French given names per year per department

Gabriele Degola, Lucas Mello Schnorr, Jean-Marc Vincent

October, 2021

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

version
```

```
##                _
## platform       x86_64-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          3
## minor          6.1
## year           2019
## month          07
## day            05
## svn rev        76782
## language       R
## version.string R version 3.6.1 (2019-07-05)
## nickname       Action of the Toes
```

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 *dpt2020_txt.zip* (to get the **dpt2020.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 = "data/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("data/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 'data/dpt2020.csv'
## 10883 annais a double   XXXX 'data/dpt2020.csv'
## 10884 annais a double   XXXX 'data/dpt2020.csv'
## 10885 annais a double   XXXX 'data/dpt2020.csv'
## 10888 annais a double   XXXX 'data/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

Data exploration

Before answering our questions, a preliminary analysis of the data is required. During data loading, some failures reading the *annais* variable are prompted.

```
unique(FirstNames$annais)
```

```
##   [1] 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914
```

```
##  [16] 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929
##  [31] 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944
##  [46] 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959
##  [61] 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974
##  [76] 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989
##  [91] 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004
## [106] 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
## [121] 2020   NA
```

The dataset contains information about the given names from 1900 to 2020. However, the year information is not available for some entries.

```
FirstNames %>% summarise(across(everything(), ~ sum(is.na(.))))
```

```
## # A tibble: 1 x 5
##    sexe preusuel annais   dpt nombre
##   <int>    <int>  <int> <int>  <int>
## 1     0        1  37244     0      0
```

That confirms that the *annais* value is not available for some entries, while in just one entry the *preusuel* is not available. As we are interested in the relation between names usage and time, these entries are dropped.

```
FirstNames <- FirstNames %>% drop_na()
```

We can proceed exploring the other variables. We can start by checking the *sexe* information.

```
unique(FirstNames$sexe)
```

```
## [1] 1 2
```

The dataset is therefore relative to two genders, probably male and female. Let's print some names for the two genders.

```
print(nrow(FirstNames[FirstNames$sexe == 1,]))
```

```
## [1] 1712945
```

```
head(FirstNames[FirstNames$sexe == 1,] %>% group_by(preusuel) %>% summarise(num = n(),
↪  tot = sum(nombre)) %>% arrange(desc(tot)))
```

```
## # A tibble: 6 x 3
##   preusuel         num     tot
##   <chr>          <int>   <dbl>
## 1 JEAN           10639 1912848
## 2 PIERRE         11278  891170
## 3 MICHEL          8593  818001
## 4 _PRENOMS_RARES 10881  798128
## 5 ANDRÉ           7878  709568
## 6 PHILIPPE        7504  535200
```

```
nrow(FirstNames[FirstNames$sexe == 2,])
```

```
## [1] 1977364
```

```
head(FirstNames[FirstNames$sexe == 2,] %>% group_by(preusuel) %>% summarise(num = n(),
↪  tot = sum(nombre)) %>% arrange(desc(tot)))
```

