INSA TOULOUSE

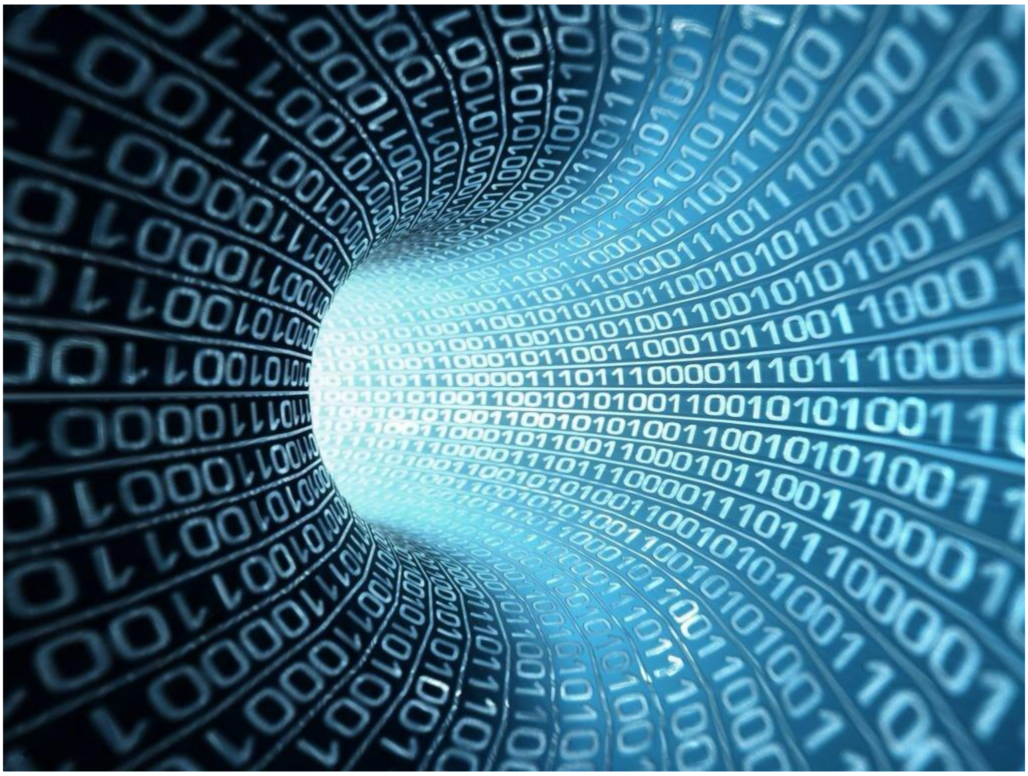
5A - ISS

January 2021

Report:

Treatment and analysis of data

By **Léo POTIERS** and **Paul SERONIE-VIVIEN**



INNOVATIVE SMART SYSTEMS

Introduction

Today, data is becoming the most valuable resource in the world, companies all over the world try to collect as much data as it is legally possible in order to improve their revenue. We can note that the most important companies that are Google, Facebook or Amazon are also the ones that probably hold the most important data centers. We also can cite the Cambridge Analytica scandal after the 2016 USA's national election or the development of artificial intelligence to point out even more the value and influence of big data.

In this course called treatment and analysis of data, we had to select a dataset and analyse it with R-Studio. We chose a free access dataset about every international soccer game in the world, since the very first game on the 30th of November of 1872 to 2020. We found interesting to look more in detail into this dataset to see the evolution of this World gathering sport over the years, and maybe to see the impact of the COVID-19 crisis on this sport.

In a first part, we will introduce the dataset and the information that it offers. Then, we will analyse it using R-Studio and expose some interesting graphics we were able to produce.

