TareaII_Parte2

Javier Aos

Parte 2: Técnicas no supervisadas

Lectura y depuración del dataset

```
elec <- readxl::read_excel(
  "G:/My Drive/1.3 Master/Modules/08_Data_mining/Datos/DatosEleccionesEspaña.xlsx")
```

```
# Ciudades con más de 100000 habitantes
elec_r2<-na.omit(elec[elec$Population >100000 & elec$Population <500000,])

# Selecciono las numéricas
elec_r<-Filter(is.numeric, elec_r2)[,-1] # Elimino Código de provincia
rownames(elec_r)<-elec_r2$Name
names(elec_r[-c(1,2,32)])
names(elec_r)<-c(
  "Pop", "Cens", "Abs", "AbsA" , "Izda",   "Dcha",   "Otr",  "IzqA",  "DechA",
"Age4",    "Age19","Age19_65",  "Age65",  "WomP", "Forei", "SameCom","SCDifProv",  "DifCom",
 "UnL25", "Un25_40" , "UnM40" , "AgrU" , "IndU"    ,  "ConsU", "ServU"  ,   "Empr",   "Indus",
"Const",    "ComHost" , "Servi", "inmuebles",  "Pob2010",    "SUPERFICIE",
"PobChange",   "PersInm","Explot" )

# Elimino las variables de población y dicótomas
elec_r<-elec_r[,-c(1,2,4,8,9,32)]
```

Seleccionamos las 10 variables con mayor MSA

```
# Eliminamos las variables que producen un error de colinealidad
x<-psych::KMO(elec_r[,c(5:12,22:29)])
sort(x$MSAi, decreasing = TRUE)
```

```
##       WomP       Age4      Const      Indus  inmuebles  PobChange    PersInm
##  0.7915638  0.7878923  0.7698950  0.7611236  0.7556504  0.7199880  0.6982107
##    ComHost      Servi  SCDifProv    SameCom      Age65      Forei      Age19
##  0.6949941  0.6820915  0.6238009  0.6199655  0.5997033  0.5781456  0.5674658
##   Age19_65 SUPERFICIE
##  0.4250948  0.3905181
```

```r
# Seleccionamos las 10 variables con un MSA más alto
elec_r<-elec_r[,c(
  'WomP','Age4','Const','Indus','inmuebles','PobChange','PersInm','ComHost',
  'Servi','SCDifProv')]
elec_r$CCAA<-elec_r2$CCAA
rownames(elec_r)<-elec_r2$Name
```

Agregamos valores por CCAA

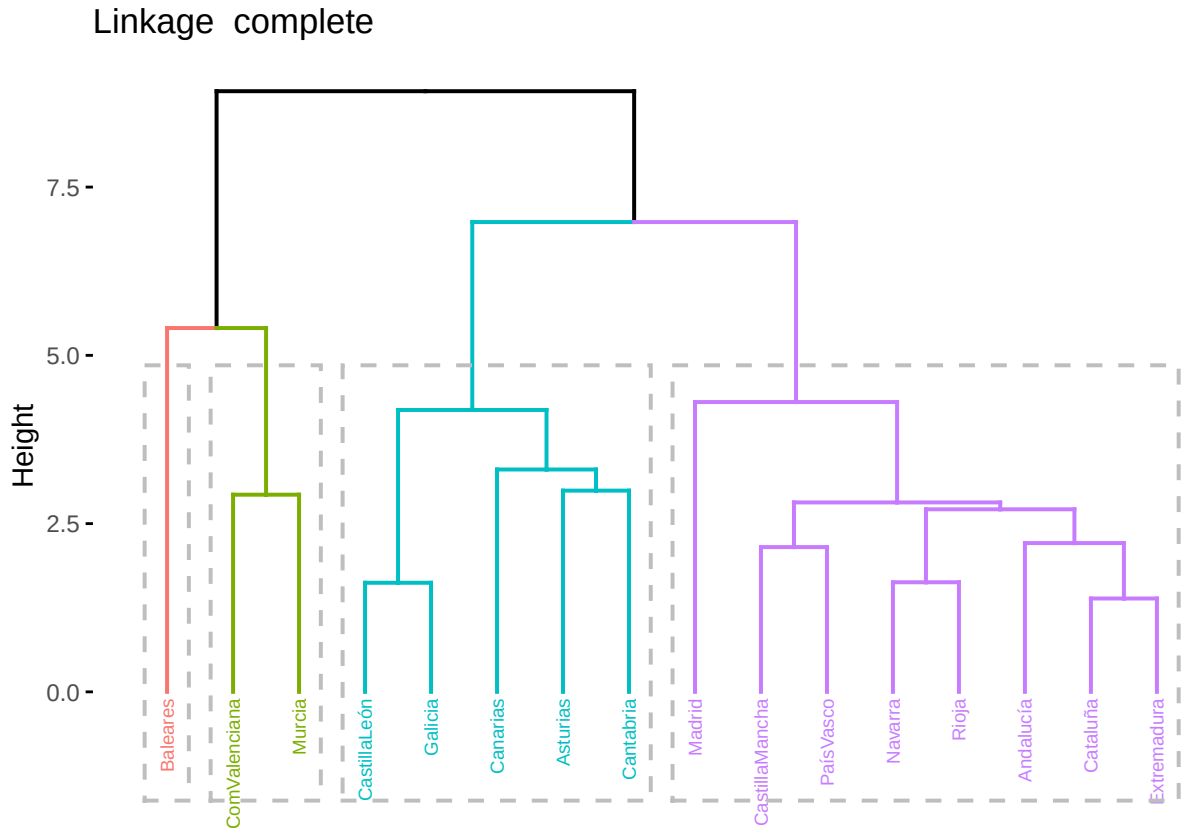
```r
ccaa<-aggregate(.~CCAA,elec_r,mean)
rownames(ccaa)<-ccaa$CCAA
ccaa<-ccaa[,-1]
```

