French given names per year per department analysis

Antoine Saget

November, 2020

1 - The data

1.1 - Building the Dataframe from the file

According to the Dataset's dictionary of variables found here (tab "DICTIONNAIRE DES VARIABLES"), the dataset is composed of the following columns :

	SEXE	PREUSUEL	ANNAIS	DPT	NOMBRE
Type	Character	Character	Character	Character	Numerical
Details	`1` for male, `2` for female	25 Chars or less	4 Chars, from 1900 to 2019 or `XXXX`	3 Chars, department or `XX`	Up to 8 digits

After looking into the documentation found here (tab "DOCUMENTATION") we understand that row with `ANNAIS = XXXX` or `DPT = XX` corresponds to a name given at least 20 times over a given period, but less than 3 times for a given year. We decide to discard them. We also decide to consider `ANNAIS`, `DPT` and `SEXE` as integers in order to restrict our parsing as much as possible and detect/discard potential error lines that doesn't respect the `integer;character;integer;integer;integer` format.

```
FirstNames <- read_delim("dpt2019.csv", delim = ";",
  col_types = "iciii")
## Warning: 72890 parsing failures.
##   row    col   expected actual          file
## 10781 annais an integer   XXXX 'dpt2019.csv'
## 10781 dpt    an integer   XX   'dpt2019.csv'
## 10782 annais an integer   XXXX 'dpt2019.csv'
## 10782 dpt    an integer   XX   'dpt2019.csv'
## 10783 annais an integer   XXXX 'dpt2019.csv'
## ..... ...... .......... ...... ............
## See problems(...) for more details.
```

We have 72890 parsing failures. We make sure every error come from either an `XXXX` year or an `XX` department :

```
parsing_failures <- problems(FirstNames)
parsing_failures_notX <- parsing_failures[parsing_failures$actual !=
  "XXXX" & parsing_failures$actual != "XX",
  ]
print(paste("Number of parsing failures different from an XXXX year or an XX department :",
  nrow(parsing_failures_notX)))
## [1] "Number of parsing failures different from an XXXX year or an XX department : 0"
```

0 parsing failures different from the expected one. Let's remove the failures from the Dataframe :

```
print(paste("Number of row before removing failures :",
  nrow(FirstNames)))
## [1] "Number of row before removing failures : 3676682"
print(paste("Expected number of rows after removing failure :",
  nrow(FirstNames) - nrow(parsing_failures)/2))
## [1] "Expected number of rows after removing failure : 3640237"
FirstNames <- FirstNames[-parsing_failures$row,
  ]
print(paste("Number of row after removing failures :",
  nrow(FirstNames)))
## [1] "Number of row after removing failures : 3640237"
```

Our Dataframe is now ready for analysis.

1.2 - Exploring the Dataframe

Before analyzing in depth, let's take an overview of the data.

The summary in Table 2 doesn't give much interesting results. We can still note that the gender mean is at 1.536, meaning that there are slightly more female rows than male (doesn't mean more female than male, just the number of row). Also we can see that for a given department, one year, 6316 people where born given the same name !

6316 Jean are born in 1946 in Paris (Table 3) !
Some general stats of the Dataset :

```
## [1] "Number of rows in the Dataframe :  3640237"
## [1] "Number of males identified :  39486980"
## [1] "Number of females identified :  37926398"
## [1] "Number of different first names :  15905"
```

We can see that even with slightly more females row, there are fewer (37926398) females identified than males (39486980).

Some male first names :

```
##  [1] "SAÏM"     "GÜNEY"    "UMUT"     "BERNADIN" "DAVENS"   "RIYAD"
##  [7] "TIERRY"   "SAMUËL"   "ONUR"     "KHELIS"
```

Some female first names :

```
##  [1] "MÉLOÉE"          "ELISETTE"        "KADIJA"          "AXELE"
##  [5] "SUSIE"           "NADIA"           "LALI"            "NOR"
##  [9] "OUMNIA"          "MARIE-CAROLINE"
```

Table 2: Summary of the FirstNames Dataframe.

Gender	First name	Birth year	Department	Number
Min. :1.000	Length:3640237	Min. :1900	Min. : 1.0	Min. : 3.00
1st Qu.:1.000	Class :character	1st Qu.:1948	1st Qu.: 31.0	1st Qu.: 4.00
Median :2.000	Mode :character	Median :1980	Median : 57.0	Median : 7.00
Mean :1.536	NA	Mean :1973	Mean :101.2	Mean : 21.27
3rd Qu.:2.000	NA	3rd Qu.:2002	3rd Qu.: 77.0	3rd Qu.: 18.00
Max. :2.000	NA	Max. :2019	Max. :974.0	Max. :6316.00

Table 3: Most attributed name in one year in one department.

