Data Wrangling

By : Sarvesh MEENOWA

Supervised by: Dr Matthieu Cisel

Bachelor of Data Science by design

October 21, 2021

Contents

List of Figures

1 Missing data

	Variable	Num	Pct_miss
1	Date.de.premiere.inscription.en.doctorat	383668	85.71
2	Identifiant.auteur	129989	29.04
3	Langue.de.la.these	63765	14.24
4	Date.de.soutenance	56746	12.68
5	Year	56746	12.68
6	Identifiant.directeur	49172	10.98
7	Identifiant.etablissement	17085	3.82
8	Mise.a.jour.dans.theses.fr	177	0.04
9	Directeur.de.these	17	0.00
10	Directeur.de.these..nom.prenom.	17	0.00
11	Titre	9	0.00
12	Discipline	5	0.00
13	Etablissement.de.soutenance	4	0.00
14	Auteur	2	0.00
15	Statut	0	0.00
16	Identifiant.de.la.these	0	0.00
17	Accessible.en.ligne	0	0.00
18	Publication.dans.theses.fr	0	0.00

Table 1: Percentage of the missingness of each variable

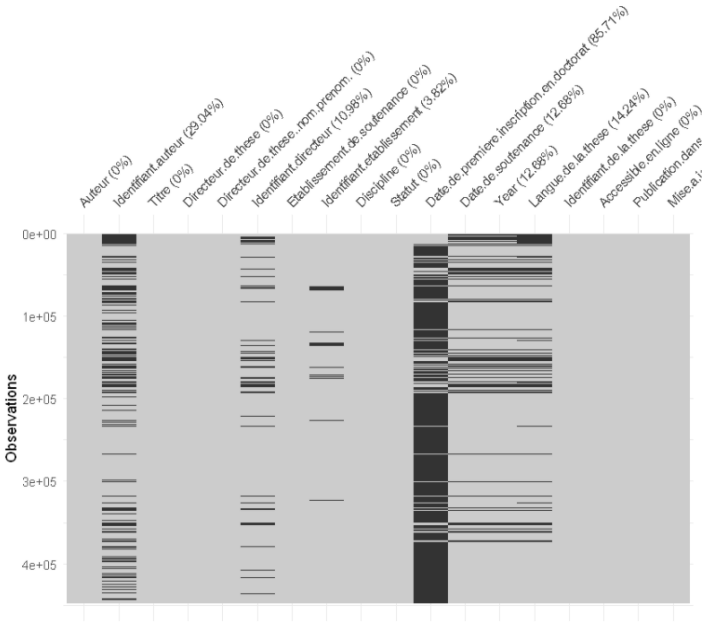

Figure 1: Shows the visual representation of the table 1

From Table 1, 8 variables have missing values, of which few variables had implicit missing data such as the author Id(Identifiant.auteur) and the supervisor Id(Identifiant.directeur). This issue was fixed while reading the document by replacing the empty cells by NA values. From Figure 1, the visualization aid in knowing where the missing values are in the data.The main observation is the relation between the missingness of the defence date and beginning date pattern. There is a symmetry between the defence date and beginning date, for all the missing defence dates, the beginning date is present and vice-versa. This might be because in more than 80% of the time,the defence dates have replaced the beginning dates.

