Analyser un jeu de données sur les thèses en France

Ibrahima LY

Chargement des librairies

In [1]:
```
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")
```

Importation de jeu de données PhD_v1

In [2]:
```
PhD_v1 = pd.read_csv("PhD_v1.csv", encoding="utf-8")
```

I. Présentation des données

1) Présentation de jeu de données PhD_v1

In [3]:
```
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
```

In [4]:
```
PhD_v1 .shape
```

Out[4]: (447644, 18)

In [5]:
```
print("Le nombre de lignes dans le jeu de données PhD_v1 est :", PhD_v1.s
      "colonnes")
```

```
Le nombre de lignes dans le jeu de données PhD_v1 est : 447644 lignes et
18 colonnes
```

In [6]:
```
print("le nombre de lignes de jeu jeu de données est de: ", len(PhD_v1))
print('Je constate que toutes les données sont bien chargées ')
```

```
le nombre de lignes de jeu jeu de données est de:  447644
Je constate que toutes les données sont bien chargées
```

2. Suppression de jeu de données PhD_v1 de la mémoire

In [7]:
```
del PhD_v1
```

Importation de jeu de données PhD_v2

In [8]:
```
PhD_v2 = pd.read_csv("PhD_v2.csv", encoding="utf-8")
```

In [9]:
```
PhD_v2_copy = PhD_v2
```

1) Présentation de jeu de données PhD_v2

In [10]:
```
# Affichage des 5 premières lignes
PhD_v2_copy.head(3)
```

Out[10]:

	Unnamed: 0	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur d these (nor prenom
0	0	Saeed Al marri	NaN	Le credit documentaire et l'onopposabilite des...	Philippe Delebecque	Delebecqu Philipp
1	1	Andrea Ramazzotti	174423705	Application de la PGD a la resolution de probl...	Jean-Claude Grandidier,Marianne Beringhier	Grandidier Jear Claude,Beringhie Mariann
2	2	OLIVIER BODENREIDER	NaN	Conception d'un outil informatique d'etude des...	Francois Kohler	Kohler Franco

3 rows × 23 columns

In [11]:
```
# On peut supprimer la variable "Unnamed: 0" car elle n'apporte aucune in
PhD_v2_copy.drop('Unnamed: 0', axis=1, inplace=True)
```

In [12]:
```
PhD_v2_copy.rename(columns={"Discipline_prÃ©di": "Discipline_predi"}, inp
```

In [13]:
```
# Get DataFrame information
PhD_v2_copy.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 448047 entries, 0 to 448046
Data columns (total 22 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   Auteur                                    448047 non-null  object
 1   Identifiant auteur                        317700 non-null  object
 2   Titre                                     448040 non-null  object
 3   Directeur de these                        448034 non-null  object
 4   Directeur de these (nom prenom)           448034 non-null  object
 5   Identifiant directeur                     448047 non-null  object
 6   Etablissement de soutenance               448046 non-null  object
 7   Identifiant etablissement                 430965 non-null  object
 8   Discipline                                448047 non-null  object
 9   Statut                                    448047 non-null  object
 10  Date de premiere inscription en doctorat  64331 non-null   object
 11  Date de soutenance                        390961 non-null  object
 12  Year                                      390961 non-null  float64
 13  Langue de la these                        448047 non-null  object
 14  Identifiant de la these                   448047 non-null  object
 15  Accessible en ligne                       448047 non-null  object
 16  Publication dans theses.fr                448047 non-null  object
 17  Mise a jour dans theses.fr                447870 non-null  object
 18  Discipline_predi                          448047 non-null  object
 19  Genre                                     448047 non-null  object
 20  etablissement_rec                         444973 non-null  object
 21  Langue_rec                                383927 non-null  object
dtypes: float64(1), object(21)
memory usage: 75.2+ MB
```

a) Dimensions de la base

In [14]:
```
PhD_v2_copy.shape
```

Out[14]:
```
(448047, 22)
```

In [15]:
```
print("Le nombre de lignes dans le jeu de données PhD_v2_copy est :", PhD
      PhD_v2_copy.shape[1], "colonnes")
```

```
Le nombre de lignes dans le jeu de données PhD_v2_copy est : 448047 ligne
s et 22 colonnes
```

Type de jeu de données

In [16]:
```
type(PhD_v2_copy)
```

Out[16]:
```
pandas.core.frame.DataFrame
```

b) Les noms des variables de jeu de données PhD_v2:

In [17]:
```
PhD_v2_copy.columns
```

Out[17]:
```
Index(['Auteur', 'Identifiant auteur', 'Titre', 'Directeur de these',
       'Directeur de these (nom prenom)', 'Identifiant directeur',
       'Etablissement de soutenance', 'Identifiant etablissement',
       'Discipline', 'Statut', 'Date de premiere inscription en doctora
t',
       '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',
       'Discipline_predi', 'Genre', 'etablissement_rec', 'Langue_rec'],
      dtype='object')
```

c) Typologie des variables

In [18]:
```
PhD_v2_copy.dtypes
```

Out[18]:
```
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 PhD_v2 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.

In [ ]:
```
```

d) Nombre de valeurs non vide

In [19]:
```
PhD_v2_copy.count().sort_values()
```

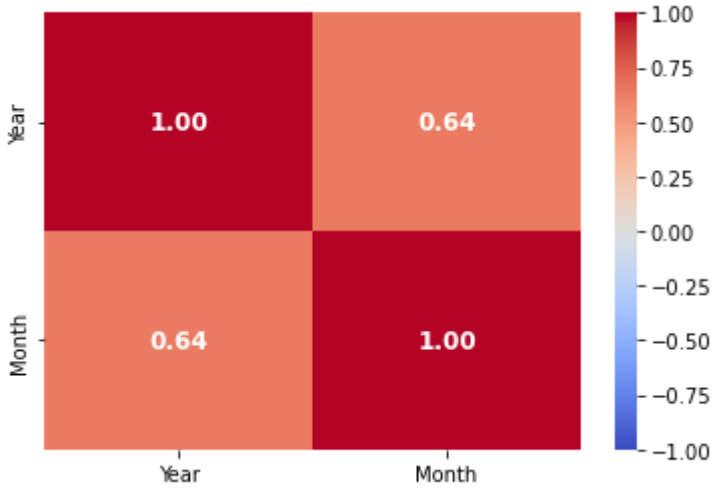
Out[19]:
```
Date de premiere inscription en doctorat     64331
Identifiant auteur                          317700
Langue_rec                                  383927
Year                                        390961
Date de soutenance                          390961
Identifiant etablissement                   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.

In [69]:
```
corr_matrix = PhD_v2_copy.corr()

sns.heatmap(corr_matrix, cmap='coolwarm', annot=True, vmin=-1, vmax=1,
            fmt='.2f', annot_kws={'fontsize': 12, 'fontweight': 'bold'})

mask = np.triu(np.ones_like(corr_matrix, dtype=bool))
corr_matrix_red = corr_matrix.where(~mask, np.nan)
corr_matrix_red = corr_matrix_red[corr_matrix_red < 0]
sns.heatmap(corr_matrix_red, cmap='Reds', annot=True, vmin=-1, vmax=0, fm
            annot_kws={'fontsize': 12, 'fontweight': 'bold', 'color': 'bl
plt.savefig('fi')
plt.show()
```



In [20]:
```
for column in PhD_v2_copy.columns:
    print(f"Contenu unique de la colonne {column} :")
    print(PhD_v2_copy[column].unique())
    print("\n")
```

```
Contenu unique de la colonne Auteur :
['Saeed Al marri' 'Andrea Ramazzotti' 'OLIVIER BODENREIDER' ...
 'Nesrine Salah' 'Ghulam sakhi Shokouh' 'Helene Labriet (Rouge-labriet)']


Contenu unique de la colonne Identifiant auteur :
[nan '174423705' '182410528' ... '244931399' '246559543' '248077481']


Contenu unique de la colonne Titre :
["Le credit documentaire et l'onopposabilite des exceptions"
```

```
 "Application de la PGD a la resolution de problemes transitoires couples
en vue de l'allegement des structures composites."
 "Conception d'un outil informatique d'etude des cinetiques observees en
toxicologie clinique"
 ...
 'Modelisation du comportement mecanique des betons avec prise en compte
des proprietes interfaciales : influence du durcissement et de la lixivia
tion'
 "Detection et classification d'objets dans des images numeriques"
 "Developpement de l'imagerie X biomedicale en contraste de phase par tav
elures"]


Contenu unique de la colonne Directeur de these :
['Philippe Delebecque' 'Jean-Claude Grandidier,Marianne Beringhier'
 'Francois Kohler' ...
 'Anne-Sophie Caro,Moulay Said El Youssoufi,Etienne Malachanne'
 'Philippe Montesinos,Baptiste Magnier' 'Sylvain Bohic,Barbara Fayard']


Contenu unique de la colonne Directeur de these (nom prenom) :
['Delebecque Philippe' 'Grandidier Jean-Claude,Beringhier Marianne'
 'Kohler Francois' ...
 'Caro Anne-Sophie,El Youssoufi Moulay Said,Malachanne Etienne'
 'Montesinos Philippe,Magnier Baptiste' 'Bohic Sylvain,Fayard Barbara']


Contenu unique de la colonne Identifiant directeur :
['29561248' '715,441,511' '57030758' ... '704,488,921' '1.82E+08'
 '156,614,561']


Contenu unique de la colonne Etablissement de soutenance :
['Paris 1'
 "Chasseneuil-du-Poitou, Ecole nationale superieure de mecanique et d'aer
otechnique"
 'Nancy 1' 'Lyon 2' 'Paris 5' 'Saint Etienne' 'La Reunion' 'Paris 8'
 'Nantes' 'Toulouse 1' 'Montpellier 3' 'Amiens' 'Paris, EHESS' 'Brest'
 'Poitiers' 'Perpignan' 'Lille 2' 'Strasbourg' 'Rennes 2' 'Angers'
 'Orleans' 'Lyon 3' 'Paris 2' 'Corte' 'Paris 9' 'Rennes 1'
 'Universite de Lorraine' 'Avignon' 'Littoral'
 'Versailles-St Quentin en Yvelines' 'Lorient' 'Lille 3' 'Toulouse 2'
 'Rouen' 'Pau' 'Artois' 'Paris 13' 'Valenciennes'
 'Cachan, Ecole normale superieure' 'Tours' 'Le Havre' 'La Rochelle'
 'Le Mans' 'Bordeaux 3' 'Jouy-en Josas, HEC' 'Dijon' 'Grenoble' 'Mulhous
e'
 'Caen' 'Lille 1' "Paris, Institut d'etudes politiques"
 'Polynesie francaise'
 'Ecole doctorale Pratiques et theories du sens (Saint-Denis, Seine-Saint
-Denis)'
 'Bordeaux 4' 'Dunkerque' 'Paris 4' 'Antilles-Guyane'
 'Paris Sciences et Lettres' "Evry-Val d'Essonne"
 'Paris, Ecole nationale des chartes' 'Paris 10' 'Toulon' 'Paris, INALCO'
 'Limoges' 'Montpellier 1' 'Montpellier' 'Reims' 'Cergy-Pontoise'
 'Cergy Pontoise' 'Toulouse 3' 'Lyon 1' 'Nice' 'Nancy 2'
 'Villeurbanne, INSA' 'Paris 6' 'Grenoble 1' 'Montpellier 2' 'Compiegne'
 'Universite de Paris-Nanterre. UFR de sciences juridiques, politiques et
administratives'
 'Vandoeuvre-les-Nancy, INPL'
 'Universite Paris-Est Creteil Val de Marne (UPEC)' 'Paris 11' 'Besancon'
 'Paris, ENMP' 'Paris Est' 'Faculte de medecine et de pharmacie (Lyon)'
 'Universite de Lyon (1896-1970)' 'Palaiseau, Ecole polytechnique'
 'Paris%202' 'Rouen, INSA'
 'Paris Ecole des hautes etudes en sciences sociales' 'Paris, EPHE'
```