Gender	First name	Birth year	Department	Number
1	JEAN	1946	75	6316

2 - First name frequency analysis

2.1 - Bruno

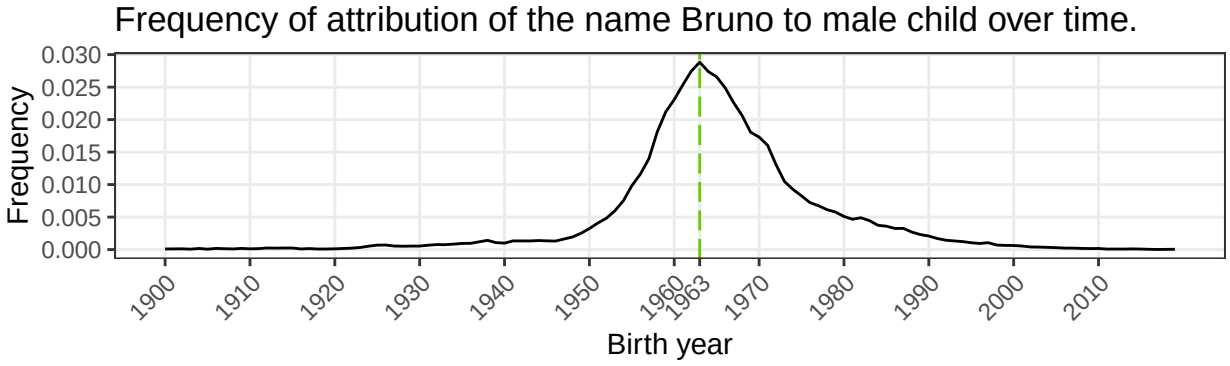
Let's analyze the frequency of the first name **Bruno**. First, we look at names close to Bruno (names containing BRUN) (Table below, *lost too much time trying to reference it / place it at the right position / caption it, but without success... still learning.*).

We can see that Bruno is sometimes used for females, also Bruneau exist. Let's check how many female Bruno have been identified :

```
## [1] "Number of female given the name Bruno :  60"
## [1] "Number of male given the name Bruno :  217332"
## [1] "Percentage of Bruno females compared to male :  0.0276075313345481"
```

The females named Bruno represents less than 0.03% of the identified people named Bruno. In the following we will focus only on the first name spelled `Bruno` and given to male.

Let's see the evolution of the frequency of parents giving their male child name `Bruno`. The frequency is important here because the number of given name (number of births) per year is not constant from one year to the other. Also, as we are looking at the male attribution of the name, it make sense to look at the frequency of attribution of the name within male attributions only. Indeed, if for a given year much more female are born (this shouldn't happen, but still), then the number of people given the name Bruno might reduce even if the frequency of attribution of the name Bruno to male child is still the same.



Gender	First name
1	BRUNEAU
1	BRUNO
1	BRUNON
1	BRUNY
1	JEAN-BRUNO

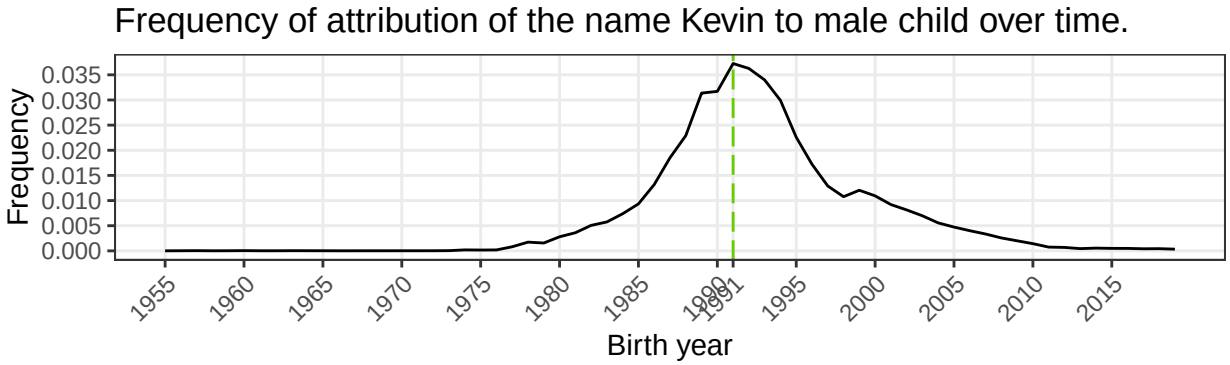
Gender	First name
2	BRUNA
2	BRUNAELLE
2	BRUNE
2	BRUNEHILDE
2	BRUNELLA
2	BRUNETTA
2	BRUNETTE
2	BRUNHILDE
2	BRUNO

```
## [1] "The named Brunos was attributed the most times to male child ( 12738 ) in 1963"
## [1] "and with the highest frequency ( 0.0288378600487649 ) in 1963"
```

