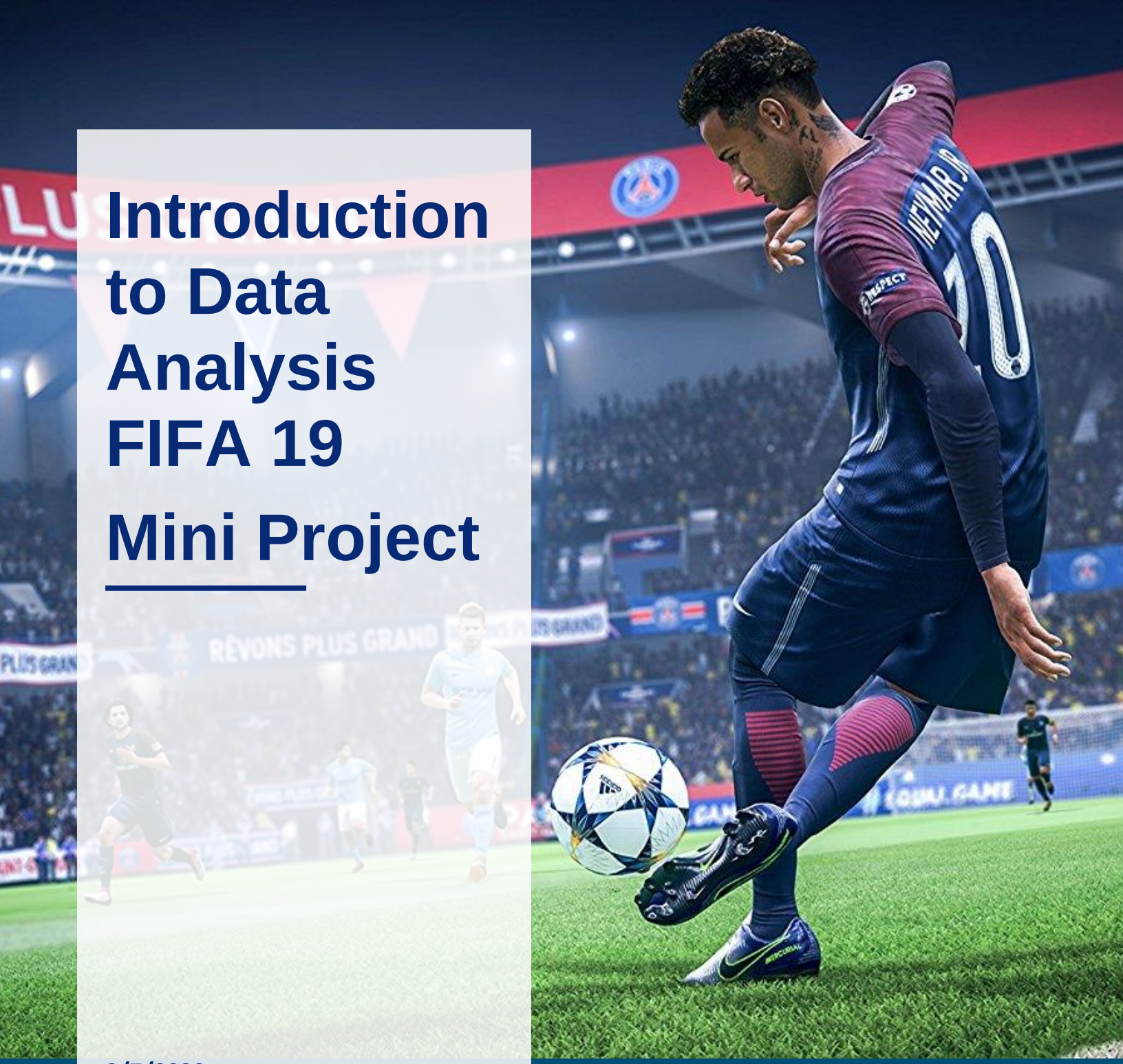


Introduction to Data Analysis FIFA 19

Mini Project

2/7/2020

Edo Cohen

Shai Ochayon

Technion – Israel Institute of Technology

Introduction to Data Analysis
FIFA 19 - Mini Project

SPRING 2020

QUESTIONS: What are the questions you wanted to explore? Why are they interesting to you?

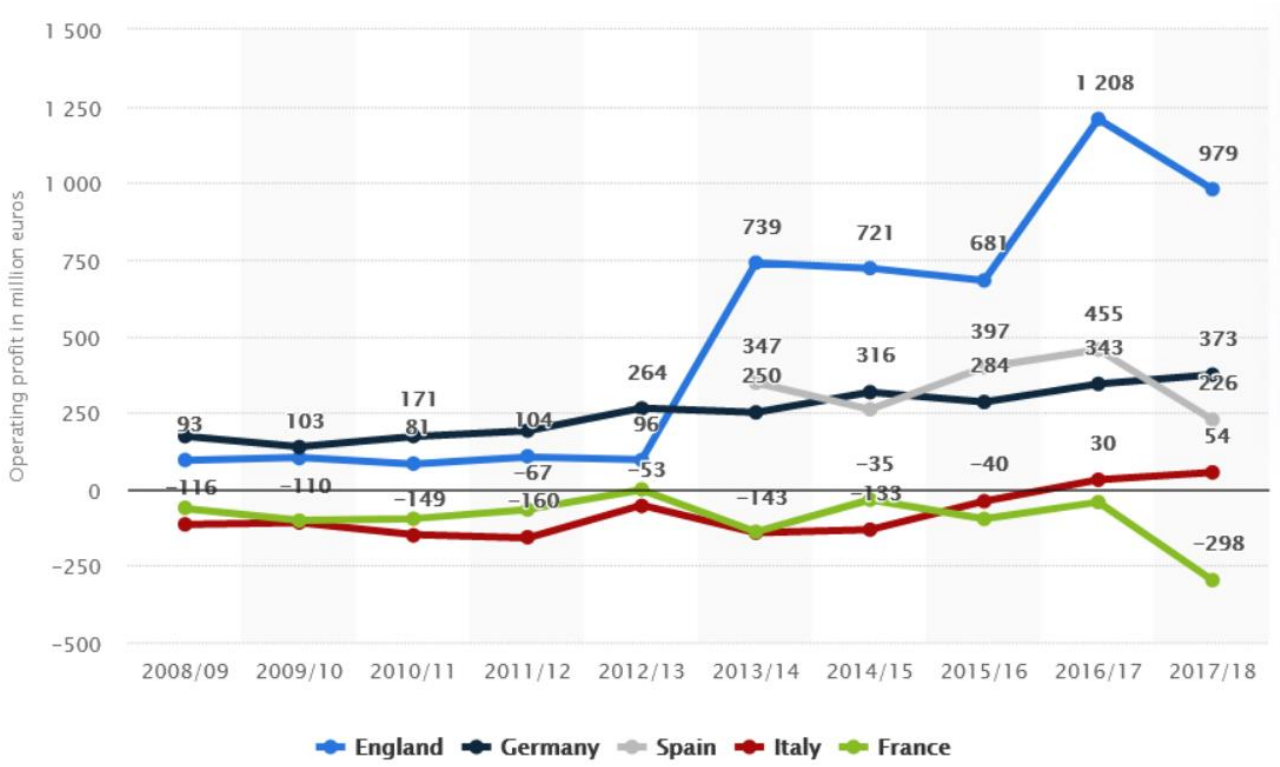
We chose to work with the FIFA 19 dataset. We chose this dataset since we are very passionate about both football and data and believe that our domain knowledge in this subject can add an extra value to our research. It was important to us to make something that can bring an actual value to football clubs and players. Hopefully, our work reflects that.

