Ejercicio de Evaluacion Parte I

Christopher Angel Rodriguez

FEB 8 2022

Cargamos Datos:

```
#Cargamos Datos de excel
Provincias <- read_excel("~/Desktop/DS UCM /Modulo 7/Evaluacion/Provincias.xlsx")

#Transformamos a dataframe
Provincias<-as.data.frame(Provincias)

#indicamos los nombres de las filas
rownames(Provincias)<-Provincias[,1]

#Eliminamos la primera columna con los nombres de Provincia
Provincias<-Provincias[,-1]
```

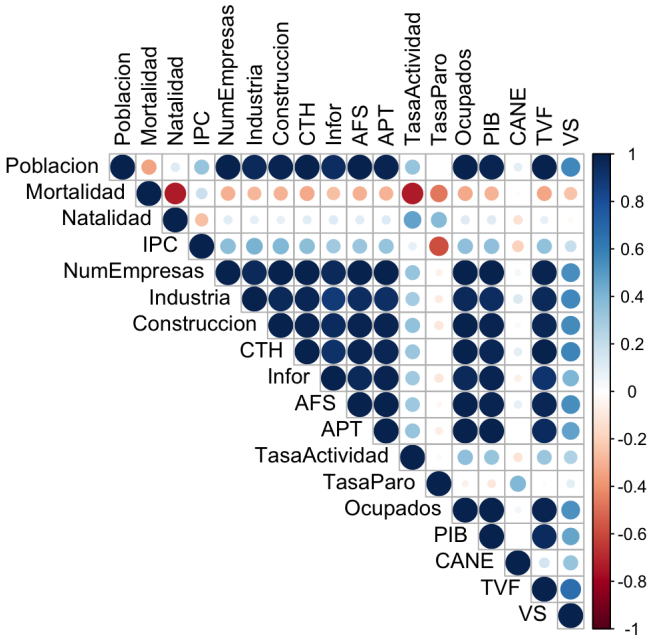
1. Calculamos Matriz de correlaciones y su representacion grafica:

```
tabla_corr<- cor(Provincias,method = 'pearson')
knitr::kable(tabla_corr,digits = 2,caption='Correlaciones')
```

Correlaciones

	Poblacion	Mortalidad	Natalidad	IPC	NumEmpresas	Industria	Construccion	CTH	Infor	AFS	APT	TasaActividad	TasaParo
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	0.39
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

```
corrplot(tabla_corr,type = 'upper', tl.col = 'black',tl.srt = 90)
```

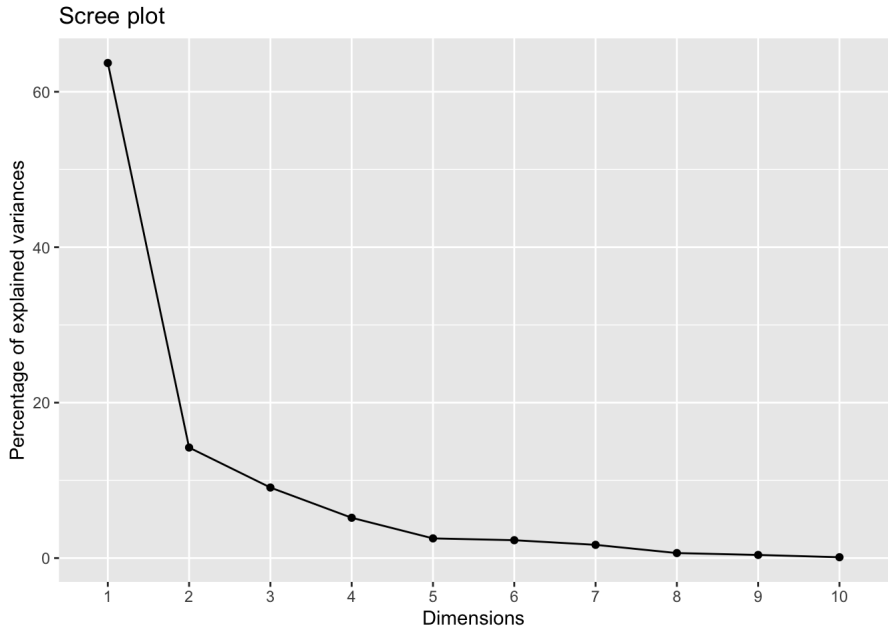


1. Vizualizamos nuestro grafico de Correlaciones e interpretamos que las variables mas correlacionadas de forma inversa son entre: (Mortalidad y Natalidad), (Motalidad y TasaActividad), (ICP y TasaParo). A pesar de que existen otras correlaciones inversas entre variables, tres variables con correlacion inversa más cercano a 1.

2. Analisis de componentes principales sobre la matriz:

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

```
# Screen plot1
fviz_eig(fit, geom="line")+ theme_grey()
```




```
knitr::kable(fit$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

	eigenvalue	percentage of variance	cumulative percentage of variance
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

2. En el ejercicio #2 he elegido continuar con cuatro componentes principales las cuales serán suficiente. Dichas variables componen una explicacion de un 92.19% de 'cumulative percentage of variance'.

3. Analisis sobre matriz de correlaciones ajustado al numero de componentes principales que he seleccionado (4).

```
fit<-PCA(Provincias,scale.unit=TRUE, ncp=4, graph=FALSE)
```

3.A Mostrar los coeficientes para obtener las componentes principales

```
knitr::kable(fit$svd$V, col.names =c("autov1","autov2","autov3","autov4"), digits =2,caption = "Autovectores")
```

Autovectores

autov1	autov2	autov3	autov4
0.29	0.00	0.05	-0.05
-0.11	-0.53	0.19	-0.16
0.04	0.50	-0.27	-0.11
0.11	-0.37	-0.26	0.44
0.29	-0.03	0.01	-0.07
0.29	-0.04	0.05	0.02
0.29	-0.05	-0.01	-0.03
0.29	-0.01	0.05	-0.03
0.28	-0.04	-0.06	-0.22
0.29	-0.02	0.04	-0.09
0.29	-0.03	-0.03	-0.14
0.11	0.33	-0.36	0.46
-0.01	0.46	0.39	-0.22
0.29	-0.02	0.00	-0.06
0.29	-0.04	-0.04	-0.13
0.02	0.10	0.66	0.28
0.29	0.00	0.10	0.04
0.17	0.05	0.29	0.57

3.A Expresion matematica para calcular el primer componente en funcion de las variables originales: (nos dejamos llevar por la lista de valores en autov1) $CP_1 = 0.29X_1^* - 0.11X_2^* + \cdots + 0.17X_{18}^*$

3.B Mostar una tabla con las correlaciones de las Variables con las Componentes Principales.

