Overall performance of top-1000 players (FIFA20 analysis)

Liana Poghosyan

20-07-2020

The problem/data description

- Why is this important/interesting
  - The performance of the players is very important in terms of business. The football is one of the most profitable spheres of sports. FIFA 20 is a football simulation video game published by Electronic Arts as part of the FIFA series. It is the 27th installment in the FIFA series.Thanks to the impressive visual design, new squads, great soundtrack and many updated features, the user will be able to reflect your talents in the most entertaining way possible. The clubs in the career mode have dynamic sleeve badges, which reflect their achievements during the game. When a player is promoted or relegated, these change, thanks to authentic career mode.The dataset provided includes the players data for the Career Mode from FIFA 20 .
- The problem statement:
  - The dataset is huge, we can have lots of assumptions, but we need to clean and make it possible to analyse. Here we can get different statistics about the football players.The problem is vital for businesses, so that the businesses can understand the connection between different variables, from the top1000 performing players and teams we can analyse the categroies important for football player

- Where does the data come from?
  - The data is from Kaggle(link)
- What was done on this data so far
  - The main advantage of the dataset is that it has no kernels. So, I am free to do any analysis I want.

Main hypotheses

Here you write what are you trying to find in the data, what are some hypotheses that you are trying to test The main reasons to use the dataset in this case are:

- The number of players in top 10 performing teams, also inlcuded in the dataset
- The average number age of players in top 20 performing teams
- Find the correlation between different numeric types(we can understand the relationship of some variables)
- To find out Relationship between some countries and their Attacking Work Rate
- Foot preferrance of player in some of top10 clubs
- The density function of hits in top-1000 best performing players.
- The analysis of dataset.

The dataset explanation

The datasets' variable explanation is as follows: **Name:** the name of the player **Country:** the country the player represents **Position:** the position the player plays **Age:** the age of the player **Overall:** overall rating of the player **Potential:** potential **Club:** the football club the player plays **Contract:** the date of contract is valid **Height: Weight: foot:** The preferred foot **Joined: Value: Wage: Release.Clause: Attacking:** values of crossing, finishing, head accuracy, short passing and volleys **Skill:** Dribbling, Curve, FK accuracy, long passing, Ball Control **Movement:** Acceleration, Sprint Speed, Agility, Reactions, Balance **Power:** Shot Power, Jumping, Stamina, Strength, Long Shots **Mentality:** Aggression, Interceptions, Positioning, Vision, Penalties, Composure **Aggression: Interceptions: Positioning: Vision: Penalties: Composure: Defending:** Defensive Awareness, Standing Tackle, Sliding Tackle **Goalkeeping:** GK Diving, GK Handling, GK Kicking, GK Positioning, GK Reflexes **W.F:** Weak Foot **SM:** Skill Moves **A.W:** Attacking Work Rate **D.W:** Defensive Work Rate **IR:** International Reputation **Hits:** overall number of hits

The plots

I read the csv file to start the analysis. In the dataset I treated empty cells as NAs, and then removed them from calculations. The dataset is a huge. Thus, I just subsetted the initial dataset to have the columns that I need in my analysis. I removed 40 columns from the initial dataset.

Now I have 19661 observations(rows) and 34 variables(columns).

```
## [1] 19661    34
```

I removed repetitions in the Names, which cause a problem during subsetting. So, in result FIFA20 dataframe has 1000 observations and 34 variables. The top 1000 best performing players are in the dataframe now.

```
## [1] 1000   34
```

The Figure 1 shows the number of members' from best performing Clubs included in the dataset. So, in top-10 from 1000 performing teams we have 13-17 memebers.