'Universite Lyon 2' 'Montpellier%203' 'Aix-Marseille'
'Paris 13 - Sorbonne Paris Cite'
"Ecole nationale des Mines d'Albi-Carmaux" 'Aix Marseille 2'
'Rennes, INSA' 'Bordeaux' 'Paris4' 'Metz' 'Toulouse, INPT' 'Paris 7'
'Nantes, Ecole nationale veterinaire' 'Paris Saclay'
'Universite de Nancy I' 'Saint-Etienne' 'Paris, CNAM' 'Bordeaux 2'
'Lyon, INSA' 'Lyon' 'Nouvelle Caledonie' 'Aix-Marseille 3'
'Universite Toulouse 1 Capitole' 'Clermont-Ferrand 1' 'Antilles'
'Universite de Rennes 1' 'Normandie' 'Sorbonne Paris Cite'
'Science de la vie et de la Sante' 'Aix-Marseille 1'
'Institut National des Langues et Civilisations Orientales'
'FernUniversitat Hagen' 'Ecole nationale superieure de chimie (Rennes)'
'Paris%20Est' 'Aix-Marseille 2'
'Rennes, Ecole nationale superieure de chimie' 'Grenoble INPG'
'Paris, Ecole normale superieure' 'Ecole centrale de Lille'
'Universite Savoie Mont Blanc' 'Universite Louis Pasteur (Strasbourg)'
'Montreal' 'Universite de Provence. Departement de sociologie'
'Universite Marc Bloch (Strasbourg)'
'Universite Henri Poincare Nancy 1. Faculte des sciences et techniques'
'Ecully, Ecole centrale de Lyon' 'Saint-Etienne, EMSE'
'Montpellier, SupAgro' 'Paris 3' 'Clermont Auvergne' 'Bordeaux 1'
'Evry, Institut national des telecommunications' 'Clermont-Ferrand 2'
'Chatenay-Malabry, Ecole centrale de Paris'
'Lyon, Ecole normale superieure' 'Paris, AgroParisTech'
'Montpellier, Ecole nationale superieure de chimie' 'Belfort-Montbeliar
d'
'Toulouse, INSA' 'Bourgogne Franche-Comte' 'Supelec' 'Strasbourg 2'
'Cergy-Pontoise, Ecole superieure des sciences economiques et commercial
es'
'Rennes, Agrocampus Ouest' 'Telecom Bretagne'
'Ecole nationale superieure Mines-Telecom Atlantique Bretagne Pays de la
Loire'
'Strasbourg 1' 'Mines Paris' 'Universite de Strasbourg'
'Universite de Recherche Paris Sciences et Lettres - PSL Research Univer
sity'
'Universite francaise du Pacifique (1987-1999)' 'Toulouse, ISAE'
"Paris, Museum national d'histoire naturelle"
"Saint-Etienne, Ecole nationale d'ingenieurs"
'Universit%C3%A9%20de%20Lorraine' 'Paris, Inalco'
'ePalaiseau, Ecole polytechnique'
'Vaulx-en-Velin, Ecole nationale des travaux publics'
'Ecole centrale de Nantes'
'Universite de Versailles-Saint-Quentin-en-Yvelines'
'Ecole Centrale de Nantes' 'Universite Bretagne Loire'
'Universite Nantes-Angers-Le Mans - COMUE' 'Paris, ENST'
'Institut de physique du globe (Paris)' 'Troyes'
'Nantes, Ecole des Mines' 'Ecole centrale de Marseille' 'Nimes'
'Bourges, INSA Centre Val de Loire' 'Universite de Bourgogne'
'Ecole normale superieure (Lyon)'
'Centre de gestion scientifique (Paris)' 'Ecole Centrale Paris'
'Lyon, Universite de Lyon' 'Sorbonne universite' 'Nutrition humaine'
'Universite de Perpignan. UFR de droit et sciences economiques'
'Universite de Franche-Comte' 'Guyane'
'Paris, Institut d&aposetudes politiques' 'Paris, ENSAM' 'Naples'
'Universite de Bretagne occidentale'
'Ecole normale superieure-Lettres et sciences humaines (Lyon 2000-200
9)'
'Universite de Perpignan' 'Universite Robert Schuman (Strasbourg)'
'Ecole nationale superieure des industries agricoles et alimentaires (Ma
ssy, Essonne)'
'Nouvelle%20Cal%C3%A9donie' 'Paris 12'
'Universite Pierre Mendes France (Grenoble)'
'Universite Stendhal (Grenoble)' 'Universite de Marne-la-Vallee'
'Universite Nancy 2' 'Rennes, Ecole normale superieure' 'ENSMP'

'Universite de Provence. Faculte des lettres et sciences humaines'
'Ecole polytechnique' 'Versailles-St Quentin-en-Yvelines'
'Faculte de Medecine de Rennes' 'PARIS 6, PITIE SALPETRIERE' 'TOULOUSE 3'
'dijon' 'clermont ferrand 1' 'Paris'
'Institut national agronomique Paris-Grignon'
'Universite de Limoges. Faculte des sciences et techniques'
'Ecole nationale superieure Mines-Telecom Lille Douai' 'Lille%203'
'Sorbonne%20universit%C3%A9' 'IMT Mines Ales'
'Universite Joseph Fourier (Grenoble)'
'Universite de Strasbourg (1538-1969). Faculte de medecine'
'Marne-la-vallee, ENPC' 'paris 8' 'SAINT ETIENNE' 'INSA de Rouen'
'Antilles Guyane' 'bordeaux 1' 'Paris. ENSAM' 'INSA ROUEN'
'Universite Paris-Sud'
'Palaiseau, Ecole nationale superieure de techniques avancees'
'PARIS Pantheon Sorbonne' 'Evry, Telecom & Management SudParis'
'Jouy-en-Josas, HEC' 'Evry, Institut National des Telecommunications'
"Paris, Institut national d'agronomie de Paris Grignon"
"Palaiseau, Institut d'optique theorique et appliquee"
'Paris, Institut agronomique, veterinaire et forestier de France'
'Laboratoire Central des Ponts et Chaussees (France)' 'Paris%2C%20EHESS'
'Clermont 1' 'AgroParisTech'
"Chasseneuil-du-Poitou, Ecole nationale superieure de mecanique et d'aeronautique"
'Jouy-en-Josas, EHEC' 'Aix-marseille' 'LYON 3' 'Ecole normale superieure'
'Paris INALCO' 'Universite des Antilles et de la Guyane'
'Rennes, Agrocampus' 'St Etienne du Rouvray, INSA' 'Savoie-Chambery'
"Cote d'Azur"
'Universite de Nancy I. UFR Sciences pharmaceutiques et biologiques'
'Ecole nationale superieure agronomique (Montpellier)'
"Ecole nationale superieure de l'aeronautique et de l'espace (Toulouse 1972-2007)"
'Universite de Tours. UFR de medecine'
"Ecole doctorale Sciences de l'homme et de la societe (Tours 1996-2018)"
'Ecole polytechnique universitaire (Tours)' 'CentraleSupelec'
'Grenoble Alpes' 'Universite de Nice' 'Poitiers I.A.E.' 'Nancy2'
'Rennes I' 'Paris1' 'Montpellier SupAgro' 'ENSAM'
'Ecole nationale superieure de ceramique industrielle (Limoges)'
'Laboratoire de Psychologie. Processus de Pensee (Angers)'
'LGIPM - Laboratoire de Genie Industriel et Production de Metz - EA3096'
'Laboratoire dInformatique de Nantes Atlantique (UMR 6241) (Nantes)'
"CIML - Centre d'Immunologie Marseille-Luminy" 'Chambery'
'Paris 2, Laval(Quebec)' 'Universie de Franche-Comte' 'Bucarest'
'Paris 08' 'Sorbonne%20Paris%20Cit%C3%A9'
'Ecole nationale superieure Mines-Telecom Lille Douai (IMT Lille Douai)'
'Ecole Centrale de Lille' 'INALCO'
'Paris, Ecole pratique des hautes etudes'
'Universite de Paris (2019-....)' 'Montpeliier 1'
'Brest, Ecole nationale superieure de techniques avancees Bretagne'
'Clermont-Ferrand' 'HEC Paris'
'Groupe de recherches socio-economiques (Toulouse)' 'Rennes 1,'
'Ecole Nationale Superieure des telecommunications' 'energetique'
'Paris, Telecom ParisTech' 'Paris,Telecom ParisTech'
'Universite du Poitiers' 'EMP'
'Laboratoire recommande APS et sciences sociales (Strasbourg)'
"Groupe de recherche et d'etude des litteratures et civilisations de la Caraibe et des Ameriques noires (Schoelcher, Martinique)"
'Universite de Fribourg (Suisse). Faculte des sciences economiques et sociales'
'Brest%2C%20%C3%89cole%20nationale%20sup%C3%A9rieure%20de%20techniques%20avanc%C3%A9es%20Bretagne'
'Palaiseau' 'Etablissement de Formation 1'

