Exploring Tunisian Emigration Patterns

Dridi Slim

2024-05-20

Contents

Introduction:

- Tunisia, a North African country with a rich history and diverse cultural heritage, has experienced significant emigration over the past few decades.
- In this project, we aim to study the various opinions on Tunisian emigration.
- The survey: The survey was created using Google Forms and distributed to respondents mainly through Gmail. The data was then collected in an Excel file and subsequently processed using the R programming language.

  survey link

Setting up the environment:

Importing packages:

```
library(readxl) #used for importing exel data
library(FactoMineR) #used for applying PCA, MCA & classification
library(factoextra) #used for applying PCA, MCA & classification
library(ggplot2) # visualization
library(dplyr) # data manipulation
library(corrplot) # plotting the corrolation
library(questionr)
library(flextable) # print pretty tables
```

Importing Data:

```
data <- read_excel("D://Analyse//PCA//final_data.xlsx")
data <- data.frame(data)
```

data overview:

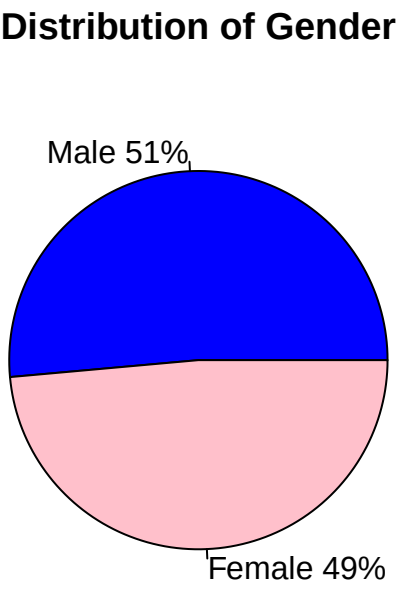
```
paste("In our data, We have",ncol(data),"variable And",nrow(data),"Observation")
```

```
## [1] "In our data, We have 43 variable And 35 Observation"
```

Descriptive statistics:

Sexe Pie chart:

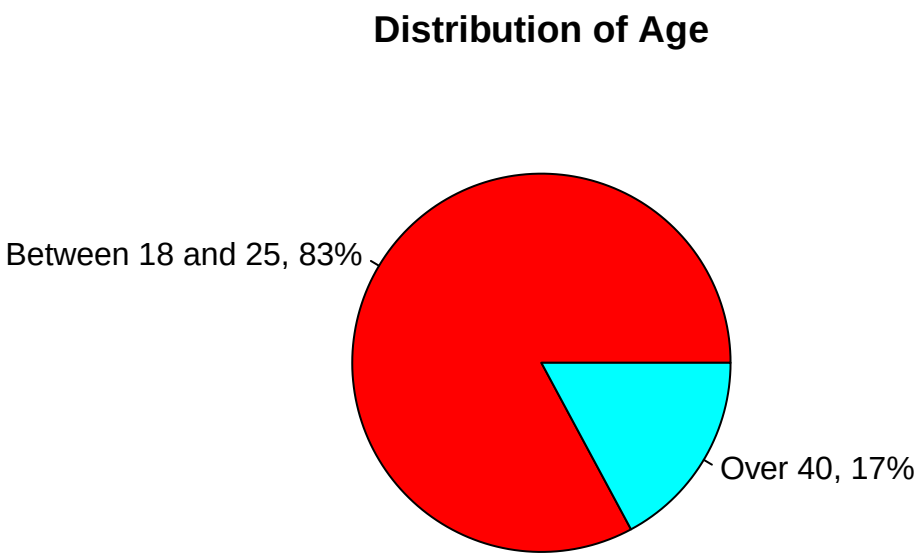
```
data$sexe<-factor(data$sexe,levels=c("Male","Female"))
freq=table((data$sexe))
perc<- round(freq/sum(freq)*100)
labls1 <- paste(c("Male","Female"),perc)
labls1 <- paste(labls1,"%",sep="")
pie(freq,labels = labls1,col=c("blue", "pink"),main="Distribution of Gender")
```



The chart effectively visualizes the nearly equal distribution between males and females, highlighting gender eaqual distribution within the dataset.

Age Pie chart:

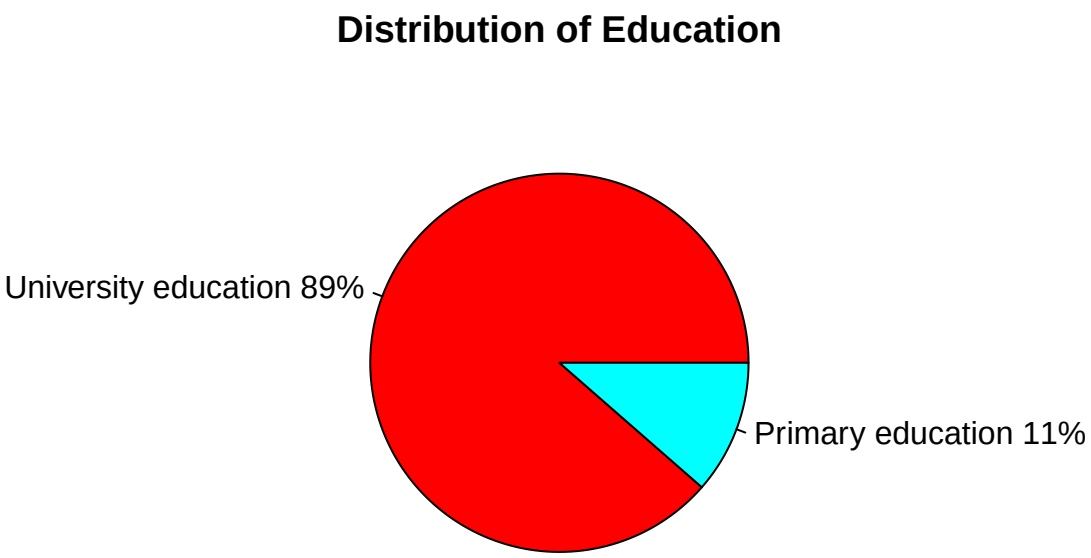
```
data$sexe<-factor(data$Age,levels=c("Between 18 and 25","Over 40"))
freq=table((data$Age))
perc <- round(freq/sum(freq)*100)
labls1 <- paste(c("Between 18 and 25,","Over 40,"),perc)
labls1 <- paste(labls1,"%",sep="")
pie(freq,labels = labls1, col=rainbow(length(labls1)),main="Distribution of Age")
```



The chart highlights a notable disparity in age distribution, with a substantial majority of individuals falling within the 18 to 25 age range.