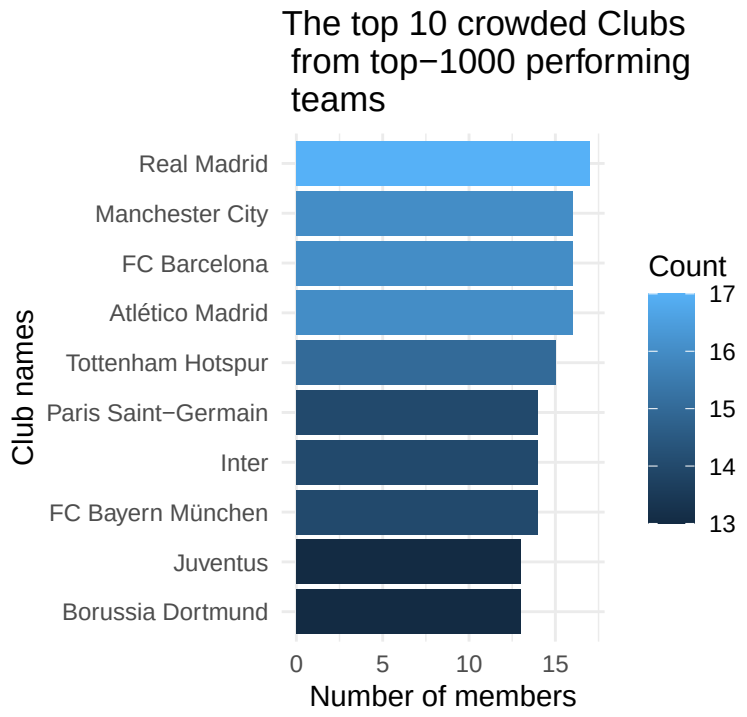


Figure 1

The Figure 2 shows that the average age of the top 20 teams out of top 1000 teams is in between 32-35.

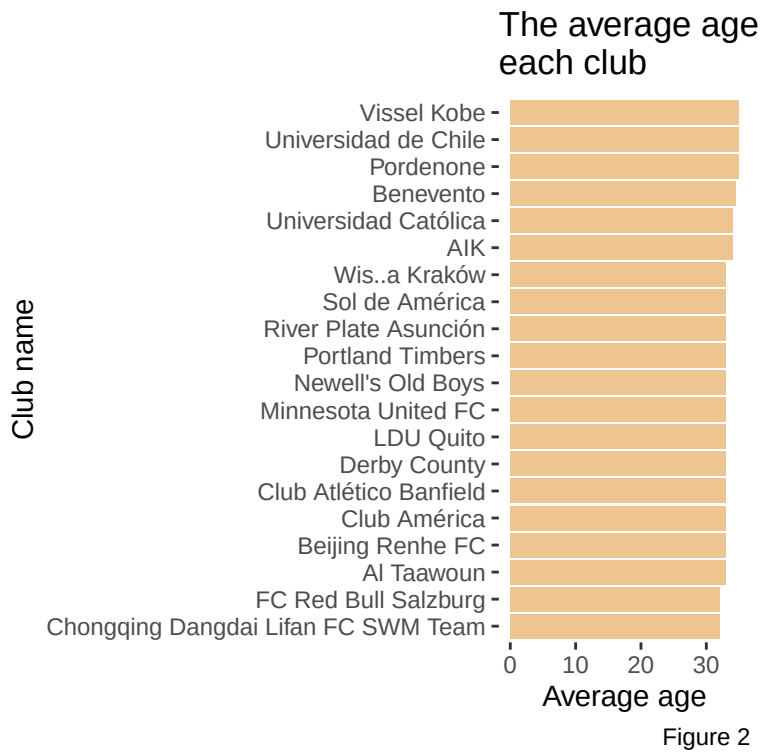
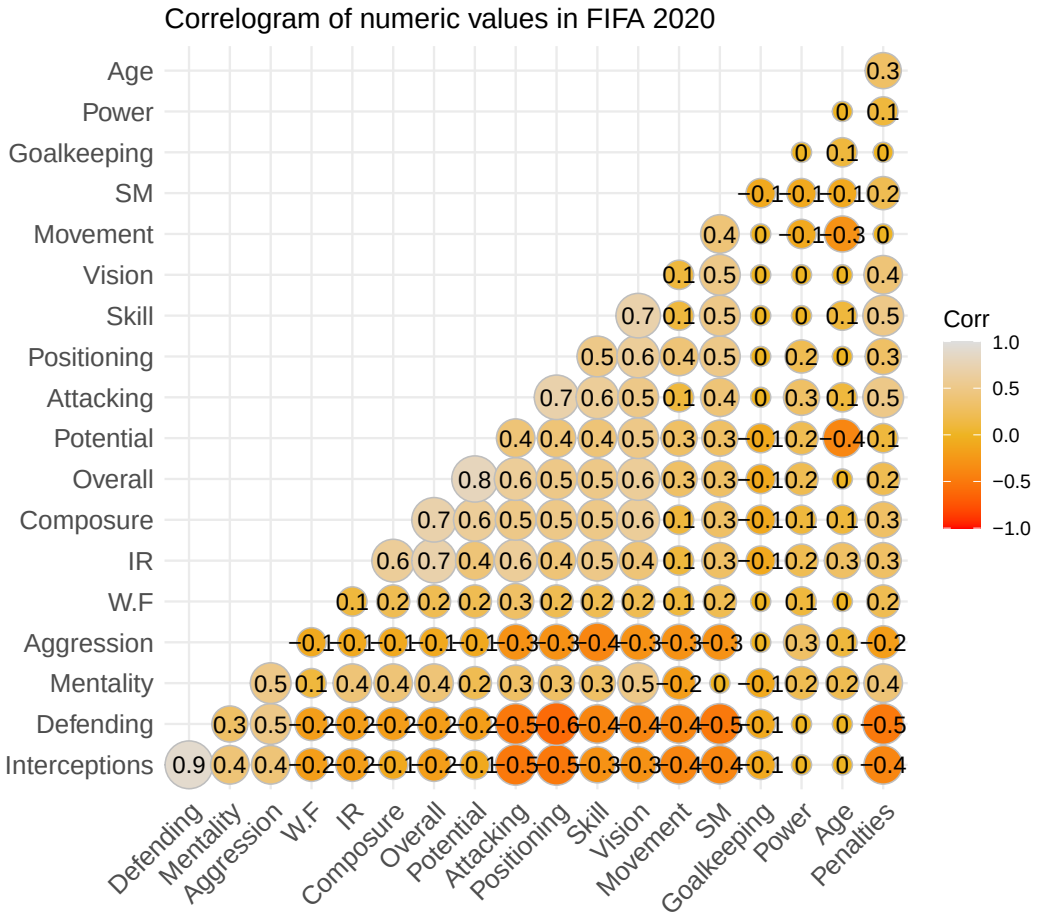


