ANALYSE EN COMPOSANTES PRINCIPALES

Audrey KIRCHER
M1 IEF

CRÉATION DU DATA SET

```r
#Installation package pour collecter les données sur DATA BANK
install.packages("WDI") #World Devlp Indicator
library(WDI)

# Création du dataset
Veolia_2019 <- WDI(indicator = c("SI.POV.DDAY","SH.H2O.SAFE.ZS","CC.EST", "PV.EST", "EG.FEC.RNEW.ZS",
                        "EN.ATM.CO2E.PC", "AG.LND.FRST.ZS", "SI.POV.GINI", "SE.SEC.ENRR" ),
         country = "all", start = 2019, end = 2019)
head(Veolia_2019)

#nom des colonnes
tx_pov =Veolia_2019$SI.POV.DDAY
corru =Veolia_2019$CC.EST
stab_pol =Veolia_2019$PV.EST
conso_re =Veolia_2019$EG.FEC.RNEW.ZS
em_co2 =Veolia_2019$EN.ATM.CO2E.PC
suf_forest =Veolia_2019$AG.LND.FRST.ZS
gini =Veolia_2019$SI.POV.GINI
tx_scolar =Veolia_2019$SE.SEC.ENRR
```

VEOLIA

MISE EN FORME + CORRÉLATION

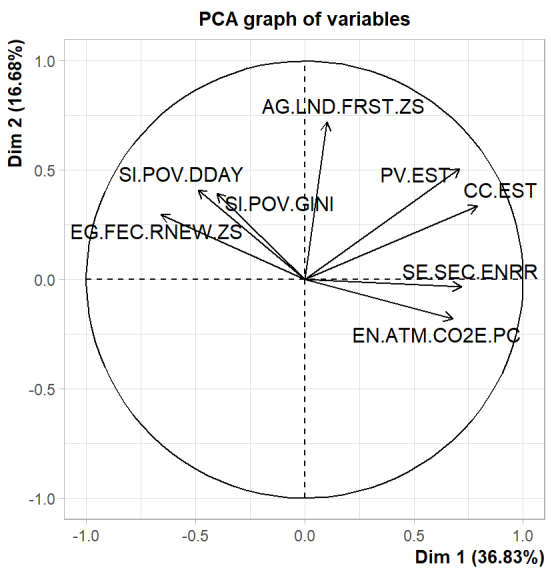
```r
#mise en forme du dataset
f =Veolia_2019[, -2:-4] #Supp les colonnes 2 et 4
country =Veolia_2019$country #Création d'un vecteur country contenant le nom des pays
row.names(f) = country #Les lignes du data frame f auront comme noms les valeurs du vecteur country
fe =f[,-1] #création nouveau data frame fe contenant à partir de f en supprimant la colonne 1


# matrice de correlation
cor(fe)
```

ACP (ANALYSE EN COMPOSANTES PRINCIPALES)

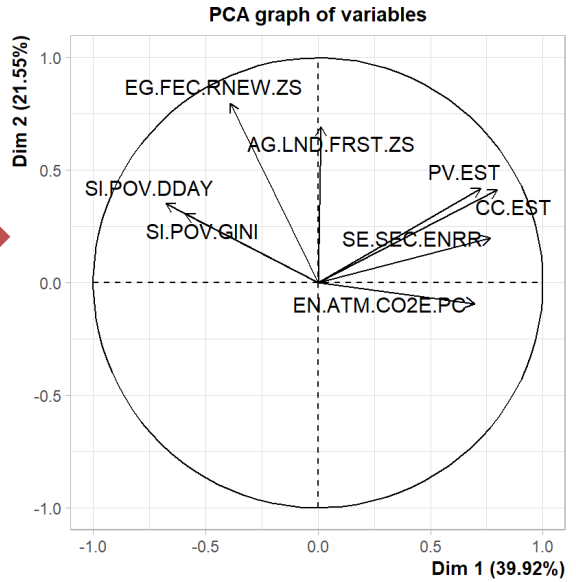
```r
#Installation package
library("FactoMineR")
bma =PCA(fe)

#supprimer les valeurs manquantes
fe_clean <-na.omit(fe) #supp NA (valeurs manquantes)
head(fe_clean)

#standardiser les données
fe_scaled =scale(fe_clean) #centre les variables sur une moyenne de 0 et une variance de 1

#nouvelle ACP (BON GRAPH)
apc_fe =PCA(fe_scaled)
```

Fonction Scale ( ) :

