Exercices - Semaine 2

Raphaël Nedellec

Préparation

Démarrez un nouveau projet rstudio intitulé TP2. Installez la library `rvest` en utilisant la commande `install.packages("rvest")`. Cette librairie sera utilisée pour lire des tables de données directement depuis le web. De même, installez le package `purrr`.

Exercice

1. Lancez la commande suivante :

```
list_tables <-
  session("https://fr.wikipedia.org/wiki/Liste_des_m%C3%A9daill%C3%A9s_olympiques") |>
  html_elements(".wikitable") |>
  html_table()
```

Que s'est-il passé ? Que contient l'objet `list_tables` ?

2. Nous allons tout d'abord nous intéresser à la première table. Créez un objet intitulé `data_medailles_sport_ete` contenant le premier élément de `list_tables`. La table n'est pas bien formattée. Supprimez la première colonne, les noms de colonnes et la première ligne. Renommez les colonnes en `c("Discipline", "Annees", "Editions", "Epreuves_2020", "Or", "Argent", "Bronze", "Total", "Athletes_medailles", "Athletes_or")`. Les colonnes `Editions`, `Epreuves_2020`, `Or`, `Argent`, `Bronze`, `Total` seront converties en colonnes d'entiers.
3. Quelles sont les 3 disciplines avec le plus de médailles distribuées depuis le début de l'histoire des jeux olympiques ?
4. Quelles sont les disciplines avec le moins d'épreuves en 2020 ?

5. La colonne `Editions` nous renseigne sur le nombre total d'apparence des disciplines aux JO d'été. Nous souhaitons vérifier ce calcul en implémentant notre propre fonction `calcul_nb_editions_int`. Dans un premier temps, la fonction `calcul_nb_editions` prendra en entrée un paramètre `depuis`, de type entier, qui représente depuis quelle année la discipline est au programme.
6. Dans certains cas, les disciplines ont été au programme de façon discontinue. Proposez une nouvelle fonction `calcul_nb_editions_str` qui prendra cette fois-ci en entrée des chaînes de caractères. Par exemple, l'appel suivant:

```
calcul_nb_editions_str("1896, 1904, depuis 1920")
```

retournera la valeur 26.

7. Définissez une fonction générique `calcul_nb_editions` et deux implémentations `calcul_nb_editions.integer` et ‘`calcul_nb_editions.character`. Quels résultats donnent les appels :

```
calcul_nb_editions(2000)
calcul_nb_editions("1904-1924, depuis 1948")
```

?

8. En Athlétisme, le Finlandais Paavo Nurmi détient le record de médailles avec 12 médailles obtenues lors des JO d'hiver.

Implémentez une fonction `calcul_medailles_individuelles` qui détermine le nombre de médaille maximal a été obtenu par un athlète lors d'olympiades. Note : s'il y a plusieurs athlètes à égalité, alors la cellule comporte plusieurs éléments, et une manipulation de la chaîne de caractères est nécessaire.

9. Quel est le top 3 des athlètes ? Vous utiliserez la fonction `lapply` pour appliquer la fonction `calcul_medailles_individuelles` à chaque élément de la colonne `Athletes_medailles`.
10. Quels sont les 3 nationalités les plus représentées, pour toutes les épreuves, au classement du nombre de médailles d'or individuelles reccueillies ?

- Pour ce faire, vous implémenterez une fonction `extraire_nationalite_athlete` qui à partir d'une chaine de charactère extraira le nombre d'athlète et leur nationalité. Par exemple, la chaine de caractère “Paavo Nurmi (FIN) (9-3-0) Carl Lewis (USA) (9-1-0)” donnera en sortie `c("FIN" = 1, "USA" = 1)`.
- Vous utilisez la fonction lapply pour appliquer la fonction à toute la colonne

- Vous agrégerez les résultats de manière à sommer toutes les lignes et à obtenir les 3 nations les plus représentées, et leur effectif.

11. Intéressez-vous désormais au deuxième tableau contenu dans `list_tables`, faisant référence aux JO d'hiver. Appliquez les fonctions `calcul_medailles_individuelles` et `extraire_nationalite_athlete` aux deux dernières colonnes, à la manière des questions 9 et 10. Cette fois-ci, vous utiliserez la fonction appropriée du package `purrr` en lieu et place de `lapply`. Quelles sont les résultats ? Quelle différence voyez-vous entre `lapply` et les fonctions de `purrr` ?