Data Wrangling & Data Processing Report

2021

Aya Ben hriz

A report presented for the analysis carried out in Data Wrangling Course.

Bachelors in Data Science
CY Tech
France

20 August 2021

Abstract

This paper will go over the data wrangling and analysis we performed on a real-life dataset of PhD defences. This dataset was scraped from theses.fr[1] and has information about PhDs that were defended in France between the years 1971 and 2020.

Contents

Chapter 1

Introduction

The process of gathering, sorting, and transforming data from its original ”raw” format to prepare it for analysis and other downstream processes is known as data wrangling. Data wrangling entails removing inaccurate and irrelevant data as well as more thoroughly transforming, reformatting, and preparing it for future needs.

In this course, we used dplyr and tidyr functions in R to perform all data wrangling tasks. For the analysis portion, we used ggplot2 and plotly to create relevant graphs. All of these tasks were performed on data regarding PhD theses defended between 1971 and 2020 in France. This paper aims to conclude insights from my analysis.

Chapter 2

Data Source

The dataset we used came from theses.fr, a website launched in the early 2000s. The dimensions of the dataset are 447644 rows and 18 columns. Each row in the dataset represents a unique thesis. The columns contain information about the thesis author, supervisor(s), institute, defence and inscription dates and publish and update dates on theses.fr. The data type and description of each column is shown in the table 2.1.

Column	Data Type	Description
Auteur	String	Author name
Identifiant auteur	Integer	Author ID
Titre	String	Title of the thesis
Directeur de these	String	Name(s) of the thesis supervisor(s)
Directeur de these nom prenom	String	Last name followed by first name of the supervisor(s)
Identifiant directeur	Integer	ID of the supervisor(s)
Etablissement de soutenance	String	Name of the Institute(University)
Identifiant etablissement	Integer	ID of the institute
Discipline	String	Thesis discipline name
Statut	String	Thesis status
Date de premiere inscription en doctorat	Date	Date of first inscription in the doctorate
Date de soutenance	Date	Date of defense of the thesis
Year	Date	Year of defense of the thesis
Langue de la these	String	Language of the thesis
Identifiant de la these	Integer	ID of the thesis
Accessible en ligne	Boolean	If the accessibility online
Publication dans theses fr	Date	Date of the publication on theses fr
Mise a jour dans theses fr	Date	Date of the last update on theses.fr

Table 2.1: Description of the columns in the dataset.

Chapter 3

Scrapping Data

We started by scrapping data from theses.fr website using the ”Beautiful Soup” library in python. In my case the scrapped data had 3000 rows and 4 columns. Each row in the dataset represents a unique thesis. The columns contain information about the thesis id, the thesis title, author and subject. The data type and description of each column is shown in the table 3.1.

Column	Data Type	Description
ID	Integer	ID of the thesis
Title	String	Title of the thesis
Author	String	Author name
Subject	String	Thesis discipline name

Table 3.1: Description of the columns in the scrapped dataset.

Since scrapping the data took a long time I decided to use the dataset provided by the instructor whose structure is described in the table 2.1 in the previous chapter to carry out all data wrangling and analysis tasks.

Chapter 4

Dealing with Missing Data

Before starting to clean our dataset we visualised the missing values in the dataset by using the vis_miss() function from the naniar library in R. The graph 4.1 shows the percentage of missing values in each column.

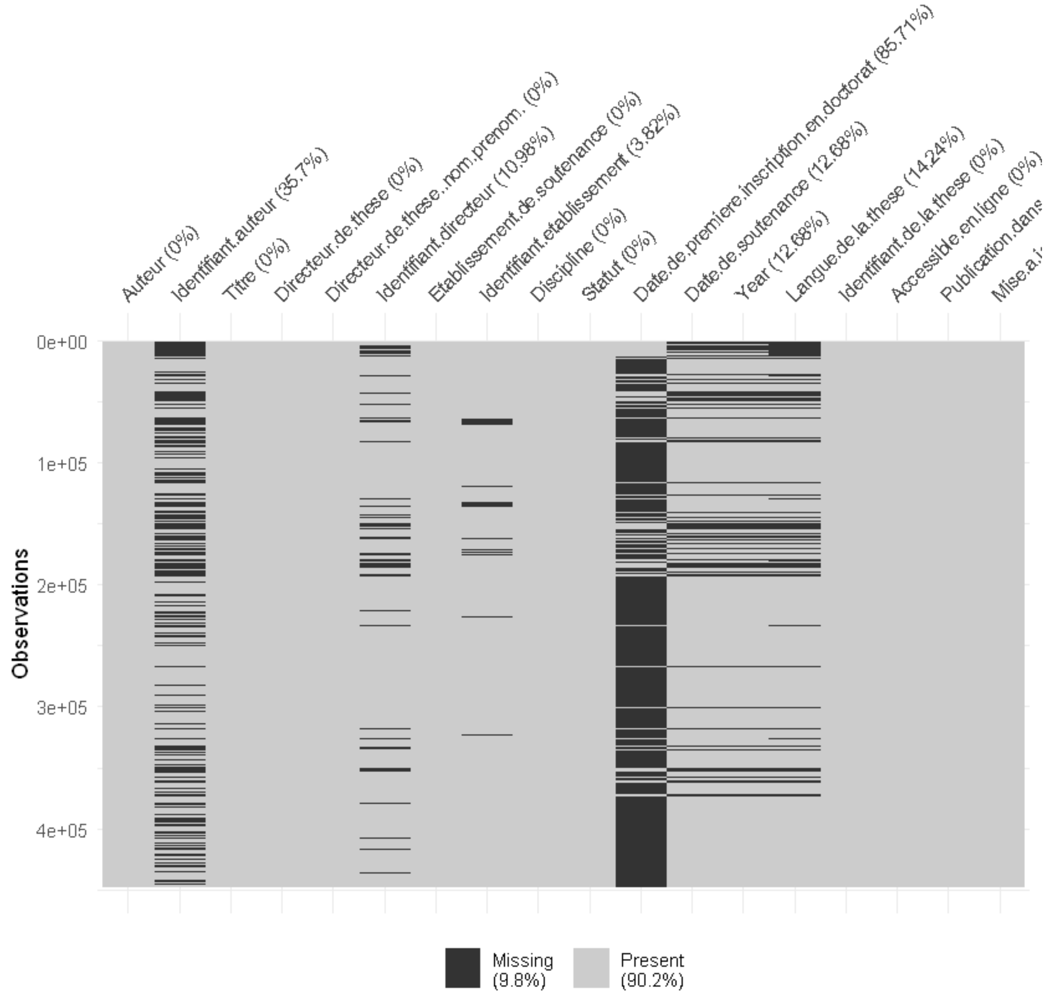


Figure 4.1: Percentage of Missing Data in Each Column

Column Name	Number of Missing Values	% of Missing Values
Date de premiere inscription en doctorat	383668	85.71
Identifiant auteur	159820	35.70
Langue de la these	63765	14.24
Date de soutenance	56746	12.68
Year	56746	12.68
Identifiant directeur	49172	10.98
Identifiant etablissement	17085	3.82
Mise a jour dans theses fr	177	0.04
Directeur de these	17	0.00
Directeur de these nom prenom	17	0.00
Titre	13	0.00
Discipline	5	0.00
Etablissement de soutenance	4	0.00
Auteur	3	0.00

Table 4.1: Number of Missing Data in Each Column

From the above figure 4.1 and table 4.1 we can say that the column ”Date de premiere inscription en doctorat” has the highest amount of missing values with 86%, followed by the ”Identifiant.auteur” column with 29% of missing data. The portion of missing data reduces to 14%-10% for the columns ”Langue de la these”, ”Date de soutenance”, ”Year” and ”Identifiant directeur”.