Choice and presentation of the dataset

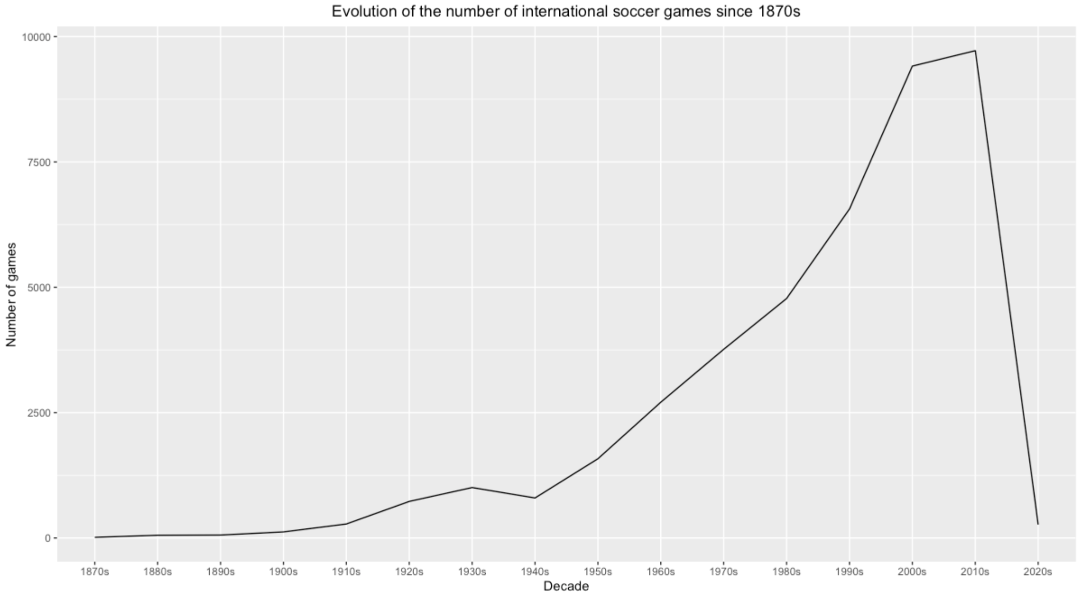
The first step of this project was picking a dataset that would ally three things. First, it needed to be easily available on the web, so we chose to do our research on the site Kaggle which contains a large variety of datasets in open source. The second criterion was to have an as recently updated as possible dataset since we wanted to treat data that had a good continuity until today. Our last will was to talk about a subject that would please the both of us. That is why we decided to go for this dataset that is referencing all the international matches that occurs between 1870 and 2020.

This dataset contains more than 40,000 matches since the very beginning of football. All the matches or defined by nine different parameters that was part at first of the dataset. First, we have the date of the match, then the two opposite teams, the home team, and the away team. After that, the scores for each of those teams. Another parameter is the tournament the match was played in when it is the case, otherwise the match is qualified as "Friendly". There is also the city and the country where the match took place. Finally, there is a Boolean saying whether the match was played on a neutral location. This parameter was key to easily recognize if a team was playing home or not. You will see later in the report that this a determinant factor in certain of our studies. We decided to add a parameter that is the year when the match was played since it was easier to treat and to analyze compared to the whole date.

