English Premier League Data Analysis

by Joseph E. Rehfus

Note: the full project, including data and code, is available at https://github.com/jrehfus1

Table of Contents

Motivation

The English Premier League (EPL) is the most-watched sports league in the world (**1**). Each season, it features 20 professional soccer clubs vying for the prestigious league title and for qualification to compete with the best amongst all of the other European clubs in continent-wide tournaments. In addition to the battle to secure glory and riches amongst the top teams, there is a perilous battle for survival between the worst clubs. The bottom three performers are relegated to a lower division at the end of the season, losing out on revenue and prestige. The EPL season is full of drama from start to finish, which is likely why its storylines attract such diverse and international viewership. As a fan myself, I decided to use match results data to answer a few exploratory questions about the EPL.

Determining the outcome of a single match

As in many competitive sports, winning soccer matches depends on scoring more often than your opponent. However, the number of points scored in a competition varies greatly depending on the sport. In this section, the final scores of EPL matches will be examined in addition to the times at which goals were scored.

Data requirement

- final scores from EPL matches
- the time at which each goal was scored for all EPL matches in a season

Data source(s)

- https://www.worldfootball.net

Pertinent script(s)

- plot_league_goals_data_v1.1.2.py

How many goals are typically scored during an EPL match?

To get a better understanding of the value of a single goal in determining the outcome of an EPL match, I plotted the distribution of total goals scored for each match of the 2019-2020 season (**Figure 1**). On average, 2.7 goals were scored per match. This value is much lower than the number of points scored in a typical American football or basketball game. Occasionally a match will produce many goals. In the 2019-2020 season, 44 games, approximately one in nine, resulted in five or more goals.

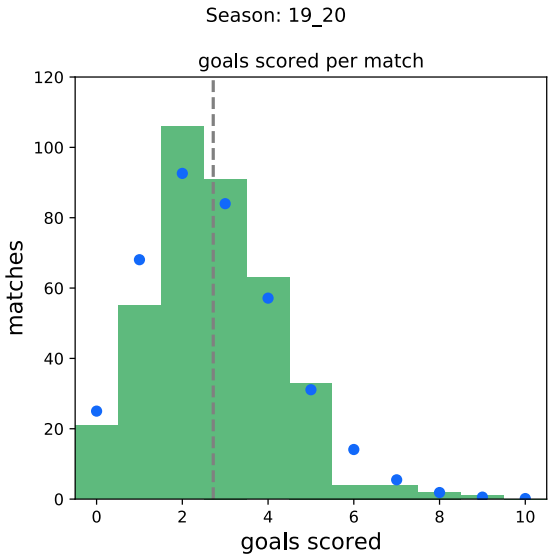

Figure 1 Histogram of the number of goals scored in matches during one EPL season. Green bars show the binned data from actual EPL games. The dashed gray line shows the average number of goals scored in a match this season (2.72). The blue points indicate the best fit Poisson probability match function to the data. ##add figure legend

What is the typical margin of victory when one team beats another?

Not all goals are equally valuable. The average difference in goals scored by either team at the end of a match was just 1.4 goals. In 55 cases during the 2019-2020 season, or ~14%, only one goal was scored: the game winner. However, the distribution of goal differences for the 2019-2020 season reveals that the majority of matches with a victor (i.e., those that did not end in a draw) were decided by more than one goal. When one team prevailed over the other, the average margin of victory was 1.91 goals. Victories achieved by more than a three-goal margin are very rare, occurring in only 26 matches, or ~7%, highlighting the competitiveness of the league.

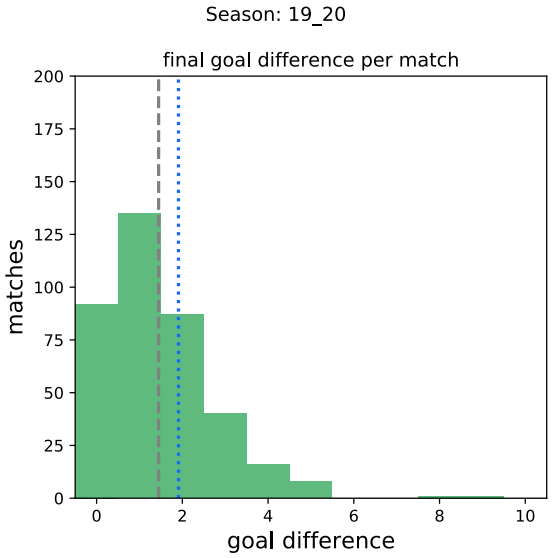

Figure 2 Histogram of the goal difference at the end of matches during one EPL season. Green bars show the binned data from actual EPL games. The dashed gray line shows the average goal difference at the end of a match this

season (1.45). The dotted blue line shows the average margin of victory, i.e. the goal difference when one team scored more goals than the other (1.91). ##add figure legend

Is home field advantage real?

