          50.43435
comp 2  4.470734e+00           2.794209e+01                          78.37643
comp 3  1.410318e+00           8.814489e+00                          87.19092
comp 4  5.555207e-01           3.472005e+00                          90.66293
comp 5  3.744041e-01           2.340026e+00                          93.00295
comp 6  3.185253e-01           1.990783e+00                          94.99374
comp 7  2.459324e-01           1.537078e+00                          96.53081
comp 8  1.832155e-01           1.145097e+00                          97.67591
comp 9  1.542500e-01           9.640625e-01                          98.63997
comp 10 1.109994e-01           6.937465e-01                          99.33372
comp 11 7.062817e-02           4.414261e-01                          99.77515
comp 12 2.610373e-02           1.631483e-01                          99.93829
comp 13 9.148175e-03           5.717609e-02                          99.99547
comp 14 4.464029e-04           2.790018e-03                          99.99826
comp 15 2.396019e-04           1.497512e-03                          99.99976
comp 16 3.872232e-05           2.420145e-04                         100.00000

barplot(daPCA$eig[,1])
```

---

Total  variability is 78%




(c) Discuss the quality of the PCA representation: provide cos2 and the contributions for each individual, [1], p.34.

---

```
> daPCA$ind$cos2
        Dim.1       Dim.2        Dim.3        Dim.4        Dim.5
1  0.909056196 0.015677314 2.273868e-02 2.561320e-02 0.0097256425
2  0.393084264 0.523594098 8.213356e-06 2.690532e-02 0.0015969037
3  0.737246495 0.138353134 6.320574e-02 7.269185e-03 0.0002250309
4  0.220606495 0.490736797 1.022092e-01 1.018465e-01 0.0001605565
5  0.061858928 0.460770676 4.138262e-01 1.032570e-03 0.0028474143
6  0.751147197 0.007331084 2.113584e-01 2.669609e-03 0.0005304083
7  0.442875379 0.050833663 1.847866e-02 3.436990e-01 0.0979862471
8  0.346485292 0.316883179 3.903280e-02 2.212642e-02 0.1451934124
9  0.003646205 0.827602943 3.988456e-02 2.161839e-04 0.0503427941
10 0.393682205 0.264505547 7.679735e-02 1.232101e-01 0.0322666152
11 0.015683578 0.810770762 9.405934e-02 2.626603e-03 0.0249749523
12 0.564743647 0.301212012 1.236615e-02 2.775828e-02 0.0542478446
13 0.201089877 0.297233198 2.498179e-01 6.295956e-03 0.0010505742
14 0.738818944 0.181887420 6.795608e-05 1.681673e-03 0.0002593462
15 0.137960513 0.290550580 1.890347e-01 2.432921e-01 0.0093158229
16 0.622062511 0.285326086 3.392780e-02 3.331273e-05 0.0069416330
17 0.114502234 0.673402834 8.809909e-04 1.517478e-02 0.0157404024
18 0.803630078 0.079470371 4.702175e-02 2.995187e-03 0.0373029239
19 0.083586852 0.050148600 5.690950e-01 4.189721e-02 0.0417404370
```

20 0.823201110 0.021420593 4.859808e-07 3.833162e-02 0.0401630679

---

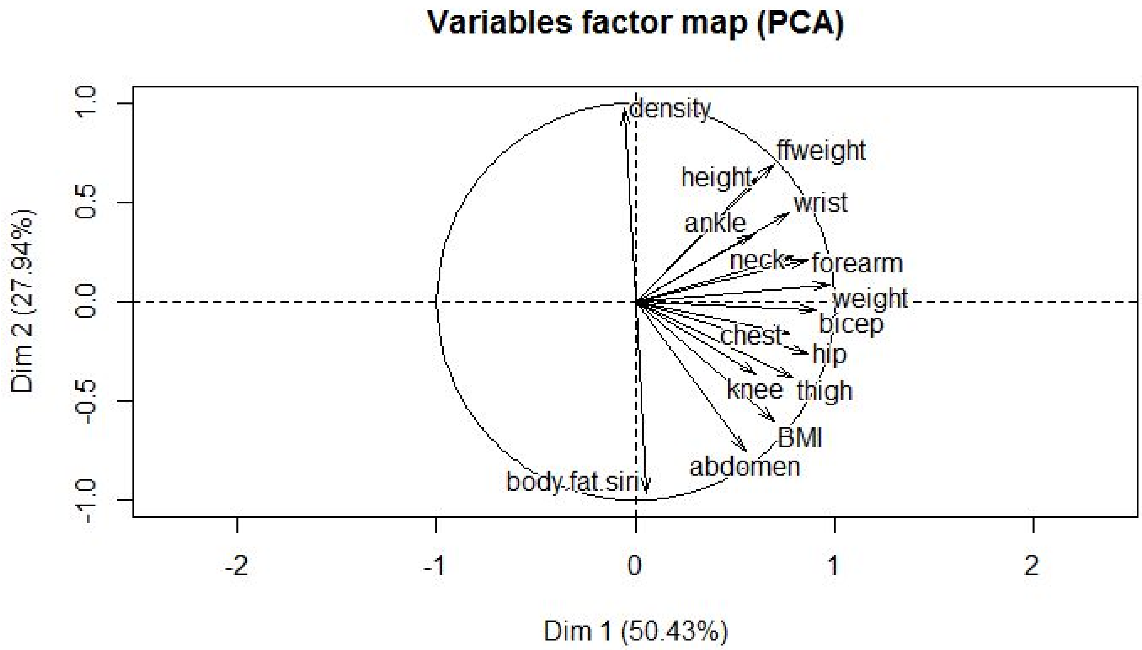
We can see that the first principal component is good for 1, 18 and 20 observations, but for huge amount of data we can observe values quite fewer than 1. It means that PCA can be not good for some observations.

(d) If there are categorical variables, paint the individuals with different colors according to the categories. Draw the confidence ellipses and interpret them, [1], p. 36.

This data frame doesn't have categorical variables.

3. Study cloud of variables, [1], pp. 36-44.

(a) Using the graphical output of pca command, discuss correlation between the variables including presence of groups of variables that are closely related.




We can see that variables density, body.fat.siri, ffweight and weight are well represented. Variable density has negative correlation, but all other variables have positive correlation.I suppose that variable body.fat.siri doesn't have correlation with over varibles.

(b) Discuss the quality of the PCA representation: provide cos2 and the contributions for variables.

---

> daPCA$var$cos2

| | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|---|---|---|---|---|---|
| body.fat.siri | 0.002500561 | 0.938236655 | 0.0269836415 | 0.0008702002 | 0.006115619 |
| density | 0.003005019 | 0.938861046 | 0.0262967687 | 0.0006504348 | 0.006908625 |
| weight | 0.949163290 | 0.007392634 | 0.0102234778 | 0.0126375265 | 0.005072404 |
| height | 0.373784887 | 0.391309542 | 0.1468190182 | 0.0023898770 | 0.008222228 |
| BMI | 0.484612462 | 0.365111038 | 0.1086145856 | 0.0088440191 | 0.001244063 |
| ffweight | 0.482264133 | 0.479974114 | 0.0002692385 | 0.0153568776 | 0.013918885 |
| neck | 0.621020742 | 0.050398739 | 0.0727969172 | 0.1305385891 | 0.019751788 |