Education Pie chart:

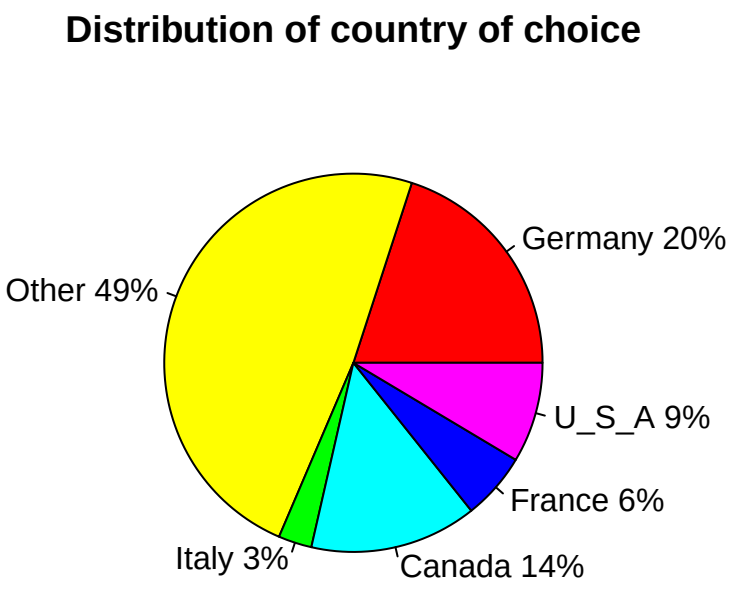
```
data$education<-factor(data$education,levels=c("University education","Primary education"))
freq=table((data$education))
perc <- round(freq/sum(freq)*100)
labls1 <- paste(c("University education","Primary education"),perc)
labls1 <- paste(labls1,"%",sep="")
pie(freq,labels = labls1, col=rainbow(length(labls1)),main="Distribution of Education")
```



This char reveals a significant educational trend.

country of choice Pie chart:

```
data$Country_of_choice<-factor(data$Country_of_choice,levels=c("Germany","Other","Italy","Canada","Fran
freq=table((data$Country_of_choice))
perc <- round(freq/sum(freq)*100)
labls1 <- paste(c("Germany","Other","Italy","Canada","France","U_S_A"),perc)
labls1 <- paste(labls1,"%",sep="")
pie(freq,labels = labls1, col=rainbow(length(labls1)),main="Distribution of country of choice")
```



Principal Component Analysis (PCA):

Perform Principal Component Analysis:

```
X2=as.matrix(data[,22:31])
res.pca2 <- PCA (X2,graph=F)
```

Calculating eigenvalues from PCA results:

```
#get the eignvalue
eig_val2 <- res.pca2$eig

#transform to a better table
eig_val2<-data.frame(eig_val2)
DIM <- paste("Dim.", 1:10, sep = "")
eig_val2 <- cbind(DIM,eig_val2)
ft <- flextable(eig_val2)
ft <- autofit(ft)
theme_zebra(ft)
```

DIM	eigenvalue	percentage.of.variance	cumulative.percentage.of.variance
Dim.1	4.8725584	48.725584	48.72558
Dim.2	1.5640172	15.640172	64.36576
Dim.3	0.8640884	8.640884	73.00664
Dim.4	0.7713012	7.713012	80.71965
Dim.5	0.7023885	7.023885	87.74354
Dim.6	0.3520669	3.520669	91.26421
Dim.7	0.2852119	2.852119	94.11633
Dim.8	0.2331341	2.331341	96.44767
Dim.9	0.2140466	2.140466	98.58813
Dim.10	0.1411867	1.411867	100.00000

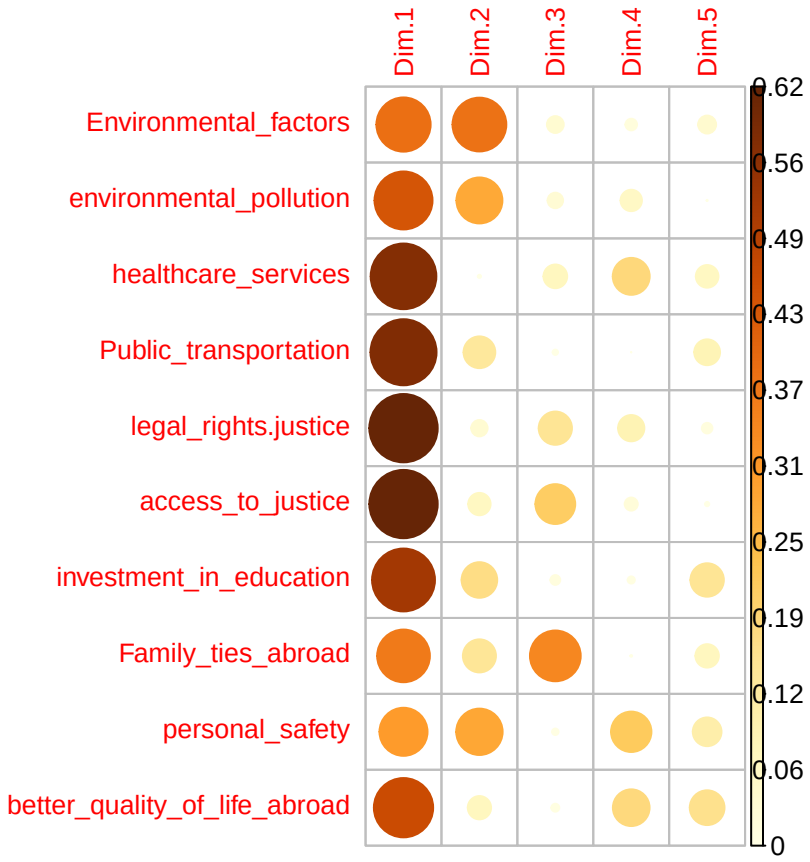
Extracting variables from PCA results:

```
var1 <- get_pca_var(res.pca2)
var1
```

```
## Principal Component Analysis Results for variables
##  ===================================================
##   Name       Description
## 1 "$coord"   "Coordinates for the variables"
## 2 "$cor"     "Correlations between variables and dimensions"
## 3 "$cos2"    "Cos2 for the variables"
## 4 "$contrib" "contributions of the variables"
```

Correlation plot using $\cos^2$:

```
corrplot(var1$cos2,tl.cex = 0.8, is.corr=F)
```



The higher the cos2, the better the variable is represented. Our variables demonstrate high correlation with the first dimension.