```
## # A tibble: 6 x 3
##   preusuel         num     tot
```

```
##   <chr>          <int>   <dbl>
## 1 MARIE          11353 2231903
## 2 _PRENOMS_RARES 11154  853451
## 3 JEANNE          9106  556897
## 4 FRANÇOISE       7170  399509
## 5 MONIQUE         5201  397739
## 6 CATHERINE       6557  391518
```

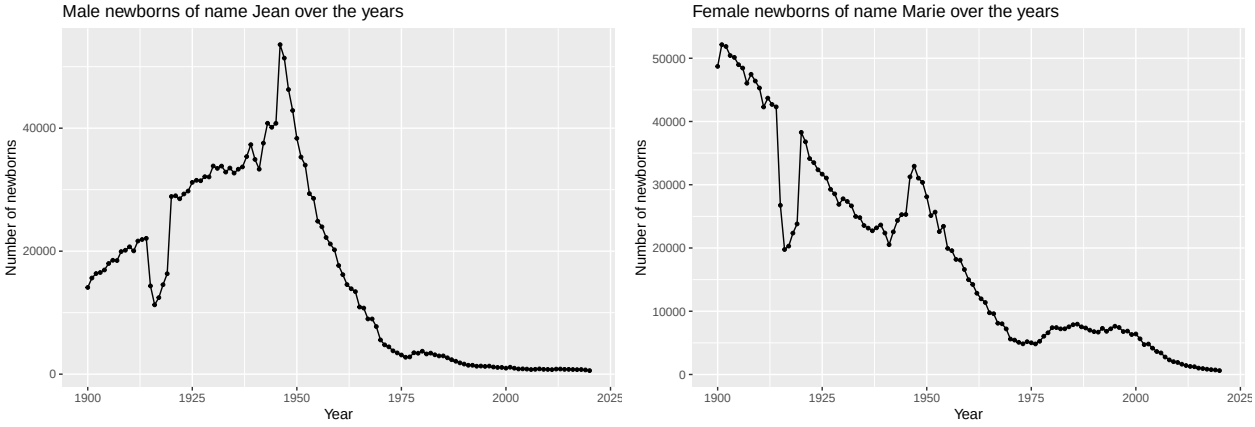
So the entries of *sexe* 1 contain male names (like Jean, Pierre and Philippe), while *sexe* 2 is relative to female names (Marie, Jeanne, Catherine). A considerable amount of entries are relative to general rare names (`_PRENOMS_RARES`). Most of the entries of the dataset are relative to female names. In total, 15271 names appear in the dataset, of which 7290 are male names and 8850 are female names. There is therefore some overlapping (names that are used for both males and females).

Now, we can analyse the firstnames appearing in the dataset.

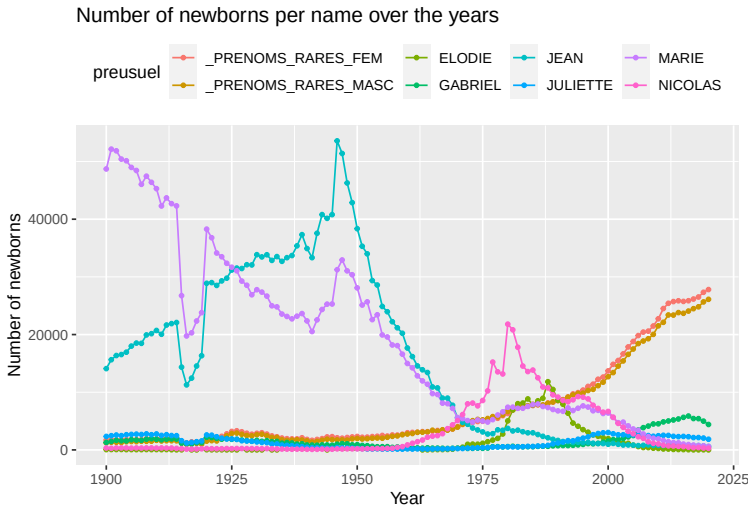
Firstnames frequencies over time

In this section, we are interested in studying the frequency of some French firstnames over the years.

As a start, the trend of the overall most used names (Jean for males, Marie for females) is analysed.

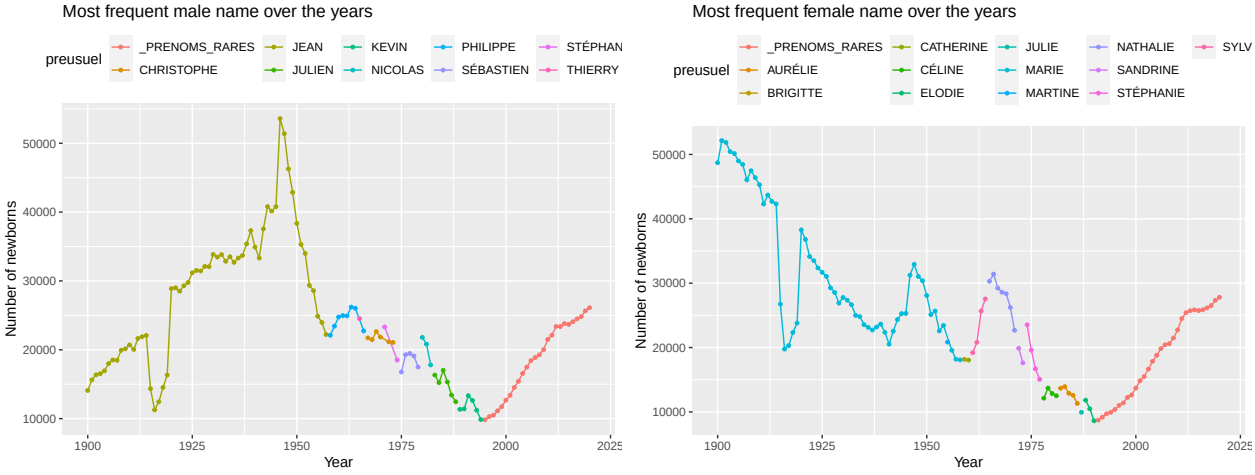