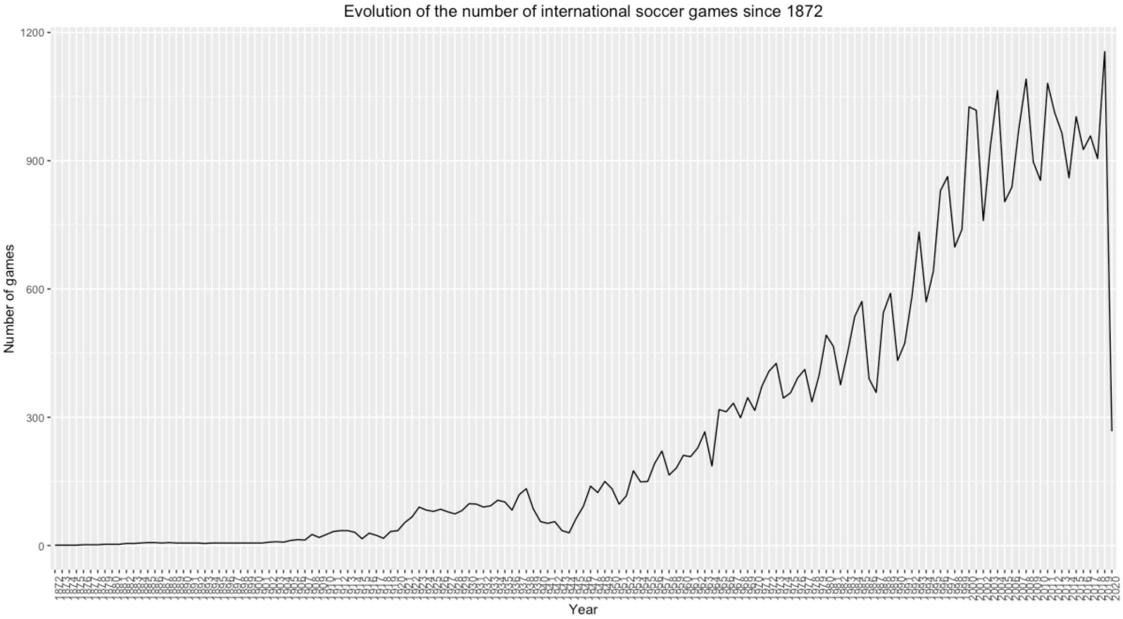
Exploitation of the dataset

1. Number of international soccer games over the time

First of all, we wanted to represent the development of international soccer by displaying the evolution of the number of games over decades:



We see that the number of games increased significantly after the World War II and has never stopped since. The final drop is only because we have just been living 1 year since the beginning of the 2020 decade. In order to display the impact of the actual sanitary crisis, we displayed the same number of games, but over the years:

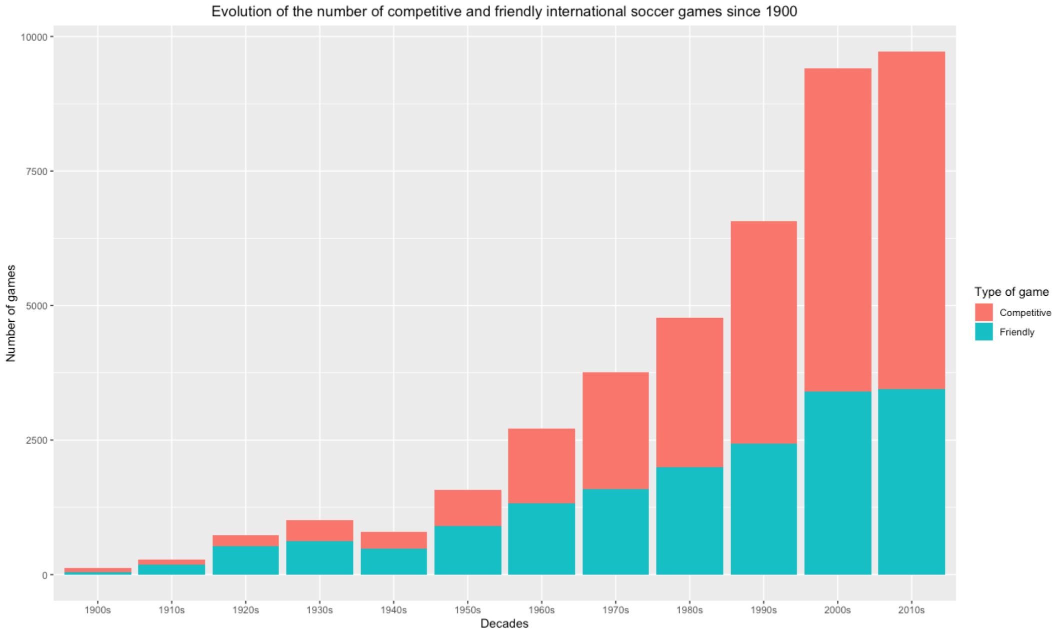


Here, we want to draw your attention on two characteristics of that curve. First, the considerable gap between every year, and 2020. This huge fall represents how historical is the situation we are living now. Secondly, since 1970 we see that the curve started to oscillate. These pics that happen every FIFA World Cup of Euro year, points out the influence of these events in this sport.

<u>Remark:</u> The R-scripts of these plots are joined to this report, "*Script 1.R*" will display the first graphic and "*Script 1.1.R*" the second.

2. Evolution of the number of competitive games over friendly games

After having showed that the international soccer developed a lot after the 1940s, we wanted to check if a similar statement was observable on the ratio of friendly or competitive game.



It is clear the increasing importance of competitive games over friendly games since the 1950s. We also notice the slight number of competitive games during the 1940s because of the cancelation of the 1942's FIFA World Cup due to the war.

Script: "*Script 2.R*"

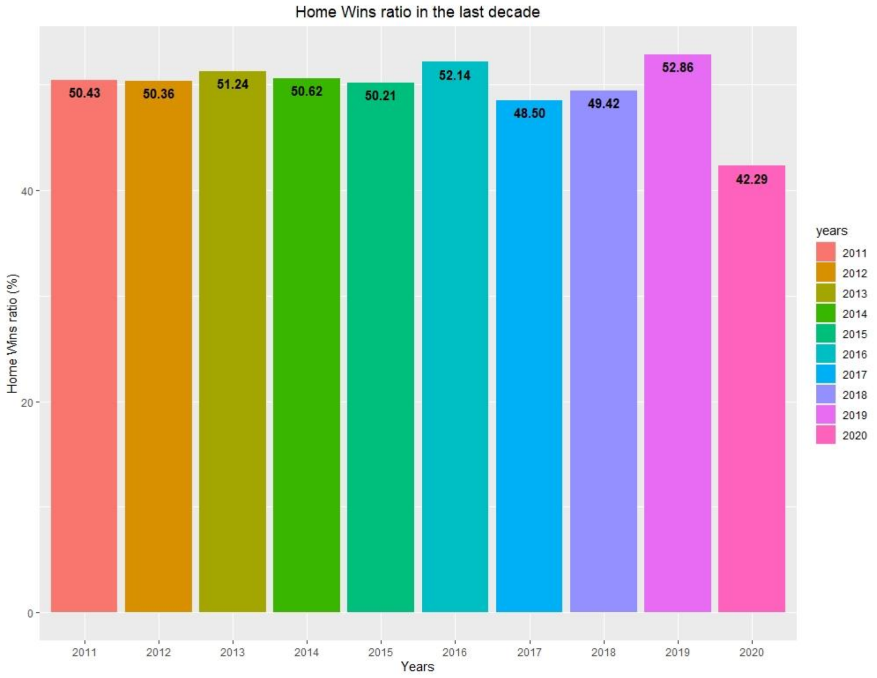
3. Home wins ratio over decades

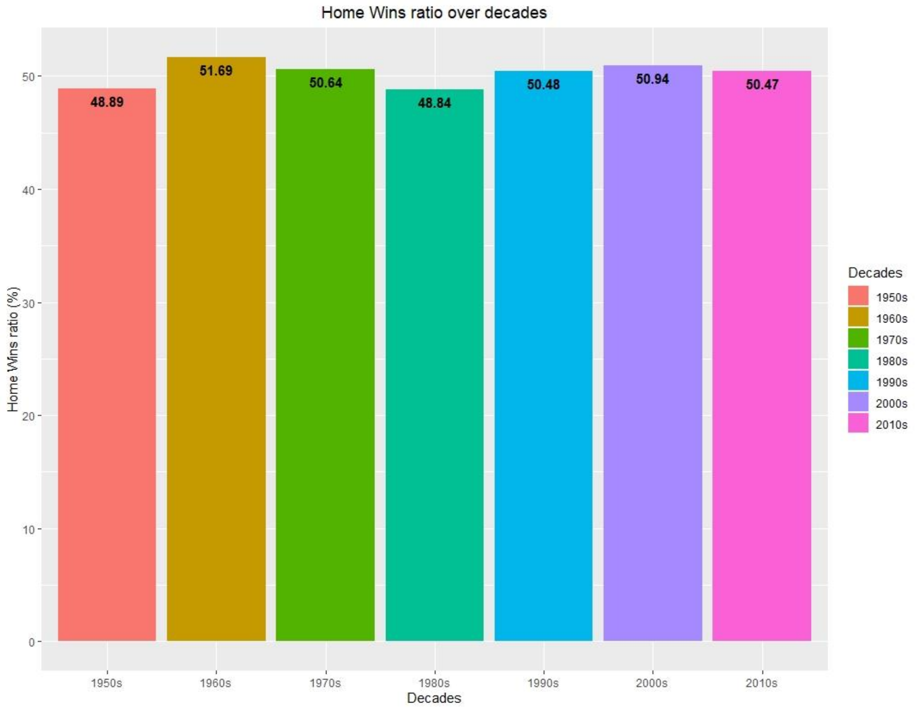
For this part, the topic is interesting since it is supposed to answer a key question for football fans: do they make a difference when it comes to the outcome of a game? To demonstrate that there might be a difference, the first thing was to exclude the matches that was played on a neutral field. Then, we decided to only treat the matches that were played after 1950 and the incredibly special period of the world wars. It appeared to us that it could be interesting if the fans had more impact at a certain period in this sport's history. Now, let us observe the evolution of the home wins ratio.

This diagram clearly shows that playing at home make a big difference. What is key in the analysis here is that there is a certain proportion of matches that end up in draws, this is not a game or lose situation. Another interesting point is that there is not a decade that is clearly above or below the rest, the ratio stayed stable for over 60 years now with a small peak during the 60s but always around 50%.

Script: "*Script 3.R*"

4. Home wins ratio and global pandemic





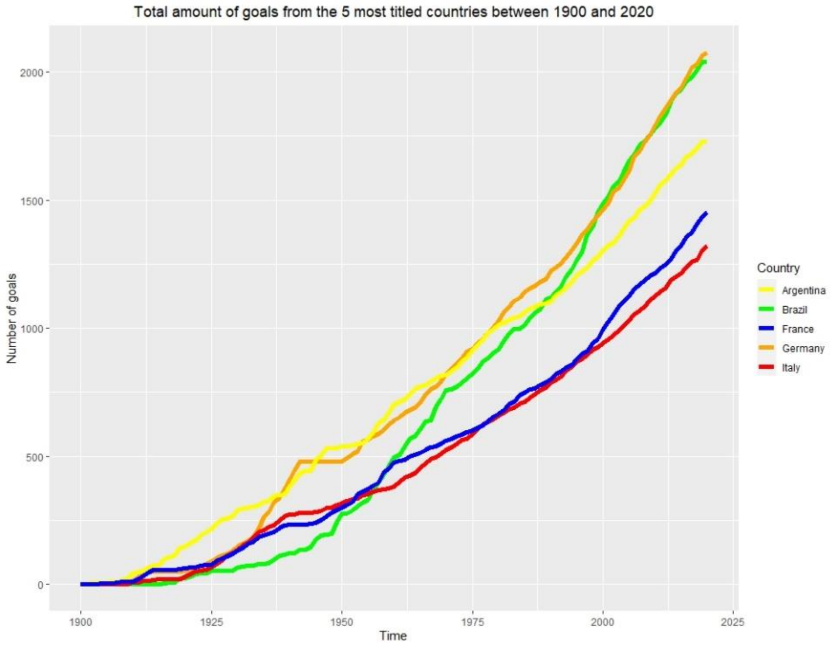
After treating the post war history of this home wins ratio, we wanted to focus on the past year, 2020, that has been special for all football fans around the world. The sanitary situation implied a higher number of matches played without any public and it seemed key to us to determine if the presence of a public is what make a difference more than playing in its own country, without long travels and jet lags. The same sorting of the matches was made but this time we just looked at the past decade year by year.

This result was beyond the expectations we had at first. There is a significant drop of the home wins ratio by almost 8% in the year 2020. What is even more alerting is that this ratio was stable during the whole decade, as suggested by the study we made on its history since 1950. From our point of view, it is unlikely that this drop happened at random and we are convinced that it is related to the lack of public in stadiums, reducing the difference between playing at home or not.

Script: “*Script 4.R*”

5. Goals from the top 5 most titled country between 1900 and 2020

In this section, we decided to put in concurrence the five biggest football nations regarding the titled they earned during their history. The chosen nations were Brazil, Germany, Italy, Argentina, and France. We needed a criterion to evaluate these teams and we opted for the goals scored by each team during the chosen period, from 1900 to 2020. This parameter is not perfect but the reason why people love this sport is mostly because of the excitement it brings whenever they see a beautiful goal scored at the last minute. This evolution is shown in the graphic below.



In the first years of the 20th century, Argentina was the first nation in our race probably because it was not as much involved in the wars as the European countries. They rapidly got caught up by Germany that took the lead just before world war two when Argentina took back its first place before losing for good after the war ended. Then came the golden age of Brazil, that was being pretty discrete so far, symbolized by it overtaking both France and Italy. After that, Brazil took Argentina's second place and is still fighting today with Germany. France and Italy went at a similar race during a long period but since the beginning of the 21st century, the gap between the two nations is growing, in favour of France.

Script: "*Script 5.R*"

Conclusion

The study of such a popular sport like football was very interesting because it transcribes every historically significant fact. Moreover, by doing this analysis we developed our ability to use R-Studio and to write R scripts.