French given names per year per department

Edited by : **Oussama Oulkaid**

October, 2021

Introduction

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, the analysis cannot be done by hand, the structure is simple.

We will use the *tidyverse* for this analysis. The file **dpt2019.csv** contains the data.

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

Build the Dataframe from file

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

```
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
```

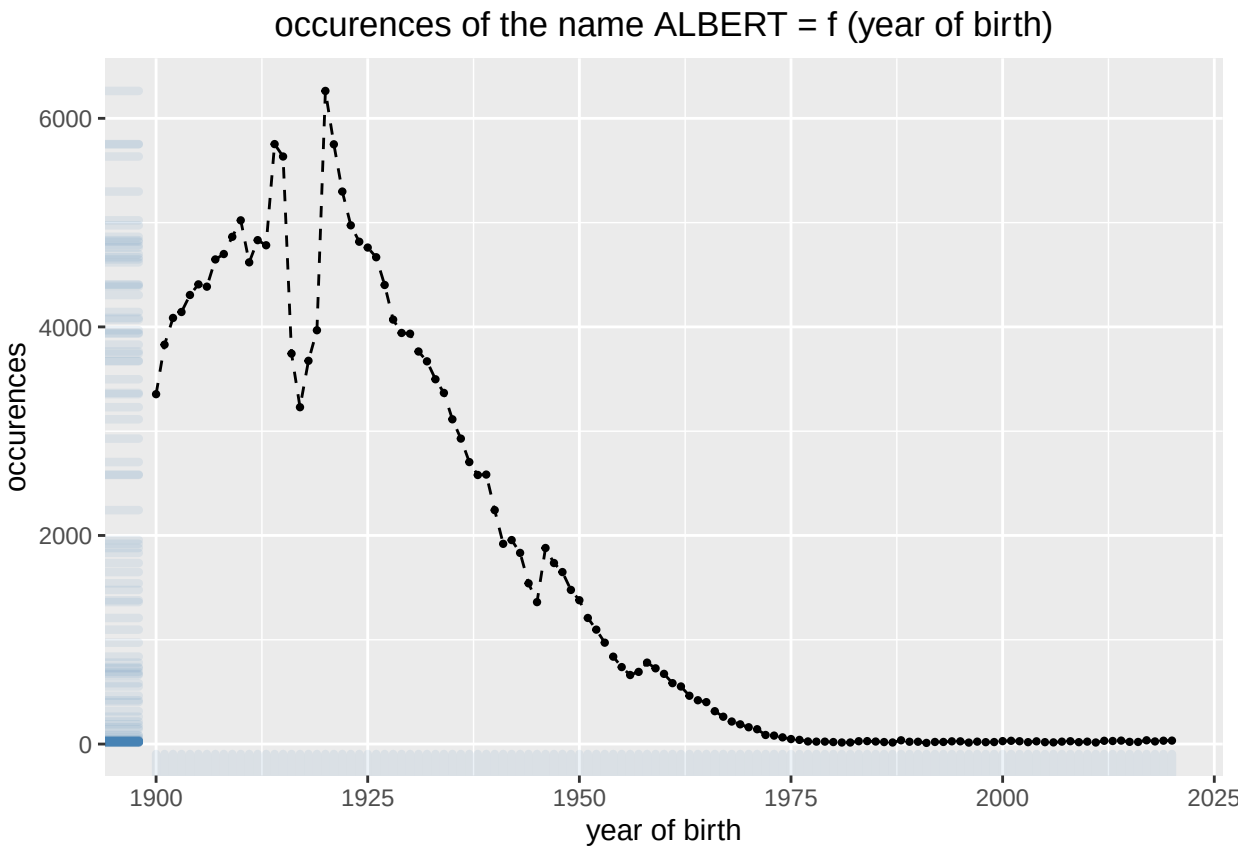
```
FirstNames[match(unique(FirstNames$sexe), FirstNames$sexe), c('sexe')]
```

Analysing Firstnames' frequencies

We first choose an example Firstname (ALBERT), and we analyse its frequency.

```
albert <- subset(FirstNames, preusuel == 'ALBERT') %>% filter(annais!='XXXX')
albert <- subset(albert[,c('annais', 'nombre')])
albert <- transform(albert, nombre = as.numeric(nombre), annais = as.numeric(annais))
albert <- setNames( aggregate(albert$nombre, by=list(annais=albert$annais), FUN=sum) ,
```

