CY Tech Sciences et Techniques

Data Wrangling

Author
Anh Thu DOAN

October 21, 2021

Contents

List of Figures

1 Missing data

Missing data is a very common problems that can make a strong effect on the process. Before any analysis, first we need to making sure that the data set is as accurate as possible. In Python, we have the library names "missingno" that provides a tool to help us represent our missing data in a very easy way. Missing data visualizations is the best way to help us have a over view of the distribution of missing value in our data set.

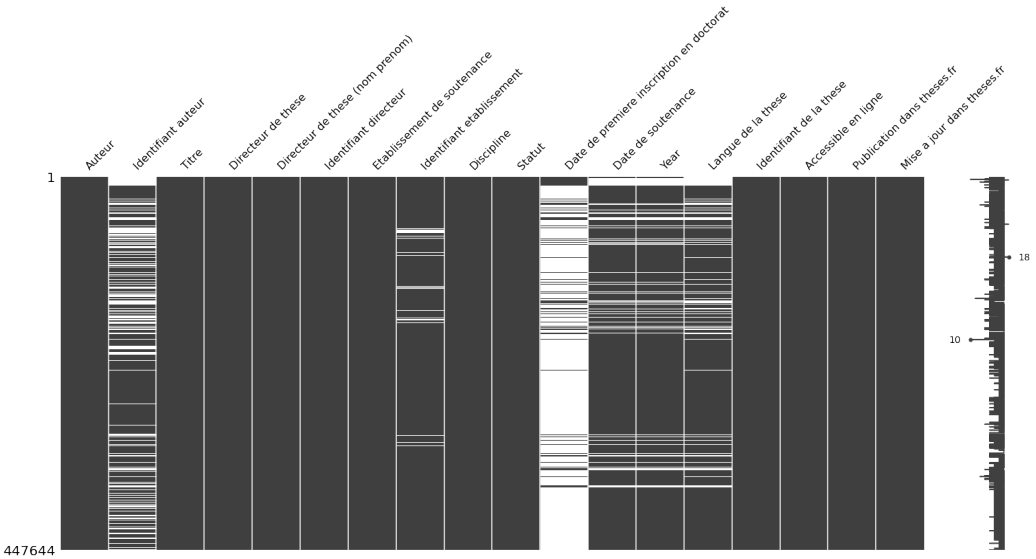

Figure 1: Missing data

First of all, less than half of the columns have the missing value. Then, the "Date of the premiere inscription en doctorat" is the column has the most missing value in it, along with that it is easy to see that there is an opposition between "Date of the premiere inscription en doctorat" and "Date de soutenance" which can be assumed as the input source data only take one out of the two information.
Secondly, the three variables such as "Date de soutenance", "Year", "Language de la these" have a major common in the missing graph. So if one these is not had one variable, they may also lack of these others two.
Finally, the spark-line on the right side summarizes the overall shape of the data completeness and shows the rows with maximum and minimum NULL values in the data set.

