French Names exercise

Carlos Vargas

October, 2021

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

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

```
library(ggplot2)

version
```

```
##                _
## platform       x86_64-conda_cos6-linux-gnu
## arch           x86_64
## os             linux-gnu
## system         x86_64, linux-gnu
## status
## major          3
## minor          6.1
## year           2019
## month          07
## day            05
## svn rev        76782
## language       R
## version.string R version 3.6.1 (2019-07-05)
## nickname       Action of the Toes
```

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)
```

All of these following questions may need a preliminary analysis of the data, feel free to present answers and justifications in your own order and structure your report as it should be for a scientific report.

1. Choose a firstname and analyse its frequency along time. Compare several firstnames frequency
2. Establish, by gender, the most given firstname by year.
3. Make a short synthesis
4. Advanced (not mandatory) : is the firstname correlated with the localization (department) ? What could be a method to analyze such a correlation.

The report should be a pdf knitted from a notebook (around 3 pages including figures), the notebook and the report should be delivered.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Results of Analysis

The first thing I notice is that the provided file name (**dpt2019.csv**) for the .csv file is wrong, this file does not exist after unzipping. I check the files in my current directory.

```
list.files()
```

```
## [1] "Correlation_Causality_exo.ipynb" "dpt2020_txt.zip"
## [3] "dpt2020.csv"                     "exo5_en.ipynb"
## [5] "names_exercise.pdf"              "names_exercise.Rmd"
```

Build the Dataframe from correct 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.
```

Overview of the data

We have 5 attributes in the dataset, 3 of them have `character` type, the other 2 are `doubles`.

```
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
```

```
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
```

At first glance, it seems that all the names are the same ('_PRENOMS_RARES'), so let's list some of the distinct values for the column `preusuel` and its count

```
distinc_names = unique(FirstNames['preusuel'])
head(distinc_names)
```

```
## # A tibble: 6 x 1
##   preusuel
##   <chr>
## 1 _PRENOMS_RARES
## 2 A
## 3 AADAM
## 4 AADEL
## 5 AADIL
## 6 AAHIL
```

```
count(distinc_names)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1 35011
```

We actually have quite a lot of different names (35011).

Lets check the range of values for the year of birth.

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

```
## # A tibble: 122 x 1
##    annais
##    <chr>
##  1 1900
##  2 1901
##  3 1902
##  4 1903
##  5 1904
##  6 1905
##  7 1906
##  8 1907
##  9 1908
## 10 1909
## # ... with 112 more rows
```

We have data from 1900 to 2020. We also have registers with no year ('XXXX').

Now, lets take a (not that random) name, say 'CARLOS', and check its frequency over the time. First, the whole dataset is filtered to get only the observations matching the given name, then this subset is grouped by year of birth and finally that result is used to create a new dataset including a new the attribute, the frequency.

```
freq_carlos <- FirstNames %>% filter(preusuel == 'CARLOS') %>% group_by(annais, preusuel) %>% summarize
```

