Minería de Datos y Modelización Predictiva (II)

Fernández Hernández, Alberto. 54003003S

1/2/2021

Table of Contents

## PARTE 1. ANÁLISIS DE COMPONENTES PRINCIPALES

### Apartado 1

**Calcular la matriz de correlaciones, y su representación gráfica ¿Cuáles son las variables más correlacionadas de forma inversa?**

De forma previa a la matriz de correlación, debemos cargar el conjunto de datos, además de renombrar los nombres de fila empleando los nombres de provincia:

provincias <- as.data.frame(read\_excel("Provincias.xlsx"))  
rownames(provincias) <- provincias$Provincia; provincias <- provincias[, -1] # Eliminamos finalmente la columna "Provincia"

Una vez cargado el conjunto de datos, calculamos la matriz de correlación:

matriz.correlaciones <-cor(provincias)  
library(formattable)  
formattable(as.data.frame(matriz.correlaciones), list(  
 'Mortalidad'= color\_tile("orange", "orange")  
))

Poblacion

Mortalidad

Natalidad

IPC

NumEmpresas

Industria

Construccion

CTH

Infor

AFS

APT

TasaActividad

TasaParo

Ocupados

PIB

CANE

TVF

VS

Poblacion

1.000000000

-0.34359937

0.11445938

0.33483095

0.99453328

0.96458267

0.98416996

0.996110046

0.94453881

0.99220221

0.980028733

0.33414642

0.009850323

0.99550003

0.980976869

0.097205159

0.99071731

0.56653976

Mortalidad

-0.343599371

1.00000000

-0.73652925

0.18846127

-0.30927237

-0.28184861

-0.29590918

-0.329333667

-0.25983697

-0.30518551

-0.295728912

-0.73348067

-0.463972419

-0.32761124

-0.295604063

0.016432872

-0.33230307

-0.24903743

Natalidad

0.114459382

-0.73652925

1.00000000

-0.25391157

0.10501778

0.09274529

0.08874307

0.101444356

0.11454957

0.10467705

0.113268704

0.47242170

0.383038405

0.11174153

0.113829452

-0.123091063

0.08444871

-0.02766720

IPC

0.334830949

0.18846127

-0.25391157

1.00000000

0.36475679

0.41971227

0.39988432

0.362397751

0.30044315

0.32267001

0.332616084

0.08970542

-0.582234824

0.35818832

0.355394973

-0.191207090

0.34095210

0.19117175

NumEmpresas

0.994533276

-0.30927237

0.10501778

0.36475679

1.00000000

0.96870767

0.99458429

0.993878918

0.96270044

0.99263984

0.990582046

0.33232416

-0.056232280

0.99834423

0.990277980

0.044834740

0.98226824

0.54044321

Industria

0.964582668

-0.28184861

0.09274529

0.41971227

0.96870767

1.00000000

0.96613670

0.976499325

0.88925300

0.94716931

0.934149162

0.29177908

-0.080353746

0.96153241

0.942717675

0.123766207

0.96533176

0.56716093

Construccion

0.984169957

-0.29590918

0.08874307

0.39988432

0.99458429

0.96613670

1.00000000

0.985702825

0.95796587

0.98427617

0.984579370

0.34122871

-0.109170963

0.99320060

0.985293522

0.029744519

0.97574923

0.55654176

CTH

0.996110046

-0.32933367

0.10144436

0.36239775

0.99387892

0.97649932

0.98570283

1.000000000

0.92902236

0.98333802

0.971043919

0.32593482

-0.007188617

0.99112993

0.971695960

0.093619748

0.99103137

0.57820586

Infor

0.944538806

-0.25983697

0.11454957

0.30044315

0.96270044

0.88925300

0.95796587

0.929022359

1.00000000

0.96891562

0.989614166

0.30815427

-0.108863576

0.96445717

0.985447530

-0.066339831

0.91150808

0.40901985

AFS

0.992202208

-0.30518551

0.10467705

0.32267001

0.99263984

0.94716931

0.98427617

0.983338024

0.96891562

1.00000000

0.991606815

0.31749496

-0.031727168

0.99462721

0.989333140

0.085739618

0.97871645

0.54055564

APT

0.980028733

-0.29572891

0.11326870

0.33261608

0.99058205

0.93414916

0.98457937

0.971043919

0.98961417

0.99160681

1.000000000

0.33229728

-0.078587345

0.99168825

0.996801753

-0.005592777

0.95694304

0.47971244

TasaActividad

0.334146420

-0.73348067

0.47242170

0.08970542

0.33232416

0.29177908

0.34122871

0.325934818

0.30815427

0.31749496

0.332297285

1.00000000

0.027812476

0.35391218

0.330715195

-0.116659302

0.32660569

0.26454726

TasaParo

0.009850323

-0.46397242

0.38303840

-0.58223482

-0.05623228

-0.08035375

-0.10917096

-0.007188617

-0.10886358

-0.03172717

-0.078587345

0.02781248

1.000000000

-0.05461326

-0.095297068

0.391013190

0.01158550

0.09520974

Ocupados

0.995500027

-0.32761124

0.11174153

0.35818832

0.99834423

0.96153241

0.99320060

0.991129933

0.96445717

0.99462721

0.991688246

0.35391218

-0.054613263

1.00000000

0.992242627

0.048739757

0.98216232

0.53953310

PIB

0.980976869

-0.29560406

0.11382945

0.35539497

0.99027798

0.94271768

0.98529352

0.971695960

0.98544753

0.98933314

0.996801753

0.33071519

-0.095297068

0.99224263

1.000000000

-0.007093683

0.95687829

0.46930178

CANE

0.097205159

0.01643287

-0.12309106

-0.19120709

0.04483474

0.12376621

0.02974452

0.093619748

-0.06633983

0.08573962

-0.005592777

-0.11665930

0.391013190

0.04873976

-0.007093683

1.000000000

0.15298830

0.33581554

TVF

0.990717314

-0.33230307

0.08444871

0.34095210

0.98226824

0.96533176

0.97574923

0.991031374

0.91150808

0.97871645

0.956943036

0.32660569

0.011585501

0.98216232

0.956878285

0.152988299

1.00000000

0.66764132

VS

0.566539758

-0.24903743

-0.02766720

0.19117175

0.54044321

0.56716093

0.55654176

0.578205863

0.40901985

0.54055564

0.479712437

0.26454726

0.095209736

0.53953310

0.469301776

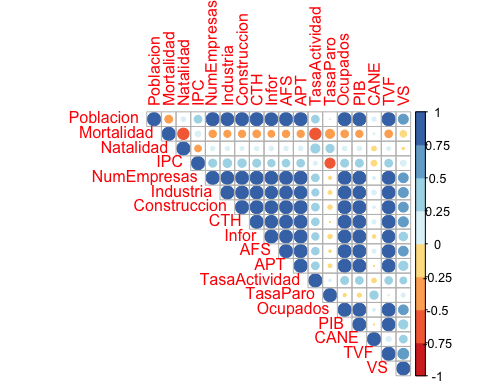
0.335815536

0.66764132

1.00000000

Así como su representación gráfica:

corrplot(matriz.correlaciones, type="upper", col=brewer.pal(n=8, name="RdYlBu"))



Analizando la gráfica de correlación, de forma inversa nos encontramos con múltiples variables aunque la mayoría de ellas con valores muy cercanos a 0. No obstante, si debemos destacar los siguientes casos:

1. **MORTALIDAD - NATALIDAD**: presenta una elevada correlación negativa (-0.74). Es decir, aquellas provincias donde es mayor la tasa de natalidad, la tasa de mortalidad se reduce.
2. **MORTALIDAD - TASA\_ACTIVIDAD**: con una correlación muy similar a la anterior (-0.73), esto es, a medida que aumenta la tasa de mortalidad en una provincia, el porcentaje de personas en “edad laboral” disminuye.
3. **MORTALIDAD - TASA\_PARO**: con una correlación bastante menor a las dos anteriores, aunque también significativa (-0.46): en provincias donde la tasa de mortalidad es elevada, el número de parados es menor (y viceversa).

En relación con la mortalidad, **podemos decir que a medida que aumenta la mortalidad en la provincia, disminuyen las tasas de natalidad, de actividad, e incluso la tasa de desempleo** (y viceversa).

1. **IPC - TASA\_PARO**: una correlación inferior a -0.5, por lo que también es significativa (-0.58): **a medida que aumenta el índice de precios al consumidor, la tasa de paro se reduce**. Dicho de otro modo, en aquellas provincias donde el coste de vida es mayor, la tasa de desempleo, por lo general, parece ser menor.

Aunque no se trata de una correlación negativa, quisiera destacar la correlación existente entre *NumEmpresas*, *Industria*, *Construccion*, *CTH*, *Infor*, *AFS* y *APT*, así como la tasa de ocupados, el *PIB* y el número de viviendas familiares, **con correlaciones muy cercanas a 1 entre sí**. Por tanto, provincias con un elevado número de empresas no sólo tendran un mayor número de empresas en cada sector (industria, construcción, comercio, hostelería, finanzas y actividades profesionales), sino además una mayor tasa de empleo, mayor PIB y un elevado número de viviendas familiares.

### Apartado 2

**Realizar un análisis de componentes principales sobre la matriz de correlaciones, calculando 7 componentes. Estudiar los valores de los autovalores obtenidos y las gráficas que los resumen. ¿Cuál es el número adecuado de componentes?**

Con los datos de provincias, realizamos el análisis con un total de siete componentes, tal y como se muestra a continuación:

analisis.comp <- PCA(provincias, scale.unit = TRUE, ncp = 7, graph = FALSE)  
# Obtenemos los autovalores  
knitr::kable(analisis.comp$eig, digits =2,caption = "Autovalores")

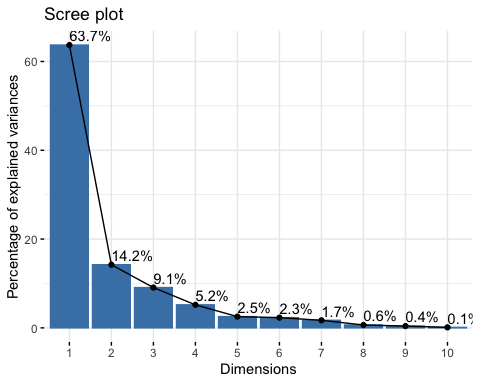
Autovalores

|  |  |  |  |
| --- | --- | --- | --- |
|  | eigenvalue | percentage of variance | cumulative percentage of variance |
| comp 1 | 11.47 | 63.70 | 63.70 |
| comp 2 | 2.56 | 14.23 | 77.93 |
| comp 3 | 1.63 | 9.08 | 87.01 |
| comp 4 | 0.93 | 5.19 | 92.19 |
| comp 5 | 0.46 | 2.54 | 94.73 |
| comp 6 | 0.41 | 2.30 | 97.03 |
| comp 7 | 0.31 | 1.71 | 98.74 |
| comp 8 | 0.12 | 0.65 | 99.39 |
| comp 9 | 0.07 | 0.41 | 99.79 |
| comp 10 | 0.02 | 0.11 | 99.91 |
| comp 11 | 0.01 | 0.05 | 99.96 |
| comp 12 | 0.00 | 0.02 | 99.98 |
| comp 13 | 0.00 | 0.01 | 99.99 |
| comp 14 | 0.00 | 0.00 | 99.99 |
| comp 15 | 0.00 | 0.00 | 100.00 |
| comp 16 | 0.00 | 0.00 | 100.00 |
| comp 17 | 0.00 | 0.00 | 100.00 |
| comp 18 | 0.00 | 0.00 | 100.00 |

Analizando la tabla anterior, podemos observar como con 7 componentes **se explica aproximadamente el 98.74 % de la variabilidad de los datos**. No obstante, ¿Es el número adecuado de componentes? Para responder a esta pregunta debemos fijarnos no solo en el porcentaje acumulado de la varianza, sino además en los **autovalores**. Entre la primera y la tercera componente nos encontramos con autovalores superiores a 1 (11.47, 2.56 y 1.63, respectivamente). Dado que la suma de los autovalores coincide con el número de variables del fichero, observamos que la componente 1 ya es capaz de explicar cerca del 63.70 % de la variabilidad total del conjunto de datos estandarizado, aproximadamente 11 de las 18 variables. Si a lo anterior le añadimos los autovalores de la segunda y tercera componente, logramos explicar cerca del 90 % de la variabilidad (cerca de 16 variables). En adelante, la variabilidad total apenas aumenta hasta la componente 7, cuyo autovalor es de 0.31, es decir, no llega a explicar ni siquiera una variable.

A primera vista, la opción adecuada sería elegir 3 componentes, dado que sus autovalores son superiores a 1. No obstante, analicemos el siguiente gráfico:

# Representamos graficamente la proporcion de varianza explicada por cada Componente Principal  
fviz\_eig(analisis.comp,addlabels=TRUE)



En función de la proporción de varianza, la componente 4 también parece aportar un porcentaje significativo con respecto a las siguientes componentes (5.2 %). De hecho, aunque su autovalor no sea 1, es muy cercano (0.93), llegando a explicar **casi** una variable. Por tanto, dado que a partir de la componente 5 la variabilidad explicada comienza a disminuir por debajo del 5 %, además de que con 4 componentes logramos explicar más del 90 % de la variabilidad, minimizamos la pérdida de información, por lo que **elegimos 4 como número adecuado de componentes**.

### Apartado 3

**Hacer de nuevo el análisis sobre la matriz de correlaciones pero ahora indicando el número de componentes principales que hemos decidido retener (Que expliquen aproximadamente el 90%). Sobre este análisis contestar los siguientes apartados.**

# Ahora con 4 componentes  
analisis.comp <- PCA(provincias, scale.unit = TRUE, ncp = 4, graph = FALSE)

#### Apartado a)

**Mostrar los coeficientes para obtener las componentes principales ¿Cuál es la expresión para calcular la primera Componente en función de las variables originales?**

En primer lugar, mostramos los coeficientes necesarios para el cálculo de la componente, esto es, **los autovectores**:

autovectores.pc1 <- t(analisis.comp$svd$V[,1])  
colnames(autovectores.pc1) <- colnames(provincias)  
knitr::kable(autovectores.pc1, digits =2,caption = "Autovectores de la primera componente")

Autovectores de la primera componente

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Poblacion | Mortalidad | Natalidad | IPC | NumEmpresas | Industria | Construccion | CTH | Infor | AFS | APT | TasaActividad | TasaParo | Ocupados | PIB | CANE | TVF | VS |
| 0.29 | -0.11 | 0.04 | 0.11 | 0.29 | 0.29 | 0.29 | 0.29 | 0.28 | 0.29 | 0.29 | 0.11 | -0.01 | 0.29 | 0.29 | 0.02 | 0.29 | 0.17 |

Por tanto, para calcular la primera Componente basta con realizar la combinación lineal de los autovectores:

Donde "\*" indican que las variables deben estar **estandarizadas**. Como podemos observar, nos encontramos con autovectores negativos en la primera componente, de los cuales la tasa de Mortalidad es la más influyente. Por el contrario, variables como la tasa de Natalidad, la tasa de paro o el CANE apenas influyen en el valor de la componente, con coeficientes muy cercanos a 0 (parece tratarse de un índice principalmente económico más que demográfico).