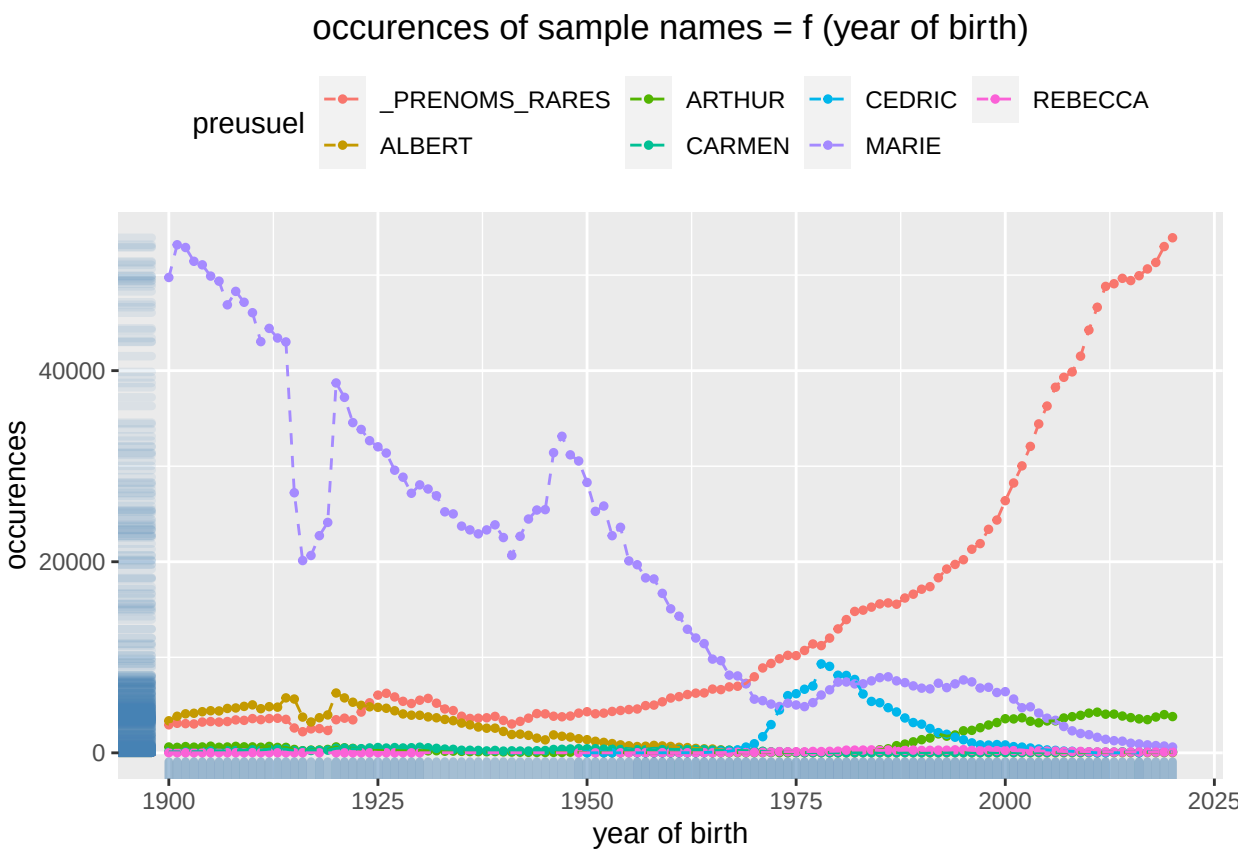
```
                    c('annais', 'sum_nombre') )

ggplot(data=albert, aes(x=annais, sum_nombre)) +
  geom_point(size=0.8) + geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) + theme(plot.title = element_text(hjust = 0.5)) +
  labs(title = "occurences of the name ALBERT = f (year of birth)", x = "year of birth", y = "occurences
```



Comment : After 1975, the number of new births with the Firstname ALBERT became very low. But the decline started since 1921.

Now, let's compare several Firstnames' frequencies.

