Does Playing at Home Have an Effect on a Football Team's Performance on the Field?

Hardit Singh (hsh310), Polina Boneva (pba590), Paulo Alting von Geusau (pau220)

February 3, 2020

1 Introduction

This paper aims to find an answer to the question, 'Does playing at home have an effect on a football team's performance on the field?', and it aims to do so by simplifying, examining, modifying and extracting the needed information from two main datasets found on Kaggle, first dataset ranging from 1993 to 2018 (dataset 1) [1], and the second ranging from 2000 to 2018 (dataset 2) [2]. These datasets are both in CSV format, and contain the full overview of dates, teams, goals, referees and seasons played in the English Premier League throughout the years 1993 to 2018. In addition to the 16000+ rows of observed and recorded football games, few more datasets were added from the website 'The World of Football', which suffice to describe the home stadiums of each team, the attendance in each season, and the comparison of financial value of the teams, which again were in CSV format. By analyzing 25 seasons of the English Premier League, we aim to reach a good understanding of the influence playing at a home stadium has on a football team, as the Premier League is one of the biggest competitions the players face every year.

2 Clean-up and Normalization

We first imported the above-mentioned datasets into Pandas Dataframes, and then started out the cleaning process by dropping columns with null values, or columns that we planned to exclude from our analysis. In cases where both the datasets contained the same columns, we chose to ignore the duplicate columns from one of the datasets as per convenience. Next, we found that the formatting of the date strings was inconsistent across the two datasets, hence we normalized the date strings using regex patterns and then parsed the dates into Datetime objects for easier use. After the parsing and cleaning of the datasets, we were left with the following: dataset 1 now contained information only about the date, home team, away team, full-time home goals, full-time away goals, full-time result, and season; and dataset 2 contained information about date, home team, away team, full-time home goals, full-time away goals, full-time result, referee, home fouls, away fouls, home yellow (cards), away yellow (cards), home red (cards), away red (cards), and season. Duplicate team names with different spellings, i.e., Middlesboro and Middlesbrough, were mapped to the same official name along with all other teams. Also, all referees' names were remapped manually, in order to use their official names as well. This unification aided all of our transformations and groupings later. Lastly, it is worth mentioning that dataset 2 contained a gap: there was no information about the season 2001-02. The final step of simplifying the data included removing the rows for season 2000-01 in order to keep the data uniform and prevent outliers and false correlations.

3 Data Exploration

In order to profile the data and investigate its quality, accuracy, and validness, we started out by looking for anomalies, patterns, and dependencies between columns. Both the datasets contain a column called the full-time result (FTR), which holds values H, A, or D, which stand for home team won, away team won, and match drawn respectively. Thus, our research starts by counting how many times does each of those letters appear in the column. Dataset 1 was grouped by column FTR and the resulting table's count values were used to fill up a pie chart, in order to visually estimate the percentage of games won by the home team, as opposed to draws or wins by the away team. Almost half of the games were won by the home team - 46.5%(Appendix 2, Fig. 2). The rest of the games' results were relatively equally divided between A and D. However, this was not sufficient to conclude that the stadium's ownership itself plays a role in those wins. To investigate further, we delved into the dynamics of the game.

In order to extract more information from our data, we merged more datasets into the initial ones.

The two collections of data we added to our main dataset, dataset 1, were about the average attendance per team for each season and the locations of all of the teams' stadiums. The data for each season's attendance was contained in a separate file; thus, we imported each of these files into separate dataframes and mapped them in a dictionary object with the corresponding seasons as the keys. Since we did not have the actual attendance figures for every match, we assumed that the average attendance of each match would be approximately equal to the average attendance of the home team for a particular season. A little nugget to mention here would be that in order to keep the computation as light as possible, we used the set_value attribute of the dataframe instead of the newer at/iat attributes wherever possible, given it proved to be upto 3 times quicker than its newer counterparts[3].

To further our analysis, we added a column for the goal difference in each of the two main datasets by subtracting the full-time away goals (FTAG) from the full-time home goals (FTHG). It served us tremendously because we were able to quickly judge if the home team won and by what margin by examining whether the number in this column was positive, negative, or zero. In order to deduce all of this data to conclude the relationship between a game's attendance and its score, we used Pearson product-moment correlation (Pearson coefficient). The result shows that the correlation between goal difference and attendance is 0.22, which shows a very slight positive correlation. It shows a high variation between the points and the line of best fit. It could be concluded that as the attendance increases, the goal difference increases in some of the games, however, Person's formula does not suffice as a proof that the opposite is not exact as well. Thus, one can say, and it will be just as valid, that as the goal difference increases, so does attendance.