#### Apartado b)

**Mostar una tabla con las correlaciones de las Variables con las Componentes Principales. Para cada Componente indicar las variables con las que está más correlacionada**

var<-get\_pca\_var(analisis.comp)  
knitr::kable(var$cor, digits =3,caption = "Correlaciones de la CP con las variables")

Correlaciones de la CP con las variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Dim.3 | Dim.4 |
| Poblacion | 0.994 | 0.004 | 0.064 | -0.052 |
| Mortalidad | -0.360 | -0.843 | 0.241 | -0.156 |
| Natalidad | 0.138 | 0.793 | -0.346 | -0.106 |
| IPC | 0.372 | -0.585 | -0.335 | 0.420 |
| NumEmpresas | 0.996 | -0.042 | 0.010 | -0.067 |
| Industria | 0.967 | -0.072 | 0.059 | 0.023 |
| Construccion | 0.993 | -0.073 | -0.015 | -0.026 |
| CTH | 0.992 | -0.017 | 0.062 | -0.027 |
| Infor | 0.953 | -0.067 | -0.083 | -0.214 |
| AFS | 0.990 | -0.026 | 0.051 | -0.089 |
| APT | 0.984 | -0.047 | -0.036 | -0.137 |
| TasaActividad | 0.387 | 0.529 | -0.464 | 0.448 |
| TasaParo | -0.047 | 0.739 | 0.495 | -0.213 |
| Ocupados | 0.997 | -0.027 | 0.003 | -0.058 |
| PIB | 0.985 | -0.058 | -0.048 | -0.129 |
| CANE | 0.060 | 0.154 | 0.840 | 0.269 |
| TVF | 0.987 | -0.003 | 0.128 | 0.043 |
| VS | 0.584 | 0.077 | 0.371 | 0.548 |

Analizando la tabla de correlaciones, destacamos las variables con las que más está correlacionada cada componente:

1. **COMPONENTE 1**: destaca la elevada correlación positiva con los campos *Poblacion*, *NumEmpresas*, *Industria*, *Construccion*, *CTH*, *Infor*, *AFS*, *APT* (tanto el total de empresas como el número de empresas de cada sector se concentran en torno a la misma componente), *Ocupados*, el *PIB*, el total de viviendas familiares o *TVF* y *VS* o viviendas secundarias. Todas ellas estan correlacionadas de forma directa. Por tanto, **valores positivos de la COMPONENTE 1 se traducen en una provincia con altos niveles de población, número de empresas, tasa de ocupación, PIB y número de viviendas** (tanto principales como de segunda residencia).

Por otro lado, nos encontramos con correlaciones menos significativas pero que también debemos tener en cuenta: correlaciones positivas con el IPC y la tasa de Actividad, aunque más bajas (0.37~0.38). Lo mismo, aunque en sentido negativo, ocurre con la Mortalidad.

1. **COMPONENTE 2**: destaca la correlación negativa entre *Mortalidad* e *IPC*, junto con una correlación positiva con los campos *Natalidad*, *TasaActividad* y *TasaParo*, principalmente. Por tanto, **valores positivos de la COMPONENTE 2 implican provincias con altos niveles de natalidad, tasa de actividad y de paro, aunque con una menor tasa de mortalidad y poder adquisitivo de las familias (IPC bajo)**. Del mismo modo, valores negativos de la COMPONENTE 2 se traducen en provincias donde la tasa de mortalidad y el poder adquisitivo de las familias es mayor (IPC alto), aunque con una menor tasa de natalidad, de actividad laboral y paro.
2. **COMPONENTE 3**: destaca la variable CANE con mayor correlación (positiva). Por otro lado, de cara la interpretación de las variables debemos tener en cuenta también correlaciones positivas (aunque menos significativas) con las tasas de paro y el número de viviendas secundarias (0.495 y 0.371), así como negativas con los valores de natalidad, IPC y tasa de actividad (-0.346, -0.335 y -0.464).

No obstante, ¿qué ocurre con la componente 4? Como podemos observar, pese a que el valor máximo de correlación corresponde con el campo *VS*, dicha variable presenta una mayor correlación con la primera componente (0.584 frente a 0.548). Por tanto, podemos asegurar que la cuarta componente **no llega a representar exclusivamente a ninguna de las variables del conjunto de datos**, pese a lograr explicar una variabilidad del 92 % tal y como vimos en el apartado anterior. Como conclusión, **definitivamente realizaremos el análisis sobre la matriz de correlación empleando tan solo las tres primeras componentes**:

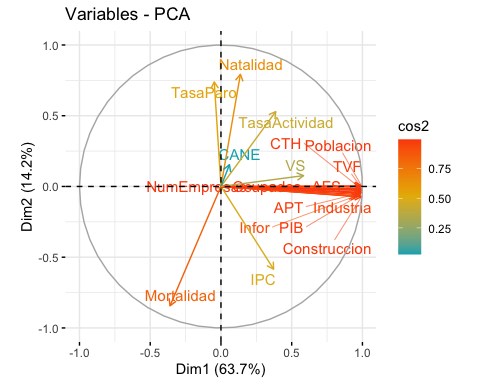
analisis.comp<-PCA(provincias,scale.unit=TRUE, ncp=3, graph=FALSE)  
var<-get\_pca\_var(analisis.comp)

#### Apartado c)

**Comentar los gráficos que representan las variables en los planos formados por las componentes, intentando explicar lo que representa cada componente**

Inicialmente, comenzamos representando el gráfico de correlación entre las componentes 1 y 2:

fviz\_pca\_var(analisis.comp, axes = c(1, 2), col.var="cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)

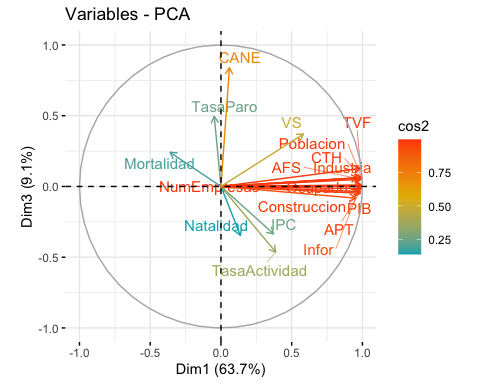


Como primer análisis, podemos observar cómo las variables que estaban correlacionadas positivamente entre sí (tal y como comprobamos en el gráfico de correlación), presentan prácticamente las mismas direcciones (Población, Ocupados o NumEmpresas, entre otros). Por el contrario, variables correlacionadas negativamente se sitúan en posiciones opuestas (cuadrantes opuestos), como por ejemplo la Natalidad y la Tasa de Actividad con la Mortalidad, o la Tasa de Paro y el IPC. A continuación, relacionamos cada componente con el conjunto de variables más representativo:

1. **COMPONENTE 1**: **podemos asegurar que la componente 1 es un índice que recoge principalmente la actividad económica de la provincia (número de empresas + PIB + tasa de ocupación)**, aunque también incluye información demográfica como la población total o el número de viviendas (aunque el número de viviendas secundarias en menor medida).
2. **COMPONENTE 2**: **corresponde con un índice principalmente demográfico**, contrastando especialmente los valores de Natalidad y Mortalidad de la Provincia, aunque también incluye información económica de la región como la proporción de población activa, nivel de desempleo y variación de los precios (aunque con menor correlación).

A continuación, analizamos la tercera componente:

fviz\_pca\_var(analisis.comp, axes = c(1, 3), col.var="cos2",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)



1. **COMPONENTE 3**: **representa principalmente al Censo Agrario del Número de Explotaciones (CANE)**, con el que mantiene una elevada correlación positiva.

#### Apartado d)

**Mostrar la tabla y los gráficos que nos muestran la proporción de la varianza de cada variable que es explicado por cada componente. ¿Cuál de las variables es la que está peor explicada?**

En primer lugar, representamos la tabla con los cosenos al cuadrado:

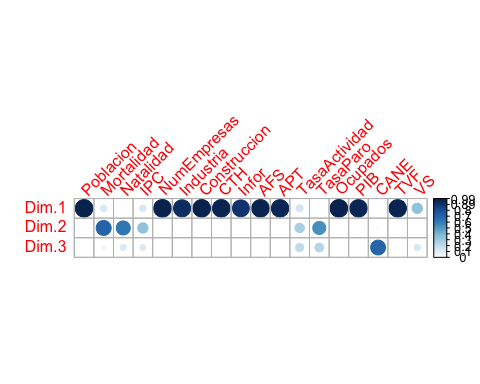
knitr::kable(var$cos2, digits =2,caption = "Cosenos al cuadrado")

Cosenos al cuadrado

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Dim.3 |
| Poblacion | 0.99 | 0.00 | 0.00 |
| Mortalidad | 0.13 | 0.71 | 0.06 |
| Natalidad | 0.02 | 0.63 | 0.12 |
| IPC | 0.14 | 0.34 | 0.11 |
| NumEmpresas | 0.99 | 0.00 | 0.00 |
| Industria | 0.94 | 0.01 | 0.00 |
| Construccion | 0.99 | 0.01 | 0.00 |
| CTH | 0.98 | 0.00 | 0.00 |
| Infor | 0.91 | 0.00 | 0.01 |
| AFS | 0.98 | 0.00 | 0.00 |
| APT | 0.97 | 0.00 | 0.00 |
| TasaActividad | 0.15 | 0.28 | 0.22 |
| TasaParo | 0.00 | 0.55 | 0.25 |
| Ocupados | 0.99 | 0.00 | 0.00 |
| PIB | 0.97 | 0.00 | 0.00 |
| CANE | 0.00 | 0.02 | 0.70 |
| TVF | 0.97 | 0.00 | 0.02 |
| VS | 0.34 | 0.01 | 0.14 |

Para visualizar mejor los resultados, lo representamos gráficamente:

corrplot(t(var$cos2), is.corr = FALSE, tl.srt=45)



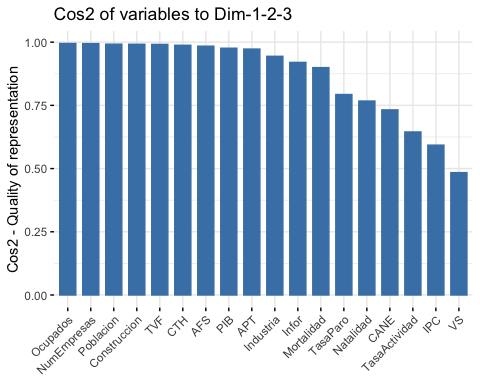
Dado que los cosenos al cuadrado nos indican la proporción de varianza, de cada variable, que es explicada en cada componente, observamos que la primera es **la que mejor explica la variación de sus variables más representativas**. Aunque bien es cierto que el número de viviendas secundarias o *VS* está peor explicada (con solo un 34 % de la varianza explicada), con el resto de variables llega a explicarse más del 90 % de la variabilidad.

En relación con la segunda componente, las variables mejor explicadas corresponden con las tasas de Mortalidad y Natalidad con un 71 y 63 % de la varianza, respectivamente. Por el contrario, están peor explicados los campos TasaParo, IPC y TasaActividad, en orden descendente (55, 34 y 28 %). De este modo, **podemos volver a comprobar como la información contenida en la primera componente es principalmente económica, mientras que la segunda contiene información demográfica**.

Por último, la componente 3 explica, principalmente, el campo CANE (con un 70 % de la varianza total), pudiendo considerarse como un índice económico adicional.

No obstante, ¿Cuál es la variable peor explicada por las componentes?

fviz\_cos2(analisis.comp, choice = "var", axes = 1:3)



Comprobamos que la variable peor explicada **corresponde con el número de viviendas secundarias, dado que el porcentaje de varianza explicada, en conjunto con las tres componentes, corresponde con un 49 %** (0.35 + 0.01 + 0.14).

#### Apartado e)

**Mostrar la tabla y los gráficos que nos muestran el porcentaje de la varianza de cada Componente que es debido a cada variable. ¿Qué variables contribuyen más a cada Componente?**

En primer lugar, mostramos la tabla con las contribuciones de cada variable a la varianza de cada componente:

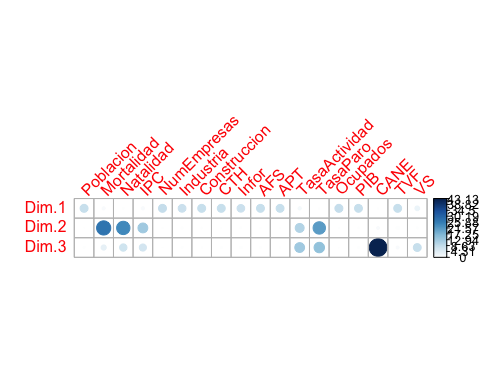
knitr::kable(var$contrib, digits = 2, caption = "Contribuciones")

Contribuciones

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Dim.3 |
| Poblacion | 8.62 | 0.00 | 0.25 |
| Mortalidad | 1.13 | 27.79 | 3.57 |
| Natalidad | 0.17 | 24.54 | 7.33 |
| IPC | 1.21 | 13.35 | 6.88 |
| NumEmpresas | 8.65 | 0.07 | 0.01 |
| Industria | 8.16 | 0.20 | 0.22 |
| Construccion | 8.60 | 0.21 | 0.01 |
| CTH | 8.58 | 0.01 | 0.24 |
| Infor | 7.92 | 0.18 | 0.42 |
| AFS | 8.55 | 0.03 | 0.16 |
| APT | 8.45 | 0.09 | 0.08 |
| TasaActividad | 1.31 | 10.93 | 13.16 |
| TasaParo | 0.02 | 21.30 | 15.00 |
| Ocupados | 8.67 | 0.03 | 0.00 |
| PIB | 8.46 | 0.13 | 0.14 |
| CANE | 0.03 | 0.93 | 43.13 |
| TVF | 8.50 | 0.00 | 1.00 |
| VS | 2.97 | 0.23 | 8.42 |

Así como su representación gráfica:

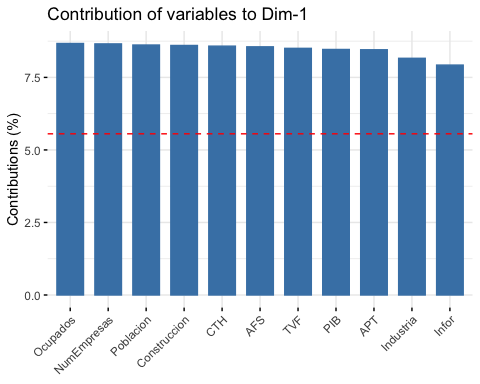
corrplot(t(var$contrib), is.corr = FALSE, tl.srt=45)



Analizando tanto la tabla como el gráfico, observamos que la componente 1 **presenta una varianza distribuida de forma equitativa sobre el conjunto de sus variables**, mientras que la varianza de las componentes 2 y 3 **es explicada principalmente por un pequeño subconjunto**.

