

# Tidyverse - Exercices

Pierre Lafaye De Micheaux

Novembre 2023

## Iris

Nous considérons le jeu de données `iris`. Répondre aux questions suivantes en utilisant les fonctions du package `dplyr` :

1. Sélectionner les variables `Petal.Width` et `Species`.
2. Construire une table qui contient uniquement les iris d'espèce `versicolor` ou `virginica`.
3. Calculer the nombre d'iris `setosa` en utilisant `summarise`.
4. Calculer la moyenne de la variable `Petal.Width` pour les iris de l'espèce `versicolor`.
5. Ajouter dans le jeu de données la variable `Sum_Petal` qui correspond à la somme de `Petal.Width` et `Sepal.Width`.
6. Calculer la moyenne et la variance de la variable `Sepal.Length` pour chaque espèce.

## Aviation

Nous considérons la table `hflights` qui contient des informations sur les vols au départ des aéroports *Houston George Bush Intercontinental Airport* (IATA: IAH) et *William P. Hobby Airport* (IATA: HOU),

```
library(hflights)
hflights <- as_tibble(hflights)
```

1. Sélectionner les variables qui se situent entre `Origin` et `Cancelled` de différentes façons.
2. Sélectionner les variables `DepTime`, `ArrTime`, `ActualElapsedTime`, `AirTime`, `ArrDelay` et `DepDelay`.
3. Ajouter une variable `ActualGroundTime` qui correspond à `ActualElapsedTime` moins `AirTime`, et ne conserver que ces trois variables.
4. Ajouter une variable `AverageSpeed` qui donne la vitesse moyenne du vol, ne conserver que les variables `Origin`, `Dest`, `Distance`, `AirTime` et `AverageSpeed`, puis ordonner la table selon les valeurs décroissantes de cette variable.
5. Sélectionner les vols à destination de `JFK`, et ne conserver que les variables `FlightNum`, `Origin` et `Dest`.
6. Calculer le nombre de vols à destination de `JFK`.
7. Créer un résumé de `hflights` qui contient :
  - `n` : le nombre total de vols ;

- `n_dest`: le nombre total de destinations ;
- `n_carrier` : le nombre total de compagnies.

8. Créer un résumé de `hflights` qui contient, pour les vols de la compagnie **AA** :

- `n` : le nombre total de vols ;
- `n_cancelled` : le nombre total de vols annulés ;
- `mean_delay` : la valeur moyenne de `ArrDelay` (attention à la gestion des NA).

9. Calculer pour chaque compagnie :

- `n` : le nombre total de vols ;
- `mean_air_time` : la valeur moyenne de `AirTime`.

10. Ordonner les compagnies en fonction des retards moyens au départ.

## Tennis

Nous considérons les données sur les résultats de tennis dans les tournois du Grand Chelem en 2013. Les données, ainsi que le **descriptif des variables**, se trouvent à l'adresse suivante :

<https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics>

Nous considérons d'abord le tournoi masculin de Roland Garros. Utiliser les verbes de `dplyr` pour répondre aux questions suivantes.

1. Importer les données.
2. Afficher le nom des adversaires de Roger Federer.
3. Afficher le nom des demi-finalistes. (7: Finale, 6: Demi-finale, ...)
4. Combien y a-t-il eu de points disputés en moyenne par match ? Il faudra penser à ajouter dans la table une variable correspondant au nombre de points de chaque match (verbe `mutate`).
5. Combien y a-t-il eu d'aces par match en moyenne ?
6. Combien y a-t-il eu d'aces par match en moyenne à chaque tour ?
7. Combien y a-t-il eu de doubles fautes au total dans le tournoi ?
8. Importer les données pour le tournoi de Wimbledon masculin de 2013.
9. Concaténer les tables en ajoutant une variable permettant d'identifier le tournoi. On pourra utiliser `bind_rows()` avec l'option `.id`.
10. Afficher les matchs de Federer pour chaque tournoi.
11. Comparer les nombres d'aces par matchs à chaque tour pour les tournois de Roland Garros et Wimbledon.

## Expressions régulières

À l'aide de la fonction `str_extract()` et d'une expression régulière, extraire la partie d'une chaîne de caractères située entre deux parenthèses. Par exemple, si la chaîne de caractères est `J'aime (beaucoup) R`, il faudra obtenir `beaucoup`.