

# French given names per year per department

Simon Pierre Bienvenue BIEND BIEND

November, 2021

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 =";")

## Rows: 3727553 Columns: 5

## -- Column specification -----
## Delimiter: ";"
## chr (3): preusuel, annais, dpt
## dbl (2): sexe, nombre

##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

## inspect the data

We see that we have one year which is in unknown format 'xxxx'

```
100*length(firstNames$annais[firstNames$annais=="XXXX"])/nrow(firstNames)
```

```
## [1] 0.9991541
```

We see that we have almost 1% of rows with this value in our dataset so we can delete it because the ratio is not significant

We see that we have a uncommon noun "PRENOM\_RARE" let see it

```
nrow(newfirstNames %>%filter(preusuel=="_PRENOMS_RARES"))
```

```
## [1] 22035
```

*## With the following command we group the dataframe by preusuel and by annais and we count the total number of newborns*

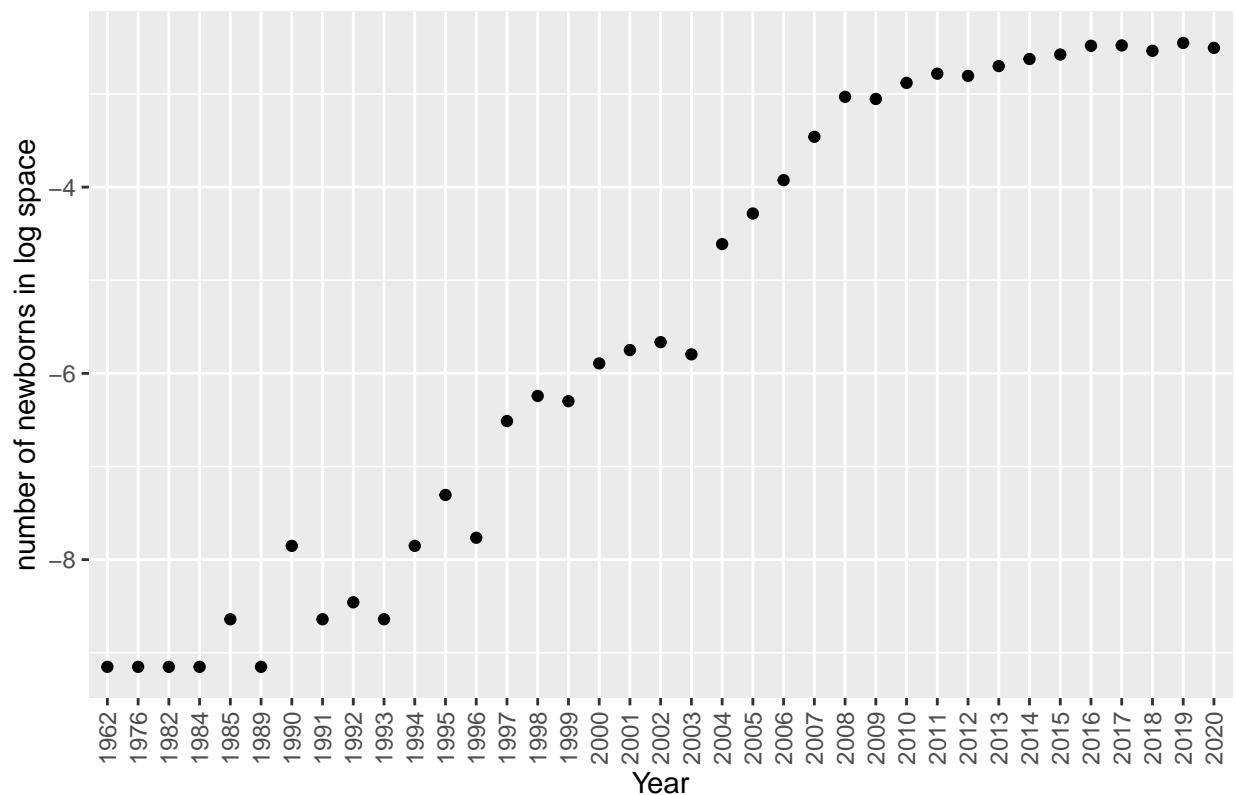
```
newfirstNames<- newfirstNames %>% group_by(preusuel,annais) %>% summarise(frequency = sum(nombre)) %>%
```

