Mini projet

EL Hadrami N'DOYE

29/12/2020

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'gplots'
## The following object is masked from 'package:stats':
##
##     lowess
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
## corrplot 0.84 loaded
```

Soient les noms suivants qui remplacera les variables du jeux de données dans la partie contribution des varaibles

Dsc : Droit, sciences politiques

Seg : Sciences economiques, gestion

Aec : Administration economique et sociale

Lsla : Lettres, sciences du langage, arts

Shs : Sciences humaines et sociales

L : Langues

P2lsh : Pluri lettres langues sciences humaines

Sfa : Sciences fondamentales et applications

Dsp : Droit, sciences politiques

Staps : Sciences et techniques des activites physiques et sportives

Pre-traitement

```
etudiants <- read.csv("data/etudiants.csv",header = TRUE,sep = ";")
filiere <- etudiants[,1]
etudiants <- as.matrix(etudiants[2:13])
rownames(etudiants) <- filiere
```

```r
etudiants.active <- as.data.frame(etudiants[,1:6])
head(etudiants.active,4)
```

```
##                                     Licence.F Licence.H Master.F Master.H
## Droit, sciences politiques              69373     37317    42371    21693
## Sciences economiques, gestion           38387     37157    29466    26929
## Administration economique et sociale    18574     12388     4183     2884
## Lettres, sciences du langage, arts      48691     17850    17672     5853
##                                     Doctorat.F Doctorat.H
## Droit, sciences politiques                4029       4342
## Sciences economiques, gestion             1983       2552
## Administration economique et sociale         0          0
## Lettres, sciences du langage, arts        4531       2401
```

```r
summary(etudiants.active)
```

```
##    Licence.F       Licence.H        Master.F        Master.H
##  Min.   : 1779   Min.   :  726   Min.   : 1963   Min.   :  811
##  1st Qu.:19570   1st Qu.:15566   1st Qu.: 5910   1st Qu.: 3948
##  Median :31352   Median :19570   Median :15132   Median : 7155
##  Mean   :38901   Mean   :25490   Mean   :18238   Mean   :14341
##  3rd Qu.:59225   3rd Qu.:37277   3rd Qu.:26518   3rd Qu.:21382
##  Max.   :94346   Max.   :54861   Max.   :43016   Max.   :48293
##    Doctorat.F       Doctorat.H
##  Min.   :   0.0   Min.   :    0.0
##  1st Qu.: 600.8   1st Qu.:  472.8
##  Median :3006.0   Median : 2476.5
##  Mean   :3041.8   Mean   : 3424.0
##  3rd Qu.:4500.0   3rd Qu.: 5009.5
##  Max.   :7787.0   Max.   :11491.0
```

Realisation d'un test de Khi-deux

