Mini Projet

RAMDÉ Ismaïl

10, avril, 2021

Chargement des données

```
prix_nobel <- read.csv("prixnobel.csv", header=TRUE, sep=";",fileEncoding="latin1", row.names=1)
```

1. Statistique descriptive de base

Aperçu et visualisation des données

- Aperçu

```
# Visualisation
str(prix_nobel)
```

```
## 'data.frame':    13 obs. of  7 variables:
##  $ Chimie       : int  24 4 8 23 1 6 4 51 1 56 ...
##  $ Economie     : int  1 3 3 6 1 0 3 43 0 47 ...
##  $ Littérature  : int  8 2 11 7 6 2 5 8 5 18 ...
##  $ Médecine     : int  18 4 12 26 5 3 2 70 3 78 ...
##  $ Paix         : int  5 1 10 11 1 1 3 19 8 25 ...
##  $ Physique     : int  24 4 9 20 5 11 10 66 2 70 ...
##  $ Mathématiques: int  1 1 11 4 1 3 9 13 1 15 ...
```

```
# Affichage de quelques lignes
head(prix_nobel)
```

```
##           Chimie Economie Littérature Médecine Paix Physique Mathématiques
## Allemagne     24        1           8       18    5       24             1
## Canada         4        3           2        4    1        4             1
## France         8        3          11       12   10        9            11
## GB            23        6           7       26   11       20             4
## Italie         1        1           6        5    1        5             1
## Japon          6        0           2        3    1       11             3
```

```
# Résumé
summary(prix_nobel)
```

```
##      Chimie         Economie      Littérature       Médecine
##  Min.   : 1.00   Min.   : 0.00   Min.   : 0.00   Min.   :  2.00
##  1st Qu.: 4.00   1st Qu.: 1.00   1st Qu.: 5.00   1st Qu.:  4.00
##  Median : 8.00   Median : 3.00   Median : 7.00   Median :  9.00
##  Mean   :22.46   Mean   :10.38   Mean   :12.38   Mean   : 26.69
##  3rd Qu.:24.00   3rd Qu.: 6.00   3rd Qu.:10.00   3rd Qu.: 26.00
##  Max.   :94.00   Max.   :47.00   Max.   :79.00   Max.   :110.00
##       Paix          Physique      Mathématiques
##  Min.   : 0.00   Min.   :  2.00   Min.   : 1.000
##  1st Qu.: 1.00   1st Qu.:  5.00   1st Qu.: 1.000
```

```
##  Median : 8.00   Median : 11.00   Median : 4.000
##  Mean   :11.62   Mean   : 26.54   Mean   : 7.846
##  3rd Qu.:16.00   3rd Qu.: 24.00   3rd Qu.:11.000
##  Max.   :51.00   Max.   :103.00   Max.   :34.000
```

La pemière impression qu'on a de notre jeu de données est qu'elle est composée essentiellement de variables quantitatives (7 variables) et de 13 observations (pays/continents). A travers le résumé des différentes variables, on constate qu'il n'y a aucune valeur manquante (NA).

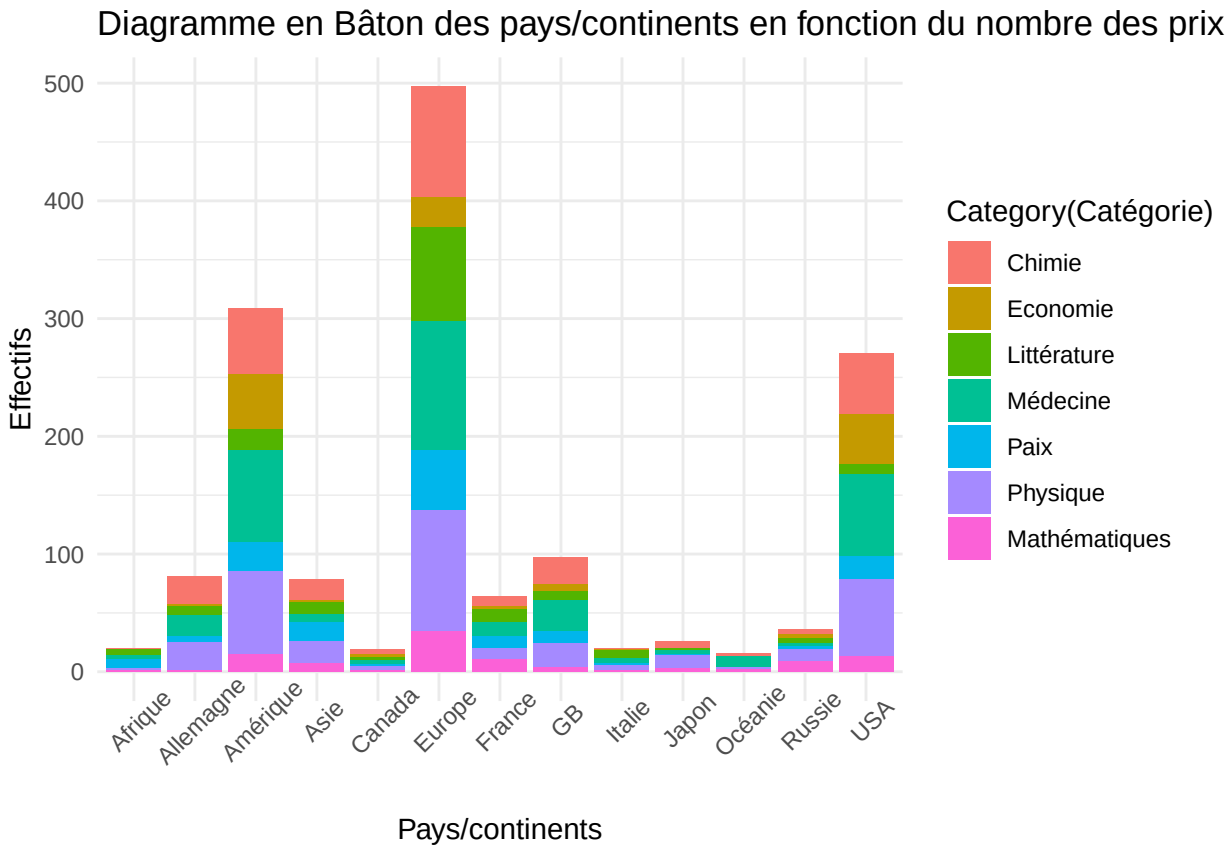
Notre base de données semble être propre, donc pas de nettoyage a apporter de notre part. Elle est prête a être utilisée.

