French given names per year per department

Lucas Mello Schnorr, Jean-Marc Vincent

October, 2022

Build the Dataframe from file

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
first_names <- read_delim("dpt2021.csv",delim=";", show_col_types = FALSE)
```

To make this analysis easier, I'll remove useless values such as "_PRENOMS_RARES" from the dataset. In addition, The year field will be converted to a proper date format, which should make visualization clearer with GGPLOT.

Drop useless values and fix formatting in the dataframe

```
first_names = subset(first_names, preusuel != '_PRENOMS_RARES' & annais != 'XXXX')
first_names <- first_names %>% rename(name = preusuel, year = annais, quantity = nombre)
first_names$year <- as.POSIXct(first_names$year, format="%Y")
```

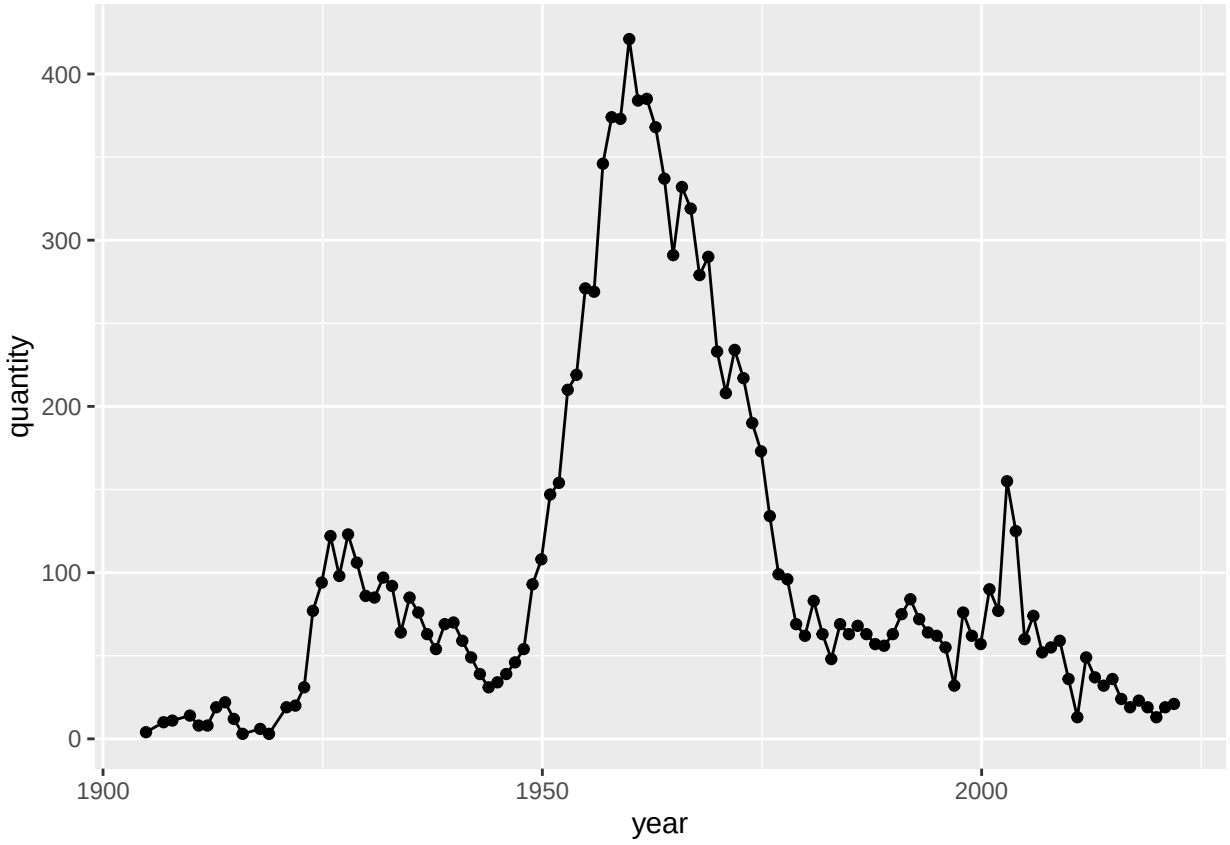
Analysis

1. Choose a firstname and analyse its frequency along time

I'm considering this just a warm-up for the next sections. Let's analyse the popularity of the name Mario over time:

```
mario_popularity = first_names %>%
                   filter(name == "MARIO") %>%
                   group_by(name, year) %>%
                   summarise_at(vars(quantity), list(quantity = sum))

ggplot(data=mario_popularity, aes(x=year, y=quantity, group=1)) +
  geom_line() +
  geom_point()
```



Even at its most popular, the first name Mario didn't really have an impressive frequency. In section 2, we will see that the most popular names have much higher frequencies.

2. Most popular first names by gender

Male

```
popular_male_names = first_names %>%
                   filter(sexe == 1) %>%
                   group_by(year, name) %>%
                   summarise(popularity = sum(quantity)) %>%
                   filter(popularity == max(popularity))
```

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

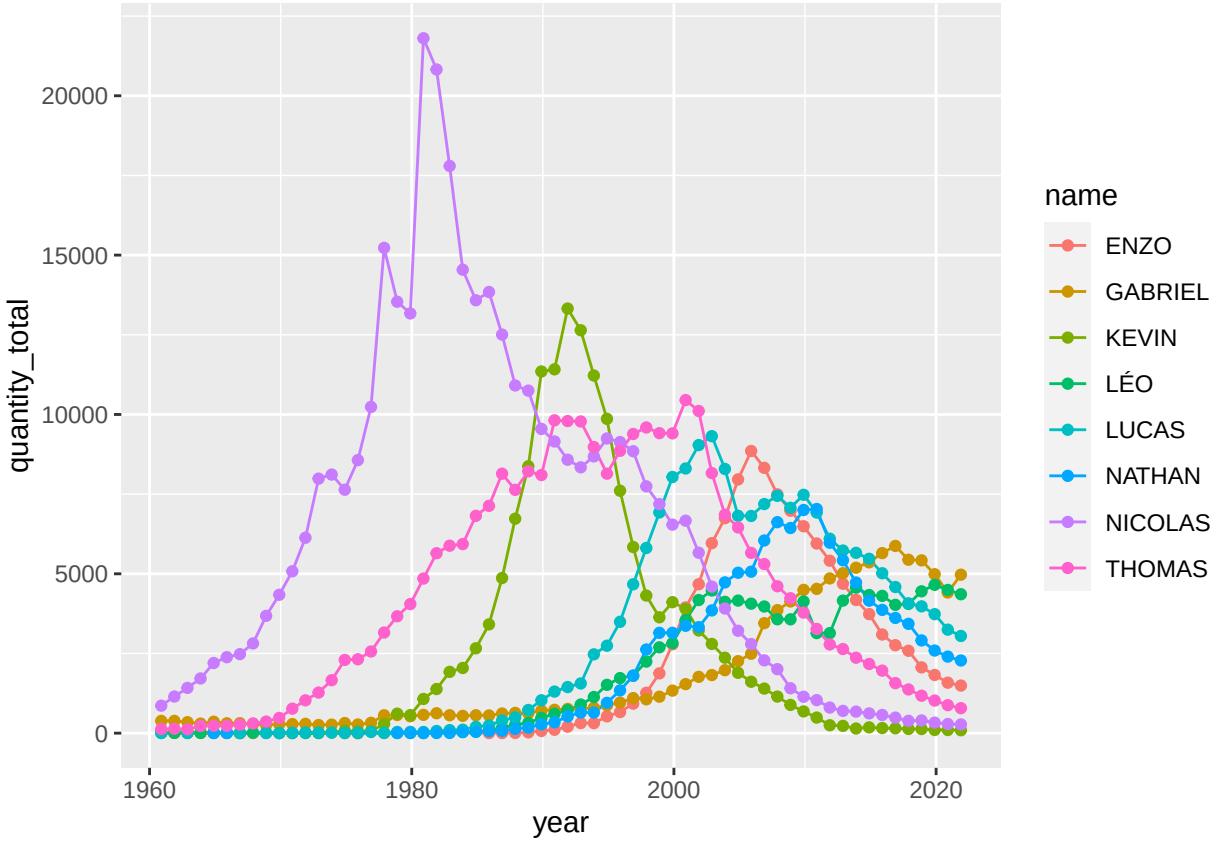
Now let's see the progression of those names over time, starting from 1960. What we want to analyze here is whether those recent popular names had already been popular sometime in the past and also whether their popularity is short-lived or not.