From the above figure 4.1 we can also see that the columns ’Date de soutenance’ and ’Date de premiere inscription en doctorat’ are complementary to each other meaning that if the date for ’Date de soutenance’ is known then the ’Date de premiere inscription en doctorat’ date is missing and vice versa. From this we can conclude that once a PhD thesis was defended the date of inscription is removed and the date of defence is added.

After this we created a new variable using the rnorm() function in R to simulate the number of pages per thesis for, having mean of 200 pages and standard deviation of 50 pages, for 80% of the dataset. The remaining 20% of the data was imputed with the mean number of pages. Since the data is centered around the mean the imputation does not have a large effect on the overall man and standard deviation.

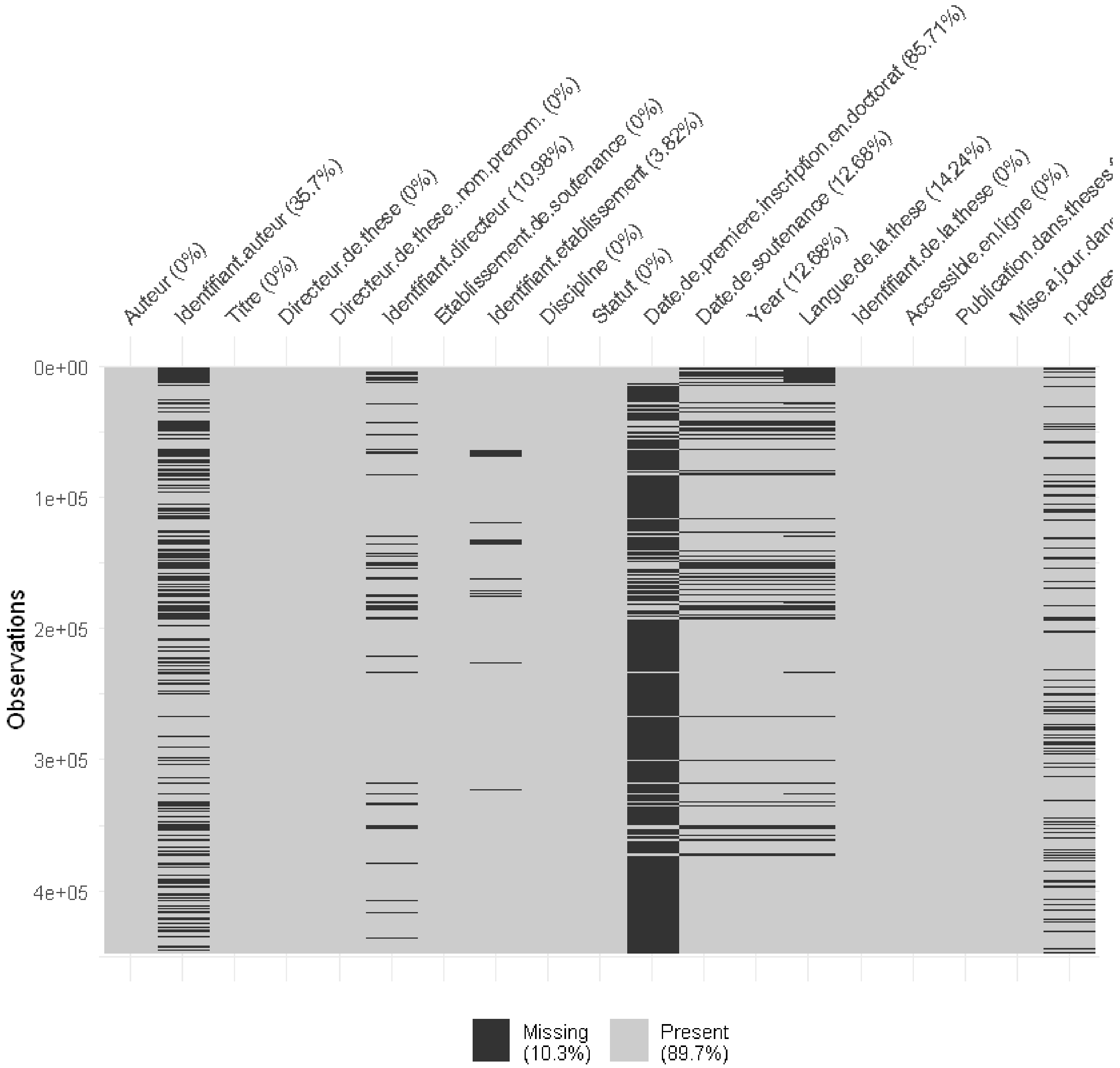


Figure 4.2: Missing Data in the new number of pages column (Before Imputation)

