Frequency of french names by department and year

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

Download Raw Data from the website

```
file = "dpt2019_txt.zip"
if(!file.exists(file)){
  download.file("https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2019_csv.zip",
    destfile=file)
}
unzip(file)
```

Build the Dataframe from file. We see that the data has rows with unexpected values.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
FirstNames <- read_delim("dpt2019.csv",delim=";");
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   sexe = col_double(),
##   preusuel = col_character(),
##   annais = col_double(),
##   dpt = col_character(),
##   nombre = col_double()
## )
```

```
## Warning: 36445 parsing failures.
##   row    col expected actual          file
## 10781 annais a double   XXXX 'dpt2019.csv'
## 10782 annais a double   XXXX 'dpt2019.csv'
## 10783 annais a double   XXXX 'dpt2019.csv'
## 10784 annais a double   XXXX 'dpt2019.csv'
## 10787 annais a double   XXXX 'dpt2019.csv'
## ..... ...... ........ ...... ............
## See problems(...) for more details.
```

Show structure of the loaded data and clean from invalid values

```
str(FirstNames)
```

```
## tibble [3,676,682 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ sexe    : num [1:3676682] 1 1 1 1 1 1 1 1 1 1 ...
##  $ preusuel: chr [1:3676682] "_PRENOMS_RARES" "_PRENOMS_RARES" "_PRENOMS_RARES" "_PRENOMS_RARES" ...
##  $ annais  : num [1:3676682] 1900 1900 1900 1900 1900 1900 1900 1900 1900 1900 ...
##  $ dpt     : chr [1:3676682] "02" "04" "05" "06" ...
##  $ nombre  : num [1:3676682] 7 9 8 23 9 4 6 3 11 7 ...
##  - attr(*, "problems")= tibble [36,445 x 5] (S3: tbl_df/tbl/data.frame)
##   ..$ row     : int [1:36445] 10781 10782 10783 10784 10787 10789 10790 12190 12191 12193 ...
##   ..$ col     : chr [1:36445] "annais" "annais" "annais" "annais" ...
##   ..$ expected: chr [1:36445] "a double" "a double" "a double" "a double" ...
##   ..$ actual  : chr [1:36445] "XXXX" "XXXX" "XXXX" "XXXX" ...
##   ..$ file    : chr [1:36445] "'dpt2019.csv'" "'dpt2019.csv'" "'dpt2019.csv'" "'dpt2019.csv'" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   sexe = col_double(),
##   ..   preusuel = col_character(),
##   ..   annais = col_double(),
##   ..   dpt = col_character(),
##   ..   nombre = col_double()
##   .. )
```

Clean data from wrong types

```
unique(select(problems(FirstNames), col, actual))
```

```
## # A tibble: 1 x 2
##   col    actual
##   <chr>  <chr>
## 1 annais XXXX
```

Removing data where the year and name are unknown