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

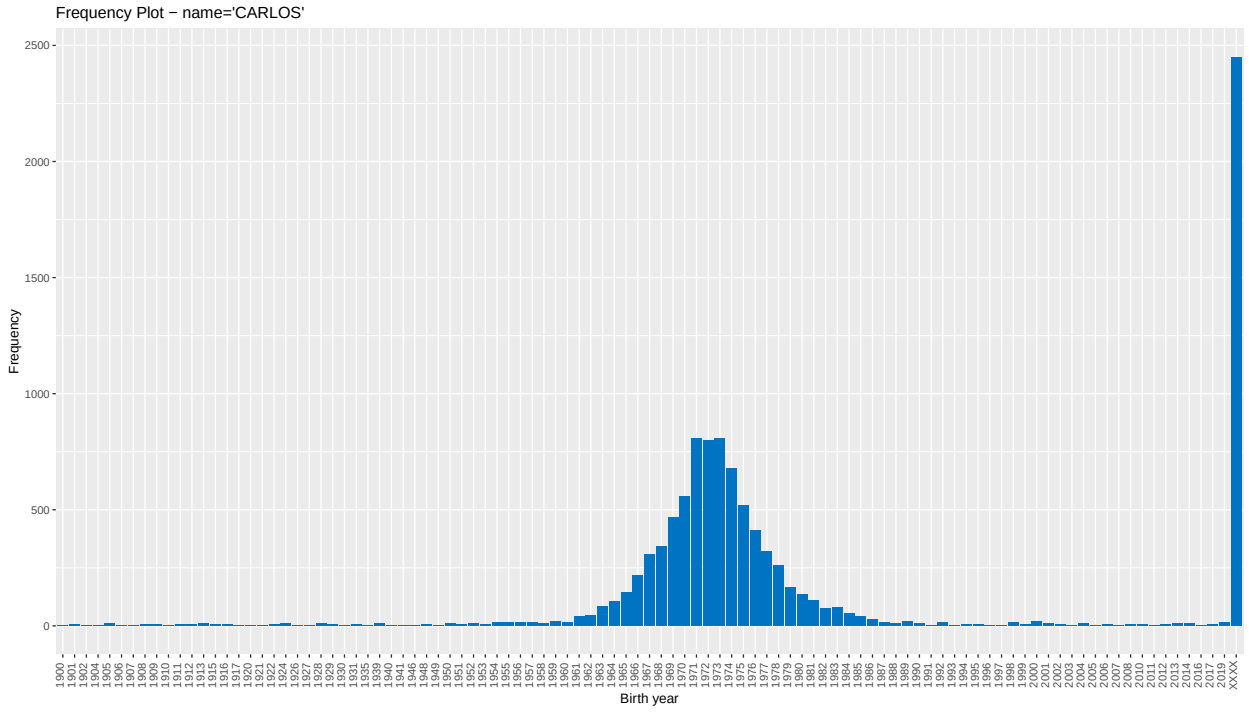
```
freq_carlos
```

```
## # A tibble: 101 x 3
## # Groups:   annais [101]
##    annais preusuel frequency
##    <chr>  <chr>        <dbl>
##  1 1900   CARLOS           3
##  2 1901   CARLOS           6
##  3 1902   CARLOS           4
##  4 1904   CARLOS           3
##  5 1905   CARLOS          10
##  6 1906   CARLOS           3
##  7 1907   CARLOS           4
```

```
##  8 1908   CARLOS           8
##  9 1909   CARLOS           7
## 10 1910   CARLOS           4
## # ... with 91 more rows
```

Plot the frequency of name 'CARLOS'

```
library(ggplot2)
ggplot(freq_carlos, aes(x = annais, y = frequency)) +
  geom_bar(fill = "#0073C2FF", stat = "identity") +
  #geom_text(aes(label = Frequency), vjust = -0.3, angle = 90) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.1, hjust=.5)) +
  ggtitle("Frequency Plot - name='CARLOS'") +
  xlab("Birth year") + ylab("Frequency")
```



By regardind the plot, the data describes a gaussian distribution in the frequencies. Between years 1961 to 1986 we observe a considerably increase in people named 'CARLOS', this could have many explanations related with relevant political, economical or social events happened in that period of time. We also observe a large number of people (~2500) named Carlos for which there is no record of the year of birth. We have this in mind for next analysis.

Now lets make the same analysis but for a different name, say CHARLES.

```
freq_charles <- FirstNames %>% filter(preusuel == 'CHARLES') %>% group_by(annais, preusuel) %>% summari
```

