Manipulation et prétraitement de données

Analyser un jeu de données réel sur les thèses en France

Ibrahima LY

29 AVRIL 2023

Résumé

Le traitement et la préparation des données sont des étapes cruciales dans toute analyse de données. En effet, elles peuvent représenter jusqu'à la moitié du travail de l'analyste. La première partie de cette analyse consiste à manipuler et analyser le jeu de données en effectuant des tâches telles que l'importation de données, la réalisation de premières visualisations et l'identification de données manquantes, de problèmes et d'outliers. La seconde partie est consacrée à la visualisation des données pour faciliter leur interprétation. En combinant ces deux parties, nous pourrons obtenir des informations significatives et pertinentes à partir des données analysées.

Table des matières

1 Introduction

L'analyse de données constitue un outil puissant pour comprendre les tendances et les évolutions dans divers domaines. Dans ce rapport, nous allons nous pencher sur un ensemble de données portant sur les thèses soutenues en France.

Le jeu de données contient des informations sur plus de 448 000 thèses, dont les titres, les noms des auteurs, les directeurs de thèse, les établissements de soutenance, les disciplines, les dates de soutenance, les langues de la thèse, entre autres. Les données ont été collectées à partir de la plateforme Theses.fr et couvrent une période allant de 1964 à 2021.

Nous allons examiner les différentes caractéristiques des thèses et des auteurs, ainsi que les tendances et les évolutions dans les disciplines, les langues de la thèse et les établissements de soutenance. Nous allons également utiliser des techniques d'analyse de données pour découvrir des relations cachées entre les différentes variables du jeu de données.

Notre objectif est de fournir une analyse approfondie et des informations utiles pour les chercheurs, les institutions d'enseignement et les décideurs dans le domaine de la recherche.

2 Présentation des données

2.1 Présentation de jeu de données PhD1

Chargement des librairies

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import missingno as msno
import datetime
import calendar
import math
import warnings
warnings.filterwarnings("ignore")
```

```
PhD_v1 .info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 447644 entries, 0 to 447643
Data columns (total 18 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   Auteur                                  447644 non-null  object
 1   Identifiant auteur                      317655 non-null  object
 2   Titre                                   447635 non-null  object
 3   Directeur de these                      447629 non-null  object
 4   Directeur de these (nom prenom)         447629 non-null  object
 5   Identifiant directeur                   447644 non-null  object
 6   Etablissement de soutenance             447640 non-null  object
 7   Identifiant etablissement               430559 non-null  object
 8   Discipline                              447639 non-null  object
 9   Statut                                  447644 non-null  object
 10  Date de premiere inscription en doctorat  63976 non-null   object
 11  Date de soutenance                      390898 non-null  object
 12  Year                                    390898 non-null  float64
 13  Langue de la these                      383879 non-null  object
 14  Identifiant de la these                 447644 non-null  object
 15  Accessible en ligne                     447644 non-null  object
 16  Publication dans theses.fr              447644 non-null  object
 17  Mise a jour dans theses.fr              447467 non-null  object
dtypes: float64(1), object(17)
memory usage: 61.5+ MB
```

Le nombre de lignes dans le jeu de données PhDv1 est : 447644 lignes et 18 colonnes

2.2 Présentation du jeu de données PhD2

Le nombre de lignes du jeu de données PhD2 est : 448047 lignes et 22 colonnes.

Les différentes colonnes du dataFrame

- Auteur
- Identifiant auteur
- Titre
- Directeur de these
- Les critères de choix de nouveaux jeux vidéo
- Directeur de these (nom prenom)
- Identifiant directeur

— Etablissement de soutenance
— Identifiant etablissement
— Discipline
— Statut
— Date de premiere inscription en doctorat
— Date de soutenance
— Year
— Langue de la these
— Identifiant de la these
— Accessible en ligne
— Publication dans theses.fr
— Mise a jour dans theses.fr
— Disciplinepredi
— Genre
— etablissementrec
— LangueR

Typologie des variables.

```
Auteur                                  object
Identifiant auteur                      object
Titre                                   object
Directeur de these                      object
Directeur de these (nom prenom)         object
Identifiant directeur                   object
Etablissement de soutenance             object
Identifiant etablissement               object
Discipline                              object
Statut                                  object
Date de premiere inscription en doctorat object
Date de soutenance                      object
Year                                   float64
Langue de la these                      object
Identifiant de la these                 object
Accessible en ligne                     object
Publication dans theses.fr              object
Mise a jour dans theses.fr              object
Discipline_predi                        object
Genre                                   object
etablissement_rec                       object
Langue_rec                              object
dtype: object
```

On constate que le jeu de données PhD2 contient principalement des variables de type object. On constate aussi que les variables Year, Date de premiere inscription en doctorat et Date de soutenance sont respectivement de type float et object. Enfin, il y a plusieurs variables de type object qui contiennent des identifiants tels que : Identifiant auteur, Identifiant directeur et Identifiant etablissement.

Nombre de valeurs non vide :