Further, the collection of data about the stadiums' locations around the country was constructed of columns for each home team, their city, their stadium name, its latitude and longitude. We used the geographical coordinates of all the stadiums to find the physical distance between them and the city from which the away team had supposedly travelled, which would later be used to see if there is any correlation between the distance that the teams and their supporters have to travel for a match and the results of the match. We calculated the distance using the geo-coordinates that we had through the Vincenty method[4]. Later, we added the StadiumDistance column to dataset 1 in order to gauge the distance the away team had to travel for each game. This information, naturally, is not guaranteed because it is possible that the away team travelled from somewhere besides their home town. Using the Pearson coefficient once more to find a correlation between goal difference and stadium distance, we arrived at the number 0.04, which shows no correlation. A scatterplot proves this by showing us a completely dispersed group of points all over the graph (Appendix 2, Fig. 1). Both when stadium distance is large and when it is small or none, the goal difference ranges from negative to a positive value.

Similar to the technique we used to create a column to exemplify the goal difference, we added columns for the difference in fouls, yellow cards, and red cards for each game in dataset 2. Those numbers were calculated by subtracting the number of home fouls, home yellow cards, and home red cards from the number of away fouls, away yellow cards, and away red cards, respectively. The differences are about the team which played at their home stadium in a particular game, meaning that a positive number shows how much more the away team has been sanctioned in comparison to them. The following iterations were done using those new columns after dataset 2 was grouped by the home team. By aggregating the grouped dataset 2 four times by four different columns, each time extracting the sum and the count for each value, we managed to figure out the average values, in the span of all 16 seasons for each team, of the goals difference, fouls difference, yellow cards difference, and red cards difference. The columns used were GoalsDifference, FoulDifference, YellowDifference, and RedDifference, respectively.

The resulting dataset helped us arrive at some intriguing conclusions. To start, the number of teams that score more goals at home and win the game is almost the same as the number of teams that do not. The average scoring margin is close to 0, namely 0.13, and while 21 teams on an average win by scoring more at home, 20 teams, on the other hand, had a negative average score and do not win at home (Appendix 2, Fig. 4). In contrast, the sanctions received by each team's opponent - here team being the home team and opponent being the away team - in a particular game, were much more than the ones received by the home team itself. A bigger part of home teams were not charged with as many fouls as their opponents, as opposed to the ones who were charged with as many or more fouls than the away team (Appendix 2, Fig. 5). Additionally, the gap quantity of home teams that have been given less yellow and red cards than their game opponents, as opposed to those received more than their opponents, is much larger. Only six of the forty-one teams have received fewer yellow cards than their 'away' opponents, and only seven of them, red (Appendix 2, Fig.

6 and 7). Those outcomes insinuate a referee bias, which we decided to investigate further. In the light of getting reliable results from valuable data, we filtered dataset 2 to get only those rows which had experienced referees - those with experience of more than ten games in total. To determine a referee bias score, we used a formula to calculate a value for each game, which we inserted into a new dataframe under a column titled RefBias. That is, per game we took the sum of yellow cards for the away team (AY) and twice the number of red cards for the away team (AR) and subtracted the sum of yellow cards for the home team (HY) and twice the number of red cards for the home team (HR). A positive difference signifies that the referee has given more yellow and red cards to the away team. Interestingly, it turned out that out of all the referees we looked at, only one has a negative bias score (Appendix 2, Fig. 8).

After creating the additional columns and features, we reached a stage where the data was ready to be visualized to ease our analysis. Naturally, other than the pie chart we constructed to picture the overall home team wins against away team wins and draws, we also plotted a bar graph to show the number of wins of the home team for each season. Each of them showed almost twice the advantage of a home team, with slight variations (Appendix 2, Fig. 3). Furthermore, we had the exact dates of each game in all 25 seasons, which meant we could include the stakes held in front of both teams by comparing their performance in the beginning and middle of the season versus at the end. We assumed that the higher stakes would lead to better performance for both teams and home advantage should not vary too much. On the contrary, the plot shows a big increase in home team win rate during the end of the season for all teams which rank higher in the Premier League and for all teams which are in the upper part of the lower half of the overall ranks (Appendix 2, Fig. 9). All the teams in the middle of the rank and the very bottom of it did not show a difference in win rate when they played at home. We can conclude from this observation that teams who have a lot at stake may pump their game up at home a lot more than expected and gain much better results overall. However, the graph is not consistent, thus the possible influence of a home stadium on play cannot be used as a reference to explain those results.