Players, managers, and fans often lament the difficulty of traveling to a hostile site and winning a match. Is so-called "home field advantage" a real phenomenon? If so, what is the magnitude of its impact?

In the 2019-2020 season, 288 matches out of 380 (75.8%) produced a winner and a loser. If home field advantage made no difference, then we would expect that half of the matches with a winner (144) were won by the home team, while the other half were won by the away team. Instead, we find that 172 matches were won by the home team, and only 116 were won by the away team. A chi-square test of the null hypothesis, that the home and away teams are equally likely to win a given match, produced a p-value of 0.00097, indicating that it is very unlikely that both teams have an equal chance of winning. Instead, the home team is the more likely victor, demonstrating that there really is an advantage to playing at home.

This result is further reflected in the histogram of home team goal difference (defined as home team score – away team score) shown below (**Figure 3**). Clearly, the home team outscores the away team more often. Instead of 0.0, the average home team goal difference is 0.31 goals, which indicates that the home team averages slightly more goals than the away team on a per-match basis.

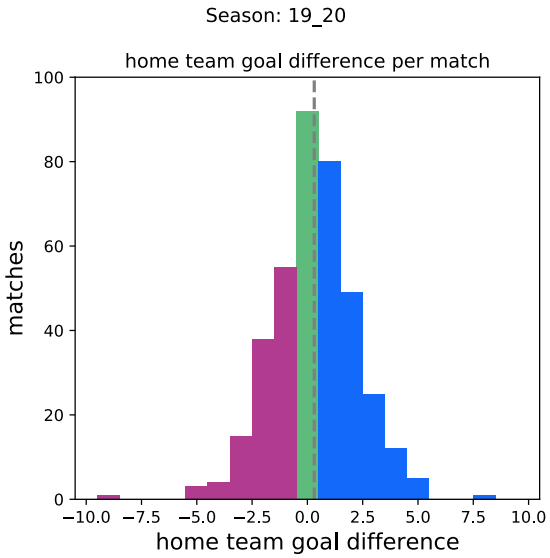

Figure 3 Histogram of the goal difference in a match from the perspective of the home team. Positive values (blue) indicate that the home team won, negative values (purple) indicate the home team lost, and a value of zero (green) indicates that the home team drew. The dashed gray line shows the average home team goal difference (0.31). ##add figure legend

When are goals scored during a match?

A soccer match consists of two 45-minute halves. The final minute of each half can be extended at the referee's discretion to make up for stoppages of play that have occurred. Goals can be scored at any point from kickoff to the final whistle, but are there some periods that, on aggregate, produce more goals than others? To find out, the times at which every goal of the 2019-2020 EPL season was scored were binned and used to produce a histogram (**Figure 4**).

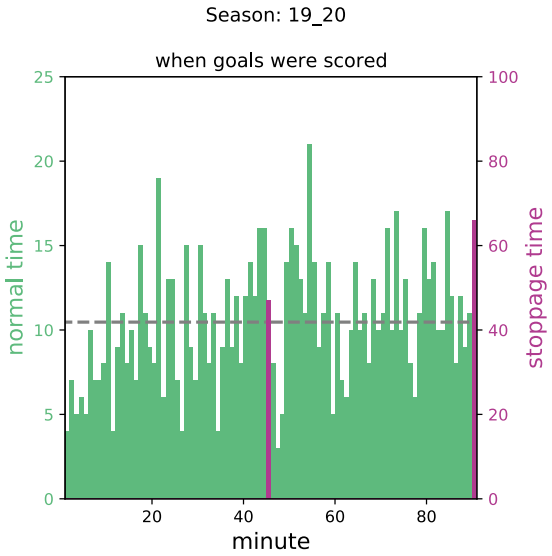

Figure 4 Histogram of the time at which goals were scored during matches in one EPL season. Both sets of bars show the binned data from actual EPL games. The green bars indicate goals scored during normal time (left axis) while the purple bars show goals scored during stoppage time at the end of either half (right axis). The dashed gray line shows the average number of goals scored during cumulative normal time minute (10.47). ##add figure legend