Escalamos datos ya que hay distintas unidades de medida

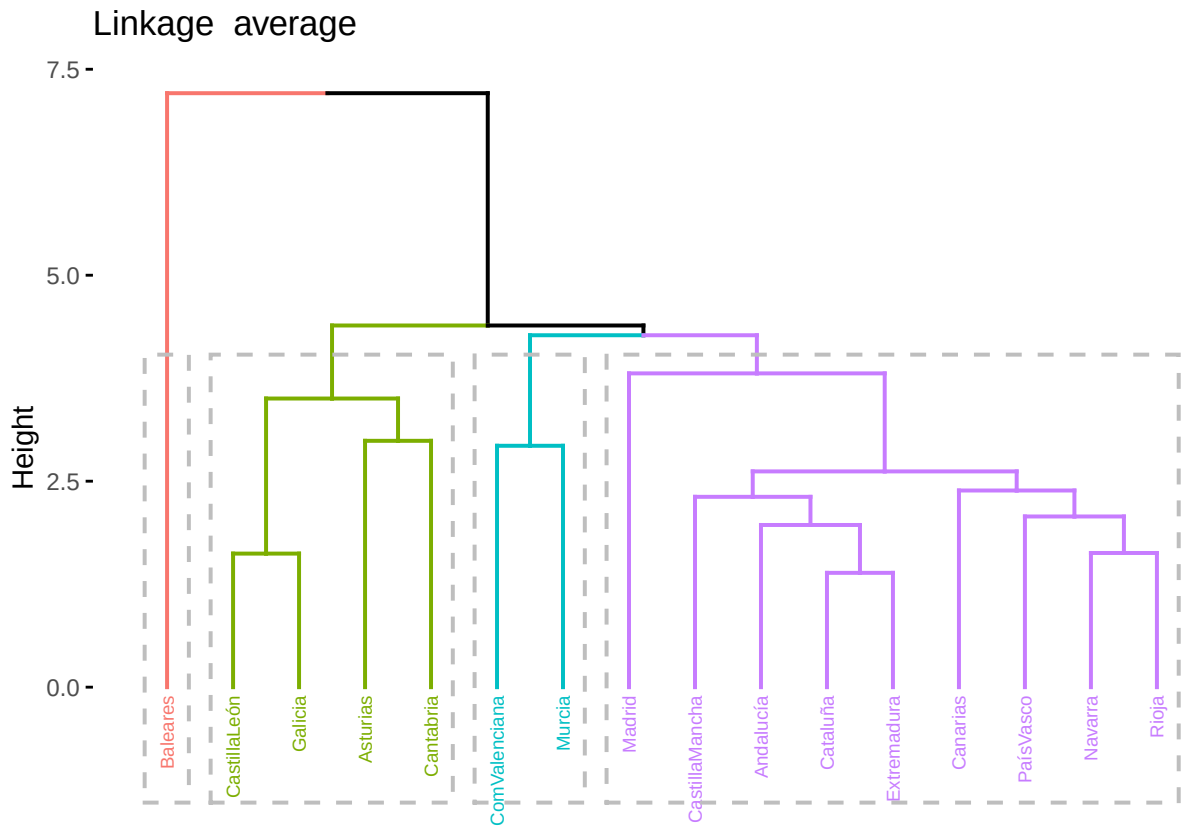
```r
ccaa.sc<-scale(ccaa)
```

Exploramos clustering jerárquico con distintos Linkages

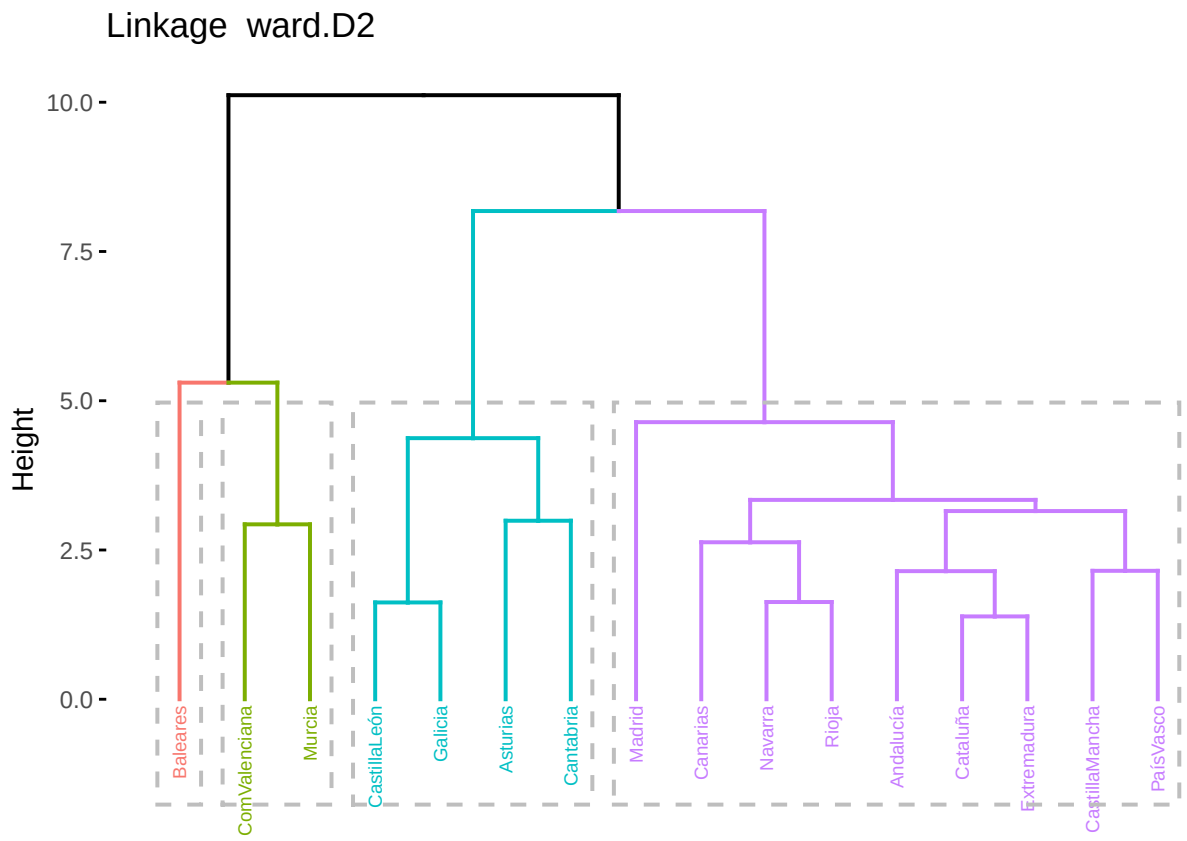
```r
methods<-c("complete","average",'ward.D2')
hclist<-list()
val.hc<-c()
for (i in 1:length(methods)){
  hc=hclust(dist(ccaa.sc),method =methods[i])
  hclist[[i]]<-hc
 print(fviz_dend(hc,k = 4, cex = 0.5, color_labels_by_k = T, rect = T)+
         ggtitle(paste('Linkage ', methods[i])))
 #Validación interna
 cl<-cutree(hc, k = 4)
 md.val<-medidasVal(ccaa.sc,cl,cl,methods[i])

 # Generar vector de medidas de validación
 val.hc<- rbind(val.hc,md.val)
}
```




```
## Indice VI  complete = 0
## Indice Rand  complete = 1
## Silueta media  complete = 0.2496225
## Within SS  complete = 52.80084
```




```
## Indice VI  average = -4.440892e-16
## Indice Rand  average = 1
## Silueta media  average = 0.2766576
## Within SS  average = 51.31597
```

