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  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
library(ggplot2)

version
```

```
##                _
## platform       x86_64-apple-darwin15.6.0
## arch           x86_64
## os             darwin15.6.0
## system         x86_64, darwin15.6.0
## status
## major          3
## minor          6.3
## year           2020
## month          02
## day            29
## svn rev        77875
## language       R
## version.string R version 3.6.3 (2020-02-29)
## nickname       Holding the Windsock
```

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 =";")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sexe = col_double(),
##   preusuel = col_character(),
##   annais = col_double(),
##   dpt = col_character(),
##   nombre = col_double()
## )
```

```
## Warning: 37244 parsing failures.
##    row    col expected actual           file
## 10882 annais a double   XXXX ’dpt2020.csv’
## 10883 annais a double   XXXX ’dpt2020.csv’
## 10884 annais a double   XXXX ’dpt2020.csv’
## 10885 annais a double   XXXX ’dpt2020.csv’
## 10888 annais a double   XXXX ’dpt2020.csv’
## ..... ...... ........ ...... .............
## See problems(...) for more details.
```

```
FirstNames
```

```
## # A tibble: 3,727,553 x 5
##     sexe preusuel       annais dpt   nombre
##    <dbl> <chr>           <dbl> <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
```

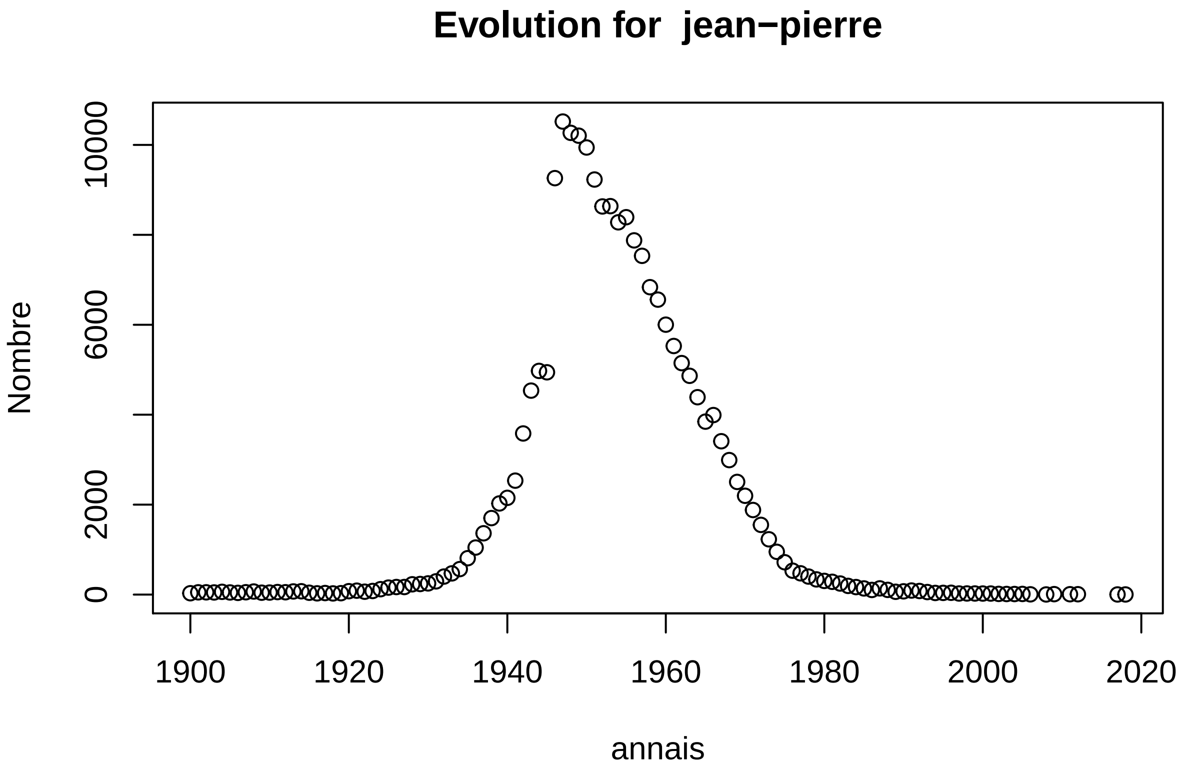
Translation in english of variables names: sexe -> gender preusuel (prénom usuel) -> Firstname annais (année de naissance) -> Birth year dpt (département) -> department (administrative area unit) nombre -> number

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.

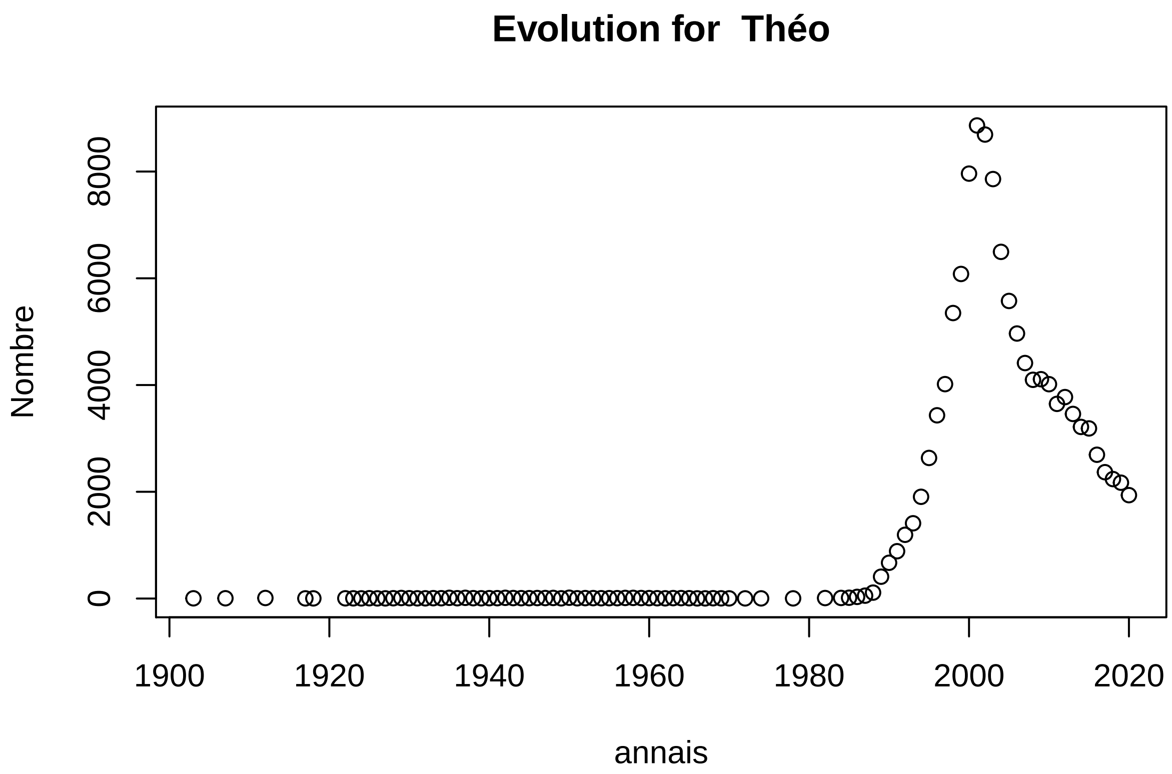
```r
# Let's replace accents by noramlized letters
normalize_text <- function(text) {
  text = toupper(iconv(text,from="UTF-8",to="ASCII//TRANSLIT"))
  text = str_replace_all(text, "[^[:alnum:]]", "")
  return(text)
}
FirstNames$preusuel_norm = normalize_text(FirstNames$preusuel)
```

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

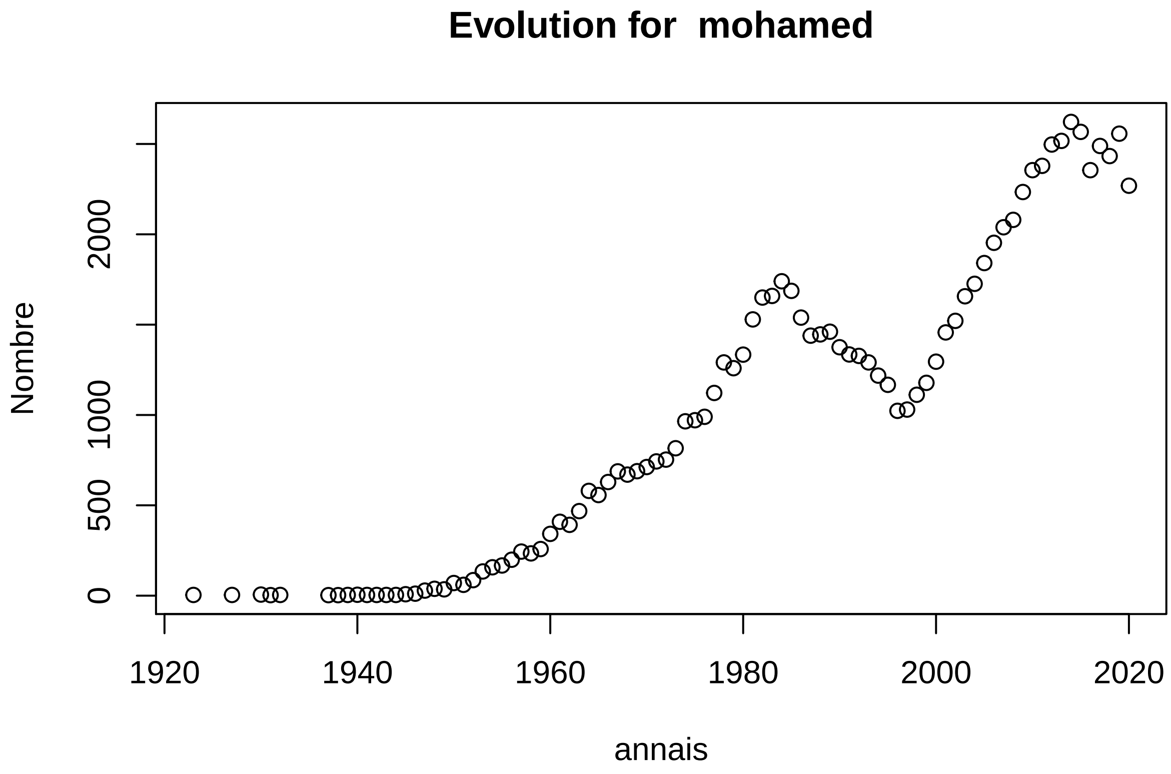
```r
plot_evolution <- function (firstname_to_analyze) {
  firstname_to_analyze_norm = normalize_text(firstname_to_analyze)
  n_occurences <- FirstNames[FirstNames$preusuel_norm == firstname_to_analyze_norm,] %>% group_by(annais
  plot(n_occurences, main=paste("Evolution for ", firstname_to_analyze))
}
plot_evolution("jean-pierre")
```



```
plot_evolution("Théo")
```



```
plot_evolution("mohamed")
```



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

```
most_given_male_name <- FirstNames[FirstNames$sexe == 1,] %>%
  filter(preusuel_norm != "PRENOMSRARES") %>%
  filter(annais != "XXXX") %>%
  drop_na(annais) %>%
  group_by(annais, preusuel_norm) %>%
  summarize(Nombre = sum(nombre)) %>%
  slice(which.max(Nombre))
```

```
## `summarise()` has grouped output by 'annais'. You can override using the `.groups` argument.
```

```
most_given_female_name <- FirstNames[FirstNames$sexe == 2,] %>%
  filter(preusuel_norm != "PRENOMSRARES") %>%
  filter(annais != "XXXX") %>%
  drop_na(annais) %>%
  group_by(annais, preusuel_norm) %>%
  summarize(Nombre = sum(nombre)) %>%
  slice(which.max(Nombre))
```

```
## `summarise()` has grouped output by 'annais'. You can override using the `.groups` argument.
```

```
print(most_given_male_name)
```

```
## # A tibble: 121 x 3
## # Groups:   annais [121]
##    annais preusuel_norm Nombre
##     <dbl> <chr>          <dbl>
##  1   1900 JEAN           14097
##  2   1901 JEAN           15634
##  3   1902 JEAN           16364
##  4   1903 JEAN           16535
##  5   1904 JEAN           16944
##  6   1905 JEAN           17998
##  7   1906 JEAN           18522
##  8   1907 JEAN           18475
##  9   1908 JEAN           19935
## 10   1909 JEAN           20152
## # ... with 111 more rows
```

```
print(most_given_female_name)
```

```
## # A tibble: 121 x 3
## # Groups:   annais [121]
##    annais preusuel_norm Nombre
##     <dbl> <chr>          <dbl>
##  1   1900 MARIE          48713
##  2   1901 MARIE          52150
##  3   1902 MARIE          51857
##  4   1903 MARIE          50424
##  5   1904 MARIE          50131
##  6   1905 MARIE          48981
##  7   1906 MARIE          48447
##  8   1907 MARIE          46048
##  9   1908 MARIE          47460
## 10   1909 MARIE          46398
## # ... with 111 more rows
```

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.