```r
chisq.test(etudiants.active)
```

```
##
##  Pearson's Chi-squared test
##
## data:  etudiants.active
## X-squared = 170789, df = 45, p-value < 2.2e-16
```

Le test de khi-deux donne un p-value < 0.05 donc elle y a une dependance significative entre les variables sur les differents individus

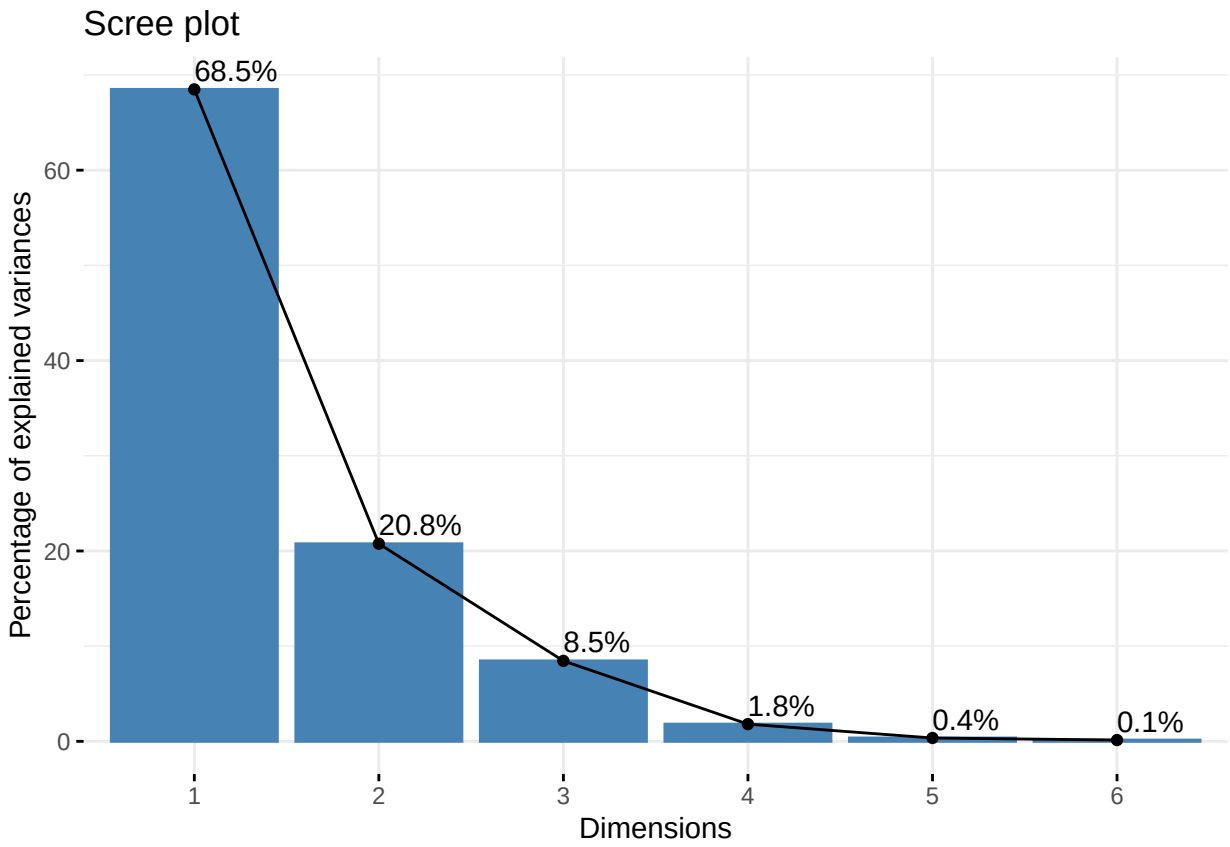
Realisation d'une ACP

```r
#etudiants.active <- etudiants[,2:12]
res.acp <- PCA(etudiants.active,scale.unit = TRUE,graph = FALSE)
res.acp$eig
```

```
##         eigenvalue percentage of variance cumulative percentage of variance
## comp 1 4.109124790             68.4854132                           68.48541
## comp 2 1.245412990             