TP2

Taniel Rémi

24 mars 2019

Analyse en composantes principales

QUESTION 1

On commence par charger les données stockées dans le fichier ocde.txt et on les stocke dans la variable ocde :

```
ocde <- read.table("ocde.txt", header=TRUE, sep = " ")
```

Analyse descriptive

On utilise la commande dim() pour obtenir les dimensions (nombre d'éléments et nombre de variables) de notre dataframe ocde:

```
dim(ocde)
```

```
## [1] 17 13
```

Notre dataframe contient donc 17 éléments pour 13 variables.

Pour analyser nos données, on se décide d'afficher les premières lignes de celle-ci, pour cela on utilise la commande head(), Voici donc les premieres lignes de notre dataframe :

```
head(ocde)
```

```
##   PAYS YEAR NATA CHOM APRI ASEC  PIB FBCF INFL RECC MINF PROT NRJ
## 1   AL 1975   97   41   73  460 6870  211   60  409  197   58 394
## 2   AU 1975  123   17  125  409 4990  267   78  391  205   56 307
## 3   BE 1975  121   42   36  399 6350  220   94  407  146   62 426
## 4   CA 1975  157   69   61  293 6990  242   83  374  155   65 878
## 5   DA 1975  142   49   98  315 7010  199   99  450  104   66 350
## 6   ES 1975  188   47  219  385 2870  232  139  231  121   50 173
```

On remarque qu'il n'y a qu'une seule variable qualitative, la variable 'PAYS', on décide donc de remplacer le numéro des lignes par la variable 'PAYS' afin de mieux identifier les points dans les prochaines analyses :

```
rownames(ocde) <- ocde$PAYS
ocde <- ocde[,-1]
```

Analyse descriptive univariée

Pour chaque variable, on décide d'afficher les valeurs minimales et maximales, la médianne, la moyenne ainsi que le 1er et 3e quartile, on utilise pour cela la commande summary() :

```
summary(ocde)
```

```
##       YEAR           NATA            CHOM            APRI      
##  Min.   :1975   Min.   : 97.0   Min.   :16.00   Min.   : 27.0  
##  1st Qu.:1975   1st Qu.:127.0   1st Qu.:23.00   1st Qu.: 64.0  
##  Median :1975   Median :142.0   Median :41.00   Median :102.0  
##  Mean   :1975   Mean   :147.8   Mean   :41.88   Mean   :116.8  
##  3rd Qu.:1975   3rd Qu.:157.0   3rd Qu.:49.00   3rd Qu.:149.0  
##  Max.   :1975   Max.   :216.0   Max.   :83.00   Max.   :282.0  
##       ASEC            PIB            FBCF            INFL      
##  Min.   :290.0   Min.   :1550   Min.   :136.0   Min.   : 60.0  
##  1st Qu.:336.0   1st Qu.:4070   1st Qu.:207.0   1st Qu.: 85.0  
##  Median :361.0   Median :5950   Median :220.0   Median : 96.0  
##  Mean   :364.5   Mean   :5368   Mean   :228.6   Mean   :108.3  
##  3rd Qu.:399.0   3rd Qu.:6990   3rd Qu.:242.0   3rd Qu.:138.0  
##  Max.   :460.0   Max.   :8460   Max.   :354.0   Max.   :169.0  
##       RECC            MINF            PROT            NRJ       
##  Min.   :230.0   Min.   : 83.0   Min.   :34.00   Min.   : 88.0  
##  1st Qu.:347.0   1st Qu.:105.0   1st Qu.:55.00   1st Qu.:297.0  
##  Median :395.0   Median :146.0   Median :59.00   Median :363.0  
##  Mean   :381.9   Mean   :155.7   Mean   :58.06   Mean   :401.5  
##  3rd Qu.:409.0   3rd Qu.:184.0   3rd Qu.:66.00   3rd Qu.:473.0  
##  Max.   :536.0   Max.   :379.0   Max.   :73.00   Max.   :878.0
```

On remarque que la variable 'YEAR' possède toujours la même valeur: 1975, nous pouvons donc la retirer notre dataframe ocde :

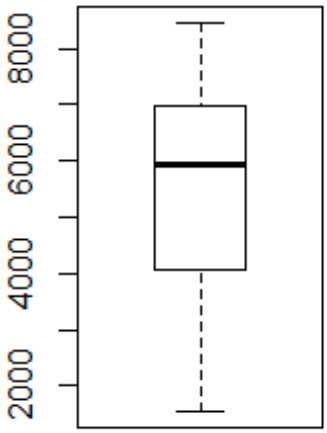
```
ocde <- ocde[,-1]
```

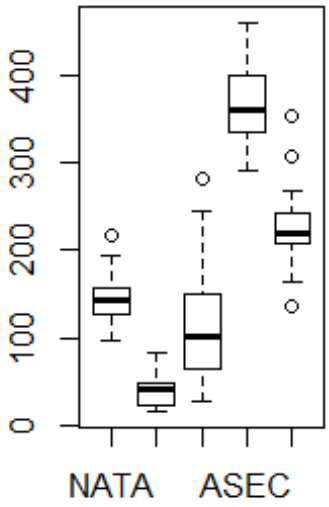
On remarque également que la variable 'PIB' possède des valeurs plus bien supérieurs aux restes des variables, cela risque donc de poser problème pour les prochains graphiques.