```
var<-get_pca_var(fit)
knitr::kable(var$cor,digits = 3,caption = 'Correlaciones de la CP con las variables')
```

Correlaciones de la CP con las variables

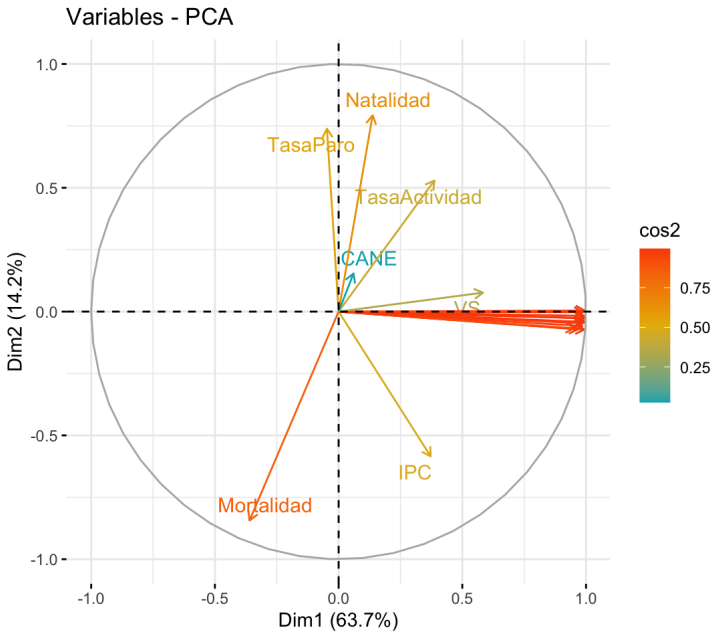
	Dim.1	Dim.2	Dim.3	Dim.4
Poblacion	0.994	0.004	0.064	-0.052
Mortalidad	-0.360	-0.843	0.241	-0.156
Natalidad	0.138	0.793	-0.346	-0.106
IPC	0.372	-0.585	-0.335	0.420
NumEmpresas	0.996	-0.042	0.010	-0.067
Industria	0.967	-0.072	0.059	0.023
Construccion	0.993	-0.073	-0.015	-0.026
CTH	0.992	-0.017	0.062	-0.027
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

3.B La componente 1 esta más correlacionada positivamente con [Poblacion,NumEmpresas,Industria,Construccion,CTH,Infor,AFS,APT,Ocupados,PIB,TVF] La componente 2 esta más correlacionada negativa importnte con [mortalidad] La componente 3 esta más correlacionada positivamente con [CANE] La componente 4 tiene correlacion (aunque debil es la mayor en la componente) con [VS]

3.C Comentar los gráficos que representan las variables en los planos formados por las componentes

Representacion grafica de variables

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



3.C La primera componente recoge los valores de [Poblacion,NumEmpresas,Industria,Infor,AFS,APT,PIB,TVF,Construccion,CTH,Ocupados] todas estas variables y tienen correlaion positiva con la componente 1. La componente 2 solo tiene 1 variable con alta correlacion negativa con [Mortalidad] Mientras que [CANE] es la variable peor representada etre las componentes.

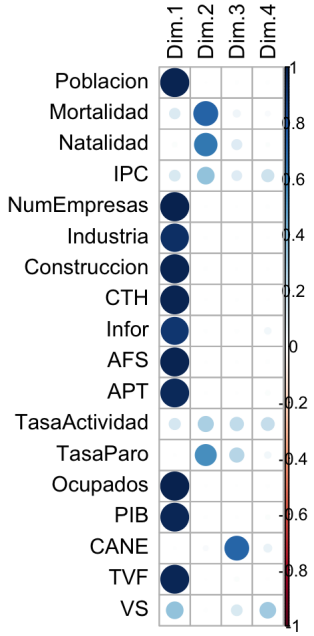
3.D Mostrar la tabla y los gráficos que nos muestran la proporción de la varianza de cada variable que es explicado por cada componente.

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

Cosenos al cuadrado

	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

```
corrplot(var$cos2,is.corr= TRUE,tl.col = 'black')
```



3.D De todas la peor explicada es VS con un Coseno al cuadrado de 0.49.

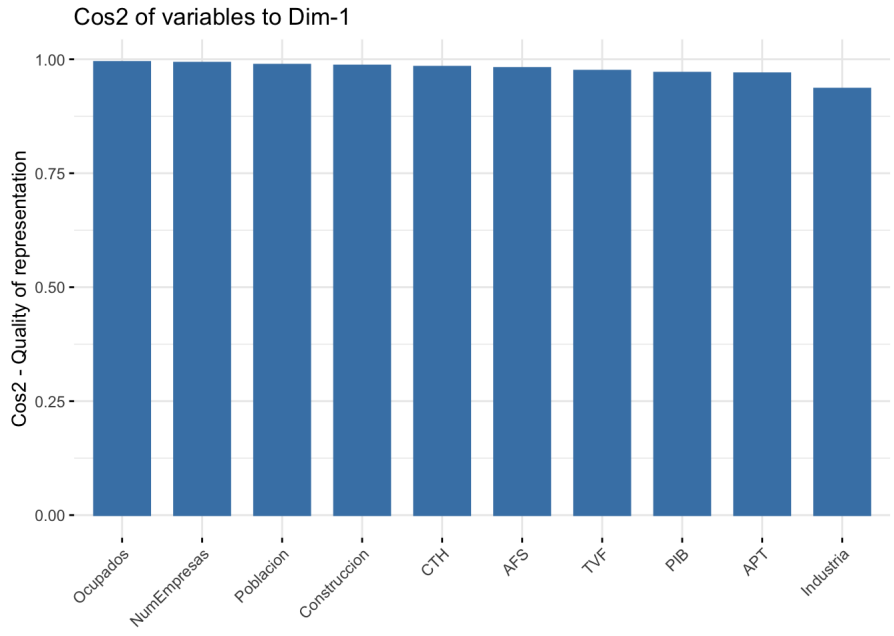
3.E Mostrar la tabla y los gráficos que nos muestran el porcentaje de la varianza de cada Componente que es debido a cada variable.