```
Date de premiere inscription en doctorat     64331
Identifiant auteur                          317700
Langue_rec                                  383927
Year                                        390961
Date de soutenance                          390961
Identifiant etablissement                   430965
etablissement_rec                           444973
Mise a jour dans theses.fr                  447870
Directeur de these                          448034
Directeur de these (nom prenom)             448034
Titre                                       448040
Etablissement de soutenance                 448046
Statut                                      448047
Identifiant directeur                       448047
Langue de la these                          448047
Identifiant de la these                     448047
Accessible en ligne                         448047
Publication dans theses.fr                  448047
Discipline_predi                            448047
Genre                                       448047
Discipline                                  448047
Auteur                                      448047
dtype: int64
```

On peut voir que certaines colonnes ont un nombre important de valeurs manquantes, telles que "Date de première inscription en doctorat" qui n'a que 64 331 valeurs non nulles sur un total de 448 047.D'autres colonnes comme "Langue de la these" ont également un grand nombre de valeurs manquantes.

Cela peut poser des problèmes lors de l'analyse des données, car les observations manquantes peuvent fausser les résultats des analyses. Il peut donc être important de prendre en compte ces valeurs manquantes lors de l'analyse et de les gérer de manière efficace et appropriée.

Statistiques descriptives :

	Year	Annee_premiere_inscription	Annee_de_soutenance
count	390961	0	0
unique	44	0	0
top	2012-01-01 00:00:00	NaN	NaN
freq	13991	NaN	NaN
first	1971-01-01 00:00:00	NaN	NaN
last	2020-01-01 00:00:00	NaN	NaN

On constate que la variable "Year" ne présente qu'une seule valeur unique qui est 1970, cela signifie qu'il n'y a pas de variation dans les années de soutenance de doctorat dans les données analysées.

De plus, la variable "Year" ne contient qu'une seule valeur unique, à savoir

"1970", ce qui est étrange et peut être considéré comme une anomalie ou une incohérence dans les données, surtout si la variable représente l'année d'obtention du doctorat.

la variable "Date de soutenance" contient des valeurs allant jusqu'à l'année 2 072, ce qui semble peu probable et peut également être considéré comme une anomalie à vérifier.

En ce qui concerne les variables "Date de premiere inscription en doctorat" et "Date de soutenance", le nombre de valeurs uniques pour chaque variable est assez élevé, ce qui peut indiquer une certaine variabilité dans les dates. Cependant, la fréquence de la valeur la plus courante dans la variable "Date de premiere inscription en doctorat" est assez faible par rapport au nombre total d'observations, ce qui peut indiquer une grande variabilité dans la date d'inscription. Par ailleurs, la date de soutenance la plus courante est en 1994.

Enfin, le fait que la première et la dernière date pour la variable "Year" soit identique à 1970 peut indiquer que les données ont été collectées à partir d'un certain point dans le temps, probablement après 1970.

D'après ces statistiques descriptives, on peut constater que :

Summary Objet	**Total/ fréquence**
Le nombre total de thèses recensées est de	448 047
Le nombre d'auteurs uniques est de	64
La majorité des thèses qui ont été soutenues est de	381 360
La langue de la majorité des thèses est le français	
Le directeur de thèse le plus fréquent est	Directeur de thèse inconnu
L'établissement de soutenance le plus fréquent est	Paris 6