Due to the additional stoppage time, goals are scored much more frequently in the 45th and 90th minutes than during any normal time minute. On average, ~10.5 goals were scored per aggregate minute of normal time for the entire season. A Chi-square test reveals that goal times were not evenly distributed this season ($p = 0.016$), perhaps due to the dearth of goals scored in the first few minutes of each half. Such a result seems plausible, as each half begins with a kickoff, the only time at which both teams restricted to their own side of the pitch. Thus, the action in all matches is most closely synchronized at the very beginning of a half. Once a half is underway, the game is not stopped again for any scheduled break until halftime or the full-time whistle, and the action in each individual match proceeds asynchronously.

Determining league position

EPL teams are assigned positions from 1st place to 20th based on their performance throughout the entire season. There are 38 weeks in an EPL season, during which each team plays each other team twice: once at home and once away. Every team begins the season with no points. League positions are determined based primarily on league points accrued. Ties in league points are broken by goal difference. If multiple teams are still tied, the third and final tiebreaker is total goals scored. I decided to explore how each of these three metrics develop week-by-week throughout an entire EPL season.

Data requirement

- final scores from EPL matches

Data source(s)

- www.worldfootball.net

Pertinent script(s)

- calculate_league_points_data_v1.1.2
- plot_league_points_data_v1.1.2
- calculate_average_league_points_data_v1.1.1
- plot_average_league_points_data_v1.1.1

How does league position change throughout a season?

Ultimately, the most important outcome that a team is interested in is league position, the culmination of league points, goal differential, and total number of goals scored. Each of these parameters were determined on a week-by-week basis for each EPL team using the final scores from their matches. The average weekly league position of teams that ended up finishing the season 1st, 2nd, 3rd, etc. down to 20th throughout the past five EPL seasons are plotted below (**Figure 5**).

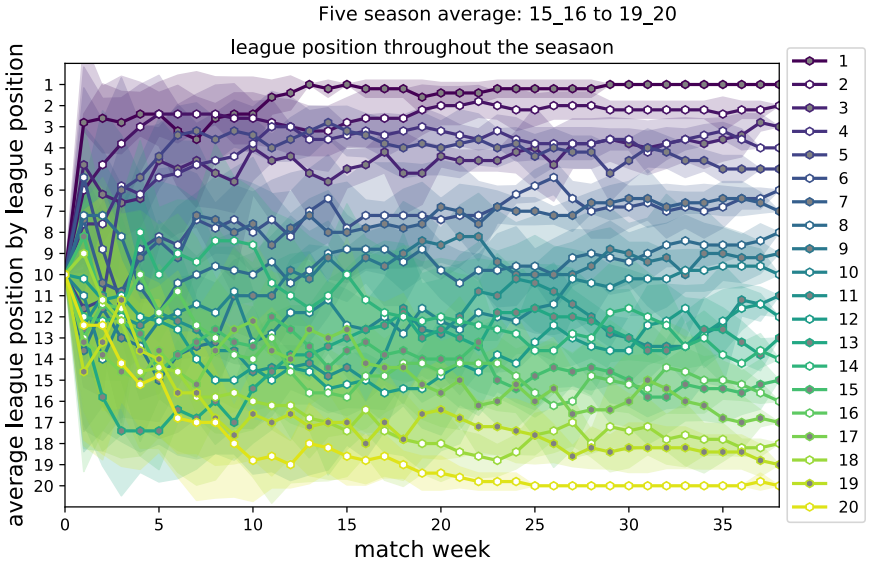

Figure 5 Average weekly league position of EPL teams that ended the season in the 1st, 2nd, 3rd, etc. league position according to the legend at right, over the past five seasons. Shading indicates ± standard deviation.

Interestingly, the teams that have finished the season in first place over this time period stayed in first place for the final 10 weeks of their seasons. Sadly, the fate of the worst team in the league has historically been sealed even earlier. The team finishing last has been in last place for the final 14 weeks of each season, except once when that team moved up into second to last place in the penultimate week, only to fall back down on the last day of the season. The final league positions for every team that did not finish first or last are much more difficult to predict, even close to the end of the season.

