French given names per year per department

November 20, 2021

0.0.1 Author: Folajimi Olaniyan

0.1 Introduction

In this work, we will like to study the evolution of French first names across the departments of France. In particular, we focus on the *INSEE Given Name Data set* and then use R to analyse and answer questions about the dataset.

[ ]:
```
library(ggplot2)
library(tidyverse)
library(dplyr)
```

0.1.1 Download Raw Data from the website

[2]:
```
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)
```

0.1.2 Build the Dataframe from file

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

[4]:
```
summary(FirstNames)
```

```
      sexe         preusuel            annais              dpt
 Min.   :1.000   Length:3727553     Length:3727553     Length:3727553
 1st Qu.:1.000   Class :character   Class :character   Class :character
 Median :2.000   Mode  :character   Mode  :character   Mode  :character
 Mean   :1.536
 3rd Qu.:2.000
 Max.   :2.000
     nombre
 Min.   :   3.00
 1st Qu.:   4.00
 Median :   7.00
 Mean   :  23.23
```

```
 3rd Qu.:  19.00
 Max.   :6310.00
```

0.2 Data cleaning

Our data set contains some rows where either the year is missing or the department is missing. In particular, when the year is missing, it is represented as 'XXXX', and when the department is missing, it is represented as 'XX'. Our first step will be to filter these from the data set

[5]:
```
FirstNames <- filter(FirstNames, annais !='XXXX' & dpt != 'XX')
```

0.2.1 20 Most common male names

Let us look at the 20 most common male names

[6]:
```
FirstNames %>% filter(sexe==1) %>% group_by(preusuel) %>% summarise(count =␣
↪sum(nombre)) %>% arrange(desc(count)) %>% top_n(20)
```

```
Selecting by count
```

A tibble: 20 × 2

preusuel <chr>	count <dbl>
JEAN	1912848
PIERRE	891170
MICHEL	818001
_PRENOMS_RARES	798128
ANDRÉ	709568
PHILIPPE	535200
LOUIS	523561
RENÉ	514553
ALAIN	504103
JACQUES	480161
BERNARD	466884
MARCEL	466167
DANIEL	432581
ROGER	421803
PAUL	420306
ROBERT	417010
CLAUDE	408989
HENRI	402493
CHRISTIAN	402452
GEORGES	402421

0.2.2 20 Most common female names

Let us look at the 20 most common female names

[12]:
```
FirstNames %>% filter(sexe==2) %>% group_by(preusuel) %>% summarise(count =␣
↪sum(nombre)) %>% arrange(desc(count)) %>% top_n(20)
```

```
Selecting by count
```

A tibble: 20 × 2

preusuel <chr>	count <dbl>
MARIE	2231903
_PRENOMS_RARES	853451
JEANNE	556897
FRANÇOISE	399509
MONIQUE	397739
CATHERINE	391518
NATHALIE	379691
ISABELLE	374129
JACQUELINE	370277
ANNE	362614
SYLVIE	361407
MARTINE	317470
MADELEINE	301884
NICOLE	290993
SUZANNE	286507
HÉLÈNE	279677
CHRISTINE	277489
LOUISE	268583
MARGUERITE	268063
DENISE	261567

Jean and Marie are the most common first names in the data set.

0.2.3 Unique names in data set

[18]:
```
length(unique(FirstNames$preusuel))
```

15271

There are 15271 unique names in the data set.