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

We see that there are not missing values, in the next chunk we will visualise graphics

```
aaron_df$log_norm_freq <- log(x = aaron_df$frequency/(sum(aaron_df$frequency)))
a1 <- ggplot(data = aaron_df) + geom_point(mapping = aes(x = annais, y = frequency)) + theme(axis.text.x = "none")
a2 <- ggplot(data=aaron_df, aes(x=annais, y=frequency))+geom_bar(stat="identity", fill="steelblue", width=1)
## We plot the data in the log-space
a3 <- ggplot(data = aaron_df) + geom_point(mapping = aes(x = annais, y = log_norm_freq)) + theme(axis.text.x = "none")
a3
```

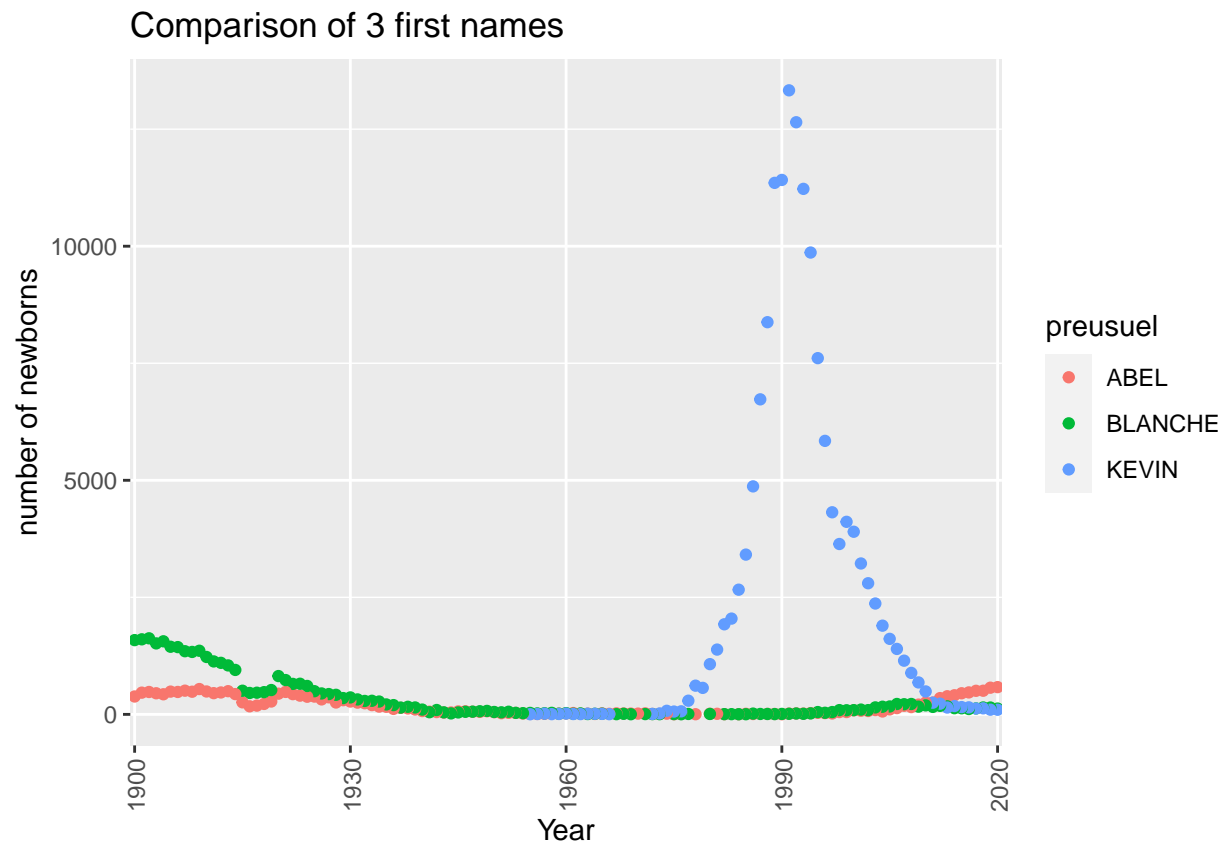
Evolution of the number of newborns with the first name AARON



```

newfirstNames <- drop_na(newfirstNames)
multiple_names <- newfirstNames%>%filter(preusuel=="ABEL" | preusuel=="BLANCHE" |preusuel=="KEVIN")
labels <- seq(1900, 2020, length.out=5)
a4 <- ggplot(data = multiple_names, aes(x = annais, y = frequency, color = preusuel))+ geom_point() + t
a4

```



```

multiple_names$log_norm_freq <- log(x = multiple_names$frequency/(sum(multiple_names$frequency)))
a5 <- ggplot(data = multiple_names, aes(x = annais, y = log_norm_freq, color = preusuel))+ geom_point()
a5

```



*## We can easily compare frequency of birth in the log space because*

With this graph we can see that the first name Kevin has been most popular than the other ones between 1975 and ~2005 but after this it decreases in popularity and the first name ABEL become most popular

```
names_per_year <- firstNames %>% group_by(preusuel, annais, sexe) %>% summarise(frequency = sum(nombre))
```

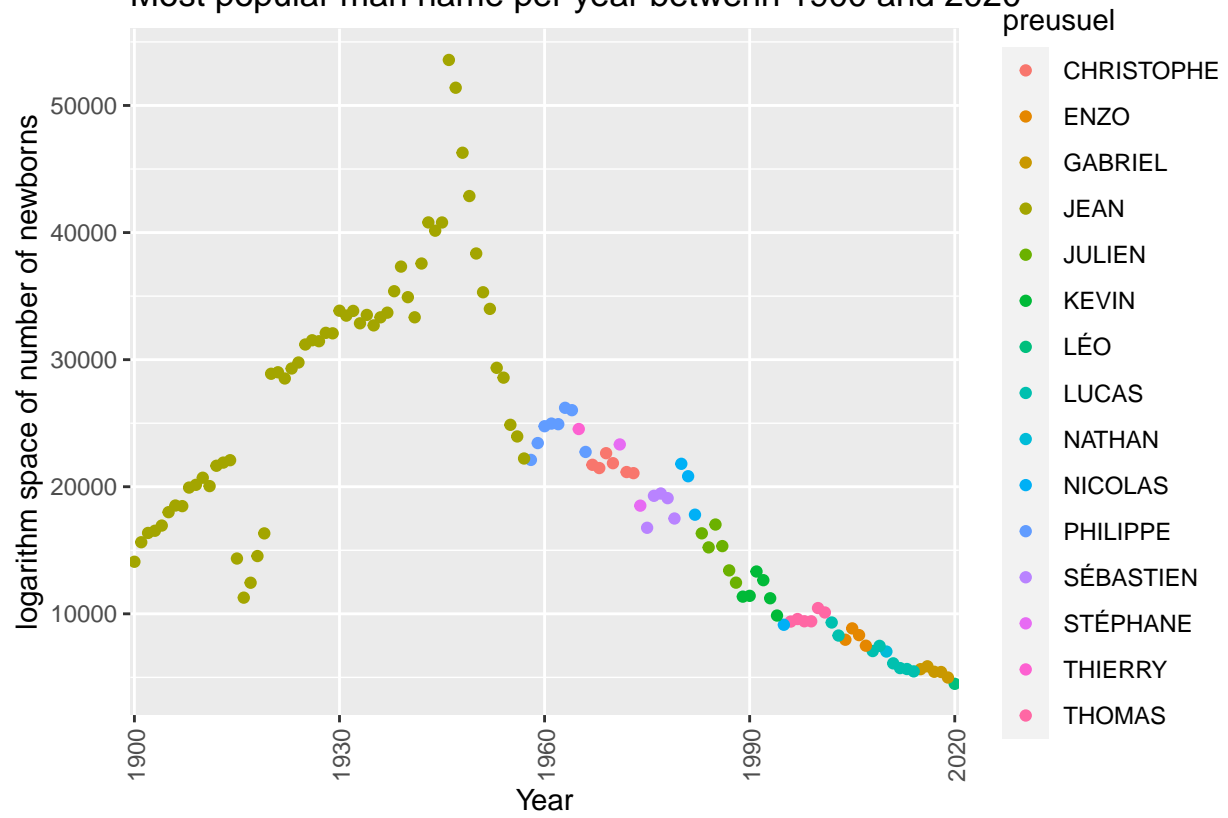
## 'summarise()' has grouped output by 'preusuel', 'annais'. You can override using the '.groups' argument

```
names_per_year <- names_per_year %>% filter(annais != "XXXX")
names_per_year <- names_per_year %>% filter(preusuel != "_PRENOMS_RARES")
names_per_year <- names_per_year %>% group_by(annais, sexe) %>% top_n(n = 1)
```