The frequency of the attribution of the name Bruno to male child increase from 1900 to 1963 with a peak of 12738 attributions (almost 3% of the attributed male name) in 1963 and decrease from this point. It seems that the name follow a kind of "trend pattern" where the name is more and more used until it's used too much and replaced by probably another trendy name at that time.

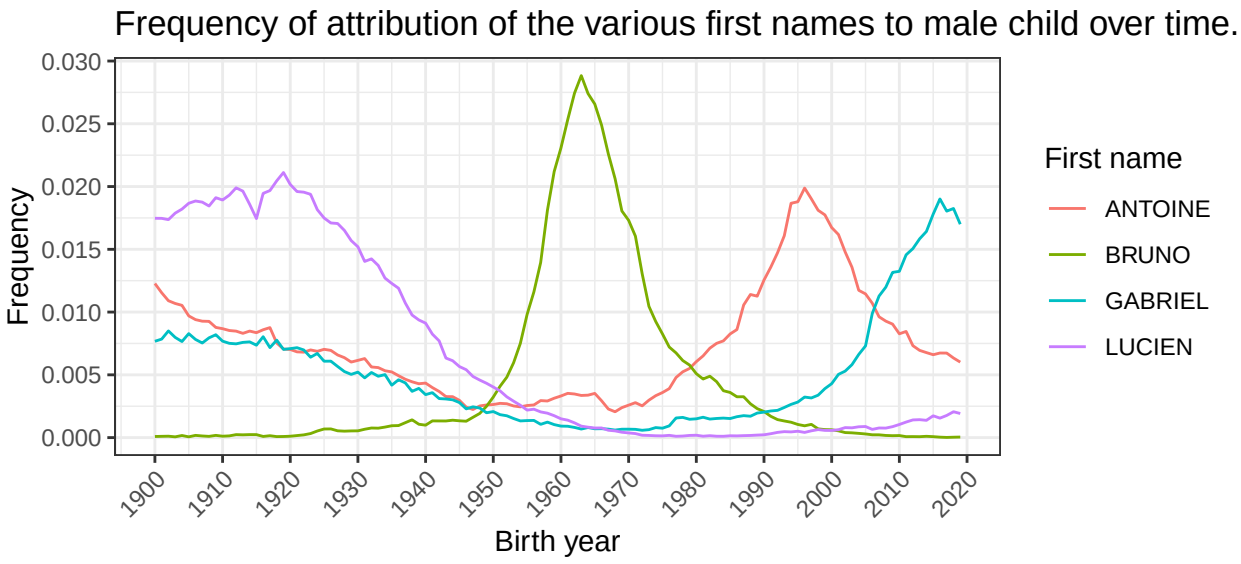
2.2 - Bruno compared to other first names

With the "trend" hypothesis in mind, let's look at the evolution of the attribution of the name Kevin which was (in our opinion) very trendy in the 2000's.



As we can see, Kevin follow the same evolution of Bruno but much later. Indeed, Kevin has never been attributed - enough time to be registered by the Insee - before 1955 and peak only in 1991.