[ ]:
```
nrow(unique(FirstNames['annais']))
```

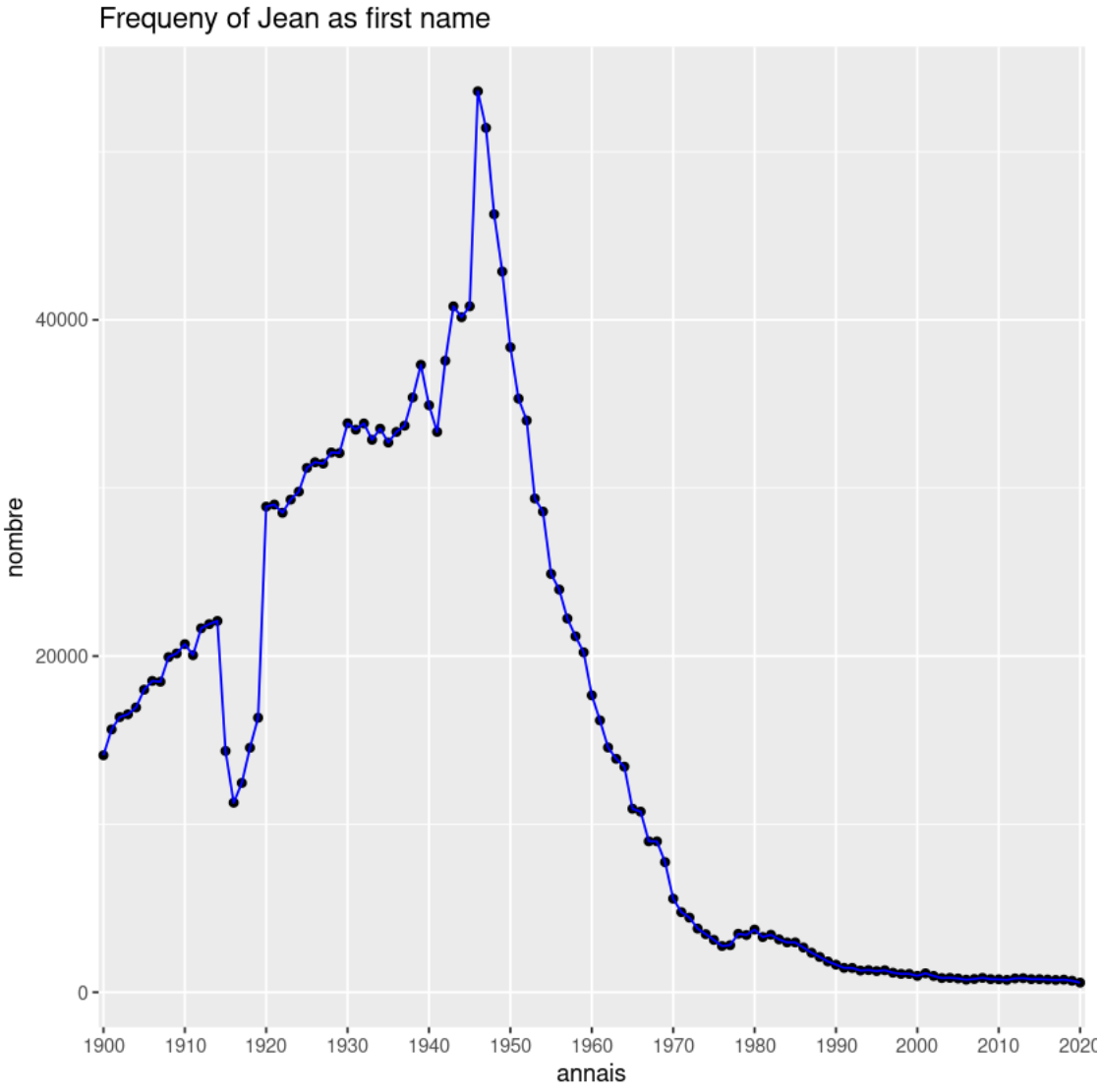
In total, there are 121 different years in the data set.

0.3 Frequency of the first name JEAN over time

[20]:
```
get_name_freq <- function(data, name) {
  freq <- filter(data, preusuel == name) %>% group_by(annais) %>%␣
↪summarise(nombre = sum(nombre))
  return(freq)
```

```
}
```

