EXAMEN TERMINAL SEPTEMBRE 2024 - PROF MONSAN

UFHB UFR MI - DATASCIENCE

CHERIF Mohamed Lamine

GOULIA Junias

2024-09-17

1 Exercice 0: rappels

La classification non supervisée regroupe des méthodes qui visent à regrouper des individus en groupes homogènes sans utiliser d'étiquettes prédéfinies.

1.1 1. K-means

- **Principe** : Regroupe les individus en k clusters en minimisant la distance intra-cluster.
- **Avantages** :
  - Simple à comprendre et rapide à calculer pour des grands ensembles de données.
  - Particulièrement efficace pour des groupes sphériques.
- **Inconvénients** :
  - Il faut choisir un nombre de clusters k à l'avance, ce qui n'est pas toujours évident.
  - Sensible aux valeurs aberrantes et à la forme des clusters (ne fonctionne pas bien pour des clusters de formes complexes).

1.2 2. Clustering hiérarchique (CAH - Classification Ascendante Hiérarchique)

- **Principe** : Crée une hiérarchie de groupes en fusionnant progressivement des individus ou des groupes d'individus selon une certaine distance.
- **Avantages** :
  - Ne nécessite pas de définir le nombre de clusters à l'avance.
  - Produit une dendrogramme qui permet d'explorer les niveaux de regroupements possibles.
- **Inconvénients** :
  - Peut être computationalement coûteux pour de grandes bases de données.
  - Choix arbitraire du seuil de coupure pour déterminer les clusters finaux.

1.3 3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Principe** : Regroupe les points denses, en fonction de la densité des individus dans une zone, et identifie les points isolés comme des outliers.
- **Avantages** :
  - Capable d'identifier des clusters de formes irrégulières.
  - Insensible aux outliers et ne nécessite pas de spécifier le nombre de clusters.
- **Inconvénients** :
  - Le choix des paramètres de densité (ε et MinPts) est délicat.
  - Moins performant lorsque les densités de clusters varient beaucoup.

2 Exercice 2 : cas pratique

2.1 Contexte

Le radiologue a réalisé une IRM et a obtenu 108 variables sur une région d'intérêt (ROI). Il cherche à comprendre comment exploiter ces données pour mieux regrouper les patients ou extraire des informations pertinentes sur les variables elles-mêmes.

2.2 Problématique

Comment peut-on regrouper les patients en fonction des 108 variables issues de l'IRM ? Peut-on aussi classer les variables elles-mêmes pour en tirer des conclusions utiles ?

2.3 Objectif et intérêt

Notre objectif est de proposer une méthode de classification des patients (clustering) en groupes homogènes, tout en explorant la structure des variables pour détecter d'éventuelles corrélations ou regroupements significatifs entre elles.

2.4 Méthodologie

On va aborder la solution en plusieurs étapes :

1. **Analyse exploratoire** pour comprendre les données.
2. **Analyse en composantes principales (ACP)** pour réduire la dimension et visualiser les données.
3. **Classification ascendante hiérarchique (CAH)** pour regrouper les patients.
4. **K-means** pour une autre approche de classification.