```
knitr::kable(var$contrib,  digits =2,caption = "Contribuciones de las variables")
```

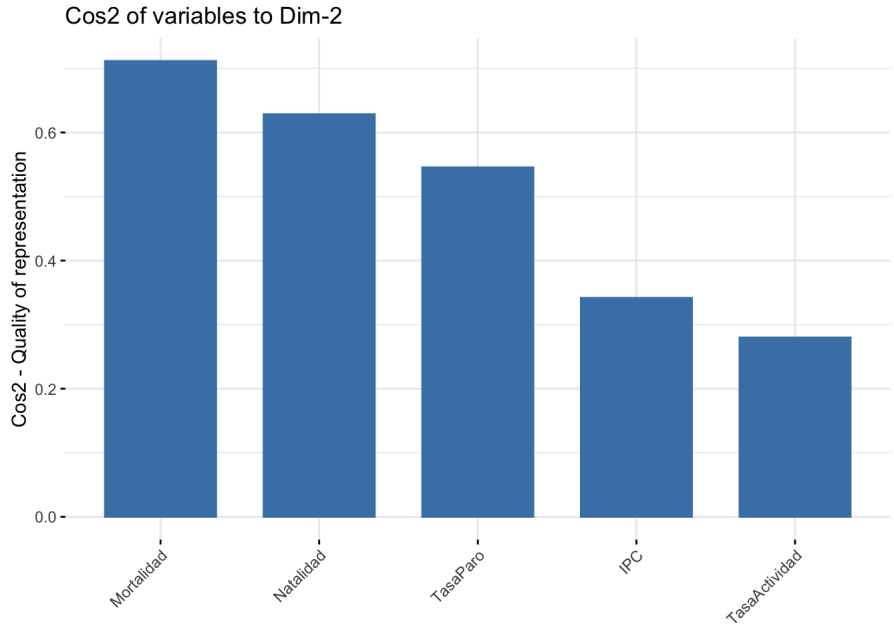
Contribuciones de las variables

	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
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

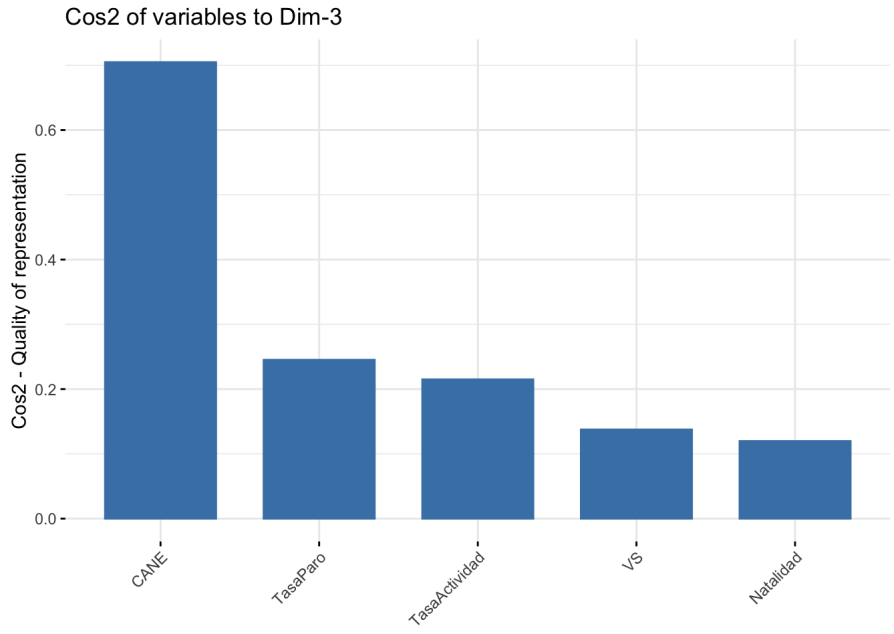
```
fviz_cos2(fit,choice="var",axes=1,top = 10)
```



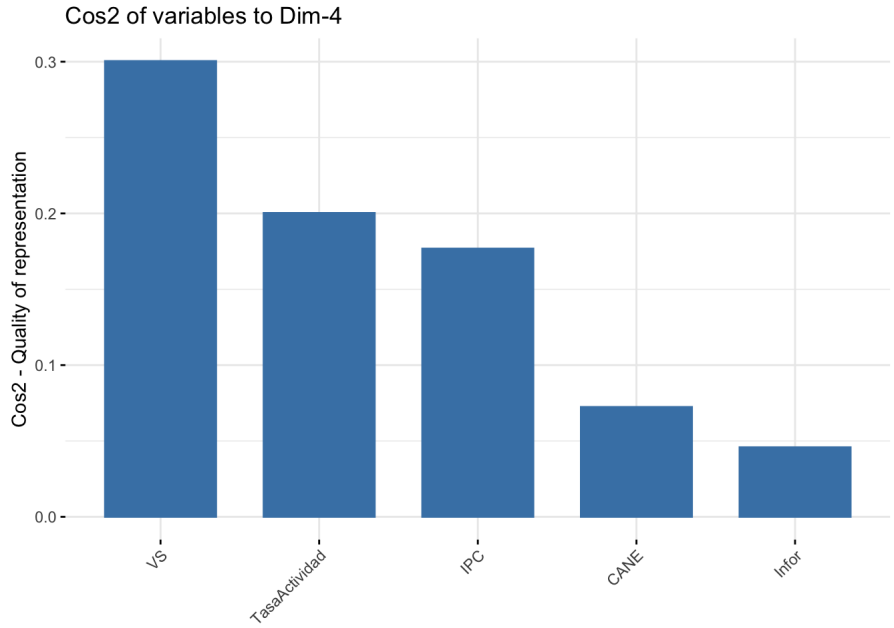

```
fviz_cos2(fit,choice="var",axes=2,top = 5)
```




```
fviz_cos2(fit,choice="var",axes=3,top = 5)
```



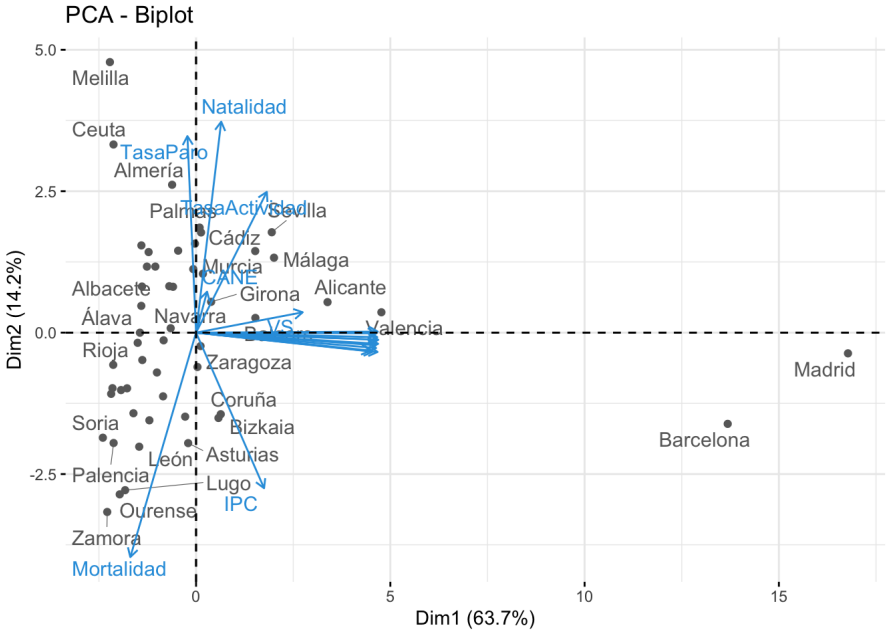

```
fviz_cos2(fit,choice="var",axes=4,top = 5)
```



3.E Las variables que atribuyen mas a la componente 1 son [Construccion,Ocupados,NumEmpresas,CTH,PIB,Poblacion,APT,AFS,Industria y TVF] Las variables que atribuyen mas a la componente 2 son [Natalidad,Mortalidad,TasaParo,IPC,TAsaActividad] Las variables que atribuyen mas a la componente 3 son [CANE,TasaParo,TasaActividad,VS,Natalidad] Las variables que atribuyen mas a la comonente 4 son [VS,TasaActividad,IPC,CANE,Infor]

3.F Comentar las provincias que tienen una posición más destacada en cada componente, en positivo o negativo

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



3.F El grafico nos demuestras que las provincias con indices altos de catidad de empresas, construccion que sucede en la ciudad,industra poblcion y el turismo tienen alto indice de opucacion que trabajan en las empresas por lo tanto esta poblacion gracias a las actividades socioeconomicas tendra un Producto Interno Bruto alto. Estas Provincias son [Madrid,Barcelona,Valencia,Alicante,Balears]. Dado que estas provincias tienen alto nivel de desarrollo,existira alta demanda para vivir en ellas. En contraste de Zamora,Ourse y Lugo que tie tienen indices los indices de mortalida mas altos, se debe a la baja natalidad en la provincia y a la baja tasa de actividad.

3.G Si tuviéramos que construir un índice que valore de forma conjunta el desarrollo económico de una provincia.

una combinación lineal de todas las variables. ¿Cuál sería el valor de dicho índice en Madrid? ¿Cual sería su valor en Melilla?