Figure 2

The Figure 3 shows the correlation between the numeric values in the dataframe. We can see that the

correlation between (Defending and Interseptions), (Potential and Overall Rating), (Positioning and Attacking), (Skill and Vision) are positive. The positive correlation shows that both variables change in the same direction.This is a high level of correlation. Also, I need to denote that Correlation is not Causation. So, in analysis we need to observe other factors as well.



In Figure 4 we have faceting. I factorized the Attacking Work Rate(the same as A.W) with levels of "Low","Medium","High". I tried to understand the number of people from "Brazil","Germany" and "Spain" having the A.W of each type. The plot shows that the Attacking Work Rate of "high" type is the highes in Spain. The "medium" type is the highest in Brazil and Spain(with equal amounts), and the "low" type is the highest in Spain.

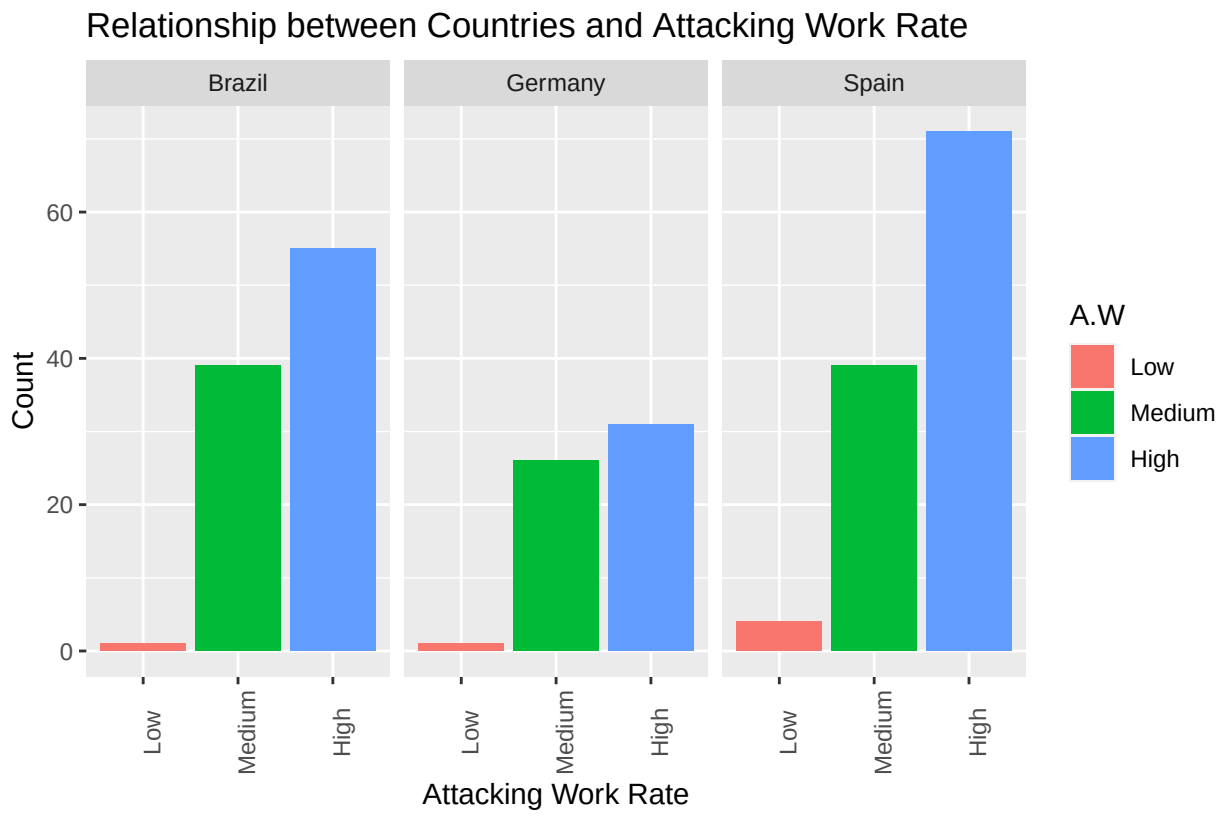


Figure 4

To make the Figure 5 we need to get some of the top 10 performing teams. Then I chose "Liverpool","Juventus","Real Madrid","FC Barcelona","Manchester City" from the Clubs and made a fecting with players preferred foot. From chosen teams the most Right foot players are in Real Madrid and the most Left foot preferring players are from Manchester City. Juventus has the least "Right" foot prefering footballers, and Liverpool has least Left foot players.

```
##  [1] "Guangzhou Evergrande Taobao FC" "Liverpool"
##  [3] "Dalian YiFang FC"               "FC Barcelona"
##  [5] "Juventus"                       "Cagliari"
##  [7] "Real Madrid"                    "FC Bayern München"
##  [9] "Paris Saint-Germain"            "Manchester City"
```

