**Université Cergy Paris**


# Diplôme universitaire : Data Analyst

## UE 1 : Manipulation et prétraitement de données

1er mai 2022

**Haury Fabien**

# Table des matières

# Chapitre 1

# Présentation des données

## 1.1 PhD V2

### 1.1.1 Présentation

Le jeu de données PhD V2 est fondé sur des informations récupérées sur le site theses.fr. Il regroupe un ensemble d'informations centrées sur les thèses telles que l'auteur, directeur de thèse, langue etc.

Les variables sont bien présentes séparément chacune pour un total de 447 644 lignes et 18 colonnes.

| Variable | Nb distinct | Variable | Nb distinct |
|---|---|---|---|
| Auteur | 430277 | Statut | 2 |
| Identifiant auteur | 313775 | Date premiere inscription soutenance | 4010 |
| Titre | 446815 | Date de soutenance | 3992 |
| Directeur de these | 159019 | Year | 45 |
| Directeur de these (nom prenom) | 159021 | Langue de la these | 206 |
| Identifiant directeur | 98907 | Identifiant de la these | 447572 |
| Etablissement de soutenance | 568 | Accessible en ligne | 2 |
| Identifiant etablissement | 573 | Publication dans theses.fr | 2765 |
| Discipline | 24263 | Mise a jour dans theses.fr | 2634 |

TABLE 1 – Nombre distinct par variables du jeu de données basé sur le site theses.fr

La Table 1 montre le nombre distinct d'observations pour chacune des variables du jeu de données. Cela nous permet d'avoir une vue d'ensemble des données.

### 1.1.2 Nature des variables

- Auteur : Class "Character"
- Identifiant auteur : Class "Character"
- Titre : Class "Character"
- Directeur de these : Class "Character"
- Directeur de these (nom prenom) : Class "Character"
- Identidiant directeur : Class "Character"
- Etablissment de soutenance : Class "Character"
- Identifiant etablissment : Class "Character"
- Discipline : Class "Character"
- Status : Class "Character"
- Date de premiere inscription en doctorat : Class "Character"
- Date de soutenance : Class Character
- Year : Class "Double"
- Langue de la these : Class "Character"
- Identifiant de la these : Class "Character"
- Accessible en ligne : Class "Character"
- Publication dans theses.fr : Class "Character"

Nous constatons que toutes les variables sauf une sont de Class "Character". Il fait sens pour la variable Year d'être de Class "Double". Cependant, pour une facilité d'utilisation, convertir les variables **Date de premiere inscription en doctorat, Date de soutenance, Year, Publication dans thèses.fr et Mise a jour dans theses.fr** en Class "Date" est préférable. De même, les variables **Statut, Langue de la these, Accessible en ligne** peuvent être changées en factor.

# Chapitre 2

# Données manquantes

## 2.1 Visualisation de l'ensemble des données

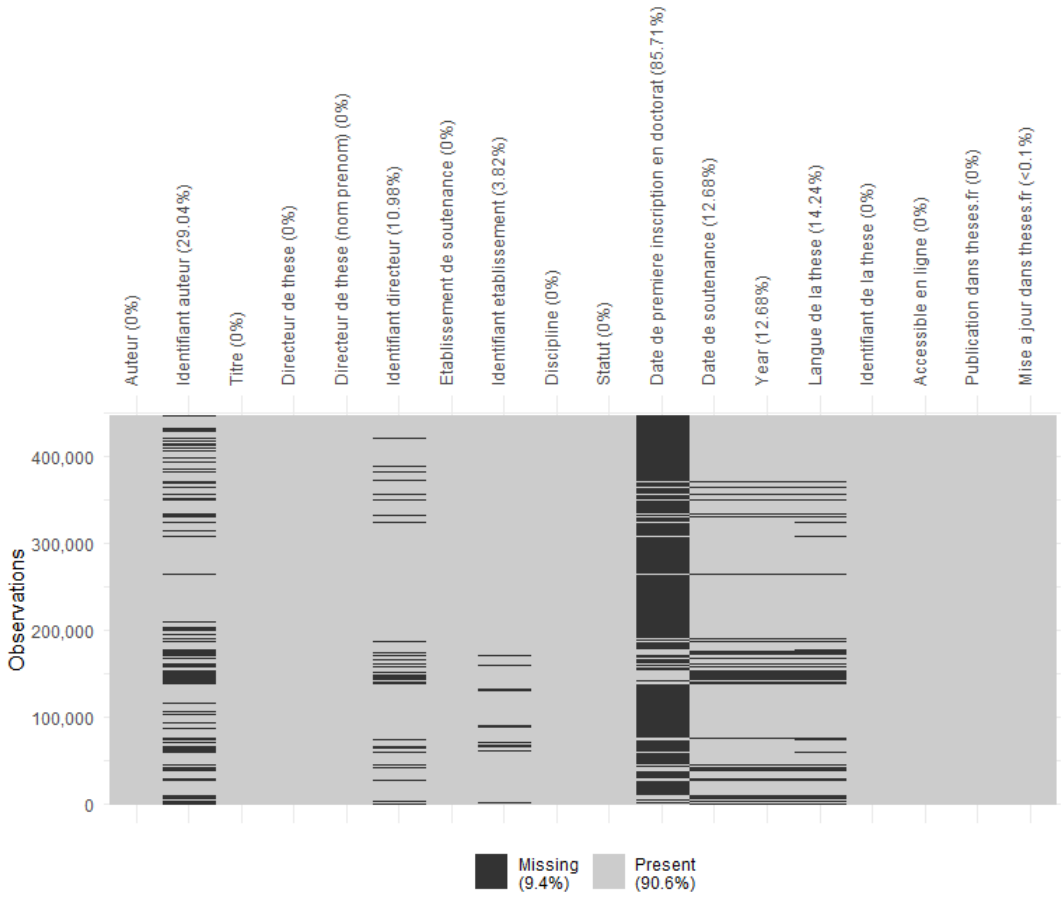



FIGURE 1 – Visualisation des données manquantes sur la totalité du jeu de données

La Figure 1 représente une visualisation des données manquantes. Il y a $9,4\%$ de données manquantes. La variable contenant le plus de données manquantes est **Date de premiere inscription en doctorat** avec $85,71\%$ .
Un pattern semble se dessiner entre **Date de premiere inscription en doctorat, Date de soutenance, Year, Langue de la thèse**. Il semble que pour une donnée présente dans **Date de premiere inscription en doctorat**, les données de **Date de soutenance, Year, Langue de la thèse** sont manquantes.

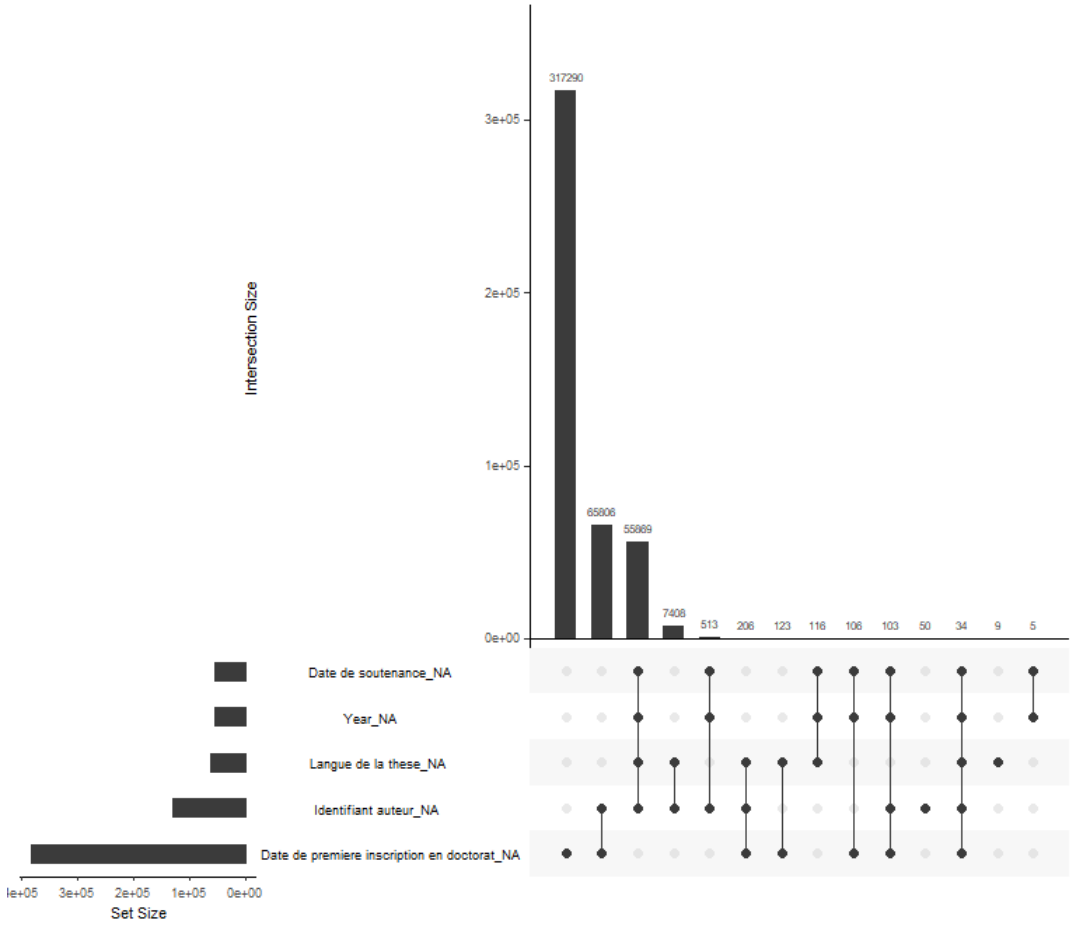



FIGURE 2 – Modèle des données manquantes

La Figure 2 nous permet de visualiser les liens entre données manquantes de différentes variables. Nous pouvons confirmer l'hypothèse d'un lien entre **Date de premiere inscription en doctorat, Date de soutenance, Year, Langue de la thèse**

## 2.2 Visualisation des données pour chaque valeur de la variable statut

Comme nous venons de le voir, il existe un lien entre **Statut** et **Date de premiere inscription en doctorat, Date de soutenance, Year, Langue de thèse**. Nous allons séparer **Statut** en deux : enCours et soutenue.

La figure 3 représente une visualisation des données manquantes pour **Statut** : enCours. 54% des données sont manquantes dont une grande majorité sont imputables aux variables **Date de soutenance, Year, Langue de la thèse**. On peut constater que peu de données de **Date de premiere inscription en doctorat** sont manquantes, a contrario de **Date de soutenance, Year, Langue de la thèse** pour lesquels très peu de données sont présentes.

La figure 4 représente une visualisation des données manquantes pour **Statut** : soutenue. 20% des données sont manquantes dont une grande majorité sont imputables aux variables **Date de premiere inscription en doctorat**.