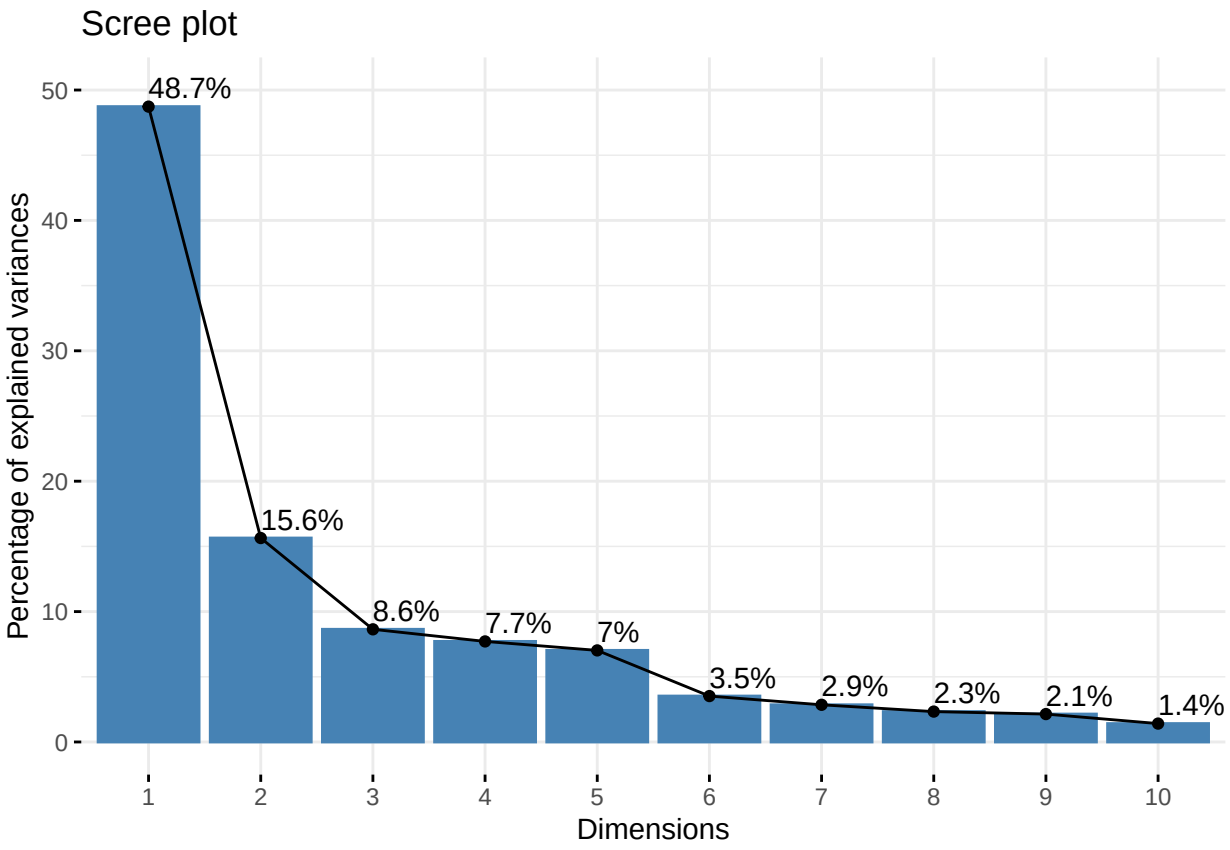
Coordinates of variables:

```
s<-res.pca2$var$coord
s<-data.frame(s)
Question <- paste("Q.", 1:10, sep = "")
s <- cbind(Question,s)
ft <- qflextable(s)
ft <- autofit(ft)
theme_zebra(ft)
```

Question	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Q.1	0.6230262	-0.61839615	0.192847731	-0.133179701	0.20620532
Q.2	0.6665378	-0.52845977	-0.177717284	0.245623500	-0.01368835
Q.3	0.7550749	0.03664609	0.271895785	-0.424221346	-0.26001239
Q.4	0.7593689	-0.36451374	-0.003391608	0.005584952	0.29618434
Q.5	0.7858247	0.18773369	-0.381839069	-0.301493335	0.12066604

Question	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Q.6	0.7855554	0.25672899	-0.457397363	-0.149838926	-0.04881122
Q.7	0.7229593	0.40895478	0.111885154	0.081595308	-0.38363534
Q.8	0.6056028	0.37957597	0.583170606	-0.020432341	0.26900539
Q.9	0.5521912	0.53294066	-0.074231686	0.463941635	0.32956141
Q.10	0.6814598	-0.26761735	0.090807118	0.421392350	-0.39906813

```
fviz_eig(res.pca2, addlabels = TRUE, ylim = c(0, 50))
```

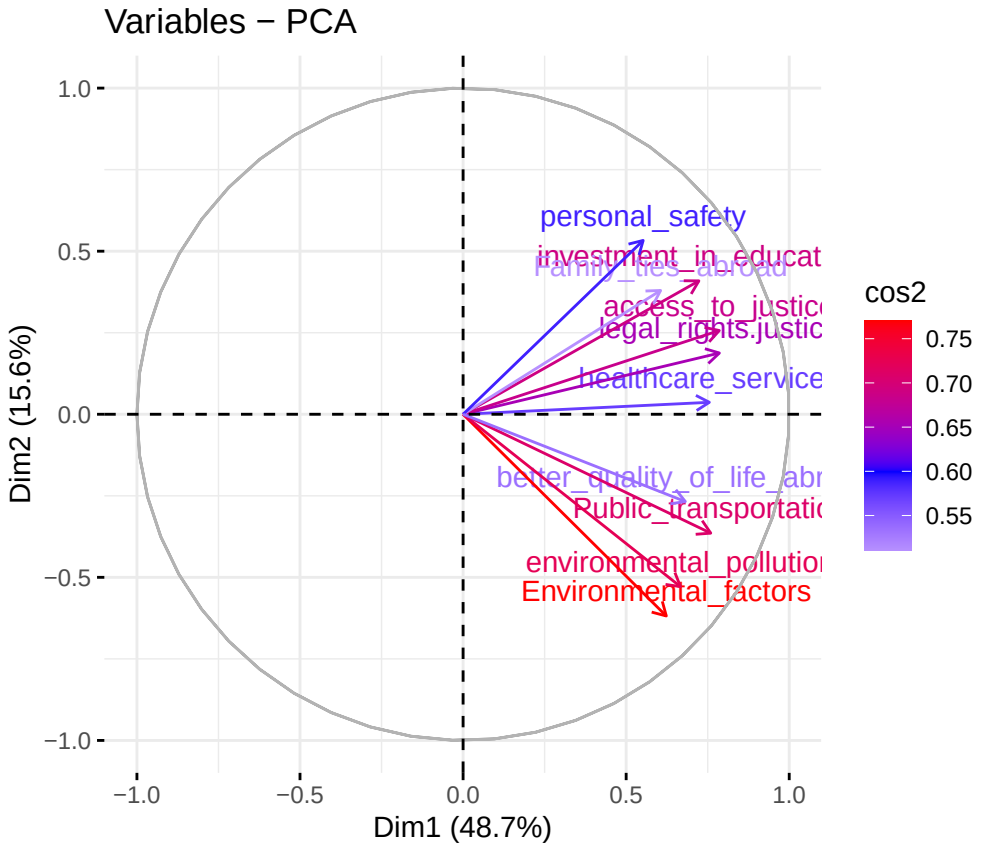


According to the Elbow criterion, we observe a bend (elbow) followed by a decrease. We select the axes before the elbow. Based on the scree plot presented above, the number of axes to retain is 3.

- Kaiser Criterion: We have 2 eigenvalues greater than 1.
- Elbow Criterion: The elbow is at the level of the third axis.
- Cumulative Inertia Criterion: The first two axes have a cumulative inertia rate of 64.36%. **Decision**: We retain the first two axes.