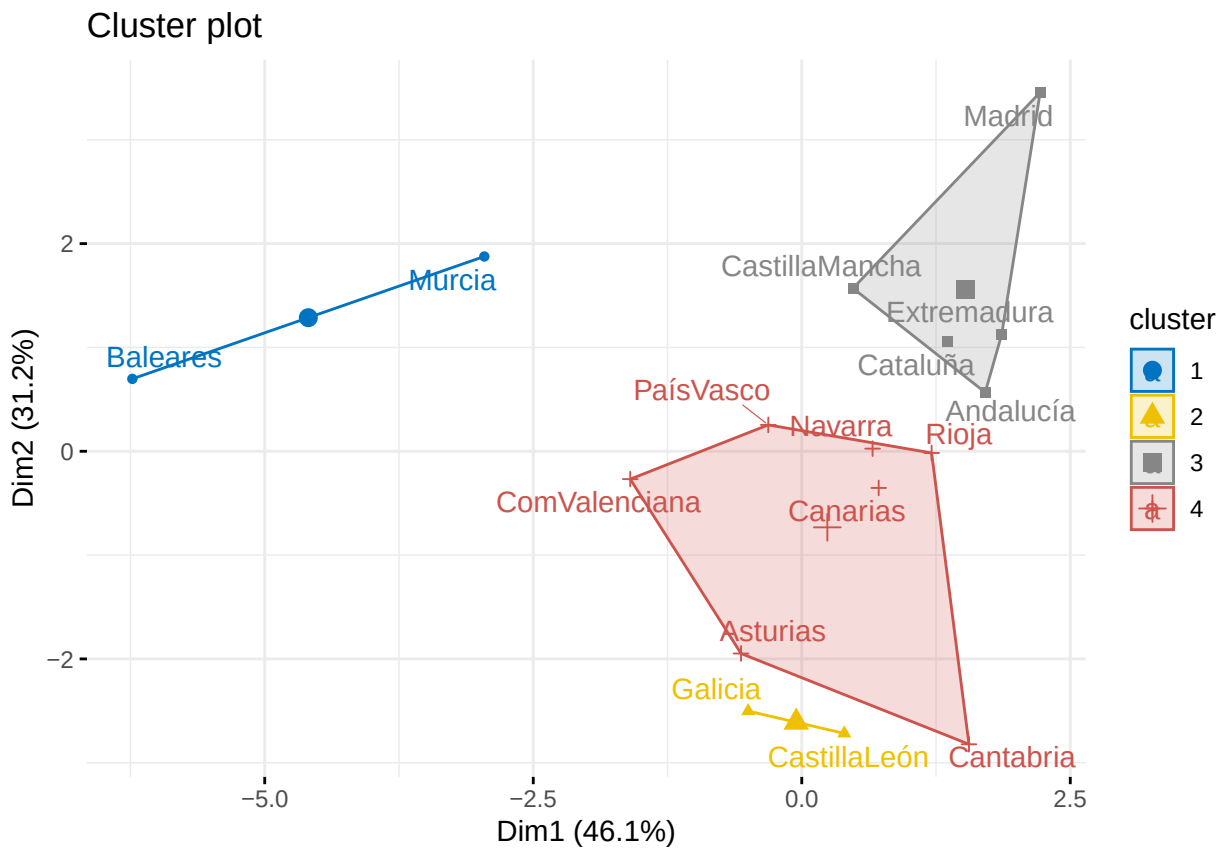



```
## Indice VI  ward.D2 = -4.440892e-16
## Indice Rand  ward.D2 = 1
## Silueta media  ward.D2 = 0.2766576
## Within SS  ward.D2 = 51.31597
```

```
names(hclist) <- rownames(val.hc)<-methods
```

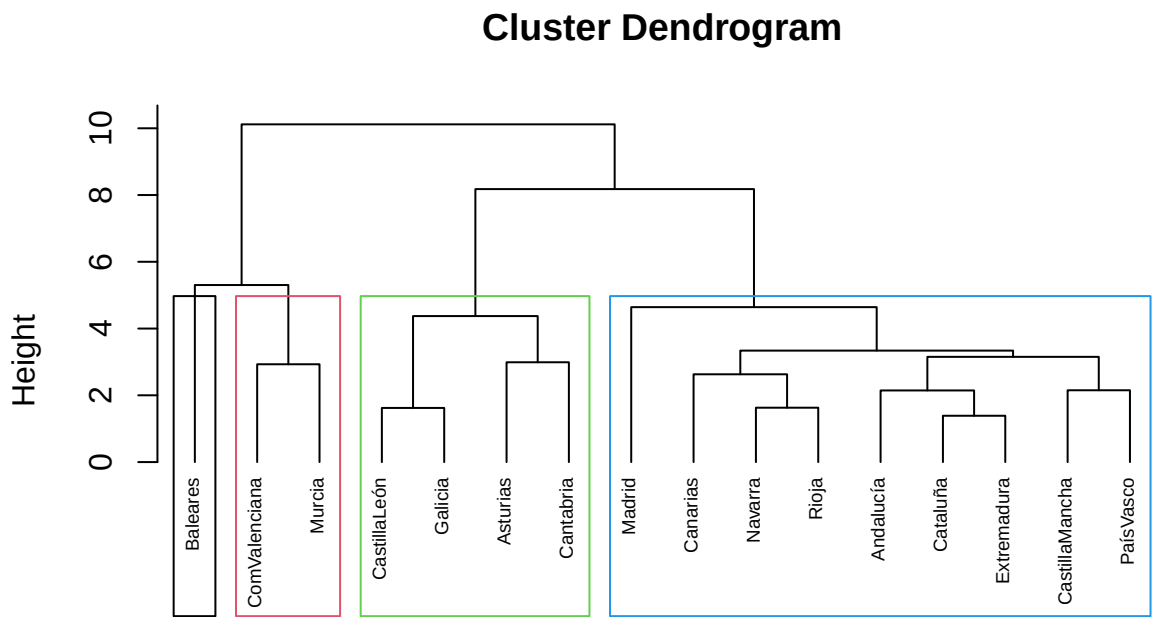
Exploramos k-means con 4 grupos

```
km.out=kmeans(ccaa.sc,4)
fviz_cluster(km.out, data = ccaa.sc,  ellipse.type = "convex", palette = "jco",repel = TRUE,
             ggtheme = theme_minimal())
```

