CY Tech Sciences et Techniques

Data Wrangling

Author
Buu NGUYEN

October 21, 2021

Contents

List of Figures

1 Missing data

Missing data is a widespread problem that can have a substantial effect on the investigation process. Before any analysis, we need to make sure that the dataset is as accurate as possible. Missing data matrix is the best visual way to help us understand the distribution of missing values.

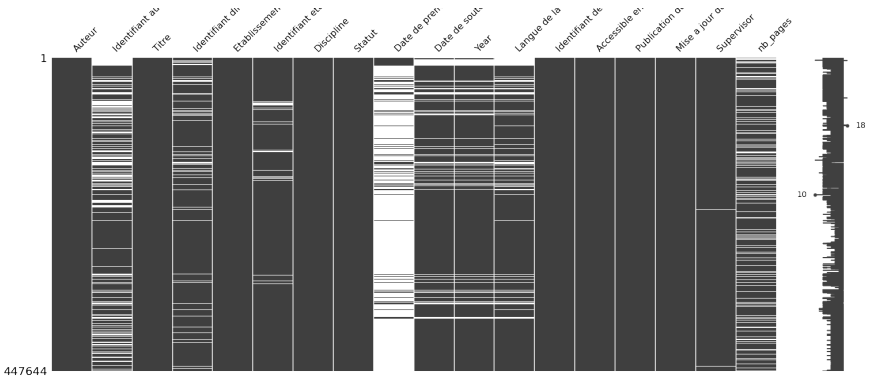

Figure 1: Missing data matrix

We can observe a contrast in missing data between the "Date of the premiere inscription en doctorat" and "Date de soutenance". We can assume that the website removes the inscription date after the defense day.

Feature	Missing Percentage
Identifiant auteur	29.14%
Identifiant directeur	10.98%
Identifiant etablissement	3.82%
Date de premiere inscription en doctorat	85.71%
Date de soutenance	12.68%
Langue de la these	14.24%

Table 1: Missing percentage of key features

There are substantial missing data in the Author ID variable. This can be a real obstacle when we try to analyze the data. Of course, it is possible to use the Author Name instead, but we have to deal with the homonym problem.

2 Common issues

71.89% of thesis in the dataset was defended on the first of January.

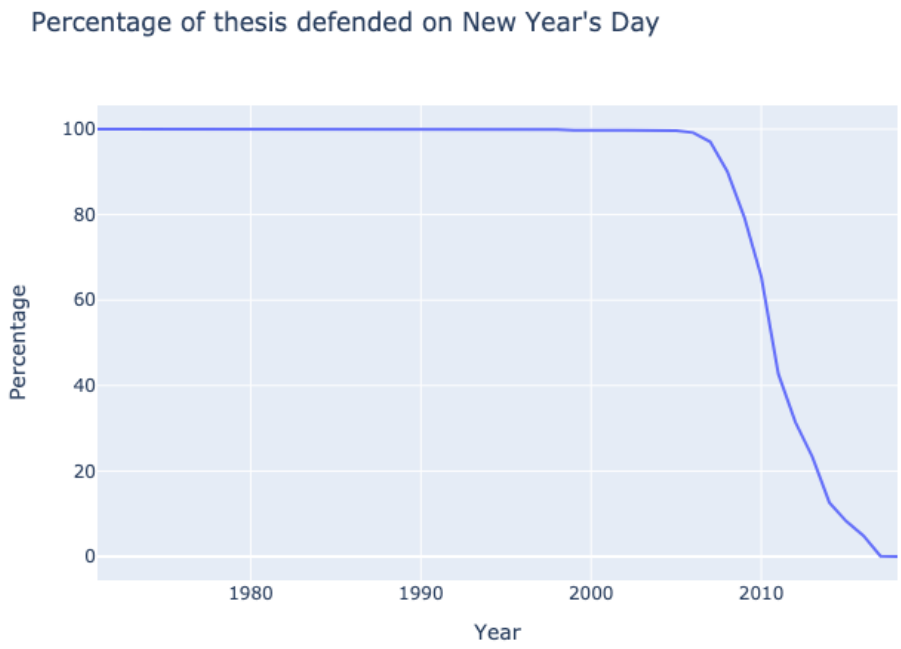

Figure 2: Thesis defended on New Year's Eve

Almost all thesis defended before 2010 have New Year's Eve as the defense day. It dropped significantly from 79.4% in 2009 to 4.88% in 2016. It has been below 1% since 2017.
The website *theses.fr* was launched in 2010. Therefore all the thesis defended before 2010 may be imported from libraries and archives which do not have details about the defense year. So when they were uploaded to the website's database, the date was set as the first of January.

7.66% of thesis have duplicated author name.

Identifiant auteur	Auteur	Year	Supervisor
203208145	Cécile Martin	2017	Jullier Laurent
81323557	Cécile Martin	2000	Lossouarn Jean
179423568	Cécile Martin	2014	Dormont Brigitte
81323557	Cécile Martin	2001	Antonini Gerard
81323557	Cécile Martin	1991	Mironneau Jean
81323557	Cécile Martin	1994	Briand Yves
182118703	Cécile Martin	1989	Vautherin Dominique

Table 2: Cécile Martin

Author name *Cécile Martin* appears in 7 records. *Cécile Martin*, with ID *81323557*, published four theses in 1991, 1994, 2000, and 2001. Three others *Cécile Martin* have three different ID which implies they are different people. It is hypothesized that *Cécile Martin* is a common name, so many authors in the dataset share it.

Drop in the number of PhD defended in 2019 and 2020.

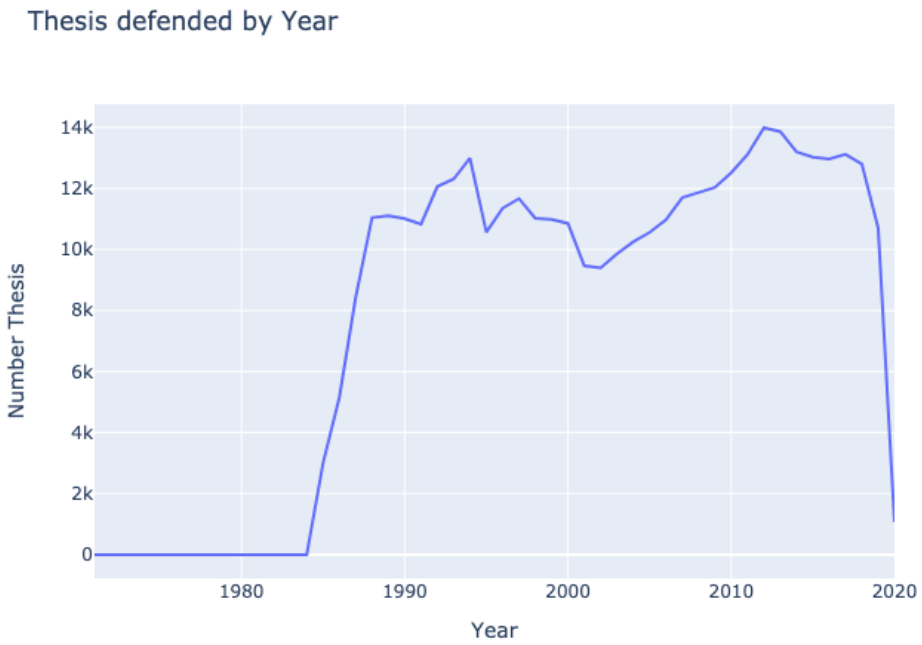

Figure 3: Thesis defended by Year

Over the past decade, the number of defended thesis has been over 12000 yearly since 2009. It dropped to 10712 in 2019 and 1070 in 2020. The last registered thesis on the website was defended on 6th December 2020. It is hypothesized that there is a significant delay window between the defense day the day the thesis shows up on the website. Another explanation is because of the effect of the COVID-19 pandemic.

3 Outliers

Supervisor ID	Number of mentor thesis
1	1057
7	718
3	712
8	618
6	557
2	517
9	284
59375140	208
26730774	205
26756625	193

Table 3: Supervisor IDs that have the most mentor thesis

Top 7 IDs that have the most mentor thesis supervised 4587 theses. However, in these theses, there are 4504 different supervisor names. Therefore, these are outliers because of input errors.

Author ID	Number of thesis
05990190X	12
69413916	7
85924660	6
069632472	6
60151013	6
56833776	5
27013340	5
34296565	5
78079365	5
66761999	5

Table 4: Author IDs that have the most thesis

These are the IDs that have the most defenses. Among 12 theses attached to ID 05990190X, *Philippe Ascher* is the co-author, but other co-authors are different. Therefore this ID may only belong to *Philippe Ascher*, not the

group of authors of these theses. We can deduce that It is an outlier because of input errors.

4 Preliminary Results

The evolution in choice of the manuscript's language over the last two decades.

