Cluster analysis of Spain's provinces

Javier Eloy Martinez Ramos

2/28/2022

Introduction

The source data file contains socioeconomic information regarding the provinces of Spain (the names of variables have been changed). We are going to reduce and try to find relations both between variables and also provinces.

Here I list the meaning of the acronyms in the variables :

CPI = Consumer Price Index CTH = Commerce, Transport and Hostelry Infor = Information and Communications IFA = Insurance and Financial Activities PTA = Professional and Technical Activities GDP = Gross Domestic Product ACNH = Agrarian Census Number of Holdings TFH = Total Family Homes (2011 Census)

```
distritos <- read_excel("Provincias.xlsx")
datos <- as.data.frame(distritos)
rownames(datos) <- datos[,1]
datos_n <- datos[, -c(1, 19)]
c()
```

```
## NULL
```

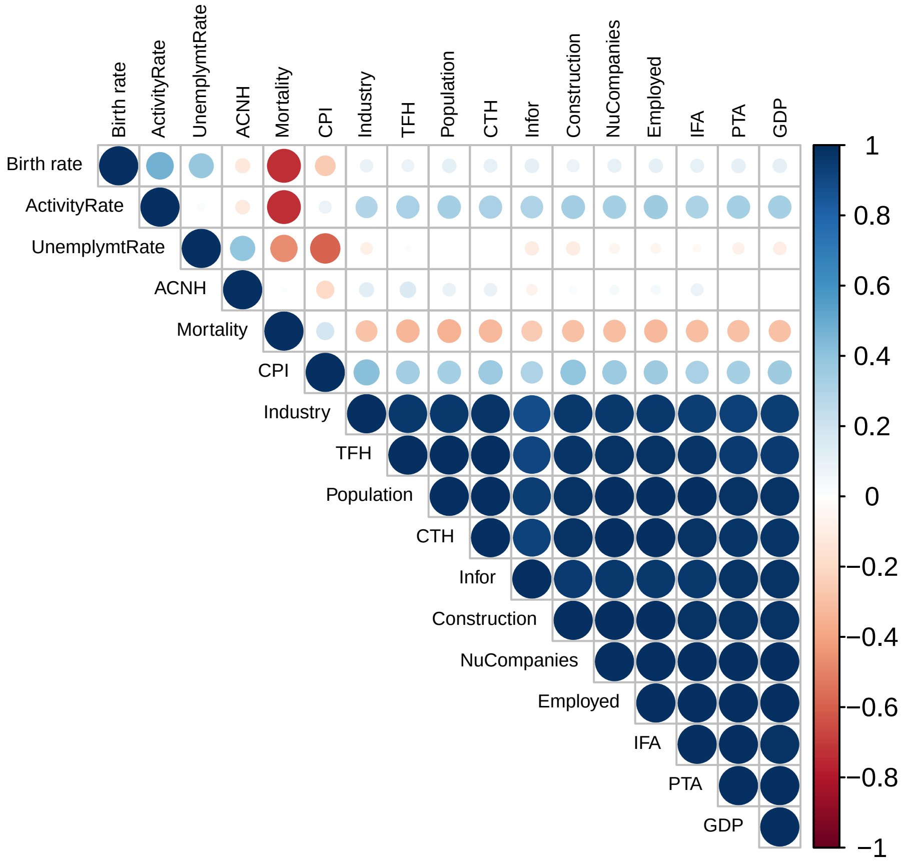
```
# Variable names in spanish:
# "Poblacion", "Mortalidad", "Natalidad", "IPC", "NumEmpresas", "Industria",
# "Construccion", "CTH", "Infor", "AFS", "APT", "TasaActividad", "TasaParo",
# "Ocupados", "PIB", "CANE", "TVF"

colnames(datos_n) <- c(
  "Population", "Mortality", "Birth rate", "CPI",
  "NuCompanies", "Industry", "Construction", "CTH",
  "Infor", "IFA", "PTA", "ActivityRate", "UnemplymtRate",
  "Employed", "GDP", "ACNH", "TFH"
  )
```

1. Correlation matrix calculation and representation

Which are the variables that are more inversely correlated?

```
R <- cor(datos_n, method = "pearson")
corrplot(R, type = "upper", order = "hclust", tl.col= "black", tl.cex = 0.6, tl.srt = 90)
```



Analysis:

According to the graphic, the most inversely correlated variables are:

1. Birth Rate / Mortality (strong inverse correlation)
2. Mortality / Activity Rate (strong inverse correlation)
3. Unemployment Rate / CPI (Consumer Price Index) (considerable inverse correlation)
4. Unemployment Rate / Mortality (medium inverse correlation)

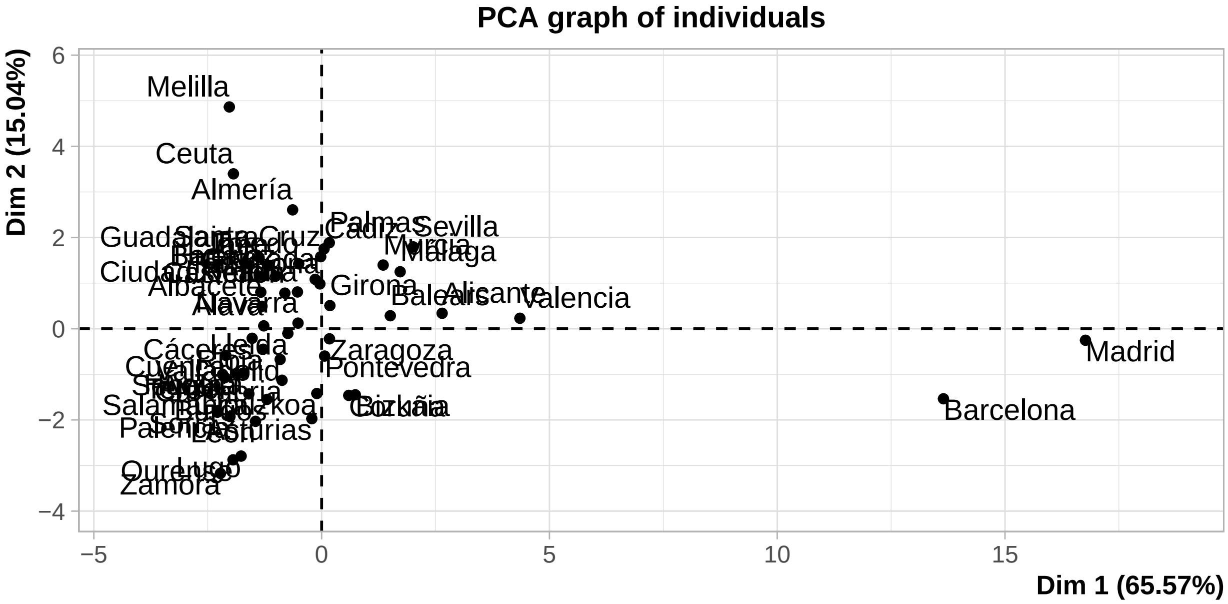
We can also see that the next variables are very strongly correlated:

1. Industry
2. TFH (Total Family Homes)
3. Population
4. CTH (Commerce, Transport and Hostelry)
5. Infor (Information and Communications)
6. Construction
7. New Companies
8. Employed
9. IFA (Insurance and Financial Activities)
10. PTA (Professional and Technical Activities)
11. GDP (Gross Domestic Product)

2. Main component analysis

Which is the right number of components to take into account in order to group the subjects?

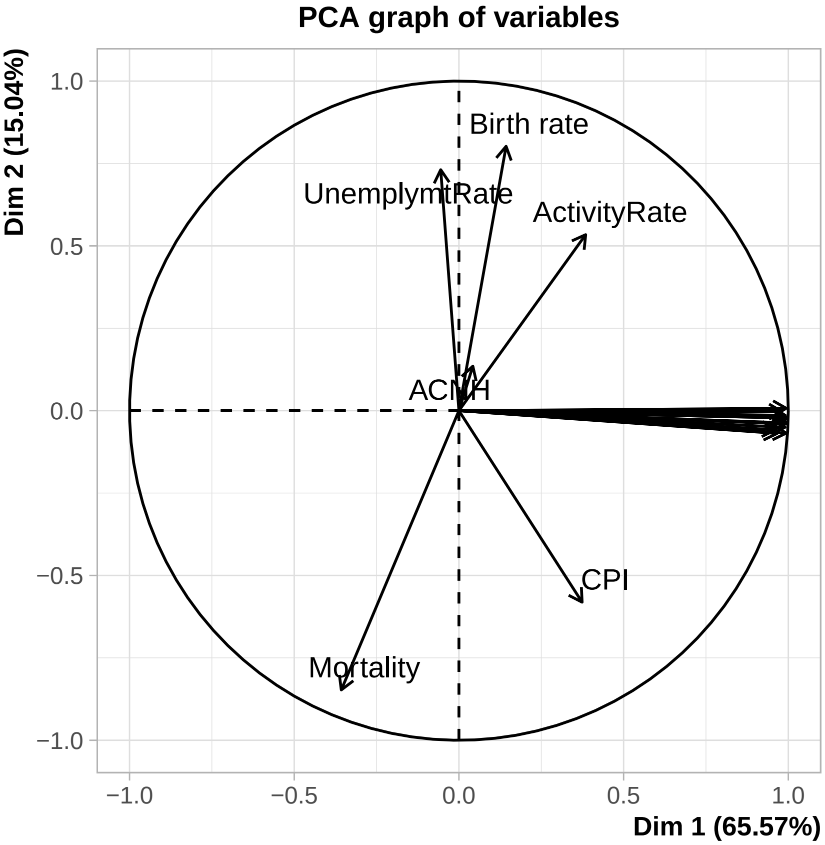
```
fit <- PCA(datos_n, scale.unit = TRUE, ncp = 7, graph = TRUE)
```



```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

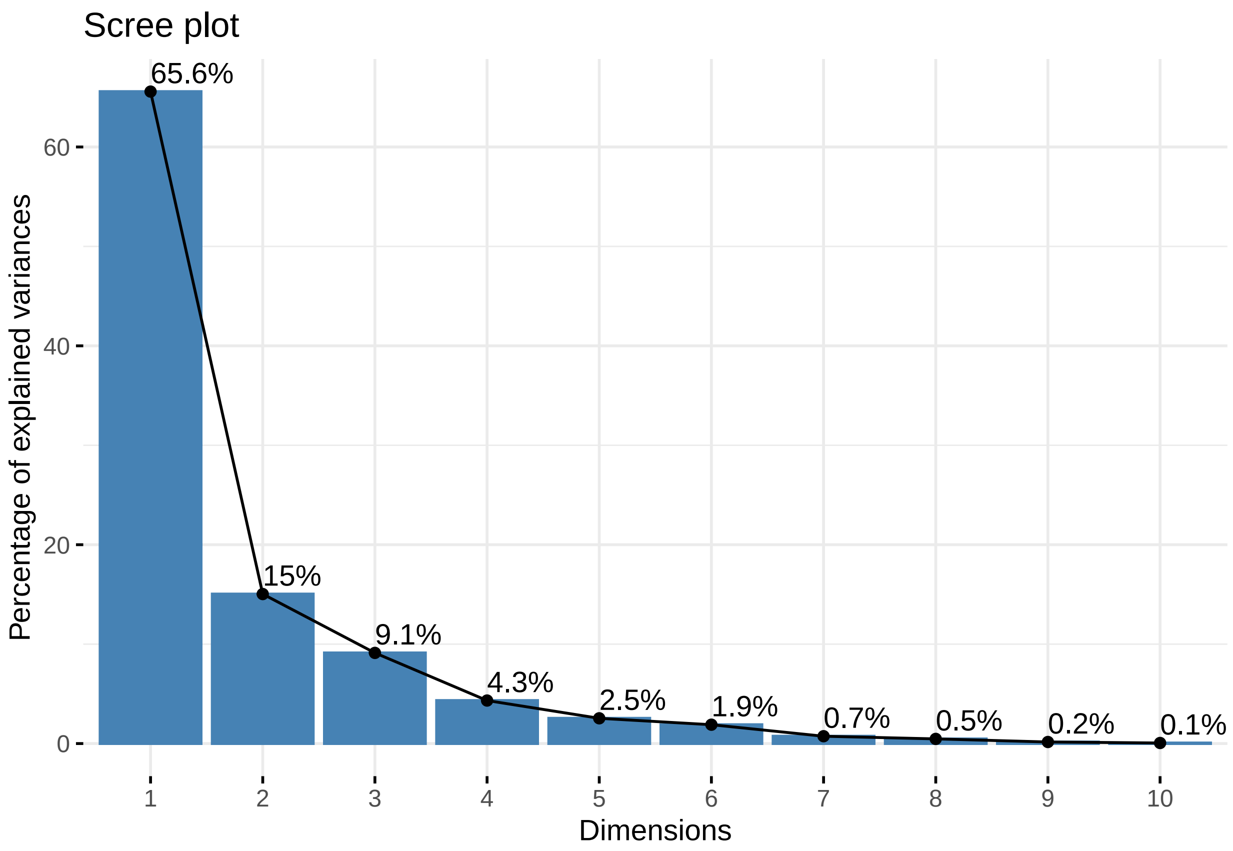
Table 1: Autovalues

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	11.15	65.57	65.57
Dim.2	2.56	15.04	80.60
Dim.3	1.55	9.12	89.72
Dim.4	0.74	4.33	94.05
Dim.5	0.43	2.54	96.59
Dim.6	0.32	1.90	98.49
Dim.7	0.13	0.74	99.23
Dim.8	0.08	0.47	99.70
Dim.9	0.03	0.16	99.86
Dim.10	0.01	0.06	99.92
Dim.11	0.01	0.04	99.96
Dim.12	0.00	0.02	99.98
Dim.13	0.00	0.01	99.99
Dim.14	0.00	0.00	100.00
Dim.15	0.00	0.00	100.00
Dim.16	0.00	0.00	100.00
Dim.17	0.00	0.00	100.00



