French given names per year per department

Edited by : **Oussama Oulkaid**. Template authors : **Lucas Mello Schnorr, Jean-Marc Vincent**

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: 793,681 x 5
##     sexe preusuel       annais dpt   nombre
##    <dbl> <chr>          <chr>  <chr> <chr>
##  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 793,671 more rows
```

Analysing first names frequencies

Let's choose a first name (ALBERT), and analyse its frequency.

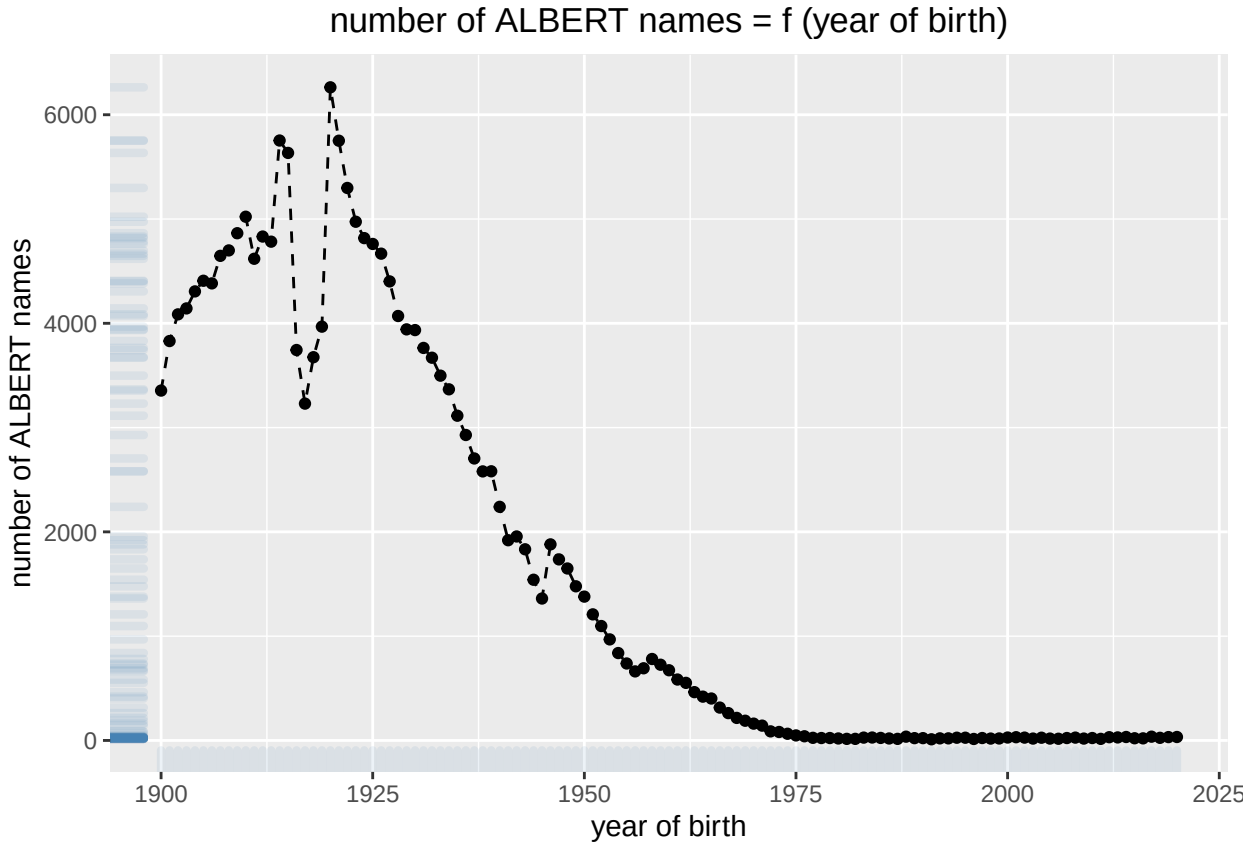
The data set contains a single gender :

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