```
datos_PIBpc <- Provincias
datos_PIBpc$PIBpc <- (datos_PIBpc$PIB/datos_PIBpc$Poblacion)
PIBpc_data <- data.frame(
  provincia = rownames(datos_PIBpc),
  PIBpc = datos_PIBpc$PIBpc
)

PIBpc_data
```

provincia <chr>	**PIBpc** <dbl>
Albacete	18.22727
Alicante	17.20118
Almería	16.97291
Álava	33.28660
Asturias	20.50418
Badajoz	15.31469
Balears	23.71315
Barcelona	26.37289
Bizkaia	27.80319
Burgos	26.40203

1-10 of 52 rows Previous **1** 2 3 4 5 6 Next

3.G Hemos calculado para introducir el indice de Producto Interno Bruto per capita. En resumen, indica el ingreso neto que tiene cada ciudadano en la provincia. El valor del PIB per capita para madrid es 30.7 mientras que en Melilla es 16.5

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

Calculamos las distancias con los valores sin estandarizar

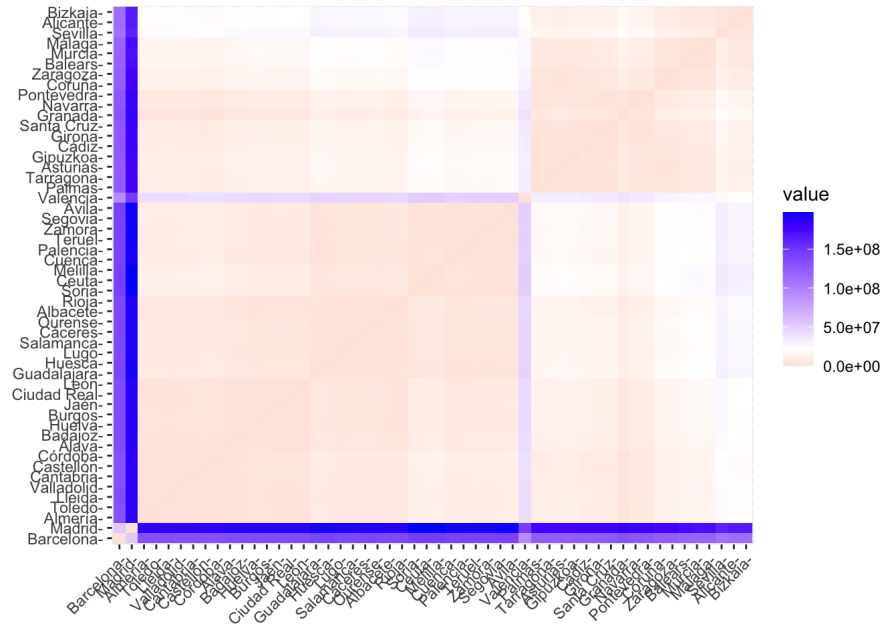
```
d <- dist(Provincias, method = "euclidean") # distance matrix

d6<-as.matrix(d)[1:6, 1:6]

knitr::kable(d6, digits =2,caption = "Distancias")
```

Distancias

	Albacete	Alicante	Almería	Álava	Asturias	Badajoz
Albacete	0	24971242	4687257	3481471.5	14555211	3362039.7
Alicante	24971242	0	20284171	21510859.3	10424490	21610942.1
Almería	4687257	20284171	0	1277176.4	9869791	1328833.4
Álava	3481472	21510859	1277176	0.0	11088936	452219.3
Asturias	14555211	10424490	9869791	11088935.7	0	11197887.0
Badajoz	3362040	21610942	1328833	452219.3	11197887	0.0

Matriz de distancias

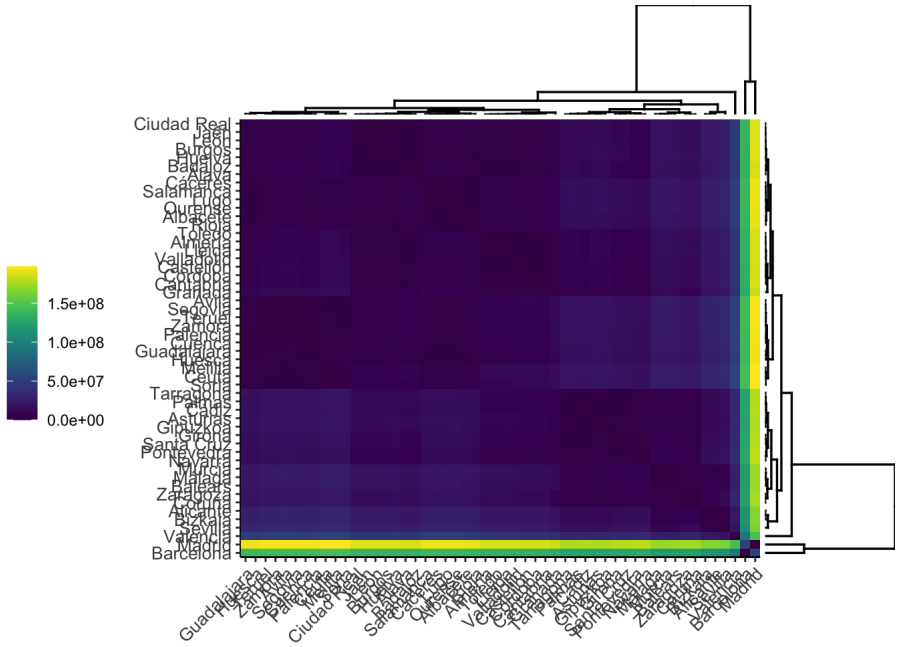
```
fviz_dist(d, show_labels = TRUE)
```



Reordenamos para agrupar las observaciones

Mapa de calor con los datos sin estandarizar

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



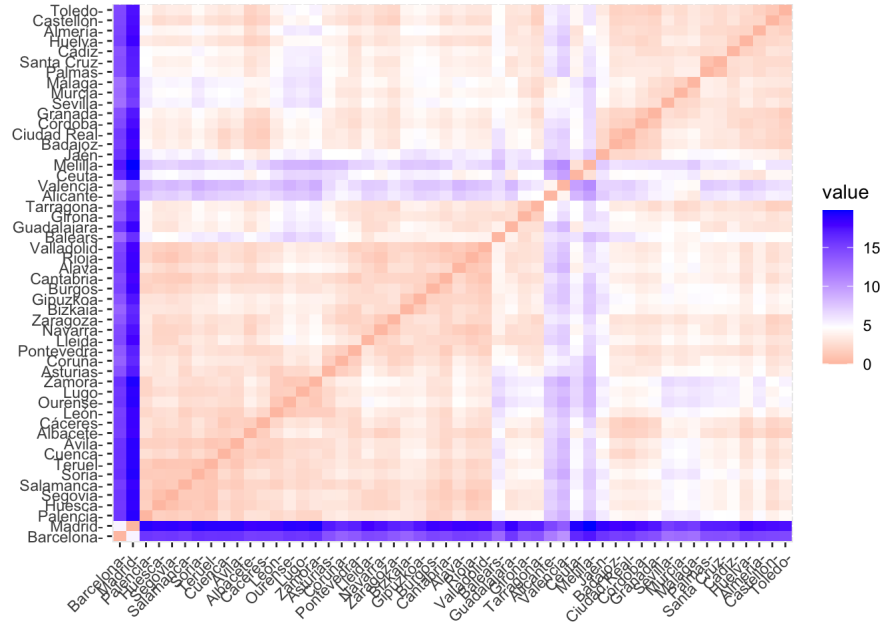
Calculamos las distancias con los valores estandarizados

