Projet Stat Descriptive

Paul Slisse | Guillaume Staub

2024-04-15

On commence par charger les librairies dont on aura besoin pour réaliser les différentes manipulations des données.

I. Description du jeu de données

```
#Importation des données renew
renew <- read.table("renew.txt", sep="", stringsAsFactors=TRUE)
flextable(head(renew,5),cwidth=1, col_keys=c("Entity", "Year", "Access.elec", "Rnw.energy.share",
"Elec.fossil", "Elec.rnw"))
```

Entity	Year	Access.elec	Rnw.energy.share	Elec.fossil	Elec.rnw
Algeria	2,001	98.96687	0.43	24.96	0.07
Algeria	2,002	98.95306	0.51	25.94	0.06
Algeria	2,003	98.93401	0.47	27.54	0.26
Algeria	2,004	98.91208	0.44	29.14	0.25
Algeria	2,005	98.88961	0.58	31.36	0.55

```
nbvar <- length(head(renew))
nbinv <- dim(renew)[1]
str(renew)
```

```
## 'data.frame':    343 obs. of  10 variables:
##  $ Entity              : Factor w/ 20 levels "Algeria","Argentina",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Year                : int  2001 2002 2003 2004 2005 2009 2010 2011 2012 2013 ...
##  $ Access.elec         : num  99 99 98.9 98.9 98.9 ...
##  $ Rnw.energy.share    : num  0.43 0.51 0.47 0.44 0.58 0.31 0.26 0.18 0.18 0.13 ...
##  $ Elec.fossil         : num  25 25.9 27.5 29.1 31.4 ...
##  $ Elec.rnw            : num  0.07 0.06 0.26 0.25 0.55 0.3 0.17 0.5 0.62 0.33 ...
##  $ Prim.energy.consumpt: num  9962 10180 10510 10759 11114 ...
##  $ co2                 : num  78650 82400 88190 89490 94190 ...
##  $ GDP.capita          : num  1741 1782 2103 2610 3113 ...
##  $ Continent           : Factor w/ 4 levels "Africa","America",..: 1 1 1 1 1 1 1 1 1 1 ...
```

On voit qu'on a bien extrait les données Il y a donc 10 variables et 343 individus.

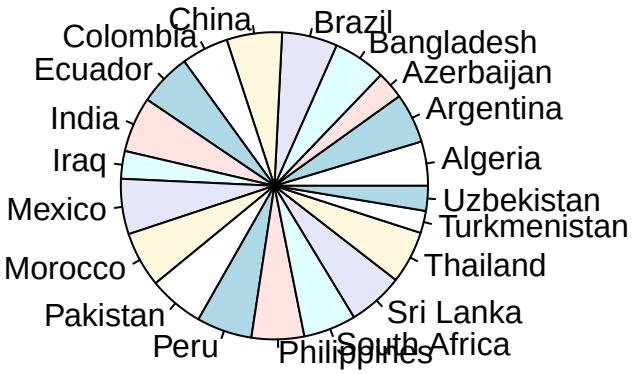
- La variable *Entity* est qualitative nominale et représente le pays auquel on s'intéresse.
- La variable *Year* est quantitative discrète et représente l'année à laquelle on s'intéresse.
- La variable *Access.elec* est quantitative continue et représente le pourcentage d'accès à l'électricité de la population dudit pays lors de ladite année.
- La variable *Rnw.energy.share* est quantitative continue et représente le pourcentage d'énergie renouvelables dans la consommation totale d'énergie du pays lors de l'année étudiée.
- La variable *Elec.fossil* est quantitative continue et représente l'électricité produite à partir d'énergie fossile dans ledit pays lors de ladite année en TWh.
- La variable *Elec.rnw* est quantitative continue et représente l'électricité produite à partir d'énergie renouvelable dans ledit pays lors de ladite année en TWh.
- La variable *Prim.energy.consumpt* est quantitative continue, et représente la consommation totale d'énergie par habitant par pays et par an en kWh.
- La variable *co2* est quantitative continue et représente les émissions de CO2 par habitant dans un pays, par année et exprimé en tonnes/personne.
- La variable *GDP.capita* est quantitative continue et représente le PIB par habitant dudit pays à ladite en année en USD.
- Enfin, la variable *Continent* est qualitative nominale et représente le continent pays étudié.

II. Analyse uni- et bi-variées

1. Analyse univariée

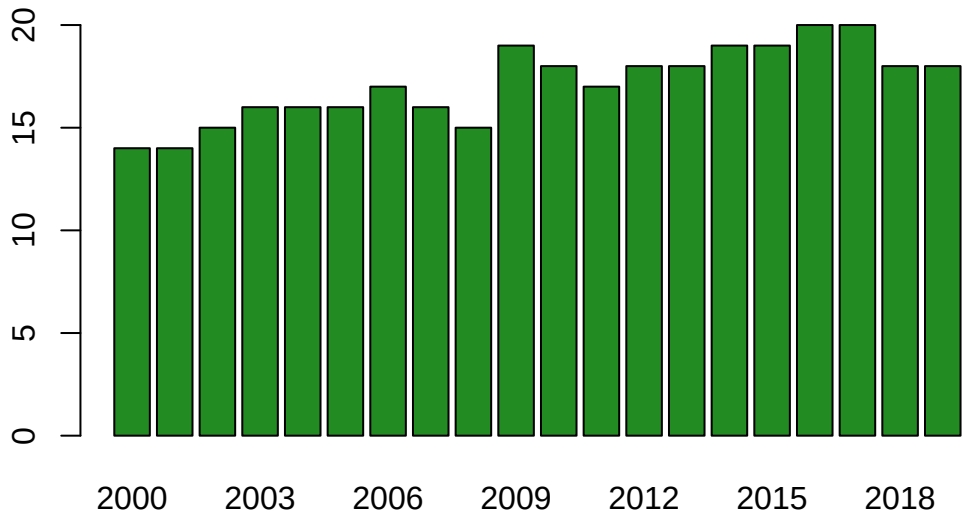
```
#Pays
Pays<-table(renew$Entity)
Pays<-data.frame(Eff=c(Pays),Freq=c(Pays)/sum(Pays))
pie(Pays$Freq,labels=row.names(Pays),main= "Representation des differents pays dans le jeu de données")
```

Representation des differents pays dans le jeu de données

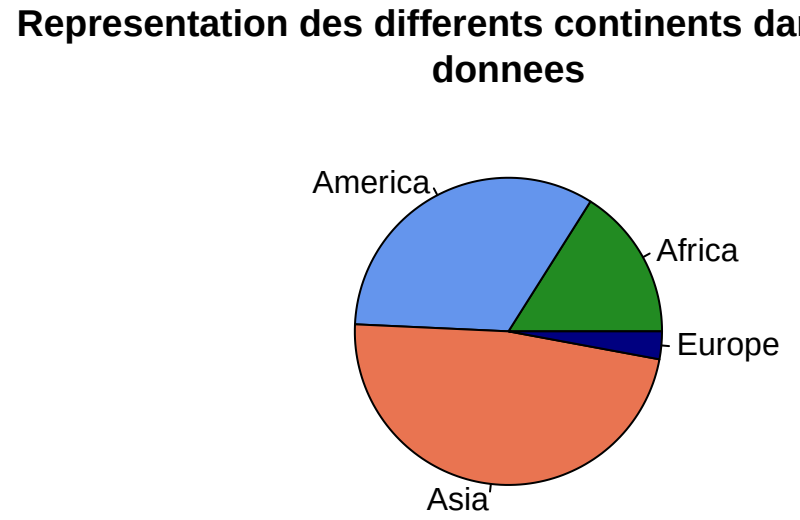



```
#Années
barplot(table(renew$Year),
main="Representation des annees dans le jeu de données", col="#228B22")
```