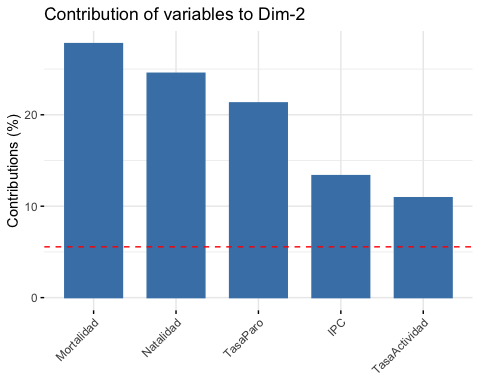
Como observamos en el siguiente gráfico, en relación con la primera componente la variable *Infor* es la que menor porcentaje aporta a la varianza (7.92 %). El resto de variables significativas presentan una aportación superior al 8 %, distribuyéndose en torno a variables económicas:

fviz\_contrib(analisis.comp,choice="var",axes=1,top=11)



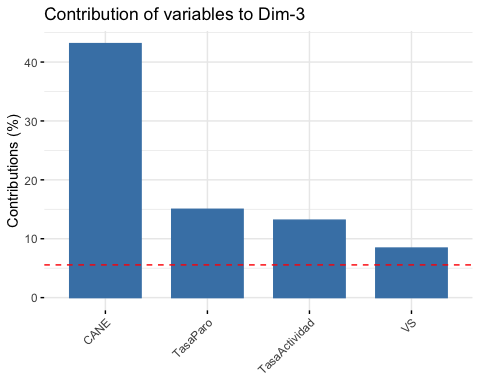
En relación con la segunda componente, los porcentajes de contribución se concentran mayoritariamente en torno a la información demográfica de la provincia, destacando los campos *Mortalidad*, *Natalidad*, como incluso *TasaParo*, cuyas contribuciones superan el 20 %. Por el contrario, los campos como el IPC y la Tasa de Actividad superan ligeramente el 10 % de la contribución:

fviz\_contrib(analisis.comp,choice="var",axes=2,top=5)



Por último, el porcentaje de contribución a la tercera componente se centra principalmente en torno a la variable *CANE*, acumulando el 43.13 % de su varianza total, siguiéndole en menor medida campos de menor correlación como la Tasa de Paro, Actividad y el número de viviendas secundarias:

fviz\_contrib(analisis.comp,choice="var",axes=3,top=4)

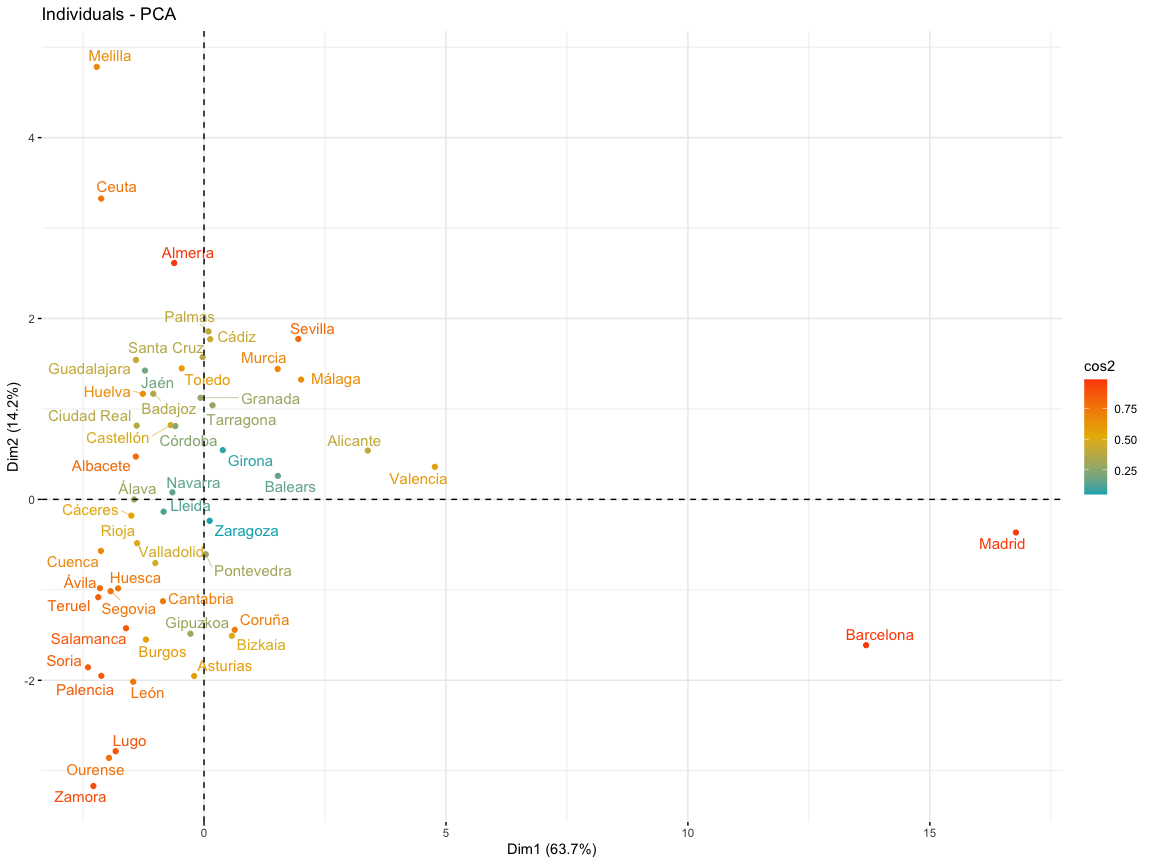


#### Apartado f)

**Sobre los gráficos que representan las observaciones en los nuevos ejes y el gráfico Biplot., teniendo en cuenta la posición de las provincias en el gráfico Comentar las provincias que tienen una posición más destacada en cada componente, en positivo o negativo, ¿Qué significa esto en términos socioeconómicos para estas provincias?**

Inicialmente, comenzamos con el gráfico con las observaciones en los nuevos ejes:

fviz\_pca\_ind(analisis.comp, axes = c(1, 2), col.ind = "cos2",  
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)



En un primer análisis, podemos observar ciertos grupos o *clusters* formados entre diferentes provincias. A modo de ejemplo:

1. **Las grandes ciudades como Madrid y Barcelona**, con valores positivos en la primera componente, lo que según la información que representa suponen altos niveles de población, viviendas familiares, tasa de ocupación y un elevado número de empresas; además de valores negativos en la segunda componente, esto es, mayor tasa mortalidad e IPC.
2. **Las provincias de Lugo, Ourense y Zamora**, con valores negativos en la segunda componente, es decir, altas de mortalidad e IPC, además de valores negativos en la primera componente, lo que se traduce en una menor población, número de viviendas y actividad económica.
3. **Las ciudades autónomas de Ceuta y Melilla**, ubicadas en valores positivos de la segunda componente: altas tasas de natalidad, desempleo y actividad laboral; aunque con valores negativos en la primera componente: menor población, número de viviendas y actividad económica.

Sin embargo, también podemos encontrar provincias que no destacan en ninguna de las componentes, esto es, los valores de las componentes se sitúan cerca de la media (dado que los datos están estandarizados, la media es cero). Como primera impresión, analizamos qué provincias se sitúan más cerca del centro del gráfico, a través de una intersección entre los gráficos de la Componente 1 - Componente 2 y la Componente 1 - Componente 3, filtrando aquellas provincias cuya distancia al centro del gráfico sea muy pequeña en ambos casos (a modo de ejemplo, menor a 0.5):

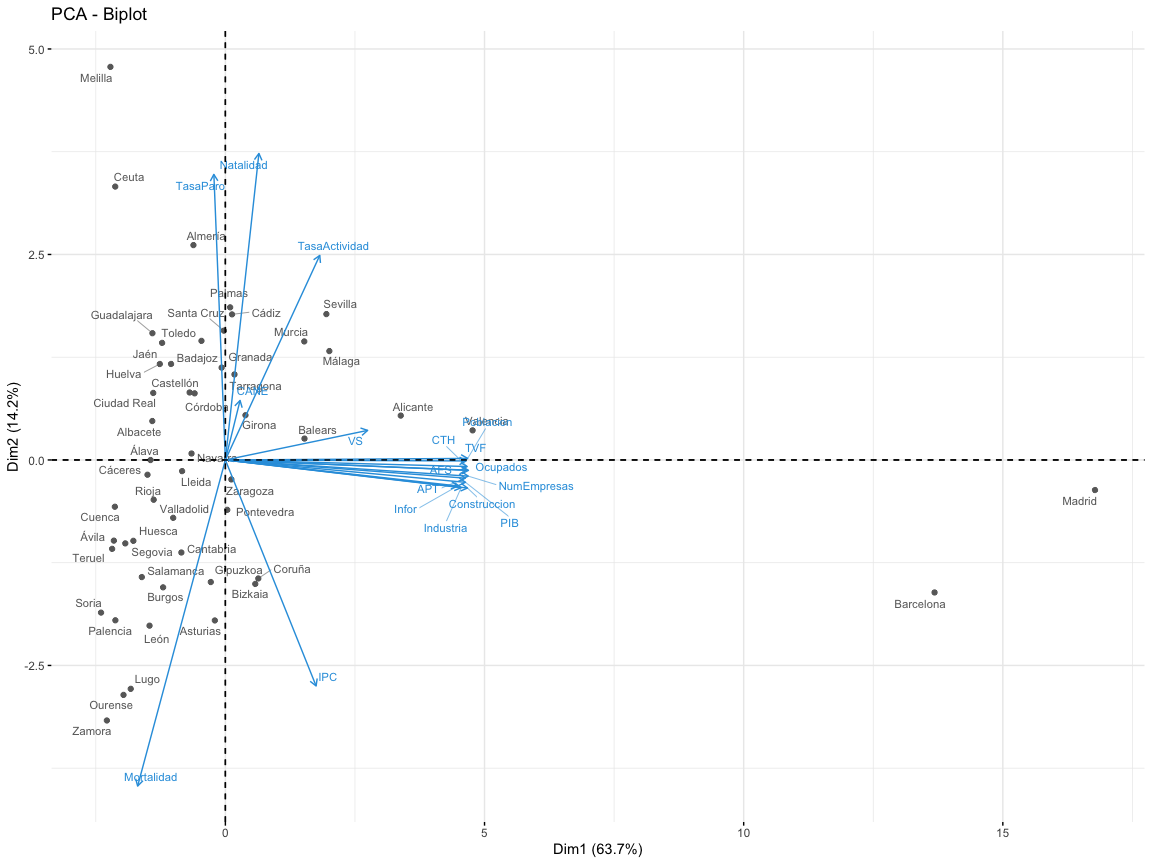
# ¿Que provincias no parecen destacar especialmente en ninguna de las componentes?  
Reduce(intersect,   
 list(rownames(observaciones.1.2$data[(abs(observaciones.1.2$data$x) < 1 & abs(observaciones.1.2$data$y) < 1), ]),  
 rownames(observaciones.1.3$data[(abs(observaciones.1.3$data$x) < 1 & abs(observaciones.1.3$data$y) < 1), ]),  
 rownames(observaciones.2.3$data[(abs(observaciones.2.3$data$x) < 1 & abs(observaciones.2.3$data$y) < 1), ])))

## [1] "Castellón" "Pontevedra" "Zaragoza"

Como podemos comprobar, las provincias de Castellón, Pontevedra y Zaragoza presentan valores en las componentes muy cercanos a 0, por lo que no parecen destacar en ninguna de las componentes, es decir, **su valores se sitúan en torno a la media**.

No obstante, nos resulta mucho más fácil analizar las variables empleando el biplot, ya que nos permite tener las variables originales. Comenzamos con las componentes 1 y 2:

fviz\_pca\_biplot(analisis.comp, repel = TRUE, col.var = "#2E9FDF", col.ind = "#696969", labelsize = 3)



Comenzamos con las variables más relevantes de la **COMPONENTE 1**:

1. **NUM\_EMPRESAS, INDUSTRIA, CONSTRUCCIÓN, INFOR, AFS, APT, POBLACION, OCUPADOS, PIB, TVF**: caben destacar grandes ciudades como **Valencia, pero en especial Barcelona y Madrid**, regiones que se caracterizan por una mayor actividad económica, lo que se traduce en un mayor número de empresas (tanto total como en cada sector); además de un mayor número de habitantes, tasa de ocupación laboral, producto interior bruto y mayor número de viviendas familiares, es decir, **son las provincias con mayor población y desarrollo económico**. Por el contrario, regiones como **Álava, Cáceres, Rioja, Cuenca y Albacete** destacan por su menor actividad empresarial y demográfica.

No obstante, destaca el gran contraste económico entre ambos extremos: **en sentido negativo a los vectores de la primera componente nos encontramos con mayor número de provincias que en sentido positivo**. Por tanto, a la vista de la disposición de los datos podemos asegurar que el nivel de población y de actividad empresarial en España se concentra mayoritariamente en torno a un pequeño grupo de provincias como Murcia, Sevilla, Málaga, Alicante, Madrid o Barcelona, entre otros. En relación al resto, la actividad económica es mucho menor.

1. **VS**: en relación con el número de viviendas secundarias, destacan las **Islas Baleares y especialmente Alicante**, en contraste con provincias de interior como **Cáceres, Rioja o Cuenca**, principalmente. Si nos fijamos en el *biplot*, Valencia se sitúa “a la par” entre las variables de la primera componente, es decir, no solo destaca por una elevada población o por una mayor actividad económica, sino también por un elevado número de viviendas secundarias; en contraste con Madrid o Barcelona, más alejados de la variable *VS*.

**COMPONENTE 2**:

1. **TASA\_PARO, NATALIDAD**: ambas variables presentan una correlación positiva (ángulo agudo), por lo que regiones del sur como **Santa Cruz, Cádiz, Palmas, Almería, y en especial las ciudades Autónomas de Ceuta y Melilla**, no sólo se caracterizan por un mayor porcentaje de desempleo, sino además por altas tasas de natalidad. Por el contrario, provincias del norte peninsular como **Gipuzkoa, Burgos, Bizkaia, Coruña o Asturias** presentan un menor porcentaje de paro y de natalidad. A *grosso modo*, salvo determinadas provincias como Álava, Navarra o Lleida (donde los valores de la componente 2 son más cercanos a cero), **la distribución de las provincias permite comprobar que, por lo general, las regiones del norte peninsular presentan menores tasas de paro y de natalidad, en comparación con las regiones del sur-levante**.
2. **TASA\_ACTIVIDAD**: **Murcia, Málaga y Sevilla** destacan por tasas de actividad elevadas, mientras que provincias como **Palencia, Soria y León** presentan una menor proporción de población activa. Curiosamente, llama la atención la menor correlación que existe entre la tasa de paro y la tasa de actividad (mayor ángulo), es decir, **no por el hecho de tener una mayor tasa de desempleo implica un mayor porcentaje de población activa**. De hecho, si comparamos las provincias de Cádiz (misma dirección que la tasa de paro) y Sevilla (situada más cerca de la tasa de actividad):

knitr::kable(cbind(get\_pca\_ind(analisis.comp)$coord[c("Cádiz", "Sevilla"), c(1,2)],provincias[c("Cádiz", "Sevilla"),c("TasaParo", "TasaActividad")]), digits = 2, caption = "Contribuciones")

Contribuciones

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | TasaParo | TasaActividad |
| Cádiz | 0.13 | 1.77 | 37.18 | 58.22 |
| Sevilla | 1.95 | 1.77 | 31.23 | 59.99 |

En el caso de Cádiz, hay una mayor tasa de paro, mientras que en Sevilla existe una mayor tasa de Actividad, pese a tener un valor prácticamente idéntico en la segunda componente. Sin embargo, el valor de la componente 1 es superior en el caso de Sevilla, dado que el valor del autovector *TasaActividad* es superior al de la *TasaParo* como pudimos comprobar en el apartado a). Esto último implica que Sevilla (al tener una mayor tasa de actividad) presenta un mayor valor de la Componente 1. Por tanto, **de todas las provincias con altas tasas de paro, Murcia, Sevilla y Málaga son las regiones donde más población se encuentra en activo**.