```
## # A tibble: 1 x 1
##    sexe
##   <dbl>
```

```
## 1     1
```

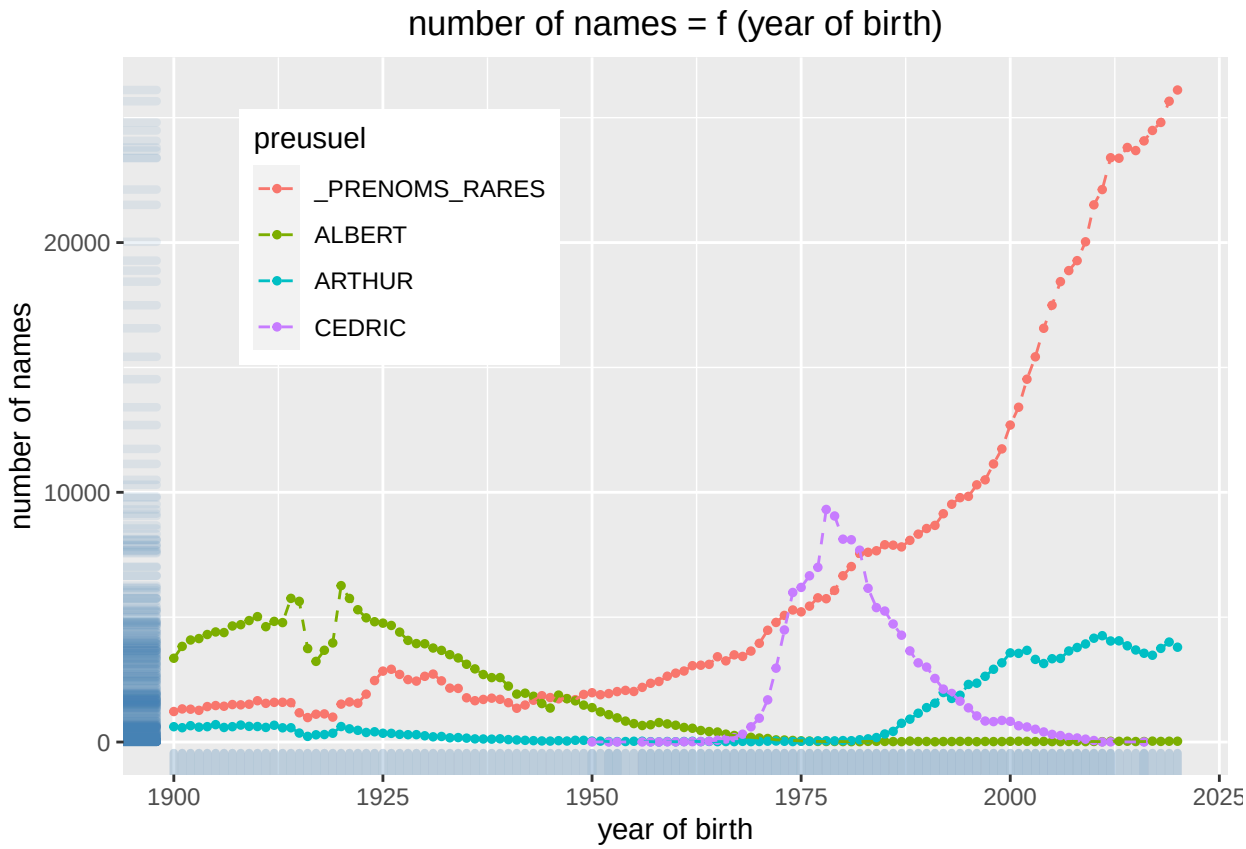
```
albert <- subset(FirstNames[,c('preusuel', 'annais', 'nombre')], preusuel == 'ALBERT') %>% group_by(anna
albert <- subset(albert[,c('annais', 'nombre')])
albert <- transform(albert, nombre = as.numeric(nombre), annais = as.numeric(annais))
albert <- setNames( aggregate(albert$nombre, by=list(annais=albert$annais), FUN=sum) , c('annais', 'sum_
ggplot(data=albert, aes(x=annais, sum_nombre)) +
  geom_point() +
  geom_line(linetype = "dashed") +
  geom_rug(col="steelblue",alpha=0.1, size=1.5) +
  labs(title = "number of ALBERT names = f (year of birth)", x = "year of birth", y = "number of ALBERT
  theme(plot.title = element_text(hjust = 0.5))
```



Comment : After 1975, the number of new births with the name ALBERT was very low.

Now, let's compare several names' frequencies.

