Original Research Article

Data Analyses of European Soccer

Yiou Wang[*]

Chengdu Foreign Languages School, Baicao Road, Chengdu 611731, Sichuan, China. E-mail: wangyiou06@gmail.com

Abstract: Using European soccer data sets, which contain data related to common European soccer leagues, players basic information, and teams' goals, etc., this paper analyzes the characteristics of European soccer and players, explores data visualization regarding European soccer, and makes predictions of results of matches. Based on Python 3 and some of the packages inside, such as numpy, the author improves the data set to make it clear and user-friendly. Visualizations of data and basic statistics, including Poisson Distribution, are then utilized to determine the results. Finally, this paper analyzes the attacking and defending abilities of different leagues and teams in Europe, ascertains distributions of players' attributes, and predicts match results by using Poisson distribution and Skellam Distribution. Generally, this paper analyzes data from leagues to matches to players. All these analyses are meaningful for the public to understand the characteristics of European soccer and the world behind the numbers.

Keywords: Soccer; Data Analytics; Python; Statistics

1. Introduction

Scholars and experts have tried hard to explore data in the sports arena. Python 3 has been gaining a tremendous amount of popularity over the past few years, and is the language of choice for many data scientists across the world. It is no accident that the language is also gaining popularity amongst sports scientists, who have to work with a lot of data on a day-to-day basis.

Data visualization is a significant part of analyzing data. Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data[1]. With the help of data visualization, we have the chance to analyze several characteristics of European soccer, such as a comparison of players and their abilities, a comparison of the different leagues, and the relative age effect.

The relative age effect in soccer is manifested by early educators and trainers, who group their players by chronological age to ensure equal opportunities for success. As we know, age-related cut-off dates are put into place to determine the age-range of our children's sports leagues as well as their classrooms. The idea here is to cordon-off groups of adolescents based on similar levels of physical and mental development. A result of these cut-off periods is a phenomenon known as "the relative age effect"[2]. Previous findings of skewed birth date distributions among sports professionals have been interpreted as evidence for systematic discrimination against children born shortly before the cut-off date for each age grouping[3]. By using Python 3, the author could explore the age distribution of European soccer players and find the potential rules for this distribution.

Predicting match results is also a hot topic around the world. One way of doing predicting matches is by using Pio-

doi: 10.18282/iss.v2i1.339

sson Distribution, which is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event[4]. Based on European soccer data of the past 30 years from the database, the author calculates the possibility of a specific number of goals and thus predicts the match results by comparing goal differentials (to be precise, we use Skellam Distribution, which is the distribution of the differences between two Poisson Distribution plots).

2. Methods

2.1 Data visualization

The author connects the database 'soccer_database.sqlite', and extracts some of the data that are essential to the data analyses in the next steps, such as players, leagues, countries, etc. The following step is to process the data to make sure that it is clear, and that it can be perfectly fitted in models the author creates. Then, the author needs to get the different numbers of players corresponding to the characteristics that the author is going to analyze. At this time, groupby and count() method is needed to determine the times that the corresponding data appears. Finally, based on matplotlib.pyplot package, the author illustrates plots of distributions.

2.2 Match results prediction of home teams and away teams

The essential technique that the author needs to predict the match result based on data of past matches is Poisson Distribution and Skellam Distribution. The probability function of Poisson distribution is:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}, k=0,1,\dots$$

The Poisson distribution is suitable for describing the number of random events occurring per unit time. The expectation and variance of the Poisson distribution are λ, in which the parameter refers to the average number of random events per unit time (or unit area). The eigenfunction is

$$\psi(t)=exp\{\lambda(e^{it}-1)\}$$[5]

The first step is quite similar to that of data visualization: extracting data that is essential to the next-step analyses. Then the data needs to be clear and well-formed, which ensures that the data can be fitted in the model the author is going to use. To predict the results by Poisson Distribution, first the author imports package poisson from scipy.stats. Getting the average number of goals is quite significant if we want to predict match results through Poisson Distribution. So the author needs the average number, lambda. The final step is to use the packages imported previously, as well as matplotlib.pyplot package, to get the plot of the Poisson distribution of the goal differences. If we want to gain the prediction result, we can find the probability that the goal difference is larger than 0, equal to 0, or less than 0, which signifies whether the home teams win, the two teams drew, or the home teams lose.