Visualize variables plot for axes 1 and 2:

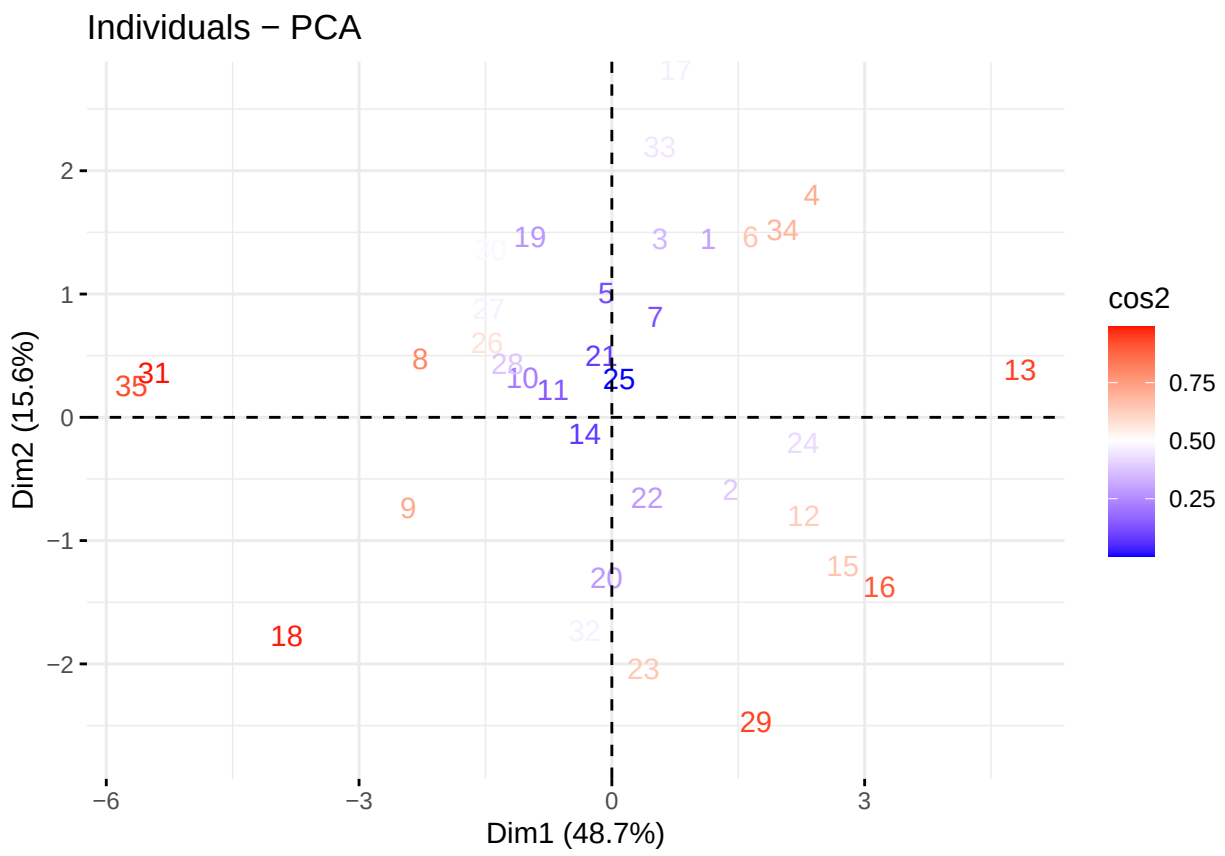
```
fviz_pca_var(res.pca2, axes=c(1,2), col.var="cos2") +
scale_color_gradient2(low="white", mid="blue",
                      high="red", midpoint=0.6)
```



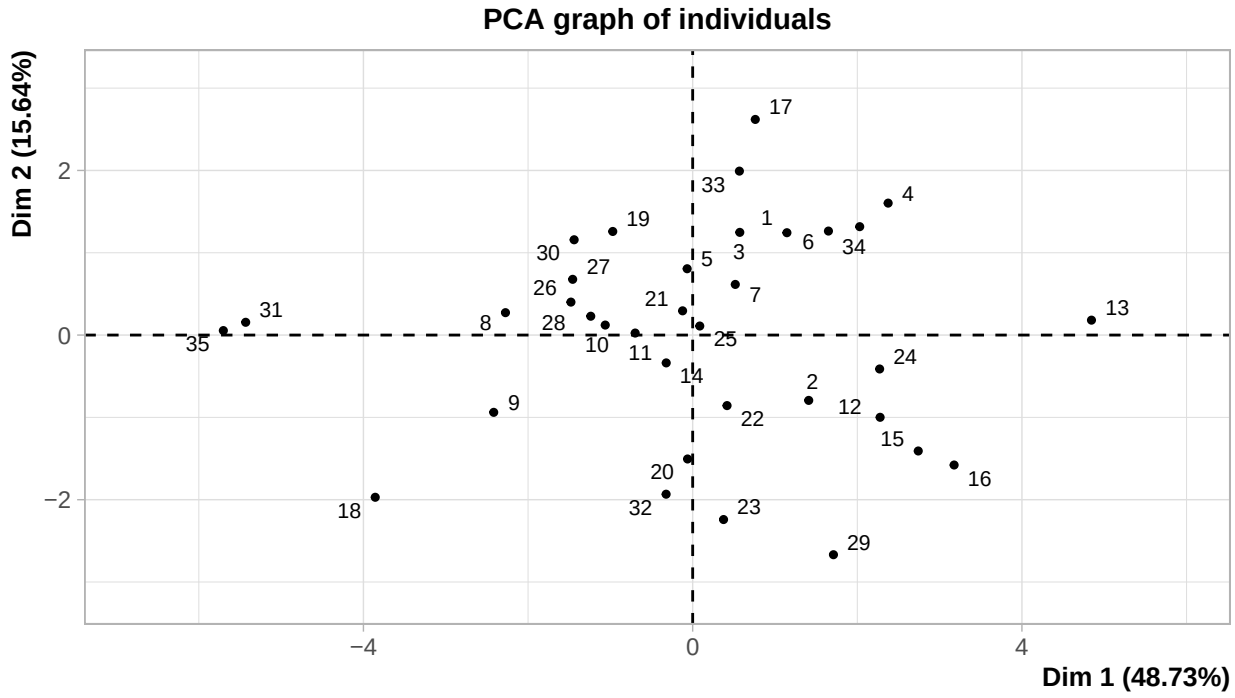
- The first two dimensions contain 64.36% of the total inertia.
- The further from the origin indicate better presentation like environmental factors.
- Variables : personal safety, investment in education, access to justice, legal rights and justice, and healthcare services are closely clustered, indicating a strong correlation among them.
- The same for variables : better quality of life ,public transportation ,environmental pollution and environmental factors.

