Nantes Université — UFR Sciences et Techniques

Master informatique

Année académique 2024-2025

Analyse de données

Travaux Pratiques

BEN HAMOUDI MELYSSA
YADEL GAYA

10/01/2025

Sommaire

Introduction:

L’analyse exploratoire des données est une étape essentielle pour mieux comprendre et interpréter des jeux de données complexes. Ce travail pratique se concentre sur l'application de deux techniques clés de l’analyse multivariée : l’Analyse en Composantes Principales (ACP) et les méthodes de classification. Ces approches permettent de réduire la dimensionnalité des données tout en conservant les informations essentielles, ainsi que de regrouper des individus ayant des caractéristiques similaires.

Dans le cadre de ce TP, les données utilisées proviennent d’un fichier nommé `euro.csv`, qui contient des informations socio-économiques sur différents pays européens. L'objectif est de mieux comprendre les similitudes et les différences entre ces pays à travers des méthodes visuelles et algorithmiques, tout en comparant les résultats obtenus avec différentes approches.

Objectifs:

1. Explorer les données et vérifier leur qualité.
2. Appliquer une Analyse en Composantes Principales (ACP) afin d’identifier les axes principaux de variation et de simplifier la visualisation des données.
3. Mettre en œuvre des techniques de classification, notamment K-means et la Classification Ascendante Hiérarchique (CAH), pour partitionner les pays en groupes homogènes.
4. Comparer les résultats des deux approches de classification afin de comprendre leurs différences et leurs complémentarités.
5. Analyser et interpréter les résultats dans une zone restreinte de l’espace factoriel pour étudier les proximités perçues entre certains pays européens.

Analyse en composantes principales:

1. Dans l’environnement R, chargez les données euro. Assurez-vous qu’elles sont correctement interprétées. Faites en particulier attention au nom des lignes et des colonnes.

```
> euro <- read.csv("~/TP AED/euro.csv", sep=";")
>  View(euro)
```

J'ai utilisé la fonction `read.csv()` pour lire le fichier CSV. J'ai spécifié le chemin du fichier et le séparateur utilisé dans le fichier

Affichage des données : Une fois le fichier chargé dans l'objet `euro`, j'ai souhaité vérifier visuellement son contenu. Pour cela, j'ai utilisé la commande `View(euro)` dans RStudio.

Action dans RStudio : Après avoir exécuté la commande `View(euro)`, une nouvelle fenêtre s'est automatiquement ouverte dans l'interface RStudio, affichant les données sous forme de tableau. Dans cette fenêtre, j'ai pu examiner les noms des colonnes et des lignes pour vérifier que les données avaient été correctement importées.

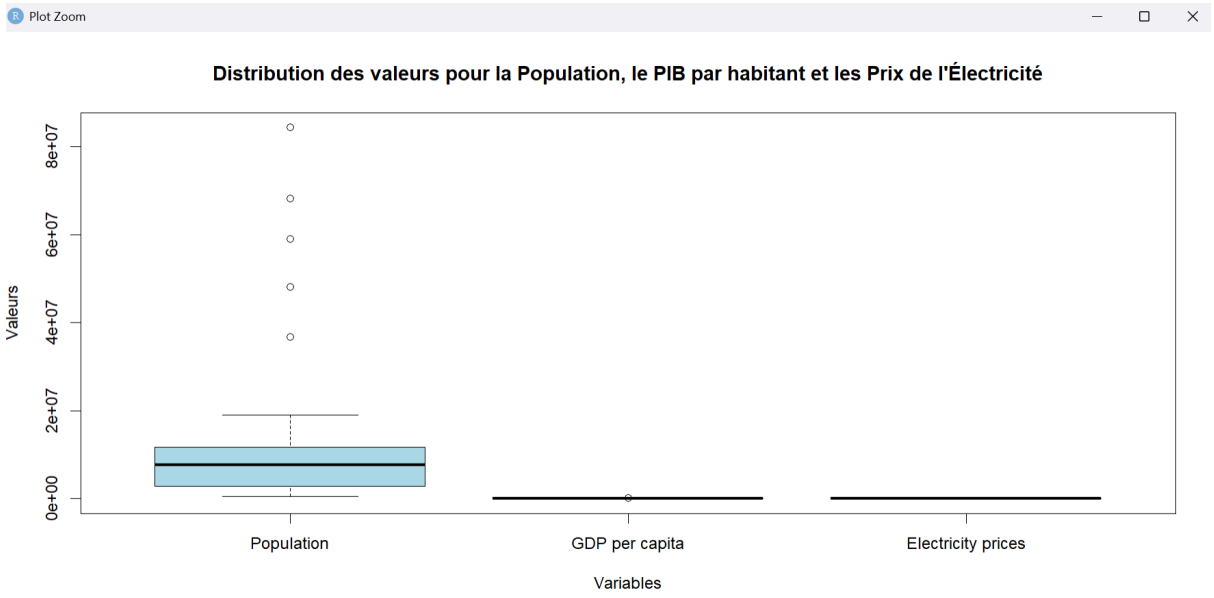
	X	Population	Youth.population	First.time.asylum.applicants	Gender.pay.gap
1	Austria	9104772	16.9	56135	18.4
2	Belgium	11742796	17.8	29260	5.0
3	Bulgaria	6447710	13.2	22390	13.0
4	Croatia	3850894	15.9	1635	12.5
5	Cyprus	920701	19.8	11660	10.2
6	Czechia	10827529	15.1	1130	17.9
7	Denmark	5932654	19.1	2355	13.9
8	Estonia	1365884	15.2	3980	21.3
9	Finland	5563970	17.2	4450	15.5
10	France	68172977	17.5	145095	13.9
11	Germany	84358845	16.0	329035	17.7
12	Greece	10413982	15.2	57895	15.0
13	Hungary	9599744	16.4	30	17.5
14	Ireland	5271395	18.4	13220	9.3
15	Italy	58997201	15.0	130565	4.3

---

2. A l'aide d'une visualisation par boîte à moustache pour chaque variable, faites vos premiers commentaires sur la distribution des valeurs sur chaque variable. Vous pourrez également utiliser d'autres méthodes d'exploration statistique que vous avez étudié l'année dernière.

```
> boxplot(euro$Population, euro$GDP.per.capita, euro$Electricity.prices,
+             names = c("Population", "GDP per capita", "Electricity prices"),
main = "Distribution des valeurs pour la Population, le PIB par habitant et
les Prix de l'Électricité",
+             xlab = "Variables",
ylab = "Valeurs",
col = c("lightblue", "lightgreen", "lightpink"))
```

Résultat:



Pour analyser la boîte à moustaches , nous allons interpréter les valeurs relatives à la population,au PIB par habitant (GDP per capita) et aux prix de l'électricité (Electricity prices). Ces variables représentent des données quantitatives qui ont des échelles très différentes, comme l'indique la boîte à moustaches.

Observations principales à partir du boxplot :

1. **Population :**
   - La boîte à moustaches montre une large dispersion des valeurs de la population parmi les pays.
   - On observe la présence de plusieurs valeurs aberrantes (outliers), qui pourraient correspondre aux pays les plus peuplés comme l'Allemagne, la France.
   - La médiane semble être située près du bas de la boîte, indiquant une asymétrie (positivement décalée).
2. **PIB par habitant :**
   - Les valeurs du PIB par habitant sont beaucoup plus concentrées.
   - La boîte et les moustaches sont étroites, ce qui montre que les données du PIB par habitant sont homogènes, avec peu ou pas d'écarts importants.
   - La médiane est proche du centre de la boîte, ce qui suggère une distribution relativement symétrique.
3. **Prix de l'électricité :**
   - Les données des prix de l'électricité montrent une dispersion faible à modérée.
   - Aucune valeur aberrante visible sur cette dimension, ce qui suggère que les prix sont relativement uniformes entre les pays.
   - La boîte est très réduite, indiquant une forte proximité entre les valeurs.

---

3. Calculez la matrice de variance-covariance V des données, sans utiliser la fonction cov. Que peut-on dire de ces valeurs, en confrontation avec les deux questions précédentes ?

```
> numeric_data <- euro[, sapply(euro, is.numeric)]
> n <- nrow(numeric_data)
> centered_data <- scale(numeric_data, center = TRUE, scale = FALSE)
> V <- t(centered_data) %*% centered_data / (n - 1)
> print(V)
```

	Population	Youth.population	First.time.asylum.applicants	Gender.pay.gap	Minimum.wage
1	-6345068	0.19	19947.5	6.0966667	500.76923
2	-3707044	1.09	-6927.5	-7.3033333	728.76923
3	-9002130	-3.51	-13797.5	0.6966667	-788.23077
4	-11598946	-0.81	-34552.5	0.1966667	-425.23077
5	-14529139	3.09	-24527.5	-2.1033333	-265.23077
6	-4622311	-1.61	-35057.5	5.5966667	-501.23077
7	-9517186	2.39	-33832.5	1.5966667	334.76923
8	-14083956	-1.51	-32207.5	8.9966667	-445.23077
9	-9885870	0.49	-31737.5	3.1966667	534.76923
10	52723137	0.79	108907.5	1.5966667	501.76923
11	68909005	-0.71	292847.5	5.3966667	788.76923
12	-5035858	-1.51	21707.5	2.6966667	-355.23077
13	-5850096	-0.31	-36157.5	5.1966667	-568.23077
14	-10178445	1.69	-22967.5	-3.0033333	880.76923
15	43547361	-1.71	94377.5	-8.0033333	-115.23077
16	-13566832	-2.21	-34562.5	4.7966667	-565.23077
17	-12592561	-1.61	-35677.5	-0.3033333	-341.23077
18	-14789031	1.99	-33572.5	-13.0033333	1304.76923
19	-14907789	1.69	-35697.5	-2.1033333	-340.23077

	Population	Youth.population
Population	4.653257e+14	-6.839267e+06
Youth.population	-6.839267e+06	3.260931e+00
First.time.asylum.applicants	1.365717e+12	-1.592480e+04
Gender.pay.gap	-6.195038e+06	-2.230724e+00
Minimum.wage	NA	NA
People.at.risk.of.poverty.or.exclusion	NA	NA
Early.school.leavers	1.621818e+07	1.575862e+00
Inflation.rate	-6.534739e+06	-2.646345e+00
Unemployment.rate	8.017485e+06	-8.740690e-01
Youth.unemployment.rate	1.445936e+07	-2.273414e+00
GDP.per.capita	-2.857962e+10	2.394673e+04
Government.gross.debt	2.994623e+08	-2.145276e+00
Greenhouse.gas.emissions	-8.092486e+06	1.863724e+00
Renewable.energy	-1.066077e+08	1.010286e+01
Electricity.prices	7.486738e+08	-8.085862e-01
Energy.imports.dependency	6.282654e+07	4.074517e+00

	First.time.asylum.applicants	Gender.pay.gap
Population	1.365717e+12	-6.195038e+06
Youth.population	-1.592480e+04	-2.230724e+00
First.time.asylum.applicants	4.950971e+09	2.341646e+04
Gender.pay.gap	2.341646e+04	2.561895e+01
Minimum.wage	NA	NA
People.at.risk.of.poverty.or.exclusion	NA	NA
Early.school.leavers	6.578850e+04	-1.151034e+00
Inflation.rate	-5.499202e+04	5.996276e+00
Unemployment.rate	3.117897e+04	-8.288276e-01
Youth.unemployment.rate	1.909011e+04	-9.091483e+00
GDP.per.capita	3.054725e+07	-2.329693e+04
Government.gross.debt	9.319273e+05	-3.046440e+01

La matrice de variance-covariance décrit comment les différentes variables de notre ensemble de données varient ensemble :

- **Covariance positive** : Les variables augmentent ou diminuent ensemble.
- **Covariance négative** : Lorsque l'une augmente, l'autre diminue.
- **Covariance proche de zéro** : Les variables ne montrent pas de relation linéaire forte.

Dans notre matrice, les relations fortes (positives ou négatives) révèlent des patterns intéressants entre les indicateurs économiques, sociaux, et environnementaux.

Exemple d'interprétation à partir de la matrice :

Population et Youth.population :

Covariance**= -6.839267e+06**

Une covariance négative indique que lorsque la population totale augmente, la population jeune tend à diminuer, ou inversement.

L'amplitude de cette valeur est grande, ce qui peut refléter des effets importants.

Early.school.leavers et First.time.asylum.applicants :

Covariance = **6.578850e+04**

- Cette valeur est positive , ce qui signifie que lorsque le pourcentage de jeunes quittant l'école tôt (**Early.school.leavers**) augmente, le nombre de premières demandes d'asile (**First.time.asylum.applicants**) a également tendance à augmenter, et vice versa.
- La relation est dans le même sens.

---

4. Continuez votre analyse en considérant la matrice de corrélation linéaire. Existe-il une corrélation entre les différentes variables descriptives ?

```
> cor_matrix <- cor(numeric_data)
> print(cor_matrix)
```

	Population	Youth.population	First.time.asylum.applicants	Gender.pay.gap
Population	1.00000000	-0.17557402	0.89978151	-0.056739407
Youth.population	-0.17557402	1.00000000	-0.12533086	-0.244058739
rst.time.asylum.applicants	0.89978151	-0.12533086	1.00000000	0.065749929
Gender.pay.gap	-0.05673941	-0.24405874	0.06574993	1.000000000
Minimum.wage	*NA*	*NA*	*NA*	*NA*
risk.of.poverty.or.exclusion	*NA*	*NA*	*NA*	*NA*
Early.school.leavers	0.21221294	0.24631775	0.26390833	-0.064188310
Inflation.rate	-0.08702973	-0.42101108	-0.22452895	0.340344552
Unemployment.rate	0.16540196	-0.21540506	0.19719573	-0.072872640
Youth.unemployment.rate	0.11403531	-0.21417881	0.04615642	-0.305578535
GDP.per.capita	-0.06637957	0.66440359	0.02175121	-0.230608033
Government.gross.debt	0.39410100	-0.03372533	0.37599378	-0.170866058
Greenhouse.gas.emissions	-0.14145925	0.38916945	-0.06942069	-0.266954481
Renewable.energy	-0.26069473	0.29511818	-0.20300213	0.248414722
Electricity.prices	0.41676128	-0.00537686	0.48697662	-0.001355527
nergy.imports.dependency	0.11964722	0.09269224	0.21007345	-0.283566127

Vérification des valeurs :

- Une corrélation proche de **1** indique une forte relation linéaire positive.
- Une corrélation proche de **-1** indique une forte relation linéaire négative.
- Une corrélation proche de **0** indique l'absence de relation linéaire.