20.7568832                           89.24230
## comp 3 0.507217178              8.4536196                           97.69592
## comp 4 0.108686680              1.8114447                           99.50736
## comp 5 0.021274439              0.3545740                           99.86193
## comp 6 0.008283924              0.1380654                          100.00000
```

Graphe des valeurs propres

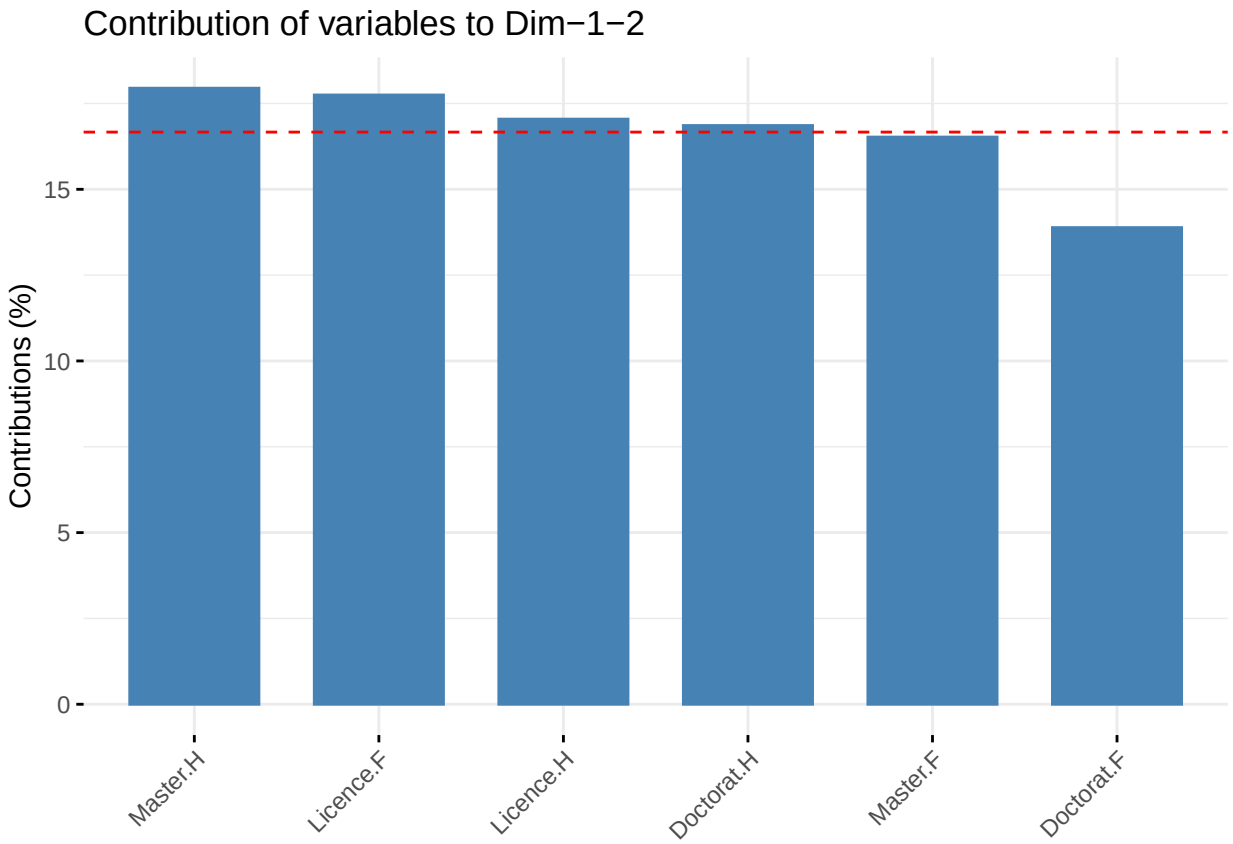
```
fviz_eig(res.acp, addlabels = TRUE)
```



Les deux premières composantes principales expliquent 89.24% de la variation,donc les deux premiers axes peuvent etre acceptés pour la suite de l'analyse.

Contributions des variables