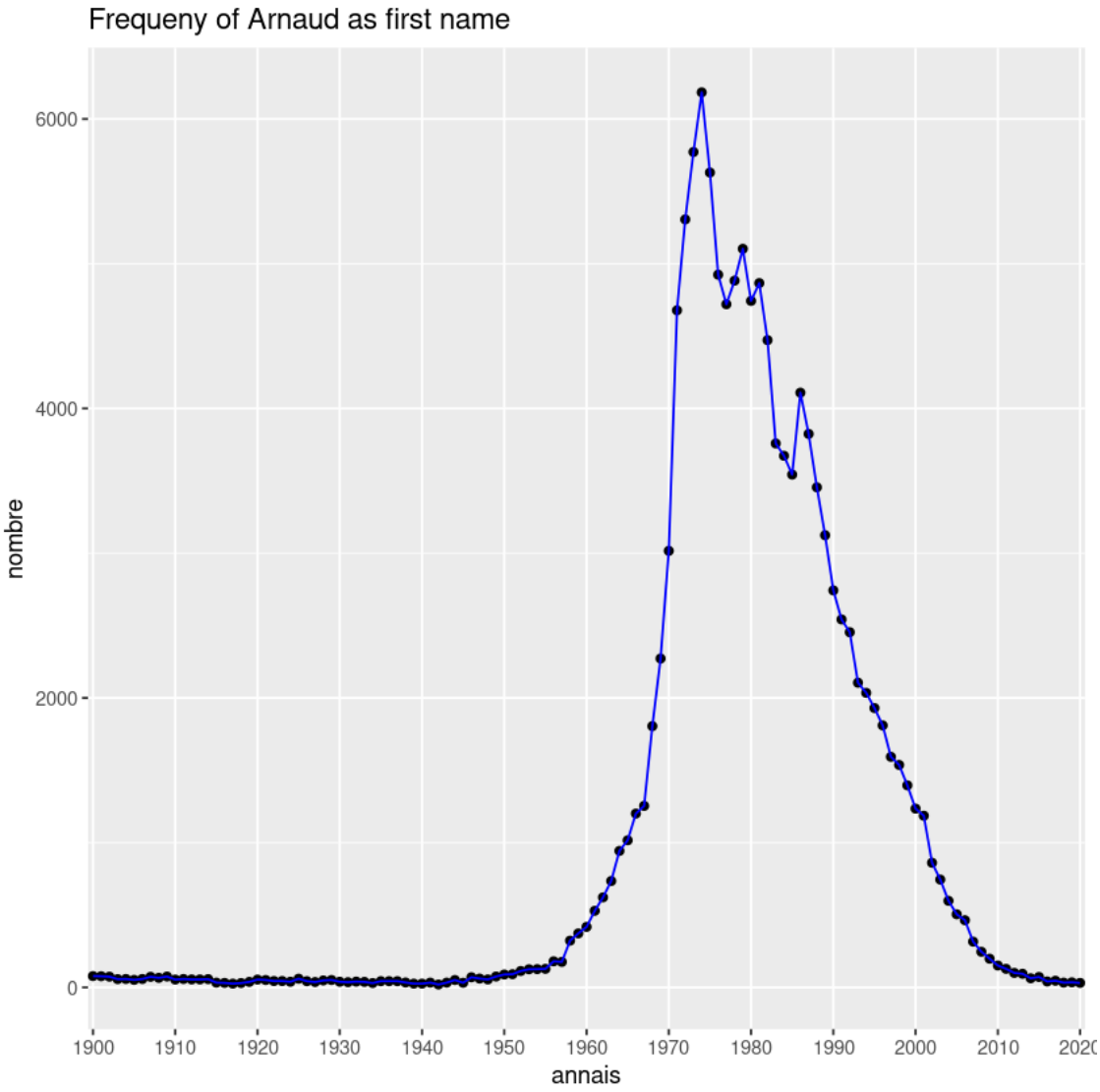
[70]:
```
jean_freq <- get_name_freq(FirstNames, 'JEAN')
scale <- scale_x_discrete(breaks = round(seq(min(jean_freq$annais),␣
 ↪max(jean_freq$annais), by = 10),1))
ggplot(data=jean_freq, aes(x=annais, y=nombre, group = 1)) + ggtitle("Frequeny␣
 ↪of Jean as first name") + geom_point() + geom_line(color="blue") + scale
```



The graph shows that while the name Jean enjoyed popularity from the 1900's until a peak in 1945, it has continues to drop in popularity consistently since them.

0.4 Frequency of the first name ARNAUD over time

[43]:
```
freq_by_year<- filter(.data = FirstNames, preusuel == 'ARNAUD') %>%␣
 ↪group_by(annais) %>% summarise(nombre = sum(nombre))
ggplot(freq_by_year, mapping = aes(x=annais, y=nombre, group = 1)) +␣
 ↪ggtitle("Frequeny of Arnaud as first name")+  geom_point() +␣
 ↪geom_line(color="blue") + scale
```