Analyse en boite à moustache

Pour chaque variable de notre dataframe ocde, on se décide d'afficher sa boite à moustache :

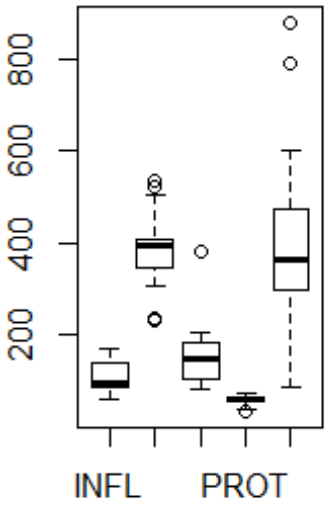
```
par(mfrow=c(1,2))
boxplot(ocde$PIB, xlab="PIB (par habitant)")
boxplot(ocde[,-c(5,7,8,9,10,11)], xlab="Autres (NATA, CHOM, APRI, ASEC,
FBCF)")
```





PIB (par habitant) utres (NATA, CHOM, APRI, ASEC,

```
boxplot(ocde[,-c(1,2,3,4,5,6)], xlab="Autres (INFL, RECC, MINF, PROT, NRJ)")
```

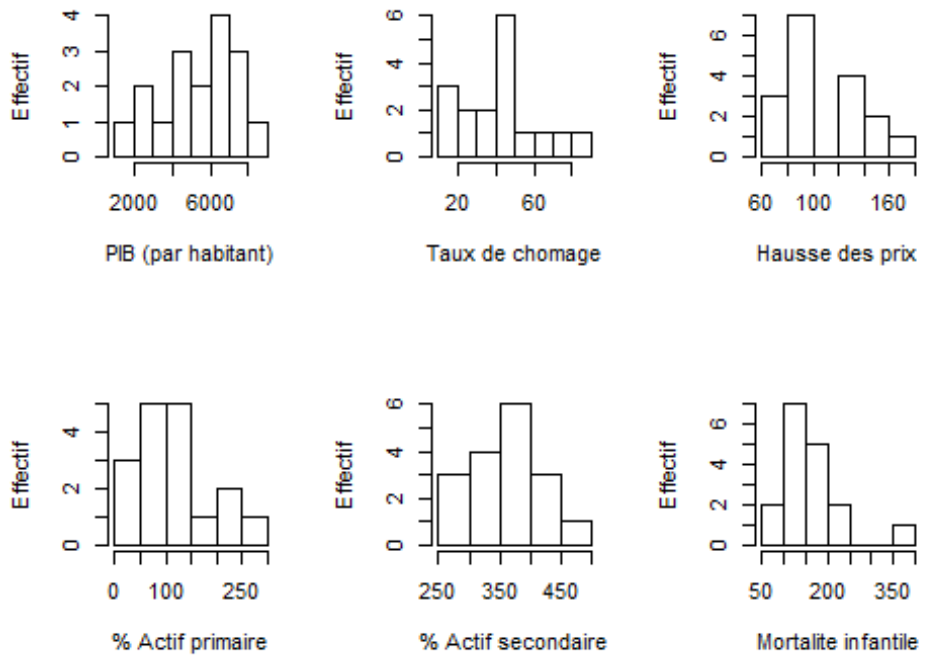


Autres (INFL, RECC, MINF, PROT,

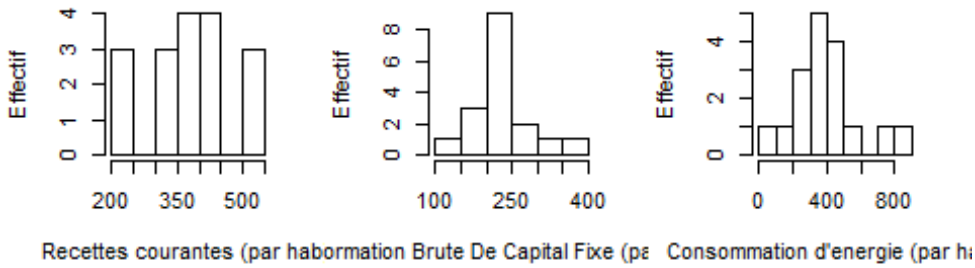
Analyse en histogramme

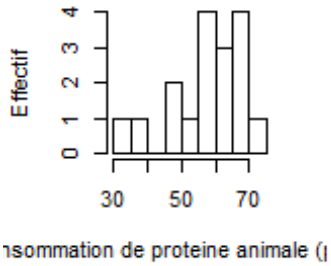
Pareil que précedemment mais cette fois-ci en histogramme :

```
par(mfrow=c(2,3))
hist(ocde$PIB, ylab="Effectif", xlab="PIB (par habitant)", main="")
hist(ocde$CHOM, ylab="Effectif", xlab="Taux de chomage", main="")