```
fviz_contrib(res.acp, choice = "var", axes = 1 :2)
```

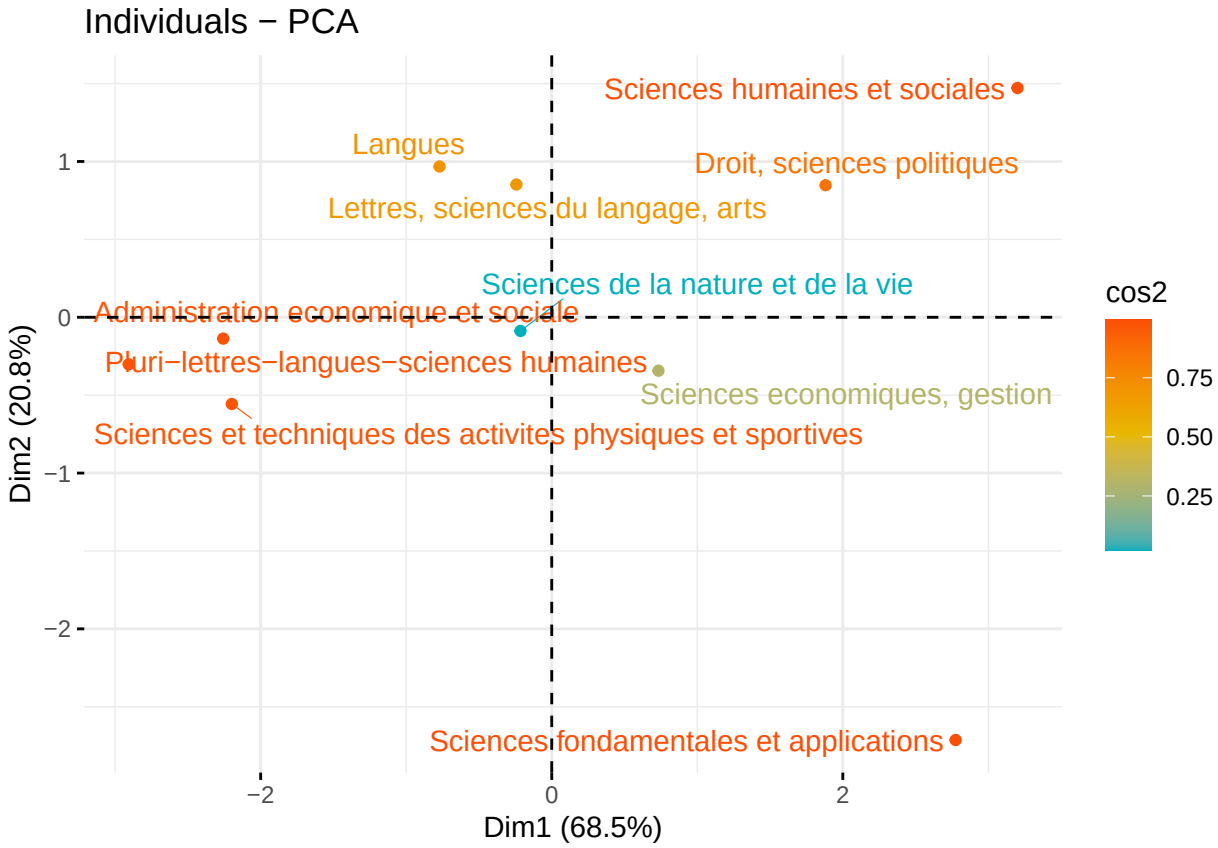


Observations

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue. Donc les variables les plus contributives sont Master.H,Licence.F,Licence.H,Doctorat.H.

Graphiques des individus

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

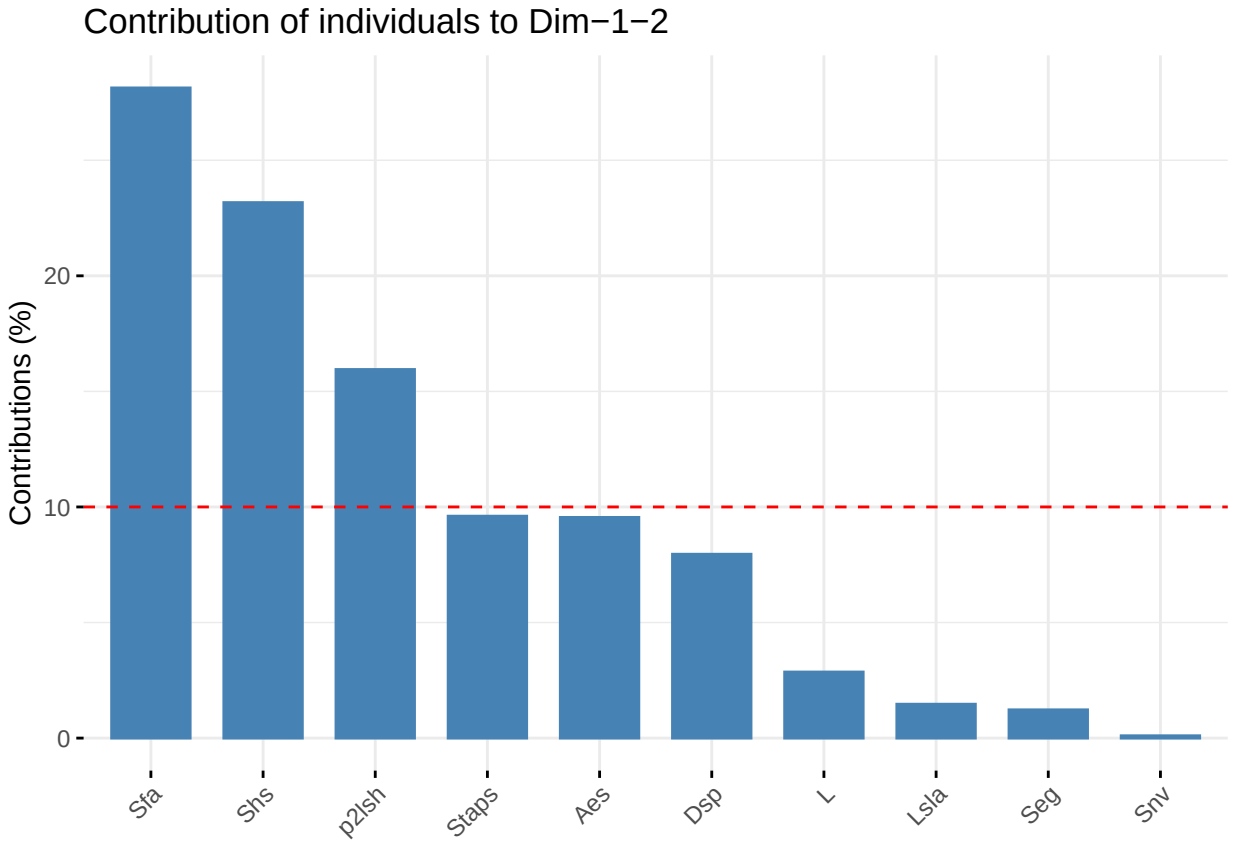


Observation:

Un cos2 élevé indique une bonne représentation de l'individus sur les axes principaux en considération(comme on peut le voir dans le graphe ci dessus).

- Les individus ayant choisis les formations sciences humaines sociales et Droits sciences politique sont representés sur le plan par des points de couleur semblable et sont proche l'un a l'autre, donc il forment un regroupement a deux individus, idem pour Aes,P2lsh et staps ce qui forme un autre regroupement de trois individus.
- Les individus ayant choisis les formations Langues et Lsla sont representés sur le plan par des points de meme couleur et sont proche l'un a l'autre ce qui forme un autre regroupement de deux individus.
- Les individus ayant choisis les formations Snv, Seg et Sfa forment chacun un regroupement d'un seul individus.

```
filiere <- c("Dsp","Seg","Aes","Lsla","L","Shs","p2lsh","Sfa","Snv","Staps")
rownames(etudiants.active) <- filiere
acp <- PCA(etudiants.active,scale.unit = TRUE,graph = FALSE)
fviz_contrib(acp, choice = "ind", axes = 1 :2)
```

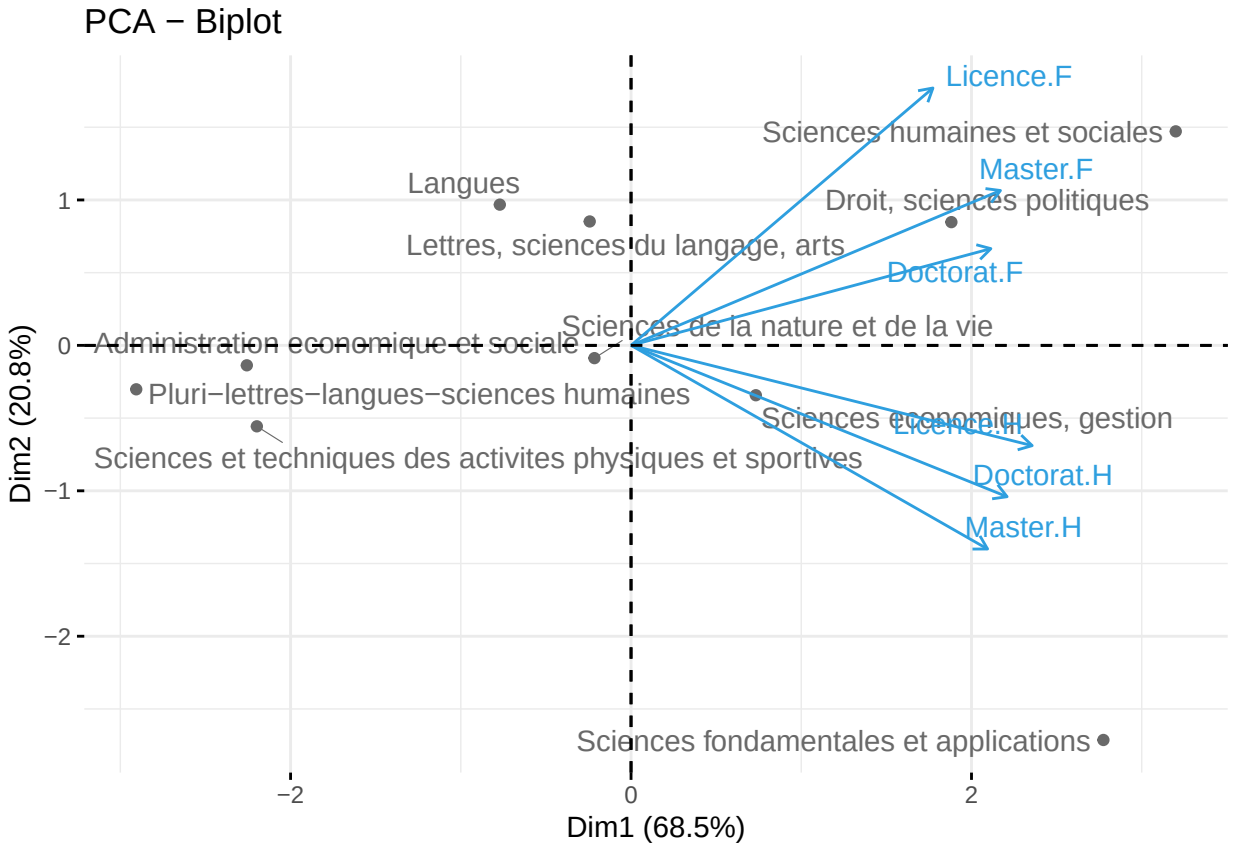


Observation:

La ligne en pointillé rouge, sur le graphique ci-dessus, indique la contribution moyenne attendue. Donc les individus qui contribue le plus sont Sfa,Shs,p2lsh cela veut dire qu'il y a plus d'etudiants diplomés dans ces filieres que les autres filieres.

Graphiques des individus et des variables