2 Common issues

2.1 How common are the defences on the first of January

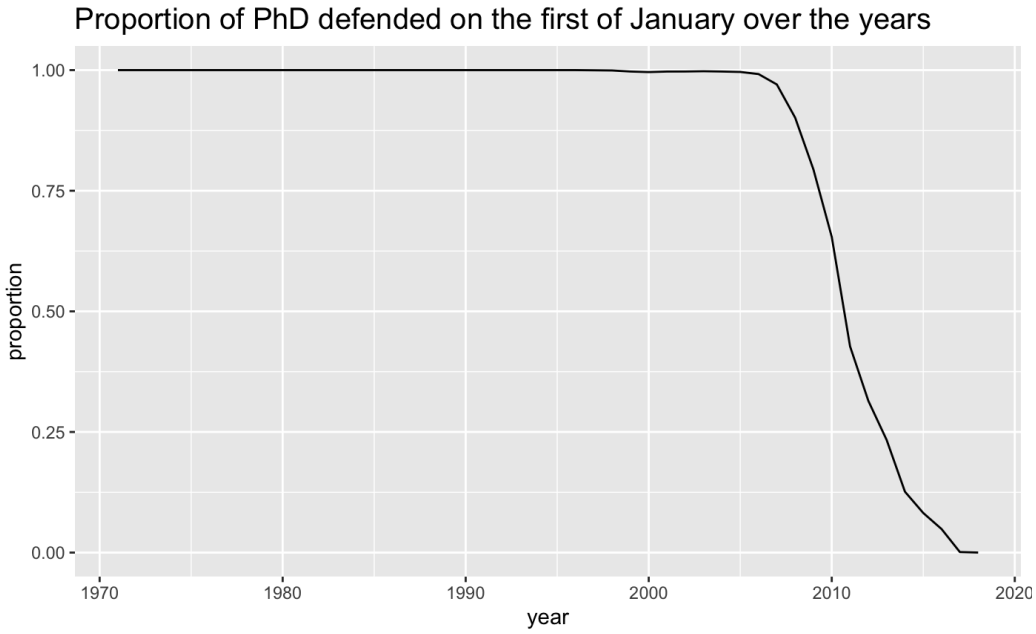

Figure 2: The proportion of PhD defended on the first of January over years

From the graph above, we have figured that the proportion of PHD which have defence date on the 1st of January of each year from 2007 to 2020, has dramatically made the statistic drop. The table data clearly shows that the downward trend of the proportion of defenses at the first of January only started from 2006 (before that it was always 1) and has significant dropped off in 2009 from around 0.73 and continued to decrease until 2020.

2.2 Cécile Martin

It is apparently seen from Figure 4 that there are 7 rows named Cécile Martin but they're only 4 different ID so that we can conclude that there are only 4 different Cécile Martin, but only one of them has defended 4 theses.

Description: df [9 × 4]

	year <dbl>	freq.x <int>	freq.y <int>	proportion <dbl>
30	2006	10885	10975	0.9917995
31	2007	11349	11697	0.9702488
32	2008	10686	11854	0.9014679
33	2009	9554	12033	0.7939832
34	2010	8190	12516	0.6543624
35	2011	5605	13110	0.4275362
36	2012	4398	13985	0.3144798
37	2013	3237	13868	0.2334151
38	2014	1666	13202	0.1261930

9 rows

Figure 3: The Table of the Proportion of PhD defended on the first of January from 2006 to 2014

Auteur <chr>	Identifiant auteur <chr>
Cecile Martin	203208145
Cecile Martin	81323557
Cecile Martin	179423568
Cecile Martin	81323557
Cecile Martin	81323557
Cecile Martin	81323557
Cecile Martin	182118703

Figure 4: The table of Auteur name Cécile Martin

	length <int>	freq <int>
1	1	4587
2	2	49309
3	8	255680
4	9	78960
5	11	59108

Figure 5: The table of The supervisor's ID length summary

2.3 The supervisor's ID

According to the table in Figure 5, it can be clearly observed that there is an issue with the supervisor's ID. Some of them only have one or two digits, whereas some others have even eight, nine, or eleven digits. We assume that those can be different ways to reference ID. This issue can be caused by the input source from where it has not been unify by the form of the ID. Thus it may leads to ambiguous analysis.

