Soccer Data Analysis: Report

by Joseph E. Rehfus

Table of Contents

Motivation

The English Premier League (EPL) is the most-watched sports league in the world (##REF). Each season, it features 20 professional soccer clubs vying for the prestigious league title and for qualification to compete with the best 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, 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.

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)

- www.transfermarkt.us

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 value of goals in 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. If you enjoy scoring, the EPL might be the league for you!

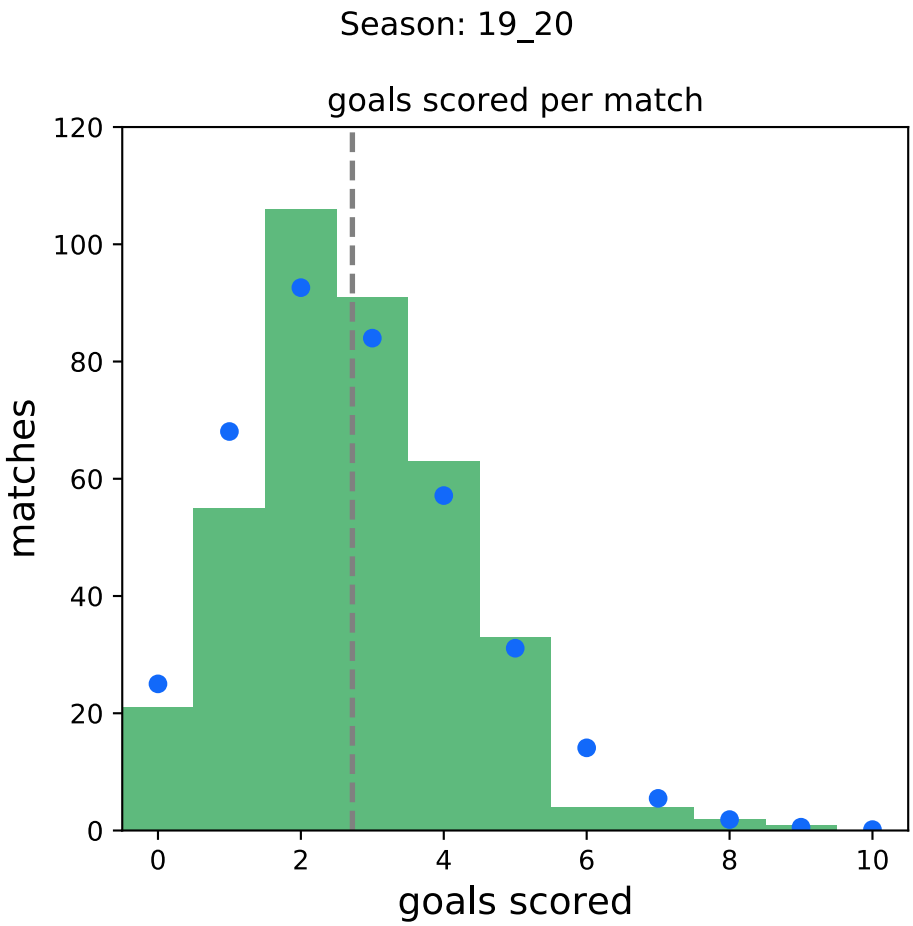

Figure 1 Histogram of the number of goals scored in matches during one EPL season. Green bars show the binned data from actual EPL games. The dashed gray line shows the average number of goals scored in a match this season (2.72). The blue points indicate the best fit Poisson probability match function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

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. 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 (ie, 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: only 26 matches, or ~7%, would be considered blowouts.