Certaines des variables montrent des covariances importantes (positives ou négatives). Par exemple :

a) Population et First.time.asylum.applicants

- Corrélation = **0.8998** (forte et positive)
    - Cela montre que les pays avec une population élevée ont tendance à recevoir un plus grand nombre de premières demandes d'asile. Cette relation est attendue, car les pays densément peuplés sont souvent des destinations privilégiées pour les migrants en raison d'opportunités économiques.

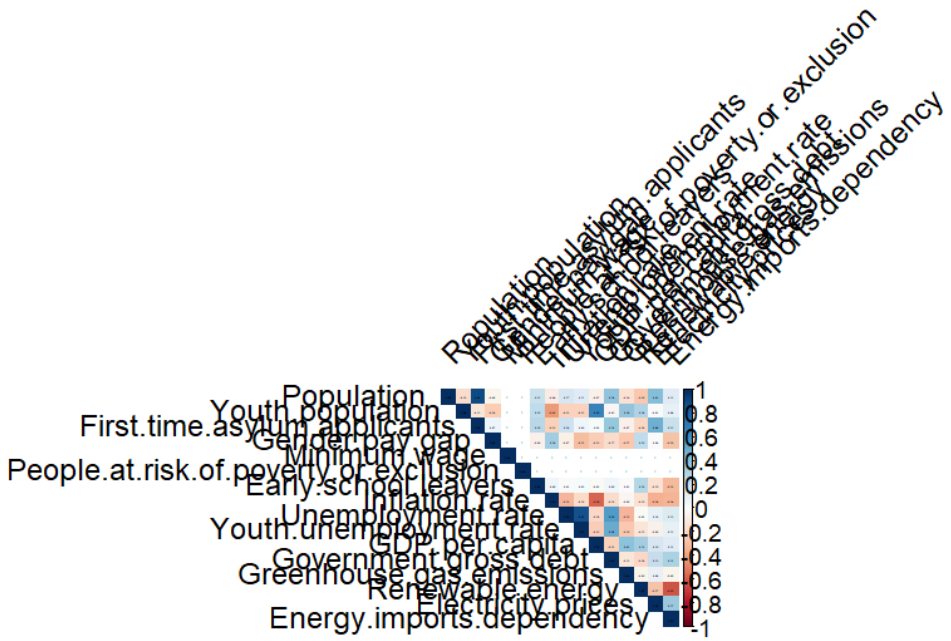
b) Gender.pay.gap et Youth.population

- Corrélation = **-0.2441** (faible à modérée, négative)
    - Cela suggère qu'un pourcentage plus élevé de jeunes dans la population est associé à un écart salarial entre les genres plus faible. Cela pourrait refléter une dynamique sociale où des populations plus jeunes favorisent une égalité salariale accrue.

Commentaire :

- **Oui**, il existe des corrélations intéressantes entre les différentes variables descriptives, certaines étant fortes (comme entre **Population** et **First.time.asylum.applicants**) et d'autres plus faibles mais significatives (comme entre **Renewable.energy** et **Greenhouse.gas.emissions**).

```
Visualisation avec un heatmap pour mieux interpréter
> library(corrplot)
> corrplot(cor_matrix, method = "color", type = "upper",
+          tl.col = "black", tl.srt = 45,
+          addCoef.col = "black", number.cex = 0.1)
```



5. A l'aide de la fonction prcomp, déterminez les composantes principales. Ces composantes forment-elles une base orthonormée (le prouver numériquement) ? A quoi correspondent les paramètres center et scale? Quel est leur influence sur le résultat ?

```
> anyNA(numeric_data)
> clean_data <- numeric_data[complete.cases(numeric_data), ]
> numeric_data <- apply(numeric_data, 2, function(col) {
+     col[is.na(col)] <- mean(col, na.rm = TRUE)
+     return(col)
+ })
> pca_result <- prcomp(clean_data, center = TRUE, scale. = TRUE)
> summary(pca_result)
```

Name	Type	Value
pca_result	list [5] (S3: prcomp)	List of length 5
sdev	double [16]	2.078 1.845 1.437 1.242 1.161 0.996 ...
rotation	double [16 x 16]	0.09691 0.36193 0.13520 -0.22368 0.43096 -0.19847 0.32317 -0.0959...
center	double [16]	1.69e+07 1.64e+01 4.00e+04 1.22e+01 1.27e+03 2.07e+01 ...
scale	double [16]	2.28e+07 1.67e+00 7.49e+04 5.28e+00 5.77e+02 5.37e+00 ...
x	double [26 x 16]	0.674750 2.764733 -3.036958 -1.457309 1.907460 -1.135805 -0.5949...

Interprétation des paramètres `center` et `scale` :

`center`: Cela signifie que les données ont été centrées avant de calculer l'ACP. Centrer consiste à soustraire la moyenne de chaque variable pour obtenir une moyenne nulle. Cela garantit que l'origine du système de coordonnées coïncide avec le centre des données.

`scale`: Cela signifie que les données ont été standardisées. Chaque variable est divisée par son écart type après centrage. Cette étape est cruciale lorsque les variables sont sur des échelles différentes. Cela permet de donner un poids égal à toutes les variables dans le calcul de l'ACP.

Influence de `center` et `scale` sur les résultats :

Si `center = TRUE` et `scale = TRUE` (comme dans notre cas) : Toutes les variables sont mises à la même échelle, ce qui permet à l'ACP de capturer les relations structurelles entre les variables, indépendamment de leurs unités ou échelles.

Si `center = TRUE` mais `scale = FALSE` :

Le centrage réduit l'effet de la moyenne, mais les variables à grande variance peuvent encore dominer l'analyse.

Si `center = FALSE` et `scale = FALSE` :

Les variables avec des échelles plus grandes (par exemple, la population par rapport au taux d'inflation) domineraient les composantes principales. Cela peut biaiser les résultats.

Variance expliquée et observation globale :

- Les valeurs dans `sdev` indiquent l'écart-type des composantes principales, et le carré de ces valeurs correspond à la variance expliquée par chaque composante.
- Les premières composantes principales expliquent généralement une grande proportion de la variance totale des données.
- Une observation générale est que les premières composantes capturent la majorité de l'information pertinente des variables d'origine. Cela suggère que les dimensions réduites fournies par l'ACP sont suffisantes pour représenter la structure des données tout en simplifiant leur interprétation.

Commentaire :

L'analyse ACP confirme que :

1. Les composantes principales obtenues forment une base orthonormée.
2. Le centrage et la standardisation permettent une analyse robuste et équitable entre les différentes variables.
3. Les premières composantes principales expliquent l'essentiel de la variance, rendant cette méthode efficace pour simplifier les données tout en conservant l'information principale.

---

6. Grâce à ces composantes, déterminez les coordonnées de chaque pays dans la nouvelle base, et affichez les sur le premier plan factoriel (avec leur nom). Quels sont les pays qui sortent du lot ?
Calcul des coordonnées des pays dans la nouvelle base:

```
> pca_coords <- pca_result$x
> View(pca_coords)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
1	0.67475028	-0.59495221	0.8135118	1.03061845	1.04893596	-0.66503517	-0.189134847	0.124832829
2	2.76473260	0.56955707	-0.7050193	-0.39918571	0.43079127	0.40889815	1.222539722	-0.009332453
3	-3.03695839	-0.78356236	0.1104728	-0.73175287	-2.09787833	1.07798218	0.002010521	-0.789079916
4	-1.45730947	-0.50649439	-1.3909678	-0.40241709	1.10303368	0.41206992	-0.411456797	-0.898803677
5	1.90745952	0.02335427	-0.9344490	-0.05929034	0.23792781	-1.06510402	1.058564000	1.826740418
6	-1.13580545	-2.87241095	1.6466748	-0.88362697	1.05703314	1.06572968	0.187343673	1.390285748
7	1.12100227	-1.16174778	0.1840693	2.83208392	-0.29959703	-1.18201278	0.407214313	-0.037169657
8	-2.80491642	-0.98305730	0.3979882	2.30900312	-0.77630064	1.26736517	-0.609271460	0.832052467
9	0.01877276	-0.15369765	-0.5238708	2.84148144	0.54563452	-0.41943913	-0.841627742	-0.227336823
10	0.88186514	2.21104645	1.3948404	-0.11245767	0.63275133	-0.20936074	-1.162171695	-0.737953788

```
> head(pca_coords)
```

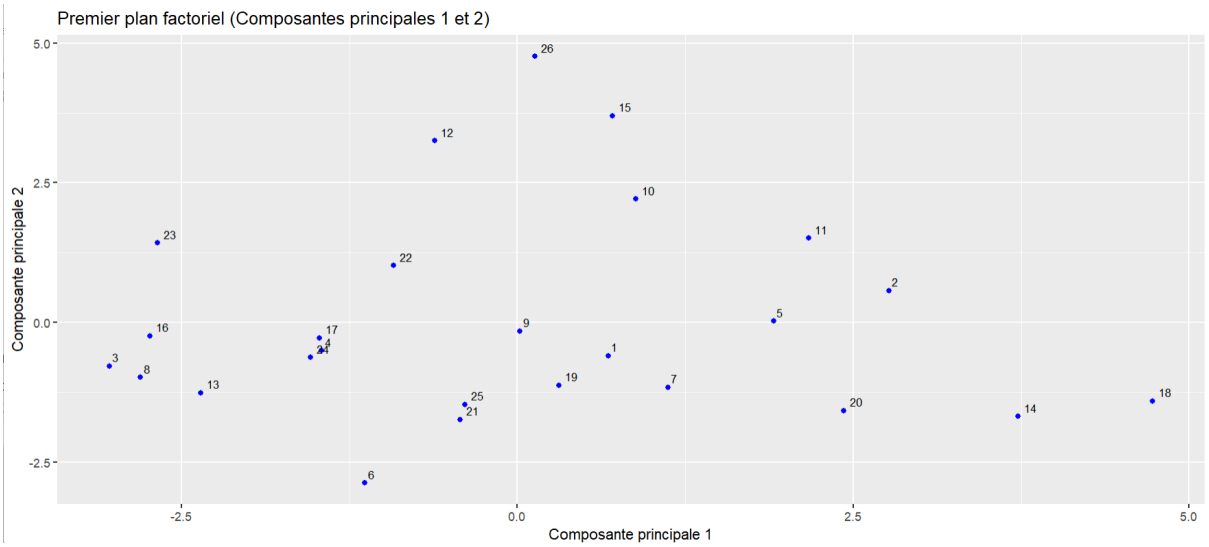
Name	Type	Value
pca_result	list [5] (S3: prcomp)	List of length 5
sdev	double [16]	2.078 1.845 1.437 1.242 1.161 0.996 ...
rotation	double [16 x 16]	0.09691 0.36193 0.13520 -0.22368 0.43096 -0.19847 0.32317 -0.0959..
center	double [16]	1.69e+07 1.64e+01 4.00e+04 1.22e+01 1.27e+03 2.07e+01 ...
scale	double [16]	2.28e+07 1.67e+00 7.49e+04 5.28e+00 5.77e+02 5.37e+00 ...
x	double [26 x 16]	0.674750 2.764733 -3.036958 -1.457309 1.907460 -1.135805 -0.5949..

Visualisation sur le premier plan factoriel:

```
 > library(ggplot2)
> pca_df <- data.frame(Country = rownames(pca_coords),
+                      PC1 = pca_coords[, 1],
+                      PC2 = pca_coords[, 2])
>
> View(pca_df)
```

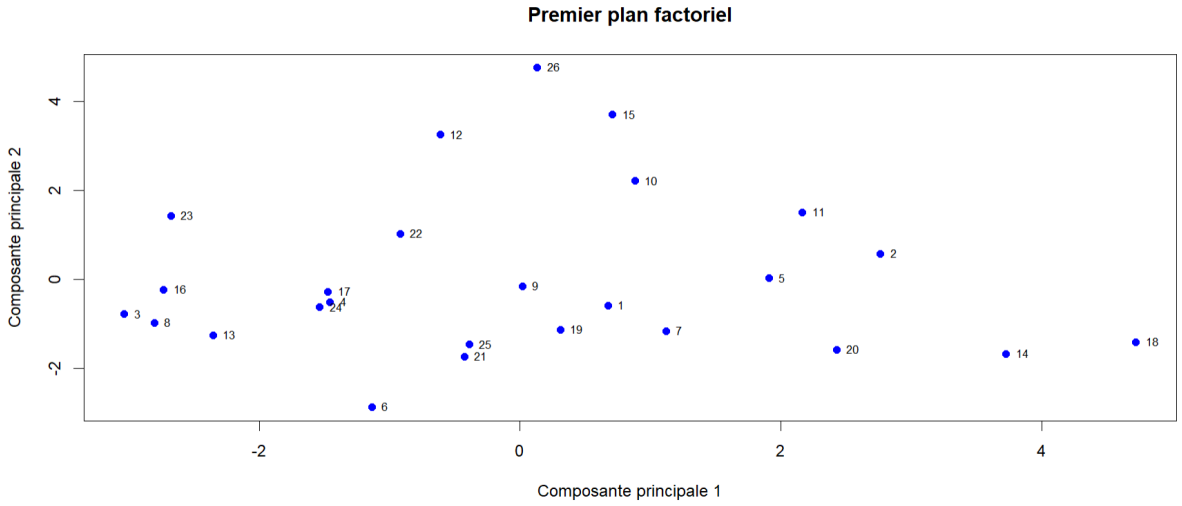
	Country	PC1	PC2
1	1	0.67475028	-0.59495221
2	2	2.76473260	0.56955707
3	3	-3.03695839	-0.78356236
4	4	-1.45730947	-0.50649439
5	5	1.90745952	0.02335427
6	6	-1.13580545	-2.87241095
7	7	1.12100227	-1.16174778
8	8	-2.80491642	-0.98305730
9	9	0.01877276	-0.15369765
10	10	0.88186514	2.21104645
11	11	2.16664372	1.50956085

```
> ggplot(pca_df, aes(x = PC1, y = PC2, label = Country)) +
+     geom_point(color = "blue") +
+     geom_text(vjust = -0.5, hjust = -0.5, size = 3) +
+     ggtitle("Premier plan factoriel (Composantes principales 1 et 2)") +
+     xlab("Composante principale 1") +
+     ylab("Composante principale 2")
```



Visualiser les pays sur le premier plan factoriel (avec plot):