'Ceske vysoke uceni technicke (Prague)'
'Universite de Franche-comte. UFR des sciences et techniques'
"Centre de recherches sur Hegel et l'idealisme allemand (Poitiers)"
'Universitat politecnica de Catalunya' 'Universitat de Girona'
'Technische Universitat (Chemnitz, Allemagne)' 'Normandie-Universite'
'Chirurgie dentaire'
"Centre de recherches historiques de l'Ouest (Rennes)"
'PARIS 6, DENTAIRE' 'Paris 5 Montrouge' 'PARIS 6, BROUSSAIS'
'Centre hospitalier universitaire Saint-Antoine (Paris)'
'Universit%C3%A9%20de%20Paris%20(2019-....)'
"Centre d'etudes et de recherches sur l'urbanisation du monde arabe (Tou
rs)"
'Ecole nationale des ponts et chaussees (France)'
"Institut d'etudes politiques de Paris"
"Ecole superieure d'interpretes et de traducteurs (Paris)"
'Universite de Limoges. Faculte de medecine et de pharmacie'
'Universite de Provence. U.E.R. de sociologie-ethnologie'
"Universite d'Aix-Marseille (1409-1973)"
'Universite Aix-Marseille. Apprentissage, didactique, evaluation, format
ion'
'Migrations internationales, espaces et societes (Poitiers)'
'Laboratoire de biologie neurovasculaire et mitochondriale integree (Ang
ers)'
'Institut du droit de la paix et du developpement (Nice)'
'Institut de recherche mathematique avancee (Strasbourg)'
'Versailles-Saint Quentin en Yvelines' 'Paris%203'
'Fondation nationale des sciences politiques (France)'
"Universite de Franche-Comte. UFR des Sciences du langage, de l'homme et
de la societe"
'Ecole nationale superieure agronomique de Rennes (1961-2004)'
"Centre d'etudes et de recherches caraibeennes (Pointe-a-Pitre, Guadelou
pe)"
'Ecole nationale du genie rural, des eaux et des forets (Paris Nancy)'
'Universite de Nantes. Faculte de droit et des sciences politiques'
'Universidad de La Laguna' 'Nancy II'
'Universite de Poitiers. UFR des sciences fondamentales et appliquees'
'Institut polytechnique de Paris' 'Universite Paris-Saclay (ComUE)'
'Paris, HESAM'
'Universite de Paris-Sud. Faculte de pharmacie (Chatenay-Malabry, Hauts-
de-Seine)'
'Universitat Regensburg'
'Universite de Paris-Sud. Faculte Jean-Monnet. UFR Droit, Economie, Gest
ion'
'Grenoble 2' "Institut d'optique quantique (Hanovre, Allemagne)"
"Universite de Paris-Sud. Faculte des sciences d'Orsay (Essonne)"
'Yale university (New Haven, Conn.)' 'Bordeaux%203'
'Observatoire de Paris' 'Universite de Metz' 'eSorbonne Paris Cite'
'Universite de Toulouse (1896-1968)'
'Universite de Nancy I. Faculte des sciences'
'Universite des sciences sociales (Grenoble)'
"Departement d'etudes germaniques (Aix-en-Provence)"
'Universite de Limoges. Faculte de droit et des sciences economiques'
'Universitatea de Vest din Timisoara'
'Technische Universitat (Dresde, Allemagne)' 'Strasbourg 3'
'Toulouse, ENSAE' 'Toulouse' 'lyon 3' 'Inalco' 'LYON 1' 'Lyon1'
'Paris-Sud' 'Paris, INA' 'Universite de Paris 5' 'PARIS 13' 'Grenoble 3'
'Universite Paris 4' 'Lyon, Ecole normale superieure (sciences)'
'Besancob' 'Paris10' 'Stasbourg 2' 'ENMP' 'Paris 6, Saint-Antoine'
'Paris 6, Pitie-Salpetriere' 'Paris 6, Pitie Salpetriere'
'Universite Lumiere - Lyon 2' 'Universite Paris-Dauphine'
'Chatenay-Malabry, Ecole centrale Paris' 'Sciences de la vie'
"Histoire de l'Art" 'Paris11' 'ENMP, Paris' 'Nancy1' 'Paris, ENGREF'
'Lettres Modernes' 'Victoria University of Wellington' 'ORLEANS'
'Pacifique' 'Etudes italiennes' 'VersaillesSaint-Quentin-en-Yvelines'

'Bordeaux3' 'Nancy'
"Centre d'etudes sur la cooperation juridique internationale"
'GIK institut of engineering sciences and technology (Topi, Pakistan)'
"Institut d'urbanisme de Paris (Creteil, Val-de-Marne)"
'Universite de Poitiers. Departement de geographie'
'[Amiens], Universite de Picardie - Jules Verne, Ecole doctorale en scie
nces humaines et sociales'
"[Amiens], Universite Picardie - Jules Verne, Ecole doctorale de Lettres
et Sciences humaines, UFR d'Economie et de gestion"
'Paris, ENC' 'Physique des polymeres' 'Nante' 'Paris, ENGR'
'Paris, Institut de physique du globe' 'Clermont -Ferrand 2'
'Jouy-en-Josas' 'Informatique' 'Universite Pantheon-Sorbonne (Paris)'
'Roumanie, Universitatea din Craiova'
'INMED - Institut de Neurobiologie de la Mediterranee (Marseille)'
'Chimie Physique' 'Fontenay-aux-Roses, Ecole normale superieure'
'Rennes, ENSA' 'nice' 'Alger' 'Lyon 1,' 'Lund (Suede)' 'Etudes anglaise
s'
'EHESS, Paris' 'Marne-la-Vallee' 'Montpellier, ENSA' 'Renne 1'
'Paris, MNHN' 'Rennes1' 'Universite Paris 8'
'Ecole Normale Superieure de Lyon' 'EHESS'
'Ecole Nationale Superieure des Technologies et Industries du Bois (Epin
al)'
'Universita degli studi di Urbino Carlo Bo (Urbino, Italie)'
'Universita degli studi (Lecce, Italie). Dipartimento di filosofia e sci
enze sociali'
"Centre de recherches sur l'action politique en Europe (Rennes)"
'Cracovie (Pologne), Uniwersytet Jagiellonski'
"Universite d'Aix-Marseille II. Faculte de pharmacie (1970-2011)"
"Institut d'histoire des relations internationales contemporaines (Pari
s)"
'Universitatea de medicina si farmacie Victor Babes (Timisoara, Roumani
e)'
'Universite de Nancy I. Faculte de medecine'
"Universite d'Orleans. Faculte de droit, d'economie et de gestion"
"Laboratoire Transformations de l'appareil productif et structuration de
l'espace social (Nice)"
'Anthropologie bio-culturelle, droit, ethique et sante (Marseille)'
'Ecole polytechnique de Montreal' nan 'Institut de geographie (Rouen)'
'Universite de Nice. Faculte de droit et science politique'
"Universite de Tours. UFR de droit, d'economie et des sciences sociales"
"Institut d'art et d'archeologie (Paris)"
"Universite Paul Cezanne (Aix-Marseille). Faculte d'economie appliquee"
'NICE' 'Ecole des Hautes Etudes en Sciences Sociales' 'Nancy I'
'Saint Etienne du Rouvray, INSA' 'Antilles-guyane' 'Massy, ENSIA'
'Tunis 2' 'Paris,ENST' 'Ecole Nationale d&aposIngenieurs de Brest'
"Ecole Nationale d'Ingenieurs de Brest" 'Nice)'
'Univ. Blaise Pascal - Clermont-Ferrand 2'
'Universite Jean Monnet (St-Etienne)' 'Universite de Besancon' 'Coimbra'
'sDijon' 'Dijon. Histoire du droit' 'Universite Rennes 2' 'Rennes2'
'Mathematiques appliquees' 'Paris, Engref' 'bordeaux 3'
'Laboratoire Information, Milieux, Medias, Mediations (Toulon (Var) Nic
e (Alpes-Maritimes) 2004-2017)'
'Nancy, ENGREF' 'Bordeaux, ENSAM'
"Villeneuve-d'Ascq, Ecole centrale de Lille et Ecole nationale d'ingenie
urs de Tunis"
'Paris, EMP' 'Paris, TELECOM ParisTech' 'Clermont Ferrand 1'
'Chatenay-Malabry, Ecole Centrale de Paris' 'Montpellier III'
'Rennes%2C%20INSA' 'Paris, Ecole des Hautes Etudes en Sciences Sociales'
'paris 5 Necker' 'Aix-en-Provence'
'Universite de Ouagadougou (Burkina-Faso)' 'Paris I'
'Laboratoire de Geographie Physique et Environnementale'
'Ecole Polytechnique Universitaire (Marseille)' 'Ouagadougou'
'INSA Rennes' 'Observatoire Paris' 'Paris X'
"Paris, Institut d'etudes politiques :"

'Universite Grenoble Alpes (ComUE)' "Universite Cote d'Azur (ComUE)"
"Universite Cote d'Azur" 'Rennes%202' '[Grenoble INPG]'
'optoelectronique' 'poitiers'
'Ecole Nationale Superieure des Telecommunications(Paris)' 'Stasbourg 1'
'Universite Louis Pasteur, Strasbourg 1'
'Versailles-St Quentin en Yvelines' 'Universite de Montreal'
'universite Paris-Saclay' 'RENNES 1' 'Marne-La-Vallee' 'Paris-Grignon'
'Paris, Ecole nationale superieure des telecommunications'
'Universite de la Nouvelle-Caledonie' 'Bordeaux III'
'Universita degli studi di Bari' 'PARIS 3' 'Vallenciennes' 'ENGREF'
'tours' 'Poitiers CEAT' 'ENSIA'
'Paris, Ecole nationale superieure des mines' 'St Etienne'
'Universite Paris 13' 'Universite Grenoble Alpes' 'Univ. de Nantes'
'Universite Paris VII' 'brest' 'Pau et Adour'
'Cluj-Napoca, Roumanie, Universitatea Babes-Bolyai' "Evry-Val D'Essonne"
'neurophysiologie' 'Montpelllier 1' 'valenciennes' 'ENSA MONTPELLIER'
'caen' 'Paris, ENSMP' 'Clermond-Ferrand 2'
"Universite du Littoral-Cote d'Opale" 'Clermont-Ferrant 2' 'lyon, INSA'
'Paris, ENPC' 'Clermont- Ferrand 2' 'PARIS 11'
'Montpellier 2 et universita degli studi di Camerino (Italie)'
'Nice-Sophia Antipolis' 'Clermont-Ferrand, Universite Blaise Pascal'
'eReims' "Universita Ca' Foscari di Venezia" 'LITTORAL'
"Villeneuve d'Ascq, Ecole centrale de Lille" 'Franche Comte'
'Saint-Etienne, ENSM' 'Histoire ancienne' 'Philosophie' 'CDhambery'
'Strasbourg1' 'UPPA' 'Paris- EHESS' 'Paris, Ecole des mines'
'Ecole nationale superieure des mines de Paris' 'rouen'
'Clermont Ferrand 2' 'ENST']

Contenu unique de la colonne Identifiant etablissement :
['27361802' '28024400' nan '02640334X' '26404788' '28209966' '26404451'
'26403552' '26403447' '26404354' '26404702' '26403714' '26374889'
'26403021' '26403765' '26403692' '26404389' '131056549' '54447658'
'26402920' '26402971' '26404494' '26403145' '29473284' '27787109'
'02778715X' '157040569' '26369044' '30969379' '03082057X' '05017746X'
'26404524' '26403994' '26403919' '26403668' '34634894' '02640463X'
'26404079' '28237080' '26404478' '31308570' '35375043' '26404435'
'27548392' '27321118' '02819005X' '30327202' '26403250' '26403064'
'26404184' '27918459' '67101925' '110349164' '34137823' '26403633'
'26603136' '182292592' '30820529' '26375052' '26403587' '31122337'
'26388715' '26403315' '28032837' '183316401' '26403838' '03463486X'
'26404672' '26402823' '27787087' '26403412' '26404214' '77713486'
'26403498' '26388812' '28021037' '26404664' '26403188' '26375249'
'190456396' '121855465' '169816079' '28025261' '27309320' '33364346'
'26375478' '15863621X' '35022116' '50228064' '26570564' '175206562'
'26403366' '26402882' '26388820' '27542084' '157779092' '26403390'
'188120777' '27404978' '26403005' '52444724' '190915757' '60121076'
'26403153' '187841578' '190906332' '19077990X' '26403781' '92642845'
'81711883' '67331246' '31738419' '26388804' '27964361' '26369125'
'26404540' '30883717' '26438763' '05989136X' '33894221' '28028694'
'117553956' '27361837' '196200032' '27548341' '30138787' '26403102'
'27960250' '149154992' '139408088' '27956768' '28032829' '68859813'
'26388766' '200716271' '26524031' '28029429' '147800374' '31235409'
'202743233' '29981735' '159330114' '26394944' '30267676' '27941426'
'03063525X' '187401039' '191639044' '88458393' '134103211' '26375273'
'98046829' '73428159' '26569477' '50522604' '33236720' '163078998'
'115588701' '203592077' '28232224' '26408805' '221333754' '86146017'
'188204024' '02637515X' '77755138' '26404311' '34751386' '02640432X'
'26404125' '30820499' '17864577X' '59078995' '30142946' '26387859'
'59358041' '20073511X' '32486111' '26404796' '34137181' '67331149'
'203342011' '77512944' '74262955' '78023629' '35533838' '28139577'
'129112798' '110047702' '33592497' '196213428' '26390310' '123405327'
'134528239' '112461301' '78615151' '08757201X' '58567992' '15328434X'
'77195876' '58570993' '29483638' '151548587' '154236152' '81217188'