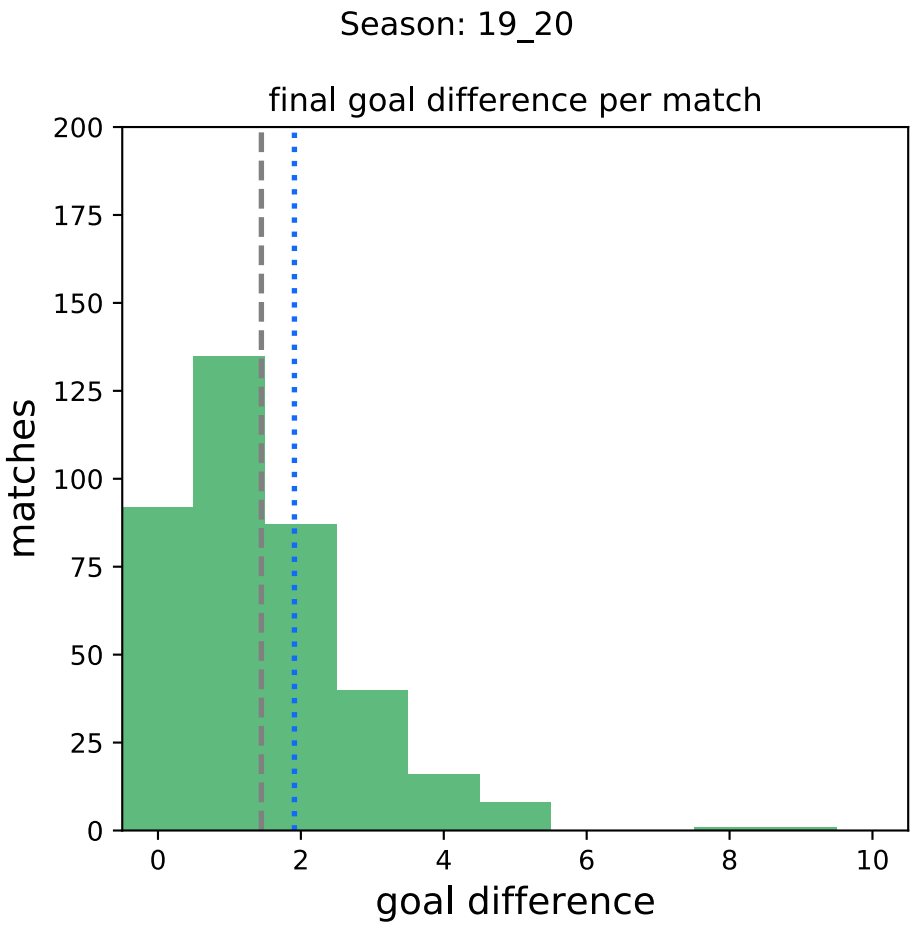

Figure 2 Histogram of the goal difference at the end of matches during one EPL season. Green bars show the binned data from actual EPL games. The dashed gray line shows the average goal difference at the end of a match this season (1.45). The dotted blue line shows the average margin of victory, i.e. the goal difference when one team scored more goals than the other (1.91).

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 0.0097, 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 zero, the average home team goal difference is 0.31, which

indicates that the home team averages slightly more goals than the away team on a per match basis.