La discipline la plus fréquente est	Médecine
La discipline la plus prédite pour ces thèses est	Biologie
L'établissement de rattachement le plus fréquent est	Sorbonne Université
La langue la plus fréquente pour les résumés est	le français
Le genre le plus représenté est	masculin

En analysant rapidement les stats descriptives, on peut remarquer qu'il y a des valeurs manquantes (par exemple, l'identifiant de l'établissement manque dans près de 17 000 observations, ce qui pourrait être un problème si l'identification des auteurs est importante pour notre analyse.) et des doublons dans certaines variables (par exemple, il y a 16 auteurs qui ont soumis plus d'une thèse). Il est important d'évaluer l'impact de ces anomalies sur l'analyse avant

de procéder à une interprétation des résultats.¶

De plus, il y a des valeurs étranges dans la colonne "Titre", avec 17 titres identifiés comme "NAME ?". Il serait donc important de vérifier la qualité des données dans chaque colonne et de s'assurer que les valeurs sont cohérentes avec l'objectif de notre analyse.

3 Données manquantes

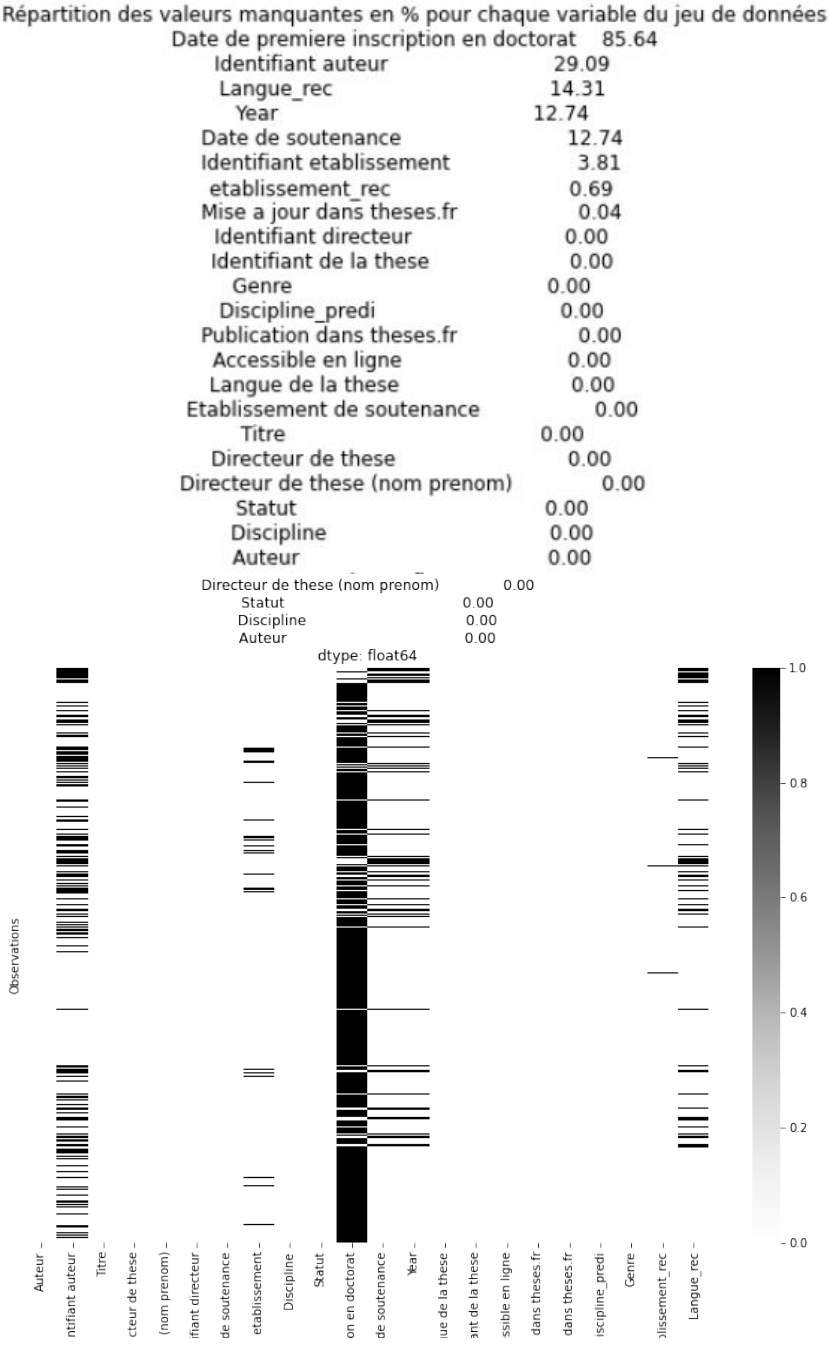
```
Date de premiere inscription en doctorat    383716
Identifiant auteur                          130347
Langue_rec                                   64120
Year                                         57086
Date de soutenance                           57086
Identifiant etablissement                    17082
etablissement_rec                             3074
Mise a jour dans theses.fr                     177
Directeur de these (nom prenom)                 13
Directeur de these                              13
Titre                                            7
Etablissement de soutenance                      1
Identifiant directeur                            0
Publication dans theses.fr                       0
Genre                                            0
Discipline_predi                                 0
Langue de la these                               0
Accessible en ligne                              0
Identifiant de la these                          0
Statut                                           0
Discipline                                       0
Auteur                                           0
dtype: int64
```

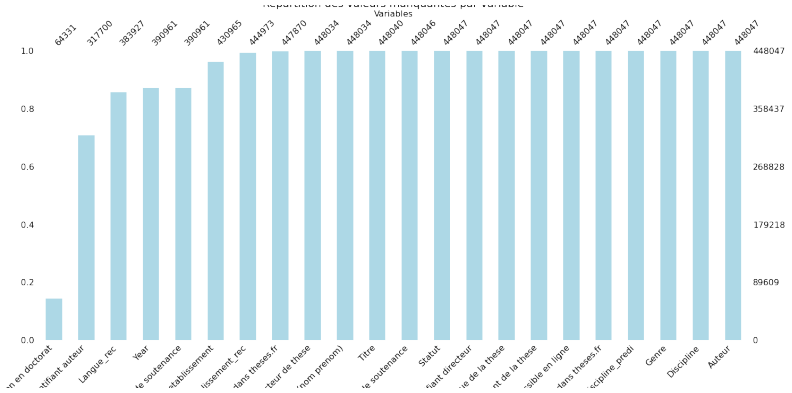
3.1 Figure 1 : Répartition des valeurs manquantes en % pour chaque variable du jeu de données PhD2

En analysant la répartition des données manquantes, on peut voire que les variables Date de première inscription en doctorat, Date de soutenance, Year, Identifiant auteur, Identifiant etablissement, etablissementRec et LangueRec ont un grand nombre de valeurs manquantes. En observant la heat map, on peut constater que pour les variables "Date de première inscription en doctorat" et Date de soutenance, les valeurs manquantes sont plus fréquentes pour les thèses en cours que pour les thèses soutenues. Cela peut être expliqué par le fait que

les thèses en cours n'ont pas encore atteint leur date de soutenance.

Figure 1: Répartition des valeurs manquantes en % pour chaque variable du jeu de données PhD_v2:

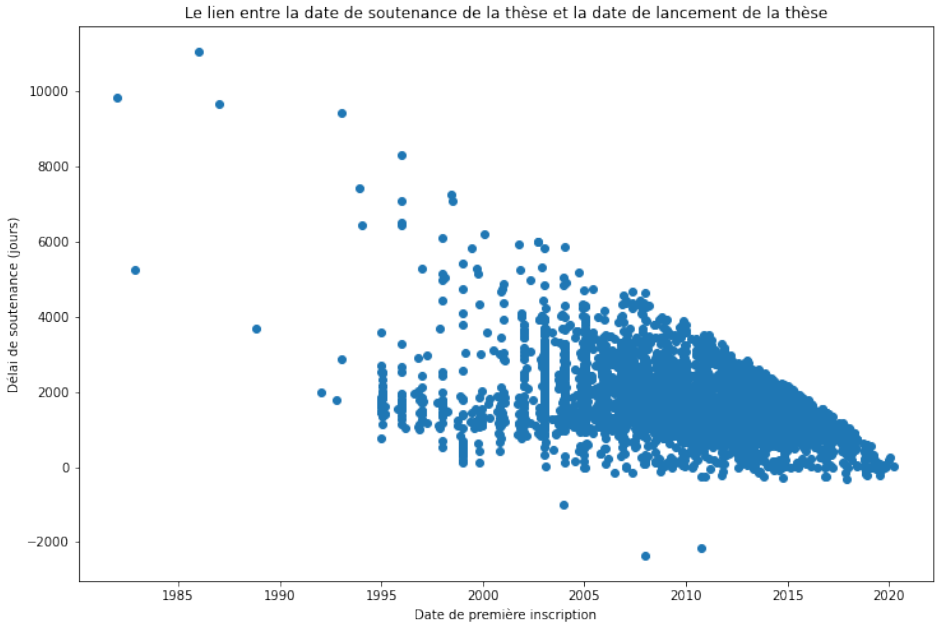




3.2 Fig2 : Lien entre le délai de soutenance et la date de première inscription

On constate sur ce graphique une certaine corrélation négative entre la date de lancement de la thèse et le délai de soutenance. En effet, on peut remarquer une tendance à une augmentation du délai de soutenance pour les thèses inscrites plus tard. Cela peut s'expliquer par plusieurs facteurs, comme :

- Des changements de direction
- De sujets de recherche
- Des contraintes personnelles des doctorants
- Etc...



3.3 La matrice de correlation

4 Qualité des données : Détection et traitement des anomalies

4.1 Gestion des doublons

