FirstNames analysis

Maša Hadži-Nikolić

December 2023

Description

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.
`https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2021_csv.zip`, 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 dpt2020txt.zip (to get the dpt2020.csv). Read in R with this code. Note that you might need to install the 'readr' package with the appropriate command.

Download raw data

```
file = "dpt2021_csv.zip"
if(!file.exists(file)){
  download.file("https://www.insee.fr/fr/statistiques/fichier/2540004/dpt2021_csv.zip",
   destfile=file)
}
unzip(file)
```

Import the libraries and build the dataset

```
library(tidyverse)
library(dplyr)
library(ggplot2)

FirstNames <- read_delim("dpt2021.csv",delim=";")
head(FirstNames)
tail(FirstNames)
```

sexe	preusuel	annais	dpt	nombre
<dbl>	<chr>	<chr>	<chr>	<dbl>
1	_PRENOMS_RARES	1900	02	7
1	_PRENOMS_RARES	1900	04	9
1	_PRENOMS_RARES	1900	05	8
1	_PRENOMS_RARES	1900	06	23
1	_PRENOMS_RARES	1900	07	9
1	_PRENOMS_RARES	1900	08	4

A tibble: 6 × 5

sexe	preusuel	annais	dpt	nombre
<dbl>	<chr>	<chr>	<chr>	<dbl>
2	ZYA	2018	59	3
2	ZYA	2021	35	5
2	ZYA	XXXX	XX	278
2	ZYNA	2013	93	3
2	ZYNA	XXXX	XX	68
2	ZYNEB	XXXX	XX	125

Cleaning the data

For better understanding, we will change the names of the columns from french to english.

```
colnames(FirstNames) <- c("Sex", "Firstname", "Year", "Departement", "Number")
head(FirstNames)
```

Sex	Firstname	Year	Departement	Number
<dbl>	<chr>	<chr>	<chr>	<dbl>
1	_PRENOMS_RARES	1900	02	7
1	_PRENOMS_RARES	1900	04	9
1	_PRENOMS_RARES	1900	05	8
1	_PRENOMS_RARES	1900	06	23
1	_PRENOMS_RARES	1900	07	9
1	_PRENOMS_RARES	1900	08	4

From the subset shown above, in certain rows we can notice values XX and XXXX for the attributes of age and department.

```
   count_cols <- sum(FirstNames$Year == "XXXX" | FirstNames$Departement == "XX")
print(count_cols)
```

38479

Also, as the number of such rows is not too big compared to our dataset and in order to avoid difficulties later in the analysis, we will drop those rows.

```
   FirstNames <- FirstNames %>%
  filter(Year != "XXXX")
FirstNames <- FirstNames %>%
  filter(Departement != "XX")

tail(FirstNames)
```

Sex	Firstname	Year	Departement	Number
<dbl>	<chr>	<chr>	<chr>	<dbl>
2	ZYA	2013	44	4
2	ZYA	2013	59	3
2	ZYA	2017	974	3
2	ZYA	2018	59	3
2	ZYA	2021	35	5
2	ZYNA	2013	93	3

Since the Year column is charachter type, we will convert it to the numeric type so we can analyze data easier later.

```
   FirstNames$Year <- as.numeric(FirstNames$Year)
head(FirstNames)
```

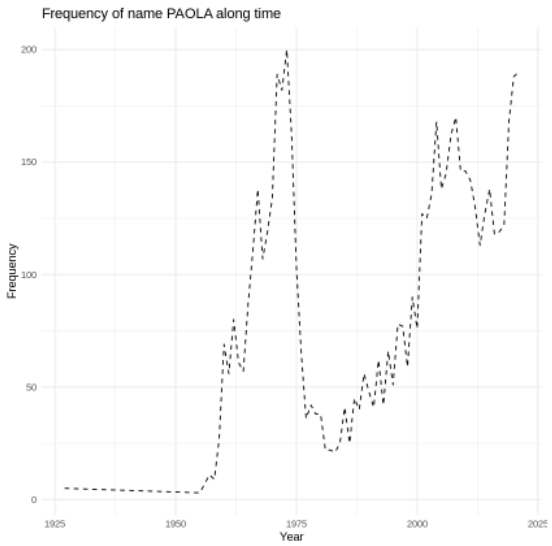
Sex	Firstname	Year	Departement	Number
<dbl>	<chr>	<dbl>	<chr>	<dbl>
1	_PRENOMS_RARES	1900	02	7
1	_PRENOMS_RARES	1900	04	9
1	_PRENOMS_RARES	1900	05	8
1	_PRENOMS_RARES	1900	06	23
1	_PRENOMS_RARES	1900	07	9
1	_PRENOMS_RARES	1900	08	4

1. Choose a firstname and analyse its frequency along time. Compare several firstnames frequency

We will analyze the frequency along time of the firstname "PAOLA".

```
paola <- subset(FirstNames, Firstname == 'PAOLA') %>%
  select(Year, Number) %>%
  group_by(Year) %>%
  summarise(sum_number = sum(Number, na.rm = TRUE))

plot <- ggplot(data = paola, aes(x = Year, y = sum_number)) +
  geom_line(linetype = "dashed", color = "black") +
  labs(title = "Frequency of name PAOLA along time",
       x = "Year",
       y = "Frequency") +
  theme_minimal()
print(plot)
```



From the graph above, we can see that until 60's, name PAOLA was not common at all. After that, the name increases rapidly and reaches its peak around the year of 1975. After the peak, there is a rapid decrease which is followed by gradual growth that starts in the begginging of the 80's.

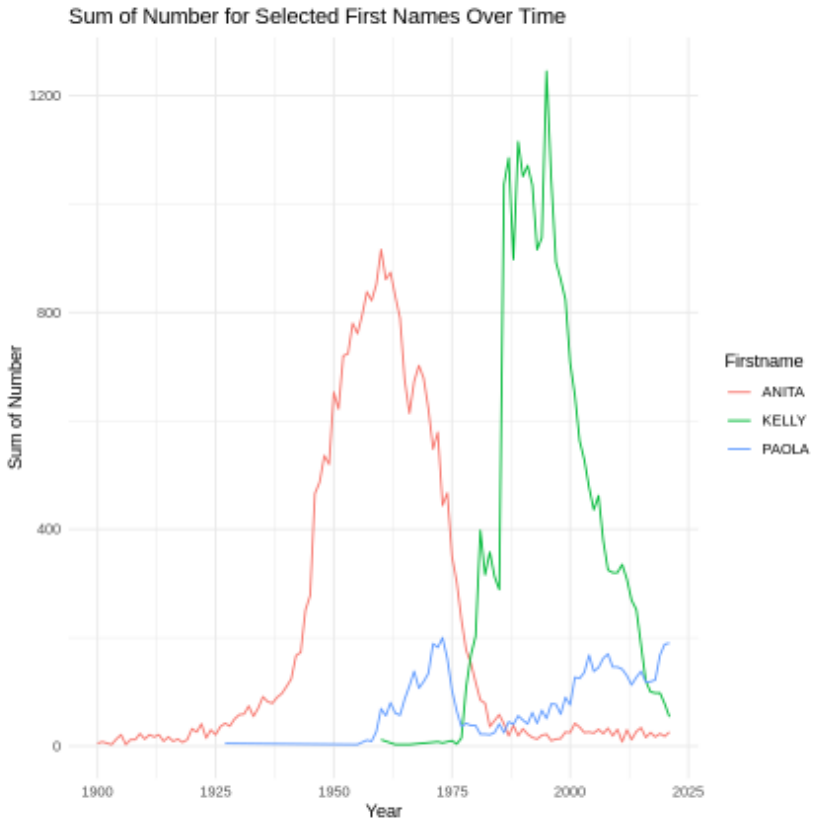
Now we will compare frequency of the following Firstnames : "KELLY", "PAOLA", and "ANITA".

```
names_to_compare <- c("KELLY", "PAOLA", "ANITA")

selected_name_data <- FirstNames %>%
  filter(Firstname %in% names_to_compare) %>%
  group_by(Firstname, Year) %>%
  summarise(sum_number = sum(Number, na.rm = TRUE),.groups = 'drop')

plot <- ggplot(selected_name_data, aes(x = Year, y = sum_number, color = Firstname)) +
  geom_line() +
  labs(title = "Sum of Number for Selected First Names Over Time",
       x = "Year",
       y = "Sum of Number") +
  theme_minimal()

print(plot)
```



From the graph above, we can conclude that both ANITA and KELLY were popular in some period of time and their distrubution remainds of a normal distribution. ANITA reaches its peak around 60's while KELLY around the year of 2000. On the other side, PAOLA was never quite common compared to the others but we can see that its frequency slightly increases in 60's and again in 2000's.

2. Establish by gender the most given firstname by year

First, we will drop rows with "PRENOMS RARES" because it doesn't give us any adequate information about the frequency of the names and then we will count the total number of each name in every year. Then we will determinate the most common name in each year and the results can be shown in the tabels below.

```
male_names_by_year <- FirstNames %>%
  filter(Firstname != "_PRENOMS_RARES") %>%
  filter(Sex == 1) %>%
  group_by(Year, Firstname) %>%
  summarise(sum_number = sum(Number), .groups = 'drop')

most_given_male <- male_names_by_year %>%
  group_by(Year) %>%
  slice(which.max(sum_number))

head(most_given_male)
tail(most_given_male)
```

Year	Firstname	sum_number
<dbl>	<chr>	<dbl>
1900	JEAN	14097
1901	JEAN	15634
1902	JEAN	16364
1903	JEAN	16535
1904	JEAN	16944
1905	JEAN	17998

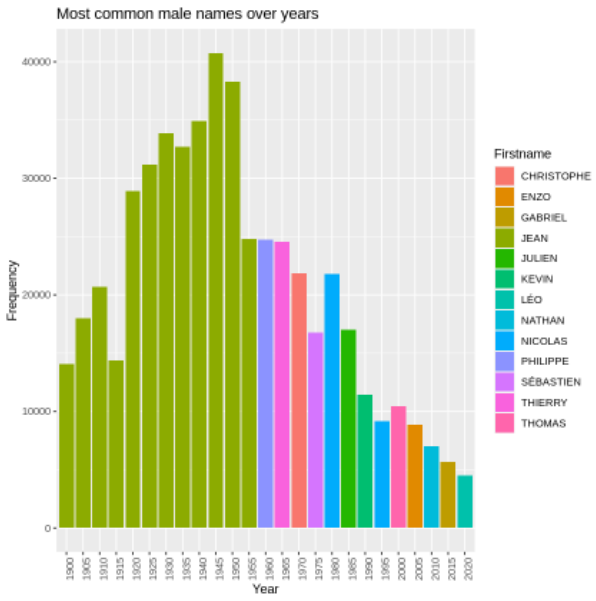
A grouped_df: 6 × 3

Year	Firstname	sum_number
<dbl>	<chr>	<dbl>
2016	GABRIEL	5875
2017	GABRIEL	5442
2018	GABRIEL	5422
2019	GABRIEL	4987
2020	LÉO	4494
2021	GABRIEL	4972

The graphical representation below can describe it more closely.

```
   most_given_male <- most_given_male %>%
  filter(Year %% 5 == 0)

ggplot(most_given_male, aes(x = factor(Year), y = sum_number, fill = Firstname)) +
  geom_col(position = "dodge") +
  labs(title = "Most common male names over years",
       x = "Year",
       y = "Frequency") +
  theme(axis.text.x.bottom = element_text(angle = 90))
```



We will do the same for female names.

```
   female_names_by_year <- FirstNames %>%
  filter(Firstname != "_PRENOMS_RARES") %>%
  filter(Sex == 2 ) %>%
  group_by(Year, Firstname) %>%
  summarise(sum_number = sum(Number), .groups = 'drop')

most_given_female <- female_names_by_year %>%
  group_by(Year) %>%
  slice(which.max(sum_number))

head(most_given_female)
tail(most_given_female)

most_given_female <- most_given_female %>%
  filter(Year %% 5 == 0)

ggplot(most_given_female, aes(x = factor(Year), y = sum_number, fill = Firstname)) +
  geom_col(position = "dodge") +
  labs(title = "Most common female names over years",
       x = "Year",
       y = "Frequency") +
  theme(axis.text.x.bottom = element_text(angle = 90))
```

Year	Firstname	sum_number
<dbl>	<chr>	<dbl>
1900	MARIE	48713
1901	MARIE	52150
1902	MARIE	51857
1903	MARIE	50424
1904	MARIE	50131
1905	MARIE	48981

A grouped_df: 6 × 3

Year	Firstname	sum_number
<dbl>	<chr>	<dbl>
2016	EMMA	4723
2017	EMMA	4815
2018	EMMA	4372
2019	EMMA	3953
2020	JADE	3816
2021	JADE	3798