```
eig <- get_eigenvalue(fit)
knitr::kable(eig, digits = 2, caption = "Autovalues")
```

```
fviz_eig(fit, addlabels = TRUE)
```

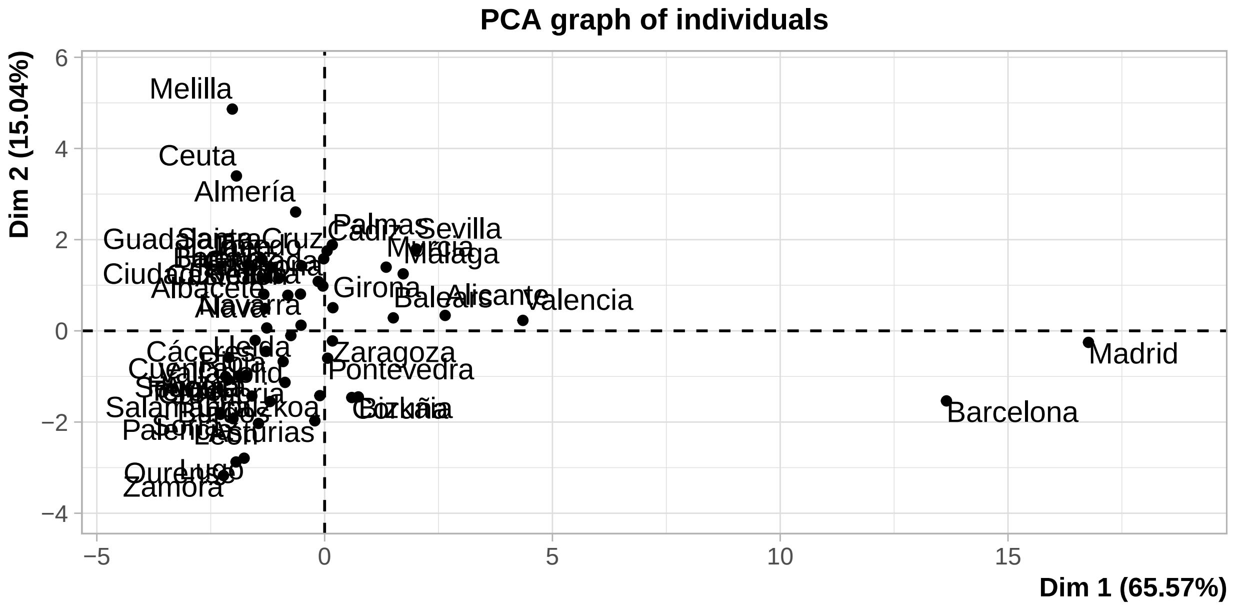


Analysis:

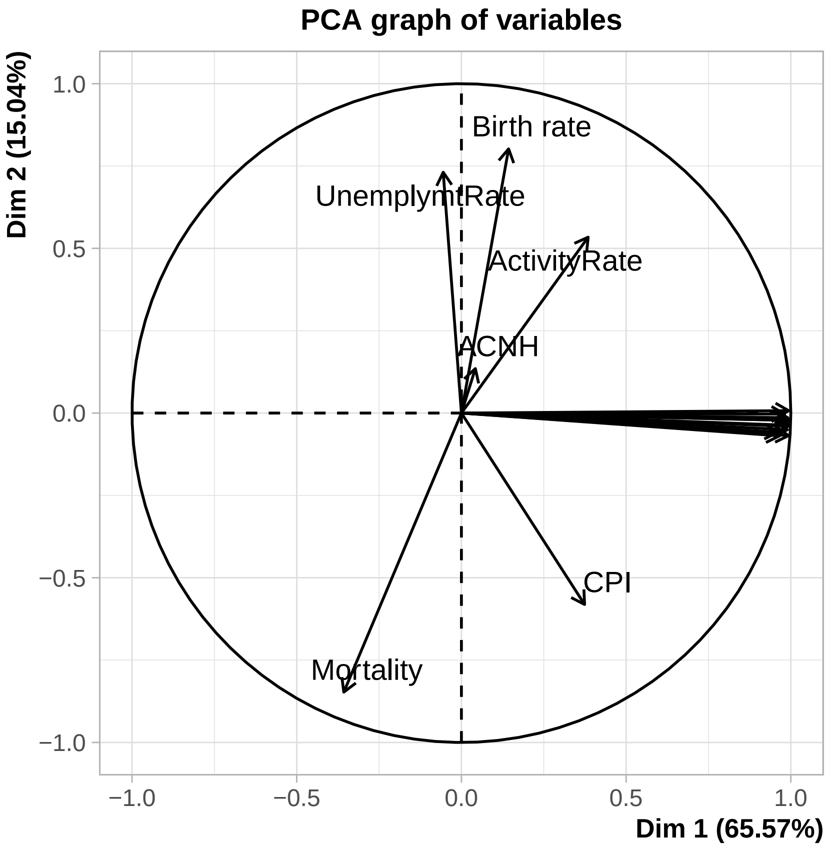
The charts show that the first 3 components explain the 89.7% of the relations between the subjects in the database.

3. New correlation matrix analysis (with the aforementioned 3 components)

```
fit3 <- PCA(datos_n, scale.unit = TRUE, ncp = 3, graph = TRUE)
```



```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



3.a. Main components coefficients

Which is the expression to calculate the first component according to the original variables?

```
kable_styling(kable(fit3$svd$V, digits = 3, caption = "Autovectors"))
```

Expression:

$$CP_1 = 0.298 Population^* - (-0.107 Mortality^*) + ... - 0.013 ACNH^* - 0.294 TFH^*$$

3.b. Variable correlation with the main components

We are going to indicate which variables are correlated with each component.

```
var <- get_pca_var(fit3)
kable_styling(kable(var$cor, digits = 2, caption = "Correlaciones de la CP con las variables"))
```

```
var$cor
```

Table 2: Autovectors

0.298	0.005	0.070
-0.107	-0.530	0.196
0.043	0.502	-0.244
0.112	-0.363	-0.319
0.299	-0.023	0.029
0.289	-0.043	0.056
0.298	-0.043	0.001
0.297	-0.009	0.064
0.288	-0.036	-0.023
0.297	-0.014	0.065
0.296	-0.025	0.004
0.115	0.334	-0.412
-0.017	0.457	0.431
0.299	-0.014	0.023
0.296	-0.032	-0.006
0.013	0.084	0.646
0.294	-0.002	0.100

Table 3: Correlaciones de la CP con las variables

	Dim.1	Dim.2	Dim.3
Population	0.99	0.01	0.09
Mortality	-0.36	-0.85	0.24
Birth rate	0.14	0.80	-0.30
CPI	0.37	-0.58	-0.40
NuCompanies	1.00	-0.04	0.04
Industry	0.97	-0.07	0.07
Construction	0.99	-0.07	0.00
CTH	0.99	-0.01	0.08
Infor	0.96	-0.06	-0.03
IFA	0.99	-0.02	0.08
PTA	0.99	-0.04	0.00
ActivityRate	0.38	0.53	-0.51
UnemplymtRate	-0.06	0.73	0.54
Employed	1.00	-0.02	0.03
GDP	0.99	-0.05	-0.01
ACNH	0.04	0.14	0.80
TFH	0.98	0.00	0.12

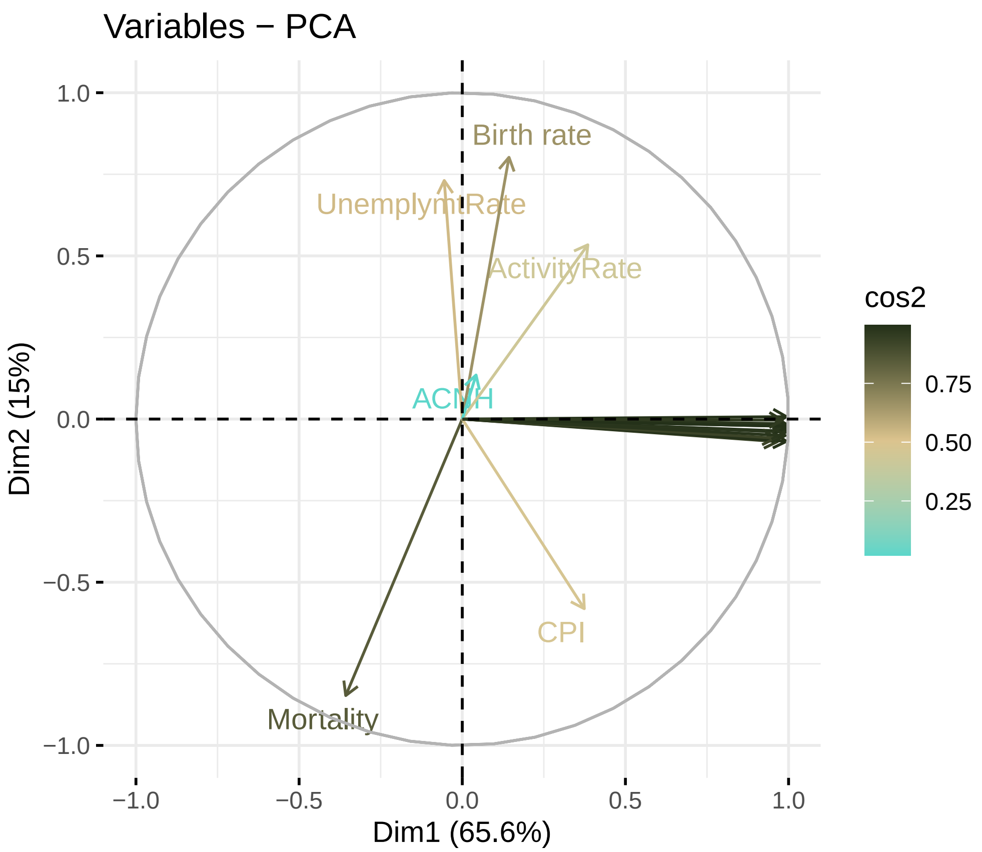
```
##                     Dim.1        Dim.2        Dim.3
## Population     0.99359491  0.007324455  0.087088297
## Mortality     -0.35711993 -0.847297481  0.243780355
## Birth rate     0.14321070  0.802086986 -0.303755469
## CPI            0.37396112 -0.581066874 -0.397165734
## NuCompanies    0.99740970 -0.037115956  0.036192031
## Industry       0.96600889 -0.069140364  0.069883580
## Construction   0.99342685 -0.068268421  0.001186581
## CTH            0.99068248 -0.014126251  0.080064005
## Infor          0.96041168 -0.058132239 -0.028161883
## IFA            0.99134716 -0.022089680  0.080379594
## PTA            0.98868621 -0.040140193  0.004571786
## ActivityRate   0.38479374  0.533993313 -0.512347931
## UnemplymtRate -0.05525336  0.731055281  0.536464849
## Employed       0.99830352 -0.022110425  0.028761124
## GDP            0.98988072 -0.050522418 -0.006987248
## ACNH           0.04229856  0.135016690  0.804095948
## TFH            0.98132311 -0.003165646  0.124602597
```

3.c. Component explanation

We are going to generate some graphs in order to explain what each component represents.

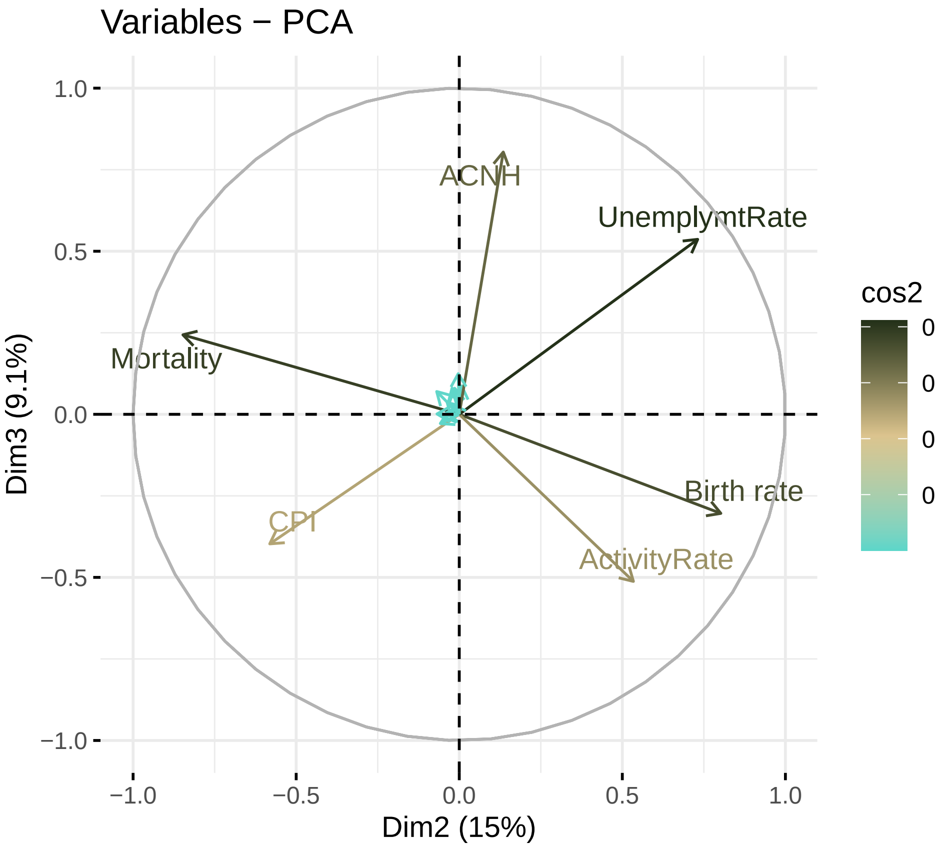
```
fviz_pca_var(
  fit3, axes = c(1, 2), col.var = "cos2",
  gradient.cols = c("#5CD6CA", "#DCC48E", "#243119"), repel = TRUE
  )
```

```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```r
fviz_pca_var(
  fit3, axes = c(2, 3), col.var = "cos2",
  gradient.cols = c("#5CD6CA", "#DCC48E", "#243119"), repel = TRUE
  )
```

```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Analysis:

Component 1

Component 1 has a positive and high correlation with population rate, the number of companies, the number of industrial companies, construction, commerce, transport and hostelry, information and commmunications, insurance and financial activities, professional and technical activities, the number of employed people, GDP and total family homes. All of these variables are represented in the first component.

Component 2

In the case of component 2, we can see a positive and high correlation with birth rate, unemployment rate and activity rate and a low correlation with the number of agrarian holdings, and a high negative correlation with mortality rate and a medium negative correlation with the consumer price index.

Component 3

In component 3 we have a high positive correlation with the number of agrarian holdings.

3.d. Proportion of variable variance explained by each component

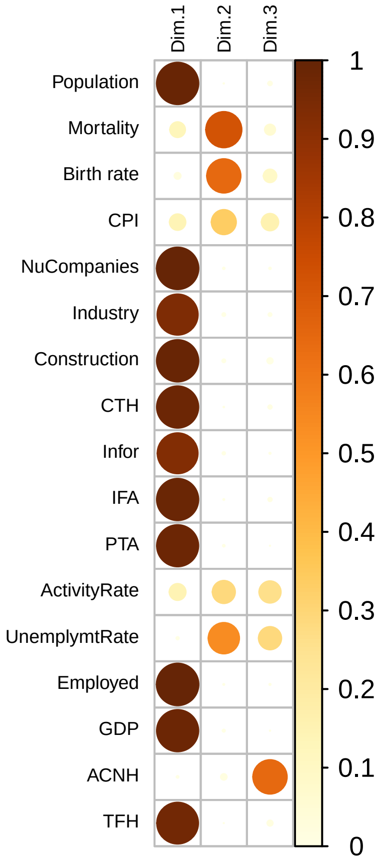
Which variable is more poorly explained?

```
kable_styling(kable(var$cos2, digits = 2, caption = "Cosenos al cuadrado"))
```

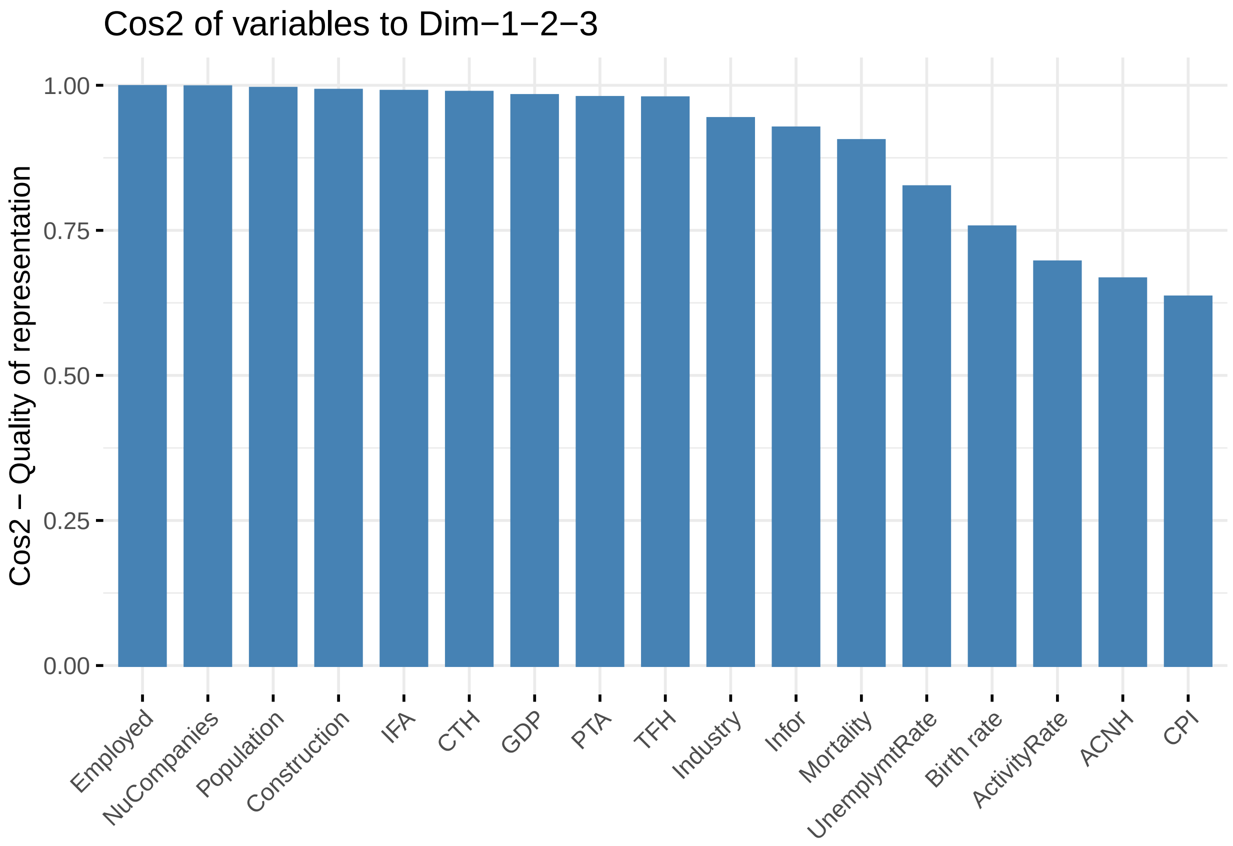
Table 4: Cosenos al cuadrado

	Dim.1	Dim.2	Dim.3
Population	0.99	0.00	0.01
Mortality	0.13	0.72	0.06
Birth rate	0.02	0.64	0.09
CPI	0.14	0.34	0.16
NuCompanies	0.99	0.00	0.00
Industry	0.93	0.00	0.00
Construction	0.99	0.00	0.00
CTH	0.98	0.00	0.01
Infor	0.92	0.00	0.00
IFA	0.98	0.00	0.01
PTA	0.98	0.00	0.00
ActivityRate	0.15	0.29	0.26
UnemplymtRate	0.00	0.53	0.29
Employed	1.00	0.00	0.00
GDP	0.98	0.00	0.00
ACNH	0.00	0.02	0.65
TFH	0.96	0.00	0.02

```
corrplot(var$cos2, is.corr = FALSE, tl.cex = 0.6, tl.col = "black", cl.ratio = 1)
```



```
fviz_cos2(fit3, choice = "var", axes = 1:3, tl.cex = 0.6)
```



According to the graph, the worst explained variable is CPI (Consumer Price Index)

3.e. Variance percentage of each component due to each variable.

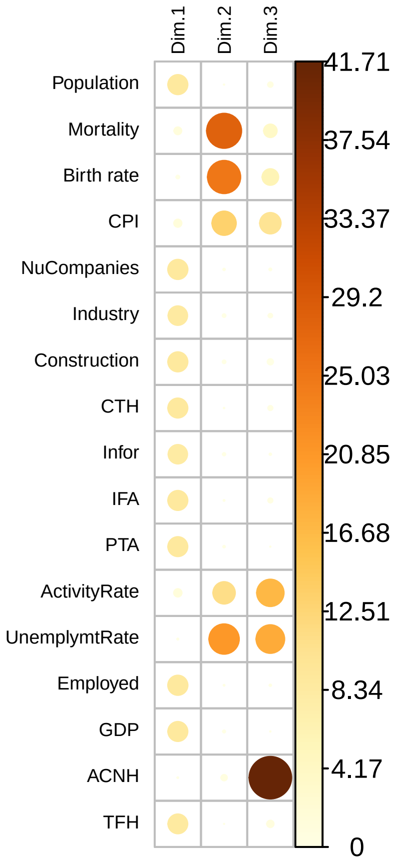
Which variables contribute the most to each component?

```
kable_styling(kable(var$contrib, digits = 2, caption = "Contribuciones"))
```

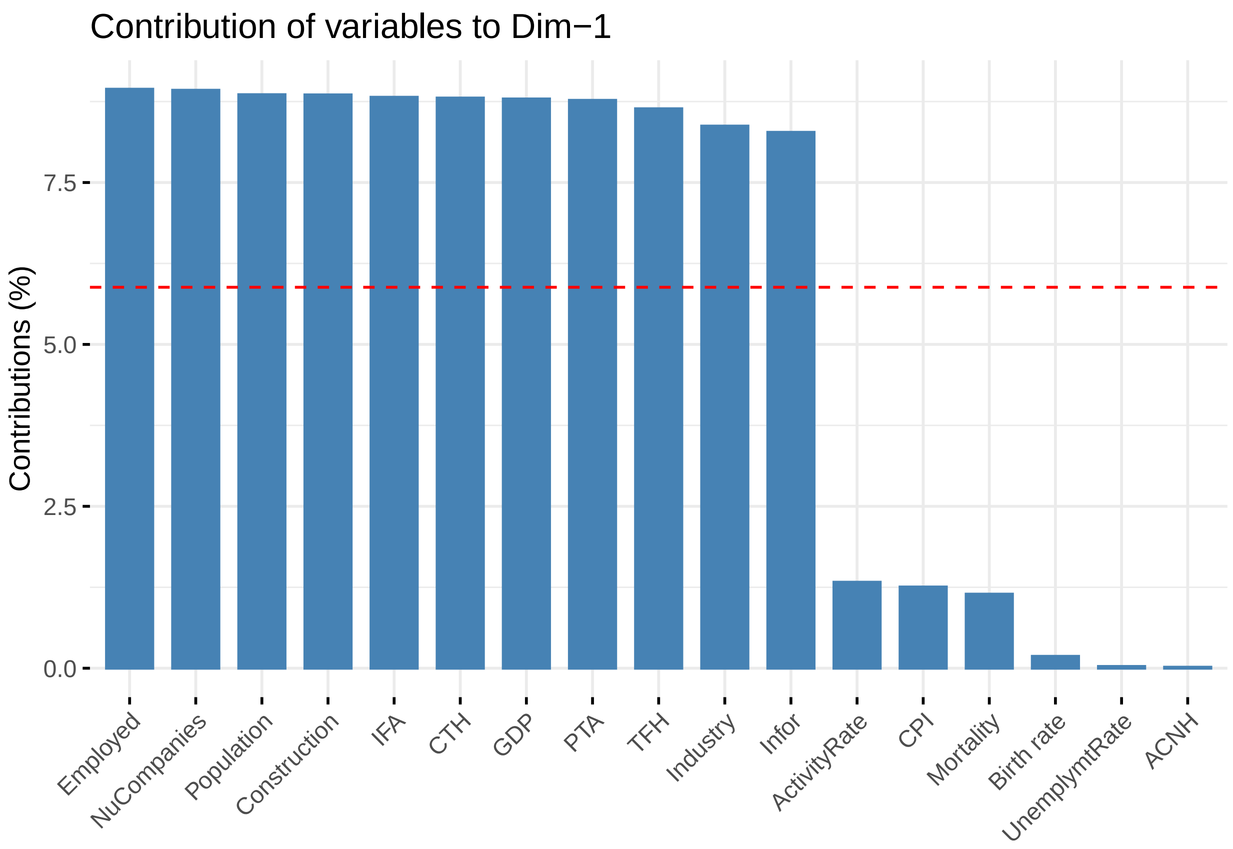
```
corrplot(var$contrib, is.corr = FALSE, tl.cex = 0.6, tl.col = "black", cl.ratio = 1)
```

Table 5: Contribuciones

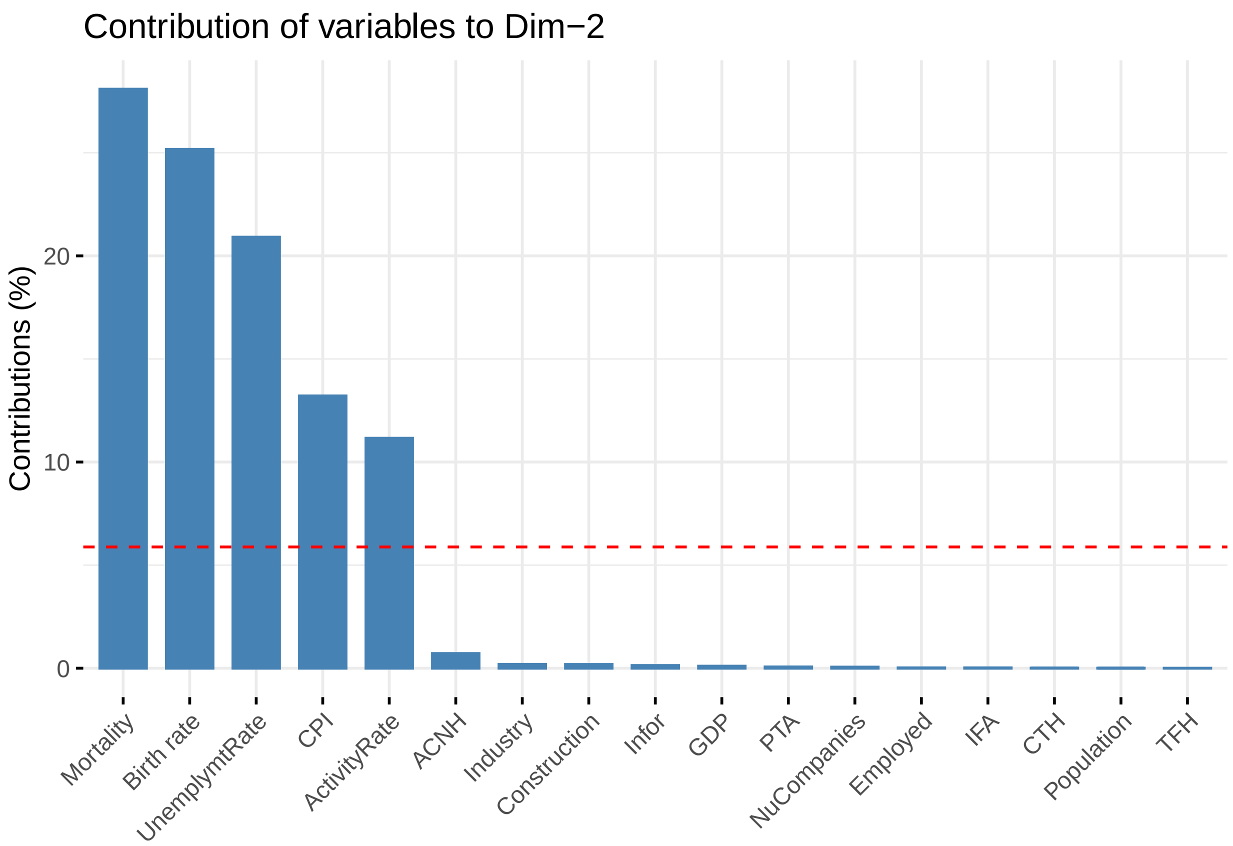
	Dim.1	Dim.2	Dim.3
Population	8.86	0.00	0.49
Mortality	1.14	28.08	3.83
Birth rate	0.18	25.17	5.95
CPI	1.25	13.21	10.18
NuCompanies	8.93	0.05	0.08
Industry	8.37	0.19	0.32
Construction	8.85	0.18	0.00
CTH	8.81	0.01	0.41
Infor	8.28	0.13	0.05
IFA	8.82	0.02	0.42
PTA	8.77	0.06	0.00
ActivityRate	1.33	11.15	16.93
UnemplymtRate	0.03	20.91	18.57
Employed	8.94	0.02	0.05
GDP	8.79	0.10	0.00
ACNH	0.02	0.71	41.71
TFH	8.64	0.00	1.00



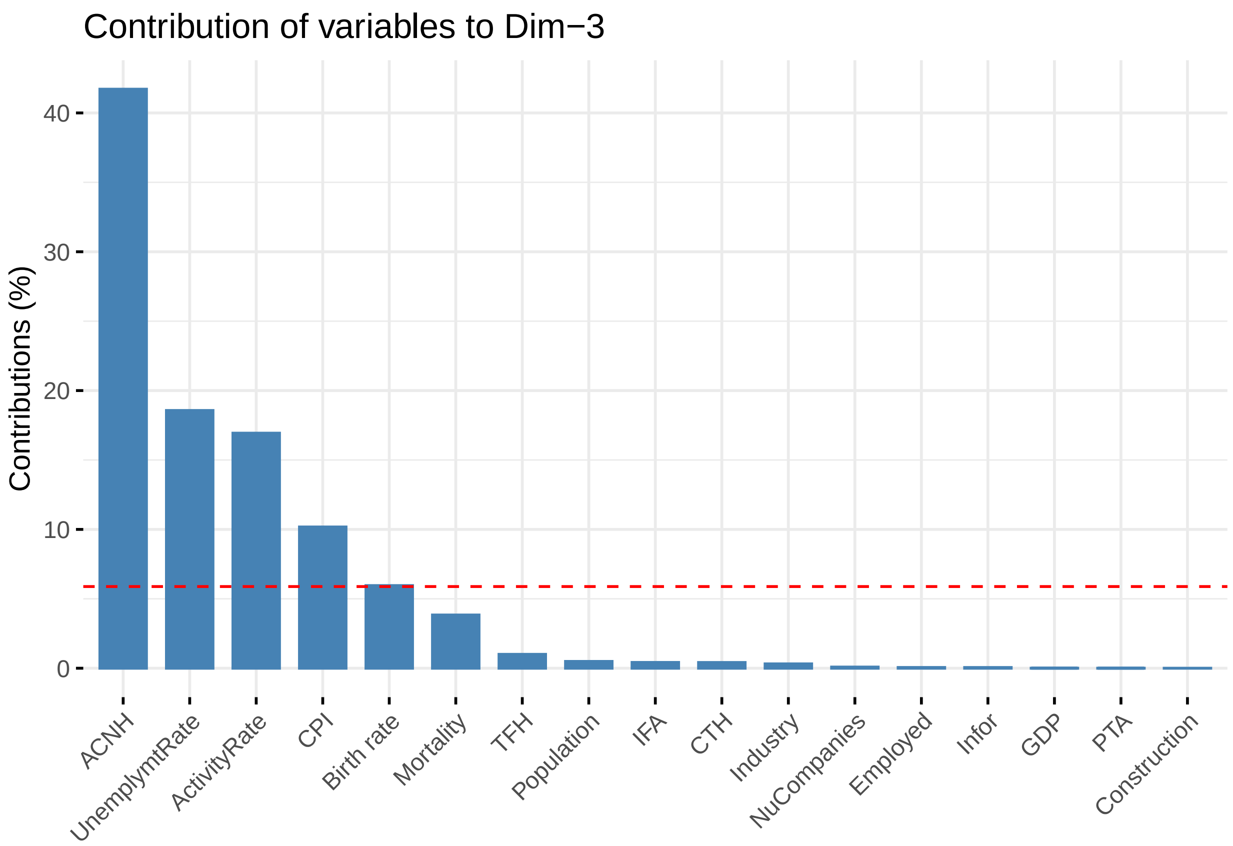
```
fviz_contrib(fit3, choice = "var", axes = 1, tl.cex = 0.6)
```