Event though the trends for the two names are different, a few common patterns can be highlighted. The usage of both names constantly decreased since 1950 until today with some small fluctuations. For both names, the usage dropped around 1918 and returned normal a few years later. We can compare the trends for other names (Jean, Marie, Nicolas, Juliette, Gabriel, Elodie). The trend of rare firstnames is also plotted.

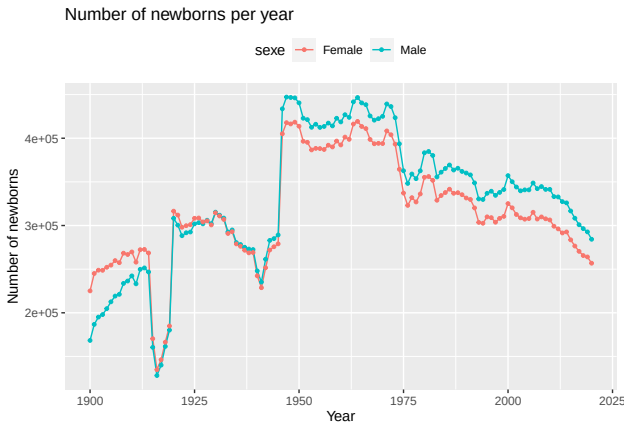


While Jean and Marie had been the most frequent name for a long time, in the last decades they have given way to other names with various trends. Recently, the so-called rare names have taken the lead for both males and females, even though several names are included in the category.

Most given firstname per year



From the two graphs, one can think that, while at the beginning of XX century only a few names were used, in the last decades a moltitude of names are instead present. The total number of newborns per year can also be plotted.

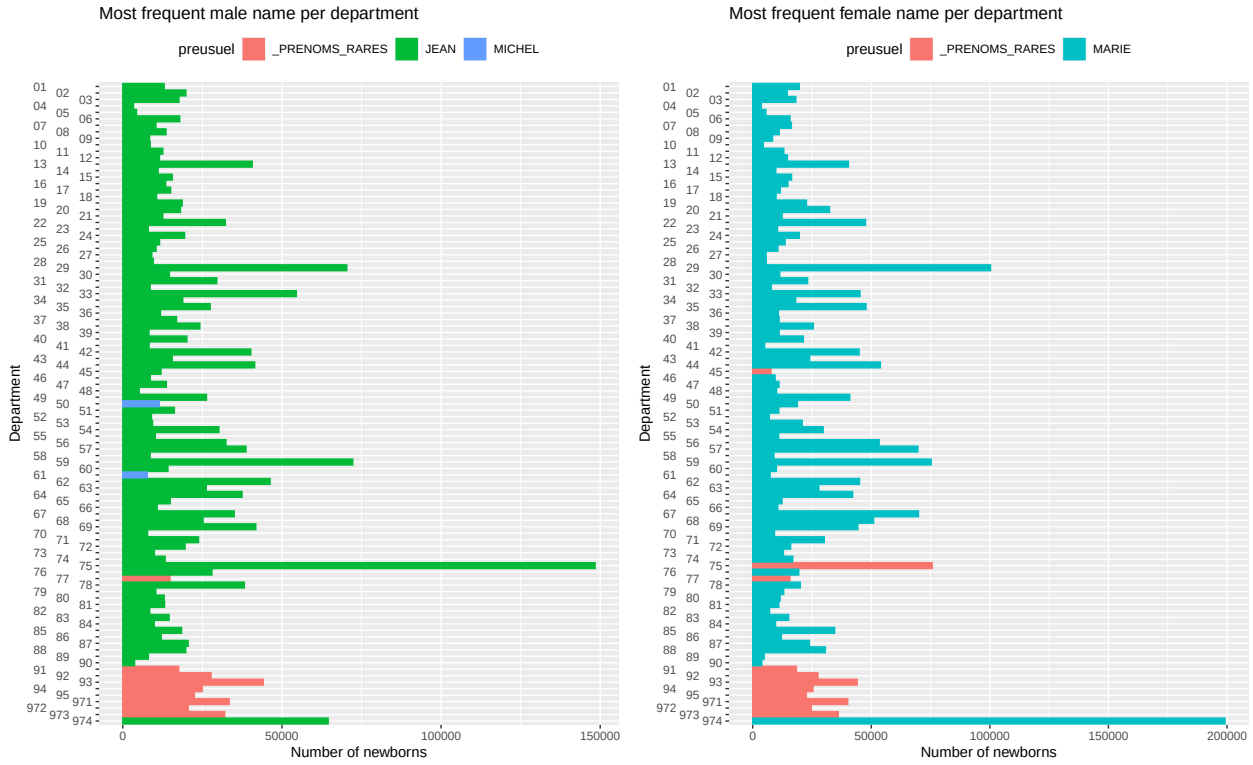


From the last graph, the drop mentionned before are more evident and involve all newborns. Another small drop can be highlighted around 1940. They are probably due to the two world wars. Lately, the number of newborns constantly decreases while the frequency of rare firstnames does increase.

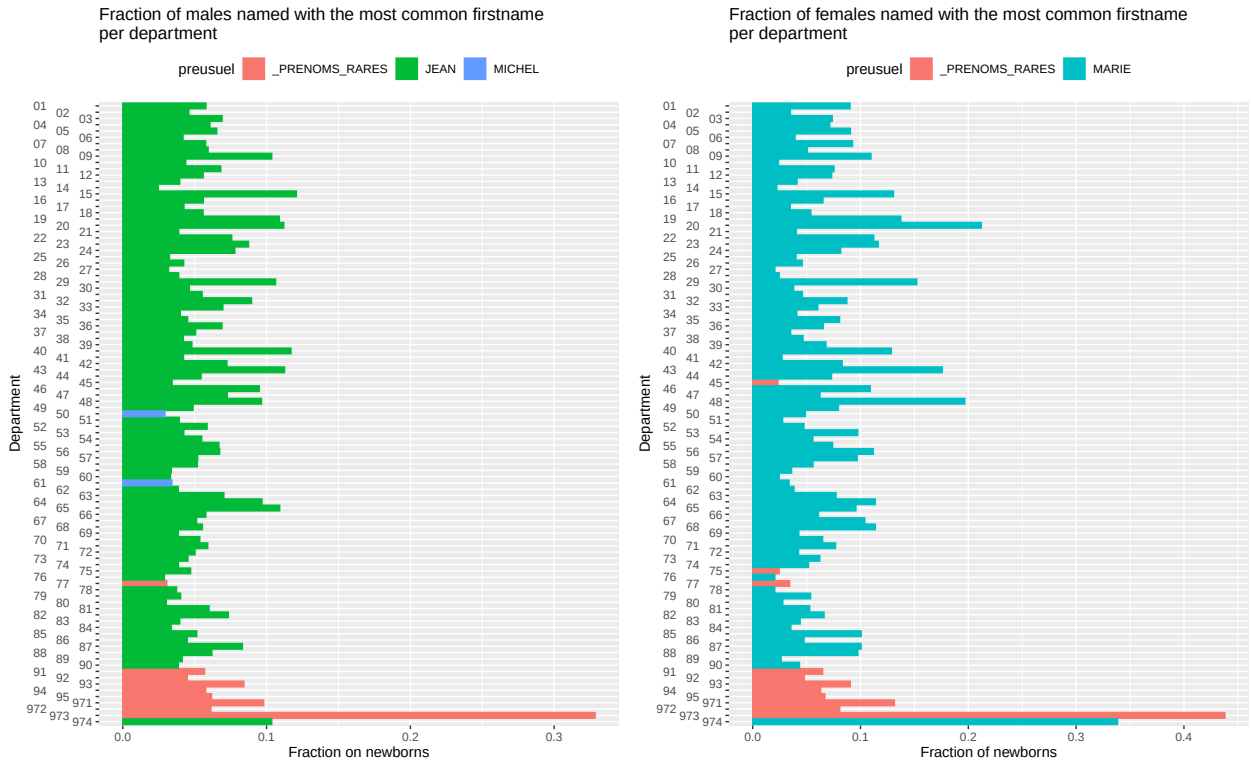