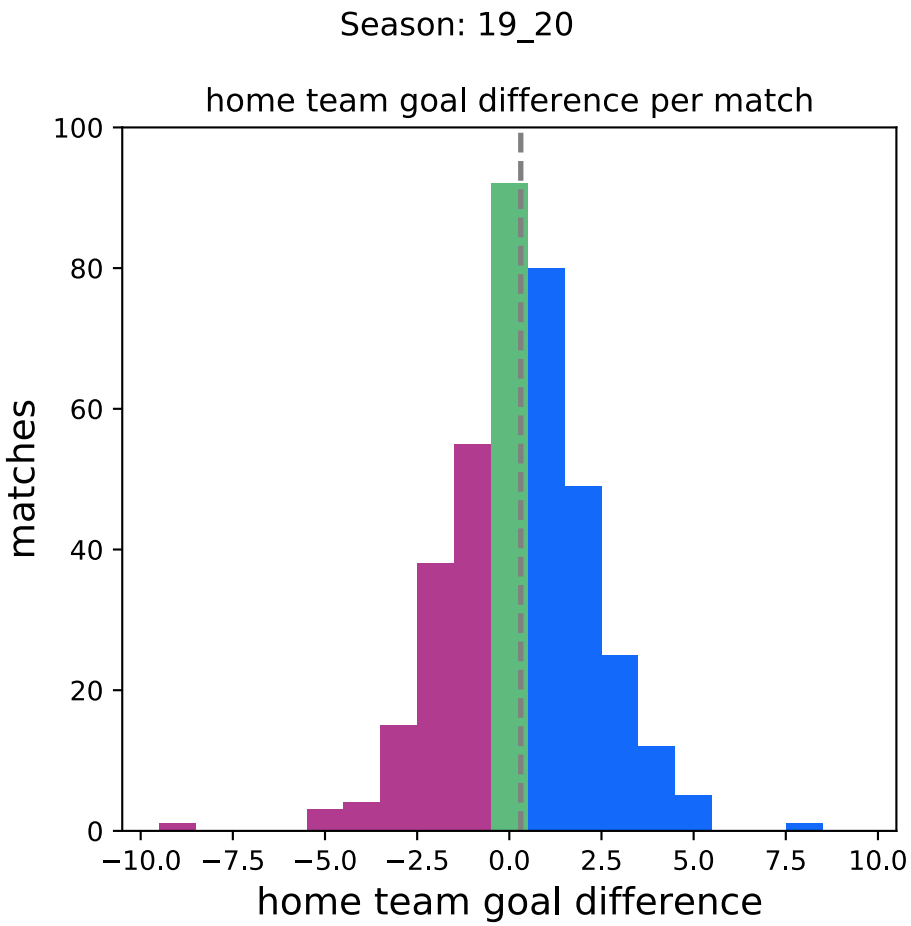

Figure 3 Histogram of the goal difference in a match from the perspective of the home team. Positive values (blue) indicate that the home team won, negative values (purple) indicate the home team lost, and a value of zero (green) indicates that the home team drew. The dashed gray line shows the average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

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**). 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, with 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.

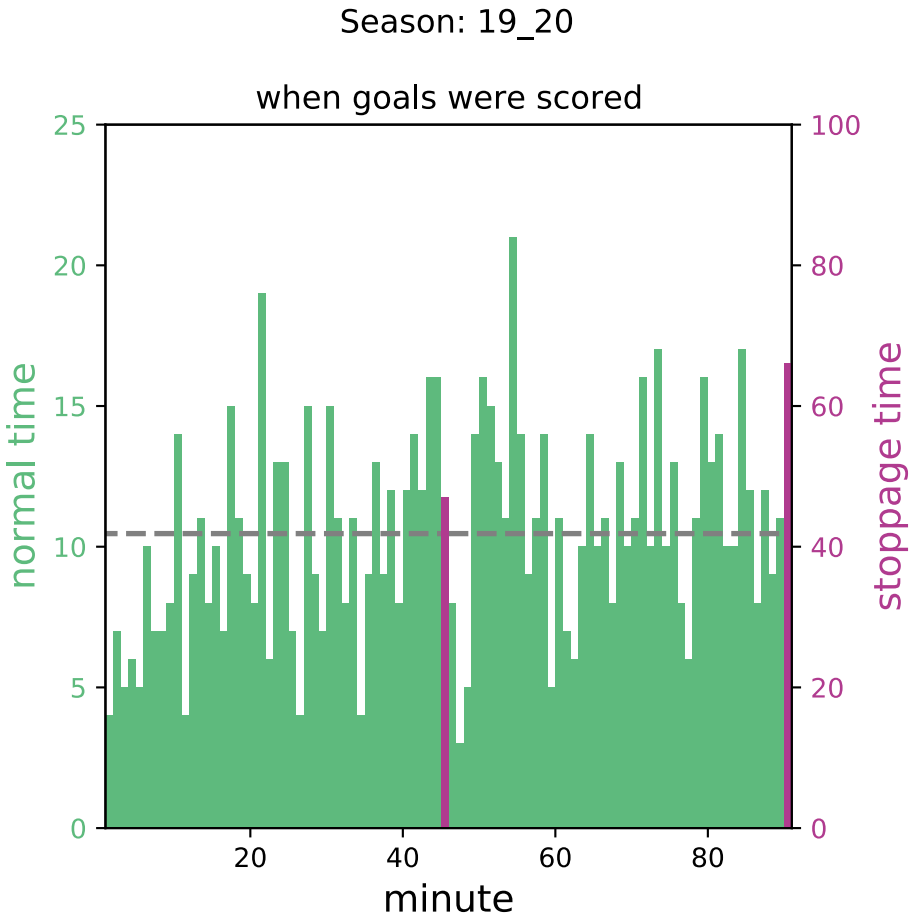

Figure 4 Histogram of the time at which goals were scored during matches in one EPL season. Both sets of bars show the binned data from actual EPL games. The green bars indicate goals scored during normal time (left axis) while the purple bars show goals scored during stoppage time at the end of either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

Determining League Position

EPL teams are assigned positions from 1st place to 20th based on their performance throughout the entire season. 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 cumulation of league points, goal differential, and total number of goals scored. 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**).

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. 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.