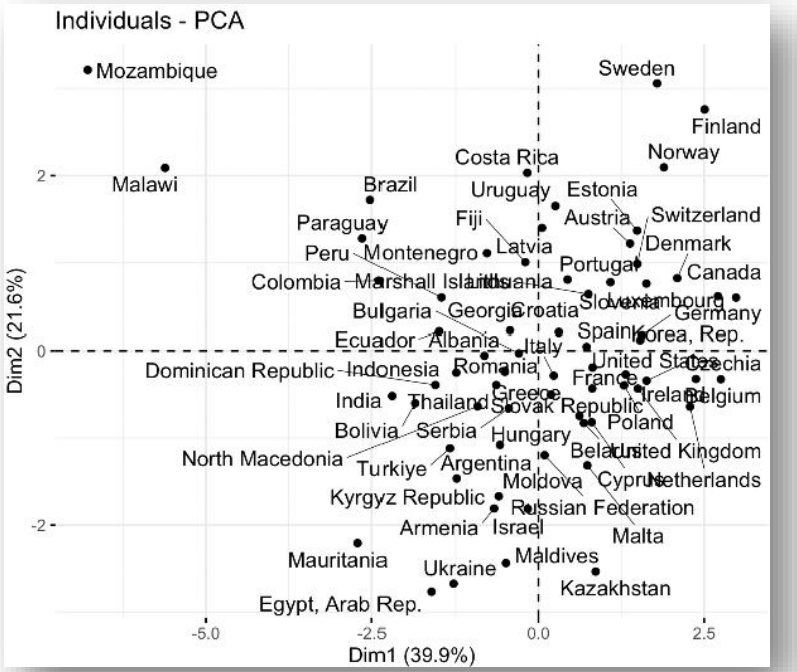
Cela garantit que chaque variable contribue de manière équitable à la construction des composantes principales, sans que l'une domine en raison de son échelle plus élevée

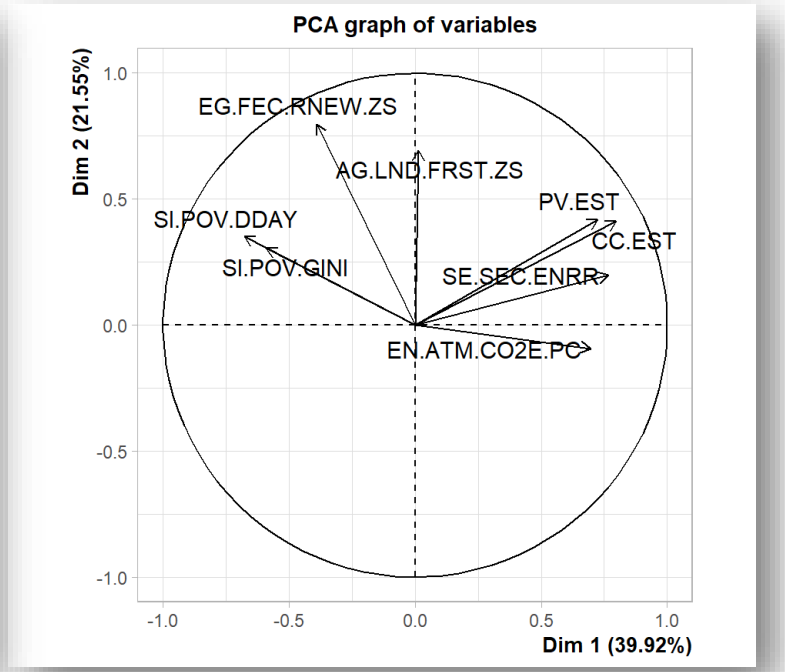


Standardisation



VISUALISATION ACP





Interprétation :

La dimension 1 explique 39,92% de la dispersion totale de l'échantillon. La dimension 2 = 21,55%.

Dim 1 : Emission CO2 + taux scolarité
Dim 2 : Surfaces forestières

ATTENTION au sens inverse de lecture !

→ Corrup élevé = indique peu de corruption

→ stab_pol élevé = Indique une forte stabilité politique et une absence de violence

```r
#visualisation
install.packages("ggplot2")
library(ggplot2)
install.packages("factoextra")
library(factoextra)

fviz_pca_ind(apc_fe, label ="ind", repel =TRUE) # affiche les noms des pays dans le cercle
```

Légende :

tx_pauv =SI.POV.DDAY
corrup =CC.EST
stab_pol =PV.EST
conso_re=EG.FEC.RNEW.ZS
em_co2=EN.ATM.CO2E.PC
suf_forest=AG.LND.FRST.ZS
gini =SI.POV.GINI
tx_scolar=SE.SEC.ENRR

VEOLIA

CONCLUSION : OÙ INVESTIR ?

