# French given names per year per department

Lucas Mello Schnorr, Jean-Marc Vincent

October, 2021

```
# The environment
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

The aim of the activity is to develop a methodology to answer a specific question on a given dataset.

The dataset is the set of Firstname given in France on a large period of time. given names data set of INSEE, we choose this dataset because it is sufficiently large, you can't do the analysis by hand, the structure is simple

You need to use the *tidyverse* for this analysis. Unzip the file *dpt2019_txt.zip* (to get the **dpt2019.csv**). Read in R with this code. Note that you might need to install the `readr` package with the appropriate command.

## Download Raw Data from the website

```
file = "dpt2020_txt.zip"
if(!file.exists(file)){
  download.file("https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2020_csv.zip",
    destfile=file)
}
unzip(file)
```

## Build the Dataframe from file

```
data <- read_delim("dpt2020.csv", delim =";")
```

```
## Rows: 3727553 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (3): preusuel, annais, dpt
## dbl (2): sexe, nombre
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Translation in english of variables names: sexe -> gender preusuel (prénom usuel) -> Firstname annais (année de naissance) -> Birth year dpt (département) -> department (administrative area unit) nombre -> number

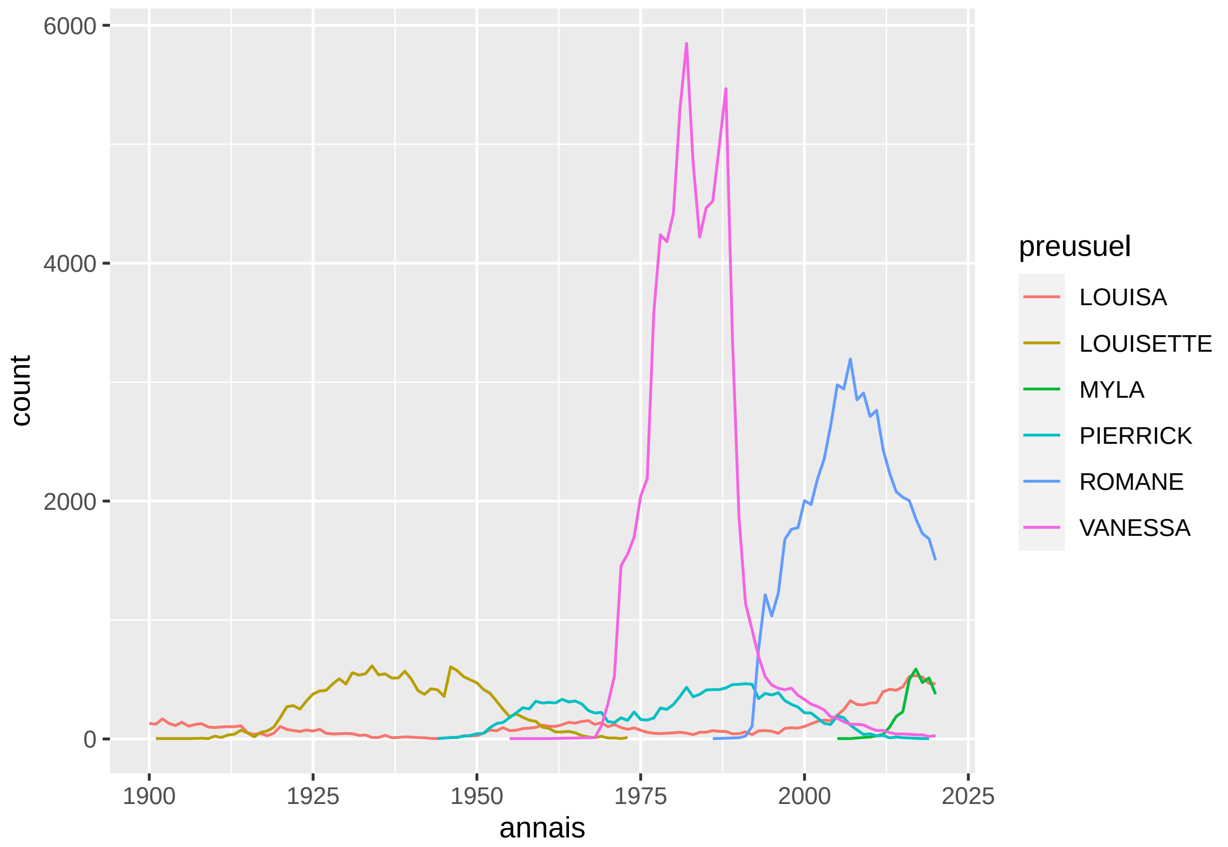
All of these following questions may need a preliminary analysis of the data, feel free to present answers and justifications in your own order and structure your report as it should be for a scientific report.

