FIFA Players Rating and Wage Analysis

- Yuchen Wu and Xinmeng Zhang -

CSE 163

Hunter Schafer

March 2022

Content

Research Questions

1. Compare the attributes such as Passing, Shooting, Pace, Dribbling, Defending, and Physic of the top 5% players over the years. Has the indicator of good players changed? I.e. Can we observe which attributes have become more/less important indicators to a good player?

   On average, the defending, passing and dribbling ratings have increased among the top 5% players, indicating that soccer players value defending techniques and teamwork more than before. Among these attributes, dribbling remains to be the most important indicator of a good player.

2. What is the most preferred foot of players in 2022?

   In 2022, the most preferred foot of players is right foot (about 76%), and approximately 24% of players are left-footed.

3. What is the distribution of FIFA player's nationality and what are the Top 10 nations that have the most players in 2022?

   Overall, England has the most players in the world. Germany and Spain rank 2nd and 3rd place respectively in 2022. Other nations that contribute the most players are France, Argentina, Brazil, Japan, Netherlands, United States, and Poland.

4. What is the trend of average height and weight of players in the Big Five League?

   From 2015 to 2022, the average height of players increased, while the average weight of players decreased. However, the range of change is minor: within 1cm of average height and within 1kg of average weight.

5. What is the age distribution of current players in 2022?

   Players of the age between 20 to 25 years old have the largest number, while the number of players older than 30 years old decrease significantly.

6. What is the wage distribution by nationality of players in the Big Five League (England, Spain, Germany, Italy, and France) in 2022?

   The English Premier League has the highest median wage, while French Ligue 1 has the lowest median wage. Interestingly, English Premier League also has the largest interquartile range while Spain has the lowest interquartile range.

7. Has Messi and C. Ronaldo's ability on each attribute improved over the years? Summarize their personal characteristics based on the attribute value distribution.

   Both Messi and Ronaldo have decreased in pace and increased in defending over the years. Messi also has significantly increased in passing and shooting, while Ronaldo has decreased rating in all attributes except defending and shooting. When compared together, Messi has better dribbling and passing techniques, while Ronaldo outperforms in physics.

8. Can the wage amount paid to the players reflect their true abilities?

   When we look at the wages of overall players, wages seem to adequately reflect their true abilities, however, if we look at the top 20 players, wages cannot accurately reflect their true abilities. By top 20 we mean the 20 players with the highest overall rating scores. Our predicted model generates a very large MSE as we see in the results of our test file, meaning that some players with similar abilities can have very different wages.

Motivation and Background

FIFA has been one of the most famous and popular soccer video games in the world. It is well known not only for its excellent gameplay, but also for its large player dataset, which includes over 20,000 players around the world. In this dataset, the player ratings are well based on the actual performance of soccer players, which can reflect the ability of these soccer players in the real world.

Therefore, besides the basic information that we want to collect from the FIFA players and summarize, such as the distribution of preferred food and age etc., we also would like to dive into this dataset and try to get an insight into the real world such as the trend of most valued attributes of players or the reasonability of the players' wage. For example, is it true that speed is more valued than other skills of players compared with before? Also, does the players' wage match their ability or is the wage just unreasonably inflated? Answering these questions is not only fun, but also offers a better understanding of the situation in the real soccer world.

Datasets

We used the sofifa dataset on Kagel to complete our analysis whose original download link is here. We selected the 2015 - 2022 player's data and can be found in the 'Dataset' folder of our github page. The major variables in the dataset include player's name, player id, overall rating, value, wage, league, nationality and their attribute data such as Attack, Power and Shooting etc.

Methodology

Numbers are corresponding to that of research questions.

1. Use pandas to select the top 5% players by their overall ratings from 2015 to 2022, then use seaborn and matplotlib to visualize the change of some key attributes over years in a line chart. Attributes include “pace”, “physic”, “shooting”, “passing”, “dribbling”, and “defending”.

2. Use pandas to filter 2022 data by the column “preferred_foot” and then plot a Pie Chart using seaborn library to see the distribution of the most preferred foot used by players.

3. Use the wordcloud library to first make a cloudmap reflecting the frequency of the nationality of the FIFA players, referring to the column “nationality_name” of the 2022 dataframe. Note that when we convert the “nationality_name” column to a single string text, wordcloud counts the collocations such as “Brazil Brazil”. Therefore, we need to convert representative nationalities consisting of multiple words into one, such as “United States” to “UnitedStates”. Next, plot a histogram to see the top 10 countries of that year that have the most players.

4. Use pandas to first simply merge the 2015 - 2022 dataframes into one without using left on or right on. At the same time add a column “year” on each of the 15-22 dataset with input of that year. Then group each year’s data by the column “year” and only select players from the Big Five League. Then calculate the mean of height and weight and plot a line chart using matplotlib with Year as x-axis and Weight(kg) & Height(cm) as two y-axis, color hue by height and weight.