```
fviz_contrib(fit3, choice = "var", axes = 2, tl.cex = 0.6)
```



```
fviz_contrib(fit3, choice = "var", axes = 3, tl.cex = 0.6)
```



Analysis:

We can see that the variables contributing the most to component 1 are Employed (number of employed people), NumCompanies (number of companies), Population, Construction (Construction companies), IFA (Insurance and Financial Activities), CTH (Commerce, Transport and Hostelry), GDP (Gross Domestic Product), PTA (Professional and Technical Activities), TFH (Total Family Homes), Industry and Infor (Information and Communications).

The variables contributing the most to component 2 are Mortality, Birth rate, Unemployment rate, CPI (Consumer Price Index) and Activity rate.

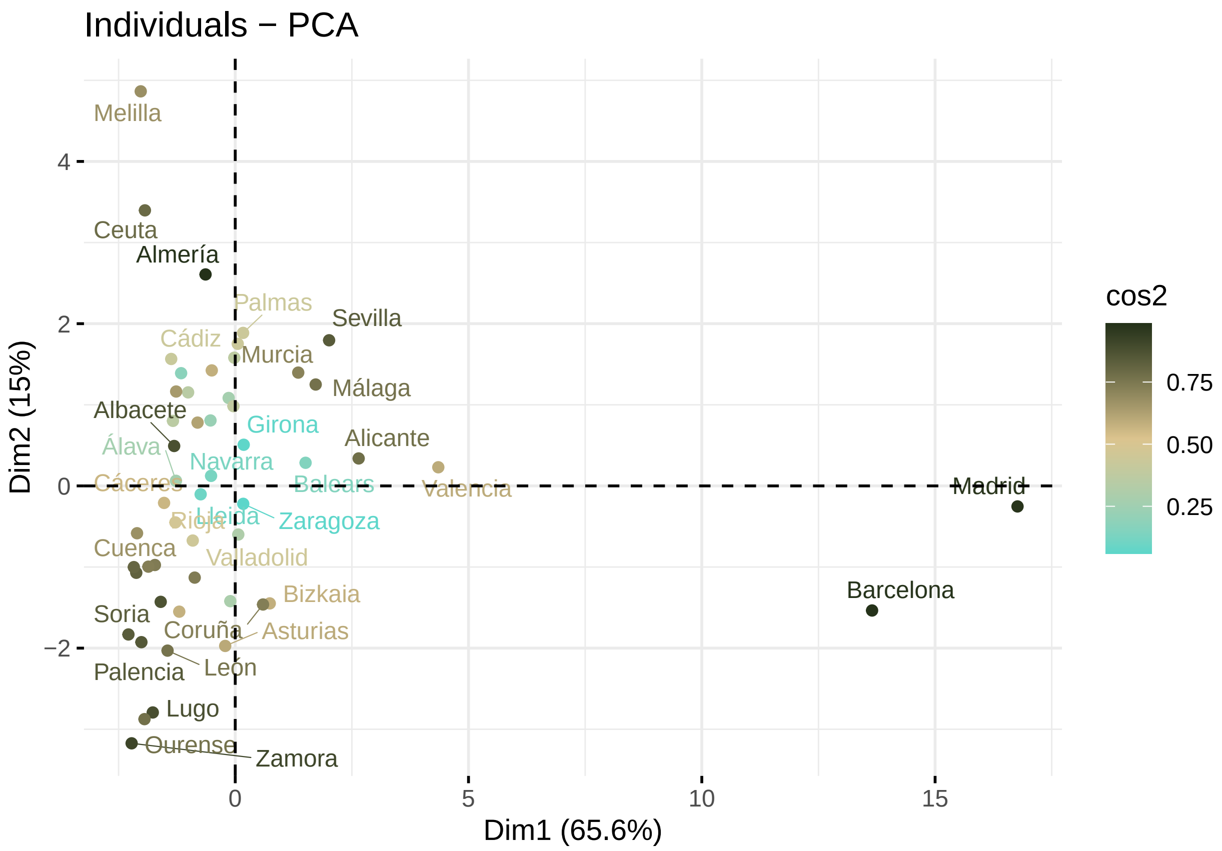
The variables contributing the most to component 3 are ACNH (Agrarian Census Number of Holdings), Unemployment rate, Activity rate, CPI (Consumer Price Index) and Birth rate.

**3.f. New axis representation graphs

What does the position in each graph represents for these provinces?

```
fviz_pca_ind(
  fit3, axes = c(1, 2), col.ind = "cos2", col.cex = 0.2,
  gradient.cols = c("#5CD6CA", "#DCC48E", "#243119"), repel = TRUE,
  labelsize = 3
  )
```

```
## Warning: ggrepel: 20 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



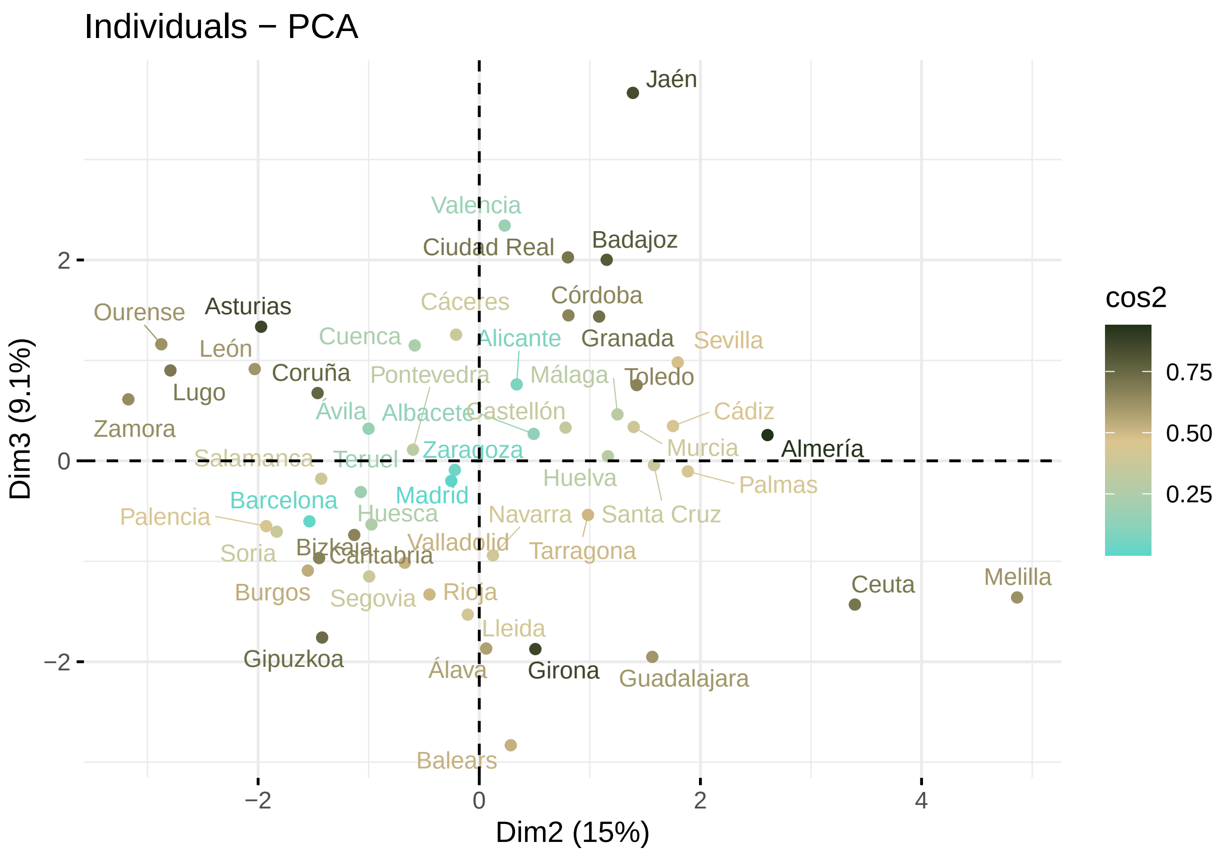
Analysis:

Graph #1 (Components 1 and 2)

In this graph we can see that Madrid and Barcelona are in the far right side of the graphic, which indicates that they have the highest values of employed people, number of companies, population, and all kinds of industries, having a great contrast with the other provinces that are located at the left side of the table.

In the dimension 2 (second component) we can see how provinces like Melilla, Ceuta and Almería show a high mortality rate, birth rate, activity and CPI, but also a great unemployment rate, unlike provinces like Zamora, Ourense, Lugo, Palencia and León, despite being in the same spectrum regarding dimension 1 (first component).

```
fviz_pca_ind(
  fit3, axes = c(2, 3), col.ind = "cos2", col.cex = 0.2,
  gradient.cols = c("#5CD6CA", "#DCC48E", "#243119"), repel = TRUE,
  labelsize = 3
  )
```



Graph #2 (Components 2 and 3)

In this table we can see the relation between components 2 and 3, and we can see that provinces like Jaén, Valencia, Ciudad Real and Badajoz have a high value in ACNH (Agrarian Census Numer of Holdings), Unemployment rate, Activity rate and CPI (Consumer Price Index), unlike provinces like Balears, Gipuzkoa, Álava, Girona and Guadalajara.