```
1 duplicates = PhD_v2_copy.duplicated()
2 print(duplicates)
```

```
0         False
1         False
2         False
3         False
4         False
          ...
448042    False
448043    False
448044    False
448045     True
448046    False
Length: 448047, dtype: bool
```

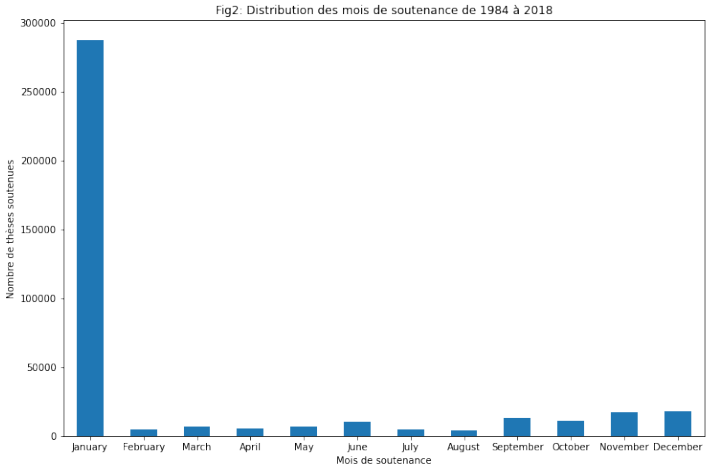
```
1 num_duplicates = duplicates.sum()
2 print(f"Le jeu de données contient {num_duplicates} doublons.")
```

```
Le jeu de données contient 412 doublons.
```

Suppression des doublons :