At this instant, we had gathered enough data and had made enough iterations to think about calculating a 'home advantage score', somewhat trying to replicate how the football website 'FootyStats' [5] did, if possible. After some hit and trial, we arrived at this formula: the average of the scoring advantage and defense advantage, each of which were calculated by a combination of goals scored and conceded by a home team respectively (explained in Appendix 1). At this instant, we had the home advantage score for every team which had played at the Premier League in the last 25 seasons, and we decided to plot the two which had stood out as outliers in our preliminary analysis of the home advantage scores: Leicester City and Manchester United (Appendix 2, Fig. 10). Interestingly enough, in 2013-14, at Manchester United's lowest point of home advantage score over the years, they also experienced their worst season since the end of the 1980s. However, at Leicester's lowest home advantage score after their return in the League since 2014-15, they won the Premier League in the season 2015-16. Consequently, the score of home advantage is an arbitrary number that may or may not be taken into account when rooting for a team, but cannot suffice to predict which team will have the most significant advantage at their home stadium.

Out of curiosity, we dug into the two outliers from our last analysis. In the case of Manchester United, their lowest point mentioned earlier was also the same year they played with a new manager after their previous one, Sir Alex Ferguson, left the position after 27 years of service[6]. The Leicester City on the other hand performed their best during the season with their lowest home advantage score, and went on to win the Premier League during that season. After research, we attributed this win to the then all-new data-analytics system that Leicester City implemented in their home stadium[7], which most other football clubs have done ever since.

4 Conclusion

All things considered, as of now we can only conclude that playing at home might influence a football's team performance in terms of referee biases and likelihood to win, given that more games were won by the home team overall in the past. However, the attendance, the distance to the stadium, and the home advantage score cannot be used to measure for sure whether the team who is playing at home will win. The number of factors that influence a team's performance is immense and cannot be covered solely by physical conditions from the outside. Moreover, an increasing number of teams use live data-analytics at their home base during a game, which might be the reason why Leicester City went back to their usual self in the season following the phenomenal win. Finances, weather conditions, technology, stadium atmosphere and psychological factors are all factors that influence each game. Therefore, a solid conclusion cannot be made using the analysis that we conducted on the acquired data.

References

[1] Lawson, Sam. “EPL Results 1993-2018.” Kaggle, 15 July 2018,
`www.kaggle.com/thefc17/epl-results-19932018`

[2] The dataset consists of statistics for all the English Premier League matches between 2000 and 2018,
`https://www.kaggle.com/devinharia/epl-dataset#EPL%202000-2018.csv`

[3] "Set value for particular cell in pandas DataFrame using index", StackOverflow, 16 Sep. 2017,
`https://stackoverflow.com/a/13842286`

[4] "Vincenty’s Formulae", Wikipedia,
`https://en.wikipedia.org/wiki/Vincenty%27s_formulae`

[5] "Footystats: Home Advantage Score", Footystats,
`https://footystats.org/england/premier-league/home-advantage-table`

[6] "Sir Alex Ferguson retires: Reaction to Man Utd announcement", BBC, 08 May 2013,
`https://www.bbc.com/sport/football/22448168`

[7] "Data Analysis at Leicester City", OptaPro,
`https://www.optasportspro.com/case-study/leicester-city/`

[8] “Data Wrangling.” SpringerLink,
`https://link-springer-com.vu-nl.idm.oclc.org/referenceworkentry/`

[9] Datopian. “English Premier League (Football).” DataHub,
`https://www.datahub.io/sports-data/english-premier-league`

[10] England Football Results Betting Odds: Premiership Results and Betting Odds,
`https://www.football-data.co.uk/englandm.php`

[11] Football Results, Statistics and Soccer Betting Odds Data,
`www.football-data.co.uk/data.php`

[12] Mart. “International Football Results from 1872 to 2019.” Kaggle, 21 Nov. 2019,
`www.kaggle.com/martj42/international-football-results-from-1872-to-2017`

[13] Mathien, Hugo. “European Soccer Database.” Kaggle, 23 Oct. 2016,
`https://www.kaggle.com/hugomathien/soccer`