5. Since there are rarely players over 40 years old, we only look at the age distribution of players under 40 years old. We first use pandas to filter the players under age 40, and then plot a histogram to see the distribution.

6. Since the wages of players in the Big Five League and age between 20 to 40 are the most representative, we use pandas to filter these players of that age range. Then select the column “league_name” and “wage” of 2022 data, and use seaborn and matplotlib to make a boxplot of player wage by league name.

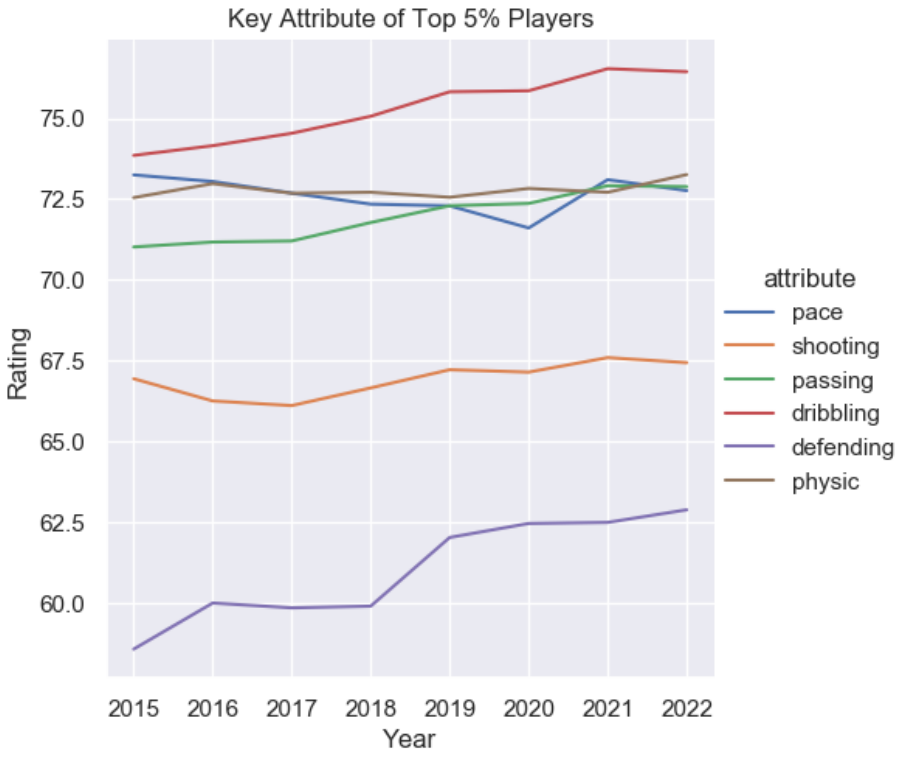
7. Use pandas to filter the row of Messi and C. Ronaldo of the merged 2015 - 2022 data using the column “short_name”. Then filter by their names and year to create a radar chart using plotly library to compare their six attributes each in 2015 and 2022. We also drew a polar graph to compare the attributes of Messi and Ronaldo in 2022. We can then summarize their characteristics based on these graphs.

8. Build a Decision Tree Regressor model using scikit-learn library in pandas, with wage as labels, overall rating, potential rating and player's value as features, using the 2015 - 2021 data to fit the model. Test the MSE. Then use the model to predict the wage of the 2022 data and see how well it predicts by looking at the mean squared error. If MSE is large, then we might be able to conclude that the wage is inflated and does not imply a corresponding improvement of the players' abilities.

Results

Has the indicator of good players changed over years?

