Exploiter {janitor} pour nettoyer les données

Marie Vaugoyeau

12 November 2024

Table of contents

1 import des packages

```
library(tidyverse)
library(janitor)
```

2 import des données

Les données ont été créées pour l'occasion.

```
data <- readxl::read_xlsx("data/donnees.xlsx")
```

3 Regardons un peu les données

```
glimpse(data)
```

```
Rows: 22
Columns: 8
$ `Prénom Patient.e` <chr> "Paula", "Pierre", "Antoine", "Adrien", "Alice", "S~
$ `Sexe / genre`     <chr> "F", "M", "M", "M", "F", "F", "M", "F", "M", "F", "~
$ date               <chr> "2024-01-01", "2024-01-16", "2024-01-31", "2024-02-~
$ `Album in\r\ng/dL` <dbl> 3.6, 3.9, 3.6, 3.9, 4.1, 3.8, 3.7, 3.7, 3.7, 3.4, 3~
$ `Fructose mg/dL`   <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Glucose mg/dL`    <dbl> 92.85714, 89.14286, 89.88571, 85.42857, 96.57143, 8~
$ `Na (mmol/L)`      <dbl> 150, 150, 150, 150, 150, 150, 150, 150, 150, 150, 1~
$ `Globulin g/dL`    <dbl> 2.3, 2.0, 2.0, 2.0, 1.9, 2.0, 2.2, 2.1, 2.1, 2.6, 2~
```

```
summary(data)
```

```
 Prénom Patient.e   Sexe / genre           date            Album in\r\ng/dL
 Length:22          Length:22          Length:22          Min.   :3.400
 Class :character   Class :character   Class :character   1st Qu.:3.700
 Mode  :character   Mode  :character   Mode  :character   Median :3.800
                                                          Mean   :3.809
                                                          3rd Qu.:3.900
                                                          Max.   :4.200
 Fructose mg/dL Glucose mg/dL    Na (mmol/L)  Globulin g/dL
 Mode:logical   Min.   :81.71   Min.   :150   Min.   :1.900
 NA's:22        1st Qu.:84.87   1st Qu.:150   1st Qu.:2.000
                Median :86.91   Median :150   Median :2.100
                Mean   :87.39   Mean   :150   Mean   :2.109
                3rd Qu.:89.70   3rd Qu.:150   3rd Qu.:2.175
                Max.   :96.57   Max.   :150   Max.   :2.600
```

4 Améliorer les noms des colonnes

```
donnees <- readxl::read_xlsx("data/donnees.xlsx") |>
  clean_names(case = "big_camel")
```


```
donnees <- readxl::read_xlsx("data/donnees.xlsx") |>
  clean_names() |>
  mutate(
    prenom_patient_e = make_clean_names(prenom_patient_e)
  )
```

5 retirer les colonnes vides

```
donnees <- readxl::read_xlsx("data/donnees.xlsx") |>
  clean_names() |>
  mutate(
    prenom_patient_e = make_clean_names(prenom_patient_e)
  ) |>
  remove_constant()
```

6 traiter les dates excel

```
donnees <- readxl::read_xlsx("data/donnees.xlsx") |>
  clean_names() |>
```

```
  mutate(
    prenom_patient_e = make_clean_names(prenom_patient_e)
  ) |>
  remove_constant() |>
  mutate(
    date =
      case_when(
      str_detect(date, "-") ~ date,
      TRUE ~ date |>
        as.numeric() |>
        excel_numeric_to_date() |>
        as.character()
    ) |>
      ymd()
  )
```

7 créer des tableaux résumés rapidement

```
tabyl(donnees, date)
```

```
       date n    percent
 2024-01-01 1 0.04545455
 2024-01-10 1 0.04545455
 2024-01-12 1 0.04545455
 2024-01-13 1 0.04545455
 2024-01-15 1 0.04545455
 2024-01-16 3 0.13636364
 2024-01-18 1 0.04545455
 2024-01-19 2 0.09090909
 2024-01-31 1 0.04545455
 2024-02-15 1 0.04545455
 2024-02-16 1 0.04545455
 2024-02-17 1 0.04545455
 2024-02-18 1 0.04545455
 2024-03-01 1 0.04545455
 2024-03-02 1 0.04545455
 2024-03-03 1 0.04545455
 2024-03-04 1 0.04545455
 2024-04-11 1 0.04545455
```

```
2024-04-16 1 0.04545455
```

```
tabyl(donnees, sexe_genre, date) |>
  adorn_totals(where = c("row", "col")) |>
  adorn_percentages() |>
  adorn_pct_formatting(digits = 0) |>
  adorn_ns() |>
  adorn_title()
```

```
                date
 sexe_genre 2024-01-01 2024-01-10 2024-01-12 2024-01-13 2024-01-15 2024-01-16
          F     8% (1)     0% (0)     0% (0)     0% (0)     8% (1)    17(2)
          M     0% (0)    10% (1)    10% (1)    10% (1)     0% (0)    10(1)
      Total     5% (1)     5% (1)     5% (1)     5% (1)     5% (1)    14(3)

 2024-01-18 2024-01-19 2024-01-31 2024-02-15 2024-02-16 2024-02-17 2024-02-18
     8% (1)    17% (2)     0% (0)     0% (0)     0% (0)     0% (0)     8(1)
     0% (0)     0% (0)    10% (1)    10% (1)    10% (1)    10% (1)     0(0)
     5% (1)     9% (2)     5% (1)     5% (1)     5% (1)     5% (1)     5(1)

 2024-03-01 2024-03-02 2024-03-03 2024-03-04 2024-04-11 2024-04-16     Total
     8% (1)     8% (1)     8% (1)     0% (0)     0% (0)     8% (1) 100 (12)
     0% (0)     0% (0)     0% (0)    10% (1)    10% (1)     0% (0) 100 (10)
     5% (1)     5% (1)     5% (1)     5% (1)     5% (1)     5% (1) 100 (22)
```