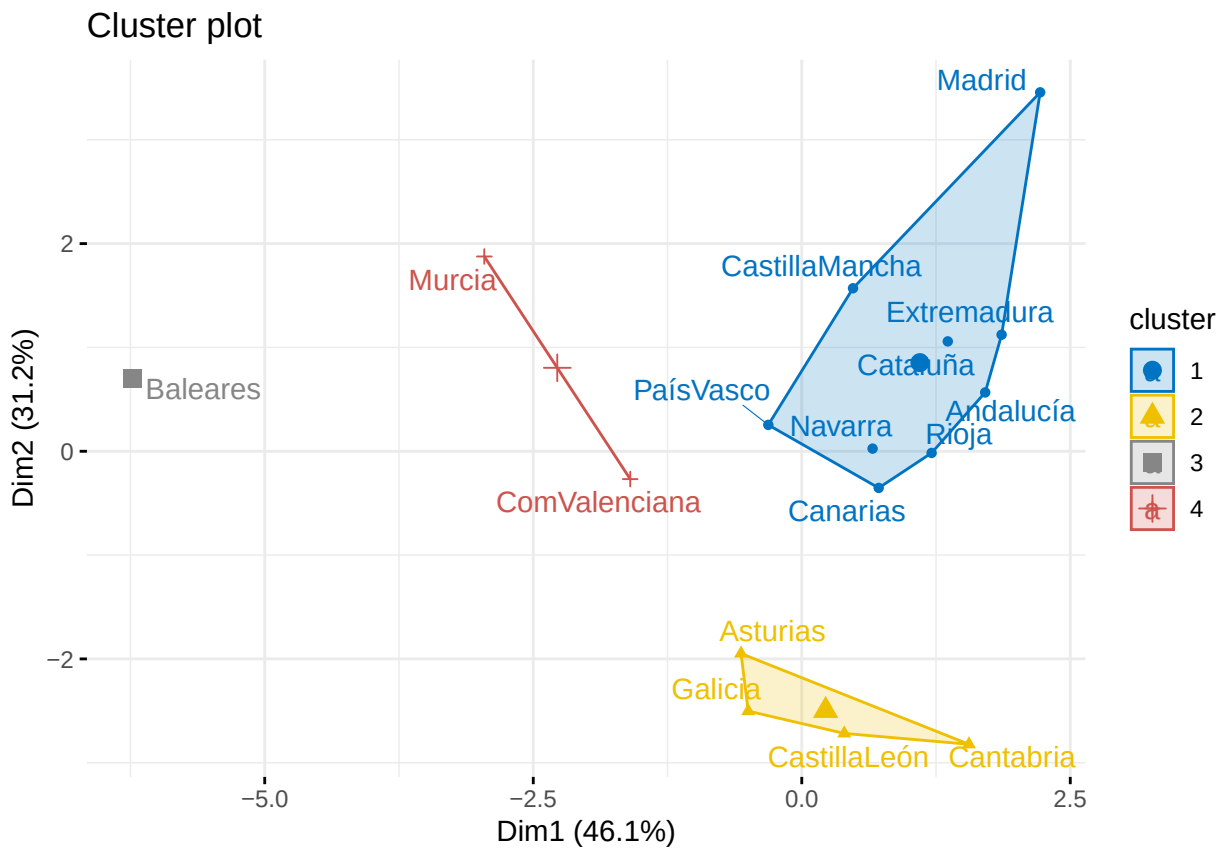


Intentamos unir fuerzas con método híbrido (hkmeans)

```
hk.out=hkmeans(ccaa.sc,4)

hkmeans_tree(hk.out, cex = 0.6)
```




```
fviz_cluster(hk.out, data = ccaa.sc,  ellipse.type = "convex", palette = "jco",repel = TRUE,
             ggtheme = theme_minimal())
```



Medidas de validación

```
md.km<-medidasVal(ccaa.sc,km.out$cluster,km.out$cluster,'kmeans')
```

```
## Indice VI  kmeans = 0
## Indice Rand  kmeans = 1
## Silueta media  kmeans = 0.2264591
## Within SS  kmeans = 53.06853
```

```
md.hk<-medidasVal(ccaa.sc,hk.out$cluster,hk.out$cluster,'hkmeans') ## Son iguales
```

```
## Indice VI  hkmeans = -4.440892e-16
## Indice Rand  hkmeans = 1
## Silueta media  hkmeans = 0.2766576
## Within SS  hkmeans = 51.31597
```

```
ValT<-rbind(val.hc,md.km,md.hk) ## El mismo clustering con varias técnicas
ValT
```

```
##                       vi rand silhouette      wss
## complete  0.000000e+00    1  0.2496225 52.80084
## average  -4.440892e-16    1  0.2766576 51.31597
```

```
## ward.D2  -4.440892e-16    1  0.2766576 51.31597
## md.km     0.000000e+00    1  0.2264591 53.06853
## md.hk    -4.440892e-16    1  0.2766576 51.31597
```

Average, ward.D2 y md.hk tiene exactamente el mismo silhouette y wss.