PCA Plot of Individuals with Cos2 Values:

```
fviz_pca_ind(res.pca2,geom ="text",col.ind="cos2")+
scale_color_gradient2(low="blue", mid="white",
high="red", midpoint=0.5)
```




```
plot.PCA(res.pca2, axes=c(1, 2), choix="ind", cex=0.7)
```



Multiple Correspondence Analysis (MCA):

Performing Multiple Correspondence Analysis:

```
res.mca <- MCA (data[,3:8], graph = FALSE)
```

Calculating eigenvalues from MCA results:

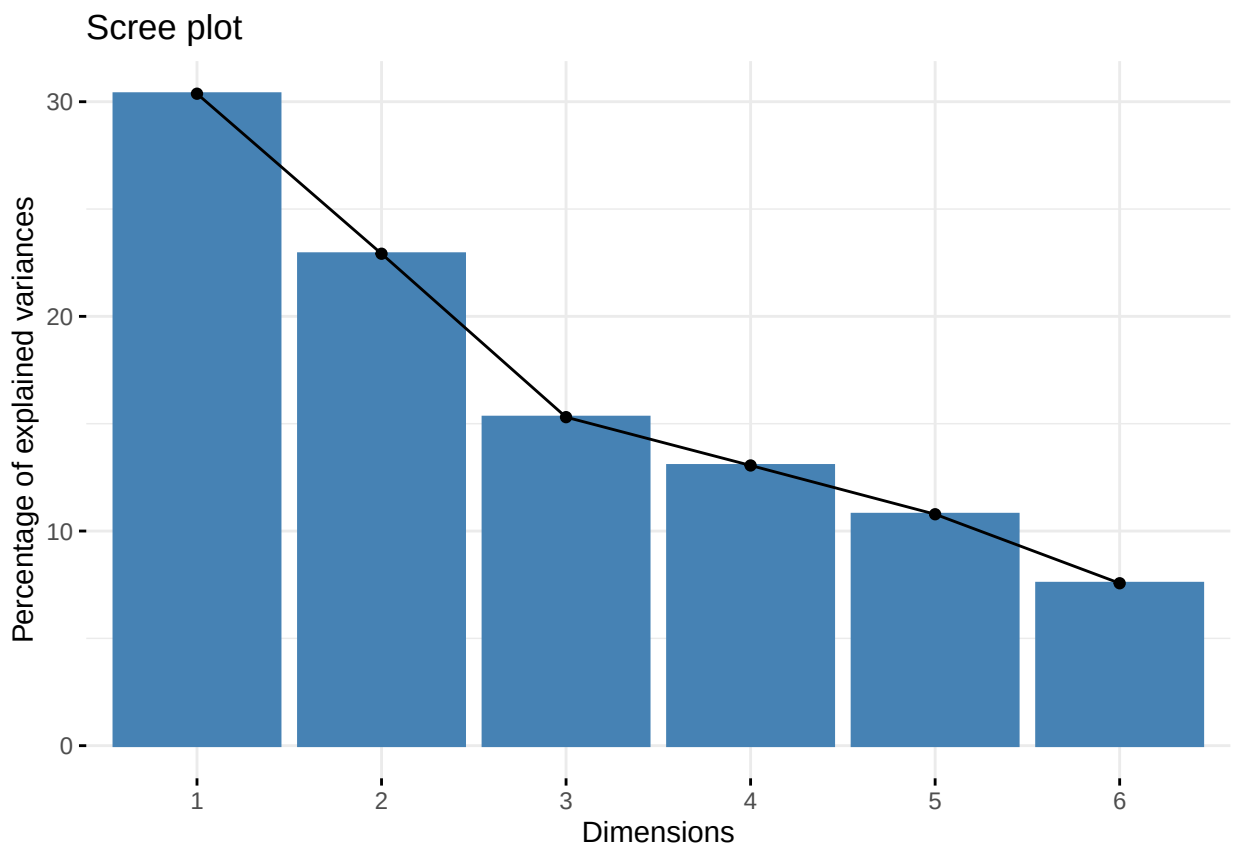
```
d<-data.frame(res.mca$eig)
DIM <- paste("Dim.", 1:6, sep = "")
d <- cbind(DIM,d)
ft <- flextable(d)
ft <- autofit(ft)
theme_zebra(ft)
```

DIM	eigenvalue	percentage.of.variance	cumulative.percentage.of.variance
Dim.1	0.30372968	30.372968	30.37297
Dim.2	0.22918314	22.918314	53.29128

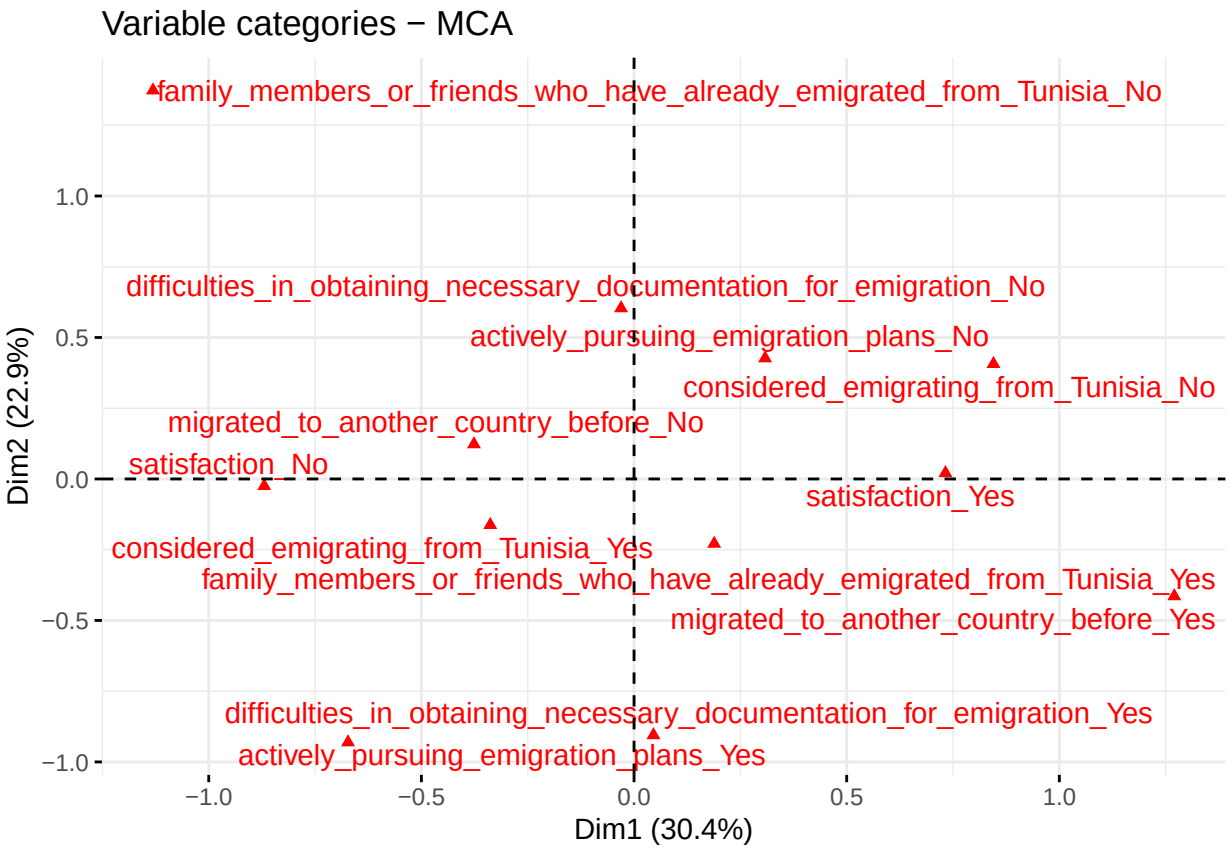
DIM	eigenvalue	percentage.of.variance	cumulative.percentage.of.variance
Dim.3	0.15304834	15.304834	68.59612
Dim.4	0.13054340	13.054340	81.65046
Dim.5	0.10782506	10.782506	92.43296
Dim.6	0.07567039	7.567039	100.00000

