Data Mining report

Franck Sirguey

April 2023

Contents

1 Introduction

France, through its history, has made sure that it is the nation where higher education is accessible to all (regardless of gender or social class). It is also one of the countries that shines the most in terms of its universities, the level of studies required and its will to train qualified people.

But where do French students go? What are the trends in higher education in France since 2001? What kind of students are they? What are the different French fields study?

We will answer these questions by analyzing the French government's higher education data (data.gouv.fr), which will allow us to understand the trends and answer the various questions.

2 Problem Understanding

From the government database, I will try to sort the different data in order to have the distribution by sex of the different fields, the share of the students in the private versus the public, the number of students as well as to make a prevision of this number (with the help of a linear regression) and to finish to make a map to see the distribution of the students in France. We will use data mining and the language R for answer at these question.

3 Data Description

We have a database composed of 371 195 data and 16 features (size of database = 82Mo). Among these features the most important are :

- the academic year (from 2001-2002 to 2020-2021), it will allow us to have our results according to the year ,
- the geographical level (if it is a value that can be assigned to a municipality, a department, a region or the country), we will therefore sort our data according to the necessary geographical level.
- geo name (gives the name of the place), very useful to map the distribution of students
- the grouping (allows us to know the field of study), which will allow us to analyze trends in the training of students
- sector (public or private) and gender (male or female)
- the effectif who will allow us to have the number of students associated to these different values.

4 Data Understanding

4.1 Analysis of the number of students per year.

4.1.1 The graph

In the first step, we plotted the number of students per year in France.

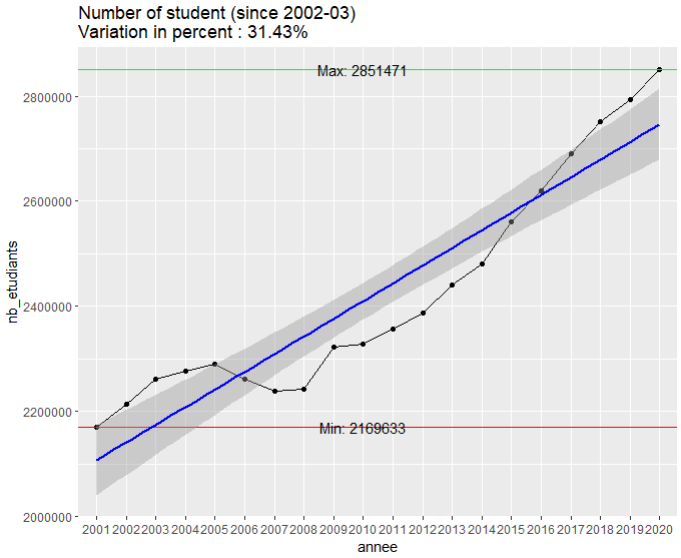
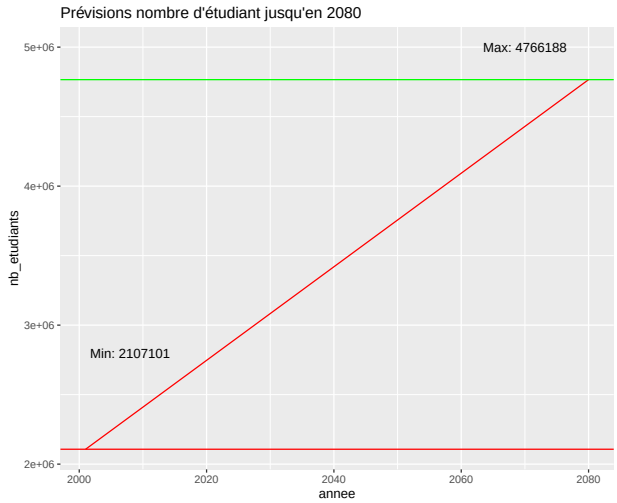


Figure 1: Number of students per year in France

As can be seen, there has been a significant increase in the number of students since 2001, with even more pronounced growth in the last 10 years (there is a total increase of 32% in 20 years). The blue line represents the linear regression line from the current years and will be used to predict future years. It would have been interesting to have data on the different sectors of work in France (primary, secondary, tertiary) to explain the causes of this increase because it represents 3.4% of the French population in 2001 against 4.2% in 2020.

4.1.2 The regression linear

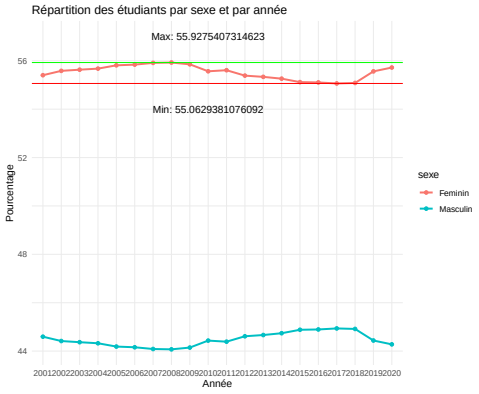
Like i said before i used a linear regression to predict the number of students that there will be in the next 60 years in France. For that i used the data of the number of students between 2001 and 2020, and with the coeficient I obtained the number of students estimated on the next 60 years (in the case where the increase of the number of students follows the same trend)



So if we follow the same tendency we will arrive at 4766188 students in 2080, we will have multiplied the number of students by 2.3.