- Visualisation

```
# charger / générer nos données
prix_nobel <- data.frame(prix_nobel)
prix_nobel$Category  <- row.names(prix_nobel)
# Transformation des données au format selon les besoins pour ggplot
pnobel <- melt(prix_nobel, value.name="Count", variable.name="Variable", na.rm=TRUE)
head(pnobel)
```

```
##    Category Variable Count
## 1 Allemagne   Chimie    24
## 2    Canada   Chimie     4
## 3    France   Chimie     8
## 4        GB   Chimie    23
## 5    Italie   Chimie     1
## 6     Japon   Chimie     6
```

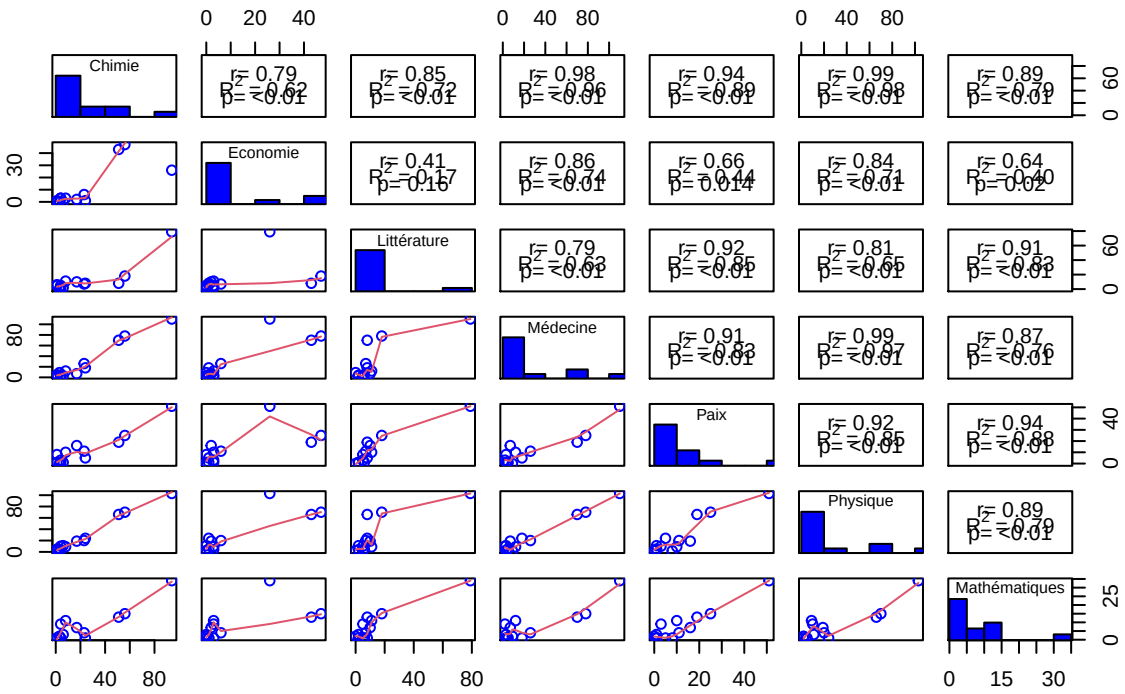
```
p<-ggplot(pnobel, aes(x=pnobel$Category, y=Count, fill=Variable)) +
  geom_bar(stat="identity",aes(x=pnobel$Category))+theme_minimal()+
  theme(axis.text.x = element_text(angle=45))+
  ggtitle("Diagramme en Bâton des pays/continents en fonction du nombre des prix") +
  xlab("Pays/continents") +
  ylab("Effectifs") +
  labs(fill = "Category(Catégorie)")
p
```



Ce digramme nous permet de visualiser l'effectif des différents prix obtenus en fonction de chaque pays/continent. On remarque que de façon générale l'Europe est le continent qui a obtenu le plus grand nombre de prix tant-disque l'Océanie et l'Afrique totalisent les plus petits effectifs. En ce qui concerne les pays ce sont les USA qui ont obtenus le plus de prix tant-disque le Canada a le plus petit nombre de prix.

Matrice de scatter plot

```
# Retour à la base de donnée initiale après l'après l'ajout de la variable "Category"
prix_nobel <- subset(prix_nobel, select = -Category)
# matrice scatterplot et test
pairs(prix_nobel, pch = 1, lower.panel=panel.smooth, upper.panel=panel.cor,diag.panel=panel.hist, col =
```



Cette matrice nous montre qu'il y a de fortes corrélations s entre les différentes variables deux à deux. On voit aussi que leurs distributions en diagonale indiquent qu'elles ne suivent pas une loi normale, d'où la nécessité plus tard de normaliser (centrer et réduire) les données avant l'ACP, CAH et K-Means.