'58562400' '185433669' '26568209' '154784788' '26523477' '26523493'
'83328904' '80481965' '112943365' '184443237' '184668794' '122545273'
'59861959' '55339174' '91474469' '132782618' '137062508' '150044909'
'155669850' '30170494' '74454935' '131156977' '170721175' '172235278'
'221693157' '06944787X' '26402955' '32568819' '70571791' '26570467'
'83551239' '236453505' '177263660' '30099501' '59946555' '147347793'
'29446953' '95304061' '29688582' '33700222' '50516795' '05855968X'
'66778646' '29757479' '34016872' '137160054' '123456789' '118441272'
'26510014' '82002525' '02636526X' '26375060' '103376216' '28350545'
'27950751' '28003691' '28025253' '145350355' '28129547' '225321319'
'26408228' '26571641' '166292745' '26449196' '26567369' '26637065'
'26580756' '26436930' '26403463' '29966744' '238327159' '238277429'
'27961087' '29089301' '30603552' '96157011' '99429861' '83633456'
'83363564' '08862191X' '60275456' '113795041' '111756227' '111322111'
'109890450' '110297369' '26433540' '26550520' '183954645' '234200383'
'50540467' '27297519' '59313307' '103961852' '59054255' '26403331'
'06057111X' '108904784' '130907820' '125980604' '68936710' '139879862'
'60704632' '34755837' '34565728' '69999511' '26364700' '33725578'
'26365642' '111398819' '03500844X' '81104197' '127263322' '27412482'
'117921912' '26568071' '02640320X' '35073632' '77806239' '32613822'
'147974542' '67331556' '67344763' '71411119' '74537547' '158144252'
'55743072' '28866665' '59350849' '07418301X' '53501675' '94833613'
'03276166X' '28115635' '155971247' '158031083' '02888129X' '121577880'
'121593355' '120113546' '69638195' '68754787' '26404605' '26567121'
'08416185X' '122903234' '87214997' '08361821X' '26537095' '33532710'
'55666027' '120027526' '15070772X' '70148902' '95815554' '121888371'
'120117584' '86220322' '147502489' '34788034' '121423433' '88947556'
'150563469' '30053234' '123395348' '121760081' '12175006X' '92243878'
'122629590' '122903943' '149799942' '121420698' '97518417' '77932757'
'150500378' '81930402' '08552896X' '148086187' '83878572' '122902440'
'121704521' '119930897' '188152962' '129538299' '12199659X' '111093406'
'110113195' '26403935' '26429802' '147289319' '116357045' '59079800'
'26402939' '26404265' '60249633' '58591508' '02659840X' '27973840'
'59432896' '77450191' '78887763' '26430843' '136539033' '122579305'
'76955869' '154787973' '26366460' '11792282X' '26569140' '199046190'
'59521058' '28956834' '26390884' '131227238' '70203008' '72236027'
'26409259' '60713674' '26639432' '26403161' '139542027' '117840793'
'33699879' '78836743' '84519142' '153885416' '113157584' '34370943'
'59946598' '150076142' '28170318' '07901657X' '86020420' '26386437'
'129896349' '84589256' '26567598' '26627388' '128800895' '151384134'
'129908134' '93128517' '109083881' '176236198' '155325655' '120796899'
'121713970' '128500484' '96289600' '84626178' '132171783' '120116235'
'151665834' '02792467X' '67306144' '28084772' '69538514' '34461760'
'154400017' '94365229' '29471257' '81826656' '58928472' '02640317X'
'26568586' '102422672' '111300878' '74452789' '85810444' '155563394'
'105756938' '112083625' '136469523' '74457519' '143382454' '78164087'
'03538526X' '26403501' '147968348' '113380690' '77550226' '60780274'
'26386283' '83172955' '168612100' '103162178' '82734798' '241035694'
'50497901' '74314807' '59925264' '35056339' '03315550X' '26390388'
'30411815' '75740192' '26431467' '30659469' '241345251' '110147456'
'06038882X' '32949871' '26404346' '71061827' '34107835' '79006671'
'121386120' '26434687' '240648315' '77329015' '113196377' '113362811'
'26507323' '69364605' '81807821' '90154681' '114976155' '111755220'
'139878777' '123037891' '76411451' '122901924' '29936314' '120064243'
'26570408' '88757552' '59547855' '161182011' '97580112' '61332755'
'79315852' '31428630' '79900704' '34680985' '84577592' '26367025'
'121755150' '119934531' '113645341' '117922242' '122903633' '26429705'
'30091896' '94596425' '95969160' '29102464' '189038950' '122902807'
'153579226' '188799702' '83865713' '121327086' '95648542' '122904117'
'33656525' '03153211X' '81752261' '75173042' '32435649' '121602079'
'26375133' '34405224' '88525236' '75444046' '33629129' '104632887'
'03277396X' '26617773']

Contenu unique de la colonne Discipline :
['Driot prive'
 'Mecanique des solides, des materiaux, des structures et des surfaces'
 'Medecine' ...
 'Sciences humaines et humanites nouvelles specialite Sociologie - Travai
l social'
 "Histoire. Histoire de l'art. Archeologie" 'Archeologie. Paleoecologie']

Contenu unique de la colonne Statut :
['enCours' 'soutenue']

Contenu unique de la colonne Date de premiere inscription en doctorat :
['30-09-11' '01-10-12' nan ... '01-09-20' '03-03-05' '07-07-20']

Contenu unique de la colonne Date de soutenance :
[nan '01-01-93' '24-11-08' ... '01-07-20' '31-03-20' '07-07-20']

Contenu unique de la colonne Year :
[nan 1993. 2008. 2005. 2009. 2013. 2011. 2010. 2007. 2012. 2006. 2004.
 2001. 2015. 2014. 2016. 1995. 1997. 1986. 1992. 1991. 1987. 1988. 1998.
 1999. 1985. 1996. 1994. 2002. 2000. 1990. 1989. 2003. 1982. 1972. 1971.
 1976. 1973. 2017. 1984. 2018. 2019. 2020. 1980. 1979.]

Contenu unique de la colonne Langue de la these :
['na' 'fr' 'en' 'ro' 'es' 'de' 'FR' 'zh' 'bs' 'it' 'co' 'fren' 'enfr'
 'enzh' 'pt' 'frensl' 'fres' 'zhen' 'esen' 'itfr' 'frel' 'cs' 'frpten'
 'hu' 'enfrde' 'esenfr' 'frpl' 'elfr' 'frhu' 'frar' 'itfren' 'frit' 'frz
h'
 'enfrzh' 'frvi' 'frenes' 'ru' 'defr' 'pl' 'bg' 'frkm' 'kkenfr' 'frcs'
 'ptfren' 'el' 'enfreu' 'pten' 'frpt' 'esfr' 'ptfr' 'rufr' 'br' 'cafr'
 'sr' 'enru' 'ensl' 'frde' 'frja' 'frla' 'ar' 'ca' 'fraf' 'eu' 'enesfr'
 'freu' 'rofr' 'frgrc' 'enpt' 'encsfr' 'he' 'enptfr' 'enit' 'fresen'
 'enpl' 'sv' 'eufr' 'enfrpt' 'frbr' 'ukfren' 'frruen' 'arfr' 'hyfrru'
 'csfrsk' 'frhe' 'frru' 'brfr' 'akfr' 'zhfr' 'frms' 'enfrpl' 'frmn'
 'enfrit' 'envi' 'frro' 'frfy' 'frsl' 'enfres' 'frqu' 'as' 'frln' 'frenp
t'
 'abfr' 'uk' 'roen' 'plen' 'frenzh' 'frendees' 'enfrcs' 'itzh' 'cofr'
 'frtch' 'eufres' 'ja' 'enbo' 'frsa' 'itfrhe' 'froc' 'frkmsa' 'hyfr'
 'zhfrit' 'nl' 'nlen' 'csfr' 'enesfrca' 'enar' 'amfr' 'frtr' 'frsakm'
 'enarfr' 'itlafr' 'frff' 'itfrla' 'sq' 'entr' 'deenfr' 'itla' 'enro'
 'plfr' 'endefr' 'ptenfr' 'enes' 'frka' 'fafr' 'akes' 'frmg' 'elenfr'
 'csenfr' 'sqfr' 'zhenfr' 'ee' 'bgenfr' 'frfr' 'enfrsk' 'encs' 'enfrro'
 'fris' 'bgfr' 'frgl' 'frko' 'enhu' 'deenfrit' 'azfr' 'fritla' 'enth'
 'frth' 'arenfrit' 'frfa' 'frty' 'eo' 'frid' 'frsv' 'aefr' 'csfrla'
 'enfrhu' 'zhfrug' 'myfr' 'enfrja' 'deen' 'amarfr' 'aafr' 'frsi' 'enla'
 'frlaptes' 'ares' 'arenfr' 'frnv' 'enitfr' 'frplen' 'ad' 'frund' 'enfrv
i'
 'ab' 'frhi' 'fritlaoc' 'am' 'enfrru' 'frukr' 'hy' 'enzhfr' 'enfrar' 'et'
 'id' 'cafres' 'frruuk' 'fi' 'nlenfr']

Contenu unique de la colonne Identifiant de la these :
['s69480' 's98826' '1993NAN19006' ... 's244358' 's244354' 's192344']

Contenu unique de la colonne Accessible en ligne :
['non' 'oui']

Contenu unique de la colonne Publication dans theses.fr :
['26-01-12' '22-11-13' '24-05-13' ... '27-04-19' '15-09-19' '08-07-20']

Contenu unique de la colonne Mise a jour dans theses.fr :
['26-01-12' '22-11-13' '17-11-12' ... '06-07-20' '07-07-20' '08-07-20']

Contenu unique de la colonne Discipline_predi :
['Droit et Science Politique' 'Materiaux, Milieux et Chimie' 'Medecine'
 'SHS' 'Biologie' 'Langues et Litteratures' 'Psychologie'
 'Economie Gestion' 'Informatique' "Science de l'ingÃ©nieur" 'Poubelle'
 "Sciences de l'education" 'MathÃ©matiques' 'Science de la Terre'
 'Mathematiques et Informatique']