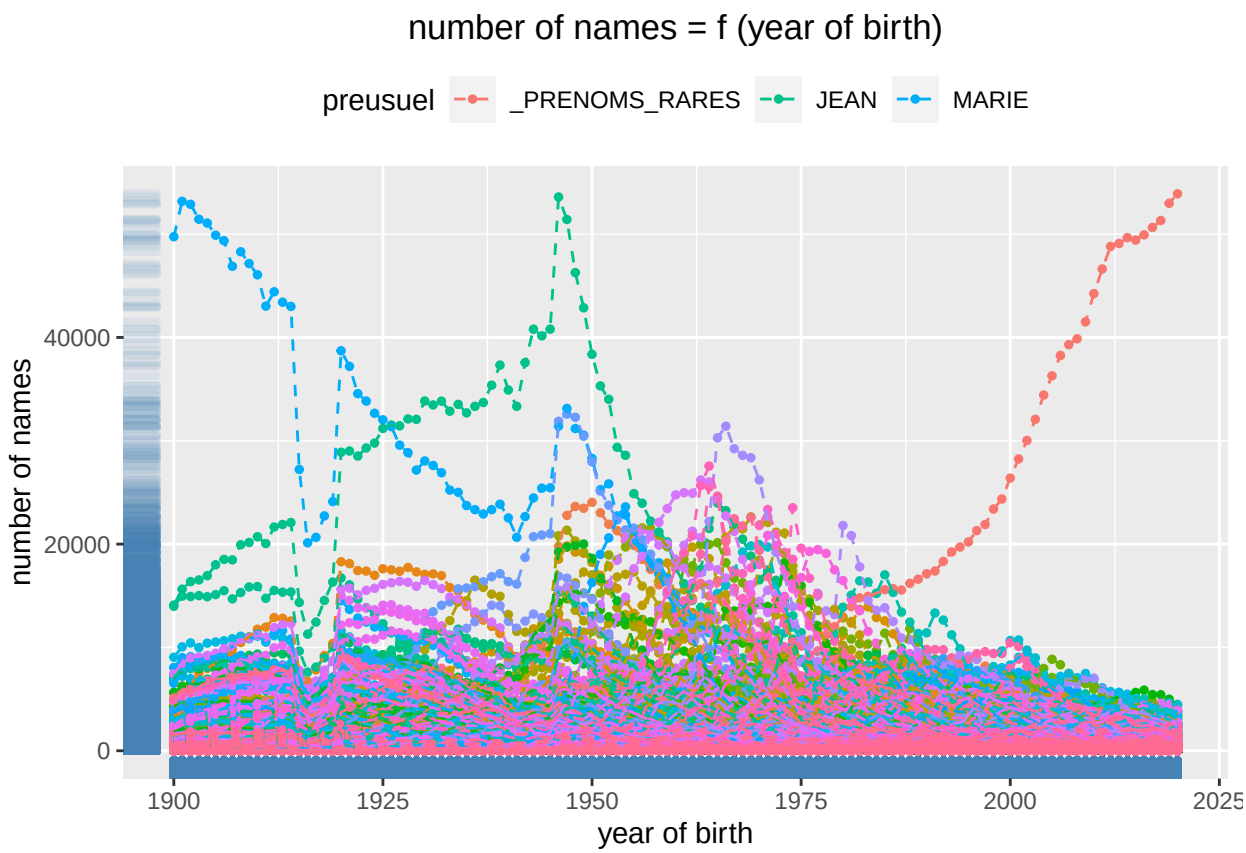
```
sample_names <- subset( FirstNames[,c('preusuel', 'annais', 'nombre')],
                        preusuel=='ALBERT' | preusuel=='CEDRIC' |preusuel=='_PRENOMS_RARES' | preusuel=
                        preusuel=='ARTHUR' | preusuel=='CARMEN' | preusuel=='MARIE') %>% filter(annais!=

sample_names <- transform(sample_names, nombre = as.numeric(nombre), annais = as.numeric(annais))
sample_names <- setNames( aggregate(sample_names$nombre, by=list(preusuel=sample_names$preusuel, annais=
                          c('preusuel', 'annais', 'sum_nombre'))

ggplot(data=sample_names, aes(x=annais, sum_nombre, colour=preusuel)) +
  geom_point(size=1) + geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "occurences of sample names = f (year of birth)", x = "year of birth", y = "occurences")
  theme( plot.title = element_text(hjust = 0.5), legend.position = "top" )
```



Comment : It appears that the set of rare Firstnames is getting diversified, thus the increasing curve of the total number of this set. We note here that there are many samples for which the year of birth is not known (labeled XXXX), but we've ignored them in this analysis.

We can also plot all the names in the data set, in order to have an exhaustive overview over the distribution (event though not readable). We'll identify the three most used Firstnames since 1900 :