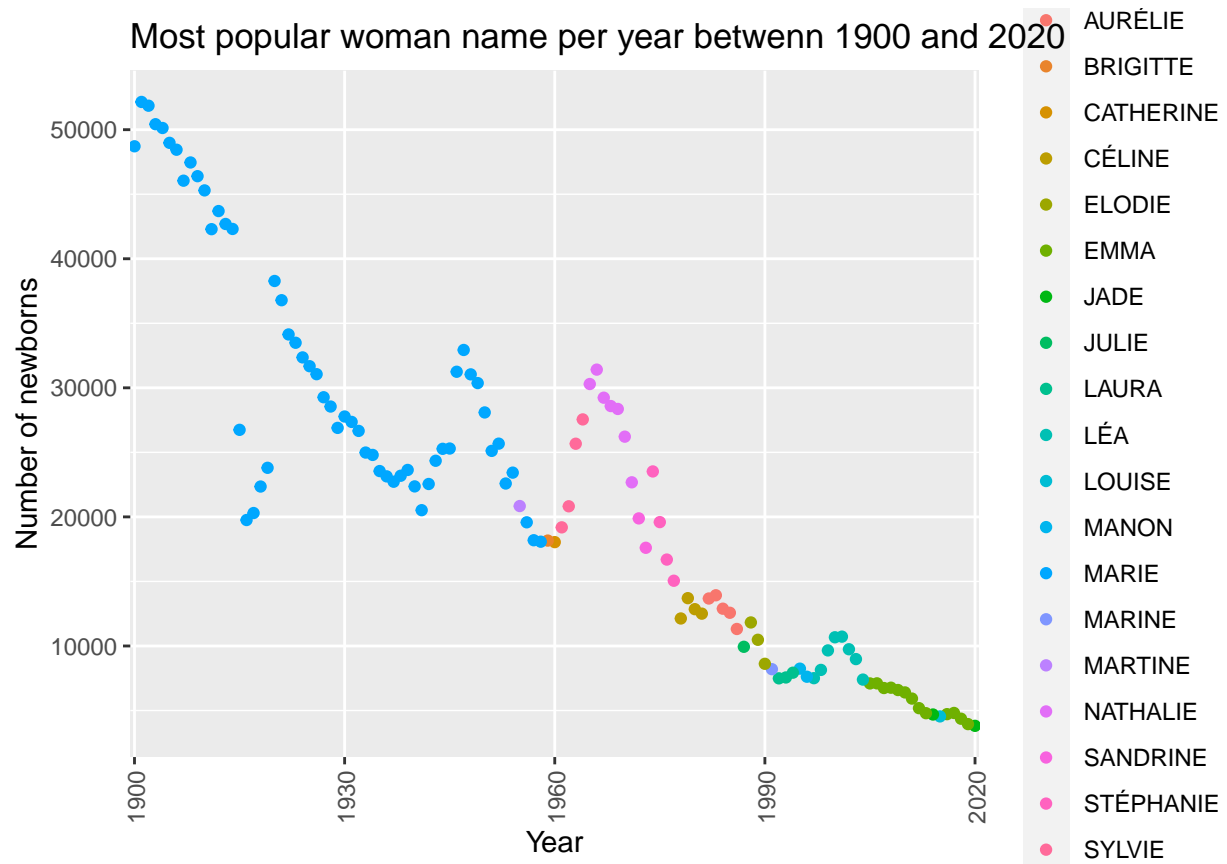
## Selecting by frequency

```
man_per_year <- names_per_year %>% filter(sexe == 1)
a6 <- ggplot(data = man_per_year, aes(x = annais, y = frequency, color = preusuel, fill=preusuel))+ geom_point()
woman_per_year <- names_per_year %>% filter(sexe == 2)
a7 <- ggplot(data = woman_per_year, aes(x = annais, y = frequency, color = preusuel, fill=preusuel))+ geom_point()
a6
```

Most popular man name per year between 1900 and 2020



a7



We see with this graph that between 1900 and almost 1960 Jean was the most popular man name and Marie for the woman side, another remark is that the number of newborns with the most popular name decreases from 1960 both for man and woman side ““