A SHORT ANALYSIS OF FRENCH FIRST NAMES FROM 1900 TO 2015

Kim Domptail - April 2017

Introduction

This is a short analysis of French first names between 1900 and 2015. The analysis is solely based on one database. It is not a sociological study of names. I was still pleasantly surprised by how many insights could be derived from a simple database with only 4 columns!

The database can be downloaded from the French National Institute of Statistics and Economic Studies (INSEE)'s website.

It contains four features:

- Gender
- First name
- Year of birth, between 1900 and 2015
- Number of people born with given name, gender, and year of birth

INSEE manages the database since 1946 and used the following criteria to include a name or not:

1. Between 1900 and 1945, the name has been given at least 20 times to either males or females.
2. Between 1946 and 2015, the name has been given at least 20 times to either males or females.
3. For a given year, the name has been given at least 3 times to males or females.

Names that do not comply with conditions 1 and 2 are grouped by sex and year of birth under one entry with the value 'PRENOMS_RARES' (rare names) in the ‘name’ column.

Names that comply with condition 2 but not condition 3 are grouped by sex and name under one entry with the value 'XXXX' in the ‘year’ column.

There are 589,411 entries, corresponding to almost 83 million people, and over 31,000 unique names.

Methodology

This is a personal project to practice programming so all the data manipulation and most of the visualization was done in Python 3.6. The treemaps were prepared using Microsoft visualization software Power BI.

Throughout the project, I practiced data manipulation (pandas dataframe), data visualization (pyplot), and some natural language processing (nltk, ngrams).

There are still many possible improvements to the code (e.g., factoring). However, since the code is only used for a ‘one-time’ analysis, and not to be used repeatedly, full optimization might not be necessary.

The code and graphs are available on my GitHub page (Jupyter notebook).

Results

Below is a summary of some of the most interesting results.

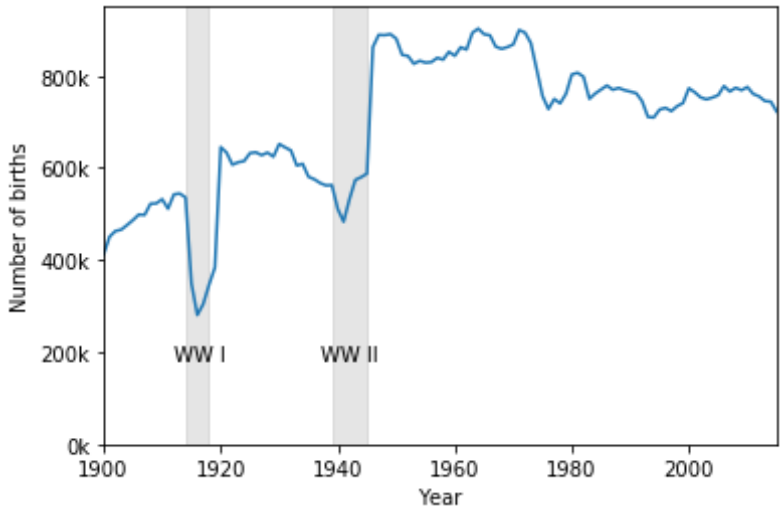
Dips in births during WWI & WWII and the Baby Boom

We can use the database to show the annual number of births in France between 1900 and 2015. (There might be some discrepancies with the actual number of births estimated by INSEE, especially for years prior to 1946.)

We can easily notice on the graph:

- the dip in the number of births during WWI, from 548,000 in 1913 down to 283,000 in 1916 (-48% in 3 years)!
- the dip in the number of births during WWII
- the Baby Boom after WWII, from 592,000 births in 1945 up to 893,000 in 1947 (+51% in 2 years)!

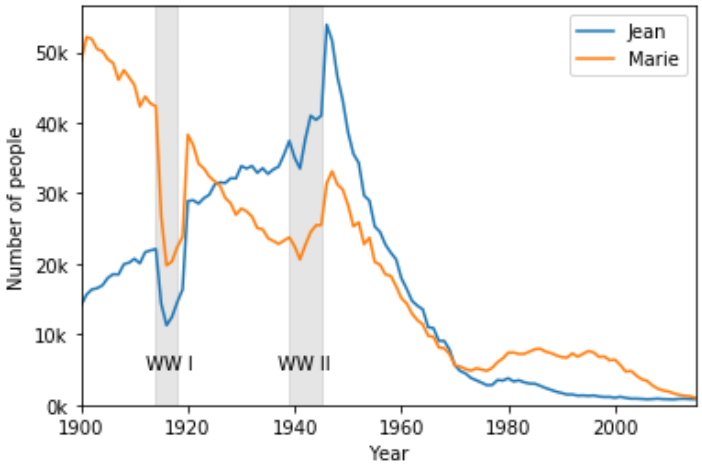
Number of births per year



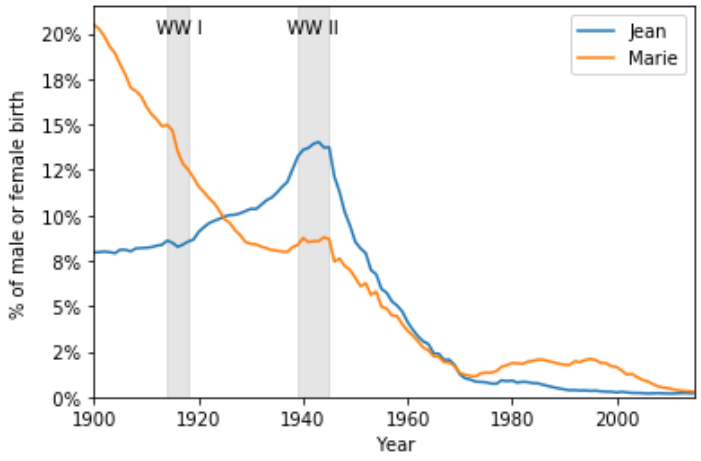
The impact of the two wars can be seen on the number of people named Marie and Jean, the two most given names overall throughout the century. We can control for the effects of the wars by calculating instead the percentage of births of the same gender these names represent each year.

Trend of the two most popular names overall Marie et Jean

In number of people



In percentage of male and female births respectively



Correlations between name trends and world events

My first name, Kim

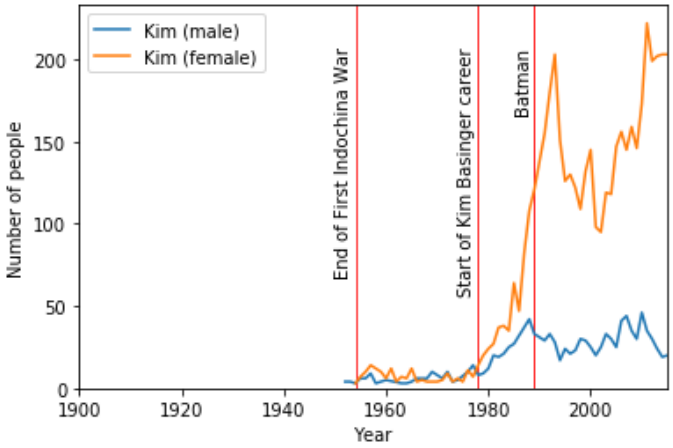
I grew up in France and had only ever met one other person named Kim before moving to the US. Indeed, there were only 79 people named Kim the year I was born (1986) and there have been less than 6000 people named Kim born in France in total between 1900 and 2015. Looking at the curve, we notice that:

- There were almost no Kims at all in France before 1952 (males) and 1954 (females).
- The number of Kims increases around 1978, and reaches a local maximum (231 Kims) in 1993.
- The number of Kims increased again in 2001 from 118 to an absolute maximum of 257 in 2011.

The trend of the name Kim is probably correlated to the following events:

- Immigration from Vietnam following the First Indochina War (1946-1954). Kim is a Sino-Vietnamese name (金) meaning gold (metal).
- The American actress, Kim Basinger, started her acting career around 1978, and gained mainstream exposure in 1989 (Batman).
- There are other famous Kims, such as the English singer Kim Wilde, popular in France since 1981, and Kim Kardashian popular since the 2007 reality show.

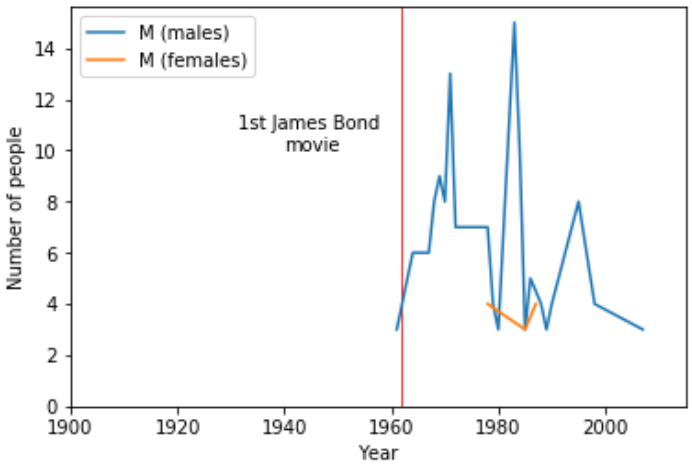
Trend of the name Kim

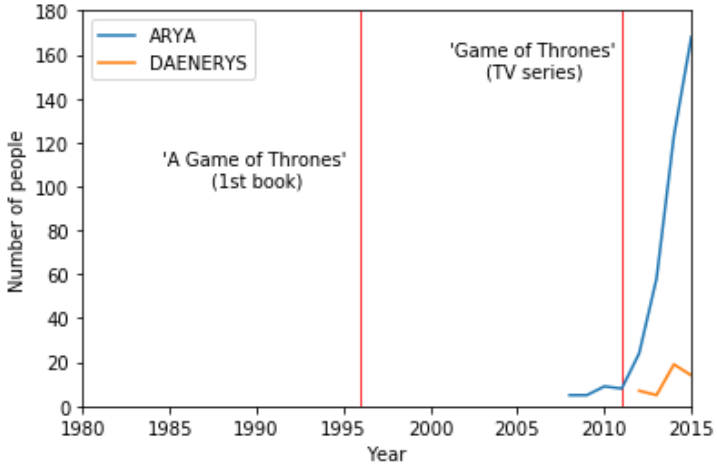


Names appearing after movies

Some names are particular (e.g., M, Arya and Daenerys) and can be assumed to have been popularized by movies.

Genesis of the names M, Arya and Daenerys



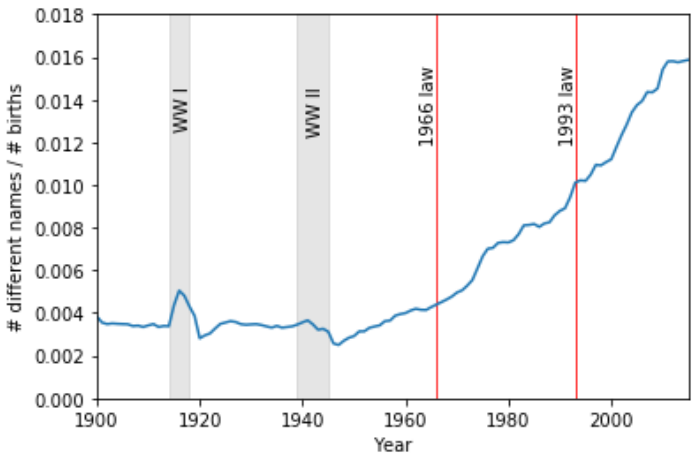


Increasing diversity in names

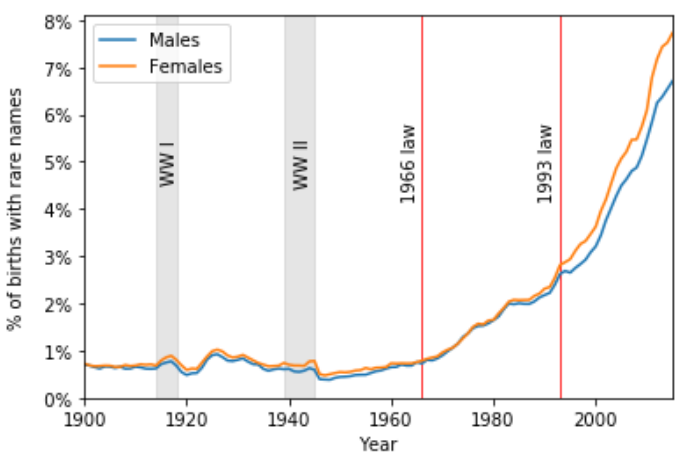
The number of different names increased from 1,600 in 1900 to 12,400 in 2015! This is due in part to the increase in the number of births. However, the ratio of names to births also quadrupled between 1900 (0.004) and 2015 (0.016). Another indicator of name diversity is the number of people with rare names, which jumped from 2,980 people in 1900 (0.7% of births) to 56,107 in 2015 (7.2% of births).

The increasing name diversity can be correlated to the different French 'Naming Laws'. The Law of 1803 limited names to those found in calendars of saints and 'historic names'. The Law of 1966 extended the list of authorized names (e.g., mythology, regional, and foreign names). Finally, the Law of 1993 authorizes almost any name, as long as it is not contrary to the interests of the child.

Number of different names per birth



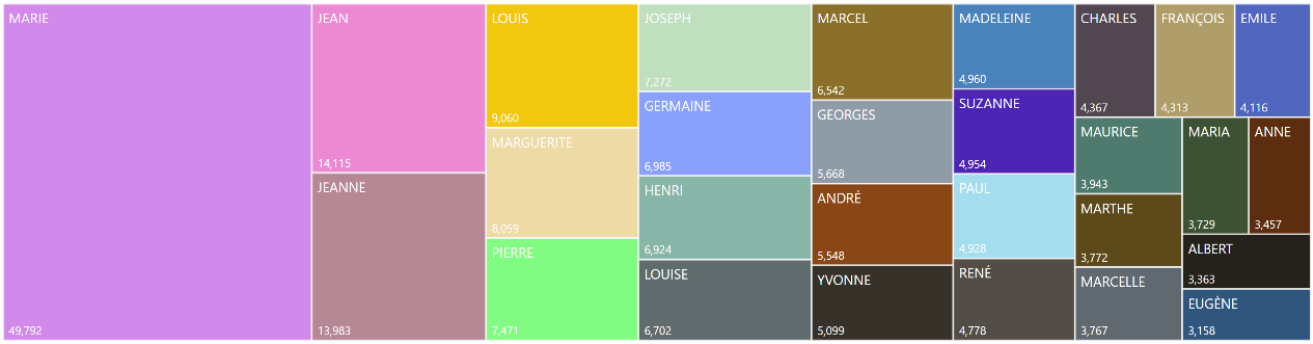
Percentage of births with 'rare names' per gender



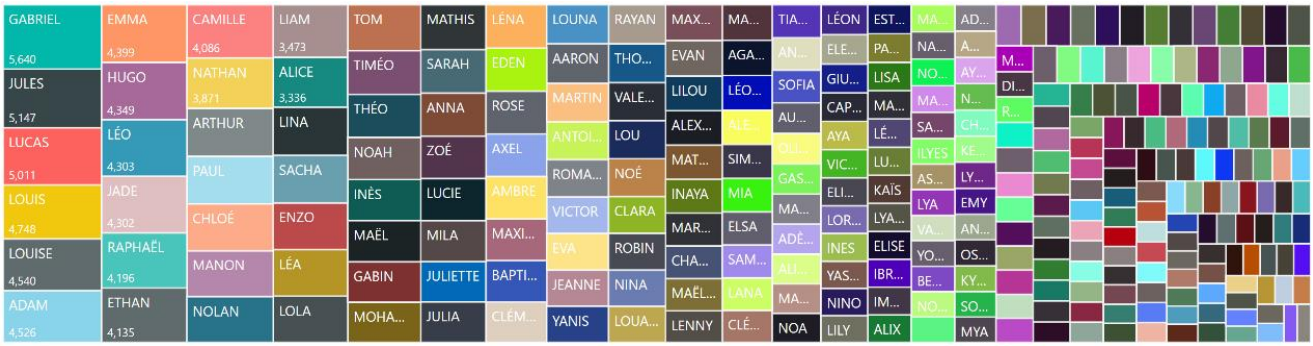
Treemaps: Half of the population was represented by only 28 names in 1900, but 297 names in 2015. Similarly, the top 20 names represented 43% of total births in 1900 but only 11% in 2015.

Treemaps of names representing 50% of the population

1900



2015

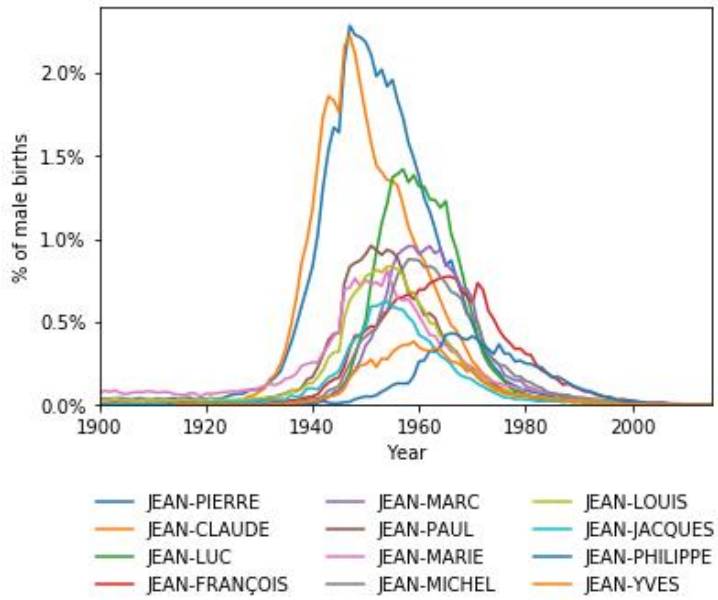


Compound names most popular in 1940-1965

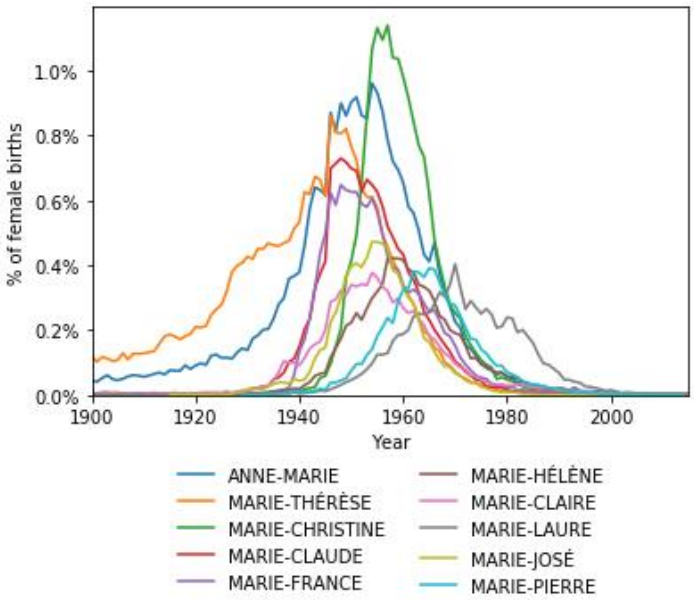
In general, both male and female compound names were most popular between 1940 and 1965. We can also note that most compound names comprised the names 'JEAN' and ‘MARIE’ respectively!

Trend of compound names with at least 10,000 people overall

Male



Female



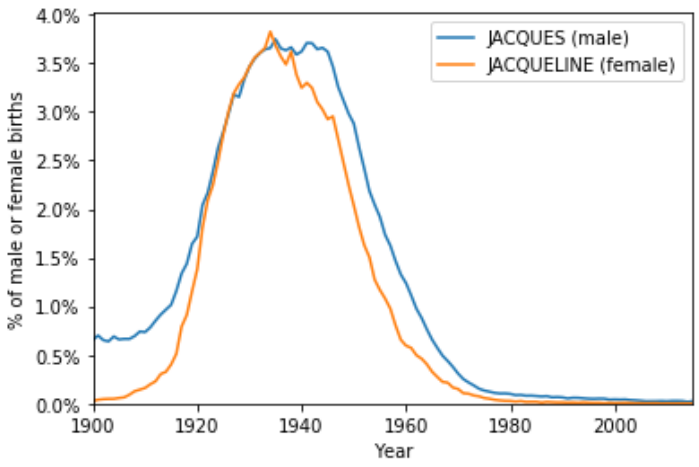
We can also observe common trends for similar types of names (similar types of names were identified using trigrams):

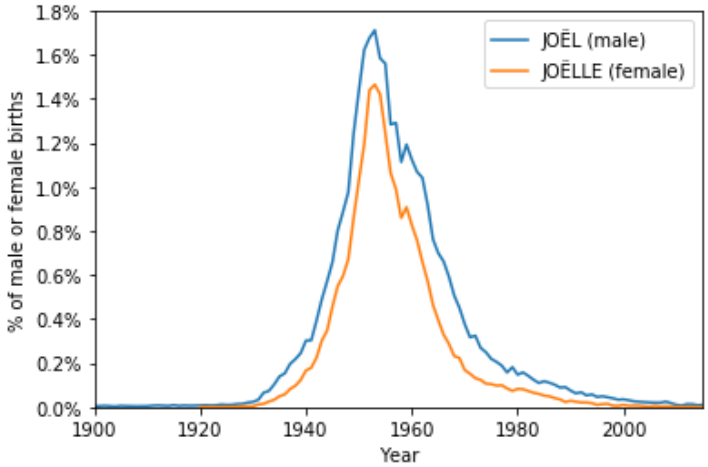
- Names ending in -ETTE (e.g., Paulette, Yvette, Odette, Colette) were most popular in 1920-1940.
- Names ending in -IANE (e.g., Christiane, Josiane, Liliane) were most popular in 1940-1960.
- Names containing 'CLAUD' (e.g., Claude, Jean-Claude, Claudine) were most popular in 1930-1960.
- Names ending in '-INE' (e.g., Catherine, Martine, Christine, Sandrine) were most popular between 1950 and 1980.

Parallel trends for male/female name pairs

N-grams (in this case five-grams) can help us find pairs of male/female names (e.g., Julien/Julie, Christian/Christiane, Louis/Louise, Patrice/Patricia, André/Andrée). In general, both names follow the same trend.

Examples of parallel trends for male/female name pairs



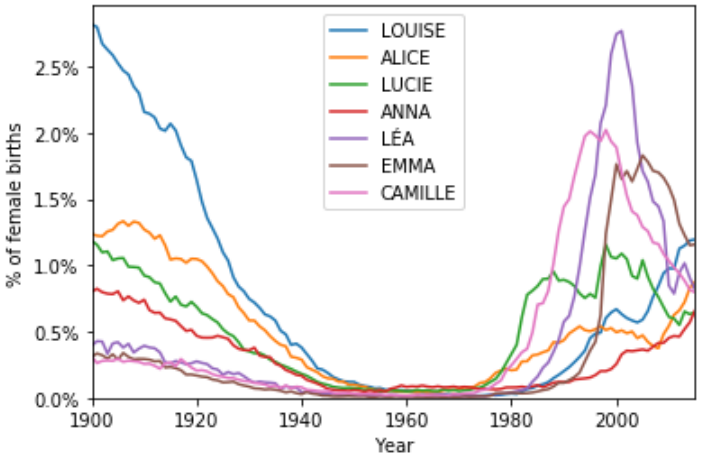


Old names and recent names in the Top 15

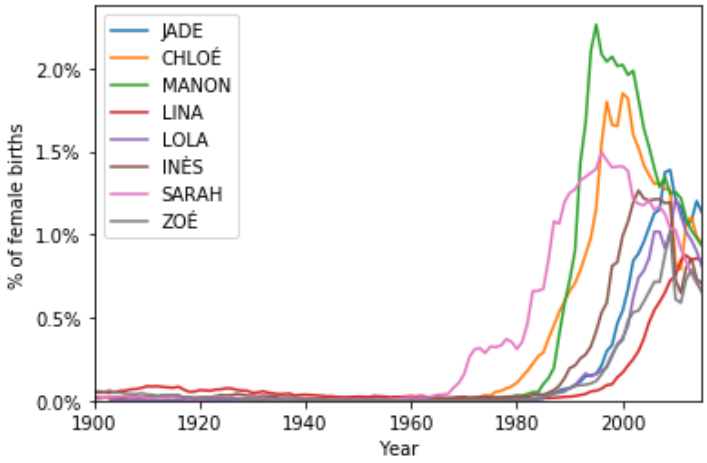
The top 15 names in 2015 comprised both 'old' names becoming popular again (e.g., Louis, Paul, Jules for males and Louise, Alice, Lucie for females), and 'recent' names (e.g., Lucas, Adam, Hugo for males, and Jade, Lina, Lola for female).

Top 15 female names in 2015

Old names becoming popular again



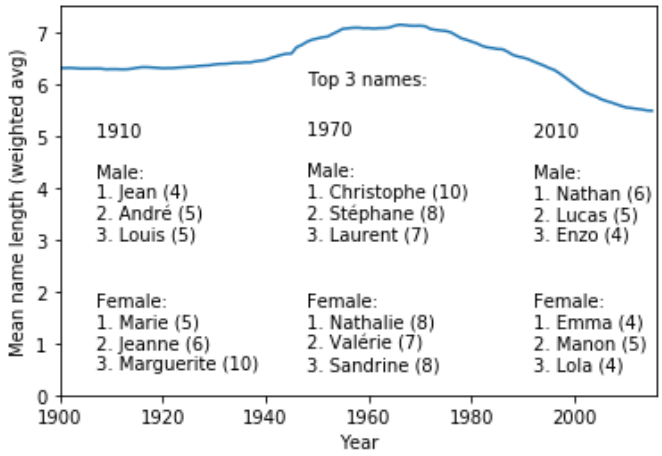
'Recent' names



Shorter names than in the past

The mean name length went from 6.3 letters in 1910 to 7.1 in 1970 and 5.5 in 2010. Throughout the century, names range from one letter (A, L, N, M) to 19 letters (Guillaume-Alexandre, François-Christophe)!

Evolution of mean name length (weighted average by number of births)



Additional Questions

Below is a non-exhaustive list of other topics I would be interested in studying:

- **Impact of movies**: Some correlations with movies/books (characters/actors) were shown anecdotally with some names (e.g., M, Arya, Daenerys). A more thorough analysis could be conducted using for example IMDb database.
- **Regional specificities**: A more detailed analysis could show differences/similarities between France's regions. A file with regional information is available on INSEE's website. See for example the analysis done by Le Monde.
- **Soundex**: The Soundex phonetic algorithm could be used to identify homophone names and further the analysis based on ngrams. For example, new spellings of the same names may also explain the increasing name diversity.
- **Classification of curves**: Classifying trend curves may enable us to predict lifetime and periodicity of names.