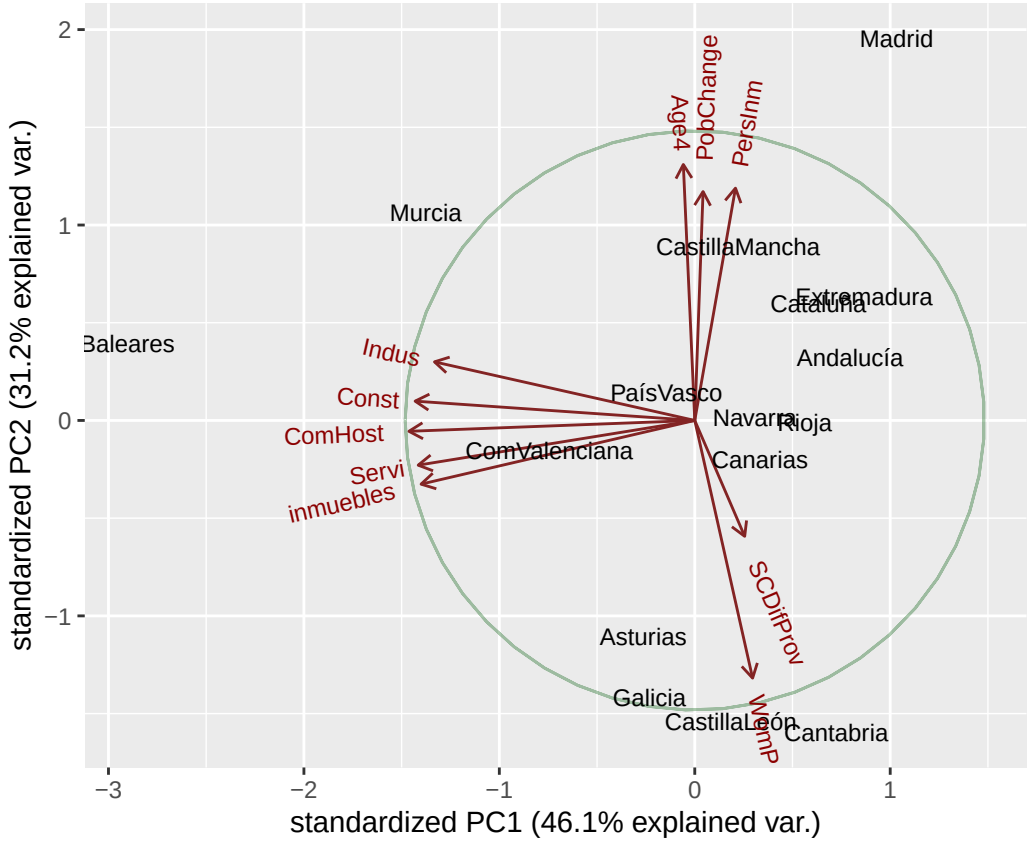
Para una mejor interpretabilidad hacemos una reducción por componentes principales

```
pr.out=prcomp(ccaa.sc, scale. = T)
# Varianza explicada
summary(pr.out)$importance[3,2]
```

```
## [1] 0.77311
```

Con la reducción por componenetes principales conseguimos explicar un 77.31% de la varianza.

```
# Biplot en plano de componentes
ggbiplot::ggbiplot(pr.out,labels=rownames(ccaa),
                   ellipse = TRUE, circle = TRUE)
```



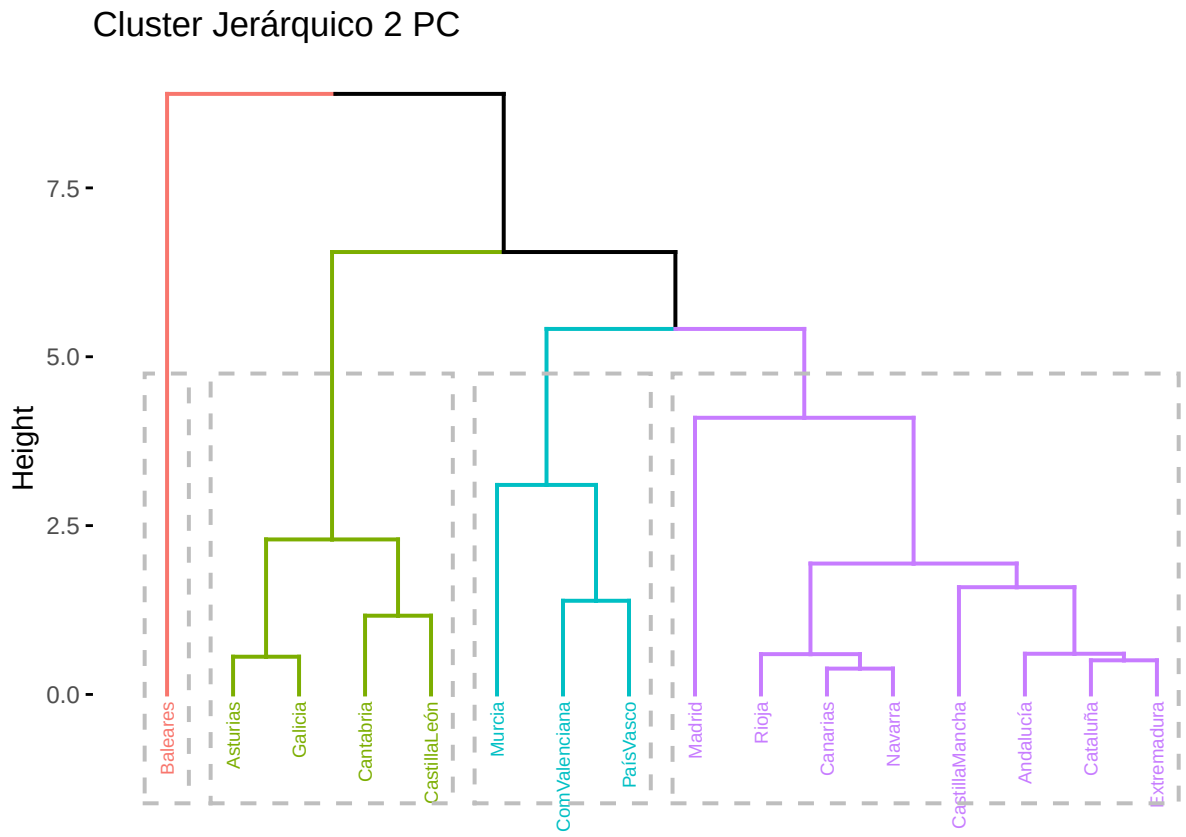
Vemos que el PC1 tiene mucha carga de variables como Indust, Const, ComHost, Servi o Inmuebles. Por lo tanto en el eje horizontal se están clasificando las CCAA según su infraestructura/desarrollo industrial

en el momento de la medición (parte izquierda) frente a el no desarrollo industrial (parte derecha). Vemos que por ejemplo, Baleares es un CA con mucho peso industrial, cantidad de inmuebles, hostelería, servicios, construcción etc. Por lo tanto, vemos que podría ser una CA que todavía tiene capacidad de crecimiento en su construcción. En cambio vemos a Madrid situada en la parte derecha, ya que es una CA con mucho desarrollo, quedando "poco espacio" para la construcción de por ejemplo nuevos inmuebles, comercios de hostelería, servicios etc. es decir, es una CA con infraestructura ya establecida.

Si nos fijamos en el PC2 vemos que tiene una gran influecia de variables como Age4, PobChange o PersInm, pudiendo interpretasarse como una natalidad y cambio poblacional alto, frente a la parte de abajo que representa la presencia de gente más mayor y una natalidad más baja con variables como WomP (muy relacionada con gente de edad avanzada debido a la esperanza de vida de las mujeres) o SCDifProv. Vemos ciudad como Cantabria o Casatilla y León con más población mayor, frente a Madrid que presenta una mayor natalidad y cambio poblacional.

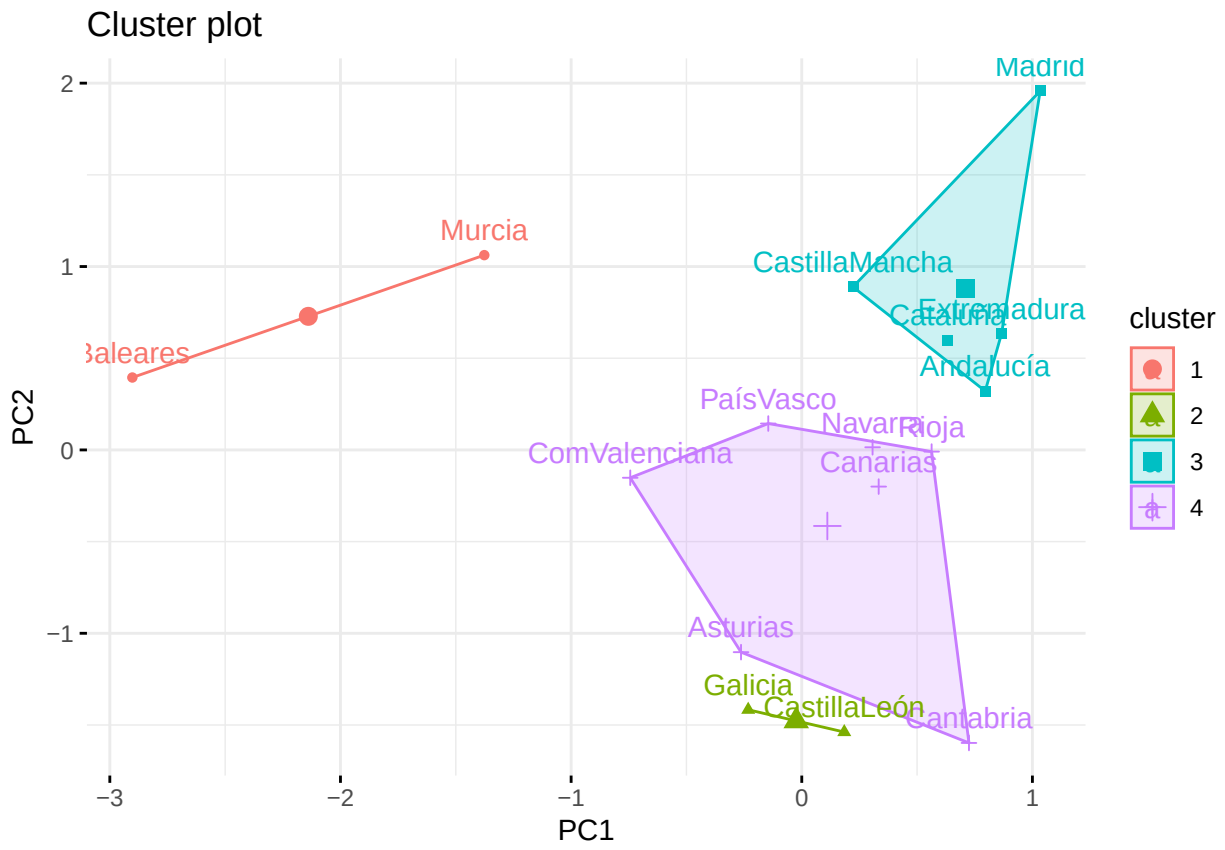
Ajustamos el cluster jerárquico a la solución de dos componentes

```
hc.pr=hclust(dist(pr.out$x[ ,1:2]))
fviz_dend(hc.pr,k = 4, cex = 0.5, color_labels_by_k = T, rect = T)+
  ggtitle("Cluster Jerárquico 2 PC")
```




```
cl.hc.pr<-cutree(hc.pr, k = 4)

km.out.pr=kmeans(pr.out$x[ ,1:2],4)
```

```
fviz_cluster(km.out, data = pr.out$x[ ,1:2],
             ggtheme = theme_minimal())
```




```
km.out.pr$centers
```

```
##          PC1         PC2
## 1  0.1351742 -0.07161157
## 2  1.5249365  1.55501378
## 3 -4.5943728  1.28671378
## 4  0.2220481 -2.49760965
```

Medidas de validación

```
md.km<-medidasVal(pr.out$x[ ,1:2],km.out.pr$cluster,km.out.pr$cluster,'kmeans PCA')
```

```
## Indice VI  kmeans PCA = 0
## Indice Rand  kmeans PCA = 1
## Silueta media  kmeans PCA = 0.3629683
## Within SS  kmeans PCA = 21.45545
```

```r
md.hk<-medidasVal(pr.out$x[ ,1:2],cl.hc.pr,cl.hc.pr,'hclus PCA')
```

```
## Indice VI  hclus PCA = 0
## Indice Rand  hclus PCA = 1
## Silueta media  hclus PCA = 0.4100688
## Within SS  hclus PCA = 22.50895
```

```r
ValT<-rbind(md.km,md.hk) ## Un poco mejor el k means
ValT
```

```
##       vi rand silhouette      wss
## md.km  0    1  0.3629683 21.45545
## md.hk  0    1  0.4100688 22.50895
```

Md.hk funciona algo mejor, aunque no hay mucha diferencia.

Conclusión

En conclusión, vemos que el cluester kmedias ha agrupado 4 grupos teniendo el verde (arriba derecha) representado por CCAA con un desarrollo de la insutria y construcción bien establecido y una alta natalidad y cambio poblacional, encontrado CCAA como Madrid o Cataluña. En el grupo rojo (centro) vemos CCAA que se acercan al centro y por tanto serían "neutras" no destacando en ninguna de las variables. En el grupo morado (abajo derecha) vemos CCAA con una buena infraestructura pero una población predominante envejecida, encontrado algunas Cantabria o Castilla y León. Por último en el grupo azul (izquierda) vemos CCAA que aún tienen una gran capacidad de mejora en su infraestructura como baleares o Murcia.