```
chest        0.590071357 0.025341610 0.0840546663 0.2083855454 0.056686053
abdomen      0.303132616 0.564268767 0.0135808361 0.0001898566 0.004790774
hip          0.750807842 0.069204165 0.0092608420 0.0480116887 0.021502439
thigh        0.626909769 0.147810348 0.0215682644 0.0489523829 0.068846092
knee         0.366882384 0.131094457 0.3853352836 0.0263995025 0.013044141
ankle        0.349934166 0.116659021 0.3735316098 0.0037936449 0.000302936
bicep        0.822642549 0.002112558 0.0702707062 0.0038994475 0.020935153
forearm      0.746271249 0.043507474 0.0605295279 0.0306340607 0.001275266
wrist        0.596492733 0.199451592 0.0001827893 0.0139670812 0.125787654
```

---

As we can see variable weight has good representation in the first PCA. Variables bode.fat.siri and density have good representation in the second PCA.

(c) Use dimdesc function to summarize the variables. Comment on the p-values.

---

```
> dimdesc(daPCA)
$Dim.1
$Dim.1$quanti
         correlation      p.value
weight     0.9742501 4.305716e-13
bicep      0.9069964 3.513505e-08
hip        0.8664917 7.804820e-07
forearm    0.8638699 9.205193e-07
thigh      0.7917763 3.190378e-05
neck       0.7880487 3.688202e-05
wrist      0.7723294 6.598178e-05
chest      0.7681610 7.640833e-05
BMI        0.6961411 6.515759e-04
ffweight   0.6944524 6.801589e-04
height     0.6113795 4.180976e-03
knee       0.6057082 4.649000e-03
ankle      0.5915523 6.008691e-03
abdomen    0.5505748 1.188267e-02


$Dim.2
$Dim.2$quanti
              correlation      p.value
density         0.9689484 2.277364e-12
ffweight        0.6928016 7.091187e-04
height          0.6255474 3.179393e-03
wrist           0.4466000 4.838213e-02
BMI            -0.6042442 4.776592e-03
abdomen        -0.7511783 1.348924e-04
body.fat.siri  -0.9686262 2.496186e-12
body.fat       -0.9688345 2.352687e-12
```

```
$Dim.3
$Dim.3$quanti
      correlation     p.value
knee    0.6207538 0.003493088
ankle   0.6111723 0.004197358
```

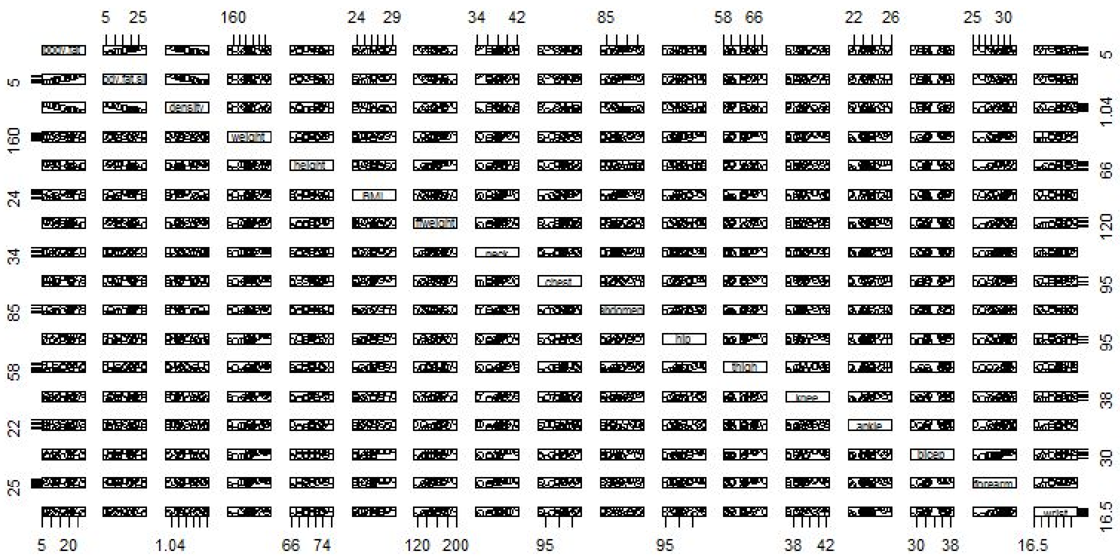
---

I suppose that in the first PCA and in the second we have significant variables.
(d) Plot the correlations between variables using pairs function. Compare the result with that of 3a.

---

```
> pairs(da)
```

---



I’m sorry for that unclear plot. When I prepared my homework I observed this plot with zoom. We can say that body.fat.siri and body.fat are positive strongly correlated. Body.fat and body.fat.siri negative strongly correlated with variable density. Variable Density hasn’t good indicators about correlation with other variables.  All observers are similar to figure 1.3.

**2 Assignment on Correspondence Analysis**

1. Get the multivariate data.
Dataset: femart.dat

Source: T.H. Bradford and L. Idleman (1991), "Feminist Art Theory
in Atlanta: The Politcal Climate", Art Journal, Volume 50, #2, pp 14-18.

Description: Education Level and Use of Feminist Art Theory in Work
Among Memebers of Atlanta art community.

Variables/Columns:

Education  8 /* 1=HS, 2=Bacheor's, 3=Master's, 4=Ph.D.
Use of Feminist Art Theory   16  /* 1=No, 2=Somewhat, 3=Yes

---