2.4 The number of PhD defended in 2019 and 2020

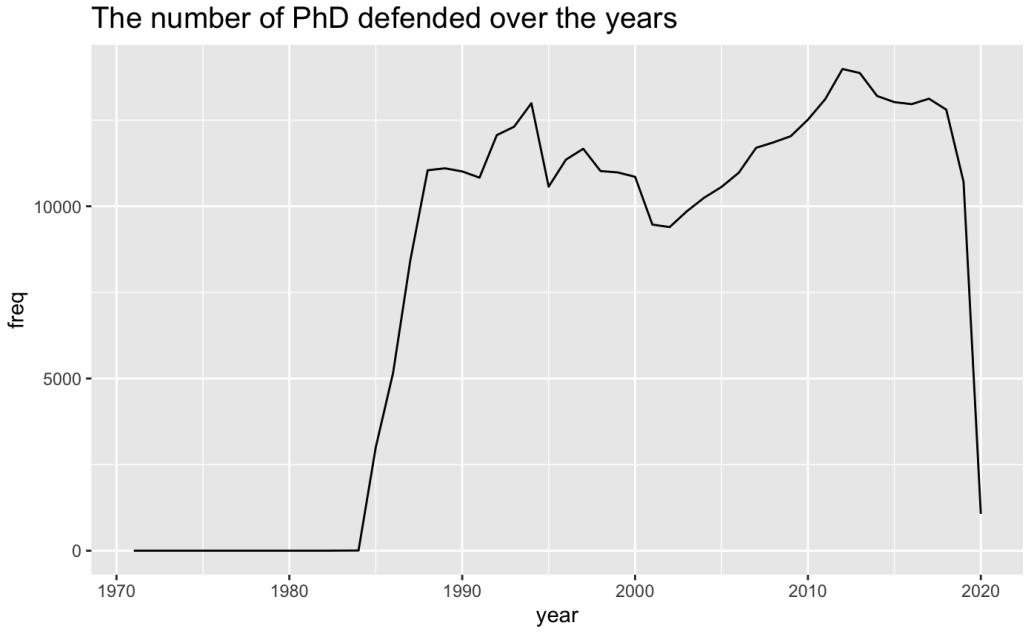

Figure 6: The number of PhD defended over year

It can be concluded from the graph Figure 6, there was a huge reduction that hit the lowest point for the last three decades. Apparently, the reason of the falling trend is due to the covid outbreak that impeded the theses defences during that period. Secondly, due to the potential pandemic threats, it has lead to most of the universities to closure with a virtually attendance for all the countries that extended the deadline period for theses defences. Last but not least, the experience of trying to complete the degree during the pandemic without any physical attendance with supervisor make it even harder.

3 Outliers

3.1 Supervisor

Directeur de these <chr>	Identifiant directeur <chr>	n <int>
Jean-Michel Scherrmann	59375140	208
Francois-Paul Blanc	26730774	201
Pierre Brunel	26756625	193
Philippe Delebecque	29561248	178
Michel Bertucat	98531891	173
Guy Pujolle	27084868	172
Bernard Teyssie	27158578	146
Bruno Foucart	26870177	132
Henry de Lumley	26997894	132
Jean-Claude Chaumeil	58552499	131

1-10 of 10 rows

Figure 7: The table of The supervisor's ID and name summary

The given table above shows the name of supervisor and how many theses they had mentored. And the one has a surprisingly large number of PhD candidates is Jean - Michel Scherrmann, who has 208 theses he have mentored over the period.