1. **IPC**: curiosamente, provincias como **Bizkaia y Coruña** se caracterizan por altos porcentajes del IPC, es decir, provincias donde la “cesta familiar” se encarece (menor poder adquisitivo), en contraposición con provincias del sur como Almería, pero en especial **Ceuta y Melilla**, donde hay un mayor decremento en los precios y servicios. En el caso de Barcelona, observamos que se sitúa “a la par” entre la primera y la segunda componente (valores tanto positivos en la primera como negativos en la segunda), es decir, **de las grandes provincias, Barcelona destaca por ser la región donde más cara resulta la cesta de la compra**. Además, y al igual que ocurría entre las tasas de paro y actividad, no todas las provincias con elevadas tasas de mortalidad presentan un IPC alto (direcciones diferentes en los vectores). A modo de ejemplo, si comparamos las provincias de Coruña y Salamanca:

knitr::kable(cbind(get\_pca\_ind(analisis.comp)$coord[c("Coruña", "Salamanca"), c(1,2)],provincias[c("Coruña", "Salamanca"),c("Mortalidad", "IPC")]), digits = 2, caption = "Contribuciones")

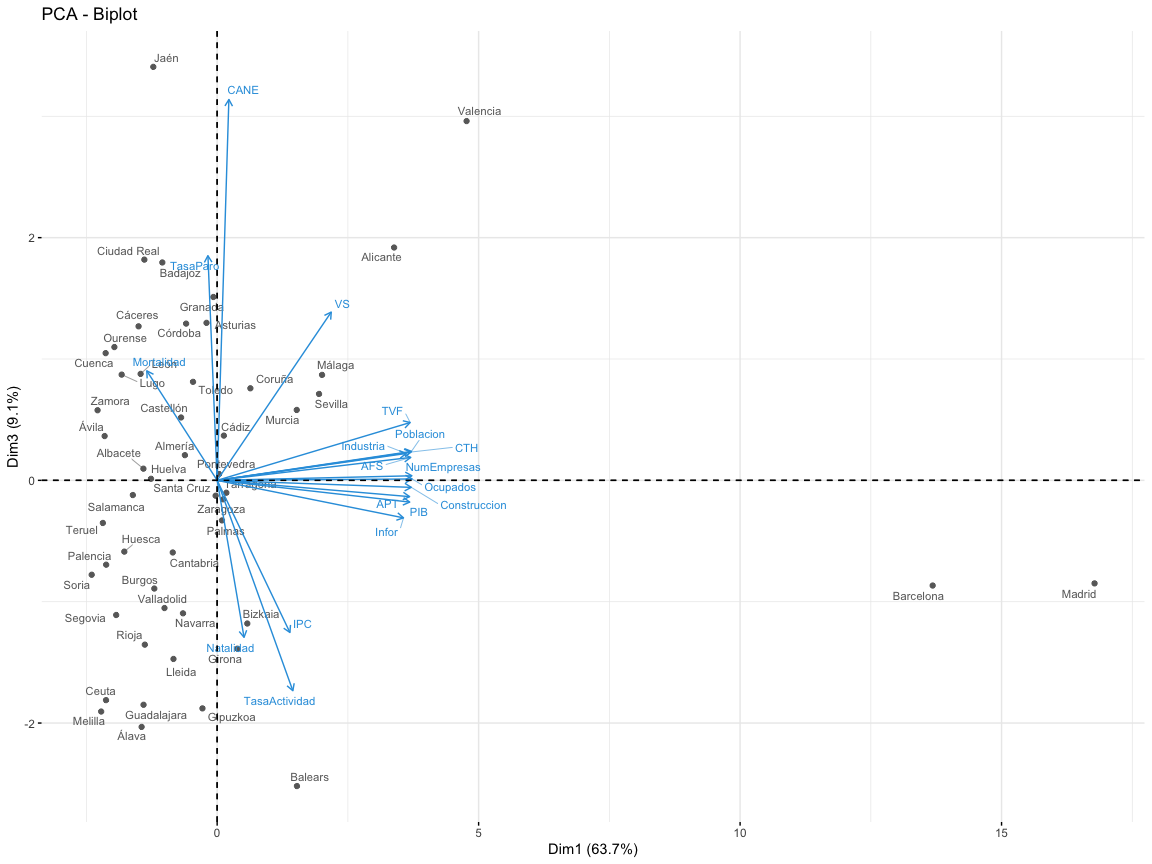
Contribuciones

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Dim.1 | Dim.2 | Mortalidad | IPC |
| Coruña | 0.64 | -1.44 | 10.65 | 102.76 |
| Salamanca | -1.61 | -1.42 | 10.80 | 102.18 |

Pese a tener unos valores de la componente 2 similares, el valor de la componente 1 es superior en el caso de Coruña, dado que el valor del autovector *IPC* es superior al de la *Mortalidad* como pudimos comprobar en el apartado a).

**COMPONENTE 3**:

fviz\_pca\_biplot(analisis.comp, repel = TRUE, axes = c(1,3), col.var = "#2E9FDF", col.ind = "#696969", labelsize = 3, mean.point = FALSE)



1. **CANE**: en relación con las explotaciones agrarias, destacan provincias como **Granada, Badajoz, Ciudad Real, pero en especial Jaén**, donde existe un elevado número de explotaciones agrarias, en comparación con provincias como **Gipuzkoa, Álava, Guadalajara o regiones desérticas como Ceuta y Melilla, e incluso las Islas Baleares**, con un menor número de explotaciones. En conclusión, y en función de la distribución de las provincias, las regiones de Andalucía, Extremadura, Levante, Galicia y Asturias concentran el mayor número de explotaciones agrícolas en España.

Además, podemos observar el menor ángulo que forman la tasa de paro con la variable CANE, por lo que conforme aumenta la tasa de paro en la provincia, generalmente también aumenta el número de explotaciones agrícolas (aunque no es así en todas las provincias, como en el caso de Asturias, con una tasa de desemplo muy baja).

#### Apartado g)

**Si tuviéramos que construir un índice que valore de forma conjunta el desarrollo económico de una provincia, como se podría construir utilizando una combinación lineal de todas las variables. ¿Cuál sería el valor de dicho índice en Madrid? ¿Cual sería su valor en Melilla?**

Dado que el objetivo es medir el desarrollo económico de una provincia, **debemos buscar el o las componentes cuyas variables contengan información principalmente económica**. Por un lado, la primera componente contiene prácticamente toda la información relevante a la actividad económica, incluyendo el número de empresas totales y de cada sector, así como el PIB. Por otro lado, pese a que la segunda componente contega información económica de la provincia como el IPC, se trata de una variable principalmente demográfica, dado que su mayor correlación se concentra en torno a las tasas de Natalidad y Mortalidad. Por último, de la tercera componente podríamos considerar el Censo de Número de Explotaciones Agrícolas como una posible variable económica.

Por tanto, dado que la segunda componente contiene información principalmente demográfica y nuestro objetivo es recopilar información económica, **se ha decidido construir un índice a partir de la primera y tercera componente**. Dado que la primera componente contiene un mayor número de variables, al valor obtenido le asignaremos un peso mayor (85 %), mientras que al valor obtenido en la segunda le asociados una importancia menor, del 15 % (ya que solo representa una variable):

autovectores.t <- data.frame(analisis.comp$svd$V[,c(1,3)], row.names = rownames(var$contrib))  
colnames(autovectores.t) <- c("PC1", "PC3")  
knitr::kable(t(autovectores.t), digits =4,caption = "Autovectores")

Autovectores

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Poblacion | Mortalidad | Natalidad | IPC | NumEmpresas | Industria | Construccion | CTH | Infor | AFS | APT | TasaActividad | TasaParo | Ocupados | PIB | CANE | TVF | VS |
| PC1 | 0.2935 | -0.1063 | 0.0406 | 0.1099 | 0.2942 | 0.2856 | 0.2933 | 0.2929 | 0.2815 | 0.2925 | 0.2907 | 0.1143 | -0.0140 | 0.2944 | 0.2909 | 0.0178 | 0.2916 | 0.1723 |
| PC3 | 0.0499 | 0.1889 | -0.2707 | -0.2623 | 0.0080 | 0.0464 | -0.0121 | 0.0488 | -0.0646 | 0.0396 | -0.0281 | -0.3627 | 0.3874 | 0.0024 | -0.0375 | 0.6567 | 0.0998 | 0.2901 |

Por tanto, el nuevo índice construido presenta la siguiente combinación lineal:

Donde es la tercera componente:

Una vez definido el índice, ¿Qué valor tomará en el caso de Madrid? ¿Y en Melilla?

valores.pca <- get\_pca\_ind(analisis.comp)$coord  
indice.economico <- tibble::tibble("Municipio" = c(rownames(provincias)), "Indice" = valores.pca[,1] \* 0.85 + valores.pca[,3] \* 0.15, "Peso Comp. 1" = valores.pca[,1] \* 0.85, "Peso Comp. 3" = valores.pca[,3] \* 0.15)

Indice económico de cada provincia (ordenado)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Municipio | Indice | Peso Comp. 1 | Peso Comp. 3 | Posicion |
| Madrid | 14.13 | 14.26 | -0.13 | 1 |
| Melilla | -2.17 | -1.89 | -0.29 | 52 |

Podemos observar que **la provincia de Madrid ocupa la primera posición, es decir, se trata de la provincia con mayor peso económico en España**, concentrando el mayor nivel de empresas, PIB y población (junto con viviendas familiares). Por el contrario, la ciudad autónoma de Melilla **ocupa la última posición en el puesto, con un índice negativo**, esto es, la provincia con menor desarrollo económico. Curiosamente, en ambos casos influye negativamente el número de explotaciones, ya que ambas provincias se situaban en sentido negativo a la tercera componente.

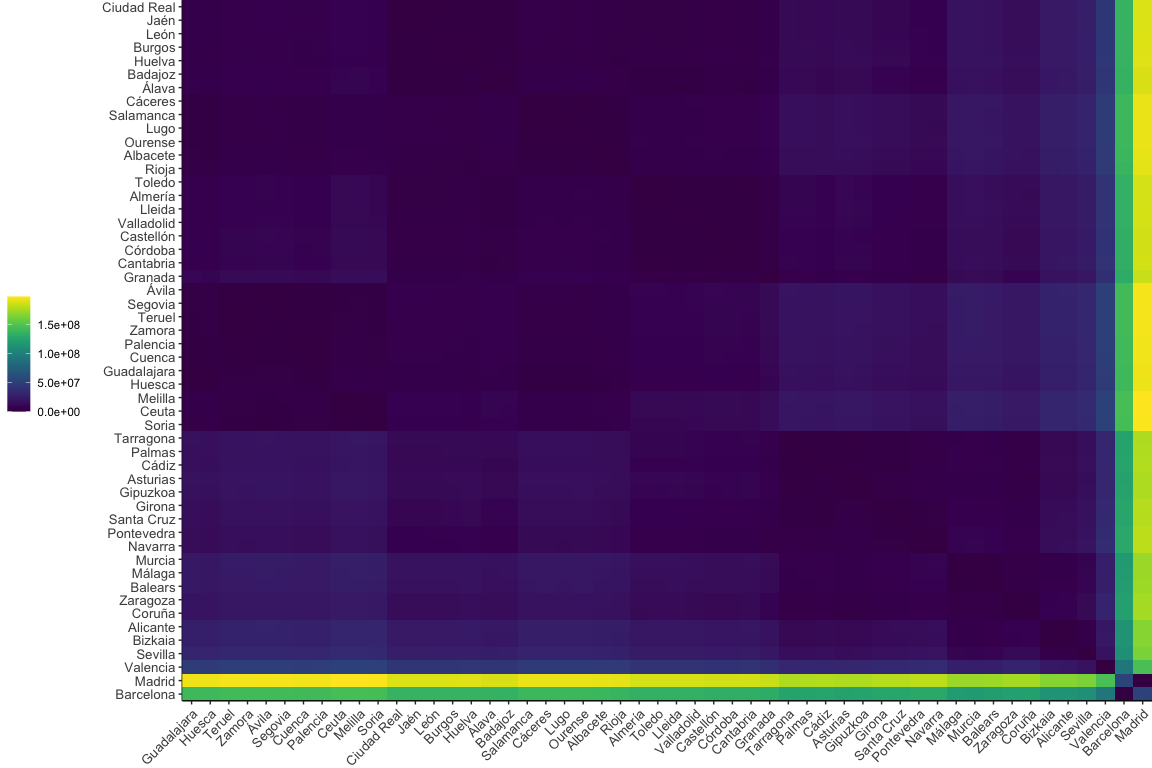
## PARTE 2. CLUSTERING

### Apartado 4

**Representar un mapa de calor de la matriz de datos, estandarizado y sin estandarizar para ver si se detectan inicialmente grupos de provincias.**

Inicialmente, representamos un mapa de calor con la matriz de datos **sin estandarizar**:

matriz.distancias <- dist(provincias, method = "euclidean")  
ggheatmap(as.matrix(matriz.distancias), seriate="mean", show\_dendrogram = c(FALSE, FALSE))

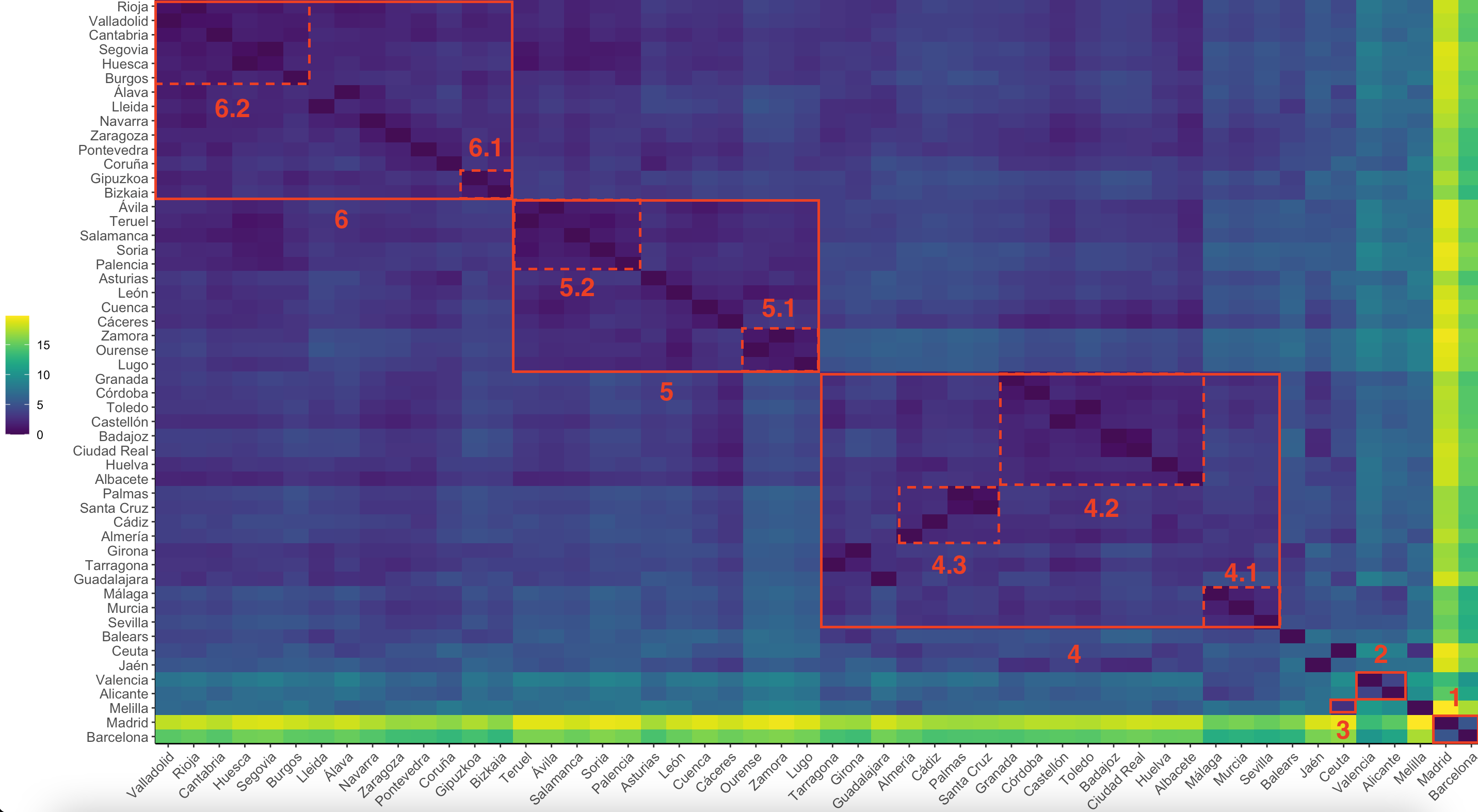


A la vista de la matriz obtenida, el principal grupo que podemos destacar corresponde con las grandes ciudades de Madrid y Barcelona, e incluso con Valencia, cuyas distancias entre si son menores en comparación al resto de provincias. En relación con el resto de observaciones, es posible detectar ciertos grupos en función de la tonalidad del color, pero aun así resulta bastante difícil, **ya que los datos están sin estandarizar**, lo que provoca que las distancias entre los individuos dependan de las unidades en las que están medidas.

