03_Names-Methodo2022-exercise

Augusta Mukam

October, 2021

Description of the work

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

Environment packages

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

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     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)
library(readr)
library(dplyr)
```

Description of the dataset

The dataset is the set of Firstname given in France on a large period of time.

We 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)
}
```

We unzip the file *dpt2020_csv.zip* file to get the **dpt2020.csv** file.

```
unzip(file)
```

Build the Dataframe from file

```
FirstNames <- 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.
```

Description of the dataset

To retrieve the full column specification for this data, we use spec()

```
FirstNames
```

```
## # A tibble: 3,727,553 x 5
##     sexe preusuel       annais dpt   nombre
##    <dbl> <chr>          <chr>  <chr>  <dbl>
##  1     1 _PRENOMS_RARES 1900   02         7
##  2     1 _PRENOMS_RARES 1900   04         9
##  3     1 _PRENOMS_RARES 1900   05         8
##  4     1 _PRENOMS_RARES 1900   06        23
##  5     1 _PRENOMS_RARES 1900   07         9
##  6     1 _PRENOMS_RARES 1900   08         4
##  7     1 _PRENOMS_RARES 1900   09         6
##  8     1 _PRENOMS_RARES 1900   10         3
##  9     1 _PRENOMS_RARES 1900   11        11
## 10     1 _PRENOMS_RARES 1900   12         7
## # ... with 3,727,543 more rows
```

Scientific report.

1. **Choose a firstname and analyse its frequency along time. Compare several firstnames frequency**

First, we count the different firstnames of the dataset by decreasing order

We see that **Camille** is the mos frequent firstName, so we decide to study its frequency per year.

Il faut absolument respecter la case sur les noms.

```
frequencies=FirstNames %>% group_by(annais) %>%count(preusuel) %>% arrange(desc(frequency))
```

```
frequencies=FirstNames%>% group_by(annais) %>% filter(preusuel=="CAMILLE") %>% summarise(n = n()) %>% a
frequencies
```

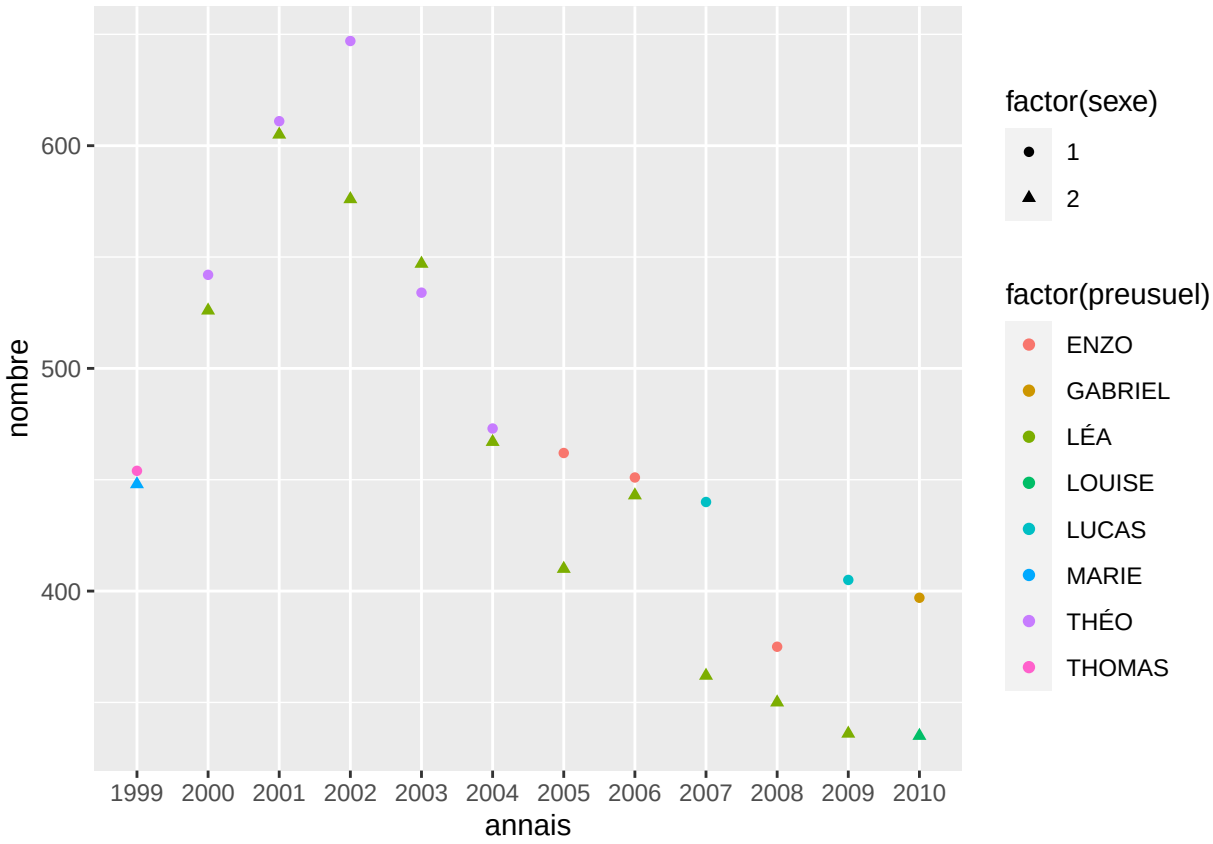
```
## # A tibble: 122 x 2
##    annais     n
##    <chr>  <int>
##  1 2017     179
```

```
##  2 2015     178
##  3 2013     174
##  4 2014     174
##  5 2019     174
##  6 2011     173
##  7 2016     173
##  8 2020     173
##  9 2018     170
## 10 1990     169
## # ... with 112 more rows
```

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

```
library(dplyr)
most_given_by_year_and_gender=FirstNames %>% filter( preusuel != "_PRENOMS_RARES") %>%  group_by(sexe,a

ggplot(data = most_given_by_year_and_gender %>% filter(1999<=as.numeric( annais) & as.numeric( annais) 
```

```
## Warning in mask$eval_all_filter(dots, env_filter): NAs introduits lors de la
## conversion automatique

## Warning in mask$eval_all_filter(dots, env_filter): NAs introduits lors de la
## conversion automatique

## Warning in mask$eval_all_filter(dots, env_filter): NAs introduits lors de la
## conversion automatique

## Warning in mask$eval_all_filter(dots, env_filter): NAs introduits lors de la
## conversion automatique
```



3.

Make a short synthesis

The entries prenoms rares dominate in term of frequency but there are many other rare first name, I mean names that occur no more than once. They should be in the prenoms rares category.

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.