```
all_names <- subset(FirstNames[,c('preusuel', 'annais', 'nombre')]) %>% group_by(preusuel) %>% filter(an
sample_names <- subset( all_names[,c('preusuel', 'annais', 'nombre')], preusuel=='ALBERT' | preusuel=='

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=

ggplot(data=sample_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 = c(0.25, 0.75) )
```



Comment : TODO

Let's try to plot all the names :p

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

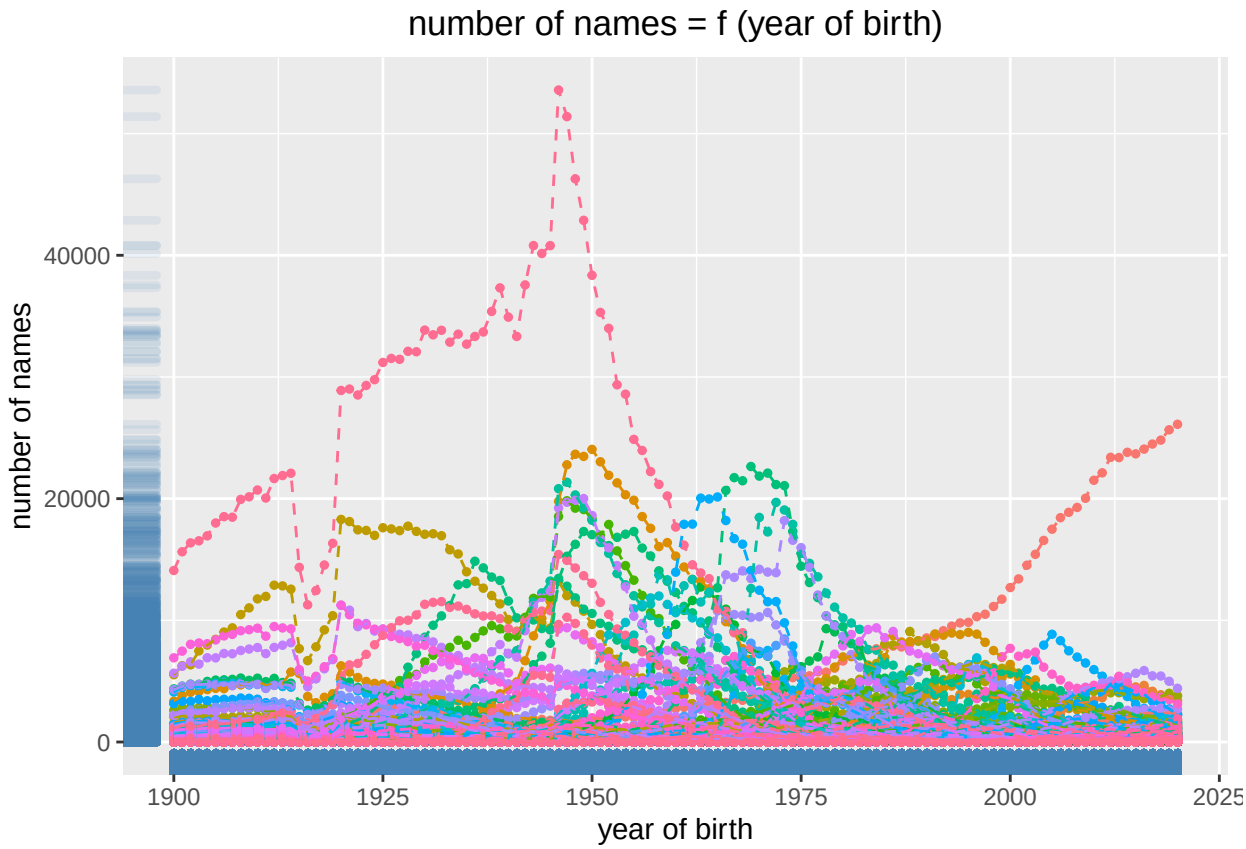
```
all_names <- transform(all_names, nombre = as.numeric(nombre), annais = as.numeric(annais))
```