1. Choose a firstname and analyse its frequency along time. Compare several firstnames frequency
2. Establish, by gender, the most given firstname by year.
3. Make a short synthesis
4. Advanced (not mandatory) : is the firstname correlated with the localization (department) ? What could be a method to analyze such a correlation.

The report should be a pdf knitted from a notebook (around 3 pages including figures), the notebook and the report should be delivered.

#Solution Proposal ## 1. Choose a firstname and analyse its frequency along time. Compare several firstname's frequency

```r
names = c("VANESSA", "PIERRICK", "LOUISETTE", "LOUISA", "MYLA", "ROMANE")
random_data = filter(data, preusuel %in% names & annais != "XXXX")
random_data = random_data %>% group_by(preusuel, annais) %>%
  summarize(count=sum(nombre), .groups = 'drop') %>%
  ungroup()
random_data$annais = as.numeric(random_data$annais)
ggplot(random_data, aes(annais, count, colour = preusuel)) +
  geom_line()
```



**Note: For the next treatments that will be applied on the data, we will apply the following pre-treatments:**

- remove the samples containing 'PRENOMS_RARES' as the firstname.
- delete all the samples with the corresponding year symbolized as 'XXXX'

## 2. Establish, by gender, the most given firstname by year.

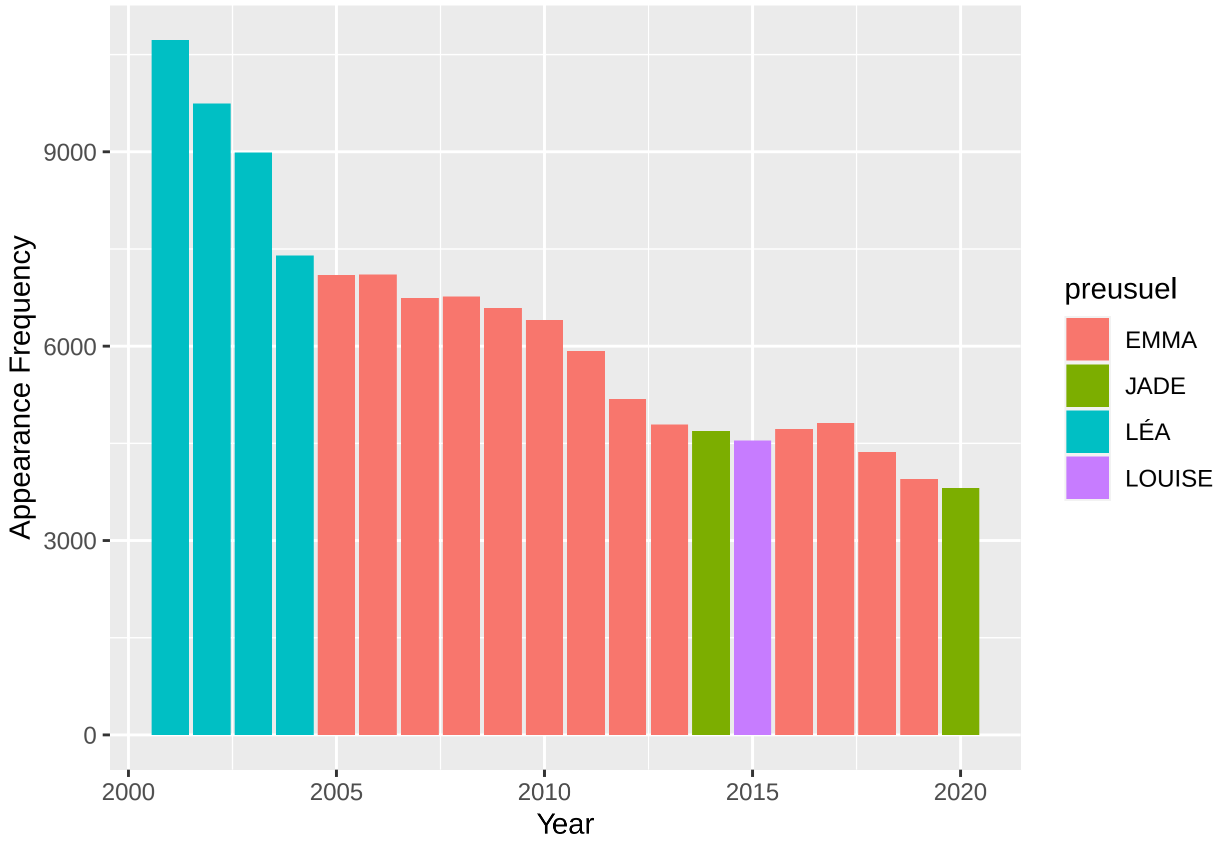
```
filtered_data = filter(data, preusuel != "_PRENOMS_RARES")
filtered_data = filter(filtered_data, annais != "XXXX")
filtered_data$annais = as.numeric(filtered_data$annais)
male = filter(filtered_data, sexe == 1)
female = filter(filtered_data, sexe == 2)
```

```
male_data_plot = male %>% group_by(annais, preusuel) %>%
  summarise(count=sum(nombre), .groups='drop') %>%
  group_by(annais) %>%
  filter(count == max(count)) %>%
  ungroup()
male_data_plot = tail(male_data_plot, n=20)
```

