Big Data processing

Elie Taillardat & Jonathan Malique - **INSA Toulouse**

PTP ISS – 21/12/18S

INTRODUCTION

During our last year studying at INSA Toulouse, we were working on data processing and more especially on Big Data processing. For this, we used the **R programming language**. We had to choose a dataset and an associated problem in order to make a small analysis and illustrate the lectures we had in class.

We personally choose the FIFA 2018 dataset with its **17,981** football players. Our analysis is going to deal with **players overall performance rating according to their different countries and characteristics**.

1. FIFA FOOTBALLERS REPARTITION AROUND THE WORLD

First, we wanted to know how the FIFA football players were dispatched **around the world**. To do that, we chose to draw two charts, one representing the repartition of players by continents, and then a map showing in which country there are the most number players **[Figure 1]**.

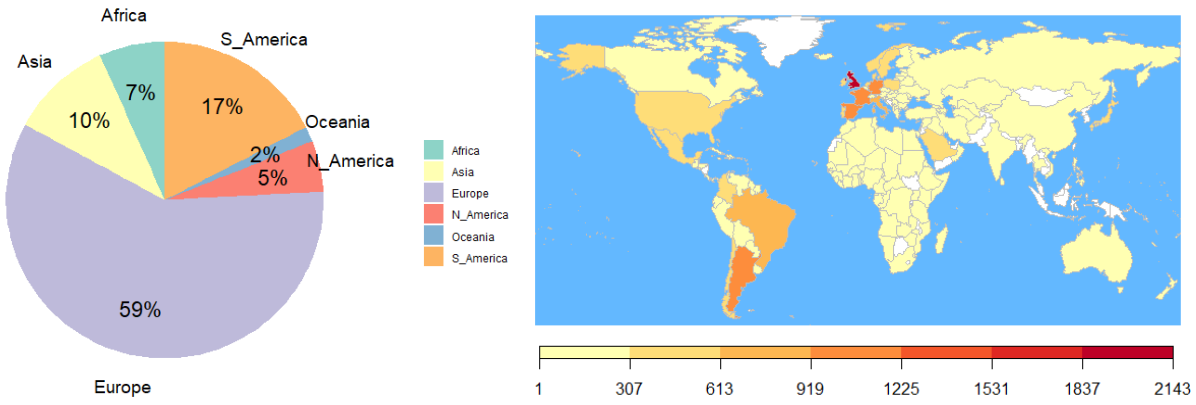

Figure 1: Player repartition in the world by country and continents

As the graphs show, we can see that the large part of the players come from **Europe** (around 60%) which seem to be normal because the main best leagues are in this continent. The map reveals that **England** is the country with the larger number of players, Premier League being one of the best (or the best) football league.

We tried next to observe from which country the top footballers are coming, based on the overall performance rating. The following chart represents the number of players by country, with the **Top 50** best players **[Figure 2]**.

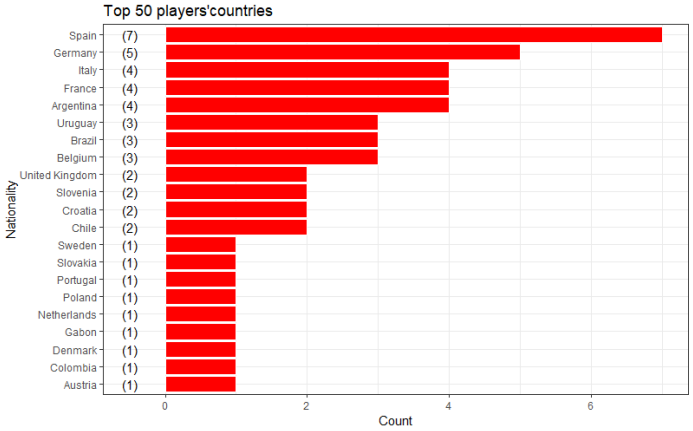

Figure 2: Best players nationalities

We can clearly notice that **Spain** is ahead with **7 top players**, and the four first countries are part of the top 5 European leagues. Nevertheless, the presence of Argentina, Uruguay and Brazil emphasizes here the fact that top players are also coming from **South America**, where the football is nearly a **religion**. The next section will present how three key parameters – Age, Height, Weight – are correlated with the player overall performance rating in the game.

2. Correlation of Age, Height & Weight with Performance Rating

The first thing was to analyse the influence of the player age with its performance. As we can observe below **[Figure 3]**, most of the football players have a better rating (noted on 100) when they are **between 25 and 30 years old**.

