Optimizing Data Processing of the Medically Assisted Reproduction (MAP) Survey :

A study for INED

ined

INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES

1 Introduction

My name is **Hiatini Tekohuotetua**, I'm passionate about everything to do with data, and I'm applying for a Data Analyst position.

We are solicited for the following objective :

- **Completion and documentation of INED survey data**

To do this, we're going to use R for data wrangling operations.

Short database analysis

We'll start with a brief presentation of the database, which contains information for each individual on the type of treatment taken and the year associated with it.
Each line represents a single observation in the context of assisted reproduction.

```
> head(data,5)
  id traitement_tentative.1 annee_tentative.1 traitement_tentative.2 annee_tentative.2 traitement_tentative.3
1  1          traitement.1              2018          traitement.1              2021                    <NA>
2  2          traitement.1              2022                  <NA>              <NA>                    <NA>
3  3          traitement.2              2020          traitement.2         2005-2016           traitement.3
4  4          traitement.1              2018          traitement.2              2022           traitement.1
5  5          traitement.4              2018          traitement.4              2018           traitement.3
```

Following the audit, there was a reduction in the frequency of treatments and attempts. This leads to a qualitative database, potentially more relevant for in-depth analysis.

Clearance of data

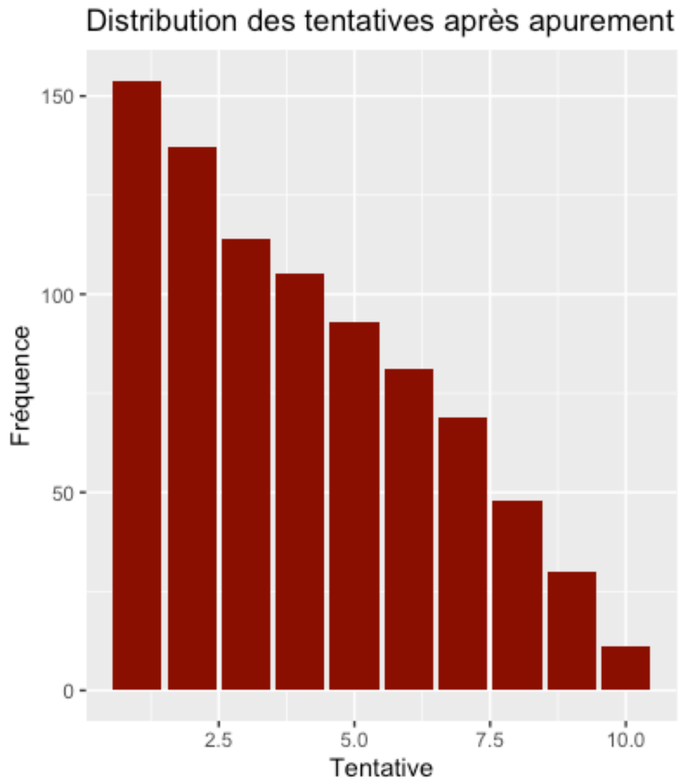
When the data has been cleared, the R code looks like this:

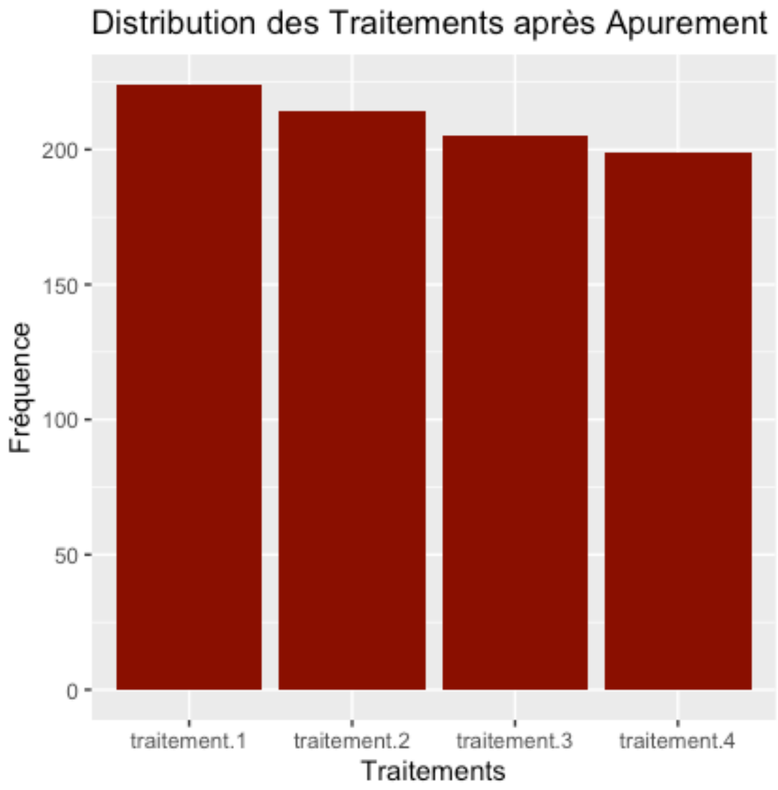
```r
donnees_apurees <- data %>%
  gather(key = "tentative", value = "traitement", -id) %>%
  separate(tentative, into = c("variable", "tentative_num"), sep = "_") %>%
  mutate(tentative_num = as.numeric(stringr::str_extract(tentative_num, "\\d+"))) %>%
  arrange(id, tentative_num) %>%
  group_by(id) %>%
  pivot_wider(names_from = variable, values_from = traitement) %>%
  filter(!is.na(annee) | !is.na(traitement))
```

- Separation of "attempt" columns:
  - Reason: Allows information to be better structured and more easily accessible.
- Mutation of "attempt_num":
  - Reason: Ensures appropriate representation of the attempt number as a numeric variable, facilitating subsequent calculations.
- Observation filtering :
  - Reason: Eliminates observations that guarantee the quality of the data analyzed.

Results analysis

Following the audit, we see a reduction in the frequency of treatments and attempts. This leads to a more relevant, qualitative database.





Conclusion

In conclusion, this process of optimizing the processing of data from the AMP-sans-Frontières survey has been both rewarding and rigorous. By applying solid technical skills, notably in the use of R and specialized packages, I was able to bring a slight clarity and coherence to the data collected.

I'm convinced that my approach would have been even more thorough, and my analyses even more relevant, had I benefited from close collaboration with the teams, accompanied by constructive feedback.

In short, this experience illustrates my ability to meet the complex challenges of data wrangling, combining technical skills, a collaborative approach and meticulous documentation. I remain available to answer further questions or discuss my contribution to this project.