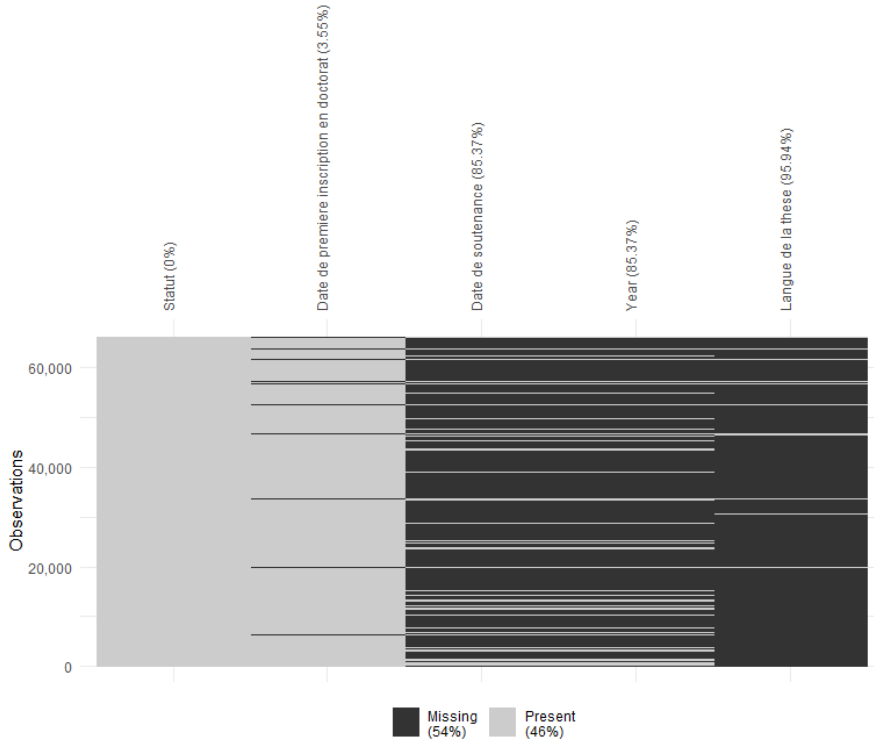



FIGURE 3 – Visualisation des données manquantes pour la valeur statut enCours

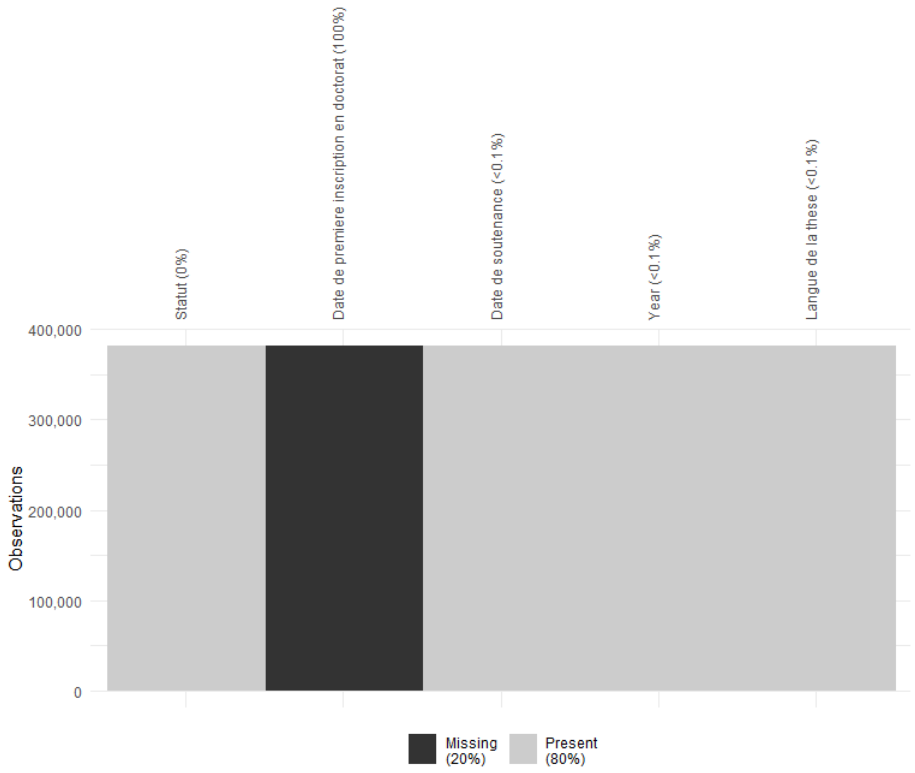



FIGURE 4 – Visualisation des données manquantes pour la valeur statut soutenue

# Chapitre 3

# Principaux problèmes détectés

## 3.1 Exploration des données pour les dates de soutenance

### 3.1.1 Distribution des soutenances par mois

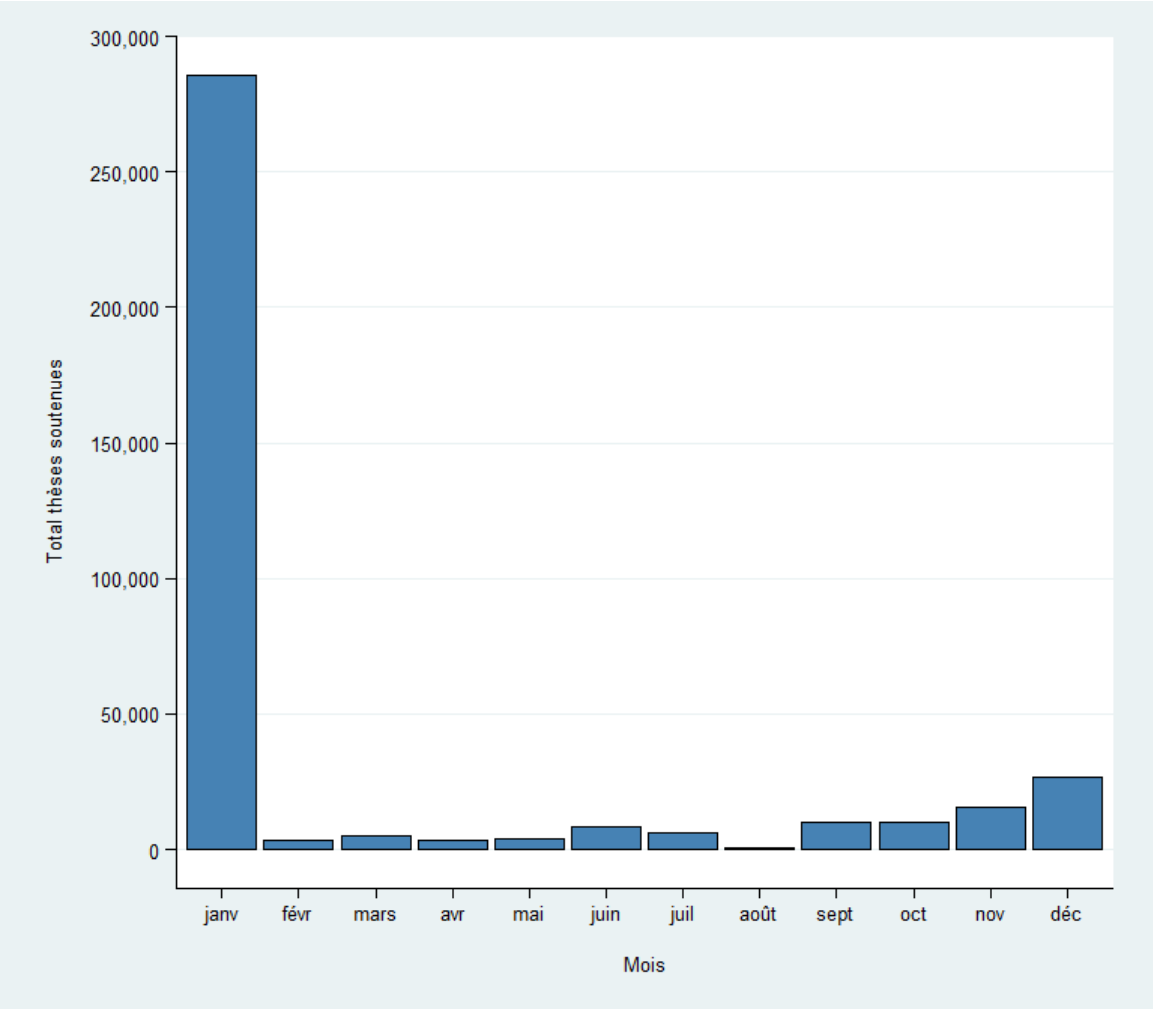



FIGURE 5 – Distribution des soutenances par mois, période 1984-2018.

La Figure 5 représente la distribution des soutenances de thèses par mois par mois pour la période de 1984 à 2018. Le mois de janvier est le mois de soutenance le plus représenté avec environ 275 000 thèses soutenues pour ce mois. La période allant de février à juillet est quasi-uniforme avec quelque milliers de thèses soutenues. Le mois d'août représente un creux, enfin la période de septembre à décembre montre un

accroissement des soutenances avec au mois de septembre, environ 12 000 thèses soutenues et en décembre, environ 25 000 thèses soutenues.

### 3.1.2 Distribution des soutenances pour chaque année

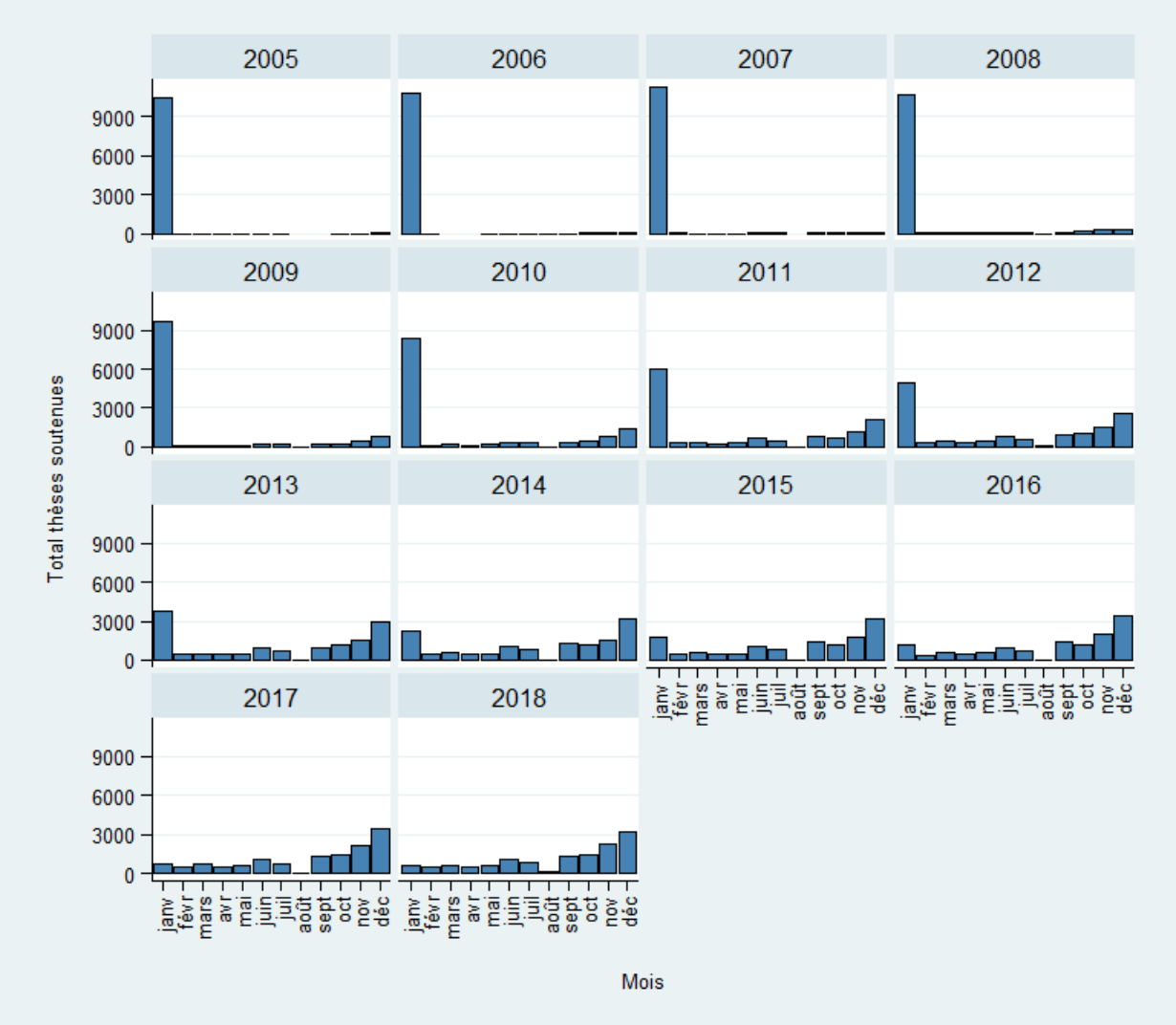



FIGURE 6 – Distribution des soutenances par mois par année respective, période 2005-2018

La Figure 6 représente la distribution des soutenances par mois par année respective pour la période de 2005 à 2018. La période de 2005 à 2010 est fortement centrée sur le mois de janvier avec ± 9000 thèses en moyenne. La période suivante de 2011 à 2018 montre la chute du mois de janvier à ± 800 thèses alors que les autres mois augmentent tous, spécialement le mois de décembre qui passe de ± 1500 thèses à plus de 3000 thèses soutenues.

