Janitor

Benjamin GUIGON

January 2021

1 Lien

Article écrit par Olfa Lamti

https://github.com/OlfaLmt/PSBX/tree/main/Janitor

2 Introduction

Janitor a pour fonction principale d'examiner et de nettoyer des jeux de données de manière rapide. Avant toute application de modèle de data science, il ne faut jamais négliger un pré processing précis. Ce package Janitor va nous permettre de faire du ménage assez facilement.

3 Fonctions principales

Une fonction très utilse est la fonction *clean*. Elle va permettre de rendre plus lisible et accessible les données. Elle ne va pas faire de miracle si le data set est vraiment défini n'importe comment, mais elle va permettre de gagner en compréhension pour un pré processing plus efficace.

```
x = janitor::clean_names(mymsa)
data.frame(mymsa = colnames(mymsa), x = colnames(x))
```

Une fois le data set clean au niveau des noms des colonnes, il est bien d'avoir un a priori sur valeurs de ces catégories en colonne. Janitor offre la possibilité de connaitre les les fréquences :
tabyl(x, meat_colour)
Cela nous renvoie un dataframe dans lequel est associé le tag de la valeur, le nombre de fois ou la valeur est apparue, et la % par rapport au reste du data set.

Pour gagner en flexibilité on peut en effet utiliser l'opérateur pipe % > % de la librairy *dplyr*.

Une fonction primordiale pour le pré processing est la fonction *remove_empty*, elle va permettre de supprimer certaines colonne vides qui pourraient géner le modèle. Il existe un pipeline de commande qui permet à partir d'un fichier brouille de le transformer avec toutes les commmandes précédentes :

x = read_excel("Name_fichier.xlsx") % ¿ % clean_names() % ¿ % remove_empty().

Il y'a tout un panoplie de fonction assez pratique comme : *adorn_totals* pour avoir une colonne total, *adorn_percentages* pour avoir les pourcentages, *get_dupes* pour avoir les doublons, *excel_numeric_to_date* pour avoir l'encodage de date comme etc...

4 Avis

Il existe plusieurs package R qui servent à l'exploration et le processing des données. Certains se suffisent et d'autre s'utilisent en synergie comme *Janitor* et *dplyr*. En effet l'utilisation des fonctions de *Janitor* plus le pipe de *dplyr* on peut faire un pré processing très efficace, rapide et flexible.

Cet article est bien rédigé et montre les fonctions les plus importantes pour nettoyer des données. Une bonne idée aurait été de prendre un data set que nous voulions étudier avec une technique de machine learning définie et montrer comment avec *Janitor* on peut le pré processer rapidement pour donner un data set exploitable.