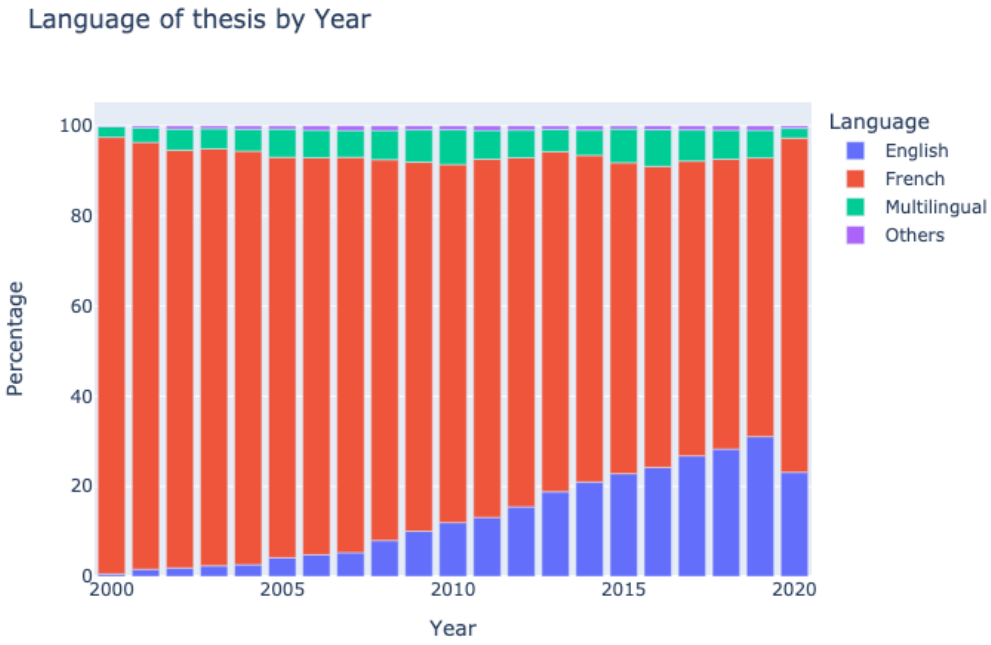

Figure 4: Language of manuscript by Year

At the beginning of the 2000s, 97% of the manuscript was written in French, only 0.6% of them were English. Over the years, the percentage of English manuscripts has increased enormously and peaked at 31% in 2019. There is also an increase in Multilingual manuscript: from 2.4% in 2000 to around 6% since 2006.

At what period of the year do PhD candidates tend to defend?

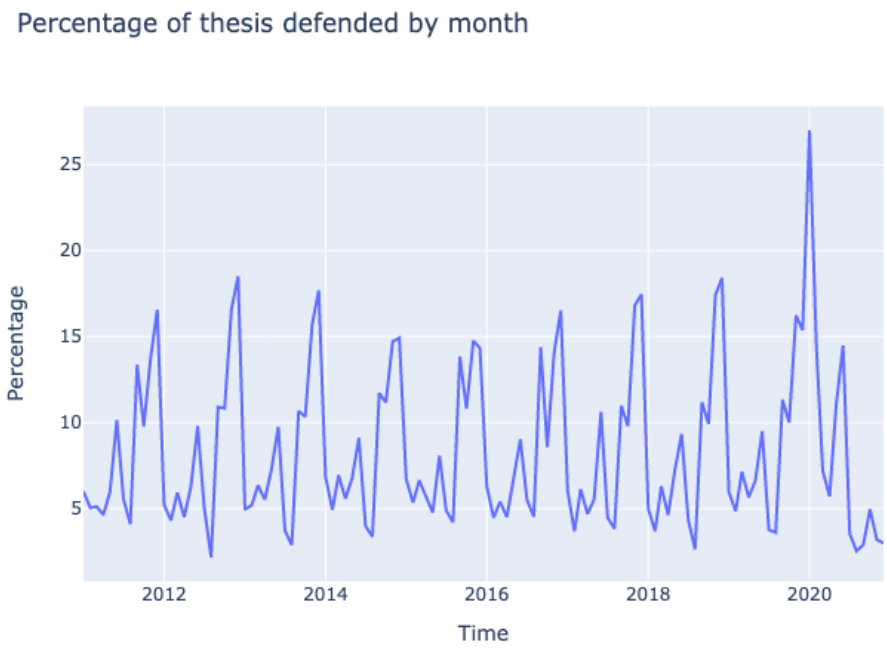

Figure 5: Percentage of defenses

From Figure 5, we can observe three peak times in a year that Ph.D. candidates tend to defend their thesis.