[14] “Stadiums.” Worldfootball.net, 1 Feb. 2020,
`https://www.worldfootball.net/`

[15] Tanersekmen. “Serie A League Player Analysis- Fifa 19.” Kaggle, Kaggle, 21 Jan. 2020,
`https://www.kaggle.com/tanersekmen/serie-a-league-player-analysis-fifa-19`

Appendix 1

Formula for home advantage

$$X = GoalsScoredAtHome/TotalMatchesPlayed$$

$$Y = GoalsScoredWhileAway/TotalMatchesPlayed$$

$$ScoringAdvantage = (X - Y)/(X + Y)$$

$$X = GoalsConcededAtHome/TotalMatchesPlayed$$

$$Y = GoalsConcededWhileAway/TotalMatchesPlayed$$

$$DefenceAdvantage = (Y - X)/(X + Y)$$

$$HomeAdvantage = (ScoringAdvantage + DefenceAdvantage)/2$$

Appendix 2

Figure 1: StadiumDistance vs GoalDifference distribution

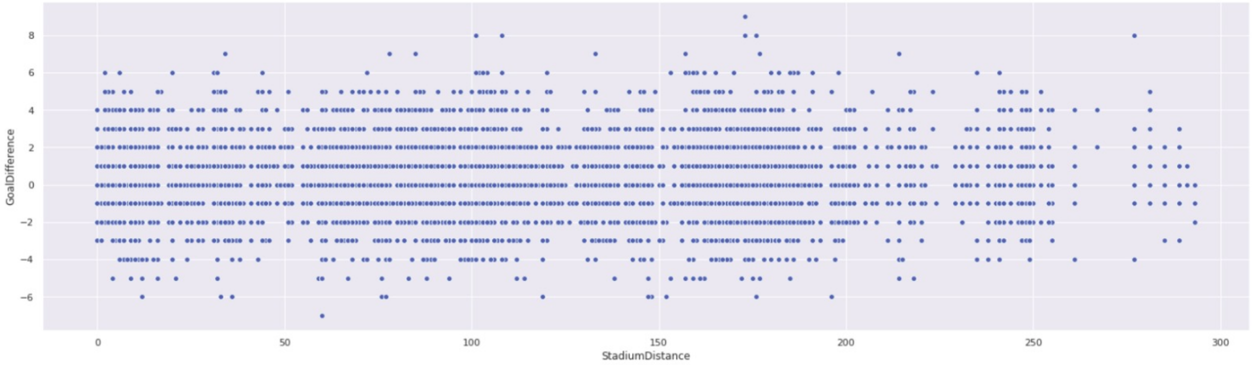
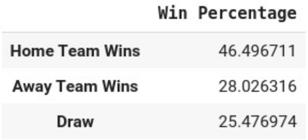