In our research we wanted to take a closer look at the financial side of football and the links between the players ability to their wage and value. We will especially look at the English Premier League (EPL) since it is the most profitable league out of Europe top leagues, according to the Deloitte Football Money League[1].

The "Deloitte Football Money League"[2] is a ranking of football clubs by revenue generated from football operations. By the report of 2019[3], 13 out of 30 European teams with the most revenue were English, with total revenue of three times more than the total revenue of the teams from Italy in the seconds place. This, to say that the smaller football clubs in England earn more than top teams in other league making the EPL very competitive. Moreover, this has been the trend in the last decade.

[1] **https://en.wikipedia.org/wiki/Premier_League#Finances**
[2] **https://en.wikipedia.org/wiki/Deloitte_Football_Money_League#2019**
[3] **https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/sports-business-group/deloitte-uk-deloitte-football-money-league-2019.pdf**

Operating profit of "Big five" European football league clubs from 2009 to 2018, by league[4]:



[4] https://www.statista.com/statistics/1022194/european-football-operating-profitr-profitability-by-league/

DATASET: Describe the dataset you use; Explain why it is appropriate for answering these questions.

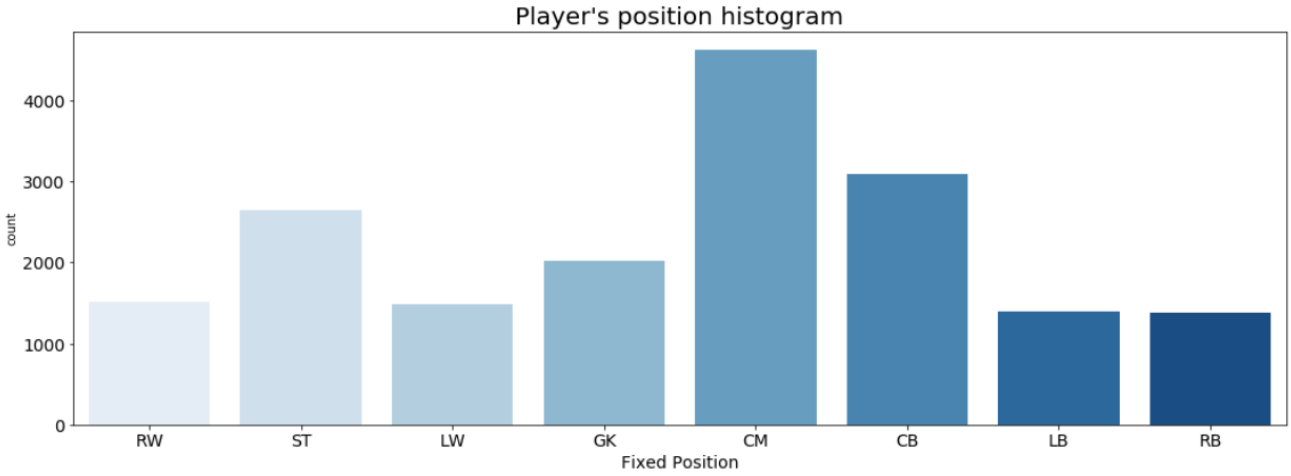
This dataset provides detailed attributes for every player registered in the latest edition of FIFA 19 database. Every record represents a real player which has numerical attributes which describes his ability, wage, market value, and other categorical attributes such as the player's club, position, preferred foot etc.

This data is appropriate for answering our questions since it provides detailed and accurate information regarding players ability, as well as value and wage which was critical for our work. As mentioned before, we tried to bring meaningful insights for the real world out of this data.

