What makes teams strong?*

Exploring statistics of Premier League 2018/19

Dai Moroi

April 6th 2022

Abstract

This paper analyzes the basic statistics of 2018-19 season England proffesional soccer league, Premier League, to figure out what makes the team strong, and also examines cause and effect of home advantage. The result indicates that there are storng correlations between strength of the teams and average possessions, the number of shots, goals scored, goals conceded, and the percentage of goal scored out of total number of shots. As for home advantage, although storong home advantage is observed, its cause is not explained in terms of the crowd factors.The linear regression model is used to analyze the data.

Keywords: football, soccer, home advantage, linear regression, Premier League

Contents

*Code and data are available at: github.com/moroidai/Final-Paper.

1 Introduction

What makes the team strong? What kind of characteristics does a strong team have? To answer these questions with convincing stories, statistics can play a essential role. Since the Internet enabled us to accumulate data easily, use of data has been featured in the world of sports, including soccer. This paper takes the example of 2018-19 season of England professional soccer league, Premier League, and examines its basic statistics, such as goals, possession rate, the number of shots, etc., from various aspects.

First, I explore the relationship between the number of wins and a variety of factors so that we can better understand the features of the teams winning many times. Then, the focus will be moved to the next topic, home advantage. In the sports competition, it often occurs that weak teams defeat strong teams at their home stadium. This phenomenon is called home advantage, which is defined by Courneya and Carron (Courneya KS 1992) as "the consistent finding that home teams in sports competitions win over 50% of the games played under a balanced home and away schedule." I see if the home advantage really exists in the Premier League, and then try to discover its cause from crowd factors. Thirdly, I focus on the statistics of the "Big six" Premier league teams, that is Manchester United, Liverpool, Chelsea, Arsenal, Manchester City and Tottenham Hotspur. These six teams are said to be biggest and most successful clubs in the league, and they actually all finished as top 6 teams in 2018-19 season. Looking at the similarity of these six teams and differences between them and others should be useful in learning the characteristics of strong teams.

The remainder of this paper is structured as follows: Section 2 discusses the data using graphs and tables. This section also explains the methodology that the data was collected and bias they have. Section 3 discusses the model, Section 4 presents the results, and finally Section 5 discusses the findings and weaknesses to be overcome. All the analysis is done in a reproducible way.

2 Data

2.1 Data and variables

The analysis for this paper uses the R statistical programming language (R Core Team 2021), primarily using `tidyverse` packages (Wickham et al. 2019) for data manipulation. It uses `pdftools` package for parsing the PDF (Ooms 2022), and `bookdown` package (Xie 2020), `ggpubr` package (Kassambara 2020), `patchwork` package (Pedersen 2020), and `knitr` package (Xie 2021) for making a R markdown report.

The data, in a .csv format, are available on the website, FootyStats. Please note that the data of Premier League 2018-19 season is the only free data right now. The data contains 20 observations of 735 variables in total. Although there are many variables, including corners, cards, offside, and data divided by every 15 minutes, this paper focuses on main stats, such as the number of goals scored, wins, or average possessions. All the variables and its explanation are available here. Table 1 shows the he first ten rows and five columns.

Table 1: First ten rows and five columns of the dataset

team_name	common_name	season	country	matches_played
Arsenal FC	Arsenal	2018/2019	England	38
Tottenham Hotspur FC	Tottenham Hotspur	2018/2019	England	38
Manchester City FC	Manchester City	2018/2019	England	38
Leicester City FC	Leicester City	2018/2019	England	38
Crystal Palace FC	Crystal Palace	2018/2019	England	38
Everton FC	Everton	2018/2019	England	38
Burnley FC	Burnley	2018/2019	England	38
Southampton FC	Southampton	2018/2019	England	38
AFC Bournemouth	AFC Bournemouth	2018/2019	England	38
Manchester United FC	Manchester United	2018/2019	England	38