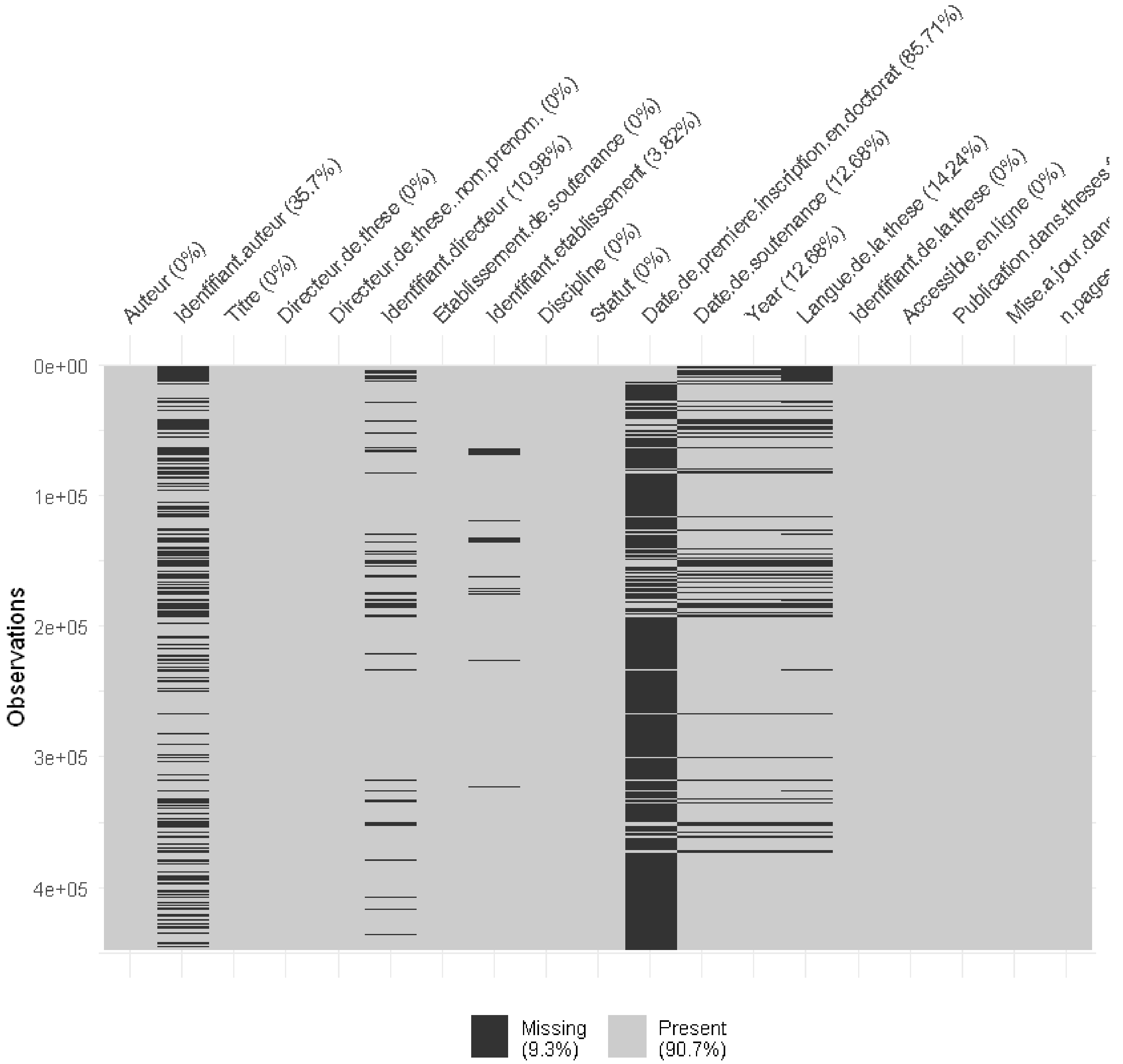


Figure 4.3: Missing Data in the new number of pages column (After Imputation)

The figures 4.2 and 4.3 the percentage of missing data in the dataset before and after the imputation is performed. We can also see that the imputation in the n.pages column has reduced the overall percentage of missing data by 1%.

Chapter 5

Checking for Common Issues

In this chapter we will be focusing on few common issues seen in the dataset.

5.1 January 1st as Defence Date

In total over the years 74.12% of the theses in the dataset had January 1st as their defence date. However, this seems very odd as new years is a national holiday in France and it is very unlikely that a thesis would be defended on this day. Thus we set out to investigate the how the proportion of defences on the first of January has evolved over the years. To achieve this we first used the lubridate library in R to split the defence date column into 3 columns with the day, month, and year. Then we used filter() and group by() functions from the dyplyr library to group and count the theses that were defended on the 1st of January for each year. We then computed the proportion of this count over the total number of theses defended each year.

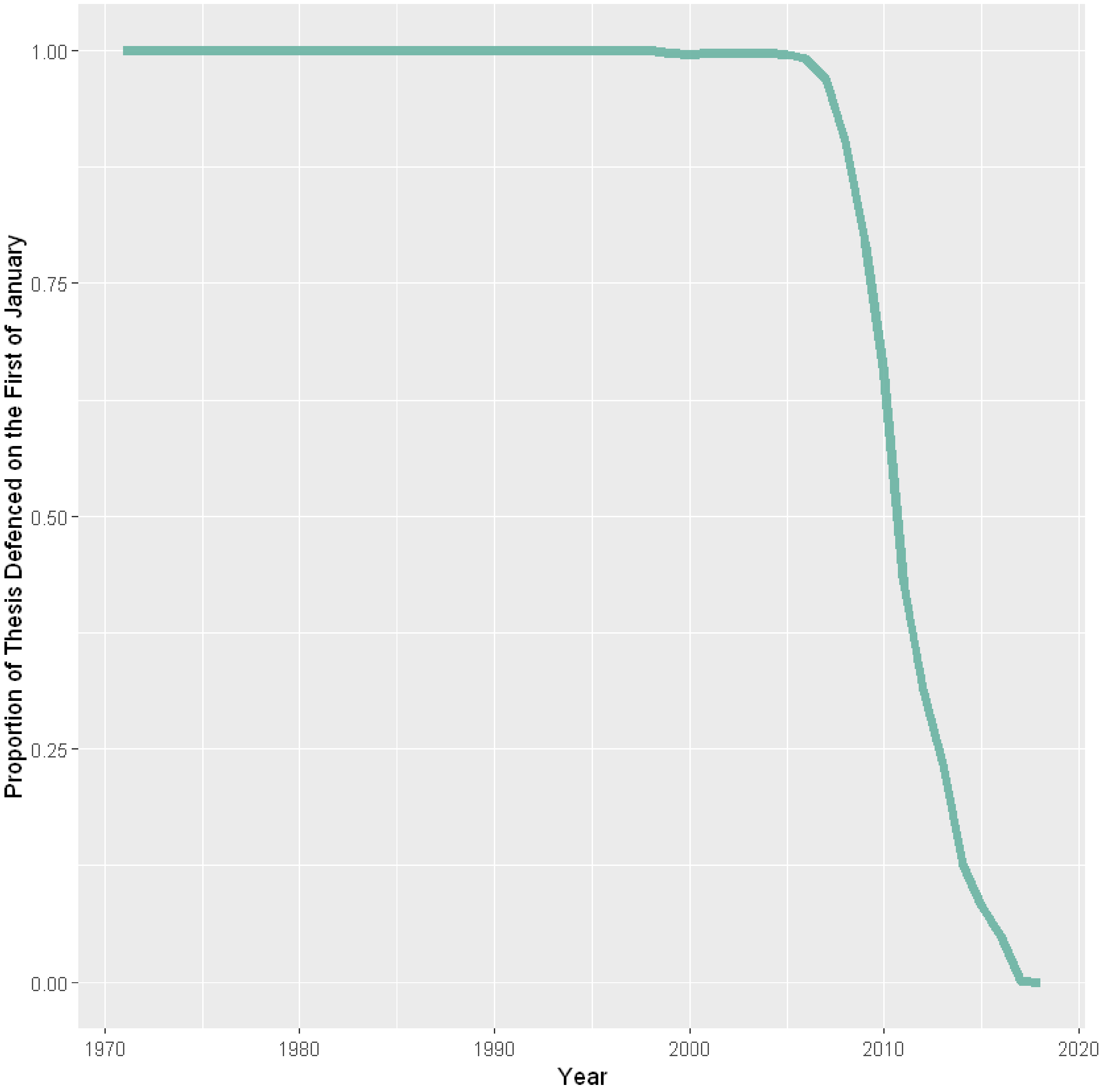


Figure 5.1: Percentage of theses defended on the first of January over years

From the figure 5.1 we can see that all theses recorded from the year 1971 to 1996 have their defence date as the 1st of January. This number slowly started to decrease and later we see a significant drop from 90% in 2008 to 79% in 2009. This further drops to 0% in 2019 and 2020.