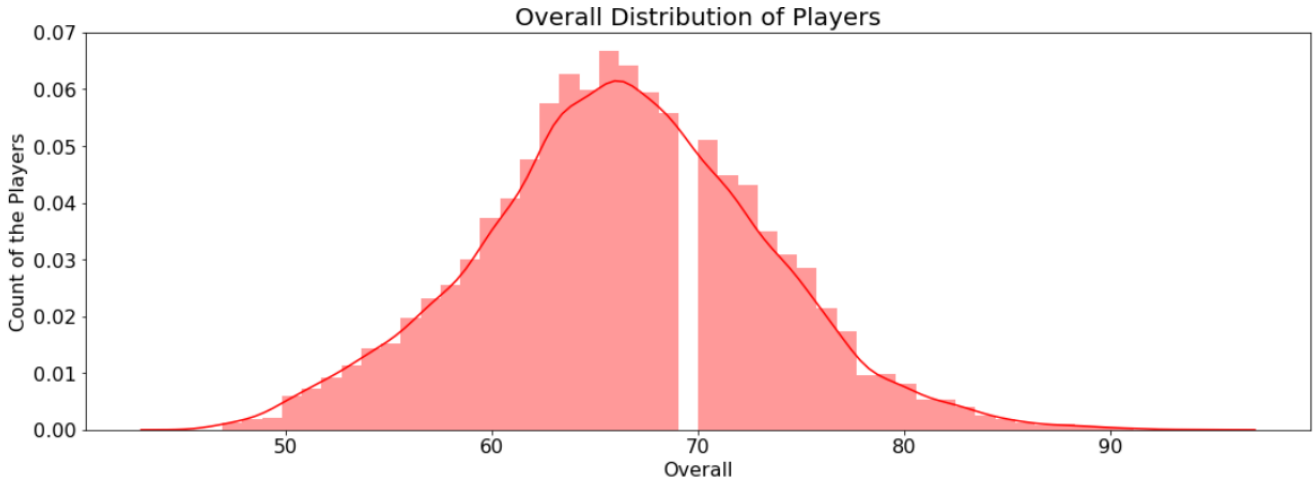
ANALYSIS & FINDINGS: What analyses did you conduct to answer your questions? What did you find?

EXPLORATORY DATA ANALYSIS

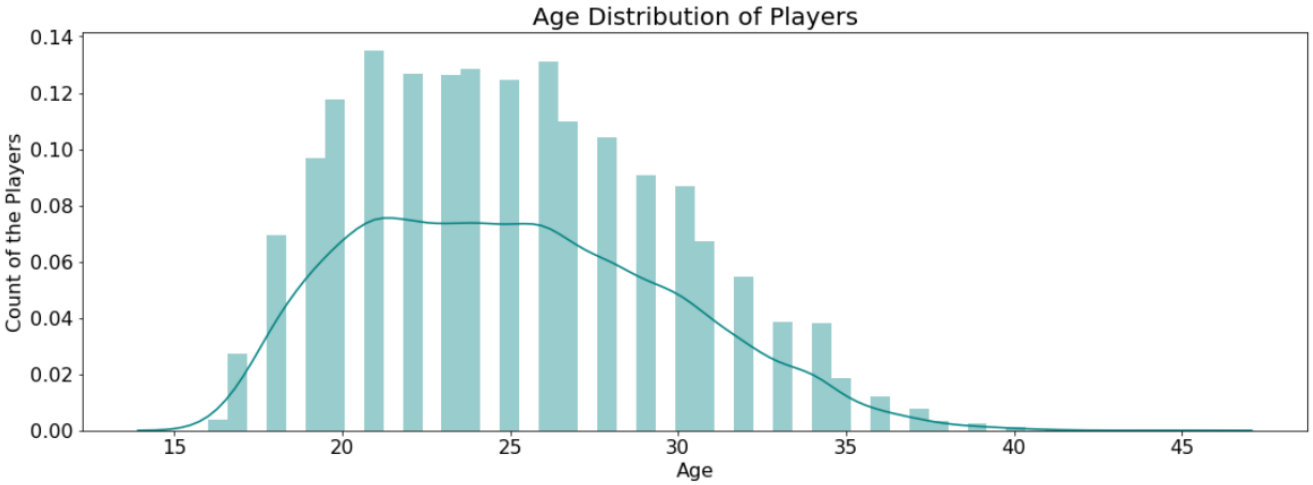
We will start off our research with some basic distributions and correlations to get a sense of the data before moving on to the heavier stuff:



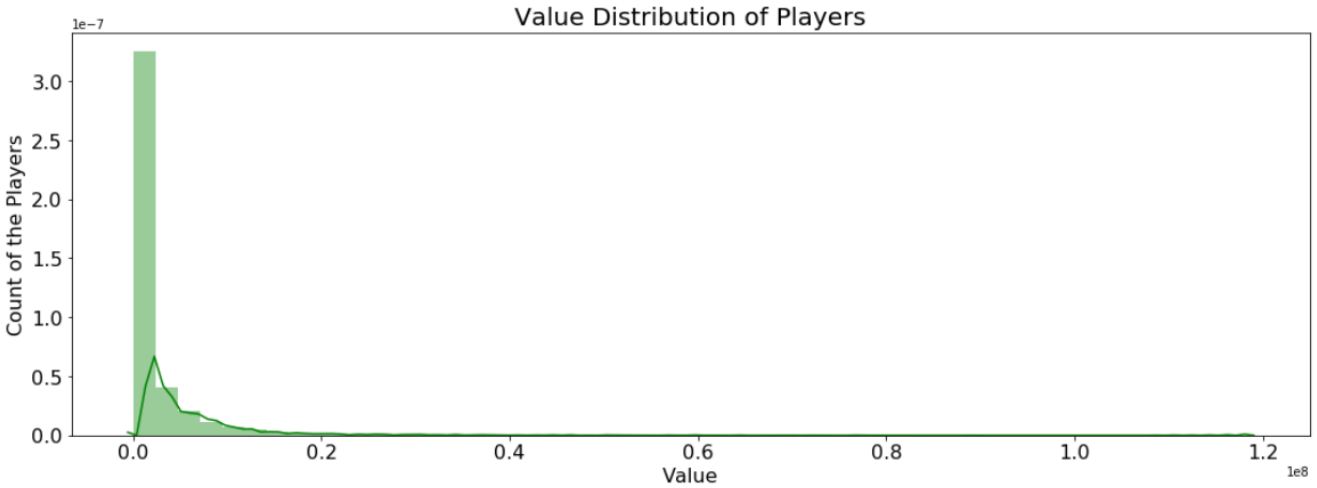
At first glance, we can see that the most common position is CM, we can assume it is because it is the most used position in most football formations.