2.2 Methodology

The dataset is based on the official open statistics by premierleague.com. Official Premier League performance data is collected and analysed by Opta, part of Stats Perform. The methodologies of data collections is described as follows (League, n.d.): "Live data is collected by a three-person team covering each match. Two highly trained analysts use a proprietary video-based collection system to gather information on what happens every time a player touches the ball, which player it was and where on the pitch the action occurred. Alongside them, a quality control analyst has the ability to rewind the video feed frame-by-frame in order to make certain that the information being distributed is as precise and consistent as possible. All the Premier League data collected is then subject to an exhaustive post-match check to ensure accuracy." Thus, the process of collecting data should be appropriate and trustworthy.

Because most of the data are clearly observed and counted by trained analysts, there is little bias in this dataset. However, there are two possible bias. The first possible bias is that some data are sometimes hard to count and can lead to overcounting or undercounting. For example, when you tell whether a shot is on target or off target, it sometimes happens that the shot is deflected from another team player. In this case, the shot is counted based on if the shot would have been on target or off target without deflection, but it is often hard to distinguish whether the original shot was on target or not. Yet, these sort of thing rarely happens, so it does not really affect the outcome of the analysis. The second possible bias is about the data called expected goals. Expected goals is a measure created recently to assess the quality of shots at goal. Before the invention of the concept of expected goals, we could just know the number of shots or chances created but not the quality of chances. Even though they are counted the same as shots, one can be better chance and another can be not a good chance. Here, expected goals try to measure the probability that shots will be scored based on many factors. Those factors include distance from goal, shot angle, type of attack, body part used to shoot, and assist type. I don't show the exact calculation method here because it is not very relevant to the paper, but I can say that the data of expected goal can be biased because of its complexity of calculation.

3 Model

This paper uses simple linear reggression model for the analysis. The mathematical equation of this model can be expressed as follows:

$$y = \beta_1 + \beta_2 x + \epsilon$$

Where β_1 is the intercept and β_2 is the slope. ϵ is residuals, the part of Y the regression model is unable to explain.

4 Results

4.1 General analysis

Firstly, using linear regression, I look at the correlations between the number of wins and four basic stats, that is goals scored, goals conceded, average possession rates, and the number of shots. Figure 1 shows the regression line and the correlations between wins and the four factors. You can see that there are very strong positive correlation between wins and goals scored, and strong negative correlation between wins and goals conceded, where correlation coefficients are 0.97 and -0.91 respectively. In addition, it is noted that the correlations are also very strong with average possession and the number of shots, where correlation coefficients are 0.84 and 0.86.

From this result, we can conclude that the strong teams scores many goals, while they rarely concede it. Also, they keep high possession rate throughout matches and shoot many times.