### 3.1.3 Pourcentage de soutenances par mois

La Figure 7 représente le pourcentage de soutenance par mois avec son écart-type sans filtrer le 1er janvier pour la période de 2005 à 2018. Le mois de janvier représente près de 50% des dates de soutenances avec une écart-type de 32.5%. Cet écart-type n'est autre que la représentation de l'évolution vue à la 6.
Le 1er Janvier étant un jour férié en France et le lendemain de réveillon, il est plausible qu'il existe un problème avec ce jour en particulier. Filtrons ce jour en particulier et regardons le résultat.

La Figure 8 représente le pourcentage de soutenance par mois avec son écart-type sans filtrer le 1er janvier pour la période de 2005 à 2018. Après filtrage du 1er janvier, il apparaît clairement qu'une majorité

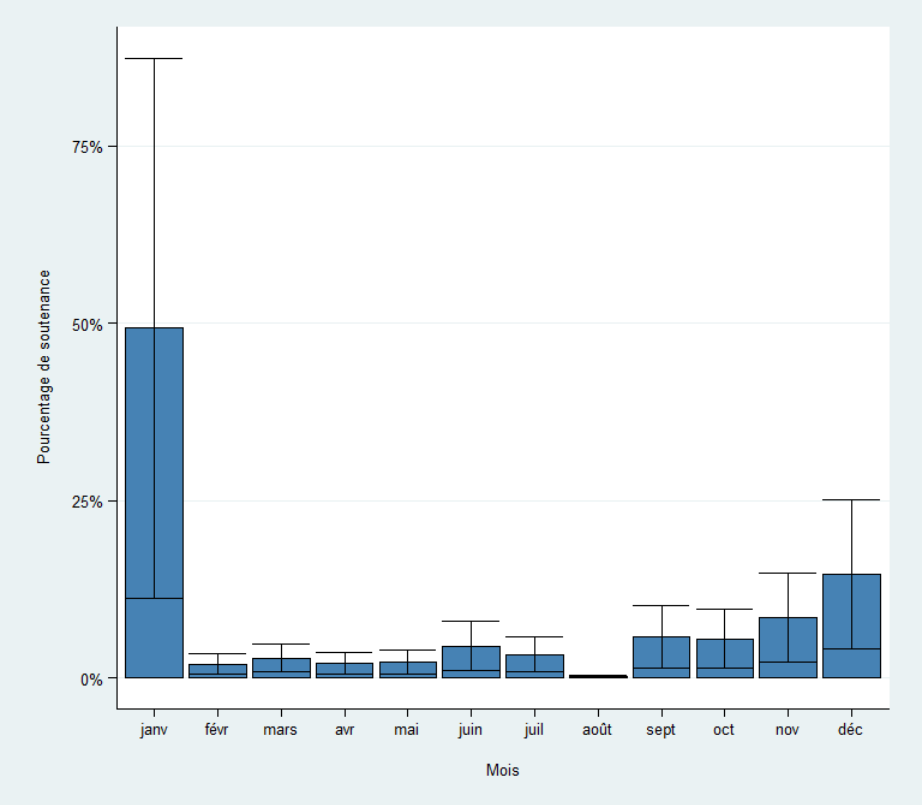



FIGURE 7 – Pourcentages des soutenances par mois avec écart-type, période 2005-2018, sans filtrer le premier janvier

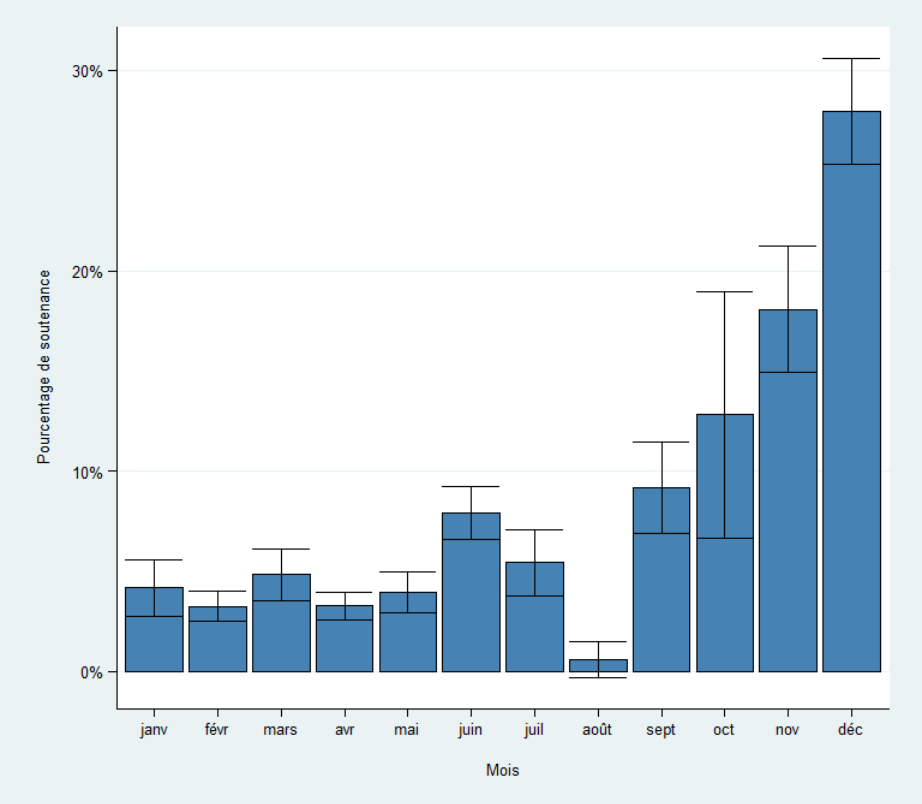



FIGURE 8 – Pourcentages des soutenance par mois avec écart-type, période 2005-2018, avec filtrage du premier janvier

de soutenance sont tenues en Décembre. Une majorité des dates sont soit situées aux mois de Juin, soit situées en automne. Plusieurs hypothèses sont envisageables pour expliquer cela :

- Pour les mois de juin, il est possible que les doctorants veuillent soit finir leur thèse avant les vacances,

soit finir leur thèse pour pouvoir être affectés à un poste d'enseignement dès la rentrée suivante.

- Pour la période d'automne, elle semble divisée en deux parties : Septembre/Octobre/Novembre et Décembre.
  - Comme nous pouvons le voir, les soutenances au mois d'août sont peu nombreuses, il est possible d'imaginer que ce mois sert de période de repos ou bien de période de relecture/finition en vue de la soutenance pour la période septembre/octobre/novembre.
  - Pour le mois de décembre, une hypothèse possible est tout simplement le fait de valider sa thèse pour pouvoir l'annoncer lors des festivités de fin d'année ou bien finir sa thèse avant la fin de l'année civile pour être éligible à des prix/récompenses durant l'année suivante.

### 3.1.4 Évolution des soutenances au premier janvier par année

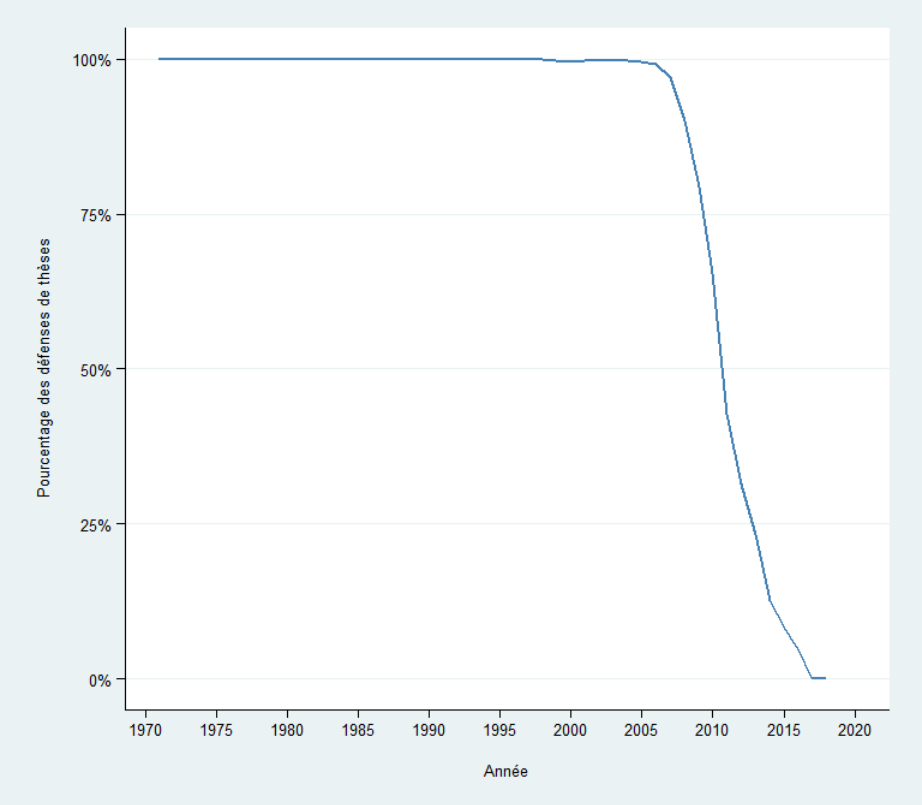



FIGURE 9 – Évolution des soutenances par mois au premier Janvier par année sur la totalité du jeu de données

La Figure 9 représente l'évolution des soutenances au premier janvier au fil des années sur la totalité du jeu de données. Pour la période allant de 1970 à 2005 la quasi totalité des soutenances ont eu lieu au premier janvier. À partir de 2005 jusqu'à 2018 la courbe décroît passant de ±100 % à environ ±0 %. Cette chute est aussi représentée à la Figure 6

## 3.2 Analyse des problèmes liés aux homonymes

### 3.2.1 Homonyme Cécile Martin

**Analyse du jeu de données.**

| Variable | Nb distinct | Variable | Nb distinct |
|---|---|---|---|
| Auteur | 1 | Statut | 1 |
| Identifiant auteur | 4 | Date de premiere inscription en soutenance | 1 |
| Titre | 7 | Date de soutenance | 7 |
| Directeur de these | 7 | Year | 7 |
| Directeur de these (nom prenom) | 7 | Langue de la these | 2 |
| Identifiant directeur | 7 | Identifiant de la these | 7 |
| Etablissement de soutenance | 7 | Accessible en ligne | 2 |
| Identifiant etablissement | 7 | Publication dans theses.fr | 3 |
| Discipline | 7 | Mise a jour dans theses.fr | 5 |

TABLE 2 – Nombre distinct des variables de PhD v2

La Table 2 montre le nombre distinct d'observation pour chacune des variables du jeu de données. Cela nous permet d'avoir une vue d'ensemble des données.
Il existe 7 Cécile Martin dans le jeu de données, cependant seules 4 sont uniques. Chacune des thèses a été faite dans un établissement différent avec un directeur différent ainsi que dans une discipline différente. Un seul statut existe : soutenue. Une thèse a été soutenue dans une langue différente des autres.