```
> a1 = fread("http://www.stat.ufl.edu/~winner/data/femart.dat")
trying URL 'http://www.stat.ufl.edu/~winner/data/femart.dat'
Content type 'text/plain' length 2771 bytes
downloaded 2771 bytes

> a2 = table(a1)
```

---

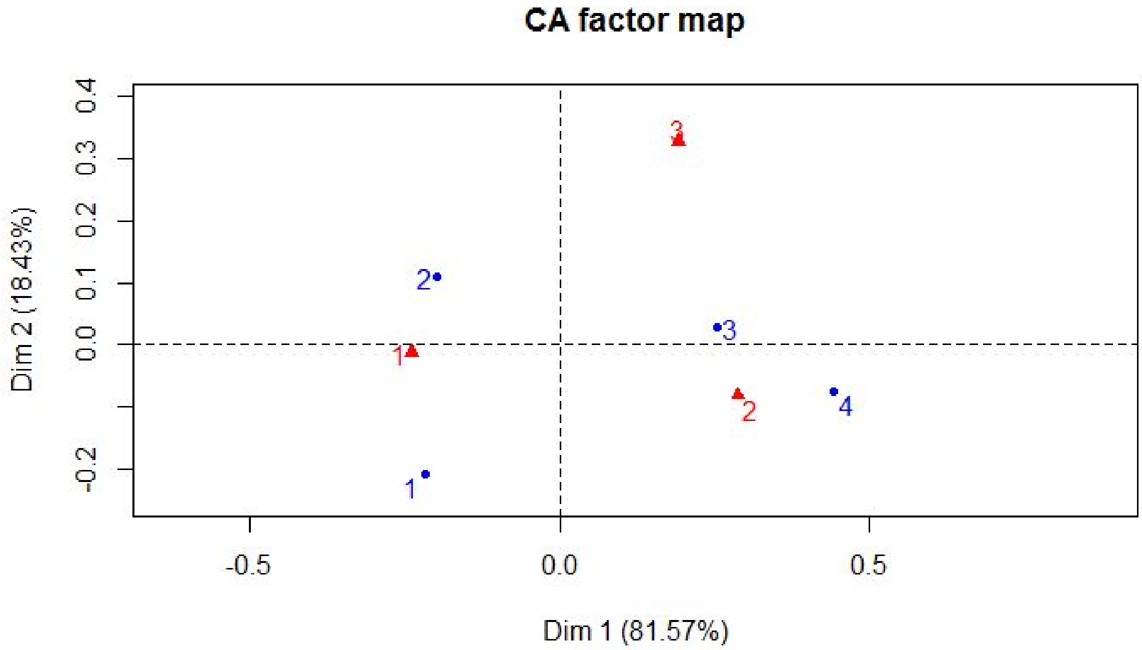
2. Use FactoMineR package to:
 (a) Do the $\chi$ 2 test for independence and interpret it, see Section 2.2.2 of [1].

---

```
> CA(a2)
**Results of the Correspondence Analysis (CA)**
The row variable has  4  categories; the column variable has 3 categories
The chi square of independence between the two variables is equal to 12.74221 (p-value =
0.04731735 ).
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$col"            "results for the columns"
3  "$col$coord"      "coord. for the columns"
4  "$col$cos2"       "cos2 for the columns"
5  "$col$contrib"    "contributions of the columns"
6  "$row"            "results for the rows"
7  "$row$coord"      "coord. for the rows"
8  "$row$cos2"       "cos2 for the rows"
9  "$row$contrib"    "contributions of the rows"
10 "$call"           "summary called parameters"
11 "$call$marge.col" "weights of the columns"
12 "$call$marge.row" "weights of the rows"
```

---

p-value is 0.04731735, that means we can reject H0 hypothesis and variables are independent.
(b) Perform the CA, get the 2D representation of row and column profiles • separately • in the same graph See p.87 of of [1]

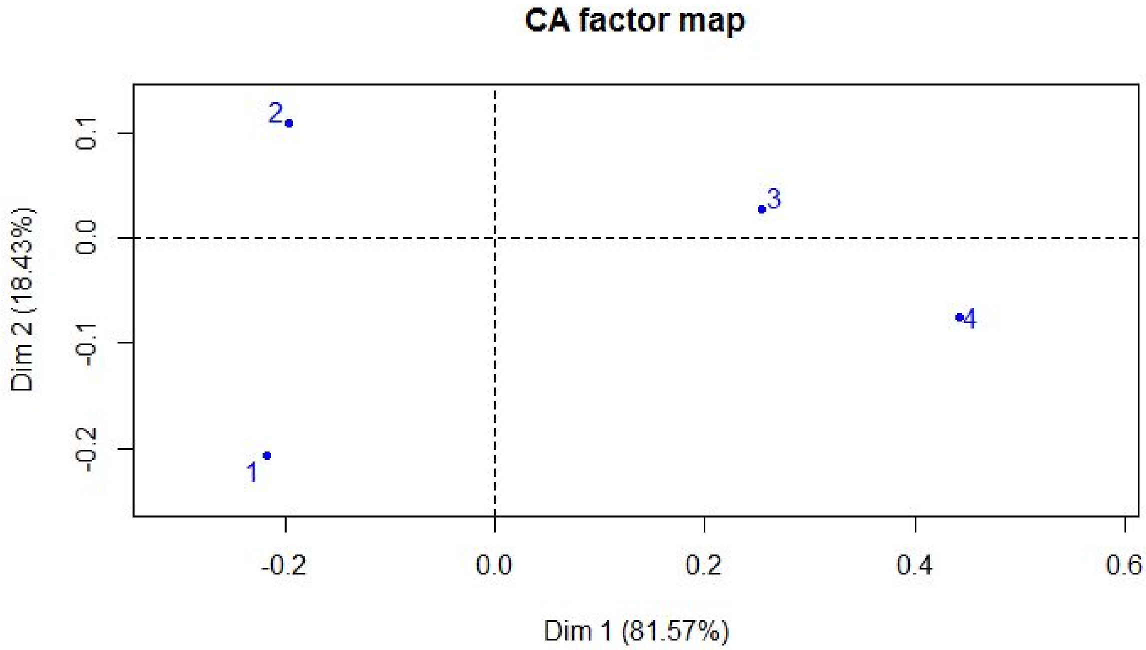


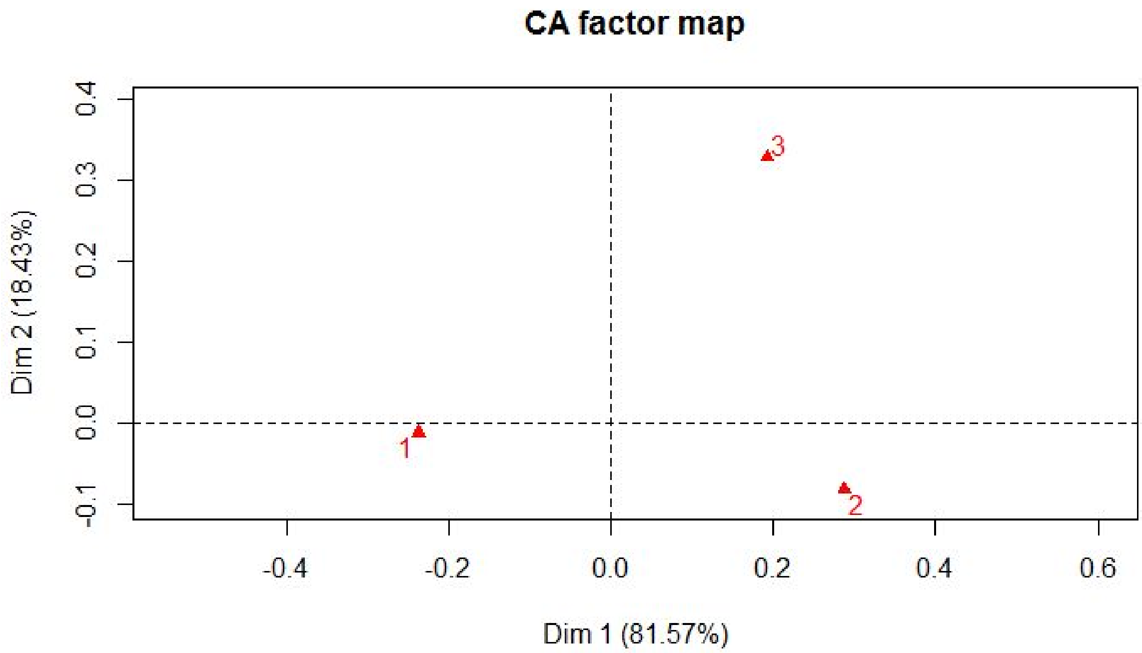

(c) Analyze the patterns obtained in item 2b. Focus on the total variability, similarities/dissimilarities and the conclusions that can be made from the simultaneous representation of rows and columns. See examples, [1], pp. 92-125.

---