```
most_popular_last_30y = tail(popular_male_names$name, 30)

top30_over_time = first_names %>%
                  filter(name %in% most_popular_last_30y & year > as.POSIXct("1960-01-01")) %>%
                  group_by(year, name) %>%
                  summarise(quantity_total = sum(quantity))
```

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

```
ggplot(data=top30_over_time, aes(x=year, y=quantity_total, color = name, group=name)) +
  geom_line() +
  geom_point()
```



It seems to be consistent that the most popular names tend to come in waves. Not only that, we cab also note that, over the last 30 years, only 8 male names have taken the top spot. Let's verify if the same phenomenon occurs for female names:

Female

```
popular_female_names = first_names %>%
                  filter(sexe == 2) %>%
                  group_by(year, name) %>%
                  summarise(popularity = sum(quantity)) %>%
                  filter(popularity == max(popularity))
```

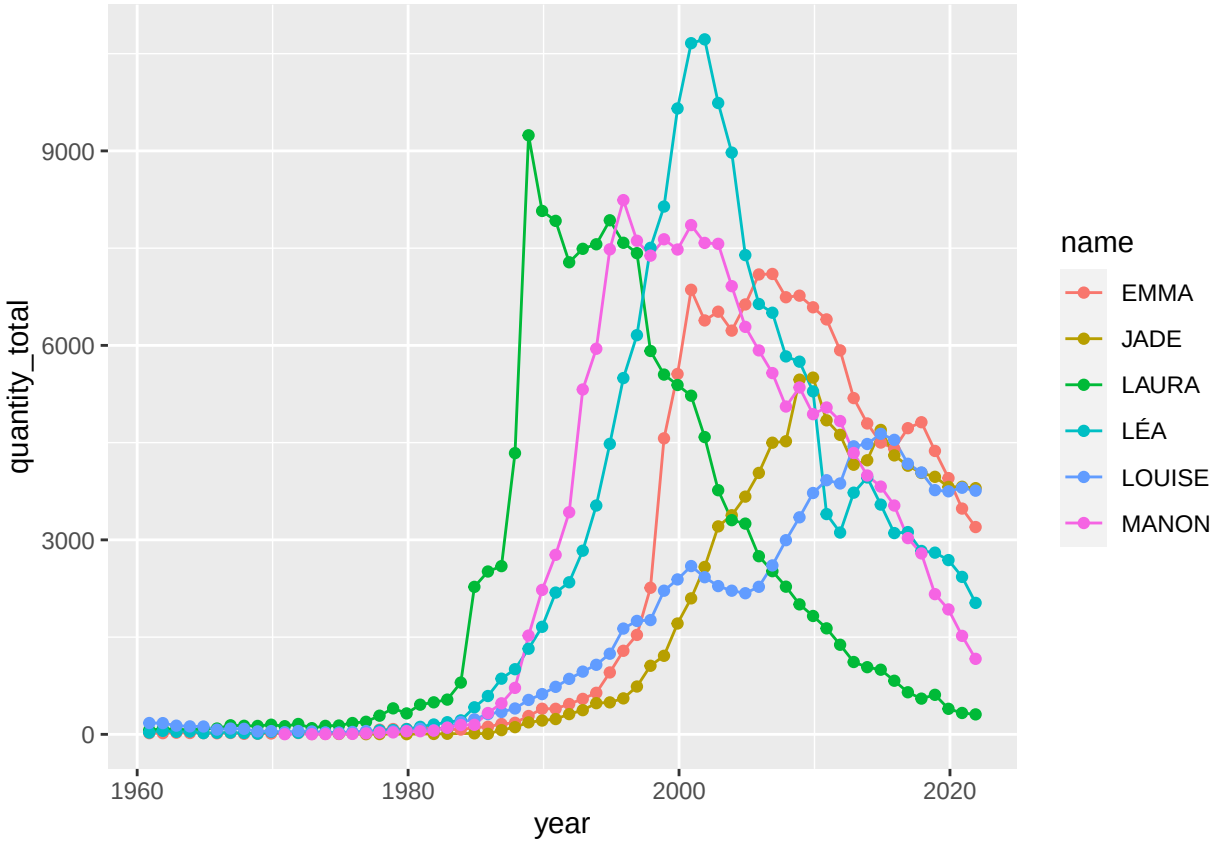
```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
most_popular_last_30y = tail(popular_female_names$name, 30)

top30_over_time = first_names %>%
                  filter(name %in% most_popular_last_30y & year > as.POSIXct("1960-01-01")) %>%
                  group_by(year, name) %>%
                  summarise(quantity_total = sum(quantity))
```

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