Pour respecter au mieux les critères ESG, il faut :

- **Pauvreté (SI.POV.DDAY)** : Un taux de pauvreté bas est souhaitable.
- **Stabilité politique (CC.EST et PV.EST)** : Des scores élevés indiquent une meilleure gouvernance.
- **Énergie renouvelable (EG.FEC.RNEW.ZS)** : Un pourcentage élevé de l'énergie renouvelable est préférable.
- **Émissions de CO_2 (EN.ATM.CO2E.PC)** : Des émissions par habitant plus faibles sont meilleures.
- **Couverture forestière (AG.LND.FRST.ZS)** : Un pourcentage élevé est positif pour l'environnement.
- **Inégalités (SI.POV.GINI)** : Un coefficient de Gini plus faible indique moins d'inégalités.
- **Éducation (SE.SEC.ENRR)** : Un taux élevé de scolarisation dans l'enseignement secondaire est souhaitable.

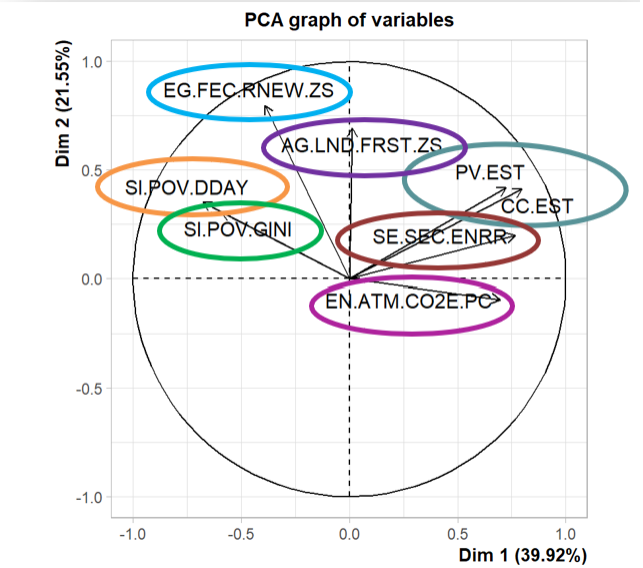
Deux catégories de pays émergent de notre analyse

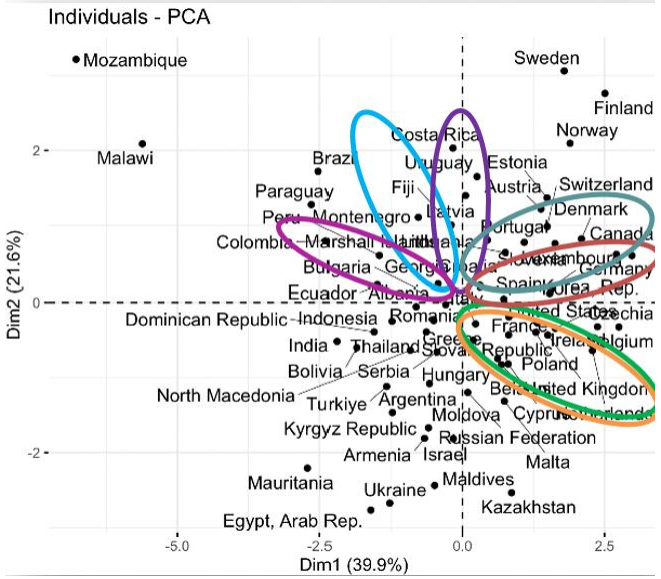
Énergie renouvelable (EG.FEC.RNEW.ZS)
Émissions de CO_2 (EN.ATM.CO2E.PC)
Couverture forestière (AG.LND.FRST.ZS)

⇒ **Pays Amérique latine**

Stabilité politique (CC.EST et PV.EST)
Éducation (SE.SEC.ENRR)
Inégalités (SI.POV.GINI)
Pauvreté (SI.POV.DDAY)

⇒ **Pays Européens**





ANALYSE DES VALEURS PROPRES

→ Ces valeurs propres représentent la quantité de variance expliquée par chaque composante principale

```
#Analyse des valeurs propres
apc_fe$eig
```

Var Comp 1 = eigenvalue/nbr de composante = 3,19/8 = 39,92%

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.1938701	39.923376	39.92338
comp 2	1.7240401	21.550502	61.47388
comp 3	1.0537293	13.171616	74.64549
comp 4	0.7341603	9.177003	83.82250
comp 5	0.5801429	7.251786	91.07428
comp 6	0.3321447	4.151808	95.22609
comp 7	0.2131177	2.663972	97.89006
comp 8	0.1687950	2.109937	100.00000

