French given names exercise

Benjamin Cathelineau

October, 2021

```
# The environment
library(tidyverse)
library(ggplot2)
```

Download Raw Data from the website

File downloaded from https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2020_csv.zip

Build the Dataframe from file

I had to change the name of the file because it wasn't the correct one.

```
FirstNames <- readr::read_delim("dpt2020.csv",delim=";")
```

1. Choose a firstname and analyse its frequency along time. Compare several first names frequency

First we can find all the different names using the following command. This will group all the entries by ***preusel***

```
table(FirstNames$preusuel)
```

```
##
## _PRENOMS_RARES              A          AADAM          AADEL          AADIL
##          22037              1              1              1              3
##          AAHIL         AAKASH        AALEYAH          AALIA         AALIYA
##              2              1              1              1              2
##  [ reached getOption("max.print") -- omitted 35000 entries ]
```

Then, using the following ***dplyr*** *pipeline we can see the one that occurs the most often.*

```
library(dplyr)
FirstNames %>% count(preusuel) %>% arrange(desc(n))
```

```
## # A tibble: 35,011 x 2
##    preusuel           n
##    <chr>          <int>
##  1 _PRENOMS_RARES 22037
##  2 CAMILLE        13822
##  3 MARIE          13302
##  4 PIERRE         11390
##  5 PAUL           10713
##  6 JEAN           10696
##  7 CLAUDE         10573
##  8 LOUIS          10126
##  9 FRANÇOIS        9977
## 10 ANTOINE         9841
## # ... with 35,001 more rows
```

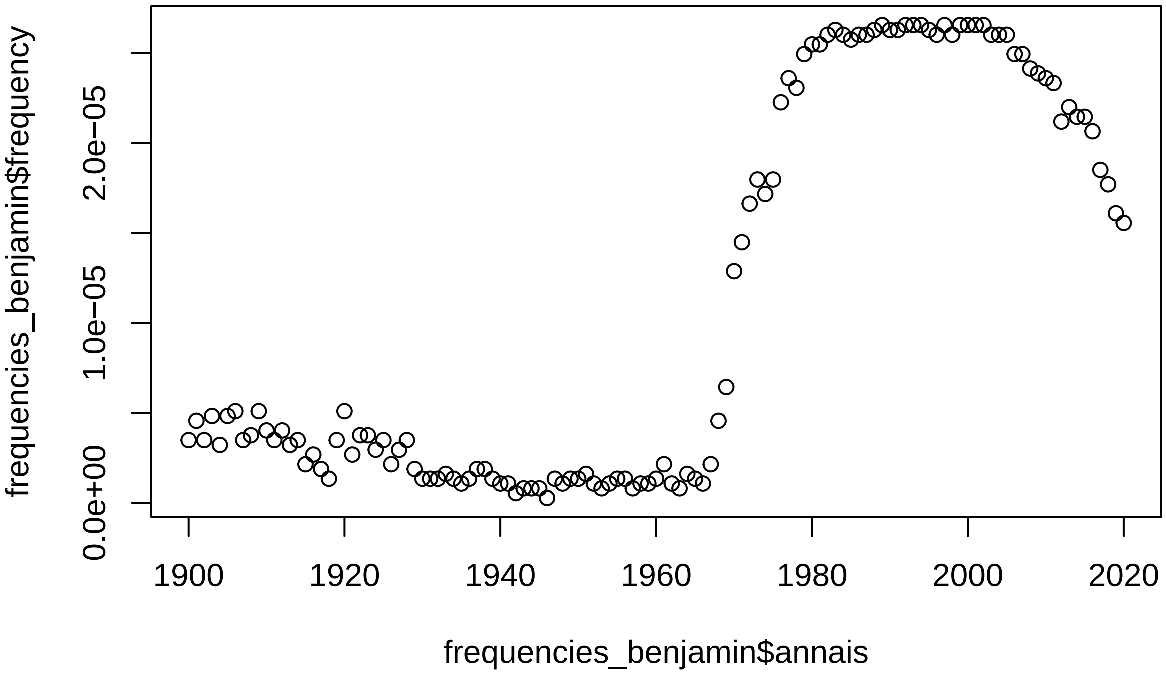
We just have to divide each "count" by the total in order to find the ***frequency*** *for that we use* ***mutate.*** *We also* ***group__by*** *year.*

```
library(dplyr)
frequencies=FirstNames %>% group_by(annais) %>%count(preusuel) %>% arrange(desc(n)) %>% mutate(frequenc
frequencies
```

```
## # A tibble: 284,258 x 4
## # Groups:   annais [122]
##    annais preusuel           n frequency
##    <chr>  <chr>          <int>     <dbl>
##  1 1994   _PRENOMS_RARES   198 0.0000531
##  2 1997   _PRENOMS_RARES   198 0.0000531
##  3 1999   _PRENOMS_RARES   198 0.0000531
##  4 2000   _PRENOMS_RARES   198 0.0000531
##  5 2002   _PRENOMS_RARES   198 0.0000531
##  6 2004   _PRENOMS_RARES   198 0.0000531
##  7 2005   _PRENOMS_RARES   198 0.0000531
##  8 2007   _PRENOMS_RARES   198 0.0000531
##  9 2009   _PRENOMS_RARES   198 0.0000531
## 10 2010   _PRENOMS_RARES   198 0.0000531
## # ... with 284,248 more rows
```