```
ggplot(data=top30_over_time, aes(x=year, y=quantity_total, color = name, group=name)) +
  geom_line() +
  geom_point()
```



Here we see the same pattern; even fewer first names have taken the top spot for female names. It can be noted, however, that the female names that have dominated over the last 30 years seem to have risen to popularity more recently than their male counterparts.

3. Optional : Which department has a larger variety of names along time ? Is there some sort of geographical correlation with the data?

First, let's prepare the data. We want to count the amount of distinct numbers in each department. We also want to have this count over time so we can plot it.

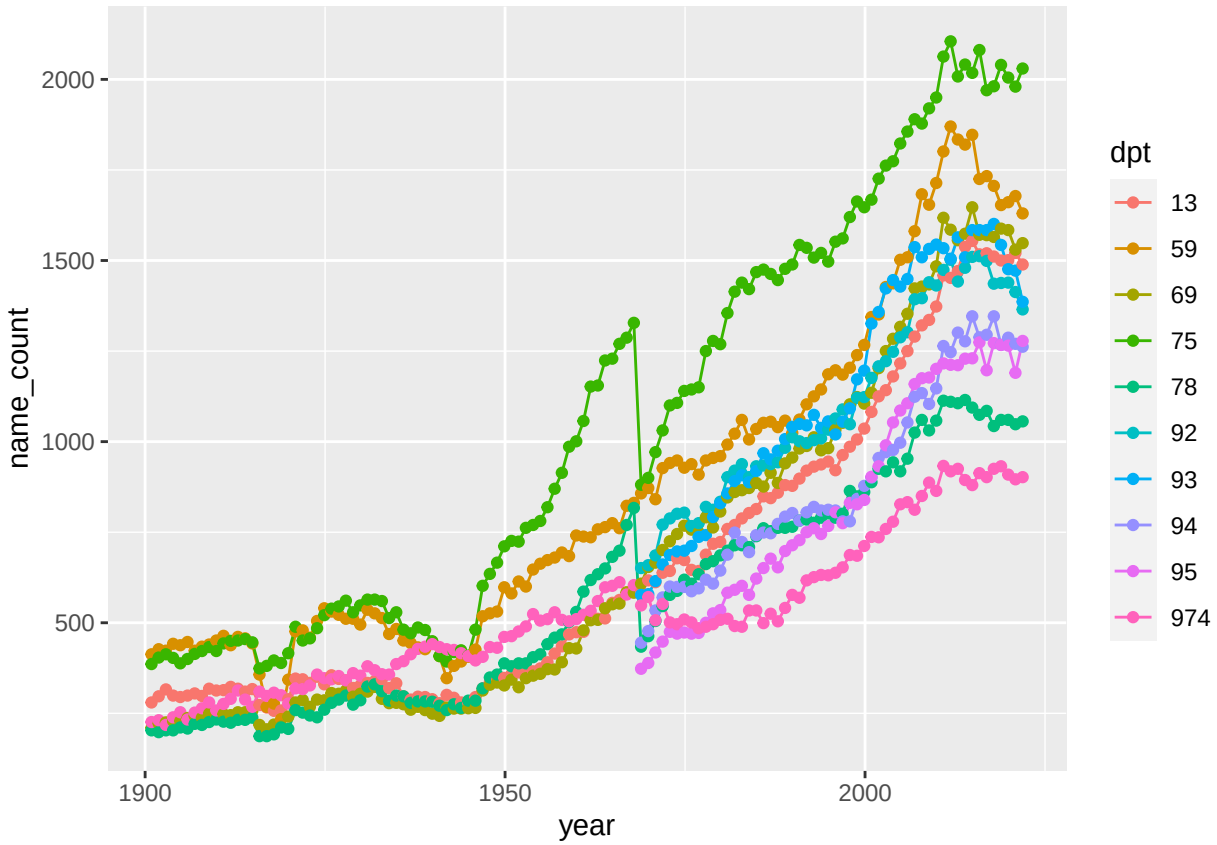
```
names_by_dpt = first_names %>%
                  group_by(dpt) %>%
                  distinct(name, dpt) %>%
                  summarize(name_count = n()) %>%
                  arrange(desc(name_count))

names_by_dpt_year = first_names %>%
                  group_by(dpt, year) %>%
                  distinct(name, dpt, year) %>%
                  summarize(name_count=n()) %>%
                  arrange(year)
```

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

Now, let's plot the top 10 departments with the highest variety of names and see how they have changed over time:

```
top_10_dpt = head(names_by_dpt, 10)$dpt
top_10_dpts_by_name_variety = names_by_dpt_year %>%
                           filter(dpt %in% top_10_dpt)
ggplot(data=top_10_dpts_by_name_variety, aes(x=year, y=name_count, color = dpt, group = dpt)) +
  geom_line() +
  geom_point()
```



The department with the most variety over time is “75”, that is, Paris. One possible explanation might be the denser population in the region and the concentration of immigrants from different parts of the world, which would naturally increase the name variety.