```
FirstNamesClean <- filter(FirstNames, annais!="XXXX" & preusuel!="_PRENOMS_RARES")
```

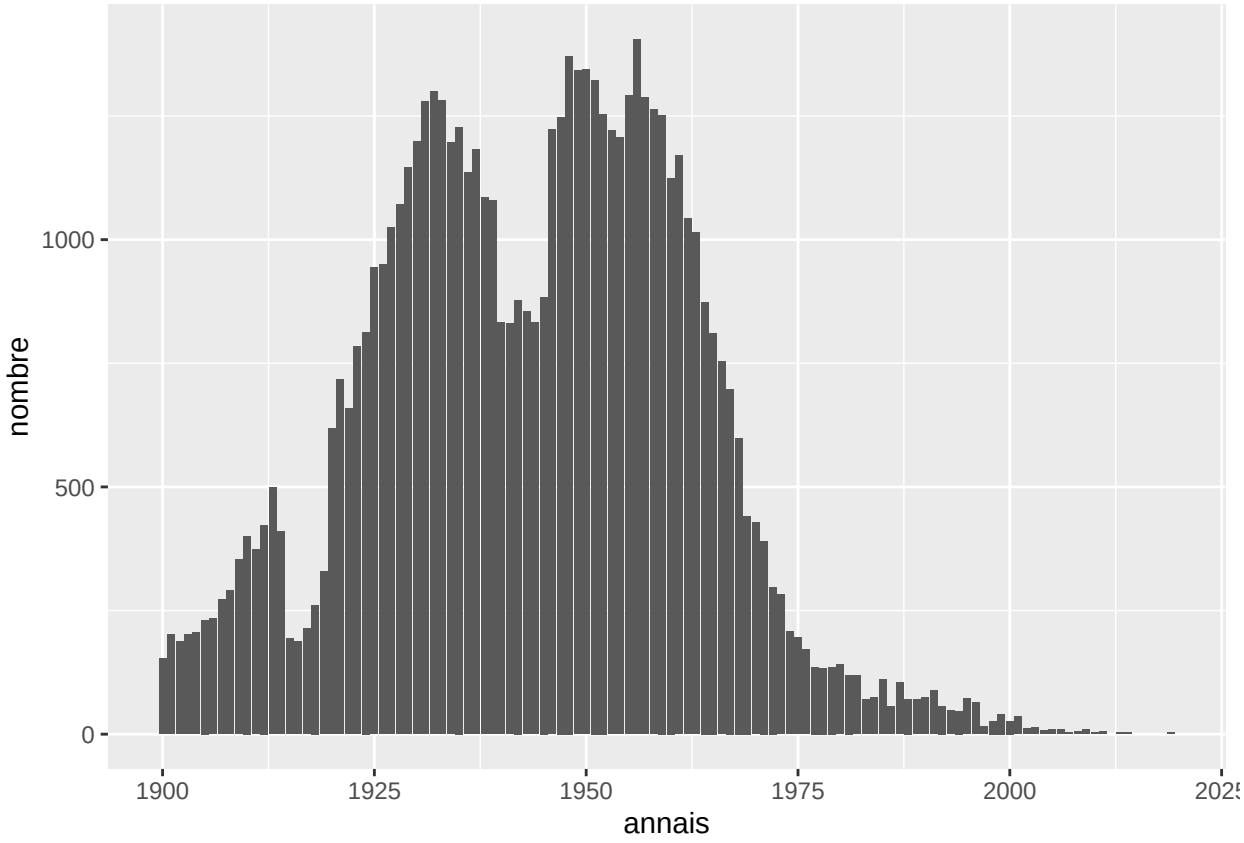
Analysis of a first name occurence over time. Comparison of frequency of several firstnames

Here we are showing frequency along the years of 3 different male names that start with the letter H. We can see that names Hubert and Henri were popular around the beginning of the century, Hugues around the middle and that the name Hugo has become popular around the end of the century.

```
Hubert <- summarise(group_by(select(filter(FirstNamesClean, preusuel=="HUBERT"), annais, nombre), annais
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