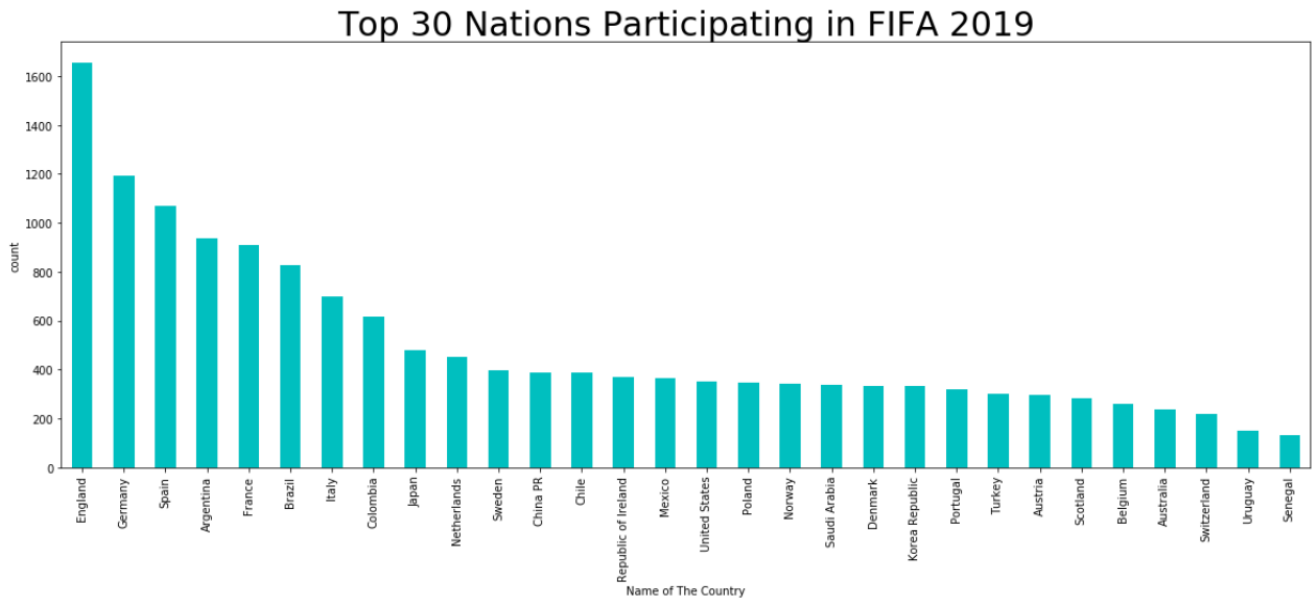
We can see that the overall rating is roughly distributed normally around 67.



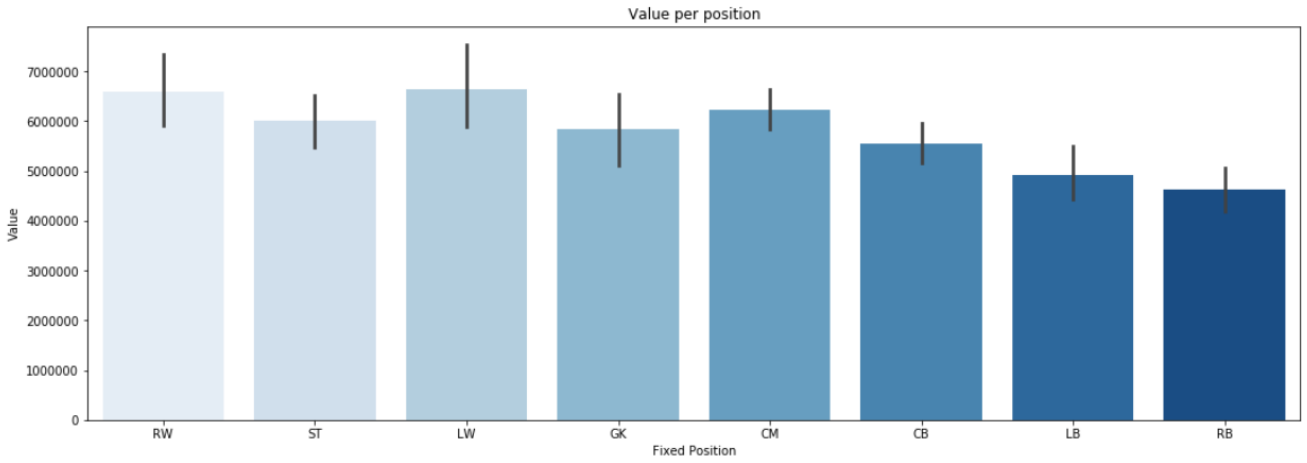
As we can expect most players are between 20 - 27 years old. The older the age the less common it is. Players over 40 are very rare.



Most of the players value at less than 10,000,000 Euros. We will dive into the value of the players deeper later, so we will keep that in mind.



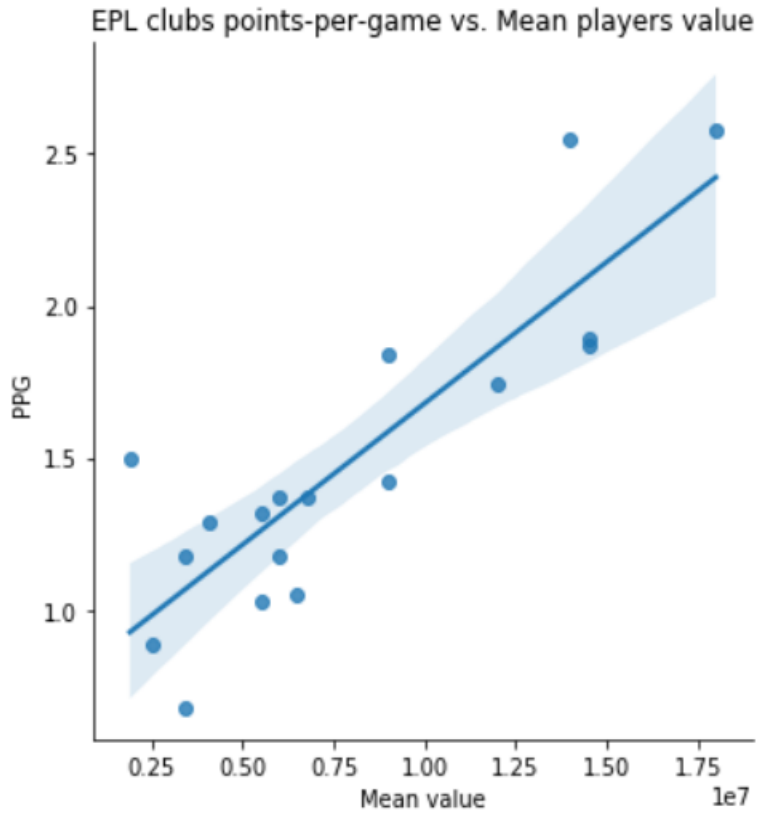
The most frequent nationality in the data is England, we assume this does not mean that there are more English footballers in general but rather that the English football is more represented in the game at the lower levels.