2.4.1 Étape 1 : Analyse exploratoire

On commence par charger et explorer les données.

```
# Charger les données
donnees_medicales <- read.csv("medical.csv",header = TRUE, sep = ";",dec = ",")

# Afficher un résumé des données
summary(donnees_medicales)
```

```
##   meanssf0RKid      sdssf0RKid    entropyssf0RKid  mppssf0RKid
##  Min.   : 38.40   Min.   :18.52   Min.   :4.250   Min.   : 40.45
##  1st Qu.: 89.01   1st Qu.:28.09   1st Qu.:4.638   1st Qu.: 89.82
##  Median :102.50   Median :34.03   Median :4.840   Median :102.81
##  Mean   :112.05   Mean   :34.90   Mean   :4.815   Mean   :113.00
##  3rd Qu.:132.05   3rd Qu.:41.19   3rd Qu.:4.992   3rd Qu.:132.18
##  Max.   :218.98   Max.   :67.27   Max.   :5.440   Max.   :218.98
##  skewnessssf0RKid kurtosisssf0RKid   meanssf2RKid      sdssf2RKid
##  Min.   :0.0000   Min.   : 0.0200   Min.   : 0.050   Min.   : 42.63
##  1st Qu.:0.2050   1st Qu.: 0.2425   1st Qu.: 2.000   1st Qu.: 62.96
##  Median :0.4200   Median : 0.6050   Median : 4.015   Median : 79.77
##  Mean   :0.5629   Mean   : 1.5007   Mean   :17.473   Mean   : 86.33
##  3rd Qu.:0.7050   3rd Qu.: 1.3825   3rd Qu.:27.325   3rd Qu.:100.88
##  Max.   :2.6500   Max.   :13.7000   Max.   :93.950   Max.   :179.31
##  entropyssf2RKid  mppssf2RKid     skewnessssf2RKid kurtosisssf2RKid
##  Min.   :5.020   Min.   : 31.64   Min.   :0.0200   Min.   :0.000
##  1st Qu.:5.367   1st Qu.: 46.40   1st Qu.:0.1500   1st Qu.:0.190
##  Median :5.610   Median : 64.06   Median :0.2400   Median :0.440
##  Mean   :5.628   Mean   : 74.39   Mean   :0.3571   Mean   :1.007
##  3rd Qu.:5.880   3rd Qu.: 95.06   3rd Qu.:0.4925   3rd Qu.:0.835
##  Max.   :6.420   Max.   :171.83   Max.   :1.4200   Max.   :9.270
```

```
# Vérifier s'il y a des valeurs manquantes
sum(is.na(donnees_medicales))
```

```
## [1] 0
```

```
# Afficher les premières lignes pour voir à quoi ça ressemble
head(donnees_medicales)
```

meanssf0RKid	sdssf0RKid	entropyssf0RKid	mppssf0RKid	skewnessssf0RKid	kurtosisssf0RKid	meanssf2RKid	sdssf2RKid	entropyssf2RKid	mppssf2RKid	skewn
132.98	31.32	4.77	133.49	0.87	3.47	3.52	64.36	5.49	50.97	0.22
130.51	42.65	4.98	131.14	1.05	0.77	40.94	106.45	5.97	98.55	0.37
94.39	36.12	4.91	95.95	0.40	0.52	7.69	75.05	5.55	61.25	0.16
91.95	27.49	4.67	92.19	0.61	0.39	22.98	79.53	5.67	70.36	0.63
64.14	33.96	4.89	65.22	0.21	0.32	36.12	99.05	5.93	94.88	0.02
135.05	40.83	5.07	135.40	0.19	0.36	2.94	95.26	5.81	76.46	0.13

```
# Vérifier la structure des données (types de variables)
str(donnees_medicales)
```

```
## 'data.frame':    68 obs. of  108 variables:
##  $ meanssf0RKid    : num  133 130.5 94.4 92 64.1 ...
##  $ sdssf0RKid      : num  31.3 42.6 36.1 27.5 34 ...
##  $ entropyssf0RKid : num  4.77 4.98 4.91 4.67 4.89 5.07 4.99 5.35 4.72 5.07 ...
##  $ mppssf0RKid     : num  133.5 131.1 96 92.2 65.2 ...
##  $ skewnessssf0RKid: num  0.87 1.05 0.4 0.61 0.21 0.19 0.42 0.93 0.16 0.24 ...
##  $ kurtosisssf0RKid: num  3.47 0.77 0.52 0.39 0.32 0.36 0.03 0.05 0.15 0.75 ...
##  $ meanssf2RKid    : num  3.52 40.94 7.69 22.98 36.12 ...
##  $ sdssf2RKid      : num  64.4 106.5 75 79.5 99 ...
##  $ entropyssf2RKid : num  5.49 5.97 5.55 5.67 5.93 5.81 6.06 6.42 5.4 6.15 ...
##  $ mppssf2RKid     : num  51 98.5 61.2 70.4 94.9 ...
##  $ skewnessssf2RKid: num  0.22 0.37 0.16 0.63 0.02 0.13 0.59 0.17 0.16 0.49 ...
##  $ kurtosisssf2RKid: num  0.69 0.36 0.52 1.02 0.22 0.28 0.22 0.16 0.22 0 ...
##  $ meanssf3RKid    : num  7.12 64.9 10.92 34.57 48.35 ...
##  $ sdssf3RKid      : num  51.3 135 64.1 92.8 96.6 ...
##  $ entropyssf3RKid : num  5.26 6.2 5.38 5.82 5.91 5.67 6.18 6.62 5.05 6.36 ...
##  $ mppssf3RKid     : num  41.6 134.2 59.6 86 98.2 ...
##  $ skewnessssf3RKid: num  0.3 0.21 0.18 0.42 0.08 0.22 0.39 0.08 0.33 0.3 ...
##  $ kurtosisssf3RKid: num  0.25 0.17 0.34 0.35 0.22 0.19 0.13 0.74 0.28 0.71 ...
##  $ meanssf4RKid    : num  7.19 86.27 9.1 41.9 57.85 ...
##  $ sdssf4RKid      : num  46 159.4 60.4 101.4 95 ...
```

2.4.2 Étape 2 : Analyse en Composantes Principales (ACP)

Utilisons l'ACP pour réduire la dimension et visualiser les données sur un plan 2D tout en gardant l'essentiel de l'information.