hist(ocde$INFL, ylab="Effectif", xlab="Hausse des prix", main="")
hist(ocde$APRI, ylab="Effectif", xlab="% Actif primaire", main="")
hist(ocde$ASEC, ylab="Effectif", xlab="% Actif secondaire", main="")
hist(ocde$MINF, ylab="Effectif", xlab="Mortalite infantile", main="")
```




```
hist(ocde$RECC, ylab="Effectif", xlab="Recettes courantes (par hab)",
main="")
hist(ocde$FBCF, ylab="Effectif", xlab="Formation Brute De Capital Fixe (par
hab)", main="")
hist(ocde$NRJ, ylab="Effectif", xlab="Consommation d'energie (par hab)",
main="")
hist(ocde$PROT, ylab="Effectif", xlab="Consommation de proteine animale (par
hab)", main="")
```





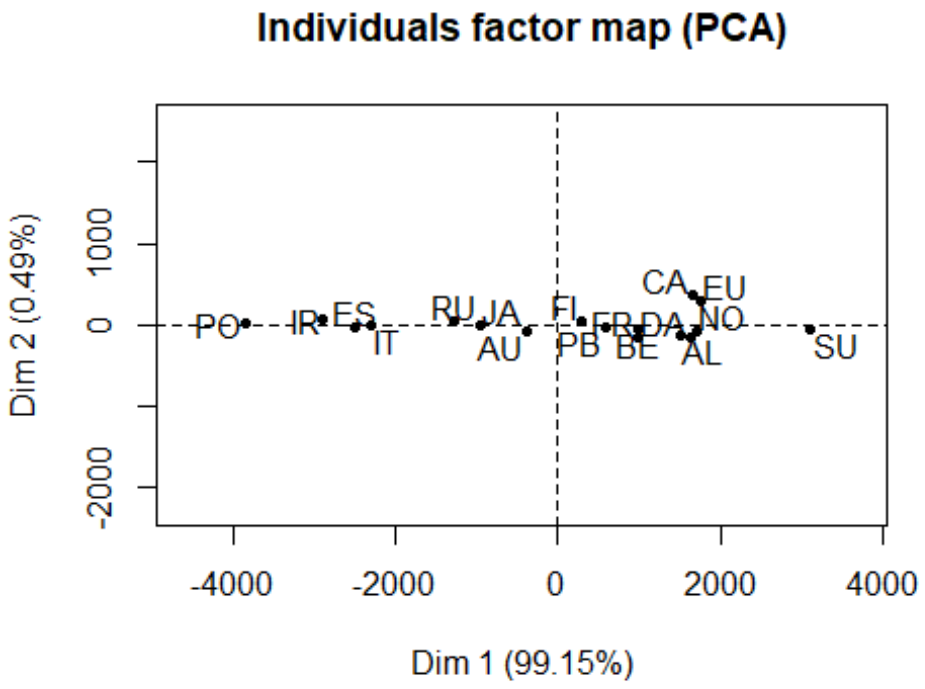
QUESTION 2

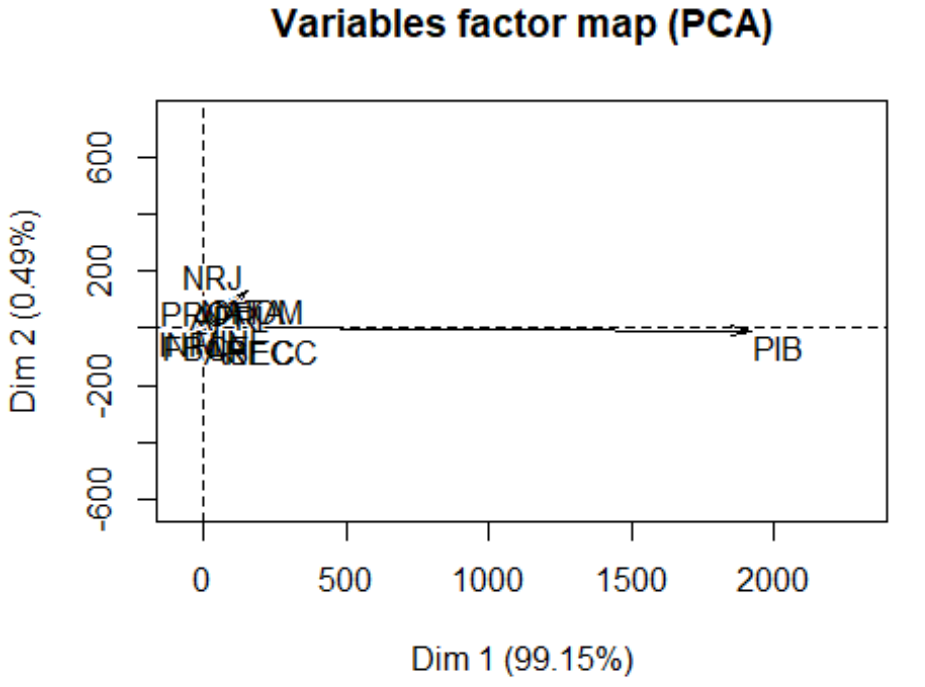
Nous devrons utiliser les packages FactoMineR et factoextra, pour les utiliser, nous devons les importer avec les commandes suivantes :

```
library(FactoMineR)
library(factoextra)
```