5.2 Homomyns in Supervisor Names

We also saw that in the 'Auteur' column of authors' names, that there are multiple theses that were defended by people with the same name. So to investigate this we grouped our dataset by the author name and author id to get the count for each unique author. Considering all the unique author names we found that only 2% of the names are homonyms meaning that for these cases there were more than

1 author with the same name.

Our aim was to investigate the authors with the name 'Cécile Martin' in the dataset. From table 5.1 below we can see that there are 7 theses associated with the name Cécile Martin which were authored by 4 different authors. Among these 4 author 1 of them has authored 4 theses whereas the remaining 3 have authored 1 thesis each.

Author	ID	Count
Cecile Martin	81323557	4
Cecile Martin	203208145	1
Cecile Martin	182118703	1
Cecile Martin	179423568	1

Table 5.1: Theses with author name Cecile Martin

5.3 Supervisor ID

We then checked for issues in the 'Identifiant directeur' column which represents the ID of the supervisor(s). First I checked for the length of the id and found that they were not the same throughout the dataset. The table 5.2 below shows the occurrence for the different lengths.

Id Length	Count	Percentage
1	4587	1.15
2	137	0.03
8	255680	64.17
9	78960	19.82
11	59108	14.83

Table 5.2: The count for different supervisor id lengths.

During further investigation we found the following issues:

- 35163 supervisors' IDs end with the character 'X'.
- 59108 supervisors' IDs have commas in the IDs.
- 4724 supervisors' IDs consist of 1 or 2 digit numbers.

I found that to treat the IDs that end with the character 'X' we can remove the 'X' from the end. For the IDs containing commas in them, the thesis had more than 1 supervisor resulting in the IDs to be incorrectly recorded.

5.4 Number of PhD defended

Finally, we investigated the trend of the total number of PhDs defended in a given year. The following figure 5.2 sums up the results.

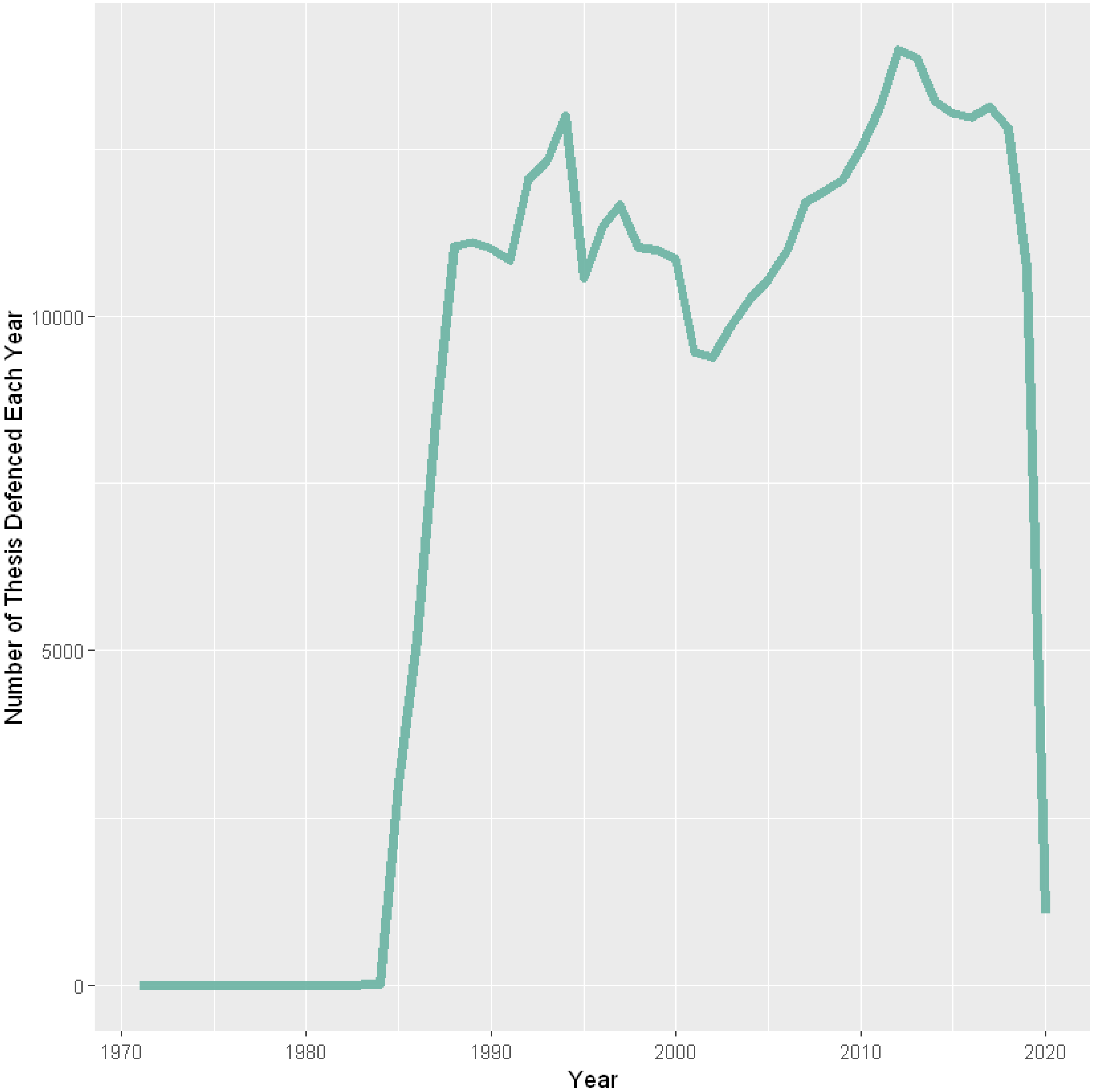


Figure 5.2: Total number of theses defended over the years

From the figure 5.2 we see a sudden rise in the total number in 1988 and a significant drop in 2019. The drop can be explained by the Covid-19 pandemic that caused the entire country into a lockdown in 2020.

Chapter 6

Investigating Outliers

After we finished investigating the common issues in the dataset our next task was to investigate the outliers in the supervisors of the theses. We started by investigating the supervisors by checking the number of theses that they have mentored. First we used group by(), filter(), count(), and arrange() functions to get number of theses supervised by each supervisor in the dataset. The table below shows the top 15 supervisors with highest number of theses supervised.

