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 3.3.5     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)

version
```

```
##                _
## platform       aarch64-apple-darwin20
## arch           aarch64
## os             darwin20
## system         aarch64, darwin20
## status
## major          4
## minor          1.1
## year           2021
## month          08
## day            10
## svn rev        80725
## language       R
## version.string R version 4.1.1 (2021-08-10)
## nickname       Kick Things
```

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

```
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.
```

```
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
```

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.

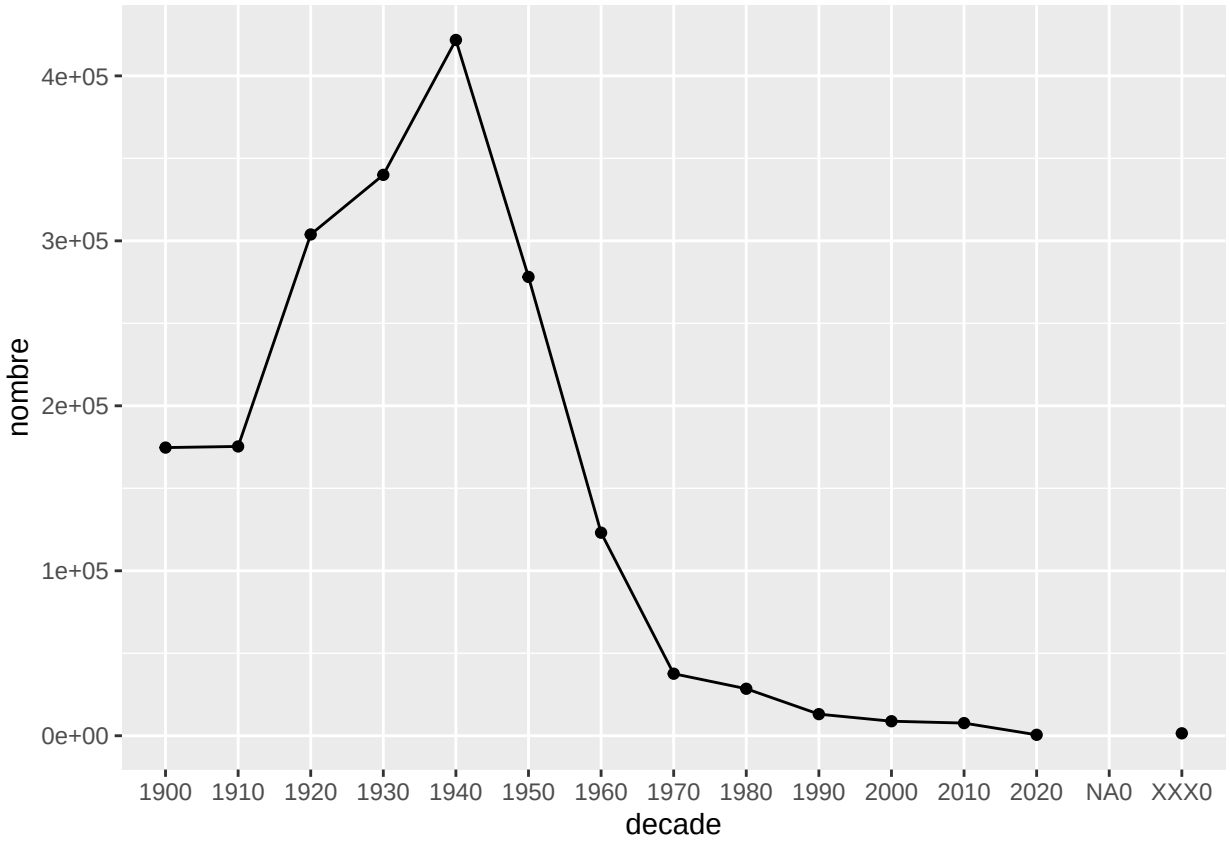
1. Choose a firstname and analyse its frequency along time. Compare several firstnames frequency

```r
#The name we choose is Jean
name = "JEAN"
jeanData = FirstNames[FirstNames$preusuel == name, ]

#We will first show the evolution of the number of child with this name for the whole country
jeanCountryData = jeanData[, c("annais", "nombre")] %>% group_by(annais) %>% summarize(nombre = sum(nomb

#we group this data by decade to do a visualisation
jeanCountryData$decade = paste0(substr(jeanCountryData$annais, start = 1, stop = 3),0)
jeanCountrydecadeData <- jeanCountryData[, c("decade", "nombre")] %>% group_by(decade) %>% summarize(nom
ggplot(jeanCountrydecadeData, aes(x=decade, y=nombre, group=1)) + geom_line() + geom_point()
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```r
#jean01Data = jeanData[jeanData$dpt == "01", ]
#ggplot(jean01Data, aes(x = annais, y = nombre)) + geom_line()
#jean01Data$decade <- jean01Data$annais / 10
#summary(jean01Data)
#jean01Data$decade = paste0(substr(jean01Data$annais, start = 1, stop = 3),0)
#jean01decadeData <- jean01Data[, c("decade", "nombre")] %>% group_by(decade) %>% summarize(nombre = n(
#ggplot(jean01decadeData, aes(x=decade, y=nombre, group=1)) + geom_line() + geom_point()
```

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.