**Distribution des années de soutenance pour l'identifiant auteur unique.**

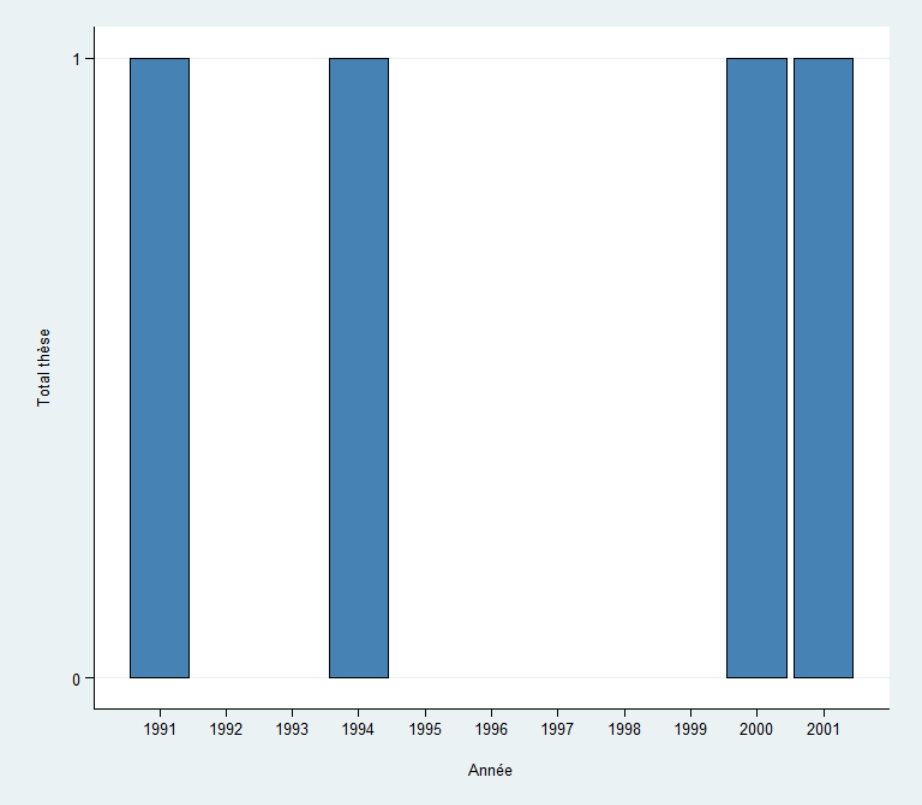



FIGURE 10 – Histogramme des années de soutenance pour l'identifiant auteur unique

La Figure 10 représente la distribution des années de soutenance pour l'identifiant auteur unique. Nous constatons que les données sont séparées en trois périodes, 1991, 1994 et 2000/2001. Si le graphique ne nous

permet pas de conclure, une exploration du jeu de données "à la main" peut nous permettre de tirer des conclusions.

| Discipline | Date de soutenance |
|---|---|
| Neurosciences | 1991-01-01 |
| Sciences biologiques et fondamentales appliquees. Psychologie | 1994-01-01 |
| Sciences biologiques fondamentales et appliquees. Sciences medicales | 2000-01-01 |
| Genie des procedes industriels | 2001-01-01 |

TABLE 3 – Discipline et Date de soutenance pour l'identifiant unique

La Table 3 contient quatre disciplines différentes ainsi que les dates de soutenances respectives. Nous pouvons conclure que la discipline liée au génie des procédés industriel n'a aucun lien avec les trois autres. Nous voyons aussi que deux des trois thèses restantes ont été soutenues dans le même domaine : Sciences biologiques fondamentales et appliquées. Leurs dates de soutenance sont suffisamment espacées pour rendre possible l'hypothèse que ces deux thèses ont été soutenues par une seule et même personne.

## 3.3 Conclusion

Nous nous sommes posé comme problème d'étudier les homonymes Cécile Martin. Durant cette analyse, nous avons constaté un problème de doublon et avons poussé l'analyse.
Nous avons potentiellement résolu le problème des doublons homonymes Cécile Martin en montrant qu'il est plausible qu'un de ces doublons soit effectivement la même personne en se fondant sur la discipline de la thèse et leurs date de soutenance.
Mais plus encore, cela nous a démontré l'utilité de faire ces analyses en gardant les autres variables afin de pouvoir discerner des patterns.

# Chapitre 4

# Outliers

## 4.1 Présentation

Le jeu de données contient les informations pour la période allant de 1984 à 2018. Il y a un total de 308 587 lignes, 66 148 directeurs distincts, 56 680 identifiants directeurs distincts et un total de 1,9% de données manquantes. Il apparaît donc que des doublons existent étant donné le nombre de directeurs par rapport au nombre de lignes. De même il apparaît que certains identifiants directeurs sont en doublon ou sont inexistants.

## 4.2 Recherche des outliers

Cherchons à partir de combien de thèses dirigées nous pouvons considérer cela comme un outlier. Commençons par un histogramme pour voir la distribution.

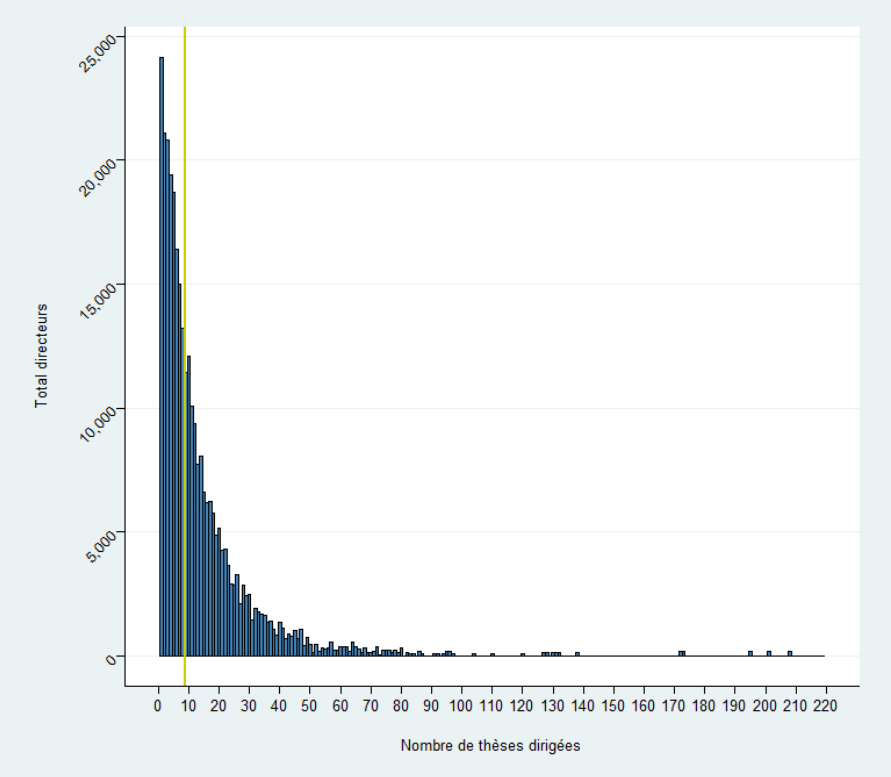



FIGURE 11 – Distribution du total de directeurs par nombre de thèses dirigées, totalité du jeu de données, avec médiane

La Figure 11 représente la distribution du total de directeurs par nombre de thèses dirigées sur la totalité du jeu de données avec ligne médiane. La ligne médiane nous permet de voir que 50% des directeurs ont dirigé entre une et neuf thèses. Il est clair qu'il existe des outliers, un certain nombre de directeurs ont dirigés plus de 100 thèses !

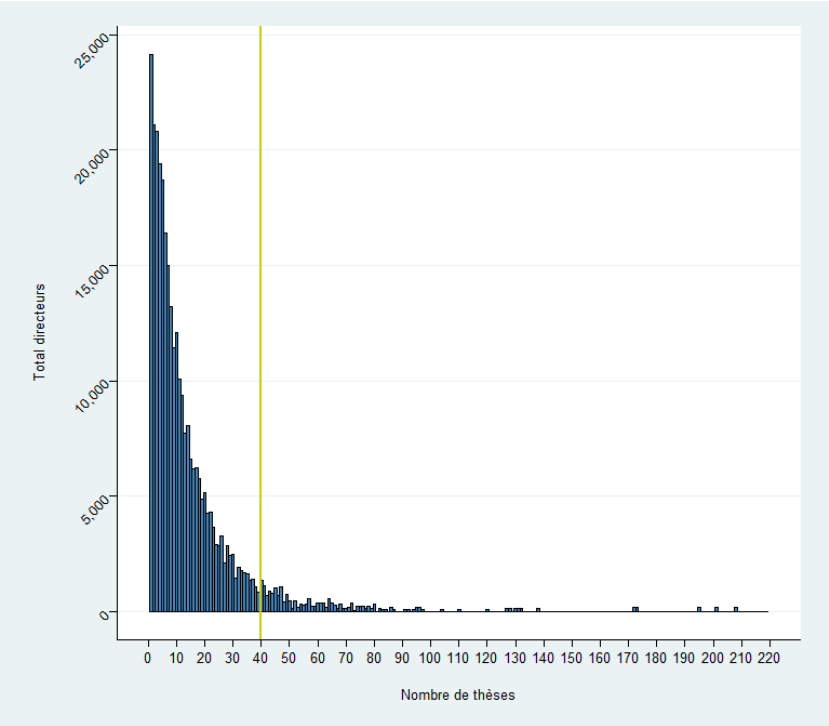



FIGURE 12 – Distribution des directeurs par nombre de thèses dirigées, avec ligne représentant le début des outliers

La Figure 12 représente un histogramme du nombre de thèses dirigées sur la totalité du jeu de données avec une représentation par une ligne verticale marquant la séparation entre les données non-outliers et données outliers. Cette ligne a été calculée par la méthode des Interquartile Range (IQR).

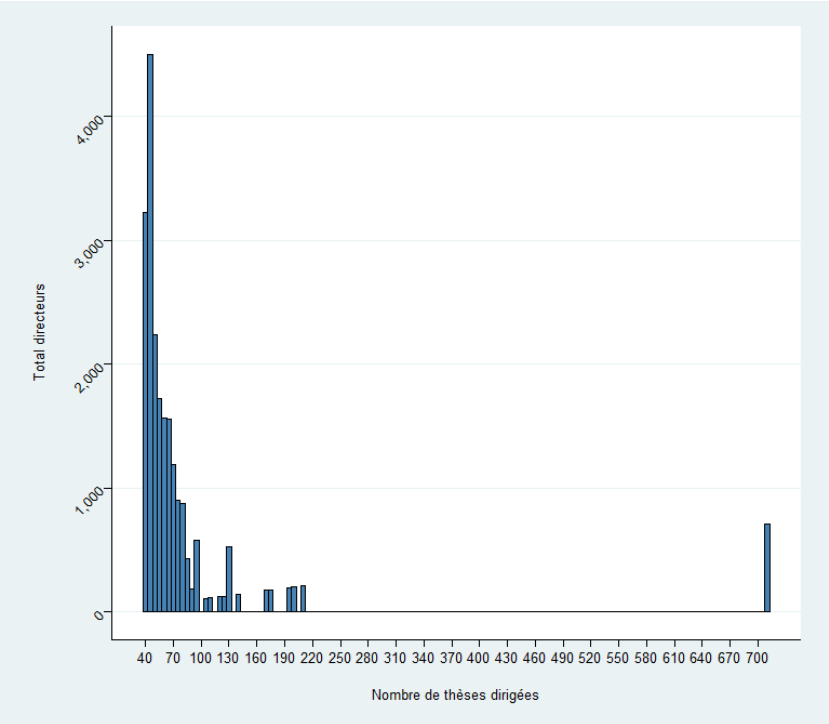



FIGURE 13 – Distribution du total des directeurs outliers par nombre de thèses dirigées

La Figure 13 représente la distribution du total des directeurs outliers par nombre de thèses dirigées. Maintenant que nous avons ciblé les anomalies, allons les voir de plus près. Pour plus de clarté, les outliers seront divisés en deux parts : de 40 à 140 thèses dirigées, de 140 à 250 thèses dirigées.

## 4.3 Analyse des outliers

### 4.3.1 Analyse des directeurs outliers entre 40 et 140 thèses dirigées

Il y a 20 068 lignes pour 367 directeurs distincts ainsi que 447 identifiants de directeurs distincts. Cela nous permet de constater qu'un grand nombre de directeurs sont dupliqués mais aussi qu'un certain nombre de directeurs partage le même identifiant ce qui n'est pas possible.

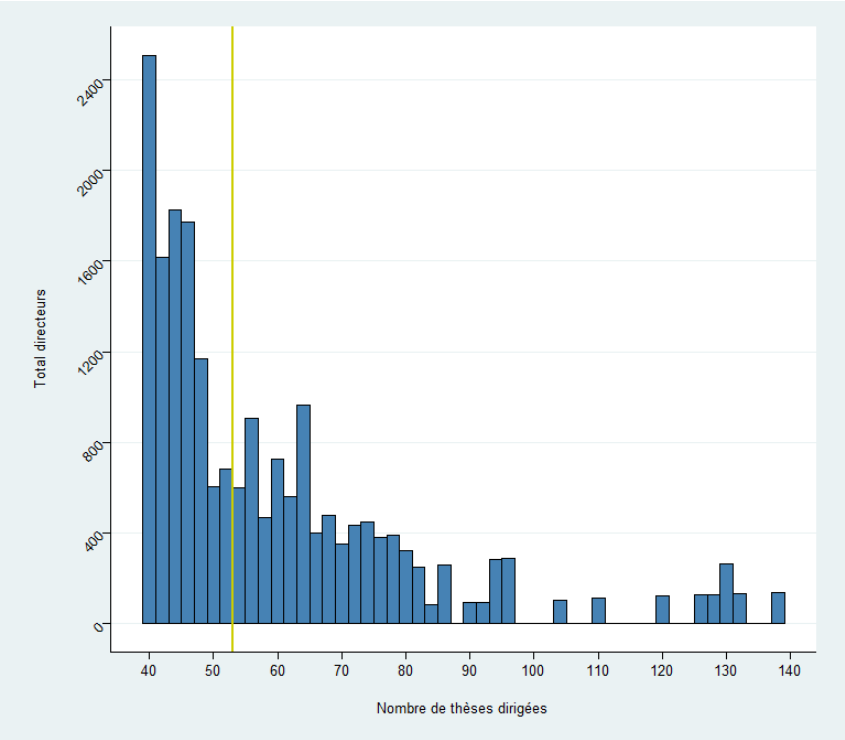



FIGURE 14 – Distribution outliers entre 40 et 140 thèses dirigées

La Figure 14 représente la distribution des outliers compris entre 40 et 140 thèses dirigées avec ligne médiane.

### 4.3.2 Analyse des directeurs outliers supérieur à 140 thèses dirigées

Il y a 1660 lignes pour six directeurs distincts ainsi que huit identifiants de directeurs distincts. Cela nous permet de constater qu'un grand nombre de directeurs sont dupliqués mais aussi qu'un certain nombre de directeurs partage le même identifiant ce qui n'est pas possible. Il existe un groupe d'anomalie extrême avec un total de thèses supérieur à 700. Allons voir plus en détail ce qu'il contient.
Il y a 711 lignes pour un directeur distinct ainsi que un identifiant de directeur distinct. Il semblerait que cette anomalie soit le résultat d'un seul directeur ! Le nom du directeur unique est : "Directeur de thèse inconnu".
En combinant les deux information citées ci-dessus il est évident que la corrélation entre les deux variables est forte mais il est aussi plausible d'émettre l'hypothèse que ce directeur unique est en fait une multitude de directeurs différents. Lors des saisies des informations soit celle-ci était inconnue, soit il s'agit du nom par défaut , soit une erreur de saisie à était effectuée.

## 4.4 Conclusion

Nous nous sommes posé comme problème d'étudier les directeurs de thèses afin de trouver les outliers. Durant cette analyse, nous avons constaté un grand nombre de doublons au sein des variables choisies, ainsi que des différences dans les outliers même.
Nous venons de voir les outliers uniquement sous le rapport des variables **Directeur de thèse (nom prenom), Identifiant directeur**. Si cela nous a permis de mieux voir la distribution et les problèmes liés à cette variable, refaire le même procédé pour la variable **Directeur de thèse** et comparer les résultats peuvent mettre en valeur les différences entre ces deux variables. Cependant, quelque soit la variable que nous choisirons dans le futur, son analyse devra être faite en gardant le jeu de données entier afin de pouvoir discerner des patterns à travers celui-ci. (i.e - lien entre les outliers extrêmes et leur identifiant établissement pour déterminer si le problème vient d'un établissement en particulier ?)

# Chapitre 5

# Résultats préliminaires

## 5.1 Présentation

Ce chapitre portera principalement sur l'évolution des langues utilisées dans les thèses. Pour ce faire le nom **thèse de la langue** est renommée en **Langue** et les observations sont regroupées et mises sous forme de factor comme suit :

- "fr" est renommé "Français"
- "en" est renommé "Anglais"
- "enfr" et "fren" sont renommés en "Bilingue"
- les données manquantes sont renommées en "NA" (notons que nous leur supprimons leur statut de NA)
- toutes les autres langues sont renommées "Autres"

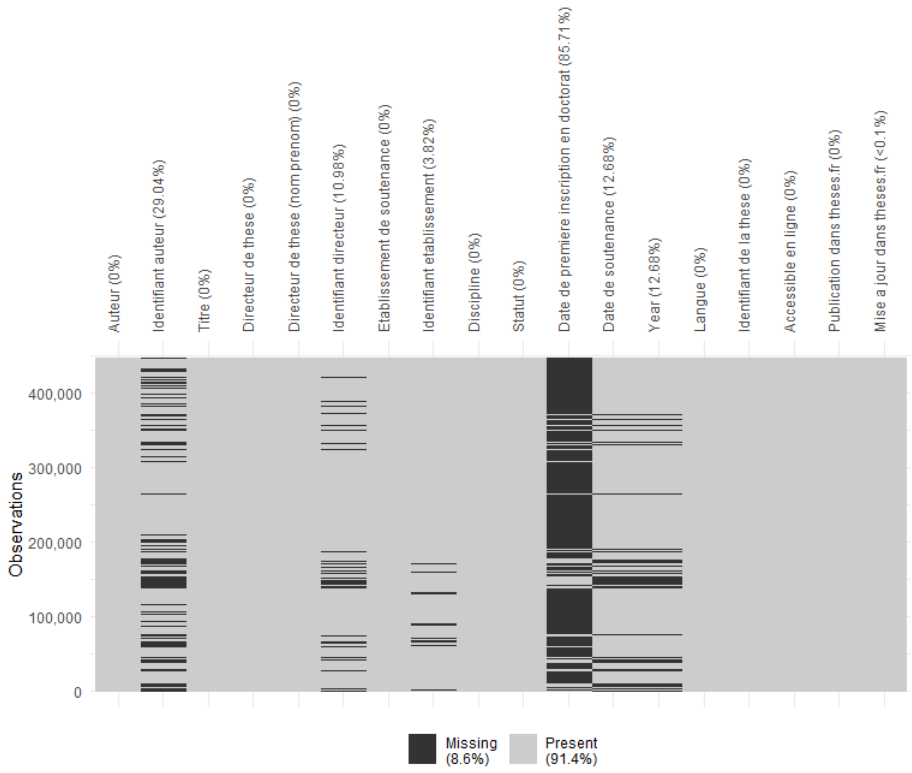



FIGURE 15 – Visualisation données manquantes

## 5.2 Évolution des langues dans le temps

### 5.2.1 Jeu de données entier

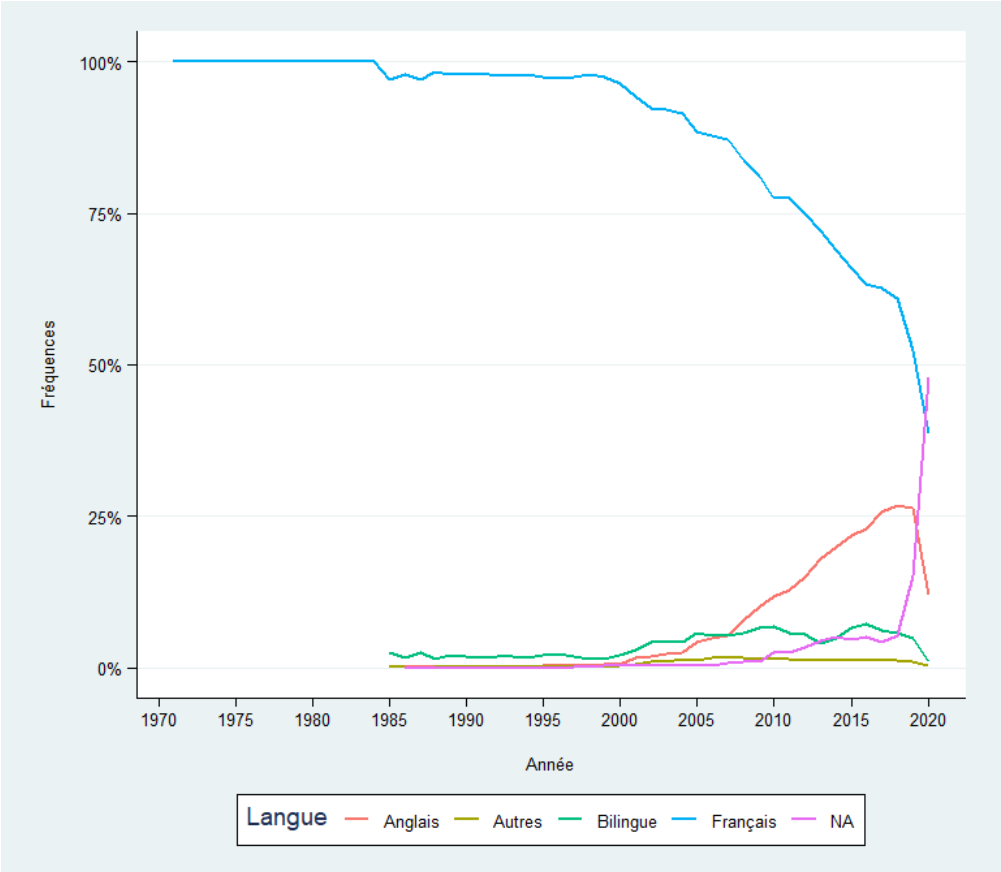



FIGURE 16 – Évolution des fréquences d'utilisation des différentes langues au cours des années

La Figure 16 représente l'évolution d'utilisation de différentes langues au fil des années pour la totalité du jeu de données. Nous pouvons voir quatre périodes, la première de 1971 à 1985 nous permet de constater l'absence de langue non française. La seconde de 1985 à 2000 montre l'apparition d'autres langues avec le français évoluant à ± 95%. La troisième période de 2000 à 2018 montre une croissance de l'anglais passant de ±0% à plus de 25% et une chute du français de 95% à 60%, enfin de 2019 à 2020 montre une chute de toutes les langues et une augmentation des données manquantes.
Cette augmentation de l'anglais est probablement dûe au fait que l'anglais soit devenu la langue mondiale en terme de communication.