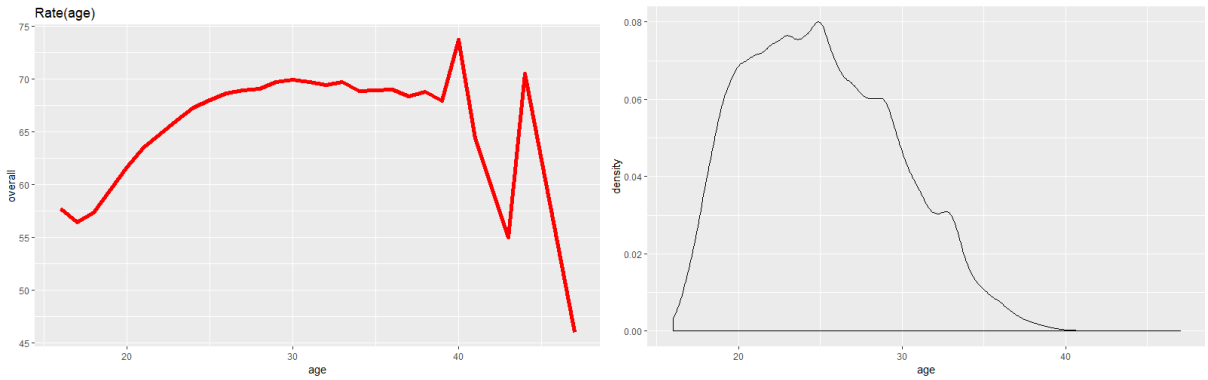

Figure 3: Age influence on overall rating and density of players

The graph on the left presents two pics around 40 years old, which is not so relevant due to the **poor** number of data there, and the fact that it is often players who were the best ones some years ago.

On the second graph of this figure, the curve simply shows that the largest population of FIFA football players are **between 23 and 26 years old**, with an important fall after these ages, representative of high level athletes.

We focused now on the **weight** parameter. As we can see on the left part of **[Figure 4]**, players with a more important weight have a higher overall performance rating than the others. This information is nevertheless a bit wrong when weights around 85kg are reached.

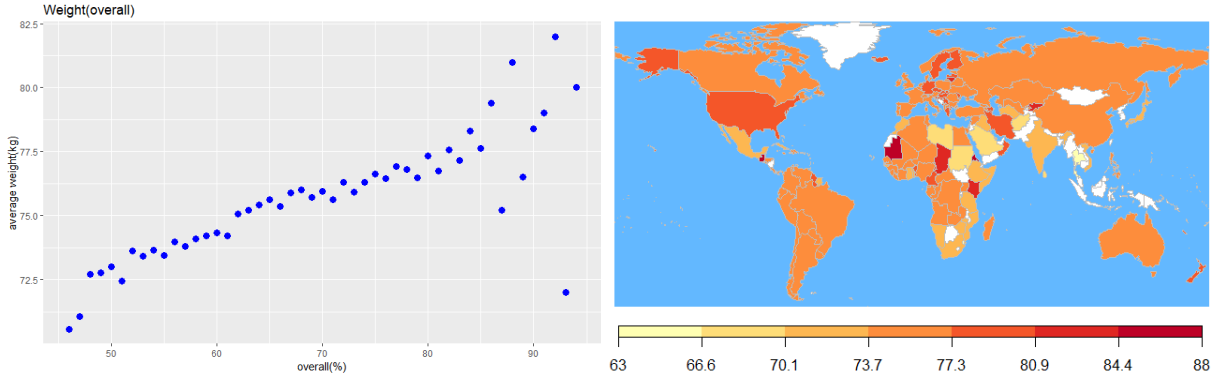

Figure 4: Weight influence on overall rating and repartition by country

Indeed, we can see at these points that **we lose the linearity** of the graph. It is due to the difference between players, such as Zlatan Ibrahimović who is mighty and so heavier but has great performances on the pitch. Moreover, to analyse in more details this weight parameter, we draw a map with players weight around the world by country. Players coming from some countries in **Africa** and from the **United States** are heavier, with weight around the previous problem case, that can explain the weakest performance in technique of their football players.

We did the same thing with the **height** parameter **[Figure 5]**. The first graph has quite the same shape, except the fact that this dispersion at the end is not so important, meaning that be tall for a player is quite a good thing, in a reasonable limit.

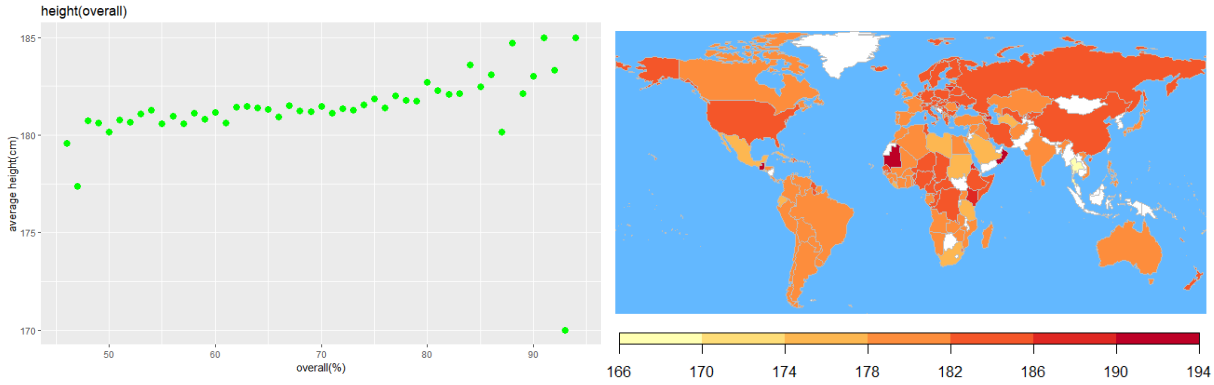

Figure 5: Height influence on overall rating and repartition by country

The map we draw again on the right side shows us player height repartition around the world, and we can recognize clearly same countries as earlier, which seems to represent football leagues with poorer technique level as we may guess, with more physical attributes.

3. Player Nationalities in the Top 5 European Leagues

Finally, we decided to analyse the different player nationalities in the Top 5 European leagues: Ligue 1 (France), Premier League (England), Bundesliga (Germany), Liga (Spain) and Serie A (Italy).

We chose this final analysis because it is a very interesting topic to deal with. As we can see in the graph below **[Figure 6]**, each league has a majority of players coming from this country, but there are a **large and important diversity** of other nationalities.

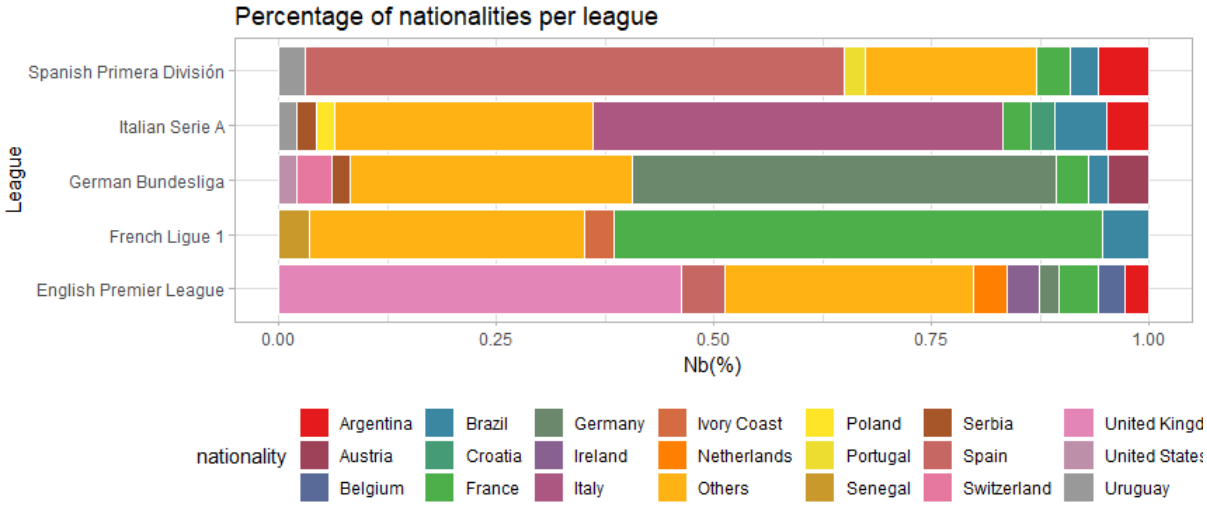

Figure 6: Player nationalities in the Top 5 European leagues

We can observe the effective presence of **South America** (Argentina, Brazil, Uruguay) players in these leagues, joining our previous observations. We can also distinguish players coming from **Africa** countries.

CONCLUSION

To conclude this short report, we can conclude that some countries are really turned towards football and that talents have more chance to emerge from these ones, with in particular the **high quality of European first leagues**. It was interesting to analyse this dataset with a world provenance vision.

However, our analysis was nevertheless **restrictive** with not all parameters taken into accounts. For example, analyses about players **technique** or **speed** would have been another contribution for this brief study.

Finally, we would like to say that it was quite nice to experiment the **R language** in a new context for us, the **Big Data processing**, and use it for this particular dataset of FIFA18 which was funny to discover.

Files attached:

- **FootballCorrectDataset.csv**
- **DisplayFIFA.R**