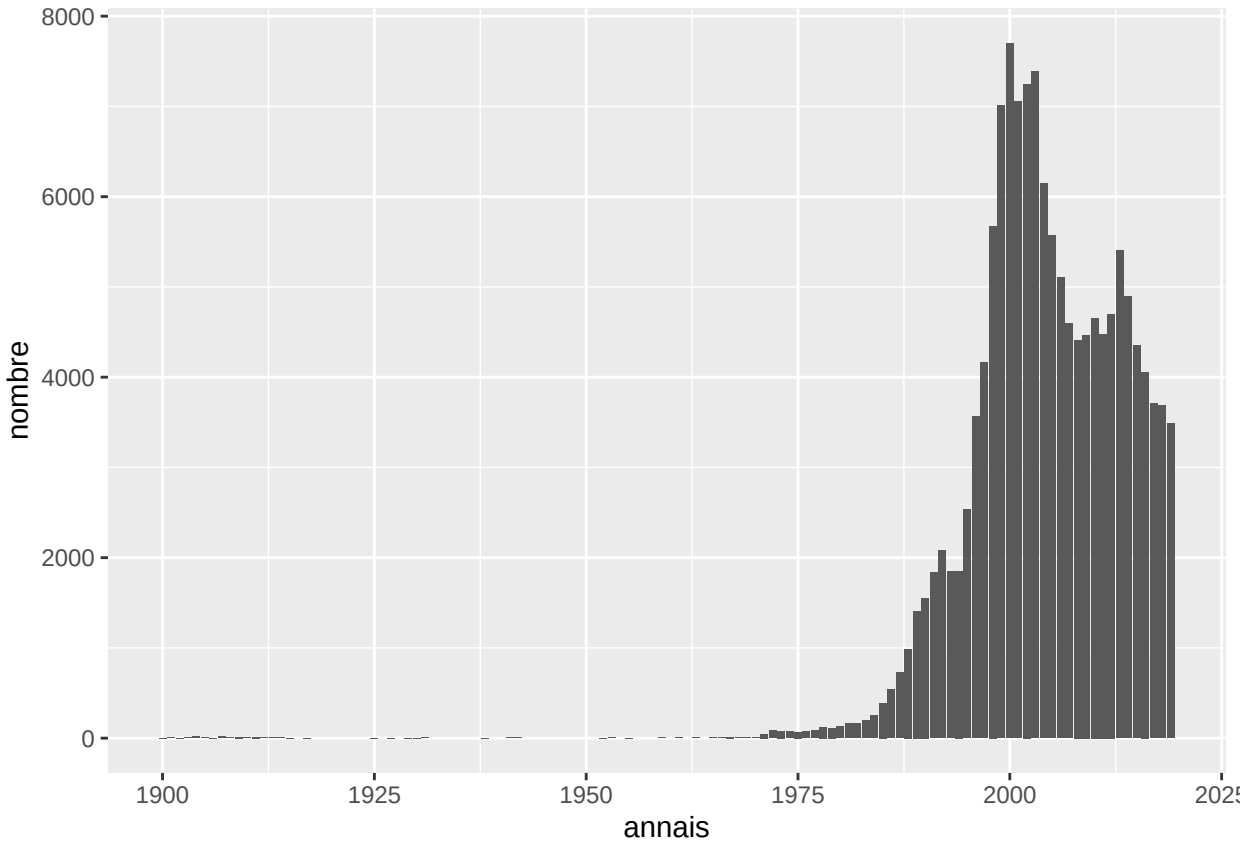
```
ggplot(data=Hubert, aes(x=annais, y=nombre))+geom_bar(stat="identity")
```




```
Hugo <- summarise(group_by(select(filter(FirstNamesClean, preusuel=="HUGO"), annais, nombre), annais), n
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

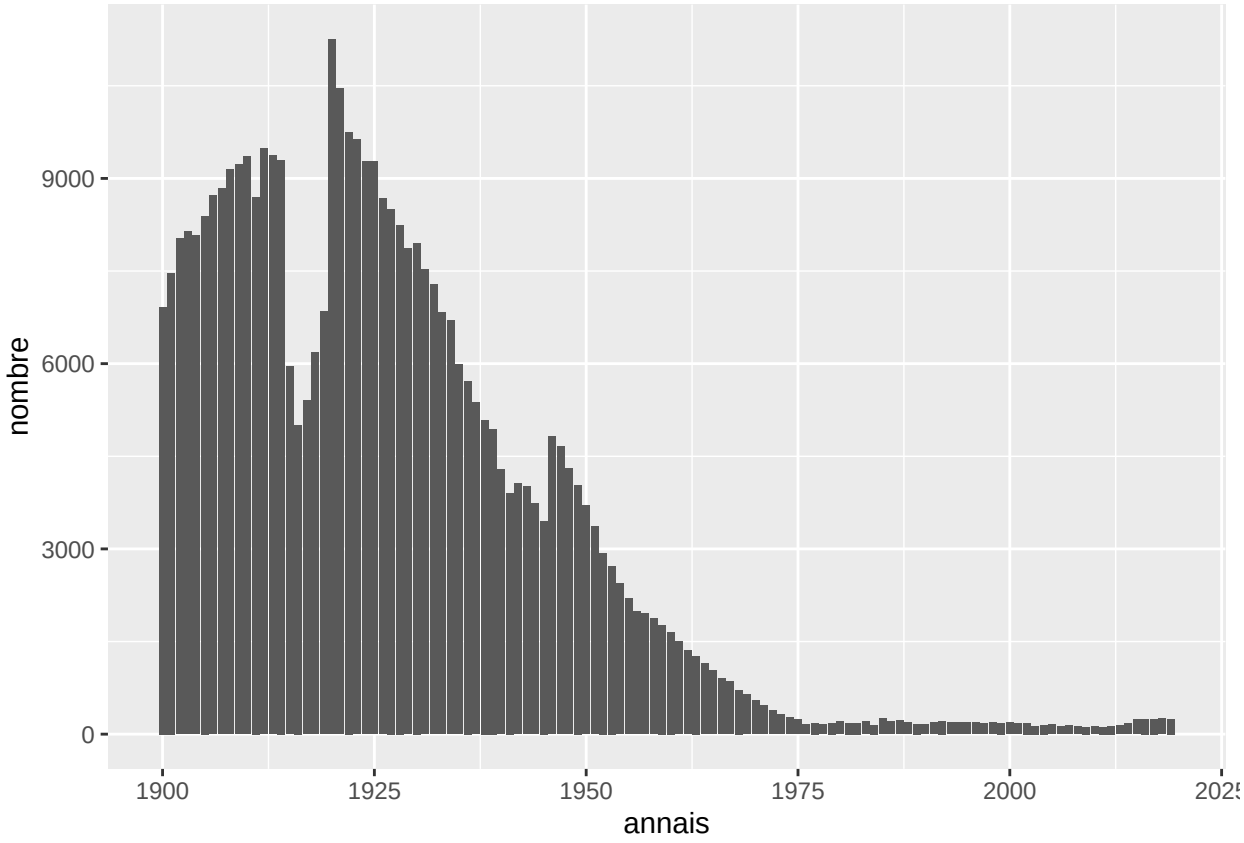
```
ggplot(data=Hugo, aes(x=annais, y=nombre))+geom_bar(stat="identity")
```