Figure 2: Win Percentage



	Win Percentage
Home Team Wins	46.496711
Away Team Wins	28.026316
Draw	25.476974

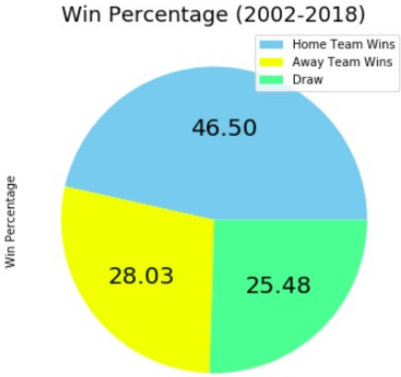


Figure 3: Home Wins, Away Wins and Draws

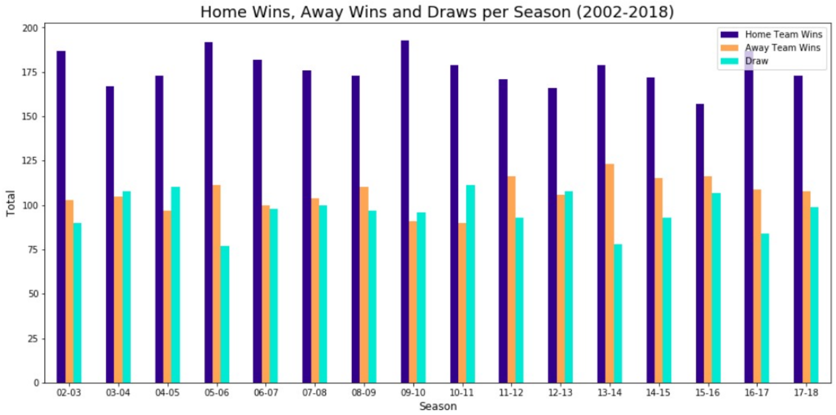


Figure 4: Scoring Margin

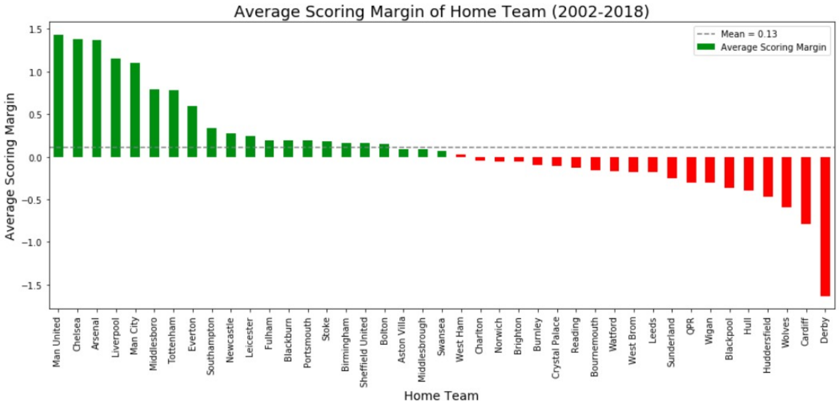


Figure 5: Foul Margin

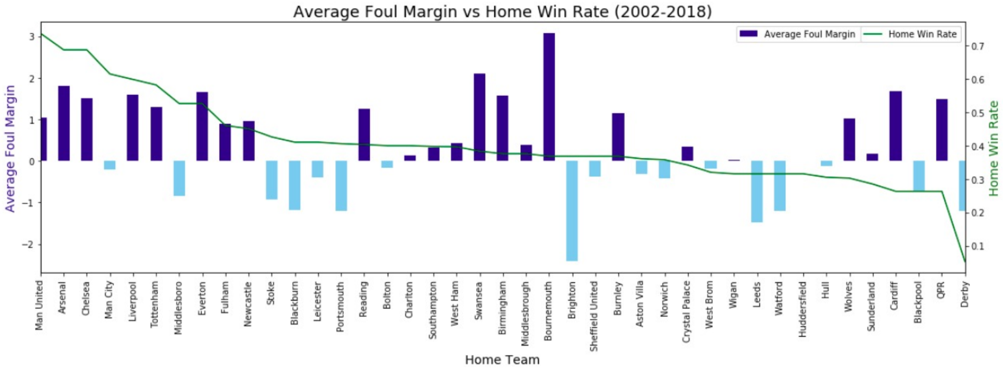


Figure 6: Yellow Card Margin