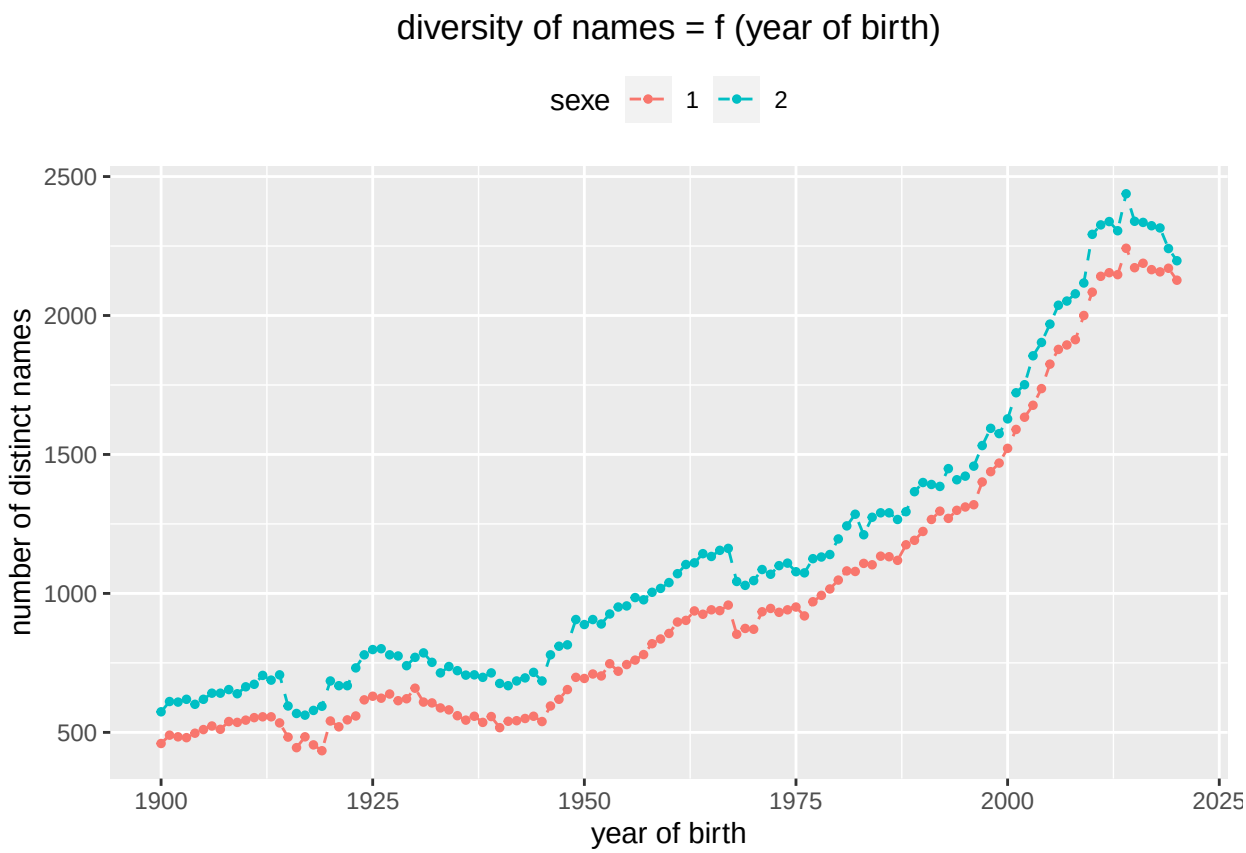
```
all_names <- subset(FirstNames[,c('preusuel', 'annais', 'nombre')]) %>% filter(annais!='XXXX')

all_names <- transform(all_names, nombre = as.numeric(nombre), annais = as.numeric(annais))
all_names <- setNames( aggregate(all_names$nombre, by=list(preusuel=all_names$preusuel, annais=all_names

#getting the most used 3 names of all time (since 1900)
most_names <- setNames( aggregate(all_names$sum_nombre, by=list(preusuel=all_names$preusuel), FUN=max)
first_name <- subset( most_names, sum_nombre == sort(most_names$sum_nombre, decreasing = TRUE)[1] )
second_name <- subset( most_names, sum_nombre == sort(most_names$sum_nombre, decreasing = TRUE)[2] )
third_name <- subset( most_names, sum_nombre == sort(most_names$sum_nombre, decreasing = TRUE)[3] )

ggplot(data=all_names, aes(x=annais, sum_nombre, colour=preusuel)) +
  geom_point( size = 1 ) + #shape = "."
  geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "number of names = f (year of birth)", x = "year of birth", y = "number of names") +
  theme( plot.title = element_text(hjust = 0.5), legend.position = "top" ) +
  scale_colour_discrete(breaks = c(first_name$preusuel, second_name$preusuel, third_name$preusuel))
```



Comment : JEAN and MARIE are the most used Firstnames since 1900. The sum of all sets of rare names end up to form the biggest set of birth names. What we conclude from this graph is that the global set of names is getting diversified.

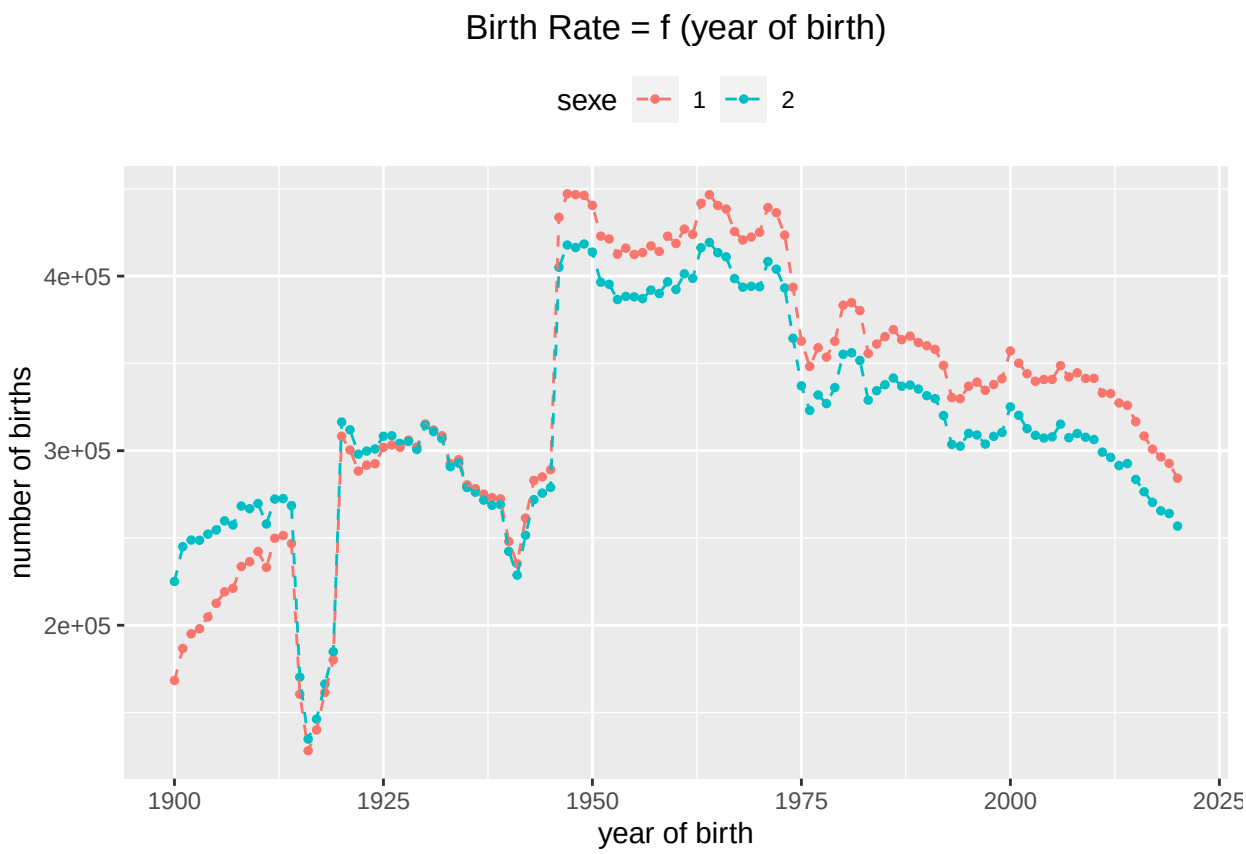
This might lead to think about the evolution of the diversity of birth names (how many distinct names). As shown on the following graph :

```
diversity <- subset(FirstNames[,c('sexe','preusuel', 'annais')]) %>% filter(annais!='XXXX')
diversity <- distinct(diversity, sexe, preusuel, annais)
diversity <- setNames( count(diversity, sexe, annais), c('sexe', 'annais', 'n_preusuel') )

diversity <- transform(diversity, annais = as.numeric(annais))
diversity <- transform(diversity, sexe = as.character(sexe))

ggplot(data=diversity, aes(x=annais, n_preusuel, colour = sexe)) +
  geom_point( size = 1 ) +
  geom_line(linetype = "dashed") +
  labs(title = "diversity of names = f (year of birth)", x = "year of birth", y = "number of distinct na
  theme( plot.title = element_text(hjust = 0.5), legend.position = "top" )
```



