Chapitre 3: Manipulation des données avec R

Aboubacar HEMA

1/12/2022

qu'est ce que c'est dplyr

dplyr est une grammaire de manipulation de données, fournissant un ensemble cohérent de verbes qui vous aident à résoudre les défis de manipulation de données les plus courants:

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

dplyr::select

Sélectionner (et renommer éventuellement) des variables dans un data.frame, en utilisant un mini-langage concis qui facilite la référence aux variables en fonction de leur nom (par exemple, sélectionner toutes les colonnes de a à gauche à f à droite). Vous pouvez également utiliser des fonctions de **prédicat** comme **is.numeric** pour sélectionner des variables en fonction de leurs propriétés.

Les sélections de dplyr implémentent un dialecte de R où les opérateurs facilitent la sélection variables:

** : pour sélectionner une plage de variables consécutives.

** ! pour prendre le complément d'un ensemble de variables.

** & et | pour sélectionner l'intersection ou l'union de deux ensembles de variables.

** c () pour combiner des sélections.

De plus, vous pouvez utiliser des aides à la sélection. Certains assistants(helpers) sélectionnent des colonnes spécifiques:

** everything(): correspond à toutes les variables.

** last_col(): Sélectionnez la dernière variable

Ces assistants sélectionnent des variables en faisant correspondre des modèles dans leurs noms:

starts_with(): commence par un pre x.

ends_with(): se termine par un su x.

contains(): contient une chaîne littérale.

matches(): correspond à une expression régulière.

num_range(): correspond à une plage numérique telle que x01, x02, x03.

Ces assistants sélectionnent des variables à partir d'un vecteur de caractères:

all_of(): correspond aux noms de variables dans un vecteur de caractères. Tous les noms doivent être présents, sinon, une erreur hors limites est émise.

all_of(): Identique à tout (), sauf qu'aucune erreur n'est renvoyée pour les noms qui ne exister.

Cet assistant sélectionne les variables avec une fonction:

where(): applique une fonction à toutes les variables et sélectionne celles pour lesquelles la fonction renvoie TRUE.

dplyr::filter

La fonction filter () est utilisée pour sous-ensemble un bloc de données, en conservant toutes les lignes qui satisfont votre conditions. Pour être conservée, la ligne doit produire une valeur TRUE pour toutes les conditions. Remarque que lorsqu'une condition est évaluée à NA, la ligne sera supprimée, contrairement au sous-ensemble de base avec [.

Fonctions de filtre utiles Il existe de nombreuses fonctions et opérateurs utiles lors de la construction des expressions utilisé pour filtrer les données:

** ==, >, >= etc

** &, |, !, xor()

** is.na()

** between(), near()

dplyr::mutate

mutate () ajoute de nouvelles variables et préserve celles existantes; transmute () ajoute de nouvelles variables et supprime ceux existants. Les nouvelles variables écrasent les variables existantes du même nom. Les variables peuvent être supprimées en définissant leur valeur sur NULL.

Useful mutate functions +, -, log(), etc., for their usual mathematical meanings lead(), lag() dense_ank(), min_ank(), percent_rank(), row_number(), cume_dist(), ntile() cumsum(), cummean(), cummin(), cummax(), cumany(), cumall() na if(), coalesce() if else(), recode(), case when()

dplyr::group_by & summarise

La plupart des opérations sur les données sont effectuées sur des groupes définis par des variables. group by () prend un data.frame existant et le convertit en un data.frame groupé où les opérations sont effectuées "par groupe".

ungroup () supprime le regroupement

summarise () crée un nouveau data.frame. Il aura une (ou plusieurs) lignes pour chaque combinaison des variables de regroupement (group_by); s'il n'y a pas de variables de regroupement, la sortie aura une seule ligne résumant toutes les observations dans l'entrée. Il contiendra une colonne pour chaque regroupement variable et une colonne pour chacune des statistiques récapitulatives que vous avez spécifiées. summary () et summary () sont des synonymes.

Les jointures (mutate-join & filter-join)

Les jointures de type mutate-join ajoutent des colonnes de y à x, correspondant aux lignes en fonction des clés:

inner_join(): inclut toutes les lignes de x et y.

left_join():: inclut toutes les lignes de x.

right_join():inclut toutes les lignes de y.

full_join (): inclut toutes les lignes de x ou y.

Si une ligne de x correspond à plusieurs lignes de y, toutes les lignes de y seront renvoyées une fois pour chaque ligne correspondante dans x.

tidyr::gather & spread