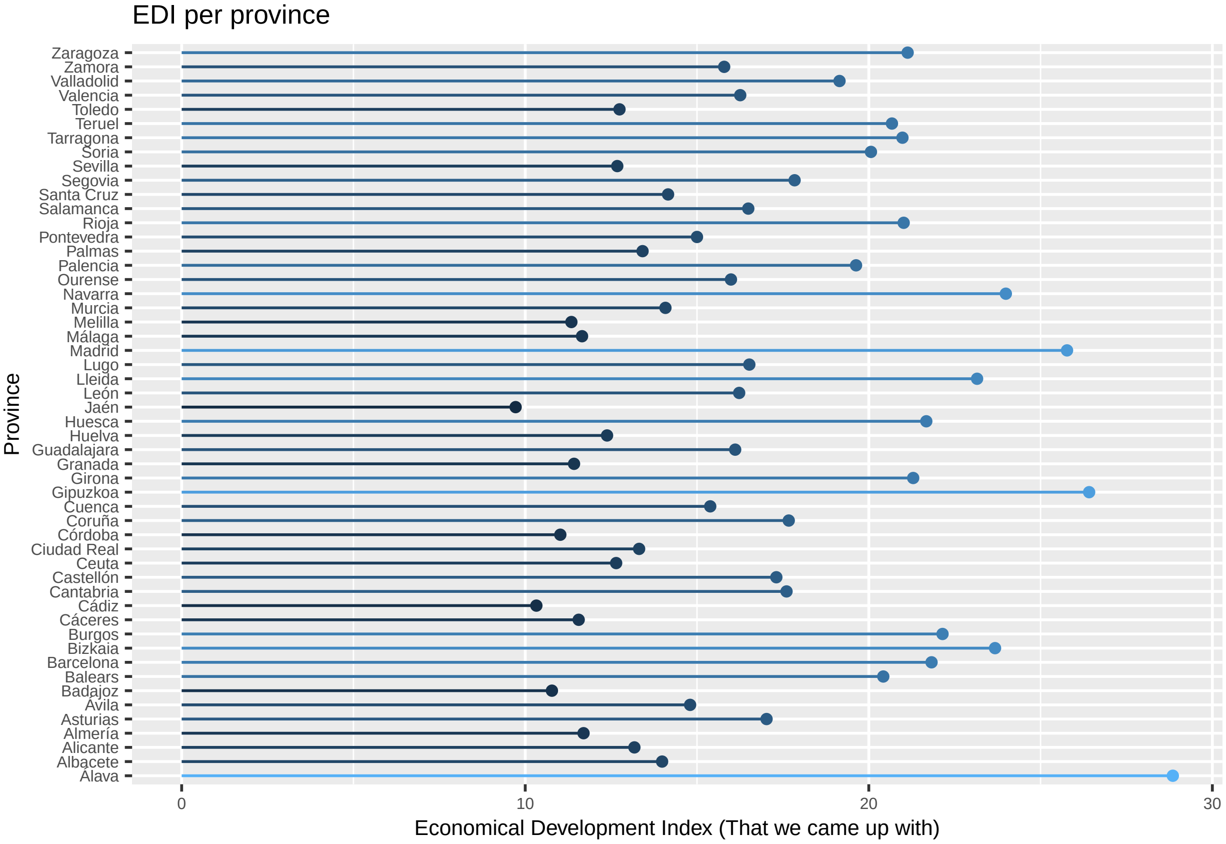
3.g. Coming up with an index that measures the economic development of a given province

We are going to choose and use some variables to measure the economic development of these provinces.

```
datos_ide <- datos_n
datos_ide$EDI <- (
  (datos_ide$GDP/datos_ide$Population)-((datos_ide$GDP/datos_ide$Population)*(datos_ide$UnemplymtRate/1
)
ide_data <- data.frame(
  province = rownames(datos_ide),
  ide = datos_ide$EDI
)
```

```
## <ScaleContinuous>
##  Range:
```

```
##  Limits:    0 --    1
```



What we did:

Using the variables we have available in this dataset, we took the population number, the GDP and the unemployment rate to interact in order to get our EDI (our made-up Economical Development Index).

We are going to divide the GDP and the population number, and to the result of that we are going to take out the percentage in the unemployment rate variable, so we adjust that number to better represent the real benefit from the perspective of the inhabitants of that province (imagine we have a place where very few people make a lot of money, but most of the population is unemployed… if we don't do this we would have a high EDI in a place where people don't really benefit from the money made there).

The formula for our EDI looks like this:

$$EDI = \left(\frac{GDP}{TotalPopulation}\right) - \left(\left(\frac{GDP}{TotalPopulation}\right) * \left(\frac{UnemploymentRate}{100}\right)\right)$$

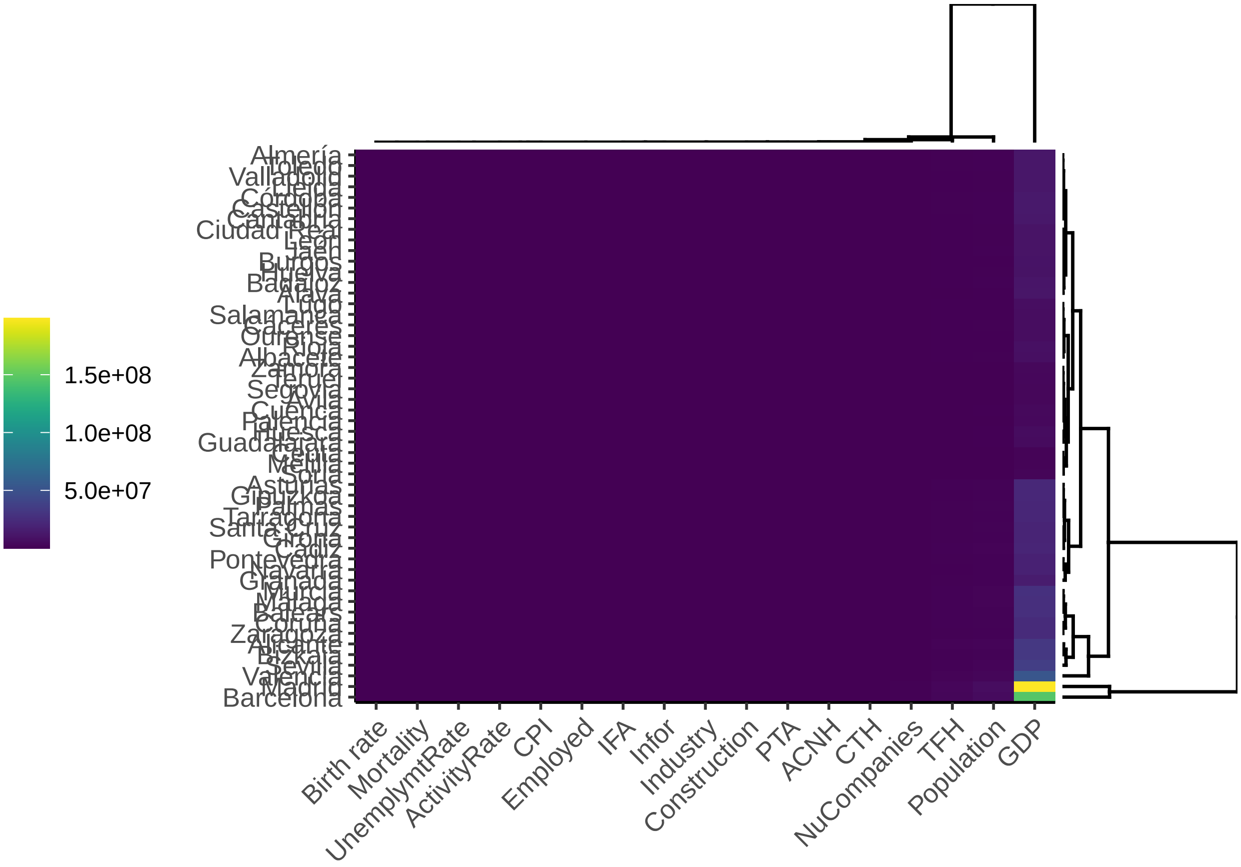
When applying this formula, the EDI for Madrid is 25.77 and Melilla's is 11.34. The highest EDI is in Álava, with 28.84, and the lowest is in Jaén, with 9.72.

4. Data matrix heat map

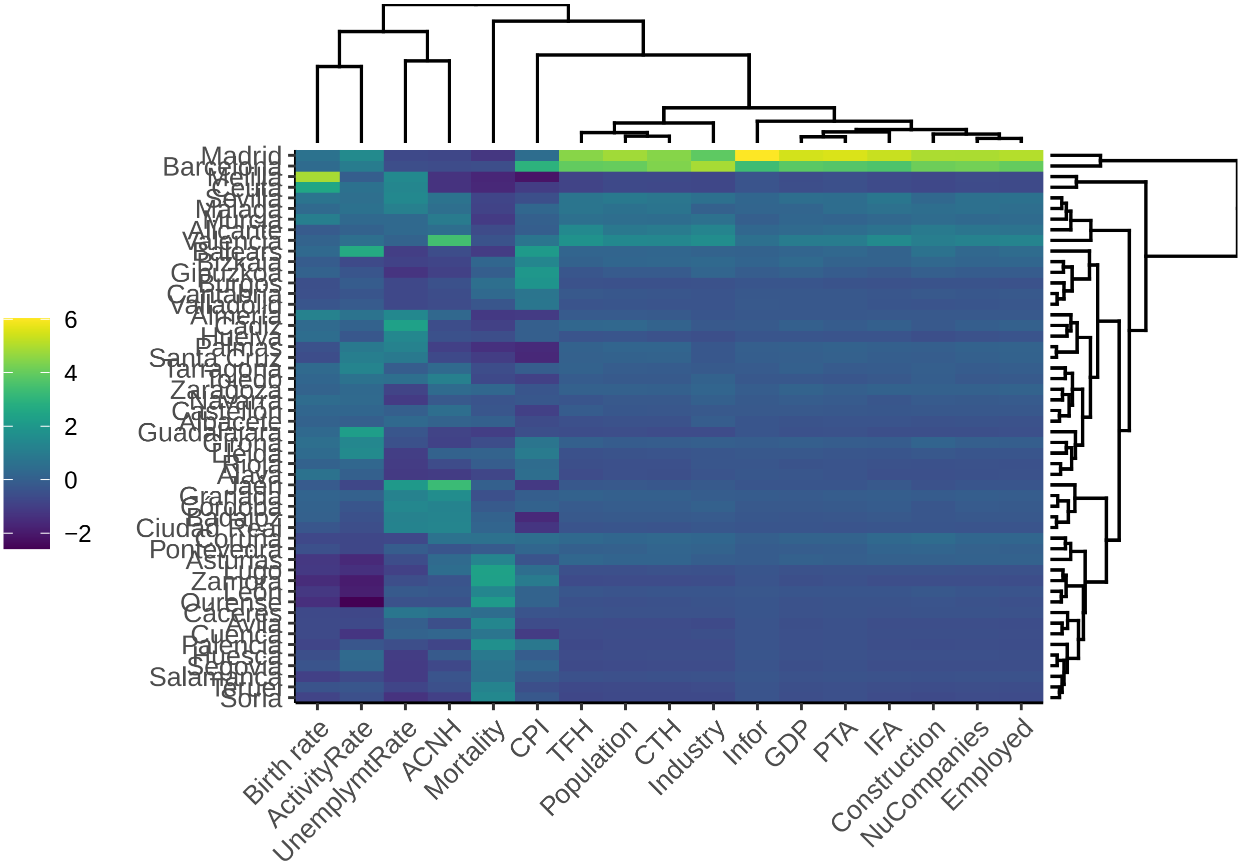
We are going to try to find groups of provinces using heat maps, using raw and standardized data.

```
datos_ST <- scale(datos_n)

ggheatmap(datos_n, seriate = "mean")
```



```
ggheatmap(datos_ST, seriate = "mean")
```



There is a clear difference between the group formed by Madrid and Barcelona and all the other provinces, which have very similar colors between them.

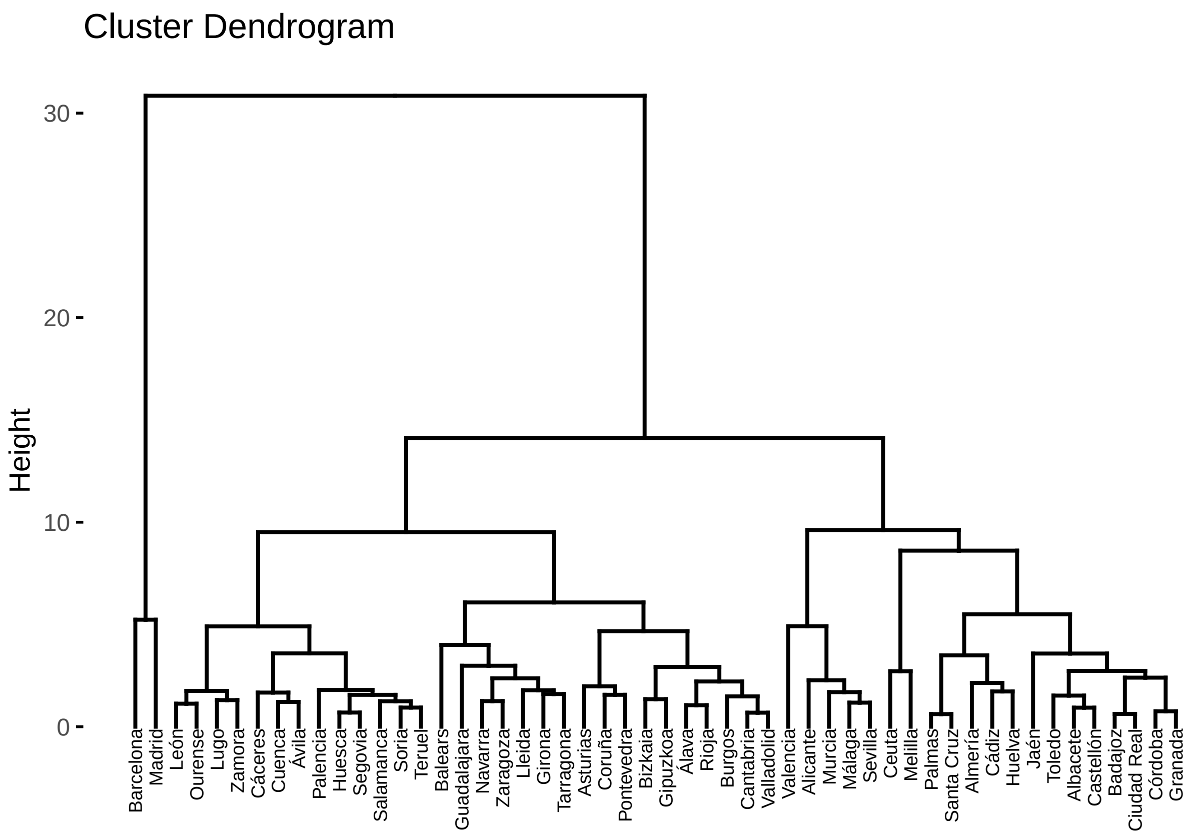
5. Hierarchic analysis of clusters

We are going to analyze clusters to determine if there are groups of provinces with similar behaviours.

