Pràctica 7. Clústers jeràrquics

La majoria d'algoritmes jeràrquics aglomeratius s'ajusten al procediment següent:

0. *Inici:* Es parteix d'una matriu de dissimilaritats $d(i,j)$ inicials donada referida a n casos. Cada cas individual és un clúster, de manera que tenim $k = n$ clústers inicials amb 1 element cada clúster ($n_i = 1, \forall i$).

I. *Aglomeració:* S'uneixen els dos clústers més similars ($d(i,j)$ mínima). Siguin p i q aquests dos clústers i $t = p \vee q$ el resultant de la seva unió.

II. *Actualització:* El nombre de clústers baixa ($k \to k-1$) i s'actualitza la matriu de dissimilaritats $d_a(,)$:

$$\begin{cases} d_a(i,j) = d(i,j) & \text{si} \quad i \neq t \\ d_a(t,j) = \varphi(\{d(i,j)\}) \end{cases}$$

on la funció φ depén del mètode d'enllaç escollit. També s'actualitza el nombre d'elements del nou clúster: $n_t = n_p + n_q$. *Nota:* depenent de mètode d'enllaç potser caldrà actualitzar algun altre paràmetre.

III. Si $k = 1$, s'acaba el procediment. En cas contrari es torna al punt **I**.

Els mètodes jeràrquics tenen una complexitat algorísmica elevada, tant en temps com en memòria. Es redueix significativament la complexitat, si la funció φ es pot escriure com una *fórmula de Lance-Williams*:

$$d_a(t,j) = \alpha_p d(p,j) + \alpha_q d(q,j) + \beta d(p,q) - \gamma |d(p,j) - d(q,j)| \tag{1}$$

Els coeficients per als 4 mètodes d'enllaç més utilitzats són:

- Enllaç complet: $\alpha_p = \alpha_q = 0.5$, $\beta = 0$, $\gamma = -0.5$. *Comprova que és la distància entre els veïns més llunyans.*
- Enllaç simple: $\alpha_p = \alpha_q = 0.5$, $\beta = 0$, $\gamma = 0.5$. *Comprova que és la distància entre els veïns més propers.*
- Enllaç de Ward: $\alpha_p = \frac{n_p+n_j}{n_p+n_q+n_j}$, $\alpha_q = \frac{n_q+n_j}{n_p+n_q+n_j}$, $\beta = \frac{-n_j}{n_p+n_q+n_j}$, $\gamma = 0$.
- Enllaç de Ward.2: Aplica la fórmula de Ward a $d^2(t,j)$ i després fa l'arrel quadrada. Amb aquest mètode, *a cada pas s'uneixen els clústers que menys incrementen les distàncies al quadrat* ***dins*** *de clústers.* Observació, en el punt inicial les distàncies dins de clústers són zero. Opcional: Per saber més dels mètodes de Ward https://arxiv.org/pdf/1111.6285.pdf

Paquets: `mclust, FactoMinerR, factoextra`

Exercici 1. Considera la matriu de distàncies següent:

	A	B	C	D	E	F	G
A	0.00	0.50	0.43	1.00	0.25	0.63	0.38
B	0.50	0.00	0.71	0.83	0.67	0.20	0.78
C	0.43	0.71	0.00	1.00	0.43	0.67	0.33
D	1.00	0.83	1.00	0.00	1.00	0.80	0.86
E	0.25	0.67	0.43	1.00	0.00	0.78	0.38
F	0.63	0.20	0.67	0.80	0.78	0.00	0.75
G	0.38	0.78	0.33	0.86	0.38	0.75	0.00

1. Guarda la matriu en `D`. Posa noms a files i columnes. Guarda-ho com a distàncies (`as.dist()`).
2. Utilitza la funció `hclust()` de la llibreria `mclust` per obtenir els resultats per a diversos mètodes: digues `hc` quan apliquis el mètode *complete*, `hs` *single*, `hw` *ward.D* i `hw2` *ward.D2*:
3. Fes els dendrogrames de les 4 solucions amb `plot(hc)`, etc. Prova-ho amb i sense l'opció `hang=-1`.
4. Per a la solució `hc`, mostra els objectes `merge`, `height`. Explica amb precisió quina informació dóna cadascun.
5. Per a la solució `hc`, mostra els objectes `order` i `labels`. Canvia les etiquetes per lletres minúscules i comprova que es modifica el dendrograma. Permuta l'ordre D,B,F,A,E,C,G per G,C,E,A,F,B,D, i comprova que es modifica el dendrograma. Fes alguna altra permutació 'raonable' (que no entrecreui clústers).
6. Per a la solució `hc`, aplica el criteri del *"colze"* en el *diagrama de sedimentació o scree graph* de les altures o distàncies d'enllaç (`height`) per decidir el nombre de clústers k. Quin valor de k tries? Ajuda't amb la visualització del dendrograma i fixa't com s'ha d'entendre l' scree graph: comença per la dreta, si talles sota el primer punt és $k = 2$, sota del segon punt $k = 3$, etc. Cal tallar on el salt de les altures no sigui menyspreable, i sempre abans de sedimentar. colorredHi pot haver més d'una solució que convé validar a posteriori (això es farà l'Exercici 2).