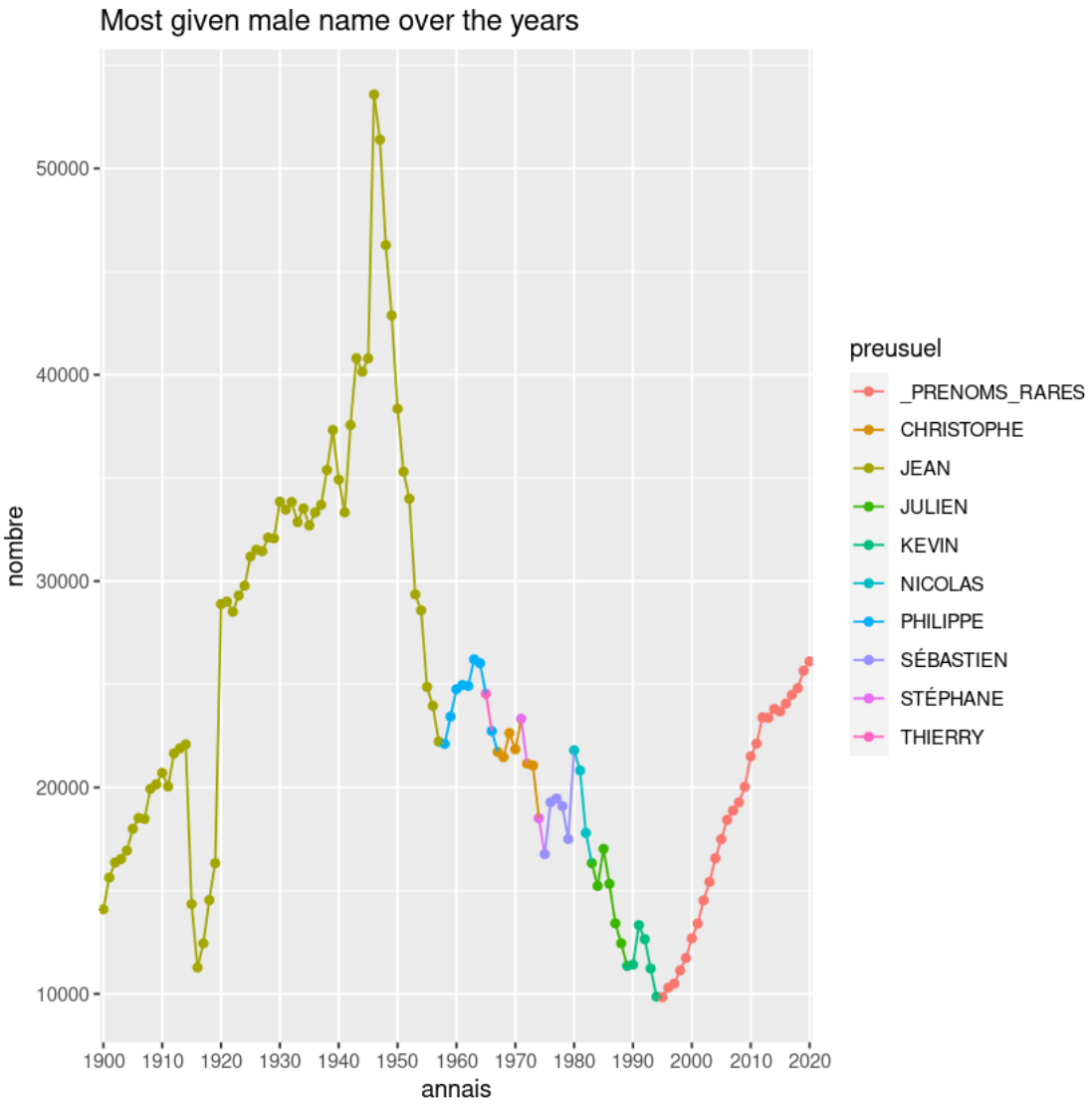


Again the name 'Arnaud' enjoyed popularity from 1960 until a peak in 1975 when it started to decline.

0.5 Most common male names over the years

[66]:
```
x <- FirstNames %>% filter(sexe == 1) %>% group_by(annais, preusuel) %>%␣
↪summarise(nombre = sum(nombre)) %>% top_n(1, nombre)
ggplot(data=x, aes(x=annais, y=nombre, colour=preusuel, group = 1)) +
  geom_point() + geom_line() + scale + ggtitle("Most given male name over the␣
↪years")
```