5.a. Cluster number recommendation by dendrogram analysis

```
# Distancias
d_st <- dist(datos_ST, method = "euclidean")
# Agrupamos y dibujamos dendrograma
res.hc_st <- hclust(d_st, method = "ward.D2")
fviz_dend(res.hc_st, cex = 0.5)
```

```
## Warning: ‘guides(<scale> = FALSE)‘ is deprecated. Please use ‘guides(<scale> =
## "none")‘ instead.
```



```
grp <- cutree(res.hc_st, k = 5)
```

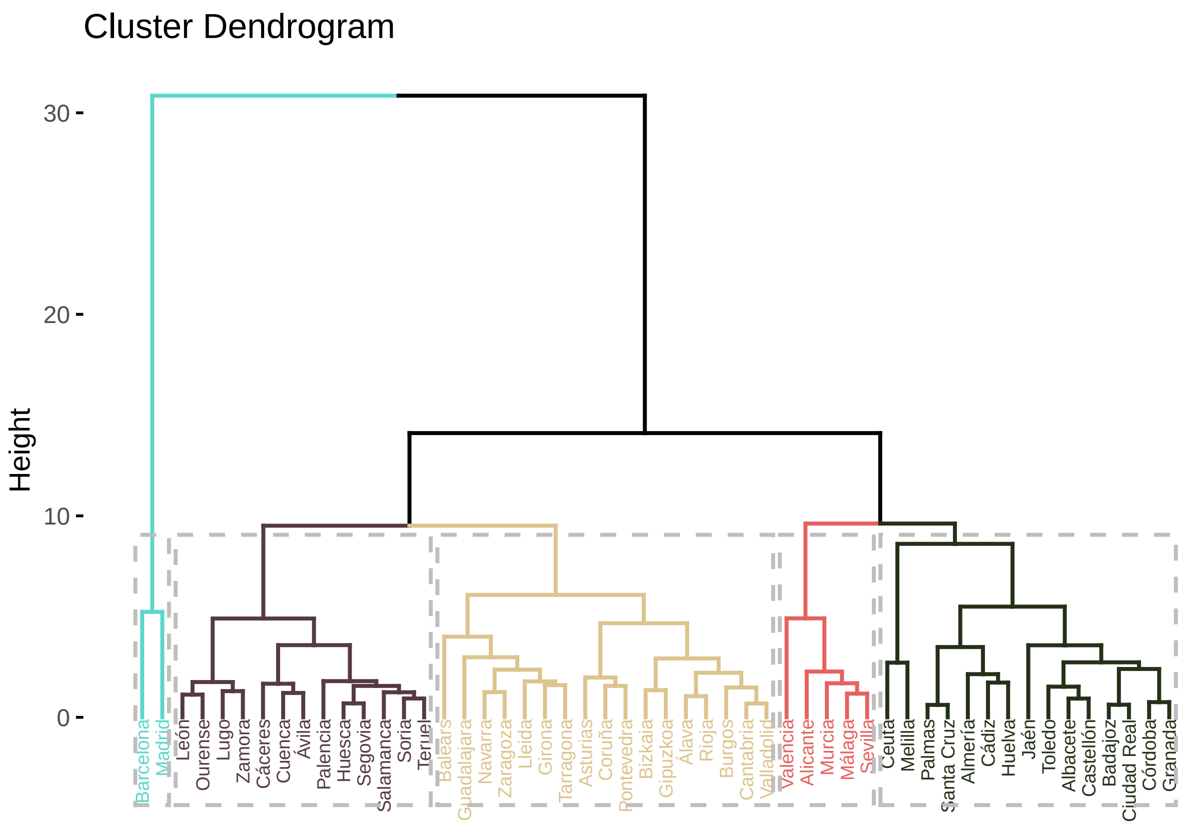
Analysis:

Up to this point, I would recommend 5 clusters.

5.b. Representation of the grouped individuals (according to the chosen number of clusters)

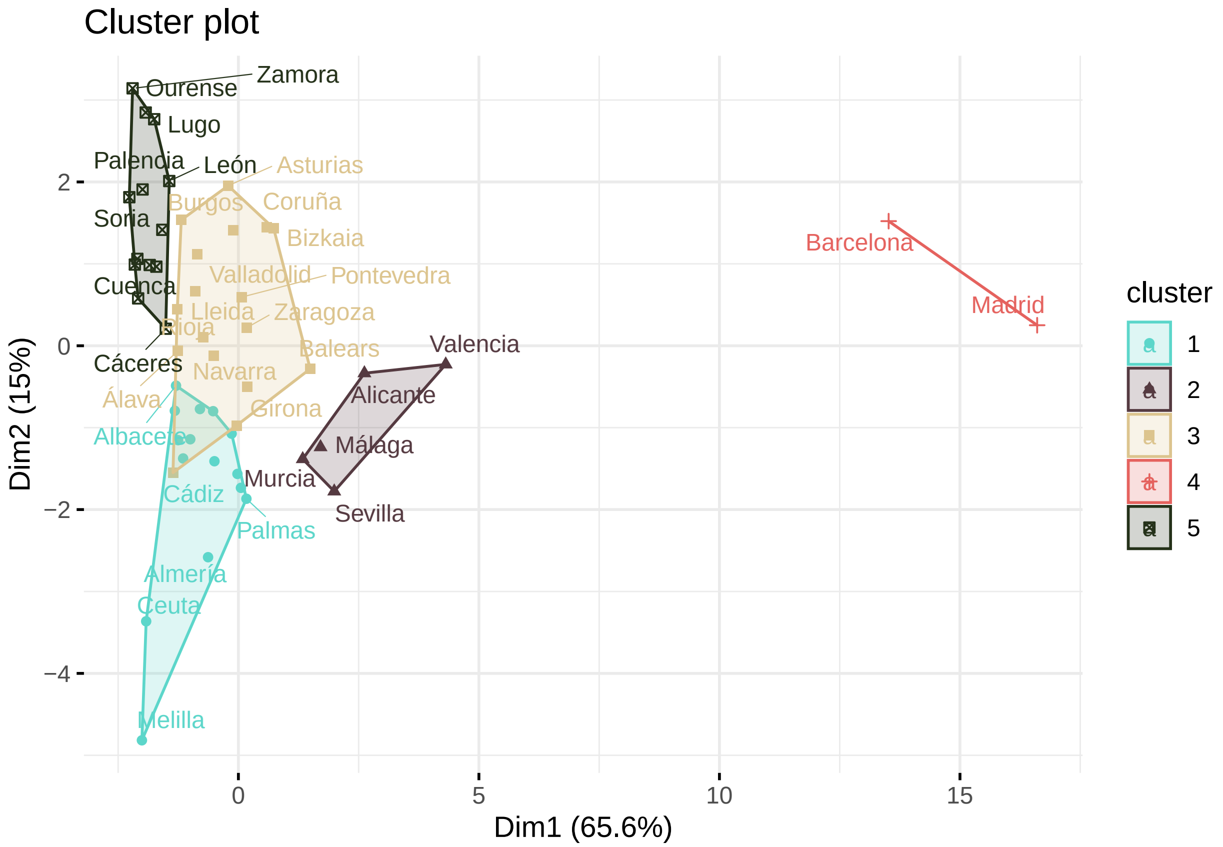
```
fviz_dend(res.hc_st, k = 5, cex = 0.5, color_labels_by_k = TRUE, rect = TRUE,
          palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
)
```

```
## Warning: ‘guides(<scale> = FALSE)‘ is deprecated. Please use ‘guides(<scale> =
## "none")‘ instead.
```



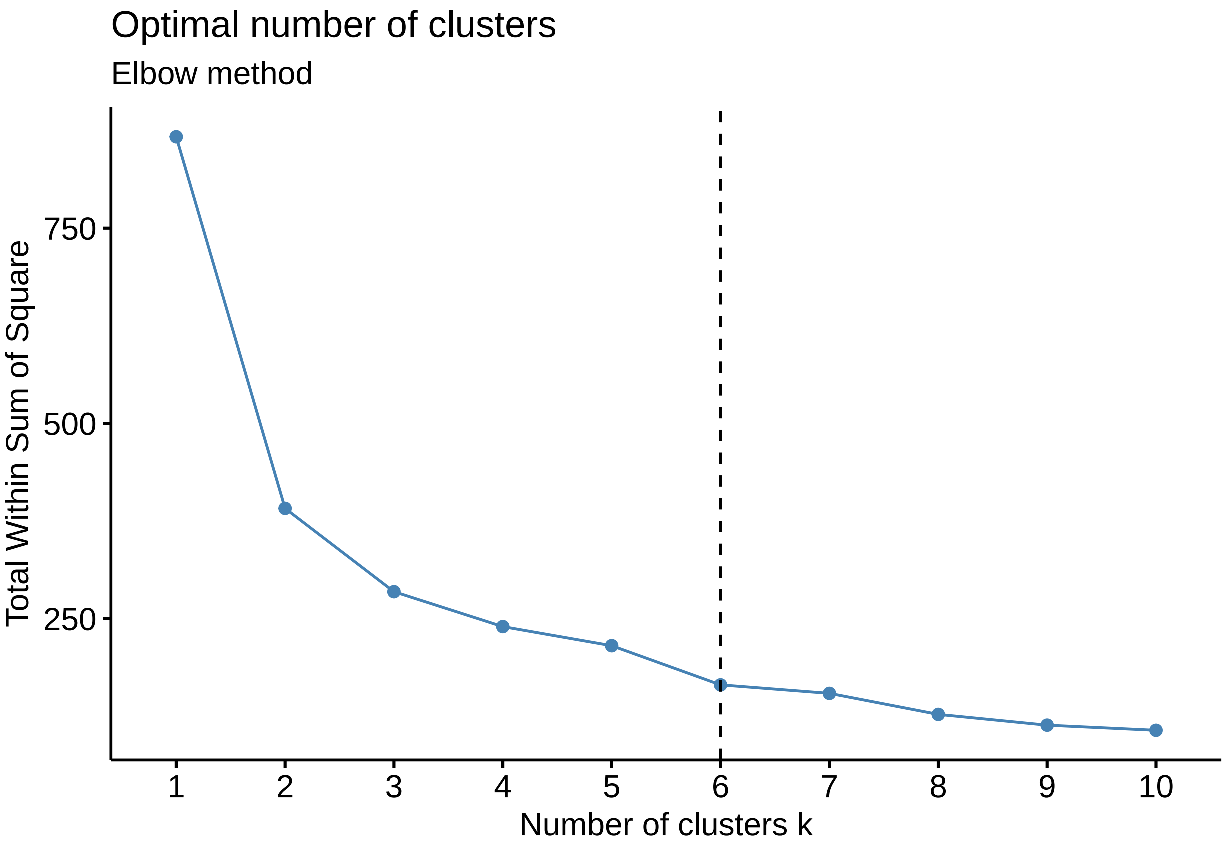
```
fviz_cluster(list(data = datos_ST, cluster = grp),
             ellipse.type = "convex", repel = TRUE, show.clust.cent = FALSE,
             ggtheme = theme_minimal(), labelsize = 9,
             palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
             )
```

```
## Warning: ggrepel: 18 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

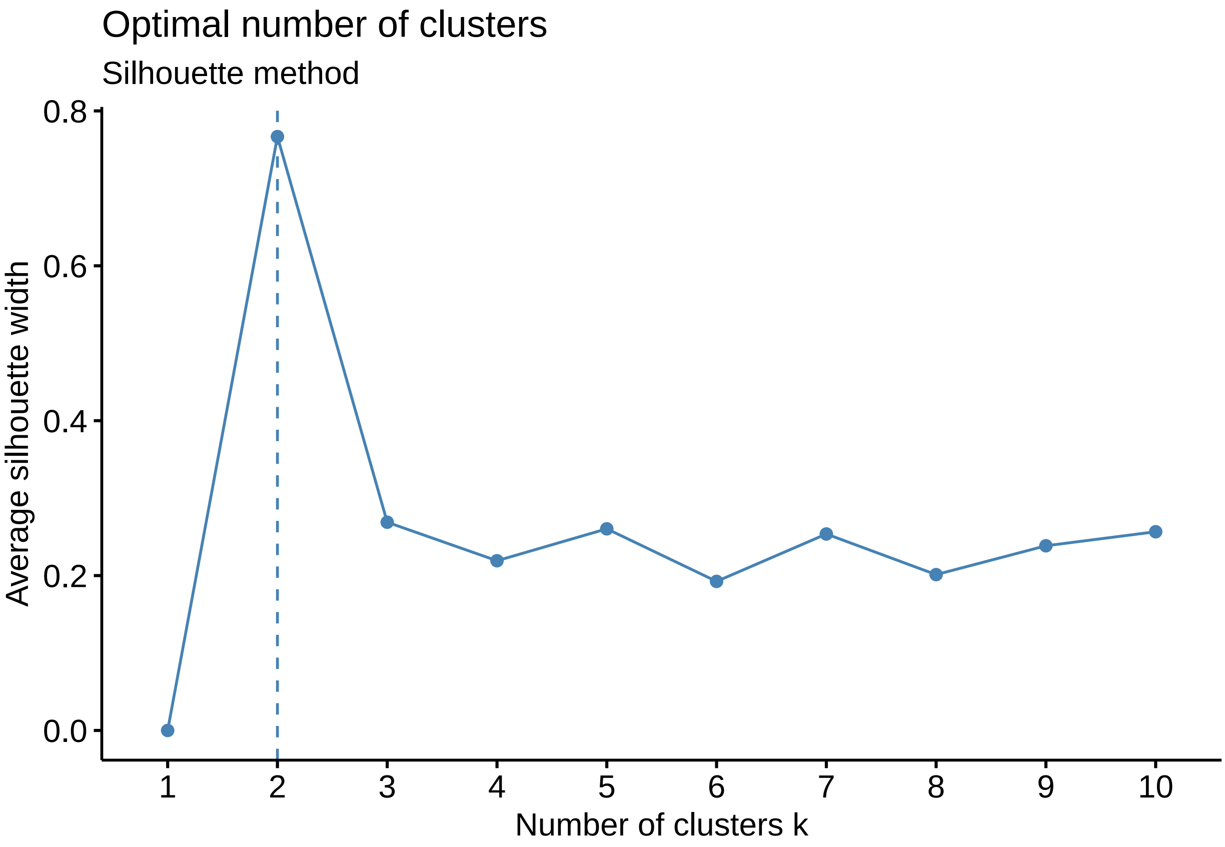


5.c. Optimal number of clusters suggested by the Elbow and Silhouette methods