3. Experiments

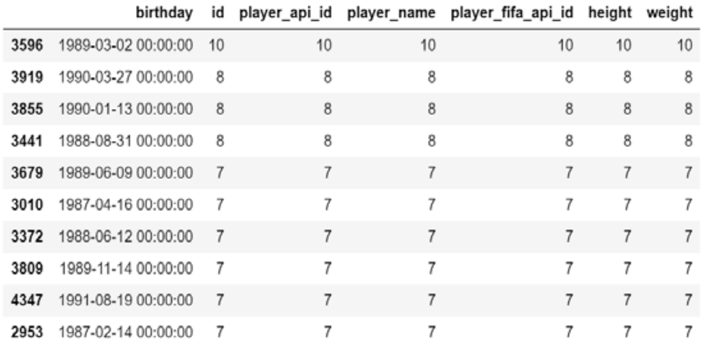
3.1 Data visualization

3.1.1 Top ten defensive and offensive leagues/teams

The author imports several packages, including numpy, matplotlib.pyplot, pandas, seaborn, and sqlite. To elicit the specific data, the author extracts league, country, team and match data by using the sqlite method. Since these data sets are given in separate columns, they need merging for a complete form. So merge method is used at this time. The merged form is called "match". Inside "match", the author finds the total goals (home team goals plus away team goals), which is an important indicator of the defensive abilities of a league. For example, if the number of total goals of a league is very large compared to that of other leagues, this league has relatively poor defensive ability (or good offensive ability). So, the author picks up columns "home_team_goals" and "away_team_goals", and gets the column "total_goal". The following step is to g*et al*l league names from the form "match", and use the names as ordinates. Thus, the author uses a for-loop to loop over the league names. By now, the author has already obtained all essential data: goal differences and league names. Then, by using matplotlib.pyplot and seaborn, the author obtains the bar-plot. Finally, the author provides a visualization of the data, which makes it clear to see which league has the best defending ability.

The process of getting the plot of the best ten defensive and offensive teams is roughly the same as that of getting the plot of the best ten defensive leagues. But there are some subtle differences. In order to obtain the best ten defensive and offensive home/away teams, the data about goals per game is quite important. First, the author gets the average home team stats and away team stats, such as the home team goals, away team goals, total goals and goal difference by using the method sort_values, which sorts the data the author needs. Then the author calculates goals per game, including goals scored and goals conceded per game. By using matplotlib.pyplot package, the author can gain the bar-plot of the top ten defensive/offensive teams.

3.1.2 Relative age effect



	birthday	id	player_api_id	player_name	player_fifa_api_id	height	weight
3596	1989-03-02 00:00:00	10	10	10	10	10	10
3919	1990-03-27 00:00:00	8	8	8	8	8	8
3855	1990-01-13 00:00:00	8	8	8	8	8	8
3441	1988-08-31 00:00:00	8	8	8	8	8	8
3679	1989-06-09 00:00:00	7	7	7	7	7	7
3010	1987-04-16 00:00:00	7	7	7	7	7	7
3372	1988-06-12 00:00:00	7	7	7	7	7	7
3809	1989-11-14 00:00:00	7	7	7	7	7	7
4347	1991-08-19 00:00:00	7	7	7	7	7	7
2953	1987-02-14 00:00:00	7	7	7	7	7	7

Figure 1

The process is quite similar to that mentioned previously. First, the author extracts the columns that contain information on players' birthdays, and makes them clear and well-fitted in drawing plots. Later, the author gets the months of players' birthdays, and uses group_by method to make sure that the data is grouped by months, and count() method to obtain the corresponding occurrence times of each month. Then a for-loop is used to get the names of all the months. Finally, by using matplotlib.pyplot package, the author draws a subplot of the distribution of the months of the players' birthdays.

3.1.3 Players' abilities comparison

To get clear comparisons, the author needs to use bar plots. First, as previously described, the author uses a for-loop to get all the relevant characteristics of the players. Then the author writes a function, for which there are two parameters—player 1 and player 2. The data that corresponds to the characters are grouped by players' names. Then the author obtains the average values of players' abilities. Finally, by using matplotlib.pyplot package, the author elicits a comparison bar plot of the two players' various abilities.

3.2 Match result prediction of home teams and away teams

First, the author extracts all the goals scored by using a previously described method to select specific columns: home team goals and away team goals. Since the database 'soccer_database.sqlite' is huge, containing data from all aspects of European soccer of the past few decades, it meets the conditions of Poisson Distribution, particularly that the total occurrences in relevant matches is extremely large. Next, by using method mean(), the author calculates the average goals of home teams and away teams, and saves them separately into "lambda_home" and "lambda_away". Additionally, "total_lambda", which represents the total average goals of all teams, is calculated.

Finally, by using matplotlib.pyplot and poisson, the author derives different plots. The author obtains the Poisson Distribution of home team goals, away team goals, and total goals. Most significantly, the author obtains the Skellam Distribution of the goal difference of home teams and away teams, which is used to predict the probability of winning, drawing, or losing.

4. Conclusion

4.1 Data visualization

The author deduced several conclusions pertaining to leagues and teams. **Figures 2** & **3** illustrate that home teams and away teams score the most goals in the Dutch Eredivisie than any other teams in the other ten leagues during the 2008/09 season to the 2015/16 season. Home teams in France’s Ligue 1 and Poland’s Ekstraklasa score the least goals compared to any other league during the 2008/09 season to the 2015/16 season. Also, the number of goals that home teams score is larger than that of goals that away teams score in all eleven leagues in Europe. Thus, we can connect these results to home advantages and away advantages. Home advantage is the psychological and physiological advantage that the home team has over the visiting team and it is prevalent in all sports, including soccer[6].We can properly speculate that home teams and away teams in Netherlands Eredivisie has more home or away advantage (Bundesliga also had an away advantage). Home teams and away teams in Poland Ekstraklasa and France Ligue 1 have the worst home or away advantage. Plus, these two figures demonstrate that the home team has a clear advantage during any given match.

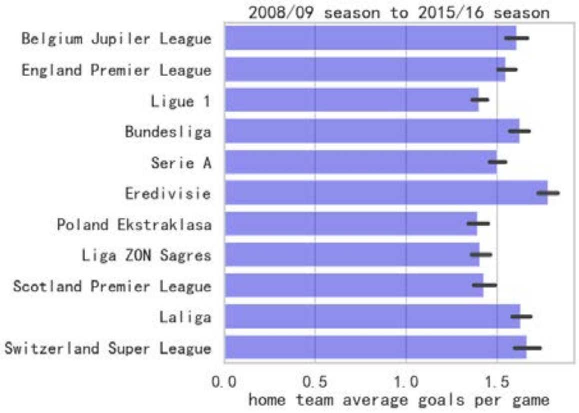


Figure 2

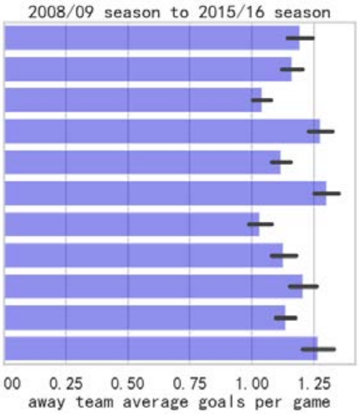


Figure 3

The distribution of total goals in the 11 leagues in Europe indicates that teams in the Netherlands Eredivisie have more goals than any other league. On the contrary, France Ligue 1 and Poland Ekstraklasa have the least total of goals than all other leagues in Europe. Therefore, we can draw the conclusion that teams have the best attacking ability in the Dutch Eredivisie, and teams have the worst attacking ability in France’s Ligue 1 and Poland’s Ekstraklasa. Admittedly, the conclusion may have some flaws and it may be due to defensive abilities rather than offensive abilities. But the method and the trends are accurate with significant approximation.

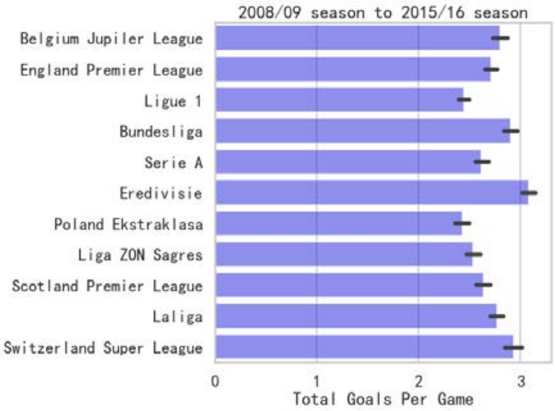


Figure 4

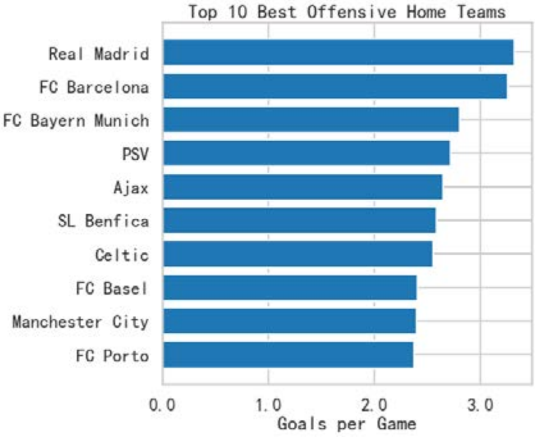


Figure 5

The author can also draw conclusions on the top 10 home/away teams which have the best attacking or defensive abilities. According to **Figures 5**, **6**, **7**, and **8**, Real Madrid is the home team that has the best offensive ability, whereas FC Barcelona is the away team that has the best offensive ability[6]. Meanwhile Glasgow Rangers is the away team that has the best defensive ability; and FC Porto is the home team that has best defensive ability.

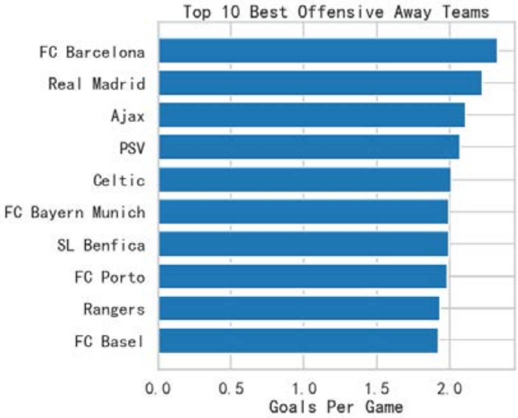


Figure 6

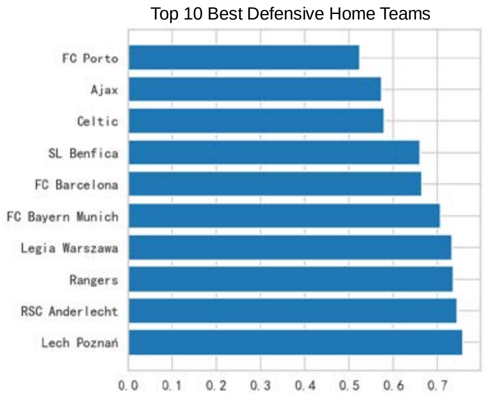


Figure 7

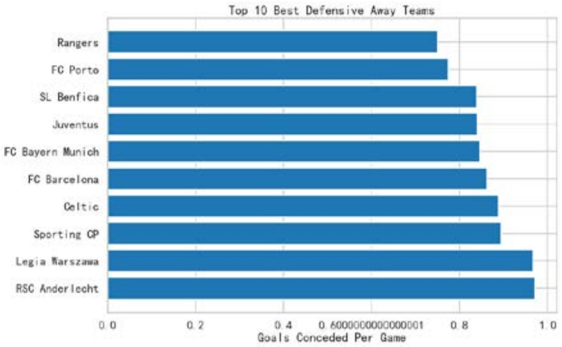


Figure 8

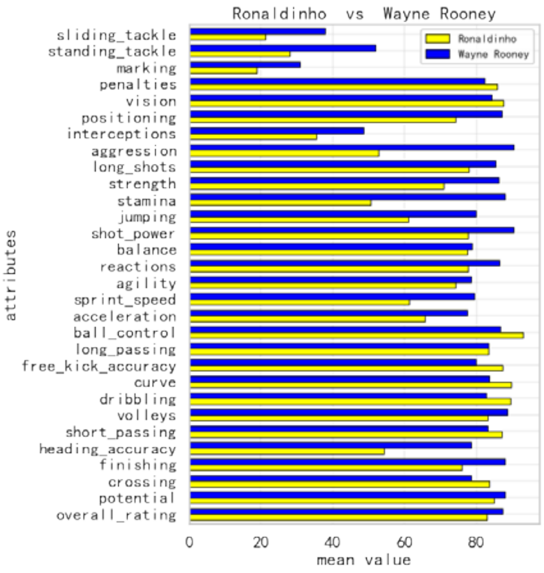


Figure 9

The author then focused on the abilities of individual players and how they compare with others. From the figure below, we can see the radar charts of different players, which show their abilities, their versatility, and the pattern of different characteristics. For example, from radar charts of eight famous play-makers, they all tend to have weaknesses in slide-tackling and intercepting ability. For most of the player-makers, they have weaknesses in heading accuracy and jumping ability (except Cristiano Ronaldo). And these eight players generally have competitive advantages in agility and shooting. Also, we find that Lionel Messi and Cristiano Ronaldo have the most extraordinary attacking abilities, such as ball control and agility; Wayne Rooney is relatively versatile in all aspects as well. Thus, we can properly determine that player-makers are generally weak in defending, such as successful tackles and interceptions, but they excel at ball control, speed, and reaction. If coaches want to train player-makers to make them more complete players, then they would do best to improve their defensive abilities.

If we want to compare two players closely, we can refer to the bar plot. For example, **Figure 10** compare the ability values of Ronaldinho and Wayne Rooney, two of the best players in the world.

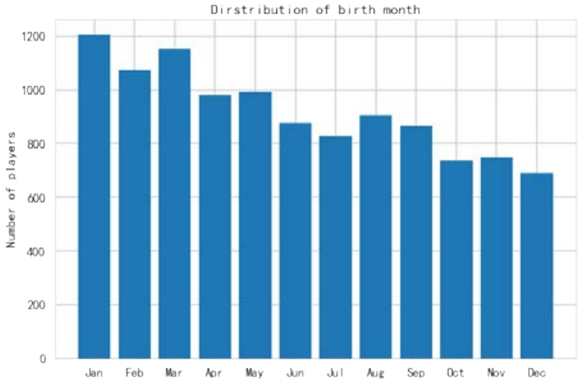


Figure 10

Generally, Ronaldinho is better at ball control, curve, dribbling, and other aspects of the game needing finetuned and intricate techniques. Wayne Rooney is better at aspects related to power and shooting such as slide tackling, standing tackling, aggression, jumping, and so on[7]. Thus, by comparing two different players through a bar plot, we can clearly see strengths and weaknesses, and how the values of pertaining to different abilities vary.

The relative age effect is also covered in the author’s research. According to **Figure 12**, there are more players born in the beginning months of the calendar year than the middle or last months. This can be attributed to a requirement for players to join a club: they must be under 18 by January 1 of that year.

Thus, players born in the beginning months of the year, who meet the requirements and are the closest to 18, would be more likely to join in the club since they are physically more advantaged than other players who compete with them. Thus, the conclusion drawn from the plot of the distribution of players’ birth months accurately confirms the relative age effect.

4.2 Match result prediction of home teams and away teams

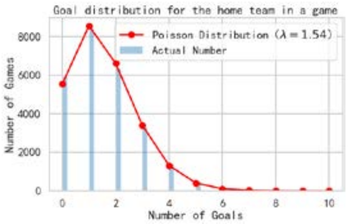


Figure 11

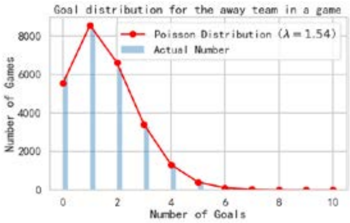


Figure 12

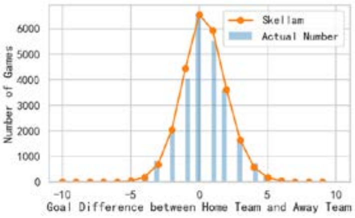


Figure 13

The author uses Poisson Distributions to obtain the distributions of number of goals that home teams or away teams could make. According to **Figures 11** & **12** above, the distributions of both home teams and away teams are roughly the same in that they are both skewed to the right, and their summits are both at one goal. However, these two figures also indicate that home teams can score more goals than away teams, which confirms the conclusion the author deduced previously: home teams have an advantage in scoring goals over away teams.

Looking at the distributions of goals scored by home teams and away teams can predict the match result by referring to the probability of various goal differences[8]. **Figure 15** demonstrates that the Skellam distribution of home teams and away teams' goal differences is more or less normal. Thus, the most probable match result is that both teams would tie. However, a closer look shows that the area of number of goals larger than 0 is larger than that of goals smaller than 0. This indicated that home teams have a greater probability of winning the game even though the probability is only a little greater than that of losing the match. Thus, these results further consolidate the theory that home teams, indeed, have a greater chance of winning the game compared to away teams.

References

1. Data Visualization Beginner's Guide: A Definition, Examples, and Learning Resources [Internet]. Available from: https://www.tableau.com/lean/articles/data-visualization.
2. Kids Born in These Months Are More Likely to Become Pro Footballers, Mia Kessler [Internet]. Available from: https://the18.com/soccer-entertainment/youth-soccer-relative-age-effect.
3. Hay R, Musch J. The relative age effect in soccer: Cross-cultural evidence for a systematic discrimination against children born late in the competition year. Sociology of Sport Journal 1999; 16: 54-64.
4. Konefa M, Chmura P, Andrzejewski M, *et al.* Analysis of match performance of full-backs from selected European soccer leagues. Central European Journal of Sport Sciences and Medicine 2015; 11(3): 45–53.
5. Zhang S. Home advantage in soccer. PIT Journal 2015; 6.
6. Haight FA. Handbook of the Poisson distribution. New York, NY, USA: John Wiley & Sons; 1967.
7. Katz A, Hayes A, Suresh T. Poisson distribution [Internet].
   Available from: https://brilliant.org/wiki/poisson-distribution/.
8. Rein R, Memmert D. Big data and tactical analysis in elite soccer: Future challenges and opportunities for sports science. Springerplus 2016; 5(1): 1410.