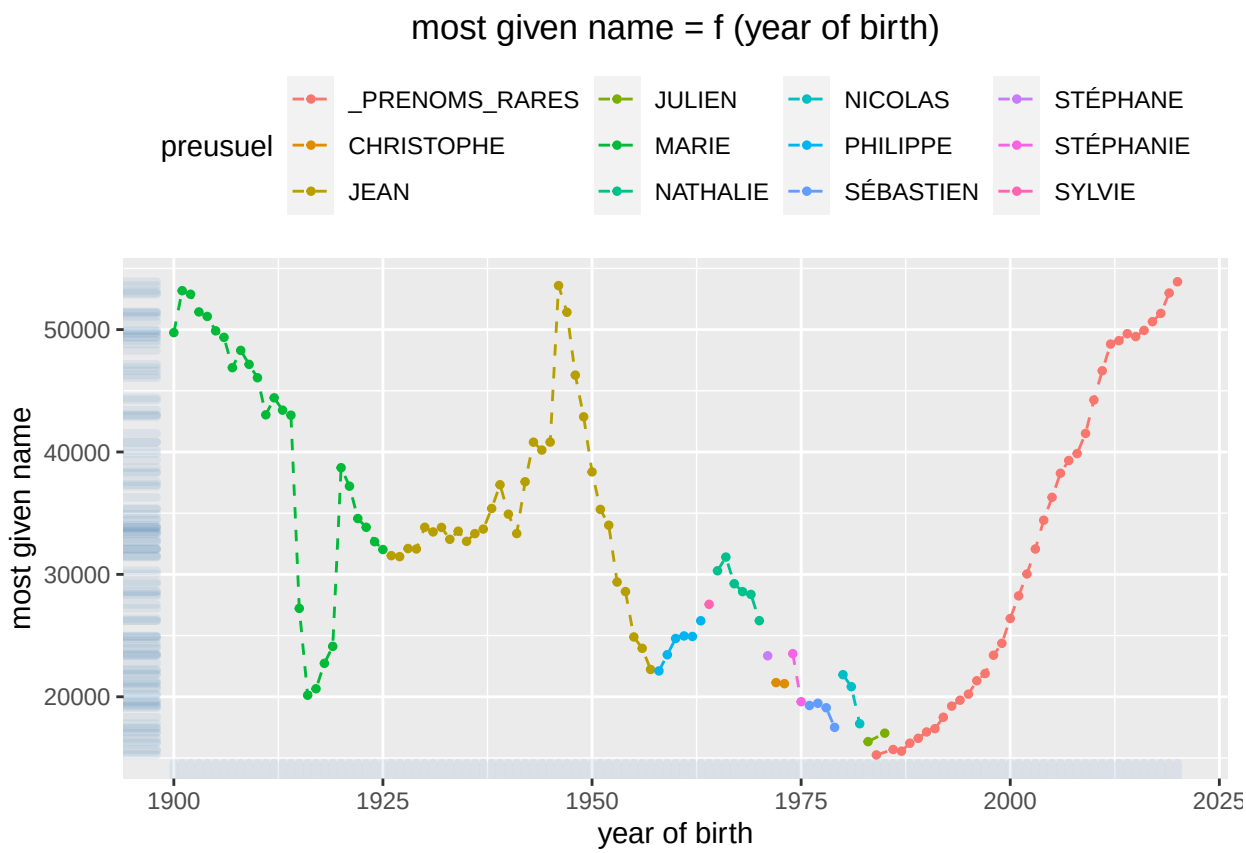
Birth Rate :

```
birth_rate <- subset(FirstNames[,c('sexe','annais','nombre')]) %>% filter(annais!='XXXX')
birth_rate <- setNames( aggregate(birth_rate$nombre, by=list(sexe=birth_rate$sexe, annais=birth_rate$an
                        c('sexe', 'annais', 'sum_nombre'))

birth_rate <- transform(birth_rate, annais = as.numeric(annais))
birth_rate <- transform(birth_rate, sexe = as.character(sexe))

ggplot(data=birth_rate, aes(x=annais, sum_nombre, colour = sexe)) +
  geom_point( size = 1 ) +
  geom_line(linetype = "dashed") +
  labs(title = "Birth Rate = f (year of birth)", x = "year of birth", y = "number of births") +
  theme( plot.title = element_text(hjust = 0.5), legend.position = "top" )
```