```
datos_estan <- scale(Provincias)


d_st <- dist(datos_estan, method = "euclidean") # distance matrix
d_st6<-as.matrix(d_st)[1:6, 1:6]
knitr::kable(d_st6, digits =2,caption = "Distancias")
```

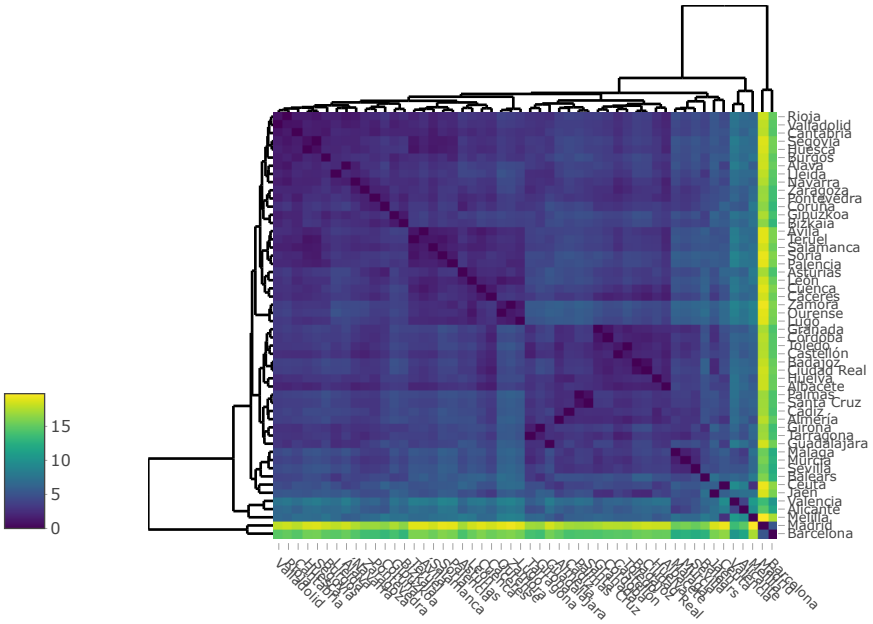
Distancias

	Albacete	Alicante	Almería	Álava	Asturias	Badajoz
Albacete	0.00	6.60	2.37	2.45	3.11	1.92
Alicante	6.60	0.00	6.07	7.35	5.90	6.35
Almería	2.37	6.07	0.00	3.68	4.77	2.39
Álava	2.45	7.35	3.68	0.00	4.20	4.22
Asturias	3.11	5.90	4.77	4.20	0.00	3.37
Badajoz	1.92	6.35	2.39	4.22	3.37	0.00

```
#Matriz de distancias
fviz_dist(d_st)
```




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



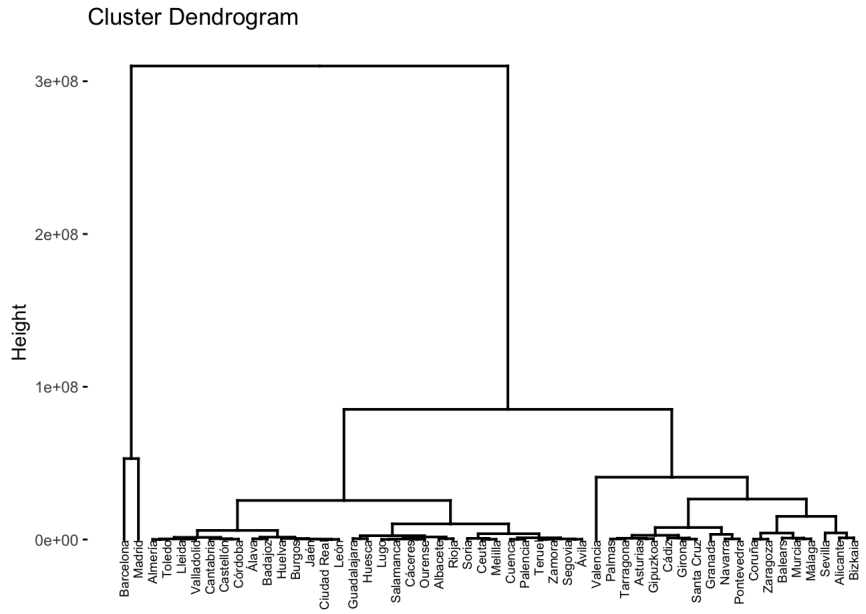
4. En el caso de Barcelon y Madrid las distancias son cortas entre las provincias en comparacion con las otras provincias. Del grafico aprecioamos entre tres y cuatro agrupaciones entre la similitud entre los colores entre las otras provincias.

5.A ¿Cuántos clusters recomendarías?

```
### Analisis Jerarquico

#Agrupamos las observaciones según el criterio de ward
res.hc <- hclust(d, method="ward.D2")
```

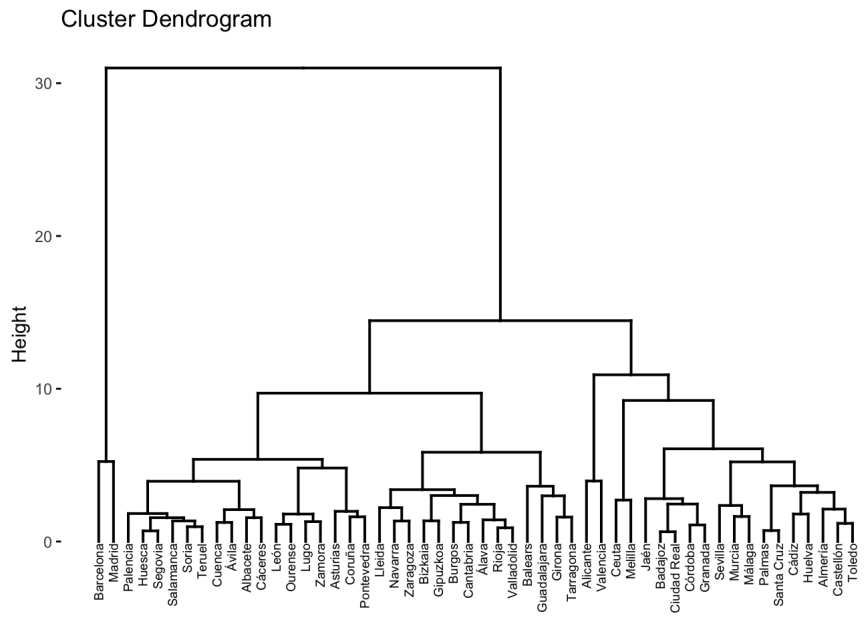
Dendrograma con los datos sin estandarizar