```
# Determinación de clusters metodo Elbow
fviz_nbclust(datos_ST, kmeans, method = "wss") +
  geom_vline(xintercept = 6, linetype = 2) +
  labs(subtitle = "Elbow method")
```



```
# Determinación de clusters metodo Silhouette
fviz_nbclust(datos_ST, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```



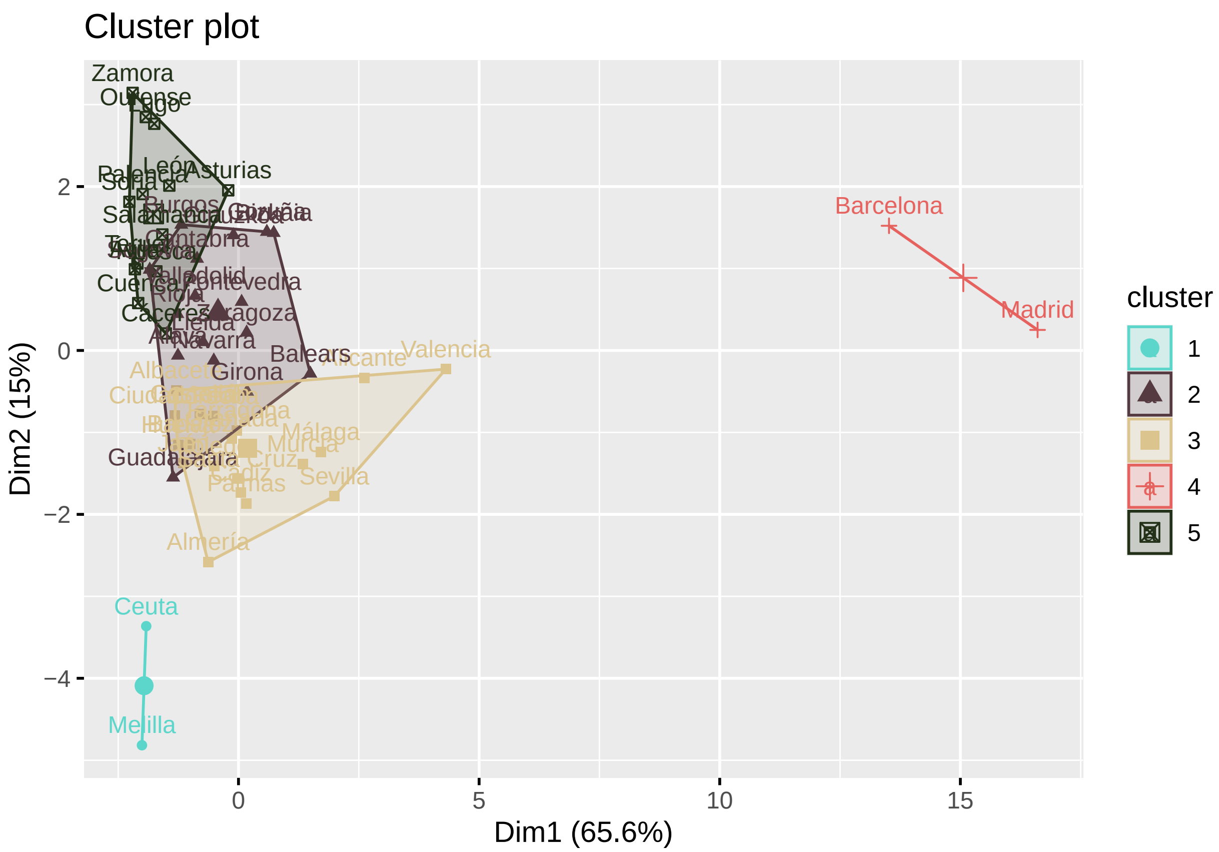
The Elbow method suggests 6 clusters, whilst the Silhouette method suggests 2. We are going to stick with the 5 we chose, since when we plot 6 clusters some of them overlap and then we know that they are not very different from each other, so we have a more diverse representation using only 5.

5.d. Non hierarchical grouping

5.d.i. Cluster representation in the main component dimensions.

Table #1 (Components 1 and 2):

```
RNGkind(sample.kind = "Rejection")
set.seed(1234)
km.res5 <- kmeans(datos_ST, 5)
fviz_cluster(
  km.res5,
  datos_ST,
  axes = c(1, 2),
  labelsize = 9,
  palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
)
```



In this graph we can see how cluster #1 is the lowest in number of companies, employed people, family homes, activity and the mortality rate is higher than the birth rate, but the unemployment rate is not that low.

Cluster #2 has a low number of companies, of employed people, of family homes, activity rate and the unemployment rate is medium-high.

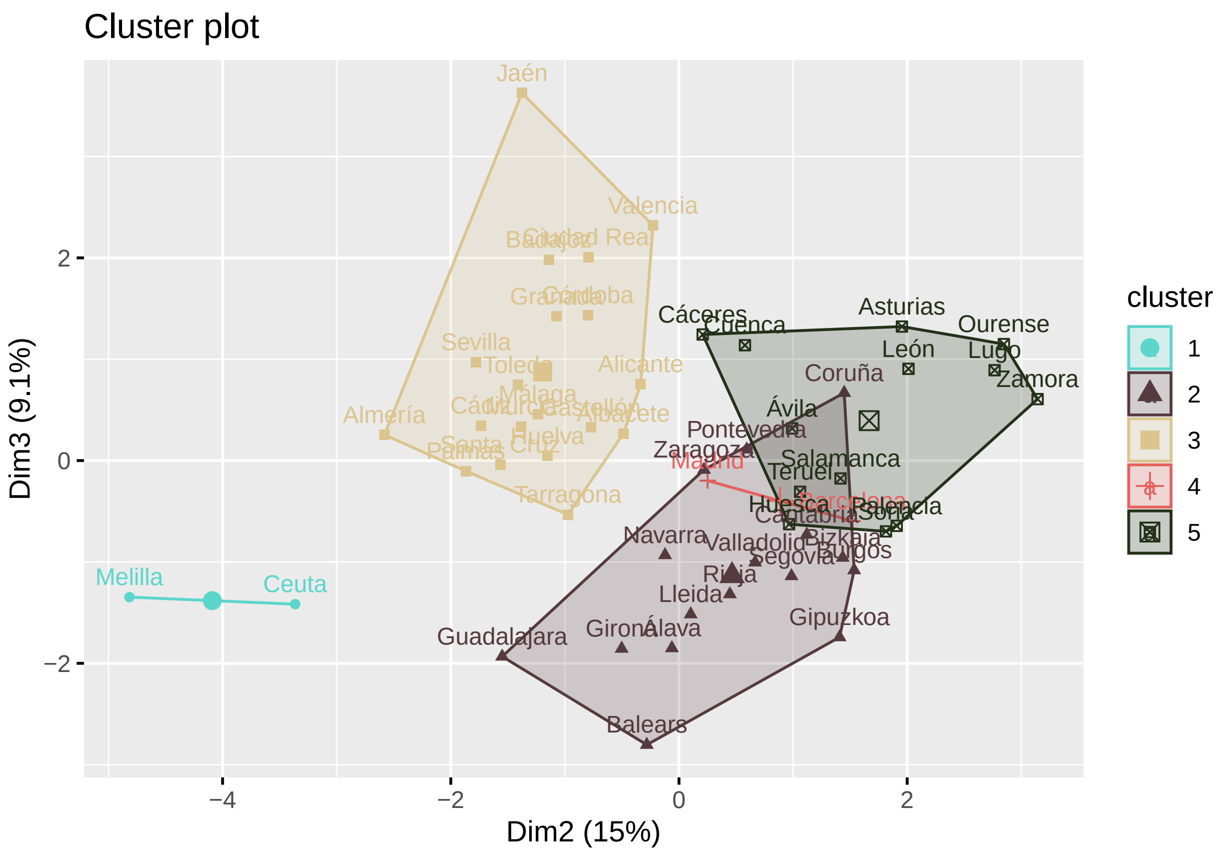
Cluster #3 has a low company number, of employed people, of family homes, activity rate is also low and the unemployment rate is medium.

Cluster #4 has a high population number, a high number of companies, of employed people, a high GDP and also a high number of family homes, a high activity rate and birth rate.

Cluster #5 has a low number of companies, of employed people, of family homes, a low activity rate, and the unemployment and birth rates are the highest.

Table #2 (Components 2 and 3):

```
fviz_cluster(
  km.res5,
  datos_ST,
  axes = c(2, 3),
  labelsize = 9,
  palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
)
```



In this graph we can see that cluster #1 has the highest CPI and the lowest unemployment rate, the mortality rate is higher that the birth rate and it also has the lowest ACNH number (farms or Agrarian Holdings).

Cluster #2 has the most balanced birth and mortality rates, the ACNH number is medium-low and the activity rate is medium.

Cluster #3 has the highest ACNH number and the mortality rate is higher than the birth rate, though that difference is not as big as in cluster #1.

Cluster #4 has a birth rate slightly higher than the mortality rate, and the ACNH is medium, like unemployment and activity rates.

Cluster #5 has the highest unemployment rate and the ACNH is medium-high.

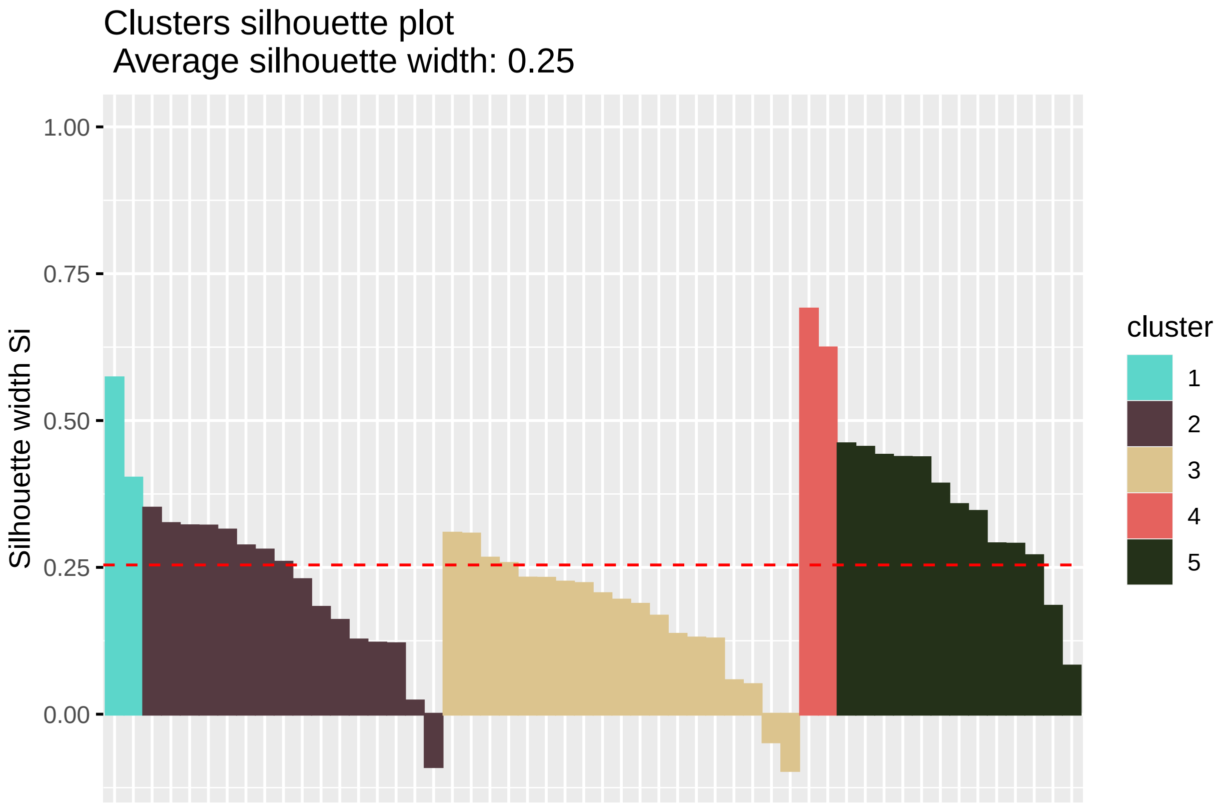
5.d.ii. Cluster number quality evaluation

```
sil <- silhouette(km.res5$cluster, dist(datos_ST))
rownames(sil) <- rownames(datos)
head(sil[, 1:3])
```

```
##           cluster neighbor   sil_width
## Albacete        3        5 -0.04700607
## Alicante        3        2  0.13600495
## Almería         3        1  0.22249485
## Álava           2        5  0.31357339
## Asturias        5        2  0.26989721
## Badajoz         3        5  0.23178580
```