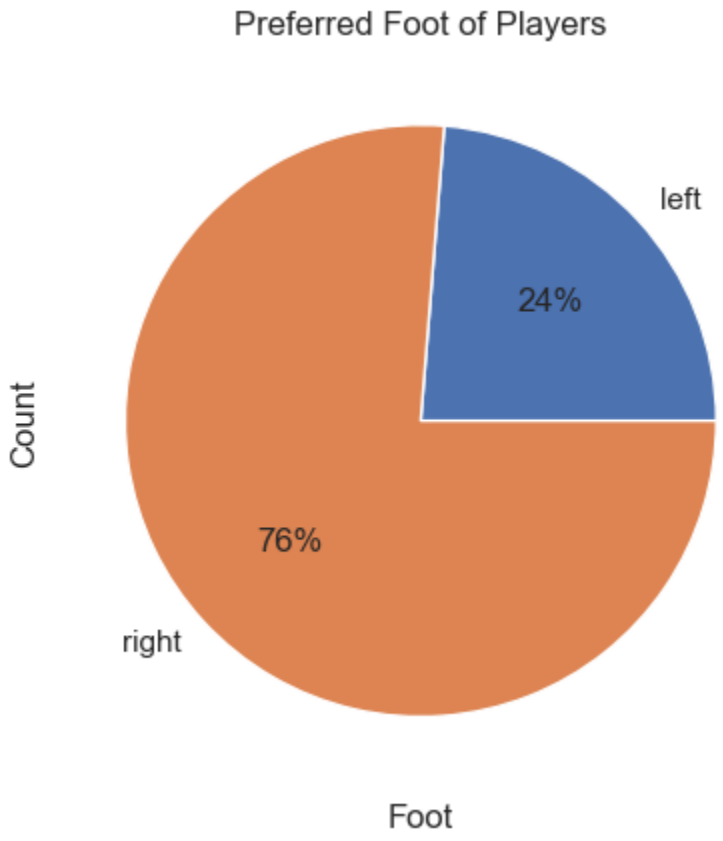
If we look at the average of the six attributes over the years, we can find a significant increase in the rating of defending. This indicates that clubs now value the defending techniques more than before. Besides, the average ratings of passing and dribbling have also increased. This may indicate that better dribbling skills and better teamwork are more valued characteristics of players. Since the player's list are changing over the years, we didn't merge the data by the player's id, but simply combined them together, we cannot be sure whether such an increase is due to individual improvement of the same group of players or because of newly hired ones. Among the six attributes, defending ranks the lowest. This might be due to the player's position on the football field. Usually, only 4 or 5 out of 11 players of the football team are mainly responsible for defending, so it's reasonable that the defending score is low for most players, especially for forwards. However, since the overall score of defending of the top 5% players are increasing over the years, we might be able to say that the requirement of the comprehensive quality for excellent players is increasing.



Graph 1

What is the most preferred foot of players in 2022?

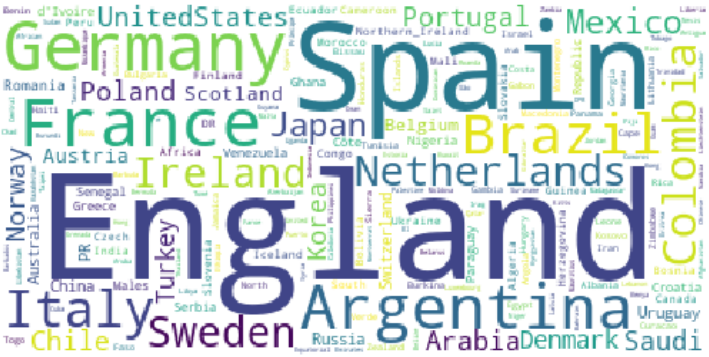
The pie chart shows that most of the players are right-footed. Specifically, 76% of players are right-footed and only 24% are left-footed. This number is quite interesting when compared with the overall ratio of left-footed people around the world. A study found that only 8.2 percent of people are left-footed (Tran and Voracek, 2016). The reason why soccer players have a relatively larger ratio of left-footers remains unclear. Some researchers propose that left-footed players have a better balance over their bodies, for example, Lionel Messi is a left-footed player. However, this hypothesis lacks solid evidence. We think that the reason for the high ratio of left-footed players in soccer is related to soccer itself. In most cases, the formation of a team is symmetrical, meaning that at least half of players in a team play on the left side of the pitch, and left-footed players often have a better performance on the left side when dribbling and passing.



Graph 2

What are the Top 10 nations that have the most FIFA players?

We can get an overall impression of the distribution of the nationalities by looking at the word cloud map (Graph 3). We can see that counting the frequencies of all nationalities from 2015 - 2022, players from England, Spain, Germany, Argentina and Brazil take a big portion of the overall players, matched with our usual impression of the highly developed football areas.