```
#Dibujamos el dendrograma correspondiente

fviz_dend(res.hc, cex = 0.5)
```



Dendrograma con los datos estandarizados

```
res.hc_st <- hclust(d_st, method="ward.D2")

fviz_dend(res.hc_st, cex = 0.5)
```



5.A Segun los dendogramas se propone 4 clusters.

5.B Representar los individuos agrupados según el número de clusters elegido.

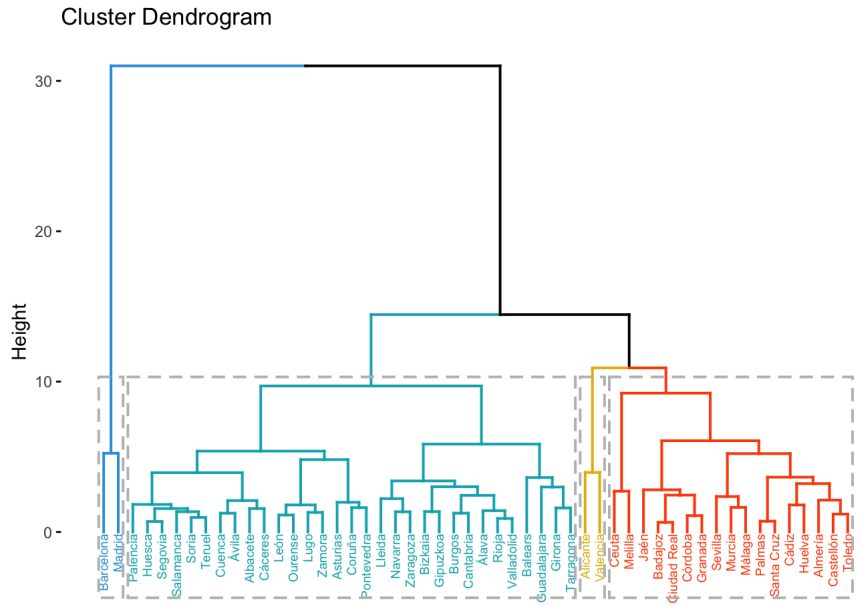
```
# Seleccionamos 3 clusters
grp <- cutree(res.hc_st, k =4 )

# Number of members in each cluster
knitr::kable(table(grp), caption = "Número de individuos por cluster")
```

Número de individuos por cluster

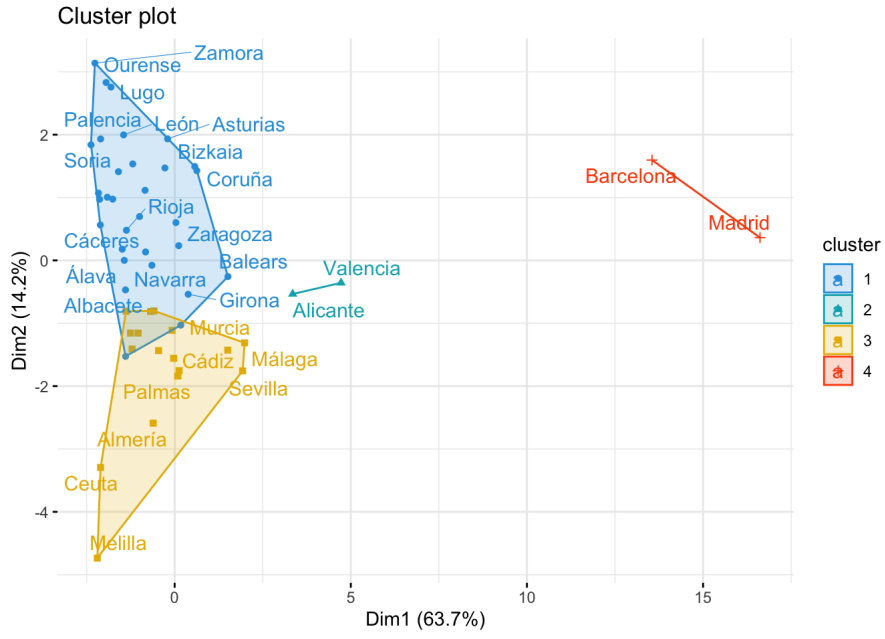
grp	Freq
1	31
2	2

grp	Freq
3	17
4	2

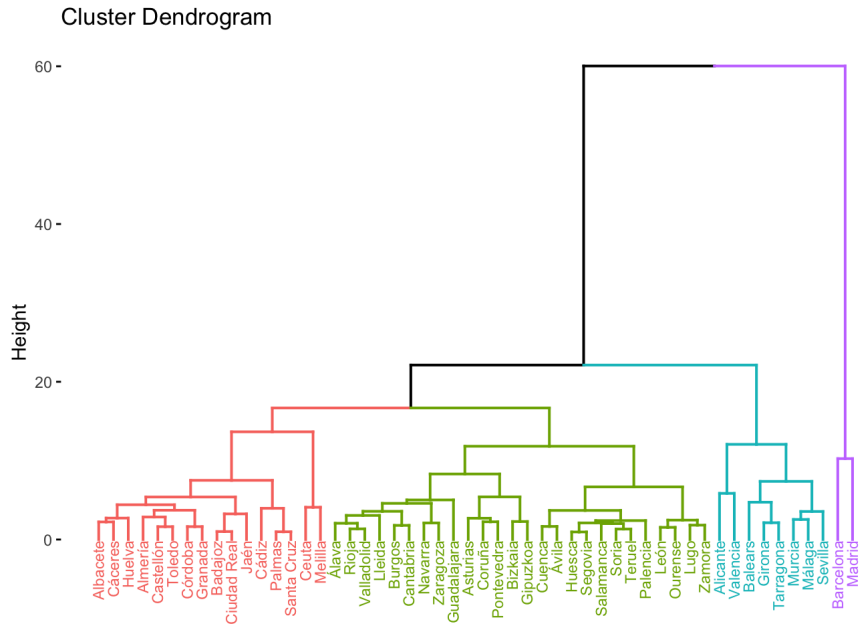
```
# Esta función representa el dendrograma con el número de clusters decidido
fviz_dend(res.hc_st, k = 4, # Cuatro Clusters
cex = 0.5, #tamaño
k_colors = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
color_labels_by_k = TRUE, #Diferentes colores a los clusters
rect = TRUE) #añade un rectángulo alrededor
```




```
#Visualizamos los clusters
fviz_cluster(list(data = datos_estan, cluster = grp),
palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07","#FC4E09"),
ellipse.type = "convex", # Concentration ellipse
repel = TRUE, # Avoid label overplotting (slow)
show.clust.cent = FALSE, ggtheme = theme_minimal())
```