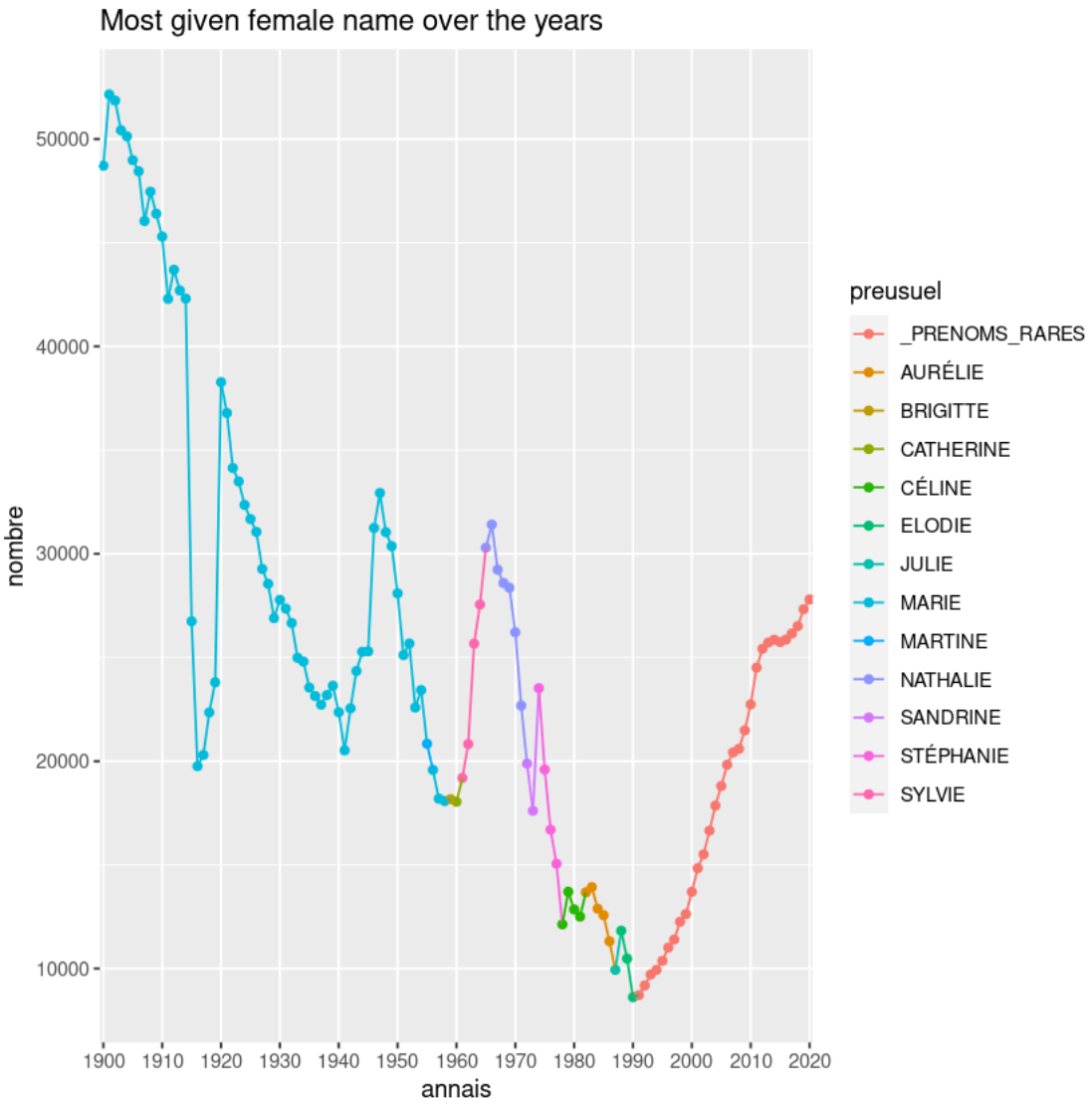
```
`summarise()` has grouped output by 'annais'. You can override using the
`.groups` argument.
```



0.6 Most common Female names over the years

[68]:
```
y <- FirstNames %>% filter(sexe == 2) %>% group_by(annais, preusuel) %>%␣
 ↪summarise(nombre = sum(nombre)) %>% top_n(1, nombre)
ggplot(data=y, aes(x=annais, y=nombre, colour=preusuel, group = 1)) +
  geom_point() + geom_line() + scale + ggtitle("Most given female name over the␣
 ↪years")
```

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



0.7 Synthesis

A common trend in the data is that a lot of names that used too be common the past have lost popularity with new names emerging with more popularity. However, there is a clear trend where rare firstnames are now more common, showing that no single name dominates.

[ ]: