

Présentation des jeux de données

Nous aurons à utiliser lors de cet atelier 3 jeux de données pour pratiquer les différentes analyses. Nous présentons ici chaque jeu de données ainsi que le moyen de l'installer ou le charger.

Le jeu de données penguins

A propos



Figure 1: Palmer penguins hex sticker (Artwork by [allison_horst](#))

Le jeu de données Penguins est un jeu de données collectées et mises à disposition par le [Dr. Kristen Gorman](#) et la station [Palmer, Antarctica LTER](#), membre du [Long Term Ecological Research Network](#) (réseau de recherche écologique à long terme) et mise à disposition de la communauté R au travers du package `palmerpenguins`.

Le jeu de données s'appelle `penguins`, mais fait références en français à des *manchots* et non à des pingouins. Pour rappel, il y a deux différences fondamentales entre les pingouins et les manchots: leur répartition géographique et leur (in)capacité à voler. Les pingouins vivent dans

l'hémisphère nord et ils peuvent voler! Quant aux manchots, ils ne peuvent pas voler et ils vivent dans l'hémisphère sud. Cependant, lors de ce atelier nous allons faire reference a ce jeu de données en utilisant le terme `penguins` pour garder le nom original du jeu de données.

Le jeu de données contient des données de 344 manchots. Il y a 3 espèces différentes de manchots dans ce jeu de données Figure 2, collectées sur 3 îles de l'archipel de Palmer, en Antarctique.

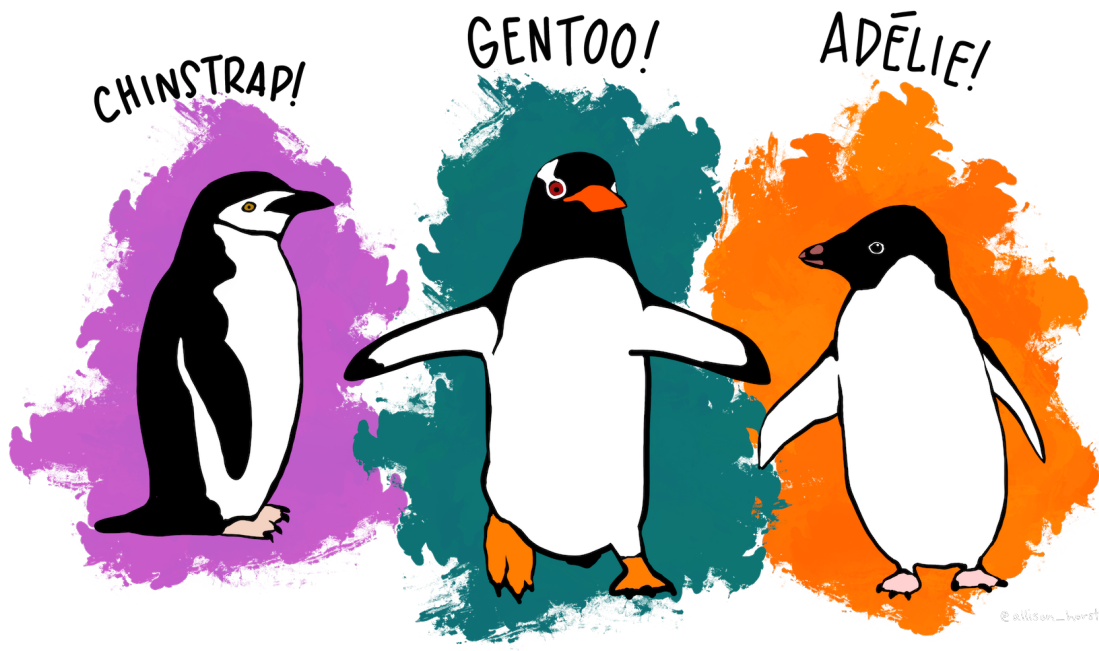


Figure 2: Les espèces de manchots dans palmerpenguins

Installation et description courte

Le package est disponible sur le CRAN et peut être installé à partir de la console R en utilisant la commande ci-dessous:

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

Le jeu de donnée est composé de 344 observations et de 8 variables:

```
str(penguins)
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species      : Factor w/ 3 levels "Adelie","Chinstrap",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ island       : Factor w/ 3 levels "Biscoe","Dream",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm  : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g    : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex           : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year           : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

Les différentes variables sont l'espèce, l'île (lieu de collecte des données), la longueur du culmen (mm), la profondeur du culmen (mm), la longueur de la nageoire (mm), le poids (g), le sexe et l'année de l'étude. Le culmen est appelé *bill* dans le jeu de donnée. En zoologie, le culmen est l'arête dorsale de la mandibule supérieure des oiseaux Figure 3.

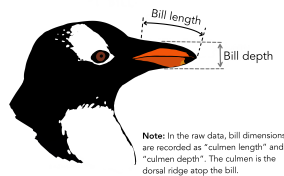


Figure 3: Description du culmen des manchots

Pour rappel, la description complète du format du jeu de donnée est disponible directement dans R en utilisant la commande `?penguins`.