Full-back (RB/LB) are the players with the lowest value, while Wingers (RW/LW) are the players with the highest value. It is fair to assume that people love beautiful goals and are generally more attracted by the attacking side of football, and as we can see it is translated into the players' worth.

USING EXTERNAL DATA:

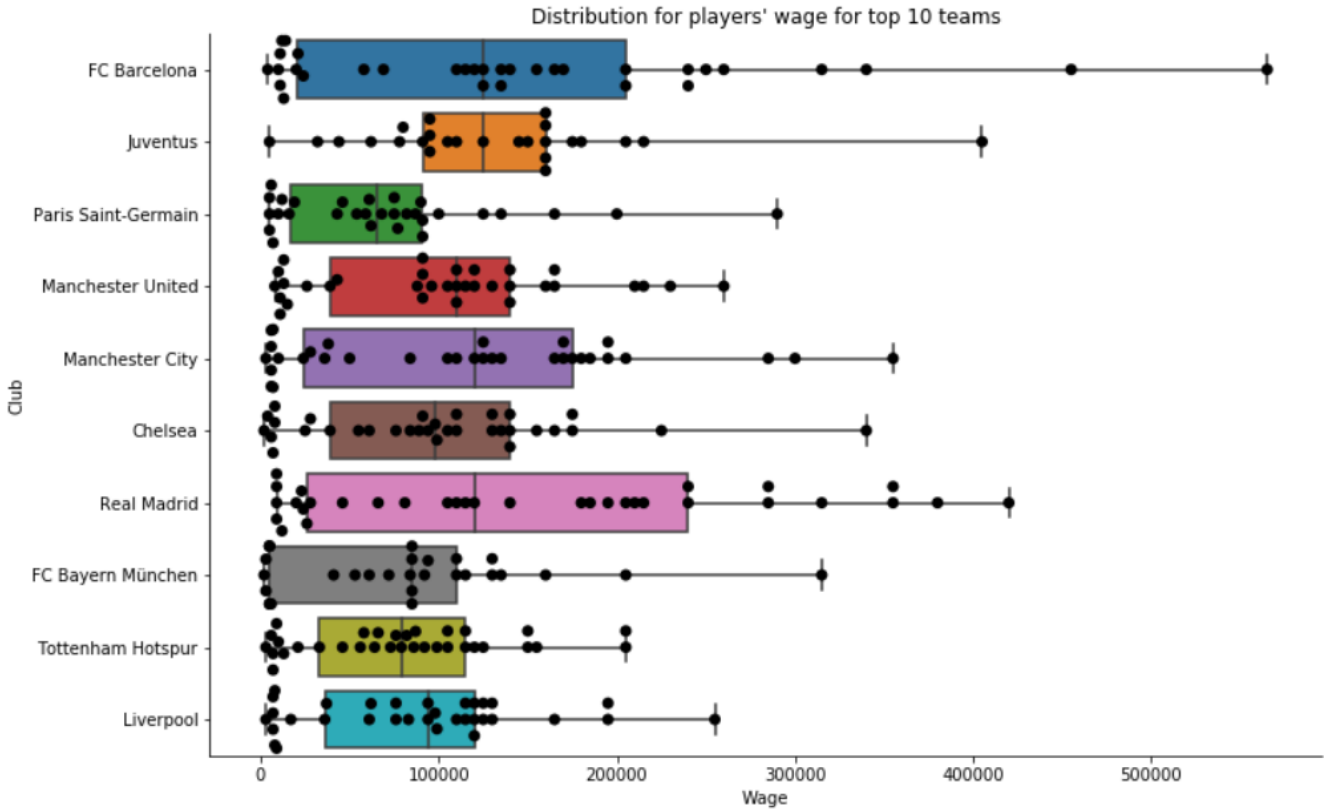
One question which was especially important to us is to know how much players' value is correlated with the team success. Unfortunately, our dataset does not provide the information required to answer this question. As a result, we had to import an external data which presents the real-world data regarding the EPL success for the 18/19 season. Out of this data we looked at the Point-Per-Game (PPG) for each club and the correlation to the mean value of its players.



Club	Value	PPG
Arsenal	9000000.0	1.84
Bournemouth	3400000.0	1.18
Burnley	6500000.0	1.05
Cardiff City	2500000.0	0.89
Chelsea	14500000.0	1.89
Crystal Palace	4050000.0	1.29
Everton	9000000.0	1.42
Fulham	3400000.0	0.68
Leicester City	6000000.0	1.37
Liverpool	14000000.0	2.55
Manchester City	18000000.0	2.58
Manchester United	12000000.0	1.74
Newcastle United	6000000.0	1.18
Southampton	5500000.0	1.03
Tottenham Hotspur	14500000.0	1.87
Watford	5500000.0	1.32
West Ham United	6750000.0	1.37
Wolverhampton Wanderers	1900000.0	1.50

We can see that in most cases the higher the mean value of the player is the more points the team won during the season. One might mistakenly think that spending more money is causing the team to be more successful. It is important to emphasize that correlation does NOT equal causation.