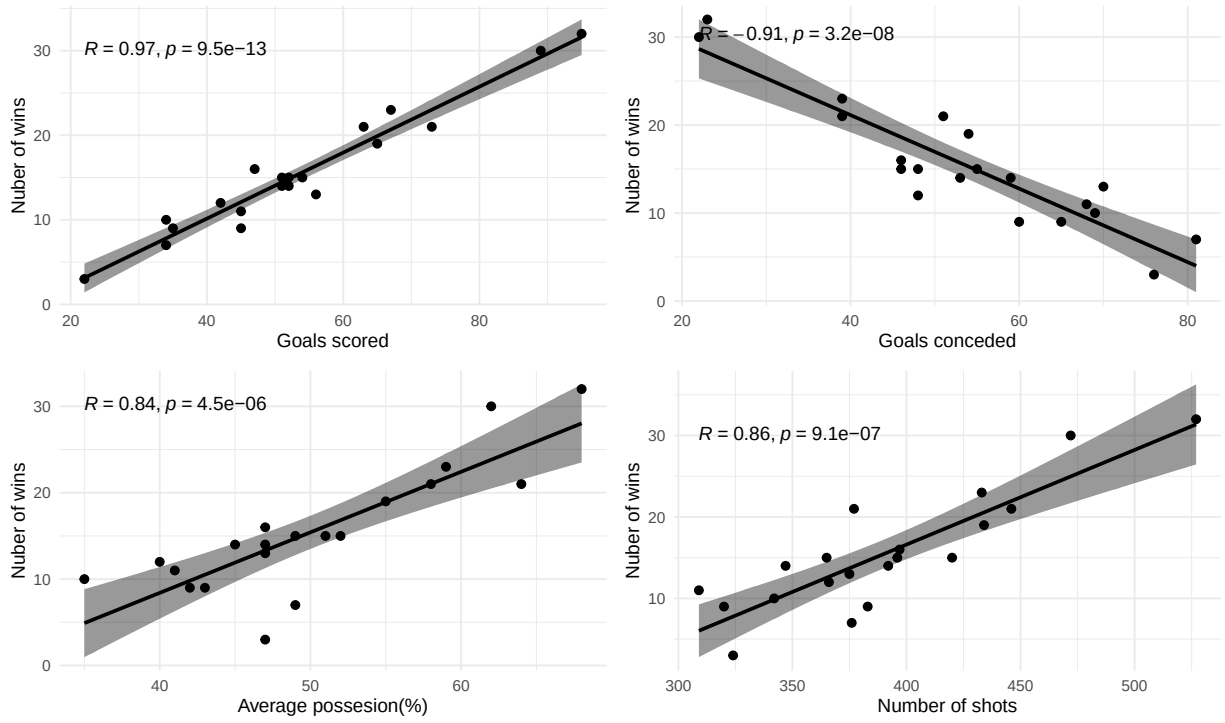

Figure 1: Correlations between wins and four factors

4.2 Home advantage

As stated in the Introduction, there is a concept called home advantage. Let's see if we can observe home advantage in Premier League. Table 2 shows the winning percentages of all teams in Premier League 2018-19 when they play at home stadium or away stadium, and the difference between the two. Although there is a exception of Crystal Palace FC, where they are 20% more likely to win at away, almost all teams have significant home advantage, where the average of difference in winning percentage is over 15%. According to the definition of home advantage, home team should win over 50% of matches, but the results still shows that there is noticable home advantage in Premier League.

But what causes home advantage? You probably think this phenomenon is strange because normally stronger team should win no matter where they play. To figure out the reasons behind it, I focus on the crowd factors.

Table 2: Home advantage of Premier League 2018/19

Team name	Win at home (%)	Win at away (%)	Difference
Manchester City FC	94.7	73.7	21.1
Liverpool FC	89.5	68.4	21.1
Arsenal FC	73.7	36.8	36.8
Tottenham Hotspur FC	63.2	57.9	5.3
Chelsea FC	63.2	47.4	15.8
Everton FC	52.6	26.3	26.3
Manchester United FC	52.6	47.4	5.3
Wolverhampton Wanderers FC	52.6	31.6	21.1
West Ham United FC	47.4	31.6	15.8
Leicester City FC	42.1	36.8	5.3
AFC Bournemouth	42.1	26.3	15.8
Watford FC	42.1	31.6	10.5
Newcastle United FC	42.1	21.1	21.1
Burnley FC	36.8	21.1	15.8
Cardiff City FC	31.6	21.1	10.5
Fulham FC	31.6	5.3	26.3
Brighton & Hove Albion FC	31.6	15.8	15.8
Crystal Palace FC	26.3	47.4	-21.1
Southampton FC	26.3	21.1	5.3
Huddersfield Town FC	10.5	5.3	5.3

My idea is that enthusiastic atmosphere that a huge crowd makes cheers players, which lead to make the most of their ability. If that is true, the number of crowds being at their home stadium matters. Figure 2 shows the correlation between the number of fans at stadiums and home advantage measured by winning percentage difference. The figure indicates that we can not recognize the correlation because the correlation coefficient is only 0.23.