2. Analyse en Composantes Principales (ACP)

- Normalisation des données et calculer l'ACP sur les individus/variables

```
res.pca <- PCA(prix_nobel, graph = FALSE)
```

- Valeurs propres et la proportion de variances

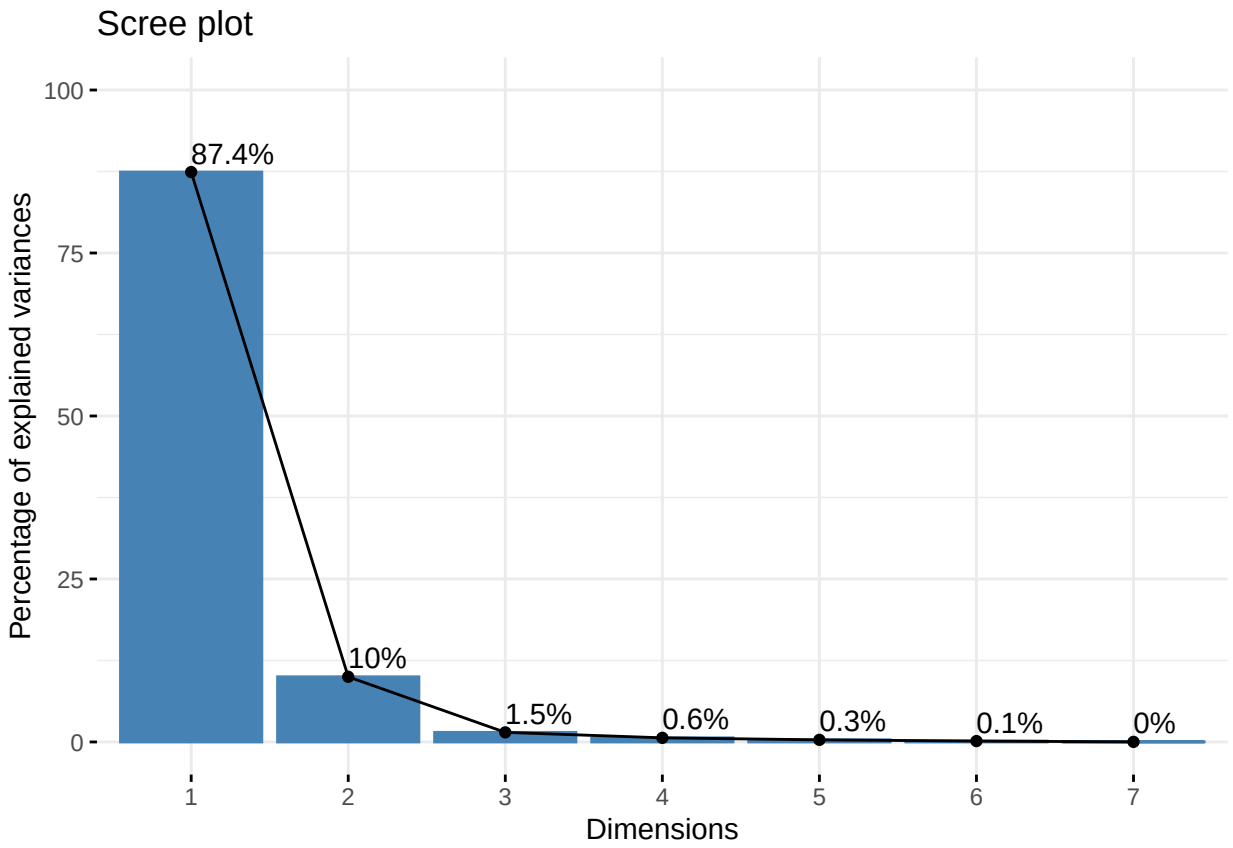
```
eig.val <- get_eigenvalue(res.pca)
eig.val
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1 6.119019424      87.41456320                     87.41456
## Dim.2 0.699650711       9.99501016                     97.40957
## Dim.3 0.103167301       1.47381859                     98.88339
## Dim.4 0.044439232       0.63484617                     99.51824
## Dim.5 0.022895499       0.32707856                     99.84532
## Dim.6 0.009784213       0.13977448                     99.98509
## Dim.7 0.001043619       0.01490885                    100.00000
```

On voit qu'environ 97.40957% de la variance totale est expliquée par les deux premières valeurs propres. Les deux premières composantes principales expliquent 97.40957% de la variation.

- Graphique des valeurs propres

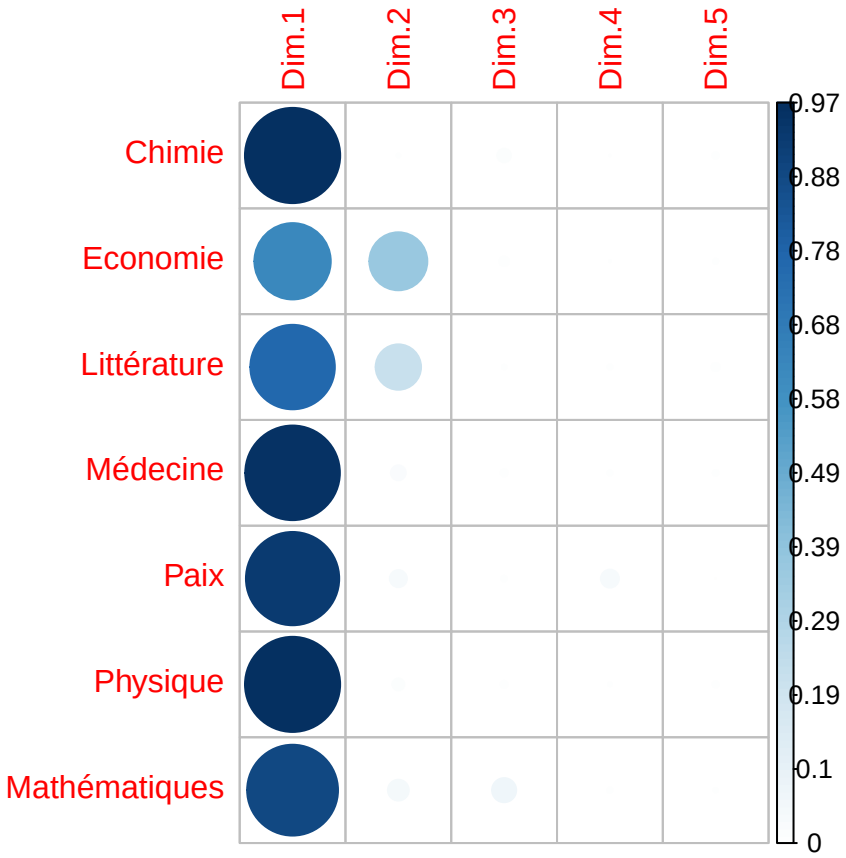
```
fviz_eig(res.pca, addlabels = TRUE, ylim = c(0, 100))
```