```
Auteur                                    0.0
Identifiant auteur                        0.0
Titre                                     0.0
Directeur de these                        0.0
Directeur de these (nom prenom)           0.0
Identifiant directeur                     0.0
Etablissement de soutenance               0.0
Identifiant etablissement                 0.0
Discipline                                0.0
Statut                                    0.0
Date de premiere inscription en doctorat  0.0
Date de soutenance                        0.0
Year                                      0.0
Langue de la these                        0.0
Identifiant de la these                   0.0
Accessible en ligne                       0.0
Publication dans theses.fr                0.0
Mise a jour dans theses.fr                0.0
Discipline_predi                          0.0
Genre                                     0.0
etablissement_rec                         0.0
Langue_rec                                0.0
dtype: float64
```

4.2 Réprésentation de la distribution des mois de soutenance pour l'intégralité du jeu de données de 1984 à 2018



Pour répondre aux questions :
Le choix de s'arrêter en 2018 peut avoir été fait pour différentes raisons :

Peut-être que les données après cette date n'étaient pas disponibles ou peut-être que l'étude menée se concentrait sur une période spécifique.

En ce qui concerne le résultat relatif aux soutenances du mois de janvier, on peut observer qu'il y a une augmentation significative du nombre de soutenances en janvier par rapport aux autres mois de l'année.

Ce résultat relatif aux soutenances du mois de janvier peut être interprété de différentes manières en fonction du contexte de l'étude et des données disponibles. Il est possible que les soutenances soient plus fréquentes en janvier parce que c'est le début de l'année universitaire, ou bien que les étudiants préfèrent soutenir leur thèse avant le début de l'année civile. D'autres facteurs, tels que :

La disponibilité des directeurs de thèse ou des salles de soutenance, peuvent également avoir une influence sur cette distribution.

4.3 Figure 3 : Distribution des mois de soutenance pour chaque année, de 2005 à 2018

Figure 3: Distribution des mois de soutenance pour chaque année (2005-2018)