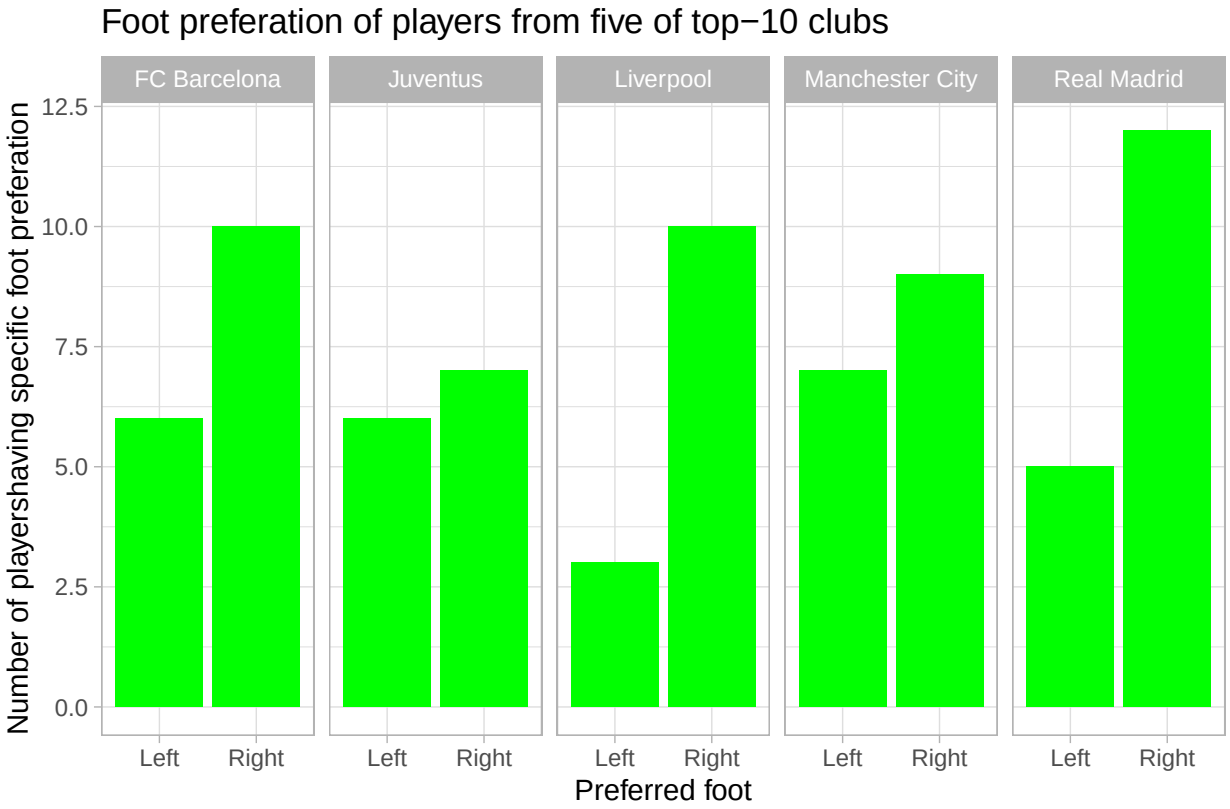


Figure 5

To have the Figure 6 I subsetted as follows:

Hits number	Naming
above 400	Very High
[300;400)	High
[200;300)	Medium
[100,200)	Low
[0,100)	Very Low

The Figure 6 shows the density of Hits in every Category. We see that the highest is in above 400 range. And the density decreases with the number of hits.