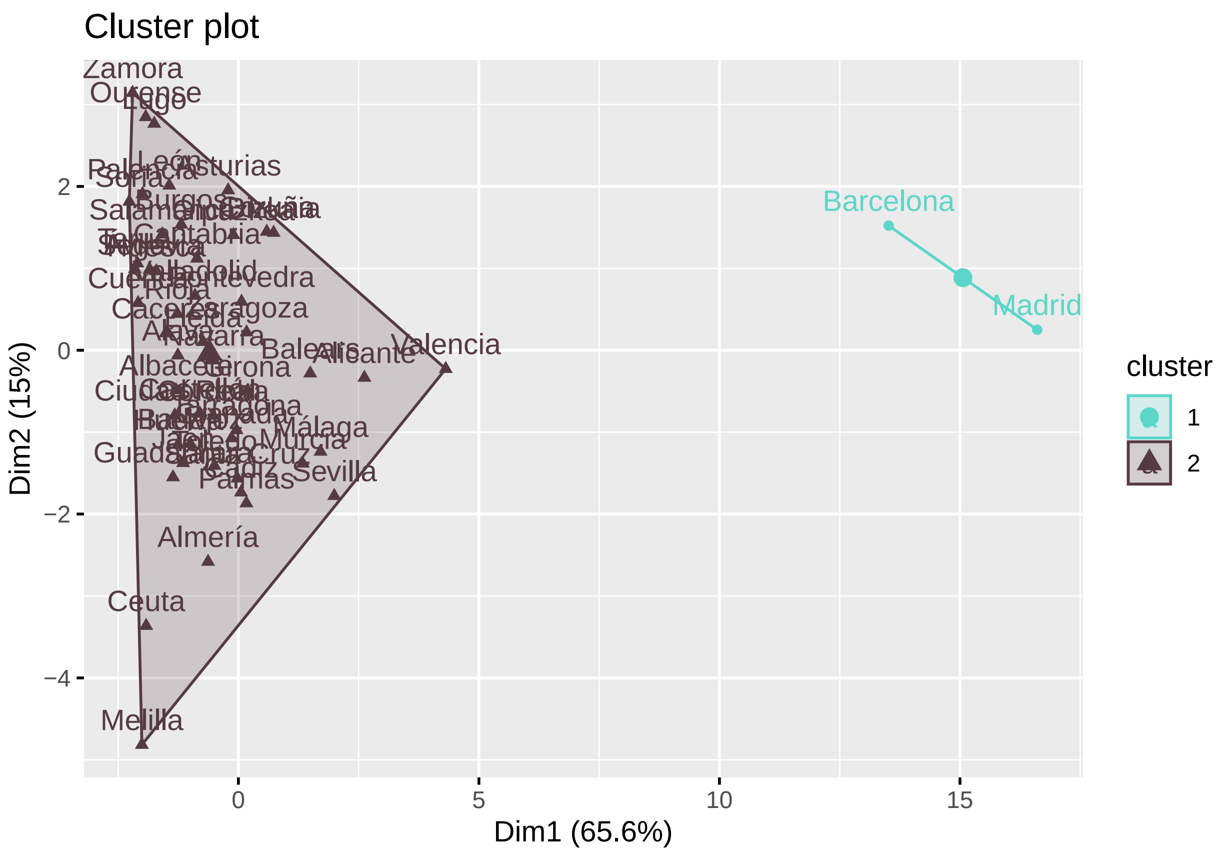
```
fviz_silhouette(sil, palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119"))
```

```
##   cluster size ave.sil.width
## 1       1    2          0.49
## 2       2   16          0.21
## 3       3   19          0.17
## 4       4    2          0.66
## 5       5   13          0.34
```

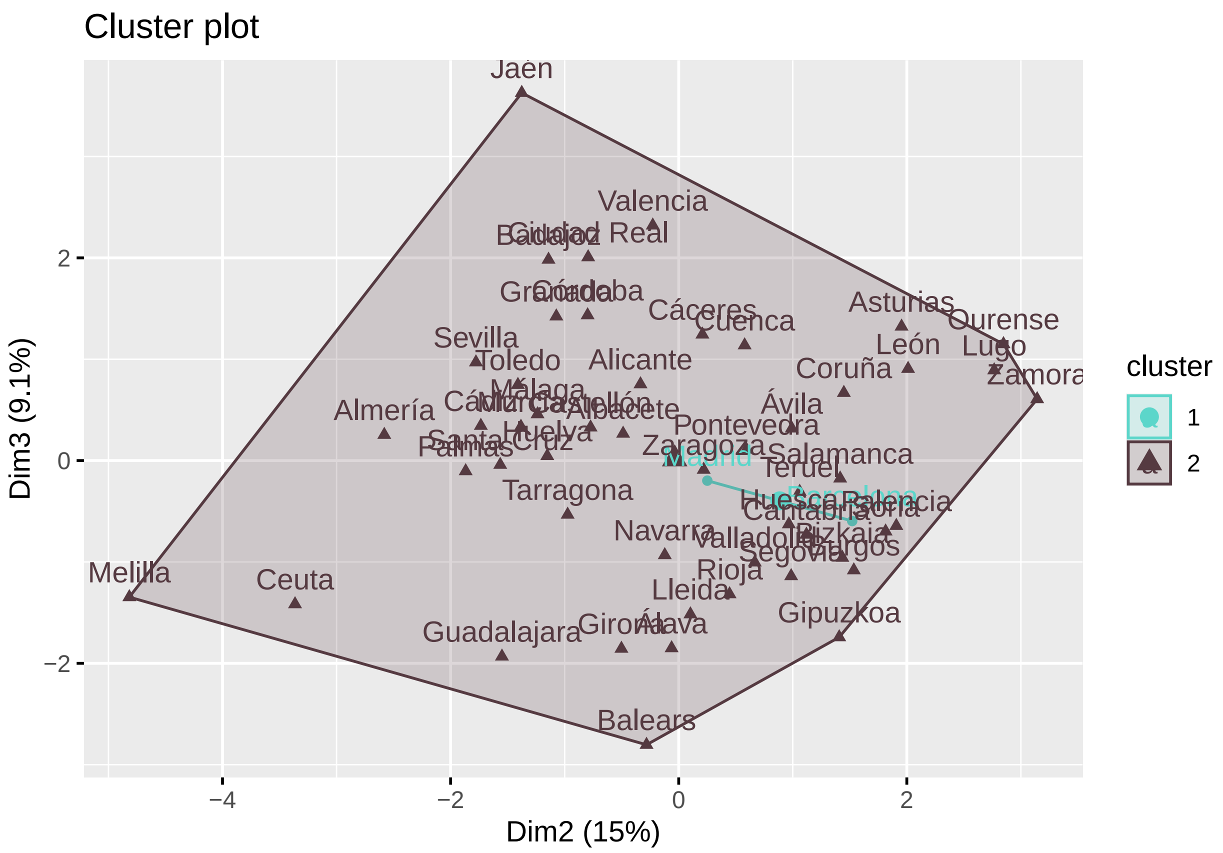


Let's try using only 2 clusters, which the Silhouette method suggested:

```
RNGkind(sample.kind = "Rejection")
set.seed(1234)
km.res2 <- kmeans(datos_ST, 2)
fviz_cluster(km.res2, datos_ST,
             axes = c(1, 2),
             palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
             )
```

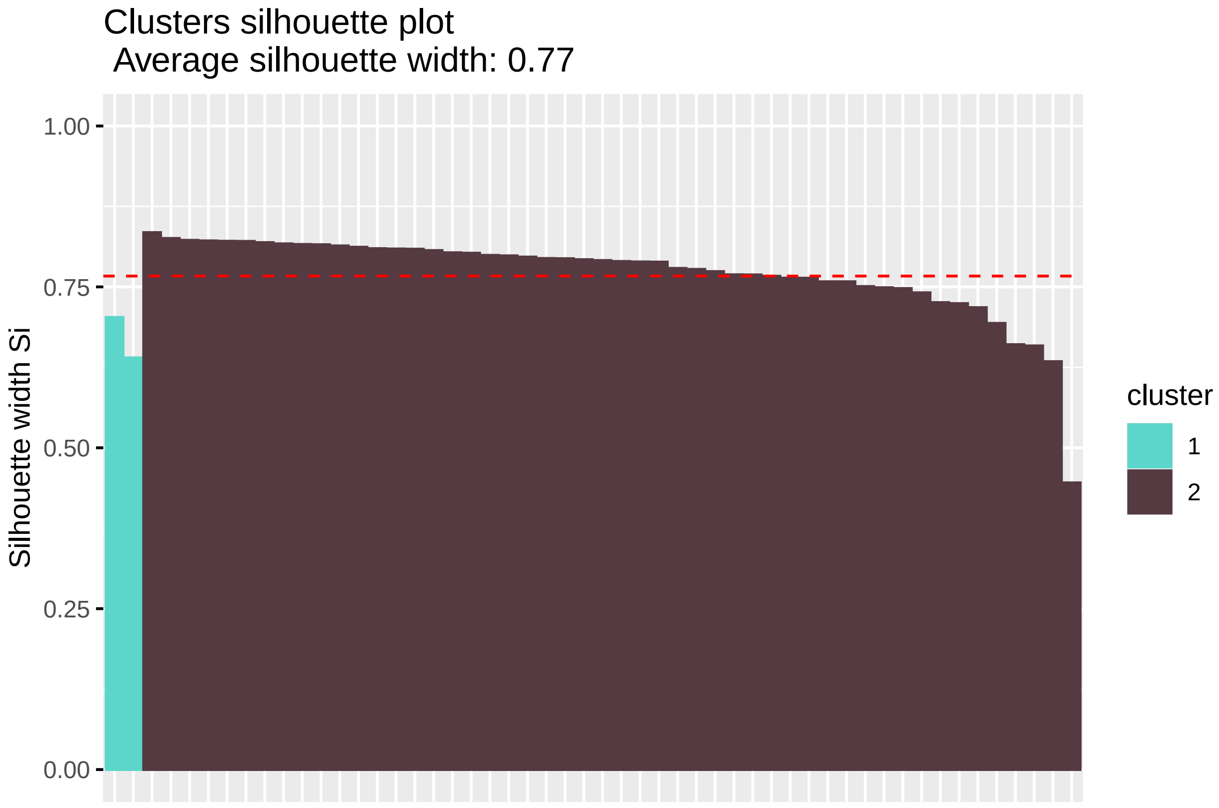


```
fviz_cluster(km.res2, datos_ST,
             axes = c(2, 3),
             palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119")
             )
```



```r
sil2 <- silhouette(km.res2$cluster, dist(datos_ST))
rownames(sil2) <- rownames(datos)
fviz_silhouette(sil2, palette = c("#5CD6CA", "#553A41", "#DCC48E", "#E5625E", "#243119"))
```

```
##   cluster size ave.sil.width
## 1       1    2          0.67
## 2       2   50          0.77
```



Here we can see that the two clusters are very different from each other, but we don't get much detail about the cluster #2 since it has every province except Madrid and Barcelona, and that cluster occupies a very big space in the graph (meaning there can be two provinces very different from each other that can be in the same cluster).

5.e. Socioeconomical characteristics of each cluster (and its provinces)

Analysis:

Cluster 1:

Provinces: Ceuta and Melilla

These are small provinces (in fact, Ceuta and Melilla are two cities that belong to Spain but are physically in the northern shore of Africa), so that's why they have such a small number of companies, of family homes and activity rates but its unemployment rate is not too high. Since they are very small, the ACNH cannot be too high. They also have the lowest EDI.

Mean EDI: 12.0

Cluster 2:

Provinces: Álava, Balears, Bizkaia, Burgos, Cantabria, Coruña, Gipuzkoa, Girona, Guadalajara, Lleida, Navarra, Pontevedra, Rioja, Segovia, Valladolid and Zaragoza.

What we see that defines this group is the balance between birth and mortality rates, the number of companies is still low and the ACNH goes from low to medium, although its mean EDI is the second best after cluster #4 which corresponds to Madrid and Barcelona.

Mean EDI: 21.0

Cluster 3:

Provinces: Albacete, Alicante, Almería, Badajoz, Castellón, Ciudad Real, Cádiz, Córdoba, Granada, Huelva, Jaén, Murcia, Málaga, Palmas, Santa Cruz, Sevilla, Tarragona, Toledo and Valencia.

This group of provinces is the one with the most agrarian activity (ACNH), but it has a low number of employed people and companies, of family homes and it has a considerable unemployment rate. It is important to mention that this cluster has the second lowest EDI after cluster #1 which corresponds to Ceuta and Melilla.

Mean EDI:: 13.2

Cluster 4:

Provinces: Barcelona y Madrid

In cluster #4 we have the catalan community capital and also the capital of all Spain. This cluster has a birth rate slightly higher then the mortality rate and the ACNH is medium, but what definitely differentiates it its the high number of companies of all kinds, its high population number, employed people, activity rate and also the highest GDP by a lot.

This can be explained by the centralized nature of Spain's economy.

Mean EDI: 23.8

Cluster 5:

Provinces: Asturias, Cuenca, Cáceres, Huesca, León, Lugo, Ourense, Palencia, Salamanca, Soria, Teruel, Zamora and Ávila

This cluster has a low number of companies, employed people, family homes and a low activity rate, and its ACNH is medium-high. This makes us think that these are provinces with a regular economic development, despite its considerable agrarian activity.

Mean EDI: 17.1