A travers ce graphe on voit bien que les deux premières composantes contiennent suffisamment d'information (97.40957% de la variance totale).

- Qualité de représentation

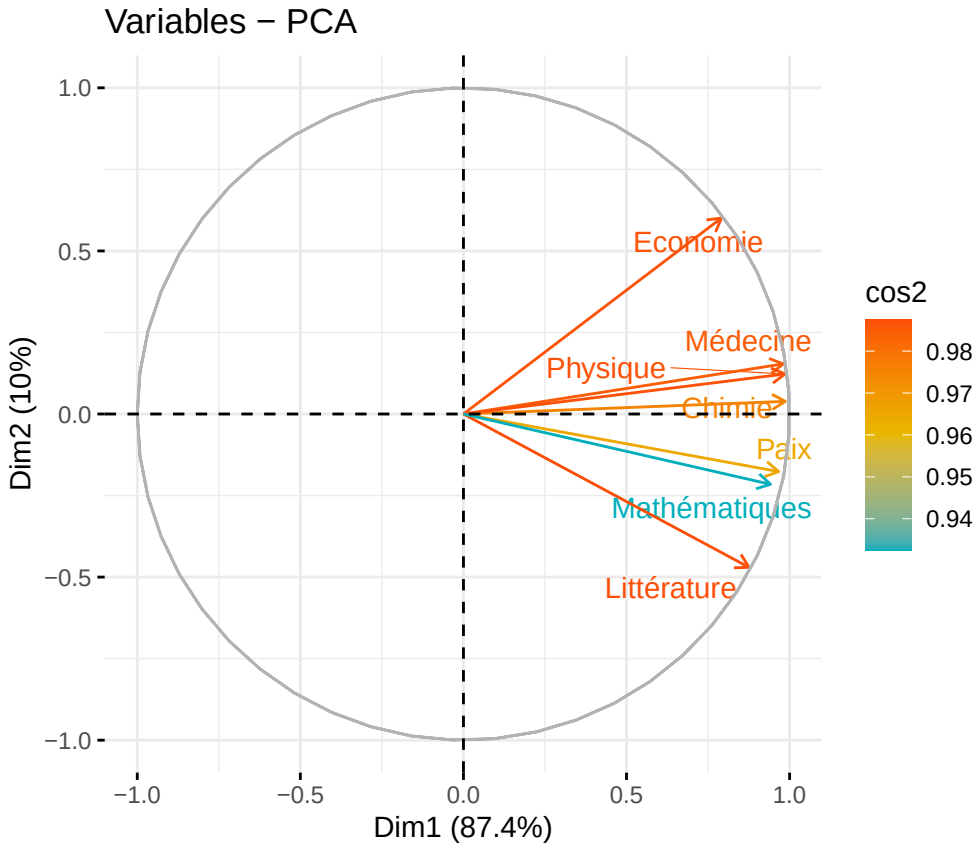
```
# Extraction des résultats, pour les variables, à partir de l'ACP
var <- get_pca_var(res.pca)
# visualiser le cos2 des variables sur toutes les dimensions
corrplot(var$cos2, is.corr=FALSE)
```



On remarque que sur la Dimension 1, les variables Chimie, Médecine, Paix, Physique et Mathématiques sont très biens représentées tan-disque les variables Economie et Littérature biens représentées. Et sur la Dimension 2 les variables Economie et Littérature sont faiblements représentées.