Comparing the graphs, it is clear that the most frequent name is always given to a small fraction of newborns, even in years of exceptional birth rate.

Correlation between firstnames and departments

As did before, one can visualize the most common name in each French department.



So, in some departments a huge number of children are named with the most frequent names (Jean, Marie or rare firstnames). However, the information is not very relevant, as some departments are bigger than others. It is therefore useful to visualize instead the fraction of newborns named with the most used name, with respect to the total number of borns.



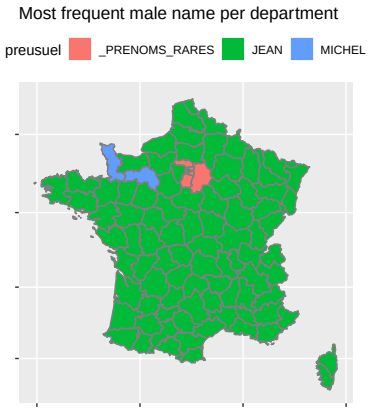
This gives a different perspective. For example, one could have been lead to think that, as around 150000 children has been named Jean, almost all the males in department 75 are named Jean. Instead, this only corresponds to the 5% of all newborns in department 75 (Paris), an highly populated region. At the same time, one female over three is actually named Marie in 974 (La Réunion). Finally, it is worth noticing that in 971, 972 and 973 (*départements d'outre mer*), the most used firstnames are the renowned rare names, that may be considered as rare in metropolitan France while being common in Central America.

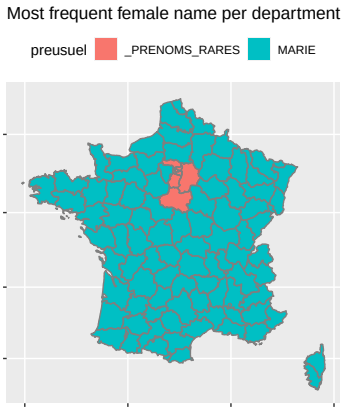
Test of map visualization

```
# Create French map object
map <- subset(map_data("france"), select = -subregion)
map <- map %>% mutate(region = tolower(region))
# Load CSV matching department codes with names
file = "data/departements-region.csv"
if(!file.exists(file)){

  ↪  download.file("https://www.data.gouv.fr/en/datasets/r/987227fb-dcb2-429e-96af-8979f97c9c84",
    destfile=file)
}
# Standardize department names as in the map
deps <- subset(read.csv(file, encoding = "UTF-8", stringsAsFactors = FALSE), select =
↪  -region_name)
deps <- deps %>% mutate(dep_name = tolower(dep_name)) %>% mutate(dep_name = gsub("[áàâ]",
↪  "a", dep_name))
deps <- deps %>% mutate(dep_name = gsub("[éèê]", "e", dep_name)) %>% mutate(dep_name =
↪  gsub("[íìî]", "i", dep_name))
deps <- deps %>% mutate(dep_name = gsub("[óòô]", "o", dep_name)) %>% mutate(dep_name =
↪  gsub("'", "", dep_name))
# Rename columns
```

```r
deps <- rename(deps, region = dep_name)
deps <- rename(deps, dpt = num_dep)
deps$region[deps$region == "corse-du-sud"] <- "corse du sud"
# Join for department code
map <- left_join(x = map, y = deps)
# Merge Corse
map$dpt[map$dpt == "2A"] <- 20
map$dpt[map$dpt == "2B"] <- 20
# Put together map and name frequencies
freq_in_dpt_male <- freq_in_dpt[freq_in_dpt$sexe == 1,]
freq_in_dpt_female <- freq_in_dpt[freq_in_dpt$sexe == 2,]
map_male <- left_join(x = map, y = freq_in_dpt_male)
map_female <- left_join(x = map, y = freq_in_dpt_female)
```





Conclusion

To conclude, the number of newborns in France has been subjected to a lot of fluctuation in the last century. The two world wars and the economic crisis of 1970 and 2008 lead to drastic drops. This affected in particular the most traditional names, with several names emerging for some periods and then declining. Recently, the names defined as *rare* by the INSEE have taken the lead surpassing the firstnames we were used to (probably due to migration towards France).

Finally, even though it is not easy to analyse the correlation between firstnames and departments, some insights can be obtained analysing the frequency of the most common names in each French department.