	Supervisor	Supervisor ID	Count
1	Jean-Michel Scherrmann	59375140	208
2	Francois-Paul Blanc	26730774	205
3	Pierre Brunel	26756625	193
4	Philippe Delebecque	29561248	178
5	Guy Pujolle	27084868	177
6	Michel Bertucat	98531891	173
7	Bernard Teyssie	27158578	146
8	Bruno Foucart	26870177	132
9	Henry De Lumley	26997894	132
10	Jean-Claude Chaumeil	58552499	131
11	Michel Maffesoli	27001067	128
12	Roger G. Boulu	59209143	127
13	Daniel-Henri Pageaux	02705554X	124
14	Georges Molinie	02703352X	116
15	Jean Bessiere	26725916	114

From the table we can see that there are supervisors who have supervised a surprisingly large number of theses. For example, Jean-Michel Scherrmann who according to the dataset has supervised 208 theses. To statistically investigate the outliers in these numbers, I used the quantile() function to obtain the values at the 1st and 99.75th percentiles. I obtained values of 35 and 1 for the 1st and 99.75th percentiles, respectively, indicating that any number of these supervised theses less than 1 or greater than 34 is an outlier.

We performed the same method to find the outliers in the authors. After grouping the data by the author names and author IDs, we used the quantiles() function to obtain the values at the 1st and 99.75th percentiles. Suprisingly, I obtained values of 1 and 2 for the 1st and 99.75th percentiles, respectively, indicating that any author who has more than 2 theses defended different than 1 thesis, is considered as having an outlier number of theses defended.

To differentiate between multiple defences or supervisions and homonyms the data was grouped by both supervisor/author name and id.

Chapter 7

Preliminary Results

7.1 Language Categories

Following the investigation of the outliers, we proceeded to process the languages in which the theses were written. We began by recording the language data using the dplyr library into four categories: French, English, Bilingual, and Other. A thesis is considered bilingual if it is written in both English and French or if it is written in English or French ad one other language and the Other category is for any other condition that does not fall into the first three. We began by recording the languages in R using the mutate() function from the dplyr package, and then we used filter(), group by(), and count() to determine the number of theses written in each language category.

Language	Count
French	334406
English	30942
Bilingual	16488
Other	2043

Table 7.1: Frequency theses written in each language category

Table 7.1 shows the frequency of each language category in the dataset, and it's clear from the frequencies that French is the most commonly used language in writing theses, which makes sense given that these theses were written in France.

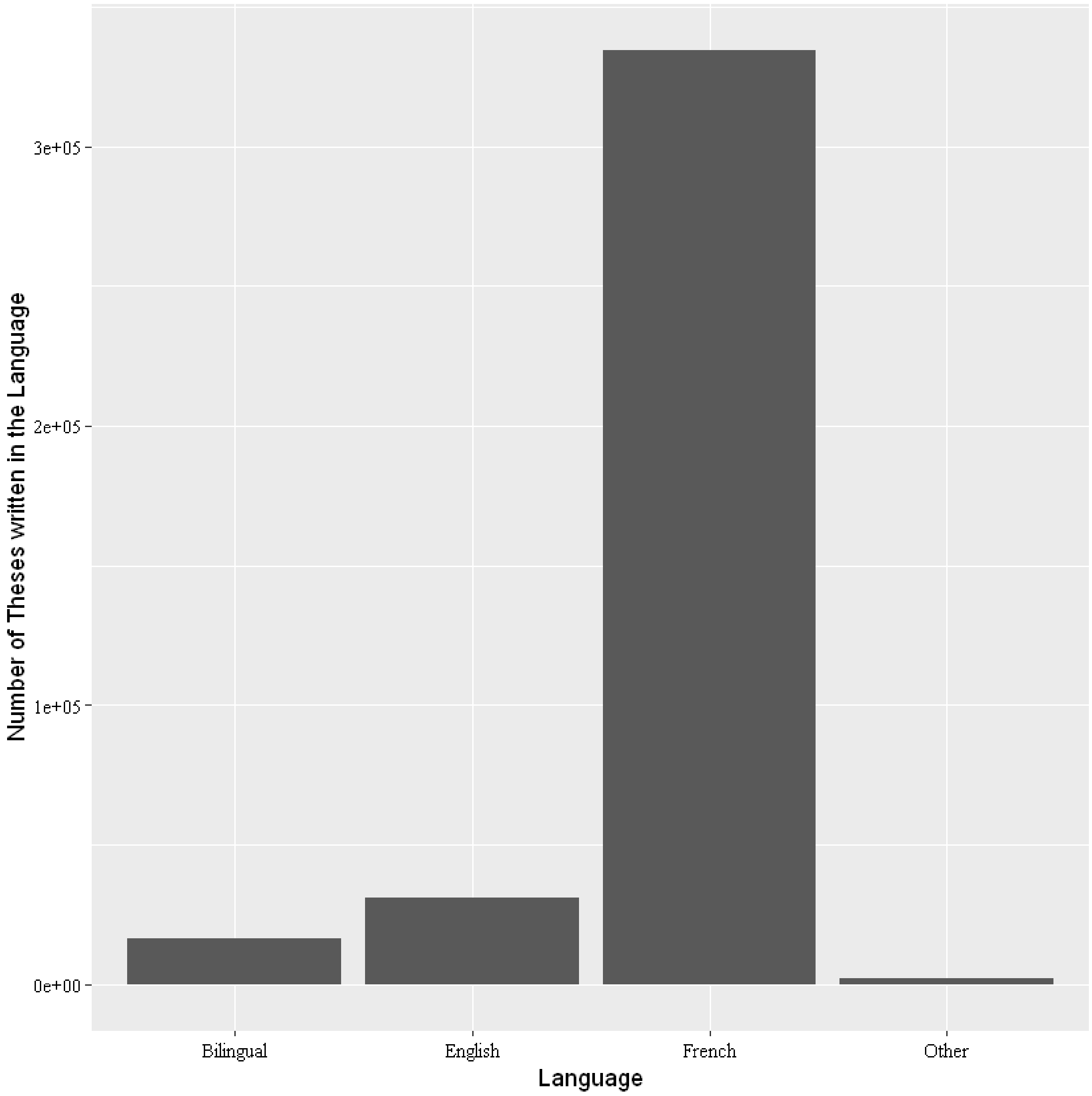


Figure 7.1: Number of theses written in each language category

We plotted a bar graph shown in the figure 7.1 above to get a better look at the frequencies of the language categories in the dataset, and it confirms our observation that the French language is the most commonly used in writing theses in France.

Furthermore, we were curious to learn how the choice of language has evolved over the years. For this we used the dplyr library to filter, group, and count the theses that were defended during the same time period. To get the evolution we computed the percentage of theses written in each language category over the total theses written for each year and then we plotted graphs to visualize this using both ggplot2 and plotly.