- Visualisons les variables

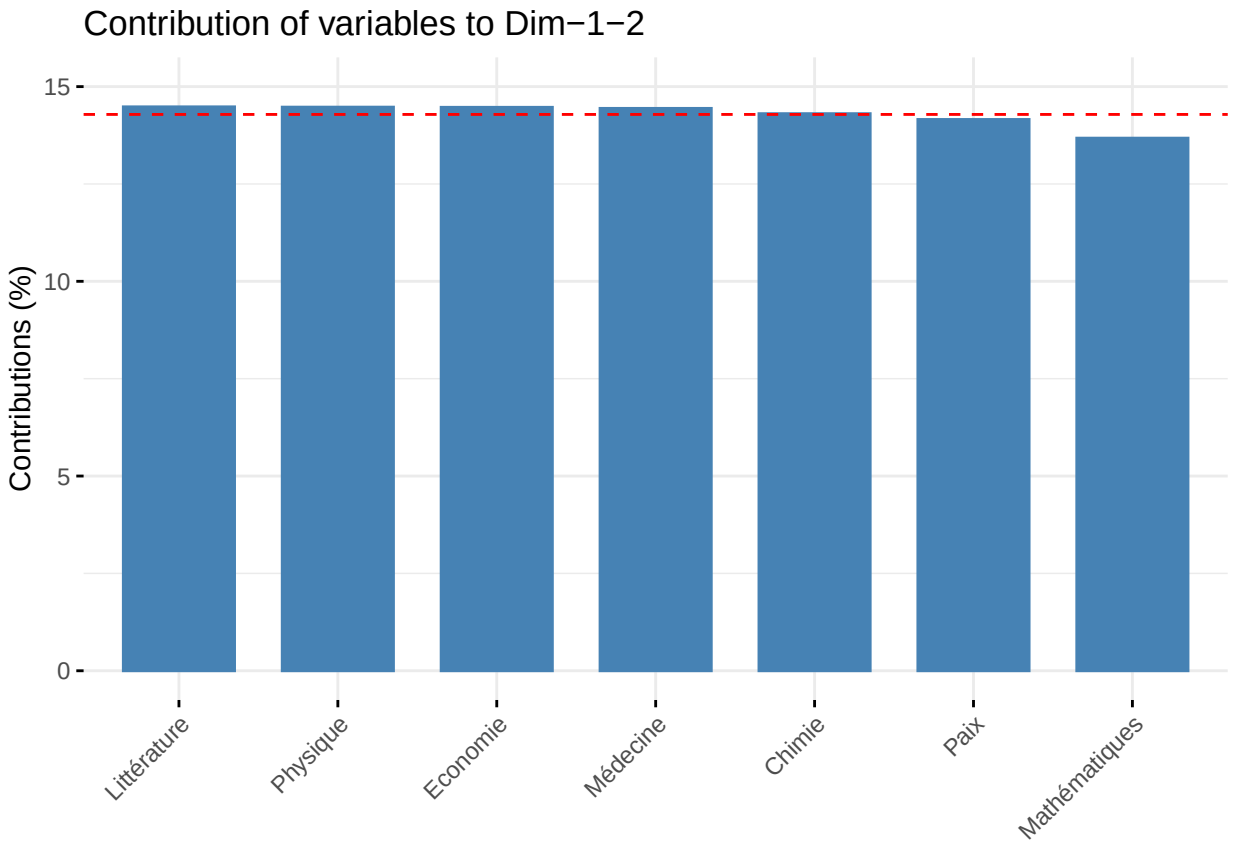
```
fviz_pca_var(res.pca, col.var = "cos2",
gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
repel = TRUE) #repel = TRUE évite le chevauchement de texte
```



On voit sur ce graphe que toutes les variables sont biens représentées car elles sont proches du cercle de corrélation. Elles sont corrélées positivement.

Contribution des variables

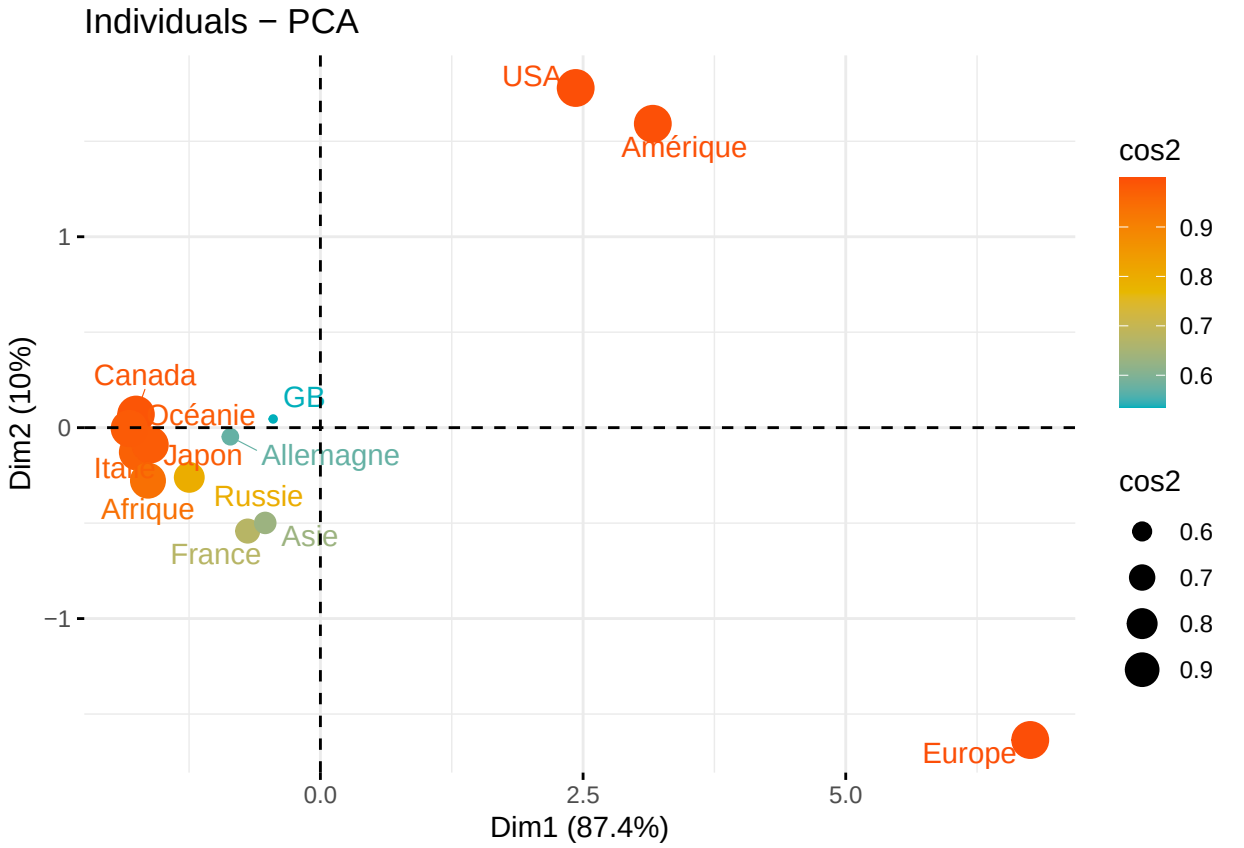
```
fviz_contrib(res.pca, choice = "var", axes = 1 :2, ylim = c(0, 15))
```



Globalement toutes les variables ont une bonne contribution aux deux dimensions. A part la variable Mathématiques qui est légèrement en-dessous des autres.