Por ello, para reflejar en mejor medida el comportamiento de las variables, **las estandarizamos**:

# Estandarizamos las variables  
provincias\_ST <- scale(provincias)  
matriz.distancias.st <- dist(provincias\_ST, method = "euclidean") # Calculamos la nueva matriz de distancias

ggheatmap(as.matrix(matriz.distancias.st), seriate="mean", show\_dendrogram = c(FALSE, FALSE))



Analizando el nuevo mapa de calor, ya podemos detectar más grupos de provincias que sin estandarizar el conjunto de datos. En general, **nos encontramos con una clara separación entre las provincias de Madrid y Barcelona (1)** con respecto al resto, del mismo modo que observamos en el Análisis de Componentes Principales. Por otro lado, junto con las dos grandes regiones detectamos agrupaciones entre diferentes provincias:

* **Valencia y Alicante**, las cuales se caracterizaban por un alto número de segundas residencias (2), en comparación con el resto de regiones.
* **Las ciudades autónomas de Ceuta y Melilla**, provincias con altas tasas de desempleo y natalidad, además de un menor desarrollo económico (3); por lo que la distancia es mayor al resto de provincias, pudiendo incluso formar un *cluster* propio.
* **Por otro lado, detectamos una gran agrupación de provincias entre las regiones de Tarragona y Sevilla (4), abarcando gran parte del sur-levante peninsular**, incluyendo la mayor parte de Castilla la Mancha, Andalucía y el litoral levantino. A su vez, incluso dentro del mismo grupo podemos detectar subconjuntos con menor distancia entre sí:
  + **4.1.** Málaga, Murcia y Sevilla, regiones caracterizadas por altas tasas de actividad, tal y como comprobamos en el análisis de componentes.
  + **4.2.** Granada, Córdoba, Castellón, Toledo, Badajoz, Ciudad Real, Huelva y Albacete.
  + **4.3.** Las provincias canarias de Palmas y Santa Cruz, junto con provincias andaluzas como Almería o Cádiz.
* A continuación (5), **desde la provincia de Ávila hasta Lugo** (es decir, regiones del norte peninsular) nos encontramos con otro grupo de provincias, del cual también podemos detectar subconjuntos más cercanos entre sí:
  + **5.1.** Las provincias de Ourense, Zamora y Lugo, regiones que, como comprobamos en el análisis de componentes, presentan altas tasas de mortalidad, esto es, valores negativos de la segunda componente.
  + **5.2.** Regiones de interior como Ávila, Teruel, Salamanca, Soria y Palencia.
* Por último, **desde la provincia de Rioja hasta Bizkaia** observamos otra posible agrupación de varios subconjuntos de provincias (6), entre las que destacan Gipuzkoa y Bizkaia (6.1), así como las regiones entre Rioja y Burgos (6.2).

En general, y desde un punto de vista gráfico, **podemos observar un total de seis posibles agrupaciones en el conjunto de datos**. No obstante, dado que *ggheatmap* agrupa las observaciones más próximas, podemos comprobar que las provincias del centro-norte peninsular (Galicia, Castilla y León, Cantábrico, Norte de Extremadura, Aragón o Norte de Cataluña) **presentan generalmente distancias mucho más pequeñas entre sí** que con las provincias del sur-levante peninsular, del mismo modo que ocurría con los valores de las provincias en la segunda componente. Esto último puede suponer que los grupos de provincias 5 y 6 (generalmente regiones del norte) puedan incluso agruparse entre sí, del mismo modo que las provincias del grupo 4 puedan juntarse con las regiones de Alicante-Valencia o Ceuta y Melilla. Es decir, **el número de posibles *clusters* podría incluso reducirse a cuatro o cinco grupos**.

### Apartado 5

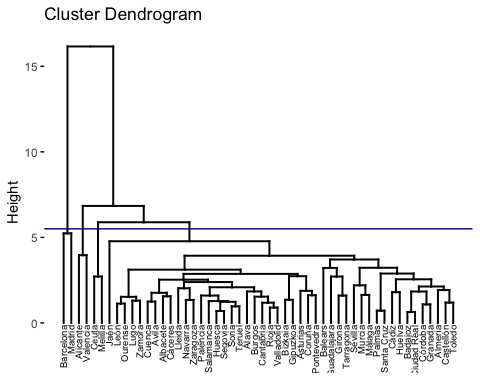
**Realizar un análisis Jerárquico de clusters para determinar si existen grupos de provincias con comportamiento similar.**

Pese a haber observado seis posibles agrupaciones, hemos hecho una observación *a priori*, visualmente, por lo que seguimos sin conocer con exactitud el número de grupos adecuado, aunque tenemos una noción de cómo están agrupados gracias al mapa de calor. Por tanto, una vez calculadas las distancias aplicamos un algoritmo de clasificación jerárquica, con el objetivo de agrupar aquellas provincias cuyas distancias sean menores.

Para ello, debemos elegir el mejor método de agrupación para el conjunto de datos. Dado que los algoritmos de enlace simple y enlace completo generan o bien muchas observaciones o bien conjuntos muy compactos, compararemos los resultados obtenidos mediante enlace medio, distancia entre centroides y distancia de *Ward*:

**ENLACE MEDIO**:

provincias.hc.media <- hclust(matriz.distancias.st, method = "average")  
fviz\_dend(provincias.hc.media, cex = 0.5) + geom\_hline(yintercept = 5.5, col = "darkblue")

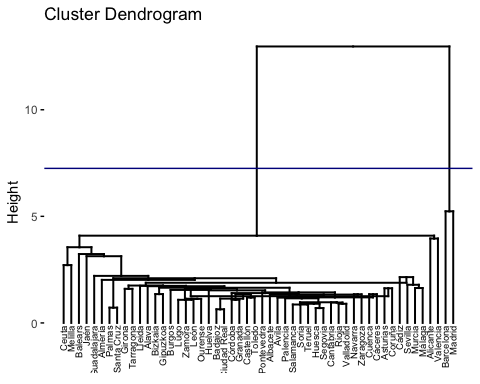


Comenzando con el enlace medio, nos encontramos con cuatro grandes agrupaciones: **las provincias de Madrid y Barcelona, Valencia y Alicante, Ceuta y Melilla, así como el resto de provincias**.

En este último caso nos encontramos ante una agrupación “excesiva” de provincias, dado que, como pudimos observar en el mapa de calor, la distancia media entre el resto de provincias y las grandes ciudades es muy grande, por lo que la media obtenida entre los diferentes grupos de provincias siempre será menor que con Madrid y Barcelona, de forma que se irán agrupando en torno a un mismo conjunto, conformando finalmente un único *cluster*. Además, una agrupación demasiado grande supondría perder información relevante en torno a las provincias concentradas dentro del mismo conjunto, especialmente entre las provincias del norte y sur peninsular.

**DISTANCIA ENTRE CENTROIDES**:

provincias.hc.centroides <- hclust(matriz.distancias.st, method = "centroid")  
fviz\_dend(provincias.hc.centroides, cex = 0.5) + geom\_hline(yintercept = 7.25, col = "darkblue")

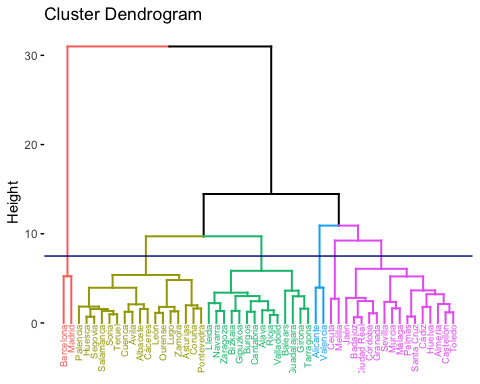


Al igual que con la media, empleando la distancia entre centroides se generan agrupaciones “excesivas”, conformando dos grandes *clusters*: Madrid y Barcelona junto con el resto de provincias.

Por el contrario, la distancia de *Ward* permite minimizar la variabilidad interna de los elementos del nuevo *cluster*, de forma que evitamos agrupar grandes grupos de provincias como en los métodos anteriores, tal y como podemos observar a continuación:

**DISTANCIA DE WARD O DE MÍNIMA VARIANZA**:

provincias.hc.ward <- hclust(matriz.distancias.st, method = "ward.D2")  
fviz\_dend(provincias.hc.ward, cex = 0.5, k = 5, color\_labels\_by\_k = TRUE) + geom\_hline(yintercept = 7.5, col = "darkblue")



#### Apartado a)

**A la vista del dendrograma ¿Cuántos clusters recomendarías?**

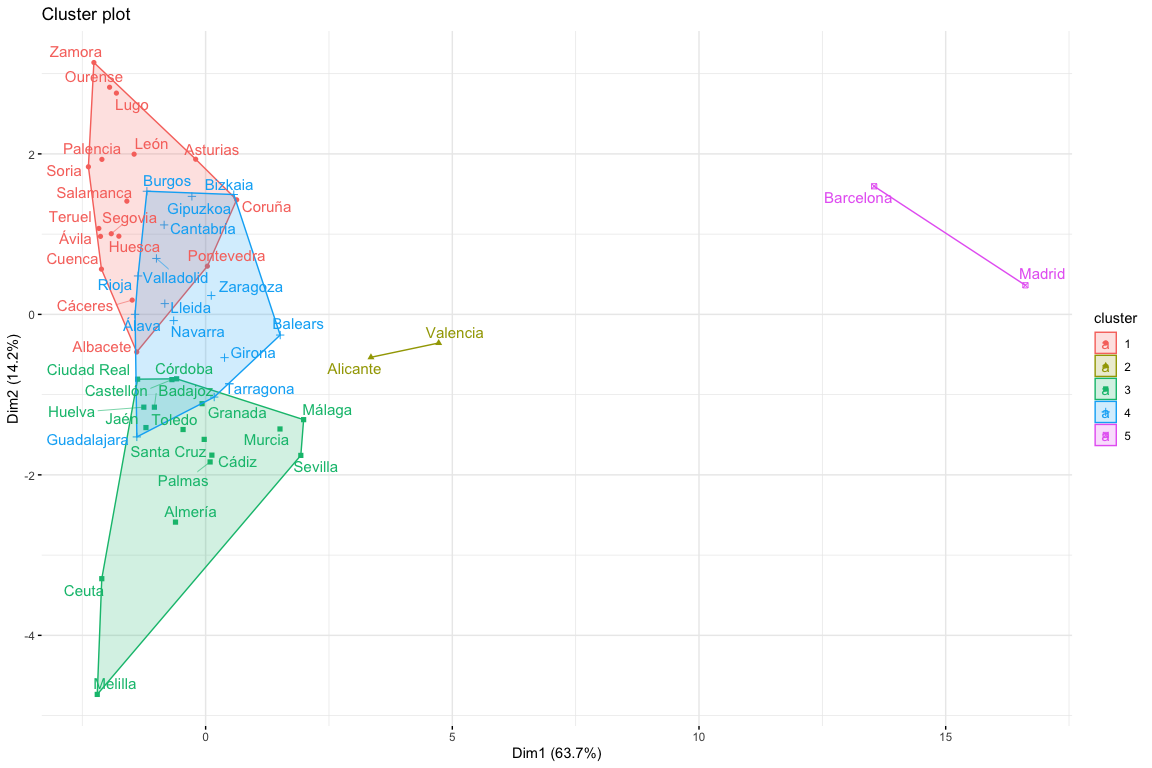
A la vista del último dendograma, **recomendaría escoger 5 grupos**, ya que podemos comprobar como la separación entre grupos es muy similar a la planteada en el mapa de calor: por un lado, las grandes provincias de Madrid y Barcelona, seguido de dos *cluster* (amarillo y verde) conformadas por regiones del norte peninsular que se agrupan en alturas diferentes, por lo que se puede ser interesante tratar ambos grupos por separado. Por otro lado, las regiones de Alicante y Valencia, conformando un *cluster* propio, seguido de provincias del sur-levante peninsular, incluyendo a Ceuta y Melilla (morado).