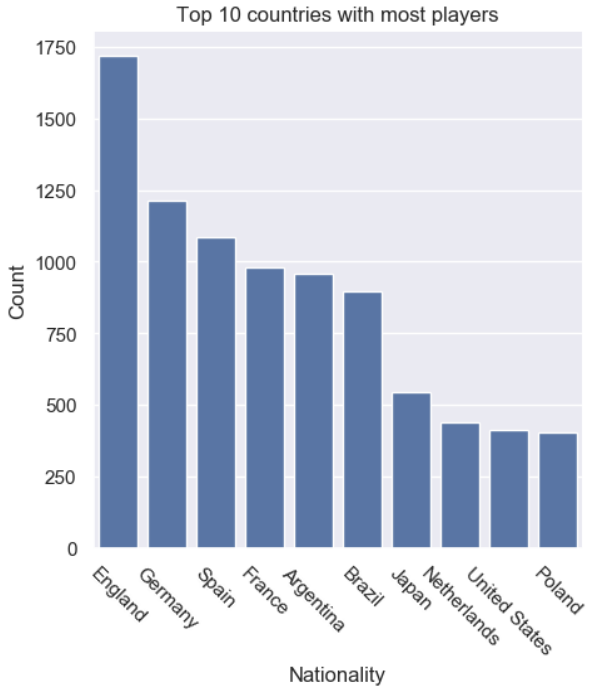
Graph 3

From Graph 4 we can see that in 2022, of the top 10 nations, England has much more players than any other nation. This might be due to the high level league in England - the English Premier League. Also notice that among the 10 countries, 6 of them are European countries, 3 of them are American countries (2 South America 1 North America), and only one of them are Asian countries. This implies that the development of football is highly unbalanced regionally. Now, let's look at Graph 5 which counts the distribution of nationality among the top 100 players. Surprisingly, though England has the most registered players, it does not have the largest number of top players, ranking only 6 while Portugal and Belgium entered the rank.

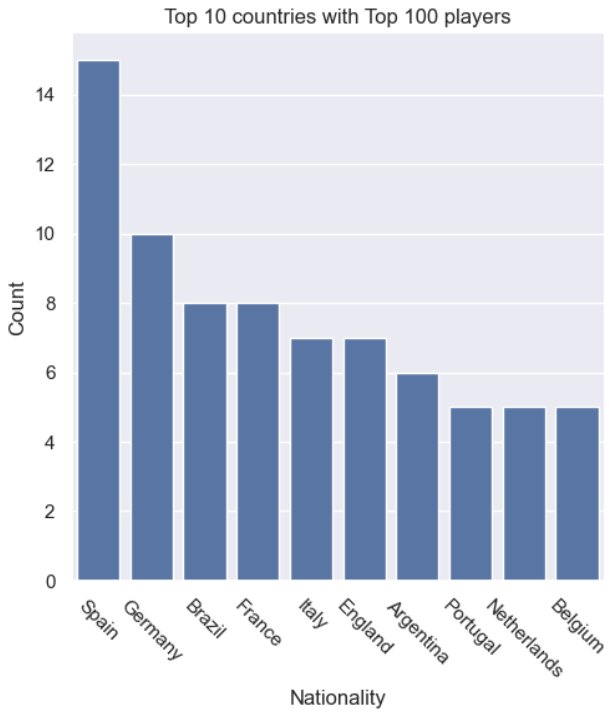
Interestingly, Asian countries with large populations such as China and India didn't have large football players' talent output, and their performances are relatively weak in both Region and

Country matches, compared with Japan and Korea which are also Asian countries but whose discourse power is much stronger. The reason behind such a phenomenon is interesting and can be discussed in further research on the local policy for talend fostering, player benefits and reward system, and cultural background.

Overall, the above findings imply that the football status may not be very correlated with a country's population but with the training and benefit system of that country and whether the policy is friendly in that area. Also, the number of registered players also does not imply a great advantage in country competitions such as England, since the player's level is also important.