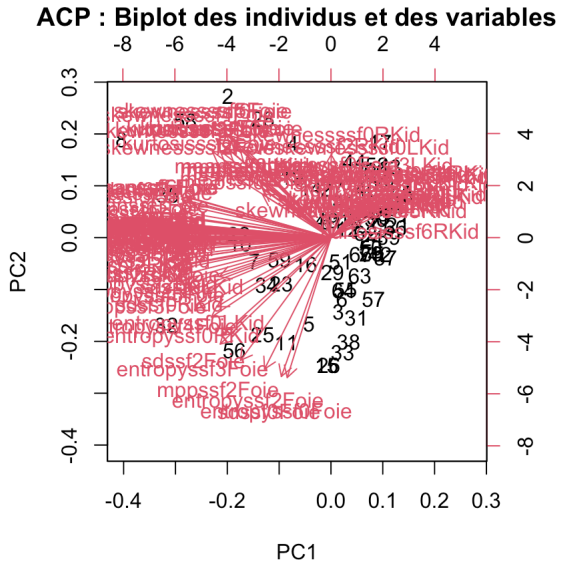
```
# S'assurer que toutes les colonnes sont numériques
donnees_medicales_numeric <- data.frame(lapply(donnees_medicales, as.numeric))

# Exécuter l'ACP
acp_resultat <- prcomp(donnees_medicales_numeric, scale. = TRUE)

# Résumé de l'ACP pour voir la proportion de variance expliquée par les composantes
summary(acp_resultat)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     7.4989 3.12525 2.68724 2.24503 2.14113 1.83184 1.70395
## Proportion of Variance 0.5207 0.09044 0.06686 0.04667 0.04245 0.03107 0.02688
## Cumulative Proportion  0.5207 0.61112 0.67799 0.72466 0.76710 0.79817 0.82506
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.63225 1.51557 1.27641 1.23597 1.16933 1.10675 1.01954
## Proportion of Variance 0.02467 0.02127 0.01509 0.01414 0.01266 0.01134 0.00962
## Cumulative Proportion  0.84973 0.87100 0.88608 0.90023 0.91289 0.92423 0.93385
##                           PC15    PC16   PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.9467 0.92842 0.8184 0.75183 0.69295 0.65441 0.61781
## Proportion of Variance 0.0083 0.00798 0.0062 0.00523 0.00445 0.00397 0.00353
## Cumulative Proportion  0.9422 0.95013 0.9563 0.96157 0.96601 0.96998 0.97351
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.59624 0.55781 0.55136 0.47798 0.43230 0.40098 0.39372
## Proportion of Variance 0.00329 0.00288 0.00281 0.00212 0.00173 0.00149 0.00144
## Cumulative Proportion  0.97680 0.97969 0.98250 0.98462 0.98635 0.98783 0.98927
##                           PC29    PC30    PC31   PC32    PC33    PC34    PC35
## Standard deviation     0.36343 0.35293 0.34726 0.3116 0.28853 0.27672 0.25773
## Proportion of Variance 0.00122 0.00115 0.00112 0.0009 0.00077 0.00071 0.00062
## Cumulative Proportion  0.99049 0.99165 0.99276 0.9937 0.99443 0.99514 0.99576
```

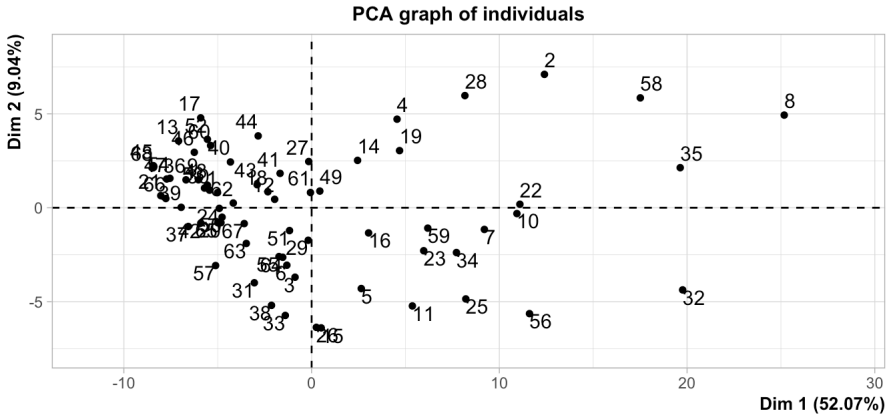
```
# Visualiser l'ACP (graphique des individus et des variables)
library(ggplot2)
biplot(acp_resultat, main = "ACP : Biplot des individus et des variables")
```