Next, we will look at the distribution of wages for Europe top 10 football clubs by revenue[5] (each point represents a player in the team):



1. We can see that both Barcelona and Real Madrid have many players who earn very high salaries, and thus also has a very big range between the highest and the lowest earners.

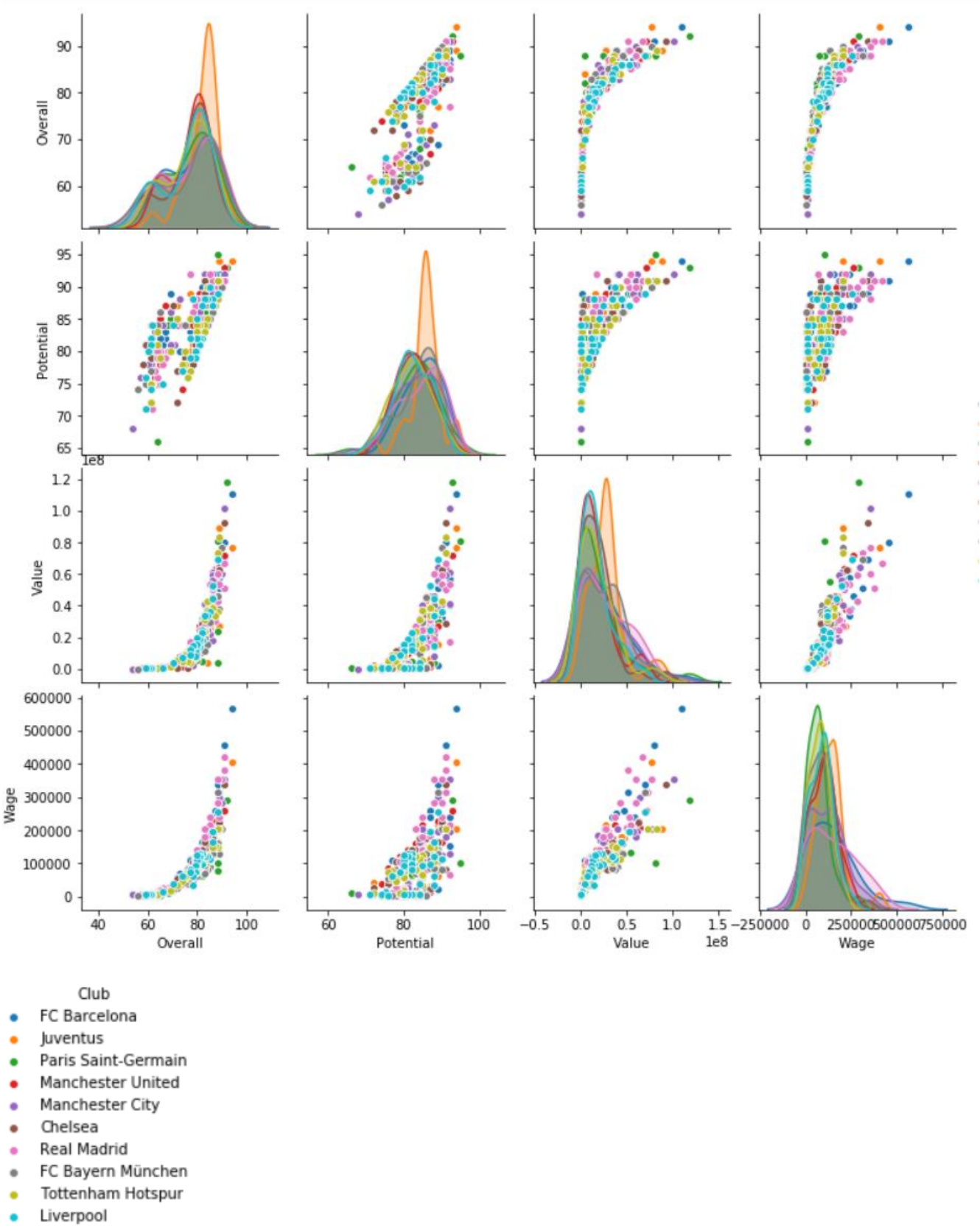
2. We can see that Barcelona, Real Madrid, Manchester City and Juventus all equal with the highest median.

3. Tottenham Hotspur has the most players clustered around the same wage.

4. The outlying point for Barcelona is Messi, which is the highest paid player by a big margin.

[5] **https://ftnnews.com/sports/38655-top-20-highest-revenue-generating-football-clubs**

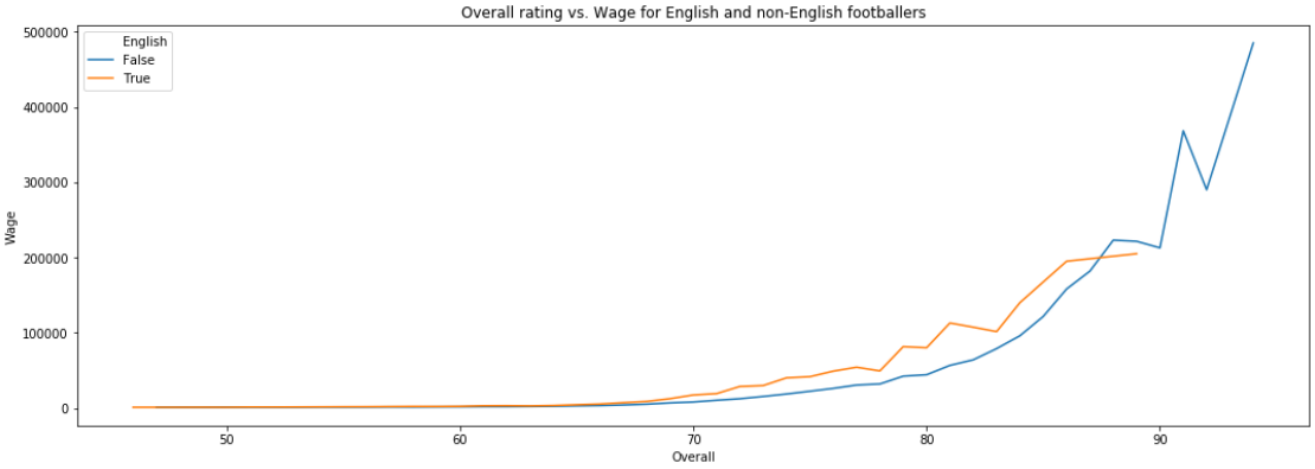
Other important correlations from the data:



Some more insights from the above correlations:

1. As the rating of the player gets higher, smaller changes in the rating increase the value by a big margin. We assume it is because the top players are very rare, and a small increment in their ability is hard to achieve, making their value significantly higher.
2. The potential is bounded from below by the overall as we would expect, since a player potential cannot be worse than his current ability.
3. Value and wage are nearly correlated linearly as a higher value is understandably correlated with the better players who earn more.