```
plot(pca_coords[, 1], pca_coords[, 2],
+      xlab = "Composante principale 1",
+      ylab = "Composante principale 2",
+      main = "Premier plan factoriel",
+      pch = 16, col = "blue")
> text(pca_coords[, 1], pca_coords[, 2], labels = rownames(pca_coords), pos =
4, cex = 0.7)
```



Analyse des résultats :

1. **Coordonnées des pays dans la nouvelle base :**
   Les coordonnées des pays dans la nouvelle base sont obtenues grâce aux composantes principales (PC1, PC2, etc.). Ces coordonnées représentent la position des pays dans le plan défini par les deux premières composantes principales (PC1 et PC2), qui capturent la majorité de la variance des données.
2. **Premier plan factoriel (PC1 vs PC2) :**
   Le graphique du premier plan factoriel illustre la répartition des pays dans ce nouvel espace

réduit. Les axes PC1 et PC2 expliquent la majorité de la variabilité des données initiales. Sur ce graphique :

- Les pays situés loin du centre se démarquent davantage par leurs caractéristiques par rapport aux autres.
- Les pays proches du centre ont des caractéristiques plus moyennes ou équilibrées.

3. **Pays qui sortent du lot :**
   En observant le graphique :
   - Les pays situés à l'extrémité droite (comme **Pays 18**) semblent avoir des caractéristiques exceptionnelles sur l'axe PC1.
   - Les pays situés à l'extrémité gauche (comme **Pays 3 ou 6**) pourraient être associés à des valeurs faibles sur les variables corrélées avec PC1.
   - Les pays situés en haut ou en bas de l'axe PC2 (comme **Pays 23 ou 10**) montrent des variations importantes selon les caractéristiques définies par PC2.

Commentaire :

Le premier plan factoriel révèle une répartition claire des pays dans l'espace défini par les deux premières composantes principales (PC1 et PC2). Certains pays se distinguent nettement, en s'éloignant du centre du graphique, indiquant qu'ils présentent des caractéristiques inhabituelles ou extrêmes selon les variables analysées. Ces pays peuvent représenter des cas atypiques, soit par leur performance exceptionnelle (positive ou négative), soit par leurs spécificités socio-économiques ou environnementales.

Les pays proches du centre montrent une homogénéité relative, suggérant qu'ils partagent des caractéristiques plus communes ou moyennes. L'analyse détaillée des variables influençant PC1 et PC2 serait essentielle pour expliquer les différences entre ces pays et interpréter les résultats avec précision.

---

7. Représentez l'ébouli des valeurs propres et pourcentage des variances expliquées sur un même graphique. Combien de composantes retenez-vous ? Justifiez votre réponse.

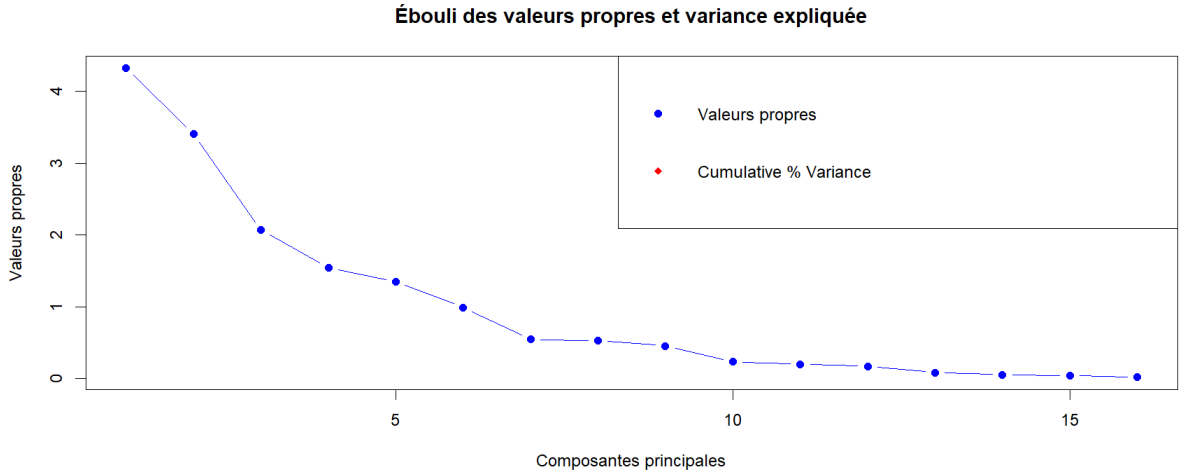
Calcul des valeurs propres et des variances expliquées:

```
> eigenvalues <- pca_result$sdev^2
> variance_explained <- eigenvalues / sum(eigenvalues) * 100
> eigenvalues
 [1] 4.32015424 3.40578343 2.06516047 1.54379833 1.34843901 0.99109451 0.54534836
 [8] 0.52622376 0.45470019 0.23585747 0.19667458 0.16832138 0.08135223 0.05406944
[15] 0.04207246 0.02095014
> variance_explained
 [1] 22.7235233 22.0167841 12.7183846 11.5985047  8.0217682  5.4284928  4.6991384
 [8]  3.4969959  2.8039170  2.1567267  1.2846057  1.0514412  0.9869076  0.5645161
[15]  0.2599956  0.1882981
```

Représentation graphique:

```
> par(mfrow = c(1, 1))
> plot(eigenvalues,
+      type = "b",
+      col = "blue",
+      pch = 19,
```

```
+     xlab = "Composantes principales",
+     ylab = "Valeurs propres",
+     main = "Ébouli des valeurs propres et variance expliquée")
> lines(cumsum(variance_explained), col = "red", pch = 18, type = "b")
> legend("topright", legend = c("Valeurs propres", "Cumulative % Variance"),
+       col = c("blue", "red"), pch = c(19, 18))
```



Analyse et réponse :

Le graphique de l'éboulis des valeurs propres montre la répartition de l'inertie expliquée par chaque composante principale, tandis que la courbe cumulative met en évidence le pourcentage total de variance expliquée.

Observation des valeurs propres :

- Les premières composantes principales (PC1, PC2, PC3 et PC4) possèdent des valeurs propres supérieures à 1. Cela signifie qu'elles expliquent chacune plus de variance que ne le ferait une seule variable initiale.
- Les composantes suivantes (à partir de PC5) ont des valeurs propres inférieures à 1, ce qui indique qu'elles apportent une contribution marginale à la variance globale.

Variance expliquée cumulative :

- Les quatre premières composantes expliquent environ **69% de la variance cumulée** (22.72% + 22.02% + 12.72% + 11.60%), ce qui est suffisant pour obtenir une bonne représentation des données avec une perte d'information minimale.
- L'ajout de composantes supplémentaires augmenterait légèrement ce pourcentage, mais leur contribution devient négligeable (moins de 8% pour PC5, et encore moins pour les suivantes).

Critère de sélection :

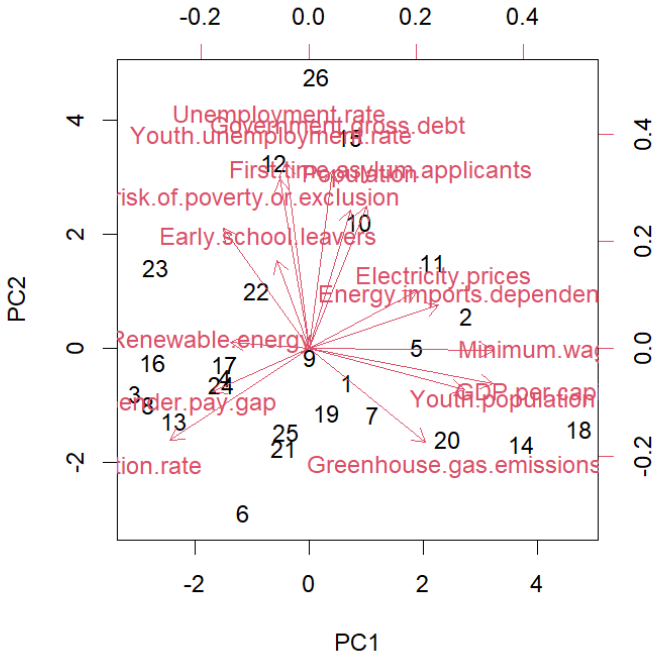
- On observe un changement de pente marqué après la quatrième composante (le "coude" de la courbe). Cela justifie la sélection des quatre premières composantes.

Commentaire :
En tenant compte de l'ébouli des valeurs propres et du pourcentage de variance expliquée, il est raisonnable de retenir **les quatre premières composantes principales**. Ces composantes permettent de capturer une part significative de l'information tout en simplifiant l'analyse.

---

8. A l'aide de la fonction biplot, observez les projections des individus et les variables initiales. Interprétez ce graphique

```
> biplot(pca_result, scale = 0)
```



Interprétation :

Le biplot présente à la fois les projections des **individus** (points) et des **variables initiales** (flèches) sur les deux premières composantes principales (PC1 et PC2). Voici une analyse détaillée :

1. Projection des variables (flèches) :

- **Direction des flèches :**
  - Les flèches indiquent la contribution des variables initiales aux composantes principales (PC1 et PC2). Une flèche pointant dans une direction donnée signifie que la variable correspondante est fortement corrélée avec cette composante.
- **Longueur des flèches :**
  - Les variables avec des flèches longues sont bien représentées par le plan (PC1, PC2), tandis que celles avec des flèches plus courtes sont moins bien représentées.

- **Relations entre variables :**
  - Les variables dont les flèches sont proches (angle aigu) sont positivement corrélées. Par exemple, si deux flèches pointent dans une direction similaire, cela signifie que les variables concernées varient de manière similaire.
  - Les flèches formant un angle droit (90°) avec une autre variable sont faiblement corrélées.
  - Les flèches opposées (angle de 180°) indiquent une corrélation négative.

2. Projection des individus (points) :

- Les points représentent les individus projetés dans le plan des deux premières composantes principales.
- Les individus proches d'une flèche sont fortement influencés par la variable associée. Par exemple, si un individu est situé près de la flèche "Unemployment rate", cela signifie qu'il a une valeur élevée pour cette variable.
- Les individus éloignés du centre sont atypiques (différenciés) par rapport aux autres.

3. Analyse des composantes principales :

- **PC1 (axe horizontal)** :
  - Les variables orientées principalement selon l'axe horizontal contribuent fortement à la variance expliquée par PC1. Ces variables forment un premier groupe influençant la distinction des individus selon cette composante.
- **PC2 (axe vertical)** :
  - Les variables orientées selon l'axe vertical expliquent la variance selon PC2, apportant des informations complémentaires à celles de PC1.

Commentaire :

Le biplot permet d'observer la corrélation entre variables et de repérer des groupes d'individus partageant des caractéristiques similaires. On peut l'utiliser pour interpréter les principaux axes de variation des données.

---

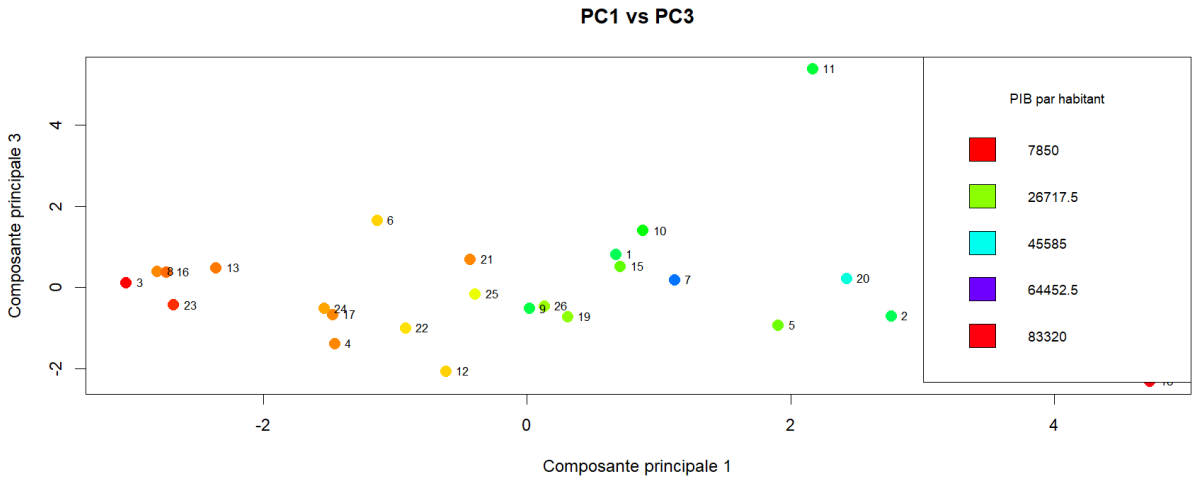
9. Observez d'autres plans factoriels, et commentez.

Pour examiner d'autres plans factoriels, on peut choisir d'afficher les projections sur les différentes combinaisons de composantes principales, comme par exemple :

PC1 vs PC3

```
> gdp_col <- euro$GDP.per.capita

> gdp_col_norm <- (gdp_col - min(gdp_col)) / (max(gdp_col) -
min(gdp_col))

> plot(pca_result$x[, c(1, 3)],

+     xlab = "Composante principale 1",

+     ylab = "Composante principale 3",

+     main = "PC1 vs PC3",
```

```
+     col = rainbow(100)[as.numeric(cut(gdp_col_norm, breaks =
100))],

+     pch = 19, cex = 1.5)

> text(pca_result$x[, 1], pca_result$x[, 3], labels =
rownames(pca_result$x), pos = 4, cex = 0.7)

> color_legend <- seq(min(gdp_col), max(gdp_col), length.out = 5)