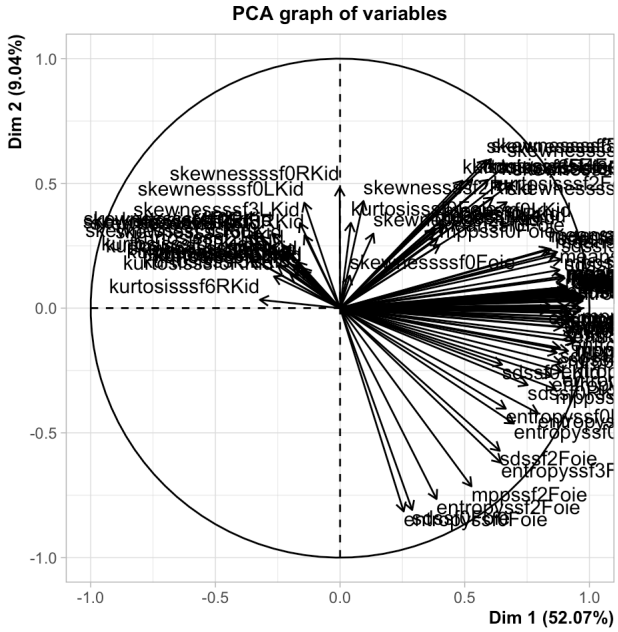



```
# Charger les librairies
library(FactoMineR)
library(factoextra)
```

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

```
# Effectuer l'ACP (sur des données numériques standardisées)
acp_fact <- PCA(donnees_medicales_numeric, scale.unit = TRUE, ncp = 5, graph = TRUE)
```