Contenu unique de la colonne Genre :
['male' 'female' 'unknown' 'andy' 'mostly_male' 'mostly_female']

Contenu unique de la colonne etablissement_rec :
['Université Paris 1 - Panthéon Sorbonne'
 "École nationale supérieure de mécanique et d'aérotechnique de Poitiers"
 'Université de Lorraine' 'Université Lumière - Lyon 2'
 'Université de Paris' 'Université Jean Monnet' 'Université de La Réunio
n'
 'Université Paris 8 - Vincennes - Saint-Denis' 'Université de Nantes'
 'Université Toulouse Capitole' 'Université Paul-Valéry - Montpellier 3'
 'Université de Picardie Jules-Verne' 'EHESS'
 'Université de Bretagne Occidentale' 'Université de Poitiers'
 'Université de Perpignan Via Domitia' 'Université de Lille'
 'Université de Strasbourg' 'Université Rennes 2' "Université d'Angers"
 "Université d'Orléans" 'Université Jean Moulin - Lyon 3'
 'Université Panthéon-Assas' 'Université de Corse Pasquale Paoli'
 'Université Paris sciences et lettres' 'Université de Rennes 1'
 'Avignon Université' 'Université du Littoral Côte d\x92Opale'
 'Université de Versailles Saint-Quentin-en-Yvelines'
 'Université Bretagne Sud' 'Université Toulouse - Jean Jaurès'
 'Université de Rouen Normandie'
 "Université de Pau et des Pays de l'Adour" "Université d'Artois"
 'Université Sorbonne Paris Nord'
 'Université Polytechnique Hauts-de-France' 'Université Paris-Saclay'
 'Université de Tours' 'Université Le Havre Normandie'
 'La Rochelle Université' 'Le Mans Université'
 'Université Bordeaux Montaigne' 'HEC Paris' 'Université de Bourgogne'
 'Université Grenoble Alpes' 'Université de Haute-Alsace'
 'Université de Caen Normandie' 'Sciences Po'
 'Université de la Polynésie Française' 'na' 'Université de Bordeaux'
 'Sorbonne Université' 'ex-Université des Antilles-Guyane'
 "Université d'Évry-Val-d'Essonne" 'Université Paris Nanterre'
 'Université de Toulon'
 'Institut national des langues et civilisations orientales'
 'Université de Limoges' 'Université de Montpellier' nan
 'CY Cergy Paris Université' 'Université Toulouse III - Paul Sabatier'
 'Université Claude Bernard - Lyon 1' "Université Côte d'Azur"
 'Université de technologie de Compiègne' 'Université Paris-Est Créteil'
 'Université de Franche-Comté'
 'Institut Mines-Télécom, au périmètre des écoles IMT Atlantique, Lille D
ouai, Albi, Alès, Mines Saint-Étienne et Institut Mines-Télécom Business
school'
 'Université Gustave Eiffel' 'Institut polytechnique de Paris'
 'Institut national des sciences appliquées de Rouen Normandie' 'EPHE'
 'Aix-Marseille Université'
 'Institut national des sciences appliquées de Rennes'

```
 "École nationale vétérinaire, agroalimentaire et de l'alimentation, Nant
es-Atlantique"
 'Conservatoire national des arts et métiers'
 'Institut national des sciences appliquées de Lyon'
 'Université de la Nouvelle-Calédonie' 'Université Clermont Auvergne'
 'Normandie Université' 'USPC'
 'École nationale supérieure de chimie de Rennes' 'INP Grenoble'
 'Centrale Lille Institut' 'Université Savoie Mont Blanc' 'Centrale Lyon'
 'Montpellier SupAgro' 'Université Sorbonne Nouvelle - Paris 3'
 'École normale supérieure de Lyon' 'AgroParisTech'
 'École nationale supérieure de chimie de Montpellier'
 'Université de technologie de Belfort-Montbéliard'
 'École supérieure des sciences économiques et commerciales'
 'Agrocampus Ouest' 'Université Française du Pacifique'
 "Muséum national d'histoire naturelle" 'Université de Lyon'
 'Centrale Nantes' 'Université de technologie de Troyes'
 'Centrale Marseille' 'Université de Nîmes'
 'Institut national des sciences appliquées Centre Val de Loire'
 'Université de Guyane' 'Arts et Métiers Sciences et Technologies'
 'École normale supérieure de Rennes'
 'École nationale des ponts et chaussées' 'ENSA Montpellier'
 'ISAE Supaero'
 'École nationale supérieure de techniques avancées Bretagne' 'HESAM'
 "École nationale d'ingénieurs de Brest"]


Contenu unique de la colonne Langue_rec :
[nan 'Français' 'Anglais' 'Autre' 'Bilingue']
```

e) Statistiques descriptives

- Stat des Dates

In [21]:
```
Summary_Date = PhD_v2_copy[["Year", "Date de premiere inscription en doct
```

In [22]:
```
Summary_Date['Annee_premiere_inscription'] = pd.to_datetime(Summary_Date[
                                                    format='%Y',
Summary_Date['Annee_de_soutenance'] = pd.to_datetime(Summary_Date['Date d
                                                    format='%Y',
Summary_Date['Year'] = pd.to_datetime(Summary_Date['Year'],
                                                format='%Y', errors='c
```

In [23]:
```
Summary_Date= Summary_Date.drop(["Date de premiere inscription en doctora
```

In [24]:
```
Summary_Date.dtypes
```

Out[24]:
```
Year                          datetime64[ns]
Annee_premiere_inscription    datetime64[ns]
Annee_de_soutenance           datetime64[ns]
dtype: object
```

In [25]:
```
Summary_Date
```

Out[25]:

	Year	Annee_premiere_inscription	Annee_de_soutenance
0	NaT	NaT	NaT
1	NaT	NaT	NaT
2	1993-01-01	NaT	NaT

3	NaT	NaT	NaT
4	NaT	NaT	NaT
...	...	...	...
448042	NaT	NaT	NaT
448043	NaT	NaT	NaT
448044	NaT	NaT	NaT
448045	2020-01-01	NaT	NaT
448046	2019-01-01	NaT	NaT

448047 rows × 3 columns

In [26]:
```
print(Summary_Date.describe(include='datetime'))
```

```
                      Year Annee_premiere_inscription Annee_de_soutenanc
e
count                390961                         0
0
unique                   44                         0
0
top     2012-01-01 00:00:00                       NaN                  Na
N
freq                  13991                       NaN                  Na
N
first   1971-01-01 00:00:00                       NaN                  Na
N
last    2020-01-01 00:00:00                       NaN                  Na
N
```

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 2072, 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.

Stat des Chaines de Caratères

In [27]:
```
Summary_Objet = PhD_v2_copy.drop(columns=['Year', 'Date de premiere inscr
```

In [28]:
```
Summary_Objet.describe(include='object')
```

Out[28]:

	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur de these (nom prenom)	Identifiant directeur	Etablissement de soutenance	Ide etablis
count	448047	317700	448040	448034	448034	448047	448046	
unique	430273	313771	446816	159019	159021	98906	567	
top	Nicolas Martin	,	#NAME?	Directeur de these inconnu	Directeur de these inconnu	na	Paris 6	27
freq	16	462	17	713	713	49488	21201	

In [ ]:
```

```

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

- Le nombre total de thèses recensées est de 448 047.
- Le nombre d'auteurs uniques est de 430 273.
- Le nombre de directeurs de thèse uniques est de 159 021.
- 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 majorité des thèses ont été soutenues (381 360 sur 448 047).
- La langue de la majorité des thèses est le français (334 443 sur 448 047).
- La discipline la plus prédite pour ces thèses est "Biologie".
- Le genre le plus représenté est "masculin".
- L'établissement de rattachement le plus fréquent est "Sorbonne Université".
- La langue la plus fréquente pour les résumés est également le français.

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.

II. Données manquantes

Sélection des valeurs manquantes dans le jeu de données PhD_v2.

In [29]:
```
PhD_v2_copy.isna()
```

Out[29]:

	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur de these (nom prenom)	Identifiant directeur	Etablissement de soutenance	Identif etablissem
0	False	True	False	False	False	False	False	F
1	False	False	False	False	False	False	False	F
2	False	True	False	False	False	False	False	-
3	False	True	False	False	False	False	False	F
4	False	True	False	False	False	False	False	F
...	...	...	...	...	...	...	...	
448042	False	True	False	False	False	False	False	F
448043	False	True	False	False	False	False	False	F
448044	False	True	False	False	False	False	False	F
448045	False	True	False	False	False	False	False	F
448046	False	True	False	False	False	False	False	F

448047 rows × 22 columns

In [30]:
```
PhD_v2[PhD_v2_copy.isna()]
```

Out[30]:

	Auteur	Identifiant auteur	Titre	Directeur de these	Directeur de these (nom prenom)	Identifiant directeur	Etablissement de soutenance	Identifi etablissem
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
...	...	...	...	...	...	...	...	
448042	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
448043	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
448044	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
448045	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
448046	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

448047 rows × 22 columns

In [31]:
```
PhD_v2_copy.isna().sum().sort_values(ascending=False)
```

Out[31]:
```
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
```

La variable Date de premiere inscription en doctorat a le plus grand nombre de valeurs manquantes, suivie de Identifiant auteur et Langue_rec.

On peut voir aussi que la variable Date de soutenance a le même nombre de valeurs manquantes que la variable Year, ce qui pourrait avoir un impact sur les analyses qui dépendent de ces variables.

1) Représentation de la répartition des données manquantes au sein du jeu de données PhD_v2.