Representation des annees dans le jeu de données



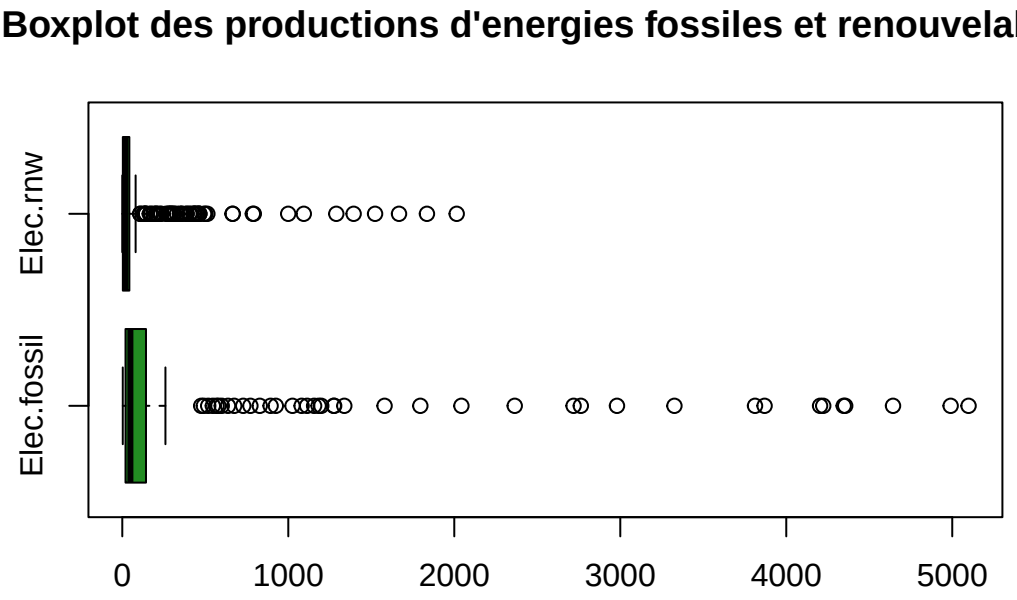
```r
#Continent
Cont<-table(renew$Continent)
Cont<-data.frame(Eff=c(Cont),Freq=c(Cont)/sum(Cont))
pie(Cont$Freq,labels=row.names(Cont),main= "Representation des differents continents dans le jeu de
donnees",col=c("#228B22","#6495ed", "#e97451","#000080"))
```



Grâce à ces trois graphiques, on voit qu'on a des pays et des années qui sont plus représentés que d'autres dans notre jeu de donnée. Cela ne sera pas à négliger dans la suite de notre étude puisque cela aura un impact sur nos résultats (par exemple Chine plus représenté que Iraq va faire grimper les indicateurs). Une solution possible pour contrer ce problème serait de faire une moyenne par pays sur la période (2000-2019) mais cela enlèverait la dépendance temporelle, l'évolution à travers les années. Les conclusions que l'on peut espérer tirer au regard de l'Europe notamment au sujet de la lutte contre le réchauffement climatique peuvent être peu représentatives. L'étude semble être plus portée sur les pays en développement d'Afrique et d'Amérique du Sud et d'Asie.

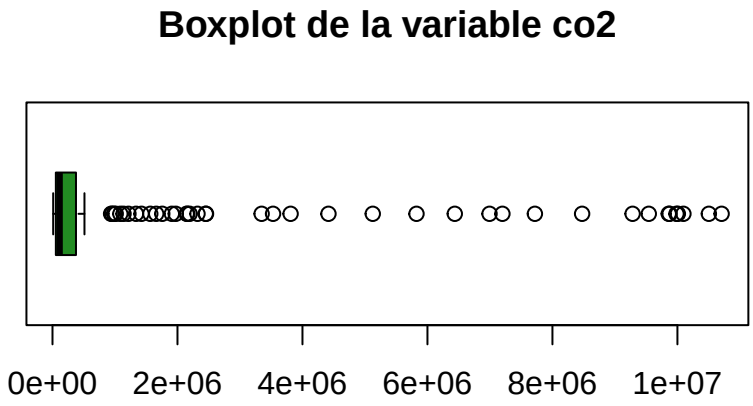
On va maintenant s'intéresser aux variables quantitatives.

```r
#Production d'énergie:Elec.fossil et Elec.rnw
boxplot(renew[5:6],main="Boxplot des productions d'energies fossiles et renouvelables", col="#228B22",
horizontal = T)
```



On voit qu'on a un grand nombre d'outliers pour les variables *Elec.fossil* et *Elec.rnw*. Cela s'explique par le fait que nous n'avons pas ramené ces variables par habitant, ainsi les pays très peuplés vont produire beaucoup plus d'électricité que les plus petit pays.

```
#CO2
boxplot(renew$co2,main="Boxplot de la variable co2", col="#228B22", horizontal = T)
```




```
M=c(mean(renew$co2),var(renew$co2))
```

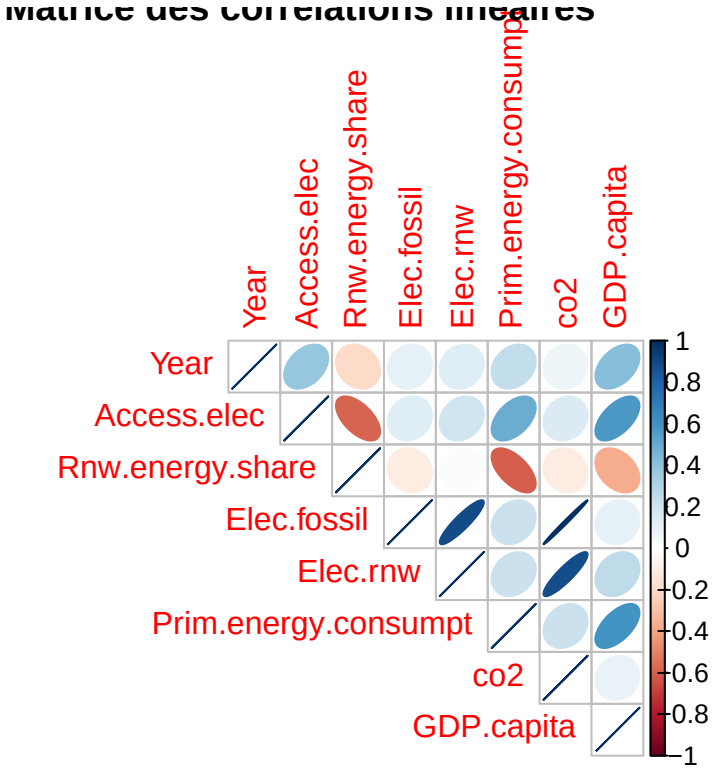
Le grand nombre d'outliers pour la variable *co*2 et l'écart-type de 1.8776447×10^6 montre une grande disparité

dans l'émission de CO2 par personne. On en déduit que l'émission de CO2 par personne va grandement varier selon le pays et l'année.

2. Analyse bi-variée

On va commencer par faire une analyse bi-variée des variables quatitatives qui se fait très facilement à l'aide du coefficient de corrélation linéaire.

```
Quanti<-renew[2:9]
Cor_Quanti=cor(Quanti)
corrplot(Cor_Quanti,method="ellipse",type="upper",title="Matrice des correlations lineaires")
```



On observe des corrélations entre multiples variables.
Premièrement, on a la production d'énergie fossile qui est liée à la production d'énergie renouvelables. Cela s'explique notamment par le fait que les pays qui produisent beaucoup d'énergies renouvelables ont un besoin énorme d'énergie et produisent donc également beaucoup d'énergie fossile.