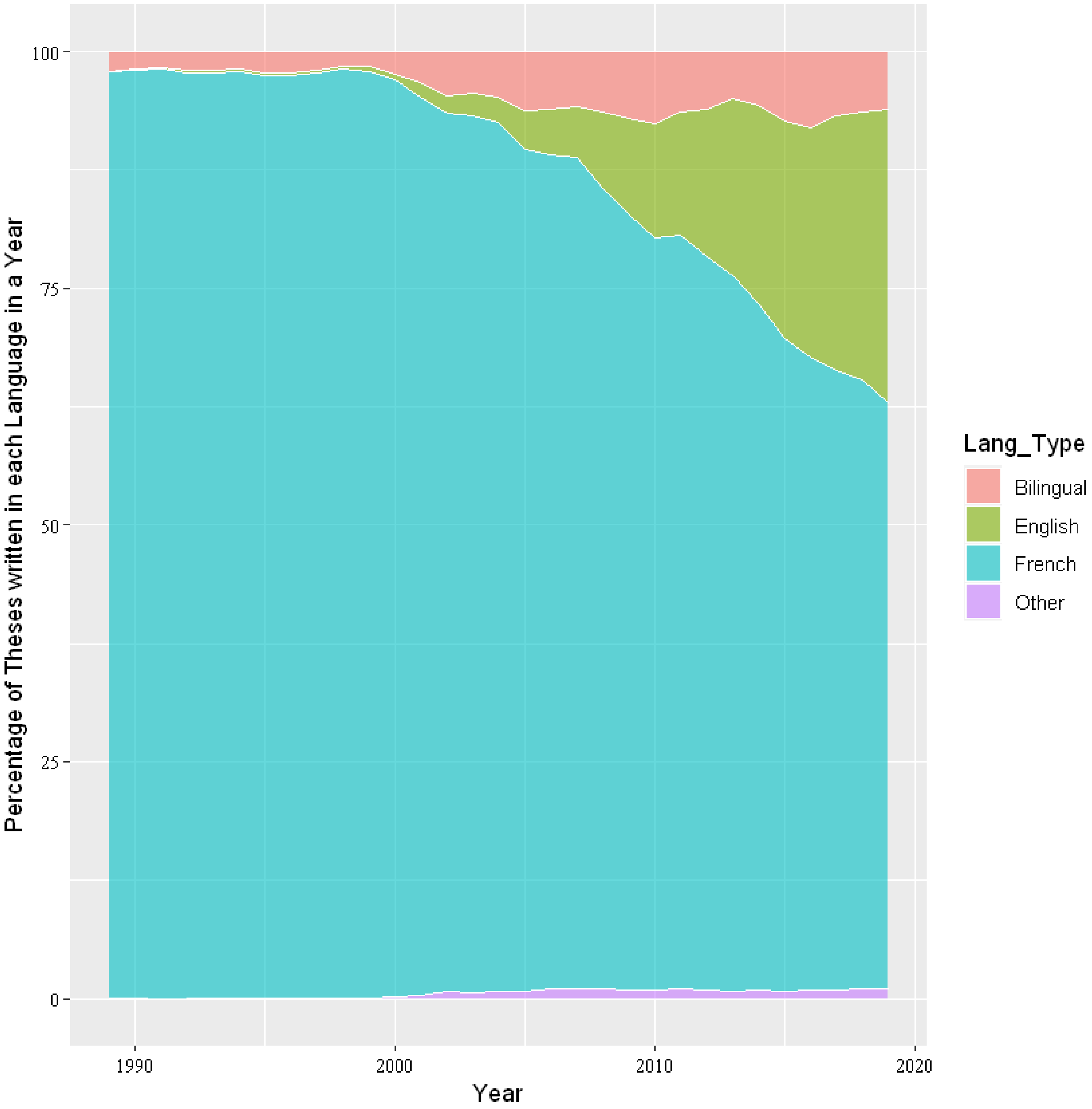
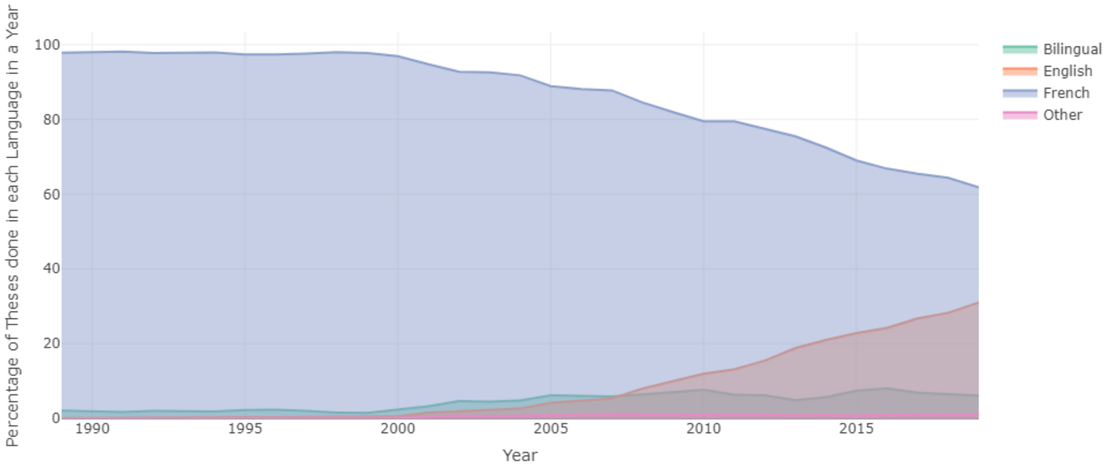
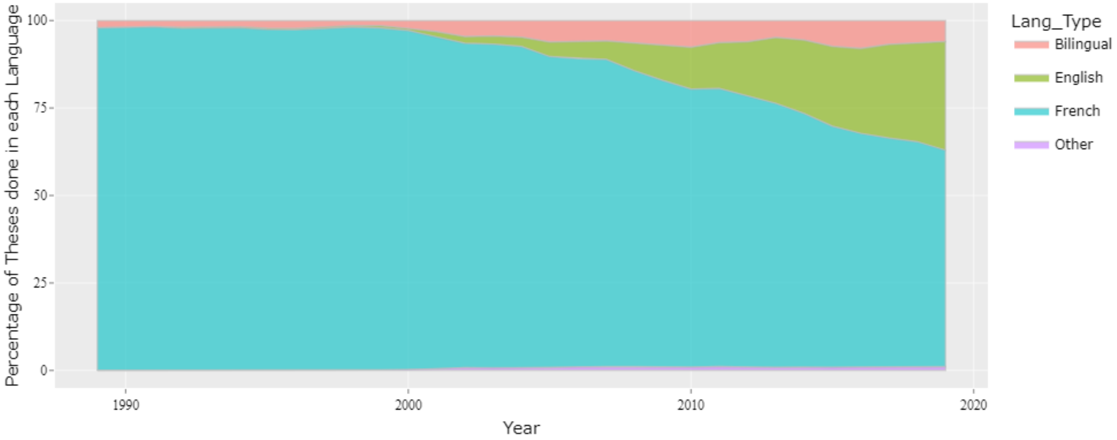


Figure 7.2: Percentage of theses written in each language category over the years (ggplot2)



(a) Percentage of theses written in each language category over the years



(b) Trend of the percentage of theses written in each language category over the years

Figure 7.3: Plotly Graphs

Surprisingly, as shown in Figures 7.2 and 7.3, the trend of writing theses in English has increased significantly and the has decreased for French over the last two decades.

7.2 Period of the year defend PhD theses

Furthermore, we investigated on which month do authors typically defend their theses. We filtered the data for theses defended We used dplyr to group, filter, and count the data, as we had done previously.

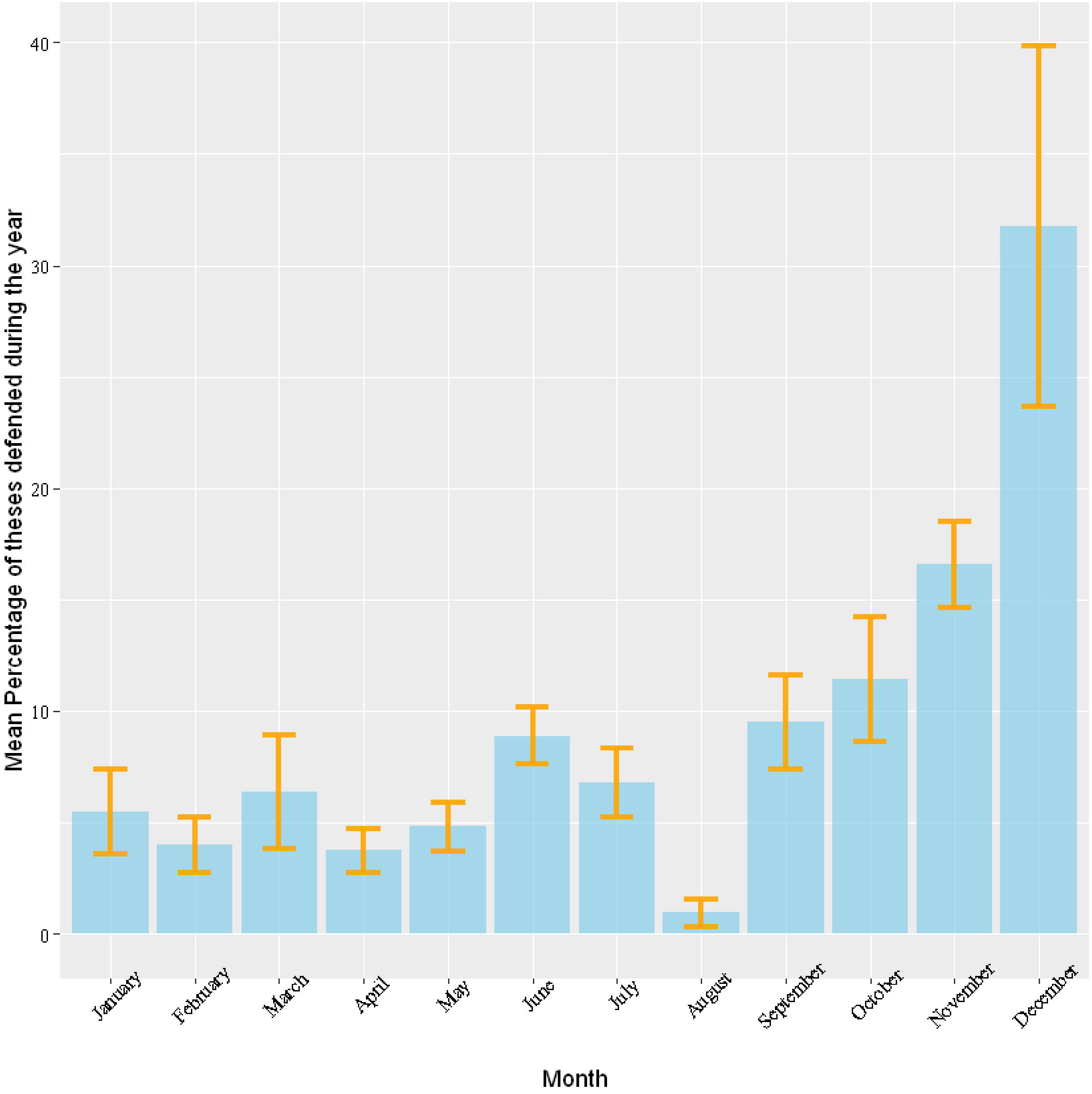


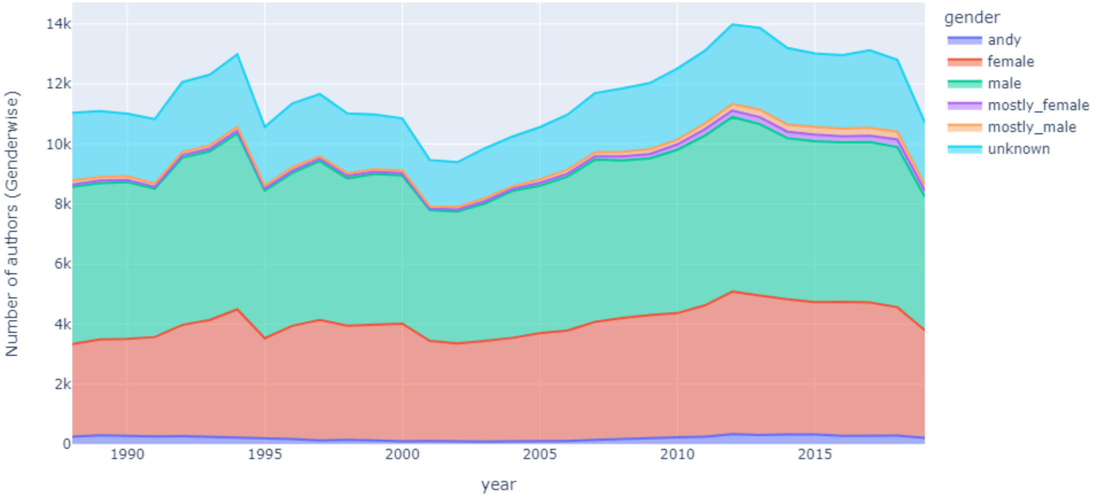
Figure 7.4: Percentage of theses written in each language category over the years (ggplot2)

Figure 7.4 shows that approximately 31.75% of authors defend their theses in December, while only 0.94% defend their theses in August. The dip in August can be explained by the fact that in August almost everyone is on vacations in France.

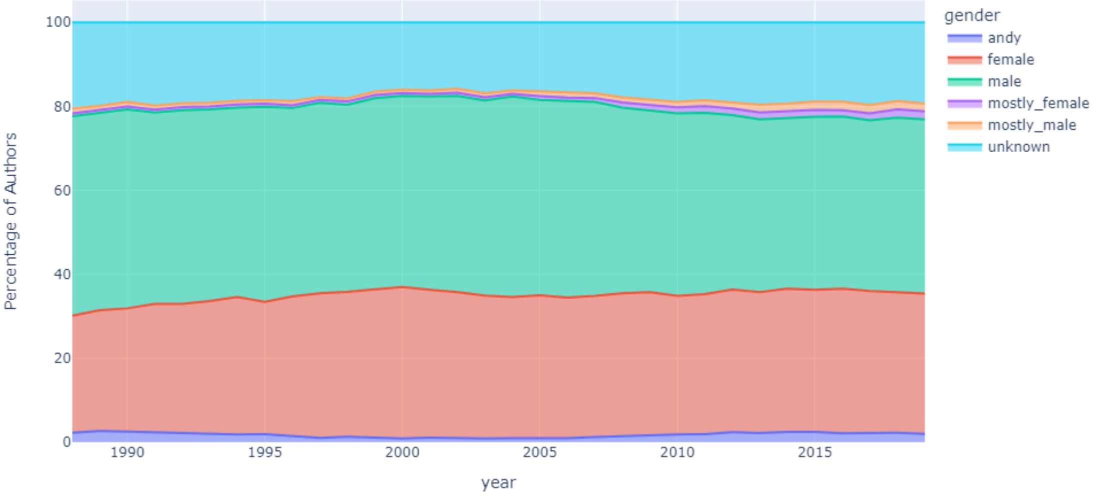
7.3 Gender

7.3.1 Authors

Finally, we investigated the trend of the gender of authors who defended their theses over the years. To obtain this information, we used Python to process the author names column and ran it through a library that detects gender based on the name. We calculated the percentage of authors in each gender category and used plot to create a stacked area plot in Python.



(a) Number of theses written by each gender category over the years



(b) Percentage of theses written by each gender category over the years

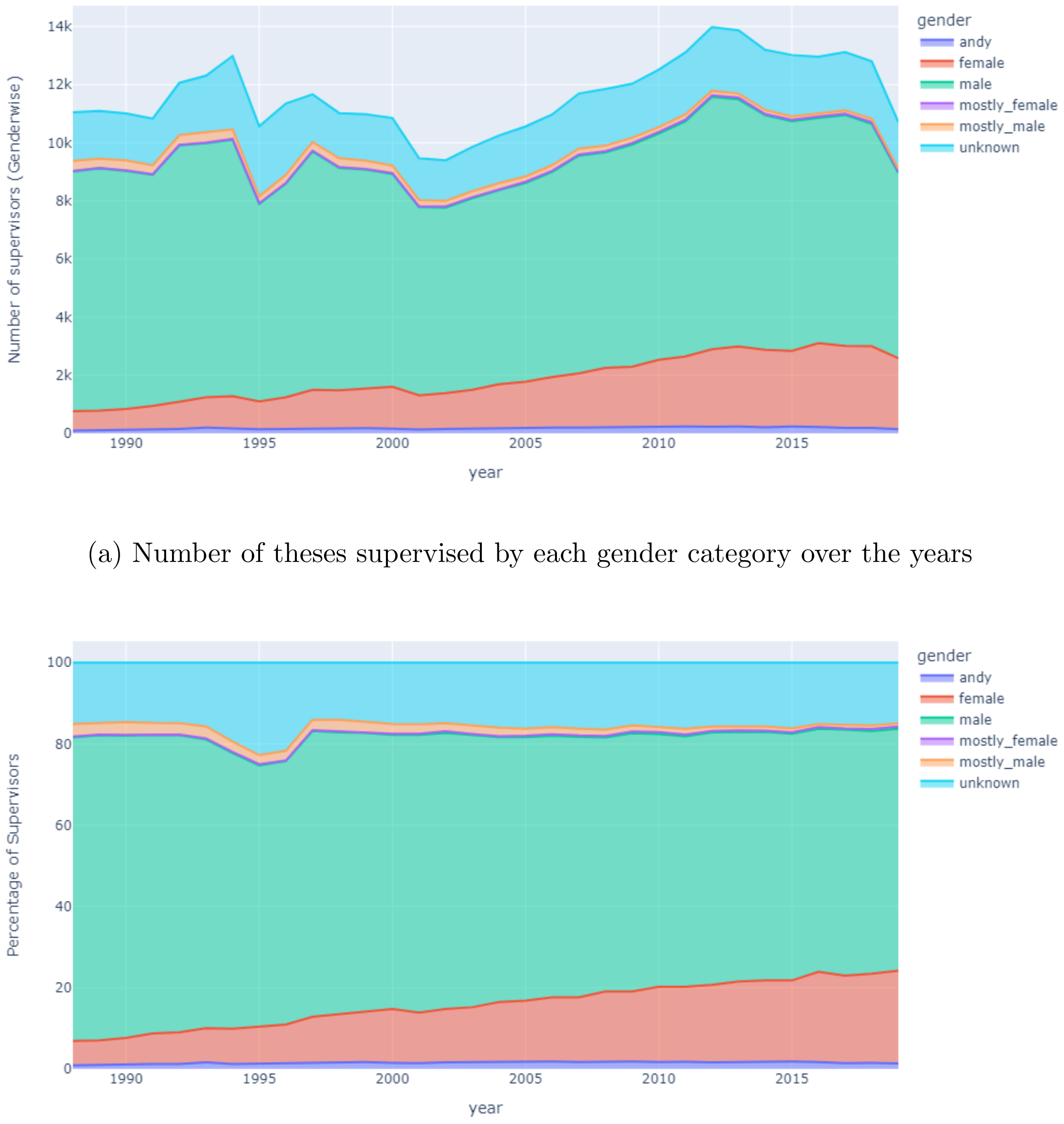
Figure 7.5: Trend of theses written by each gender category over the years

In the graphs in figure 7.5 we see that the gender detection library has categorised the names into 5 groups: andy, female, male, mostly_female, mostly_male and unknown. Here andy stands for androgynous names i.e. equally likely to be a male or a female name and unknown stands for names that the libraries dataset does not include so it does not know in which group the name lies.

We can see a trend where male and female authors have almost the same distribution of the number of theses over the years, but with fewer numbers for females.

7.3.2 Supervisors

We also investigated the trend in the number of theses supervised by each gender category over the year for the supervisors. To achieve this we repeated the same steps that were used to generate figure 7.5. The figure 7.6 shows the the trend over time has stayed the same for percentage of male supervisors except for the dip in 1995 and 1996. It also shows that the trend had been slowly increasing for percentage of female supervisors.



(a) Number of theses supervised by each gender category over the years

(b) Percentage of theses supervised by each gender category over the years

Figure 7.6: Trend of theses supervised by each gender category over the years

Bibliography

[1] Agence bibliographique de l'enseignement supérieur (ABES). Thèses. fre. Text. URL: `http://www.theses.fr` (visited on 10/20/2021).