```
> daCA = CA(a2)
> plot(daCA, invisible = "col")

> plot(daCA, invisible = "row")
```

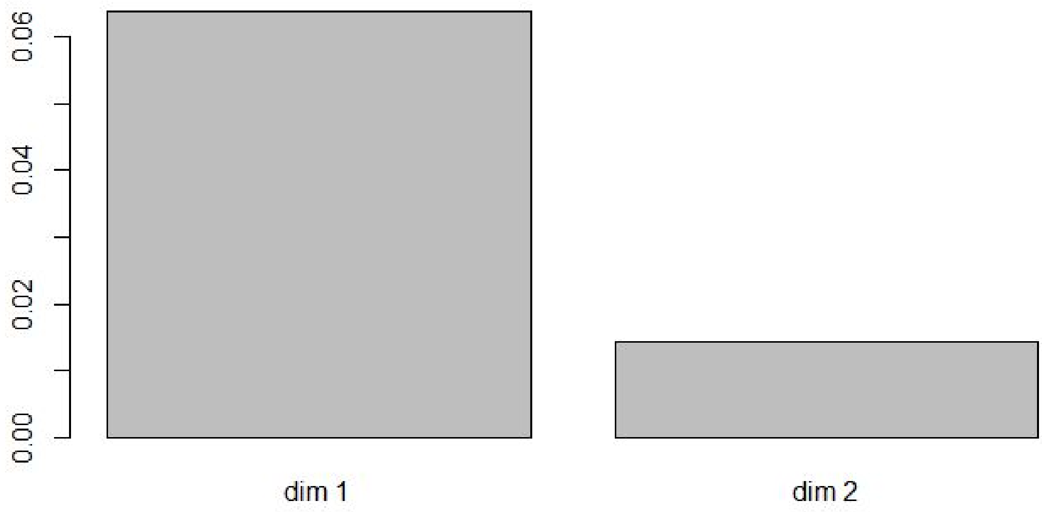
---




(d) Provide the table and graph of eigenvalues, justify the choice of principal components.

```
> daCA$eig
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.06376664            81.57113                          81.57113
dim 2 0.01440641            18.42887                         100.00000
> barplot(daCA$eig[,1])
```




(e) Discuss the quality of the CA representation based on cos2 for rows and columns, [1], p.87.

```
> daCA$col$cos2
      Dim 1      Dim 2
1 0.9972535 0.00274648
2 0.9248620 0.07513795
3 0.2555170 0.74448295
> daCA$row$cos2
      Dim 1      Dim 2
1 0.5237529 0.47624712
2 0.7627731 0.23722688
3 0.9887069 0.01129313
4 0.9721622 0.02783777
```

**3 Assignment on Multiple Correspondence Analysis**

1. Get the multivariate data.

```
> d1 = read.csv('sfo cust sat 2014 data file_WEIGHTED_flysfo.csv', header = T, sep=";")
> d2=da3[1:20, c("Q4FOOD","Q4STORE","Q4WIFI","Q4BAGS")]
Error: object 'da3' not found
> d2=d1[1:20, c("Q4FOOD","Q4STORE","Q4WIFI","Q4BAGS")]
> tail(d2)
  Q4FOOD Q4STORE Q4WIFI Q4BAGS
```

