French given names per year per department

Lucas Mello Schnorr, Jean-Marc Vincent

November, 2020

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 = "dpt2019_txt.zip"
if(!file.exists(file)){
  download.file("https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2019_csv.zip",
    destfile=file)
}
unzip(file)
```

Build the Dataframe from file

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

```
library(ggplot2)
options(dplyr.summarise.inform=F)
# FirstNames <- read_delim("dpt2019.csv",delim=";");
namedata <- read.csv(file = 'dpt2019.csv', sep = ';')
```

Filter out incomplete data

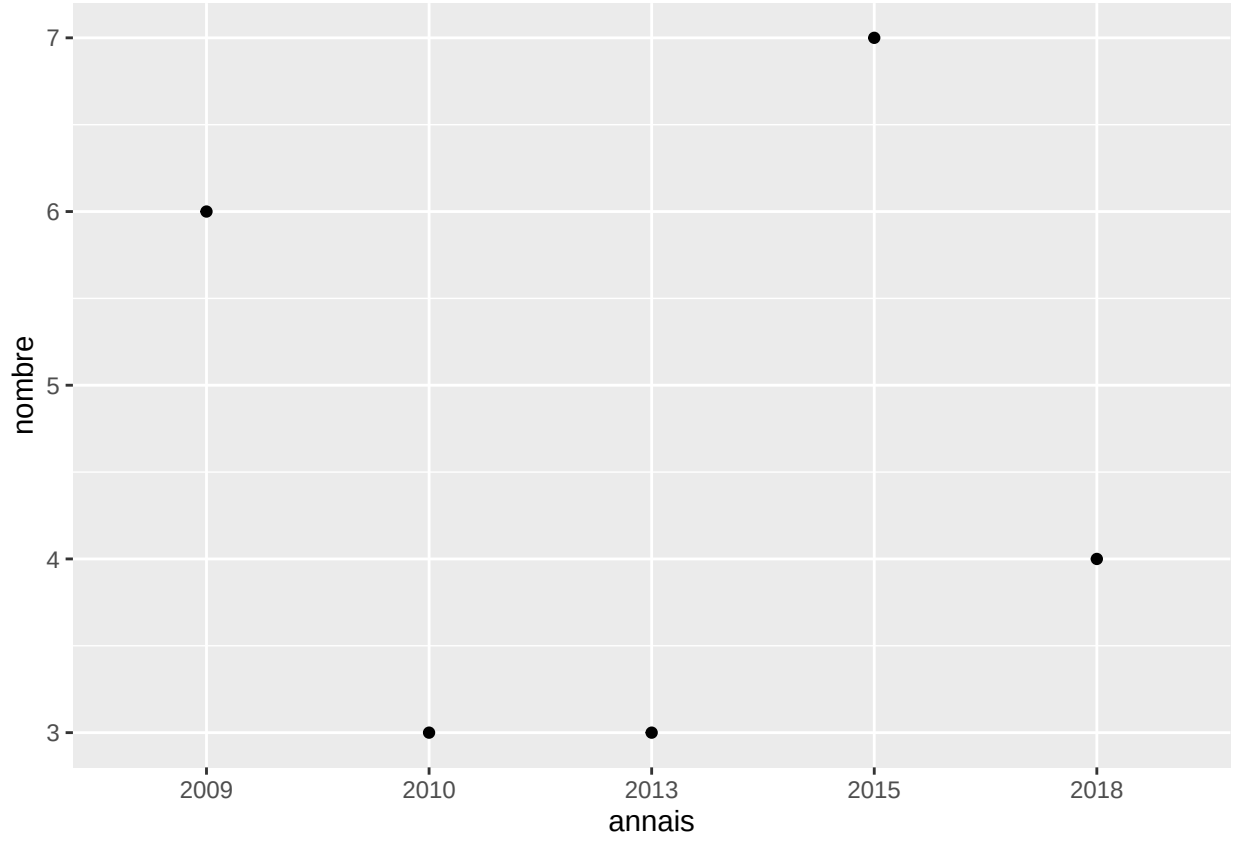
```
FirstNames = filter(namedata, annais != "XXXX" & dpt != "XX" & preusuel != "_PRENOMS_RARES")
tail(FirstNames[complete.cases(FirstNames),],10)
```