```
## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion
```

```
## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion
```

```
all_names <- setNames( aggregate(all_names$nombre, by=list(preusuel=all_names$preusuel, annais=all_name
```

```
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 = "none" )
```



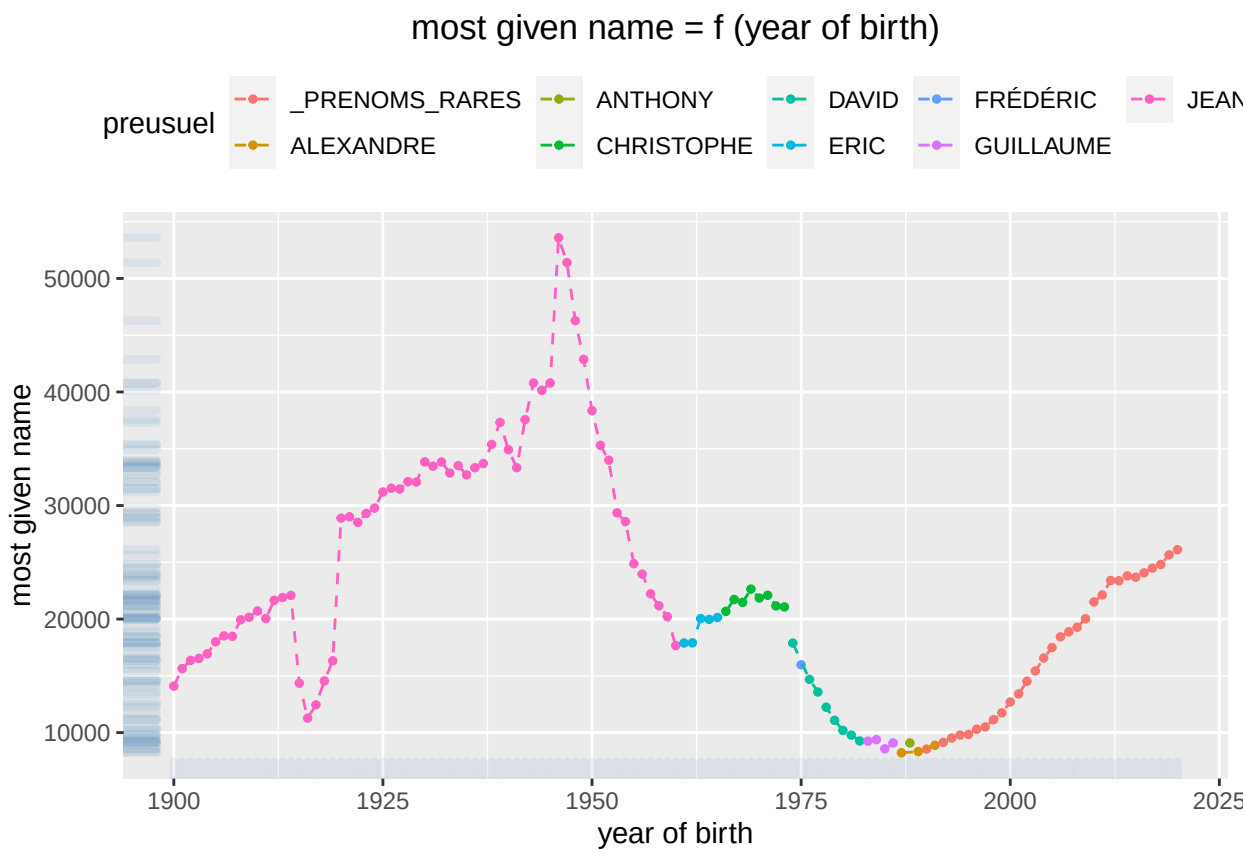
Comment : TODO

Now, let's 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))
```

```
## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion

## Warning in eval(substitute(list(...)), `_data`, parent.frame()): NAs introduced
## by coercion
```

```
all_names <- setNames( aggregate(all_names$nombre, by=list(preusuel=all_names$preusuel, annais=all_name

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" )
```