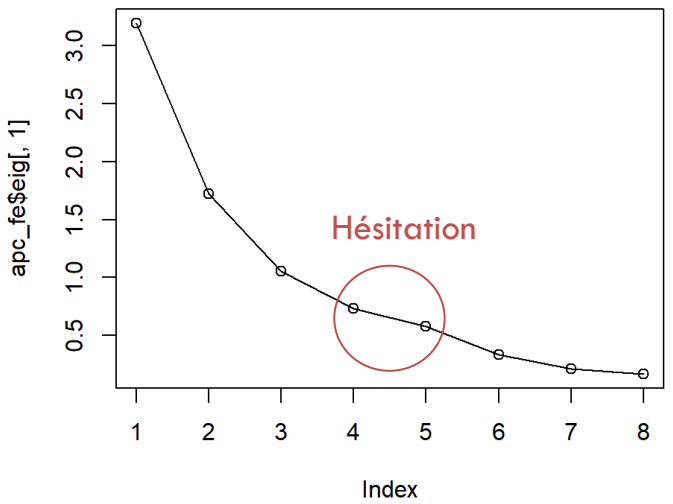
Méthode de KAISER : Identifier les valeurs qui dépassent 1

Dans notre cas, les AXES 1-2-3 ont une valeur propre > 1

```
comp 1  3.1938701
comp 2  1.7240401
comp 3  1.0537293
```

Ces composantes sont considérées comme importantes et doivent être retenues dans l'analyse, car elles apportent une information significative. (valeur seuil au dessus de 1)

Méthode du coude :



ANALYSE CORRELATION DIM

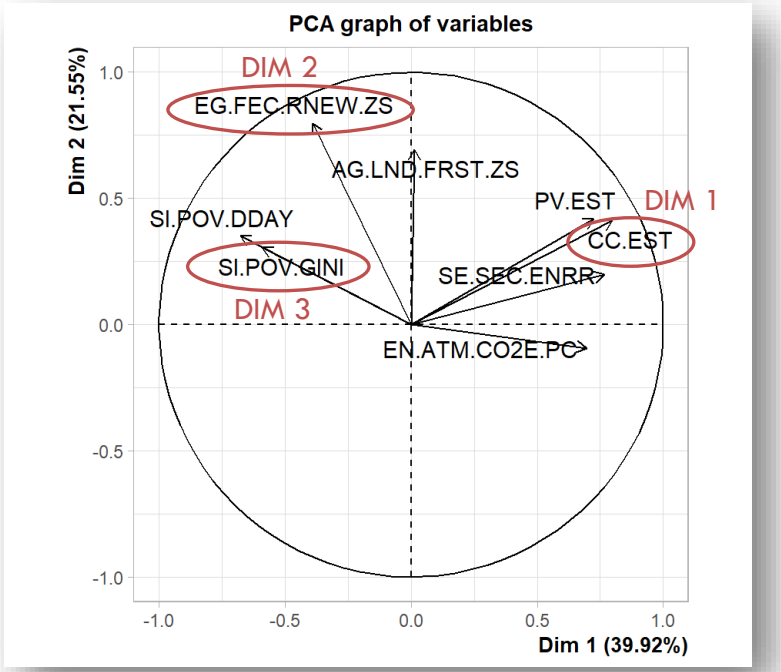
→ Cela permet de mettre en évidence les variables qui ont le plus d'impact sur chaque composante principale.

```
#Examination des corrélations des dim
round(apc_fe$var$cor,1)
round(apc_fe$var$cor,2) #pour avoir deux chiffres après la virgule
```

	Dim.1	Dim.2	Dim.3
SI.POV.DDAY	-0.68	0.35	-0.46
CC.EST	0.80	0.41	-0.19
PV.EST	0.72	0.42	-0.31
EG.FEC.RNEW.ZS	-0.39	0.80	-0.25
EN.ATM.CO2E.PC	0.70	-0.09	-0.01
AG.LND.FRST.ZS	0.01	0.69	0.49
SI.POV.GINI	-0.59	0.30	0.52
SE.SEC.ENRR	0.77	0.20	0.36

Interprétations :

- La dimension 1 est expliquée à 80% par la variable de Corrup
- La dimension 2 est expliquée à 80% par la variable de Conso_re
- La dimension 3 est expliquée à 52% par la variable de Gini