```
##         sexe preusuel annais dpt nombre
## 3618392    2  ZUZANNA   2015  94      3
## 3618393    2  ZUZANNA   2018  75      4
## 3618394    2      ZYA   2011  85      4
## 3618395    2      ZYA   2011  91      3
## 3618396    2      ZYA   2011 974      3
## 3618397    2      ZYA   2013  44      4
## 3618398    2      ZYA   2013  59      3
## 3618399    2      ZYA   2017 974      3
## 3618400    2      ZYA   2018  59      3
## 3618401    2     ZYNA   2013  93      3
```

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

- We choose a first name "zuzanna

```
ChoosenName = filter(FirstNames, preusuel == "ZUZANNA")
ChoosenName = ChoosenName %>%
  group_by(annais) %>%
  summarise(nombre = sum(nombre))
ggplot(data = ChoosenName, aes(x=annais, y=nombre))+geom_point()
```

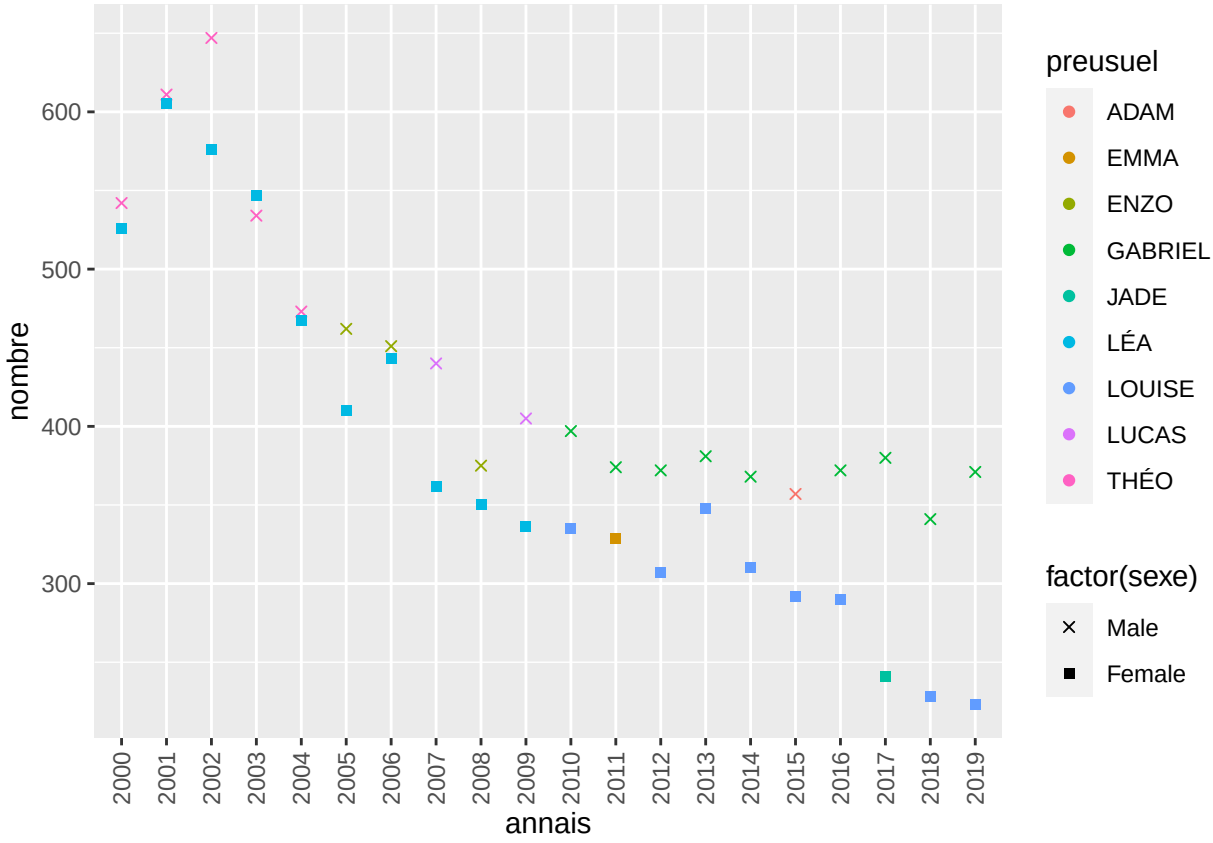


- Now we will pick another first names and compare them (we will only show the past 20 years, to make the result readable)

```
CompareNames = filter(FirstNames, as.numeric(as.character(annais)) >= 2000 & (preusuel == "ZUZANNA" | p
CompareNames = CompareNames %>%
  group_by(annais, preusuel) %>%
  summarise(nombre = sum(nombre))
q <- ggplot(data = CompareNames, aes(x=annais, y=nombre, color = preusuel))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

