Exercices - Semaine 3

Raphaël Nedellec

Préparation

Démarrez un nouveau projet rstudio intitulé TP3. Nous allons avoir besoin de plusieurs librairies aujourd'hui. Installez les en utilisant la commande suivante:

```
install.packages(c("stringr", "lubridate", "arrow", "dplyr", "tidyr",
 "jsonlite", "RSQLite"))
```

À la racine du projet, créez un dossier data. Téléchargez les données associées au TP3 et décompressez les dans le dossier data. Supprimez le fichier .zip.

Exercice

Import et lecture des données

1. Listez les fichiers présents dans le dossier data. Quelles sont les extensions des différents fichiers ?
2. Lisez le fichier .parquet en utilisant la librairie arrow. Quelles sont les dimensions de la table ? Quelles sont les colonnes présentes ?
3. Lisez le fichier .json en utilisant la librairie jsonlite. Quelles sont les dimensions de la table ? Quelles sont les colonnes présentes ?
4. Importez la librairie RSQLite, et ouvrez une connexion à la base de données sqlite en utilisant la fonction `dbConnect`. Le driver à utiliser sera `SQLite()`. Quelles sont les tables présentes dans la table ? Vous pourrez utiliser la fonction `dbListTables`.
5. Créez deux nouvelles tables dans la base de données à l'aide de la fonction `dbWriteTable`. Les tables s'appeleront respectivement `olympics_athletes` et `tokyo_athletes` pour les fichiers `olympics_athletes.json` et `tokyo_athletes.parquet`.
6. Inspectez la table `olympics_athletes` en utilisant la fonction `dbListFields`. Quelles sont les colonnes de la table ?
7. Importez cette table depuis la base de données en utilisant la fonction `dbReadTable`. Convertissez la table en tibble en utilisant la fonction `as_tibble`.

dplyr, tidyr

Dans les questions suivantes, utilisez en priorité les fonctions des packages dplyr, tidyr.

8. Convertissez la colonne `Sex` en variable catégorielle avec la fonction `mutate`.
9. Créez deux colonnes à partir de la colonne `Games`. La première colonne `Year` sera une colonne de type `integer` contenant l'année des jeux. La deuxième colonne `isSummer` sera une colonne booléenne qui indiquera si les jeux sont des jeux d'été ou d'hiver. Vous pourrez utiliser la fonction `separate_wider_delim` de tidyr notamment.

Les questions suivantes nécessitent l'application de plusieurs instructions en séquence. Essayez tant que possible de chaîner les instructions avec des pipes (`%>%` ou `|>`).

10. Calculez la moyenne d'age des athletes pour chacune des éditions des JO d'été. Quelle édition a compté les athlètes les plus jeunes ? Les plus vieux ?
11. Quelle est la discipline des JO d'été dont la taille des athlètes féminines est la plus grande ? Quelle a été cette discipline au cours de chacune des éditions ? Calculez le nombre de fois où chaque discipline a été la discipline avec les plus grandes athlètes.

stringr, lubridate

Vous disposez du texte suivant :

> Les jeux olympiques d'été se déroulent normalement tous les 4 ans, durant les mois de Juillet et Août. Les jeux de Rio ont eu lieu du 5 Août 2016 au 20 Août 2016, ceux de Tokyo du 23 Juillet 2021 au 8 Août 2021, et ceux de Paris auront lieu du 26 Juillet 2024 au 11 Août 2024. Plus de 10000 athlètes sont attendus du monde entier à Paris.

12. En utilisant les fonctions du packages `stringr`, extrayez les dates des différentes éditions des JO. Aide : définissez une expression régulière capable de capturer une séquence de caractères représentant une date telle que "26 Juillet 2024". Vous pourrez utiliser cette *regex* avec la fonction `str_extract_all` par exemple.
13. Remplacez les noms des mois par leur numéro pour rendre ces éléments convertibles en date en utilisant la fonction `str_replace`.
14. Convertissez ces 3 éléments en date en utilisant la fonction appropriée du package `lubridate`.
15. Combien de jours ont séparés les éditions de Rio et Tokyo ? Et sépareront les éditions de Tokyo et de Paris ? Faites le même calcul en semaines.