Scree Plot for Multiple Correspondence Analysis :

```
fviz_screeplot(res.mca)
```




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

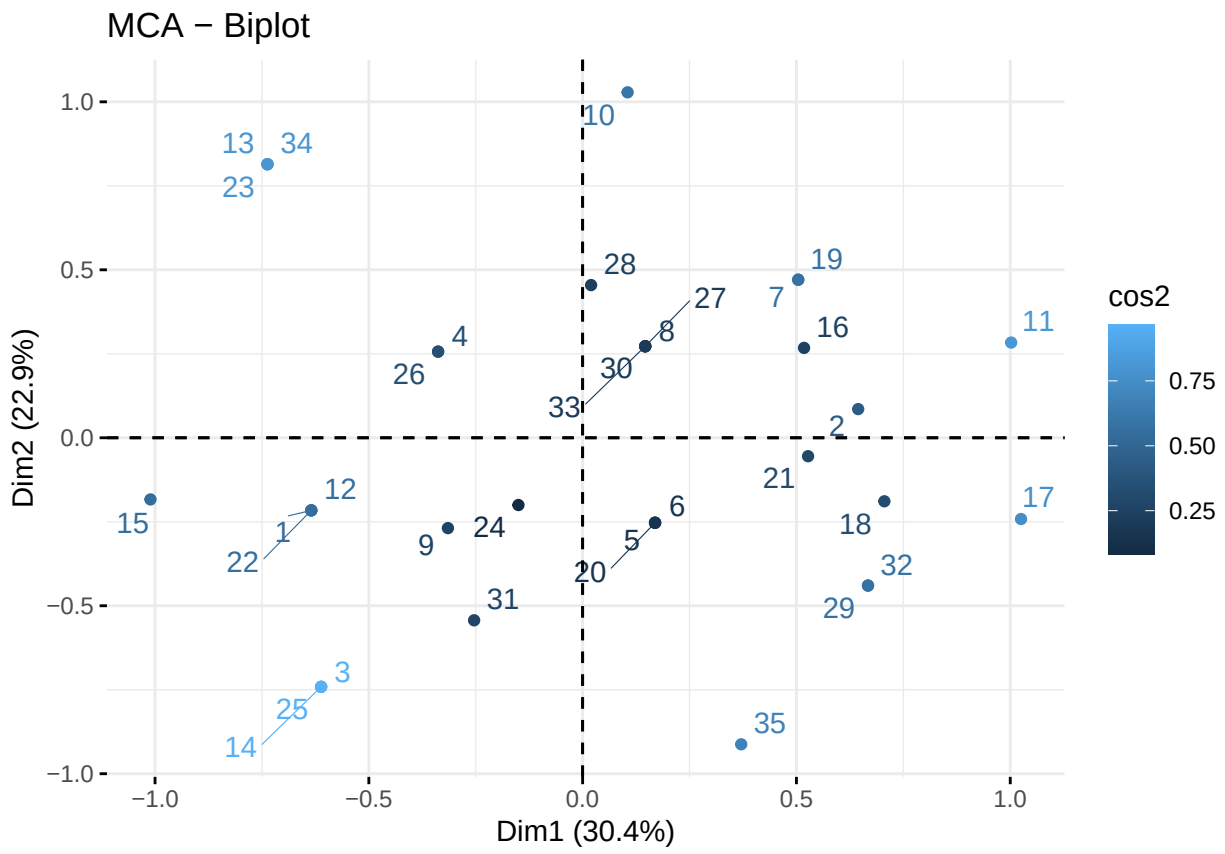


- It's just a matter of identifying the clusters of modalities that are close on the map.

- We have a first group of modalities on the left, consisting of modalities : migrated_to_another_country_before_no , satisfation_no , considered_emigration_from_tunisia_yes .
  -> whose interpretation is as follows: individuals who want to emigrate from tunisia

- Second groupe: difficulties_in_obtaining_necessary_documentation_for_emigration_no ,actively_pursuing_emigration_plans_no ,considered_emigration_from_tunisia_no ,satisfaction_yes .
  -> whose interpretation is as follows: individuals who doesn't want to emigrate from tunisia.

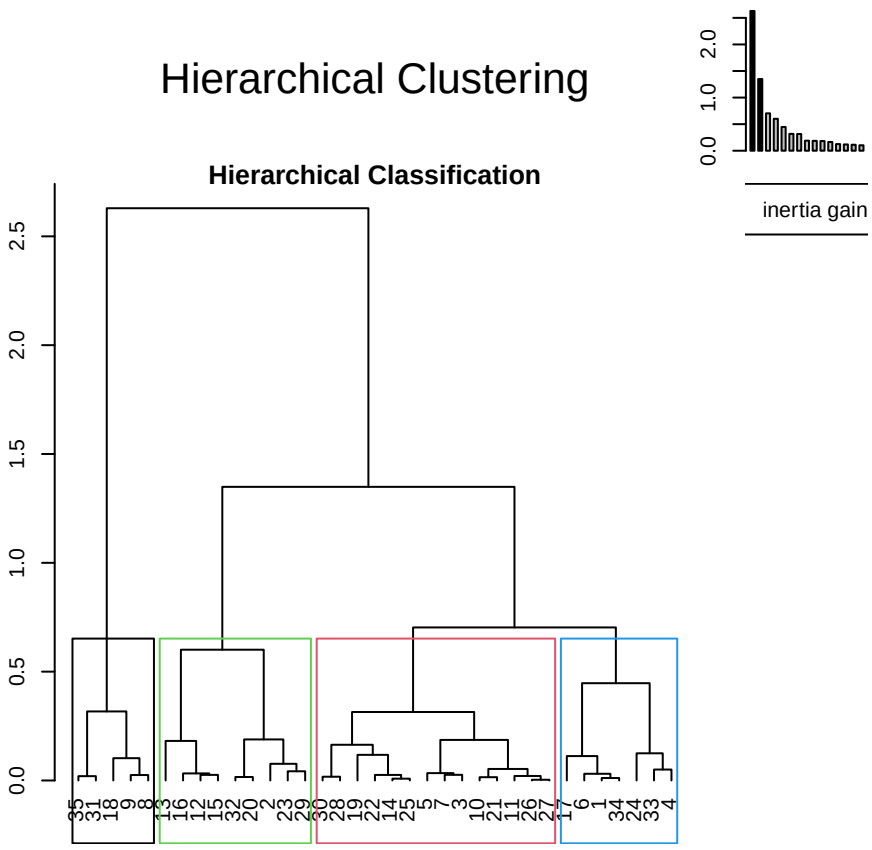
```
fviz_mca_biplot(res.mca,
col.ind="cos2",
invisible ="var",repel = TRUE,
ggtheme = theme_minimal ())
```