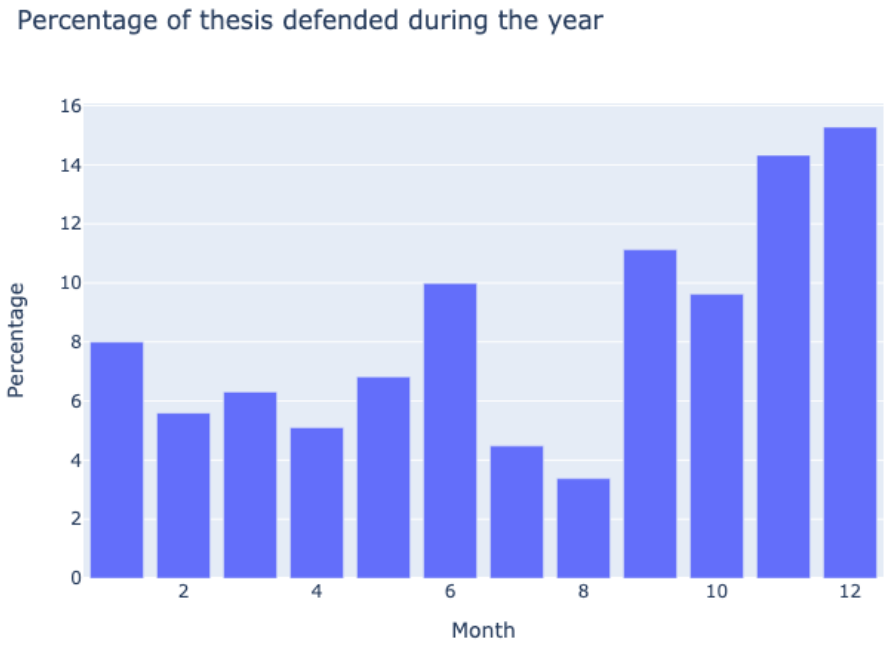

Figure 6: Percentage of defenses by Month

After investigating Figure 6, we recognize that the most picked months to defend are June with 10%, September 11%, November 14%, and December 15%.

The evolution of gender among Ph.D. candidates over the past decades.

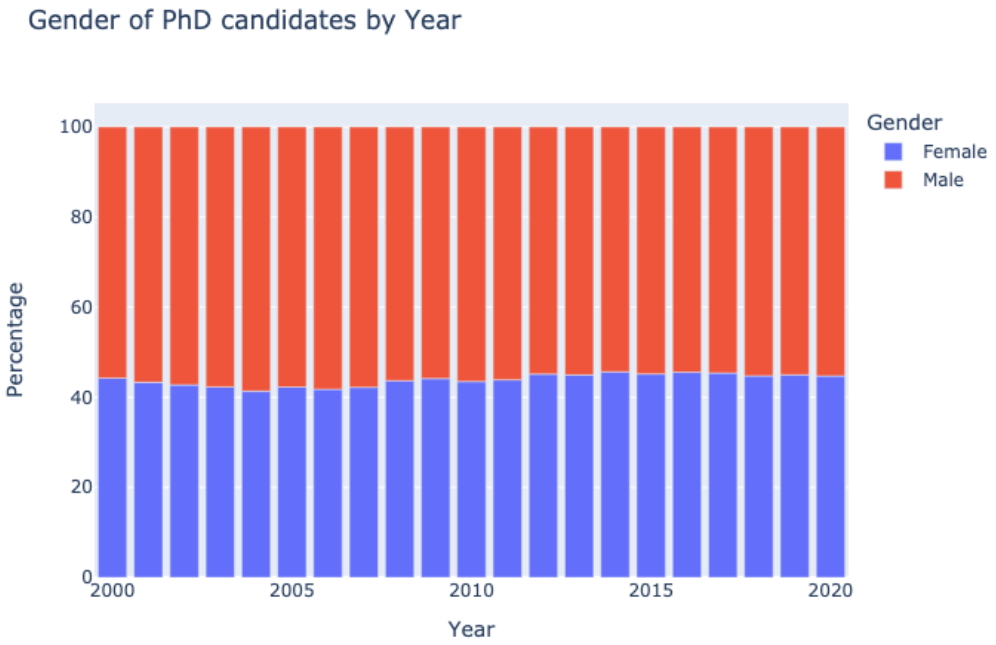

Figure 7: Gender of Author by Year

From Figure 7, we observe that the distribution of gender among Ph.D. candidates has been around 45% Female, 55% Male over the past decades.