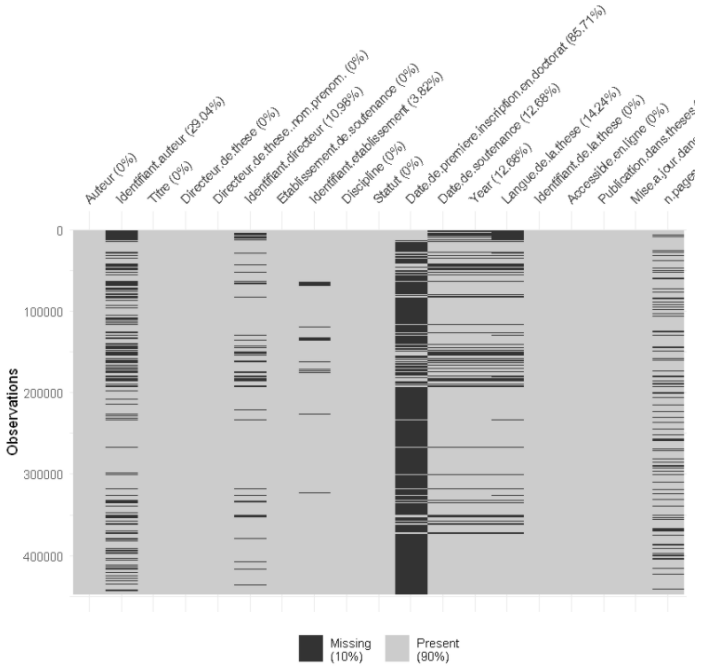

Figure 2: Shows the visual representation of the missingness of data after introducing the simulated variable n.pages

A variable, n.pages, representing the number of pages was created with a normal distribution having a mean of 200 pages and a standard deviation of 50 pages, for a sample of 80 percent of the dataset. The 20 percent that remains is filled with missing values. From Figure1 , the n.pages variable has missing values spread across the column. In this case, the type of missing data is MCAR(Missing Completely at Random), so mean imputation can be a valid approach since the sample mean is not biased. Mean imputation will fill the missing values with the univariate average of n.pages, hence keeping the whole dataset.

2 Common issues

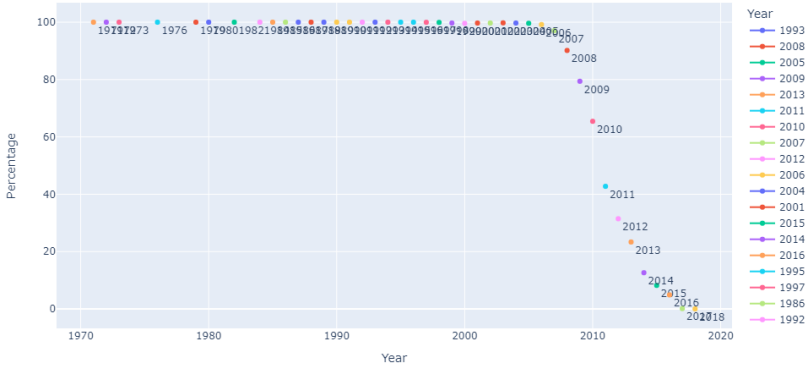

Figure 3: Shows how the proportion of defences at the first of January evolved over the years

From Figure 3, in the last decade, the proportion of defences decreased at a decelerating rate to 0. At the start till around 2005, the proportion of defences plateaued at 100 percent, which means that all theses in that given period were defended on the first of January. This is an unlikely scenario considering that the first of January is a public holiday, so this can be explained by the fact these.fr [1] opened in July 2011 and stores all theses defended from 2006 in the institutions which have chosen to abandon the submission of the paper thesis in favour of electronic support. In this case, they might have filled the defence dates with the first of January for the years prior to 2006.

Auteur	Identifiant.auteur	Discipline	Year
Cecile Martin	81323557	Sciences biologiques fondamentales et appliquees. Sciences medicales	2000
Cecile Martin	81323557	Genie des procedes industriels	2001
Cecile Martin	81323557	Neurosciences	1991
Cecile Martin	81323557	Sciences biologiques et fondamentales appliquees. Psychologie	1994

Figure 4: Shows the case of Cécile Martin grouped by her name and unique ID

Moreover, homonyms are pretty common in the dataset, by grouping author names together, there were slightly more than 14000 occurrences, and after grouping by the author names and their unique IDs, there were around

