

# ANÁLISIS DE COMPONENTES PRINCIPALES Y CLUSTER

3/4/2021

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

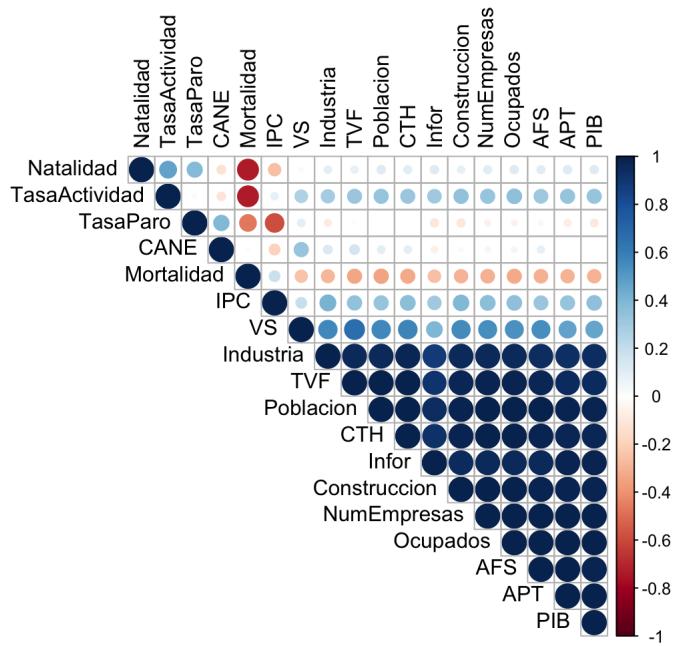
```
# 1.1)
PROVINCIAS <- read_excel("Provincias.xlsx")
datos <- as.data.frame(PROVINCIAS)
rownames(datos)<-datos[,1]
prov<-datos[,-1]

# 1.2)
R<-cor(prov, method="pearson")
knitr::kable(R, digits =2,caption = "Correlaciones")
```

Correlaciones

	Poblacion	Mortalidad	Natalidad	IPC	NumEmpresas	Industria	Construccion	CTH	Infor	AFS	APT	TasaActividad	TasaParc	
Poblacion	1.00	-0.34	0.11	0.33		0.99	0.96	0.98	1.00	0.94	0.99	0.98	0.33	0.01
Mortalidad	-0.34	1.00	-0.74	0.19		-0.31	-0.28	-0.30	-0.33	-0.26	-0.31	-0.30	-0.73	-0.46
Natalidad	0.11	-0.74	1.00	-0.25		0.11	0.09	0.09	0.10	0.11	0.10	0.11	0.47	0.38
IPC	0.33	0.19	-0.25	1.00		0.36	0.42	0.40	0.36	0.30	0.32	0.33	0.09	-0.58
NumEmpresas	0.99	-0.31	0.11	0.36		1.00	0.97	0.99	0.99	0.96	0.99	0.99	0.33	-0.06
Industria	0.96	-0.28	0.09	0.42		0.97	1.00	0.97	0.98	0.89	0.95	0.93	0.29	-0.08
Construccion	0.98	-0.30	0.09	0.40		0.99	0.97	1.00	0.99	0.96	0.98	0.98	0.34	-0.11
CTH	1.00	-0.33	0.10	0.36		0.99	0.98	0.99	1.00	0.93	0.98	0.97	0.33	-0.01
Infor	0.94	-0.26	0.11	0.30		0.96	0.89	0.96	0.93	1.00	0.97	0.99	0.31	-0.11
AFS	0.99	-0.31	0.10	0.32		0.99	0.95	0.98	0.98	0.97	1.00	0.99	0.32	-0.03
APT	0.98	-0.30	0.11	0.33		0.99	0.93	0.98	0.97	0.99	0.99	1.00	0.33	-0.08
TasaActividad	0.33	-0.73	0.47	0.09		0.33	0.29	0.34	0.33	0.31	0.32	0.33	1.00	0.03
TasaParo	0.01	-0.46	0.38	-0.58		-0.06	-0.08	-0.11	-0.01	-0.11	-0.03	-0.08	0.03	1.00
Ocupados	1.00	-0.33	0.11	0.36		1.00	0.96	0.99	0.99	0.96	0.99	0.99	0.35	-0.05
PIB	0.98	-0.30	0.11	0.36		0.99	0.94	0.99	0.97	0.99	0.99	1.00	0.33	-0.10
CANE	0.10	0.02	-0.12	-0.19		0.04	0.12	0.03	0.09	-0.07	0.09	-0.01	-0.12	0.38
TVF	0.99	-0.33	0.08	0.34		0.98	0.97	0.98	0.99	0.91	0.98	0.96	0.33	0.01
VS	0.57	-0.25	-0.03	0.19		0.54	0.57	0.56	0.58	0.41	0.54	0.48	0.26	0.10

```
# 1.3)
corrplot(R, type="upper", order="hclust", tl.col="black", tl.srt=90)
```



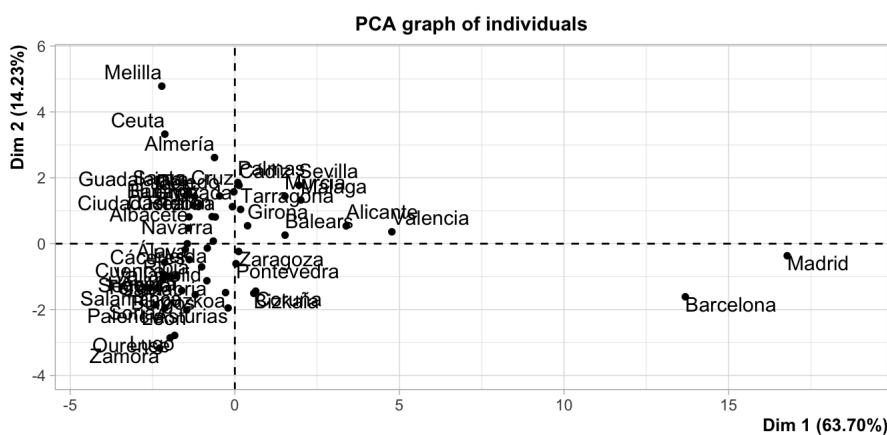
Usando los datos del fichero Excel 'Provincias.xlsx', que contiene 18 variables, calculamos la matriz de correlaciones.

Asimismo, analizamos las correlaciones mediante una salida gráfica. Observamos en primer lugar que las correlaciones positivas se muestran en azul y las correlaciones negativas en rojo. La intensidad del color y el tamaño del círculo son proporcionales a los coeficientes de correlación.

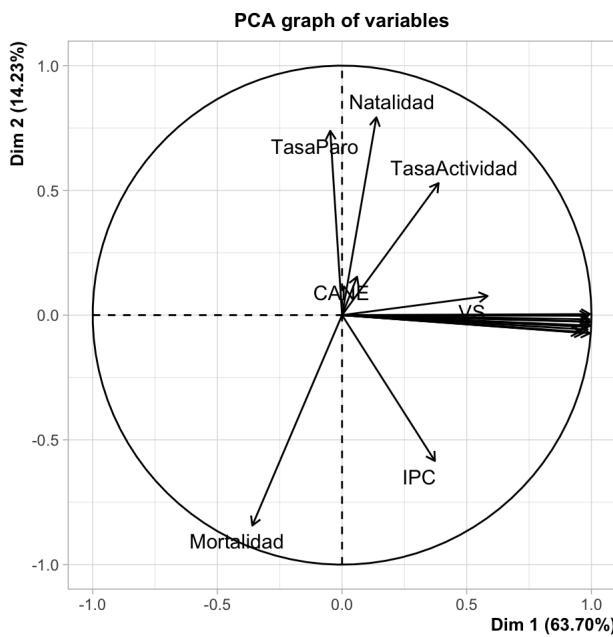
De acuerdo con la matriz de correlación y la representación gráfica, observamos que las variables más correlacionadas de forma inversa son: Industria, TVF, Poblacion, CTH, Infor, Construcción, NumEmpresas, Ocupados, AFS, APT y PIB.

## 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?

```
#2.1)
fit<-PCA(prov,scale.unit=TRUE,ncp=7,graph=TRUE)
```



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



```
summary(fit)
```

```

## Call:
## PCA(X = prov, scale.unit = TRUE, ncp = 7, graph = TRUE)
##
## Eigenvalues
##           Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance    11.466  2.561  1.634  0.934  0.457  0.414  0.307
## % of var.  63.702 14.225  9.078  5.189  2.536  2.302  1.707
## Cumulative % of var. 63.702 77.927 87.006 92.195 94.731 97.033 98.740
##           Dim.8   Dim.9   Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance    0.117  0.073  0.020  0.009  0.004  0.002  0.001
## % of var.  0.648  0.406  0.113  0.050  0.020  0.013  0.004
## Cumulative % of var. 99.387 99.794 99.907 99.957 99.977 99.990 99.994
##           Dim.15  Dim.16  Dim.17  Dim.18
## Variance    0.001  0.000  0.000  0.000
## % of var.  0.003  0.002  0.001  0.000
## Cumulative % of var. 99.997 99.999 100.000 100.000
##
## Individuals (the 10 first)
##           Dist   Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3
## Albacete | 1.652 | -1.410  0.334  0.729 | 0.473  0.168  0.082 | 0.096
## Alicante | 5.395 |  3.384  1.920  0.393 | 0.540  0.219  0.010 | 1.919
## Almería | 2.704 | -0.617  0.064  0.052 | 2.614  5.132  0.935 | 0.208
## Álava    | 2.681 | -1.444  0.350  0.290 | -0.001  0.000  0.000 | -2.032
## Asturias | 2.560 | -0.204  0.007  0.006 | -1.953  2.863  0.582 | 1.298
## Badajoz  | 2.611 | -1.048  0.184  0.161 | 1.168  1.024  0.200 | 1.796
## Balears  | 3.940 |  1.526  0.391  0.150 | 0.260  0.051  0.004 | -2.519
## Barcelona| 13.930| 13.683 31.399  0.965 | -1.612  1.952  0.013 | -0.867
## Bizkaia  | 2.289 |  0.576  0.056  0.063 | -1.508  1.708  0.434 | -1.180
## Burgos   | 2.583 | -1.202  0.242  0.217 | -1.550  1.804  0.360 | -0.892
##           ctr   cos2
## Albacete | 0.011  0.003 |
## Alicante | 4.332  0.126 |
## Almería | 0.051  0.006 |
## Álava    | 4.858  0.574 |
## Asturias | 1.982  0.257 |
## Badajoz  | 3.794  0.473 |
## Balears  | 7.469  0.409 |
## Barcelona| 0.884  0.004 |
## Bizkaia  | 1.638  0.266 |
## Burgos   | 0.935  0.119 |
##
## Variables (the 10 first)
##           Dim.1   ctr   cos2   Dim.2   ctr   cos2   Dim.3   ctr
## Poblacion | 0.994  8.616  0.988 | 0.004  0.001  0.000 | 0.064  0.249
## Mortalidad | -0.360 1.130  0.130 | -0.843 27.786  0.711 | 0.241  3.567
## Natalidad  | 0.138  0.165  0.019 | 0.793 24.544  0.628 | -0.346 7.327
## IPC        | 0.372  1.208  0.139 | -0.585 13.346  0.342 | -0.335 6.881
## NumEmpresas| 0.996  8.654  0.992 | -0.042  0.068  0.002 | 0.010  0.006
## Industria  | 0.967  8.158  0.935 | -0.072  0.200  0.005 | 0.059  0.215
## Construccion| 0.993  8.600  0.986 | -0.073  0.206  0.005 | -0.015 0.015
## CTH        | 0.992  8.577  0.983 | -0.017  0.011  0.000 | 0.062  0.239
## Infor      | 0.953  7.924  0.909 | -0.067  0.177  0.005 | -0.083 0.418
## AFS        | 0.990  8.554  0.981 | -0.026  0.027  0.001 | 0.051  0.156
##           cos2
## Poblacion | 0.004 |
## Mortalidad | 0.058 |
## Natalidad  | 0.120 |
## IPC        | 0.112 |
## NumEmpresas| 0.000 |
## Industria  | 0.004 |
## Construccion| 0.000 |
## CTH        | 0.004 |
## Infor      | 0.007 |
## AFS        | 0.003 |

```

```

# 2.2)
eig<-get_eigenvalue(fit)
knitr::kable(eig, digits =2,caption = "Autovalores")

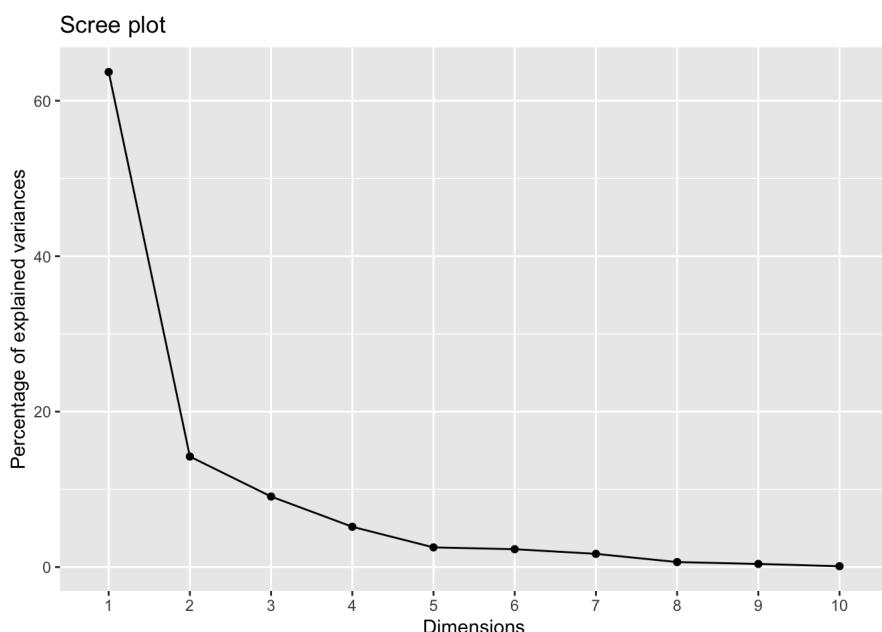
```

## Autovalores

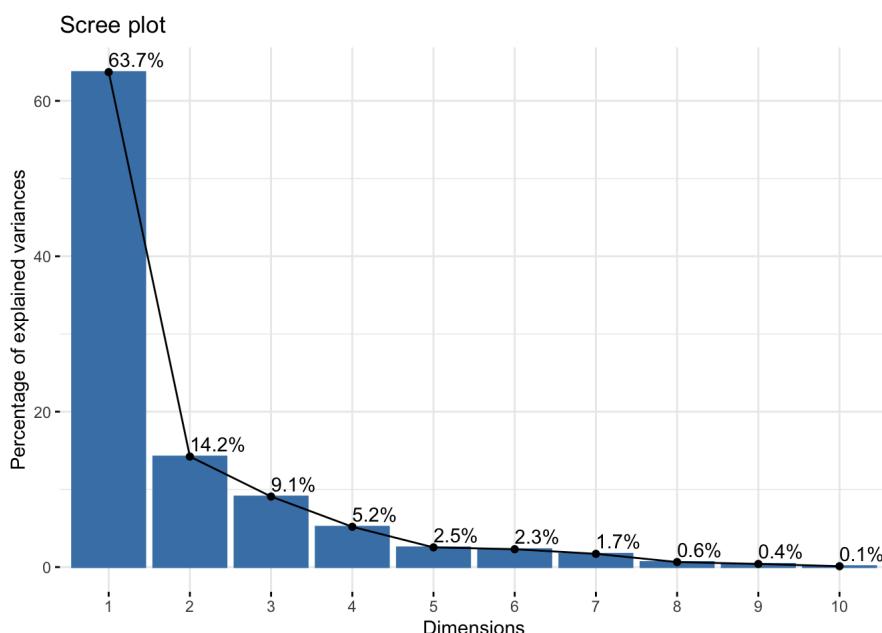
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	11.47	63.70	63.70
Dim.2	2.56	14.23	77.93
Dim.3	1.63	9.08	87.01
Dim.4	0.93	5.19	92.19
Dim.5	0.46	2.54	94.73
Dim.6	0.41	2.30	97.03

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.7	0.31	1.71	98.74
Dim.8	0.12	0.65	99.39
Dim.9	0.07	0.41	99.79
Dim.10	0.02	0.11	99.91
Dim.11	0.01	0.05	99.96
Dim.12	0.00	0.02	99.98
Dim.13	0.00	0.01	99.99
Dim.14	0.00	0.00	99.99
Dim.15	0.00	0.00	100.00
Dim.16	0.00	0.00	100.00
Dim.17	0.00	0.00	100.00
Dim.18	0.00	0.00	100.00

```
# 2.4)
fviz_eig(fit, geom="line") + theme_grey()
```



```
# 2.5)
fviz_eig(fit, addlabels=TRUE)
```



Recordamos que el número de Componentes determina la proporción de variabilidad a explicar que se considere suficiente, como mínimo el 70%, pero si es posible en torno al 80-90%. Así que analizamos los eigenvalues para determinar el número adecuado de componentes.