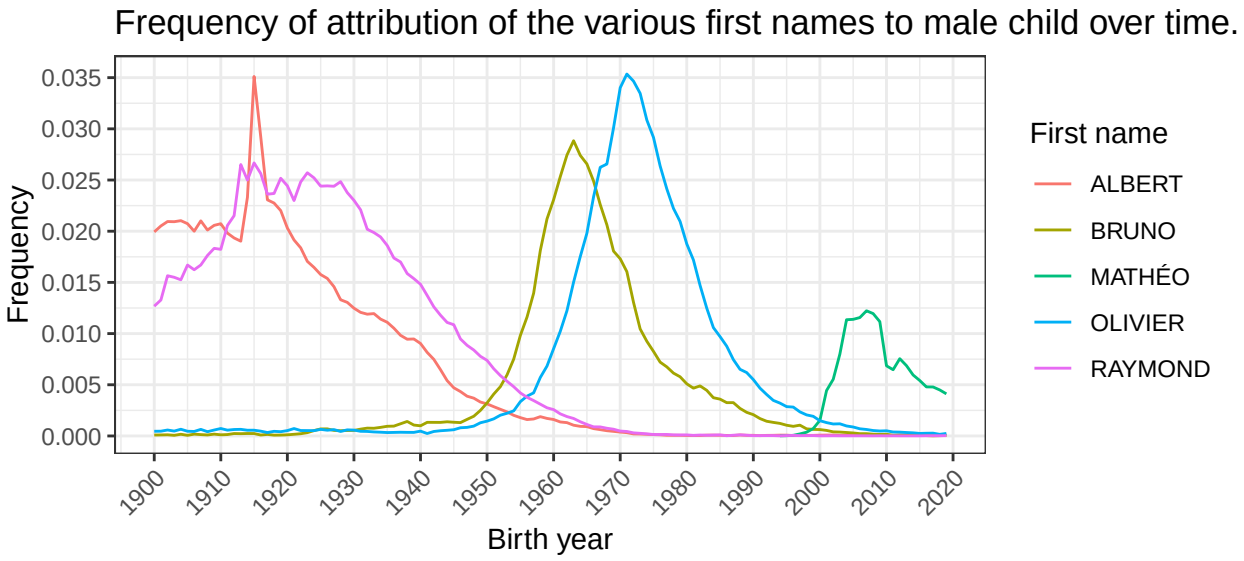
Now let's compare several first names frequencies :



Some name like Lucien were often attributed in the past but are old-fashioned now. Others, like Gabriel or Antoine were quite used in the past, decreased in popularity and became trendy again in the 2000's.

We've seen earlier that more than 15000 unique names where identified in the dataset but we can't come up with that many name ideas, so let's compare Bruno against random names and see if we find interesting results.

We often have very rare names leading to result that are difficult to interpret so we'll only take random names that are relatively frequent.



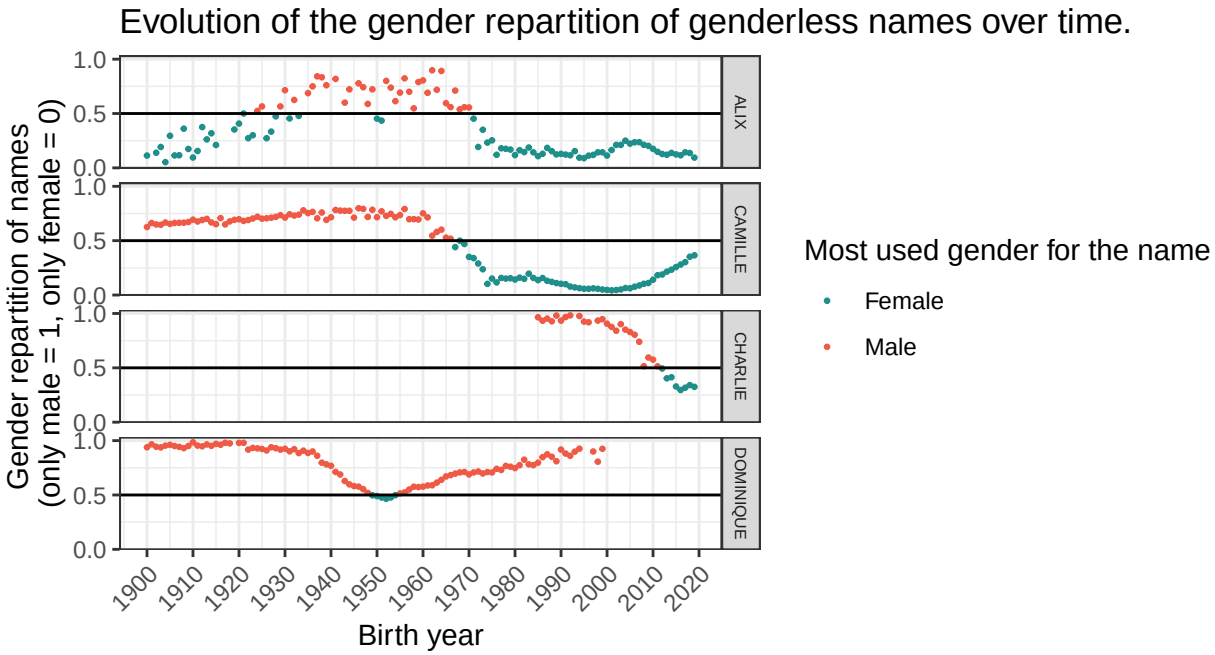
We can see here that Albert was more popular than Bruno in the past and that Mathéo is a very recent name that is already decreasing in popularity.