```
15     2     1     1     2
16     2     1     2     2
17     1     1     2     2
18     2     1     1     2
19     1     0     0     0
20     1     2     1     2
> str(d2)
'data.frame':    20 obs. of  4 variables:
 $ Q4FOOD : int  1 2 2 0 2 1 1 2 1 1 ...
 $ Q4STORE: int  2 2 2 1 2 2 2 2 1 2 ...
 $ Q4WIFI : int  2 2 1 1 2 1 1 2 2 1 ...
 $ Q4BAGS : int  2 1 2 1 2 2 2 2 2 2 ...
> d2$Q4FOOD=as.factor(d2$Q4FOOD)
> d2$Q4STORE=as.factor(d2$Q4STORE)
> d2$Q4WIFI=as.factor(d2$Q4WIFI)
> d2$Q4BAGS=as.factor(d2$Q4BAGS)
> str(d2)
'data.frame':    20 obs. of  4 variables:
 $ Q4FOOD : Factor w/ 3 levels "0","1","2": 2 3 3 1 3 2 2 3 2 2 ...
 $ Q4STORE: Factor w/ 3 levels "0","1","2": 3 3 3 2 3 3 3 3 2 3 ...
 $ Q4WIFI : Factor w/ 3 levels "0","1","2": 3 3 2 2 3 2 2 3 3 2 ...
 $ Q4BAGS : Factor w/ 3 levels "0","1","2": 3 2 3 2 3 3 3 3 3 3 ...
```

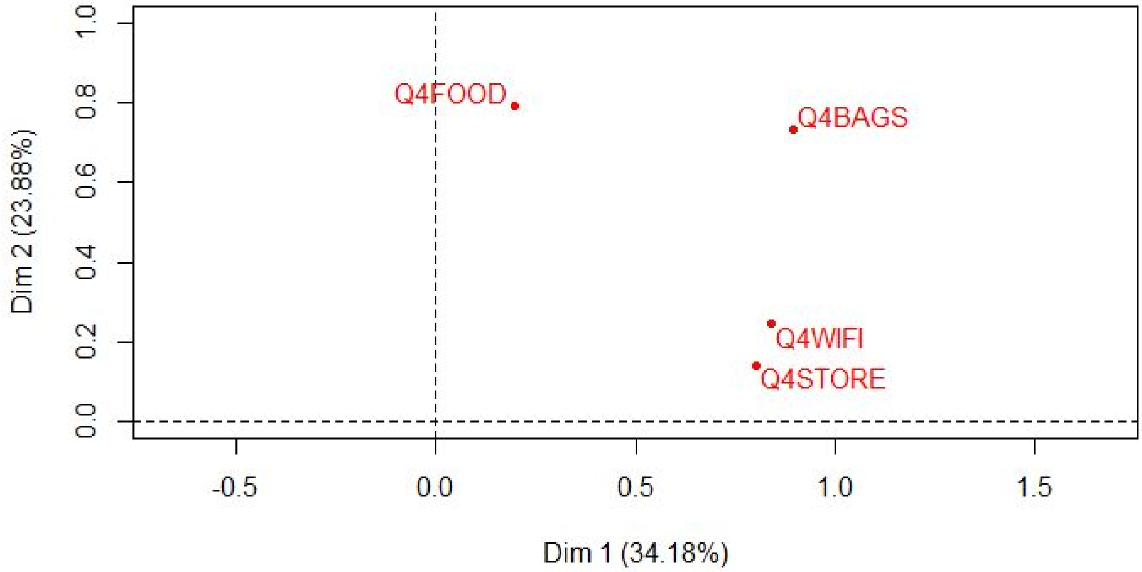
---

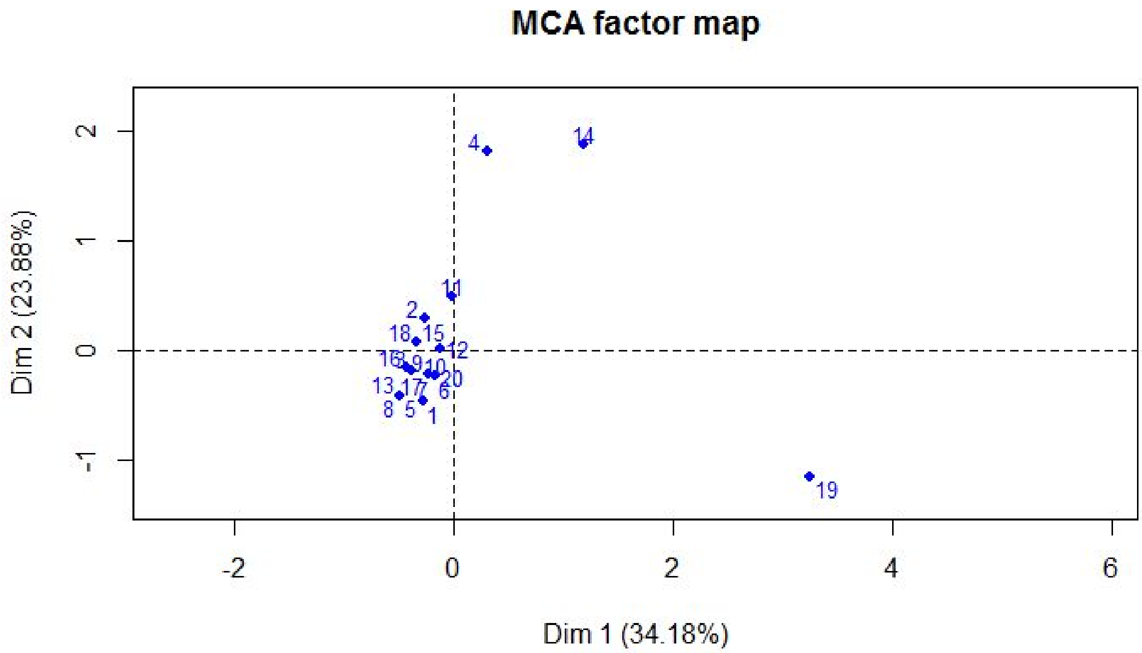
2. Use FactoMineR package:
 (a) Conduct the MCA. Visualize individuals and categories, see Section 3.6 of [1].

---

```
> MCA(d2)
**Results of the Multiple Correspondence Analysis (MCA)**
The analysis was performed on 20 individuals, described by 4 variables
*The results are available in the following objects:

   name              description
1  "$eig"            "eigenvalues"
2  "$var"            "results for the variables"
3  "$var$coord"      "coord. of the categories"
4  "$var$cos2"       "cos2 for the categories"
5  "$var$contrib"    "contributions of the categories"
6  "$var$v.test"     "v-test for the categories"
7  "$ind"            "results for the individuals"
8  "$ind$coord"      "coord. for the individuals"
9  "$ind$cos2"       "cos2 for the individuals"
10 "$ind$contrib"    "contributions of the individuals"
11 "$call"           "intermediate results"
12 "$call$marge.col" "weights of columns"
13 "$call$marge.li"  "weights of rows"
```

---




(b) Provide a detailed interpretation of the obtained patterns. Focus on variability of individuals and categories, comment on the extreme cases.

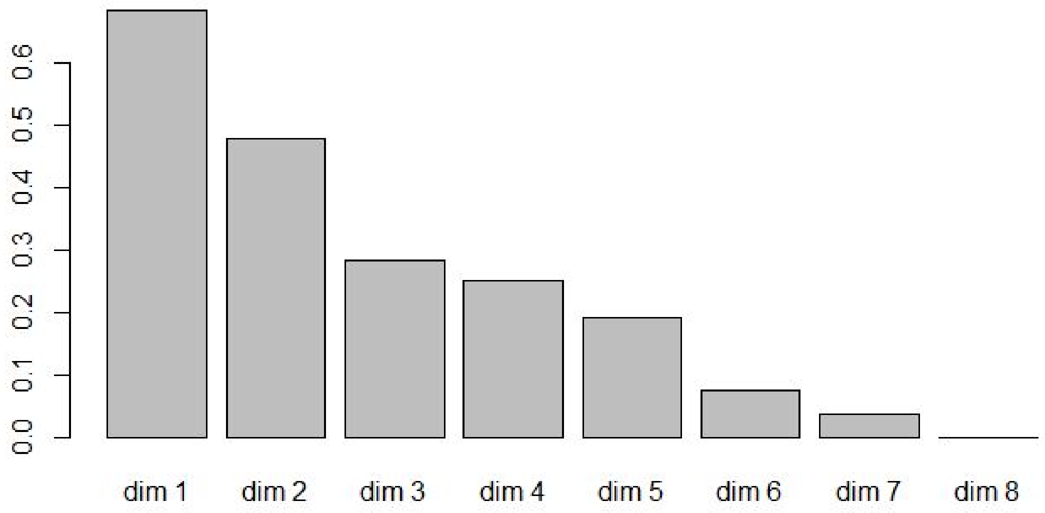
---

```
> d2[4,]
  Q4FOOD Q4STORE Q4WIFI Q4BAGS
4      0       1      1      1
> d2[19,]
   Q4FOOD Q4STORE Q4WIFI Q4BAGS
```

19 1 0 0 0

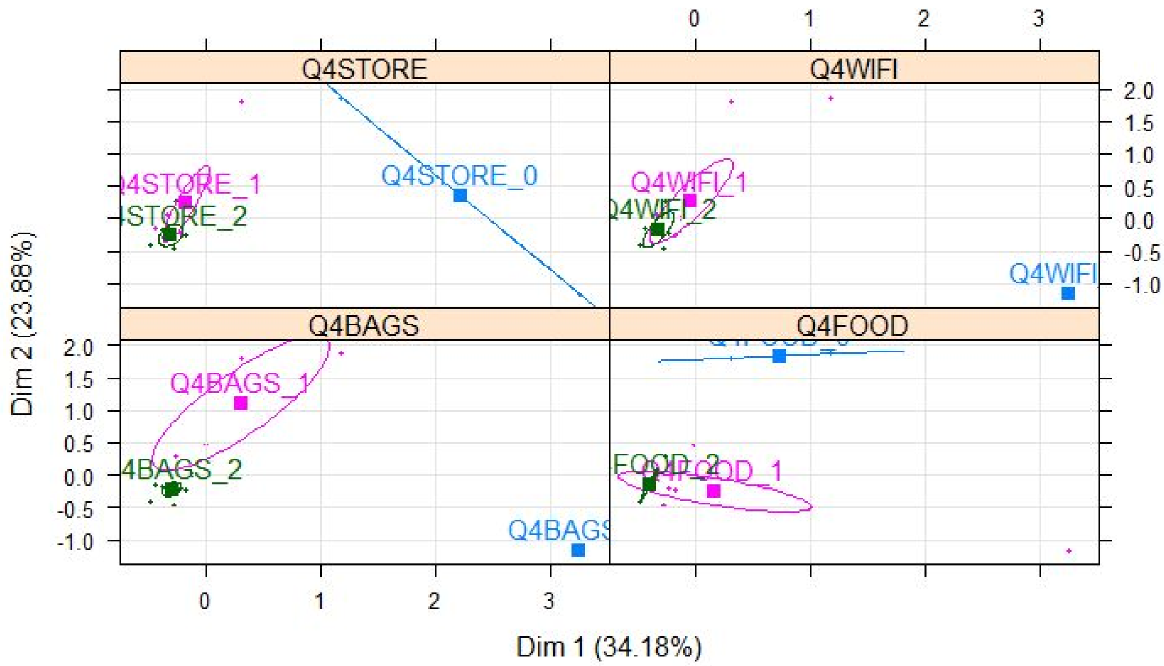
(c) Provide a table of eigenvalues, comment on the values of the largest ones and justify the choice of principal components. Do you need to look at the PCs other than the first two ones?

```
> d4=MCA(d2)$eig
> barplot(d4[,1])
```




(d) Draw the confidence ellipses around the categories and interpret the results, p.147 of [1].

```
> plotellipses(MCA(d2))
```



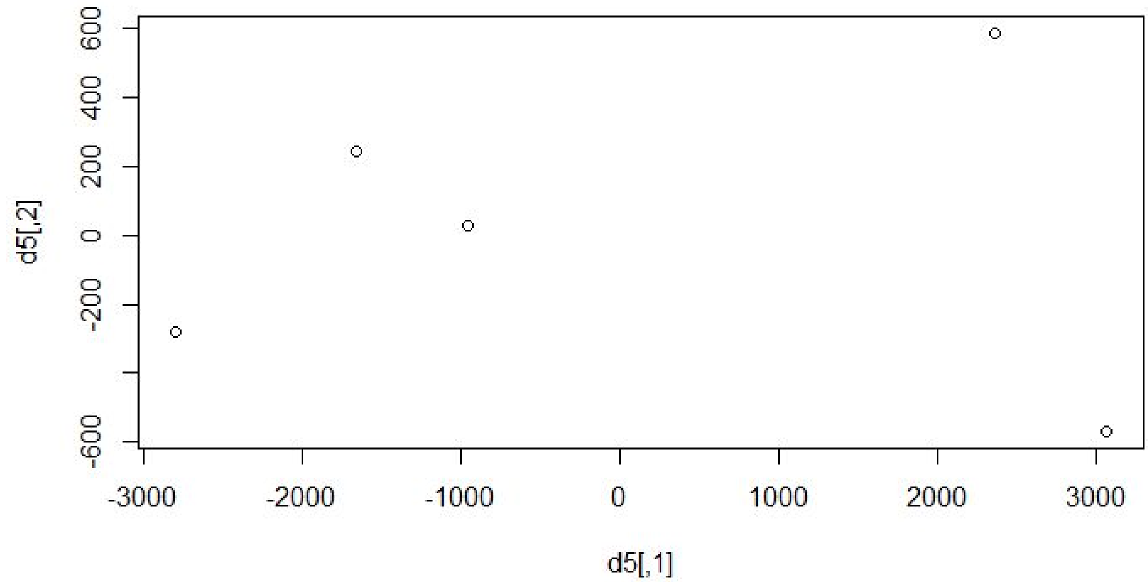

**4 Assignment on Multidimensional Scaling**

Get the distances between 10-12 Russian cities. You can retrieve this information at https://www.avtodispetcher.ru/distance/table/c172-rossiya/