La proporción de varianza explicada por cada autovalor se da en la segunda columna en porcentaje. Observamos en este caso que el 63.7% (Dim.1) de la proporción de variabilidad se explica por este autovalor. Por lo tanto, el número adecuado de componentes es 4, ya que el 92.19% de la proporción de variabilidad total se explica por los 4 primeros porcentajes de los eigenvalues.

También se aprecia con los puntos de la gráfica que se nivelan. Recordemos que se puede trazar una recta que agrupe en su entorno a los autovalores más pequeños y todos los que queden por encima corresponderían a las Componentes Principales retenidas.

Por otra parte, podemos determinar el número adecuado de componentes analizando las gráficas de eigenvalues. Llegamos a la misma observación: las 4 componentes principales explicarían el 92.19% de la varianza de las variables iniciales. Por lo tanto, parece adecuado retener las 4 primeras Componentes Principales.

### 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.

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

```
# a.1)
res.desc <- dimdesc(fit, axes = c(1,2,3), proba = 0.05)
res.desc$Dim.1
```

```
## $quanti
##          correlation      p.value
## Ocupados        0.9969955 3.201168e-57
## NumEmpresas     0.9961625 1.438712e-54
## Poblacion       0.9939432 1.264155e-49
## Construccion    0.9930557 3.819079e-48
## CTH            0.9917052 3.195364e-46
## AFS            0.9903502 1.382069e-44
## TVF            0.9873051 1.267521e-41
## PIB            0.9850716 7.100277e-40
## APT            0.9844547 1.939991e-39
## Industria      0.9671781 2.062516e-31
## Infor           0.9532238 1.229340e-27
## VS              0.5835900 5.601988e-06
## TasaActividad   0.3871698 4.574551e-03
## IPC             0.3721837 6.588117e-03
## Mortalidad      -0.3599267 8.771012e-03
##
## attr(", "class")
## [1] "condes" "list"
```

```
# a.2)
knitr::kable(fit$svd$V, digits = 3, caption = "Avalores propios")
```

```
## Warning in kable_pipe(x = structure(c("0.294", "-0.106", "0.041", "0.110", : The
## table should have a header (column names)
```

Avalores propios

0.294	0.002	0.050	-0.053
-0.106	-0.527	0.189	-0.161
0.041	0.495	-0.271	-0.110
0.110	-0.365	-0.262	0.435
0.294	-0.026	0.008	-0.069
0.286	-0.045	0.046	0.023
0.293	-0.045	-0.012	-0.026
0.293	-0.011	0.049	-0.028
0.282	-0.042	-0.065	-0.222
0.292	-0.016	0.040	-0.092
0.291	-0.029	-0.028	-0.142

0.114	0.331	-0.363	0.463
-0.014	0.462	0.387	-0.220
0.294	-0.017	0.002	-0.060
0.291	-0.036	-0.037	-0.134
0.018	0.096	0.657	0.278
0.292	-0.002	0.100	0.044
0.172	0.048	0.290	0.567

Observamos en las tablas que la primera dimensión está representada al inicio mayormente por Ocupados, NumEmpresas Poblacion y Construccion.

Asimismo, la expresión para calcular el primer componente basado en las variables originales es la siguiente:  $PC1 = 0.29Poblacion^* + -0.10Mortalidad^* + 0.04Natalidad^* + 0.11IPC^* + 0.29NumEmpresas^* + 0.28Industria^* + 0.29Construccion^* + 0.29CTH^* + 0.28Infor^* + 0.29AFS^* + 0.29APT^* + 0.11TasaActividad^* + -0.01TasaParo^* + 0.29Ocupados^* + 0.29PIB^* + 0.01CANE^* + 0.29TVF^* + 0.17VS^*$

## 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

```
# b)
var<-get_pca_var(fit)
var$coord
```

	Dim.1	Dim.2	Dim.3	Dim.4
## Poblacion	0.99394316	0.003989744	0.063848580	-0.05160463
## Mortalidad	-0.35992669	-0.843475898	0.241432113	-0.15590654
## Natalidad	0.13757149	0.792749686	-0.346016593	-0.10599223
## IPC	0.37218373	-0.584573656	-0.335325313	0.42045677
## NumEmpresas	0.99616251	-0.041840166	0.010165492	-0.06681923
## Industria	0.96717807	-0.071587054	0.059274514	0.02266458
## Construccion	0.99305566	-0.072618036	-0.015444759	-0.02550892
## CTH	0.99170521	-0.016989690	0.062434517	-0.02665589
## Infor	0.95322382	-0.067235670	-0.082630799	-0.21409732
## AFS	0.99035020	-0.026329340	0.050562974	-0.08860621
## APT	0.98445474	-0.047097804	-0.035880121	-0.13710328
## TasaActividad	0.38716977	0.529035826	-0.463689989	0.44752178
## TasaParo	-0.04739450	0.738572741	0.495165389	-0.21281673
## Ocupados	0.99699549	-0.026944763	0.003105467	-0.05803848
## PIB	0.98507163	-0.057858046	-0.047926363	-0.12912629
## CANE	0.06023588	0.153910432	0.839526520	0.26905683
## TVF	0.98730511	-0.002559620	0.127527357	0.04266413
## VS	0.58359002	0.076521234	0.370894920	0.54810069

```
knitr::kable(var$cor, digits =2,caption = "Correlaciones de la CP con las variables")
```

Correlaciones de la CP con las variables

	Dim.1	Dim.2	Dim.3	Dim.4
Poblacion	0.99	0.00	0.06	-0.05
Mortalidad	-0.36	-0.84	0.24	-0.16
Natalidad	0.14	0.79	-0.35	-0.11
IPC	0.37	-0.58	-0.34	0.42
NumEmpresas	1.00	-0.04	0.01	-0.07
Industria	0.97	-0.07	0.06	0.02
Construccion	0.99	-0.07	-0.02	-0.03
CTH	0.99	-0.02	0.06	-0.03
Infor	0.95	-0.07	-0.08	-0.21
AFS	0.99	-0.03	0.05	-0.09
APT	0.98	-0.05	-0.04	-0.14
TasaActividad	0.39	0.53	-0.46	0.45
TasaParo	-0.05	0.74	0.50	-0.21
Ocupados	1.00	-0.03	0.00	-0.06
PIB	0.99	-0.06	-0.05	-0.13

	Dim.1	Dim.2	Dim.3	Dim.4
CANE	0.06	0.15	0.84	0.27
TVF	0.99	0.00	0.13	0.04
VS	0.58	0.08	0.37	0.55

Observemos que la primera componente (alrededor de 0,99), tiene una correlación alta con las variables Población, NumEmpresas, Construcción, CTH, AFS, Ocupados, PIB y TVF.

Para la segunda componente, observamos que tiene un gran correlación solo con Natalidad (0,79) y TasaParo (0,73). También tiene una correlación negativa importante con Mortalidad (-0,84).

Para la tercera componente, vemos que la variable más correlacionada es CANE, con 0,84.

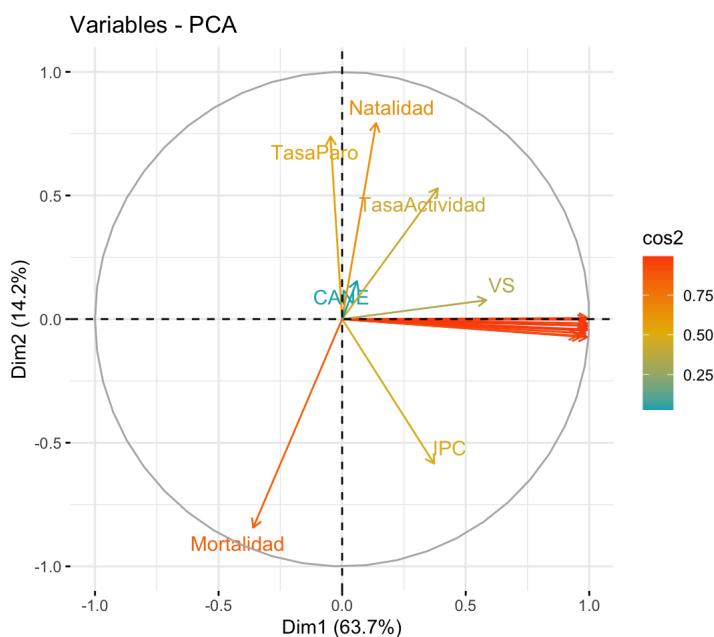
Finalmente, la cuarta componente no tiene variables correlacionadas significativas, aunque VS (0,55) y TasaActividad (0,45) tienen una cierta relevancia.

La siguiente representación gráfica nos puede ayudar a entender como están recogidas nuestras variables iniciales en las nuevas componentes, puesto que representa el coeficiente de correlación entre las variables y las nuevas componentes.

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

```
#c)
fviz_pca_var(fit, axes = c(1,2), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```

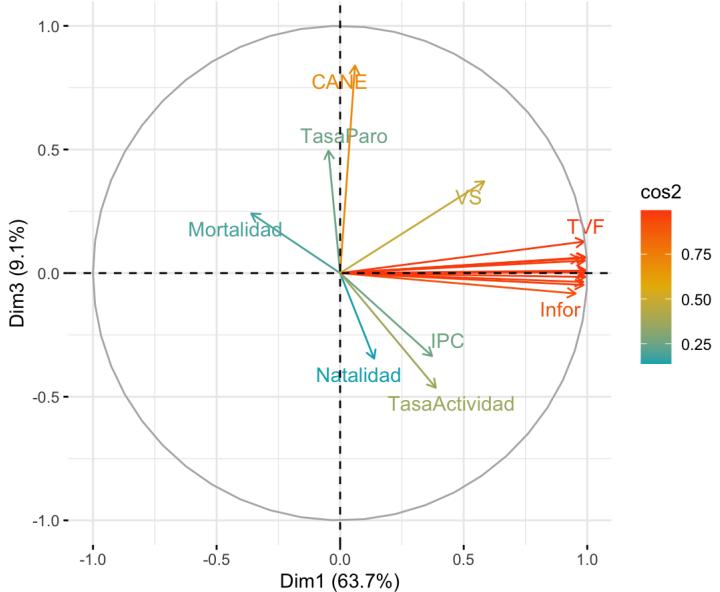
```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_var(fit, axes = c(1,3), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```

```
## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

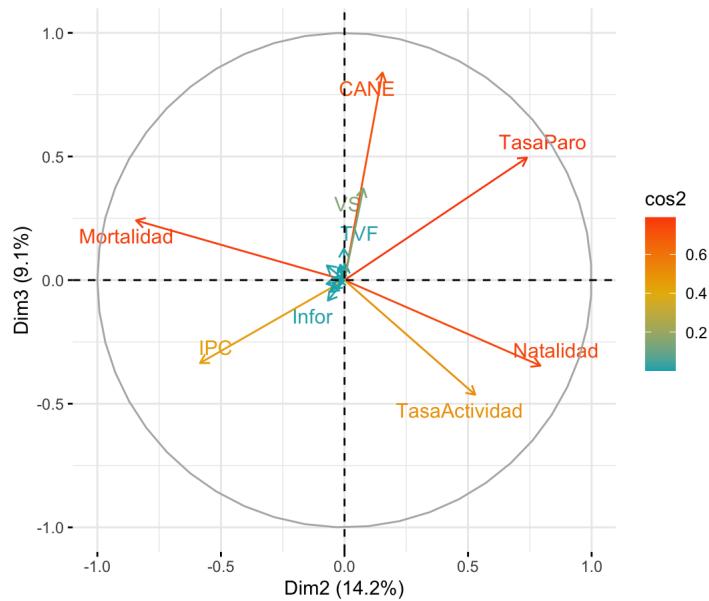
Variables - PCA



```
fviz_pca_var(fit, axes = c(2,3), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```

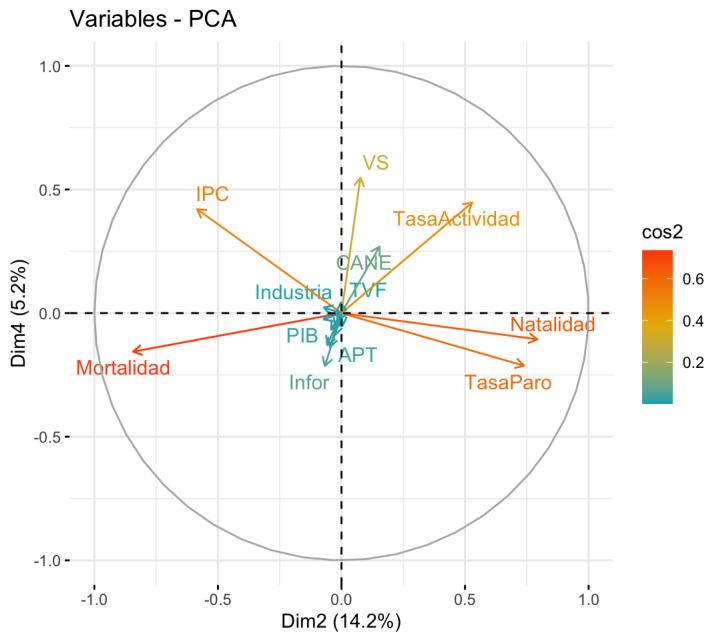
```
## Warning: ggrepel: 9 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

Variables - PCA



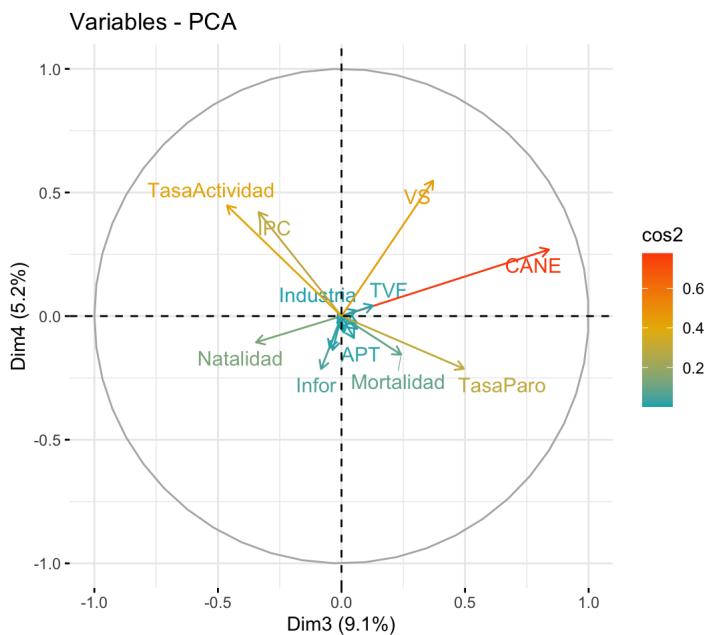
```
fviz_pca_var(fit, axes = c(2,4), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```

```
## Warning: ggrepel: 6 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_var(fit, axes = c(3,4), col.var="cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),repel = TRUE )
```

```
## Warning: ggrepel: 7 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



- La componente 1 representa la Población, el NumEmpresas, la Industria, la Construcción, el CTH, el PIB, el Infor, el AFS, el APT, los Ocupados y el TVF.
- La componente 2 representa a la Mortalidad, Natalidad y la TasaParo.
- La componente 3 representa sobre todo la información de la variable CANE.
- La componente 4 representa sobre todo la información de la variable IPC.

**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?**

```
# d.1)
knitr::kable(var$cos2, digits =2,caption = "Cosenos al cuadrado")
```

Cosenos al cuadrado

Dim.1

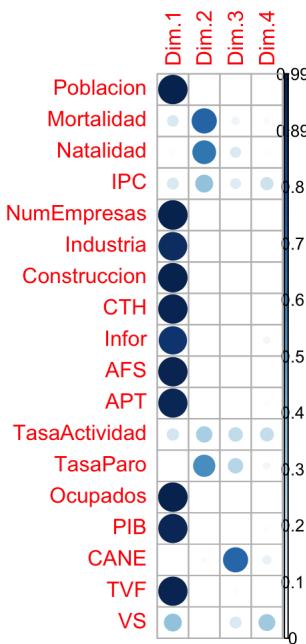
Dim.2

Dim.3

Dim.4

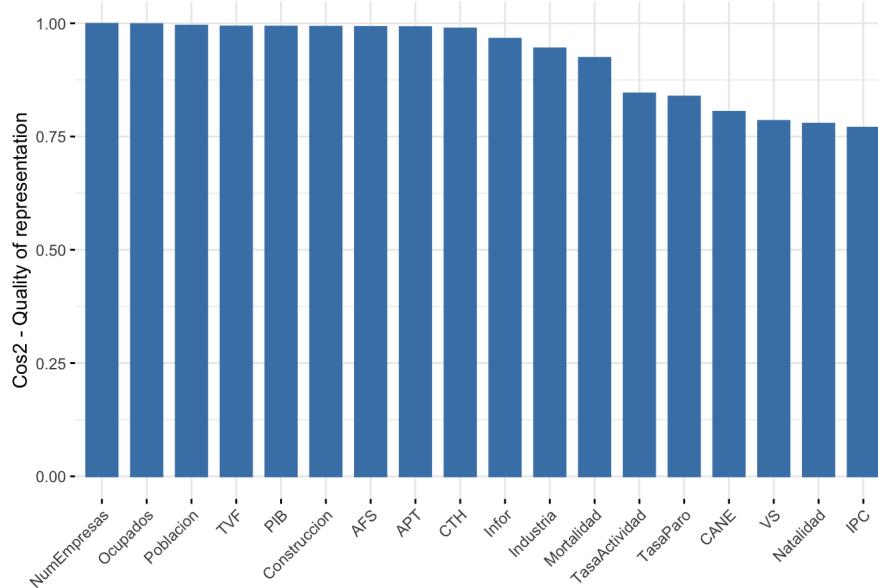
	Dim.1	Dim.2	Dim.3	Dim.4
Poblacion	0.99	0.00	0.00	0.00
Mortalidad	0.13	0.71	0.06	0.02
Natalidad	0.02	0.63	0.12	0.01
IPC	0.14	0.34	0.11	0.18
NumEmpresas	0.99	0.00	0.00	0.00
Industria	0.94	0.01	0.00	0.00
Construccion	0.99	0.01	0.00	0.00
CTH	0.98	0.00	0.00	0.00
Infor	0.91	0.00	0.01	0.05
AFS	0.98	0.00	0.00	0.01
APT	0.97	0.00	0.00	0.02
TasaActividad	0.15	0.28	0.22	0.20
TasaParo	0.00	0.55	0.25	0.05
Ocupados	0.99	0.00	0.00	0.00
PIB	0.97	0.00	0.00	0.02
CANE	0.00	0.02	0.70	0.07
TVF	0.97	0.00	0.02	0.00
VS	0.34	0.01	0.14	0.30

```
# d.2)
corrplot(var$cos2, is.corr=FALSE)
```



```
# d.3)
fviz_cos2(fit, choice="var", axes=1:4)
```

### Cos2 of variables to Dim-1-2-3-4



```
# d.4)
var$contrib
```

```
##           Dim.1      Dim.2      Dim.3      Dim.4
## Poblacion 8.61580869 6.216756e-04 2.494737e-01 0.28510690
## Mortalidad 1.12979868 2.778556e+01 3.567076e+00 2.60231048
## Natalidad  0.16505544 2.454404e+01 7.326830e+00 1.20275871
## IPC        1.20805785 1.334604e+01 6.881053e+00 18.92662230
## NumEmpresas 8.65432772 6.836923e-02 6.323813e-03 0.47800598
## Industria   8.15804006 2.001441e-01 2.150098e-01 0.05499531
## Construccion 8.60042932 2.059505e-01 1.459770e-02 0.06966501
## CTH         8.57705398 1.127314e-02 2.385458e-01 0.07607058
## Infor       7.92433313 1.765523e-01 4.178363e-01 4.90741583
## AFS          8.55363151 2.707409e-02 1.564543e-01 0.84054061
## APT          8.45209684 8.663137e-02 7.878255e-02 2.01245468
## TasaActividad 1.30730170 1.093060e+01 1.315763e+01 21.44167757
## TasaParo     0.01958975 2.130396e+01 1.500454e+01 4.84888573
## Ocupados    8.66880711 2.835454e-02 5.901685e-04 0.36063054
## PIB          8.46269279 1.307379e-01 1.405631e-01 1.78508842
## CANE         0.03164342 9.251454e-01 4.313115e+01 7.75029908
## TVF          8.50111166 2.558733e-04 9.952426e-01 0.19487506
## VS           2.97022035 2.286851e-01 8.418293e+00 32.16259720
```

En los gráficos que nos muestran la proporción de la varianza de cada variable que es explicada por cada componente:

- Componente 1: Gráficamente se muestra que las variables (en Dim1) representan principalmente a las variables Población, NumEmpresas, Industria, Construcción, CTH, Infor, AFS, APT, Ocupados, PIB y TVF.
- Componente 2: La segunda (Dim2) explica Mortalidad y Natalidad y podríamos incluir TasaParo.
- Componente 3: La tercera (Dim3) representa principalmente CANE.
- Componente 4: La cuarta (Dim4) representa sobre todo IPC

Finalmente, la variable que está peor explicada es VS, lo que se observa también en el corplot y en las tablas.

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

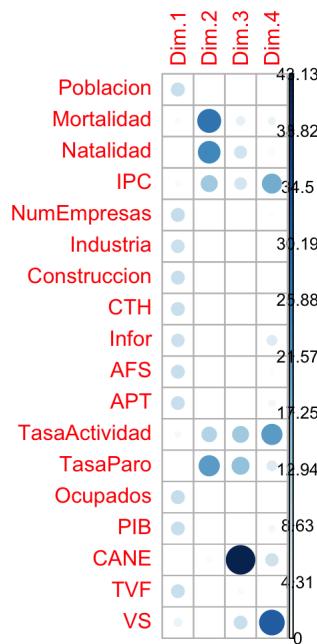
```
# e.1)
knitr:::kable(var$contrib, digits =2,caption = "Contribuciones")
```

Contribuciones

	Dim.1	Dim.2	Dim.3	Dim.4
Poblacion	8.62	0.00	0.25	0.29
Mortalidad	1.13	27.79	3.57	2.60
Natalidad	0.17	24.54	7.33	1.20
IPC	1.21	13.35	6.88	18.93

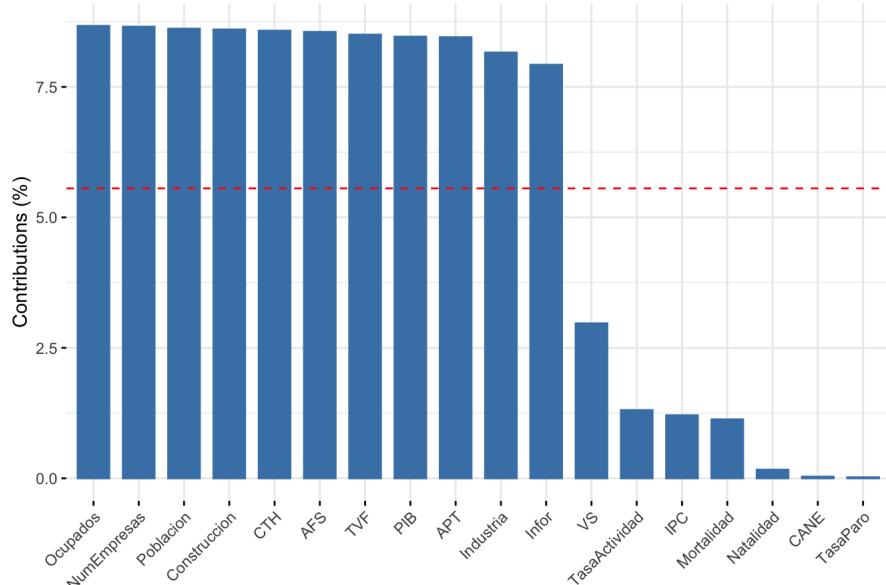
	<b>Dim.1</b>	<b>Dim.2</b>	<b>Dim.3</b>	<b>Dim.4</b>
NumEmpresas	8.65	0.07	0.01	0.48
Industria	8.16	0.20	0.22	0.05
Construccion	8.60	0.21	0.01	0.07
CTH	8.58	0.01	0.24	0.08
Infor	7.92	0.18	0.42	4.91
AFS	8.55	0.03	0.16	0.84
APT	8.45	0.09	0.08	2.01
TasaActividad	1.31	10.93	13.16	21.44
TasaParo	0.02	21.30	15.00	4.85
Ocupados	8.67	0.03	0.00	0.36
PIB	8.46	0.13	0.14	1.79
CANE	0.03	0.93	43.13	7.75
TVF	8.50	0.00	1.00	0.19
VS	2.97	0.23	8.42	32.16

```
# e.2)
corrplot(var$contrib, is.corr=FALSE)
```



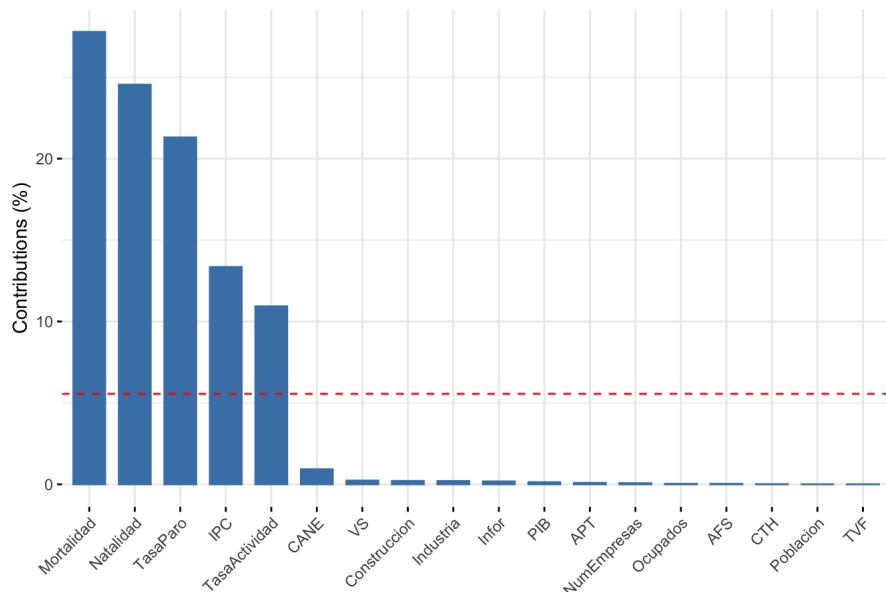
```
# e.3)
fviz_contrib(fit, choice="var", axes=1)
```

### Contribution of variables to Dim-1



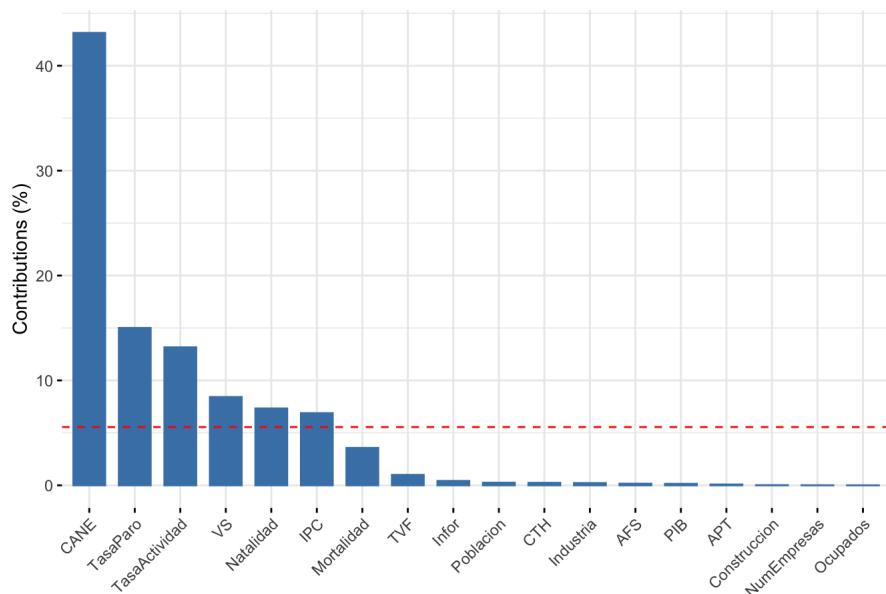
```
fviz_contrib(fit, choice = "var", axes = 2)
```

### Contribution of variables to Dim-2

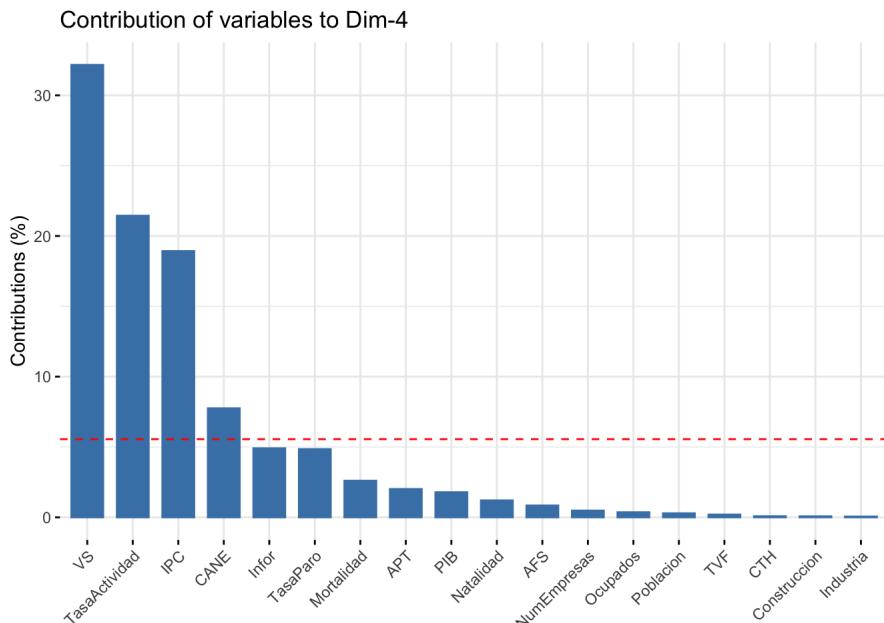