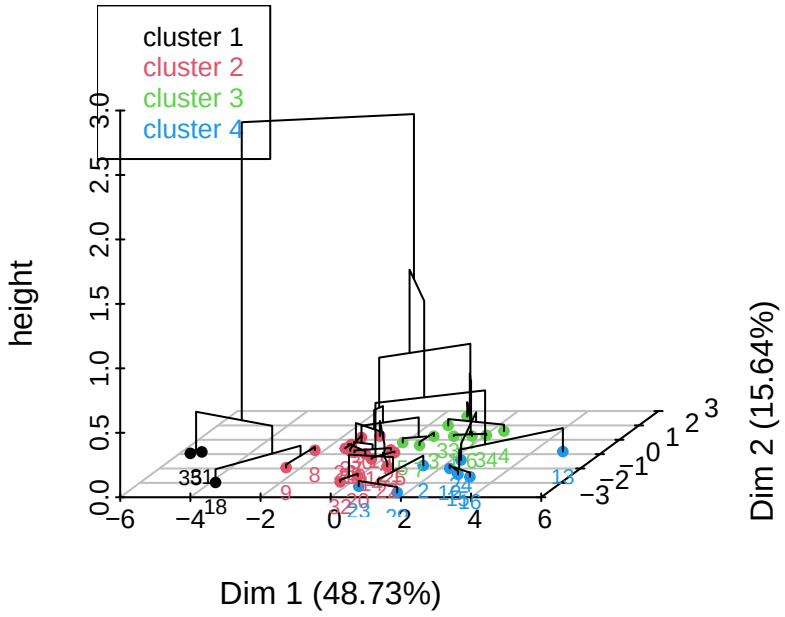
Classification:

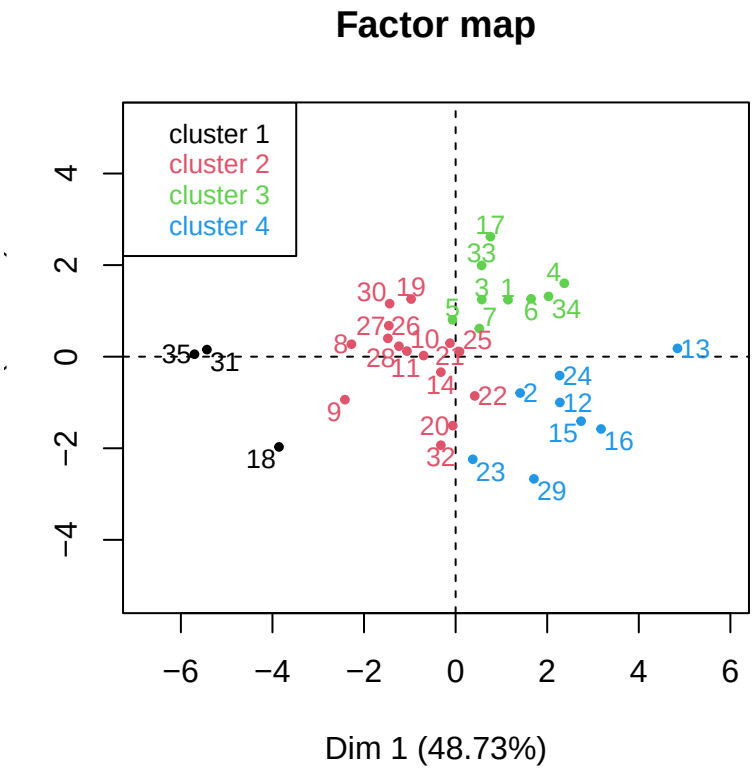
The idea of classifying various elements of a dataset into multiple categories is to group the data based on their similarity. Since the data share common characteristics, it is easier to predict their behavior. ## Performing classification:

```
res.pca <- PCA(X2, ncp = 4, graph = FALSE)
res.hcpc <- HCPC(res.pca,nb.clust = 4, graph = FALSE)
HCPC (res.pca, nb.clust = 4, graph = TRUE)
```


Hierarchical Clustering
Hierarchical Classification
0.0
0.5
1.0
1.5
2.0
2.5
35 31 18 9 8 13 16 12 15 32 20 2 23 29 30 28 19 22 14 25 5 7 3 10 21 11 26 27 17 6 1 34 24 33 4
0.0
1.0
2.0
inertia gain

Hierarchical clustering on the factor map






```
## **Results for the Hierarchical Clustering on Principal Components**
##    name
## 1  "$data.clust"
## 2  "$desc.var"
## 3  "$desc.var$quanti.var"
## 4  "$desc.var$quanti"
## 5  "$desc.axes"
## 6  "$desc.axes$quanti.var"
## 7  "$desc.axes$quanti"
## 8  "$desc.ind"
## 9  "$desc.ind$para"
## 10 "$desc.ind$dist"
## 11 "$call"
## 12 "$call$t"
##    description
## 1  "dataset with the cluster of the individuals"
## 2  "description of the clusters by the variables"
## 3  "description of the cluster var. by the continuous var."
## 4  "description of the clusters by the continuous var."
## 5  "description of the clusters by the dimensions"
## 6  "description of the cluster var. by the axes"
## 7  "description of the clusters by the axes"
## 8  "description of the clusters by the individuals"
## 9  "parangons of each clusters"
## 10 "specific individuals"
## 11 "summary statistics"
```