THe following command gives the frequency ***by year*** *for a single name*

```
frequencies_benjamin=frequencies %>% filter(preusuel=="BENJAMIN")
plot(frequencies_benjamin$annais,frequencies_benjamin$frequency )
```



*We can even check our result with **sum** which should be equal to 1*

```
# To get the frequency
library(dplyr)
sum( frequencies$frequency)
```

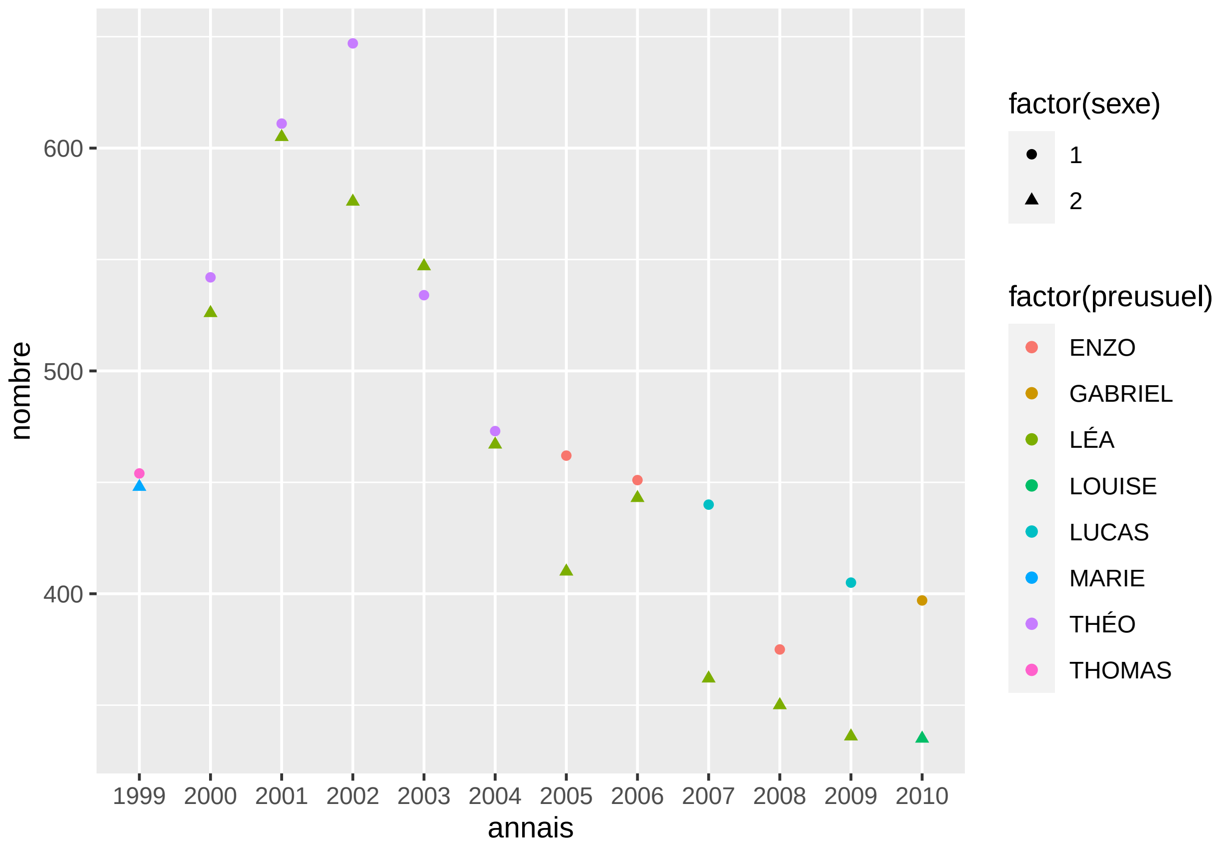
```
## [1] 1
```

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

*We use **group by** for that, and **filter*** *to only keep the maximum. We also remove the “_PRENOMS_RARES” category because it actually regroups all rares names*

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

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



3. Make a short synthesis

There is definitively something strange going on with this data, specifically with the entries prenoms rares, which could be translated to rare First names.

Indeed theses entries dominate in term of frequency, be it in the total, or even when grouping by gender and year

I'm not sure what rare first name exactly means when there are many other rare first name, that is names that occur no more than once

In my opinion theses name are also rare and should be in the prenoms rares category

```
library(dplyr)
# https://community.rstudio.com/t/how-do-i-generate-a-count-in-r-within-mutate/64655
counted_data= FirstNames %>% group_by(preusuel) %>%mutate (count_preusuel =n())
mutated_rare_name=counted_data %>% filter(count_preusuel==1) %>% mutate(preusuel="_PRENOMS_RARES")
no_new_rare_name = counted_data %>% filter(count_preusuel>1)