```
Henri <- summarise(group_by(select(filter(FirstNamesClean, preusuel=="HENRI"), annais, nombre), annais)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

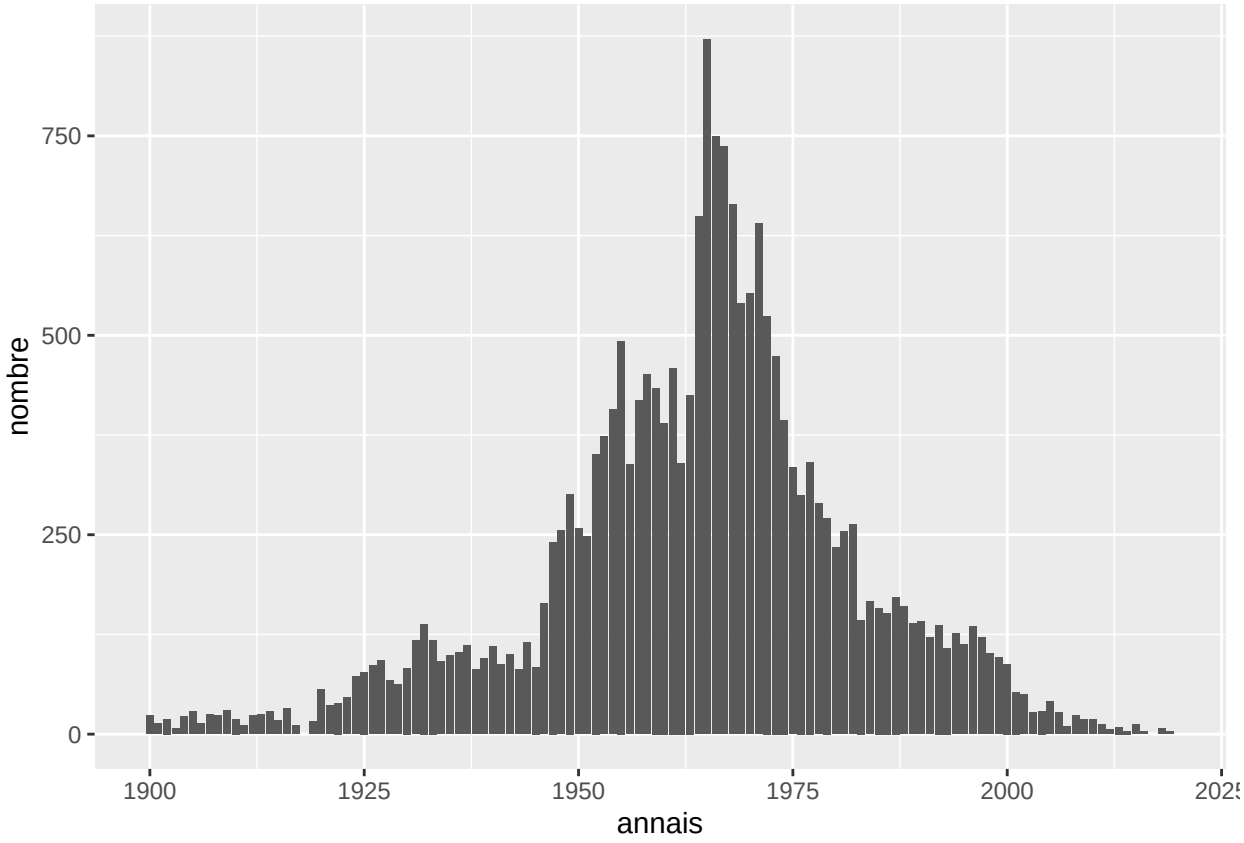
```
ggplot(data=Henri, aes(x=annais, y=nombre))+geom_bar(stat="identity")
```




```
Hugues <- summarise(group_by(select(filter(FirstNamesClean, preusuel=="HUGUES"), annais, nombre), annais
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
ggplot(data=Hugues, aes(x=annais, y=nombre))+geom_bar(stat="identity")
```



Establish by gender the most given firstname by year. Analyse the evolution of the most frequent firstname.

```
HighestOccurence <- unique(filter(group_by(summarise(group_by(FirstNamesClean, annais, preusuel, sexe),
```

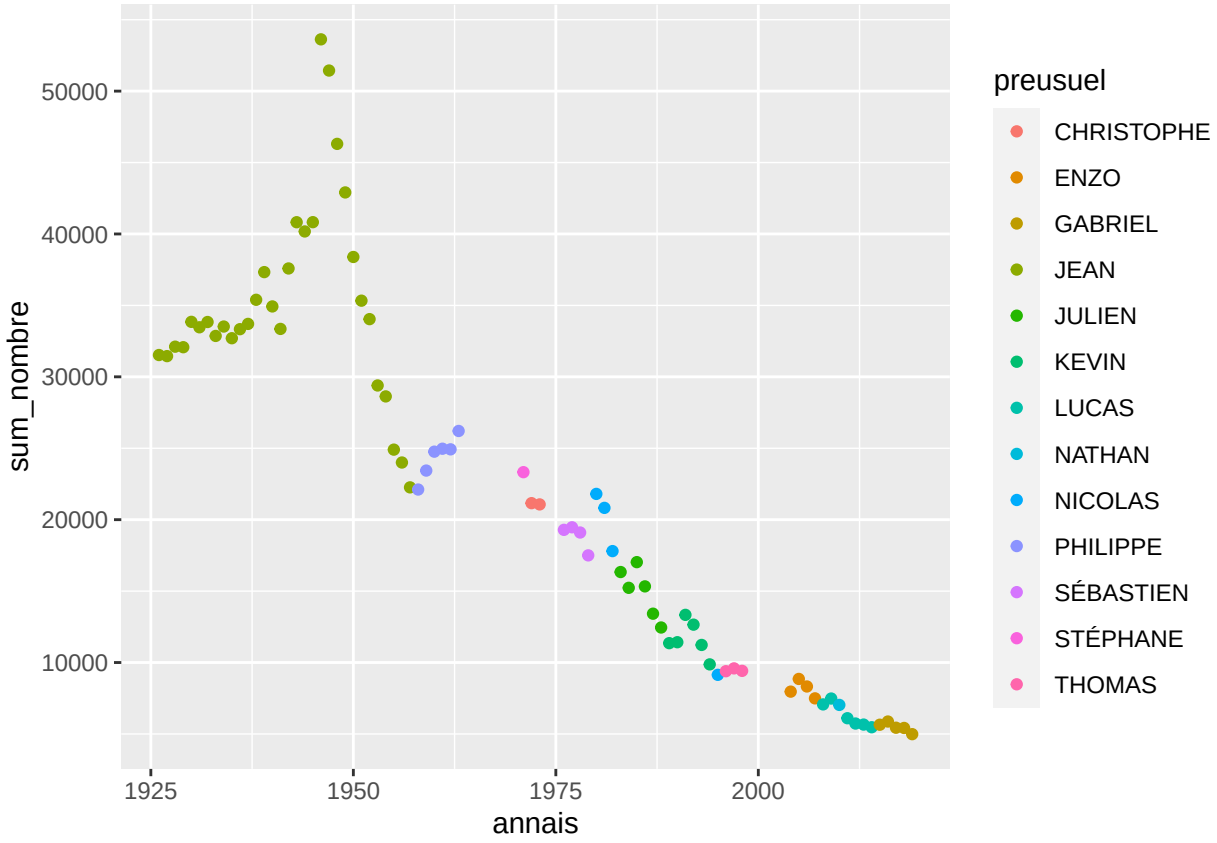
```
## `summarise()` regrouping output by 'annais', 'preusuel', 'sexe' (override with `.groups` argument)
```

```
  #filter(), n == max(sum_nombre))

HighestOccurenceWomen <- filter(HighestOccurence, sexe==2)
HighestOccurenceMen <- filter(HighestOccurence, sexe==1)
```