4.2 Analysis of gender of student per year

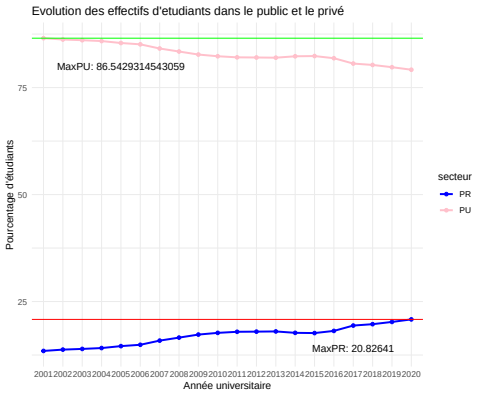
I made a graph considering the year and gender of the different student populations in higher education.



This graph shows us that there is a significant disparity between men and women in higher education, but more importantly, this disparity remains extremely stable (55% regardless of the year, with a difference of only 1%).

4.3 Analysis of sector of student per year

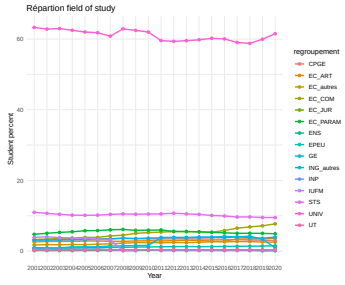
I have followed the exact same process for the sector, comparing the percentage of people who work in the public sector versus those in the private sector.



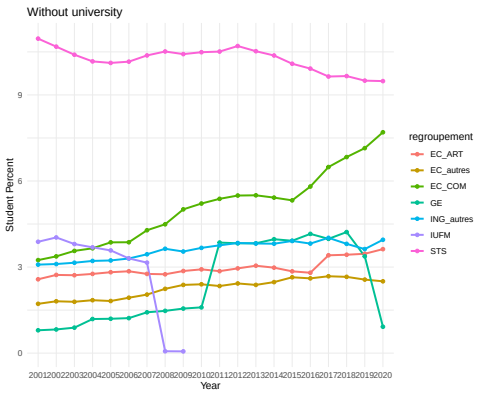
The graph is quite significant and shows that more and more people are pursuing higher education in the private sector, although the public sector (through universities) remains overwhelmingly dominant (86% in 2001, 80% in 2020). This represents a decrease of approximately 7% over the 19-year period.

4.4 Distribution by field of study

We will draw the different curves that represent the percentage of people in the different fields.



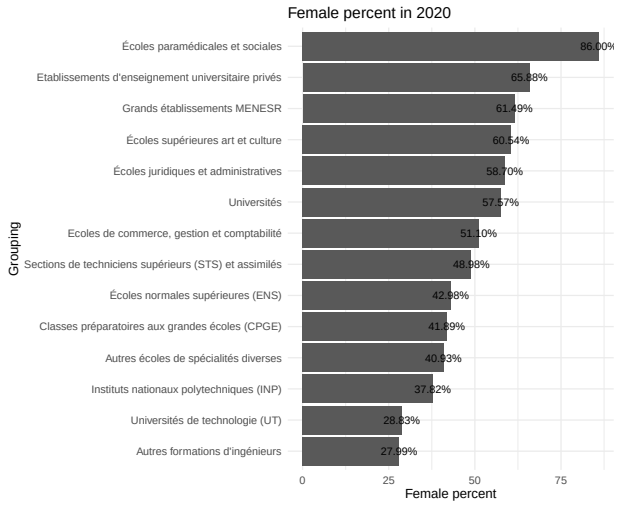
As you can see, the university dominates all the other formation, that's why I decided to remove it as well as some formation that are not very interesting, here is what you get:



Here these different curves are extremely interesting we can notice several things: The first one is a progressive and continuous decrease of the 2nd biggest training (Sections of Superior Technicians). The second one is the continuous and progressive rise of the business schools (passed from 3 to 7.5 %). The last is An explosive rise of the percentage of people in the "grandes écoles"in 2010, followed by a sharp decline in the last three years.

4.5 Gender distribution by field of study

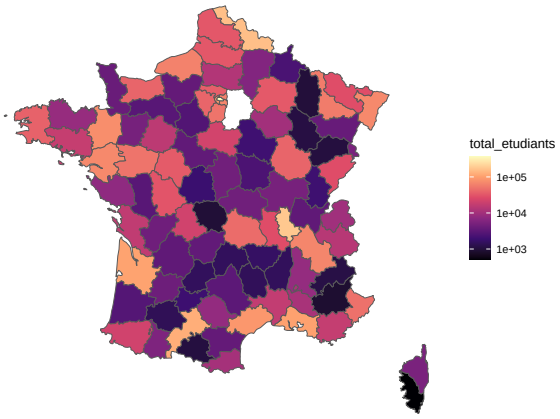
Now, we are going to look at the different gender representation in the various fields.



We can observe two things: women are in the majority in fields such as art, law, and healthcare (86% is the biggest imbalance across all fields), but they are a minority in science fields, which is even more surprising considering that they represent 55% of higher education students.

4.6 Distribution by department

To conclude, I made a distribution of students according to the department of their training using a map. Here it is:



This map shows us several things, the first is that there is a very high concentration of students in Ile de France (it's logical because Ile de France has more than 10,000,000 inhabitants), the second is that we can notice that the French coastline are generally above the average unlike the center of France and Corsica which has few students.

5 Conclusion

In conclusion, we note that France should have more and more students over the years, with a significant gender disparity in certain fields of study (men are in the majority in science, women are more present in law and medicine). We also observe that women are in the majority in higher education, that there is a larger share of people who are trained in private institutions, a decrease in the number of people in STS and an increase in the number of people in business schools and finally, a somewhat uneven geographical distribution of students (strong concentration in the ile de france and on the coastline, weaker in the center of France and in Corsica).