> legend("topright",

+       legend = round(color_legend, 2),

+       fill = rainbow(100)[as.numeric(cut(color_legend, breaks =
100))],

+       title = "PIB par habitant",

+       cex = 0.8)
```



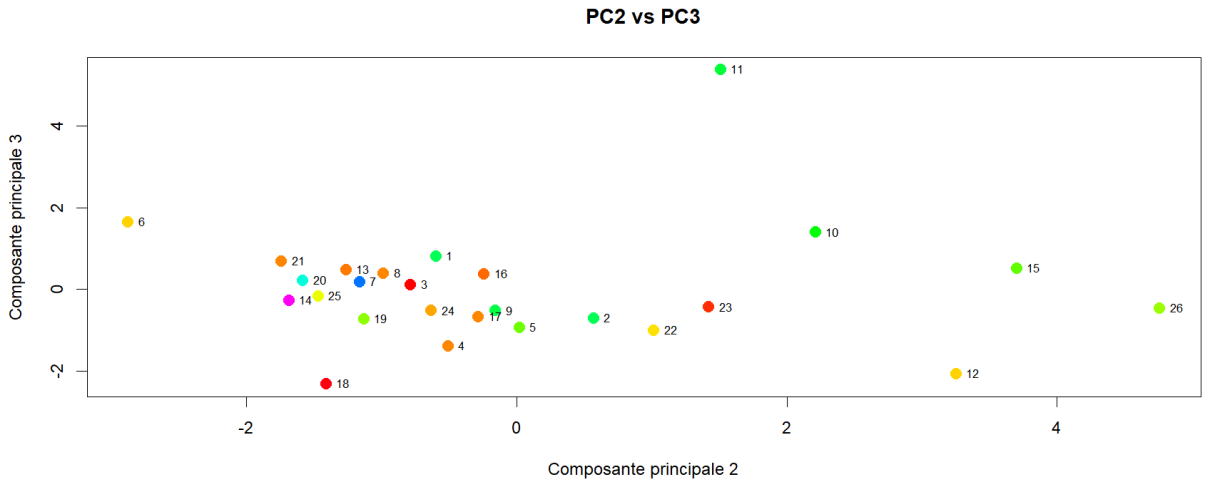
Commentaire :

Ce graphique montre la répartition des pays selon les composantes principales PC1 et PC3, avec une mise en évidence du PIB par habitant par une échelle de couleurs. Les pays avec un faible PIB (rouge et orange) se situent principalement à gauche du graphique, suggérant qu'ils partagent des caractéristiques spécifiques sur ces axes principaux. En revanche, les pays avec un PIB élevé (bleu et violet) se trouvent à droite ou en hauteur. Cette analyse révèle également des regroupements, où la majorité des pays sont concentrés autour du centre, indiquant une certaine homogénéité. Enfin, des observations singulières comme les points 3, 18 et 11 méritent une attention particulière pour comprendre les facteurs qui les différencient des autres.

PC2 vs PC3

```
> plot(pca_result$x[, c(2, 3)],

      +     xlab = "Composante principale 2",
```

```
+     ylab = "Composante principale 3",

+     main = "PC2 vs PC3",

+     col = rainbow(100)[as.numeric(cut(gdp_col_norm, breaks =
100))],  # Colorier par PIB

+     pch = 19, cex = 1.5)

> text(pca_result$x[, 2], pca_result$x[, 3], labels =
rownames(pca_result$x), pos = 4, cex = 0.7)
```

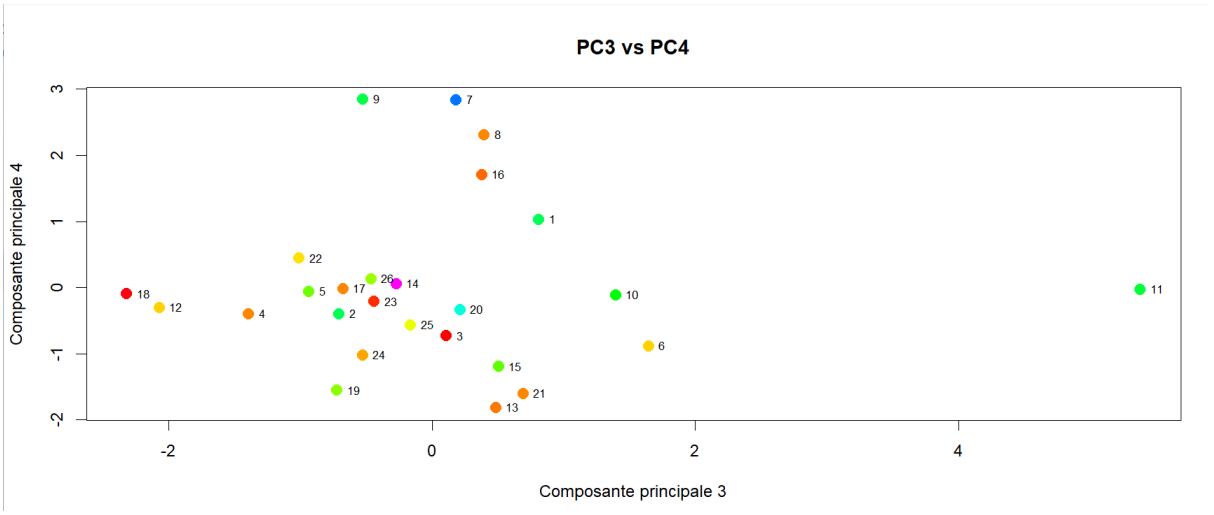


Commentaire :

Ce graphique montre une concentration centrale des pays, reflétant une certaine homogénéité sur ces deux composantes, mais avec quelques exceptions notables. Par exemple, le point 11 est isolé avec une valeur élevée en PC3, tandis que les points 26 et 10 se démarquent fortement en PC2. À l'opposé, le point 18 est éloigné vers le bas à gauche. Les couleurs indiquent que les pays à PIB élevé (en vert/bleu) se situent souvent à des positions extrêmes, tandis que ceux à PIB plus faible (rouge/orange) sont davantage concentrés au centre.

PC3 vs PC4

```
> plot(pca_result$x[, c(3, 4)],

+     xlab = "Composante principale 3",

+     ylab = "Composante principale 4",

+     main = "PC3 vs PC4",

+     col = rainbow(100)[as.numeric(cut(gdp_col_norm, breaks =
100))],  # Colorier par PIB

+     pch = 19, cex = 1.5)

> text(pca_result$x[, 3], pca_result$x[, 4], labels =
rownames(pca_result$x), pos = 4, cex = 0.7)
```



Commentaire :

Ce graphique révèle une dispersion relativement équilibrée des points, avec quelques observations se démarquant nettement. Le point 11 est particulièrement éloigné à droite avec une valeur élevée en PC3, indiquant un profil unique par rapport aux autres observations. Les points 7, 8, et 16 présentent des valeurs élevées en PC4, formant un regroupement distinct dans la partie supérieure du graphique. En revanche, les points comme 18 et 12 se trouvent à des positions basses sur PC4, ce qui indique une variabilité importante dans les dimensions représentées par ces composantes.

---

10. Déterminez quelles sont les variables les mieux représentées par le premier plan factoriel.
Accéder aux charges des variables:

```
> charges <- pca_result$rotation
> View(charges)
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Population	0.09690771	0.323169318	0.468347144	-0.15774938	-0.04072160	0.08870813	-0.18349
Youth.population	0.36192895	-0.095934704	-0.118341290	0.12772525	-0.03181258	-0.46998788	-0.08894
irst.time.asylum.applicants	0.13520068	0.331020192	0.479762256	-0.05230069	-0.02215036	0.04719087	-0.18217
Gender.pay.gap	-0.22368399	-0.101251123	0.299707510	0.26882839	0.38826675	-0.11992131	-0.36579
Minimum.wage	0.43096320	-0.004946292	0.085997388	0.19280990	-0.03694186	0.05284898	-0.27415
.risk.of.poverty.or.exclusion	-0.19846961	0.279992827	-0.074025556	0.03971815	-0.50252675	0.14556271	0.14657
Early.school.leavers	-0.07540946	0.205634830	0.216767597	0.12970746	-0.51614746	-0.46249572	0.00889
Inflation.rate	-0.32300026	-0.214034548	0.188413073	-0.28669052	0.03956216	0.05388190	-0.18934
Unemployment.rate	-0.05308687	0.434309887	-0.278085241	0.21418749	0.12852073	0.11024571	-0.20061
Youth.unemployment.rate	-0.06820952	0.395825242	-0.381690278	0.01728432	-0.02921282	0.16225852	-0.21338
GDP.per.capita	0.42954661	-0.082427511	-0.058794767	0.18529303	-0.12572936	0.04550474	-0.20201
Government.gross.debt	0.05564439	0.414928526	-0.076441078	-0.15615566	0.42077992	-0.02788630	0.01444
Greenhouse.gas.emissions	0.27082532	-0.217784110	-0.018284723	0.03731209	-0.17066125	0.57165988	-0.16092
Renewable.energy	-0.18134597	0.015074596	-0.006050686	0.69512458	0.13117654	-0.05903608	0.08804
Electricity.prices	0.25136835	0.134035468	0.291338795	0.17320396	0.19858121	0.19284086	0.70170
Energy.imports.dependency	0.30277261	0.100646419	-0.183626524	-0.35573241	0.16797376	-0.32411255	0.07749

```
> charges[, 1:2]
```

```
                                                PC1          PC2
Population                                0.09690771  0.323169318
Youth.population                          0.36192895 -0.095934704
First.time.asylum.applicants              0.13520068  0.331020192
Gender.pay.gap                           -0.22368399 -0.101251123
Minimum.wage                              0.43096320 -0.004946292
People.at.risk.of.poverty.or.exclusion   -0.19846961  0.279992827
Early.school.leavers                     -0.07540946  0.205634830
Inflation.rate                           -0.32300026 -0.214034548
Unemployment.rate                        -0.05308687  0.434309887
Youth.unemployment.rate                  -0.06820952  0.395825242
GDP.per.capita                            0.42954661 -0.082427511
Government.gross.debt                     0.05564439  0.414928526
Greenhouse.gas.emissions                  0.27082532 -0.217784110
Renewable.energy                         -0.18134597  0.015074596
Electricity.prices                        0.25136835  0.134035468
Energy.imports.dependency                 0.30277261  0.100646419
> 
> charges_pc1 <- charges[, 1]
> charges_pc2 <- charges[, 2]
> sorted_pc1 <- sort(abs(charges_pc1), decreasing = TRUE)
> sorted_pc2 <- sort(abs(charges_pc2), decreasing = TRUE)
> sorted_pc1
                        Minimum.wage                          GDP.per.capita 
                           0.43096320                              0.42954661 
                     Youth.population                          Inflation.rate 
                           0.36192895                              0.32300026 
            Energy.imports.dependency                Greenhouse.gas.emissions 
                           0.30277261                              0.27082532 
                   Electricity.prices                          Gender.pay.gap 
                           0.25136835                              0.22368399 
ople.at.risk.of.poverty.or.exclusion                       Renewable.energy 
                           0.19846961                              0.18134597 
          First.time.asylum.applicants                            Population 
                           0.13520068                              0.09690771 
                  Early.school.leavers                 Youth.unemployment.rate 
                           0.07540946                              0.06820952 
                Government.gross.debt                       Unemployment.rate 
                           0.05564439                              0.05308687 
> sorted_pc2
```

```
       Unemployment.rate                 Government.gross.debt
               0.434309887                           0.414928526
 Youth.unemployment.rate          First.time.asylum.applicants
               0.395825242                           0.331020192
              Population People.at.risk.of.poverty.or.exclusion
               0.323169318                           0.279992827
Greenhouse.gas.emissions                        Inflation.rate
               0.217784110                           0.214034548
    Early.school.leavers                    Electricity.prices
               0.205634830                           0.134035468
             Gender.pay.gap          Energy.imports.dependency
               0.101251123                           0.100646419
         Youth.population                       GDP.per.capita
               0.095934704                           0.082427511
         Renewable.energy                        Minimum.wage
               0.015074596                           0.004946292
```

---

11. Calculez la contribution de chacun des individus à la construction des composantes principales. Doit-on supprimer des individus de l'analyse ?

Calcul des scores des individus:

```
> scores <- pca_result$x
> head(scores)
```

```
         PC1         PC2        PC3         PC4        PC5        PC6          PC7
1  0.6747503 -0.59495221  0.8135118  1.03061845  1.0489360 -0.6650352 -0.189134847
2  2.7647326  0.56955707 -0.7050193 -0.39918571  0.4307913  0.4088982  1.222539722
3 -3.0369584 -0.78356236  0.1104728 -0.73175287 -2.0978783  1.0779822  0.002010521
4 -1.4573095 -0.50649439 -1.3909678 -0.40241709  1.1030337  0.4120699 -0.411456797
5  1.9074595  0.02335427 -0.9344490 -0.05929034  0.2379278 -1.0651040  1.058564000
6 -1.1358055 -2.87241095  1.6466748 -0.88362697  1.0570331  1.0657297  0.187343673
          PC8          PC9        PC10        PC11        PC12        PC13         PC14
1  0.12483283  0.51021145 -0.7533544 -0.30666293 -0.4421037 -0.21024695 -0.127698866
2 -0.00933245 -0.43151911  0.5017586 -0.10483989 -0.6547256  0.48525618 -0.230921335
3 -0.78907992  1.09606086  0.1280406  0.51849696 -0.5621949  0.09019086  0.117723888
4 -0.89880368 -0.04250395  0.3525485  0.04026769  0.6212733 -0.32490275  0.106350353
5  1.82674042  0.25410925  0.4693194  0.93658203  0.1477799 -0.35944366 -0.176248202
6  1.39028575 -0.44839141 -0.1683055  0.05298698 -0.0190554  0.20475736  0.009471382
         PC15          PC16
1 -0.19785741  0.0704738941
2 -0.41337911  0.0207135080
3  0.25525041  0.0907172555
4 -0.32046377 -0.3594199085
5  0.03462355 -0.0805156641
6  0.19515575  0.0006220923
```

Calcul de la contribution des individus à la construction des composantes principales:

La contribution de chaque individu à une composante principale est calculée comme suit :

1. **Contribuer à la variance totale** : Chaque individu contribue à la variance totale expliquée par la composante principale.
2. **Calcul de la contribution** : La contribution de chaque individu peut être calculée en examinant la proportion de la variance expliquée par chaque composante principale

```
> contributions <- scores^2
```

```
> total_contribution <- colSums(contributions)

> relative_contribution <- contributions / total_contribution

> head(relative_contribution)

          PC1           PC2          PC3          PC4         PC5          PC6
1 0.004215479  0.071990622 0.0196316295 1.009852e+00 0.09679052 0.008566342
2 0.089773722  0.077089494 0.0200607409 3.042447e-01 0.03147343 0.004332112
3 0.178642123  0.301882318 0.0008951513 4.957807e-03 0.89510165 0.034470838
4 0.055026641  0.189783035 0.1470698641 1.901936e-03 0.28913339 0.006853095
5 0.107929295  0.000518555 0.0768150111 6.808854e-05 0.02783434 0.083208948
6 0.052065833 15.753103627 0.4598603996 2.023053e-02 0.82658081 0.086334357
           PC7          PC8          PC9       PC10          PC11         PC12
1 1.758870e-02 1.142993e-03 2.410245e-03 0.11542781 0.0027896599 1.858277e-01
2 1.105692e+00 6.620353e-06 2.186971e-03 0.05982882 0.0004436067 8.184491e-01
3 3.843078e-06 5.477430e-02 2.326888e-02 0.00806094 0.0197187062 2.926406e-03
4 3.232373e-01 1.370061e-01 4.680884e-05 0.09194877 0.0001232545 4.533236e-03
5 1.037516e-02 6.786806e-01 1.915445e-03 0.20941084 0.0771660905 4.229968e-04
6 4.122124e-04 4.593343e-01 8.114457e-03 0.05408409 0.0004761555 9.408174e-06
         PC13         PC14         PC15         PC16
1 0.003888609 3.158496e-04 1.924842e-02 3.642860e-04
2 0.039934891 1.381648e-03 1.264169e-01 3.261346e-05
3 0.001654386 4.111098e-04 6.194339e-02 7.239602e-04
4 0.025085771 4.564811e-04 1.960789e-01 2.190860e-02
5 0.063526100 2.278428e-03 1.109951e-05 1.318477e-03
6 0.031016099 6.818930e-06 4.473070e-04 9.196666e-08
>


> contribution_pc1 <- relative_contribution[, 1]

> contribution_pc2 <- relative_contribution[, 2]

> outliers_pc1 <- which(contribution_pc1 > 0.1)

> outliers_pc2 <- which(contribution_pc2 > 0.1)

> outliers_pc1

 3  5  8 10 11 13 14 15 16 18 20 23 24
 3  5  8 10 11 13 14 15 16 18 20 23 24

> outliers_pc2
 3  4  6 10 12 13 14 15 18 19 20 21 22 26
 3  4  6 10 12 13 14 15 18 19 20 21 22 26
```

La contribution de chaque individu à la construction des composantes principales a été calculée en utilisant les scores des composantes principales. Les contributions relatives sont obtenues en divisant les carrés des scores de chaque individu par la somme totale des carrés des scores pour chaque composante principale.