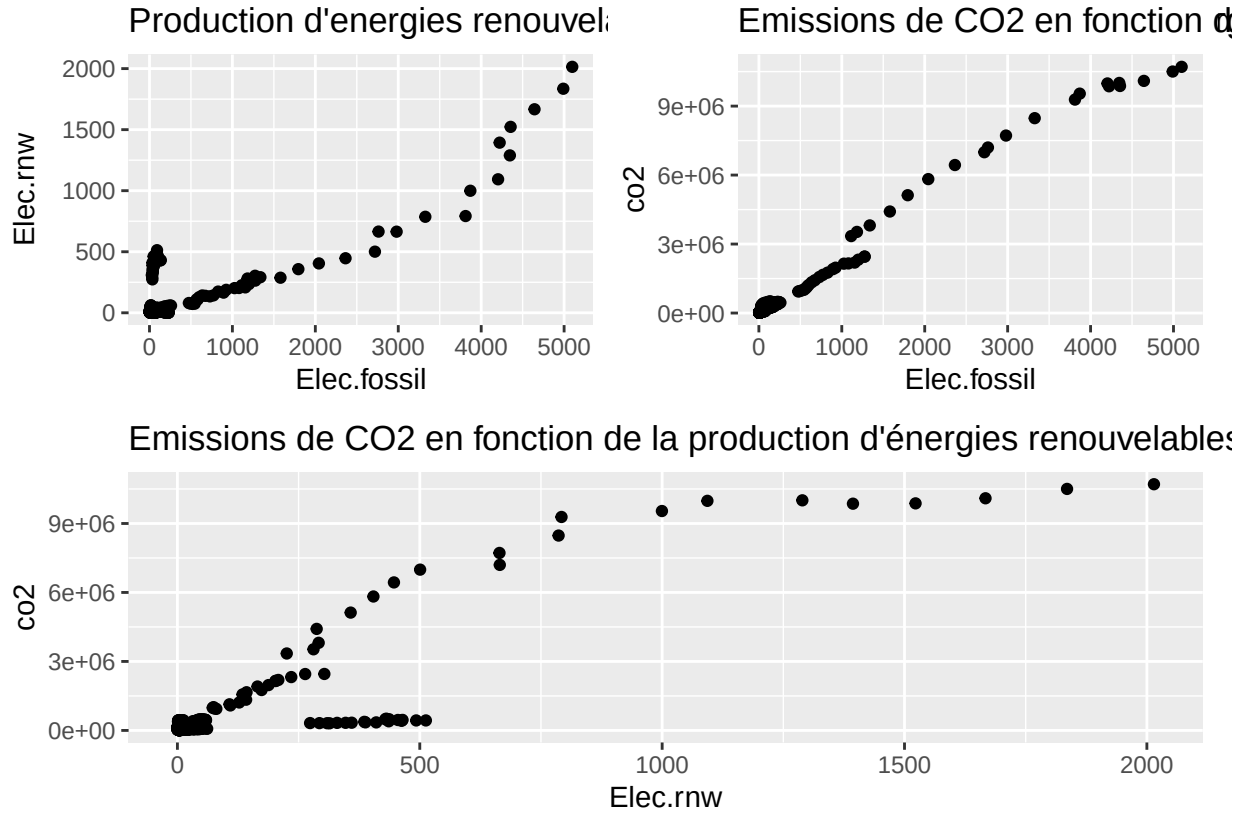
```
p1<-ggplot(renew,aes(x=Elec.fossil,y=Elec.rnw))+
  ggtitle(label="Production d'energies renouvelables en fonction de la production d'energies fossiles")+
  geom_point()
```

De plus, la production d'énergie fossile est également liée aux émissions de CO2 par habitant, ce à quoi on pouvait s'attendre puisque les énergies fossiles émettent beaucoup de gaz à effet de serre.

```
p2<-ggplot(renew,aes(x=Elec.fossil,y=co2))+
  ggtitle(label="Emissions de CO2 en fonction de la production d'energies fossiles")+
  geom_point()
```

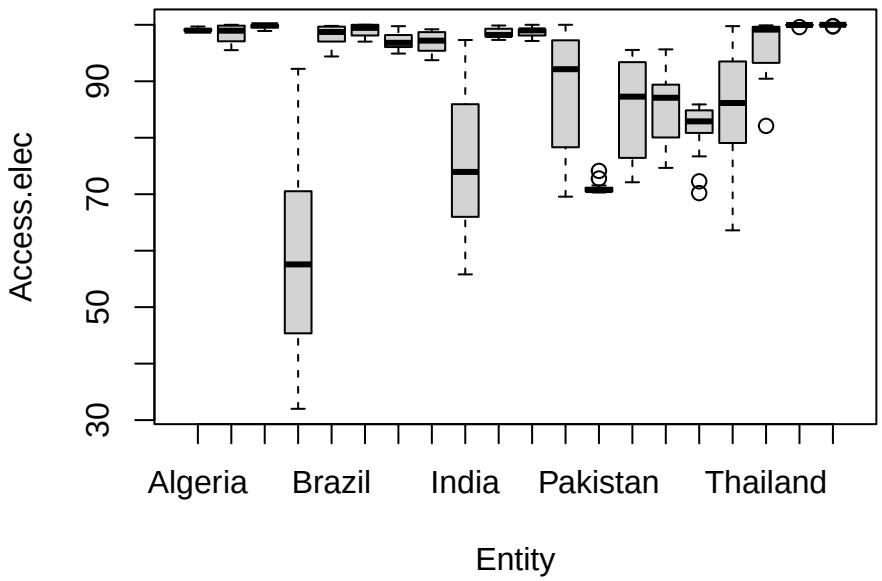
On observe également une corrélation entre la production d'énergie renouvelables et les émissions de CO2. Cette corrélation peut sembler étrange à première vue mais comme on l'a dit auparavant, c'est parce que les pays qui produisent beaucoup d'énergies renouvelables sont ceux qui ont un grand besoin d'énergie et qui produisent également beaucoup via les énergies fossiles.

```
p3<-ggplot(renew,aes(x=Elec.rnw,y=co2))+
  ggtitle("Emissions de CO2 en fonction de la production d'énergies renouvelables")+
  geom_point()
(p1+p2)/p3
```



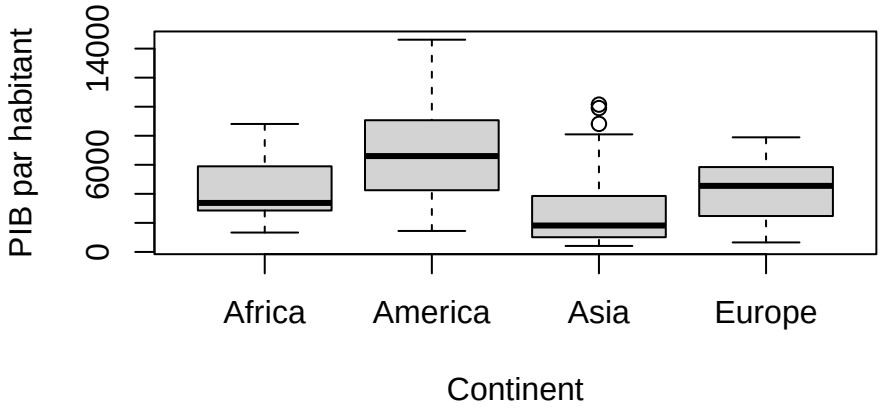
On va maintenant regarder si les variables qualitatives *Entity* et *Continent* sont liées à d'autres variables. Premièrement, forcément, ces 2 variables sont liées puisque qu'un pays appartient à un seul continent. On sait que le lien entre ces deux variables et la variable $Year$ est inintéressant car $Year$ donne juste une information sur l'année des données qui suivent pour le pays qui appartient toujours au même continent. On va calculer les rapports de corrélation η^2 pour les variables *Entity* et *Continent* avec les variables quantitatives pour voir les potentiels lien avec les variables quantitatives.

```
#Entity et Access.elec
boxplot(renew$Access.elec~renew$Entity,xlab="Entity",ylab="Access.elec")
```



On remarque que la variable *Entity* semble liée à toutes les variables quatitatives continues (coefficient $\eta^2 > 0,6$). Cela ne permet pas de réaliser des grandes déductions, cela était prévisible puisque chaque pays à sa propre politique concerant la production d'énergie et que certaines variables dépendent du nombre d'habitants et donc du pays.(A voir au nombre de pages)

```
#Continent et GDP
GDP_cont=eta2(renew$GDP.capita,renew$Continent)
boxplot(renew$GDP.capita~renew$Continent,xlab="Continent",ylab="PIB par habitant")
```



On observe une légère corrélation entre le continent et le PIB par habitant (η^2=0.3404149). Les pays avec

un PIB par habitant élevé seront plutôt ceux d'Amérique et ceux avec un PIB faible sont plutôt situé en Asie.