```
fviz_pca_biplot(res.acp,
repel = TRUE,col.var = "#2E9FDF", # Couleur des variables
col.ind = "#696969")
```



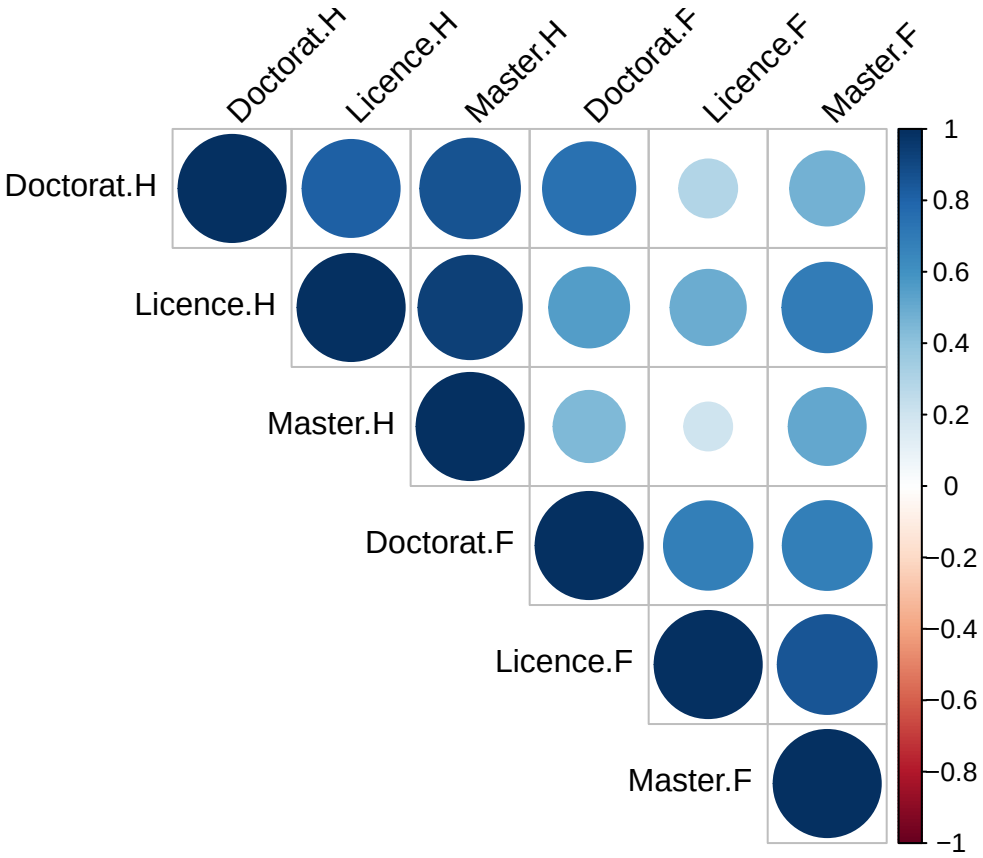
Observations

- Les individus feminins obtiennent plus de diplome de Licence,Master et Doctorat dans le domaine de sciences humaines et sociales et Droits sciences politique.
- Les individus masculins obtiennent plus de diplome de Licence , Master et Doctorat dans le domaine de Sciences economiques gestions.

Classification hiérarchique ascendante des données

- Realisation d'un correlogramme pour toute les variables

```
corr <- cor(etudiants.active,method = "pearson")
corrplot(corr, type="upper", order="hclust", tl.col="black", tl.srt=45)
```

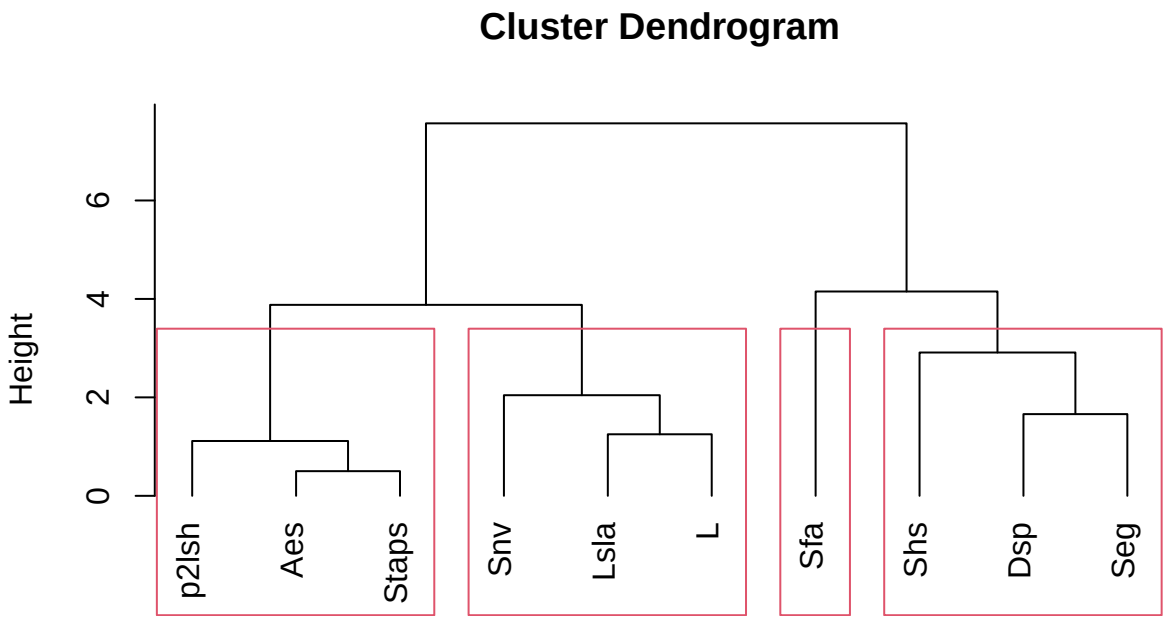


Obsevations

Avant d'entammer la classification on constate deja qu'elle y a une forte correlation entre la variable Doctorat.H et Licence.H puis Doctorat.H et Master.H donc y a une possiblité de regrouper ces trois variables.

Utilisation de la fonction hclust pour la classification

```
filiere <- c("Dsp","Seg","Aes","Lsla","L","Shs","p2lsh","Sfa","Snv","Staps")
rownames(etudiants.active) <- filiere
etudiants.cr <- scale(etudiants.active,center=T,scale=T)
d.etudiants <- dist(etudiants.cr)
tree <- hclust(d.etudiants,method = "ward.D2")
plot(tree,hang = -1)
rect.hclust(tree,k=4)
```




```
print(sort(cutree(tree,k=4)))
```

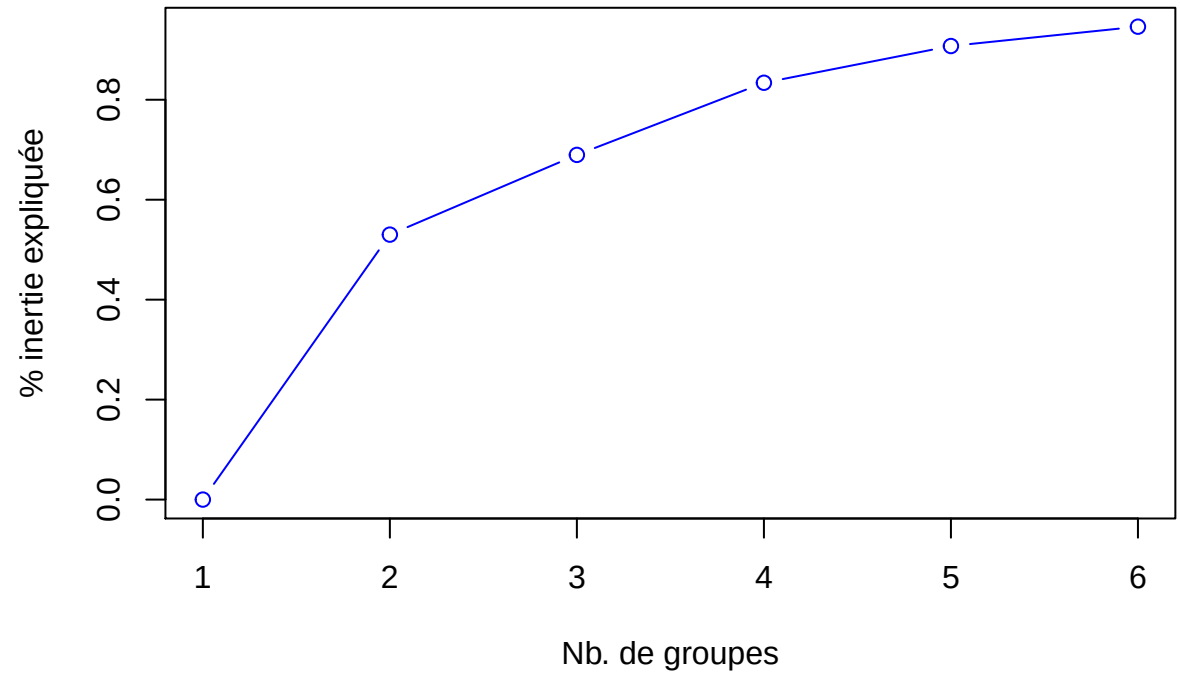
```
##   Dsp   Seg   Shs   Aes p2lsh Staps  Lsla     L   Snv   Sfa
##     1     1     1     2     2     2     3     3     3     4
```

Observations

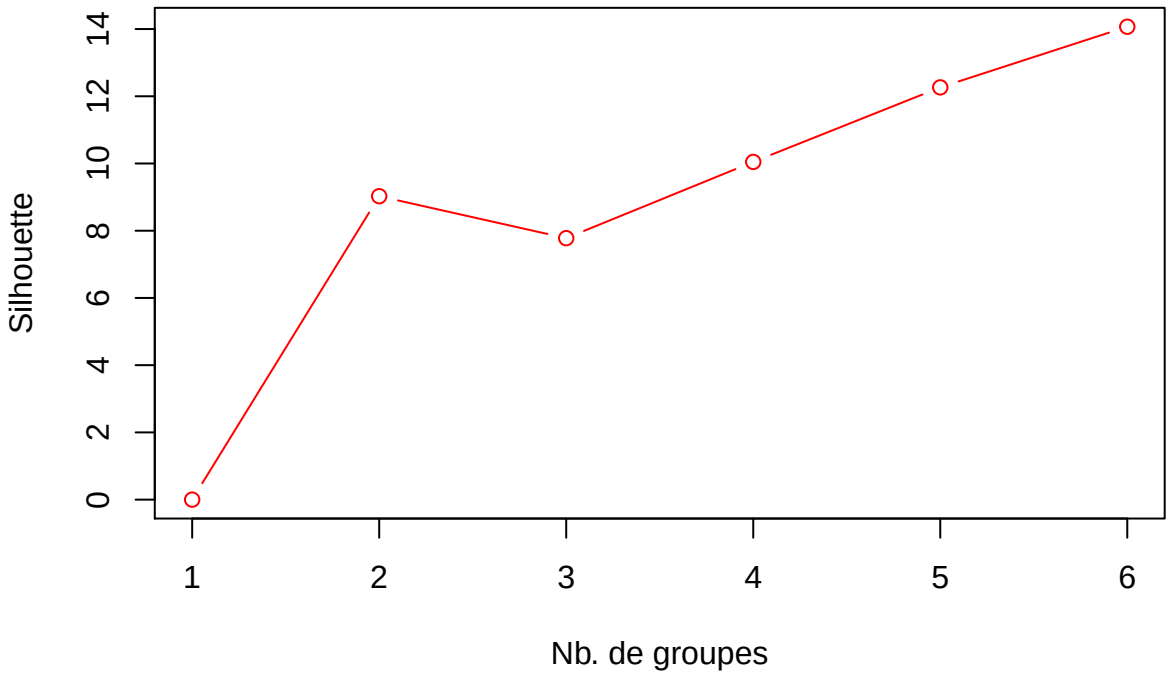
Le dendogramme suggere un decoupage en 4 groupes, nous verrons ensuite la methode de K-means pour trouver le nombre de decoupages (k) optimal afin de confirmer ou rejetter le nombre de decoupages trouvés sur la methode de classification.

Methode de K-means

```
groupes.kmeans <- kmeans(etudiants.cr,centers=4,nstart=5)
inertie <- rep(0,times=6)
for (k in 2:6){
  group <- kmeans(etudiants.cr,centers = k ,nstart=5)
  inertie[k] <- group$betweenss/group$totss
}
plot(1:6,inertie,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée",col="blue")
```




```
solkmeans <- kmeansruns(etudiants.cr,krange=2:6,criterion="ch")
plot(1:6,solkmeans$crit,type="b",xlab="Nb. de groupes",ylab="Silhouette",col="red")
```



Observations

Le graphe 1 montre l'evolution de la proportion d'inertie expliquée par la partition et le graphe 2 cherche a maximiser la valeur k pour la partition en utilisant la fonction kmeansruns du package «fpc». Dans le deux graphe on confirme donc que le K optimal est de 6 ce qui montre qu'on a six groupes qui forment des clusters.