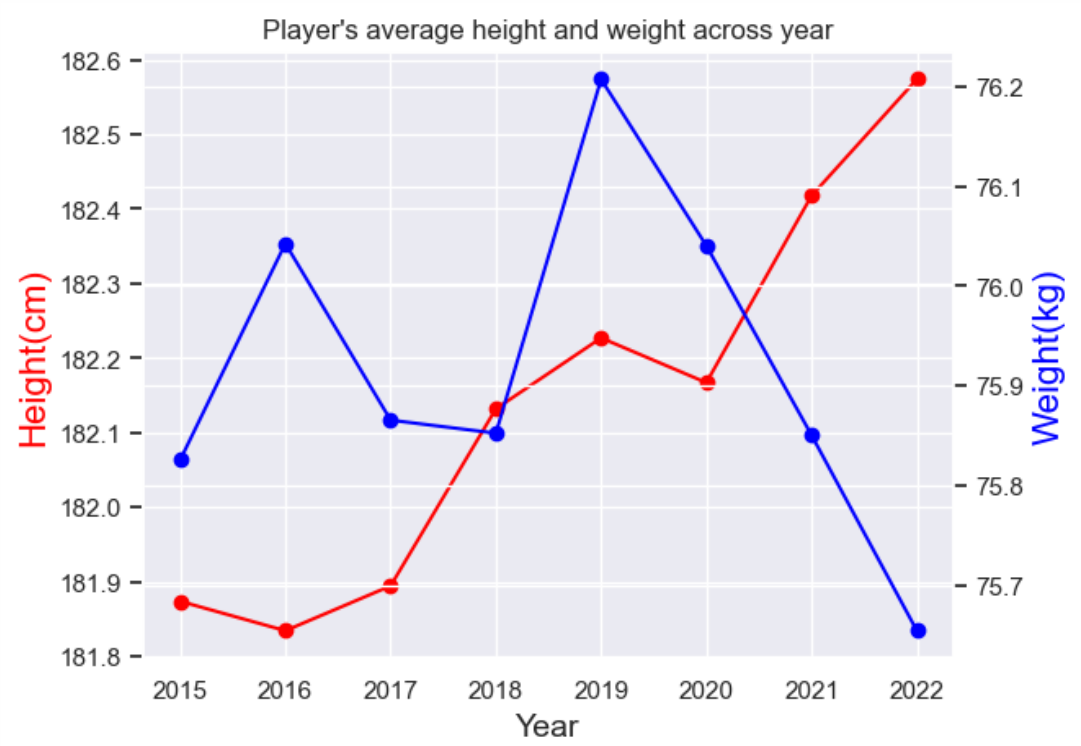
Graph 4



Graph 5

What is the trend of average height and weight of players in the Big Five League?

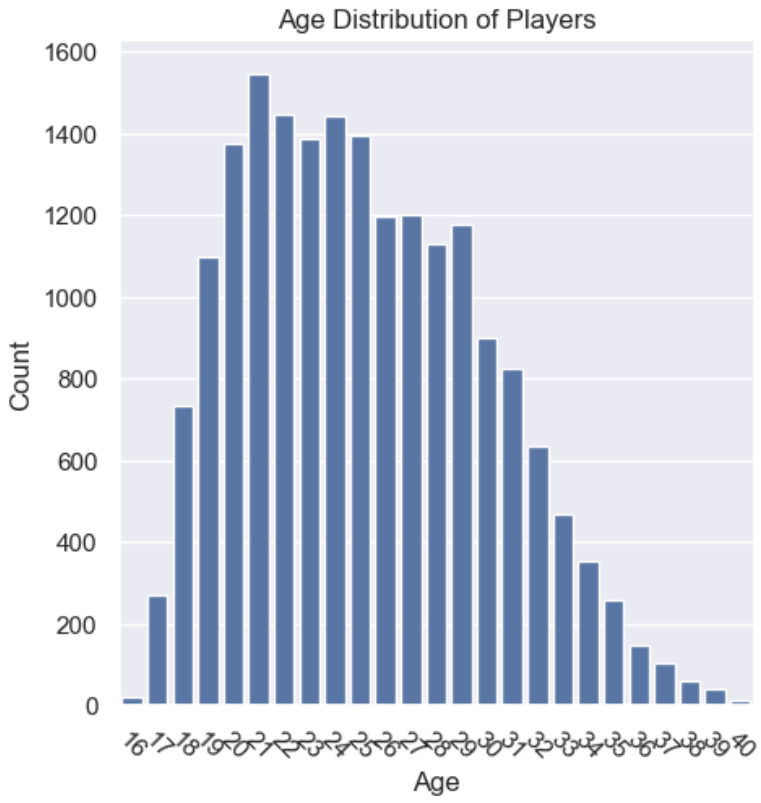
Graph 6 shows the trend of average height and weight of the FIFA players. We can observe that the average weight and height only has minor change over the past 7 years. The average height increased within 1 cm and the average weight fluctuated more, but also within 1 kg. Overall these two statistical numbers remained pretty stable, which is reasonable because, first, for the same profession adult player, his height and weight are unlikely to change very much, and secondly, this generation's height and weight are at a similar level so the newly registered players' statistics will shake the original one by much.



Graph 6

What is the age distribution of current players in 2022?

We filter the dataset by limiting the wage under 40 as most current players are under 40. The number of players between 20 to 25 is larger than other age intervals. We can see that soccer clubs now tend to have more young players in the team as the graph is left skewed. There are multiple possible reasons for that. The first factor is about soccer tactics: players between 20 to 25 years old often have significantly better stamina than other players. This allows them to cover more distance on the pitch than their opponents and brings advantage to the team. Another factor might be related with business operation: the market values of young players often have a greater potential to rise, so the clubs can benefit from the transfer of young players. It is noteworthy that the number of players olders than 30 decreases sharply, which might be due to the fast decrease in physical ability.

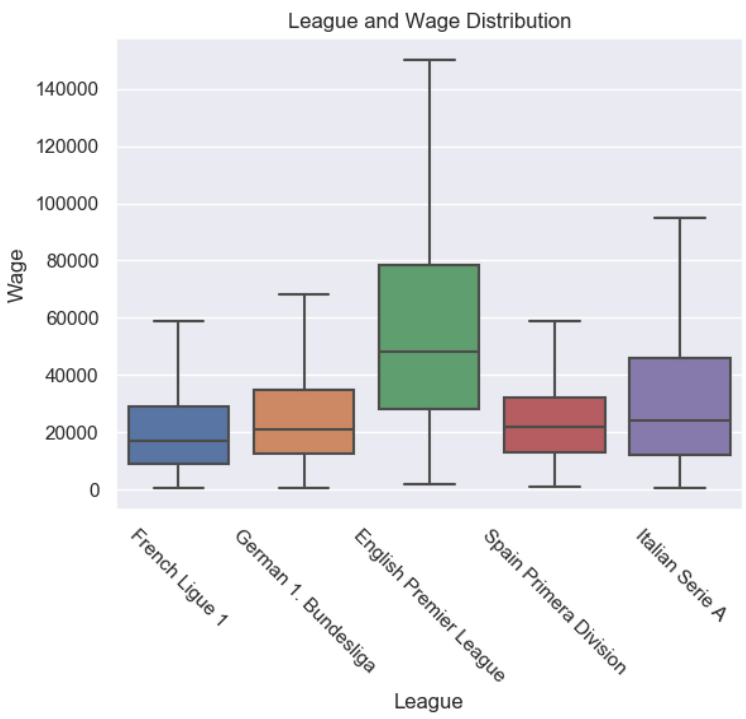


Graph 7

What is the wage distribution by nationality of players in the Big Five League (England, Spain, Germany, Italy, and France) in 2022?

Among the Big Five Leagues, English Premier League has the highest median wages than other leagues. This is not so surprising since the Premier League is the most successful league in the business field, which generated twice the revenue than the second place (Statista, 2021). Higher revenue of the league allows the clubs to pay a higher wage to their players.

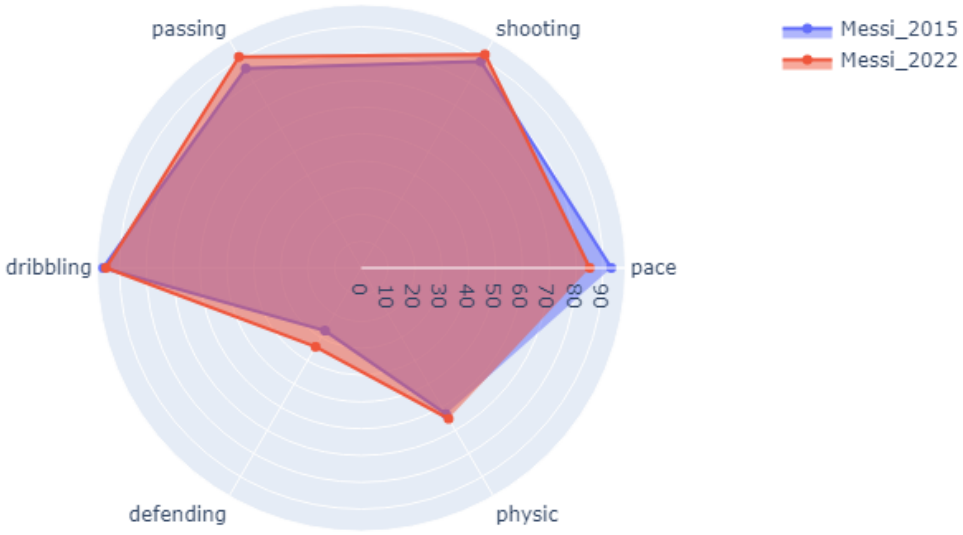
Another interesting point to look at is the interquartile range of wage for the Big Five League. Besides the highest median wage, the Premier League also has the largest interquartile range, meaning that there is a big gap between the wealthy clubs and other clubs in England. The top clubs in England sometimes can pay ten times the wages to the players than others. If we look at French Ligue 1, German Bundesliga, and Spain Primera Division, the corresponding interquartile range is much smaller, indicating that there is no big difference in the financial conditions among the clubs except one or two super clubs in the nation (i.e. Paris Saint-Germain in France, Barcelona and Real Madrid in Spain, and Bayern Munich in Germany).