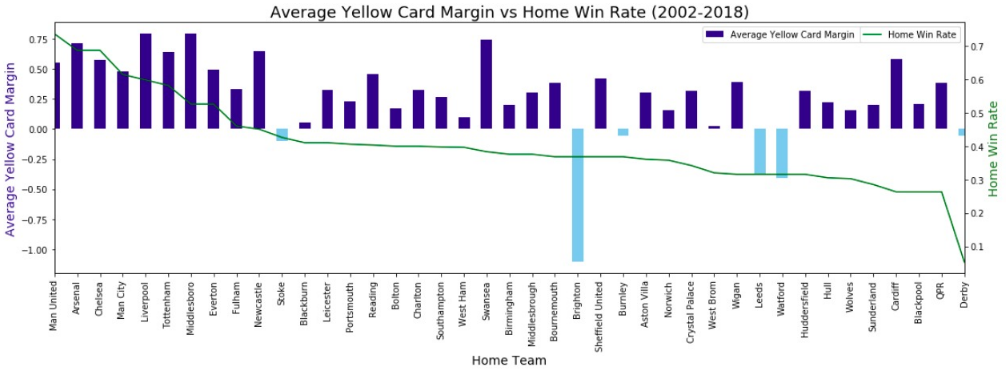


Figure 7: Red Card Margin

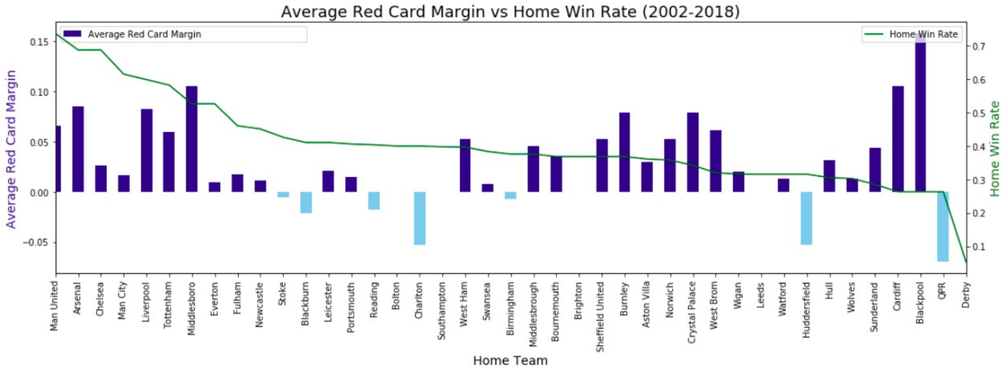


Figure 8: Referee Bias Score

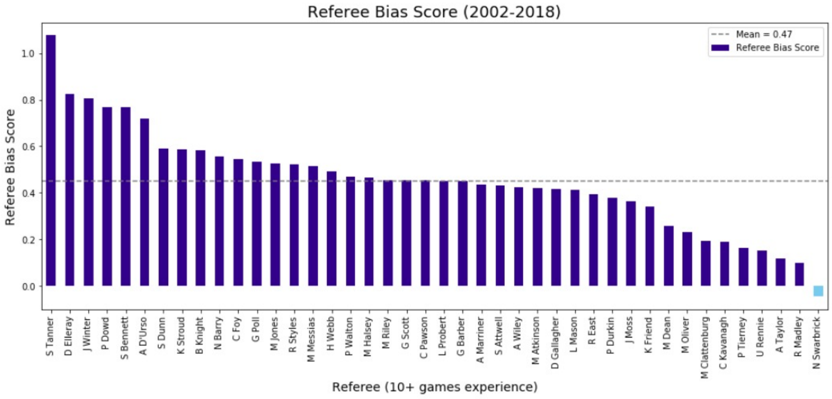


Figure 9: Ending Season

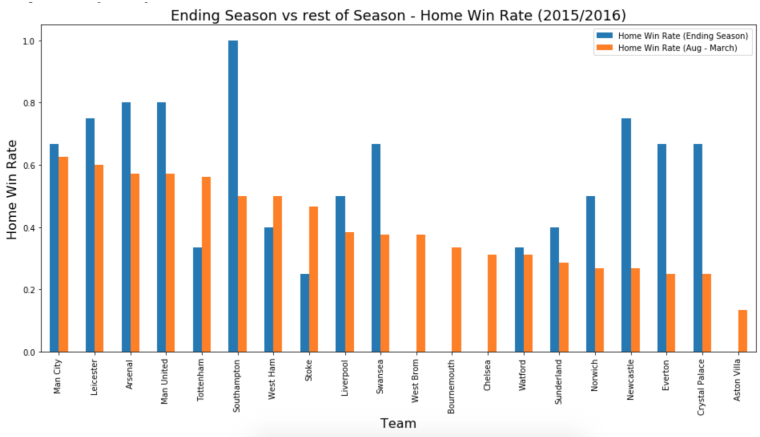
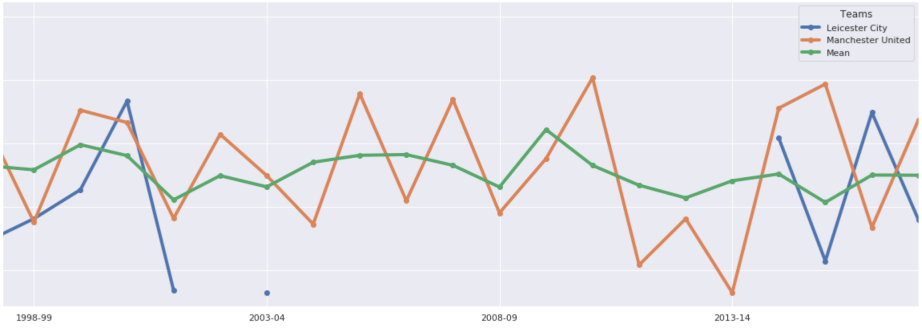


Figure 10: Home Advantage - Leicester and Manchester United


Teams
Leicester City
Manchester United
Mean