```
fviz_contrib(fit, choice = "var", axes = 3)
```

### Contribution of variables to Dim-3



```
fviz_contrib(fit, choice="var", axes=4)
```

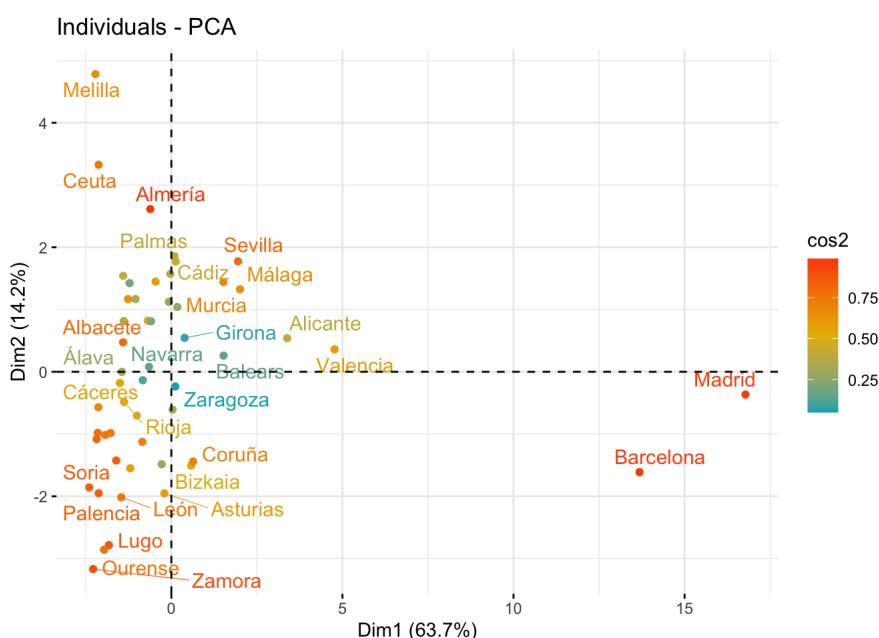


- Componente 1: observamos que Población, NumEmpresas, Industria, Construcción, CTH, Infor, AFS, APT, Ocupados, PIB y TVF contribuyen casi en la misma proporción.
- Componente 2: Mortalidad, Natalidad, TasaParo y IPC contribuyen mayormente
- Componente 3: sobre todo CANE contribuye.
- Componente 4: VS, TasaActividad y IPC contribuyen principalmente.

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?

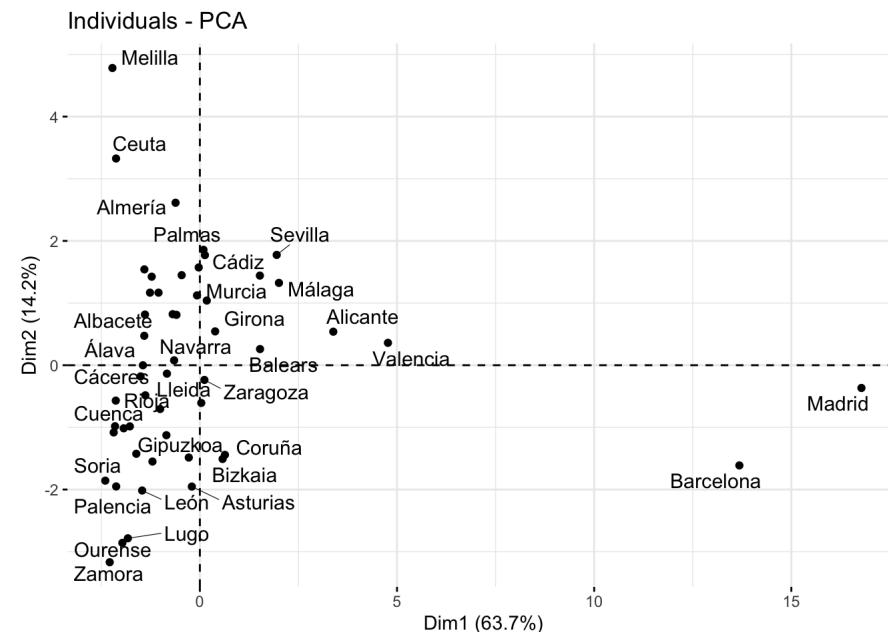
```
#f1)
fviz_pca_ind(fit, col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
```

```
## Warning: ggrepel: 23 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



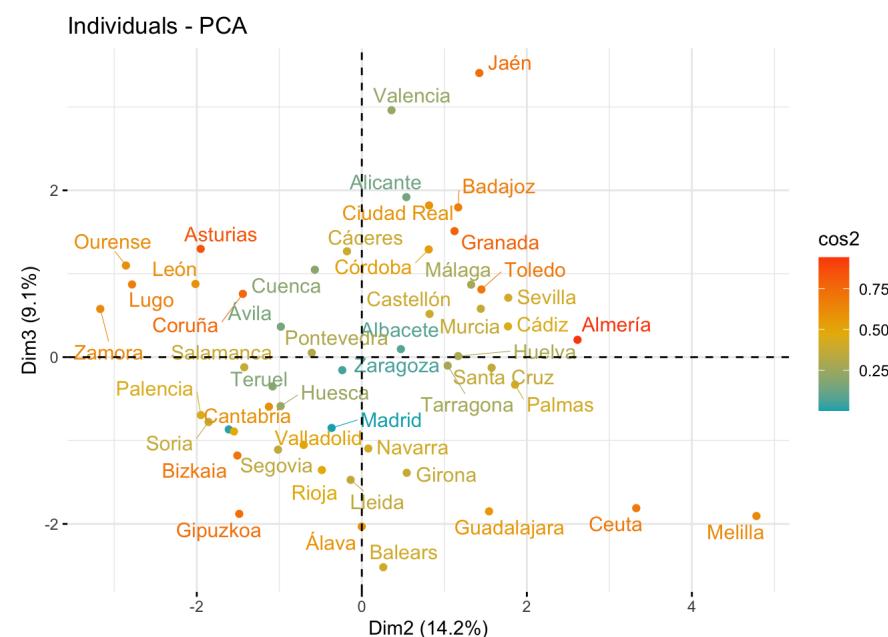
```
# f.2)
fviz_pca_ind(fit, axes = c(1, 2), gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = TRUE)
```

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



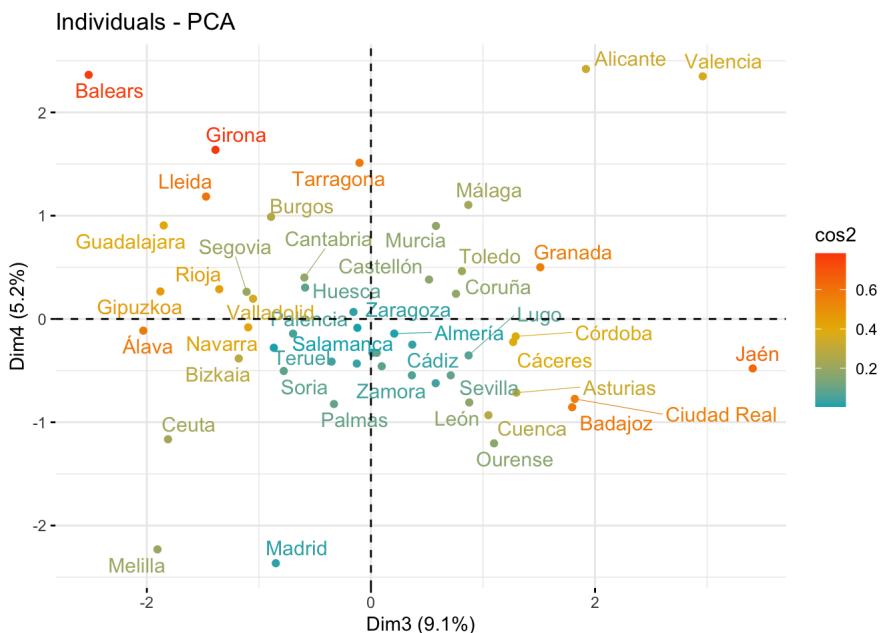
```
fviz_pca_ind(fit, axes = c(2, 3), col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = T
RUE)
```

```
## Warning: ggrepel: 2 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_ind(fit, axes = c(3, 4), col.ind = "cos2", gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"), repel = T
RUE)
```

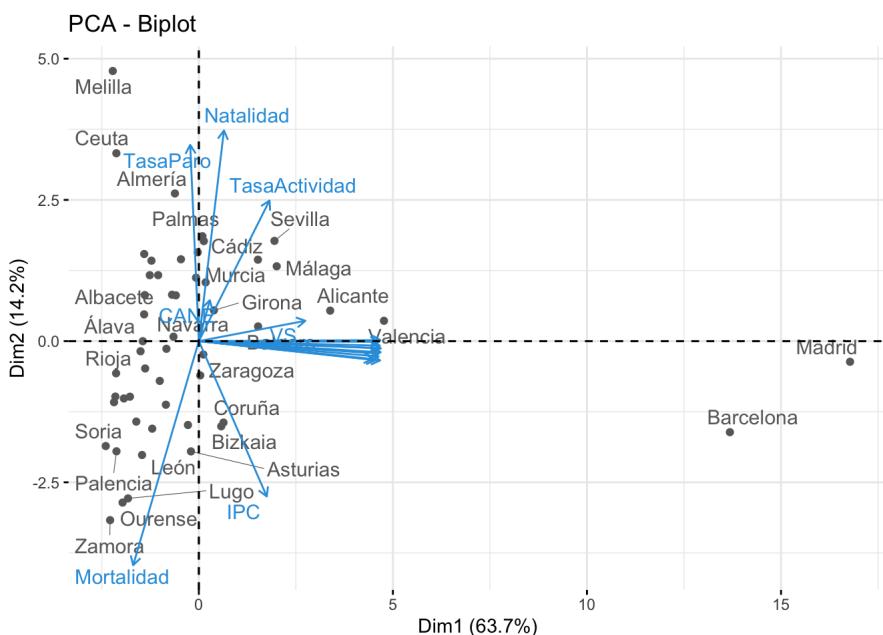
```
## Warning: ggrepel: 6 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
# f.3) Representación conjunta de los individuos y las variables en los planos de las CP
fviz_pca_biplot(fit, repel = TRUE, axes = c(1, 2), col.var = "#2E9FDF", col.ind = "#696969")
```

```
## Warning: ggrepel: 24 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

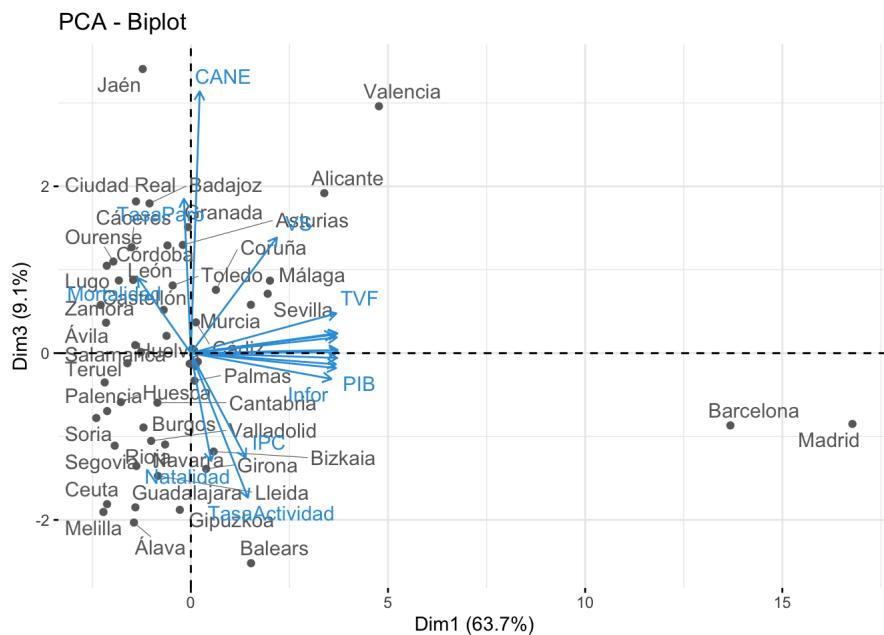
```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_biplot(fit, repel = TRUE, axes = c(1, 3), col.var = "#2E9FDF", col.ind = "#696969")
```

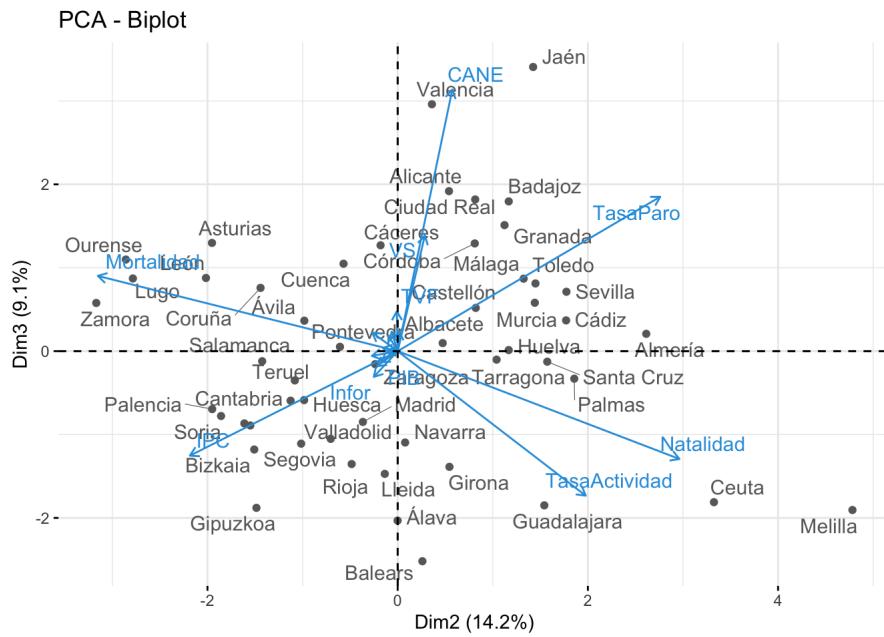
```
## Warning: ggrepel: 7 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
## Warning: ggrepel: 8 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



```
fviz_pca_biplot(fit, repel = TRUE, axes = c(2, 3), col.var = "#2E9FDF", col.ind = "#696969")
```

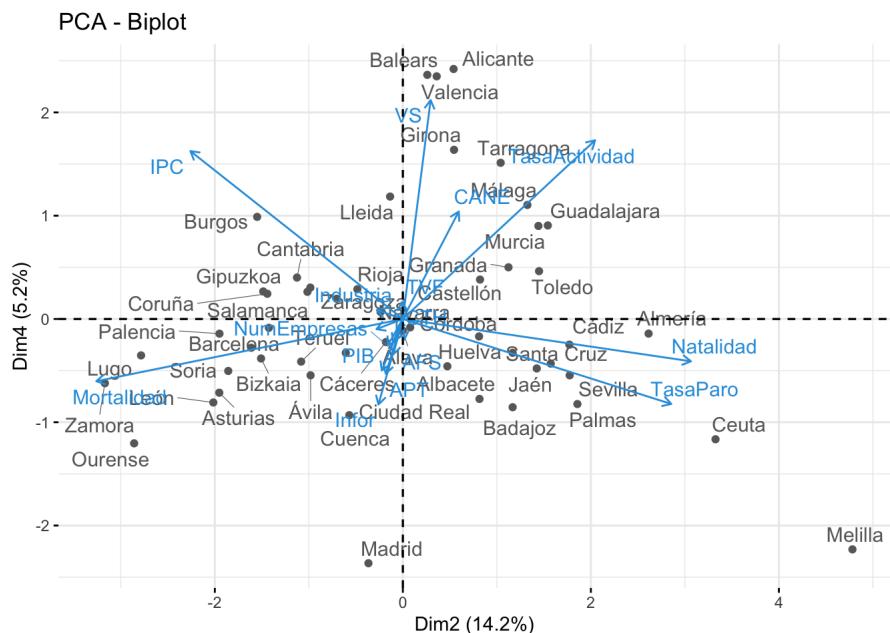
```
## Warning: ggrepel: 2 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps  
  
## Warning: ggrepel: 8 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```



```
fviz_pca_biplot(fit, repel = TRUE, axes = c(2, 4), col.var = "#2E9FDF", col.ind = "#696969")
```

```
## Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```

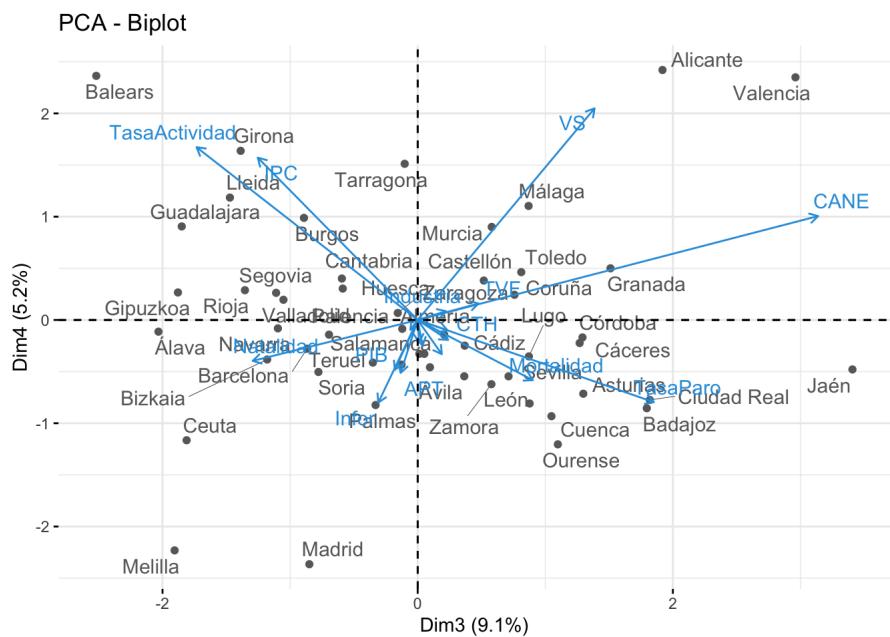
```
## Warning: ggrepel: 3 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps
```



```
fviz_pca_biplot(fit, repel = TRUE, axes = c(3, 4), col.var = "#2E9FDF", col.ind = "#696969")
```

```
## Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

```
## Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



De las provincias más destacadas en cada componente, en positivo o negativo:

Vemos que Madrid y Barcelona tienen un comportamiento similar. Tienen un valor alto de la CP1 lo que significa alto porcentaje Población, NumEmpresas, Industria, Construcción. Se trata de provincias más desarrolladas ya que emplean a la mayoría de la población del país, con un número de empresas igual de importante y una participación en el PIB considerable. Sin embargo, tienen una tasa de Natalidad negativa.

Valencia y Alicante también son similares, ya que tienen un alto porcentaje de VS (Viviendas secundarias) y una TasaActividad alta, lo que puede ser el resultado de las actividades turísticas nacionales. Además, un CANE muy importante significa que son regiones con una agricultura relevante.

Melilla y Ceuta tienen una tasa de paro importante (negativa) y al contrario una Natalidad significativa de manera positiva. Se trata de provincias pobladas pero con poca actividad económica.

Zamora y Ourense tienen mucha Mortalidad, así como una tasa de paro significativa. Se trata de provincias poco desarrolladas.

**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? ¿Cuál sería su valor en Melilla?

```
# g.1)
ind<-get_pca_ind(fit)
knitr::kable(ind$coord, digits =3,caption = "Valores de los individuos en las Cp")
```

Valores de los individuos en las Cp

	Dim.1	Dim.2	Dim.3	Dim.4
Albacete	-1.410	0.473	0.096	-0.458
Alicante	3.384	0.540	1.919	2.420
Almería	-0.617	2.614	0.208	-0.142
Álava	-1.444	-0.001	-2.032	-0.113
Asturias	-0.204	-1.953	1.298	-0.714
Badajoz	-1.048	1.168	1.796	-0.854
Baleares	1.526	0.260	-2.519	2.364
Barcelona	13.683	-1.612	-0.867	-0.279
Bizkaia	0.576	-1.508	-1.180	-0.383
Burgos	-1.202	-1.550	-0.892	0.989
Cantabria	-0.849	-1.126	-0.594	0.401
Castellón	-0.690	0.821	0.518	0.382
Ceuta	-2.125	3.326	-1.811	-1.165
Ciudad Real	-1.392	0.815	1.819	-0.774
Coruña	0.635	-1.442	0.759	0.244
Cuenca	-2.132	-0.569	1.048	-0.932
Cáceres	-1.503	-0.180	1.269	-0.223
Cádiz	0.128	1.771	0.369	-0.249
Córdoba	-0.594	0.811	1.292	-0.169
Gipuzkoa	-0.281	-1.485	-1.879	0.266
Girona	0.388	0.544	-1.387	1.638
Granada	-0.072	1.124	1.511	0.500
Guadalajara	-1.408	1.542	-1.849	0.906
Huelva	-1.265	1.168	0.013	-0.327
Huesca	-1.776	-0.984	-0.587	0.304
Jaén	-1.221	1.424	3.407	-0.479
León	-1.464	-2.016	0.877	-0.809
Lleida	-0.835	-0.136	-1.472	1.185
Lugo	-1.826	-2.785	0.871	-0.353
Madrid	16.778	-0.366	-0.849	-2.365
Melilla	-2.218	4.782	-1.905	-2.231
Murcia	1.522	1.442	0.580	0.901
Málaga	2.006	1.325	0.869	1.104
Navarra	-0.653	0.078	-1.096	-0.081
Ourense	-1.965	-2.858	1.098	-1.204
Palencia	-2.122	-1.951	-0.695	-0.142
Palmas	0.092	1.857	-0.330	-0.824
Pontevedra	0.036	-0.607	0.052	-0.328
Rioja	-1.383	-0.484	-1.354	0.288
Salamanca	-1.612	-1.425	-0.121	-0.086
Santa Cruz	-0.029	1.573	-0.126	-0.432
Segovia	-1.931	-1.015	-1.110	0.263
Sevilla	1.948	1.775	0.712	-0.546

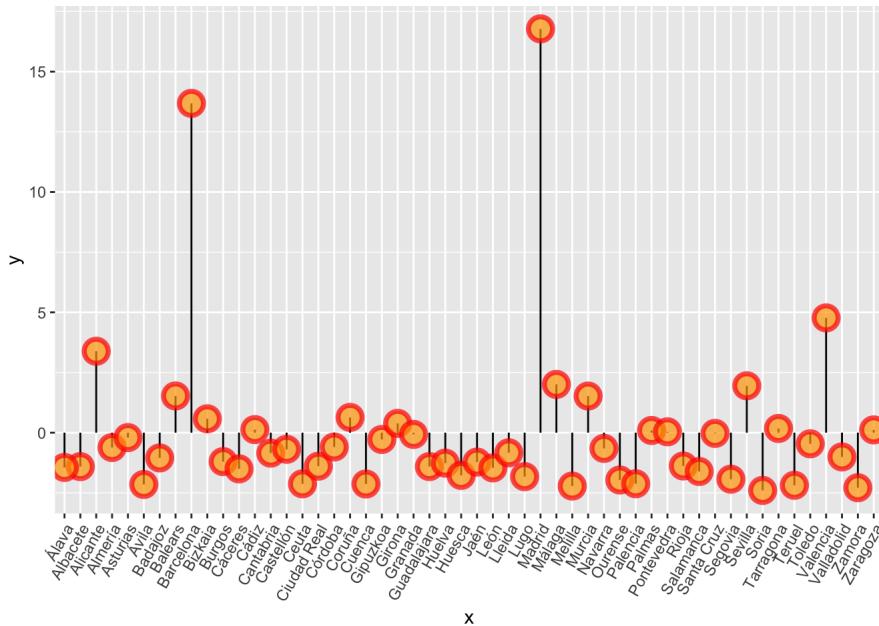
	<b>Dim.1</b>	<b>Dim.2</b>	<b>Dim.3</b>	<b>Dim.4</b>
Soria	-2.399	-1.857	-0.778	-0.503
Tarragona	0.175	1.040	-0.102	1.512
Teruel	-2.185	-1.082	-0.351	-0.413
Toledo	-0.461	1.449	0.812	0.463
Valencia	4.770	0.360	2.961	2.349
Valladolid	-1.007	-0.704	-1.052	0.196
Zamora	-2.287	-3.169	0.578	-0.622
Zaragoza	0.115	-0.237	-0.156	0.068
Ávila	-2.152	-0.982	0.365	-0.545

```
# g.2)
data <- data.frame(ind$coord[,1])
data
```

```
##      ind.coord...1.
## Albacete -1.41017686
## Alicante  3.38367603
## Almería  -0.61651226
## Álava     -1.44429594
## Asturias -0.20350137
## Badajoz   -1.04846596
## Balears   1.52621574
## Barcelona 13.68268794
## Bizkaia   0.57644373
## Burgos    -1.20182455
## Cantabria -0.84908539
## Castellón -0.68987702
## Ceuta     -2.12522644
## Ciudad Real -1.39218950
## Coruña    0.63521316
## Cuenca    -2.13228382
## Cáceres   -1.50273216
## Cádiz     0.12808408
## Córdoba   -0.59357443
## Gipuzkoa -0.28110176
## Girona    0.38750851
## Granada   -0.07240306
## Guadalajara -1.40758889
## Huelva    -1.26474244
## Huesca    -1.77555839
## Jaén      -1.22127859
## León      -1.46356388
## Lleida    -0.83527088
## Lugo      -1.82595354
## Madrid    16.77829847
## Melilla   -2.21794591
## Murcia    1.52219165
## Málaga    2.00578012
## Navarra   -0.65330322
## Ourense   -1.96532314
## Palencia  -2.12245758
## Palmas    0.09184283
## Pontevedra 0.03605515
## Rioja     -1.38317512
## Salamanca -1.61199535
## Santa Cruz -0.02919167
## Segovia   -1.93110326
## Sevilla   1.94829323
## Soria     -2.39905891
## Tarragona 0.17541169
## Teruel    -2.18541694
## Toledo    -0.46148674
## Valencia  4.76983184
## Valladolid -1.00710369
## Zamora   -2.28666239
## Zaragoza  0.11540844
## Ávila     -2.15151156
```

```
# g.3)
y=data[,1]
x= rownames(data)

# g4)
ggplot(data, aes(x=x, y=y)) +
  geom_segment( aes(x=x, xend=x, yend=0) ) +
  geom_point( size=5, color="red", fill=alpha("orange", 0.3), alpha=0.7, shape=21, stroke=2) + theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

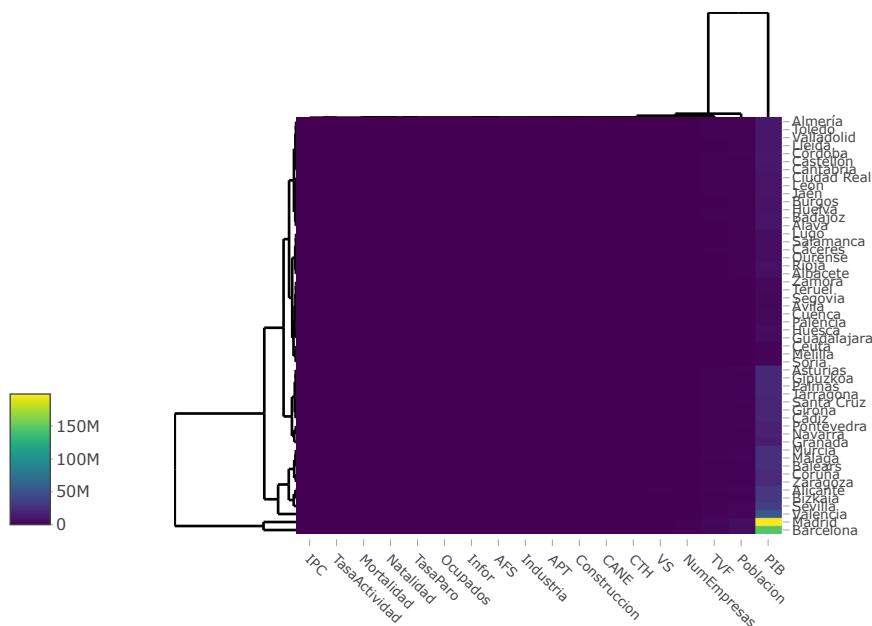


Con un índice que valore de forma conjunta el desarrollo económico de una provincia:

- El valor de dicho índice en Madrid es 16,77.
- El valor de dicho índice en Melilla es -2,21.

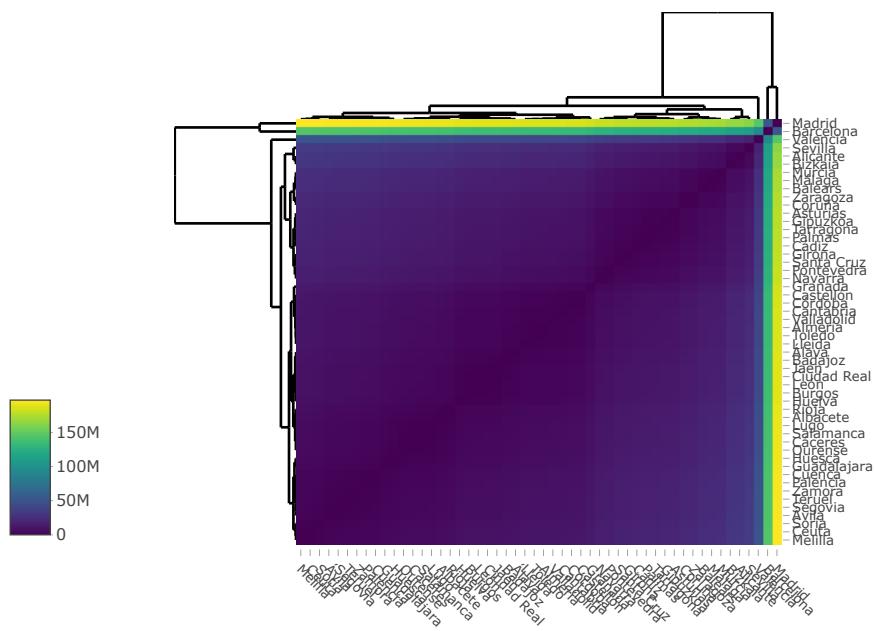
## 4. Representar un mapa de calor de la matriz de datos, estandarizado y sin estandarizar para ver si se detectan inicialmente grupos de provincias.

```
# 4.1.1)
heatmaply(prov, seriate = "mean", row_dend_left = TRUE, plot_method = "plotly")
```

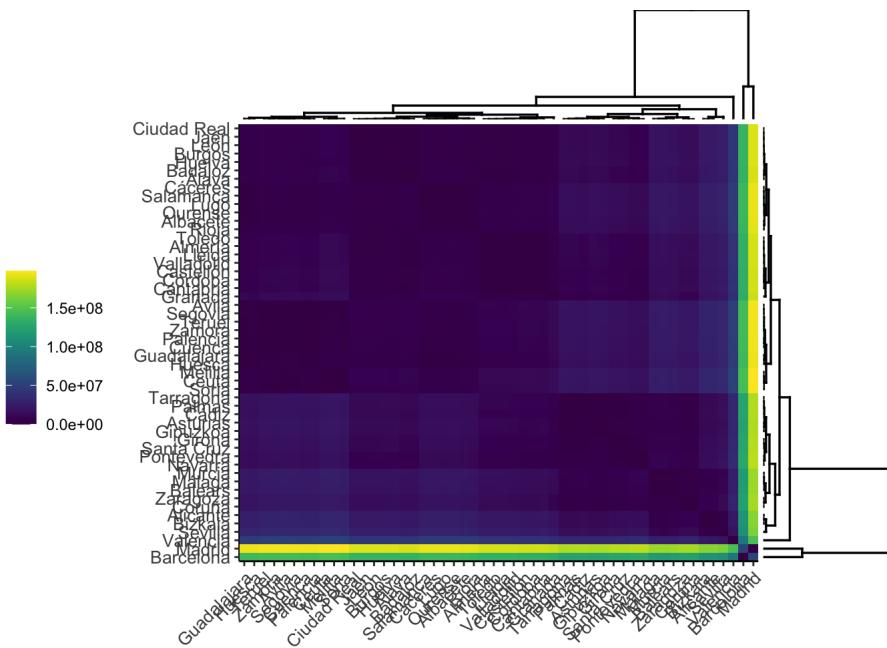


```
# 4.1.2)
d <- dist(prov, method = "euclidean")

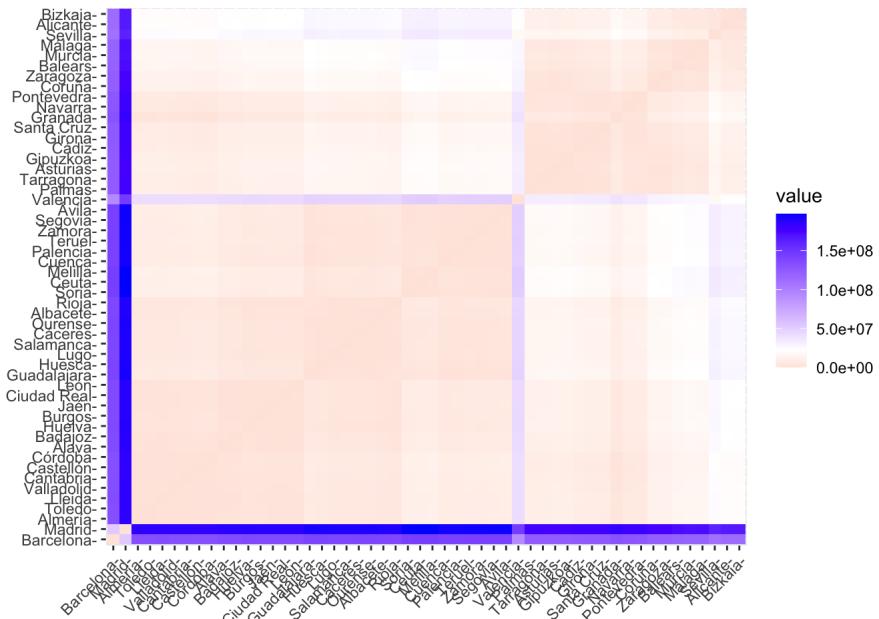
# 4.1.3)
heatmaply(as.matrix(d), seriate = "OLO", row_dend_left = TRUE, plot_method = "plotly")
```



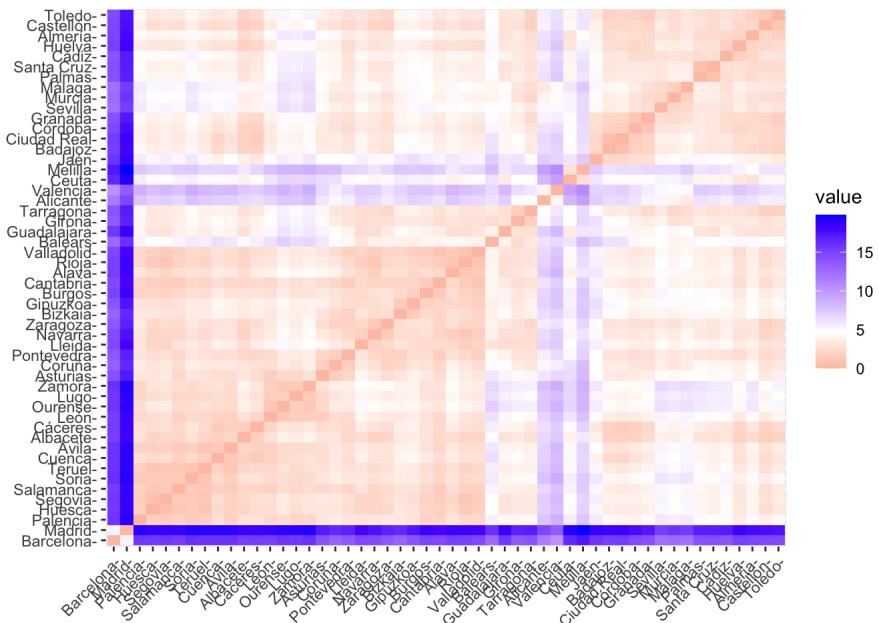
```
# 4.1.4)
ggheatmap(as.matrix(d), seriate="mean")
```



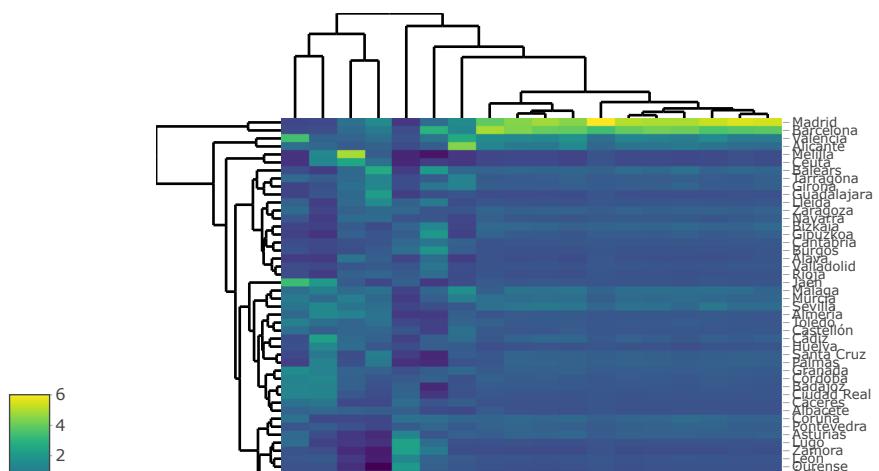
```
# 4.1.5)
fviz_dist(d, show_labels = TRUE)
```

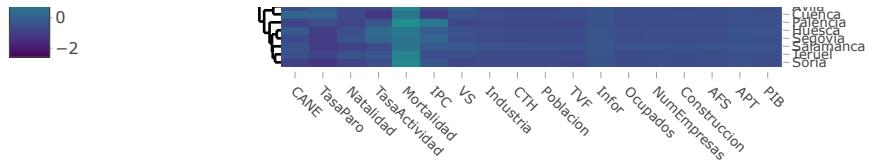


```
#Standardize the data  
datos_ST <- scale(prov)  
  
# 4.2.1)  
d_st <- dist(datos_ST, method = "euclidean")  
  
# 4.2.2)  
fviz_dist(d_st)
```

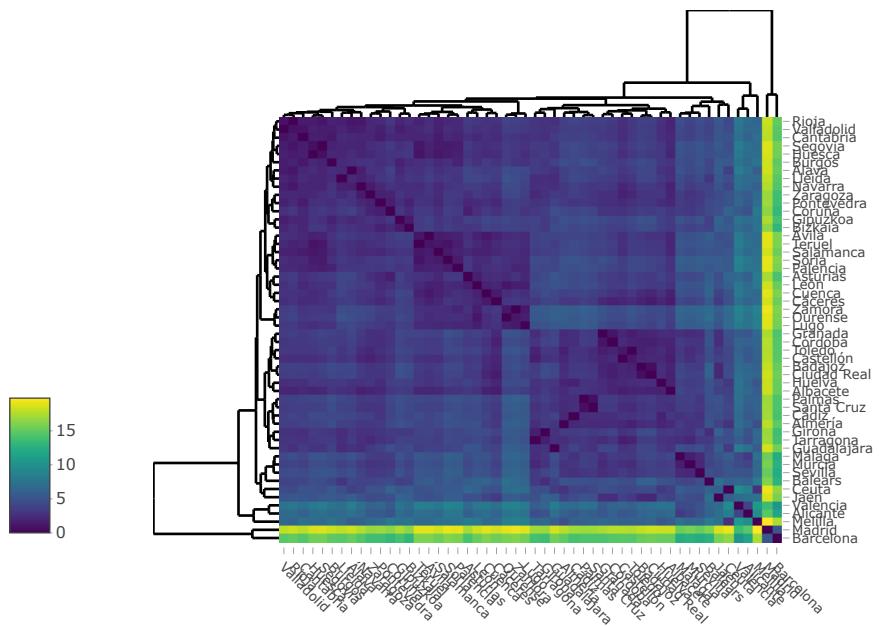


```
# 4.2.3)  
heatmaphy(as.matrix(datos_ST), seriate = "mean", row_dend_left = TRUE, plot_method = "plotly")
```

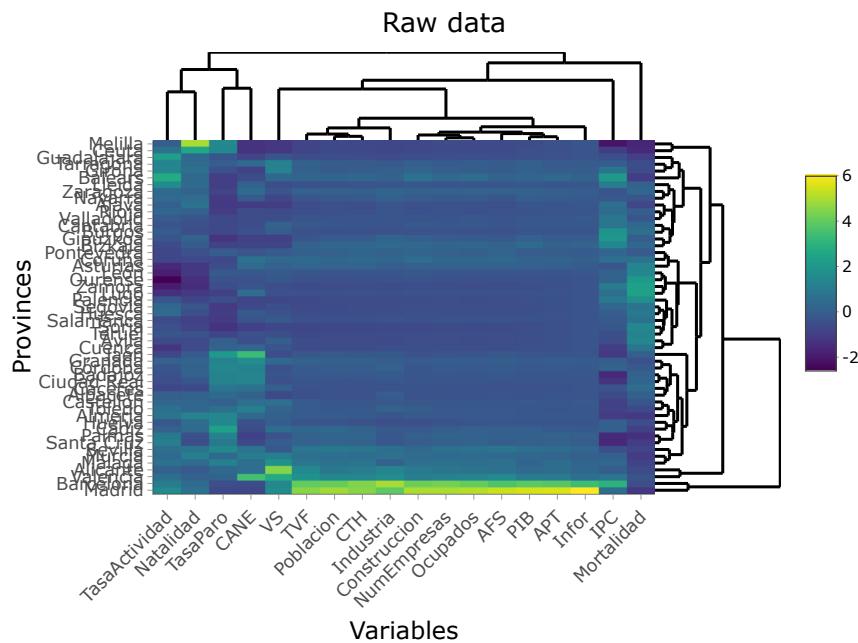




```
# 4.2.4)
heatmaply(as.matrix(d_st), seriate = "mean", row_dend_left = TRUE, plot_method = "plotly")
```



```
# 4.2.5)  
heatmaphy(scale(prov), xlab = "Variables", ylab = "Provinces", main = "Raw data")
```



En el mapa de calor de la matriz de datos, tanto estandarizado como sin estandarizar, observamos que se detectan inicialmente cuatro grupos de provincias:

- Grupo 1: Madrid y Barcelona.
  - Grupo 2: Entre otras provincias, Valencia y Alicante.
  - Grupo 3: Entre otras provincias, Cádiz y Málaga.
  - Grupo 4: Entre otras provincias, Zamora y Soria.

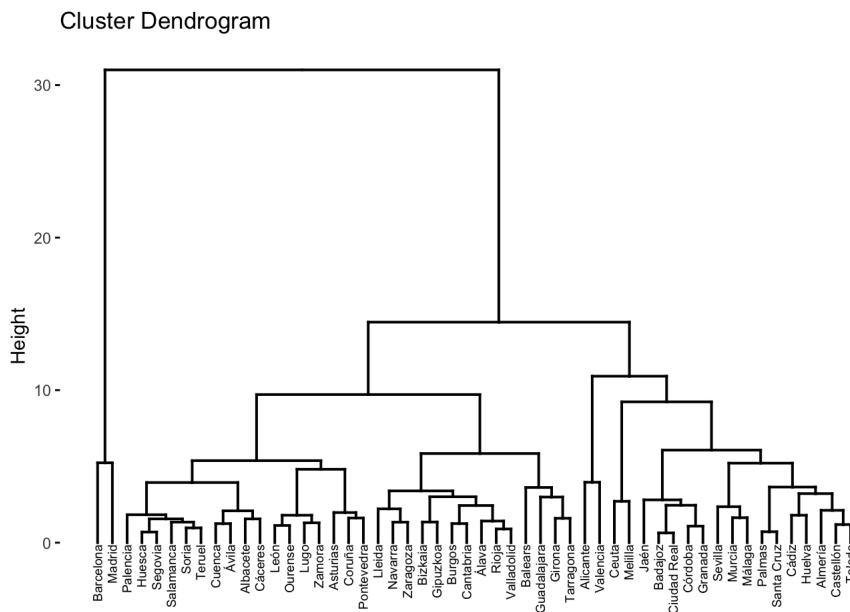
Madrid y Barcelona tienen una distancia muy superior con el resto de las provincias.

5. Realizar un análisis Jerárquico de clusters para determinar si existen grupos de provincias con

comportamiento similar.

## a. A la vista del dendrograma ¿Cuántos clusters recomendarías?

```
# a)
res.hc_st <- hclust(d_st, method="ward.D2")
fviz_dend(res.hc_st, cex = 0.5)
```



-A la vista del dendrograma, el número de clusters recomendado es 4. -Significa que en nuestro caso vemos la composición para k = 4.

## b. Representar los individuos agrupados según el número de clusters elegido.

```
# b.1)
grp <- cutree(res.hc_st, k = 4)

# b.2)
knitr::kable(table(grp), caption = "Número de individuos por cluster")
```

Número de individuos por cluster

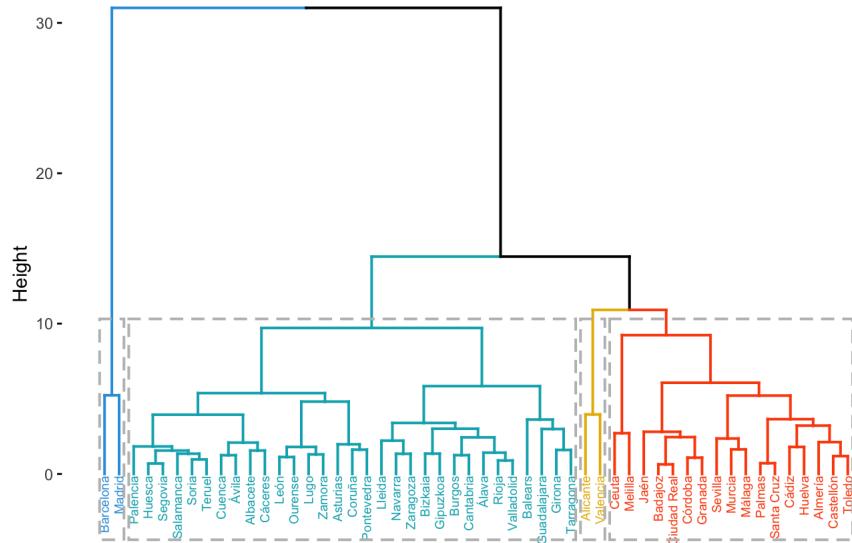
grp	Freq
1	31
2	2
3	17
4	2

```
# b.3)
rownames(prov)[grp == 4]
```

```
## [1] "Barcelona" "Madrid"
```

```
# b.4)
fviz_dend(res.hc_st, k = 4,
          cex = 0.5,
          k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
          color_labels_by_k = TRUE,
          rect = TRUE)
```

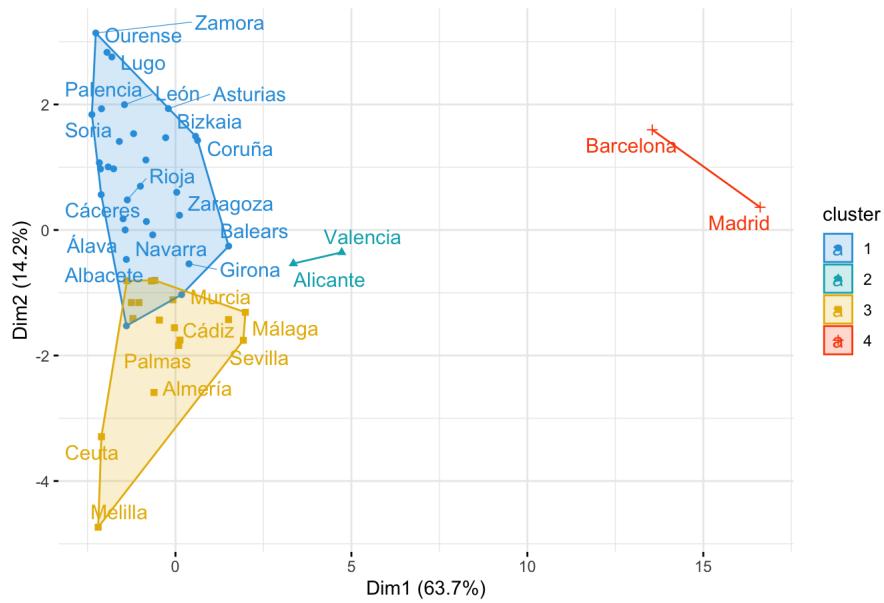
### Cluster Dendrogram



```
# b.5)
fviz_cluster(list(data = datos_ST, cluster = grp),
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type = "convex", repel = TRUE,
             show.clust.cent = FALSE, ggtheme = theme_minimal())
```

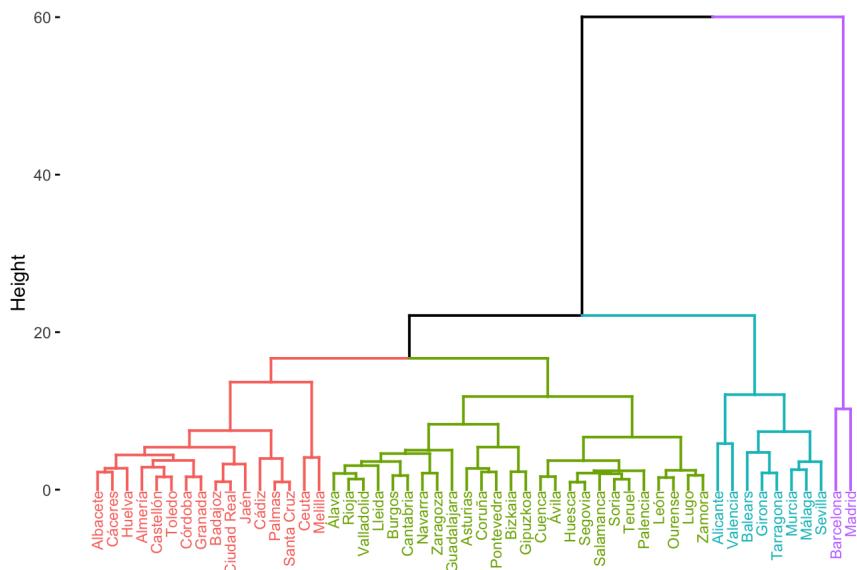
## Warning: ggrepel: 23 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps

### Cluster plot



```
# b.6)
res.agnes <- agnes(x = prov,
                     stand = TRUE,
                     metric = "euclidean",
                     method = "ward")
# b.7)
fviz_dend(res.agnes, cex = 0.6, k = 4)
```

### Cluster Dendrogram

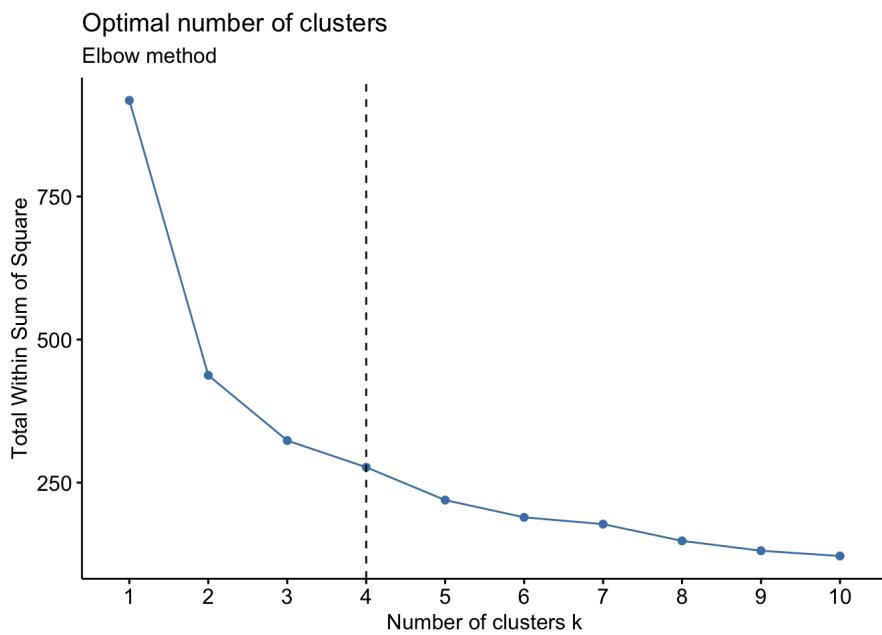


Representamos los individuos agrupados según el número de clusters elegido, destacados con colores diferentes, mediante la función fviz\_dend.

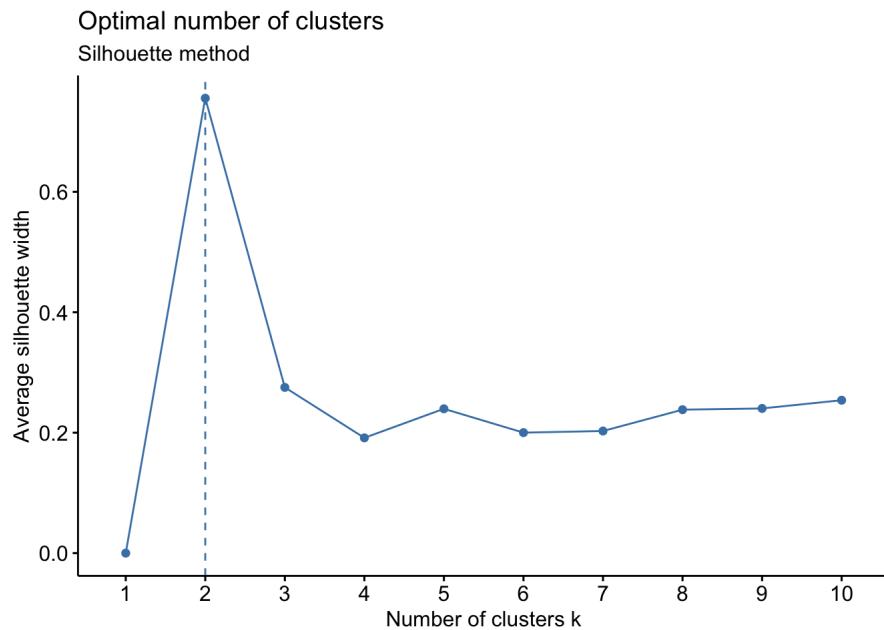
La función agnes, que directamente estandariza, calcula las distancias entre individuos y realiza el cluster jerárquico. Observemos que la función Agnes actúa directamente sobre los datos, no sobre la matriz de distancias. Por ello, la función no agrupa exactamente las mismas provincias, sobre todo en el cluster 2, ya que antes solo estaban Valencia y Alicante.

## c. ¿Qué número óptimo de clusters nos indican los criterios Silhouette y de Elbow?

```
# c.1) Elbow
fviz_nbclust(datos_ST, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2) +
  labs(subtitle = "Elbow method")
```



```
# C.2) Silhouette
fviz_nbclust(datos_ST, kmeans, method = "silhouette") +
  labs(subtitle = "Silhouette method")
```



Los criterios Silhouette y de Elbow nos indican el número óptimo de clusters.

- Basándonos en Elbow, el número óptimo de clusters es 4.
- Basándonos en Silhouette, el número óptimo de clusters es 2.

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

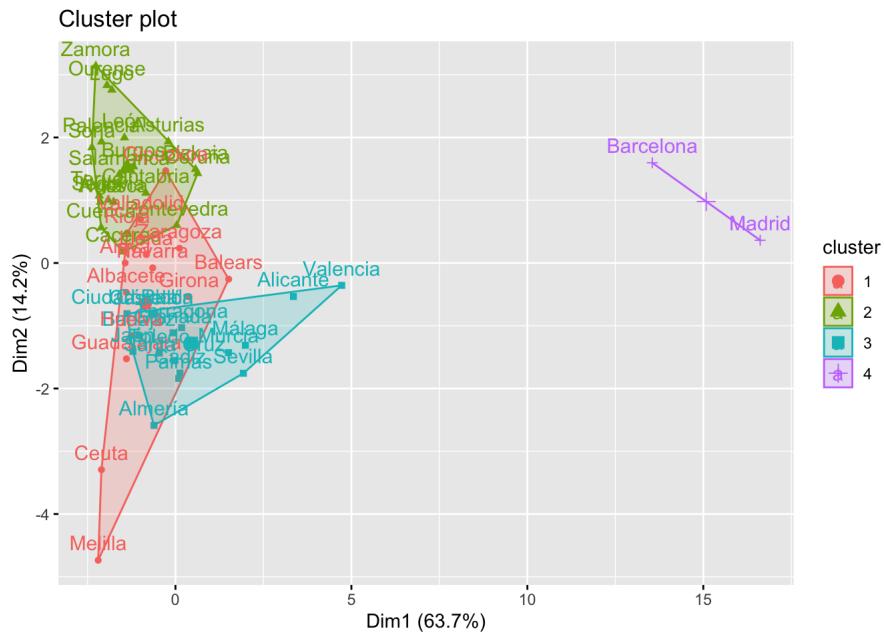
**i. Representar los clusters formados en los planos de las Componentes principales. Relacionar la posición de cada cluster en el plano con lo que representa cada componente principal.**

```
# i.1)
RNGkind(sample.kind = "Rejection")
set.seed(1234)

# i.2)
km.res <- kmeans(datos_ST, 4)
head(km.res$cluster, 20)
```

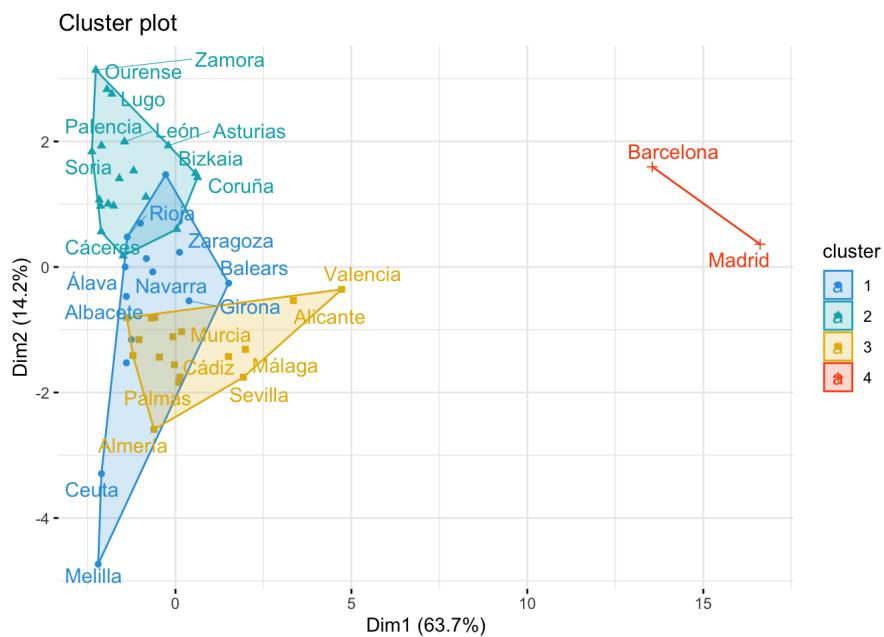
```
##    Albacete    Alicante    Almería    Álava    Asturias    Badajoz
##      1           3           3          1          2          3
##    Baleares   Barcelona   Bizkaia   Burgos   Cantabria   Castellón
##      1           4           2          2          2          3
##    Ceuta     Ciudad Real   Coruña   Cuenca   Cáceres     Cádiz
##      1           3           2          2          2          3
##    Córdoba    Gipuzkoa
##      3           1
```

```
# i.3)
fviz_cluster(km.res, datos_ST)
```



```
# i.4)
fviz_cluster(km.res, datos_ST, palette = c("#E9FDF", "#00AFBB", "#E7B800", "#FC4E07"), ellipse.type = "convex", r
epel = TRUE, show.clust.cent = FALSE, ggtheme = theme_minimal())
```

```
## Warning: ggrepel: 23 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

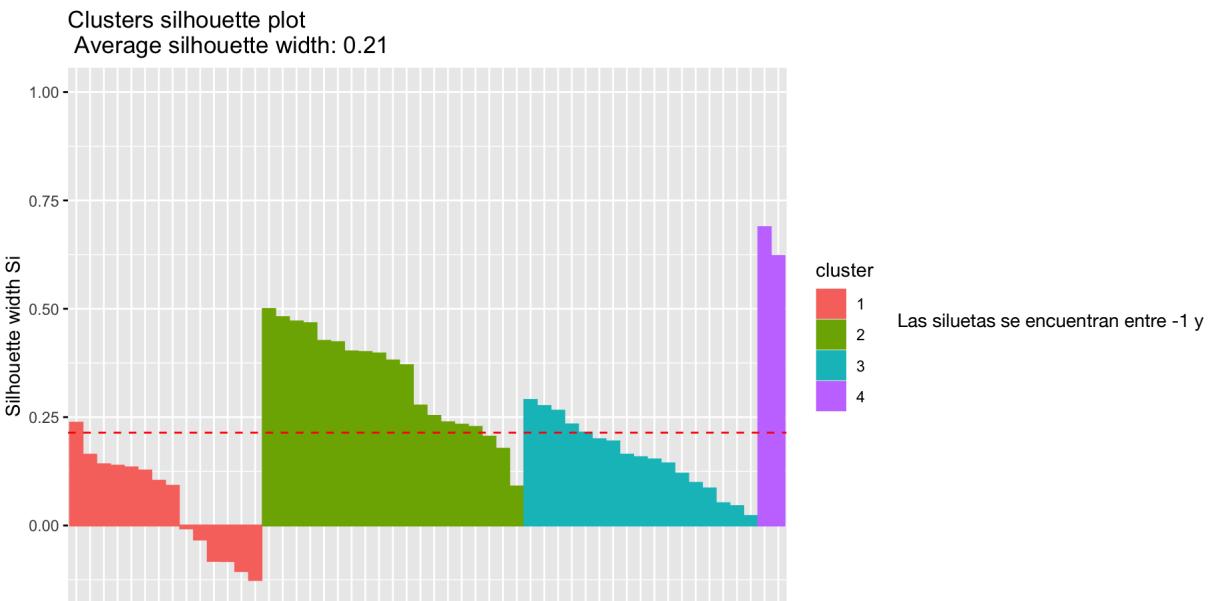


- Cluster 1 se sitúa principalmente en la parte negativa se debe al hecho de que tiene una tasa de paro importante, pero una tasa de natalidad positiva. En ciertas provincias, como Baleares, tienen una tasa de Actividad positiva, por ello el cluster se sitúa en el medio, más o menos entre el cluster 2 y 3.
- Cluster 2, que se posiciona casi al extremo, con entre otras provincias Jaén y Zamora, presenta una tasa importante de TasaParo y poca población, natalidad, TasaActividad, de acuerdo con lo que representa la componente 4. Se sitúa más cerca del primer cluster que del tercer cluster, y tiene casi todo en negativo.
- Cluster 3, con entre otras provincias, Sevilla, Valencia y Alicante, se posiciona en el medio, con una mayoridad en positivo ya que presenta una tasa importante de Natalidad, CTH, TasaCtividad, sobre todo de CANE, así como poca mortalidad. La distancia entre las provincias de este cluster y las otras es mediana aunque se posiciona más cerca del primer cluster que del cuarto, de acuerdo con lo que representa la componente 2.
- Cluster 4, con Barcelona y Madrid, presenta una alta correlación con la primer componente, lo que podría significar que las provincias tengan un número importante de empresas y contribuyan de manera significativa al PIB. La distancia entre las provincias de este cluster y las otras se destaca claramente en los gráficos.

## ii. Evaluar la calidad de los clusters.

```
# ii)
sil <- silhouette(km.res$cluster, dist(datos_ST))
rownames(sil) <- rownames(prov)
fviz_silhouette(sil)
```

```
##   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
```

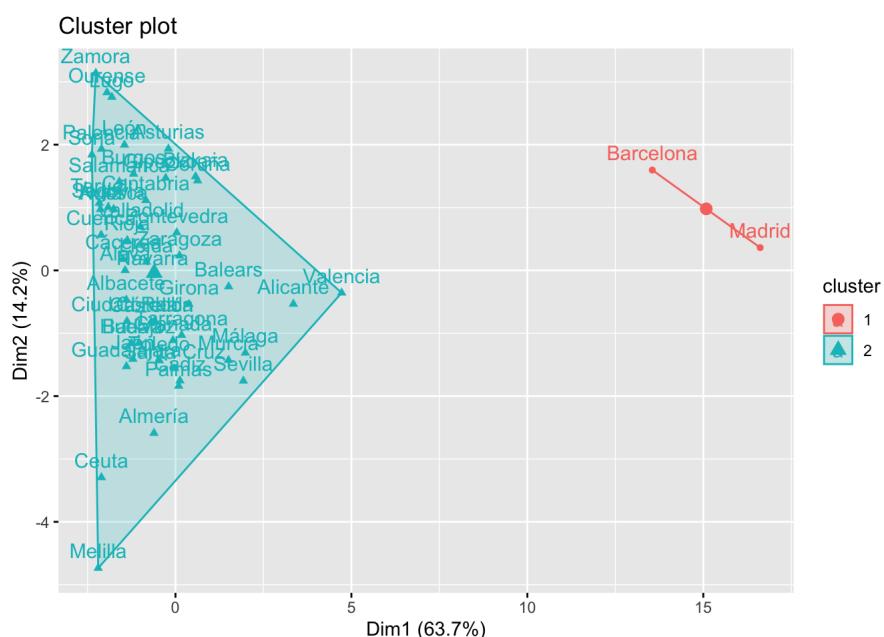


1. Si la silueta está próxima a 1 eso querría indicar que la observación se encuentra bien agrupada, mientras que si vale 0 indica que la observación podría pertenecer a su cluster actual o a otro cercano a él. Si la silueta es negativa indicaría una mala agrupación para la observación.

En nuestro ejemplo, la mayoría de las provincias están bien agrupadas, aunque haya una parte negativa en el primer cluster. De hecho, el cluster 1 está mal clasificado porque su silueta es negativa.

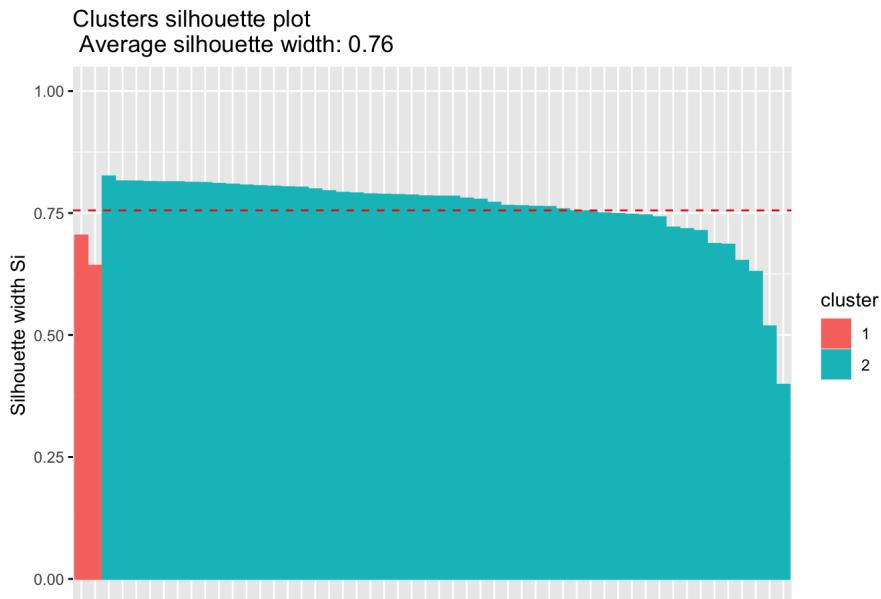
Por ello, probamos de nuevo, con 2 Clusters, que es lo que nos recomienda el criterio Silhouette.

```
RNGkind(sample.kind = "Rejection")
set.seed(1234)
km.res2 <- kmeans(datos_ST, 2)
fviz_cluster(km.res2, datos_ST)
```



```
sil2 <- silhouette(km.res2$cluster, dist(datos_ST))
fviz_silhouette(sil2)
```

```
##   cluster size ave.sil.width
## 1       1     2       0.67
## 2       2    50       0.76
```



Se aprecia que ahora no existen observaciones con valor de la silueta negativo.

**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.**

```
# e.1)
ordenado<-sort(km.res$cluster)
knitr::kable(ordenado, digits =2, caption = "Provincias y clusters")
```

Provincias y clusters

	x
Albacete	1
Álava	1
Balears	1
Ceuta	1
Gipuzkoa	1
Girona	1
Guadalajara	1
Huelva	1
Lleida	1
Melilla	1
Navarra	1
Rioja	1
Valladolid	1
Zaragoza	1
Asturias	2
Bizkaia	2
Burgos	2
Cantabria	2
Coruña	2
Cuenca	2

x

Cáceres	2
Huesca	2
León	2
Lugo	2
Ourense	2
Palencia	2
Pontevedra	2
Salamanca	2
Segovia	2
Soria	2
Teruel	2
Zamora	2
Ávila	2
Alicante	3
Almería	3
Badajoz	3
Castellón	3
Ciudad Real	3
Cádiz	3
Córdoba	3
Granada	3
Jaén	3
Murcia	3
Málaga	3
Palmas	3
Santa Cruz	3
Sevilla	3
Tarragona	3
Toledo	3
Valencia	3
Barcelona	4
Madrid	4

```
print(km.res)
```

```

## K-means clustering with 4 clusters of sizes 14, 19, 17, 2
##
## Cluster means:
##   Poblacion Mortalidad Natalidad      IPC NumEmpresas Industria
## 1 -0.3383277 -0.5613613  0.7516305  0.1748911 -0.29381802 -0.2596145
## 2 -0.3917379  1.0537407 -0.7800285  0.2113687 -0.34577315 -0.3744500
## 3  0.1977043 -0.6092292  0.1925114 -0.5790431  0.09089903  0.1157796
## 4  4.4093174 -0.9025593  0.5125104  1.6896263  4.56892924  4.3904501
##   Construccion CTH     Infor      AFS      APT TasaActividad
## 1 -0.26863161 -0.3451124 -0.22194959 -0.3222065 -0.25573380  0.6055338
## 2 -0.31869710 -0.3735719 -0.26468021 -0.3406618 -0.31756417 -0.8366415
## 3  0.04498425  0.1819546 -0.07211276  0.1219692  0.02023994  0.2904450
## 4  4.52567761  4.4181054  4.68106759  4.4549948  4.63495671  1.2405756
##   TasaParo Ocupados      PIB      CANE      TVF      VS
## 1 -0.3207477 -0.2765890 -0.23519683 -0.5932634 -0.3914030 -0.5032386
## 2 -0.5763826 -0.3655866 -0.32545922 -0.2710301 -0.3882567 -0.3051687
## 3  0.9829029  0.1052622  0.01217502  0.8729416  0.2610450  0.5807317
## 4 -0.6338058  4.5144671  4.63475280 -0.6923738  4.2093774  1.4855534
##
## Clustering vector:
##   Albacete    Alicante    Almería    Álava    Asturias    Badajoz
## 1          1          3          3          1          2          3
##   Balears    Barcelona    Bizkaia    Burgos    Cantabria    Castellón
## 1          1          4          2          2          2          3
##   Ceuta    Ciudad Real    Coruña    Cuenca    Cáceres    Cádiz
## 1          3          2          2          2          2          3
##   Córdoba    Gipuzkoa    Girona    Granada    Guadalajara    Huelva
## 3          1          1          1          3          1          1
##   Huesca     Jaén        León    Lleida      Lugo    Madrid
## 2          2          3          2          1          2          4
##   Melilla    Murcia    Málaga    Navarra    Ourense    Palencia
## 1          3          3          3          1          2          2
##   Palmas    Pontevedra    Rioja    Salamanca    Santa Cruz    Segovia
## 3          2          1          2          2          3          2
##   Sevilla    Soria    Tarragona    Teruel    Toledo    Valencia
## 3          2          3          2          2          3          3
##   Valladolid    Zamora    Zaragoza    Ávila
## 1          2          1          1          2
##
## Within cluster sum of squares by cluster:
## [1] 91.28026 58.04058 117.64328 13.70956
## (between_SS / total_SS =  69.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"       "totss"         "withinss"      "tot.withinss"
## [6] "betweenss"    "size"          "iter"          "ifault"

```

```

# e.2)
knitr::kable(km.res$centers, digits = 2,caption = "Estadísticos de los clusters, datos STD")

```

#### Estadísticos de los clusters, datos STD

Poblacion	Mortalidad	Natalidad	IPC	NumEmpresas	Industria	Construccion	CTH	Infor	AFS	APT	TasaActividad	TasaParo	Ocupados	
-0.34	-0.56	0.75	0.17		-0.29	-0.26	-0.27	-0.35	-0.22	-0.32	-0.26	0.61	-0.32	-0.28
-0.39	1.05	-0.78	0.21		-0.35	-0.37	-0.32	-0.37	-0.26	-0.34	-0.32	-0.84	-0.58	-0.37
0.20	-0.61	0.19	-0.58		0.09	0.12	0.04	0.18	-0.07	0.12	0.02	0.29	0.98	0.11
4.41	-0.90	0.51	1.69		4.57	4.39	4.53	4.42	4.68	4.45	4.63	1.24	-0.63	4.51

```

# e.3)
Est_Clus<-aggregate(prov, by=list(km.res$cluster),mean)
knitr::kable(Est_Clus, digits = 2,caption = "Estadísticos de los clusters")

```

#### Estadísticos de los clusters

Group.1	Poblacion	Mortalidad	Natalidad	IPC	NumEmpresas	Industria	Construccion	CTH	Infor	AFS	APT	TasaActivid
1	508918.1	8.19	10.45	102.49	34689.71	2553.93	4991.21	13327.64	469.79	711.86	5635.57	60
2	447266.9	11.61	7.17	102.52	29986.74	1999.32	4466.84	12468.89	342.32	673.68	4374.00	54
3	1127658.4	8.09	9.25	101.88	69514.29	4366.94	8275.94	29231.59	916.76	1630.59	11266.47	59
4	5989112.0	7.46	9.94	103.73	474865.50	25012.00	55205.50	157055.00	15096.00	10593.00	105424.00	62

- Cluster 1: Las características socioeconómicas que hacen pertenecer a las provincias a dicho cluster es el hecho de tener una fuerte Natalidad positiva y una mortalidad negativa. Tienen una TasaActividad positiva y una Tasa Paro negativa, por lo que suponemos que hay una cierta actividad económica, pero donde la población no puede permanecer por falta de empresas y de industria en las áreas. De hecho, el cluster 1 tiene datos similares a los del cluster 2.

- Cluster 2: Las características socioeconómicas que hacen a las provincias pertenecer a dicho cluster se reflejan claramente en el hecho de que se trata de las provincias de industrialización inducida y escasa, con poca natalidad y poca tasa de actividad. De hecho, la llamamos la España 'vacía'. Tiene todo en negativo; salvo en IPC y sobre todo la Mortalidad, lo que confirma que se trata de provincias con poca actividad. Por ello, el cluster 2 es el único con tantas variables negativas.
- Cluster 3: Las características socioeconómicas que hacen a las provincias pertenecer a dicho cluster son claramente el hecho de tener espacios agrícolas productivos (tasa CANE importante) y formar parte de áreas en desarrollo que contribuyen al PIB de manera positiva, aunque en poca medida. Se aprecia que la tasa de paro es positiva, lo que confirma que las provincias no están completamente desarrolladas, sino en expansión, teniendo en cuenta que la tasa Comercio, transporte y hostelería y la tasa actividad son positivas y favorables. En efecto, las medias de las variables en el cluster 3 son casi todas positivas, pero se quedan lejos de las medias del cluster 4.
- Cluster 4: El tercer cluster reagrupa las provincias con un fuerte desarrollo económico debido al número de empresas y la población importante. Con una contribución alta al PIB, una tasa de industria, construcción, CTH y Infor importante se trata de provincias dinámicas y consolidadas. Por ello, las medias de las variables en el cluster 4 son mayores que la de tres otros clusters.