Graph 8

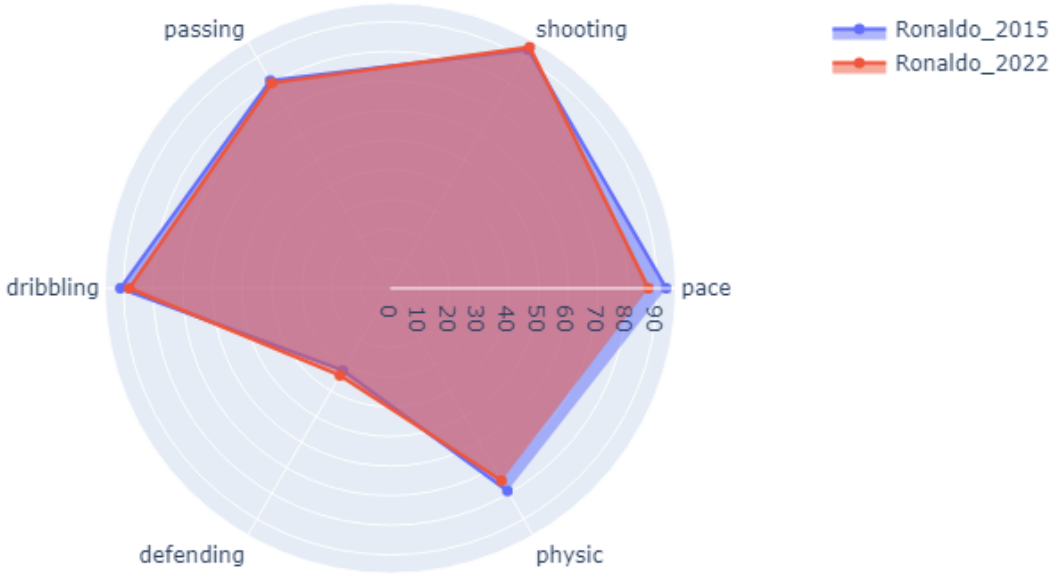
Has Messi and C. Ronaldo's ability on each attribute improved over the years? Summarize their personal characteristics based on the attribute value distribution.

Looking at Messi in 2015 and Messi in 2022, the biggest change is the decrease in pace due to aging. However, although the physical ability of Messi decreased, his passing and shooting skills progressed a lot. This shows that Messi has undergone a change of the role in the team. In 2015 Messi was the player who used the dribbling skill and speed to beat the defending players, but in 2022 Messi has transformed into a player who can connect the team using the passing skills. Messi has also become better at teamwork these years, as his defending rating increases. Messi now participates more in defense than in 2015.



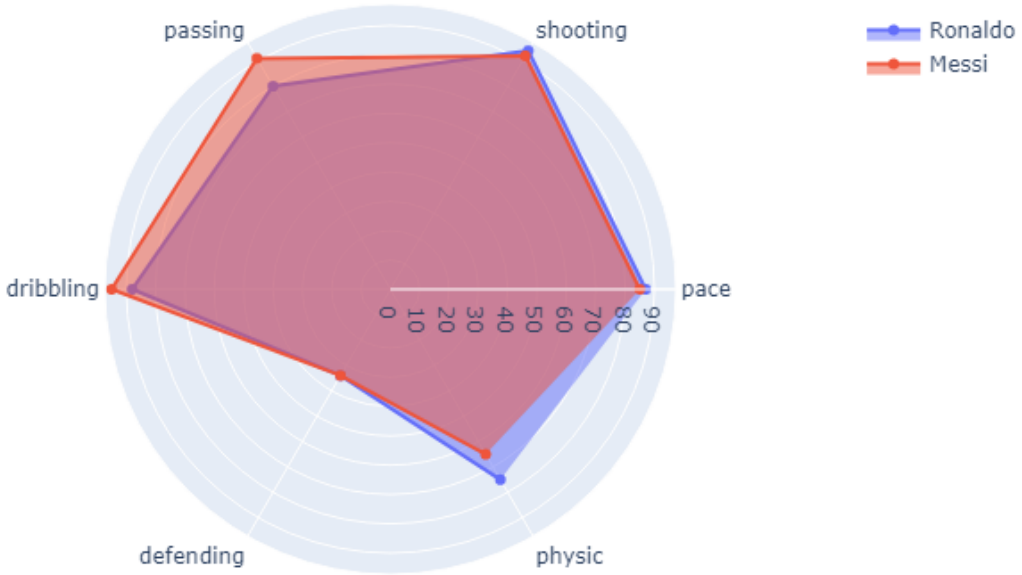
Graph 9

If we look at Ronaldo, his pace has also decreased a lot, which is not surprising since Ronaldo already turned 37 this year. However, the other attributes of Ronaldo have stayed relatively stable over these years. In fact, one reason that Ronaldo and Messi are the best players in the world is that they are highly disciplined, so they can keep a highly competitive form. Also, it's not surprising that defending is his weakest attribute since he is a famous forward player and mainly organizes attacks and shooting.



Graph 10

If we compare Messi with Ronaldo, we can rarely get a conclusion over which one is better than the other. In fact, both are the best players in the world. However, their positions on the pitch and the characteristics are very different. On the one hand, Ronaldo has a better physic rating and shooting rating, which means that he is a typical striker who uses physical condition and shooting skills to directly score for the team. Messi, on the other hand, is better at passing and dribbling, so he can attract the defending players and pass the ball to his teammates. If needed, Messi can also find the chance to score with his shooting skill.



Graph 11

Can the wage amount paid to the players reflect their true abilities?