```
# Agglomerative Nesting (Hierarchical Clustering)
res.agnes <- agnes(x =datos_estan, # datos
stand = TRUE, # Standardizamos
metric = "euclidean", # distancia entre individuos
method = "ward") # distancia entre clusters

fviz_dend(res.agnes, cex = 0.6, k = 4)
```

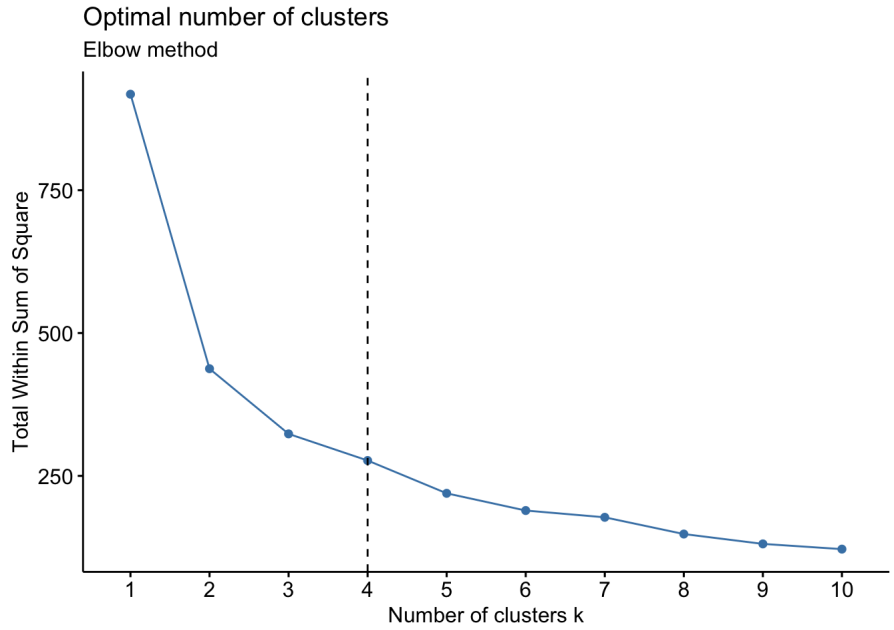


5.C Qué número óptimo de clusters nos indican los criterios Silhoutte y de Elbow?

Determinación del número óptimo de clusters

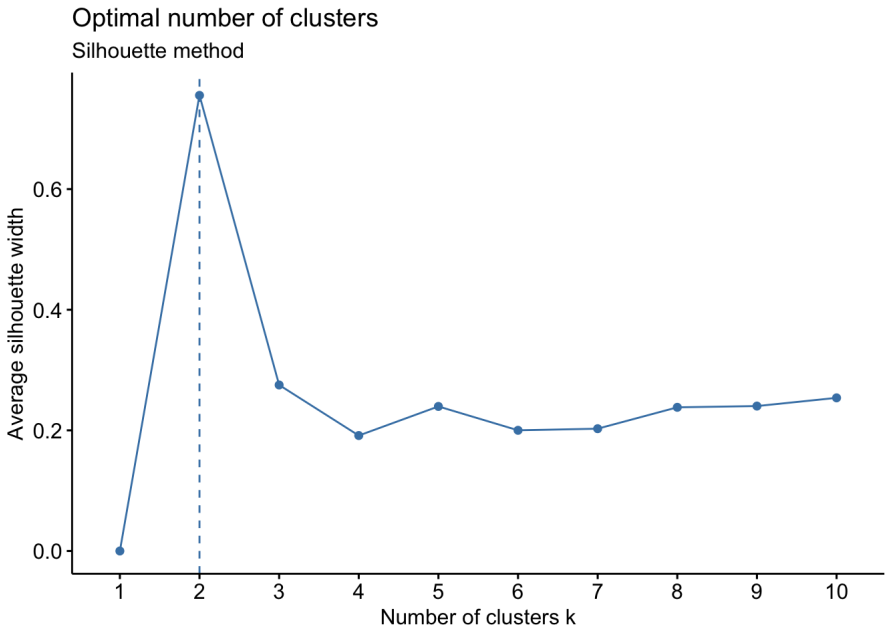
Elbow method

```
fviz_nbclust(datos_estan, kmeans, method = "wss") +
geom_vline(xintercept = 4, linetype = 2)+
labs(subtitle = "Elbow method")
```



Silhouette method

```
### Silhouette method
fviz_nbclust(datos_estan, kmeans, method = "silhouette")+

labs(subtitle = "Silhouette method")
```



5.C El metodo de Elbow nos indica 4 clusters mientras que el metodo Silhouette 2 clusters.

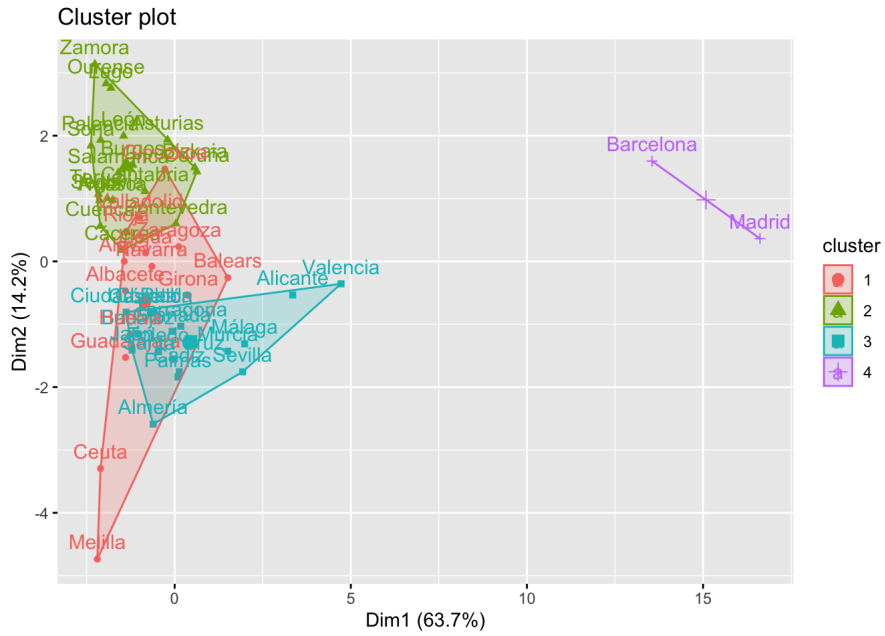
5.D Con el número de clústeres que nos indica Elbow en el apartado anterior, realizar un agrupamiento no jerárquico.

i. Representar los clústeres formados en los planos de las Componentes principales. Relacionar la posición de cada clúster en el plano con lo que representa cada componente principal.
ii. Evaluar la calidad de los clústers

```
#Nos aseguramos que tenemos todos la misma semilla
RNGkind(sample.kind ="Rejection")
set.seed(1234)
# Compute k-means
km.res <- kmeans(datos_estan, 4)
head(km.res$cluster, 20)
```

```
##    Albacete    Alicante     Almería       Álava    Asturias     Badajoz
##           1           3           3           1           2           3
##     Balears   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
```