### 5.2.2 Période 2004 à 2018

Pourquoi commencer en 2004 ? Commencer en 2004 semble un choix pertinent dû à la mise en place du système LMD (licence-master-doctorat) en France qui consiste à harmoniser le système universitaire au niveau européen. Étudier cette période permet donc de voir l'évolution des langues après sa mise en place et donc de constater les changements qui en découlent. La période choisie se finit en 2018 pour éviter les effets de bord dûs aux données manquantes après cette année.

Calculons les pourcentages d'augmentation entre 2004 et 2018 : Table 4

Nous pouvons constater en lien avec la Figure 17 les évolutions des langues.

| Langue | count 2004 | count 2018 | % changement |
|---|---|---|---|
| Français | 9371 | 7807 | -16.7 |
| Anglais | 267 | 3429 | 1184.3 |
| Bilingue | 435 | 741 | 70.3 |
| Autres | 137 | 155 | 13.1 |
| NA | 40 | 464 | 1060 |

TABLE 4 – Évolution des langues pour la période 2004 à 2018

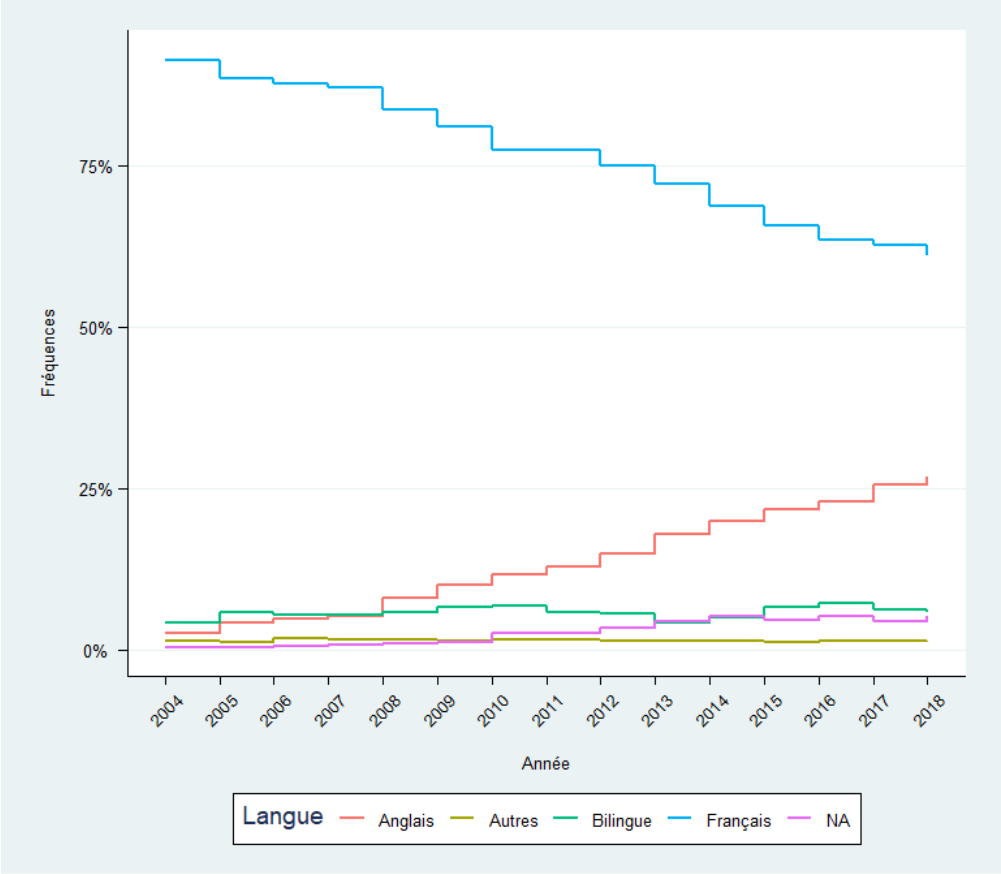



FIGURE 17 – Évolution des fréquences d'utilisation des différentes langues au cours des années pour la période de 2004 à 2018

La Figure 17 montre l'évolution des fréquences d'utilisation des différentes langues au cours des années. Nous pouvons constater que le français chute de ±80% à ±63%, l'anglais a augmenté de ±5% à ±27%. Les autres langues ainsi que les thèses soutenues en bilingue restent constantes avec ±2% et ±6% respectivement. Une augmentation de ±0% à ±6% des données manquantes est à noter.

## 5.3 Référence bibliographique

Comme le souligne Martin (2015), le signalement des thèses de doctorat : «Ce sont 10 000 doctorats environ qui sont délivrés en France chaque année. Ce chiffre augmente continuellement, de 32 % entre 2005 et 2012».[1]Nous allons donc vérifier ce chiffre.

La Table 5 représente le total des langues pour les années 2005 et 2012 et va nous servir pour calculer le pourcentage d'augmentation.

Pour calculer le pourcentage d'augmentation, il faut :

| Langue | count 2005 | count 2012 |
|---|---|---|
| Français | 9352 | 10477 |
| Anglais | 437 | 2089 |
| Bilingue | 611 | 771 |
| Autres | 121 | 184 |
| NA | 40 | 464 |
| Total | 10561 | 13985 |

TABLE 5 – Total langue pour les années 2005 et 2012

1. Calculer le coefficient d'accroissement :
$$\frac{13985}{10561} = 1.32$$
2. Multiplier par 100 pour obtenir un pourcentage
$$1.32 * 100 = 132\%$$
3. Soustraire 100% car nous posons que la valeur initiale était de 100%
$$132\% - 100\% = 32\%$$

Nous trouvons donc le même résultat de 32% pour la même période.

## 5.4 Conclusion

Nous nous sommes proposé d'étudier l'évolution des langues utilisées pour les soutenances des thèses. Durant cette analyse, nous nous somme focalisé en particulier sur la période allant de 2004 à 2018 pour étudier les changements de langues du au système LMD.
Si cette analyse ne nous permet pas d'affirmer que le système LMD a influencé le choix des langues des thèses soutenues, elle nous a permis de constater que l'anglais est en augmentation constante depuis le début des années 2000.

# Chapitre 6

# SQL

Question 1 : Quelle est la durée du plus court film produit entre 1942 et 1968 inclus ?
Open Secret en 1948 avec une durée de 68 minutes.

Question 2 : Quelle est la durée moyenne des films réalisés entre 1954 et 1967 inclus ?
La durée moyenne des films réalisés entre 1954 et 1967 est de 132.80 minutes.

Question 3 : Sur la période 1960-1970 (inclus), combien de langues distinctes ont-elles été utilisées dans les films ?
Il y a quatre langues différentes. (French, German, Italian, English)

Question 4 : Comptez combien de films ont été produits après les années 2000 en français ou en espagnol.
Il y a eut 100 films produit.

Question 5 : Identifiez le film en langue française ayant le plus rapporté d'argent entre 1990 et 1999 inclus.
The Red Violin, 1998, 9473382 dollars de recettes.

Question 6 : Donnez 5 films en anglais commençant par la lettre Z, toutes époques confondues.
Zero Effect, Zoolander, Zoom, Zodiac, Zombieland


| | INCORP_DATE | NAME | STATE_ID | CUST_ID |
|---|---|---|---|---|
| 1 | 1995-05-01 | Chilton Engineering | 12-345-678 | 10 |
| 2 | 2001-01-01 | Northeast Cooling Inc. | 23-456-789 | 11 |
| 3 | 2002-06-30 | Superior Auto Body | 34-567-890 | 12 |
| 4 | 1999-05-01 | AAA Insurance Inc. | 45-678-901 | 13 |

FIGURE 18 – Table résultat manipulation de Wampserver

# Chapitre 7

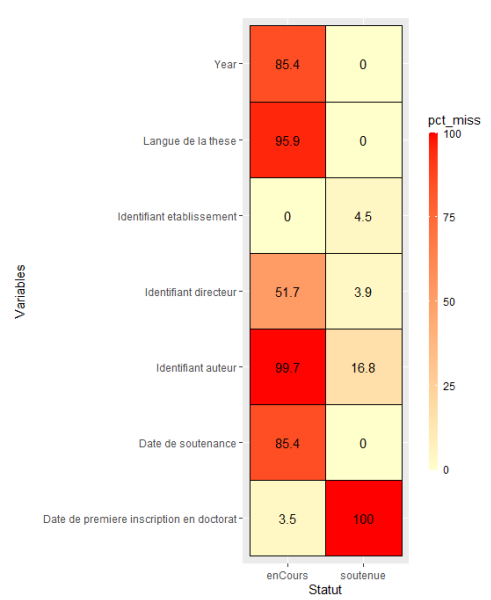
# Travail en bonus

## 7.1 Présentation

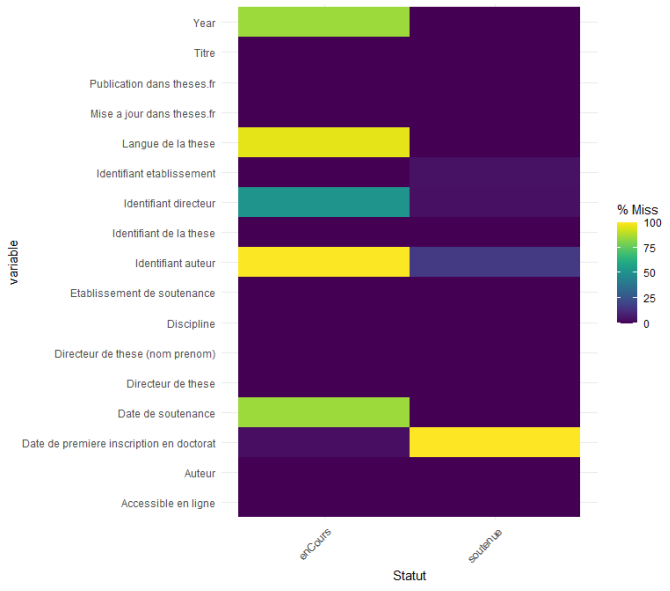
Ce chapitre sera dédié aux données manquantes avec une approche différente du chapitre 2, à l'évolution des rapports homme/femme et des langues au sein des disciplines universitaires et enfin au web scraping pour reproduire le jeu de données.
Nous utiliserons aussi un nouveau jeu de données nommé PhD_V3. Celui-ci est nettoyé et contient aussi de nouvelles variables telles que le genre ou bien un recodage des disciplines.

## 7.2 Données manquantes

Dans cette section nous allons représenter les données manquantes en utilisant des heatmap.




(a) Heatmap des données manquantes par Statut




(b) Heatmap des données manquantes par Statut pour l'ensemble du jeu de données

FIGURE 19 – Heatmap des données manquantes

## 7.3 Problèmes discipline universitaire

### 7.3.1 Problème des genres en fonction des disciplines

#### PhD V2

Le genre des auteurs n'étant pas présent de base sur le jeu de données PhD_V2, il a été ajouté en utilisant la librairie Genderguesser. Notons que seules les cinq disciplines les plus représentées sont utilisées ci-dessous.
Les deux graphiques suivants serviront d'exemple pour montrer qu'il est possible de présenter une même donnée de façon différente ayant pour résultat une lecture différente de celle-ci.

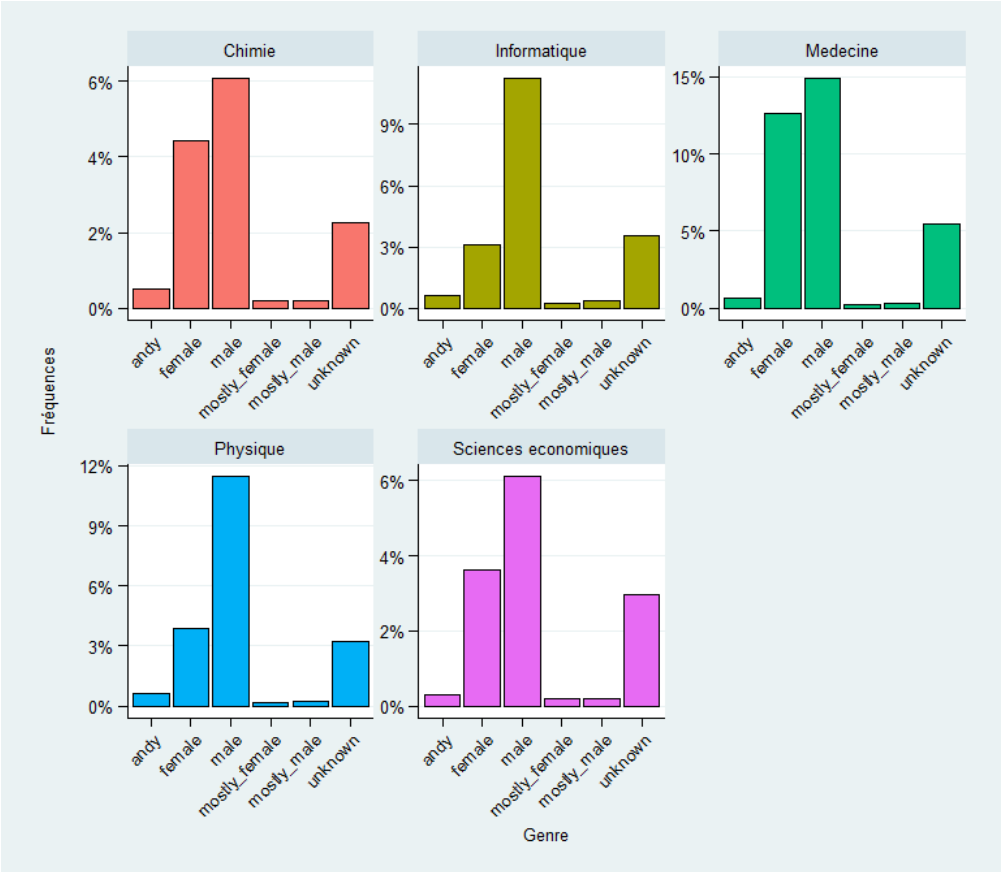



FIGURE 20 – Facet wrap des pourcentages des genres par disciplines

La Figure 20 représente le pourcentage des genres par disciplines. Nous pouvons constater une dominance des hommes dans toutes les disciplines. Notons la quasi parité homme/femme en médecine et ainsi qu'en chimie. Ce type de graphique permet d'observer les distributions de la variable cible par catégorie.

La Figure 21 représente les fréquences relatives de chaque discipline par genre. Nous pouvons voir qu'environ 46% des femmes sont en médecine. Ce type de graphique permet de voir la distribution au sein d'une même catégorie.

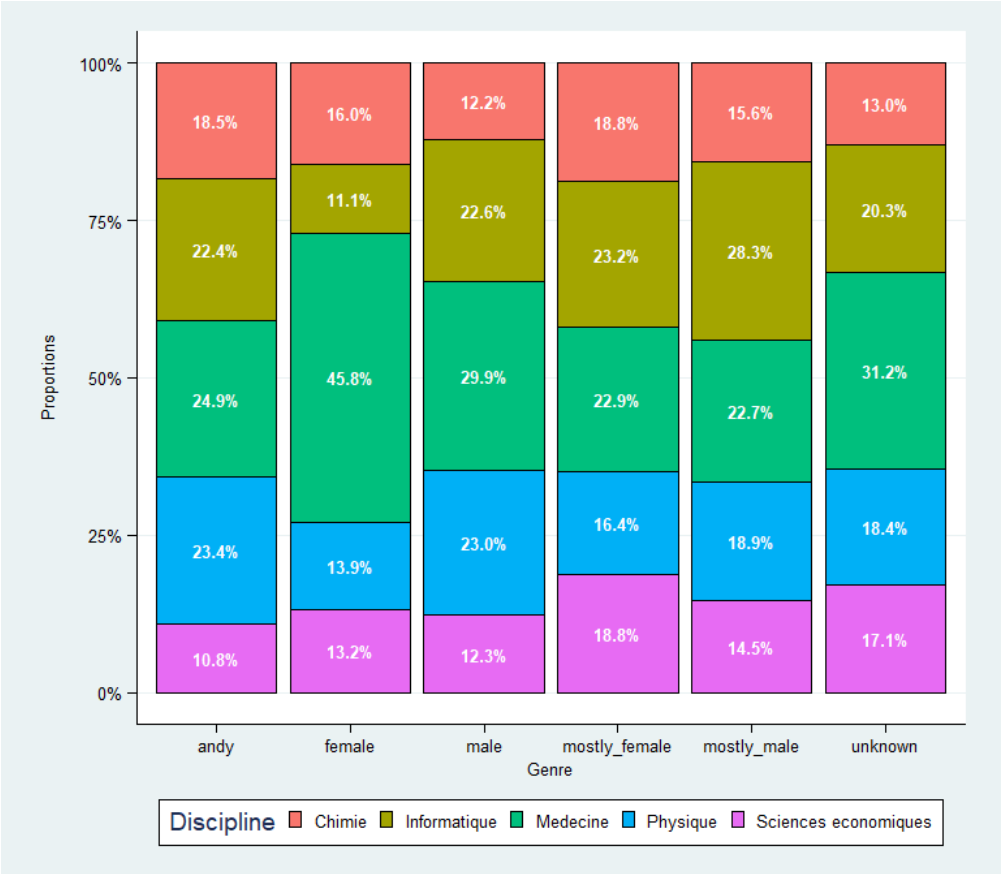



FIGURE 21 – Fréquences relatives de chaque discipline par genre

**PhD V3**

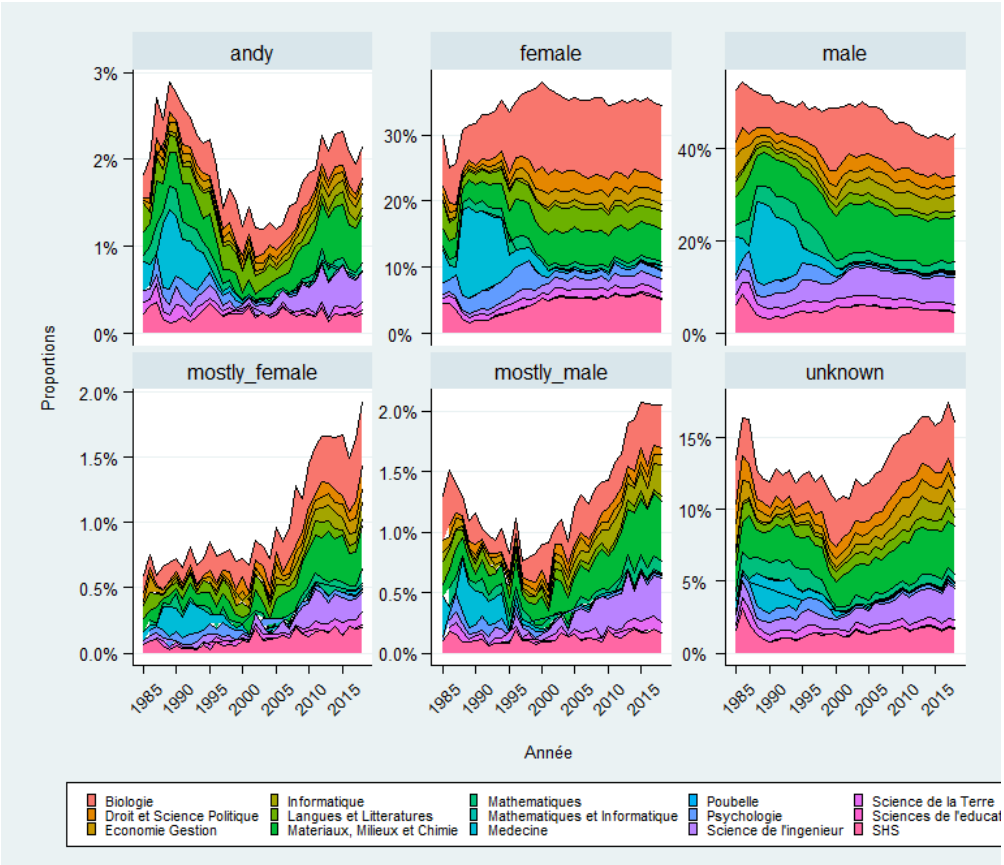



FIGURE 22 – Evolution des genres par discipline au fil des ans, période de 1985 à 2018

La Figure 22 représente l'évolution des proportions des genres par discipline au fil des ans. Nous pouvons constater que de 1990 à 1995 les femmes en médecine représente ± 12 % du total. On peut aussi constater après 1995 les femmes sont majoritairement représentées en biologie.