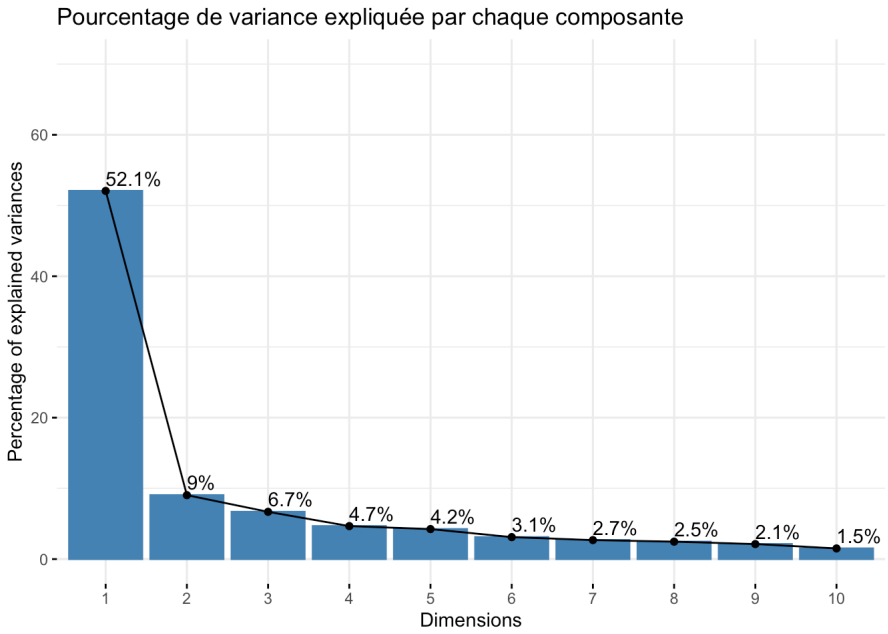




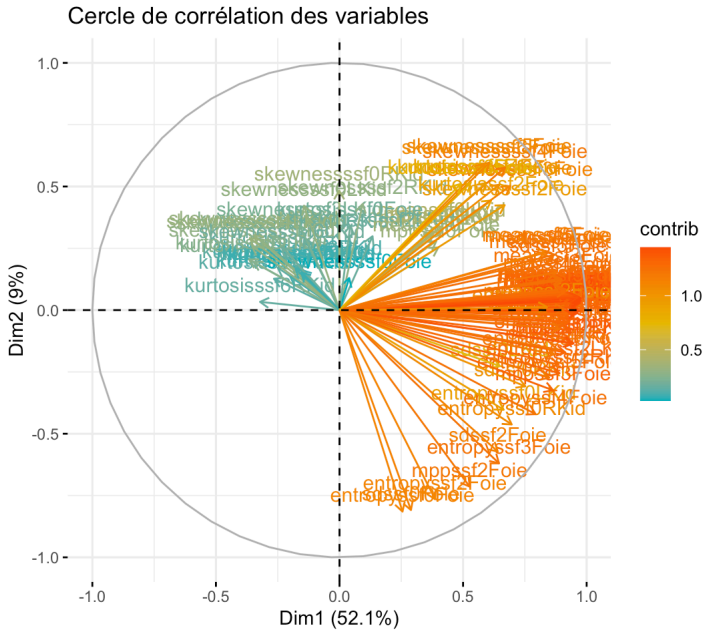

```
# Afficher le résumé de l'ACP
print(acp_fact)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 68 individuals, described by 108 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

```
# Afficher la proportion de variance expliquée par les composantes principales
fviz_screeplot(acp_fact, addlabels = TRUE, ylim = c(0, 70),
               title = "Pourcentage de variance expliquée par chaque composante")
```



```
# Cercle de corrélation pour voir quelles variables sont les plus importantes
fviz_pca_var(acp_fact, col.var = "contrib",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             title = "Cercle de corrélation des variables")
```




```
# Tableau des contributions des variables aux 2 premières composantes
print(acp_resultat$var$contrib[, 1:2])  # Contribution des variables aux axes 1 et 2
```

```
## NULL
```

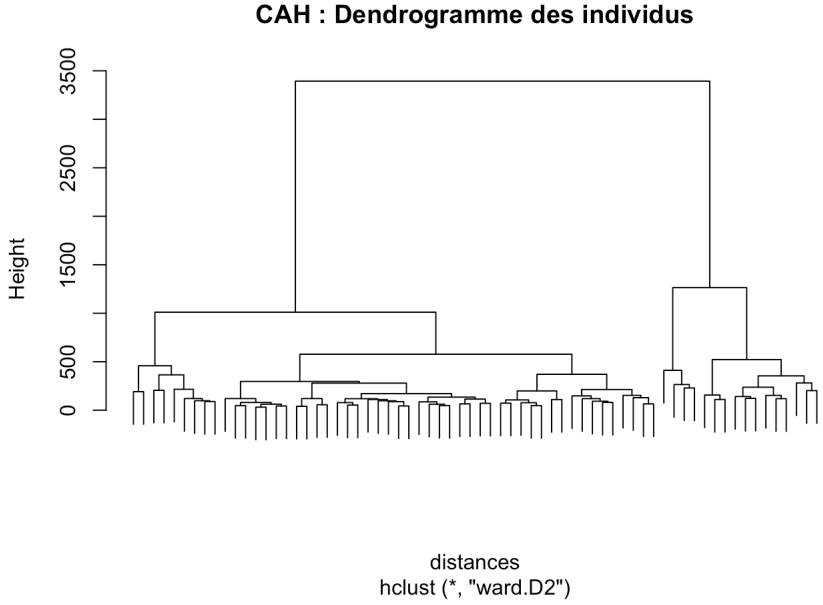
2.4.3 Étape 3 : Classification Ascendante Hiérarchique (CAH)

Utilisons la CAH pour regrouper les individus de manière hiérarchique, en créant un dendrogramme qui montre les similarités entre les individus.

```
# Calculer la matrice de distances entre individus
distances <- dist(donnees_medicales_numeric)

# Appliquer la CAH
cah_resultat <- hclust(distances, method = "ward.D2")

# Visualiser le dendrogramme
plot(cah_resultat, labels = FALSE, main = "CAH : Dendrogramme des individus")
```



```
# Découper le dendrogramme en 3 clusters
clusters_cah <- cutree(cah_resultat, k = 4)