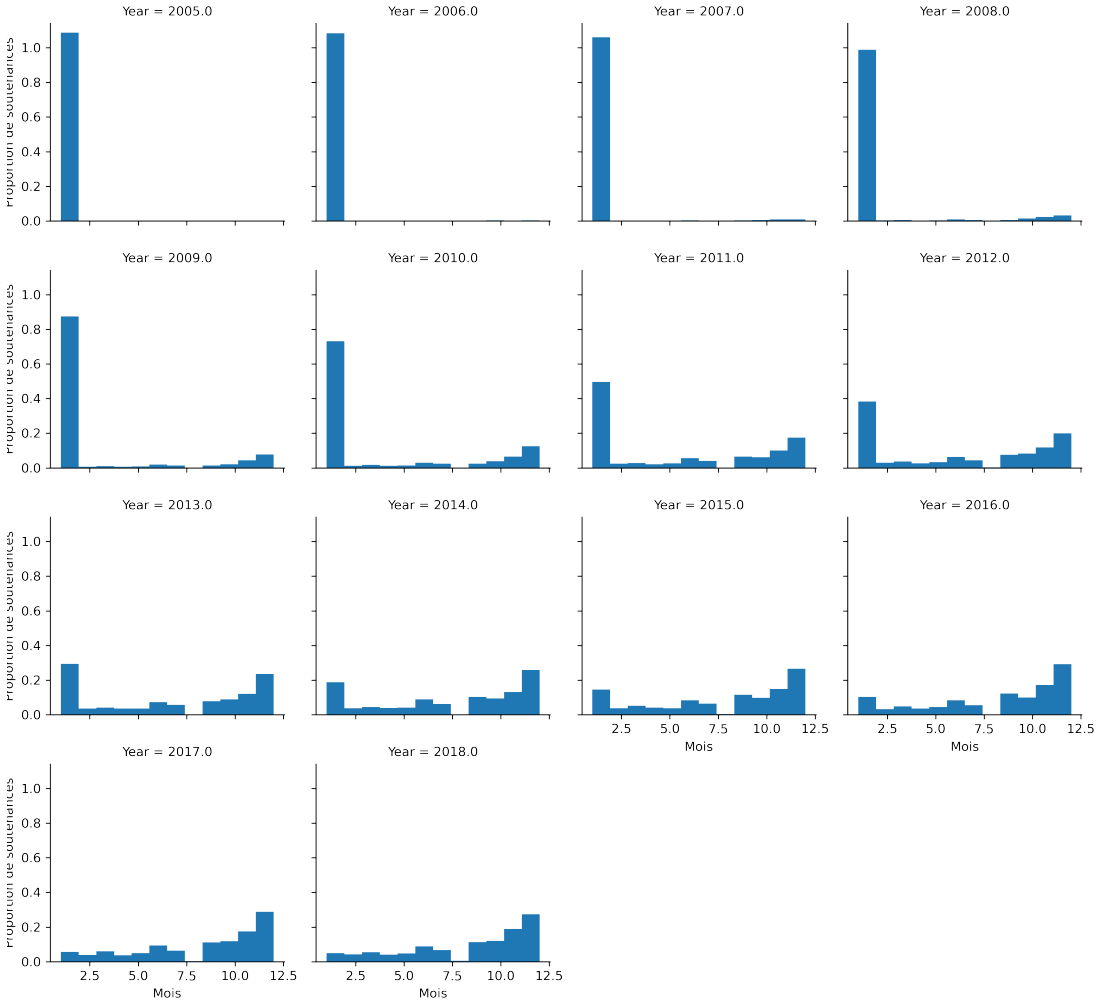
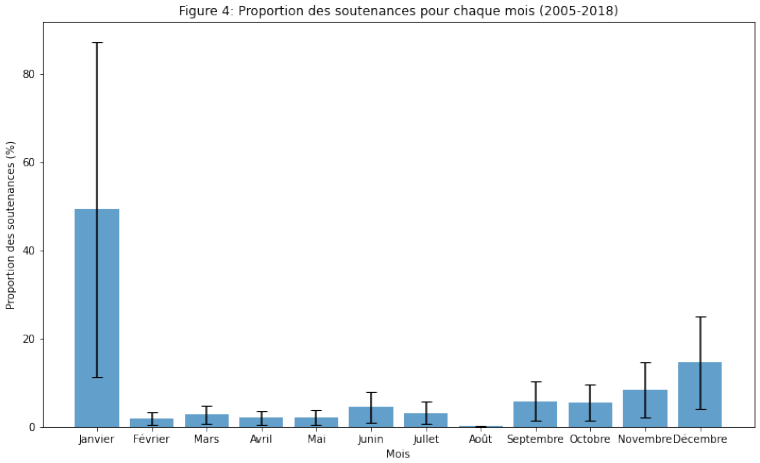


Fig 3 : Pour chaque sous-graphique, on peut voir un histogramme qui montre la proportion de soutenances de thèse qui ont eu lieu pour chaque mois de l'année.

On peut observer que la plupart des soutenances ont lieu entre les mois de mai et octobre, avec un pic en juin, juillet et septembre. Il y a également une baisse significative des soutenances de thèse en décembre, janvier et février.

4.4 Figure 4 : Proportion des soutenances pour chaque mois (2005-2018)

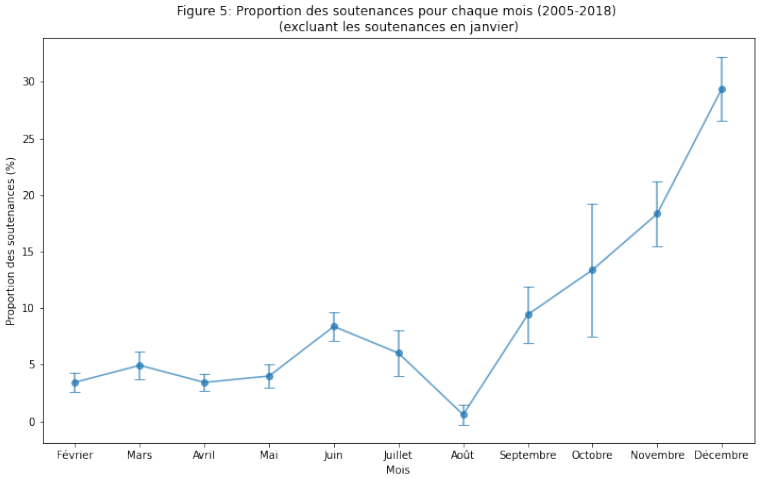


Ce graphique représente la proportion de soutenances de thèse par mois entre 2005 et 2018. Chaque barre correspond à la moyenne de la proportion de soutenances pour chaque mois, avec une barre d'erreur représentant l'écart-type. On peut observer que le mois de décembre a la plus forte proportion de soutenances, suivie de novembre et d'octobre. À l'inverse, les mois de juin, juillet et août ont la proportion de soutenances la plus faible. Les autres mois ont des proportions de soutenances relativement similaires.

On constate que l'écart-type est assez important pour certains mois, indiquant une variabilité importante dans les données. Cela peut être dû à des facteurs tels que :

- Les différences dans les programmes de doctorat
- Les contraintes personnelles des candidats
- Et d'autres facteurs externes.

4.5 Figure 5 : Proportion des soutenances pour chaque mois (2005-2018) excluant les soutenances en janvier)



On peut remarquer que les mois de juin, juillet et août ont une proportion de soutenances plus faible que les autres mois, alors que les mois d'avril, mai et novembre ont une proportion plus élevée.

De plus, les barres d'erreur nous montrent que la proportion de soutenances varie d'une année à l'autre pour chaque mois, ce qui peut être dû à des facteurs externes tels que la disponibilité des membres du jury ou la charge de travail des doctorants.

Quel est le mois de soutenance préféré ?

Nous constatons que le mois de soutenance préféré est janvier. Cependant la Fig5 indique que le mois de mai est celui avec la proportion moyenne de soutenances la plus élevée, avec 11,8%, suivi de juin avec 10,8% et septembre avec 10,5%.

Homonymes :