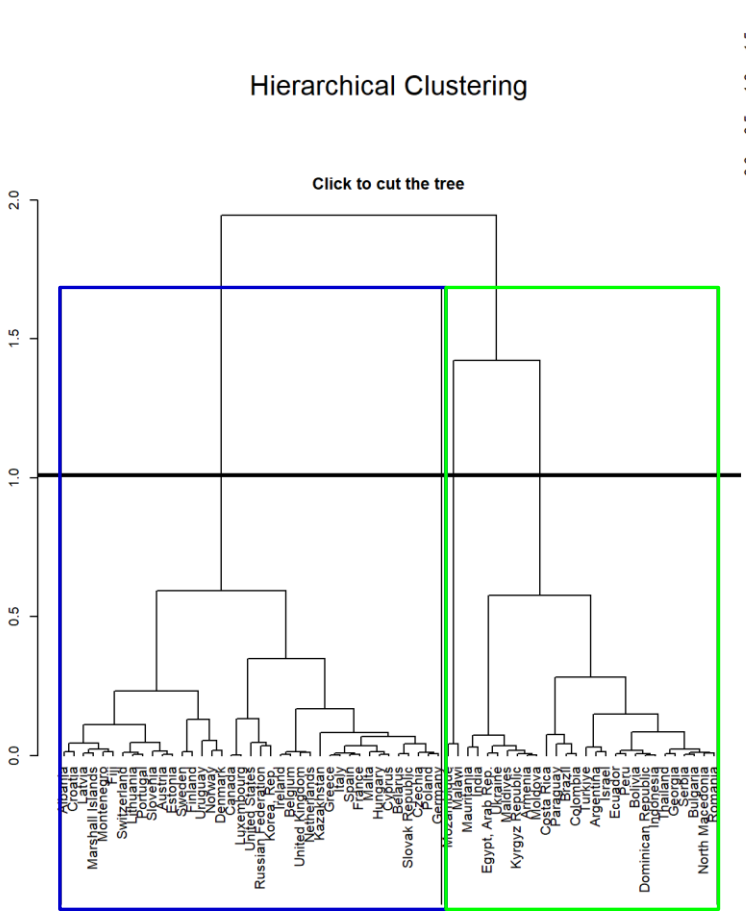


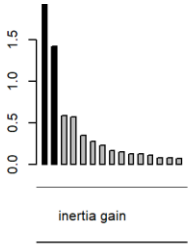
Légende :

tx_pauv =SI.POV.DDAY
corrup =CC.EST
stab_pol =PV.EST
conso_re=EG.FEC.RNEW.ZS
em_co2=EN.ATM.CO2E.PC
suf_forest=AG.LND.FRST.ZS
gini =SI.POV.GINI
tx_scolar=SE.SEC.ENRR

MÉTHODE CLUSTERING (OU REGROUPEMENT)

→ Cela permet de former des groupes au sein d'un échantillon.





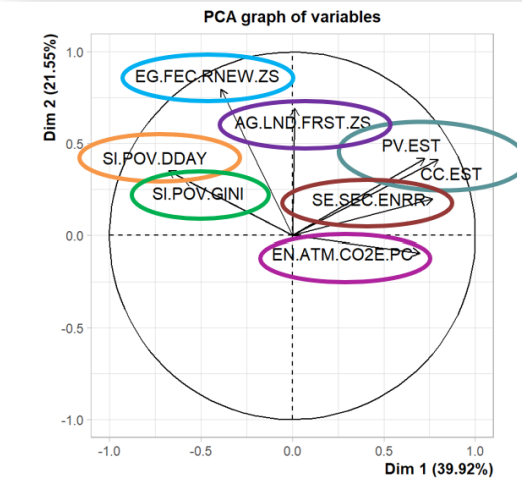
2 groupes identifiés

Stabilité politique (CC.EST et PV.EST)
Éducation (SE.SEC.ENRR)
Inégalités (SI.POV.GINI)
Pauvreté (SI.POV.DDAY)

⇒ **Pays Européens**

Énergie renouvelable (EG.FEC.RNEW.ZS)
Émissions de CO_2 (EN.ATM.CO2E.PC)
Couverture forestière (AG.LND.FRST.ZS)

⇒ **Pays Amérique latine**



Légende :

tx_pauv =SI.POV.DDAY
corrup =CC.EST
stab_pol =PV.EST
conso_re=EG.FEC.RNEW.ZS
em_co2=EN.ATM.CO2E.PC
suf_forest=AG.LND.FRST.ZS
gini =SI.POV.GINI
tx_scolar=SE.SEC.ENRR