Nous allons dans un premier temps, faire une analyse en composante principales principales de la base de donnée sans réduire les donnees, pour cela on utilise la commande PCA() avec l'option 'scale.unit = FALSE' :

```
pca <- PCA(ocde, scale.unit=F)
```





Sans réduire les données, il est difficile d'analyser les réultats etant donné que la variable 'PIB' "explose" le reste des variables,

L'axe principal monopolise donc les données (99,15% de pourcentage d'inertie), en l'état, le graphe n'est pas exploitable, tout comme le reste de l'analyse en composantes principales…

Question 4

On se propose donc de relancer une analyse en composantes principales, mais cette fois-ci en réduisant les données, on utilise donc le parametre 'scale.unit = TRUE' et on stocke le résultat de l'analyse dans la variable pca :

```
pca <- PCA(ocde[,-1], scale.unit=T, graph=F)
```

Fonctions d'analyse

Dans cette partie, je ne donnerai que les fonctions et leurs résultats, aucune analyse ne sera faite sur les résultats de ces fonctions, elles seront faîtes dans les questions suivantes.

Valeurs propres

On se décide d'afficher les valeurs propres, c'est a dire le pourcentage de variances (pourcentage d'inertie) expliqués pour chaque axe principal :

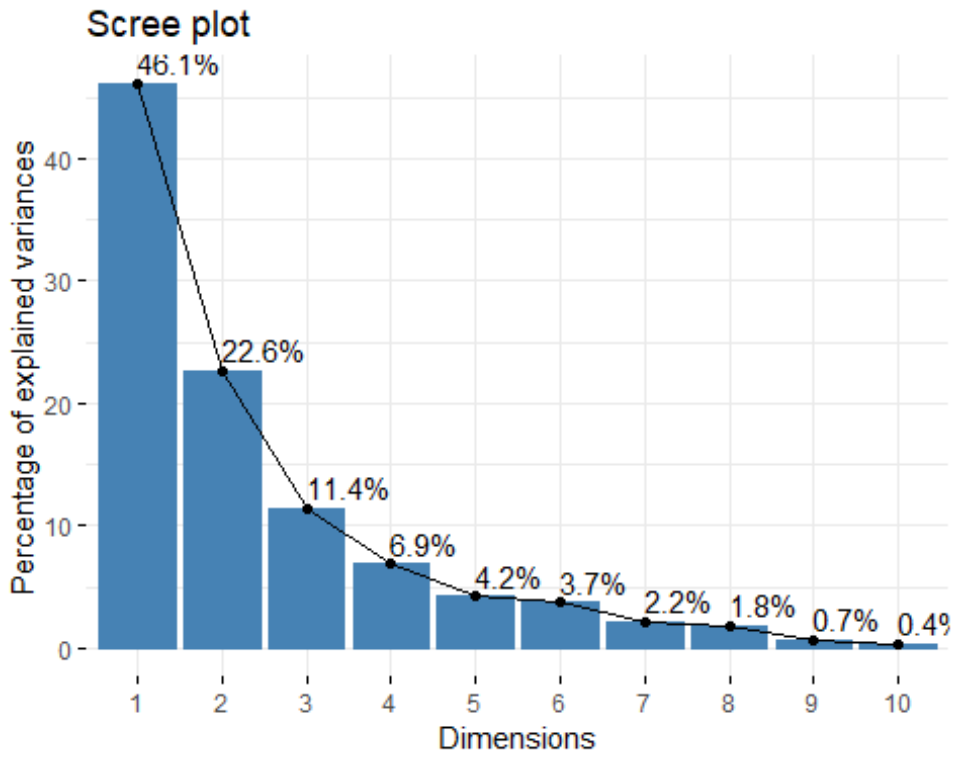
```
get_eigenvalue(pca)
```

```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1  4.60664697       46.0664697                     46.06647
## Dim.2  2.26339082       22.6339082                     68.70038
## Dim.3  1.14397643       11.4397643                     80.14014
## Dim.4  0.68600943        6.8600943                     87.00024
## Dim.5  0.42158396        4.2158396                     91.21608
## Dim.6  0.37229511        3.7229511                     94.93903
## Dim.7  0.21916595        2.1916595                     97.13069
## Dim.8  0.18252643        1.8252643                     98.95595
## Dim.9  0.06700182        0.6700182                     99.62597
## Dim.10 0.03740308        0.3740308                    100.00000
```

Part d'inertie exliquée par chaque axe

Pour afficher la partie d'inertie expliquée par chaque axe, on utilise la fonction suivante :