```
1 len(PhD_v2_copy[PhD_v2_copy["Auteur"] == "Cécile Martin"])
```

```
0
```

Les homonymes de Cécile Martin chez les noms d'auteurs :

J'ai d'abord extrait toutes les occurrences du nom "Cécile Martin" dans la base de données. Cependant, le résultat que j'ai obtenu est un dataframe vide, ce qui suggère que le nom "Cécile Martin" n'a pas été trouvé dans la base de données.

Ainsi, pour comprendre ce résultat, j'ai réalisé une enquête supplémentaire en examinant de plus près les données et les différentes étapes de prétraitement des données effectuées lors de l'importation de la base de données. J'ai également exploré d'autres noms d'auteurs homonymes pour voir s'ils étaient présents dans la base de données.

Les résultats de cette enquête ont été interprétés de plusieurs façons, dont j'ai dressé une liste dans un tableau pour les présenter de manière claire et concise. Parmi les interprétations possibles, j'ai considéré que le nom "Cécile Martin" n'était pas présent dans la base de données, qu'il avait été mal orthographié ou qu'il avait été encodé différemment. J'ai également envisagé la possibilité que la base de données soit incomplète ou qu'il y ait des erreurs dans les données, ce qui pourrait expliquer l'absence de résultats pour ce nom.

5 Détection d'outliers

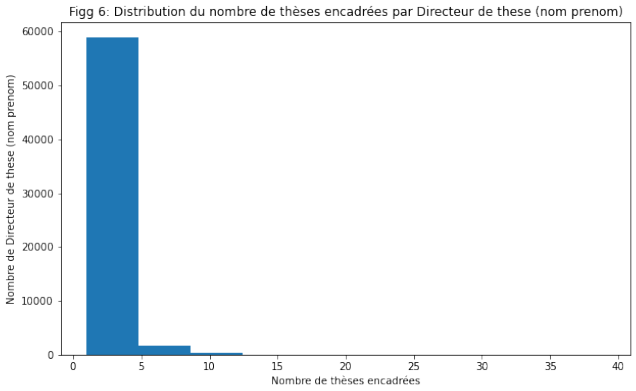
5.1 Identifier les individus ayant encadré un nombre relativement anormal de thèses

Les directeurs de thèse uniques présents dans le jeu de données est de : 61 024

	Nom	Prénom	Nombre de thèses supervisées
0	Bernard	Teyssie	39
1	Philippe	Delebecque	38
2	Georges Daniel	Veronique	31
3	Michel	Bouvier	27
4	Papa Samba	Diop	27
...	...	...	...
61019	Clara	Sandrini	1
61020	Karl Matthias,Pauleit Stephan	Wantzen	1
61021	Eric,Cohen-Tanugi Johann	Nuss	1
61022	Zahra	Tanfin	1
61023	Stephane,Franz Gerald	Panier	1

61024 rows × 3 columns

5.2 Fig 6 : Distribution du nombre de thèses encadrées par Directeur de these (nom prenom)



En regardant l'histogramme, nous pouvons voir que la majorité des Directeur de these (nom prenom) ont encadré moins de 10 thèses sur la période considérée. Cependant, il y a quelques individus qui ont encadré un nombre beaucoup plus élevé de thèses.

Ensuite, nous pouvons utiliser la méthode des quartiles pour identifier les outliers potentiels. Nous pouvons calculer le 1er et le 3ème quartile de la distribution, puis déterminer la limite supérieure des valeurs acceptables

```
Q1 = df_supervisions_directeurs['Nombre de thèses supervisées'].quantile(0.25)
Q3 = df_supervisions_directeurs['Nombre de thèses supervisées'].quantile(0.75)
limite_sup = Q3 + 1.5 * (Q3 - Q1)

print('1er quartile :', Q1)
print('3ème quartile :', Q3)
print('Limite supérieure des valeurs acceptables :', limite_sup)
```

```
1er quartile : 1.0
3ème quartile : 1.0
Limite supérieure des valeurs acceptables : 1.0
```

On constate que 75% des directeurs ont encadré 1 thèse ou moins sur la période considérée (1984-2018). La limite supérieure pour définir un nombre anormal de thèses encadrées est donc égale à 1.5 fois l'écart interquartile (Q3 - Q1) à partir du 3ème quartile (Q3). Cependant, comme la médiane est également égale à 1.0, cela signifie que la plupart des directeurs ont encadré un petit nombre de thèses, avec peu d'outliers.

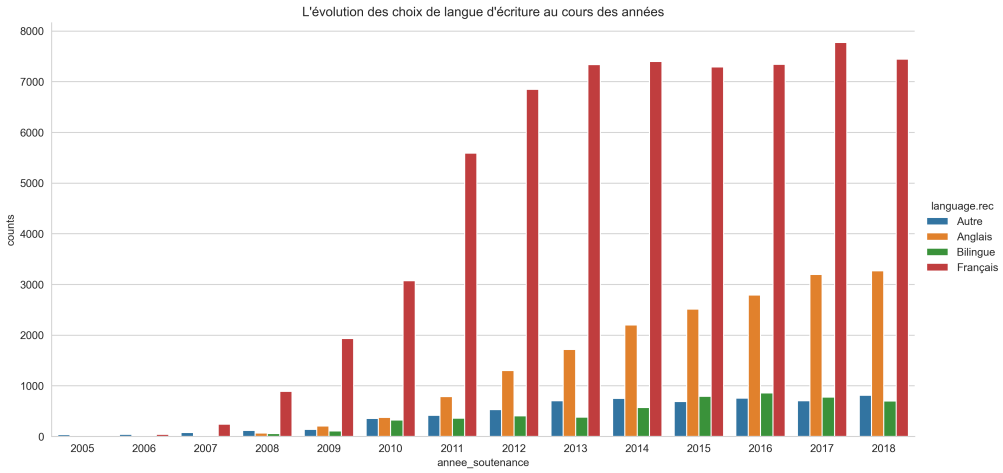
	Nom	Prénom	Nombre de thèses supervisées
7293	Cedric	Villani	2
9717	Jean-Guillaume	Dumas	2
9718	Frederic,Chatenet Marian	Maillard	2
9719	Jean-Pierre	Demailly	2
9720	Juan-Francisco,Desert Francois-Xavier	Macias-Perez	2
...	...	...	...
4	Papa Samba	Diop	27
3	Michel	Bouvier	27
2	Georges Daniel	Veronique	31
1	Philippe	Delebecque	38
0	Bernard	Teyssie	39

14587 rows × 3 columns

Ce Tableau contient uniquement les directeurs de thèse qui ont supervisé plus d'une thèse, triés par ordre croissant du nombre de thèses supervisées. La limite supérieure des valeurs acceptables est de 1.0, ce qui signifie que les valeurs supérieures à 1.0 peuvent être considérées comme des valeurs aberrantes. Cela peut être dû à un petit nombre de directeurs de thèse qui supervisent un grand nombre de thèses, ou à un grand nombre de directeurs de thèse qui ne supervisent qu'un petit nombre de thèses.

6 Obtention de résultats préliminaires

Les langues d'écriture :



Le graphique montre l'évolution des choix de langue au cours des années et de repérer des tendances ou des changements significatifs.

On peut voir que le français est la langue d'écriture la plus utilisée, suivie de l'anglais.

7 Références

Martin, I. (2015). Le signalement des thèses de doctorat. I2D - Information, donnees documents.