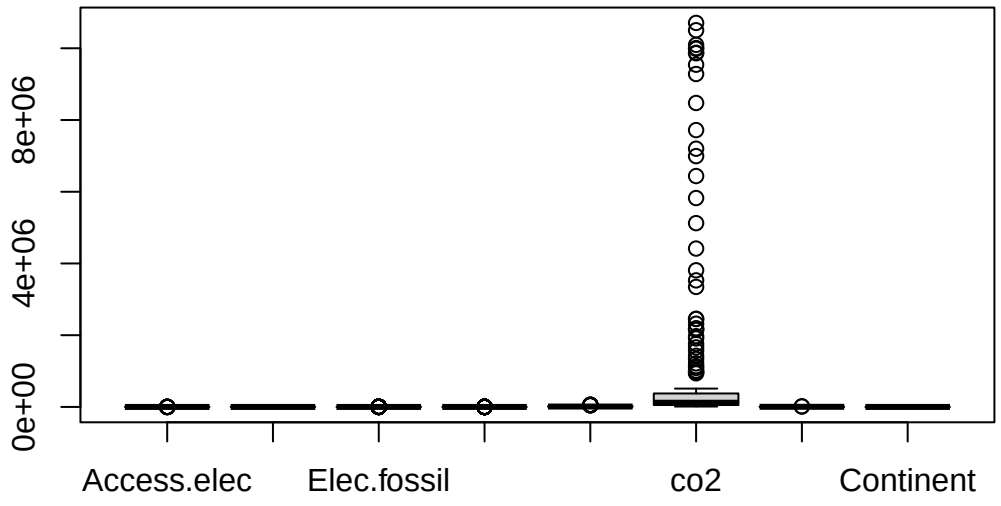
III. ACP

Tout d'abord préparons notre matrice de travail, pour cela on va mettre en nom de ligne le nom du pays et l'année. Puis nous enlèverons les variables *Entity* et *Year* afin de réaliser l'ACP.

```
#Préparation des données pour l'ACP
row.names(renew)=paste(renew$Entity,renew$Year,sep="")
renew<-renew[3:10]
flextable(head(renew,5),cwidth=1,col_keys=c("Access.elec","Rnw.energy.share","Elec_fossil","Elec.rnw",
"Prim.energy.consumpt","co2","GDP.capita"))
```

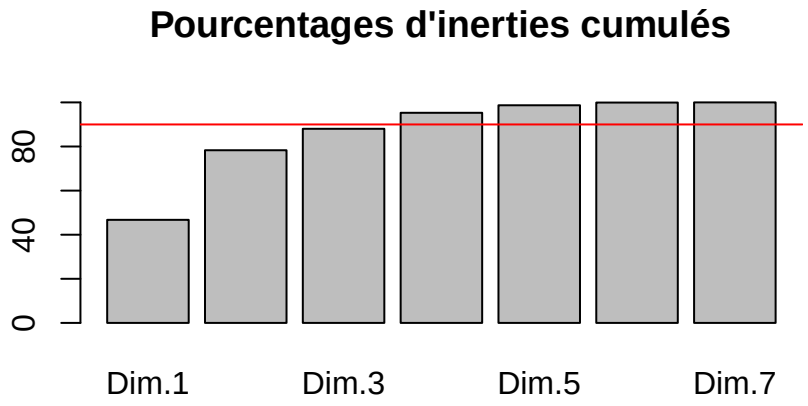
Access.elec	Rnw.energy.share	Elec.rnw	Prim.energy.consumpt	co2	GDP.capita
98.96687	0.43	0.07	9,961.64	78,650	1,740.607
98.95306	0.51	0.06	10,180.35	82,400	1,781.829
98.93401	0.47	0.26	10,510.46	88,190	2,103.381
98.91208	0.44	0.25	10,759.02	89,490	2,610.185
98.88961	0.58	0.55	11,113.72	94,190	3,113.095

```
boxplot(renew)
```



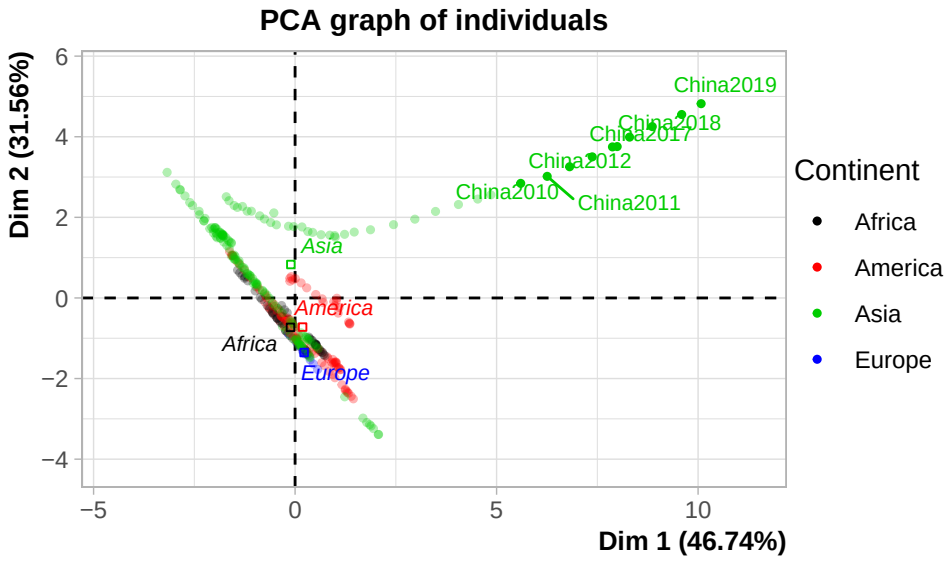
Les variables n'ont pas toutes les mêmes unités et prennent donc des valeurs différentes avec des variations plus ou moins importantes, notamment la variable $co2$ qui peut prendre des valeurs très élevées et dispersées. Ainsi, on va faire une ACP centrée réduite.

```
#Calcul de l'ACP et affichage de l'inertie
res.ACP<-PCA(renew,scale.unit = TRUE,ncp=7,quali.sup = 8,graph=FALSE)
barplot(res.ACP$eig[,"cumulative percentage of variance"],names.arg=paste("Dim",1:7,sep="."),
        main="Pourcentages d'inerties cumulés")
abline(h=90,col="red")
```

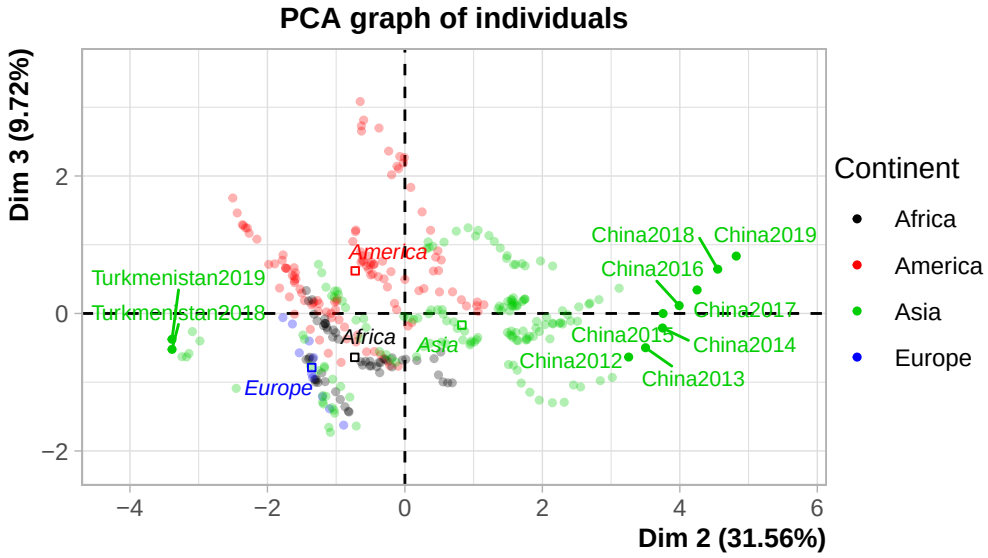


On ne va garder que 3 composantes principales, puisqu'avec 3 composantes principales, on a quasiment 90% de l'inertie totale, rajouter d'autres composantes ne fait gagner que très peu d'inertie.