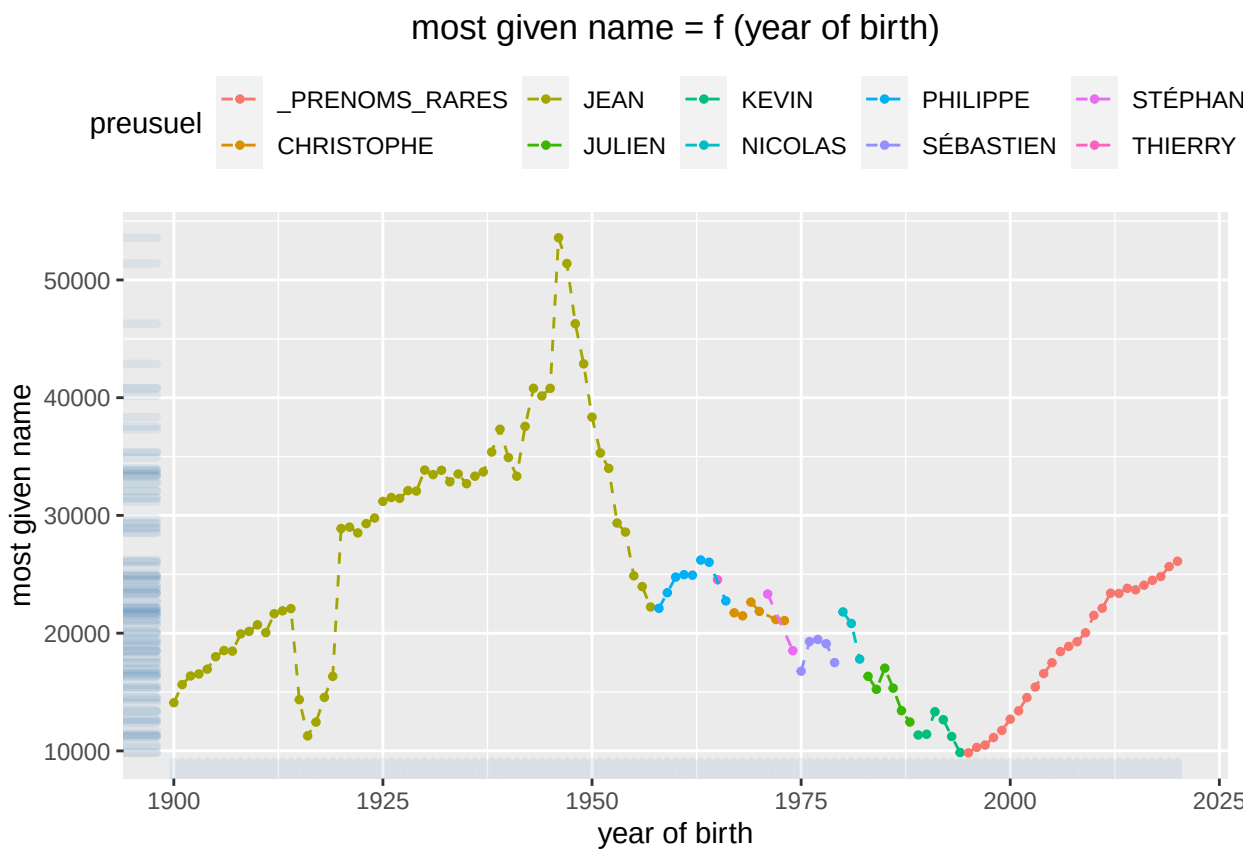


Compute the most given firstname per year

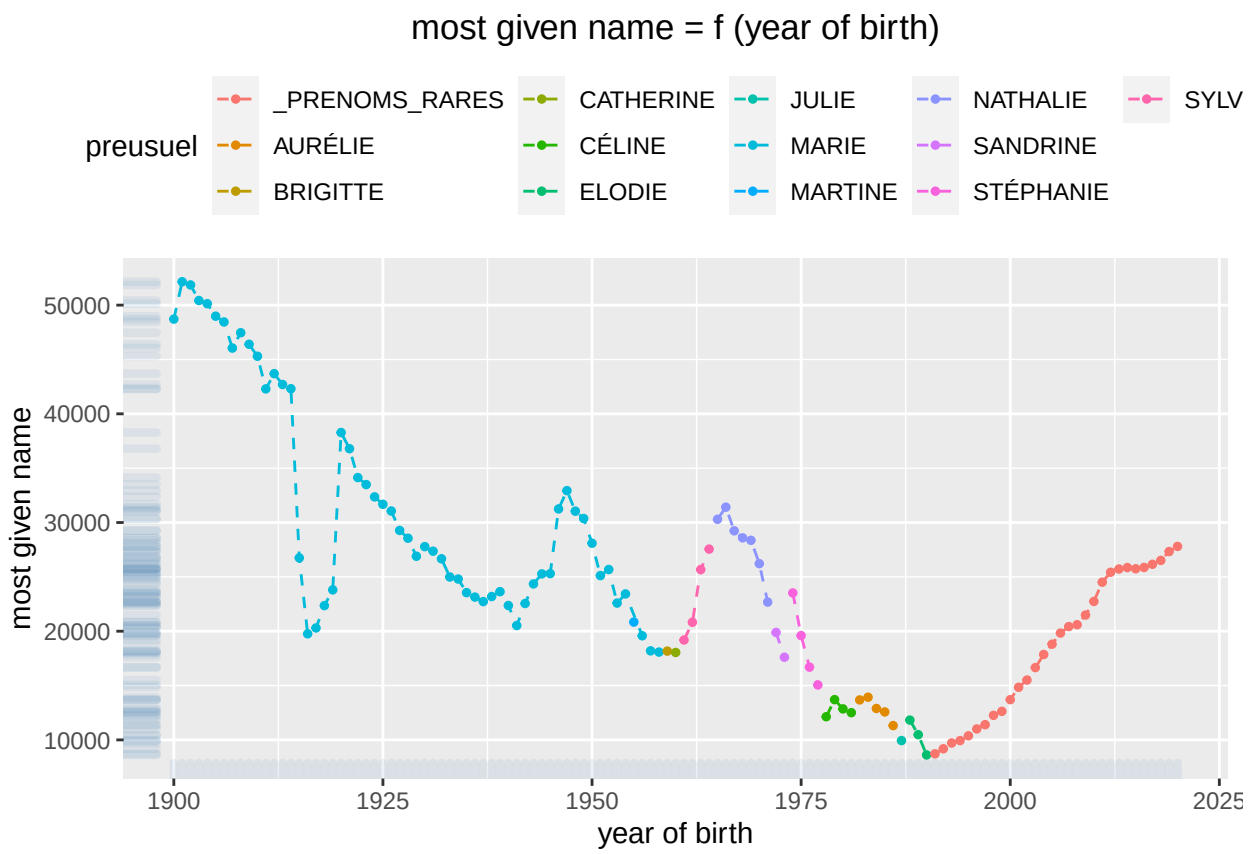
```
all_names <- subset(FirstNames[,c('preusuel', 'annais', 'nombre')]) %>% filter(annais!='XXXX')

all_names <- transform(all_names, nombre = as.numeric(nombre), annais = as.numeric(annais))
all_names <- setNames( aggregate(all_names$nombre, by=list(preusuel=all_names$preusuel, annais=all_names

all_names <- all_names %>% group_by(annais) %>% top_n(1, sum_nombre)

# extract the first occurance of preusuel
first_appearance <- all_names[match(unique(all_names$preusuel), all_names$preusuel),]

ggplot(data=all_names, aes(x=annais, sum_nombre, colour=preusuel)) +
  geom_point( size = 1 ) + #shape = "."
  geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "most given name = f (year of birth)", x = "year of birth", y = "most given name") +
  theme( plot.title = element_text(hjust = 0.5) , legend.position = "top" )
```




```
all_names <- subset(FirstNames[,c('sexe','preusuel', 'annais', 'nombre')]) %>% filter(annais!='XXXX')

all_names <- transform(all_names, nombre = as.numeric(nombre), annais = as.numeric(annais))
all_names <- transform(all_names, sexe = as.character(sexe))