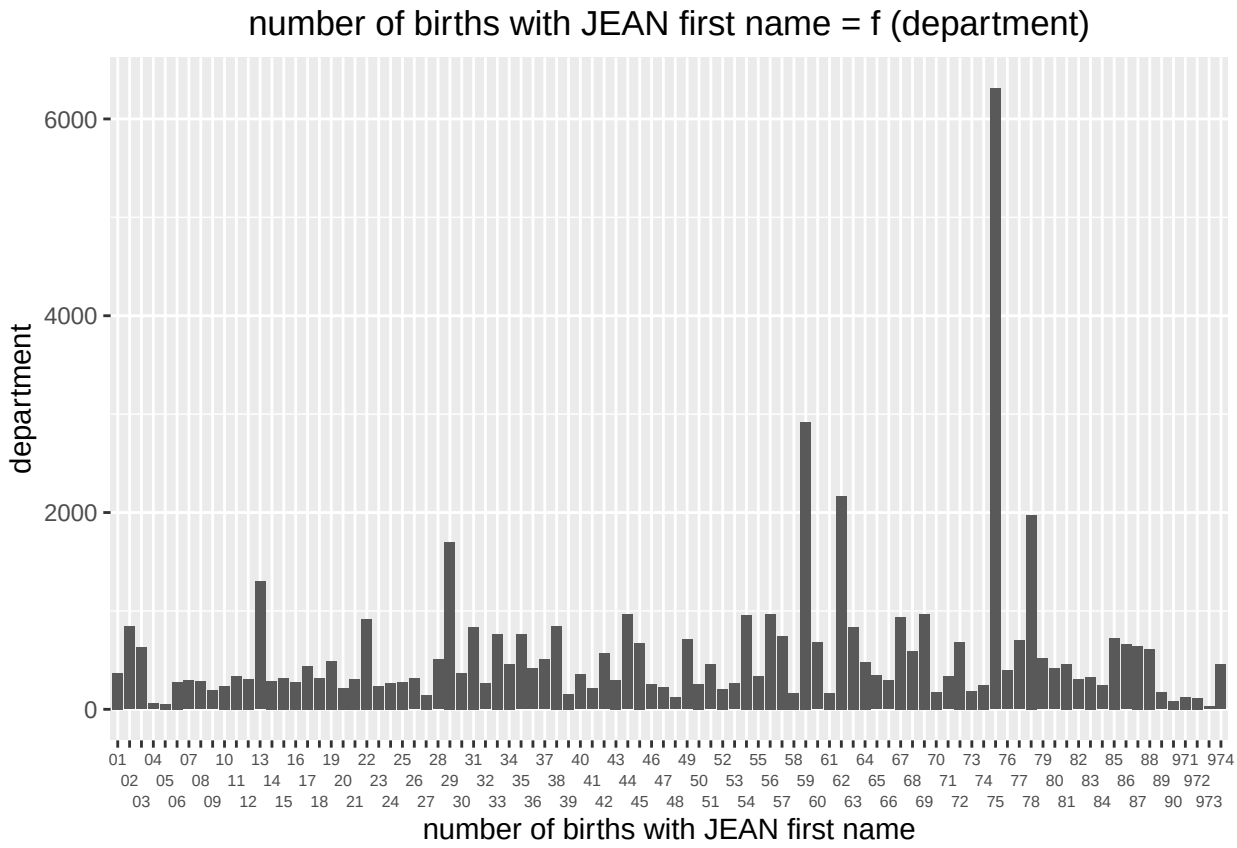


Synthesis : TODO

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