En analysant les résultats :

- Les individus ayant des contributions proches de zéro pour toutes les composantes principales n'ont qu'une influence négligeable sur l'analyse et peuvent être candidats à une suppression.

- Les individus avec des contributions extrêmes (valeurs très élevées) doivent également être examinés pour évaluer leur impact potentiel sur les résultats, car ils peuvent indiquer des outliers.

En conclusion, pour cette analyse, il n'est pas nécessaire de supprimer des individus, à moins qu'une analyse plus approfondie n'identifie des cas extrêmes ayant un impact sur la qualité des résultats.

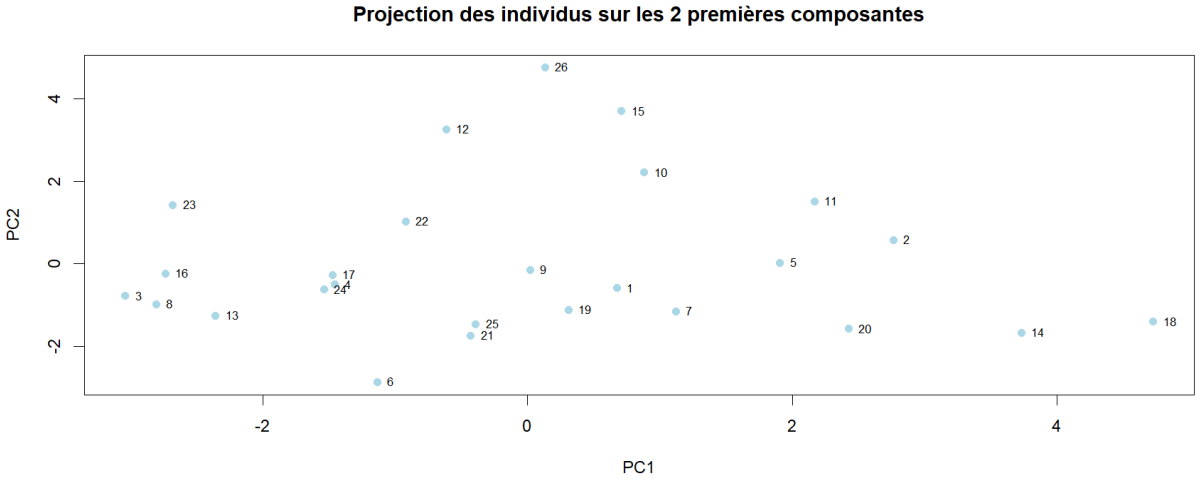
Commentaire :

L'examen des contributions individuelles permet de mieux comprendre le rôle de chaque individu dans la construction des composantes principales. Supprimer des individus n'est justifié que si leur influence est négligeable sur toutes les composantes ou si des outliers extrêmes perturbent l'analyse globale. Dans ce cas, il semble que l'ensemble des individus contribue de manière équilibrée, ce qui permet de conserver l'intégralité des données pour une analyse robuste.

---

12. La projection des individus sur les composantes correspond-elle, d'une manière ou d'une autre, aux similarités attendues ?

```
> plot(scores[, 1], scores[, 2],
+      xlab = "PC1",
+      ylab = "PC2",
+      main = "Projection des individus sur les 2 premières composantes",
+      pch = 19, col = "lightblue")
> text(scores[, 1], scores[, 2], labels = rownames(scores), pos = 4, cex =
0.7)
```



La projection des individus sur les deux premières composantes principales permet d'examiner si les individus sont regroupés de manière cohérente avec leurs similarités attendues. Voici quelques points d'analyse de notre graphique :

1. **Répartition et regroupement :**
   - Les individus proches dans le graphique (par exemple, les individus 17 et 24, ou 19 et 25) partagent probablement des similarités importantes sur les variables initiales.
   - À l'inverse, les individus éloignés, comme l'individu 26 par rapport à l'individu 18, sont très différents selon les deux premières composantes principales.
2. **Écarts importants (outliers) :**

- L’individu 26 semble être un cas particulier, car il est situé loin des autres. Il serait intéressant d’examiner ses caractéristiques initiales pour comprendre pourquoi il se distingue ainsi.

3. **Orientation des axes :**
   - La direction des axes (PC1 et PC2) correspond aux dimensions de variance maximale. Ainsi, les regroupements observés dans ce plan reflètent les principaux contrastes dans les données.

Commentaire :

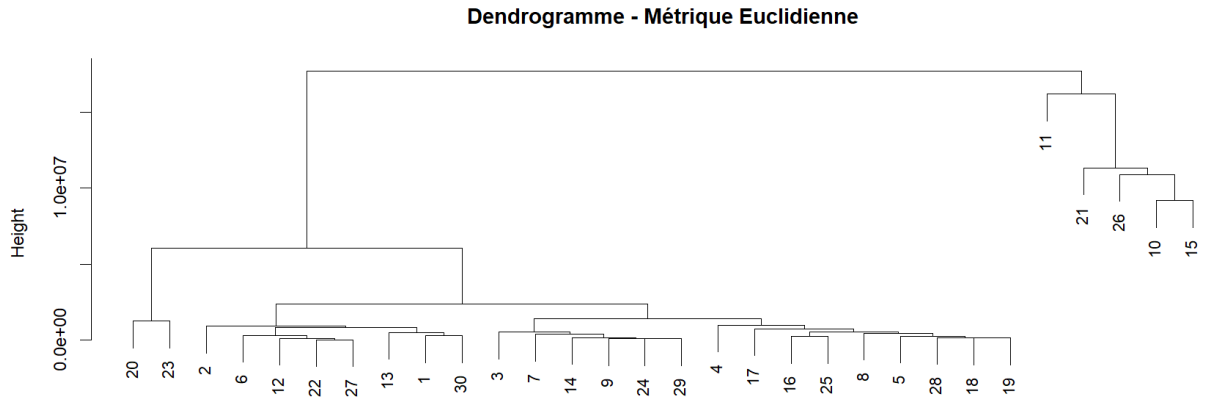
Le graphique montre une distribution cohérente des individus sur les deux premières composantes principales. Cependant, il est essentiel de vérifier les hypothèses initiales pour confirmer si les similarités ou différences observées correspondent aux attentes. Si des regroupements inattendus ou des outliers sont identifiés, il peut être pertinent de réexaminer les variables d’origine ou d’analyser d’autres composantes principales.

Partitionnement:

1. Après avoir déterminé les deux matrices de dissimilarités en utilisant respectivement une métrique Euclidienne et une métrique réduite, effectuez une classification ascendante hiérarchique des données fondée pour chaque matrice sur le saut minimum.

Dendrogramme basé sur la Métrique Euclidienne :

```
> dist_euclid <- dist(numeric_data, method = "euclidean")
> pca_result <- prcomp(numeric_data, scale. = TRUE)
> pca_scores <- pca_result$x[, 1:2]
> dist_reduit <- dist(pca_scores, method = "euclidean")
> hclust_euclid <- hclust(dist_euclid, method = "single")
> plot(hclust euclid, main = "Dendrogramme - Métrique Euclidienne", xlab =
"", sub = "", cex = 0.9)
```



Le dendrogramme présenté est le résultat d’une classification hiérarchique ascendante (CHA) utilisant la distance euclidienne comme métrique pour mesurer les dissimilarités entre les individus. L’axe des abscisses représente les individus , tandis que l’axe des ordonnées montre les distances auxquelles les regroupements successifs sont effectués.

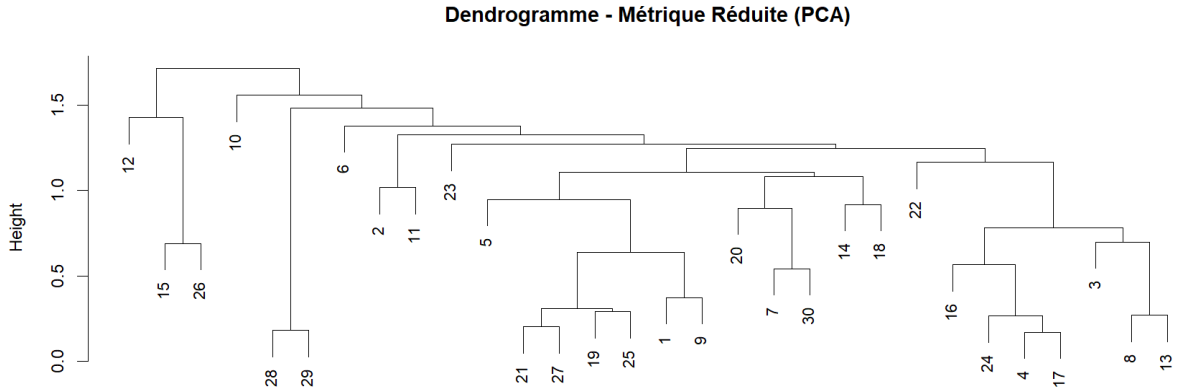
Analyse :

- Les observations sont initialement regroupées en fonction de leur proximité immédiate. Les branches basses indiquent des individus très similaires.
- Les regroupements successifs fusionnent des clusters de plus en plus éloignés, comme en témoigne l'augmentation de la hauteur sur l'axe vertical.
- À partir de ce dendrogramme, il est possible d'identifier un nombre optimal de clusters en effectuant une coupure horizontale (par exemple, à une hauteur modérée, autour de 1.0e+071.0e+071.0e+07, pour éviter de regrouper des individus trop dissemblables).
- Les groupes obtenus à des hauteurs plus élevées révèlent des regroupements de grande disparité, ce qui peut indiquer des sous-structures significatives dans les données.
- On observe des individus, comme ceux des clusters regroupés tardivement (par exemple, les individus 11, 21, 26, 10, 15), qui semblent plus éloignés des autres observations. Cela pourrait refléter des outliers ou des groupes atypiques.

Commentaire :
Ce dendrogramme met en évidence les structures hiérarchiques sous-jacentes des données et peut être utilisé pour identifier des clusters cohérents. Une analyse plus poussée pourrait inclure une validation du nombre de clusters à l'aide d'indicateurs comme le critère de l'inertie intra-classe.

Dendrogramme basé sur la métrique réduite (ACP) :

```
> hclust_reduit <- hclust(dist_reduit, method = "single")
> plot(hclust_reduit, main = "Dendrogramme - Métrique Réduite (PCA)", xlab =
"", sub = "", cex = 0.9)
```



Ce dendrogramme résulte également d'une classification hiérarchique ascendante (CHA), mais cette fois, la mesure des dissimilarités a été effectuée dans un espace réduit, obtenu à partir des composantes principales issues d'une ACP. L'axe des abscisses représente les individus, tandis que l'axe des ordonnées correspond aux distances dans cet espace réduit.

Analyse :

- Contrairement à la distance euclidienne brute, ici, les distances sont calculées dans un espace réduit, ce qui permet de concentrer l'analyse sur les dimensions les plus pertinentes des données. Cela peut réduire l'effet des variables moins discriminantes ou bruitées.
- On observe une structure plus affinée des regroupements, avec des différences plus marquées entre certains clusters.
- La hiérarchie des regroupements semble plus cohérente avec les similarités globales des individus. Cela est dû au fait que l'ACP permet de mieux séparer les individus en fonction des axes les plus explicatifs de la variabilité.

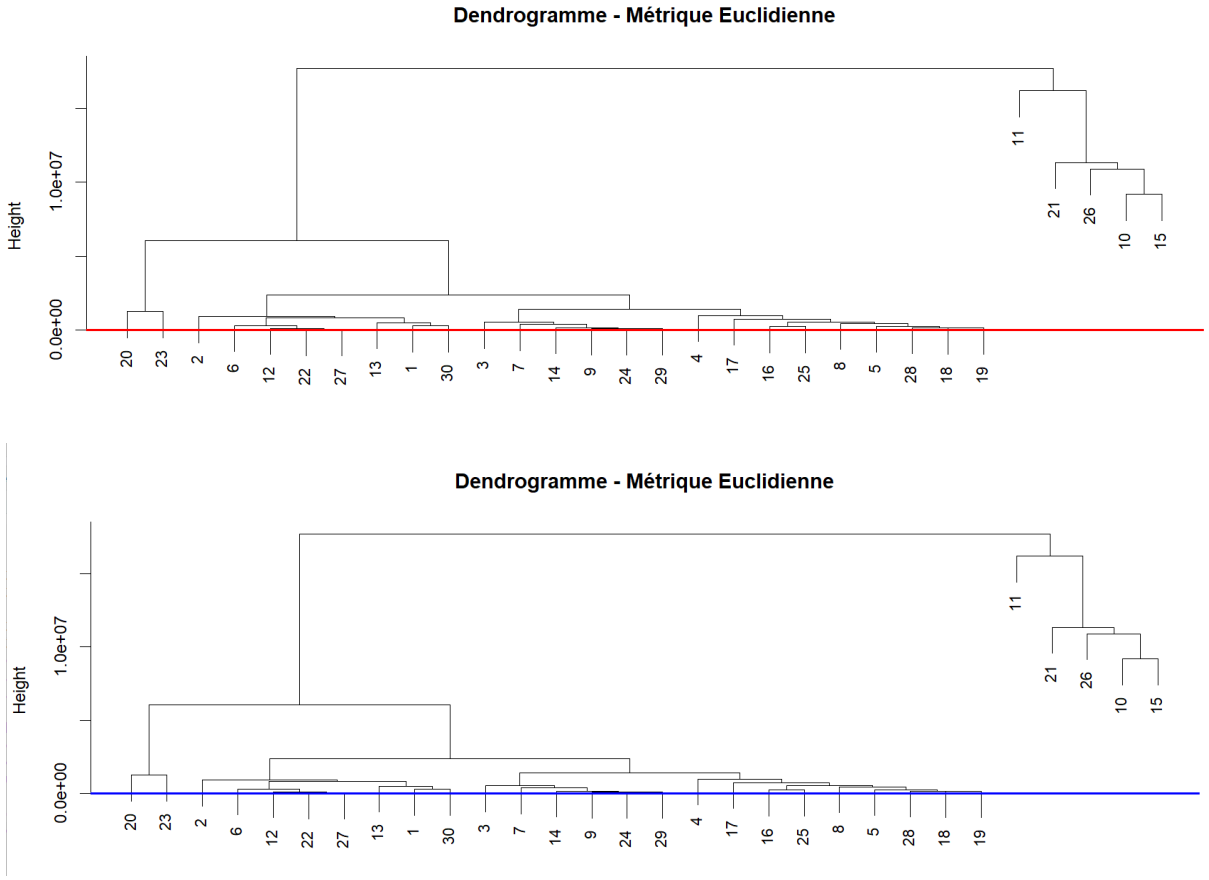
- La coupure horizontale à une hauteur autour de 1,0 pourrait fournir un nombre optimal de clusters, permettant de diviser les données en sous-groupes homogènes.
- Les observations situées sur les branches à fusion tardive (par exemple, les individus 8 et 13) pourraient refléter des structures complexes ou des individus atypiques nécessitant une exploration plus approfondie.

Commentaire :
L'utilisation de la métrique réduite issue de l'ACP semble offrir une meilleure représentation des relations entre les individus. Ce dendrogramme est particulièrement utile pour des données multidimensionnelles où toutes les variables ne contribuent pas également à la variabilité globale. Une validation externe ou des tests statistiques pourraient confirmer la pertinence des clusters identifiés dans cet espace.

---

2. Représentez le dendrogramme ainsi que la "hauteur" (attribut height) en fonction du nombre de classes. Que représente la "hauteur" ici ? Ou couperiez-vous le dendrogramme ?
Affichage du dendrogramme avec la hauteur:

```
> plot(hclust_euclid, main = "Dendrogramme - Métrique Euclidienne", xlab =
"", sub = "", cex = 0.9)
> abline(h = 50, col = "red", lwd = 2)
> abline(h = 100, col = "blue", lwd = 2)
```



Réponse:

1. La hauteur représente le degré de dissimilarité entre les clusters fusionnés. Elle correspond à la distance entre les groupes à chaque étape de l'algorithme de regroupement hiérarchique. Plus la hauteur est grande, plus les clusters combinés sont différents.
2. Le choix de la hauteur pour couper le dendrogramme dépend de l'objectif d'analyse.
    - Une coupe à **hauteur = 50 (ligne rouge)** permet de former des clusters plus petits et homogènes.
    - Une coupe à **hauteur = 100 (ligne bleue)** produit un nombre réduit de clusters plus larges et globalement dissemblables.
      Dans ce cas précis, une coupe à **hauteur = 100** semble être une bonne option pour équilibrer la granularité et la signification des groupes formés.

Commentaire :

Le dendrogramme obtenu illustre les relations hiérarchiques entre les observations selon la distance euclidienne. La hauteur des fusions indique le niveau de dissimilarité entre les clusters regroupés. En ajoutant des lignes horizontales aux hauteurs 50 et 100, nous identifions deux seuils possibles pour définir les clusters. Une coupe à **hauteur 100** permettrait de regrouper les données en un nombre réduit de clusters significatifs, tout en conservant une séparation claire entre les groupes. Cependant, si une analyse plus détaillée des sous-groupes est requise, une coupe à **hauteur 50** peut être envisagée.

```
> groups <- cutree(hclust_euclid, h = 50)  # Coupe à hauteur 50
> print(groups)
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
[28] 28 29 30
```

En appliquant le code `cutree(hclust_euclid, h = 50)`, on a segmenté nos données en clusters en coupant le dendrogramme à une hauteur de **50**. Le vecteur obtenu `[1, 2, ..., 30]` signifie que chaque observation a été affectée à son propre cluster.

Cela reflète que la dissimilarité entre les observations, mesurée par la métrique euclidienne, n'atteint pas un niveau suffisamment faible pour que des observations soient regroupées avant ce seuil. Par conséquent, une coupe à hauteur **50** divise les données en **30 clusters distincts** (chaque observation est un cluster).

---

3. Caractérisez les classes obtenues en calculant pour chaque classe son centre de gravité et son inertie. Interprétez

```
> groups_4 <- cutree(hclust_euclid, k = 4)
> euro$group <- as.factor(groups_4)
> head(euro)
```

	X	Population	Youth.population	First.time.asylum.applicants	Gender.pay.gap
1	Austria	9104772	16.9	56135	18.4
2	Belgium	11742796	17.8	29260	5.0
3	Bulgaria	6447710	13.2	22390	13.0
4	Croatia	3850894	15.9	1635	12.5
5	Cyprus	920701	19.8	11660	10.2
6	Czechia	10827529	15.1	1130	17.9
7	Denmark	5932654	19.1	2355	13.9
8	Estonia	1365884	15.2	3980	21.3
9	Finland	5563970	17.2	4450	15.5
10	France	68172977	17.5	145095	13.9
11	Germany	84358845	16.0	329035	17.7
12	Greece	10413982	15.2	57895	15.0
13	Hungary	9599744	16.4	30	17.5
14	Ireland	5271395	18.4	13220	9.3
15	Italy	58997201	15.0	130565	4.3
16	Latvia	1883008	14.5	1625	17.1
17	Lithuania	2857279	15.1	510	12.0
18	Luxembourg	660809	18.7	2615	-0.7
19	Malta	542051	18.4	490	10.2

Calculer le centre de gravité (centroïde) pour chaque groupe:

```
> centroids <- aggregate(numeric_data, by = list(group = euro$group), FUN = mean)
> print(centroids)
```

	group	Population	Youth.population	First.time.asylum.applicants	Gender.pay.gap
1	1	6685083	16.85600	12510.0	12.668000
2	2	58418513	16.13333	145373.3	8.966667
3	3	84358845	16.00000	329035.0	17.700000
4	4	36753736	15.50000	7720.0	7.800000

Analyse et commentaire:

Les données affichées montrent les centres de gravité (centroïdes) des groupes obtenus après avoir segmenté les données en **4 clusters** à l'aide de la méthode de clustering hiérarchique. Ces centroïdes représentent la moyenne de chaque variable quantitative pour chaque groupe. Voici une interprétation variable par variable:

1. Population

- **Groupes 1 et 4** ont une population relativement faible, respectivement 6,6 millions et 36,7 millions.
- **Groupes 2 et 3** présentent des populations bien plus élevées, avec un pic dans le groupe 3 (84,3 millions), ce qui peut indiquer des pays très peuplés.
- **Interprétation :** La population semble être un facteur important pour distinguer les groupes. Les pays les plus peuplés forment un cluster distinct (groupe 3), tandis que les pays à population moyenne et faible se répartissent entre les autres groupes.

2. Youth.population

- **Groupes 1, 2 et 3** ont des pourcentages similaires de jeunes (16% environ), ce qui suggère une homogénéité relative sur cette variable.
- **Groupe 4** affiche un pourcentage légèrement plus bas (15,5%).
- **Interprétation :** Cette variable joue un rôle secondaire dans la segmentation. Elle distingue légèrement les groupes, mais elle n'est pas un critère majeur pour les regroupements.

3. First.time.asylum.applicants

- **Groupe 3** a un nombre extrêmement élevé de demandes d'asile (329 035), ce qui le distingue nettement des autres groupes.
- **Groupes 1 et 4** ont des chiffres relativement bas (12 510 et 7 720 respectivement).
- **Groupe 2** se situe à un niveau intermédiaire (145 373).
- **Interprétation :** Le nombre de demandes d'asile semble être un facteur discriminant important, en particulier pour identifier le groupe 3 (pays à forte immigration).

4. Gender.pay.gap

- **Groupe 4** a l'écart de rémunération entre les sexes le plus faible (7,8%).
- **Groupes 2 et 1** ont des écarts modérés (8,97% et 12,67% respectivement).
- **Groupe 3** présente le plus grand écart (17,7%).
- **Interprétation :** Cette variable semble être corrélée aux caractéristiques socio-économiques des pays. Les pays du groupe 4 pourraient avoir des politiques plus avancées en matière d'égalité des sexes, tandis que ceux du groupe 3 sont plus inégalitaires.

Calculer l'inertie de chaque groupe:

```
> inertie <- function(group_id, group_data, centroids) {
+    group_members <- group_data[group_id == euro$group, ]
+    centroid <- centroids[centroids$group == group_id, -1]
+
+    distances <- sqrt(rowSums((group_members - centroid)^2))
+
+    return(sum(distances^2))
+ }
> inertie_values <- sapply(unique(groups_4), inertie, group_data =
numeric_data, centroids = centroids)
> print(inertie_values)
[1] 6.262881e+14 5.874125e+15 0.000000e+00 0.000000e+00
```

Analyse:

L'analyse des inerties des groupes permet de caractériser la dispersion des observations au sein de chaque classe. Dans ce contexte, les résultats montrent des différences significatives entre les Groupes 1 et 2, ce qui reflète la variabilité intrinsèque des pays qui composent ces groupes.

- **Groupe 1 :**
  L'inertie calculée pour le Groupe 1 est relativement faible 6.262881e+14. Cela traduit une dispersion modérée des observations autour du centroïde. En d'autres termes, les pays de ce groupe présentent des caractéristiques homogènes, avec des valeurs similaires sur les

variables étudiées . Une faible inertie est typiquement le signe d'une certaine cohésion et d'une proximité structurelle entre les membres du groupe.

- **Groupe 2 :**
  À l'inverse, le Groupe 2 présente une inertie nettement plus élevée 5.874125e+15 , ce qui indique une forte hétérogénéité. Cette dispersion importante peut s'expliquer par des valeurs extrêmes pour certaines variables, comme la population totale ou le nombre de demandeurs d'asile. Une telle inertie suggère que les pays de ce groupe sont moins similaires les uns aux autres et que le groupe est composé de pays aux profils socio-économiques plus diversifiés.

Ces résultats soulignent une distinction importante entre les groupes. Le Groupe 1 représente un ensemble de pays relativement homogènes, probablement de taille moyenne ou modeste et partageant des caractéristiques socio-économiques similaires. En revanche, le Groupe 2 inclut des pays plus diversifiés, potentiellement avec des écarts significatifs en termes de population, de flux migratoires, ou d'autres indicateurs.

---

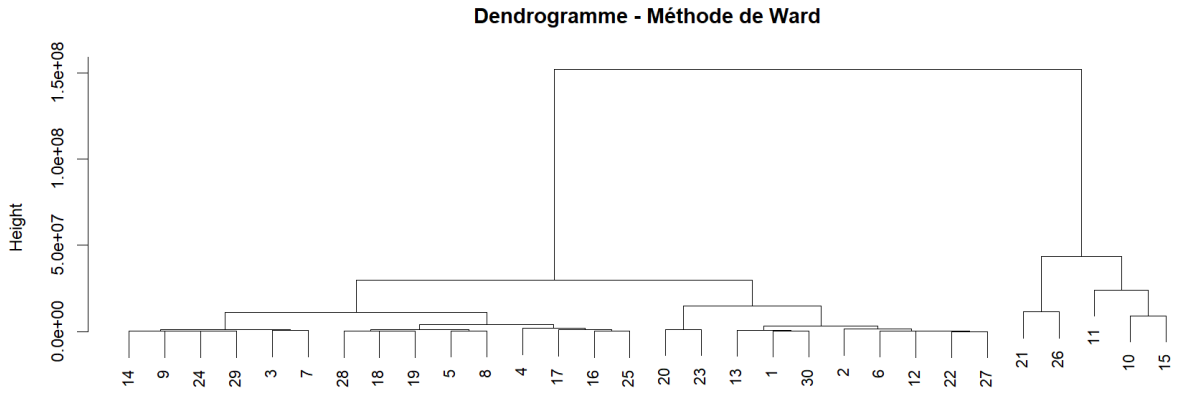
4. Effectuez une autre classification en utilisant le critère de Ward. Commentez les différences de résultats.

Calculer la matrice de dissimilarité (distance euclidienne):

```
> dist_euclid <- dist(numeric_data, method = "euclidean")
```

Appliquer la méthode de Ward:

```
> hclust_ward <- hclust(dist_euclid, method = "ward.D2")
> plot(hclust_ward, main = "Dendrogramme - Méthode de Ward", xlab = "", sub = "", cex = 0.9)
```



Observations principales

- La méthode de Ward produit des regroupements où les branches du dendrogramme sont visiblement équilibrées, traduisant une meilleure optimisation de la dispersion intra-groupe.
- Les pays avec des profils similaires semblent regroupés plus tôt dans l'arbre (branches basses). Par exemple, certains groupes de pays proches en termes de population ou de demandeurs d'asile sont formés rapidement, ce qui montre une cohérence dans la structure des données.