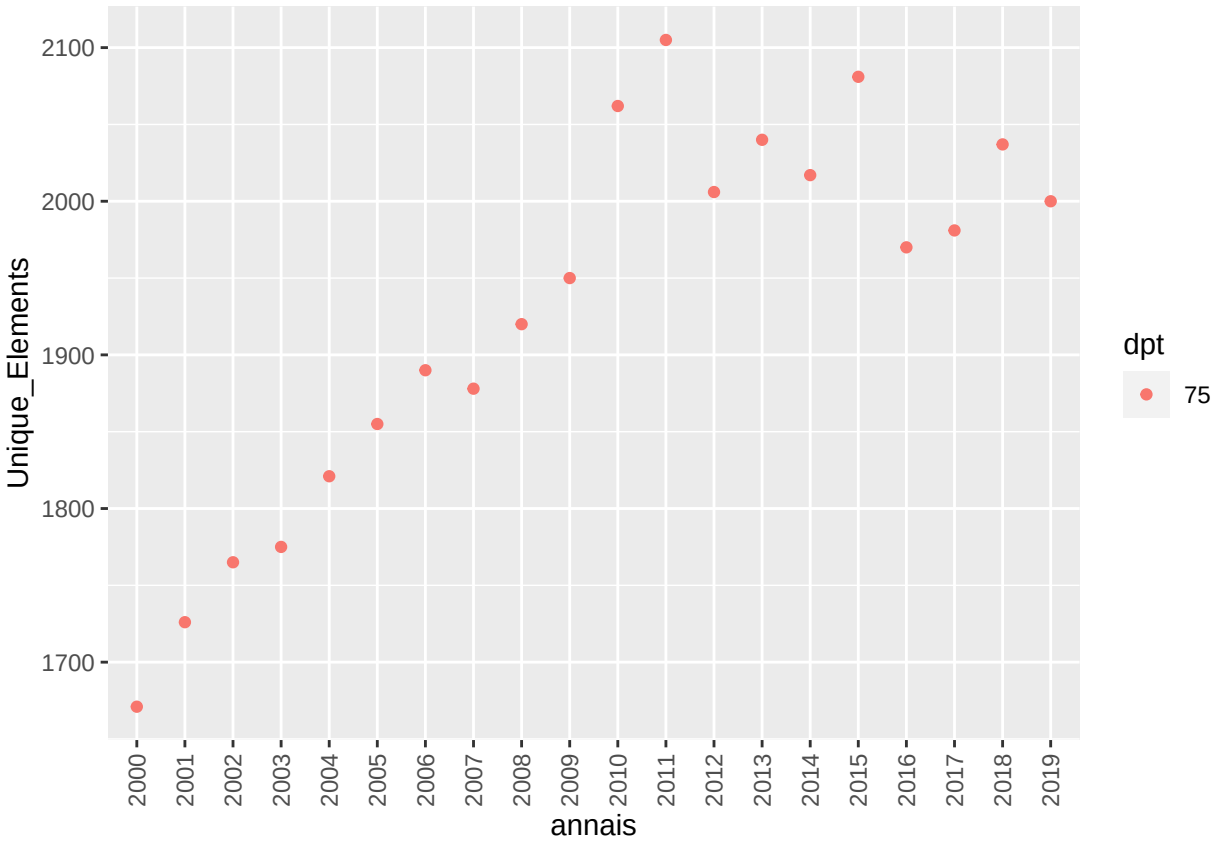
2. Establish by gender the most given firstname by year. Analyse the evolution of the most frequent firstname.

```
HighestNames = FirstNames %>%
             group_by(sexe, annais) %>%
             filter(nombre == max(nombre) & as.numeric(as.character(annais)) >= 2000)
q <- ggplot(data = HighestNames, aes(x=annais, y=nombre, shape = factor(sexe), color = preusuel))+geom_
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



3. Optional : Which department has a larger variety of names along time ? Is there some sort of geographical correlation with the data?

```
CountUniqueNames = FirstNames %>%
  filter(as.numeric(as.character(annais)) >= 2000) %>%
  group_by(annais, dpt) %>%
  summarise(Unique_Elements = n_distinct(preusuel))
CountUniqueNamesFiltered = CountUniqueNames %>%
  filter(Unique_Elements == max(Unique_Elements))
q <- ggplot(data = CountUniqueNamesFiltered, aes(x=annais, y=Unique_Elements, color = dpt))+geom_point()
q + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Department 75 has the highest variety. Yes there is a correlation because department 75 (Seine) is one of the most populated department in France