

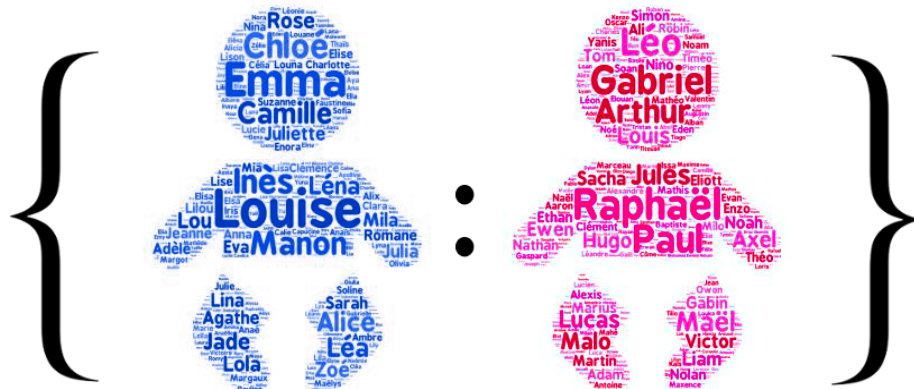


le  
campus  
numérique



# Projet Python 2 :

## Origines et évolution des prénoms en France



[ Source : plus ou moins data.gouv.fr ]

Type "help", "copyright", "credits" or "license" for more information.

```
>>> formateur.request ( " Florian Dadouchi " )
```

## Objectifs du module

L'objectif de ce projet est de **stabiliser les bases de l'analyse de données en Python**. Vous devrez également vous servir de ces deux jours pour valider les compétences liées à ce module.

Pour ce faire vous avez deux options :

1. Reprendre le projet "**Arbres de Grenoble**" débuté en Décembre. Vous connaissez déjà ce dataset. Cela vous permet de reprendre votre code et "d'aller droit au but".
2. Commencer une nouvelles analyse sur les "**Origines et évolution des prénoms en France**", leur évolution dans le temps et leurs origines.

## Modalités

- Durée du projet : 2 jours.
- Ce projet sera réalisé en binôme
- Vous produirez un notebook commun au binôme

## Dataset du projet

- Ce projet s'insère encore une fois dans le cadre de l'**Open Data**. Cette fois ci nous analyserons conjointement deux dataset :
  1. Un fichier de données nationales qui contient les prénoms attribués aux enfants nés en France (hors Mayotte) entre 1900 et 2018 et les effectifs par sexe associés à chaque prénom. Les données sont classées par département. Le fichier contient 3.5 millions de lignes. Le fichier est disponible à cette adresse :  
<https://www.data.gouv.fr/fr/datasets/fichier-des-prenoms-de-1900-a-2018/>
  2. Un second dataset issu du travail de Mike Campbell au travers de son site web "Behind the Name" : <https://www.behindthename.com/>

*Behind the Name is a website for learning about all aspects of given names. Its scope is broad: all given names from all cultures and periods are eligible to be included in the main name database. Names from mythology and fiction are also eligible. There are currently 22874 names in the database, a fraction of what the scope entails. There's still much work to be done.*

La base de données issue de ce site est également disponible sur [data.gouv.fr](https://www.data.gouv.fr) et comprend plusieurs champs, dont les origines des prénoms ! La base de données issue de ce site est disponible à cette adresse :

<https://www.data.gouv.fr/fr/datasets/liste-de-prenoms/>

Mike Campbell est software designer and information architect, et basé à Victoria, British Columbia, Canada) :

- une interview : <https://nameberry.com/blog/behind-the-namer>
- si vous voulez parler avec lui : <https://www.linkedin.com/in/mike-campbell-218b8124/>

## Contexte du projet

- L'analyse des prénoms n'a rien de nouveau en soi, c'est même un grand classique : analyse de popularité instantanée, historique, "changement de sexe" des prénoms, etc. Voici deux exemples de liens ayant réutilisé ce 1er dataset :
  - Réutilisation de type "*analyse*" (évolution dans le temps, ...) :  
<https://www.lefigaro.fr/fig-data/prenoms/> (désolé...)
  - Réutilisation de type "*podium*" des prénoms :  
Insee : <https://www.insee.fr/fr/statistiques/3532172>

Vous pourrez très facilement retrouver nombre d'analyses en allant voir les "réutilisations" et "contributions communautaires" en bas de page sur [data.gouv.fr](https://www.data.gouv.fr) par exemple.

- Le réel apport que nous pouvons proposer dans ce projet est d'inclure une dimension supplémentaire à l'analyse : les langues d'origine des prénoms.

