TD n° 7 : Tidyverse (partie 3)

Exercice 1. On considère la librairie `janitor` dont la description est disponible ici:

`https://www.rdocumentation.org/packages/janitor/versions/2.2.0`

Comprendre et reproduire les commandes associées à l'exemple du jeu de données `dirty_data.xlsx`.

Exercice 2. Le but de ce TD est de netoyer des données avec la librarie `janitor` et les acquis sur le tidyverse. Pour ce faire, on travaille avec le jeu de données `raw_data.csv` disponible ici:

`https://chesneau.users.lmno.cnrs.fr/raw_data.csv`

1. À l'aide de `read.csv2`, mettre le jeu de données dans une data frame notée `rd`, en précisant: `stringsAsFactors = TRUE`.
2. Visualiser les données et réfléchisser aux potentielles choses à nettoyer.
3. Reproduire et comprendre l'enjeu des commandes suivantes (certaines commandes :

```
library(janitor)
rd2 = rd
rd2 = clean_names(rd2)
str(rd2)

library(tidyverse)
rd2 = rd2 %>%
    rename(code_postal=code_po_stal)

names(rd2)
levels(rd2$gender)

rd3 = rd2
rd3$gender = str_to_lower(rd3$gender)
rd3$gender
rd3$gender = str_trim(rd3$gender, side="both")
rd3$gender
rd3$gender = as.factor(rd3$gender)
levels(rd3$gender)

rd3$code_postal
rd3$code_postal = str_pad(rd3$code_postal, 5, "left", "0")

rd3$code_postal
class(rd3$code_postal)

rd3 = rd3 %>%
```

```
        separate(nom_prenom,c("nom", "prenom"), sep="_")
    rd3

    rd3 = rd3 %>%
        mutate(nom=str_to_title(nom),
               prenom=str_to_title(prenom))
    rd3

    names(rd)
    names(rd2)

    rd4 = rd3
    names(rd3) = str_replace_all(names(rd3),"_",".." )
    names(rd3)

    rd4 = rd2
    rd4$serum_cholestoral = str_replace(rd4$serum_cholestoral,
    "non renseigne", "NA")
    rd4$serum_cholestoral = as.numeric(rd4$serum_cholestoral)

    str(rd4)
```

Exercice 3. Comprendre et reproduire les commandes suivantes :

```
 messy_data = data.frame(
  ID = c(1, 2, 3, 4, 5, 6, 7, 8),
  FirstName = c("John", "Jane", "Mike", NA, "Emily", "Steve", "Alice", "Bob"),
  LastName = c("Doe", "Smith", "Johnson", "Brown", "Wilson", "Adams", "Roberts",
  "Davis"),
  Age = c(25, 32, NA, 40, 28, 35, 24, "NaN"),
  Email = c("john.doe@email.com", "jane.smith@email.com", "mike.johnson@email.com",
  "invalid-email", "emily@email.com", "steve.adams@email.com",
  "alice.roberts@email.com",  "bob@email.com")
)

messy_data = rbind(messy_data, messy_data[c(1, 3, 5), ])

library(janitor)
library(tidyr)
cleaned_data = messy_data %>%
  distinct() %>%
  drop_na(FirstName, LastName) %>%
  mutate(Age = as.numeric(Age)) %>%
  filter(!is.na(Age))

# Display the cleaned dataset
cleaned_data
```