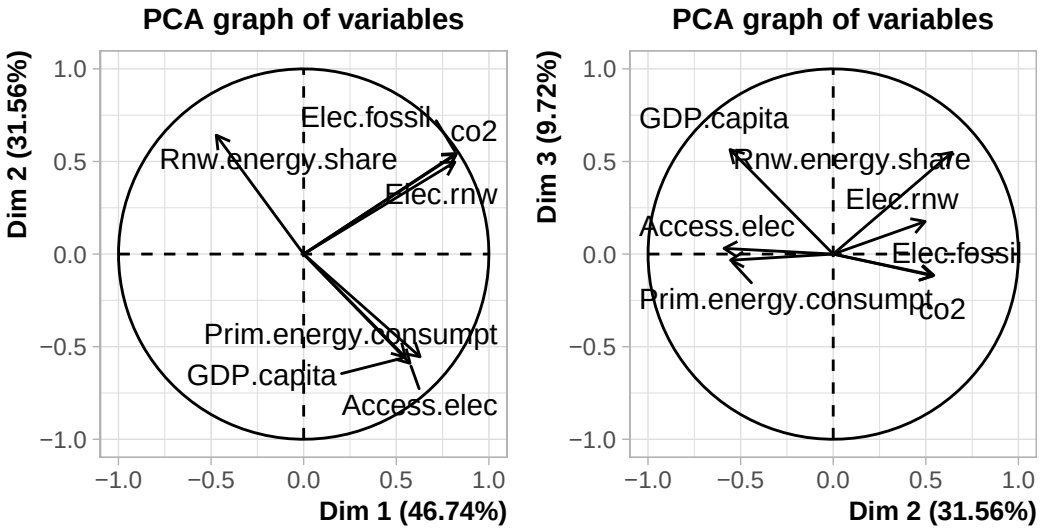
```
#Affichage des résultats de l'ACP centrée réduite
p1<-plot(res.ACP,cex=0.7,habillage=8,choix="ind",select="contrib 10")
p2<-plot(res.ACP,axes=c(2,3),cex=0.7,habillage=8,choix="ind",select="contrib 10")
p3<-plot(res.ACP,choix="var")
p4<-plot(res.ACP,axes=c(2,3),choix="var")
p1
```


PCA graph of individuals
Dim 2 (31.56%)
Dim 1 (46.74%)
China2019
China2018
China2017
China2012
China2010
China2011
Asia
America
Africa
Europe
Continent
Africa
America
Asia
Europe

```
p2
```


PCA graph of individuals
Dim 3 (9.72%)
Dim 2 (31.56%)
China2018
China2019
China2016
China2017
China2015
China2014
China2012
China2013
Turkmenistan2019
Turkmenistan2018
America
Africa
Asia
Europe
Continent
Africa
America
Asia
Europe

```
p3+p4
```



La dimension 1 va séparer les pays qui se tournent plus vers le renouvelable que les autres grâce à la variable *Rnw.energy.share* et les gros pays qui produisent beaucoup d'énergie et émettent beaucoup de CO2 surtout à gauche avec *Rnw.share*. La dimension 2 va séparer les pays qui sont plus tournés plus vers la qualité de vie des habitants avec un bon accès à l'électricité, une consommation personnelle d'energie élevée et un bon PIB par habitant et les pays tourné vers la production d'électricité. La dimension 3 va permettre de repérer les pays qui ont des politiques plus durables tourné vers les énergies renouvelables et la richesse de ses habitants.

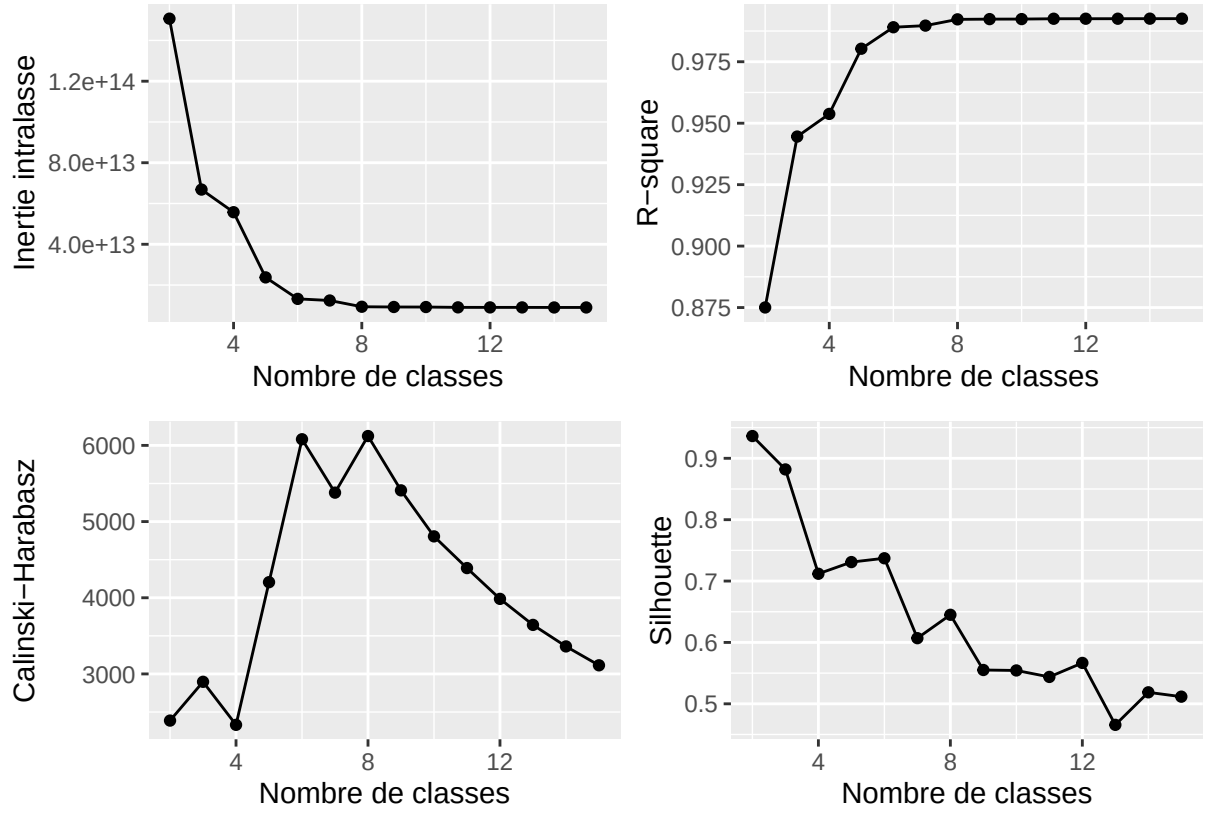
IV. Clustering

K-means

```
#Choix du nombre de classe pour les K-means
#Paramètres
Kmax<-15
reskmeanscluster<-matrix(0,nrow=nrow(renew),ncol=Kmax-1)
d<-dist(renew[,-c(8)],method="euclidean")
#Indicateurs
Iintra<-NULL
R2<-NULL
CH<-NULL
Silhou<-NULL
for (k in 2:Kmax){
  resaux<-kmeans(renew[1:7],k)
  #Inertie intra-classe
  Iintra<-c(Iintra,resaux$tot.withinss)
  #Silouhette
  aux<-index.S(d,resaux$cluster)
  Silhou<-c(Silhou,aux)
  #R-square
  R2<-c(R2,1-(resaux$tot.withinss/resaux$totss))
  #Calinsky-Harabasz
  CH<-c(CH,index.G1(renew[,-c(8)],resaux$cluster))
```