final_data=rbind(mutated_rare_name,no_new_rare_name)
final_data
```

```
## # A tibble: 3,727,553 x 6
## # Groups:   preusuel [15,677]
##     sexe preusuel       annais dpt   nombre count_preusuel
##    <dbl> <chr>          <chr>  <chr>  <dbl>          <int>
##   1      1 _PRENOMS_RARES XXXX   XX         27                1
##   2      1 _PRENOMS_RARES XXXX   XX         30                1
##   3      1 _PRENOMS_RARES XXXX   XX         56                1
##   4      1 _PRENOMS_RARES XXXX   XX         27                1
##   5      1 _PRENOMS_RARES XXXX   XX         22                1
##   6      1 _PRENOMS_RARES XXXX   XX        165                1
##   7      1 _PRENOMS_RARES XXXX   XX         44                1
##   8      1 _PRENOMS_RARES XXXX   XX         30                1
##   9      1 _PRENOMS_RARES XXXX   XX         22                1
##  10      1 _PRENOMS_RARES XXXX   XX         70                1
## # ... with 3,727,543 more rows
```

*I tried to do it but it was quite complicated and I think I lost the information for **annais** and **dpt** because of the **group_by** but I'm not sure*

Finally, we can say that some name are used both for men and women, so it's something to keep in mind.

For example the name CAMILLE is given almost equally for men and women

```
library(dplyr)
total_women= FirstNames %>% filter(preusuel == 'CAMILLE') %>% filter(sexe == 2)
total_men=FirstNames %>% filter(preusuel == 'CAMILLE') %>% filter(sexe == 1)
nrow(total_men)
```

```
## [1] 6893
```

```
nrow(total_women)
```

```
## [1] 6929
```