dplyr

Benjamin GUIGON

January 2021

1 Lien

Article écrit par Jiayue LIU, Soukaina ELGHALDY

https://github.com/liu-jiayue/psbx

2 Introduction

dplyr fait par de la très grand librairy *tidyverse* qui sert majoritairement à l'exploration de données. Ce qui est très important pour le pre-precessing des data scientists.

Les avantages de dplyr sont : sa rapidité, son ergonomie, la synergie avec tout la librairy *tidyverse*

3 Constitution

tidyverse regroupe 7 packages :

- ggplot2 (pour la visualisation des données)
- dplyr (pour la manipulation des données)
- tidyr (pour la remise en forme des données)
- purrr (pour la programmation)
- readr (pour l'importation de données)
- tibble (pour les tableaux de données)
- forcats (pour les variables qualitatives)
- stringr (pour les chaînes de caractères)

4 Nomenclature

Afin d'être plus userfriendly, *tidyverse* a créé un type de donnée appelé *"tidy data* qui est un mélange entre les DataFrame de pandas et les tables SQL.

Ces dataFrame sont appelés *tribble*. On peut facilement les transformer en DataFrame classique et inversement.

5 Manipulation

Il existe toute une panoplie de manipulation SQL'like : Slice, Filter, Select, Arrange Mutate etc...

exemple : "filter(data, age ¿= 70, age ¡= 80)", on retourve un peu les memes utilisations que sur pandas ou SQL mais avec certaines facilités en plus comme "select(data, starts_with("couleur"))" qui nous exemptent d'utiliser les RegEx.

Un point très interessant pour alleger tous les codes R en général, est l'utilisation de l'opérateur pipe % > %. Il permet de faire comprendre à R que la valeur avant le pipe sera automatiquement prise comme 1er paramètre des fonctions qui suivent, ca évite de faire des commandes aussi imbriquées que celle la : *arrange(select(filter(data, age == 46), nom, age), age)*.

6 Avis

Tidyverse est l'une des librairy les plus utilisées et l'une des plus utilise pour l'exploration des données.

Le travail est très pertinant, il explique très bien tous les aspects important notamment avec les opérations sur les *tribble*. Il résume assez rapidement tout ce qui est utilise pour gérer un dataset sur R sans difficulté. Bien que ca ne soit pas une librairy Data science pure et dure, il est important de bien maitriser toutes les bases.