# Ajouter ces clusters aux données
donnees_medicales_numeric$cluster_cah <- as.factor(clusters_cah)
head(donnees_medicales_numeric)
```

meanssf0RKid	sdssf0RKid	entropyssf0RKid	mppssf0RKid	skewnessssf0RKid	kurtosisssf0RKid	meanssf2RKid	sdssf2RKid	entropyssf2RKid	mppssf2RKid	skewn
132.98	31.32	4.77	133.49	0.87	3.47	3.52	64.36	5.49	50.97	0.22
130.51	42.65	4.98	131.14	1.05	0.77	40.94	106.45	5.97	98.55	0.37
94.39	36.12	4.91	95.95	0.40	0.52	7.69	75.05	5.55	61.25	0.16
91.95	27.49	4.67	92.19	0.61	0.39	22.98	79.53	5.67	70.36	0.63
64.14	33.96	4.89	65.22	0.21	0.32	36.12	99.05	5.93	94.88	0.02
135.05	40.83	5.07	135.40	0.19	0.36	2.94	95.26	5.81	76.46	0.13

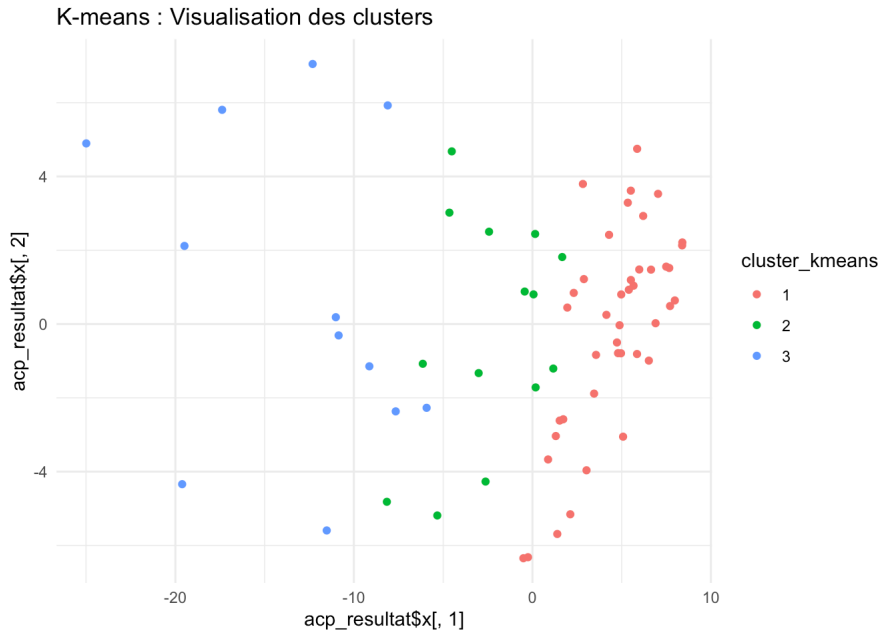
2.4.4 Étape 4 : K-means

Le K-means est une autre méthode pour regrouper les individus en fonction de leurs caractéristiques.

```
# Appliquer l'algorithme K-means avec 3 clusters
set.seed(123)  # Pour rendre les résultats reproductibles
kmeans_resultat <- kmeans(donnees_medicales_numeric, centers = 3)

# Ajouter les clusters K-means aux données
donnees_medicales_numeric$cluster_kmeans <- as.factor(kmeans_resultat$cluster)

# Visualiser les clusters obtenus avec K-means (projection sur les 2 premières dimensions ACP)
ggplot(donnees_medicales_numeric, aes(x = acp_resultat$x[,1], y = acp_resultat$x[,2], color = cluster_kmeans)) +
  geom_point() +
  labs(title = "K-means : Visualisation des clusters") +
  theme_minimal()
```



2.4.5 Conclusion

Grâce à l'ACP, nous avons réussi à réduire la dimension des données et à visualiser les individus et les variables de manière plus simple. La CAH et le K-means nous ont permis de regrouper les patients en fonction des 108 variables. Ces techniques nous offrent deux manières complémentaires de classer les individus.

2.4.6 Synthèse

- **ACP** : a aidé à simplifier l'interprétation des variables en les projetant sur deux dimensions principales.
- **CAH** : a permis de créer un dendrogramme des individus, révélant des groupes potentiels.
- **K-means** : a proposé une méthode non hiérarchique de classification, avec des résultats visuels clairs sur les clusters.

2.4.7 Recommandation

Nous recommandons d'utiliser la méthode K-means si l'on cherche des regroupements clairs et facilement visualisables. Cependant, si l'on souhaite une analyse plus détaillée des relations entre les individus, la CAH est plus adaptée. Une exploration plus approfondie des variables via une analyse en composantes principales peut aussi permettre d'identifier des patterns intéressants.