French given names exercise

Benjamin Cathelineau

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-w64-mingw32
## arch           x86_64
## os             mingw32
## system         x86_64, mingw32
## status
## major          4
## minor          1.1
## year           2021
## month          08
## day            10
##  [ getOption("max.print") est atteint -- 4 lignes omises ]
```

Download Raw Data from the website

file downloaded from https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2020_csv.zip

Build the Dataframe from file

I had to change the name of the file because it wasn't the correct one.

```
FirstNames <- readr::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.
```

```
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
```

Translation in english of variables names:

sexe -> gender

preusuel (prénom usuel) -> Firstname

annais (année de naissance) -> Birth year

dpt (département) -> department (administrative area unit)

nombre -> number

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 first names frequency *First We can find all the different names using the following command. This will group all the entries by* ***preusel***

```
table(FirstNames$preusuel)
```

```
##
## _PRENOMS_RARES              A          AADAM          AADEL          AADIL
##          22037              1              1              1              3
##          AAHIL         AAKASH        AALEYAH          AALIA         AALIYA
##              2              1              1              1              2
##  [ reached getOption("max.print") -- omitted 35000 entries ]
```

Then, using the following ***dplyr*** *pipeline we can see the one that occurs the most often*

```
# With this command, we can see which name occurs more often
library(dplyr)
FirstNames %>% count(preusuel) %>% arrange(desc(n))
```

```
## # A tibble: 35,011 x 2
##    preusuel           n
##    <chr>          <int>
##  1 _PRENOMS_RARES 22037
##  2 CAMILLE        13822
##  3 MARIE          13302
##  4 PIERRE         11390
##  5 PAUL           10713
##  6 JEAN           10696
##  7 CLAUDE         10573
##  8 LOUIS          10126
##  9 FRANÇOIS        9977
## 10 ANTOINE         9841
## # ... with 35,001 more rows
```

We just have to divide each "count" by the total in order to find the ***frequency*** *for that we use* ***mutate***

```
# To get the frequency
library(dplyr)
frequencies=FirstNames %>% count(preusuel) %>% arrange(desc(n)) %>% mutate(frequency=n/nrow(FirstNames)
frequencies
```

```
## # A tibble: 35,011 x 3
##    preusuel           n frequency
##    <chr>          <int>     <dbl>
##  1 _PRENOMS_RARES 22037   0.00591
##  2 CAMILLE        13822   0.00371
##  3 MARIE          13302   0.00357
##  4 PIERRE         11390   0.00306
##  5 PAUL           10713   0.00287
##  6 JEAN           10696   0.00287
##  7 CLAUDE         10573   0.00284
##  8 LOUIS          10126   0.00272
##  9 FRANÇOIS        9977   0.00268
## 10 ANTOINE         9841   0.00264
## # ... with 35,001 more rows
```

We can even check our result with ***sum*** *which should be equal to 1*

```
# To get the frequency
library(dplyr)
sum( frequencies$frequency)
```

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

2. Establish, by gender, the most given firstname by year. *We use* ***group by*** *for that, and* ***filter*** *to only keep the maximum*

```
library(dplyr)
FirstNames %>% group_by(sexe,annais) %>% filter(nombre==max(nombre))
```

```
## # A tibble: 245 x 5
## # Groups:   sexe, annais [244]
##     sexe preusuel       annais dpt   nombre
##    <dbl> <chr>          <chr>  <chr>  <dbl>
##  1     1 _PRENOMS_RARES 1982   75       997
##  2     1 _PRENOMS_RARES 1983   75      1069
##  3     1 _PRENOMS_RARES 1984   75      1087
##  4     1 _PRENOMS_RARES 1985   75      1109
##  5     1 _PRENOMS_RARES 1986   75      1117
##  6     1 _PRENOMS_RARES 1987   75       984
##  7     1 _PRENOMS_RARES 1988   75      1130
##  8     1 _PRENOMS_RARES 1989   75      1145
##  9     1 _PRENOMS_RARES 1990   75      1177
## 10     1 _PRENOMS_RARES 1991   75      1158
## # ... with 235 more rows
```

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.