```
# Visualize clusters using factoextra
fviz_cluster(km.res, datos_estan)
```




```
ordenado<-sort(km.res$cluster)
knitr::kable(ordenado, digits =2, caption = "Provincia y cluster")
```

Provincia y cluster

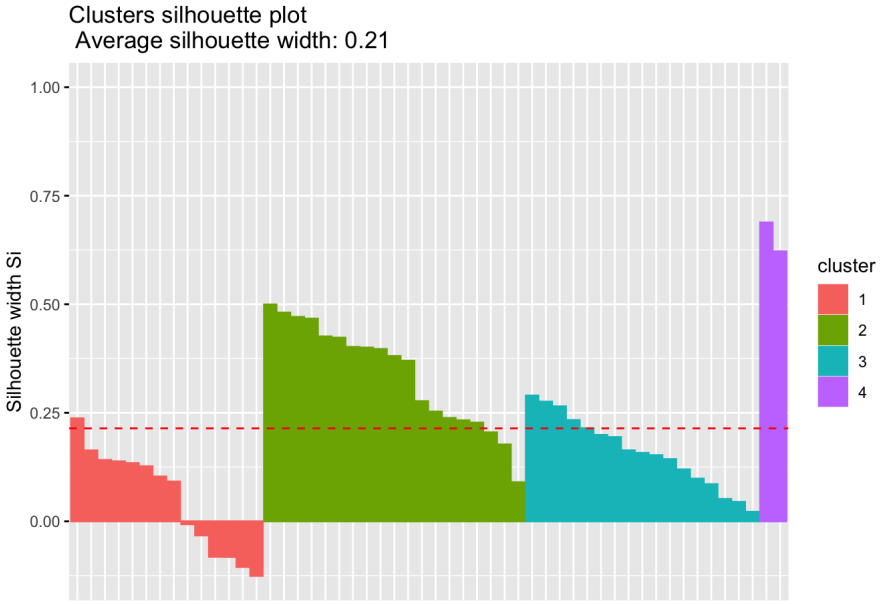
	x
Albacete	1
Álava	1
Balears	1
Ceuta	1
Gipuzkoa	1
Girona	1
Guadalajara	1
Huelva	1
Lleida	1
Melilla	1
Navarra	1
Rioja	1
Valladolid	1
Zaragoza	1
Asturias	2
Bizkaia	2
Burgos	2
Cantabria	2
Coruña	2
Cuenca	2
Cáceres	2
Huesca	2
León	2
Lugo	2
Ourense	2
Palencia	2
Pontevedra	2

	x
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

Calidad de los clusters

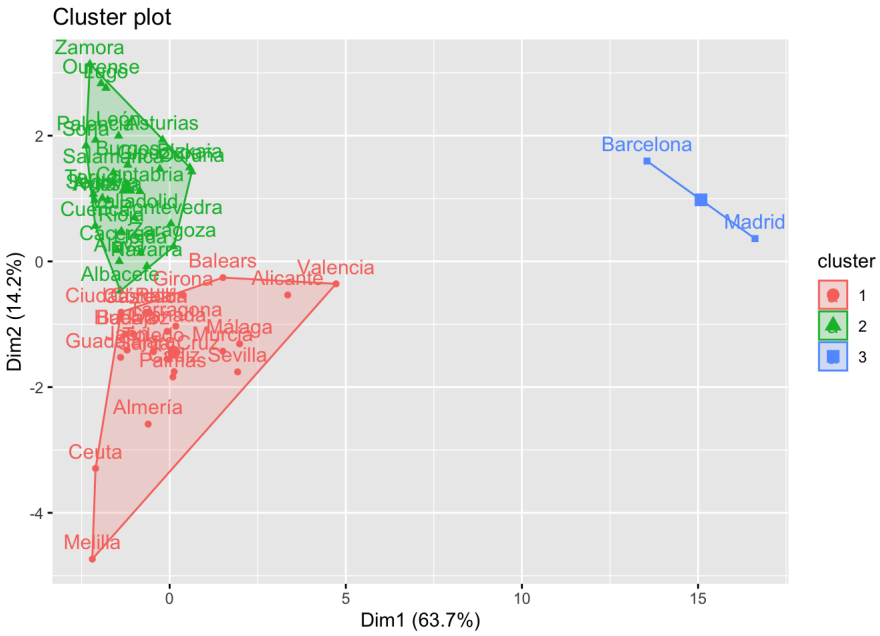
```
#Evaluación de la calidad de los clusters
sil <- silhouette(km.res$cluster, dist( datos_estan))
rownames(sil) <- rownames( datos_estan)
fviz_silhouette(sil)
```

```
##   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
```



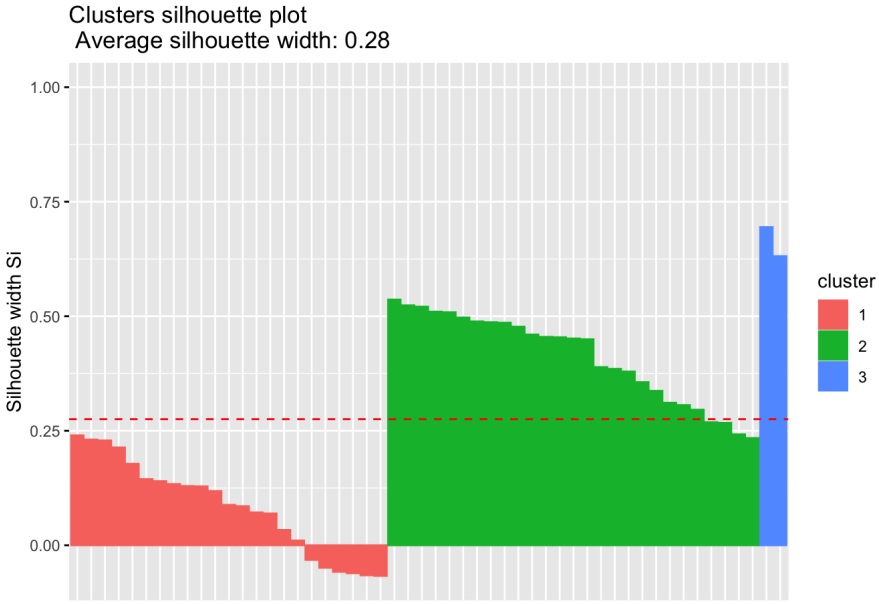
Ya que el primer cluster nos dio negativo probamos con el criterio de Silouette el cual considera 2 cluster

```
RNGkind(sample.kind = "Rejection")
set.seed(1234)
km.res3 <- kmeans(datos_estan, 3)
fviz_cluster(km.res3, datos_estan)
```




```
sil <- silhouette(km.res3$cluster, dist(datos_estan))

fviz_silhouette(sil)
```

```
##   cluster size ave.sil.width
## 1       1   23          0.08
## 2       2   27          0.41
## 3       3    2          0.66
```



5.E Explicar las provincias que forman cada uno de los clústeres

Las provincias de Barcelona y Madrid tienen indices de [Poblacion,NumEmpresas,Industria,Construcccion,Turismo,AFS,APT] mayores que el resto de las provincias y estan altamente correlacionadas entre si mismas. Estas van a ser las variables que indicaran un alto desarrollo socioeconomico en las provincias, mientras que los indices de mortalidad y tasa de paro son bajas en Madrid y Barcolona dado el desarrollo economico de estas gran ciudades demonstrado por su Producto interno Bruto

Sin embargo, se observaron provincias cuyo indice de mortalidad y tasa de paro son igualmente alto respecto a las otras provincias por el cual se le atribuye su indice de natalidad y baja tasa de actividad. Por otro lado la mayor densidad de provincias se encuentran generalmente por valores similares en indices. Hemos decidido entnces quedearnos con dos clusters de agrupacion dado que result con la representacion optima de las provincias