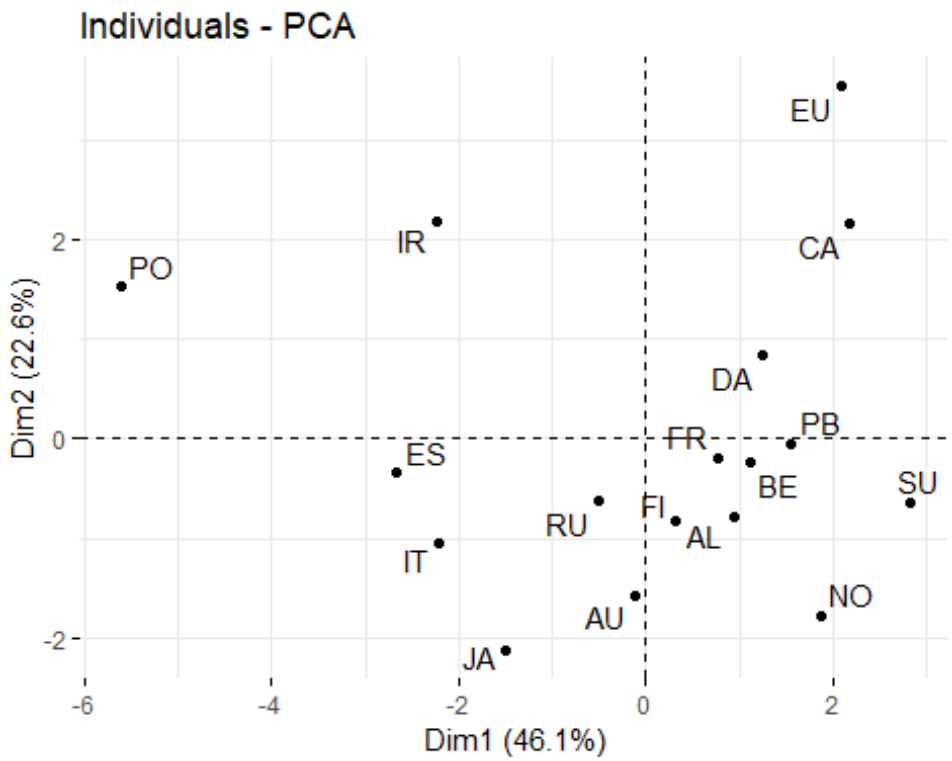
```
fviz_eig(pca, addlabels=T)
```



Coordonnées des individus dans la nouvelle base

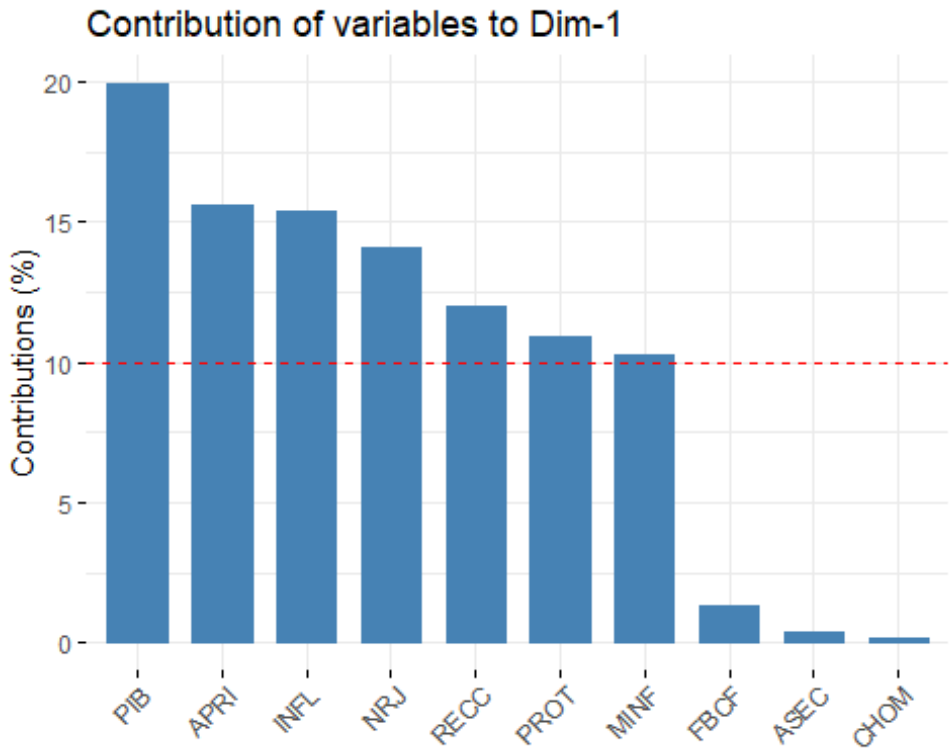
Pour afficher les coordonnées des individus dans la nouvelle base, on utilise la fonction suivante :

```
fviz_pca_ind(pca, repel=T)
```

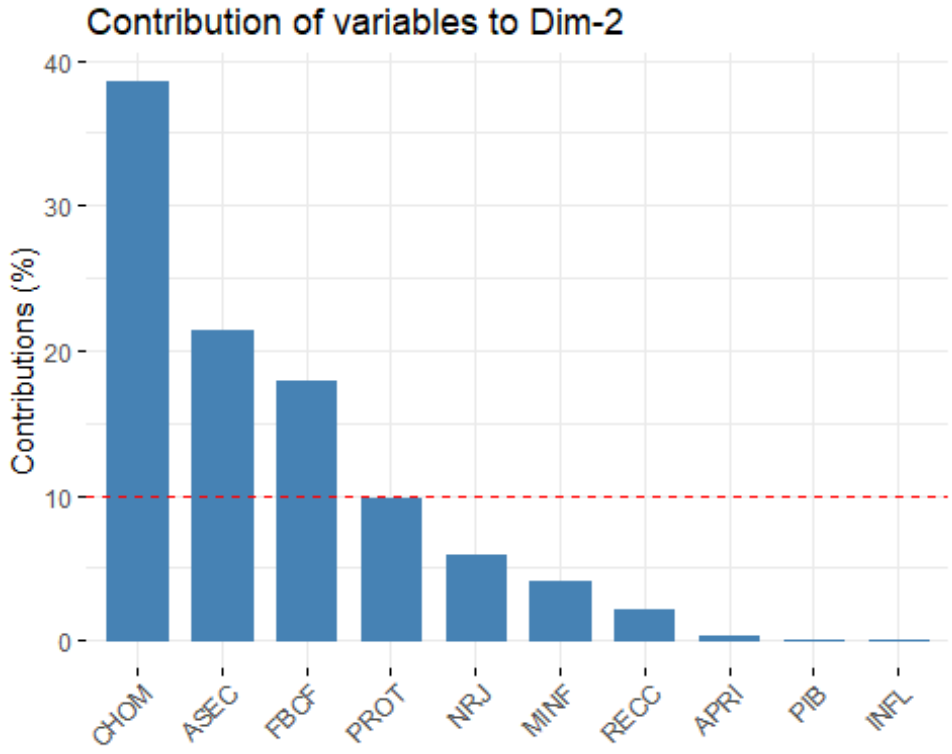


Contribution des variables sur les 2 axes principaux

```
# Contributions des variables sur l'axe 1
fviz_contrib(pca, choice = "var", axes = 1, top = 10)
```

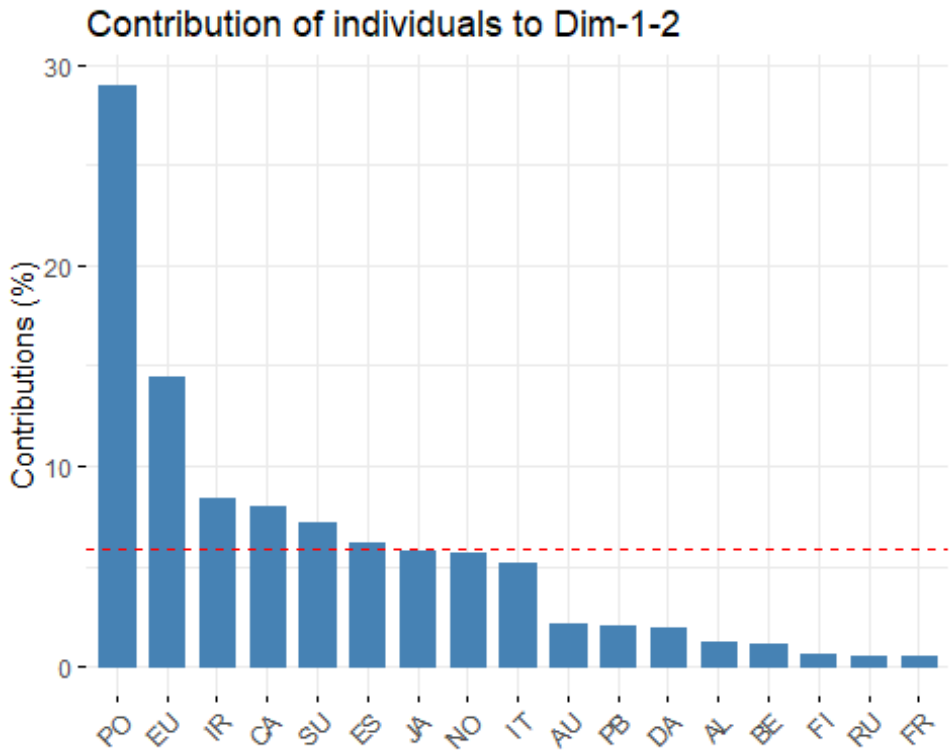



```r
# Contributions des variables sur l'axe 2
fviz_contrib(pca, choice = "var", axes = 2, top = 10)
```



Contribution des individus sur les 2 axes principaux

```
fviz_contrib(pca, choice="ind", axes=1:2)
```

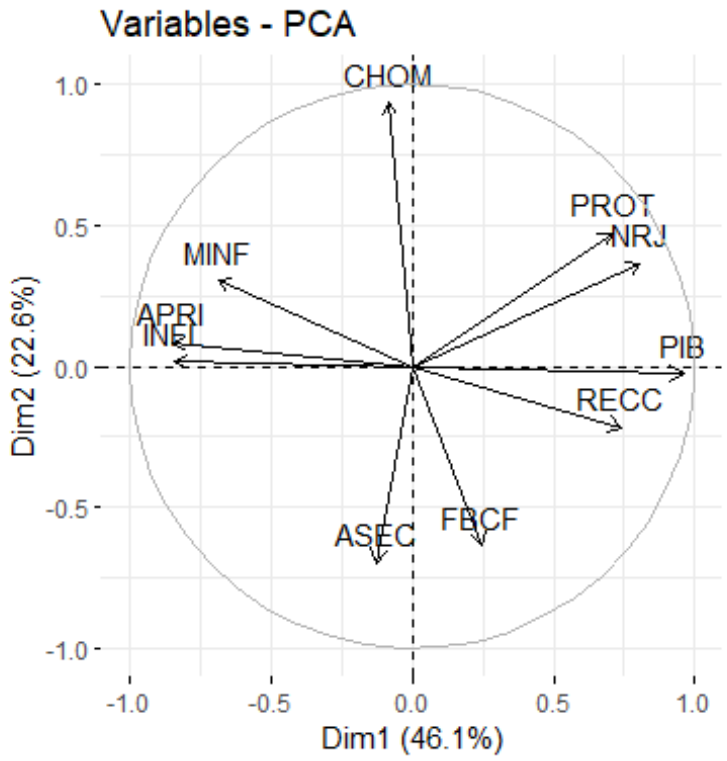


Cercle de corrélations

Graphique

Pour afficher le cercle de corrélations, on utilise la fonction suivante :

```
fviz_pca_var(pca, col.var="black")
```



Analyse

Axe 1

Les variables PROT, NRJ, PIB et RECC sont posivitement corrélés entre elles et sont négativement corrélés aux variables APRI, INF et MINF.

On peut donc déduire que plus un pays est riche (fort PIB, forte recette courante par habitant), il est beaucoup moins ? à l'inflation des prix, la mortalité infantile et possède un faible pourcentage d'actifs dans le secteur primaire.

Axe 2

Les variables ASEC et FBCF sont positivement corrélés et négativement corrélés avec la variable CHOM.

Dans ce cas, on peut déduire que plus le taux de chômage dans un pays est élevés, moins il possède un fort pourcentage d'actifs dans le secteur secondaire et de formation brute de capital fixe par habitant.

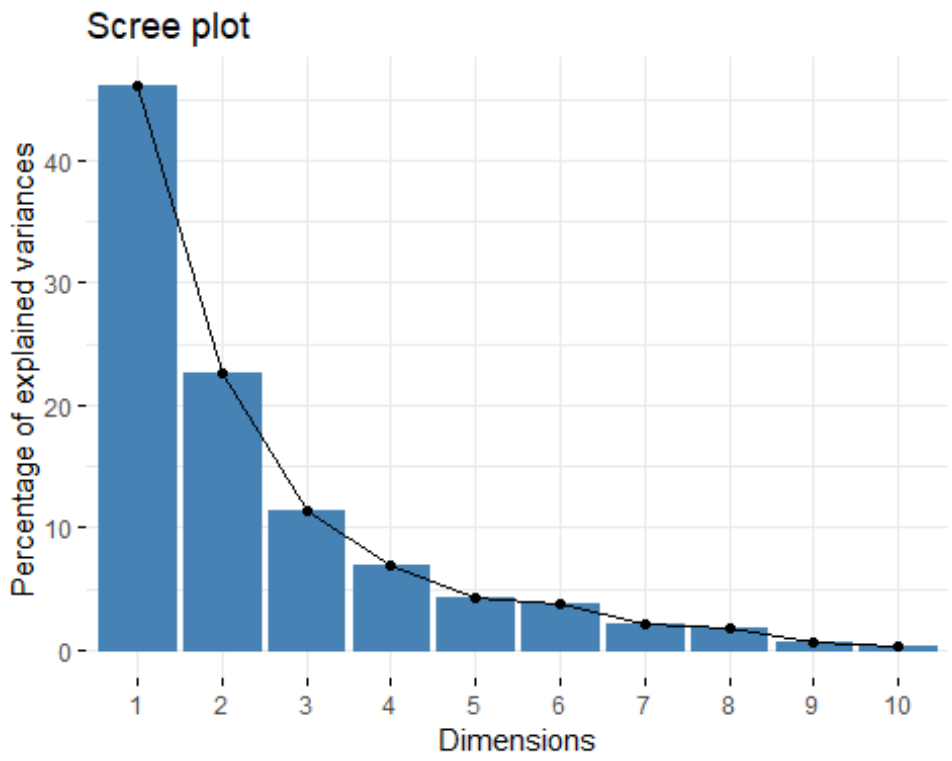
Question 5

Pour déterminer le nombre d'axes à retenir, nous pouvons utiliser 3 critères, qui sont les suivants : - Nombre d'axes à retenir pour obtenir un pourcentage cumulé d'inertie de 80% - Critère du "coude" - Régle de "Kaiser"

Critère du "coude"

Pour utiliser le critère du "coude", nous devons d'abord afficher le graphique representant la part de variance pour chaque dimension grace à la commande suivante :

```
fviz_eig(pca)
```



Avec le critere du "coude", nous retenons les 3 premiers axes.

Seuil

Concernant le critere du seuil, nous devons sommer chaque pourcentage d'inertie expliqué des x premieres dimensions jusqu'a atteindre 80%, pour cela on utilise :

```
cumsum(pca$eig[,2])
```

```
##    comp 1    comp 2    comp 3    comp 4    comp 5    comp 6    comp 7
##  46.06647  68.70038  80.14014  87.00024  91.21608  94.93903  97.13069
```

```
##    comp 8    comp 9   comp 10
##  98.95595  99.62597 100.00000
```

Pour atteindre un seuil de 80% de pourcentage d'inertie, nous devons egalement retenir les 3 premiers axes

Regle de "Kaiser"

Etant donné que notre ACP est normée, nous devons retenir les axes ayant une valeur propre supérieur a 1 :

```
sum(pca$eig[,1] > 1)
```

```
## [1] 3
```

Tout comme les 2 premiers critères, nous retenons 3 axes

Conclusion

Les 3 critères ayant donné le même nombre d'axes a retenir, nous decidons de retenir 3 axes.

Question 6

Qualité de representation

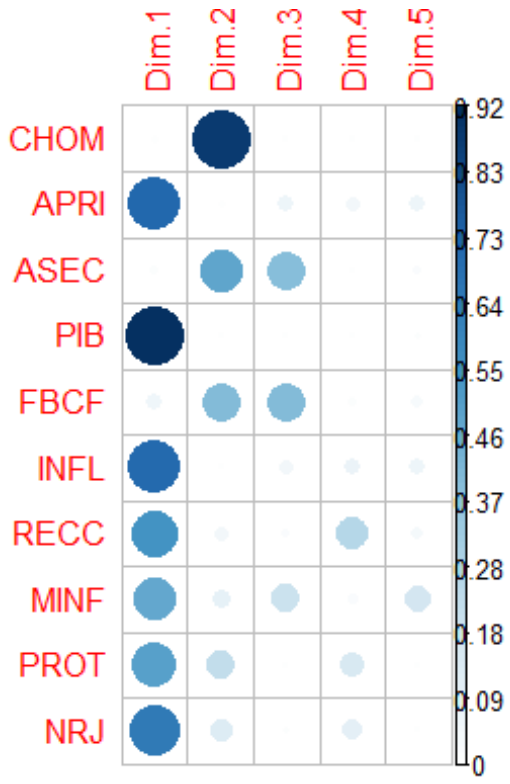
La qualité de representation des variables sur la carte de l'ACP s'appelle cos2, on peut visualiser ses valeurs avec la commande :

```
pca$var$cos2
```

```
##            Dim.1        Dim.2       Dim.3        Dim.4        Dim.5
## CHOM 0.007100345 0.8726781521 0.001820478 0.0008964709 0.0032529296
## APRI 0.718241793 0.0070731506 0.072285848 0.0477251707 0.0690951575
## ASEC 0.016311396 0.4834527822 0.387917665 0.0009412734 0.0197745259
## PIB  0.918505443 0.0006962086 0.003183338 0.0048718784 0.0094625591
## FBCF 0.061928586 0.4037704767 0.397739340 0.0146565735 0.0361077027
## INFL 0.709891793 0.0002617331 0.046998677 0.0749544421 0.0675407852
## RECC 0.551043716 0.0467649085 0.020792394 0.2623189273 0.0449947371
## MINF 0.472352407 0.0929896880 0.197155766 0.0261208112 0.1703296250
## PROT 0.503604881 0.2228254346 0.005110729 0.1496370043 0.0007015364
## NRJ  0.647666607 0.1328782836 0.010972196 0.1038868739 0.0003243977
```

On utilise le package corrplot pour visualiser le cos2 de chaque variable sur chaque axe principal :

```
library(corrplot)
corrplot(pca$var$cos2, is.corr=F)
```



Un cos2 élevé (ici en bleu foncé) indique une bonne représentation de la variable sur l'axe principal en question, dans ce cas, sur le cercle de corrélations, la variable est proche de la circonférence du cercle.

Inversement, un cos2 faible (ici en bleu clair) indique que la variable n'est pas parfaitement representé sur l'axe principal en question, dans ce cas, sur le cercle de corrélations, la variable est proche du centre du cercle.

Analyse des contributions

Contribution des variables aux axes

Dans le cas de l'axe 1, ce sont les variables PIB, APRI, INFL, NRJ, RECC, PROT et MINF, tandis que pour l'axe 2 ce sont CHOM, ASEC, FBCF et PROT qui y contribuent le plus.

Contribution des individus aux axes

test