Le travail de JEAN SOURIS **est disponible ici**

Théo Marié

20/12/2020

SYNTHESE DU TRAVAIL EN QUESTION

DPLYR est un package interessant par ses fonctionnalité de traitement et de modifications de bases de donnés, c'est indispensable dans le mode professionel du data analyst. Ici, le travail est ordonné, avec une introduction et 3 exemples de fonctionnalité Slice, Select et Rename. Sa simplicité montre son efficacité. Le code est bien réalisé, mais le pdf est illisible. Il est important de bien présenter son travail pour pouvoir être compris de tous, surtout de ceux qui ne maitrisent pas le langage python.

EXTRAIT COMMENTE DES PARTIES DU CODE

```
library("dplyr")
library("nycflights13")
data(airports)
airports %>% slice_sample(n=6)
```

```
## # A tibble: 6 x 8
##   faa   name                        lat    lon   alt    tz dst   tzone
##   <chr> <chr>                     <dbl>  <dbl> <dbl> <dbl> <chr> <chr>
## 1 KVC   King Cove Airport          55.1 -162.    155    -9 A     America/Anchor~
## 2 PTA   Port Alsworth Airport      60.2 -154.    253    -9 A     America/Anchor~
## 3 AGS   Augusta Rgnl At Bush Fld   33.4  -82.0   144    -5 A     America/New_Yo~
## 4 MGY   Dayton-Wright Brothers A~  39.6  -84.2   957    -5 U     America/New_Yo~
## 5 HNS   Haines Airport             59.2 -136.     15    -9 A     America/Anchor~
## 6 APC   Napa County Airport        38.2 -122.     35    -8 A     America/Los_An~
```

Ici, la fonction Slice est utisée dans sa version slice_sample, qui nous permet de selectionner aléatoirement. Ici, il est demandé avec n=6 de choisir 6 lignes aléatoirement dans la base de données "airports".

```
data(flights)
select(flights, -origin, -time_hour)
```

```
## # A tibble: 336,776 x 17
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      533            529         4      850            830
##  3  2013     1     1      542            540         2      923            850
##  4  2013     1     1      544            545        -1     1004           1022
##  5  2013     1     1      554            600        -6      812            837
##  6  2013     1     1      554            558        -4      740            728
##  7  2013     1     1      555            600        -5      913            854
```

```
##  8  2013     1     1      557            600        -3      709            723
##  9  2013     1     1      557            600        -3      838            846
## 10  2013     1     1      558            600        -2      753            745
## # ... with 336,766 more rows, and 9 more variables: arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, dest <chr>, air_time <dbl>,
## #   distance <dbl>, hour <dbl>, minute <dbl>
```

Il est ici question de la focntion “select”, et plus précisément de la fonctionnalité qui sert à présenter un tableau sans certaines colonnes.SI par exemple, on veut présenter un tableau général sur un produit, sans vouloir afficher son numéro d’identification, on peut avoir recours à cette fonctionnalité. Dans la table flight, il enlève la colonne “origin” et “time_hour”.

```
rename(airports, altitude = alt, time_zone = tzone)
```

```
## # A tibble: 1,458 x 8
##    faa   name                   lat    lon altitude    tz dst   time_zone
##    <chr> <chr>                <dbl>  <dbl>    <dbl> <dbl> <chr> <chr>
##  1 04G   Lansdowne Airport     41.1  -80.6     1044    -5 A     America/New_Yo~
##  2 06A   Moton Field Municipa~  32.5  -85.7      264    -6 A     America/Chicago
##  3 06C   Schaumburg Regional    42.0  -88.1      801    -6 A     America/Chicago
##  4 06N   Randall Airport        41.4  -74.4      523    -5 A     America/New_Yo~
##  5 09J   Jekyll Island Airport  31.1  -81.4       11    -5 A     America/New_Yo~
##  6 0A9   Elizabethton Municip~  36.4  -82.2     1593    -5 A     America/New_Yo~
##  7 0G6   Williams County Airp~  41.5  -84.5      730    -5 A     America/New_Yo~
##  8 0G7   Finger Lakes Regiona~  42.9  -76.8      492    -5 A     America/New_Yo~
##  9 0P2   Shoestring Aviation ~  39.8  -76.6     1000    -5 U     America/New_Yo~
## 10 0S9   Jefferson County Intl  48.1 -123.       108    -8 A     America/Los_An~
## # ... with 1,448 more rows
```

Grace à la fonctionnalité Rename, il nous explique que l’on peut renommer des noms de colonne detableau. Si une colonne que l’on utilise est le fruit d’une addition de 2 valeurs, alors il est important de renommer cette colonne pour bien comprendre ce qu’elle exprime (par exemple PRIX HT + TVA = PRIX TTC). Il utilise cette fonctionnalité ici pour renommer la colonne altitude par alt, et la colonne time_zone par tzone.

EVALUATION DU TRAVAIL EN QUESTION

Critère 1 : Visuellement apréciable sur pdf 1/4 très peu agréable à regarder

Critère 2 : idées pour faire le code 3/4 De beaux exemples, de belles idées

Critère 3 : Fonctionnalité du code 4/4 Tout fonctionne très bien

Critère 4 : lisibilité du code 4/4 Le code est très bien écrit, espacé.

Critère 5 : explications données 3/4 Les explications sont claires

CONCULSION

Ce travail est d’une efficacité remarquable, il est simple et utile. Jean nous a montré certaines fonctionnalités basiques mais très pratiques. Malgrès un pdf illisible, il nous offre ici un très bon travail.