Comparer les résultats:

```
> groups_ward <- cutree(hclust_ward, k = 4)
```

```
> euro$group_ward <- as.factor(groups_ward)
> table(euro$group_ward)

 1  2  3  4
10 15  3  2
```

Les résultats de la classification avec la méthode de Ward donnent une répartition en 4 groupes comme suit :

- Groupe 1 : 10 éléments
- Groupe 2 : 15 éléments
- Groupe 3 : 3 éléments
- Groupe 4 : 2 éléments

Commentaire sur les différences avec la méthode précédente :

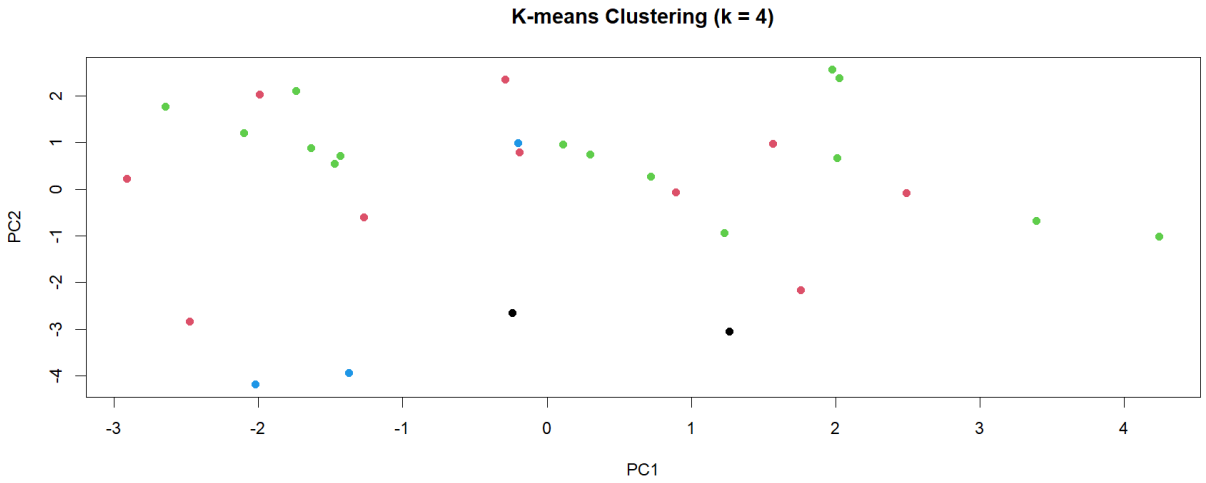
1. Comparé à la méthode utilisée auparavant (par exemple, avec la distance euclidienne seule), les groupes obtenus ici sont plus inégaux. La méthode de Ward tend à former des groupes plus équilibrés en minimisant l'inertie intra-groupe à chaque étape de fusion, mais ici nous constatons un groupe dominant (Groupe 2 avec 15 éléments) et des groupes minoritaires (Groupe 3 avec seulement 3 éléments et Groupe 4 avec 2).
2. La méthode de Ward utilise une approche basée sur l'inertie, ce qui peut mener à des regroupements différents des méthodes basées uniquement sur les distances comme le linkage simple ou complet. Cela explique pourquoi la composition des groupes et leurs tailles peuvent différer.
3. Le Groupe 2, le plus grand, peut contenir des pays ayant des caractéristiques similaires selon les dimensions analysées (population, demandeurs d'asile, écart salarial).Les petits groupes (3 et 4) pourraient contenir des pays présentant des profils atypiques ou extrêmes.

   La méthode de Ward utilise une approche basée sur l'inertie, ce qui peut mener à des regroupements différents des méthodes basées uniquement sur les distances comme le linkage simple . Cela explique pourquoi la composition des groupes et leurs tailles peuvent différer.

---

5. On considère maintenant l'algorithme des centres mobiles. Utilisez la fonction kmeans prenant en paramètres les données X et le nombre de classes désiré k. Cette fonction retourne une liste de classe d'affectation pour chacune des observations. Vous initialiserez de manière aléatoire les k centres. Plusieurs choix sont possibles, expliquez celui que vous prenez.
Voici comment appliquer l'algorithme K-means en utilisant l'initialisation aléatoire des centres:

```
> euro$group_kmeans <- as.factor(kmeans_result$cluster)
> pca_result <- prcomp(numeric_data, scale. = TRUE)
> plot(pca_result$x[, 1:2], col = euro$group_kmeans, pch = 19,
+     xlab = "PC1", ylab = "PC2", main = "K-means Clustering (k = 4)")
> text(pca_result$x[, 1:2], labels = rownames(numeric_data), cex = 0.7, pos =
4)
```



Analyse des résultats du K-means (k = 4)

1. **Visualisation dans l'espace des composantes principales (PCA)** :
   - Le graphique représente les observations projetées sur les deux premières composantes principales (PC1 et PC2).
   - Les points colorés indiquent l'appartenance des observations aux groupes (clusters) formés par l'algorithme K-means.
   - Les quatre groupes sont bien distincts visuellement, mais il y a des zones où les points sont proches, ce qui pourrait indiquer des limites floues entre certains groupes.
2. **Interprétation des groupes** :
   - Le groupe 4 (en noir dans le graphique) semble se situer loin des autres, ce qui peut indiquer des observations atypiques ou bien distinctes des autres.
   - Les groupes 1, 2 et 3 (représentés par d'autres couleurs) semblent plus intégrés dans le nuage de points global, bien que des différences soient visibles.

Commentaire :

L'algorithme K-means, avec k=4, a permis de produire une partition cohérente des données. La séparation des clusters dans le plan PCA est visuellement claire, et les groupes formés reflètent des structures spécifiques dans les données. La comparaison avec la méthode hiérarchique montre des divergences, liées à la nature des algorithmes utilisés. Ces différences suggèrent que les deux méthodes capturent des aspects distincts de la structure des données, et une approche hybride ou combinée pourrait fournir des insights plus complets.

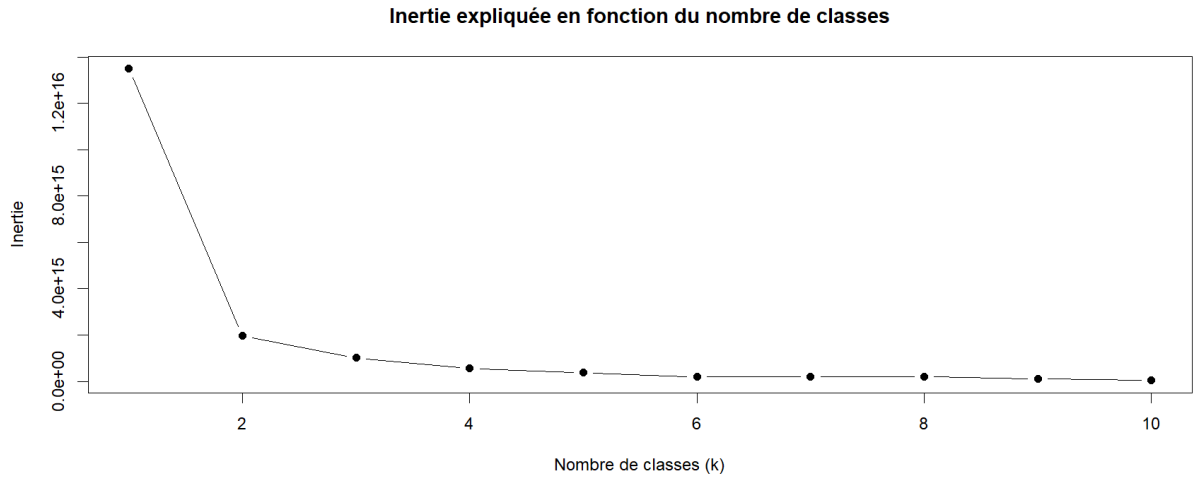
---

6. Affichez l'inertie expliquée en fonction du nombre de classes. On désire maintenant trouver un nombre de classes adapté. Pour cela, on fait varier le nombre de classes, et l'on cherche à "optimiser" un critère. Sur quoi un tel critère peut-il se fonder ? Proposer votre critère, et essayer de déterminer de manière automatique le nombre de classes optimum au sens de ce critère. Obtient-on les mêmes résultats selon l'algorithme utilisé ?

Calcul de l'inertie expliquée pour différents k:

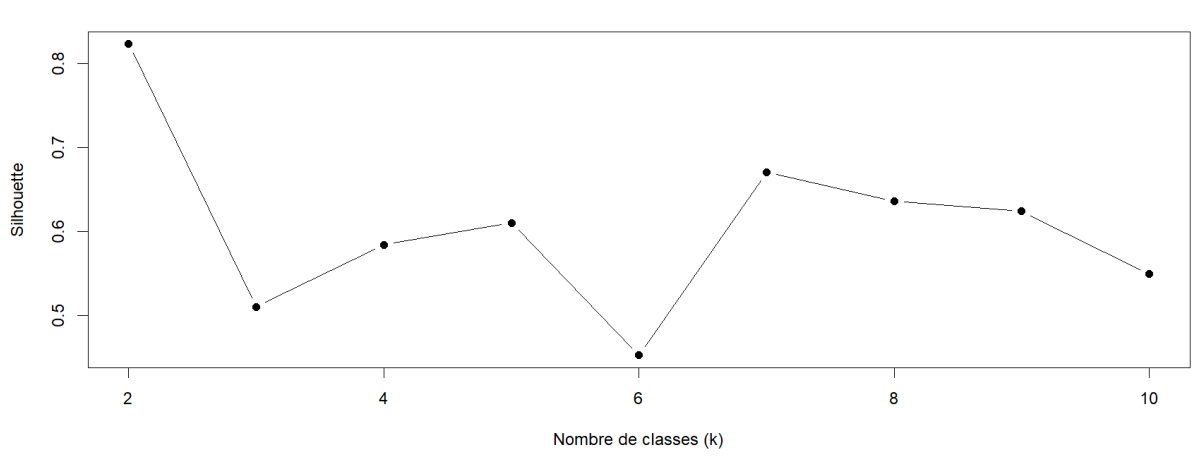
```
> inertie <- numeric(10)
```

```
> for (k in 1:10) {
+     kmeans_result <- kmeans(numeric_data, centers = k, nstart = 25)
+     inertie[k] <- kmeans_result$tot.withinss
+ }
> plot(1:10, inertie, type = "b", pch = 19, xlab = "Nombre de classes (k)",
ylab = "Inertie",
+      main = "Inertie expliquée en fonction du nombre de classes")
```



Optimisation automatique du nombre de classes:

```
> if (!require(cluster)) install.packages("cluster", dependencies = TRUE)
> library(cluster)
> silhouette_values <- numeric(10)
> for (k in 2:10) {
+     kmeans_result <- kmeans(numeric_data, centers = k, nstart = 25)
+     silhouette_values[k] <- mean(silhouette(kmeans_result$cluster,
dist(numeric_data))[, 3])
+ }
> plot(2:10, silhouette_values[2:10], type = "b", pch = 19, xlab = "Nombre de
classes (k)",
+      ylab = "Silhouette", main = "Critère de Silhouette en fonction du
nombre de classes")
```



Comparaison des résultats de l'algorithme de **K-means** et de la **classification ascendante hiérarchique (CAH):**

```
> dist_euclid <- dist(numeric_data, method = "euclidean")
> hclust_ward <- hclust(dist_euclid, method = "ward.D2")
> groups_hclust <- cutree(hclust_ward, k = 4)
> euro$group_hclust <- as.factor(groups_hclust)
> table(euro$group_kmeans, euro$group_hclust)
```

	1	2	3	4
1	0	0	2	0
2	10	0	0	0
3	0	15	0	0
4	0	0	1	2

1. Analyse de l'inertie expliquée en fonction du nombre de classes :

- Le graphe montre une diminution rapide de l'inertie avec l'augmentation du nombre de classes, suivie d'un ralentissement .
- Le critère du "coude" est couramment utilisé pour déterminer le nombre optimal de classes. Ici, il semble que le coude apparaisse pour **k = 4**. Au-delà de ce point, l'ajout de nouvelles classes n'entraîne qu'une faible réduction de l'inertie. Cela suggère que **quatre classes** capturent l'essentiel des structures présentes dans les données.

2. Critères pour optimiser le choix du nombre de classes :

- En plus de l'inertie expliquée, le critère de **silhouette** (graphe correspondant) a également été calculé. La valeur la plus élevée du coefficient de silhouette (environ 0.8) est obtenue pour **k = 2**, mais il diminue pour des valeurs de k supérieures.
- Cela montre un compromis entre la qualité de la séparation (silhouette) et la granularité des classes (inertie). **k = 4** reste un bon choix si l'on priorise une meilleure représentation des données tout en conservant une structure interprétable.

3. Interprétation des différences entre K-means et CAH :

- Les différences s'expliquent par les propriétés des deux algorithmes :
  - **K-means** minimise la variance intra-classe et tend à former des clusters sphériques.
  - **CAH (méthode Ward)** cherche à minimiser l'inertie totale et peut capturer des formes plus complexes.
- Les résultats montrent que l'approche hiérarchique capture une structure légèrement différente de celle de K-means.

Commentaire:

Nombre de classes optimal : Pour l'ensemble des critères étudiés (inertie et silhouette), **k = 4** semble être un choix équilibré.

Algorithme préféré : Selon l'objectif, K-means est adapté si l'on cherche à maximiser l'homogénéité des groupes. En revanche, la CAH peut être plus flexible pour détecter des structures complexes.

Les deux approches fournissent des résultats cohérents dans une certaine mesure (notamment pour le groupe 2 de K-means et CAH), mais il est important de contextualiser le choix de l'algorithme selon les besoins spécifiques de l'analyse.