```
 dMDS = matrix(0,5,5)
> dMDS
     [,1] [,2] [,3] [,4] [,5]
[1,]    0 3391 5229 1421 4106
[2,] 3391    0 1958 4061  717
[3,] 5229 1958    0 5899 1599
[4,] 1421 4061 5899    0 4808
[5,] 4106  717 1559 4808    0
```

1. Do the classical multidimensional scaling using command cmdscale from MASS package:
(a) Plot a two-dimensional MDS configuration representing the cities. Compare the result with the actual geographical location of the cities across the country.

```
> dist = as.dist(dMDS)
> d5 =cmdscale(dist)
> plot(d5)
```




(b) Based on the computed eigenvalues, discuss the quality of representation in the 2D space.

---

```
> eigen(dMDS)
eigen() decomposition
$values
[1] 13557.8395 -9324.8557 -2198.6313 -1380.0380  -654.3145

$vectors
           [,1]       [,2]       [,3]        [,4]          [,5]
[1,] -0.4546269 -0.4623658 -0.1142902 -0.75108843  0.0523062972
[2,] -0.3594804  0.2802791 -0.5708791  0.08543286 -0.6844413742
[3,] -0.4963807  0.4892043  0.6939363 -0.11628040 -0.1253082803
[4,] -0.5114280 -0.5622463  0.1404968  0.63472766  0.0005485956
[5,] -0.3951403  0.3901534 -0.3996847  0.11030441  0.7163110990
```

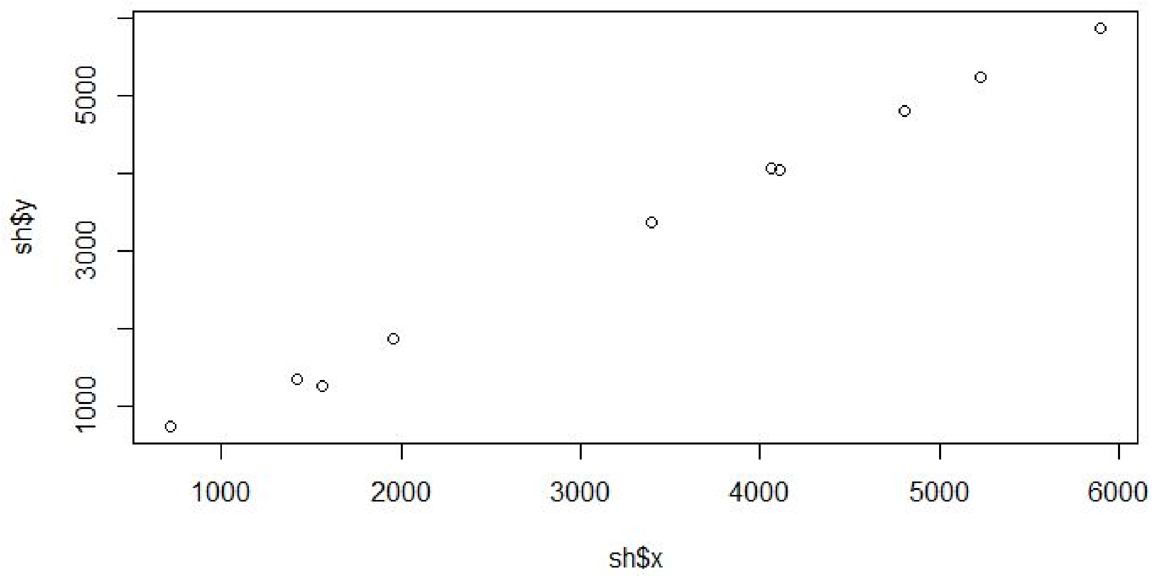
---

(c) Plot the Shepard diagram and discuss it.

---