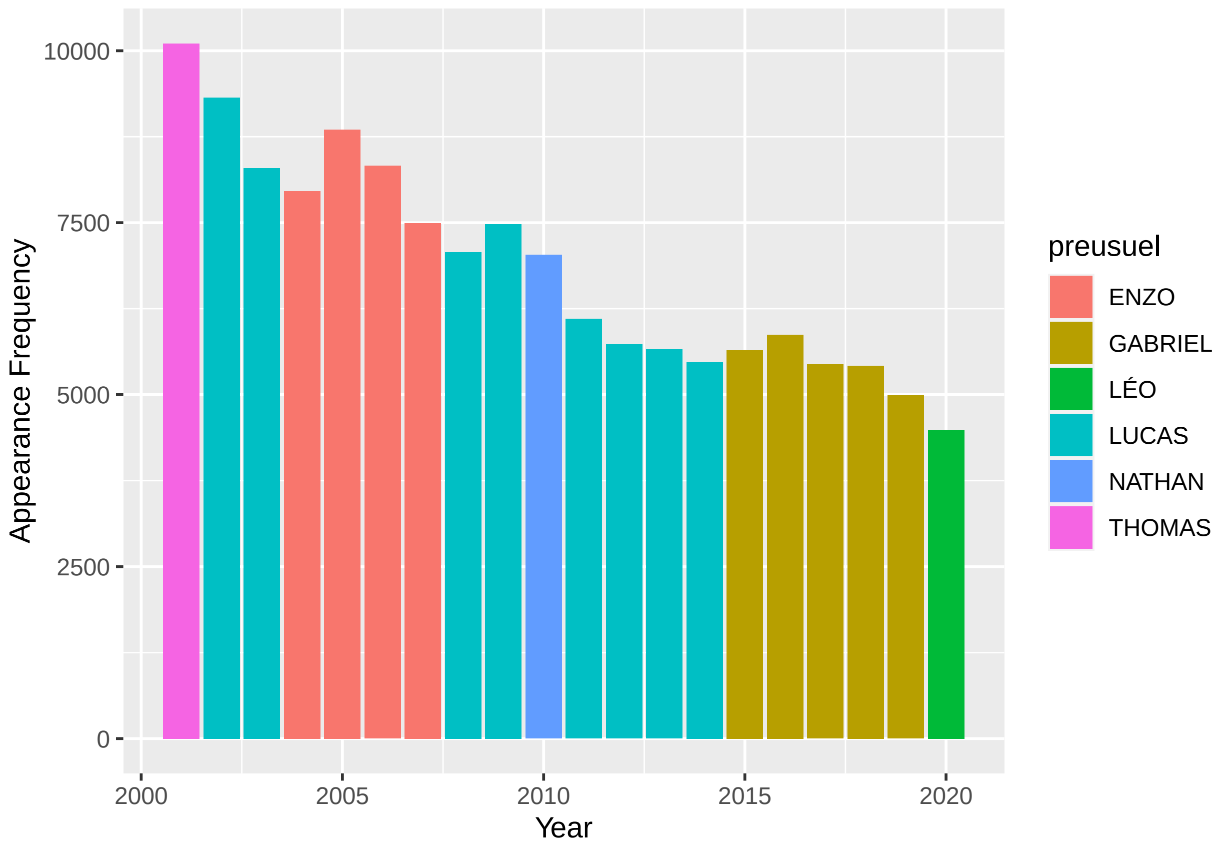
```
female_data_plot = female %>% group_by(annais, preusuel) %>%
  summarise(count=sum(nombre), .groups='drop') %>%
  group_by(annais) %>%
  filter(count == max(count)) %>%
  ungroup()
```

```r
female_data_plot = tail(female_data_plot, n=20)
```

```r
ggplot(female_data_plot, aes(annais, count, fill=preusuel)) + geom_bar(stat="identity") + xlab("Year")
```



```r
ggplot(male_data_plot, aes(annais, count, fill=preusuel)) +
  geom_bar(stat="identity") +
  xlab("Year") +
  ylab("Appearance Frequency")
```



## 3. Make a short analysis

These are the points of our analysis.

- There samples with **_PRENOMS_RARES** were not probably significant at the time of the registration but once accumulated they represent a considerable amount of unknown data.
- Similarly, the attribute **year** faces the same problem and the joint deletion of these unknown attributes lead to drastic scale down in the amount of the data available for the treatment.