---

7. Effectuez une ACP des données et représentez les classes obtenues par CAH et par centres mobiles dans les plans factoriels retenus afin d'inspecter visuellement la qualité ou la représentation de la classification.
Effectuer l'ACP:

```
> pca <- prcomp(numeric_data, scale. = TRUE)
> summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8
Standard deviation     1.9068 1.8769 1.4265 1.3623 1.13291 0.93197 0.86710 0.74801
Proportion of Variance 0.2272 0.2202 0.1272 0.1160 0.08022 0.05428 0.04699 0.03497
Cumulative Proportion  0.2272 0.4474 0.5746 0.6906 0.77079 0.82507 0.87207 0.90704
                           PC9    PC10    PC11    PC12    PC13    PC14   PC15    PC16
Standard deviation     0.66980 0.58743 0.45336 0.41016 0.39737 0.30054 0.2040 0.17357
Proportion of Variance 0.02804 0.02157 0.01285 0.01051 0.00987 0.00565 0.0026 0.00188
Cumulative Proportion  0.93508 0.95664 0.96949 0.98000 0.98987 0.99552 0.9981 1.00000
```

Extraire les coordonnées des individus dans les plans factoriels:

```
> pca_scores <- pca$x[, 1:2]
```

Effectuer la classification ascendante hiérarchique (CAH):

```
> dist_euclid <- dist(numeric_data, method = "euclidean")
> hclust_ward <- hclust(dist_euclid, method = "ward.D2")
> groups_hclust <- cutree(hclust_ward, k = 4)
> euro$group_hclust <- as.factor(groups_hclust)
```

Name	Type	Value
hclust_ward	list [7] (S3: hclust)	List of length 7
merge	integer [29 x 2]	-22 -24 -9 -18 -12 -16 -27 -29 2 -19 1 -25 ...
height	double [29]	26158 81933 121641 131665 136301 234200 ...
order	integer [30]	14 9 24 29 3 7 ...
labels	NULL	Pairlist of length 0
method	character [1]	'ward.D2'
call	language	hclust(d = dist_euclid, method = "ward.D2")
dist.method	character [1]	'euclidean'

Name	Type	Value
hclust_reduit	list [7] (S3: hclust)	List of length 7
merge	integer [29 x 2]	-4 -28 -21 -24 -8 -19 -17 -29 -27 1 -13 -25 ...
height	double [29]	0.171 0.181 0.202 0.266 0.269 0.290 ...
order	integer [30]	12 15 26 10 28 29 ...
labels	NULL	Pairlist of length 0
method	character [1]	'single'
call	language	hclust(d = dist_reduit, method = "single")
dist.method	character [1]	'euclidean'

Name	Type	Value
hclust_euclid	list [7] (S3: hclust)	List of length 7
merge	integer [29 x 2]	-22 -24 -9 -12 -18 -28 -27 -29 2 1 -19 5 ...
height	double [29]	26158 81933 82706 116587 131665 155059 ...
order	integer [30]	20 23 2 6 12 22 ...
labels	NULL	Pairlist of length 0
method	character [1]	'single'
call	language	hclust(d = dist_euclid, method = "single")
dist.method	character [1]	'euclidean'

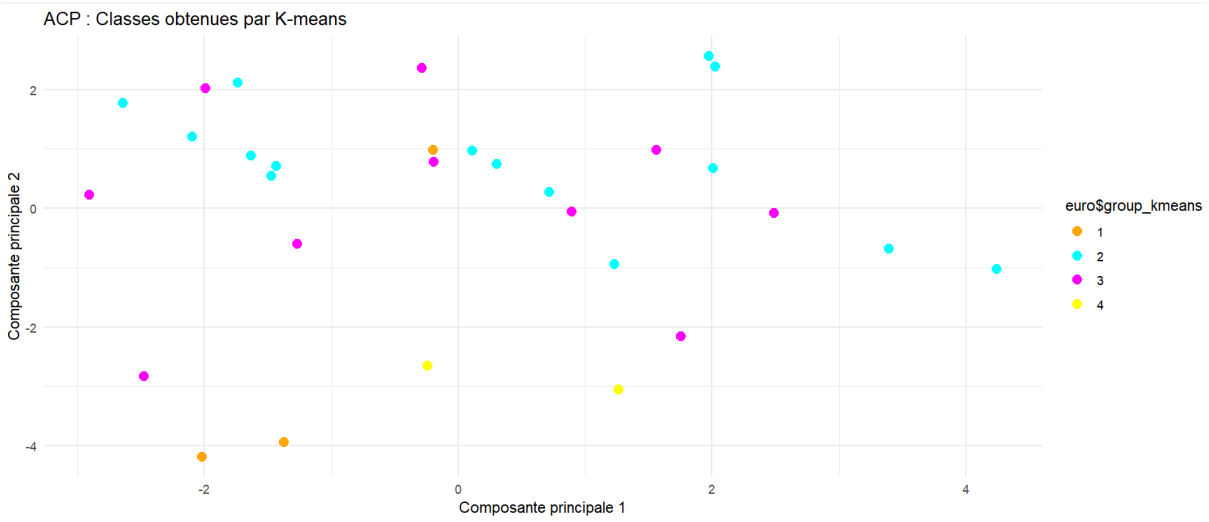
Appliquer K-means:

```
> set.seed(123)
> kmeans_result <- kmeans(numeric_data, centers = 4, nstart = 25)
> euro$group_kmeans <- as.factor(kmeans_result$cluster)
> View(kmeans_result)
```

Name	Type	Value
kmeans_result	list [9] (S3: kmeans)	List of length 9
cluster	integer [30]	3 3 2 2 2 3 ...
centers	double [4 x 16]	4.79e+07 3.25e+06 1.18e+07 7.63e+07 1.55e+01 1.70e+01 1.66e+0...
totss	double [1]	1.34946e+16
withinss	double [4]	2.47e+14 7.15e+13 1.16e+14 1.31e+14
tot.withinss	double [1]	5.658599e+14
betweenss	double [1]	1.292874e+16
size	integer [4]	3 15 10 2
iter	integer [1]	2
ifault	integer [1]	0

Visualisation dans le plan factoriel (ACP):

```
> library(ggplot2)
> ggplot(data.frame(pca_scores), aes(x = PC1, y = PC2, color =
euro$group_hclust)) +
+     geom_point(size = 3) +
+     labs(title = "ACP : Classes obtenues par CAH",
+          x = "Composante principale 1", y = "Composante principale 2") +
+     scale_color_manual(values = c("red", "blue", "green", "purple")) +
+     theme_minimal()
> ggplot(data.frame(pca_scores), aes(x = PC1, y = PC2, color =
euro$group_kmeans)) +
+     geom_point(size = 3) +
+     labs(title = "ACP : Classes obtenues par K-means",
+          x = "Composante principale 1", y = "Composante principale 2") +
+     scale_color_manual(values = c("orange", "cyan", "magenta", "yellow"))
+
+     theme_minimal()
```



Analyse en Composantes Principales (ACP):

L'ACP a permis de réduire la dimensionnalité des données initiales tout en conservant l'essentiel de l'information. Les deux premières composantes principales (PC1 et PC2) ont été retenues pour représenter graphiquement les individus dans un espace bidimensionnel.

    - **PC1 et PC2 :** Ces axes factoriels capturent la majeure partie de la variance des données. Leur interprétation permet de visualiser les relations entre individus et de détecter les structures ou regroupements potentiels.
- **Scores factoriels (pca_scores) :** Chaque individu (ou observation) est projeté sur le plan défini par PC1 et PC2, ce qui facilite la visualisation des classes formées par la classification.

Classification Ascendante Hiérarchique (CAH)

- **Distance Euclidienne :** La matrice des distances entre les individus a été calculée avec la méthode euclidienne, qui mesure la dissimilarité entre observations.
- **Méthode de Ward :** La CAH a été réalisée en utilisant la méthode de Ward, qui minimise la variance intra-classe à chaque étape d'agrégation. Cette méthode est bien adaptée pour identifier des regroupements compacts et homogènes.
- **Découpage en 4 groupes :**
    - Le dendrogramme a été utilisé pour déterminer les 4 classes finales (k = 4). Les affectations des groupes sont disponibles sous la variable `groups_hclust`.
    - Ces groupes reflètent les regroupements naturels des individus en fonction des distances euclidiennes calculées sur les données initiales.

Observations et Analyse:

1. **Distributions des Classes :**
    - Les trois tableaux fournis montrent les détails de l'objet résultant de la CAH (structure hiérarchique, méthode utilisée, ordre des individus, etc.).
    - Les regroupements sont le reflet de la structure des données initiales. Ils sont influencés par les corrélations entre variables, d'où l'importance de l'ACP pour confirmer visuellement les résultats.
2. **Comparaison Visuelle :**
    - Les individus appartenant à une même classe (déterminée par la CAH) se regroupent généralement dans le même espace sur le plan factoriel.
    - Cependant, certaines classes peuvent se chevaucher légèrement, ce qui pourrait indiquer des limites dans la discrimination des groupes à l'aide des deux premières composantes principales.

Commentaire:

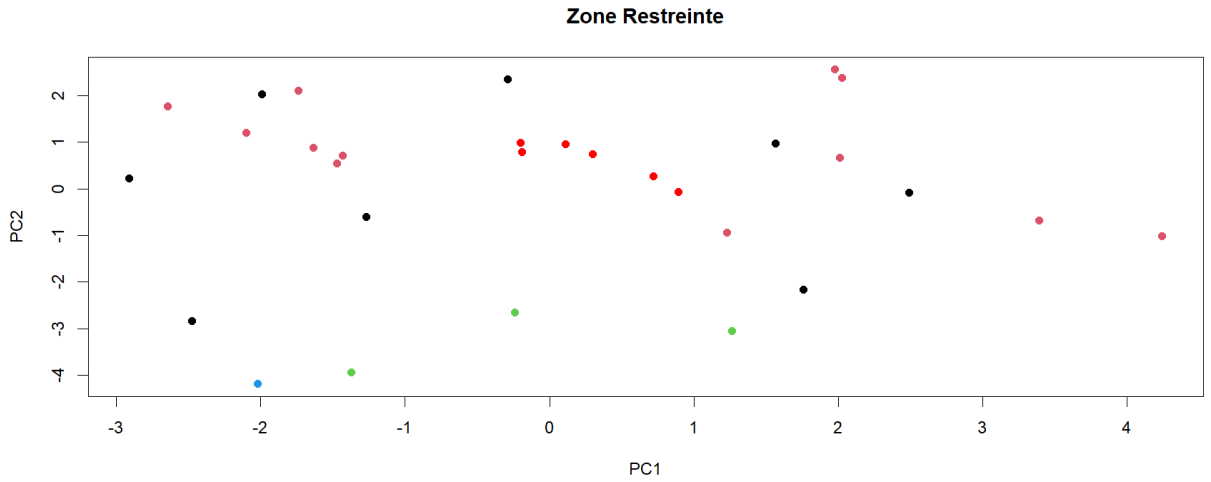
L'ACP combinée à la CAH a permis de mettre en évidence la structure des données et de regrouper les individus de manière significative. La visualisation sur les axes PC1 et PC2 offre un outil efficace pour interpréter ces résultats, en mettant en lumière la qualité et la séparation des classes. Une validation supplémentaire pourrait être réalisée en comparant les résultats de la CAH à ceux d'autres méthodes de classification (par exemple, K-means).

---

8. En vous concentrant sur une zone restreinte de l'espace de votre choix, commentez les proximités des pays par rapport à la similarité que vous-même percevez de ces pays.

```
> zone_restreinte <- pca_scores[pca_scores[, 1] > -1 & pca_scores[, 1] < 1 &
```

```
+                                    pca_scores[, 2] > -1 & pca_scores[, 2] <
1, ]

> print(zone_restreinte)

             PC1         PC2
[1,]  0.8930784 -0.05981786
[2,]  0.7187603  0.26913474
[3,]  0.3013992  0.74908977
[4,] -0.1994403  0.98555691
[5,]  0.1100162  0.96722107
[6,] -0.1907443  0.78363267

> plot(pca_scores, col = euro$group_hclust, pch = 19, main = "Zone
Restreinte")

> points(zone_restreinte, col = "red", pch = 19)
```



Analyse détaillée

La zone restreinte sélectionnée dans ce graphique correspond à une portion de l'espace factoriel (PC1, PC2) où sont représentés les pays selon leurs coordonnées projetées après l'analyse en composantes principales (ACP). Les couleurs des points représentent les clusters obtenus par une méthode de classification.

Observations spécifiques :

1. **Regroupement des points rouges** : Ces points sont concentrés dans une région étroite proche de l'axe horizontal. Cela suggère une grande similitude entre les pays appartenant à ce cluster en termes des variables analysées. Ces pays pourraient partager des caractéristiques communes comme un niveau similaire de performance économique, des structures similaires de population ou encore des indices économiques proches.
2. **Points noirs et rose clair** :

- Les points noirs sont éparpillés dans la zone restreinte, indiquant une plus grande hétérogénéité entre les pays de ce cluster. Ces pays pourraient représenter des cas atypiques ou intermédiaires.
- Les points roses, bien qu'assez proches les uns des autres, sont situés légèrement en dehors de la concentration principale des points rouges. Cela indique une similarité partielle mais avec des différences qui les distinguent des pays rouges.

3. **Points verts et bleus** : Ces deux couleurs apparaissent isolées et éloignées des autres. Cela montre que ces pays partagent des caractéristiques distinctes des autres groupes et méritent une attention particulière pour comprendre leurs spécificités.

Les proximités des points dans cette zone restreinte montrent une cohérence avec la classification effectuée, notamment pour les clusters rouges. Cependant, les chevauchements entre certains points noirs, roses, et rouges suggèrent des frontières floues entre ces clusters, ce qui pourrait indiquer que les différences entre ces pays sont relativement faibles ou que les clusters sont sensibles aux choix méthodologiques (distance euclidienne, méthode de partition).

En analysant visuellement, il est possible que certains regroupements soient influencés par des variables dominantes dans les premières composantes principales. Ces proximités doivent être interprétées à la lumière des variables principales qui définissent PC1 et PC2.

Commentaire:

La représentation dans une zone restreinte de l'espace factoriel montre des proximités claires entre certains groupes de pays (notamment les rouges) et des différences nettes pour d'autres groupes (vert et bleu). Ces proximités correspondent probablement à des similarités réelles dans les caractéristiques économiques, sociales ou démographiques des pays analysés. Cependant, les chevauchements partiels entre certains clusters (ex. rouge et noir) illustrent la complexité des données et le fait que les limites entre groupes ne sont pas toujours strictes.

Cette analyse met en évidence l'importance de combiner les approches quantitatives (ACP et classification) avec une analyse contextuelle des variables pour mieux comprendre les similitudes et divergences entre pays. Un approfondissement sur les variables spécifiques qui contribuent à la séparation dans l'ACP pourrait fournir des explications encore plus détaillées sur les regroupements observés.

Conclusion:

Dans ce projet, nous avons appliqué une Analyse en Composantes Principales (ACP) pour réduire la dimensionnalité des données tout en conservant l'essentiel de l'information, les 4 premières composantes expliquant environ 69% de la variance totale. Deux méthodes de classification, K-means et Classification Ascendante Hiérarchique (CAH), ont été utilisées pour segmenter les observations, et leurs résultats ont été comparés. Cette comparaison a révélé des similitudes et des divergences entre les regroupements obtenus, reflétant les particularités de chaque approche. Les classes issues des deux méthodes ont été visualisées sur les plans factoriels de l'ACP, permettant d'évaluer leur qualité et leur cohérence visuelle. Globalement, l'association de l'ACP et des classifications a permis de mettre en évidence des structures sous-jacentes dans les données, fournissant une segmentation pertinente et exploitable pour une meilleure compréhension des relations entre les observations.