The week-to-week variability in the aggregated league position data suggests that teams can change position dramatically throughout the course of a season. To take a closer look at the league positions of individual teams, the data from only the 2019-2020 EPL season is plotted below (**Figure 6**).

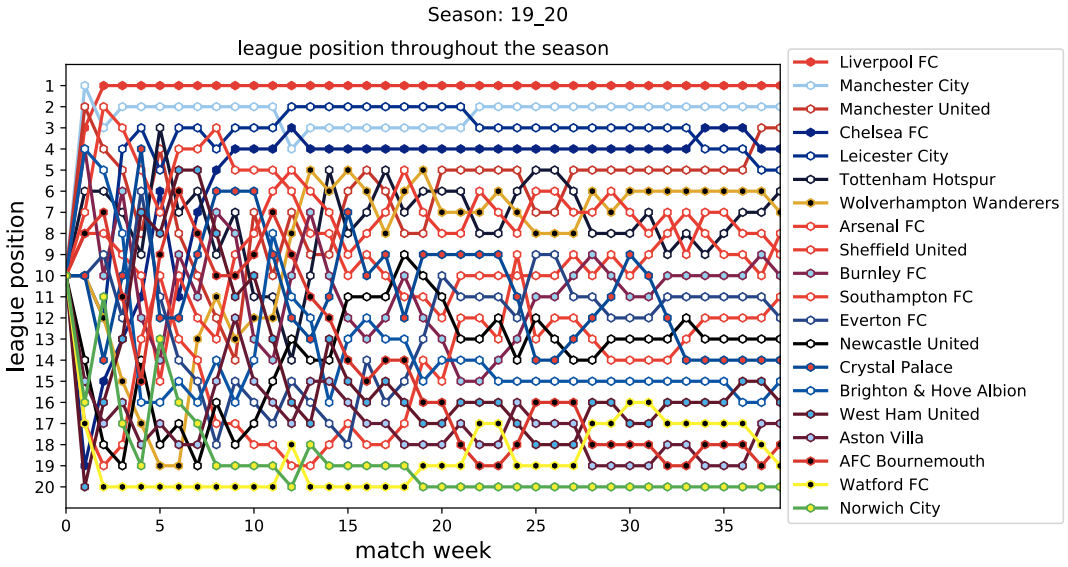

Figure 6 League position for each EPL team as a function of match week. League position is determined first by league points, then goal difference, then total goals scored.

As expected, based on the aggregated data in **Figure 5**, some teams (Crystal Palace and Southampton FC, for example) experienced fluctuations in league position that spanned multiple places over the course of just a few weeks. On the other hand, teams like Liverpool FC, Chelsea FC, and Norwich City were much steadier. The high variability in league position for many teams is likely a result of the fact that league position depends on how the teams in nearby positions perform. For example, if several teams are only separated by a few league points, they can exchange places in just one weekend. The league positions can be especially variable over a short time period when teams are only separated based on goal differential or goals scored.

Conclusions

Examining the outcome of individual matches over a single season has revealed that EPL games are generally very closely contested. Only a few goals are scored per match, and this result is reflected in the goal difference data. Goals are scored throughout matches, although the beginnings of each half seem to have fewer goals than the rest of the game. If you need to make

a snack trip, that is the time to do it! Interestingly, home field advantage is real in the EPL, although the benefit seems to be modest (victory is certainly not assured for home teams). Lastly, the league position for the very best and worst teams is relatively stable compared to those in the middle of the pack. It seems that there really are exceptionally good and bad performers.

Open questions

This work also opens the door to some exciting new areas of analysis. For example, what factors have the largest impact on home field advantage? Are goals scored at the same rate in stoppage time as normal time, or do time-wasting tactics work? Furthermore, can EPL seasons be ranked based on some sort of "excitement" metric that takes into account league position changes throughout the season? Can this metric be predicted? I aim to address these questions, and more, in future reports.

References

1) Pilger, S. (2014, February 6). *Why the Premier League is the Most Powerful League in the World*. Retrieved from https://bleacherreport.com/articles/1948434-why-the-premier-league-is-the-most-powerful-league-in-the-world