### 7.3.2 Problème des langues en fonction des disciplines

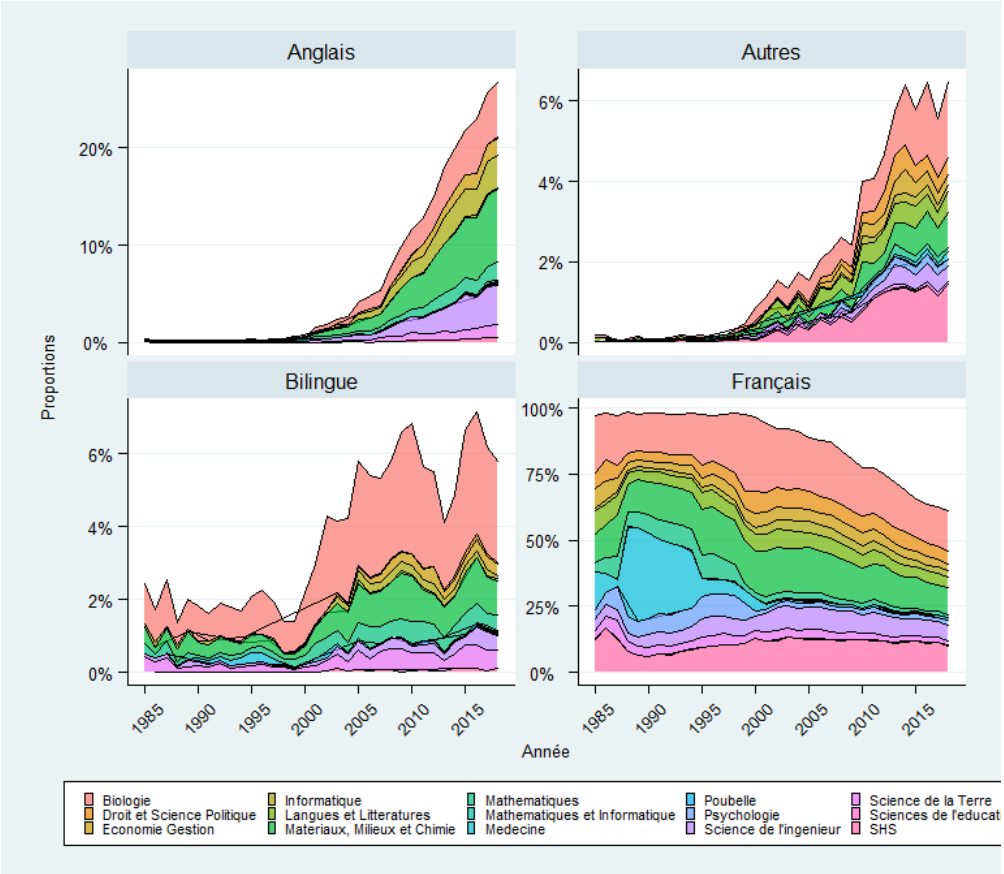



FIGURE 23 – Evolution des langues par discipline au fil des ans, période de 1985 à 2018

La Figure 23 représente l'évolution de l'utilisation des langues par discipline au fil des ans. Nous pouvons constater la chute du français, comme vue lors du chapitre cinq. La langue française est majoritairement utilisée en biologie (± 12,5%), matériaux, milieux et chimie (± 10%) et SHS (± 12,5%) après les années 2000.

## 7.4 Web scraping

Seule la table résultante du web scraping sera montrée en deux parties.

| | Auteur | Titre | Discipline | Directeur |
|---|---|---|---|---|
| 1 | Djelloul Kab | La responsabilité médicale en... | Droit privé | , |
| 2 | Hamza Tarin | Caractérisation et optimisatio... | Sciences de l'ingénieur | Abdellah Arhaliass, |
| 3 | Farah Bibi SHAIK DAWOOD | Formulation et procédés de p... | Sciences de l'ingénieur | Abdellah Arhaliass, Raphaëlle Savoire. |
| 4 | Riad EL HAMOUD | Réduction de la fatigue des é... | Sciences de l'ingénieur | Abdul-Hamid Soubra, Mourad Ait-ahmed. |
| 5 | Simon Husser | Privé et public en droit pénal | Droit pénal | Agathe Lepage |
| 6 | Marthe Cachard-Chastel | Synergie d'effets neurochimi... | Pharmacologie expérimentale et clinique | Alain Gardier |
| 7 | Alexandre Derre | Douleurs chroniques : implica... | Biologie Santé | Alexandre Pattyn |
| 8 | Lara Aldaou | Etude expérimentale et numé... | Sciences de l'ingénieur | Ali-Nordine Leklou, |
| 9 | Elise Madec | Etude multiparamétrique de l... | Physique | Amanda Silva brun |
| 10 | Diane Letourneur | Impact combiné de toxines d... | Sciences de la vie et de la sante | Amel Mettouchi |
| 11 | Clara Gandrez | Modèle comportemental d'id... | Conception (AM) | Améziane Aoussat |
| 12 | Adèle Kauffmann | Les effets du brexit sur la cito... | Droit de l'Union européenne | Anastasia Iliopoulou |
| 13 | Jeanne-Valérie Hell | Histoire, imaginaire dans "les... | Littérature française | André Daspre |
| 14 | Trang Nguyen Vinh | Analogie entre le courant éle... | Géographie et aménagement | André Dauphiné |
| 15 | Lucette Heller-Goldenberg | Histoire des auberges de jeun... | Histoire | André Nouschi |
| 16 | Alia Lakhoua | Le tissage de la soie à tunis d... | Histoire et civilisation | André Nouschi |
| 17 | Pierre Bicaba Nanye | La crise économique de 1929... | Histoire | André Nouschi |
| 18 | Monique Jacomino-Laborieux | L'algérie coloniale 1830-1962... | Histoire | André Nouschi |
| 19 | Angga Perima | Combinatorial antibiotic scre... | Chimie physique et chimie analytique | Andrew D. Griffiths |
| 20 | Patie Cendra Rakotoarimanana | Nanoscale surface engineerin... | Chimie | Anne-Marie Gonçalves |
| 21 | Renata Andrade (Da silva andrade) | Le cannibalisme dans l'art co... | Arts | Anne Creissels |
| 22 | Amelie Francois | Une sexualisation virtuelle de... | Arts | Anne Creissels |
| 23 | Clemence Canet | La visite guidée comme perfo... | Arts | Anne Creissels |
| 24 | Hiam Dahanni | Conception environnemental... | Sciences de l'ingénieur | Anne Ventura, André Orcesi. |
| 25 | Athul Kaitheri | Caractérisation des variations... | Sciences de la Planète et de l'Univers | Anthony Mémin, Frédérique Rémy. |
| 26 | Geoffroy Laurin | La rente et le droit des sûreté... | Histoire du Droit | Anthony Mergey |
| 27 | Lucas Tuduri | Croyances motivées et théori... | Sciences économiques | Antoine Billot |
| 28 | Martin Odoh | La détermination de la respo... | Droit Public | Antoine Delblond, Dodzi Kokoroko. |
| 29 | Jalal Elmir | Le secret bancaire, une instit... | Droit des affaires | Antoine Gaudemet |
| 30 | Benjamin Fontaine | Intérêt social et activisme act... | Droit des affaires | Antoine Gaudemet |
| 31 | Jérôme Beaumont | Rôle prépondérant des cellule... | Droit des affaires | Antoine Gaudemet |
| 32 | Alessandro Lauro | La "justiciabilité" du système ... | Droit public | Armel Le Divellec, |
| 33 | Mainak Sarkar | Trois articles sur l'analyse des... | Science de gestion - EM2PSI | Arnaud De Bruyn, Arnaud De Bruyn. |
| 34 | Coline Fonderflick | L'accord collectif de travail à ... | Droit social | Arnaud Martinon |
| 35 | Baptiste Bataille | Création et institutionnalisati... | Sciences de l'information et de la communication | Arnaud Mercier |
| 36 | Guillaume Perissat | Les enjeux de la communicati... | Sciences de l'information et de la communication | Arnaud Mercier |
| 37 | Clarisse NYNGONE MAYAZA | Consommateurs de soins et s... | Droit | Augustin Emane |
| 38 | Leslie-anne Merleau | Effets à l□échelle physiologiq... | Dynamique des milieux naturels et anthropises pass... | Aurélie Goutte, Olivier Lourdais. |
| 39 | Malak Dia | Nanoparticules et matière or... | Science de la Terre et de l'Environnement | Béatrice Bechet, |
| 40 | Simon Gouzy | Conditions de formation de la... | Sciences de la Terre et des planètes | Benjamin Rondeau, Vassilissa Vinogradoff. |
| 41 | Fang Zhao | Sous-spécification et analyse ... | Sciences du langage - linguistique | Benoit Crabbe |
| 42 | Clara Lahiani (Coudert) | Les datacenters à l'épreuve d... | Droit fiscal | Benoît Delaunay |

FIGURE 24 – Data frame résultant du web scraping partie une

| Etablissment | Statut | Date_inscription | Date_soutenance_ymd | Date_soutenance_y | Link_to_pdf | Langue |
|---|---|---|---|---|---|---|
| Catherine Puigelier, Corinne Pizzio-Delaporte. | en_cours | NA | 2022-02-03 | NA | NA | NA |
| Ahmed Rhallabi, Jamal Fajoui. | en_cours | 2022-02-03 | NA | NA | NA | NA |
| Nantes | en_cours | 2022-02-07 | NA | NA | NA | NA |
| Nantes | en_cours | 2022-01-06 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | NA | 2022-03-15 | NA | NA | NA |
| Université Paris XI | soutenue | NA | NA | 2007 | NA | NA |
| Montpellier | en_cours | NA | 2022-04-08 | NA | NA | fr |
| Nabil Issaadi, Ouali Amiri. | en_cours | 2022-02-25 | NA | NA | NA | NA |
| Université Paris Cité | en_cours | 2021-07-13 | NA | NA | NA | NA |
| Université Paris Cité | en_cours | 2021-07-19 | NA | NA | NA | NA |
| Paris, HESAM | en_cours | NA | 2022-03-31 | NA | NA | fr |
| Université Paris-Panthéon-Assas | en_cours | 2021-10-08 | NA | NA | NA | NA |
| Nice | soutenue | NA | NA | 1988 | NA | fr |
| Nice | soutenue | NA | NA | 1985 | NA | fr |
| Nice | soutenue | NA | NA | 1985 | NA | fr |
| Nice | soutenue | NA | NA | 1987 | NA | fr |
| Nice | soutenue | NA | NA | 1988 | NA | fr |
| Nice | soutenue | NA | NA | 1985 | NA | fr |
| Paris 6 | soutenue | NA | 2017-12-11 | NA | NA | en |
| université Paris-Saclay | en_cours | NA | 2021-11-16 | NA | NA | fr |
| Paris 8 | en_cours | 2021-10-16 | NA | NA | NA | NA |
| Paris 8 | en_cours | 2021-11-26 | NA | NA | NA | NA |
| Paris 8 | en_cours | 2021-10-27 | NA | NA | NA | NA |
| Nantes | en_cours | 2021-12-03 | NA | NA | NA | NA |
| Université Côte d'Azur | soutenue | NA | 2021-12-02 | NA | theses.fr/2021COAZ41... | en |
| Université Paris-Panthéon-Assas | en_cours | 2022-01-26 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2021-10-28 | NA | NA | NA | NA |
| Nantes | en_cours | 2020-01-13 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | NA | 2022-01-21 | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2021-10-28 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2019-09-30 | NA | NA | NA | NA |
| Luigi Benvenuti, Marco Mancini. | en_cours | 2019-09-30 | NA | NA | NA | NA |
| CY Cergy Paris Université | en_cours | 2017-09-01 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2021-11-18 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2021-10-12 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | 2022-01-06 | NA | NA | NA | NA |
| Nantes | en_cours | 2021-12-16 | NA | NA | NA | NA |
| Université Paris sciences et lettres | en_cours | 2021-09-01 | NA | NA | NA | NA |
| Denis Courtier-muras, Pierre-Emmanuel Peyneau. | en_cours | 2021-12-16 | NA | NA | NA | NA |
| Nantes | en_cours | 2021-10-20 | NA | NA | NA | NA |
| Université Paris Cité | en_cours | 2021-09-08 | NA | NA | NA | NA |
| Université Paris-Panthéon-Assas | en_cours | NA | 2022-01-15 | NA | NA | NA |

FIGURE 25 – Data frame résultant du web scraping partie deux

# Table des figures

# Liste des tableaux

# Bibliographie

[1] I Martin. « Le signalement des thèses de doctorat ».
In : *I2D - Information, données & documents* (2015), p. 46-47. DOI : 10.3917/i2d.151.0046.