```
f1 = cmdscale(dist, k = 2, eig = TRUE)
> sh = Shepard(dist, f1$points)
> plot(sh)
```

---




(d) Check whether the MDS configuration you obtained does restore the original distances in a sufficiently high dimensional space.

---

**5 Assignment on Clustering**

Get the built-in data with at least 4 quantitative (continuous) variables.

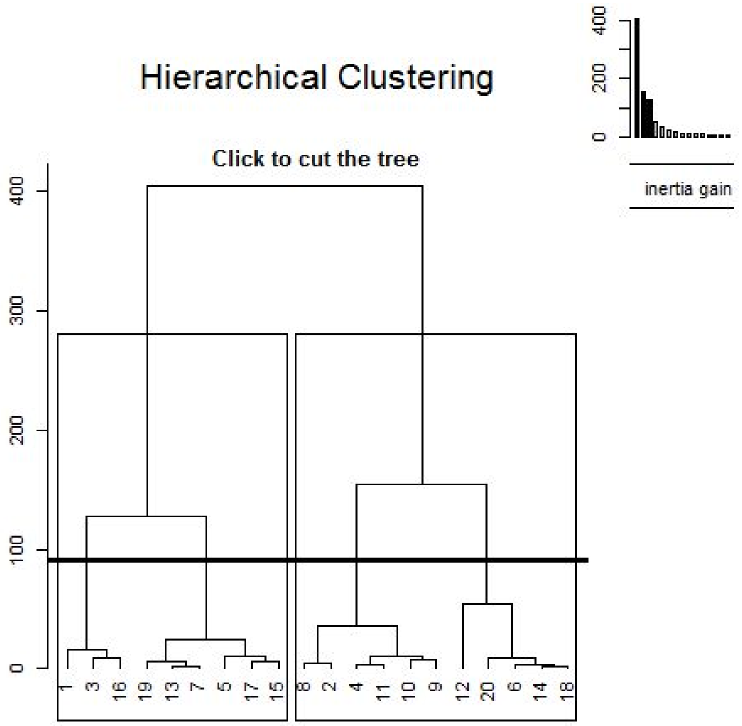
1. Do the hierarchical clustering (preceded by the PCA) using command HCPC from FactoMineR package:

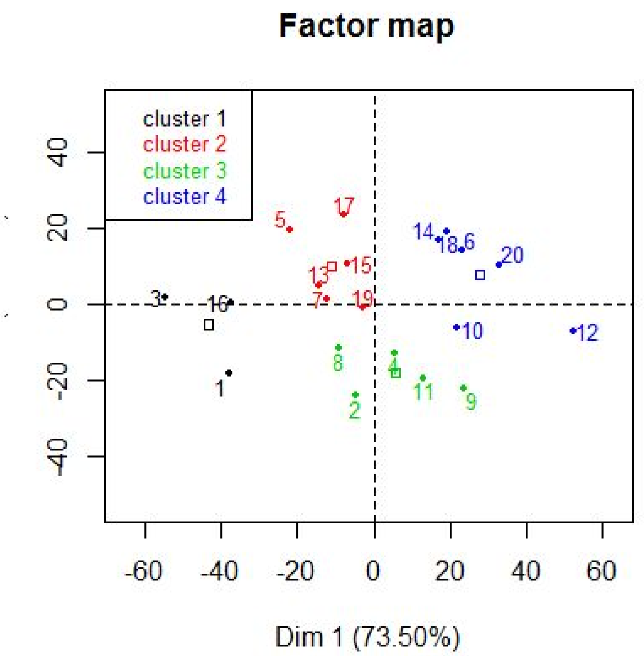
(a) Clearly name the recommended (by HCPC) clusters.

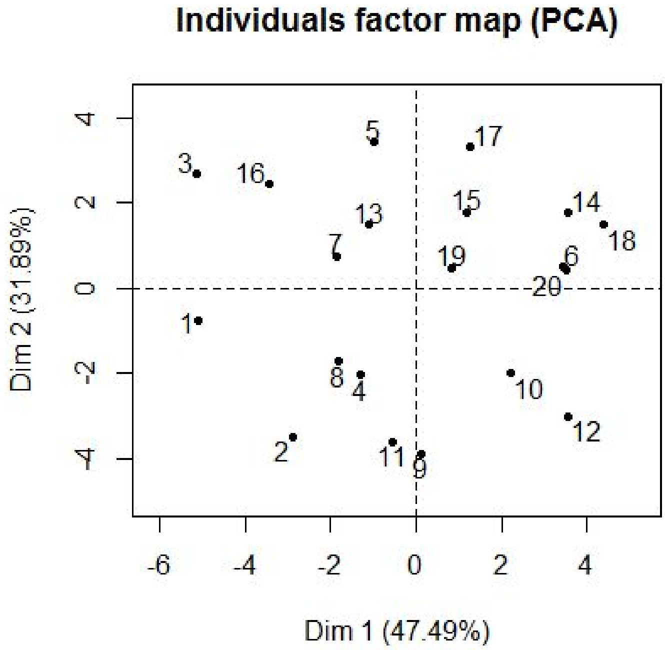
---

```
> da = fat[1:20, c('body.fat', 'body.fat.siri', 'density', 'weight', 'height', 'BMI', 'ffweight', 'neck',
'chest', 'abdomen', 'hip', 'thigh', 'knee', 'ankle', 'bicep', 'forearm', 'wrist')]
> HCPC(da, nb.clust=-1)
**Results for the Hierarchical Clustering on Principal Components**
   name                    description
1  "$data.clust"           "dataset with the cluster of the individuals"
2  "$desc.var"             "description of the clusters by the variables"
3  "$desc.var$quanti.var"  "description of the cluster var. by the continuous var."
4  "$desc.var$quanti"      "description of the clusters by the continuous var."
5  "$desc.axes"            "description of the clusters by the dimensions"
6  "$desc.axes$quanti.var" "description of the cluster var. by the axes"
7  "$desc.axes$quanti"     "description of the clusters by the axes"
8  "$desc.ind"             "description of the clusters by the individuals"
9  "$desc.ind$para"        "parangons of each clusters"
10 "$desc.ind$dist"        "specific individuals"
11 "$call"                 "summary statistics"
```

12 "$call$t"   "description of the tree"






(b) Explain the meaning of the barplot in the upper-right corner of the output.
2. Perform the K-means clustering, choosing K according to the results of hierarchical clustering.

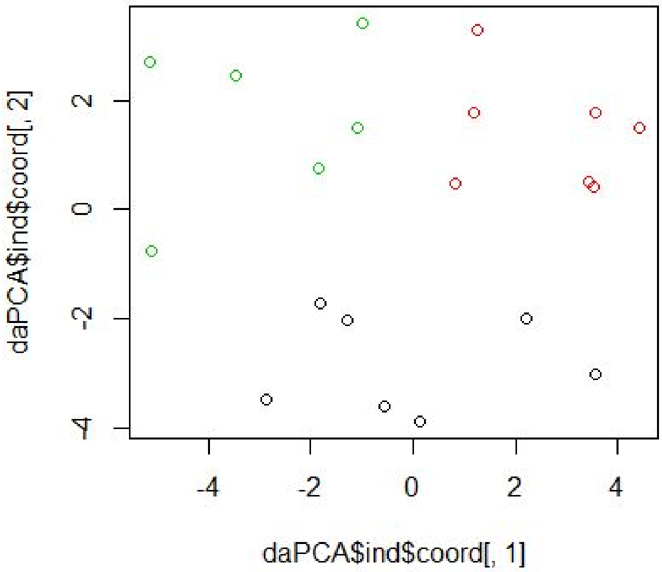
---

```
> kmeans(daPCA$ind$coord[,1:2],3)
K-means clustering with 3 clusters of sizes 7, 7, 6

Cluster means:
        Dim.1     Dim.2
1 -0.08475407 -2.817600
2  2.60526850  1.387734
3 -2.94060017  1.668177

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 3  1  3  1  3  2  3  1  1  1  1  1  3  2  2  3  2  2  2  2

Within cluster sum of squares by cluster:
[1] 35.83373 19.13446 29.70956
 (between_SS / total_SS =  68.6 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

```
> plot(daPCA$ind$coord[,1],daPCA$ind$coord[,2],col=kmeans(daPCA$ind$coord[,1:2],3)$cluster)
```

---

(a) Plot the results




(b) Compare distribution of points over clusters with that of hierarchical approach