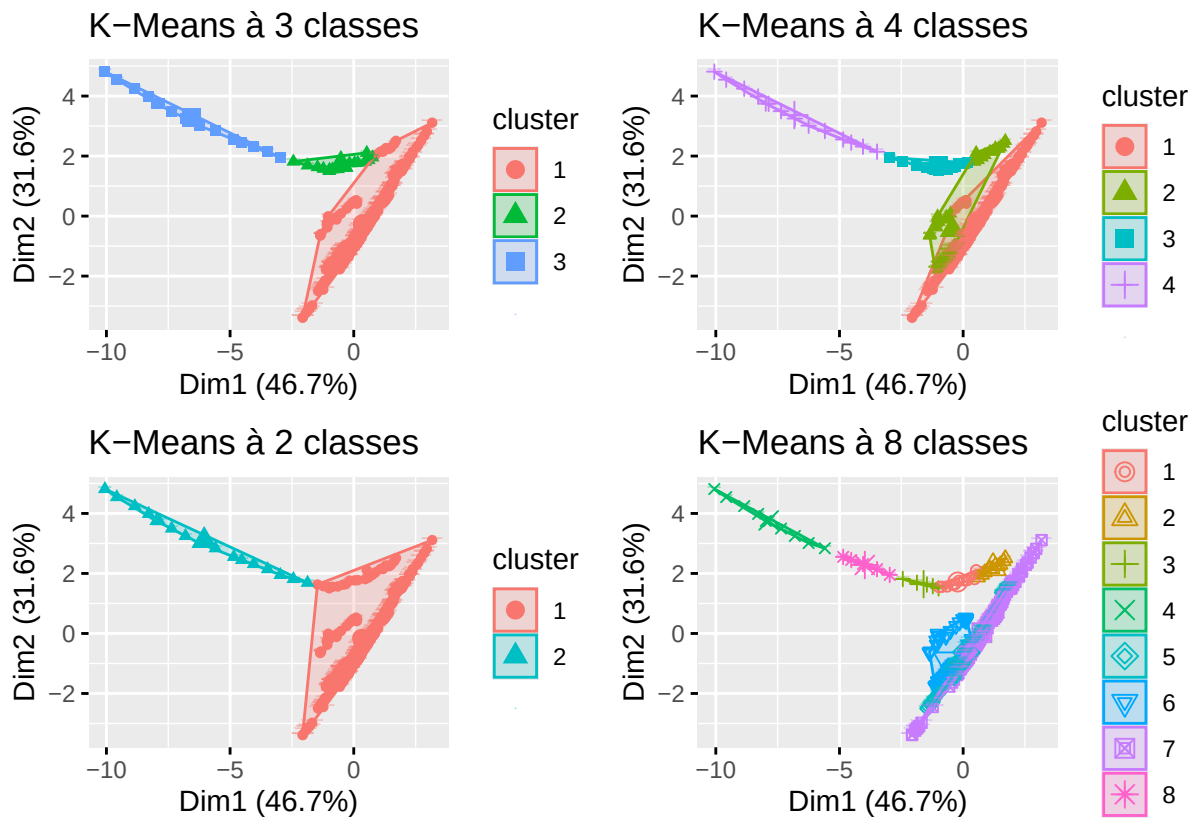
```
  reskmeanscluster[,k-1]<-resaux$cluster
}
#Inertie intra-classe
df<-data.frame(K=2:15,Iintra=Iintra)
plot1<-ggplot(df,aes(x=K,y=Iintra))+
  geom_line()+
  geom_point()+
  xlab("Nombre de classes")+
ylab("Inertie intralasse")
#Indicateur R-square
df2<-data.frame(K=2:15,R2=R2)
plot2<-ggplot(df2,aes(x=K,y=R2))+
  geom_line()+
  geom_point()+
  xlab("Nombre de classes")+
ylab("R-square")
#Calinski-Harabasz
df3<-data.frame(K=2:15,CH=CH)
plot3<-ggplot(df,aes(x=K,y=CH))+
  geom_line()+
  geom_point()+
  xlab("Nombre de classes")+
ylab("Calinski-Harabasz")
#Silhouette
df4<-data.frame(K=2:Kmax,Silhouette=Silhou)
plot4<-ggplot(df4,aes(x=K,y=Silhouette))+
geom_point()+
geom_line()+theme(legend.position = "bottom")+
xlab("Nombre de classes")+
ylab("Silhouette")

(plot1 + plot2) / (plot3 +plot4)
```



Silhouette nous dit de prendre 2 classes, Calinski-Harabasz nous dit d'en prendre 8, Iintra et R-square nous dit de prendre plutot 3 ou 4 classes. On va essayer de faire des clusterings K-Means avec 2, 3, 4 et 8 classes.

```
#Calcul de la méthode des K-means avec différentes valeurs de K
res3means1<-kmeans(renew[1:7],3,algorithm="Hartigan-Wong")
res4means<-kmeans(renew[1:7],4,algorithm="Hartigan-Wong")
res2means<-kmeans(renew[1:7],2,algorithm="Hartigan-Wong")
res8means<-kmeans(renew[1:7],8,algorithm="Hartigan-Wong")
#Affichage
p1<-fviz_cluster(res3means1,data=renew[1:7],labelsize = 1,main="K-Means à 3 classes")
p2<-fviz_cluster(res4means,data=renew[1:7],labelsize = 1,main="K-Means à 4 classes")
p3<-fviz_cluster(res2means,data=renew[1:7],labelsize = 1,main="K-Means à 2 classes")
p4<-fviz_cluster(res8means,data=renew[1:7],labelsize = 1,main="K-Means à 8 classes")
(p1 + p2 )/(p3+p4)
```



Pour 4 classes et 8 classes, on arrive pas à trouver un plan factoriel dans lequel chaque classe est bien définie/distincte des autres, on ne va donc pas conserver ces clustering.En essayant avec 3 classes, on obtient le même résultat pour les différentes méthodes de K-means (ARI=1). Choisir 3 classes donne un résultat assez satisfaisant et interprétable. Avec 2 classes,on a un résultat bien distinct mais qui fournit sûrement moins d'informations pour être exploité.

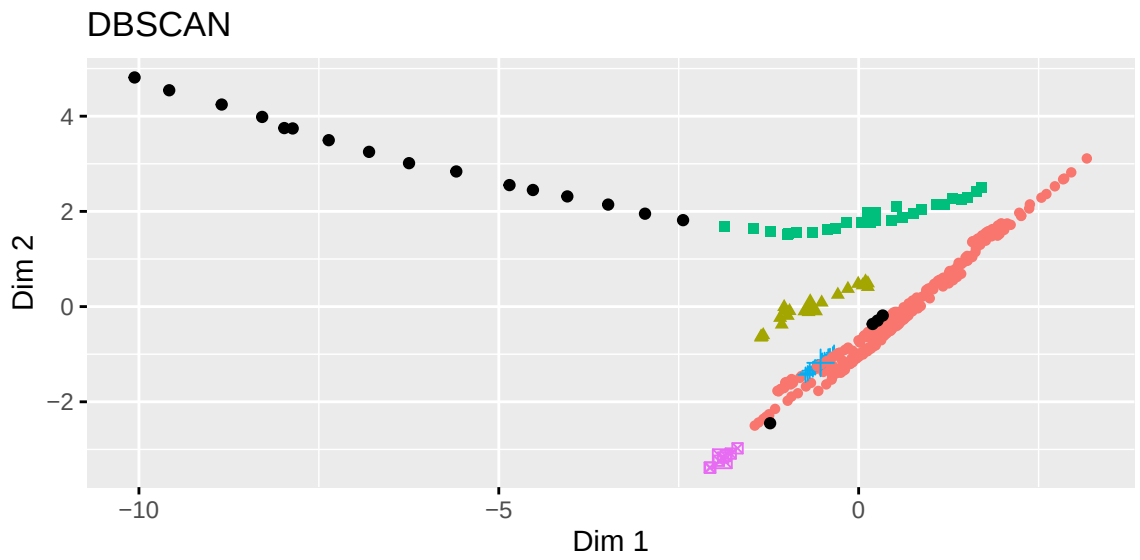
```
table(res3means1$cluster,renew$Continent)
```

```
##
##      Africa America Asia Europe
##   1      55     114  132     10
##   2       0       0   17      0
##   3       0       0   15      0
```

On essaye de voir si il peut y avoir une corrélation entre les classes et les continents, cela ne semble pas être le cas pour 3 classes avec la table qu'on obtient ci-dessus. On voit que le clustering sépare la Chine moderne (après 2005) dans une classe, l'Inde moderne (après 2008) et le reste des données chinoises et dans la classe restante, tout le reste. On en conclut qu'on a une classe avec les pays développés (qui ne sont maintenant plus en développement comme les autres), une classe avec les pays en bonne voie de développement et les pays qui sont encore vraiment en développement.