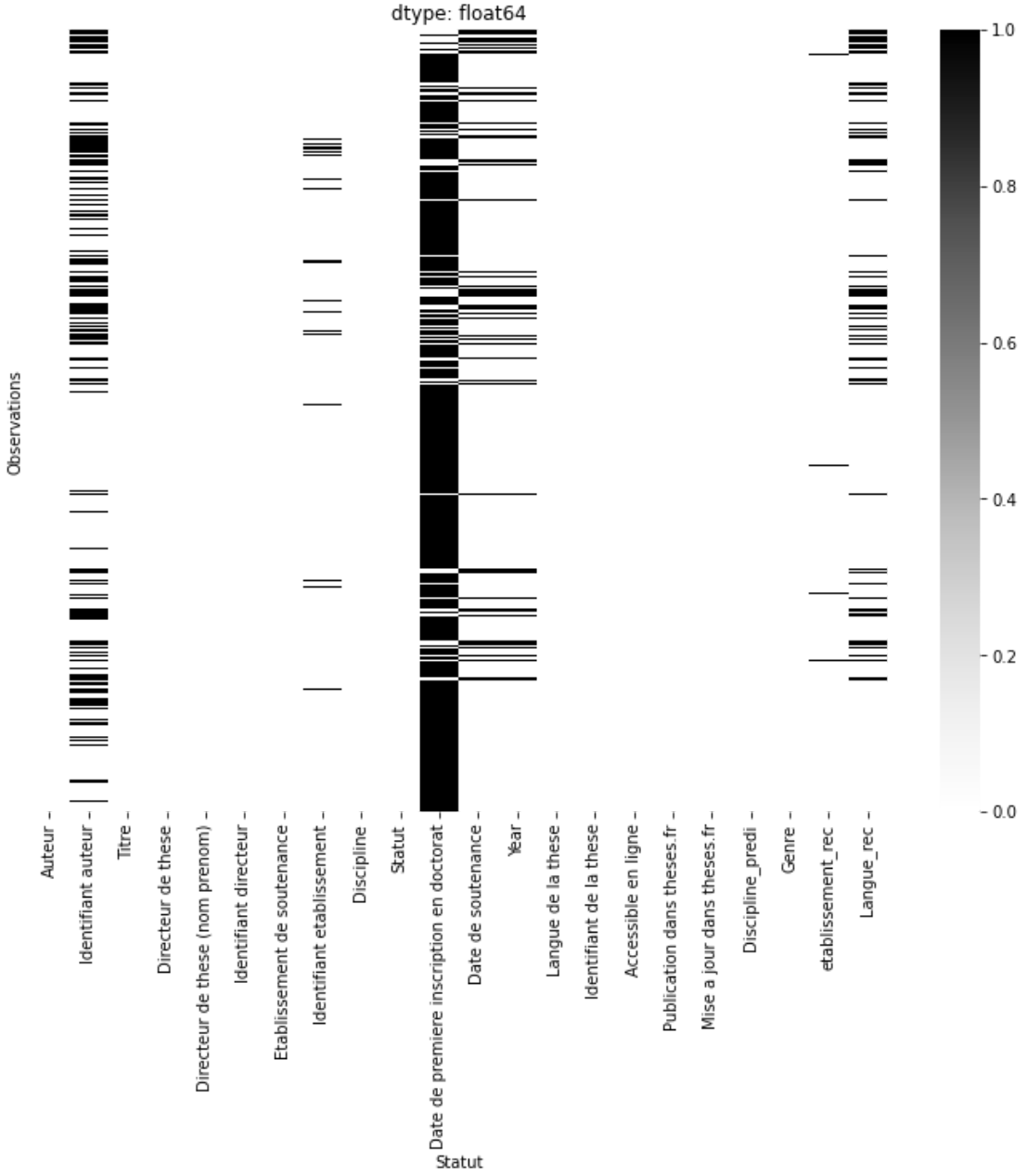
*1.a **Graphique: 1** de la Répartition des données manquantes*

In [32]:
```
plt.figure(figsize=(12, 9))
sns.heatmap(PhD_v2_copy.isnull(), cmap="Greys", cbar=True, yticklabels=Fa
plt.title("Figure 1: Répartition des valeurs manquantes en % pour chaque
          f" {(PhD_v2.isnull().mean()*100).round(2).sort_values(ascending
plt.xticks(rotation=90)
plt.xlabel('Statut')
plt.ylabel('Observations')
plt.savefig('Figure 1.png')
plt.show()
```

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

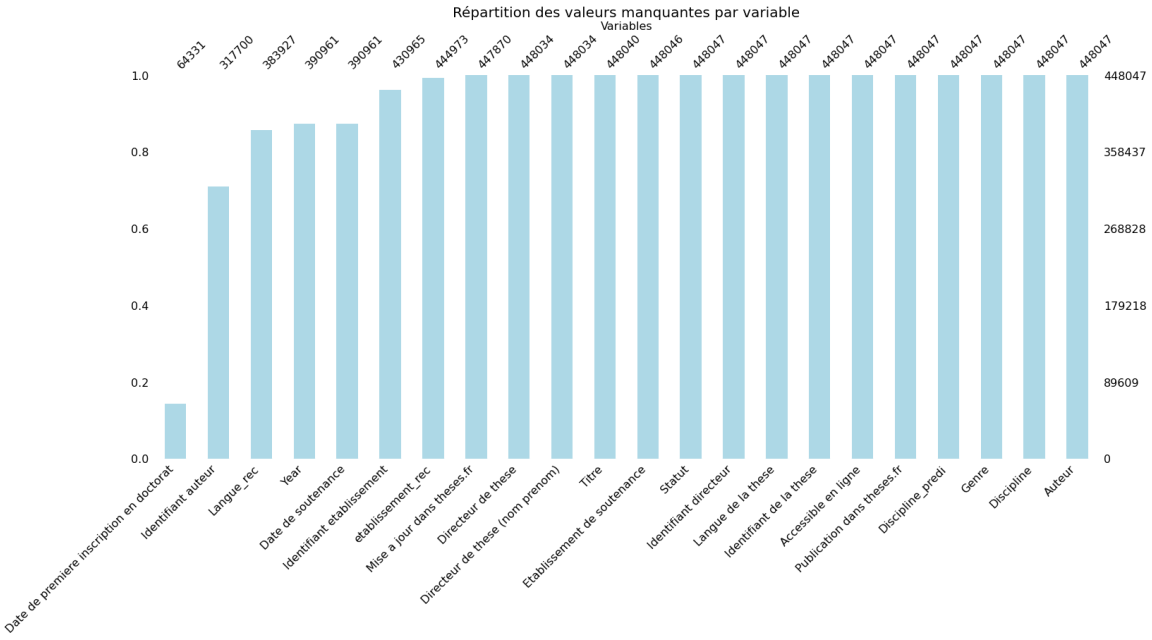
Date de premiere inscription en doctorat	85.64
Identifiant auteur	29.09
Langue_rec	14.31
Year	12.74
Date de soutenance	12.74
Identifiant etablissement	3.81
etablissement_rec	0.69
Mise a jour dans theses.fr	0.04
Identifiant directeur	0.00
Identifiant de la these	0.00
Genre	0.00
Discipline_predi	0.00
Publication dans theses.fr	0.00
Accessible en ligne	0.00
Langue de la these	0.00
Etablissement de soutenance	0.00
Titre	0.00
Directeur de these	0.00
Directeur de these (nom prenom)	0.00
Statut	0.00
Discipline	0.00
Auteur	0.00

dtype: float64



In [33]:
```
plt.figure(figsize=(10,5))
msno.bar(PhD_v2_copy, color='lightblue', sort='ascending')
plt.title("Répartition des valeurs manquantes par variable", fontsize=20)
plt.xlabel('Variables', fontsize=16)
```

```
plt.ylabel('% de valeurs manquantes', fontsize=16)
plt.savefig('Figure 1_1.png')
plt.show()
```



En examinant le graphique de 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, etablissement_rec et Langue_rec 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.*

Pour voir le lien entre la date de lancement de la thèse et la date de soutenance de la thèse, on peut commencer par créer un nouveau dataframe (df) qui ne contient que ces deux colonnes, en éliminant les valeurs manquantes

In [34]:
```
df = PhD_v2_copy[['Date de premiere inscription en doctorat', 'Date de so
```

Transformation les dates en objets de type datetime pour faciliter la manipulation

In [35]:
```
df['Date de premiere inscription en doctorat'] = pd.to_datetime(df['Date
df['Date de soutenance'] = pd.to_datetime(df['Date de soutenance'])
```

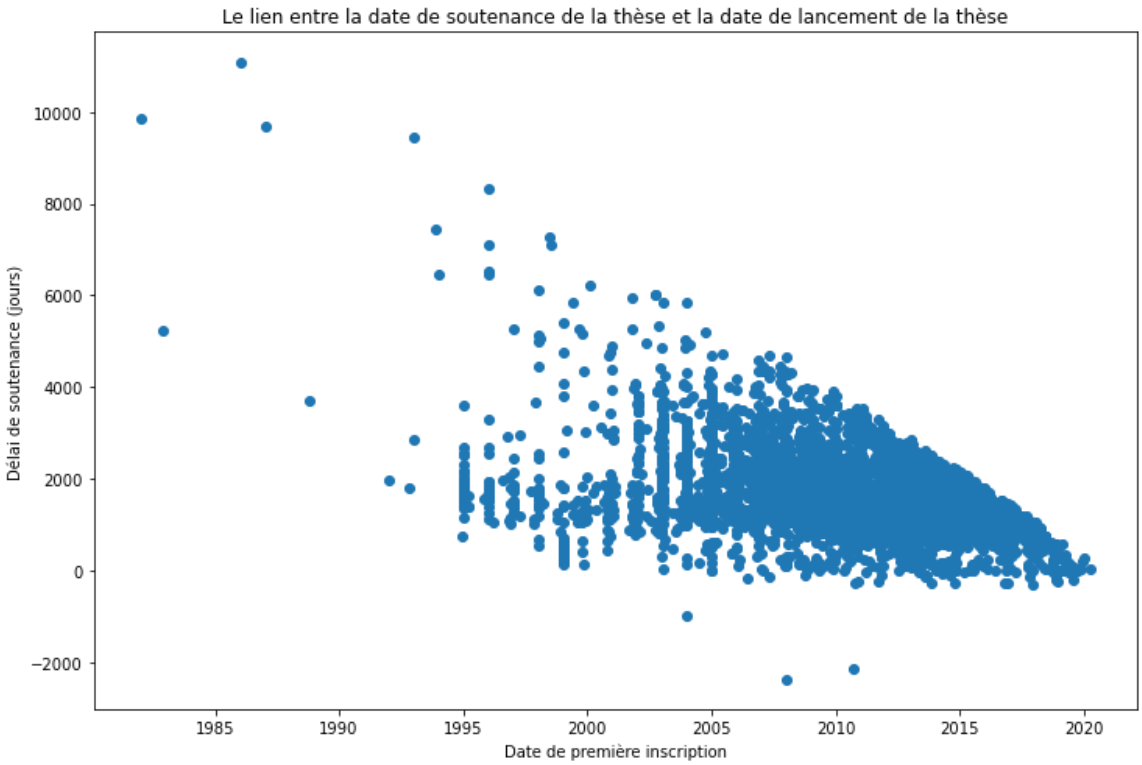
Cration d'une nouvelle colonne qui calcule la différence entre la date de soutenance et la date de première inscription en jours

In [36]:
```
df['Délai de soutenance'] = (df['Date de soutenance'] - df['Date de premi
```

Réprésentation graphique du lien entre le délai de soutenance et la date de première inscription

```

```

In [37]:
```
plt.figure(figsize=(12, 8))
plt.scatter(df['Date de premiere inscription en doctorat'], df['Délai de
plt.title('Le lien entre la date de soutenance de la thèse et la date de
plt.xlabel('Date de première inscription')
plt.ylabel('Délai de soutenance (jours)')
plt.savefig('2.png')
plt.show()
```



On constate qu'il y a 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 ou de sujets de recherche*,
- *Des contraintes personnelles des doctorants*,
- *Etc...*

In [ ]:
```
```

III.Qualité des données : Détection et traitement des anomalies

1.Gestion des doublons:

- Verification da la présence des doublons dans le jeu de données

In [38]:
```
duplicates = PhD_v2_copy.duplicated()
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
```

In [39]:
```
num_duplicates = duplicates.sum()
print(f"Le jeu de données contient {num_duplicates} doublons.")
```