7. Un cop decidit el nombre de clústers, aquests es poden enquadrar en el dendrograma amb la funció `rect.hclust()`. Aplica-la a `hc` amb $k = 3$ (has de tenir el dendrograma obert !). Esbrina què fa la funció `cutree()`, i guarda el resultat en l'objecte `clus`.

Exercici 2. el fitxer "lifeexp.dat" conté l'esperança de vida a diverses regions, per edat i sexe, a la dècada dels 60s. *Nota:* l'esperança de vida a determinada edat és la mitjana d'anys de vida restants "afegits a l'edat actual".

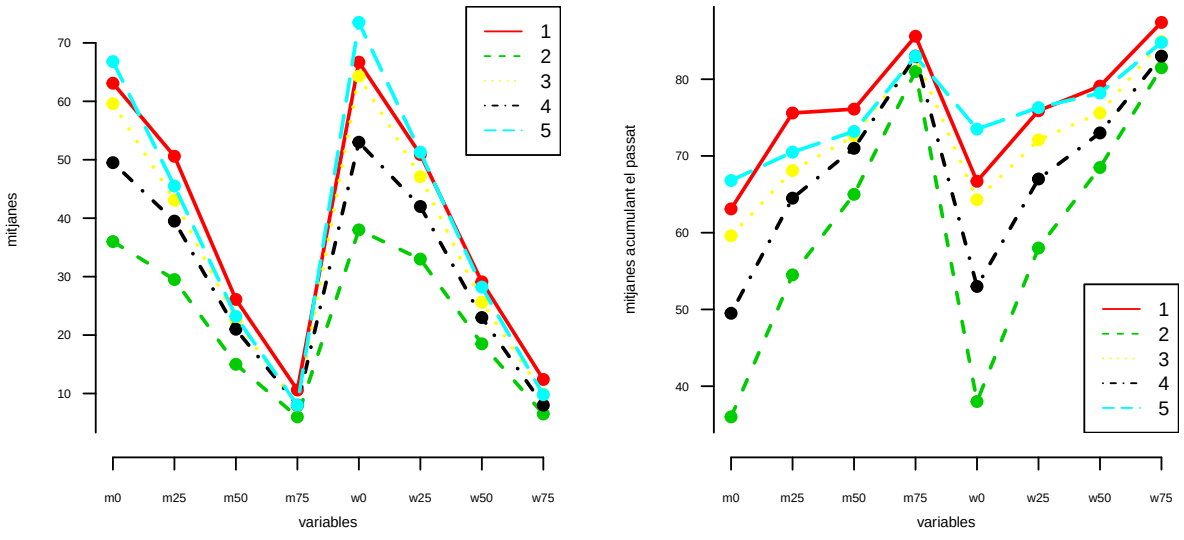
1. Carrega el fitxer amb `source()` i guarda'l en `lifeexp`. Quina classe té aquest objecte? Quins elements té? Guarda en l'objecte `life` només el primer element (`$value`). Com expliques que a Guatemala, per exemple, els valors de m0 i m25 fossin 49 i 40, respectivament?
2. Abreuja els noms de les files (països o regions), corregint manualment les abreujatures que creguis convenient.
3. Calcula la matriu de distàncies (Euclidianes) dels països (pots usar la funció `dist()`).
4. Aplica `hclust()` amb enllaç: simple, complet, ward.D i ward.D2. i guarda'ls en `hs, hc, hw, hw2`, respectivament. Dibuixa els dendrogrames en una figura 2×2 (fes que totes les branques s'ajustin a l'eix horitzontal).
5. Pel mètode de Ward, fes un scree graph de les altures i decideix raonadament quins valors de k trobes raonables segons aquest criteri. Per a $k = 3$, remarca els clústers sobre el dendrograma amb la funció `rect.hclust()` sense tancar el plot.
6. Considera la solució de Ward amb $\mathbf{k = 5}$ i fes el procés de *validació a-posteriori* basat en els passos següents:
   (a) Aplica `cutree()` per tenir el clúster de cada cas, guarda-ho en `clus` que sigui un `factor`, i afegeix la variable al dataframe `life`. Visualtiza la capçalera del fitxer.
   (b) Per a les 8 variables inicials, fes una taula que mostri les seves mitjanes en funció del clúster. Una possibilitat és usar la funció `by()` :

```
mitj<-by(life[,1:8], life$clus, function(x) round(apply(x,2,mean),1),simplify=T)
dfmitj<-as.data.frame(matrix(unlist(mitj),byrow=F,ncol=k))  # k = num clusters
rownames(dfmitj)<-names(mitj[[1]])
colnames(dfmitj)<-paste("clus",1:k,sep="")
dfmitj
```

   Interpreta els clústers en funció dels resultats de la taula.

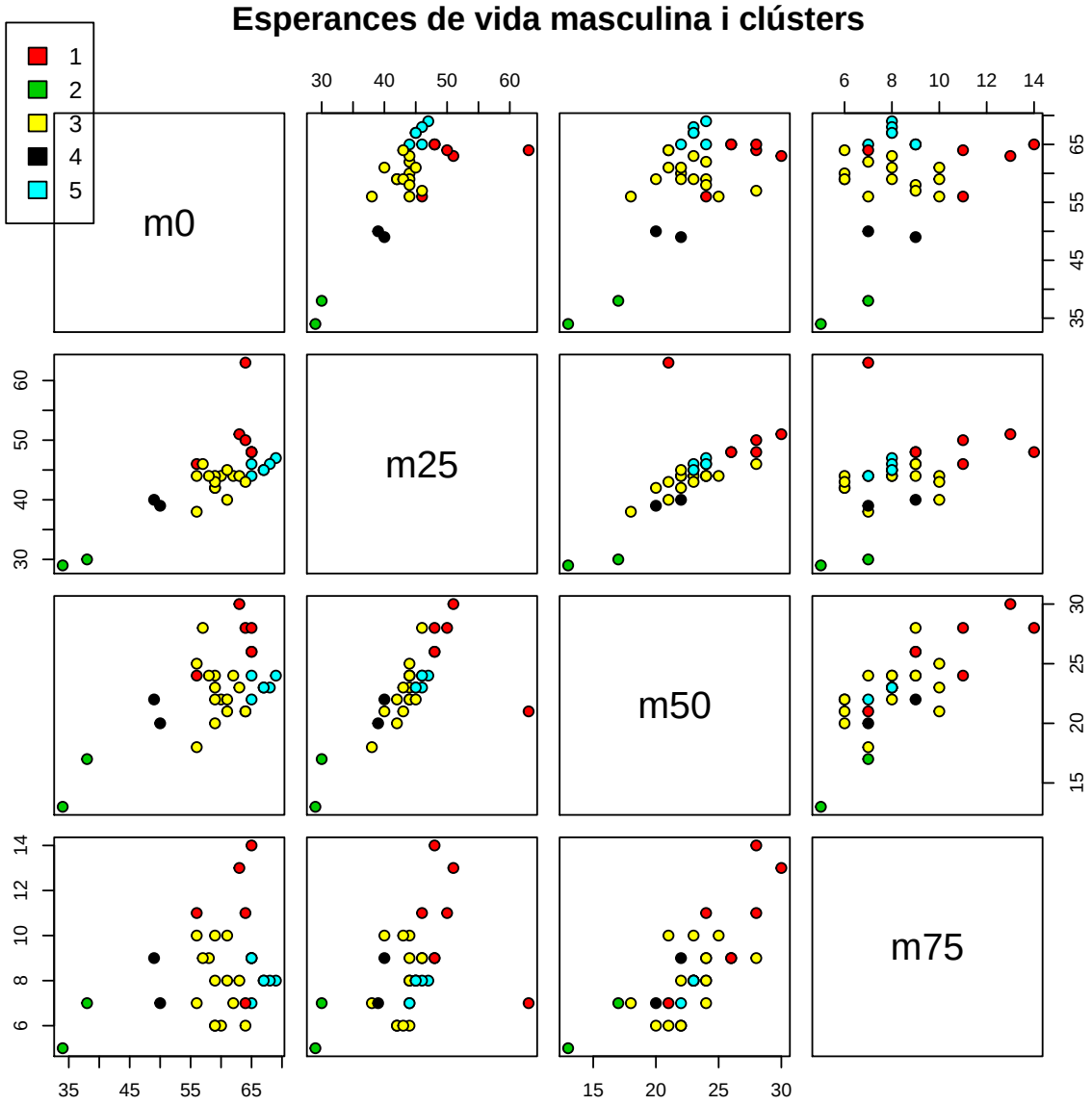
   (c) Dibuixa les mitjanes tabulades en l'apartat anterior usant la funció `matplot()`. Afegeix a cada mitjana els anys de vida anteriors i repeteix la gràfica. Per què creus que ho fem?
   Es mostra el codi de la gràfica esquerra i les dues gràfiques de costat.

```
k<-5
etiq<-rownames(dfmitj)
matplot(1:8,dfmitj,type="o",pch=19,lty=1:k,lwd=2,col=c("red","green3","yellow","black","cyan"),
        axes=F,xlab="variables",ylab="mitjanes",cex=.7)
axis(2,seq(0,70,by=10),las=1)
axis(1,1:8,labels=etiq,line=1,cex.axis=.6)
legend("topright",legend=1:k,lty=1:k,col=c("red","green3","yellow","black","cyan"))
```



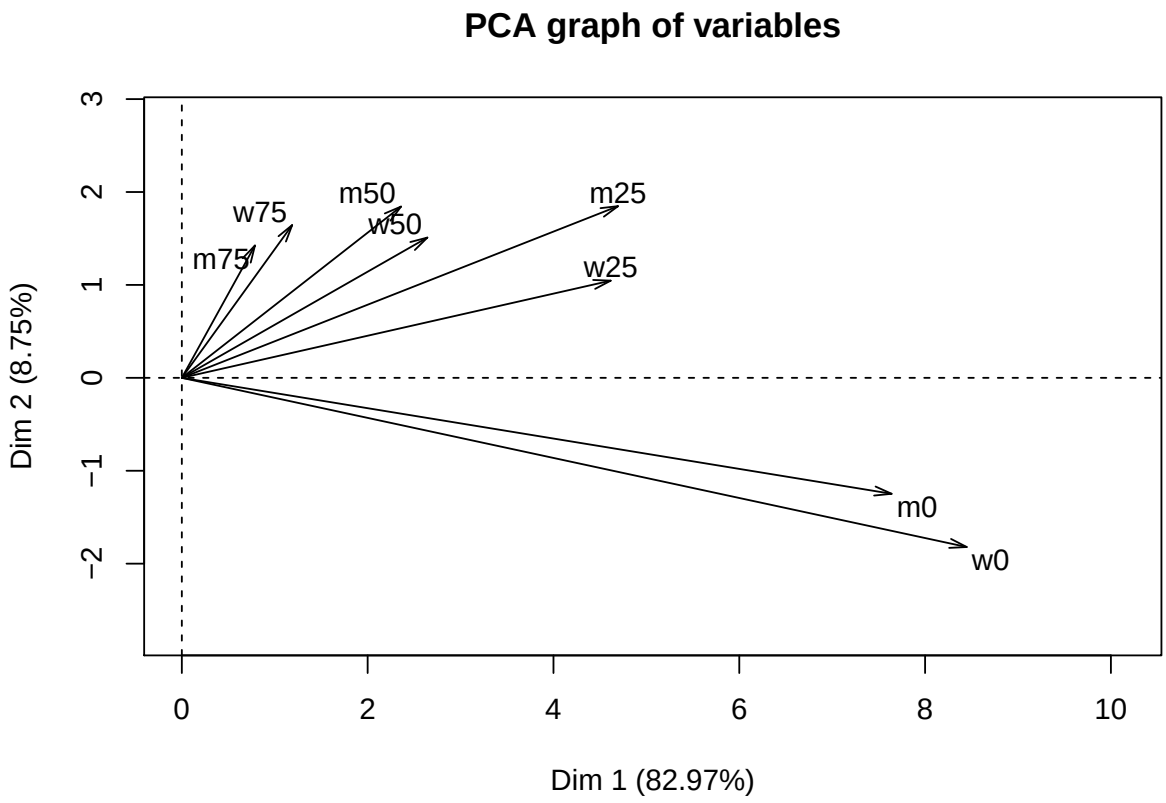
Caracteritza els clústers en funció d'aquests resultats. Quines variables semblen discriminar més entre els clústers, segons les mitjanes univariants? (Aquesta caracterització és univariant, atès que usa descriptius univariants de les variables.)

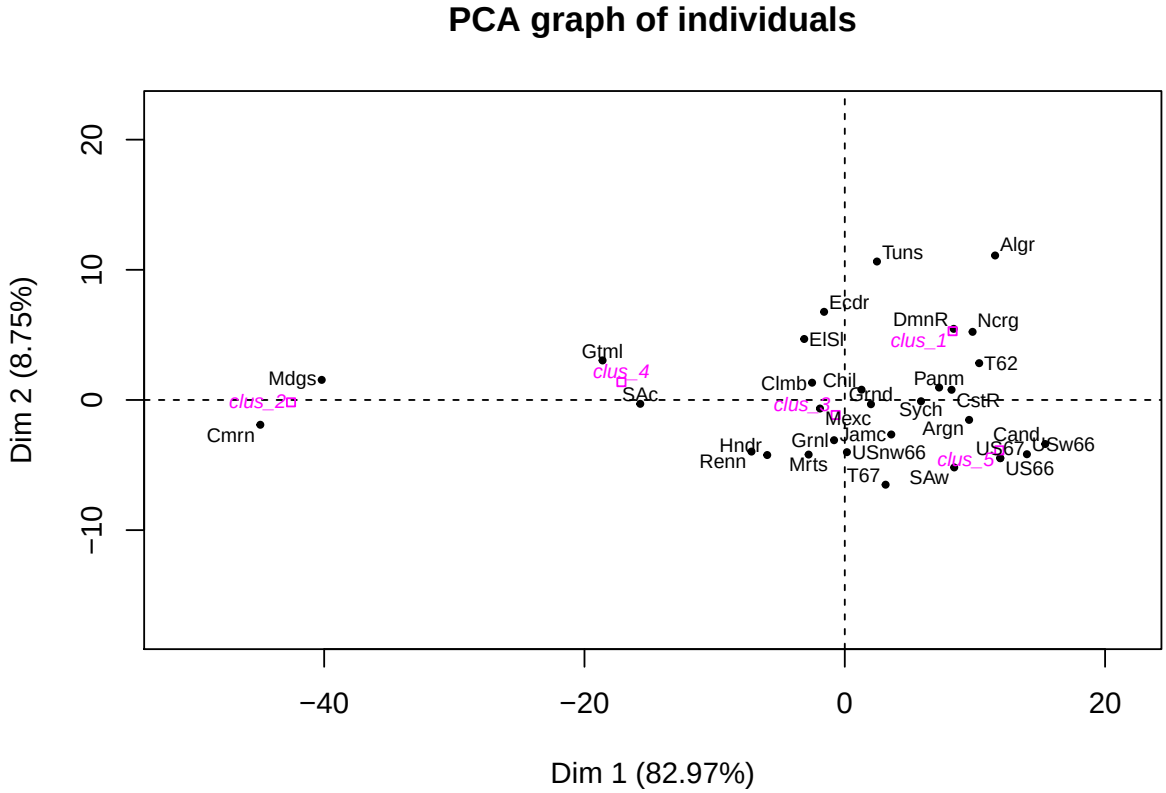
(d) Fes una matriu de gràfiques 2D de les relacions bivariades entre les variables de l'arxiu *life* amb el factor *clus*, usant `pairs()` (o si ho preferiu, `scatterplotMatrix()`).

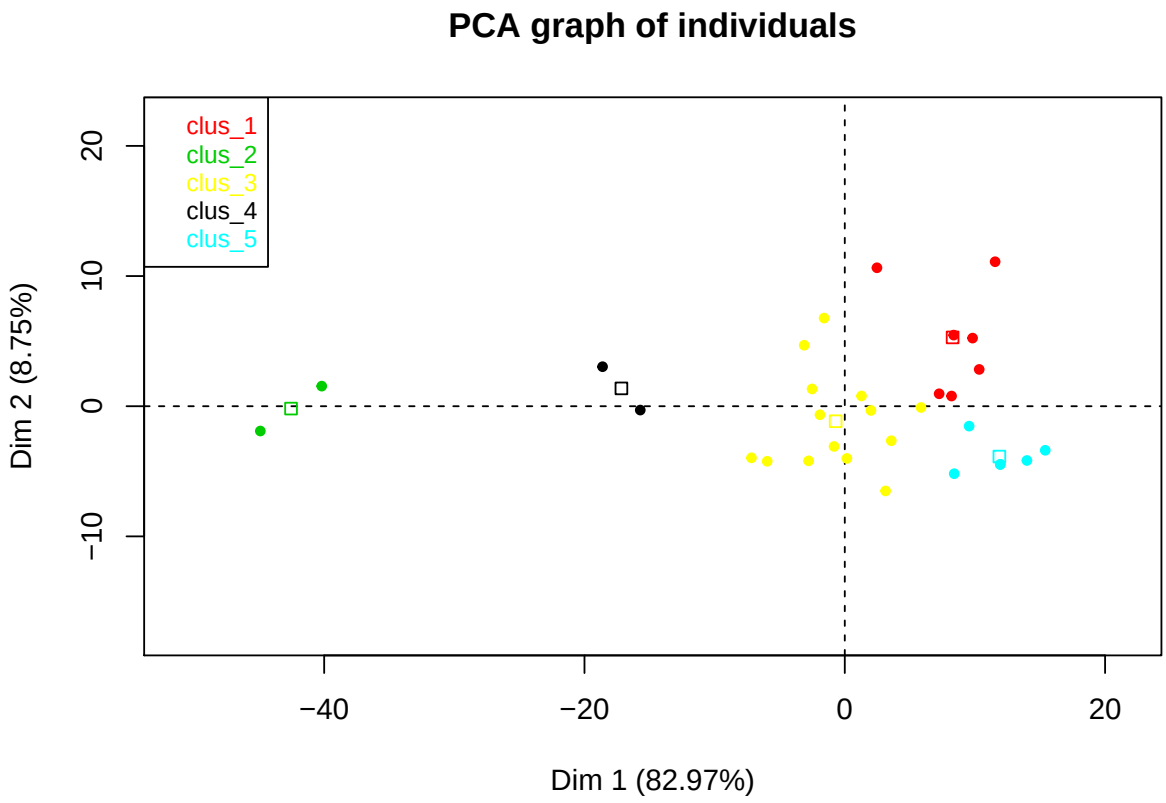


Interpreta els clústers en funció d'aquesta visualització. Quines parelles de variables discriminen més entre clústers? (Aquesta caracterització és bivariant.)

(e) Aplica **PCA** (amb 2 components,i sense tipificar (!)) a les 8 variables d'esperança de vida i representa els casos segons l'etiqueta del cúster al qual pertanyen. Interpreta els resultats, utilitzant tota la informació del resultat de components principals que faci falta. (En aquest cas, la caracterització és multivariant.) *Nota:* En aquest cas, no es tipifica per dues raons, els resultats dels clústers s'han fet sense tipificar i les variables tenen totes les mateixes unitats.


PCA graph of variables
w75
m50
m25
w50
m75
w25
m0
w0
Dim 2 (8.75%)
Dim 1 (82.97%)
3
2
1
0
−1
−2
0
2
4
6
8
10


PCA graph of individuals
Tuns
Algr
Ecdr
DmnR
Ncrg
clus_1
ElSl
Grml
clus_4
T62
Mdgs
clus_2
SAc
Clmb
Chil
Grnd
Panm
CstR
Sych
clus_3
Mexc
Argn
Cand
Jamc
Grnl
Hndr
USnw66
US67
USw66
Renn
Mrts
clus_5
US66
T67
SAw
Cmrn
Dim 2 (8.75%)
Dim 1 (82.97%)
20
10
0
−10
−40
−20
0
20



(f) Una altra opció és usar la funció `HCPC()` de la llibreria `FactoMineR`, aplicada directament sobre el resultat de comps principals. Explora: arguments de la funció, gràfiques, resultats, etc. Explora també la funció `fviz_cluster()` de la llibreria `factoextra`.

Nota: Aplicant els clústers sobre les primeres components, sempre i quan expliquin suficient variància, s'aconsegueixen uns grups que poden ser més 'robusts' perquè s'elimina soroll de les dades.

```
## Loading required package: factoextra
## Loading required package: ggplot2
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

Referències:

B. Everitt, *An R and S+ Companion to Multivariate Analysis*. (Ref. completa a la bibliografia del curs).

J. Josse (2010); *Principal component methods, hierarchical clustering, partitional clustering: why would we need to choose for visualizing data?*

https://pdfs.semanticscholar.org/0433/5d99d840ac3370f5aeb262828cf127d3ff1c.pdf?$_ga$ = 2.10827699.625578034.1589642240-901937157.1553510316