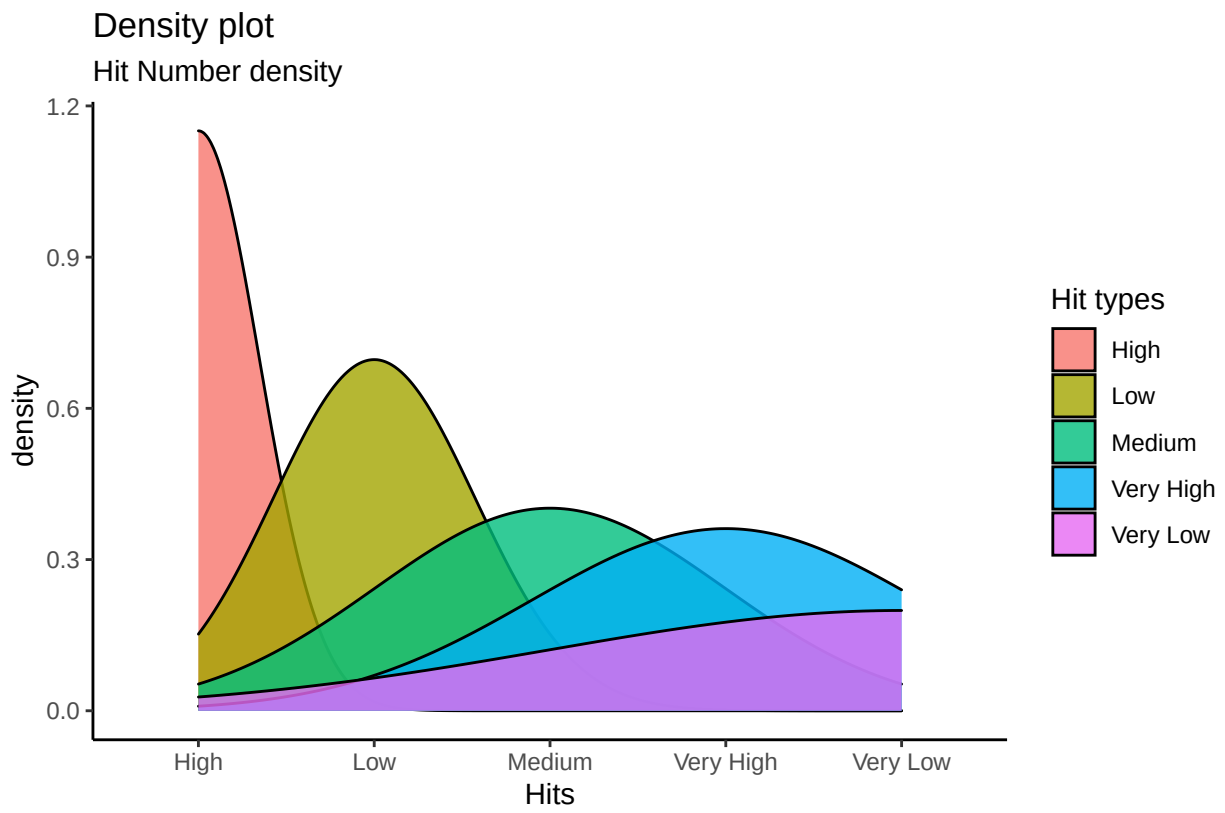
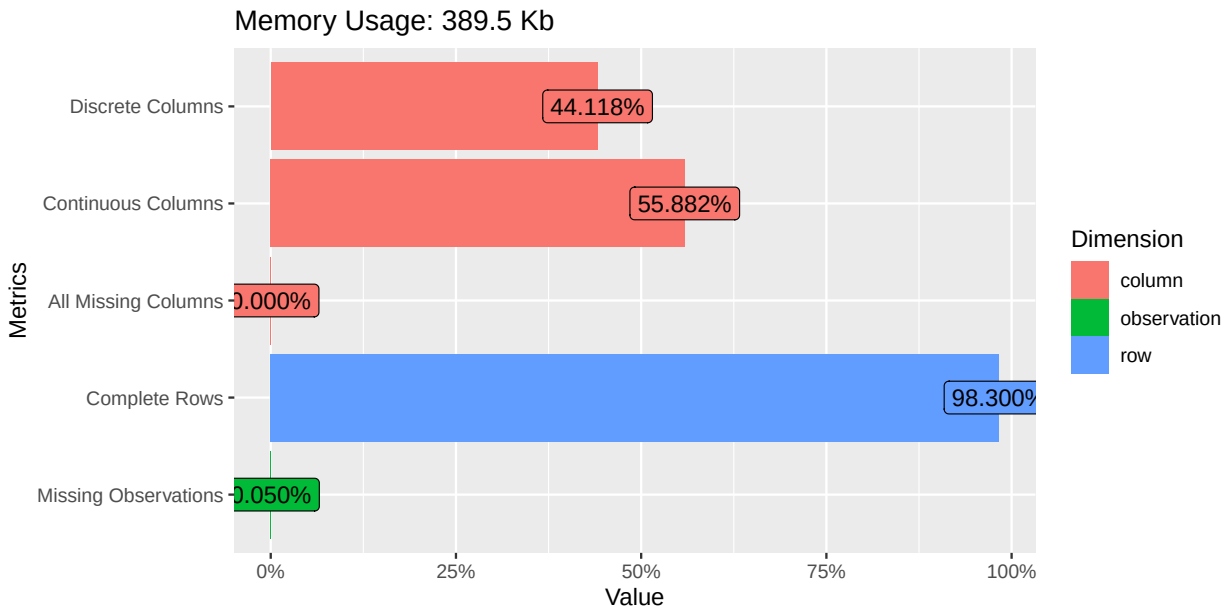


Figure 6

The table shows the number of every category of Hits and the number of their occurences in top-1000 best performing players.

```
## # A tibble: 5 x 2
##   Hits      Count
##   <chr>     <int>
## 1 Very High   369
## 2 Low         307
## 3 Medium      149
## 4 High        118
## 5 Very Low     57
```

Also, the plot describing my dataset.



Analysis

I observed that:

- On maximum the well-performing Clubs in have 13-17 memebers in FIFA20.
- The average age of the top 20 teams out of top 1000 teams is in between 32-35.
- The correlation between different numeric types is available in the Correlogram to show the relationship of different types.
- Players representing Spain are more attacking in comparison with Brazil and Germany. And Germany is more relaxed in terms of attacking in comparison with Brazil and Spain.
- From best performing clubs Manchester City has more Left foot preferring players, and Real Madrid has more Right foot prefering people.
- The Hit number in best performing teams in very high mostly.
- In my dataset I had mostl Complete Rows, Continuous Columns are more than Discrete Columns.

Summary of findings and recommendatins if any

To sum up, we can see that the best players are above 32. There is a positive correlation between Defending and Interseptions, Potential and Overall Rating, Positioning and Attacking, Skill and Vision. German people play with normal attacking rate. Another observation was that players prefer Right foot more, but it also has connection with demographic advantage of right-oriented and left-oriented people.