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

- Suppression des doublons

In [40]:
```
PhD_v2_copy.drop_duplicates(inplace=True)
```

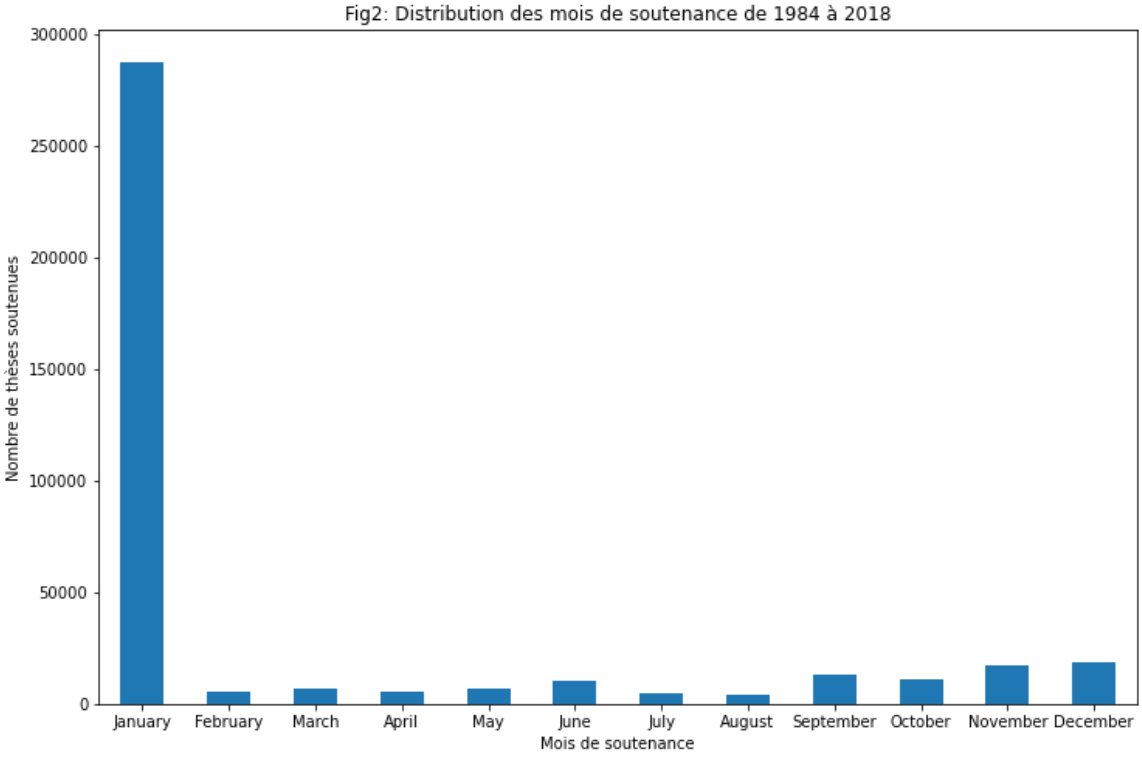
In [41]:
```
PhD_v2_copy[PhD_v2_copy.duplicated()].sum()
```

Out[41]:
```
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
```

2. La réprésentation de la distribution des mois de soutenance pour l'intégralité du jeu de données

In [42]:
```
# Convertir la date en mois
PhD_v2_copy['Month'] = pd.to_datetime(PhD_v2_copy['Date de soutenance']).

# Ordonner les mois dans l'ordre chronologique
PhD_v2_copy['Month'] = pd.Categorical(PhD_v2_copy['Month'],
                                      categories=['January', 'February',
                                                  'July', 'August', 'Sep
                                      ordered=True)

# Plot
fig, ax = plt.subplots(figsize=(12, 8))
PhD_v2_copy['Month'].value_counts().sort_index().plot(kind='bar', ax=ax)
plt.title('Fig2: Distribution des mois de soutenance de 1984 à 2018')
plt.xlabel('Mois de soutenance')
plt.ylabel('Nombre de thèses soutenues')
```

```
plt.xticks(rotation=0)
plt.savefig('Fig2.png')
plt.show()
```



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.

3. Figure 2: Distribution du mois de soutenance pour chaque année, de 2005 à 2018

In [43]:
```
# Convertir la colonne "Date de soutenance" en datetime
PhD_v2_copy["Date de soutenance"] = pd.to_datetime(PhD_v2_copy["Date de s

# Extraire les années et les mois
PhD_v2_copy["Year"] = PhD_v2_copy["Date de soutenance"].dt.year
PhD_v2_copy["Month"] = PhD_v2_copy["Date de soutenance"].dt.month
```

```python
# Filtre pour les années entre 2005 et 2018
PhD_v2_copy = PhD_v2_copy[(PhD_v2_copy["Year"] >= 2005) & (PhD_v2_copy["Y

# Créer une FacetGrid pour chaque année
g = sns.FacetGrid(PhD_v2_copy, col="Year", col_wrap=4, height=3)

# Tracer l'histogramme de la distribution des mois de soutenance pour cha
g.map(plt.hist, "Month", bins=12, density=True)

# Ajouter le titre et les étiquettes des axes
g.fig.subplots_adjust(top=0.9)
g.fig.suptitle("Figure 3: Distribution des mois de soutenance pour chaque
g.set_axis_labels("Mois", "Proportion de soutenances")
# Enregistrer la figure
plt.savefig("distribution_mois_soutenance.png", dpi=300)

# Afficher la figure
plt.show()
```

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

Ce graphique est constitué de 14 sous-graphiques, un pour chaque année de 2005 à 2018 inclusivement. Chaque sous-graphique représente la distribution des mois de soutenance des thèses pour une année donnée.

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.

Les mois sont représentés sur l'axe des abscisses et la proportion de soutenances est représentée sur l'axe des ordonnées.

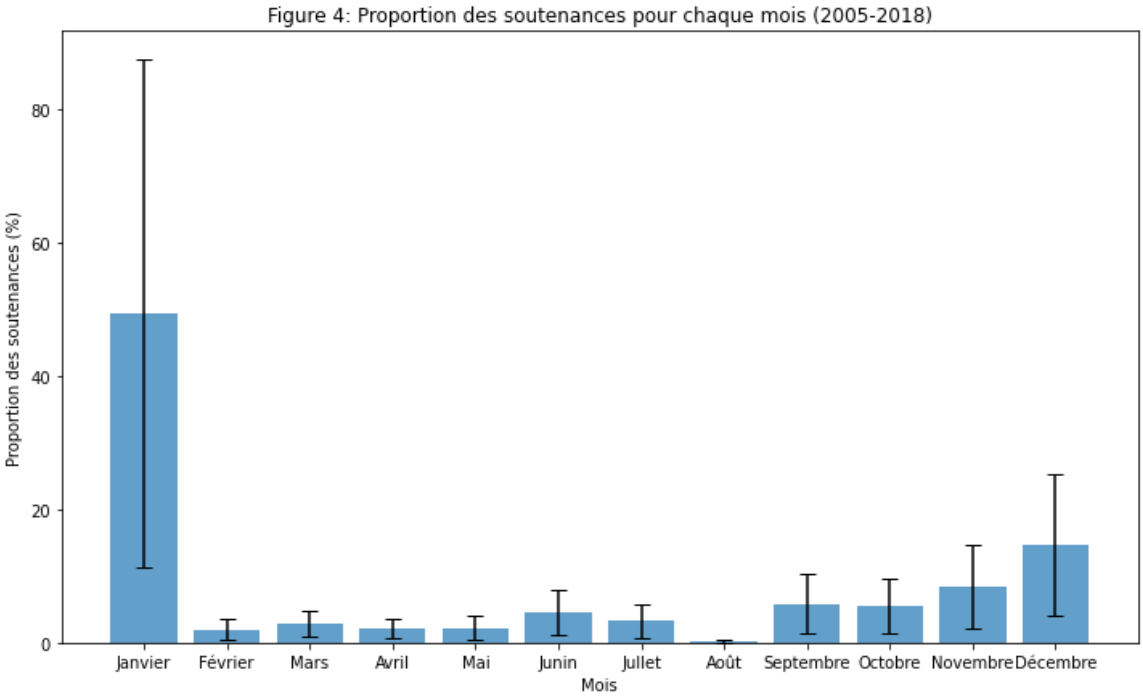
La couleur de chaque barre dans l'histogramme représente la proportion de soutenances de thèse pour un mois donné. Plus la barre est haute, plus il y a eu de soutenances de thèse pour ce mois.

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.

Ce graphique est utile pour visualiser les fluctuations annuelles dans les mois de soutenance des thèses, ce qui pourrait aider les universités à mieux planifier leurs programmes de doctorat et à répartir les ressources de manière plus efficace en fonction des pics de soutenances de thèse.

Figure 3: Proportion des thèses soutenues au fil des mois

In [71]:
```
# Convertir la colonne "Date de soutenance" en datetime
PhD_v2_copy["Date de soutenance"] = pd.to_datetime(PhD_v2_copy["Date de s

# Extraire les années et les mois
PhD_v2_copy["Year"] = PhD_v2_copy["Date de soutenance"].dt.year
PhD_v2_copy["Month"] = PhD_v2_copy["Date de soutenance"].dt.month

# Filtre pour les années entre 2005 et 2018
PhD_v2_copy = PhD_v2_copy[(PhD_v2_copy["Year"] >= 2005) & (PhD_v2_copy["Y

# Calculer la proportion de soutenances pour chaque mois et chaque année
df = PhD_v2_copy.groupby(["Month", "Year"])["Titre"].count().reset_index(
df = df.rename(columns={"Titre": "Count"})
df["Proportion"] = df["Count"] / df.groupby("Year")["Count"].transform(su

# Calculer la moyenne et l'écart-type de la proportion de soutenances pou
stats = df.groupby("Month")["Proportion"].agg(["mean", "std"]).reset_inde

# Créer la figure et les axes
fig, ax = plt.subplots(figsize=(12, 7))

# Tracer les barres de la proportion de soutenances pour chaque mois, ave
ax.bar(stats["Month"], stats["mean"], yerr=stats["std"], capsize=5, alpha

# Ajouter les noms des mois sur l'axe des abscisses
ax.set_xticks(range(1, 13))
ax.set_xticklabels(['Janvier', 'Février', 'Mars', 'Avril', 'Mai', 'Junin'
                    'Octobre', 'Novembre', 'Décembre'])

# Ajouter le titre et les étiquettes des axes
plt.title("Figure 4: Proportion des soutenances pour chaque mois (2005-20
plt.xlabel("Mois")
plt.ylabel("Proportion des soutenances (%)")
plt.savefig('Fig4.png')
# Afficher le graphique
plt.show()
```



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.