2.3 - Further experiments

From the past experiments it's seems to us that first names are more varied today than in the past, let's check this hypothesis by plotting the number of unique name attributed per year :

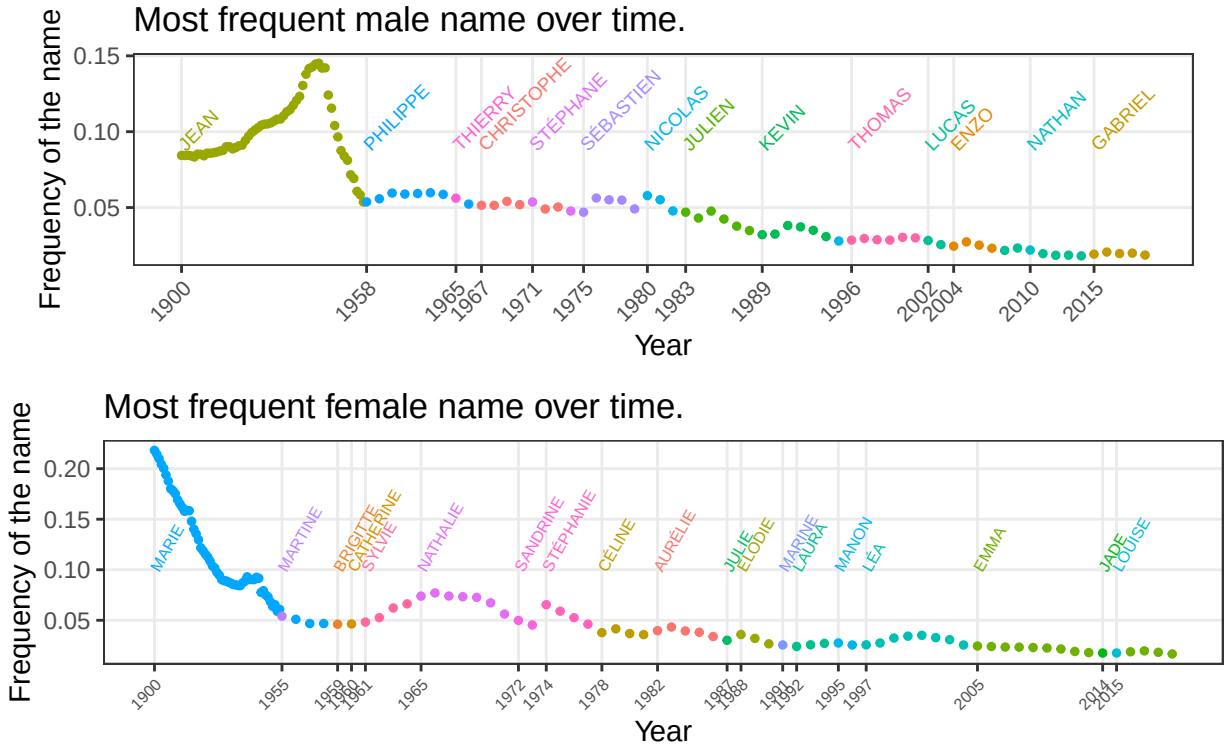
> Note that the evolution of the number of births per year in France since 1900 is quite steady (even decreasing a little), so "more varied names" cannot be explained by "more babies".

We only looked at male first names so far. Let's look at the evolution of the distribution of a genderless names over time.



We can see for example that Alix is mostly used as a female name with exception for the 1930-1970 period. Dominique on the other hand is mostly used as male name. An interesting shift happened to Camille that was mostly used as a male name before 1965 and is mostly used as a female name since.

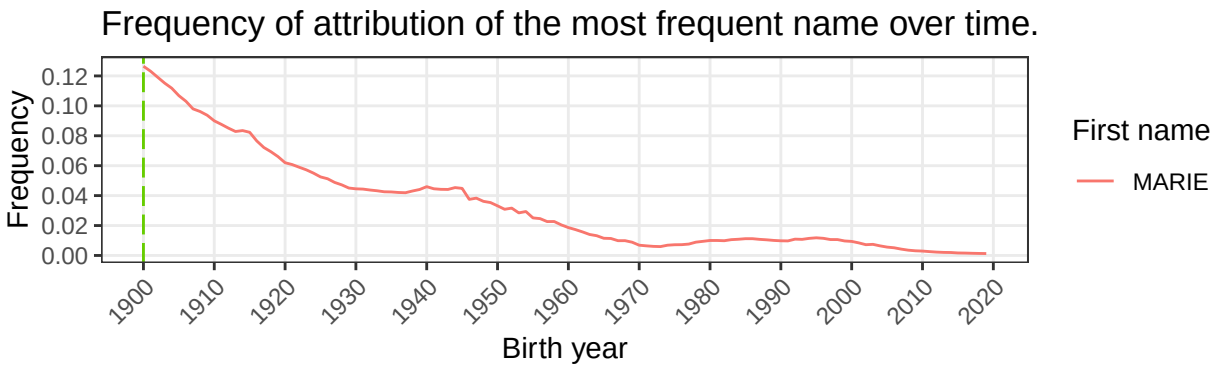
3 - Most given firstname by year



We can see that for both males and females the number of occurences of the most frequent name decrease

over time, this make sense because of the increase of the variety of names we demonstrated earlier.

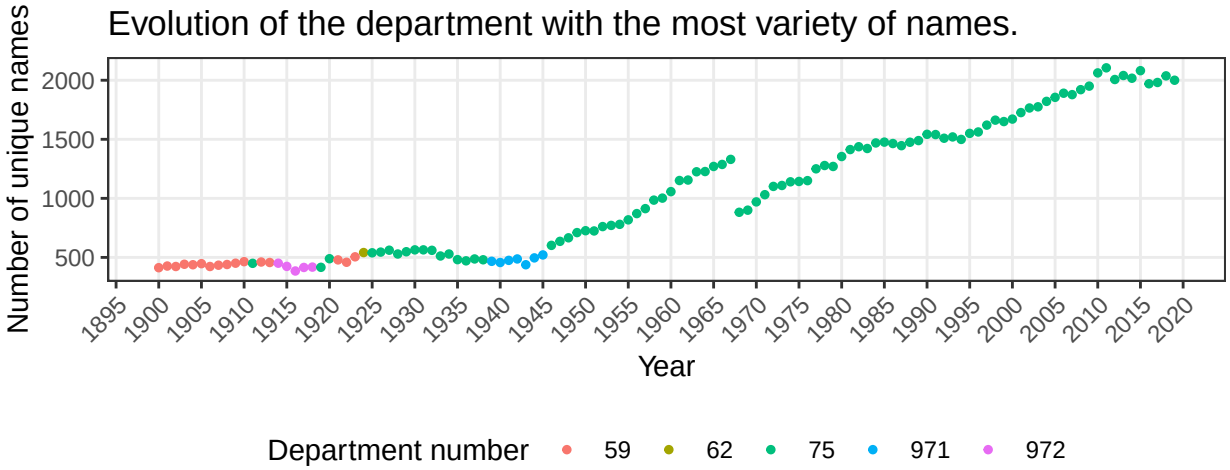
Now let's see the frequency of the most frequent name with no regards to it's gender.




```
## [1] "Number of Marie born since 1900 : 2256131"
```

With 2256131 attributions since 1900 and more than 12% of attributions in 1900, Marie is the overall most frequent name given to children since 1900 in France.

4 - Variety of name



The most varied department is Paris. This make sense as Paris concentrate a lot of population and a very diverse population as well (potentially more unique names when mixings origins).

We can see that from 1939 to 1945, Guadeloupe is the most diverse department, the same situation happens in 1914-1918 and this time Martinique is the most varied department. This can be explained in two ways. Firstly, both periods where time of war during which the birth rate significantly dropped, less babies mean less chance of varied names. Secondly, as this was time of war, it's possible that there is a lot of missing data during this period. So, Martinique and Guadeloupe didn't suddenly become more diverse, it's just that variety decreased in mainland France during those period as a result of the birth rate decreasing (and/or missing data because of the wars).