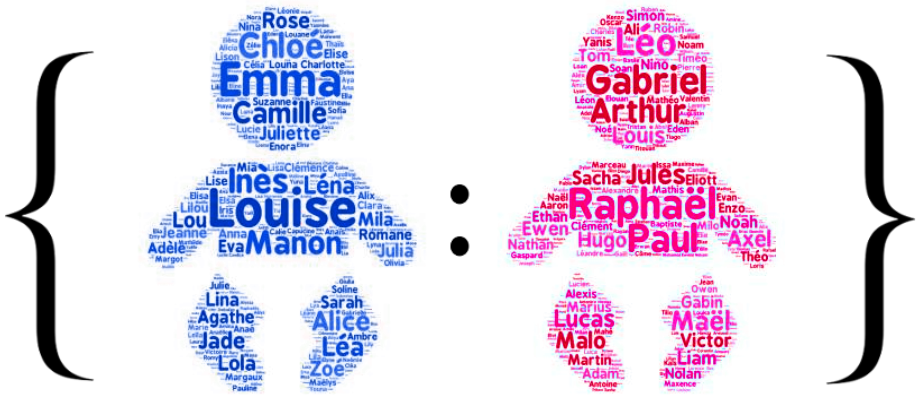
Projet Python 2 :

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** (**“ David Nadjar “**)

Objectifs du module

L’objectif de ce projet est de **stabiliser ce que vous avez appris jusqu’à présent en analyse de données avec Python**. Vous pourrez 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 nouvelle 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/ficher-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* 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/

Remarque importante !

Le but de ce projet est de vous donner un terrain de jeu pour aller expérimenter. Vous n'êtes pas tenus de respecter l'ordre des questions, ni même de les faire toutes.

Prenez le temps d'aller consulter les ressources fournies, elles sont longues mais ce sont des mines d'or. Essayez même de répondre aux questions de plusieurs manières différentes : en python pur, avec pandas, etc.

Pour résumer, no stress, enjoy the trip !

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 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 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 jupyter 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 :
  - Quelle est la proportion totale Femmes / Hommes au cours du temps depuis 1900 ?
  - Comment a-t-elle évolué au cours du temps ?
- On va maintenant essayer de retrouver (ou non!) quelques uns des résultats présentés dans l'article du Figaro (https://www.lefigaro.fr/fig-data/prenoms/):
  - Evolution du prénom Marie dans le temps (nbre de naissances de Marie par an) ?
    - **[Aide : essayez de faire ça de plusieurs manières différentes si vous le pouvez → sélecteurs pandas, list comprehension, pandas.query( ), boucles, etc.]**
  - En 2017, on recensait 13.000 prénoms différents, soit 7,6 fois plus qu'en 1900.

        - **[Aide : comment fait-on pour connaître les éléments uniques ? Pandas ? Python pur ?]**
    - " En 1900, parmi les prénoms recensés par l'Insee, le plus donné, Marie, représentait 11% des naissances."
    - " En 2017, des «prénoms rares» ont été donnés à près de 55 000 enfants, soit 10 fois plus que le prénom le plus donné (Gabriel). "
    - " 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, ….) ?
        - **[Aide : Aller voir comment traiter du texte en python → 'in' , regex, vectorized string functions de pandas, etc]**
    - " 1955-1960 : Les prénoms composés (Marie-Christine, Anne-Marie,...) prennent leur envol" ?
        - **[Aide : idem précédent]**
    - " 2015 : Les *prénoms rares* forment la première catégorie de prénoms "
    - **[Plus dur]** Combien de prénoms faut-il pour nommer 50% des bébés par année (e.g. en 1900 27 prénoms suffisent) ?
        - **[Aide : Aller voir "groupby" de pandas. Passez un peu de temps à vous familiariser avec grâce aux ressources. Regardez également comment trier des données (ordre croissant/ décroissant) et la fonction cumsum de pandas. Confusion garantie au 1er abord]**
    - La part des 10 prénoms les plus utilisés a-t-elle bien été divisée par 5 entre 1900 et aujourd'hui ?

Si vous ne savez pas par quel bout prendre ces questions, n'hésitez pas à demander au formateur. Egalement passez du temps à vous familiariser avec les ressources. Pas besoin d'aller très vite !

    - **[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), Maite (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".
    - Lisez cette base de données avec pandas (ou autre) et regardez là droit dans les yeux pour faire une idée de ce qu'il y a dedans.
        - **[Aide : Des problèmes pour lire le csv ? Allez voir le package chardet]**
    - Combien de prénoms sont analysés dans cette base ?
    - Donner les 10 prénoms les plus multiculturels recensés dans ce dataset.
        - **[Aide : allez voir les vectorized string functions de pandas, fillna( ) si nécessaire, explode( )]**

- Reprenez votre dataframe des prénoms et essayez de rajouter une colonne 'origine' dans votre dataframe des prénoms. Essayez de le faire de deux manières : en utilisant un mapping avec dictionnaire (voir ressource) et avec la fonction merge de pandas.
  - **[Aide : allez voir les vectorized string functions de pandas, fillna( ) si nécessaire, explode( ), merge de pandas, pensez à normaliser les deux bases de données sur le même format (majuscule, minuscule, etc.)]**
  - Ressources :
    - Mapping with a dictionary : **https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.map.html**
- 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 ?
  - **[Aide : Pensez à nettoyer et normaliser les deux bases de données sur le même format (doublons, majuscule, minuscule, etc.)]**
  - **[Aide : base de données de prénoms (doublons, majuscule, minuscule, etc.)**
- 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 Basques ? Bretons ? Catalans ?
- Quels sont les départements donnant le plus de prénoms multiculturels ?
- Donne-t-on plus de prénoms Espagnols ou Catalans près de la frontière espagnole? Italiens vers l'Italie? Etc?
- Tracez un histogramme décrivant l'origine de prénoms sur la France entière toutes périodes confondues.

Pour ceux qui s'ennuient :

Option #1 : Prédiction de date de naissance à partir du prénom

- Allez chercher le fichier décrivant l'espérance de vie en France depuis 1946, disponible ici : https://www.ined.fr/fr/tout-savoir-population/chiffres/france/mortalite-cause-deces/esperance-vie/ sous la rubrique " Téléchargements > Évolution et projection de l'espérance de vie à la naissance. Graphique Insee Première, n° 1320, 2010 " ou directement ici : https://www.ined.fr/fichier/s_rubrique/192/copie.de.ip1320.1.fr.xls
- Essayez de prédire l'âge d'une personne basé sur son prénom !
- Inspirez -vous de cette analyse : https://www.ekintzler.com/projects/age-prediction/

Option #2 : Origines manquantes :

- 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 supplémentaire (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! Vous ne pourrez pas tout récupérer…

Option # 3 : Le reste du monde

- 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
- 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.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

Tutoriel rapide sur pandas :

- https://pandas.pydata.org/docs/user_guide/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-osiE80TeTt2d9bfVyTiXJA-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
- 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
- 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 une API 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/