Module « Données » : les prénoms en France

Victor Boby, Anaïs Mazoué, Ariane Menu

Avril 2022

Le projet relatif aux prénoms en France vise à mettre en relation différents *datasets* grâce à des enrichissements et traitements réalisés sur ces derniers pour obtenir des visualisations de leurs données. Cette brève documentation a pour objectif de présenter tour à tour le projet et la méthode de travail suivie pour le mener à bien.

À travers ce projet, l'intention est d'établir des corrélations entre la fréquence d'utilisation d'un prénom et l'impact de différents facteurs au cours du temps. Pour y parvenir, nous nous sommes appuyés sur plusieurs *datasets*, le principal étant le *Fichier des prénoms de 1900 à 2020* réalisé par l'Institut National de la Statistique et des Études Économiques (Insee)[1] et disponible sur le catalogue des données publiques de l'administration. Les données correspondent ici aux prénoms donnés aux enfants nés en France entre 1900 et 2020. À chaque prénom sont associés l'année d'attribution, le sexe de l'enfant et le nombre d'enfants ayant été ainsi nommés au cours d'une année. Il y a alors une ligne par année et par prénom, ce qui représente un nombre important de données et par conséquent un fichier plutôt volumineux. Néanmoins, un autre set plus lourd encore est proposé par l'Insee : il s'agit des mêmes données mais à l'échelle des départements et non plus de la France dans son ensemble. Le nombre de données fournies est donc bien plus conséquent mais également plus riche et c'est pourquoi nous avons choisi celui-ci comme point de départ à certains de nos travaux.

Plusieurs pistes ont alors été envisagées pour compléter ce *dataset* et réaliser des visualisations. Après plusieurs idées qui n'ont finalement pas été retenues, nous avons décidé de débuter notre réflexion par une analyse de l'origine des prénoms, ici les origines biblique, anglaise, bretonne et occitane ont été choisies. Outre cela, nous avons également cherché à montrer dans quelle mesure une diversification dans le choix des prénoms des enfants pouvait être observée au cours du temps. Enfin, nous avons étudié la sur et sous représentation des prénoms au sein de la classe politique française des cent vingt dernières années.

[1] https://www.insee.fr/fr/statistiques/2540004?sommaire=4767262&q=pr%C3%A9noms

Chacune des approches choisies et brièvement décrites ci-dessus va désormais être documentée avec plus de précisions relatives à la récupération des données, à leur enrichissement et à la réalisation de visualisations grâce aux outils Dataiku et Tableau.

I. Visualiser les tendances culturelles des prénoms

Ce thème avait pour but de rendre compte des tendances culturelles ayant traversé les prénoms de 1900 à nos jours. Il a fallu pour ce faire trouver des données relatives aux origines culturelles des prénoms.

A. Compléter le jeu de données initial

Trouver des données d'enrichissement ne fut pas une tâche aisée, de nombreuses pistes s'étant avérées sans issue. Wikidata ne fut ainsi d'aucune aide, la *Property* P735 (*Given name*) ne comprenant pas d'attribut relatif à l'origine du prénom. Si nous pouvons parfois retrouver les attributs "*native label*", "*writing system*" ou encore "*language of work or name*", ils ne furent cependant pas exploitables[2]. D'autres pistes furent explorées, mais ou bien les sites trouvés ne mettaient pas à disposition leurs données (*Namepedia.org*), ou bien leurs informations manquaient de pertinence (par exemple *Nationalize.io*[3]). *BehindTheName.com* semblait prometteur, mais dispose d'une API limitée (en plus d'avoir des droits de réutilisation flous quant à notre projet).

Ayant à coeur de réussir à exploiter ce thème, il fut finalement choisi d'utiliser quatre pages Wikipédia recensant des prénoms : la liste de prénoms anglais, la liste des personnages de la Bible (plus longue que celle des prénoms hébraïques), la liste de prénoms bretons et la liste de prénoms occitans. Nous étudions ainsi deux tendances régionales, une tendance étrangère et une tendance religieuse. Notons cependant que de nombreux prénoms sont communs à plusieurs listes[4] (à l'instar du prénom "Aaron", recensé comme prénom anglais,

[2] "*native label*" se contenant souvent d'un "[prénom] (*multiple languages*)" générique, et "*writing system*" réalisant de trop larges regroupements ("*latin script*" par exemple). Un test fut mené en faisant une jointure avec des données portant sur l'attribut "*language of work or name*" via une requête SPARQL, mais les résultats ne furent pas probants, les prénoms ayant la valeur "néerlandais" dépassant systématiquement et de loin tous les autres.

[3] Avec ici un requêtage de l'API sur le prénom "Axel".

[4] Ceci ayant pour effet d'aplatir les résultats, voir la visualisation 3 par exemple, où les mêmes départements ont parfois des résultats anormalement similaires.

biblique et breton), tandis que la pertinence de certains pourrait laisser à désirer (par exemple le prénom “Eric”, recensé dans la liste occitane).

B. Préparer les données avec Dataiku

- **Préparation du jeu d’origine**

En vue de notamment réaliser des visualisations basées sur le département d’attribution des prénoms, il fut décidé de travailler avec le set le plus lourd, celui des départements. Le jeu fut préparé en retirant les données superflues ; les “XXXX” des colonnes “annais” (années) et les “XX” des départements furent ainsi remplacés par des valeurs NULL afin de permettre la conversion en format “*integer*”. Dans la colonne relative au sexe de l’enfant, les “1” et “2” furent rempalcés par “Masculin” et “Féminin”. Enfin, les valeurs “_PRENOMS_RARES” de la colonne des prénoms furent transformées en “Prénoms rares” pour plus de lisibilité. Ces valeurs furent donc nettoyées mais non supprimées, étant importantes pour avoir le plus d’informations relatives à la distribution des prénoms. Les diacritiques furent volontairement gardées car souvent marqueuses de l’origine d’un prénom[5].

- **Préparation des jeux secondaires**

Le contenu des listes fut importé dans Dataiku sous la forme de quatre documents csv. Les listes firent chacune l’objet d’une recette de préparation afin de nettoyer les données. Les prénoms ont pu être extraits à l’aide de commandes “*find and replace*” utilisant notamment des expressions régulières ainsi que des tokenisations[6]. Chaque liste se vit ajouter une colonne de valeur booléenne “*true*” en vue de leur agrégation prochaine, avant de subir une recette “*Distinct*” afin de retirer les prénoms doublons.

Les quatre jeux furent simultanément joints au set d’origine via un *Left join*, la condition étant le prénom[7] (insensible à la casse) ; la colonne “prénom” des jeux secondaires fut systématiquement retirée afin de ne pas avoir de colonnes doublon. Ne reste ainsi que le jeu de base enrichi de quatre colonnes booléennes indiquant par un “*true*” l’appartenance d’un prénom à une tendance culturelle.

[5] Les prénoms anglophones n’ont par exemple pas d’accent, tandis que les prénoms occitans en ont fréquemment.
[6] Les prénoms bretons posant alors problème en raison de leurs nombreuses apostrophes, ils furent finalement récupérés sur le site Observable, ayant déjà isolé les prénoms de la liste Wikipédia.
[7] Les quelques mots parasites rescapés issus des listes purent ainsi être éliminés.

Nous arrivons ainsi au *workflow* suivant (Fig. 1) :

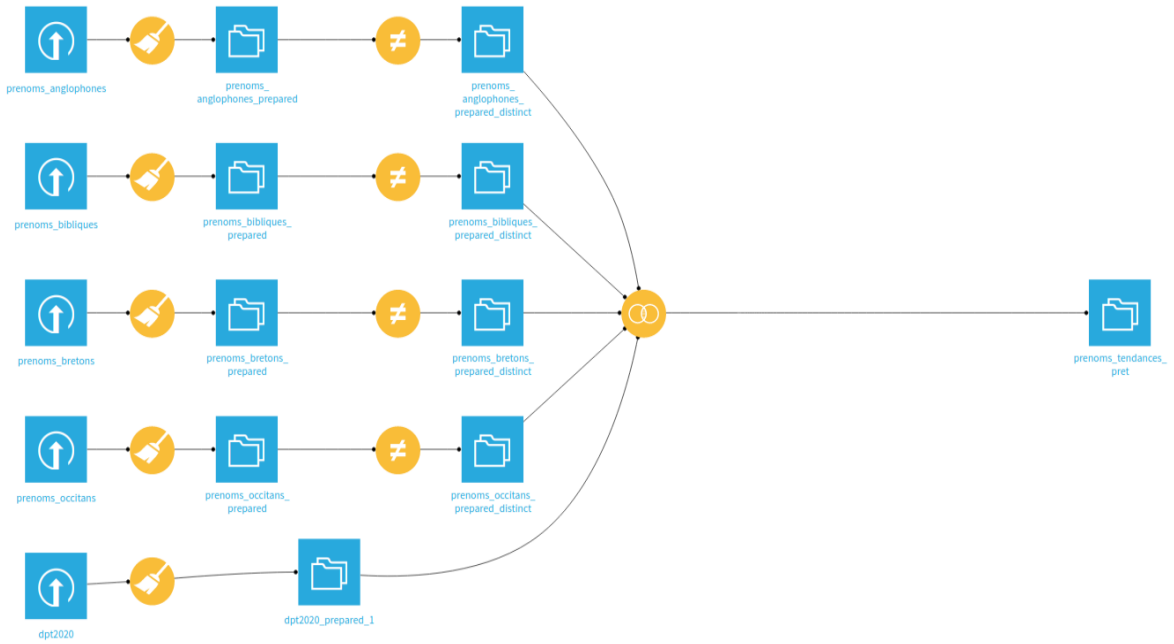


Figure 1 : Workflow de la préparation des données avec Dataiku

C. Visualiser les données avec Tableau

La visualisation comprend trois tableaux de bord, chacun composés de plusieurs dataviz.

La première visualisation (Fig. 2) est la plus généraliste et se pose ainsi en introduction. Elle présente l'évolution des quatre tendances étudiées au fil du temps et au regard du reste des prénoms de la population. En plus de rendre compte de creux flagrants lors des deux guerres mondiales – et du baby boom – , cette visualisation illustre notamment l'importance des prénoms tirés de la Bible, qui s'amenuise cependant à partir de la deuxième moitié du XXe siècle.

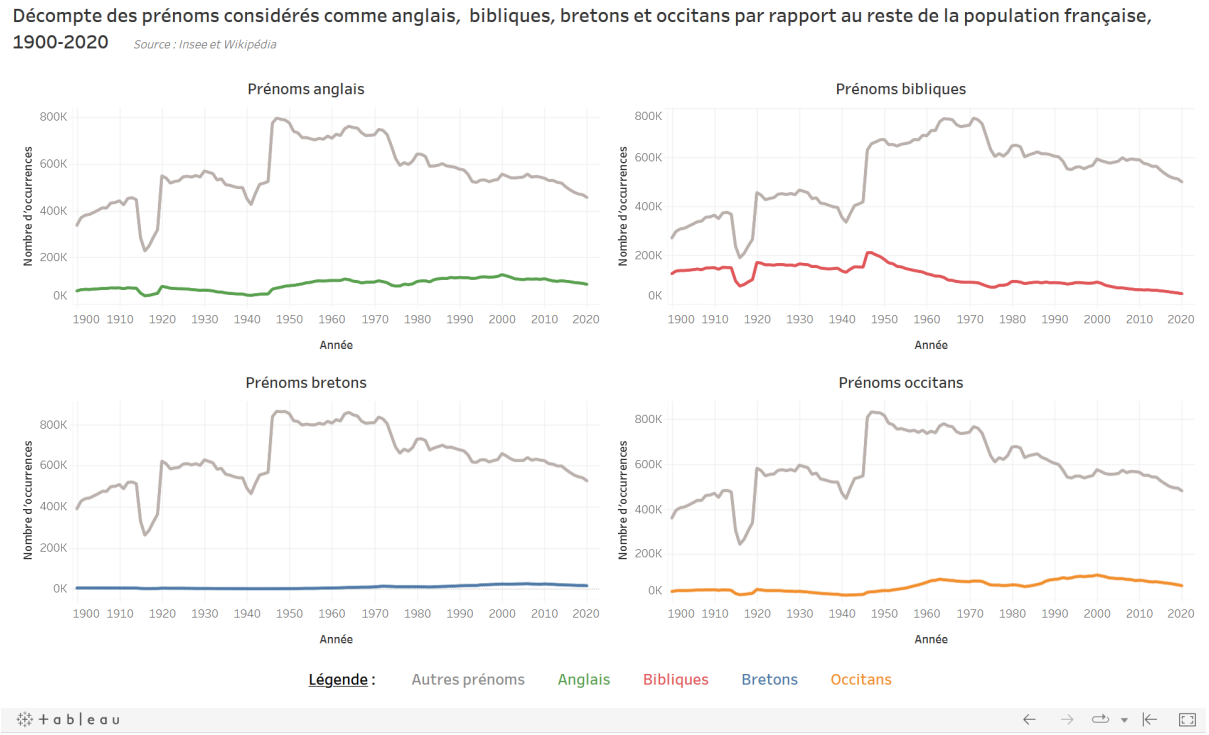


Figure 2 : Visualisation 1 ; Décompte des prénoms considérés comme anglais, bibliques, bretons et occitans par rapport au reste de la population française, 1900-2020 (Source : Insee et Wikipédia)

S'il eut été intéressant de proposer une visualisation empilant les quatre tendances afin de mieux les comparer, cela fut impossible, Tableau agrégeant systématiquement les valeurs booléennes entre elles[8].

La deuxième visualisation (Fig. 3) rend compte de la distribution des prénoms les plus populaires selon le sexe de l'enfant. Pour plus de lisibilité, seuls les prénoms représentant au moins 2% de chaque corpus sont représentés, le reste des valeurs étant regroupé sous l'appellation "autre". Cette représentation permet de se rendre compte du poids écrasant de certains prénoms, à l'instar de "Marie" dans la catégorie biblique.

[8] Ainsi, Tableau génère une courbe par combinaison possible ({True, True, False, True}, {False, True, False, True}, {False, True, True, False} etc). Les courbes non désirées sont impossibles à retirer, voir le forum Tableau pour plus d'informations.

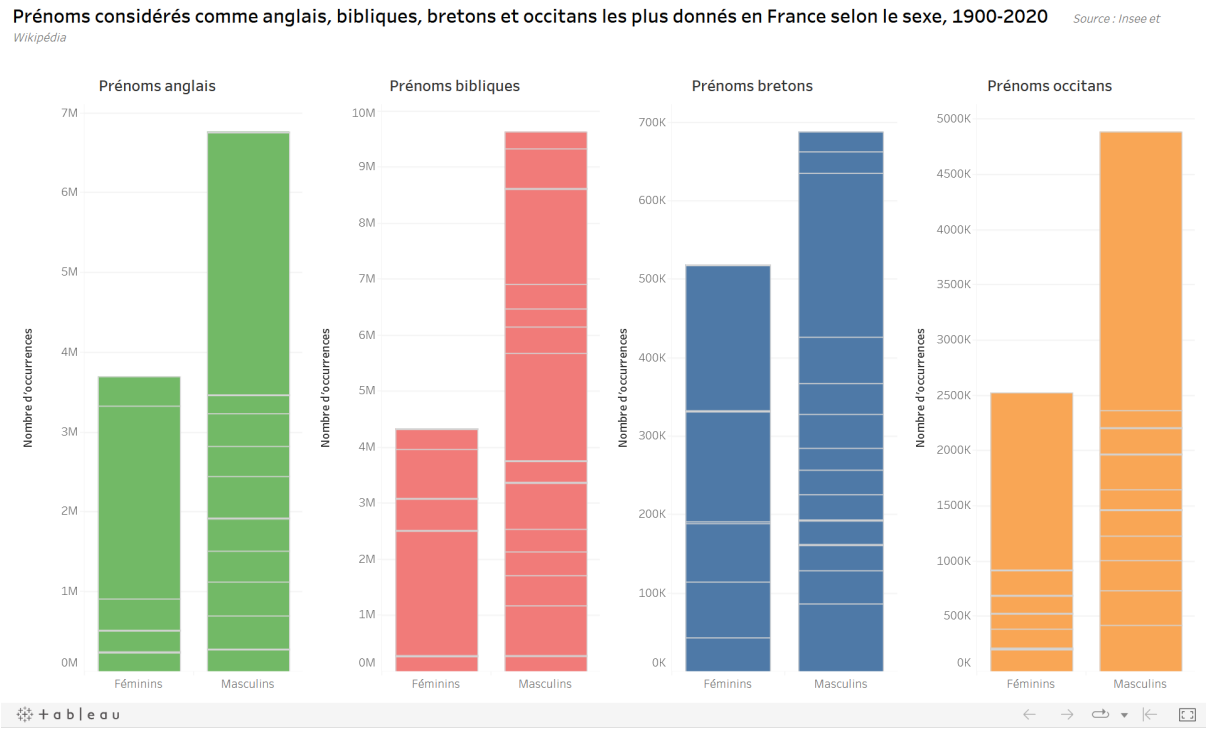


Figure 3 : Visualisation 2 ; Prénoms considérés comme anglais, bibliques, bretons et occitans les plus donnés en France selon le sexe, 1900-2020 (Source : Insee et Wikipédia)

Enfin, la troisième visualisation (Fig. 4) exploite les données relatives aux départements de notre jeu de départ en présentant la distribution des prénoms sur le territoire au fil du temps[9]. La visualisation est interactive, l'utilisateur pouvant régler l'année grâce à un curseur localisé en haut au milieu. Si la représentation est efficace pour les prénoms bretons, nous voyons qu'elle est moins pertinente pour les prénoms occitans en raison de la qualité des valeurs récupérées sur Wikipédia, problème déjà évoqué en introduction de cette partie.

[9] La Corse a toutefois dû être écartée, le département "20" n'existant pas et n'étant en conséquence pas reconnu par Tableau.

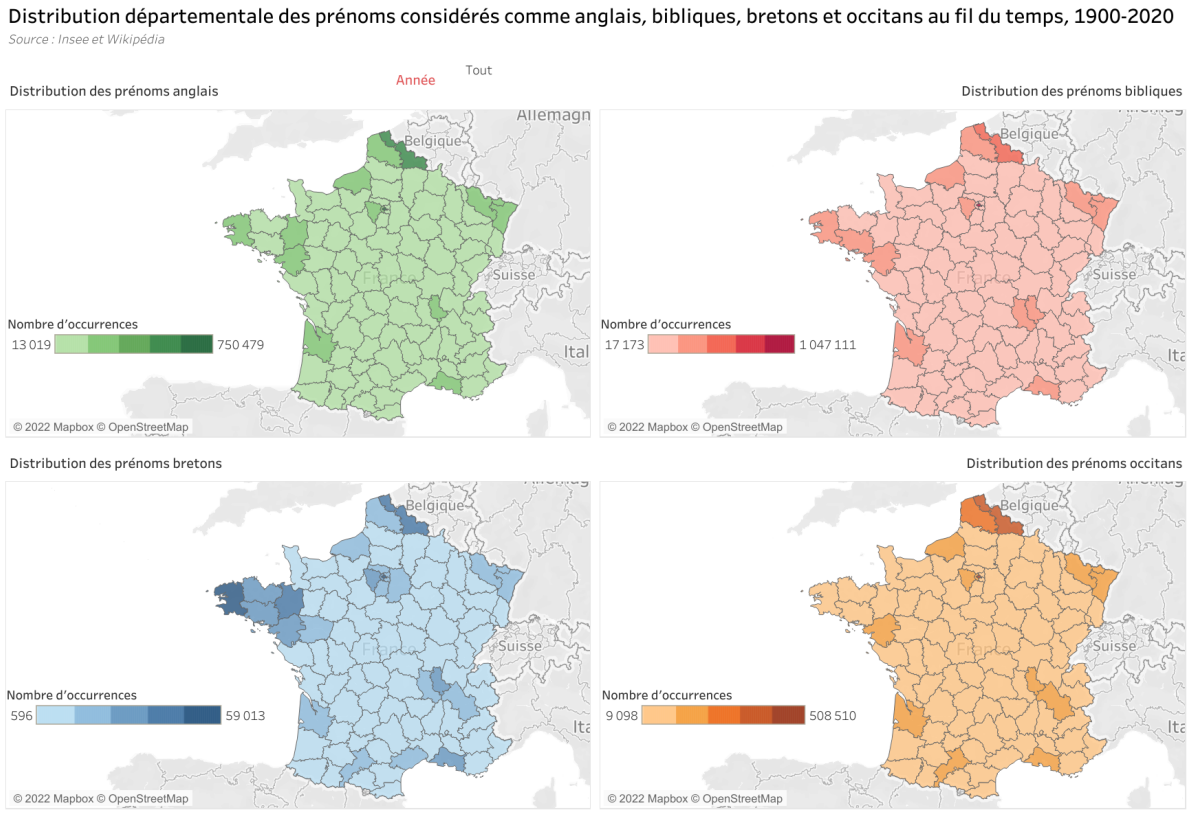


Figure 4 : Visualisation 3 ; Distribution départementale des prénoms considérés comme anglais, bibliques, bretons et occitans au fil du temps, 1900-2020 (Source : Insee et Wikipédia)

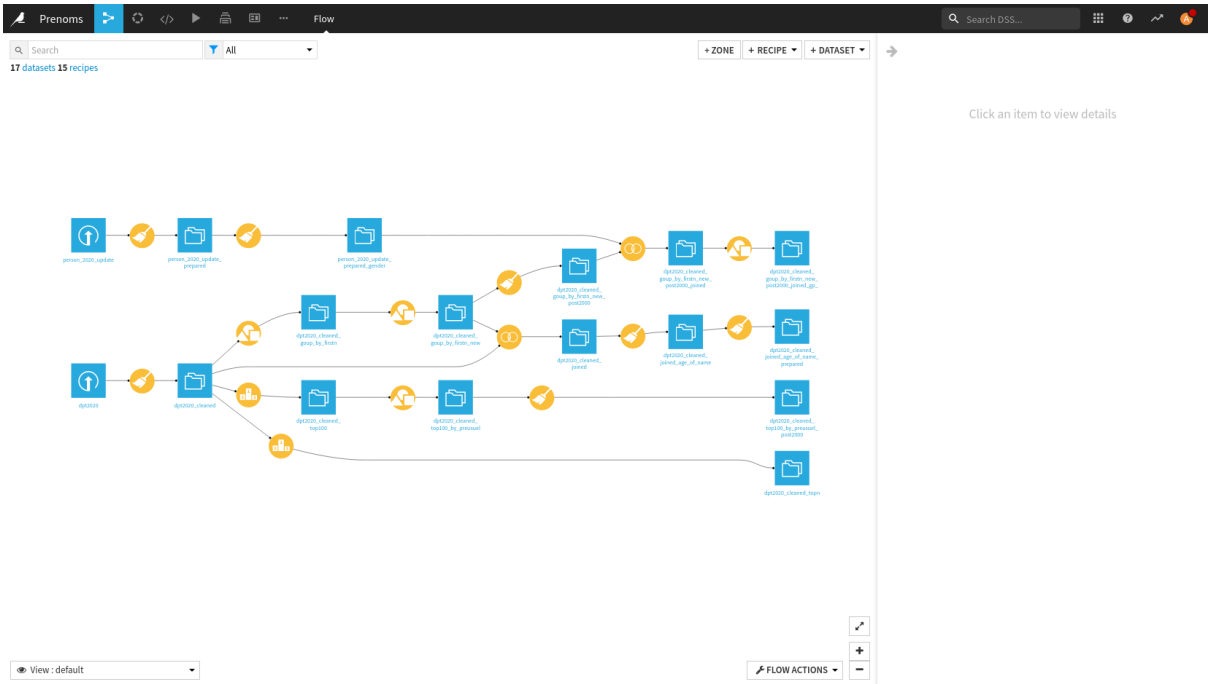
II. Vers une diversification des prénoms:

A. Dataiku:

- **Préparation du jeu d'origine**

803 lignes (sur 3,7 millions) pour lesquelles des données n'étaient pas exploitables ont été supprimées. Il s'agit notamment d'enregistrements dont le champ année de naissance (ANNAIS) prenait la valeur «XXXX» et les départements de naissance (DPT) codés en «XX» pour certains cas de prénoms rares (voir "Conditions portant sur les prénoms retenus" sur le site de l'INSEE).

Les valeurs «_PRENOMS_RARES_» du champ prénom (PREUSUEL) ont été conservées pour les visualisations obtenues exclusivement depuis ce dataset, afin de ne pas perdre cette notion de l'angle adopté sur la diversité.



Pour les 10 prénoms les plus donnés par année depuis 1900 : les 10 premiers prénoms par département par année ont été filtrés dans Dataiku (recette TopN).

Concernant l'évolution du nombre de prénoms par département depuis 1900 : le nombre de prénoms par département et par année est obtenu directement dans Tableau avec les fonctions de calcul.

D'autres pistes ont été explorées sur l'ancienneté des prénoms (détermination de la première apparition du prénom) mais n'ont pas pu être finalisées.

B. Tableau

Les data visualisations du Top10 des prénoms et l'accroissement du nombre de prénoms différents illustrent leur diversité toujours croissante, qu'il serait intéressant d'étudier plus profondément, par exemple en étudiant la proportion de prénoms récents (<= 5, 10 ans) parmi les prénoms de ces dernières décennies. Ainsi considérés ensembles, ces prénoms rivalisent en volume avec les Marie et Jean du début et du milieu du XXe siècle.

Vous pourrez retrouver ces DataViz ici :

- Top 10 des prénoms les plus donnés par année depuis 1900 : https://public.tableau.com/app/profile/victor.boby/viz/Top10Prnomspardpartementetparanne/Feuille2#1

- Evolution du nombre de prénoms par département depuis 1900 : https://public.tableau.com/app/profile/victor.boby/viz/Diversitdesprnomsdepuis1900/Prnomsdiffrentspardpartementdepuis1900

III. Visualiser la “diversité” des prénoms des personnalités politiques en France

Le dernier thème porte sur un cas d’étude visant à rendre compte de la diversité des prénoms au sein d’une profession. En cette période d'élection présidentielle, le métier de politique constituait un sujet idéal.

A. Compléter le jeu de données initial

Le jeu d'origine fut enrichi des données de Wikidata, récupérées depuis son SPARQL endpoint avec la requête suivante :

```
SELECT ?prenomLabel ?sexeLabel
WHERE {
  #On cherche des êtres humains
  ?personne wdt:P31 wd:Q5.
  #De nationalité française
  ?personne wdt:P27 wd:Q142 .
  #Qui sont des personnalités politiques
  ?personne wdt:P106 wd:Q82955 .
  #On récupère leur prénom
  ?personne wdt:P735 ?prenom.
  #On récupère leur sexe ou genre
  ?personne wdt:P21 ?sexe .

  #On récupère leur date de naissance
  ?personne wdt:P569 ?naissance .
  #On ne garde que les personnes nées depuis 1900
  FILTER (YEAR(?naissance) >= 1900)

  #On affiche les labels en français
  SERVICE wikibase:label {
      bd:serviceParam wikibase:language "fr" .
  }
}
```

Fut ainsi récupérée une liste des prénoms et sexe de plus de 61 000 personnalités politiques.

B. Préparer les données avec Dataiku

- **Préparation du jeu d'origine**

La dataviz prévue ne faisant pas usage des données géographiques, il fut cette fois décidé de travailler avec le plus léger des *datasets* primaires, celui d'échelle nationale. La colonne des prénoms fut harmonisée en remplaçant les valeurs "_PRENOMS_RARES" par "Prénoms rares" et en supprimant les diacritiques. Comme précédemment, les valeurs "1" et "2" de la colonne renseignant le sexe de la personne furent transformées en "Masculin" et "Féminin". Notre souhait étant de comparer les données additionnées sur toute la période, la colonne des années fut estimée superflue et fut supprimée.

Le nombre d'occurrences de chaque prénom fut ensuite additionné par prénom et par sexe grâce à une recette "*group*".

- **Préparation du jeu secondaire**

Les données issues de la requête SPARQL furent triées alphabétiquement pour plus de lisibilité avant d'être nettoyées. Les valeurs "masculin" et "féminin" furent capitalisées pour s'aligner sur les données du jeu primaire, tandis que l'unique valeur "femme trans" fut transformée en "Féminin". Toujours dans un but d'harmonisation avec le jeu primaire, les prénoms des politiques virent leurs diacritiques supprimées. Une colonne fut créée pour compter chaque personnalité politique.

Cette colonne permit à l'étape suivante d'additionner le nombre d'occurrences de chaque prénom (par prénom et par sexe) via une recette "*group*".

Enfin, les jeux primaire et secondaire purent être associés avec un *Left join* sur les prénoms ; une dernière recette de préparation fut créée pour réordonner les colonnes et transformer la colonne prénom du jeu secondaire, doublon, en "*true*".

Nous arrivons alors au *workflow* suivant (Fig. 5) :

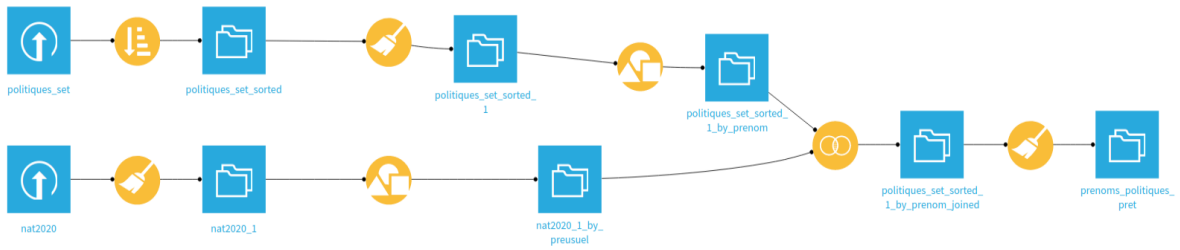


Figure 5 : Workflow de la préparation des données avec Dataiku

C. Visualiser les données avec Tableau

Cette dernière visualisation put être créée en calculant le pourcentage d'occurrences des prénoms au sein des politiques et de la population. Elle permet de rendre compte de la sur ou sous représentation des prénoms au sein de la classe politique française de ces cent vingt dernières années. La démarcation est rendue plus visible grâce à la création d'une courbe de type "x = y". Chaque prénom se trouvant sous la courbe est sur représenté, tandis que chaque prénom placé au-dessus d'elle est en sous-effectif. Nous constatons avec la visualisation qu'une poignée de prénoms masculins accapare nettement la majeure partie de l'espace politique français.

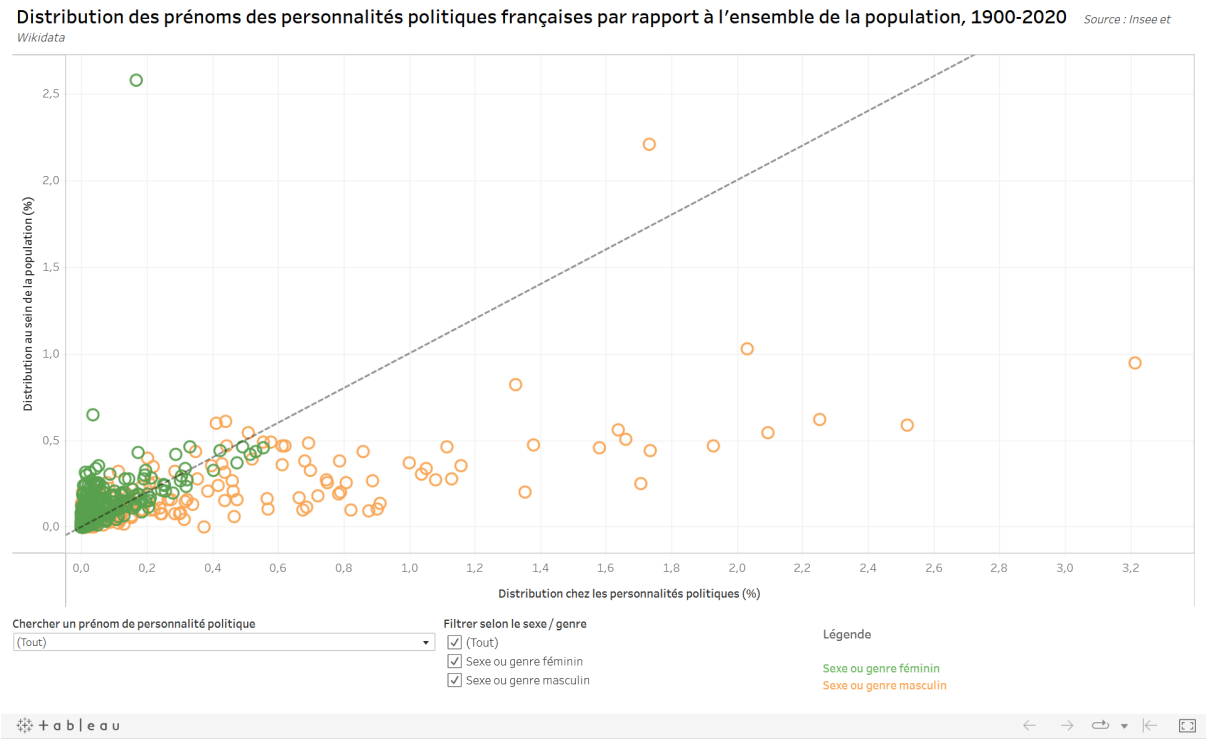


Figure 6 : Distribution des prénoms des personnalités politiques françaises par rapport à l'ensemble de la population, 1900-2020 (Source : Insee et Wikidata)

Pour terminer, notons que des filtres sur les prénoms et le sexe des politiques permettent à l'utilisateur d'affiner les résultats.

Conclusion

Les visualisations réalisées grâce à Tableau après les différents traitements appliqués sur Dataiku ont donc permis de mettre en évidence certaines tendances dans l'attribution des prénoms en France depuis 1900. Ainsi, les prénoms d'origine biblique, majoritaires au début du XXe siècle, font place à des prénoms récents ou rares dont la proportion ne fait qu'augmenter depuis le dernier quart du XXe siècle. Par ailleurs, l'apparition de ces prénoms rares ne se fait pas au même moment dans l'ensemble des départements mais progressivement entre les années 1970 et les années 1990. La comparaison d'une seule catégorie de personnes, ici les politiques, avec l'ensemble de la population française, permet aussi de mettre en lumière des divergences comme la surreprésentation de certains prénoms, masculins pour la plupart, au sein d'un groupe précis. Des constantes s'observent aussi : par exemple, les prénoms d'origine bretonne sont en immense majorité attribués en Bretagne.

Ainsi, ces quelques éléments peuvent ouvrir la voie à des réflexions plus larges sur l'impact et les évolutions du contexte socio-culturel sur l'attribution du prénom d'un enfant. Pour aller plus loin, il pourrait être envisagé d'étudier l'influence d'une œuvre de fiction et de sa diffusion sur l'évolution des prénoms en France.