DBSCAN

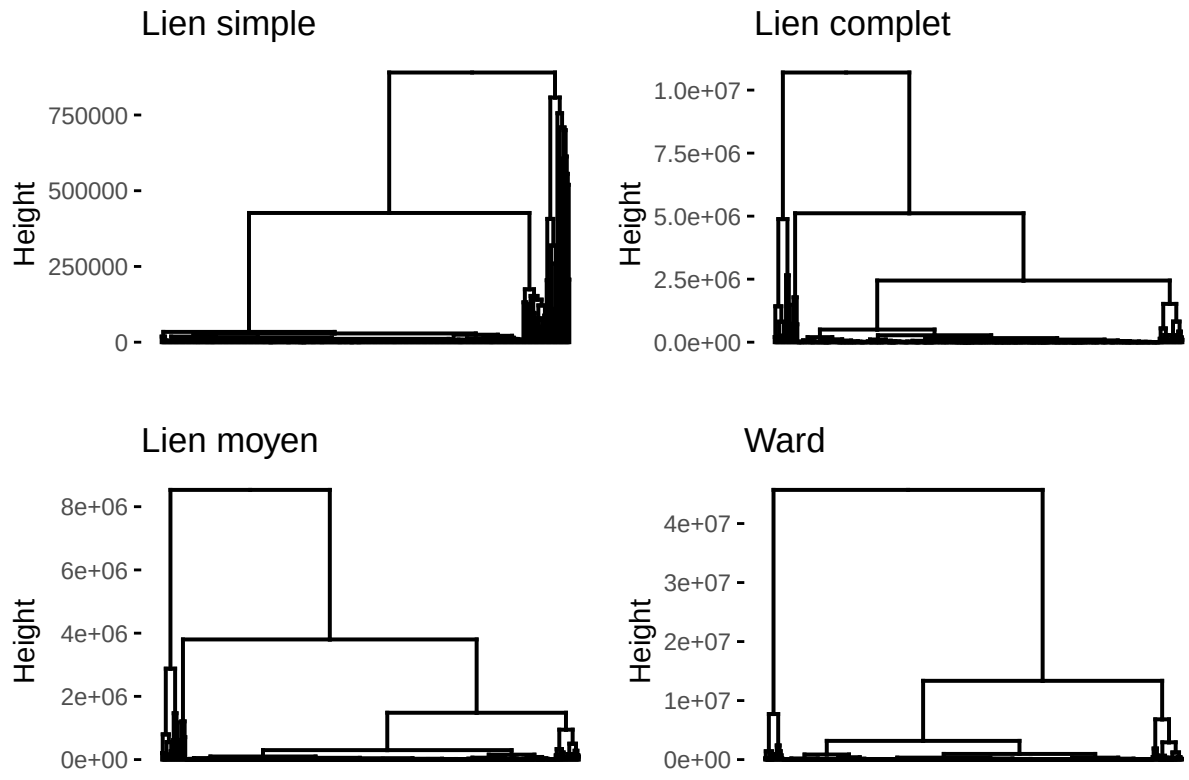
```
#Choix des paramètres
minPts<-4
eps<-0.75
#Calcul et affichage du clustering
resdbscan<-dbscan(renew[1:7],eps,minPts,scale=TRUE)
pdb<-fviz_cluster(resdbscan,renew[,-c(8)],axes=c(1,2), geom="point",ellipse="FALSE")+
theme(legend.position="none")+
xlab("Dim 1")+ylab("Dim 2")+ggtitle("DBSCAN")
pdb
```



Beaucoup de mal à trouver de bonnes valeurs de paramètres pour que les classes semblent à minima cohérentes/interprétables sur le premier plan factoriel. K-means semble beaucoup plus efficace pour ce jeu de données. La méthode DBSCAN qui est plus utilisé pour de la reconnaissance de forme n'est pas adapté dans notre contexte, il est donc normal de ne pas trouver de résultat satisfaisant.

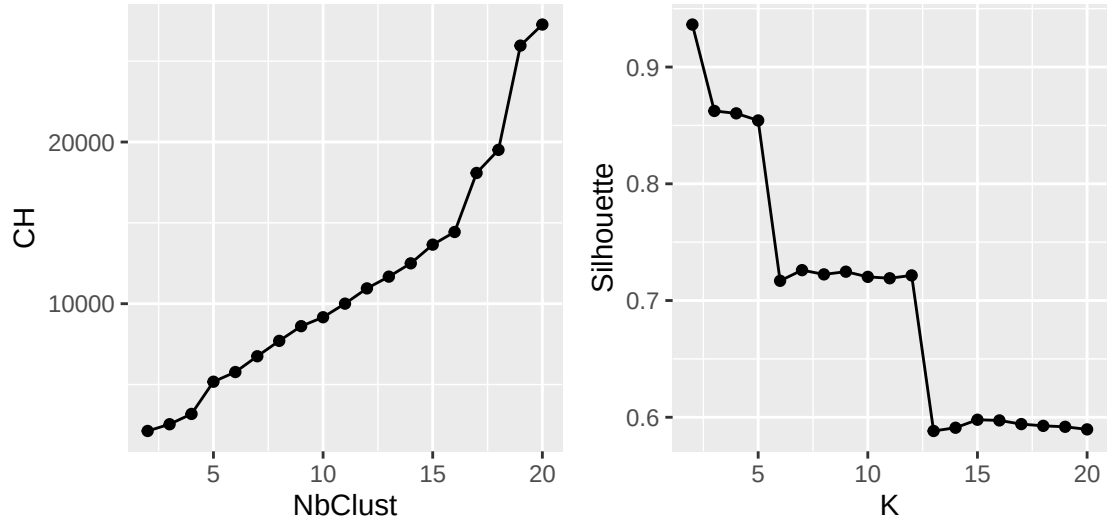
Classification ascendante hiérarchique.

```
#Choix de la méthode d'aggrégation
hsingle<-hclust(d,method="single")
hcomplete<-hclust(d,method="complete")
haverage<-hclust(d,method="average")
hward<-hclust(d,method="ward.D2")
#Affichage des dendrogrammes
p1<-fviz_dend(hsingle,show_labels = FALSE,main="Lien simple")
p2<-fviz_dend(hcomplete,show_labels = FALSE,main="Lien complet")
p3<-fviz_dend(haverage,show_labels = FALSE,main="Lien moyen")
p4<-fviz_dend(hward,show_labels = FALSE,main="Ward")
(p1+p2)/(p3+p4)
```



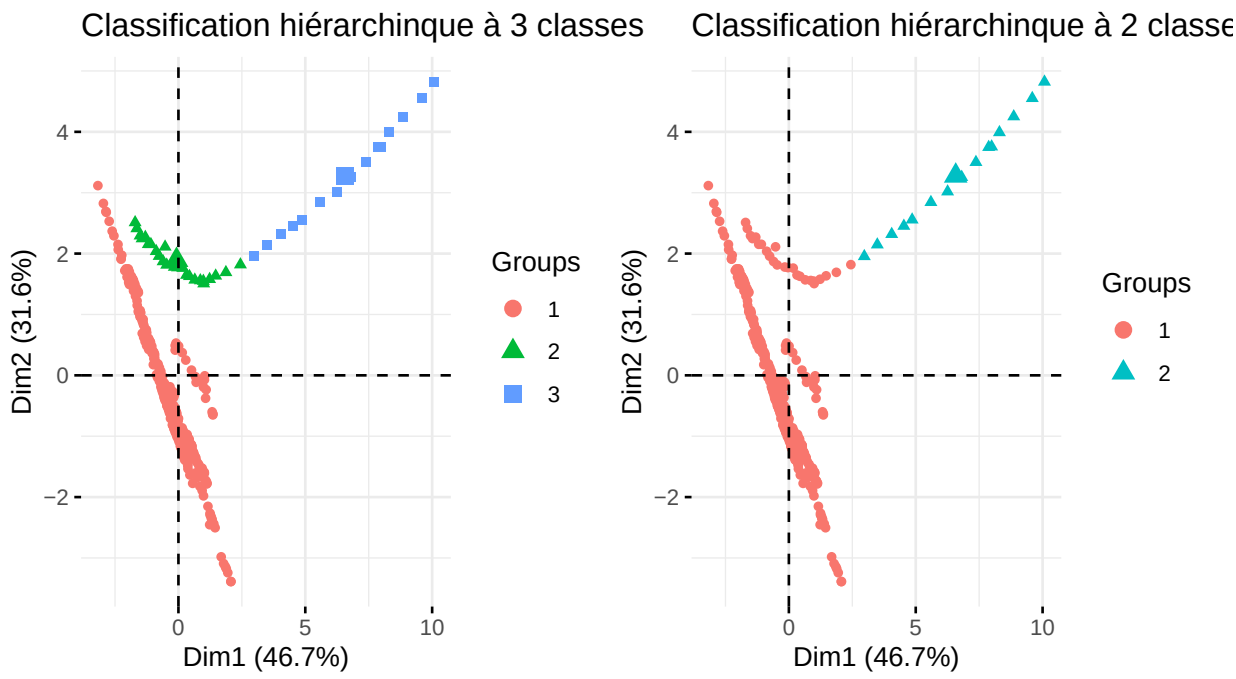
La méthode du lien simple donne un dendrogramme écrasé sur la droite et donc des classes de tailles très différentes. Celles du lien complet et moyen donnent des résultats sensiblements équivalents, toujours resserés mais sur la gauche cette fois-ci. La méthode de Ward donne un résultat moins écrasé que les autres et avec des classes de tailles plus proches. On va donc conserver la méthode de Ward.

```
#Indicateurs pour le découpage du dendogramme
CH<-NULL
Silhou<-NULL
Kmax=20
for (k in 2:Kmax){
clusters=cutree(hward, k=k)
#Calinsky-Harabasz
CH<-c(CH,index.G1(renew[,-c(8)], clusters,d,centrotypes="medoids"))
#Silhouette
Silhou<-c(Silhou,index.S(d, clusters))
}
daux<-data.frame(NbClust=2:Kmax,CH=CH)
df<-data.frame(K=2:Kmax,Silhouette=Silhou)
#Affichage
p2<-ggplot(df,aes(x=K,y=Silhouette))+
geom_point()+
geom_line()+theme(legend.position = "bottom")
p1<-ggplot(daux,aes(x=NbClust,y=CH))+geom_line()+geom_point()
p1+p2
```