It's natural to assume that players' wages are correlated with their ability, reflected as their overall score and potential score. It remains to test whether a player's value also correlates with wage. Therefore, we ran a regression plot on wage versus value on the 2015 - 2022 data and figured that there's indeed a positive correlation between these two variables as shown by Graph 12, and thus we included value as a feature. This finding is reasonable since when a player signs a contract with the club, the club usually needs to consider his value in order to determine his wage. Also note that the dispersion degree tends to increase as the amount of wage and value increases, meaning that the top players' wage fluctuates largely.

Moreover, looking at the MSE generated, we first divide the wage on each row by 10000. Then we use the merged 2015 - 2021 data to train and test the LASSO model and find out that the MSE is 5.51, which is relatively small, meaning that this model possibly has high accuracy in predicting the wage using overall rating, potential score and value. Therefore, we are confident to use the model to predict the wage in 2022 data, and found out the MSE to be 6.57, slightly larger than the first MSE, which is reasonable because the predictive power within the same dataset should be higher compared to using the model to a different dataset. The two MSE are both relatively small and thus we can conclude that the features selected indeed have predictive power on players' wage. However, notice that the MSE generated in the test file using only data from top 20 players are more than 10 times larger, which probably means that the predictive power model is weaker for the top players as their wage might be greatly influenced by other factors such as age and clubs they are in, with a similar competitive sports level.

Graph 12

Impact and Limitations

We discovered some very interesting topics that can be further studied in later research, such as analysis of football performance in regard to a country's policy on benefit system and talent fostering plans, or player's height change over the past 30 - 40 years with diet and nutrition structure change. That being said, we might be able to provide some potential research topics for people that are interested in this area for

The website only has data of the 2015 - 2022 dataset, and thus asking for the trend of some characteristics such as average height and weight may not be very informative. We can see from Graph 6 that the average height and weight are pretty stable in the stated years. If we have data across a broad range of time, for example 1980 - 2022, we might be able to observe a more obvious trend, and infer the possible influencing factors such as nutrition.

Challenge Goals

1. Messy Data and Multiple datasets

   In our project, we used multiple datasets from 2015 to 2022 and we wrote functions to combine these datasets first since we want to know the player ratings over the past few years instead of a single year. Moreover, based on the need of each research question, we need to combine and group the dataset in multiple ways, for example, which merge method we should use and how to select the top 5% by their overall ratings. Besides, the dataset contains too much information and some of the values are missing, meaning we need to filter out missing data. There are more than a hundred columns, and values include names, numbers, data, and URLs. To utilize the dataset, we need to make a lot of modifications first.

2. Machine learning

We want to use Machine Learning methods with the help of the scikit-learn library in pandas to create a LASSO model and fit the 2015 - 2021 data to see the predictability of the model on wage. Then we want to utilize the model to predict the wage in 2022 to see if players' ability and overall performance rating score can explain their wage amount, thus to tell if the wage is inflated. The difficulty is to determine what we want to predict and decide which feature and labels to use. We hesitated on whether to include the "value" column as a feature because there might be correlation but not causal relationship with players' wage. In the end, we decided …. Also, when we were trying to predict the 2022 wage, we choose the top … after serious consideration because …

3. Data Visualization

We need to choose the appropriate type of graph when plotting based on the need of research questions. Which one most efficiently and accurately conveys the information that we want to get is a key and hard task to accomplish. In this project, we experienced some new libraries and graph types such as word map and radar chart before, which generate our desired output more visually appealing and more efficiently.

Work Plan Evaluation

1) Data wrangling (2 - 3h) Actual: 4h

   Deal with all the garbleds and missing values in the data. Rearrange, filter or combine necessary datasets for time series analysis research questions.

   The data we have is very messy with more than 100 columns. We also need to merge the separate data from 2015 to 2022, which cost us more time than expected.

2) Data Visualization coding (5 - 7h) Actual: 8h

   Split the parts where need visualization into half and each one of us takes responsibility for our own parts. We can work individually at first and discuss it on Discord if we come up with any difficulties.

Some graphs that we want to plot are not taught in the class, such as polar graph and word map. We spend some extra time learning how to install and use them properly

3) Machine Learning coding (3 - 4h) Actual: 6h

   We want to do this together since we want to discuss in detail which labels to be included in the LASSO model and the interpretation of the outcome.

   We spend a lot of time discussing what factors should be included in the predicted model, and the actual testing costs longer time as well.

4) Script and interpret data analysis outcome (4 - 5h) Actual: 5h

   Do this together using collaborative documents such as Google Doc.

   We are not facing too much trouble interpreting our results and writing this report

Testing

We used a smaller data file that only included the top 20 players by overall rating. After we ran our test file and generated the results, we compared the plot we got using the smaller file with the raw data of the top 20 players, and found that our results were correct.

Collaboration

This work is done by Yuchen Wu and Xinmeng Zhang, with help from course staff Gabrielle Rackner on Machine learning such as feature selection. Questions regarding coding are resolved via internet search. Besides, all data analysis questions are done independently without outside help.