But the problem of using the raw number of attendance at stadium is that we could underestimate the power of full stadium. For example, a full stadium with 20000 fans can be more enthusiastic than a stadium with 20000 fans whose capacity is 60000. In order to take this effect into account, I use crowd density as a factor, which is calculated by dividing the average attendance by stadiums' capacity. Figure 3 shows the correlation between home advantage and crowd densities. However, the results still don't show that there is positive correlation between home advantage and crowd factors.

4.3 BIG 6

Here, I look at the characteristics of "Big six" teams, namely Manchester United, Liverpool, Chelsea, Arsenal, Manchester City and Tottenham Hotspur.

Figure 4 shows the relation between goals scored and expected goals, and blue dots shows the data of Big six teams. The 45 degree line means that expected goals equals the actual goals scored, which is supposed to be most expected to happen. You can see that all big six teams are located higher than the line, which means they scored more than expected. This happens probably because they have a lot of good strikers and those strikers are good at shooting, which resulted in scoring more than expected.

Same as the previous figure, when you look at Figure 5, which shows Relations between the league position and percentage of goals scored out of total shots, it is noted that "Big six" teams are more likely to score when they shoot.

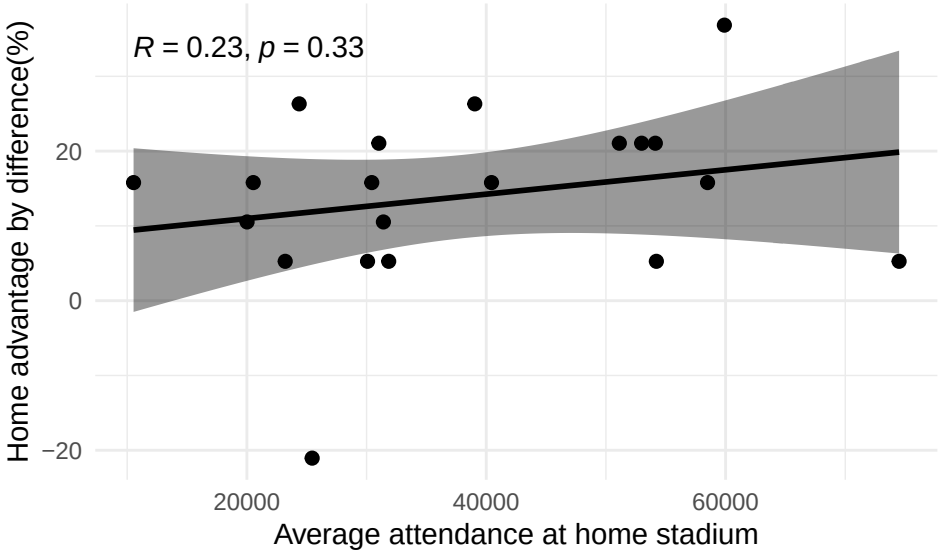

Figure 2: Correlations between home advantage and number of fans at stadiums

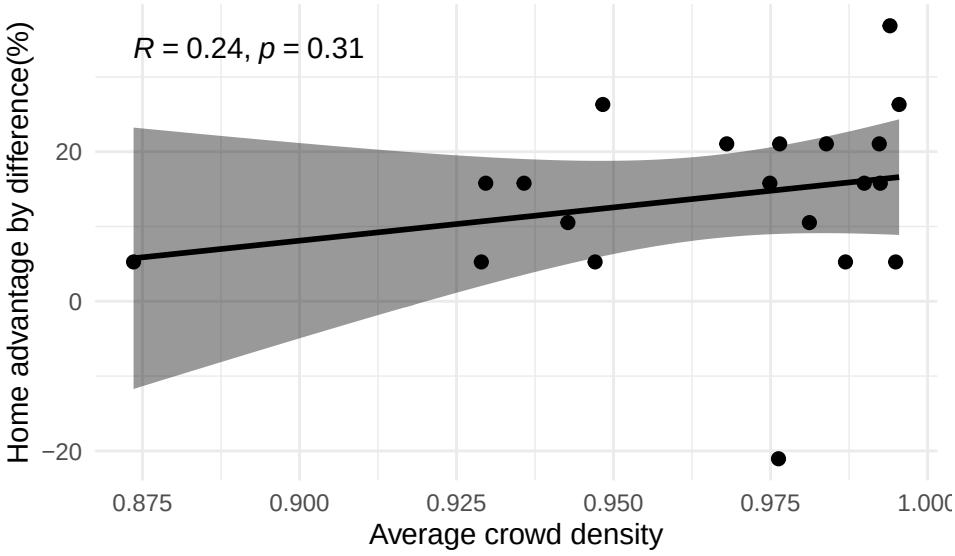

Figure 3: Correlation between home advantage and crowd densities

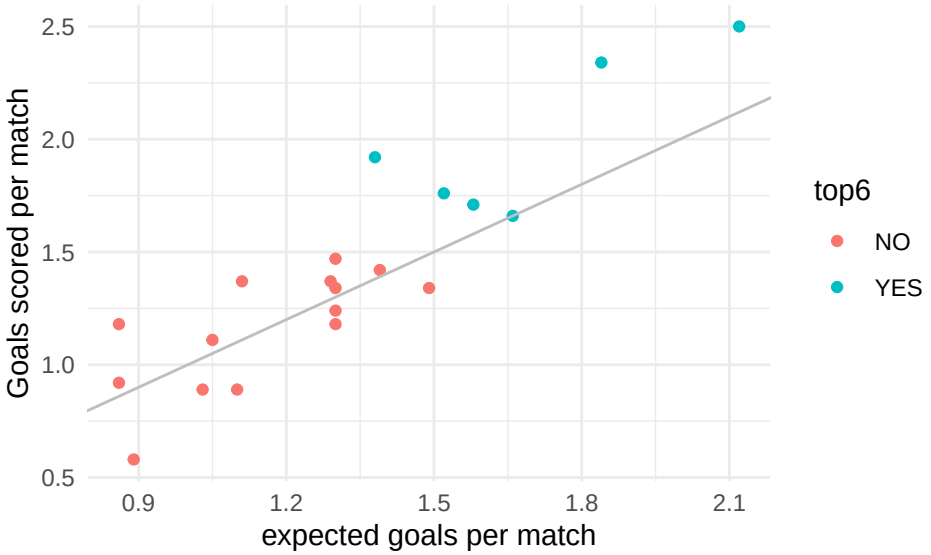

Figure 4: The relation between goals scored and expected goals, sorted by BIG6

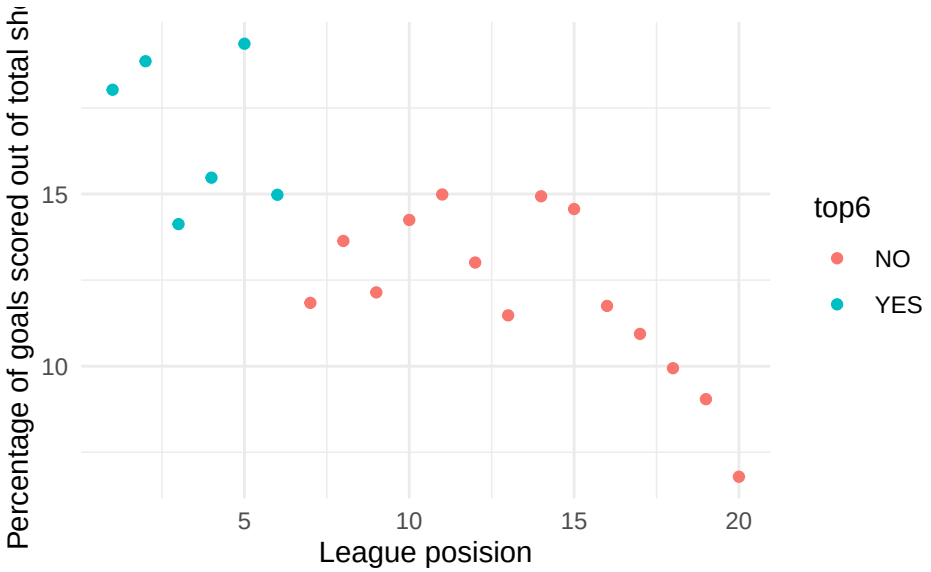

Figure 5: Relations between the league position and percentage of goals scored out of total shots, sorted by BIG6