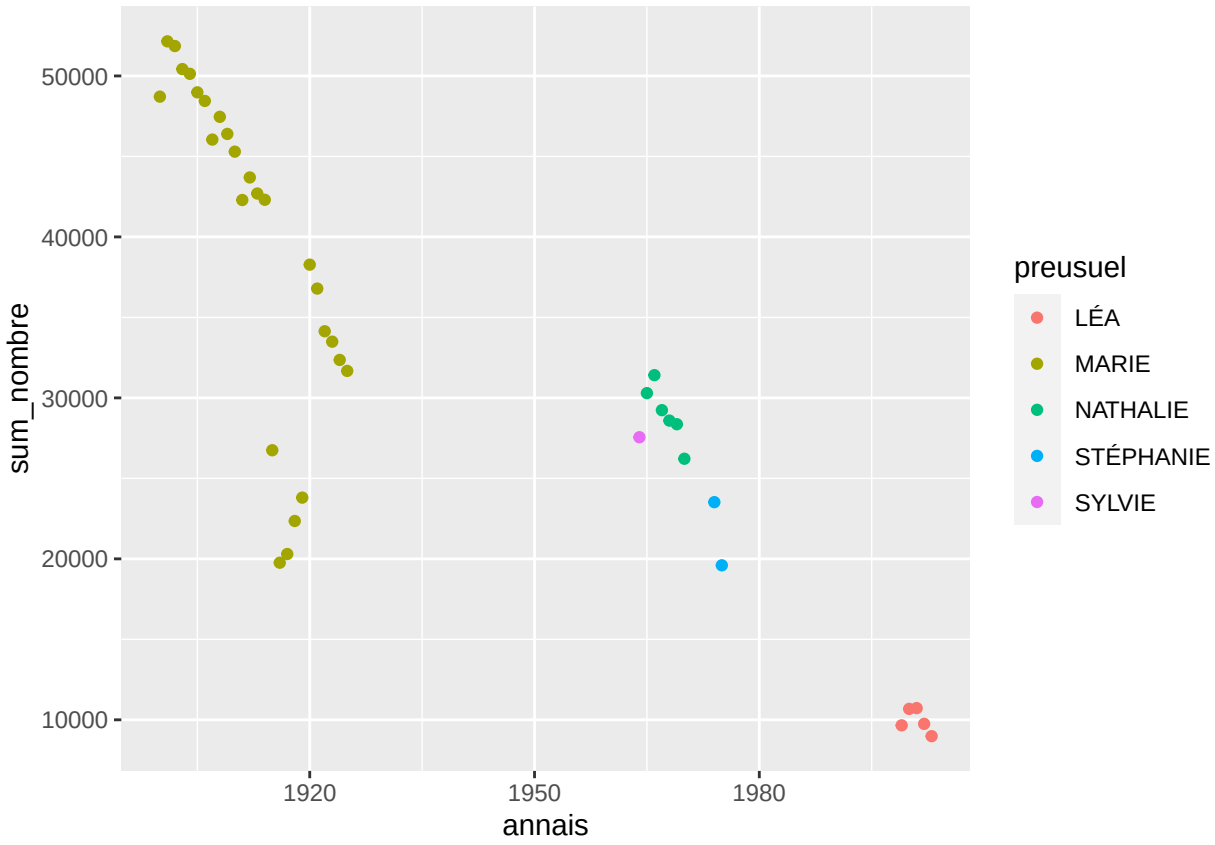
Frequency of most frequent male names by year

```
ggplot(data = HighestOccurenceMen, aes(x=annais, y=sum_nombre, color = preusuel))+geom_point()
```



Frequency of most frequent female names by year

```
ggplot(data = HighestOccurenceWomen, aes(x=annais, y=sum_nombre, color = preusuel))+geom_point()
```


50000
40000
30000
20000
10000
sum_nombre
1920
1950
1980
annais
preusuel
LÉA
MARIE
NATHALIE
STÉPHANIE
SYLVIE