- Qualité et contribution des individus

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



On remarque d'une part en rouge les individus qui on une bonne contribution (USA, Europe, Canada, ...) et au fur à mesure que l'on descend en couleur la contribution diminue aussi, c'est le cas par exemple de la GB et l'Allemagne qui n'ont pas une bonne contribution.

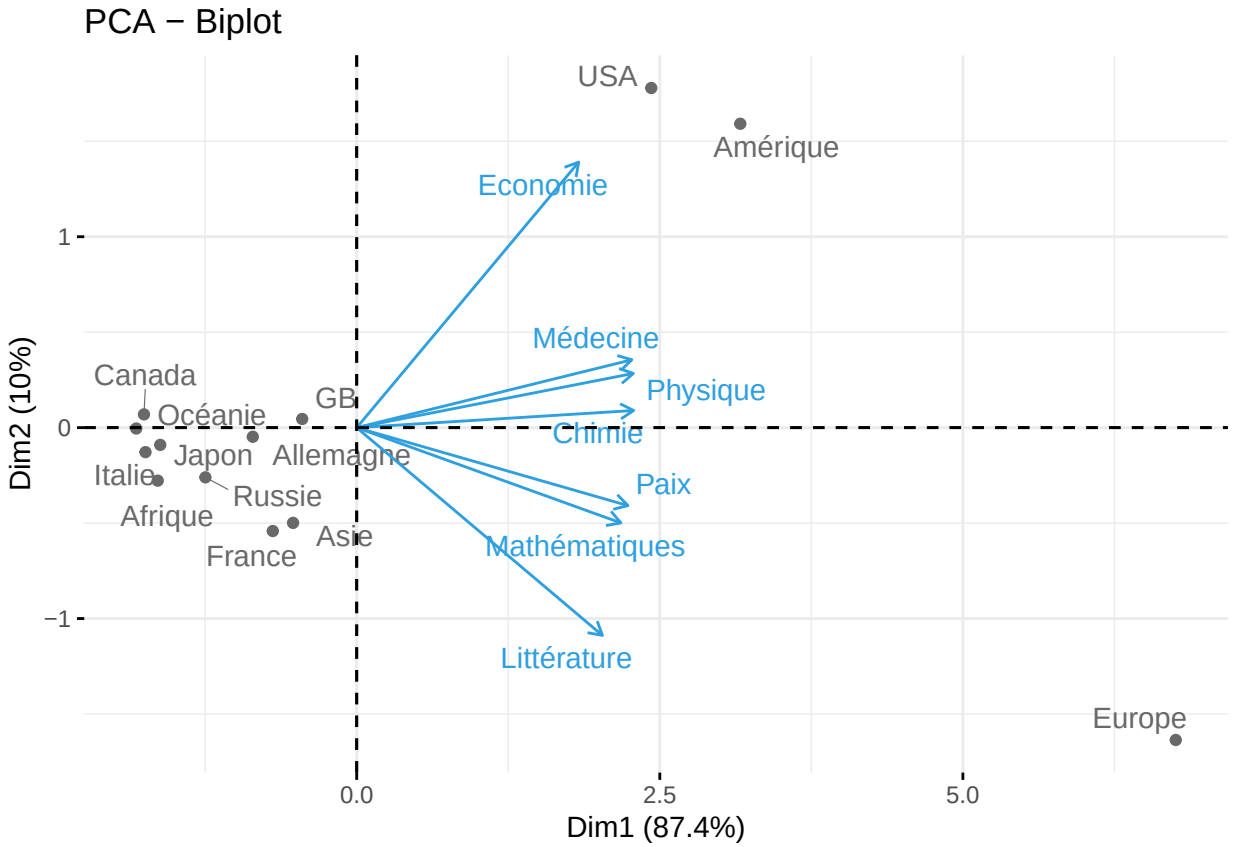
Et d'autre part des proximités entre certains individus :

- les USA et l'Amérique
- Le Canada, l'Océanie, le Japon, l'Italie et L'Afrique
- La France et l'Asie et la Russie
- La GB et l'Allemagne

Notons également un individus atypique (l'Europe) qui s'écarte fortement des autres.

- Créons un biplot des individus et des variables

```
fviz_pca_biplot(res.pca, repel = TRUE, col.var = "#2E9FDF", col.ind = "#696969")
```



On constate que les USA, l'Amérique et l'Europe sont les individus qui ont des grandes valeurs pour toutes variables (différents prix Nobel). Les autres variables ont des valeurs relativement faibles.

3. Classification Automatique CAH et K-Means

- Classification Ascendante Hiérarchique

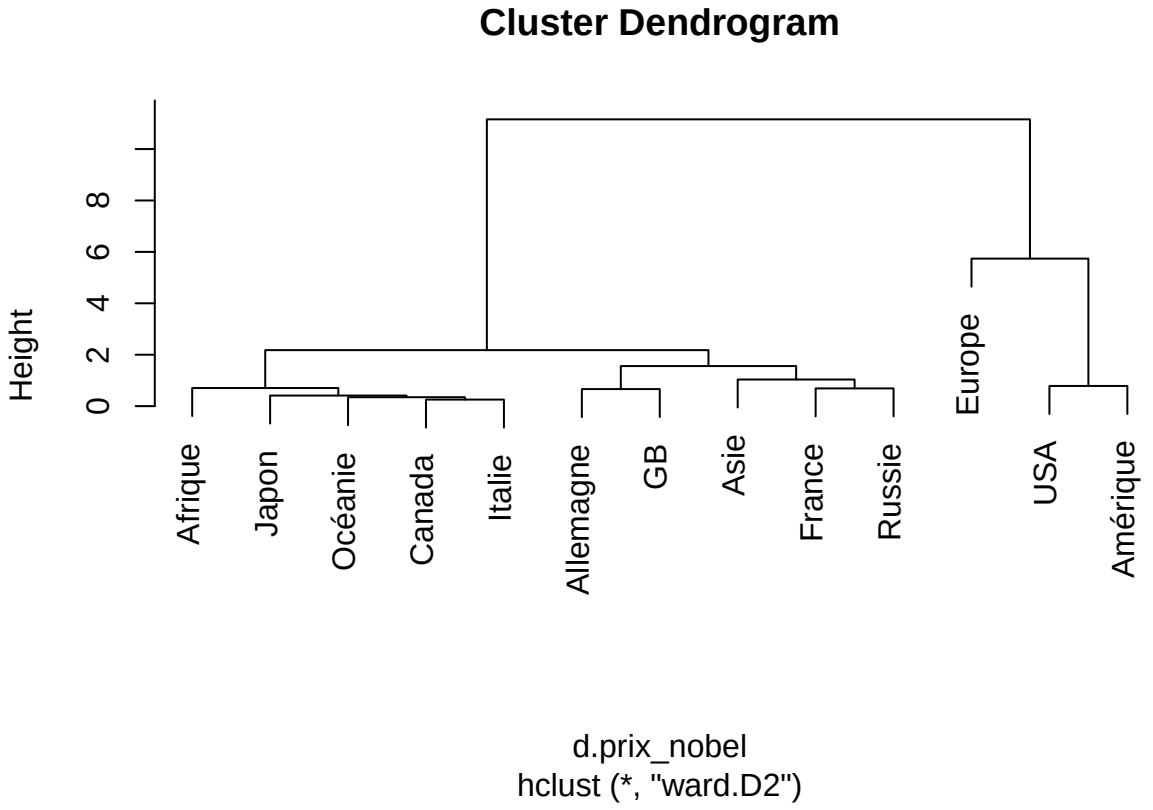
```
# hclust

# centrage réduction des données
# pour éviter que les variables à forte variance pèsent indûment sur les résultats
prix_nobel.cr <-scale(prix_nobel,center=T,scale=T)

# matrice des distances entre individus
d.prix_nobel<-dist(prix_nobel.cr)

# CAH -critère de Ward
# method= «ward.D2» correspond au vrai critère de Ward
# utilisant le carré de la distance
cah.ward<-hclust(d.prix_nobel,method="ward.D2")

# affichage du dendrogramme
plot(cah.ward)
```



De façon automatique notre dendrogramme suggère 4 groupes.

```
# découpage en 4 groupes
groupes.cah <-cutree(cah.ward,k=4)

# liste des groupes
print(sort(groupes.cah))
```

```
## Allemagne    France        GB    Russie      Asie    Canada    Italie     Japon
##         1         1         1         1         1         2         2         2
##   Afrique   Océanie       USA  Amérique    Europe
##         2         2         3         3         4
```

- Méthode des centres mobiles (K-means)

Nous allons a présent tenter de d'améliorer notre modèle, c'est à dire trouver le nombre de groupe optimal tout en nous basant sur les résultats du découpage de la méthode automatique CAH.

```
# K-means avec les données centrées et réduites
# center = 4 : nombre de groupes demandés
# nstart = 5 : nombre d'essais avec différents individus de départ
groupes.kmeans <- kmeans(prix_nobel.cr,centers=4,nstart=5)
# Affichage des résultats
print(groupes.kmeans)
```

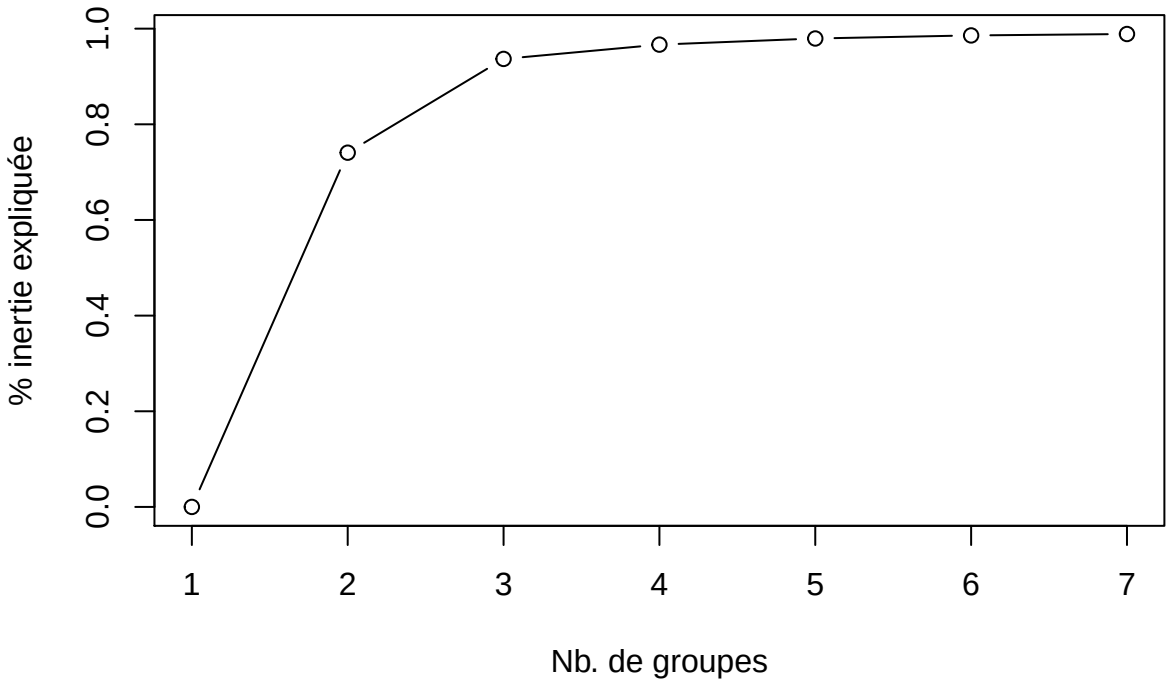
```
## K-means clustering with 4 clusters of sizes 2, 1, 4, 6
##
## Cluster means:
##       Chimie   Economie Littérature   Médecine        Paix   Physique
## 1  1.1001734  2.0561347  0.02995919  1.3310379  0.73261449  1.2899524
```

```
## 2  2.5357156  0.9275452  3.24308178  2.3439252  2.77850830  2.3788733
## 3 -0.1581414 -0.4386421 -0.16477552 -0.3078701 -0.07868822 -0.2656488
## 4 -0.6839161 -0.5475410 -0.44064968 -0.6290867 -0.65483073 -0.6493638
##   Mathématiques
## 1     0.6630429
## 2     2.8179322
## 3    -0.2258490
## 4    -0.5401037
##
## Clustering vector:
## Allemagne    Canada    France        GB    Italie     Japon    Russie       USA
##         3         4         3         3         4         4         4         1
##   Afrique  Amérique      Asie    Europe   Océanie
##         4         1         3         2         4
##
## Within cluster sum of squares by cluster:
## [1] 0.308316 0.000000 1.492824 1.008898
##  (between_SS / total_SS =  96.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# les correspondances des groupes entre CAH et K-Means :
print(table(groupes.cah,groupes.kmeans$cluster))
```

```
##
## groupes.cah 1 2 3 4
##           1 0 0 4 1
##           2 0 0 0 5
##           3 2 0 0 0
##           4 0 1 0 0
```

```
# Evaluer la proportion d'inertie expliquée
inertie.expl <- rep(0, times=7)
for (k in 2 :7){
  clus <- kmeans(prix_nobel.cr,centers=k,nstart=5)
  inertie.expl[k] <- clus$betweenss/clus$totss
}
# Graphique
plot(1 :7,inertie.expl,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")
```

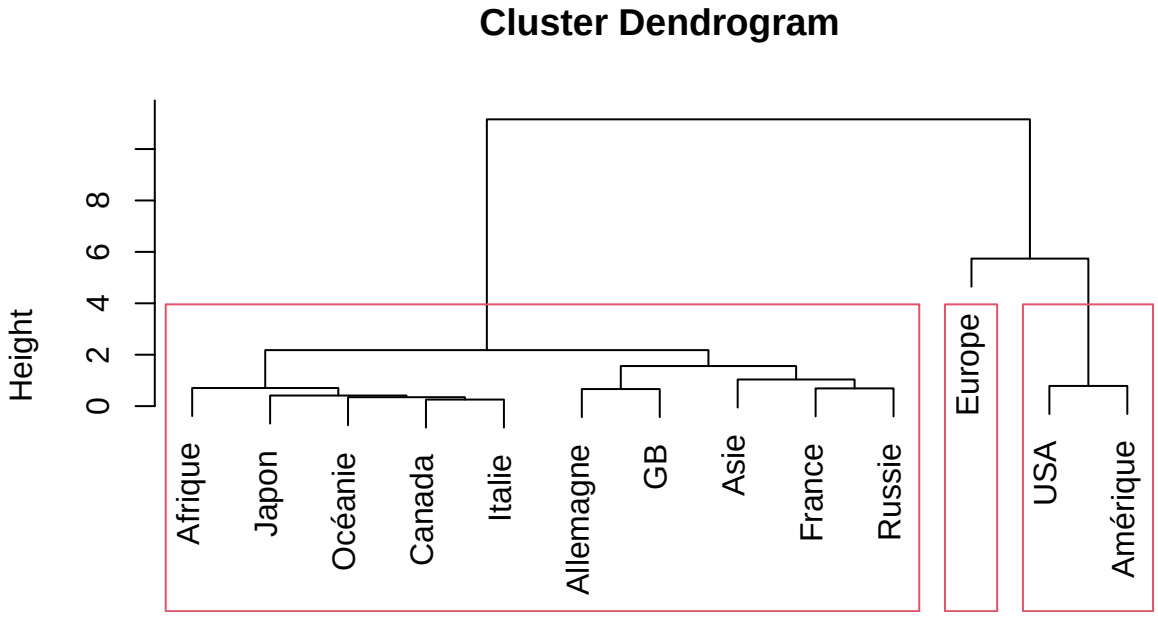


A partir de k = 3 classes, l'adjonction d'un groupe supplémentaire n'augmente pas «significativement» la part d'inertie expliquée par la partition.

Nous allons donc proceder par un decoupage en 3 groupes à l'aide de la commande rect.hclust:

```
# affichage du dendrogramme
plot(cah.ward)

# Découpage en classes/groupes
#dendrogramme avec matérialisation des groupes
rect.hclust(cah.ward,k=3)
```




```
# découpage en 3 groupes
groupes.cah2 <-cutree(cah.ward,k=3)

# liste des groupes
print(sort(groupes.cah2))
```

```
## Allemagne    Canada    France        GB    Italie     Japon    Russie   Afrique
##         1         1         1         1         1         1         1         1
##      Asie   Océanie       USA  Amérique    Europe
##         1         1         2         2         3
```