Now, we will look at the correlation between wage and overall rating. It will not surprise us to see that better players earn more money (as we have shown before), but we will also split the players into two groups: English and non-English footballers:



One might expect that players with the same overall rating, which represents the players' ability, will have similar wages. As we can see though, English players with the same overall rating seems to earn more money. Next, we will try to quantify how big this difference really is, and whether it can just happen due to randomness.

ESTIMATION AND HYPOTHESIS TESTING

We discovered before that English top-level players earn more compared to non-English players with the same ability. We would like to test whether it is due to randomness or whether the difference is significant. We find this question interesting since we are exploring in our research the financial side of English football, and we think that the conclusion of this question can have practical implications on players' negotiation of contracts. Our null and alternative hypothesis is as followed:

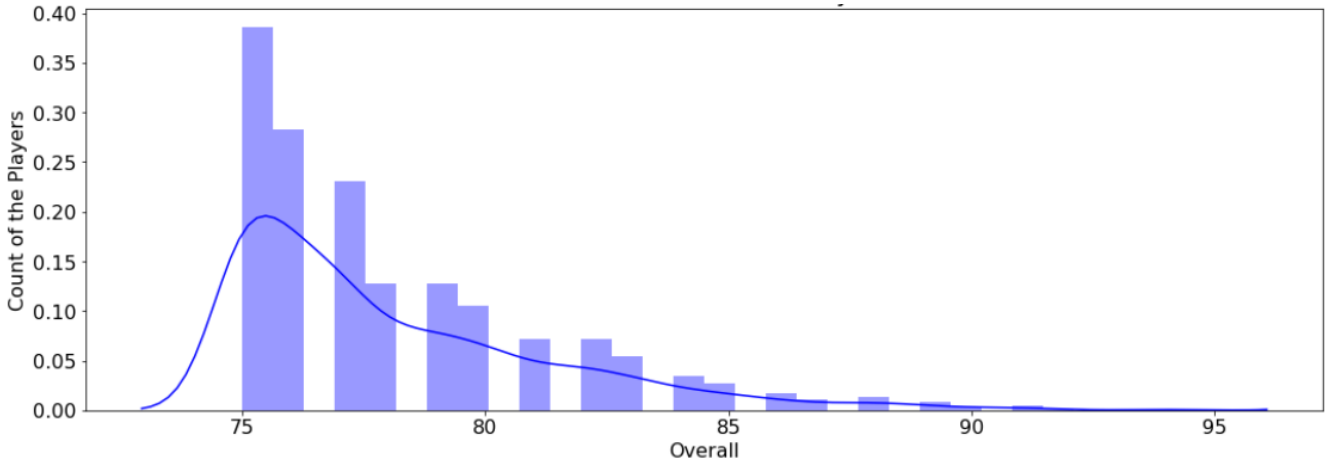
$\boldsymbol{H_0}$: Top-level (overall 75+) English football players earn on average the same as non-English football players with the same football skills.

$\boldsymbol{H_1}$: Top-level (overall 75+) English football players earn on average more than non-English football players with the same football skills.

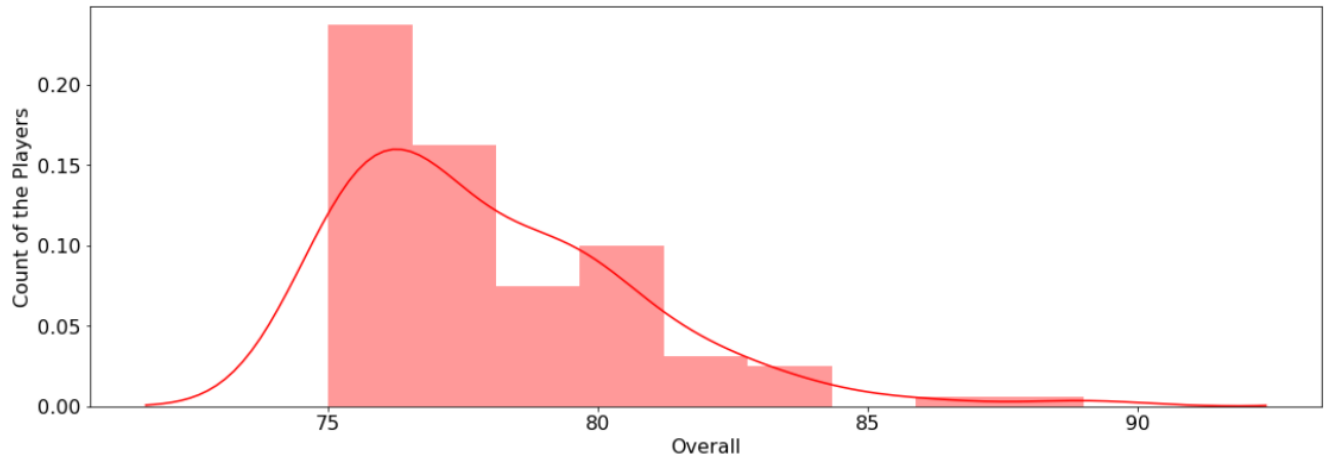
Test statistic: The difference between the averages of the before mentioned groups. First though, we want to make sure that the top English players' overall ratings are distributed the same as those of the non-English since wage is very correlated with overall which represents general ability. If the English players are simply better, it would be meaningless to show they earn more money as one can easily expect that:

Non-English median overall: 77.0
English median overall: 77.0
Non-English mean overall: 78.0972010178117
English mean overall: 77.99029126213593
Non-English mean STD: 3.3100503166225663
English mean STD: 2.688009488892081

Overall rating distribution for non-English footballers:

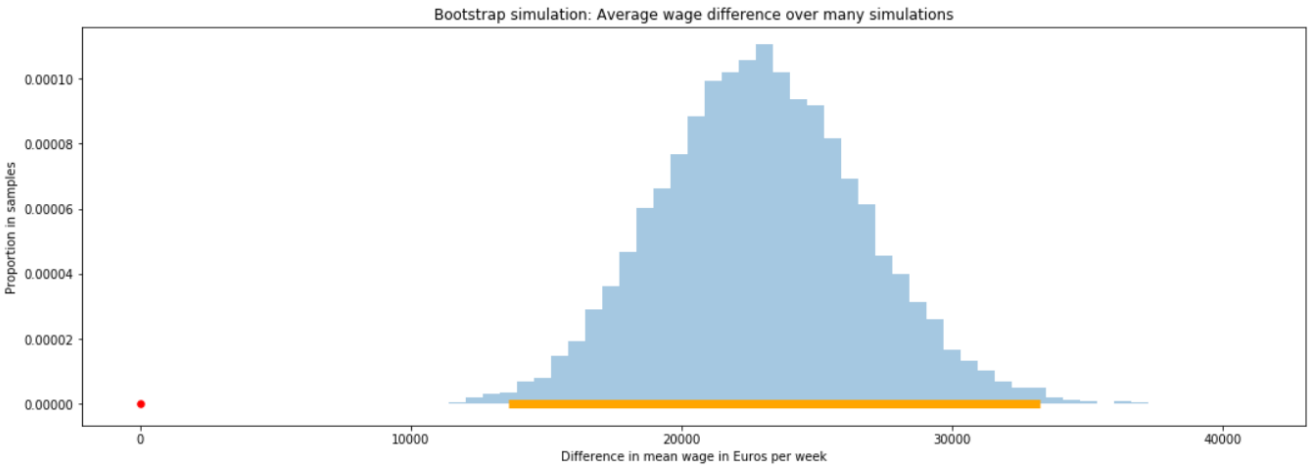


Overall rating distribution for English footballers:



We can see that the data is distributed quite evenly for both groups so we can move on for testing our hypothesis.

We will use bootstrap technique to simulate many groups of containing English and non-English players from our original data and then calculate the mean difference between the average wages. Finally, we will plot the distribution of all those differences.



The 99% bootstrap confidence interval for difference is between [14117 - 32943] Euros per-week

We found out that on average top-level English football players earn roughly 14k - 33k Euros per week more (!!!). Clearly zero is outside the 99% CI for the difference between the means, therefore we reject the null hypothesis and conclude with 99% confidence level that there is a difference between the average wage of the two groups.

It is important to emphasize that according to our findings we cannot conclude that being English leads to earning more money, but rather that there is merely a correlation between the two.

Since the mean income for top level player is around 45k Euros p/w (and the median income is even lower), we found a great difference in wages between the groups. This could come as a surprise without the proper explanation.

Among others, a likely explanation to our findings is that according to the "Homegrown Player Rule"[6] of the English Football Association (FA), each squad of the premier league teams must contain at least 8 out of 25 players who fulfill the "Home Grown Player"

[6] https://www.premierleague.com/news/1335777
https://en.wikipedia.org/wiki/Homegrown_Player_Rule_(England)

(HGP) criteria. A "Home-Grown Player" means a player who, irrespective of nationality or age, has been registered with any club affiliated to The Football Association or the Football Association of Wales for a period, continuous or not, of three entire seasons, or 36 months, before his 21st birthday (or the end of the season during which he turns 21).

It is reasonable to assume that majority of the players who fulfill the criteria are English. This combining with the previously shown data by which the EPL is the most profitable football league in the world, may explain our findings.

LIMITATIONS: What are some limitations of your analyses and potential biases of the data you used? How might these biases affect your findings?

1. A potential measurement bias and the main limitation the research might have is that it is attempting to give real world insights using data from a video game. We need to be aware to the fact that the data might be inaccurate at some cases, therefore affecting our results.
2. When cleaning the data we dropped some missing values, hence a slight selection bias might occur, though we do not believe this had a significant effect on our results since the missing data was a very small part of the whole dataset.

FUTURE DIRECTIONS: What new questions came up following your exploration of this data?

In our work, we gave a special emphasis on the value which the tools we develop bring to people and businesses. A promising tool we would like to develop is a prediction to the outcome of a match given the data we have. Additional data we need is teams latest form, players latest form, where the match is being held (home or away), weather, opening lineup, injuries, suspensions etc. Some immediate uses of this tool will be:

1) Football clubs - Managers could test different lineups in order to know which players are the best fit for a specific game in order to create an optimal team for different cases. In some cases, players with a lower overall rating might be more suitable. For instance, better passing, dribbling, physical attitude or more experience is needed in certain situations.
2) Gambling businesses – no need to expend further.

Some more question we are curious about:

1. How can we evaluate players with key contribution to a team performance? An additional data needed is the match statistics including parameters such as winning rate, goal scored, goals conceded, total shots, total tackles, key dribbles etc.
2. Is there a correlation between the social-economic state in which a player grew up to his success as a professional footballer? There are many known cases where kids from very poor places became a great success, if we can find there is no correlation between the social-economic state to future success, it would be a make a positive social image for football and a great motivation for kids to play football and become professionals. To answer this question, we will need the social-economic history of as many football players.
3. Evaluate the efficiency of various training methods to improve players skills. For this task, we will need the different statistics of a player over time as well as the time he was practiced with a certain training method.
4. Evaluating the personal and professional traits of a manager in a manner similarly to the evaluation of the players. This to predict a manager effect on his team success.