all_names <- setNames( aggregate(all_names$nombre, by=list(sexe=all_names$sexe, preusuel=all_names$preus
                                 FUN=sum) , c('sexe', 'preusuel', 'annais', 'sum_nombre'))

#all_names <- all_names %>% group_by(annais) %>% top_n(1, sum_nombre)

#1
all_names_1 <- subset(all_names, sexe=='1')
all_names_1 <- all_names_1 %>% group_by(annais) %>% top_n(1, sum_nombre)

#2
all_names_2 <- subset(all_names, sexe=='2')
all_names_2 <- all_names_2 %>% group_by(annais) %>% top_n(1, sum_nombre)

# extract the first occurance of preusuel
#first_appearance <- all_names[match(unique(all_names$preusuel), all_names$preusuel),]

ggplot(data=all_names_1, aes(x=annais, sum_nombre, colour=preusuel)) +
  geom_point( size = 1 ) + #shape = "."
  geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "most given name = f (year of birth)", x = "year of birth", y = "most given name") +
  theme( plot.title = element_text(hjust = 0.5) , legend.position = "top" )
```




```
ggplot(data=all_names_2, aes(x=annais, sum_nombre, colour=preusuel)) +
  geom_point( size = 1 ) + #shape = "."
  geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "most given name = f (year of birth)", x = "year of birth", y = "most given name") +
  theme( plot.title = element_text(hjust = 0.5) , legend.position = "top" )
```