Résultat vraiment bizarre, CH ne fait qu'augmenter et Silhouette max à 2. On devrait normalement avoir CH qui tend vers 0. On ne peut pas déterminer graphiquement où couper le dendrogramme à cause de l'échelle. Graphiquement,sur le dendrogramme, on observe bien que la plus grande diminution de la mesure de Ward s'effectue quand on passe de 1 à 2 classes. On va quand même continuer à regarder pour 3 classes comme pour les K-means.

```
#Découpage du dendrogramme et affichage des clusterings
ClustH3<-cutree(hward,3)
p1<-fviz_pca_ind(res.ACP,cex=0.5,habillage=as.factor(ClustH3),label="none",
                 title="Classification hiérarchinque à 3 classes")
ClustH2<-cutree(hward,2)
p2<-fviz_pca_ind(res.ACP,cex=0.5,habillage=as.factor(ClustH2),label="none",
                 title="Classification hiérarchinque à 2 classes")
p1+p2
```



Le clustering en 3 classes semble très proche de celui obtenu par la méthode des K-means avec 3 classes également (on va s'intéresser au lien entre les résultats obtenu avec les deux méthodes juste après). Le clustering en 4 classes et sensiblement le même que le clustering en 3 classes à une chose près, on a scindé une classe en deux. Au vu des interprétations obtenues via la méthode des K-means, on va garder un clustering en 3 classes et la même conclusion.

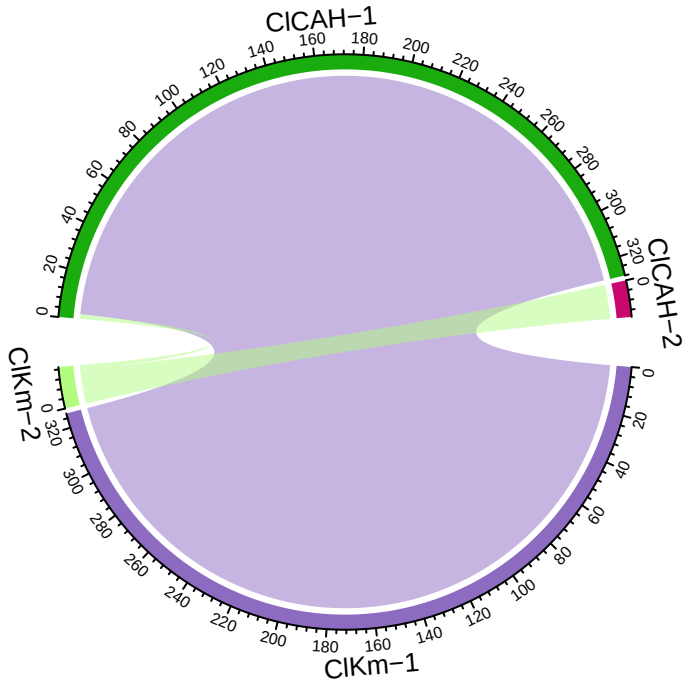
```
#Comparaison K-means/CAH
table(res3means1$cluster,ClustH3)
```

```
##    ClustH3
##       1   2   3
##   1 303   8   0
##   2   0  17   0
##   3   0   0  15
```

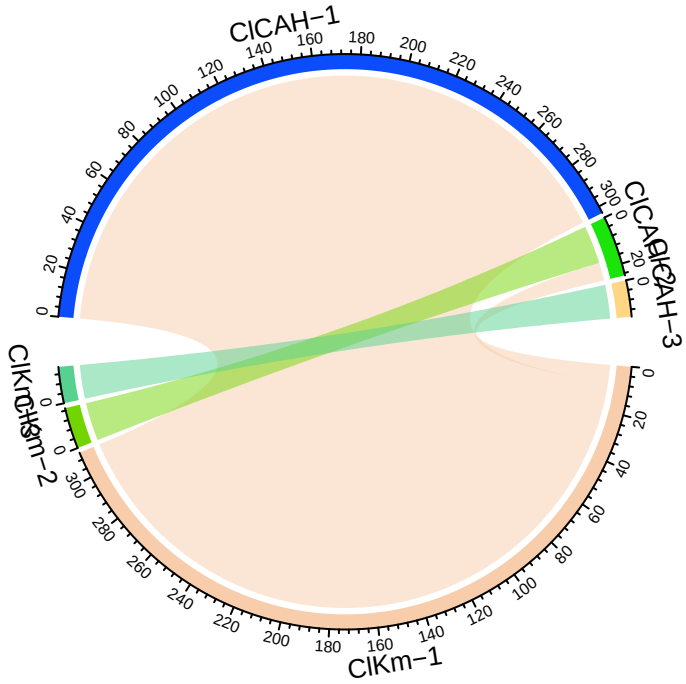
```
adjustedRandIndex(res3means1$cluster,ClustH3)
```

```
## [1] 0.8605144
```

```
clust21<-paste("ClKm-",res2means$cluster,sep="")
clust22<-paste("ClCAH-",cutree(hward,2),sep="")
clust31<-paste("ClKm-",res3means1$cluster,sep="")
clust32<-paste("ClCAH-",cutree(hward,3),sep="")
chordDiagram(table(clust21,clust22))
```




```
chordDiagram(table(clust31,clust32))
```



On observe que les clustering de 2 et 3 classes obtenues par méthode des Kmeans et classification hiérarchique sont quasi semblables (ARI proche de 1 et graphique). On tire donc les mêmes conclusions que pour la

méthode des K-means.

V. Analyse critique du jeu de données

En effectuant les premières manipulations, nous nous sommes rapidement rendus compte que les données concernant la Chine et l'Inde avaient beaucoup de poids dans les calculs et avaient donc beaucoup plus d'impact que les autres pays (ACP, clustering) car les valeurs sont plus élevées. Nous nous sommes donc posé la question : "Quels seraient les résultats si on enlevait la Chine et l'Inde du jeu de données ?" D'un point de vue des variables étudiées indépendamment des autres, on observe beaucoup moins d'outliers sur les boxplots. En réalisant la matrice de corrélation linéaire, on a plus qu'une seule corrélation, celle entre les émissions de CO2 et la production d'énergie fossile (ce à quoi on pouvait triviallement s'attendre). Les deux autres ont disparu, mais comme on l'avait dit, elles étaient dûes à des pays très peuplés (Chine et Inde). En enlevant la Chine et l'Inde, il ne nous reste que des pays vraiment en développement, beaucoup plus homogène. Ainsi, l'ACP a eu du mal à deceler des particularités, les données sont très centrées et les composantes principales très difficiles à interpréter (beaucoup de combinaisons de variables). De même pour le clustering qui marche également beaucoup moins bien faute d'hétérogénéité.