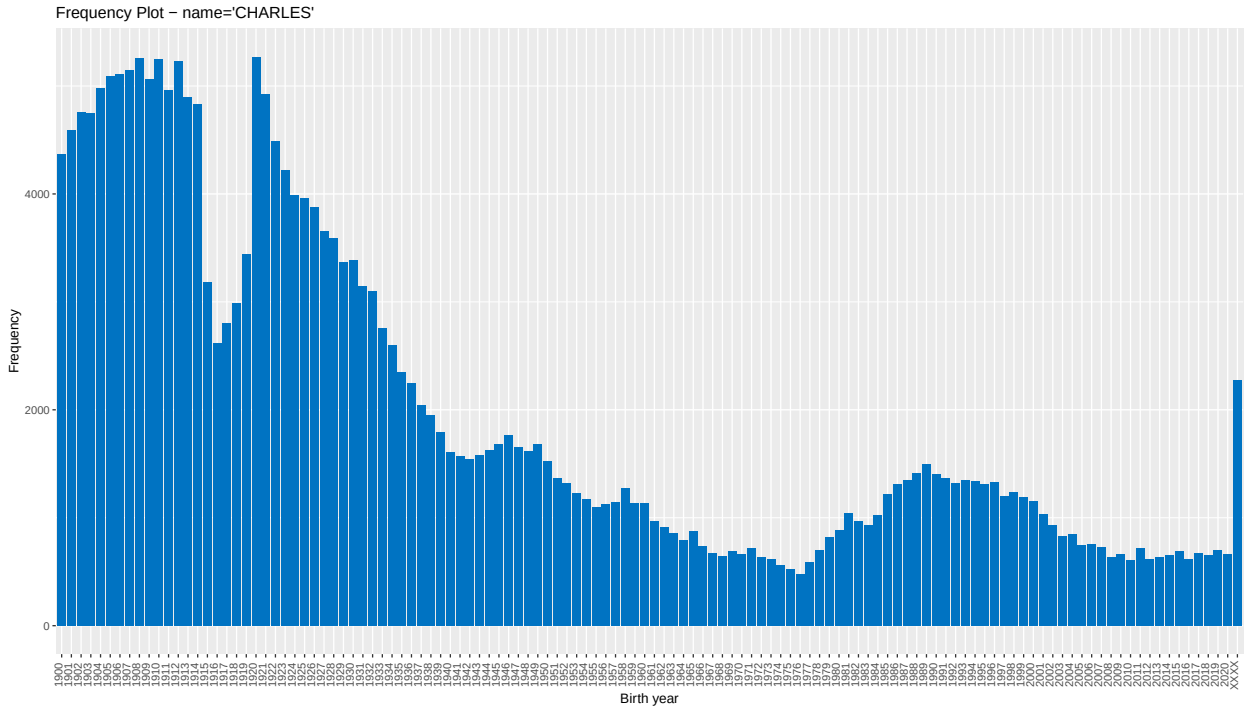
```
## `summarise()` has grouped output by 'annais'. You can override using the `.groups` argument.
```

```
freq_charles
```

```
## # A tibble: 122 x 3
## # Groups:   annais [122]
##    annais preusuel frequency
##    <chr>  <chr>        <dbl>
##  1 1900   CHARLES       4364
##  2 1901   CHARLES       4588
##  3 1902   CHARLES       4756
##  4 1903   CHARLES       4744
##  5 1904   CHARLES       4977
##  6 1905   CHARLES       5088
##  7 1906   CHARLES       5110
##  8 1907   CHARLES       5151
##  9 1908   CHARLES       5258
## 10 1909   CHARLES       5067
## # ... with 112 more rows
```

Plot the frequency of name 'CHARLES'

```
library(ggplot2)
ggplot(freq_charles, aes(x = annais, y = frequency)) +
  geom_bar(fill = "#0073C2FF", stat = "identity") +
  #geom_text(aes(label = Frequency), vjust = -0.3, angle = 90) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.1, hjust=.5)) +
  ggtitle("Frequency Plot - name='CHARLES'") +
  xlab("Birth year") + ylab("Frequency")
```



From the second plot, we observe that the frequency distribution for the name 'CHARLES' is a bit more uniform. Also, there are more indivisuals within this group compared with the previous one, this is expected because 'CHARLES' is a pretty common name in France and western Europe. We still see a considerably number of observations with no birth of year (~2000).

Most given firstnames

We will establish, by gender, the most given firstname by year.

First, we group the data by name, year of birth and sex. We also compute the frequency for the names.

```
all_names <- FirstNames %>% group_by(preusuel, annais, sexe) %>% summarize(frequency=sum(nombre))
```

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

Then we remove the observations with no year of birth, as well as those with rare names.

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

```
## Selecting by frequency
```

We filter the male names and plot the most used names per year.

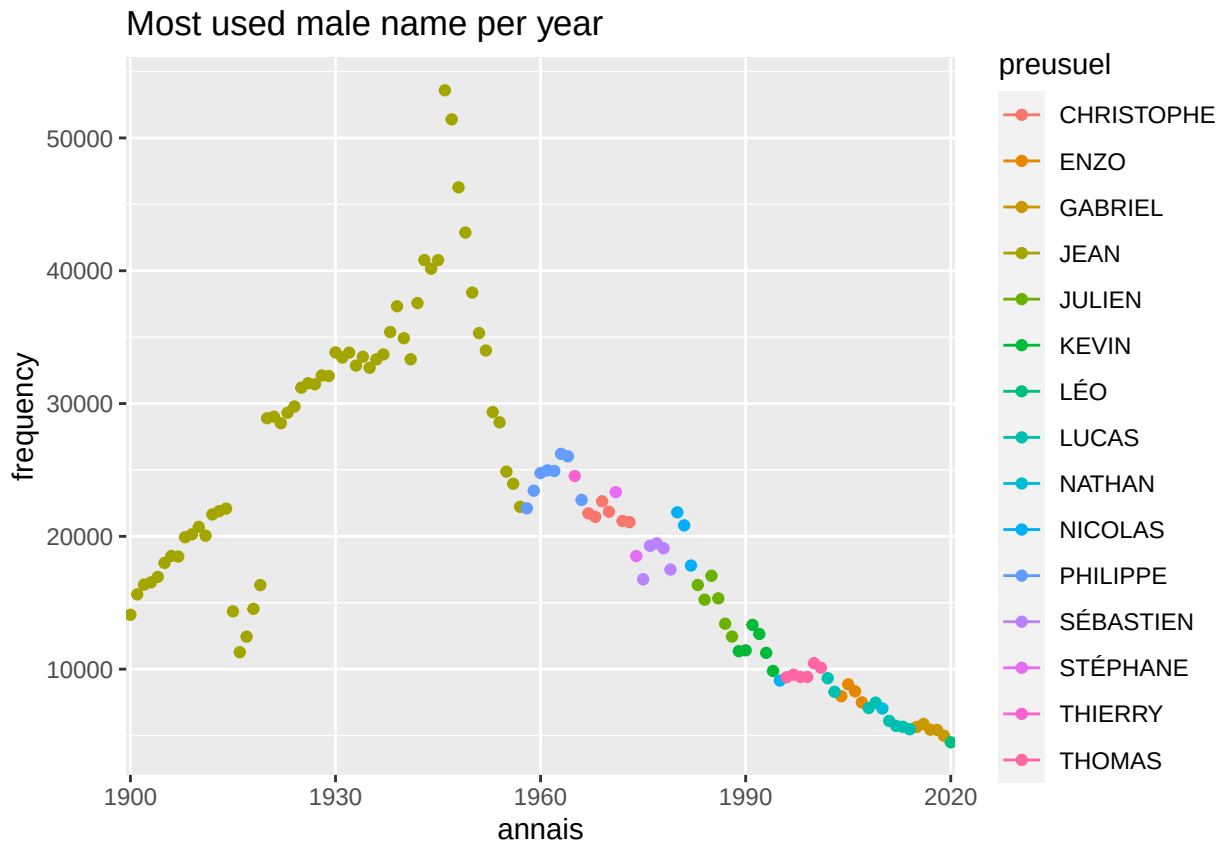
```
labels <- seq(1900, 2020, length.out=5)

most_used_man <- all_names %>% filter(sexe == 1)

plot_man <- ggplot(data=most_used_man, mapping = aes(x = annais, y = frequency, color=preusuel)) +
  scale_x_discrete(breaks = labels, labels=as.character(labels)) + geom_point() + geom_line() +
  ggtitle("Most used male name per year")

plot_man
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



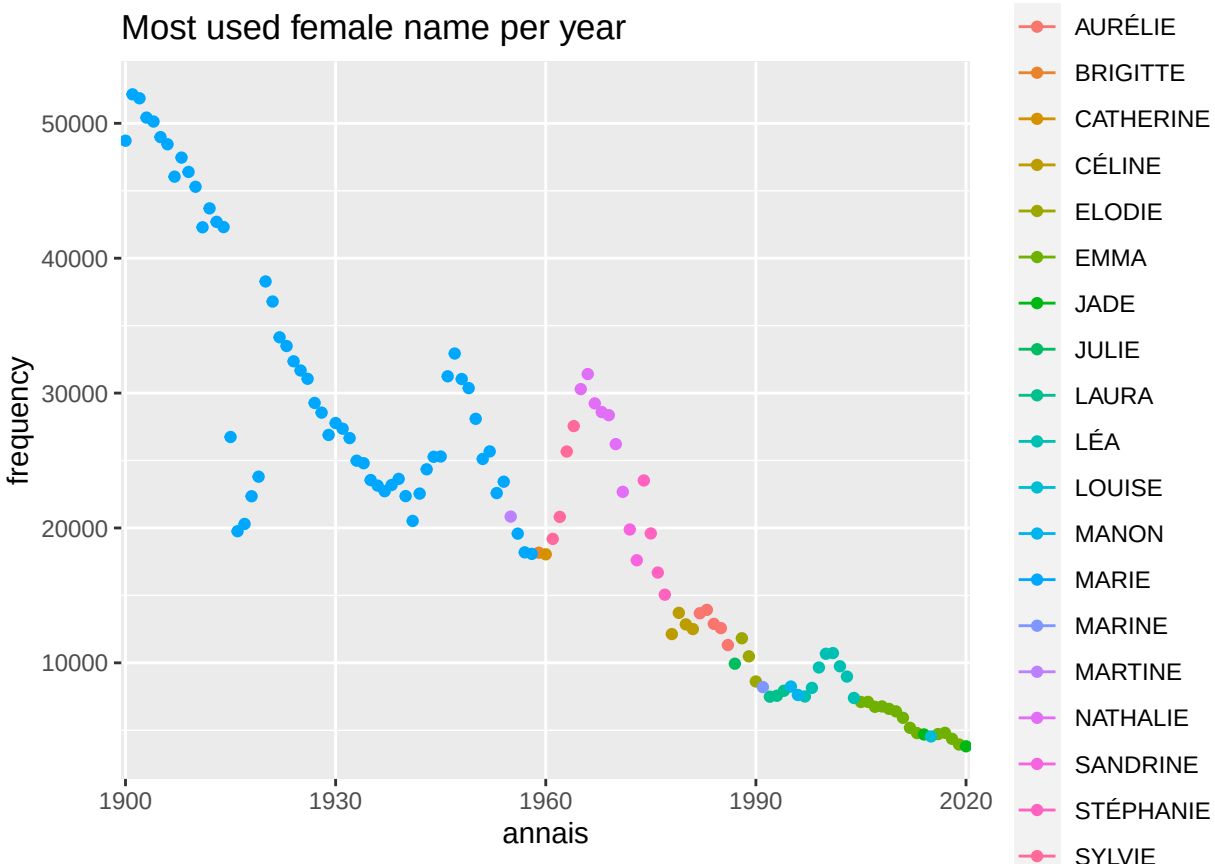
We use now the most used woman names.

```
most_used_woman <- all_names %>% filter(sexe == 2)

plot_woman <- ggplot(data=most_used_woman, mapping = aes(x = annais, y = frequency, color=preusuel)) +
  scale_x_discrete(breaks = labels, labels=as.character(labels)) + geom_point() + geom_line() +
  ggtitle("Most used female name per year")

plot_woman
```

```
## geom_path: Each group consists of only one observation. Do you need to adjust
## the group aesthetic?
```



Comments

From the las plots we can observe that for males, there are som names (e.g. Jean) that tend to be quite popular for about 50 years (1900-1950) which is weird. After 1960, other names remains popular for a couple of consecutive years but the general picture is more heterogeneous, decreasing the frequency of popular names over the years. In the case of woman names, the pattern is the same. Both plots show that in recent years the diversification of names is much more evident.