gg

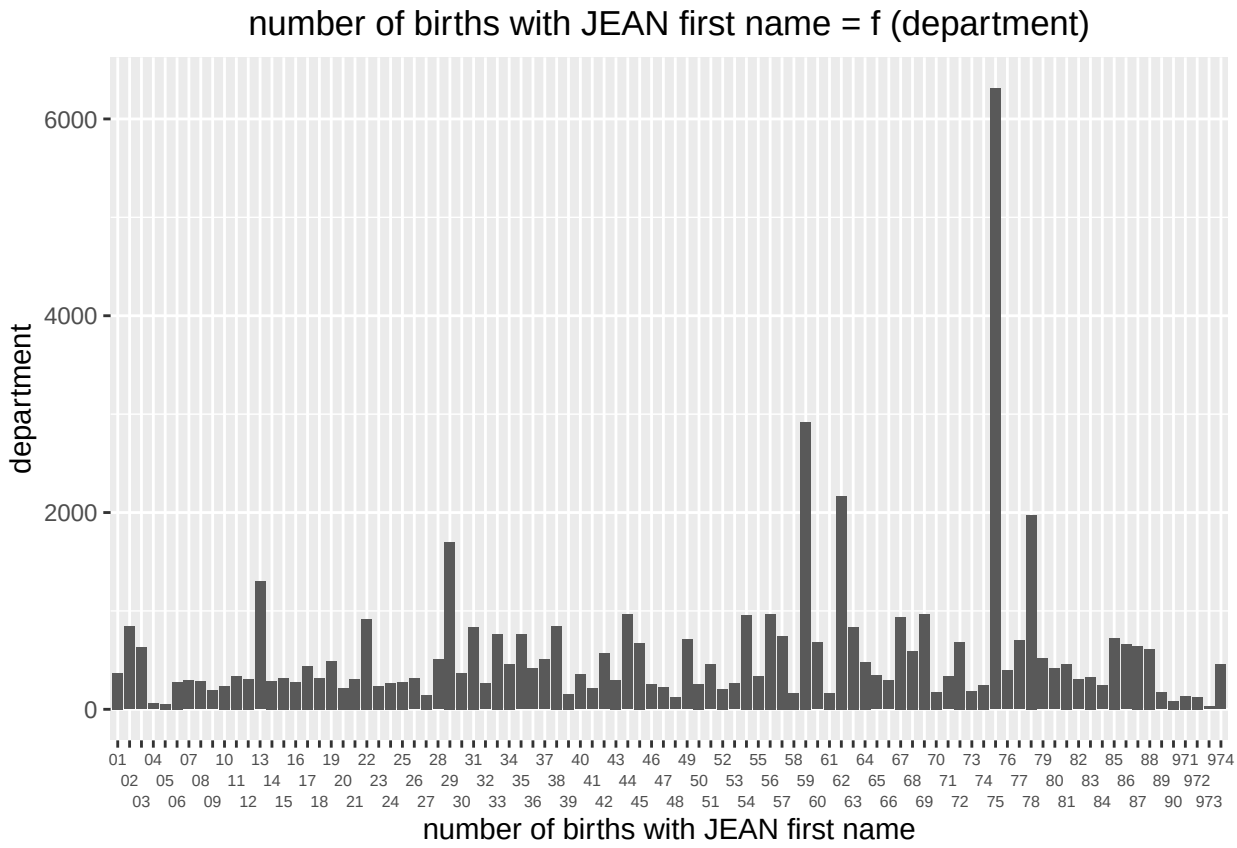
```
#library("gridExtra")
#ggarrange(plot1, plot2 + font("x.text", size = 10),
#                    ncol = 1, nrow = 2)
```

Synthesis : TODO

Any correlation between the first name and the localization (department) ?

Let's take for insatance the most used name on 1946 (JEAN) ; 53584 (the peak), and see its distribution over the departments.

```
jean <- FirstNames[,c('preusuel', 'annais', 'dpt', 'nombre')] %>% filter(preusuel=='JEAN', annais=='194
jean <- subset(jean[,c('dpt','nombre')])
jean <- transform(jean, nombre = as.numeric(nombre))

ggplot(data=jean, aes(x=dpt, nombre)) +
  geom_bar(stat="identity") +
  scale_x_discrete(guide = guide_axis(n.dodge=3), ) +
  labs(title="number of births with JEAN first name = f (department)", x="number of births with JEAN fi
  theme( plot.title = element_text(hjust = 0.5),   axis.text.x = element_text(size = 6), legend.positio
```



Comment : TODO