#### Apartado b)

**Representar los individuos agrupados según el número de clusters elegido.**

**COMPONENTES 1 Y 2**

# Cortamos el arbol en 5 clusters  
grupos.provincias <- cutree(provincias.hc.ward, k = 5)  
fviz\_cluster(list(data = provincias\_ST, cluster = grupos.provincias),  
 ellipse.type = "convex", repel = TRUE, show.clust.cent = FALSE, ggtheme = theme\_minimal())



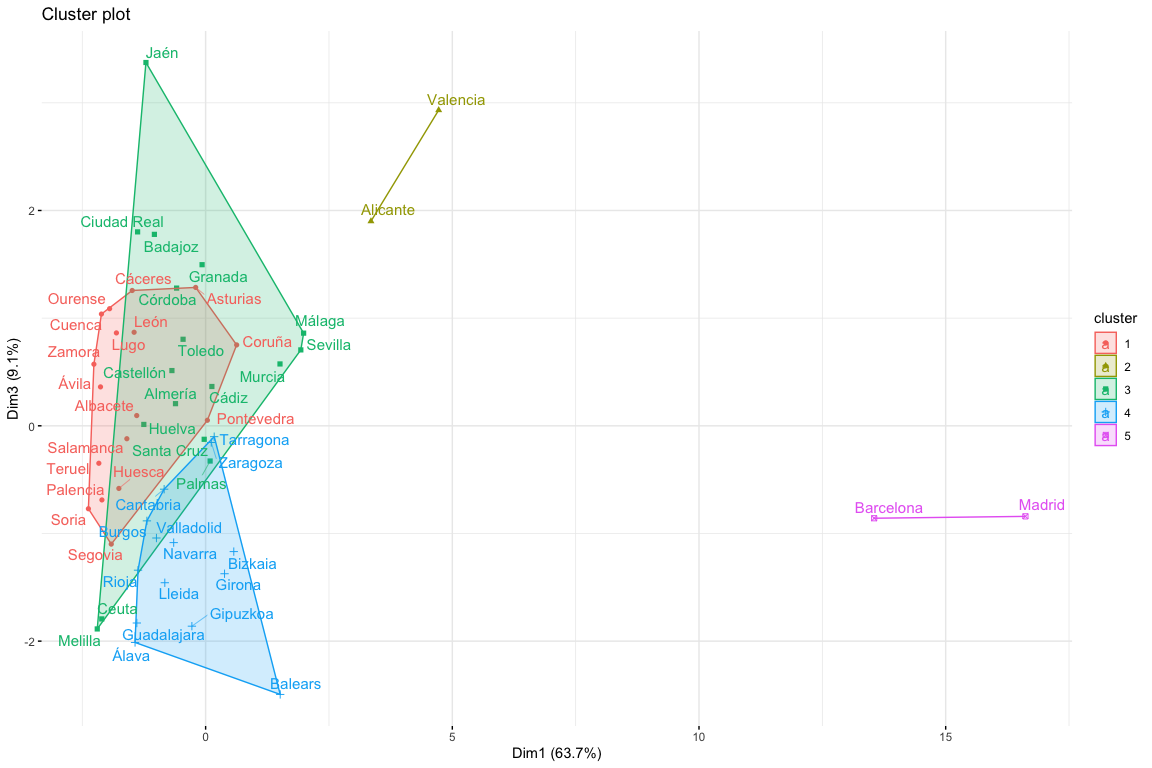
En relación con la primera componente, y como pudimos comprobar previamente, en sentido positivo **destacan los *clusters* formados por las provincias de Alicante y Valencia, pero en especial Madrid y Barcelona**, regiones con una mayor actividad económica, un elevado número de población, así como un alto porcentaje de viviendas familiares y tasa de ocupación. Por otro lado, la gran mayoría de las provincias de los *clusters* 1, 3 y 4 presentan un comportamiento similar, **sitúandose en sentido negativo**, lo que se traduce en grupos de menor población y actividad económica, especialmente en regiones como Zamora o Soria (grupo 1), Ceuta y Melilla (grupo 3), o Guadalajara (grupo 4). No obstante, caben destcar pequeños subconjuntos de provincias en dichos *cluster* con valores positivos en la primera componente como Baleares, Coruña, Málaga o Sevilla, es decir, se tratan de provincias que destacan, dentro de cada *cluster*, por una mayor población y desarrollo económico con respecto.

En relación con la segunda componente, dado que el eje está invertido, debemos interpretar los valores justo al revés. Es decir: \* **La gran mayoría de las provincias del primer *cluster*, junto con Madrid y Barcelona, presentan valores positivos**, esto es, se caracterizan por altas de mortalidad y mayor valor del IPC, lo que se traduce en un elevado coste en la “cesta de la compra”, especialmente en regiones de Castilla-León y Galicia como Zamora, Ourense o Lugo, entre otras.

* En contraposición, **tanto las provincias de Alicante y Valencia como del tercer *cluster* se sitúan en sentido negativo de la segunda componente**, es decir, grupos caracterizados por altas tasas de natalidad y de actividad, aunque con un elevado porcentaje de desempleo, especialmente en las regiones de Almería, Ceuta y Melilla, donde se intensifican los valores negativos.
* Por último, en relación con el cuarto *cluster* no parece estar tan claro cual es la distribución de las provincias, con regiones en sentido tanto positivo como negativo de la primera o segunda componente, **aunque aparentemente se encuentran más orientadas a valores positivos (mayor mortalidad e IPC)**. Sin embargo, a simple vista muchas de las provincias abarcan valores de las componentes muy cercanos a 0 (como es el caso de Zaragoza), por lo que puede tratarse de provincias que no destaquen en ninguna componente en particular (valores en torno a la media).

**COMPONENTE 3**

fviz\_cluster(list(data = provincias\_ST, cluster = grupos.provincias),  
 ellipse.type = "convex", repel = TRUE, show.clust.cent = FALSE, ggtheme = theme\_minimal(), axes = c(1,3))



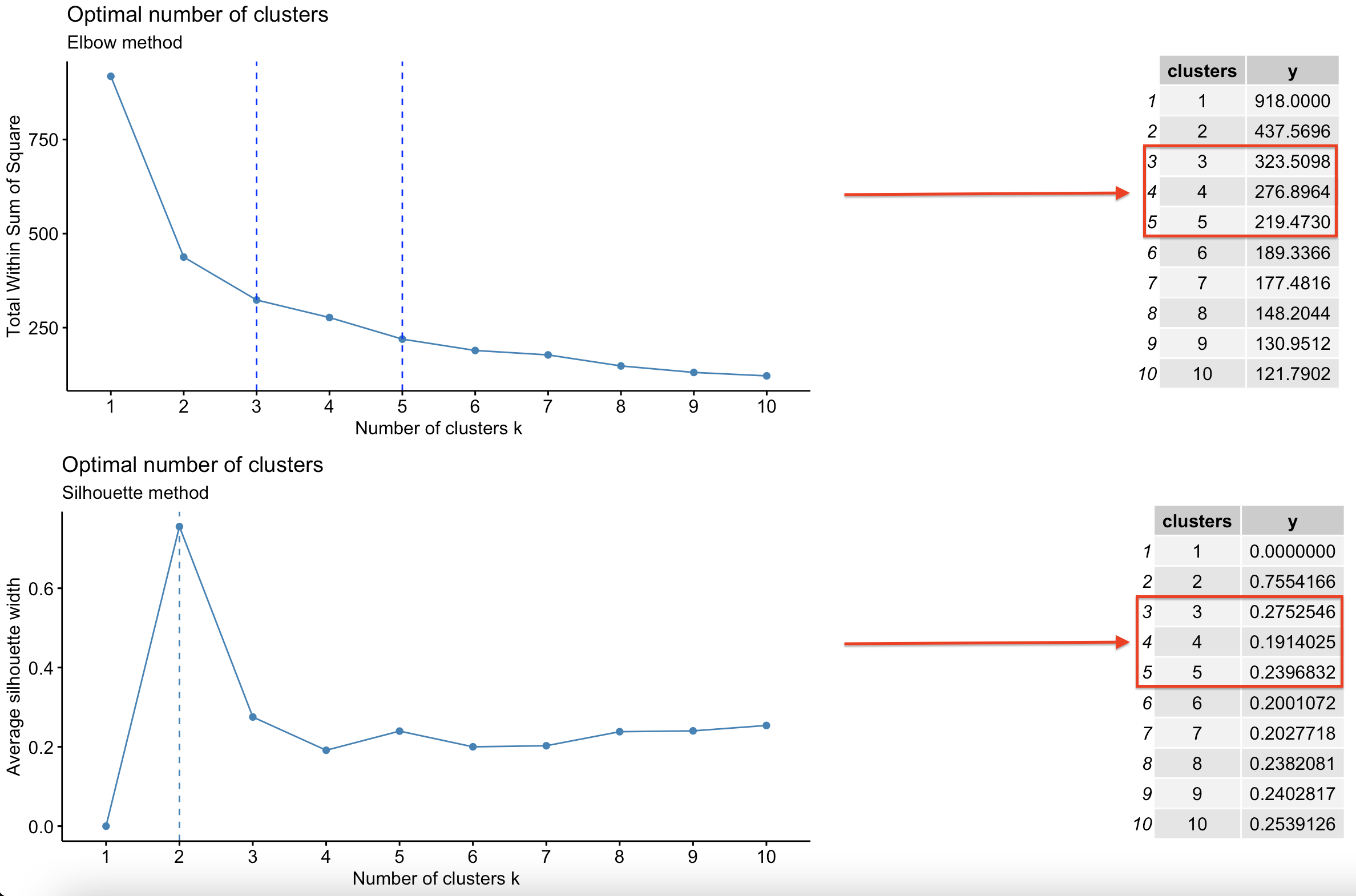
En relación con la tercera componente, **existe una clara separación entre el segundo y quinto *cluster*, con un mayor número de explotaciones agrarias en las provincias de Alicante y Valencia**. Por el contrario, el *cluster* 4, pese a no tener una distribución tan clara en las componentes 1 y 2, presenta valores negativos en la tercera componente, destacando aquellas provincias con un menor número de explotaciones agrícolas. En relación al primer y tercer *cluter*, **la separación no es tan evidente**, especialmente en el primero donde existe un contraste entre provincias con un elevado número de explotaciones (sentido positivo), en contraposición con regiones como Segovia o Soria (sentido negativo). Por otro lado, las provincias del tercer grupo están orientadas mayoritariamente a valores positivos, salvo las provincias canarias de Santa Cruz y Palmas, junto con las ciudades de Ceuta y Melilla, donde el número de explotaciones agrícolas es menor.

#### Apartado c)

**¿Qué número óptimo de clusters nos indican los criterios Silhoutte y de Elbow?**

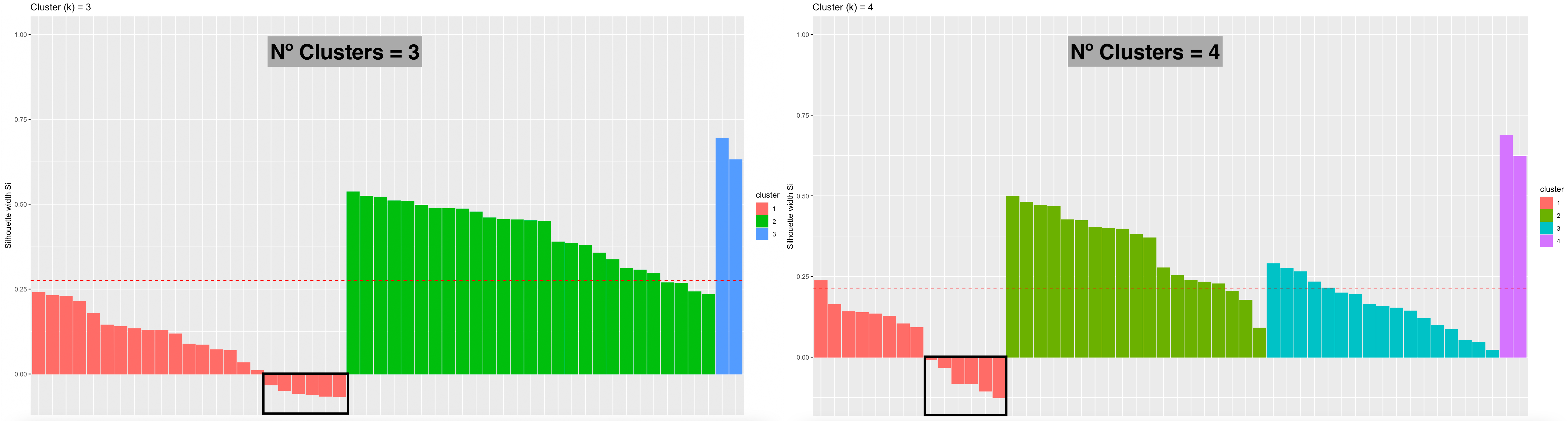
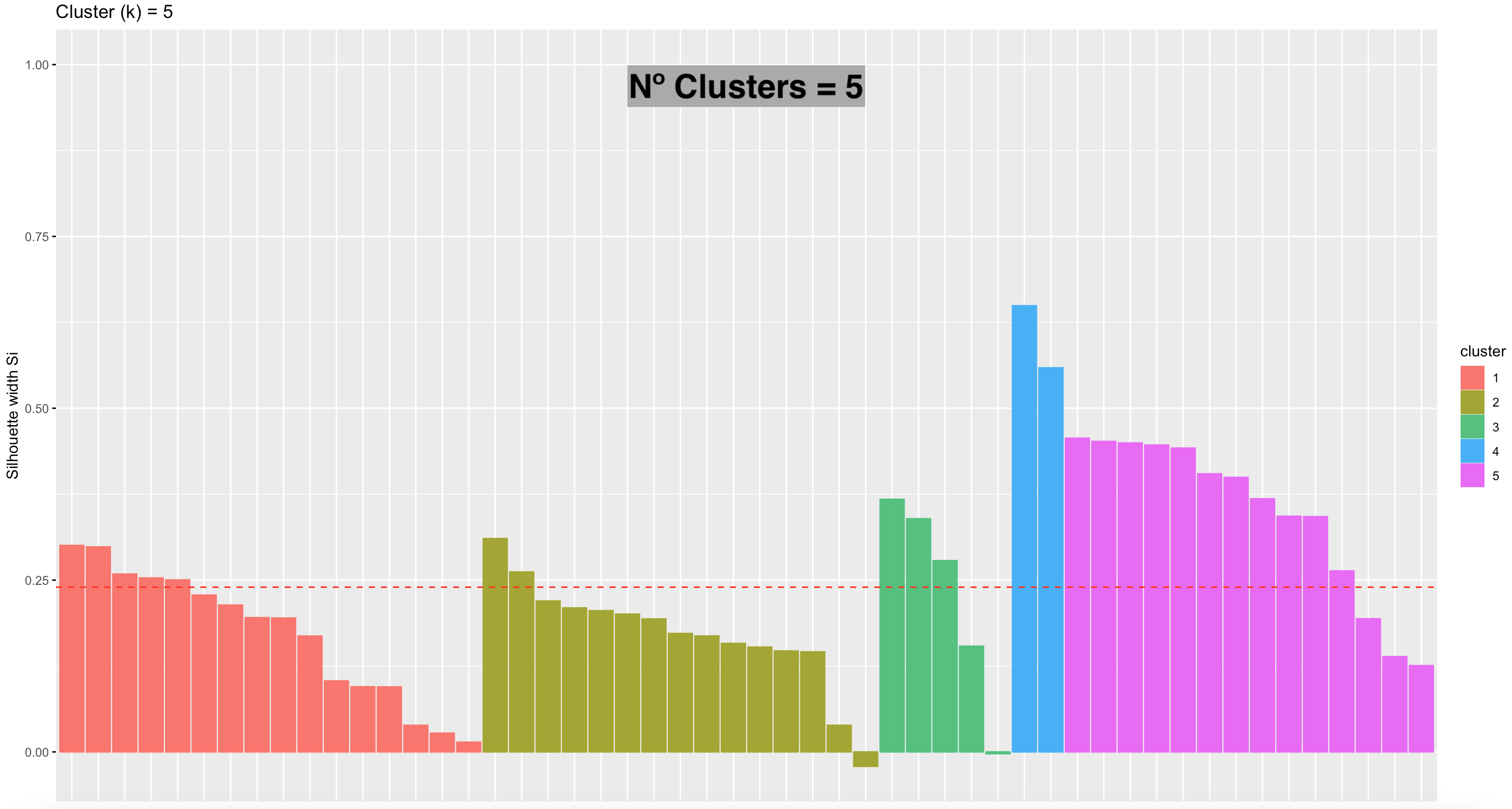
# Nos aseguramos que tenemos la misma semilla  
RNGkind(sample.kind = "Rejection")  
set.seed(1234)

fviz\_nbclust(provincias\_ST, kmeans, method = "wss", ) + labs(subtitle = "Elbow method")  
fviz\_nbclust(provincias\_ST, kmeans, method = "silhouette") + labs(subtitle = "Silhouette method")



Comenzando con la salida obtenida en el criterio de *Elbow*, podríamos considerar como número óptimo de *cluster* entre 3 y 5, dado que la variabilidad total continúa reduciéndose significativamente: de 3 a 4 *cluster* se reduce en 47, mientras que de 4 a 5 en 57. Por el contrario, a partir de seis *cluster* la reducción ya no es tan significativa.

Sin embargo, el criterio de *Silhouette* recomienda emplear tan solo dos grupos, por lo que no parece dejar tan claro cual puede ser la agrupación óptima. Por tanto, ¿Qué número de *cluster* elegimos? Por un lado, emplear tan solo dos grupos no es la mejor alternativa, dado que según el criterio de *Elbow* la varibilidad continúa reduciéndose significativamente al aumentar el número de grupos: de 2 a 3 *cluster* se reduce de 437 a 323, es decir, disminuye en más de 100; por lo que **el número óptimo estará comprendido entre 3 y 5**. Para decidir el número de grupos, realizamos la agrupación con 3,4 y 5 *cluster*, comparando la calidad de cada modelo mediante los valores de anchura o *silhouette* obtenidos:

Como podemos analizar en las gráficas anteriores, empleando 3 y 4 *cluster* observamos valores de siluetas inferiores a 1 en el primer grupo, un indicativo de una mala agrupación de las provincias. Por el contrario, con 5 *cluster* el número siluetas negativas se reduce de 6 a tan solo 2 observaciones. Además, si analizamos la silueta media en cada *cluster*:

## Numero de cluster: 3   
## cluster size ave.sil.width  
## 1 1 23 0.08  
## 2 2 27 0.41  
## 3 3 2 0.66  
## Numero de cluster: 4   
## cluster size ave.sil.width  
## 1 1 14 0.05  
## 2 2 19 0.34  
## 3 3 17 0.16  
## 4 4 2 0.66  
## Numero de cluster: 5   
## cluster size ave.sil.width  
## 1 1 16 0.17  
## 2 2 15 0.17  
## 3 3 5 0.23  
## 4 4 2 0.60  
## 5 5 14 0.34

Podemos observar que la media en el *cluster* 1 en las dos primeras opciones presenta un valor medio cercano a 0, por lo que muchas de las observaciones (en especial aquellas cuyos valores de silueta son negativos), podrían pertenecer al *cluster* actual o a otro cercano a él. Por el contrario, empleando 5 *cluster* la media mejora significativamente, con un valor mínimo en la silueta de 0.17. Por tanto, realizamos el agrupamiento no jerárquico con 5 *cluster*.

#### Apartado d)

**Con el número de clusters decidido en el apartado anterior realizar un agrupamiento no jerárquico.**

##### 1. **Relacionar la posición de cada cluster en el plano con lo que representa cada componente principal.**

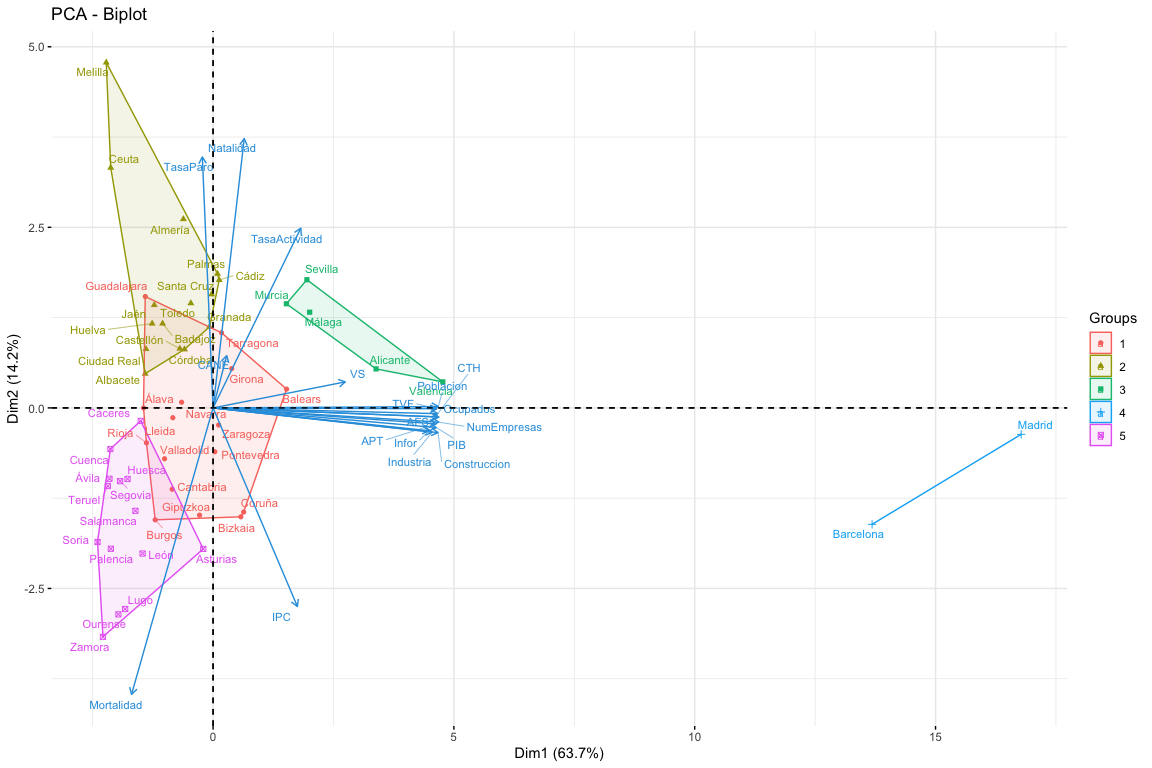
Inicialmente, representamos los *cluster* formados en los planos de las Componentes Principales.

**COMPONENTES 1 Y 2**:

set.seed(1234)  
RNGkind(sample.kind = "Rejection")  
km.res <- kmeans(provincias\_ST, 5)

Con el objetivo de visualizar en mejor medida la relación entre los *cluster* y las Componentes Principales, en lugar de la función *fviz\_clust* emplearemos el gráfico *biplot* con las direcciones originales de las variables:

fviz\_pca\_biplot(analisis.comp, habillage = provincias$cluster, repel = TRUE, col.var = "#2E9FDF", col.ind = "#696969", labelsize = 3, addEllipses = T, ellipse.type = "convex", mean.point = FALSE)



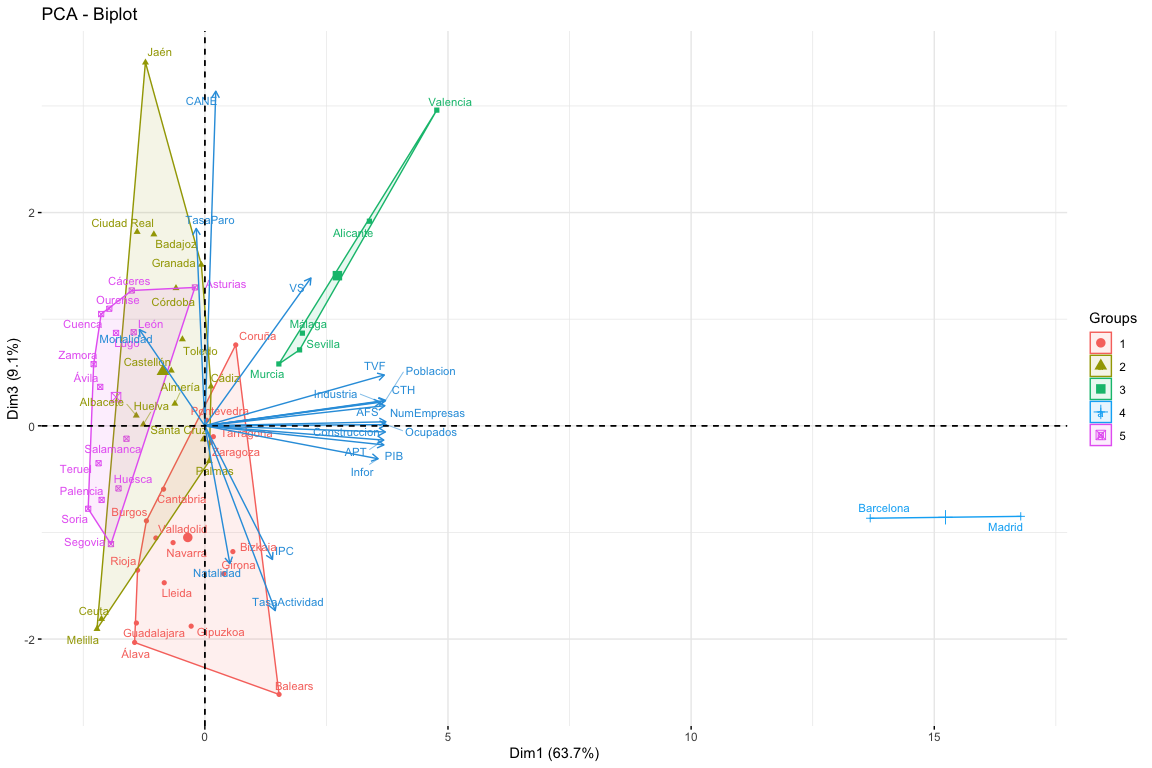
Con el agrupamiento no jerárquico, nos encontramos con una agrupación muy similar a la que pudimos observar en el agrupamiento jerárquico, salvo con ciertas diferencias en el agrupamiento de provincias. A modo de ejemplo, Málaga, Murcia y Sevilla se situaba junto con el resto de regiones del sur peninsular, y ahora conforman un único *cluster* con las provincias de Alicante y Valencia. No obstante, la disposición que presentan los grupos es bastante similar:

* *Cluster* 2: **presentan valores negativos en la primera componente como positivos en la segunda**, a excepción de las provincias de Cádiz y Palmas, con valores ligeramente positivos en la componente 1. Por lo general, el *cluster* representa:
  + **Provincias con un bajo desarrollo económico, además de una menor población y número de viviendas**.
  + En comparación con el *cluster* 5, representa aquellas regiones con **altos porcentajes de natalidad, aunque con una elevada tasa de desempleo y de actividad**.
  + Por el contrario, **presenta una menor tasa de mortalidad, además de valores en el IPC bajos**, es decir, representa aquellas provincias con un mayor abaratamiento en el precio de la cesta de la compra.
* *Cluster* 3: **al situarse en valores positivos tanto en la primera como en la segunda componente**:
  + A diferencia de los *cluster* 2 y 5, **representa a provincias con mayor población, número de empresas y viviendas, especialmente en relación a las segundas residencias**, concretamente en regiones como Alicante o Valencia, los cuales ya pudimos estudiar en el análisis de componentes.
  + No obstante, los valores positivos en la componente 2 muestran, además, **provincias con altas tasas de natalidad, desempleo y población activa**.
* *Cluster* 4: **presenta valores positivos en la primera componente, así como negativos en la segunda**, por lo que representa:
  + Grandes provincias como Madrid y Barcelona, **regiones muy pobladas, con una gran actividad económica y número de viviendas**.
  + Aunque con **mayor porcentaje de mortalidad y coste de vida** (especialmente en Barcelona, con un menor valor en la segunda componente).
* *Cluster* 5: **presentan valores negativos tanto en la primera como en la segunda componente**, es decir:
  + Representa un conjunto de provincias con **mayor despoblación y bajo desarrollo económico**.
  + Por el contrario, y a diferencia del segundo *cluster*, se tratan de provincias con **elevadas tasas de mortalidad y altos porcentajes en el IPC, es decir, provincias donde el precio de la cesta familiar se encarece**.

Por último, ¿Qué ocurre con el primer *cluster*? **A diferencia del resto de grupos, sus provincias no se sitúan en valores únicamente positivos o negativos en las componentes**. A modo de ejemplo, Coruña presenta un valor positivo en la primera componente y negativo en la segunda, lo que indica una mayor población, desarrollo económico, además de una mayor tasa de mortalidad e IPC; en comparación a lo que ocurre con la provincia de Guadalajara. No obstante, la mayor parte de las provincias se sitúan principalmente en el sentido negativo de la segunda componente, concretamente en direcciones opuestas a las tasas de paro y natalidad, aunque visualmente no parece estar tan claro qué grupo de provincias representa.

**COMPONENTE 3**:

fviz\_pca\_biplot(analisis.comp, habillage = provincias$cluster, axes = c(1,3), repel = TRUE, col.var = "#2E9FDF", col.ind = "#696969", labelsize = 3, addEllipses = T, ellipse.type = "convex")



En relación al tercer *cluster*, nos encontramos con valores positivos en la tercera componente, es decir, **representa aquellas provincias con un elevado número de explotaciones agrícolas** (especialmente en el caso de Valencia). Por el contrario, **tanto las regiones de Madrid y Barcelona como el *cluster* 1 (a excepción de Coruña) representan totalmente lo opuesto**: provincias con un menor número de explotaciones agrarias.

Por último, en relación con los *cluster* 2 y 5, presentan valores tanto positivos como negativos en la tercera componente, por lo que **no destacan en cuanto al número de explotaciones agrarias se refiere**, en especial con el *cluster* 5, donde prácticamente una mitad de las provincias presentan valores positivos mientras que la otra mitad se sitúan en sentido negativo. En el caso del segundo *cluster*, la dispersión es incluso mayor, con extremos negativos como Ceuta y Melilla, así como extremos positivos como Jaen.

#### Apartado e)

**Explicar las provincias que forman cada uno de los clusters y comentar cuales son las características socioeconómicas que las hacen pertenecer a dicho cluster.**

Un vez realizado el proceso de *clustering*, mostramos las provincias asociadas a cada grupo:

provincias.ordenado <- sort(km.res$cluster)



Para analizar las características socioeconomicas de cada grupo de provincias, estudiamos los valores medios de las componentes en cada *cluster*. Comenzamos con el índice económico de la primera componente:

knitr::kable(EsT\_Clus\_Comp1, digits =2,caption = "Valores medios de la Componente 1 (Población + Número de Empresas + Viviendas)")

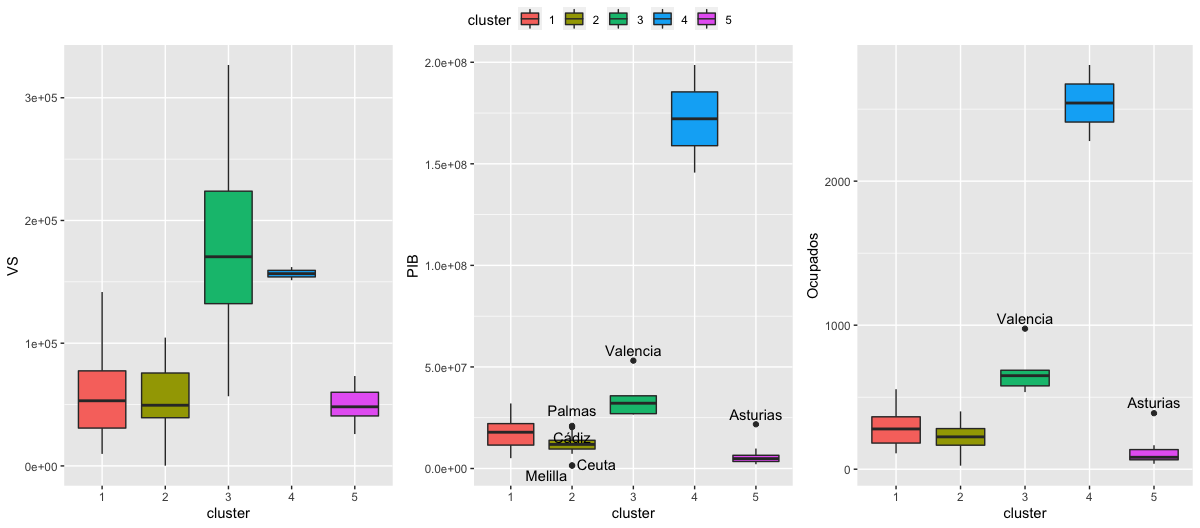
Valores medios de la Componente 1 (Población + Número de Empresas + Viviendas)

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group.1 | Poblacion | Ocupados | PIB | NumEmpresas | Industria | Construccion | CTH | Infor | AFS | APT | TVF | VS |
| 1 | 689452.6 | 285.29 | 17092150 | 47950.69 | 3259.56 | 7048.00 | 18535.50 | 674.50 | 1022.50 | 7759.56 | 380657.9 | 58739.12 |
| 2 | 667162.8 | 225.63 | 11676712 | 38314.87 | 2500.93 | 4507.07 | 16988.93 | 412.20 | 875.47 | 5778.47 | 358378.8 | 54634.47 |
| 3 | 1889495.4 | 685.20 | 34927844 | 122866.60 | 7704.60 | 14389.00 | 49801.60 | 1839.40 | 2955.80 | 21007.80 | 1056591.8 | 182003.60 |
| 4 | 5989112.0 | 2542.35 | 172165306 | 474865.50 | 25012.00 | 55205.50 | 157055.00 | 15096.00 | 10593.00 | 105424.00 | 2748888.0 | 156678.50 |
| 5 | 307639.7 | 113.49 | 6193759 | 20062.86 | 1413.57 | 3079.93 | 8573.07 | 178.14 | 443.93 | 2690.36 | 211548.5 | 49917.71 |

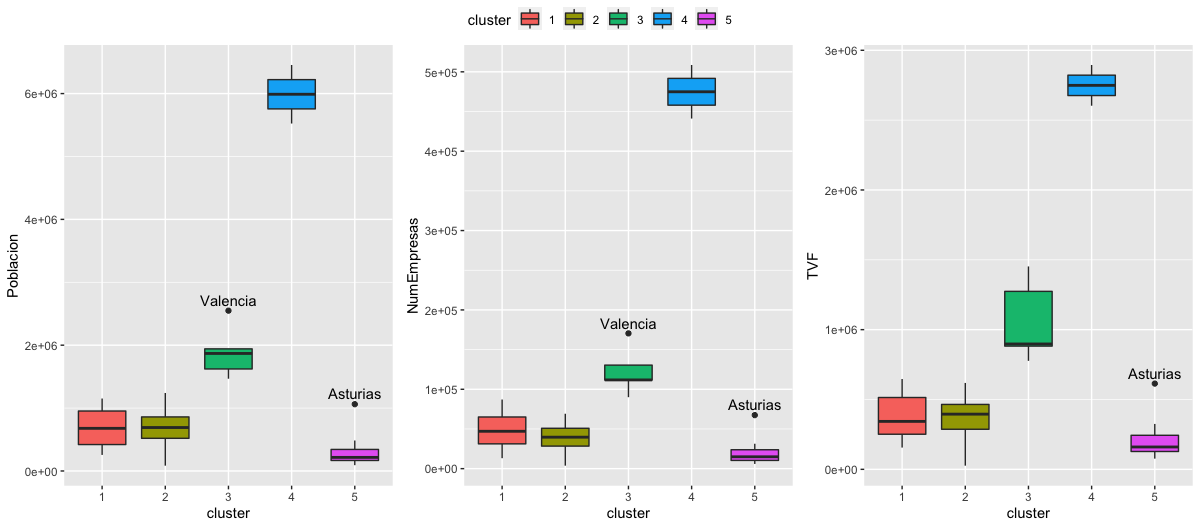
Comenzando con las características económicas de la primera componente ¿Qué diferencias existen entre cada grupo de provincias? Por un lado, **los *cluster* 3 y 4 concentran tanto la mayor parte de la población como el número de empresas y viviendas**, es decir, suponen aquellas regiones con mayor actividad económica en España, desde grandes provincias como Barcelona o Madrid, hasta Alicante, Murcia, Málaga, Sevilla y Valencia; en contraste con las provincias de los *cluster* 2 y 5 (entornos más rurales), que no destacan por su actividad económica, además de tener una población media mucho más pequeña.

No obstante, entre las provincias de los *cluster* 3 y 4 destacan dos diferencias fundamentales:

* **Mayor número de segundas residencias en las provincias del *cluster* 3**, desde regiones como Sevilla con algo más de 56.000 viviendas secundarias hasta provincias como Valencia o Alicante con más de 200.000, superando incluso a la media.
* **Mayor tasa de ocupación en el *cluster* 4 además de concentrar el PIB más alto**, lo que refleja que las provincias que lo componen (Madrid y Barcelona), concentran gran parte de la actividad económica y financiera en España, tal y como podemos comprobar en los diagramas de caja y bigotes:



Sin embargo, ¿Qué ocurre con las provincias del primer *cluster*? Si analizamos los valores medios de la tabla, se tratan de provincias que no destacan ni en valores muy bajos de población, empresas o viviendas, dado que sus valores medios son superiores a los *cluster* 2 y 5; pero tampoco se consideran demasiado altos, en comparación con las provincias de los grupos 3 y 4. Por otro lado, si nos fijamos en el diagrama de cajas, la dispersión de los valores originales tanto en población como número de empresas o viviendas es muy similar con respecto a los *cluster* 2 y 5, aunque ligeramente superior:



Es decir, **las provincias del primer *cluster* no destaca por una mayor o menor actividad económica, sino que los valores de las provincias de la componente 1 se distribuyen en torno a los valores medios**.

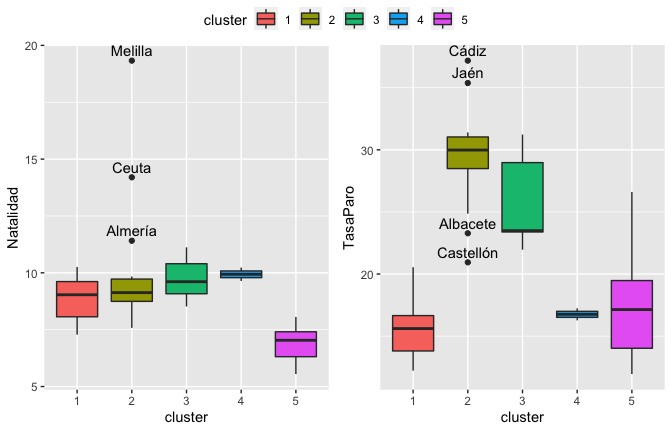
**¿Y en relación con las variables de la Segunda Componente?**

Valores medios de la Componente 2 (Natalidad + Mortalidad + Desempleo + Actividad + IPC)

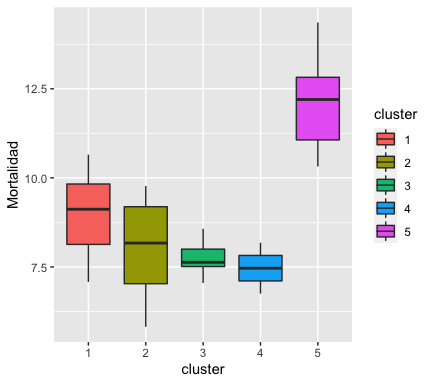
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Group.1 | Natalidad | TasaParo | TasaActividad | Mortalidad | IPC |
| 1 | 8.91 | 15.83 | 59.57 | 8.97 | 102.91 |
| 2 | 10.09 | 29.53 | 58.33 | 7.99 | 101.58 |
| 3 | 9.75 | 25.81 | 59.52 | 7.75 | 102.43 |
| 4 | 9.94 | 16.75 | 62.86 | 7.46 | 103.73 |
| 5 | 6.94 | 17.30 | 54.02 | 12.18 | 102.31 |

Por un lado, los *cluster* 2 y 5 se caracterizan por provincias con menor población y actividad económica. Sin embargo, existe un diferencia significativa entre ambos grupos:

* Por un lado, **las provincias pertenecientes al *cluster* 2 no solo corresponden con las regiones de mayor tasa de natalidad (superior al 10 % medio), sino que además sus provincias concentran las mayores de tasas de paro en comparación con el resto de grupos**, con un porcentaje medio muy cercano al 30 %; incluso (como podemos observar en el diagrama de cajas) provincias como Cádiz o Jaen presentan valores de desempleo en torno al 40 %:

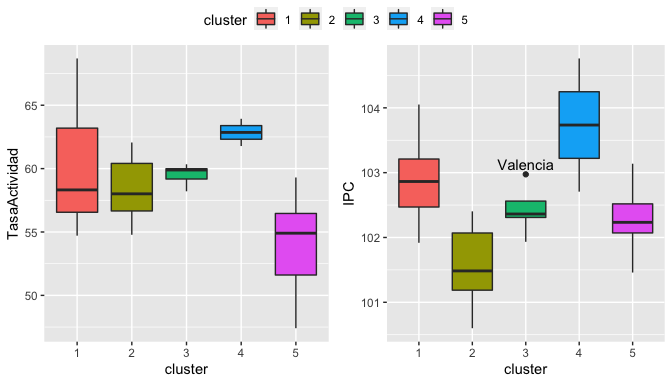


* Por el contrario, **al *cluster* 5 pertenecen las provincias con las tasas de natalidad mas bajas** (inferiores al 7 %) **y las tasas de mortalidad más altas**, superando ampliamente al resto de regiones, con un 12 % medio.



En relación a los *cluster* 3 y 4, pese a concentrar la mayor parte de la población y actividad económica, también presentan diferencias:

* Por un lado, las grandes provincias de Madrid y Barcelona presentan un IPC y una tasa de actividad media mucho más altas, en comparación con el cuarto *cluster*, donde las provincias que lo agrupan presentan una mayor tasa de desempleo. Es decir, en relación con las provincias más pobladas y de mayor actividad económica, **Madrid y Barcelona se caracterizan por un mayor coste de vida, así como un mayor porcentaje de población activa. Por el contrario, y pese a tener menos población, las provincias del *cluster* 3 presentan un mayor desempleo**, con un 25.81 % frente a un 16.75 %:



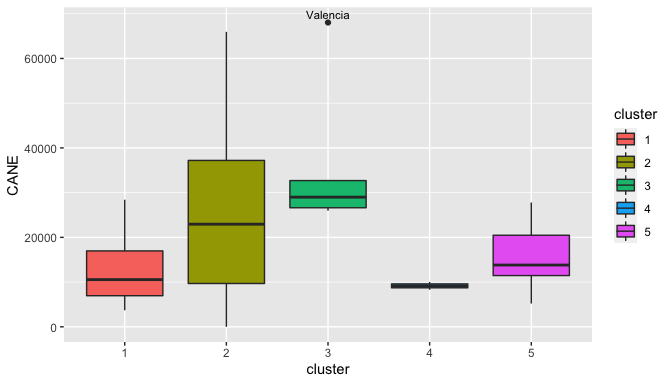
* ¿Y las provincias del *cluster* 1? Al igual que en el *cluster* 5, las provincias que lo componen se caracterizan por bajas tasas de natalidad y elevadas porcentajes de mortalidad (8.91 y 8.97 %, respetivamente). No obstante, **la diferencia fundamental reside en que las provincias del *cluster* 1 concentran las tasas de paro más bajas de todos los grupos** (alrededor del 15 %), ya que, como pudimos observar en los gráficos *biplot*, la gran parte de sus provincias se situan en sentido opuesto al vector *tasaParo*. Además, si nos fijamos en el diagrama de cajas con los valores de desempleo, las provincias del *cluster* 1 presentan una dispersión mucho menor en comparación a las provincias del *cluster* 5, también situadas en sentido negativo de la componente.

Por otro lado, el valor medio del IPC en las provincias del *cluster* 1 es mayor en comparación a las regiones del *cluster* 5, es decir, **de las provincias que presentan altas tasas de mortalidad, las pertenecientes al *cluster* 1 son las que presentan los valores de desempleo más bajos y donde el coste de vida es mayor**.

¿Y en relación con el número de explotaciones agrícolas?

Valores medios de la Componente 3

|  |  |
| --- | --- |
| Grupo | CANE |
| 1 | 13214.81 |
| 2 | 24147.80 |
| 3 | 36462.20 |
| 4 | 9156.00 |
| 5 | 15394.21 |

 En relación con el número de explotaciones agrícolas, **destacan principalmente las provincias del tercer *cluster*, donde el número medio de explotaciones es mayor**, en especial con la región de Valencia, superando más de 60.000 cultivos. En contraposición, las provincias pertenecientes al *cluster* 2 destacan por su elevado contraste entre provincias con más 60.000 explotaciones, en comparación con regiones sin prácticamente ningún registro en el censo agrario. Por el contrario, **los *cluster* 1 y 4 corresponden con las provincias con el menor número de explotaciones**.

Por tanto, resumiendo las características socioeconómicas de cada *cluster*, podemos concluir que:

* **Características socioeconómicas del *Cluster* 1**: representa provincias con crecimiento demográfico negativo, dada su menor población, y su menor acitividad económica. No obstante, sus provincias presentan las tasas de desempleo medio más bajas de todos los grupos.
* **Características socioeconómicas del *Cluster* 2**: agrupa aquellas provincias con un nivel de población y desarrollo económico bajo (en relación al número de empresas), con las mayores tasas de natalidad, es decir, provincias con un crecimiento demográfico positivo, aunque con las tasas desempleo más altas.
* **Características socioeconómicas del *Cluster* 3**: al contrario que en el *cluster* anterior, correponden con provincias mucho más pobladas y con mayor desarrollo económico, destacando por el mayor número de segundas residencias. No obstante, sus porcentajes de desempleo son muy elevados, además de un mayor número de población activa. Además, representa al grupo de provincias con el mayor número medio de explotaciones agrarias.
* **Características socioeconómicas del *Cluster* 4**: representan los principales motores económicos del país, no solo por provincias de mayor población y desarrollo económico sino además donde mayor tasa de ocupación y riqueza se genera, donde incluso el coste de vida es el más elevado (altos porcentajes en el IPC).
* **Características socioeconómicas del *Cluster* 5**: las provincias que lo compone no solo presentan un nivel de población y desarrollo económico bajo, sino que también concentran las regiones con mayor crecimiento demográfico negativo, con las mayores tasas de mortalidad y, como consecuencia, las tasas de natalidad más bajas de todos los grupos.