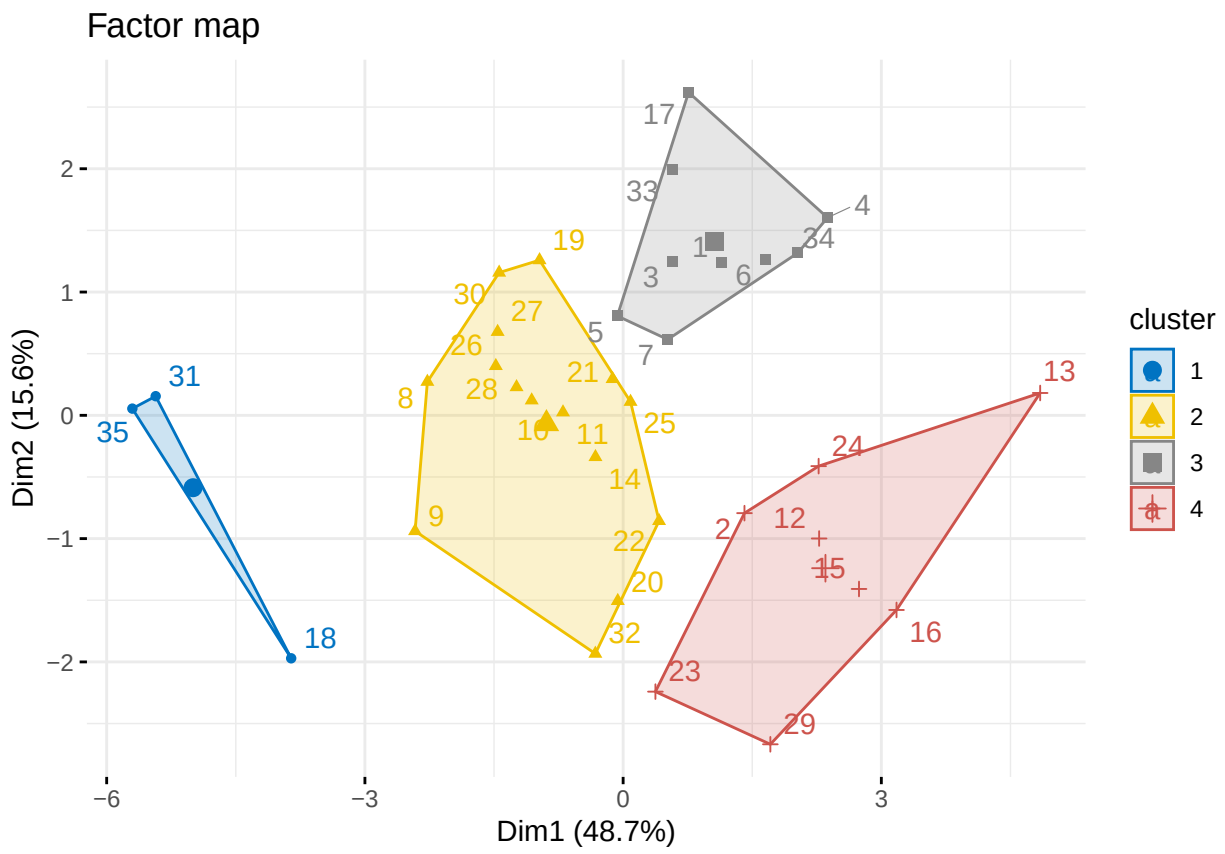
```
## 12 "description of the tree"
```

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

```
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                                    Eta2      P-value
## Public_transportation         0.6643051 1.689445e-07
## Environmental_factors         0.6172918 1.246095e-06
## investment_in_education       0.6111111 1.590070e-06
## access_to_justice             0.5233127 3.478969e-05
## legal_rights.justice          0.5210666 3.734742e-05
## environmental_pollution       0.5075383 5.684351e-05
## personal_safety               0.4637240 2.046772e-04
## better_quality_of_life_abroad 0.4620742 2.143213e-04
## healthcare_services           0.4396765 3.946356e-04
## Family_ties_abroad            0.3225417 6.551321e-03
##
## Description of each cluster by quantitative variables
## =====================================================
## $`1`
##                                  v.test Mean in category Overall mean
## environmental_pollution       -2.119533         1.666667     3.371429
## healthcare_services           -2.159275         1.666667     3.685714
## Family_ties_abroad            -2.850400         1.000000     3.314286
## Public_transportation         -2.917955         1.666667     3.942857
## personal_safety               -2.988475         1.333333     3.571429
## access_to_justice             -3.157421         1.000000     3.514286
## legal_rights.justice          -3.352191         1.000000     3.714286
## investment_in_education       -3.365396         1.000000     3.971429
## better_quality_of_life_abroad -3.440447         2.000000     4.057143
##                               sd in category Overall sd      p.value
## environmental_pollution            0.9428090   1.435981 0.0340454409
## healthcare_services                0.4714045   1.669413 0.0308288431
## Family_ties_abroad                 0.0000000   1.449560 0.0043664247
## Public_transportation              0.9428090   1.392692 0.0035233569
## personal_safety                    0.4714045   1.337070 0.0028037331
## access_to_justice                  0.0000000   1.421698 0.0015917154
## legal_rights.justice               0.0000000   1.445613 0.0008017477
## investment_in_education            0.0000000   1.576356 0.0007643401
## better_quality_of_life_abroad      0.8164966   1.067517 0.0005807540
##
## $`2`
##                         v.test Mean in category Overall mean sd in category
## healthcare_services -2.275853         2.933333     3.685714       1.339983
##                     Overall sd    p.value
## healthcare_services   1.669413 0.02285482
##
## $`3`
##                            v.test Mean in category Overall mean sd in category
## investment_in_education  3.205732         5.444444     3.971429      1.0657403
## personal_safety          3.095231         4.777778     3.571429      1.1331154
## access_to_justice        2.512639         4.555556     3.514286      1.0657403
```

```
## legal_rights.justice     1.996445         4.555556     3.714286      1.0657403
## Environmental_factors   -2.161122         2.333333     3.342857      0.8164966
##                          Overall sd     p.value
## investment_in_education    1.576356 0.001347195
## personal_safety            1.337070 0.001966597
## access_to_justice          1.421698 0.011983197
## legal_rights.justice       1.445613 0.045885471
## Environmental_factors      1.602549 0.030685914
##
## $`4`
##                                 v.test Mean in category Overall mean
## Environmental_factors         4.272383            5.500     3.342857
## Public_transportation         4.118508            5.750     3.942857
## environmental_pollution       3.875943            5.125     3.371429
## healthcare_services           2.974096            5.250     3.685714
## better_quality_of_life_abroad 2.060021            4.750     4.057143
## legal_rights.justice          1.999552            4.625     3.714286
##                               sd in category Overall sd      p.value
## Environmental_factors              0.8660254   1.602549 1.933952e-05
## Public_transportation              0.6614378   1.392692 3.813338e-05
## environmental_pollution            1.1659224   1.435981 1.062125e-04
## healthcare_services                0.9682458   1.669413 2.938534e-03
## better_quality_of_life_abroad      0.4330127   1.067517 3.939655e-02
## legal_rights.justice               1.4086785   1.445613 4.554862e-02
```

```
fviz_cluster(res.hcpc,
repel = TRUE,
show.clust.cent = TRUE,
palette = "jco",
ggtheme = theme_minimal(),
main = "Factor map"
)
```



Observed Classes:
- Cluster_1 [18-31-35]
- Cluster_2 [9-8-26-30-19-27-28-10-21-11-25-14-20-32-22-27]
- Cluster_3 [23-29-1-15-12-2-24-13]
- Cluster_4 [1-5-3-6-34-4-33-17]
- Each Cluster present a groupe of individuals.
- Each cluster seems to have distinct characteristics based on the quantitative variables.
- Cluster 1 has lower scores in categories like environmental pollution, healthcare services, and access to justice compared to other clusters.
- Cluster 2 has a significantly higher mean in healthcare services compared to the overall mean.
- Cluster 3 shows higher scores in investment in education and personal safety compared to other clusters.
- Cluster 4 exhibits higher scores in environmental factors and public transportation.
- “Public_transportation” seems to be strongly associated with Clusters 1 and 4, as indicated by their high Eta2 values and low P-values.
- Similarly, “Environmental_factors” appear to be linked with Clusters 1, 3, and 4.