## Remarques préliminaires

- Le sujet est long mais vous n'êtes absolument pas tenus d'aller jusqu'au bout !
- Je n'introduis pas de nouveaux concepts par rapport à la session précédente pour vous donner le temps de repasser sur ce que vous avez déjà vu.
- N'oubliez pas que vous êtes libres de prendre des libertés par rapport aux consignes exploratoires que je vous donne par la suite. Vous pouvez sauter des parties ou en approfondir certaines.
- Vous êtes également libres de publier vos travaux, c'est à dire les ajouter à la liste des "réutilisations" d'un dataset. Un python notebook annoté correctement fait largement l'affaire !

## Consignes

- Pour commencer il faudra lire le csv
  - Pour ceux qui ne veulent pas se rajouter de difficulté supplémentaire (au moins dans un premier temps) un second fichier contenant un peu moins d'information est disponible ici : <https://www.data.gouv.fr/fr/datasets/r/fa6a5147-7fe0-41bd-9d25-b2baef879f68>
  - Ce fichier agrège l'information relative au département. Il est donc un peu plus facile à prendre en main. Il contient environ 630.000 lignes. Cela dit c'est un bon exercice de savoir passer de l'un à l'autre. Vous ne serez pas pénalisés par le choix de l'un ou de l'autre des fichiers csv.
- Commençons par quelques sélections basiques :
  - Quelle est la proportion totale Femmes / Hommes au cours depuis 1900 ?
  - Comment a-t-elle évolué au cours du temps ?
- Pour se mettre en jambe on va d'abord essayer de retrouver quelques uns des résultats présentés dans l'article du Figaro (<https://www.lefigaro.fr/fig-data/prenoms/>):
  - En 1900, parmi les 1707 prénoms recensés par l'Insee, le plus donné, Marie, représentait 11% des naissances.
  - En 2017, on recensait 13.000 prénoms différents, soit 7,6 fois plus qu'en 1900.
  - Evolution du prénom Marie dans le temps ?
  - Un pic a été atteint en 2012, avec plus de 13.643 prénoms recensés.
  - Evolution des prénoms composés contenant Marie au cours du temps (Marie-Pierre, Marie-Paul, ....) ?
  - Combien de prénoms faut-il pour nommer 50% des bébés par année (e.g. en 1900 27 prénoms suffisent) ?
  - La part des 10 prénoms les plus utilisés a-t-elle bien été divisée par 5 entre 1900 et aujourd'hui ?

- [Plus dur] Qu'en pensez vous? Extrait du Figaro :

À l'échelle régionale, le prénom Loïc, très populaire entre 1975 et 2000, constitue un bon exemple de cette circulation des prénoms. Cantonné à la Bretagne jusque dans les années 1980, il a essaimé en France jusqu'à être davantage donné dans d'autres régions, s'effaçant en Bretagne.

- [Plus dur] Qu'en pensez vous? Extrait du Figaro :

De même, les Erwan, Enora (bretons), Maelys (occitan) courent dans toute la France, tout comme les Wilfried (anglo-saxon), Enzo, Nino (italiens), Medhi, Inès (arabes) ou Lola (espagnol) ont fait leur trou. À l'inverse, les Armel (breton), Maïte (basque), Giulia (italien) restent majoritairement limités à une région.

- Analysons maintenant cette base de données des prénoms donnés en France à la lumière des informations fournies par "Behind The Name".

- Pensez à nettoyer et normaliser la base de données des prénoms (doublons, ...)
- Quelle est la représentativité de cette base de données ? Quel pourcentage de prénoms arrive-t-on à faire correspondre à la base de données précédente ?
- Essayer de retrouver certaines tendances temporelles dans l'attribution des prénoms ? Donne-t-on plus de prénoms d'origine différente maintenant ? Quand est-ce que ça s'est diversifié ? Y a-t-il des origines qui ont disparu ? Qu'en est-il des prénoms basque ? Du bre
- Donne-t-on plus de prénoms Espagnols ou Catalans près de la frontière espagnole? Italiens vers l'Italie? Etc?
- Si vous manquez d'inspiration, voici quelques extraits d'articles que j'ai trouvé en fouillant sur internet :
  - "Pourtant, en Bretagne toujours, le recours aux prénoms bretons moins connus ne cesse de croître depuis les années 1980, selon l'analyse de Baptiste Coulmont." Qu'en pensez-vous ?
  - "Stéphanie Rapoport établit un lien entre ce phénomène et l'ère d'Internet, dans une société où les prénoms et les cultures circulent. «Certains piochent volontairement dans des inspirations étrangères "

- Extrait du Monde<sup>1</sup> :

Les émigrés d'Europe du Sud, qui sont arrivés en France entre les années 1930 et la fin des années 1970, ont surtout donné à leurs enfants des prénoms français. D'une part parce qu'ils ne pensaient pas retourner dans le pays dans lequel ils étaient nés. D'autre part parce que les prénoms de leur pays d'origine étaient souvent assez proches (Maria, Marie) ou avaient des équivalents structurels (Giuseppe, Joseph).

---

1

[https://www.lemonde.fr/idees/article/2018/10/11/baptiste-coulmont-les-prenoms-du-pays-d-origine-s-estompent-en-faveur-de-ceux-du-pays-d-accueil\\_5368061\\_3232.html](https://www.lemonde.fr/idees/article/2018/10/11/baptiste-coulmont-les-prenoms-du-pays-d-origine-s-estompent-en-faveur-de-ceux-du-pays-d-accueil_5368061_3232.html)

## Pour ceux qui vont vite :

- Vous avez probablement remarqué que la dernière date de mise à jour de la liste des prénoms et de leur origine sur data.gouv.fr remonte à 2014... Pourtant Mike et ses amis ont continué à bosser entre temps! Des données supplémentaires (quantitatives et qualitatives) existent sur leur site. Une Application Programming Interface (**API**) existe et est documentée ici : <https://www.behindthename.com/api/> Essayez d'obtenir les infos manquantes pour votre analyse!
- Pour ceux qui voudraient étendre leur analyse au niveau mondial, les listes des prénoms donnés en Belgique, USA, Pologne, .... sont centralisées ici : <https://www.politologue.com/prenoms/>

## Ressources

Quelques ressources sur les prénoms :

- [https://www.ined.fr/fichier/s\\_rubrique/29081/565.population.societes.avril2019.immigres.prenoms.france.fr.pdf](https://www.ined.fr/fichier/s_rubrique/29081/565.population.societes.avril2019.immigres.prenoms.france.fr.pdf)
- <http://coulmont.com/blog/2019/04/10/pop-soc-prenoms-blog/>
- [https://www.lemonde.fr/idees/article/2018/10/11/baptiste-coulmont-les-prenoms-du-pays-d-origine-s-estompent-en-faveur-de-ceux-du-pays-d-accueil\\_5368061\\_3232.html](https://www.lemonde.fr/idees/article/2018/10/11/baptiste-coulmont-les-prenoms-du-pays-d-origine-s-estompent-en-faveur-de-ceux-du-pays-d-accueil_5368061_3232.html)
- [https://www.liberation.fr/france/2019/04/10/patrick-simon-et-baptiste-coulmont-on-ne-peut-pas-juger-la-volonte-d-assimilation-en-ne-se-fondant-q\\_1720602](https://www.liberation.fr/france/2019/04/10/patrick-simon-et-baptiste-coulmont-on-ne-peut-pas-juger-la-volonte-d-assimilation-en-ne-se-fondant-q_1720602)
- [https://www.ined.fr/fichier/s\\_rubrique/29081/565.population.societes.avril2019.immigres.prenoms.france.fr.pdf](https://www.ined.fr/fichier/s_rubrique/29081/565.population.societes.avril2019.immigres.prenoms.france.fr.pdf)

Tutoriel rapide sur pandas :

- [https://pandas.pydata.org/pandas-docs/stable/getting\\_started/10min.html](https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html)

Python Data Science Handbook (un livre en or) :

- Chapitres à aller voir -> 1 à 3 (surtout 2 et 3)  
<https://jakevdp.github.io/PythonDataScienceHandbook/>
- Les notebooks sont dispos ici  
<https://github.com/jakevdp/PythonDataScienceHandbook>

Python tutor (pour comprendre what's going on under the hood) :

- <http://pythontutor.com/visualize.html#mode=edit>

Pour les visuels :

- MIT Open Course Ware (pour ceux qui aiment les cours live -> long) :  
<https://www.youtube.com/user/MIT/search?query=python>
- Corey Schafer (Orienté python pur)  
<https://www.youtube.com/watch?v=YYXdXT2l-Gg&list=PL-osiE80TeTt2d9bFvVytXJA-UTHn6WwU>
- Data School (orienté data)  
<https://www.youtube.com/user/dataschool/videos>

Cheat Sheets :

- Python basics :  
[https://github.com/FavioVazquez/ds-cheatsheets/blob/master/Python/Datacamp/python\\_basics.pdf](https://github.com/FavioVazquez/ds-cheatsheets/blob/master/Python/Datacamp/python_basics.pdf)

- Pandas basics cheat sheet :  
<http://datacamp-community-prod.s3.amazonaws.com/dbed353d-2757-4617-8206-8767ab379ab3>
- Data Wrangling with pandas :  
[https://pandas.pydata.org/Pandas\\_Cheat\\_Sheet.pdf](https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)
- Pandas (more advanced) cheat sheet :  
<https://github.com/FavioVazquez/ds-cheatsheets/blob/master/Python/Datacamp/pandas.pdf>

C'est quoi une API ?

- <https://medium.com/@perrysetgo/what-exactly-is-an-api-69f36968a41f>

Ok, j'ai compris mais concrètement en python on fait comment ?

- Doc officielle du package *requests*  
<https://requests-fr.readthedocs.io/en/latest/user/quickstart.html>
- DataQuest :  
<https://www.dataquest.io/blog/python-api-tutorial/>