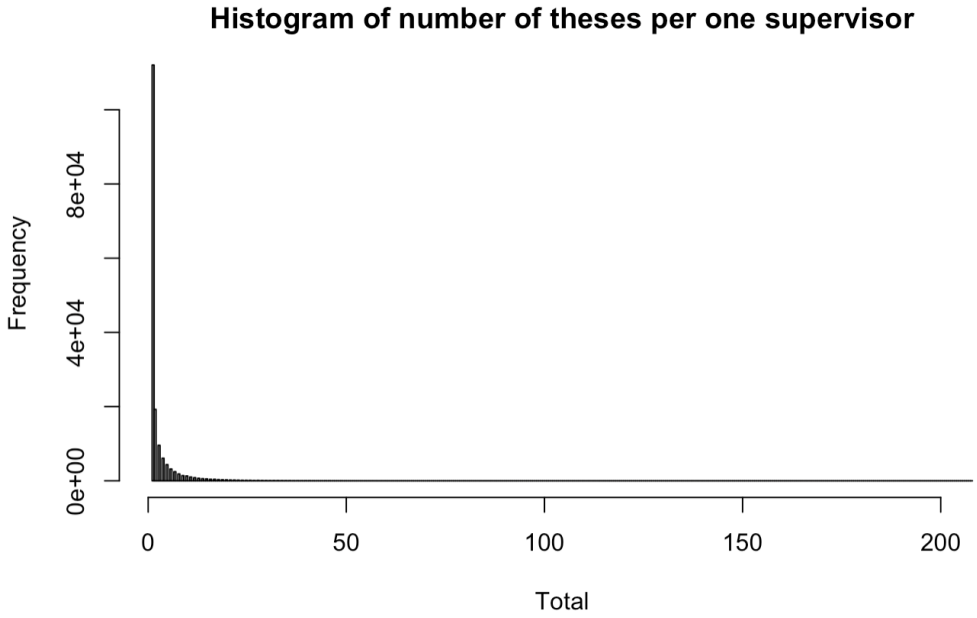

Figure 8: Histogram of number of theses per one supervisor

From the histogram 8, there seems to be some observations much higher than all other observations.

```r
library(outliers)
test <- grubbs.test(supervisor$n)
test
```

```
        Grubbs test for one outlier

data:  supervisor$n
G = 43.42271, U = 0.98884, p-value < 2.2e-16
alternative hypothesis: highest value 208 is an outlier
```

```r
test <- grubbs.test(supervisor$n, opposite = TRUE)
test
```

```
        Grubbs test for one outlier

data:  supervisor$n
G = 0.34767, U = 1.00000, p-value = 1
alternative hypothesis: lowest value 1 is an outlier
```

Figure 9: Grubbs test

Apply the Grubbs test to test whether the highest value and the smallest value is an outlier: As we see from the out put in the Figure 9, the p-value is < 0.001. At the 5% significance level, we conclude that the highest value 208 is an outlier.
On the other hand, the R output indicates that the test is now performed on the lowest value (see alternative hypothesis: lowest value 1 is an outlier). The p-value is 1. At the 5% significance level, we do not reject the hypothesis that the lowest value 1 is not an outlier.

3.2 Author

	Auteur <chr>	Identifiant.auteur <chr>	n <int>
1	Catherine Leport	69413916	7
2	Philippe Blanc	85924660	6
3	Thierry Martin	60151013	6
4	Beatrice Durand	56833776	5
5	Eric Renault	34296565	5
6	Nathalie Martin	27013340	5
7	Pascal Andre	55750931	5
8	Patrick Martin	78079365	5
9	Philippe Andre	61648493	5
10	Philippe Chevalier	66761999	5

Figure 10: Top 10 PhD have the most theses

4 Preliminary Results

4.1 Languages

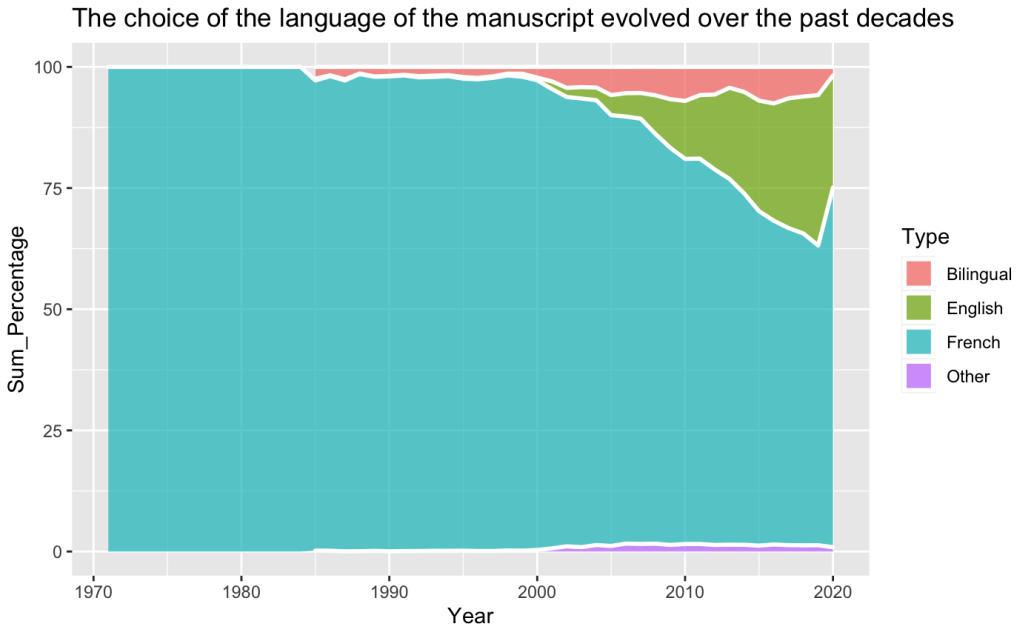

Figure 11: The choice of the language of the manuscript evolved over the past decades (ggplot)

The presented graph Figure 11 illustrates the comparison of the evolved of the languages of the manuscript over the past decades. As can be seen, the main languages was used is French in the first 3 decades from 1970 - 2000. Beside that, Bilingual which are English and French was been developed from 1990 and slight increasing. Following that, English started to be use in theses from 2000 and increased sharply around 30% in the period 2000-2020. Generally speaking, French is still the main languages is used in these in France, but English is also being developed and used more widely.

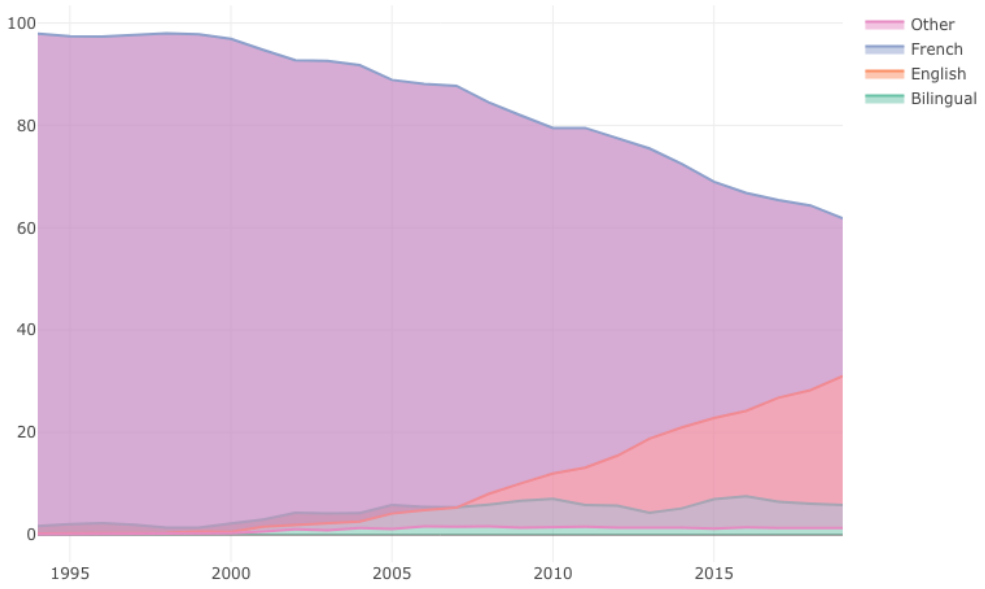

Figure 12: The choice of the language of the manuscript evolved over the past decades (plotly)

The Figure 12 represented the same with Figure 11, this was just used different tool to made.

4.2 Defended period in year

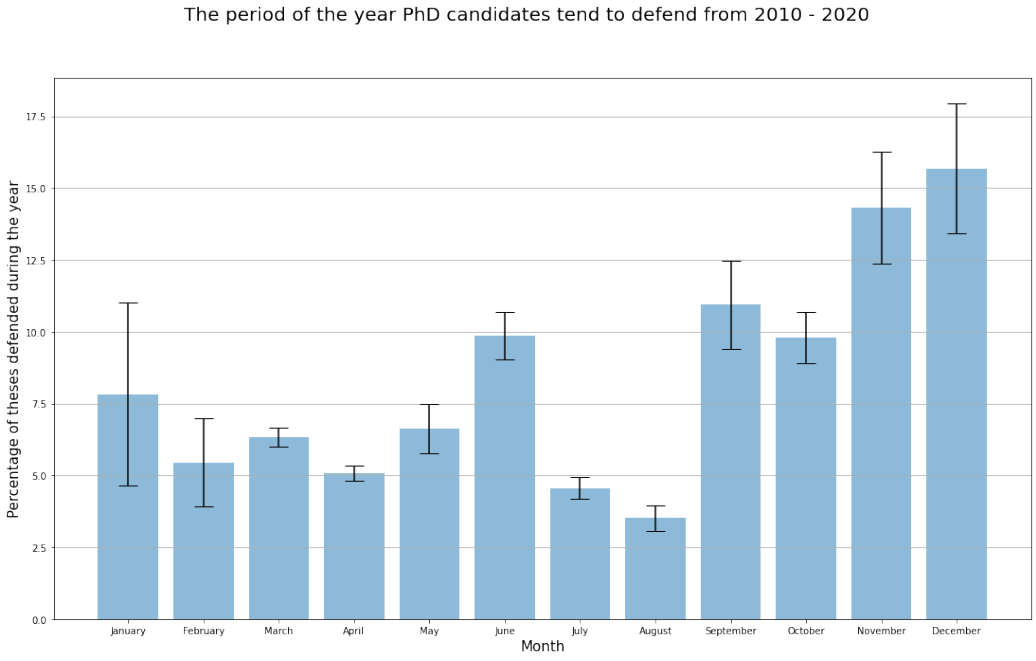

Figure 13: The period of the year do PhD candidates tend to defend

The figure 13 is the ratio of theses defended in function of the month. Indeed, from those figure, we could make a first observation on the trend, which tell us that theses defense are willing to happen within the two last months of the year. Indeed, the spike for November and December seems be very important compare to the rest of the year. According to our research in order to understand this phenomenon, it seems appear that in Frances there is a specific regulation, which grant student (especially student who defenses thesis) the right to keep their student status until the end of their thesis defenses. Nevertheless, this right is only applicable for thesis defenses which date is within the current civil year, which means before December 31 of the year the student is enrolled. Thus, we think that most of student want to benefit the maximum of time in order to prepare a good thesis defenses, and try to postpone the defenses date at the further period, without having to pay any extra fees (student insurance care, application fees and so on. . .). Finally, from this analysis, we can figure out that for an overall view, December is indeed the favorite month for thesis defenses, with an average of 15