[1] https://www.theses.fr/apropos.html#: :text=A%20l'ouverture%2C%20en%20juillet,plus%20de%206%20000%

4500 occurrences. So this would explain the difference between homonyms and people who did more than one thesis. From Figure 4, in the case of Cécile Martin, ID, 81323557 appeared 4 times, this means that Cécile Martin worked on four theses in the years 1991, 1994, 2000 and 2001 respectively. However, only three out of the four theses appear to be in a relatively similar discipline, it refers to the theses in 1991, 1994 and 2000. Furthermore, it is unlikely that she worked on two theses in the space of one year. There probably have been a mistake in linking the proper ID to the Cécile Martin for the year 2001.

	Number	**Percentage**
Supervisor IDs	398472	-
More than 1 supervisor	83436	20.9
Supevisor IDs with comma	59108	70.8
Supervisor IDs with 'X'	7479	9.0
Supervisor IDs with Exponential sign	388	0.5

Table 2: Supervisor IDs issues with relative percentages

Furthermore, there were multiple issues with the supervisor IDs. Table 2 summarizes some of the main issues with the supervisor IDs. There are 398472 supervisor IDs and from this amount, 20.9% represent theses with more than 1 supervisor. Most of the issues arose in the situation where there was more than 1 supervisor. The 20.9% represent 83436 theses, and out of this amount, 70.8% has a comma in the IDs while 9% of those, represent IDs having 'X' and only 0.5% have the exponential sign, 'E' in the IDs. For the rest of the IDs, they were numeric but of different lengths(1,2,8,9,11).

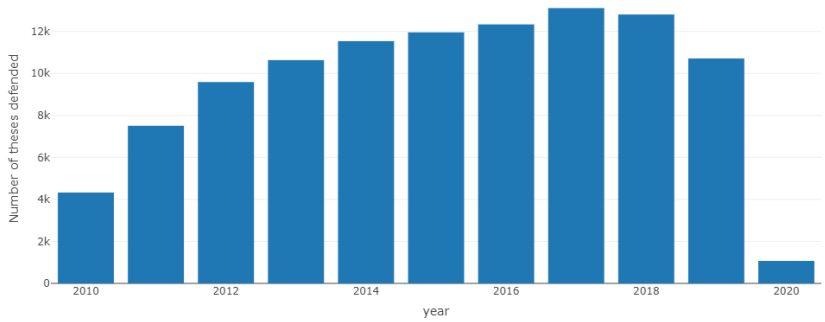

Figure 5: Bar chart to show the number of theses defended from 2010 to 2020.

From Figure 5, we can see a sudden drop in the number of PhD defended in 2019 and 2020. The hypotheses which might explain the drop are the following :

- The attractiveness of doing a PhD has decreased due to fairly average salaries considering the amount of effort it requires.
- The dataset ends in the middle of 2020, perhaps not all theses have been uploaded to the website, so it could be a case of data dredging.
- In March 2020, there was a lockdown in France and universities did not know yet how long it will last, so there might have been some rescheduling or registration of theses in late 2020 or even in 2021.

3 Outliers

In order to detect outliers, percentiles were used. With this method,a lower and an upper bound is set, so all observations outside the bounds would be considered as outliers.A lower bound is implicitly set as number of theses supervised to equal one, since this is the minimum that can be supervised, and it is also representative of 0.003rd percentile. The upper bound is set as 99.7th percentile, since by the empirical rule, 3 standard deviation from the mean would capture 99.7% of our data. The upper bound calculated consider outliers to be the number of theses supervised to be greater than 34. But in order to verify whether it is an outlier or mistake, for the supervisors having the most number of supervisions, the number was cross-checked on the supervisor page where the information about the number of supervisions [2] is provided. For instance, Jean-michel scherrmann has supervised 208 theses from the data and on website, it says that he supervised 209 these. The difference might be due to the fact that the dataset is only till mid 2020. A more robust way instead of manually checking the results would be to use a code that scraps this information(number of theses supervised) and compare it with the result of the count from the dataset. So, it appears that the numbers are not mistakes.

For authors, after grouping by author names and filtering by the number of theses greater than one, we can obtain the number of times the author names have been repeated, which is 14507, but in order to distinguish between homonyms and the case where the author defended more than one

[2]https://www.theses.fr/059375140

these, the data is grouped by author names and their IDs. The number then decreased to 4566, where the difference would indicate the amount of homonyms. We can use a scraper code to check whether the authors wrote the same number of theses as our analysis.

4 Preliminary Results

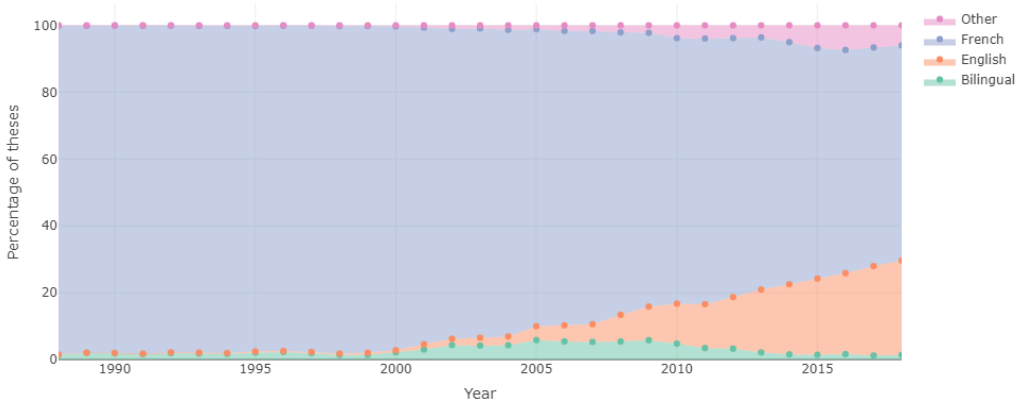

Figure 6: Shows how the proportion of the choice of the language of manuscript evolved over the past decades

Over the past decades, there has been a change in proportion of the choice of the language of theses chosen. From Figure 6, from the year 1988 to the early 2000s, French was the primary choice of the language of manuscript, accounting for more than 90% of the theses. During that period, only a minority of less than 10% used other languages(English not included in the category 'others'), English and a combination of French and English. After 2005, English as choice of manuscript gained popularity rising to a proportion of around 30% in 2019, while French language usage decreased to around 60% in 2019. The choice of using both English and French always stayed below the 10% margin and more recently very rarely used as language of manuscript at less than 5%. Choices of other languages has also increased in proportion,albeit a small increase, to around 5-10%.

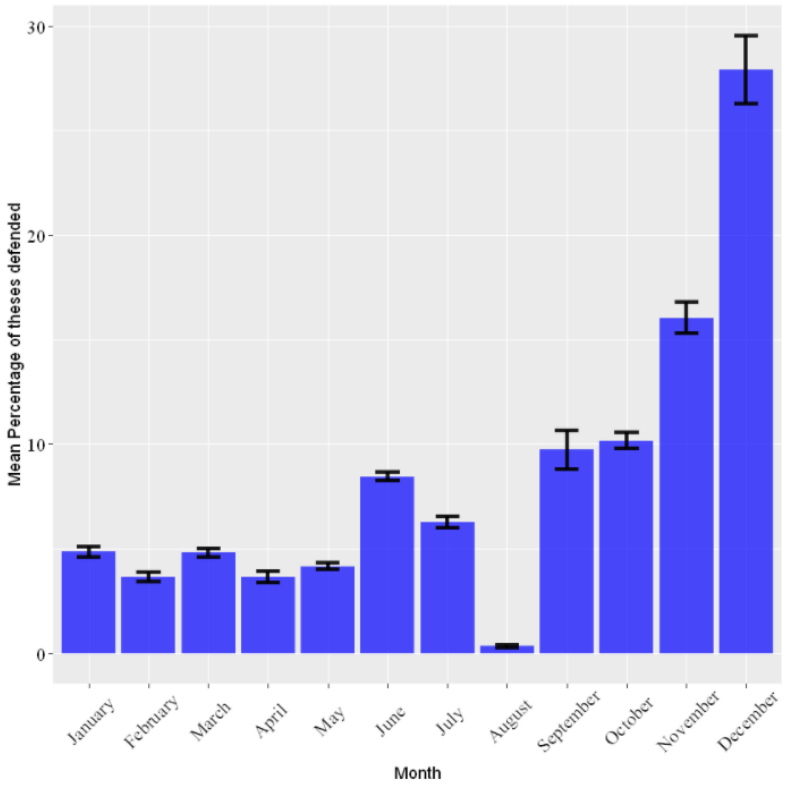

Figure 7: Shows the mean percentage of theses defended in each month from 2009 to 2019 included.

Figure 7 gives information about the period of the year that PhD candidates tend to defend their theses. The amount of theses defended appear to be consistenly low from January to May at around 10%. After almost doubling in proportion from May to June, there is a sudden dip from June to August to around 1-2%. With a sharp increase from August to September, there was again an abrupt increase in the last few months of the year, with a peak of slightly more than 25% of theses defended in December.

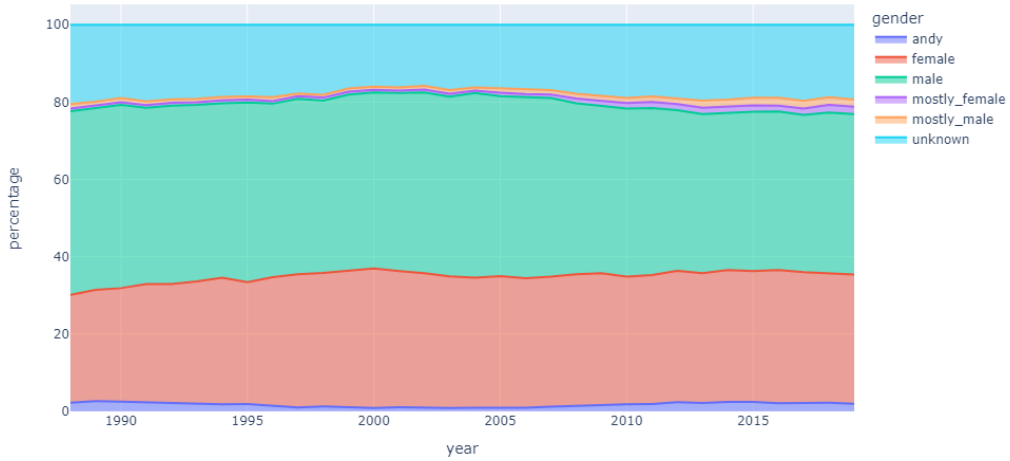

Figure 8: Shows the evolution of gender of PhD candidates from 1988 to 2019.

Using a gender derivation library which deduces the genders from author names, there were 3 major categories of gender, namely, male, female and androgynous(labelled as andy).From Figure 8, the distribution of the proportion between the 3 types of remain fairly steady from 1988 to 2019. Males account for maximum of the proportion of the PhD candidates, steadily representing around 50% for the period of 1988-2019. The amount of female PhD underwent a slight increase from 1990 to 1995 and from then till 2019, a consistent proportion of around 30% was maintained. For the androgynous candidates, they represented an almost uniform proportion of 20% from 1988 to 2019.

5 Annexes

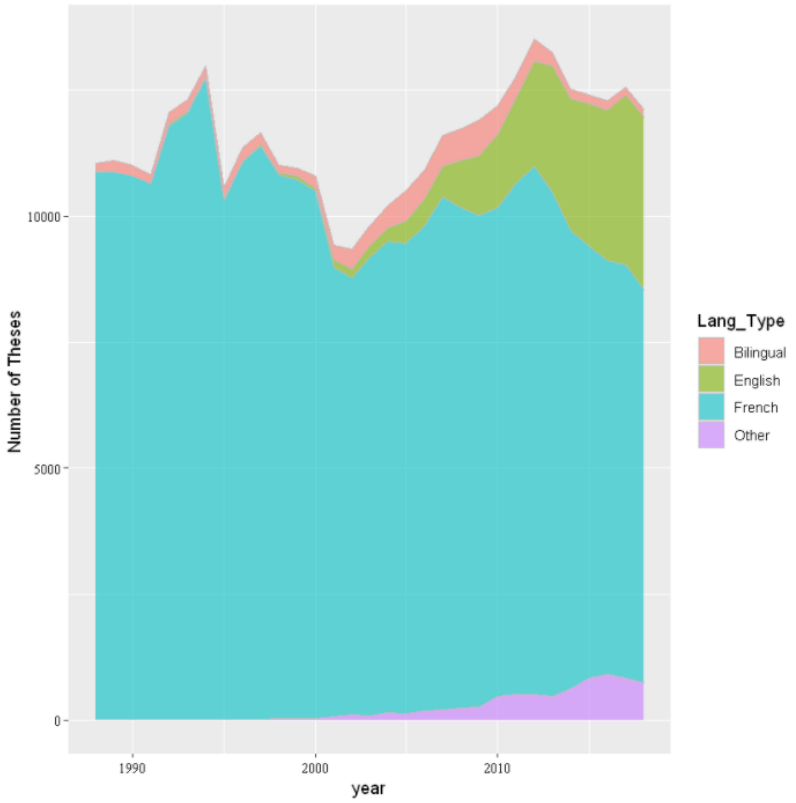

Figure 9: Shows how the trend of the choice of the language of manuscript evolved over the past decades

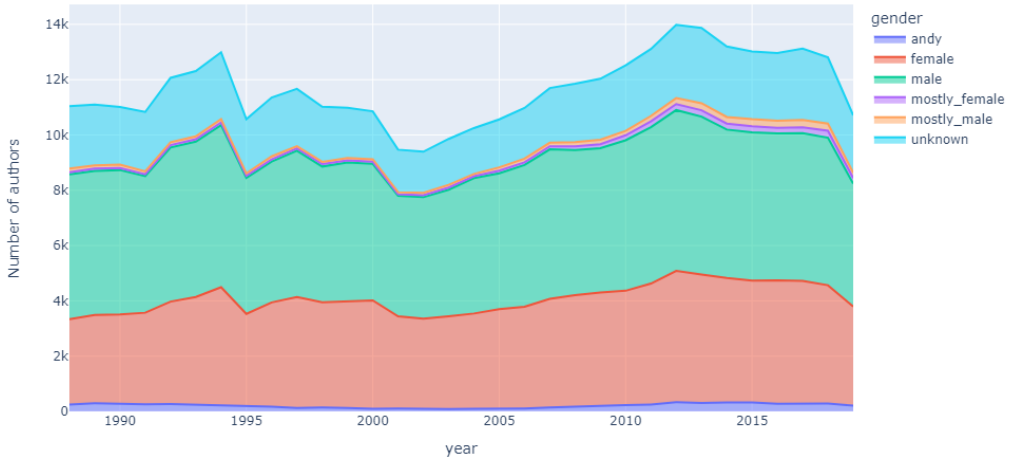

Figure 10: Shows the number of theses defended by each gender of PhD candidates from 1988 to 2019.

6 references

Thèses—A Propos de theses.fr. https://www.theses.fr/apropos.html
Jean-Michel Scherrmann - https://www.theses.fr/059375140#directeurSoutenue