In [72]:
```
# La proportion moyenne de soutenance pour chaque année au mois de janvie
jan_stats = df[df["Month"] == 1].groupby("Year")["Proportion"].mean().res
print(jan_stats)
```

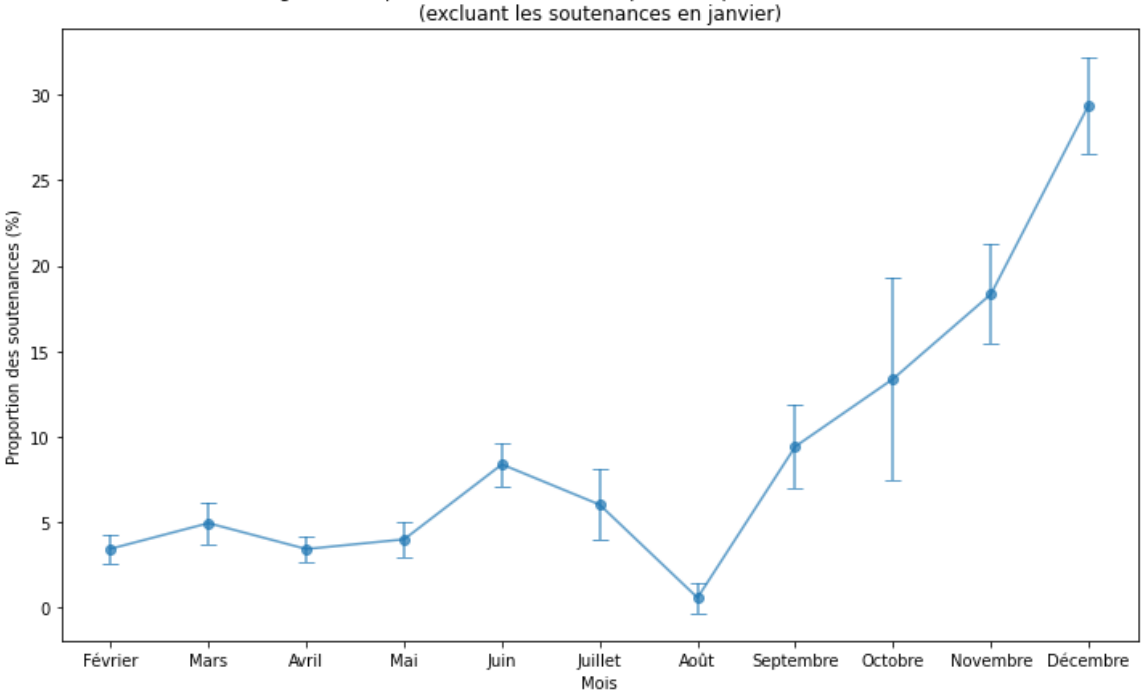
```
      Year  Proportion
0   2005.0   99.630752
1   2006.0   99.189066
2   2007.0   97.084723
3   2008.0   90.442045
4   2009.0   80.171196
5   2010.0   67.026206
6   2011.0   45.446224
7   2012.0   35.044691
8   2013.0   26.882031
9   2014.0   17.277685
10  2015.0   13.337941
11  2016.0    9.417663
12  2017.0    5.128791
13  2018.0    4.560718
```

On voit que la proportion de soutenances pour le mois de janvier diminue considérablement entre 2005 et 2018, passant de 99,6% à seulement 4,6%. Cette tendance peut être due à différents facteurs tels que:

- L'évolution des programmes de doctorat
- la disponibilité des financements
- Des changements dans la demande du marché du travail. *Cette information pourrait être utile pour les universités ou les institutions en charge de la formation doctorale pour évaluer les tendances et adapter leur offre en conséquence*.

In [73]:
```
# Convertir la colonne "Date de soutenance" en datetime
PhD_v2_copy["Date de soutenance"] = pd.to_datetime(PhD_v2_copy["Date de s

# Extraire les années et les mois
PhD_v2_copy["Year"] = PhD_v2_copy["Date de soutenance"].dt.year
PhD_v2_copy["Month"] = PhD_v2_copy["Date de soutenance"].dt.month

# Filtre pour les années entre 2005 et 2018 et les mois différents de jan
PhD_v2_copy = PhD_v2_copy[(PhD_v2_copy["Year"] >= 2005) & (PhD_v2_copy["Y

# Calculer la proportion de soutenances pour chaque mois et chaque année
df = PhD_v2_copy.groupby(["Month", "Year"])["Titre"].count().reset_index(
df = df.rename(columns={"Titre": "Count"})
df["Proportion"] = df["Count"] / df.groupby("Year")["Count"].transform(su

# Calculer la moyenne et l'écart-type de la proportion de soutenances pou
stats = df.groupby("Month")["Proportion"].agg(["mean", "std"]).reset_inde

# Créer la figure et les axes
fig, ax = plt.subplots(figsize=(12, 7))

# Tracer la courbe de la proportion de soutenances pour chaque mois, avec
ax.errorbar(stats["Month"], stats["mean"], yerr=stats["std"], capsize=5,

# Ajouter les noms des mois sur l'axe des abscisses
ax.set_xticks(range(2, 13))
ax.set_xticklabels(['Février', 'Mars', 'Avril', 'Mai', 'Juin', 'Juillet',
                    'Octobre', 'Novembre', 'Décembre'])

# Ajouter le titre et les étiquettes des axes
plt.title("Figure 5: Proportion des soutenances pour chaque mois (2005-20
plt.xlabel("Mois")
plt.ylabel("Proportion des soutenances (%)")
plt.savefig('Fig5.png')
# Afficher le graphique
plt.show()
```

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 ?

Fig3 montre que le mois de soutenance préféré est janvier. **Cependant Fig4 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

In [74]:
```
len(PhD_v2_copy[PhD_v2_copy["Auteur"] == "Cécile Martin"])
```

Out[74]:
```
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 mon 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.

IV. Détection d'outliers

Identification les individus ayant encadré un nombre relativement anormal de thèses

In [ ]:
```
```

In [75]:
```
directeurs_uniques = PhD_v2_copy['Directeur de these (nom prenom)'].uniqu
```

In [76]:
```
print("Les directeurs de thèse uniques présents dans le jeu de données es
```

```
Les directeurs de thèse uniques présents dans le jeu de données est de :
61024
```

Pour chaque directeur de thèse, compter le nombre de thèses qu'il/elle a supervisées sur la période considérée (1984-2018).

In [77]:
```
# Filtrer le DataFrame en fonction de la plage d'années
PhD_filtered = PhD_v2_copy[(PhD_v2_copy['Year'] >= 1984) & (PhD_v2_copy['

# Calculer le nombre de thèses supervisées pour chaque directeur
supervisions_directeurs = PhD_filtered['Directeur de these (nom prenom)']
```

Créer un nouveau jeu de données avec une ligne par directeur de thèse, contenant les informations suivantes : nom, prénom, nombre de thèses supervisées.

In [78]:
```
df_supervisions_directeurs = pd.DataFrame(list(supervisions_directeurs.it
df_supervisions_directeurs[['Prénom', 'Nom']] = df_supervisions_directeur
df_supervisions_directeurs = df_supervisions_directeurs[['Nom', 'Prénom',
```

In [79]:
```
df_supervisions_directeurs
```

Out[79]:

	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

Commençons par créer un histogramme pour visualiser la distribution :

In [80]:
```
plt.figure(figsize=(10,6))
plt.hist(df_supervisions_directeurs['Nombre de thèses supervisées'], bins
plt.title('Figg 6: Distribution du nombre de thèses encadrées par Directe
plt.xlabel('Nombre de thèses encadrées')
plt.ylabel('Nombre de Directeur de these (nom prenom)')
plt.savefig('fig6')
plt.show()
```

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 utiliser la formule suivante pour déterminer la limite supérieure des valeurs acceptables :

In [81]:
```
Q1 = df_supervisions_directeurs['Nombre de thèses supervisées'].quantile(
Q3 = df_supervisions_directeurs['Nombre de thèses supervisées'].quantile(
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.

In [82]:
```
df_supervisions_directeurs[df_supervisions_directeurs['Nombre de thèses s
```

Out[82]:

	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

On a un Tableau qui 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.

V. Obtention de résultats préliminaires:

In [83]:
```
PhD_v2_copy['Langue_rec'].value_counts()
```

Out[83]:
```
Français    63193
Anglais     18429
Bilingue     5351
Autre        1113
Name: Langue_rec, dtype: int64
```

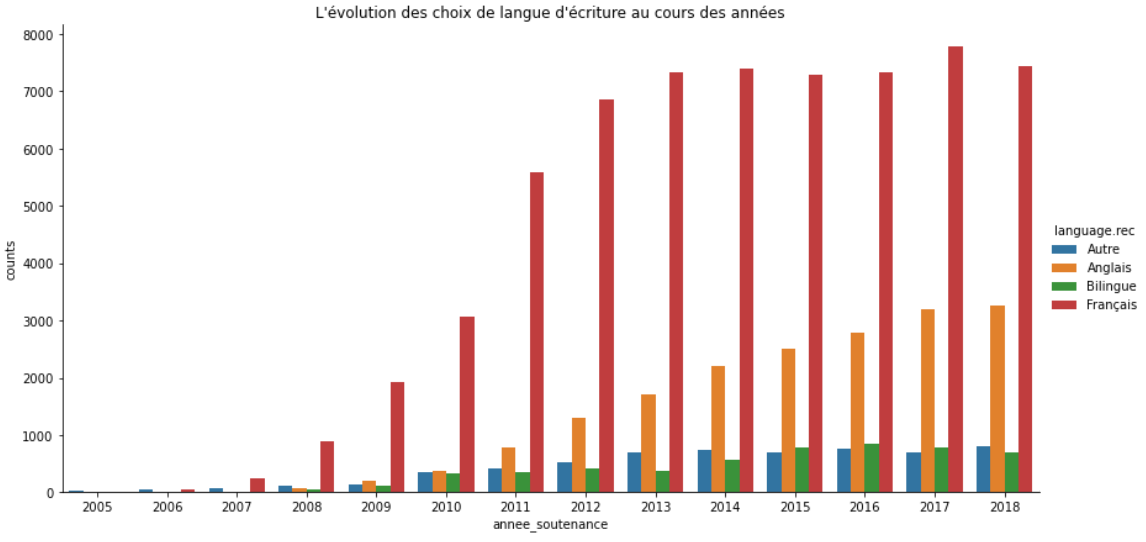
In [84]:
```
def recode_language(lang):
    if lang in ['Français', 'français']:
        return 'Français'
    elif lang in ['Anglais', 'anglais']:
        return 'Anglais'
    elif lang in ['Bilingue', 'Anglais-Français', 'Français-Anglais', 'en
        return 'Bilingue'
    else:
        return 'Autre'

PhD_v2_copy['language.rec'] = PhD_filtered['Langue_rec'].apply(lambda x:
```

In [85]:
```
PhD_v2_copy['language.rec'].value_counts()
```

Out[85]:
```
Français    63193
Anglais     18429
Autre        6140
Bilingue     5351
Name: language.rec, dtype: int64
```

In [86]:
```
PhD_v2_copy['Date_soutenance'] = pd.to_datetime(PhD_v2_copy['Date de sout
```

In [87]:
```
PhD_v2_copy['Date_soutenance'] = pd.to_datetime(PhD_v2_copy['Date de sout
PhD_v2_copy['Date_soutenance'] = pd.to_datetime(PhD_v2_copy['Date de sout

PhD_v2_copy['annee_soutenance'] = PhD_v2_copy['Date_soutenance'].dt.year

data_plot = PhD_v2_copy.groupby(['annee_soutenance', 'language.rec']).siz

sns.catplot(x='annee_soutenance', y='counts', hue='language.rec', kind='b
plt.title("L'évolution des choix de langue d'écriture au cours des années
plt.savefig("Fig7.png", dpi=300, bbox_inches='tight')
plt.show()
```



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

In [ ]: