TP ACP

PC & YM

28/07/2021

Packages

Plusieurs packages sont disponibles pour réaliser des ACP avec R. Dans ce TP nous utiliserons les packages `ade4` qui réalise les calculs et `factoextra` qui fournit des outils pour visualiser les résultats.

```
install.packages("ade4")
install.packages("factoextra")
```

Nous utiliserons aussi un package pour les données : `palmerpenguins`, de Horst AM, Hill AP, Gorman KB (2020) , et qui fournit des mesures de la morphologie de trois espèces de pengouins:

```
install.packages("palmerpenguins")
```

Une fois installés, on charge ces packages ainsi que d'autres déjà connus : `dplyr`, `ggplot2`

```
library(dplyr)
library(ggplot2)
library(ade4)
library(factoextra)
library(palmerpenguins)
```

Données

les données sont décrites sur la page github du package : https://allisonhorst.github.io/palmerpenguins/

Nous utiliserons l'objet `penguins` fourni par ce package , dont voici les 6 premières lignes:

```
data("penguins")
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6              190        3650 male
## # ... with 1 more variable: year <int>
```

Question 1 : Préparation des données

Filtrer les observations non attribuées, de façon à ce qu'il n'y ait plus de valeurs `NA` dans le dataframe.

Indice : Utiliser les fonctions `anyNA` et `na.omit`

Question 2 : Affichage des données

Commencer par afficher les noms et les types des variables du dataset `penguins`.

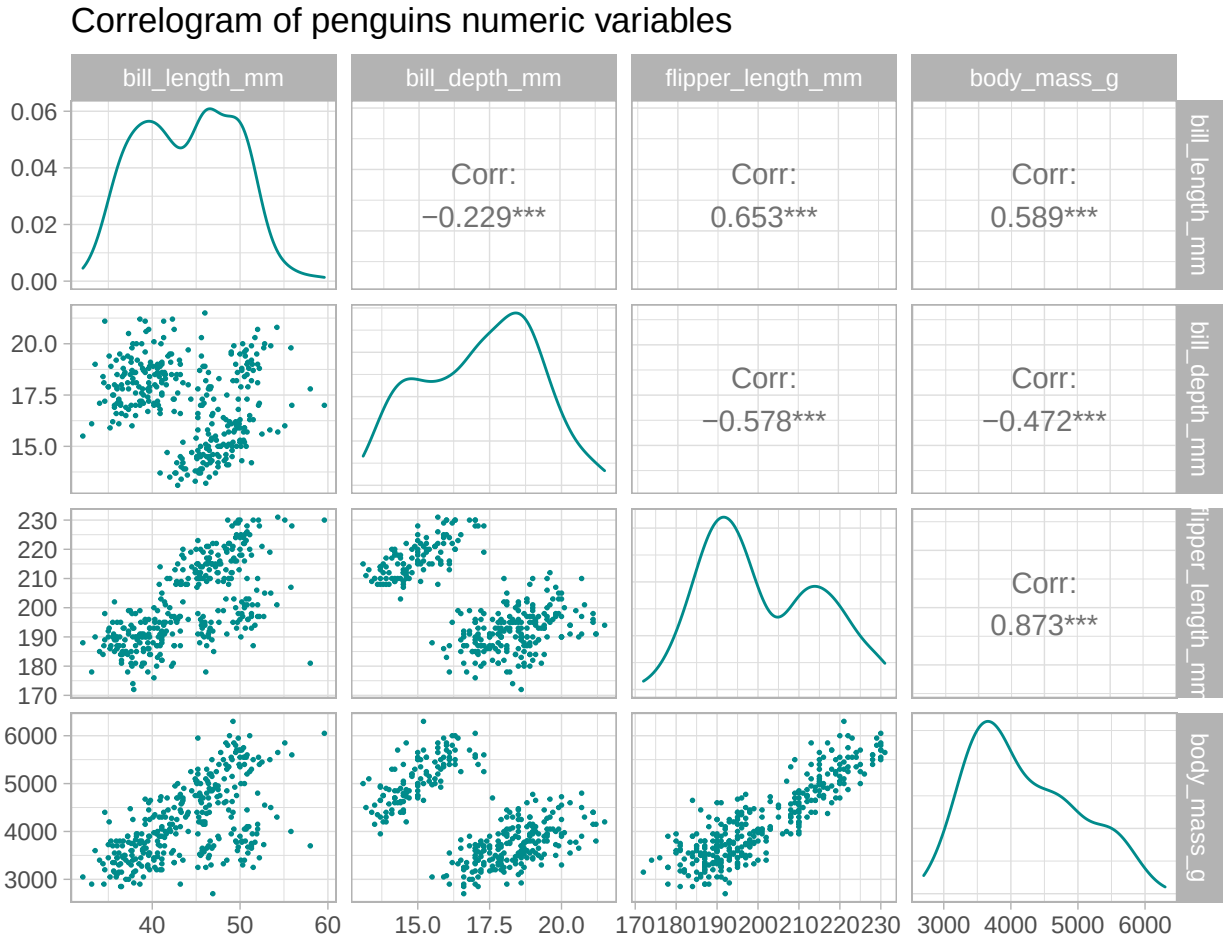
Réaliser ensuite :

- un affichage des nuages de points des variables **numériques** deux à deux
- calculer leur corrélations
- afficher la densité ou l'histogramme de chaque variable

Commenter ces graphiques : quelle structure remarquez-vous ?

Indice : Utiliser les fonctions de base `cor`, `plot` , `hist`, ou `ggpairs` du package `GGally`

Voici une version synthétique de ce que vous devriez obtenir :



Question 3 : Préparation des données

Créer un dataframe nommé `dataACP` contenant uniquement les variables numériques qui dévrivent la morphologie des pingouins.

Indice : utiliser la fonction `select` du package `dplyr` , ou l'indexation de colonne standard.

Analyse en composantes principales

Question 4 : calculer l'inertie de `dataACP` *sans les normaliser*

Indice : utiliser la fonction `var` pour calculer la variance d'un vecteur / d'une liste / d'une colonne, la fonction `diag` qui renvoie la diagonale d'une matrice carrée

Question 4 bis : Pourquoi calculer l'inertie de variables normalisées est inutile (et trivial) ?

Question 5 : Réaliser une ACP sur `dataACP` et stocker le résultat dans une variable (e.g. `result_ACP`)

Indice : utiliser la fonction `dudi.pca` du package `ade4` et les fonctions du package `factoextra` :

- `get_eigenvalue` : Extraction des valeurs propres / variances des composantes principales
- `fviz_eig` : Visualisation des valeurs propres
- `get_pca_ind`, `get_pca_var`: Extraction des résultats pour les individus et les variables, respectivement.
- `fviz_pca_ind`, `fviz_pca_var`: visualisez les résultats des individus et des variables, respectivement.

Bien que les fonctions de `factoextra` fassent le travail pour vous, il est important de bien lire la documentation de la fonction `dudi.pca` pour savoir quels attributs extraire de l'objet résultat, dans le cas d'une automatisation des traitements par exemple.

Question 5 bis : quel est le pourcentage d'inertie capturée par les deux premières composantes ? Comment l'obtenir sans lire le scree-plot ?

Question 5 ter : quelle est la coordonnée de la deuxième composante dans l'espace de départ ?

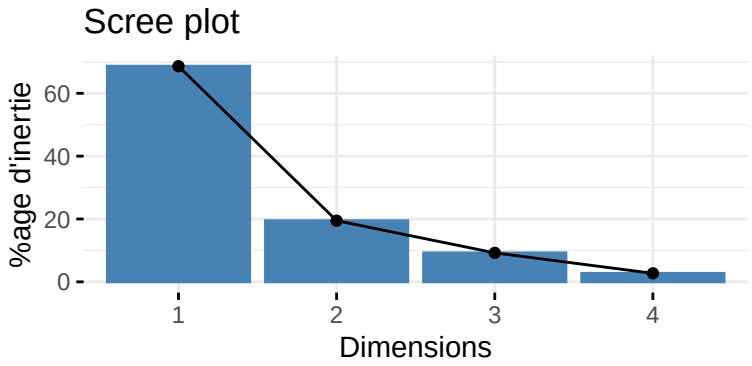
Question 6 : D'après-vous, faut-il normaliser les variables de `dataACP` lors de l'ACP ? Pourquoi ?

Pour s'en assurer, recommencer le calcul de l'ACP et comparer les résultats .

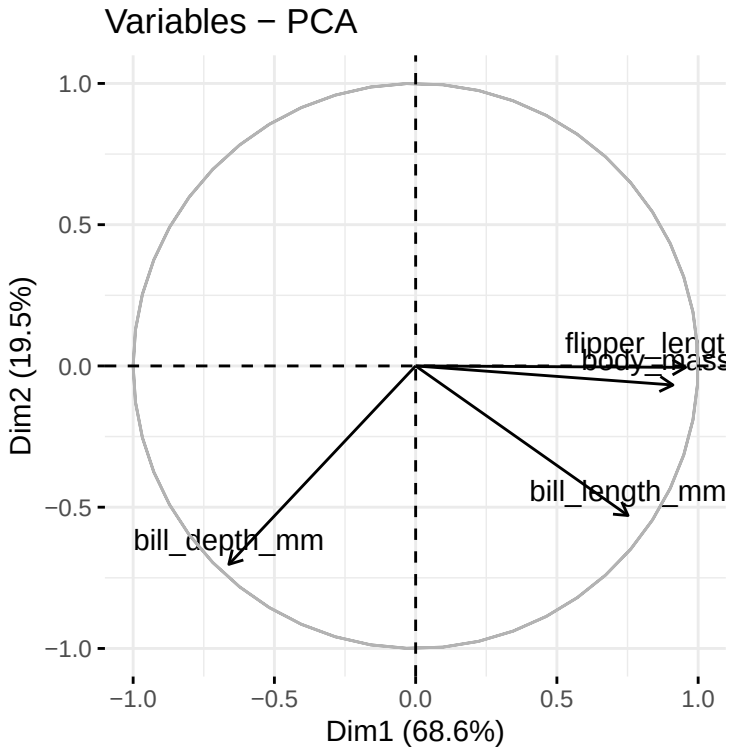
Interprétation des résultats

Question 7 : l'ACP s'est-elle bien passée ? Justifier .

Vous devriez obtenir à la question 5 un graphique à l'allure suivante :



Question 8 : Que dire des variables projetées dans le plan formé par les deux premières composantes ?



Question 9 : Quelle est la contribution des variables `bill_length` et `bill_depth` à la 3ème composante ?

Question 10 : projeter les individus dans le plan formé par les deux premières composantes . Interpréter.

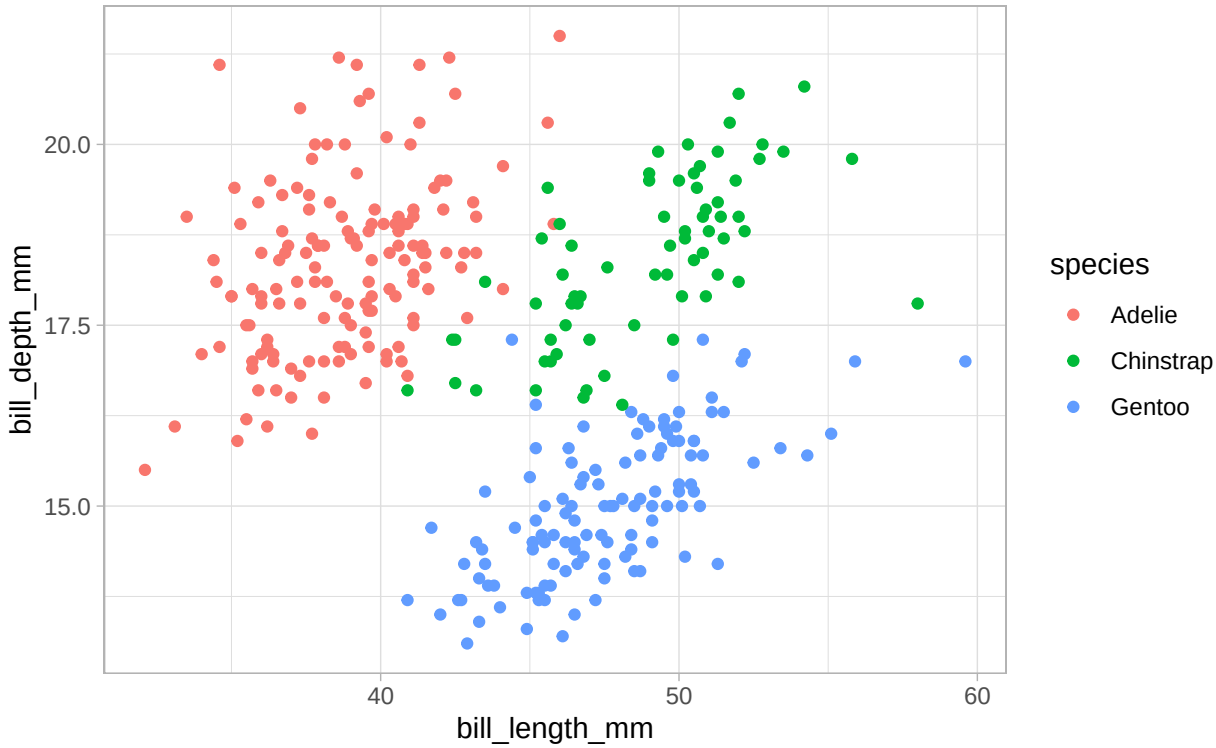
Indice : cf. la liste des fonctions de `factoextra`

Regrouppements

Nous cherchons maintenant à trouver une projection qui permette de séparer visuellement les trois espèces de pengouins.

Voici comment obtenir des nuages de points des variables des pingouins ,colorés par espèce :

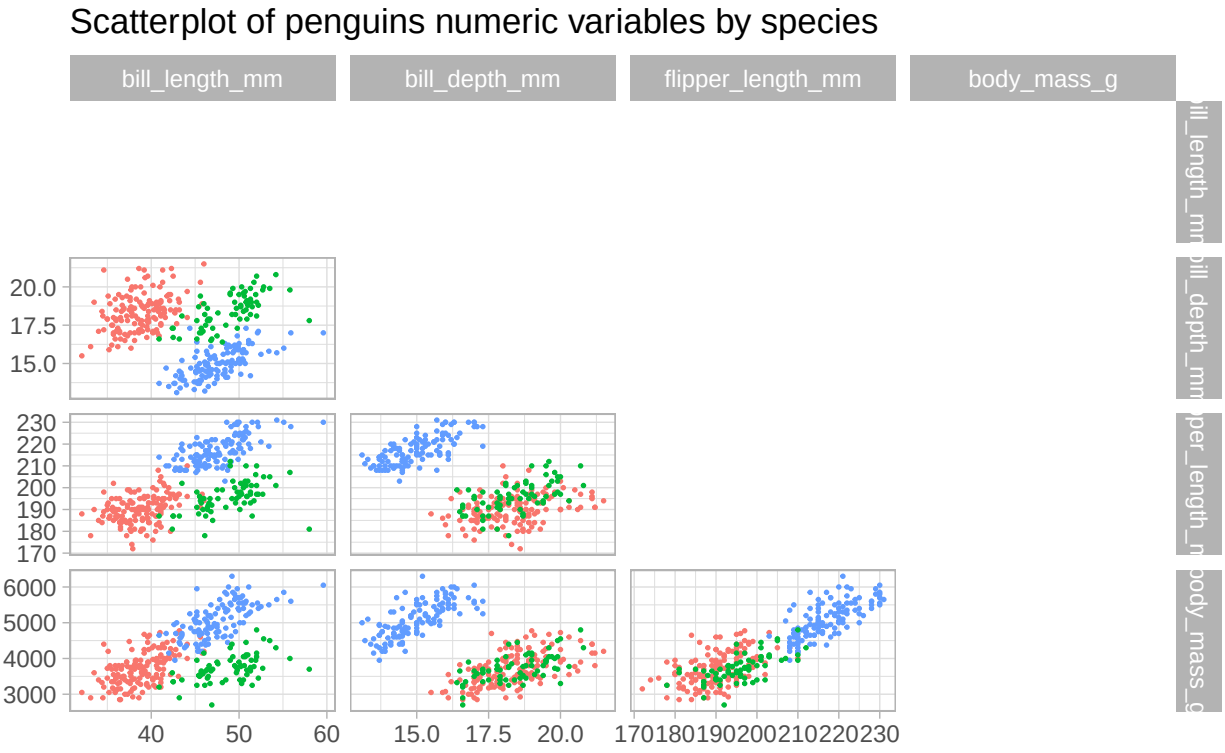
```r
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm, colour = species)) +
  geom_point() +
  theme_light()
```



Voici une version pour tous les couples de variables :

```r
ggpairs(data= penguins[,c("species" ,"bill_length_mm","bill_depth_mm", "flipper_length_mm", "body_mass_g
        columns = c("bill_length_mm","bill_depth_mm", "flipper_length_mm", "body_mass_g" ),
        #lower = list(continuous ="points", size=0.1),
        lower= list(continuous =wrap("points",  size=0.3), combo = "facethist", discrete = "facetbar", n
         #upper= list(continuous =wrap("points",  size=0.3), combo = "facethist", discrete = "facetbar"
        title="Scatterplot of penguins numeric variables by species",

        diag=NULL,
        upper=NULL,
        mapping = aes(color=species)
        ) +theme_light()
```



Question 11 : Les graphiques ci-dessus permettent-ils d'opérer cette classification visuelle ?