4.3 Gender detector by name

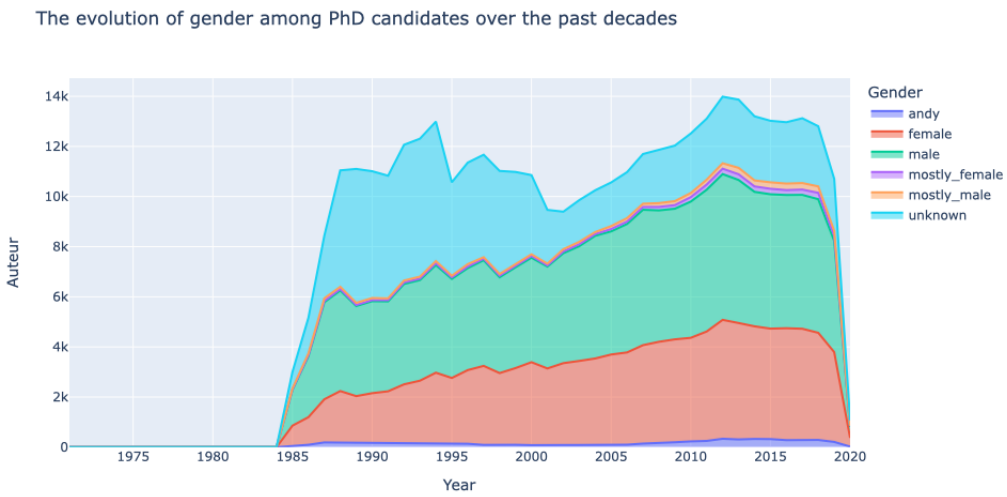

Figure 14: The evolution of gender among PhD candidates over the past decades

The figure 14 is the evolution of gender regarding PHD candidate over the past decades. Indeed, within this analysis we did not want to make a differentiation of between male or female. The main aim for us through this analysis is to monitor the amount of female who are taking PHD. Actually, we have classified the graph with several hypothesis. As the matter of fact, in order to reinforce our analysis we set some hypothesis in function of the name of the candidates. For some of them for whom we are sure that is a female name, we have classified those as female (red graph). For some other inputs, where we are not so sure, we have assume that the name is a female name as well. Nevertheless, we have figured out that even though we have set hypothesis, there is a common trend, which appears among the graph. Then the trend shows us that within years, the amount of women has increase compare to the early period. Except for the covid outbreak period, we acknowledge a significant and stable increase within years. This evolution is certainly explain by the education accessibility to women. Indeed, with the modern way of thinking, the society are more willing to accept the fact that a woman can take longer studies, without having to take care of children or being a housewife. Furthermore, in the modern society, the gender equality is an important asset, and make women feeling more confident on their career

or way of life. In order to illustrate our thought, we can quote the fact that more and more women are leading a company as CEO or director. This situation is a bit encouraging for the future, especially in terms women rights.

4.4 University

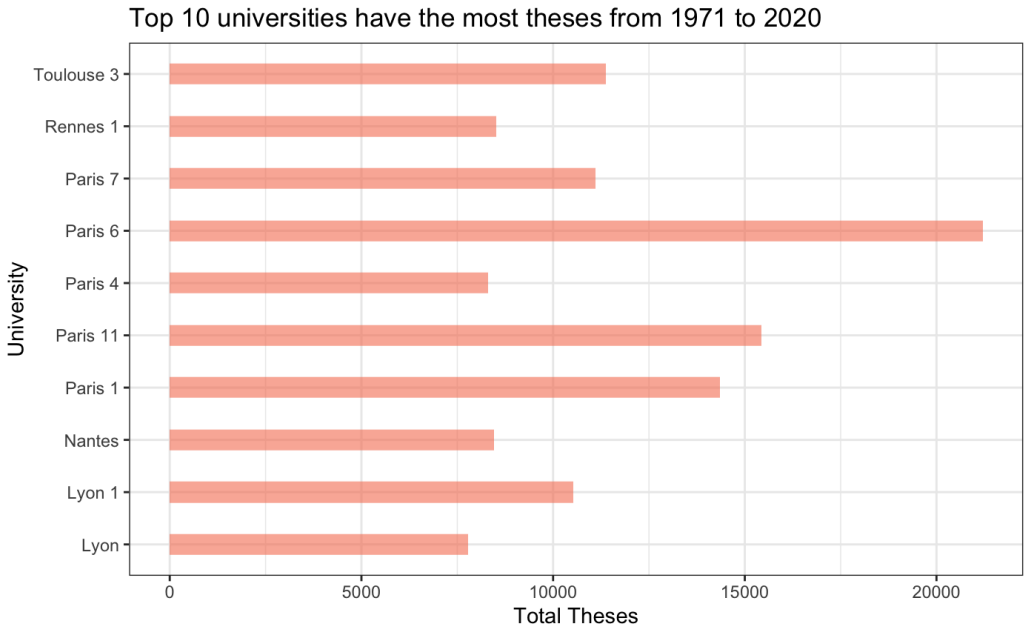

Figure 15: Top 10 universities have the most theses from 1971 to 2020

The figure 15 is the top ten of French Universities, which have the most of theses defense. According to this graph, we could observe that most of universities, which are quote, are located in the biggest cities of France. Indeed, as expected Paris remains on the Top of the list as Paris is a nerve center of France, where every big decisions are taken. Paris is a well-known place to study, and where big industries such as Social Media industry, Luxury and many others. Furthermore, among this ranking we have observed as well that the others cities which host the rest of universities are nothing else than the other French leading edge technology poles. For instance, Toulouse where the heart of European Aerospace industry has an important footprint and Lyon where the French Pharmaceutical industry is well renowned. Finally, from this analysis we could state that French higher education are more centralized and located by poles, which is easier to find.