5 Discussion

Over the analysis above, I explore the basic stats to find the characteristics of strong teams and what makes them strong. I would like to discuss two points regarding the results here. First point is about the fact that the strong correlation between wins and the average possession rate is observed. I think that this result can change over time because the trend tactics of soccer have been changing, and I expect that this correlation will decay over time. Almost 10 years ago, especially in Spanish league, strong teams preferred keeping possession, but now it is popular that they don't stick to keep possession, but instead bring the ball straight faster as possible to the goal, because hey can attack before the opponent's defender get ready if they attack fast. The second point is about crowd density. Looking at Figure 3, you probably notice that the crowd densities is pretty high in Premier league, which means all the teams played in the full stadiums, which makes it difficult to analyze home advantage based on crowd factors. Therefore, I can say that Premier League is not a good example for analyzing crowd factors because it is very popular. Leagues with less popularity should be chosen, if you want to do it.

The limitation of this paper is that it only uses the data of single season and single league. As a next step, it is recommended to compare it over time or across different leagues. Also, the dataset only contains basic stats, so using more advanced and deeper stats, such as the one using graphic data of fields, is highly recommended.

Appendix

A Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - There is no description as to why they created the dataset on the website.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - The dataset was created by the Footystats (https://footystats.org/). It is based on Premier League official stats.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - There is no description.
4. *Any other comments?*
   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The instances represent the Premier league teams.
2. *How many instances are there in total (of each type, if appropriate)?*

- 20.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - Yes, it contains all possible instances
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consists of 735 continuous variables (raw data).
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - No.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - No
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - No.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There are no errors, sources of noise, or redundancies in the dataset.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - There is no confidential data, and the dataset is publicly available.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - It doesn't contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

- No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - It is not possible to identify individuals.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No, it doesn't.
16. *Any other comments?*
    - None

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
    -
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
    -
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    -
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
    -
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, pleae describe the timeframe in which the data associated with the instances was created.*
    -
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
    -
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

-

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   -

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

   -

10.

A.1 *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

11.

A.2 *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

12.

A.3 *Any other comments?*

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

   -

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   -

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   -

4. *Any other comments?*

-

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - No.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - https://github.com/moroidai/paper4/blob/main/outputs/paper/paper.Rmd
3. *What (other) tasks could the dataset be used for?*
   - Analysis on soccer tactics, league characteristics, etc.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No.
6. *Any other comments?*
   - None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - TBD
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - TBD
3. *When will the dataset be distributed?*
   - TBD
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - TBD
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - TBD

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - TBD
7. *Any other comments?*
    - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - TBD
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - TBD
3. *Is there an erratum? If so, please provide a link or other access point.*
    - TBD
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - TBD
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - TBD
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - TBD
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - TBD
8. *Any other comments?*
    - TBD

References

Courneya KS, Carron AV. 1992. "The Home Advantage in Sport Com- Petitions: A Literature Review." *J Sport Exerc Psychol 1992* 2: 245–57.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots.* https://CRAN.R-project.org/package=ggpubr.

League, Premier. n.d. *Statistics Explained.* https://www.premierleague.com/stats/clarification.

Ooms, Jeroen. 2022. *Pdftools: Text Extraction, Rendering and Converting of PDF Documents.* https://CRAN.R-project.org/package=pdftools.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with r Markdown*. https://github.com/rstudio/bookdown.

———. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.