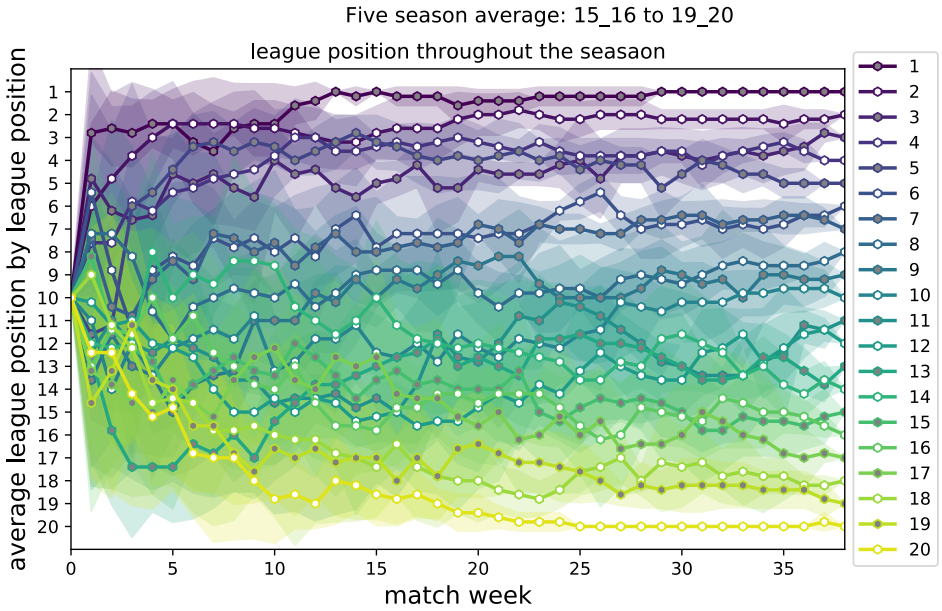

Figure 5 Average weekly league position of EPL teams that ended the season in the 1st, 2nd, 3rd, etc. league position according to the legend at right, over the past five seasons. Shading indicates ± standard deviation.

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**). As expected, some teams, such as Crystal Palace and Southampton FC, 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.

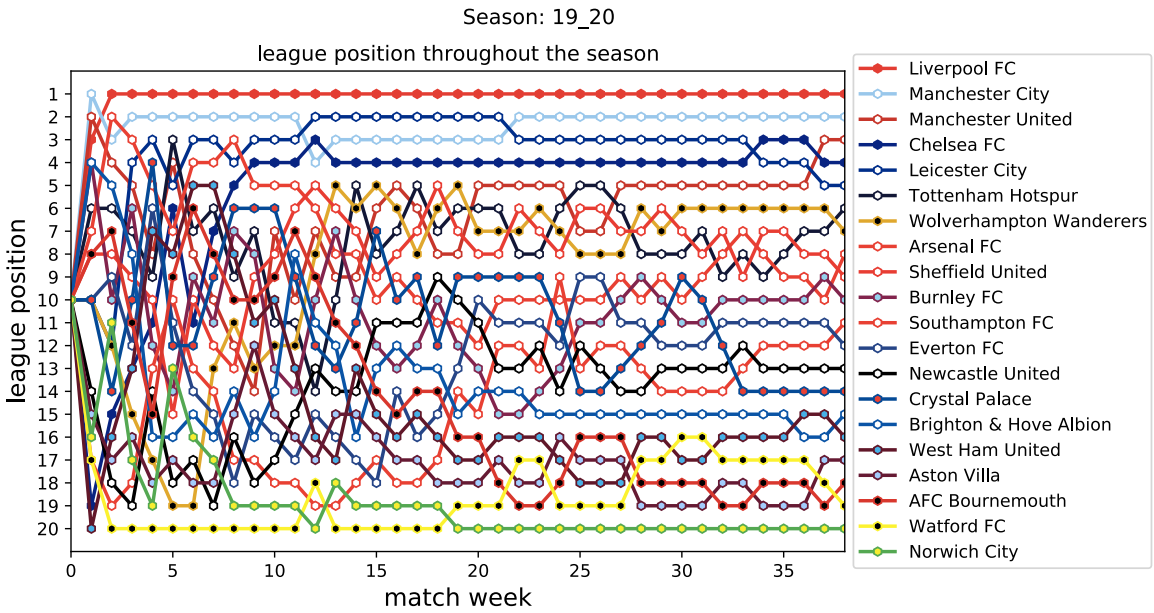

Figure 6 League position for each EPL team as a function of match week. League position is determined first by league points, then goal difference, then total goals scored.

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.

When are league points won throughout a season?

League points are the primary determinant of league position. Teams are awarded three league points for a win, one for a tie, and zero for a loss. Thus, the maximum possible number of league points that can be collected in a 38-match season is 114. For reference, the highest point total ever achieved in the modern EPL format (begun in the 1992-1993 season) is 100 points, a record set by Manchester City in the 2017-2018 season. The weekly average league points for teams grouped based on final league position over the past five seasons are plotted below (**Figure 7**).

As with the league position data presented above, there is a lot of variability in weekly league points based on final league position. For example, the team that has finished in first place has historically accrued 94±7 points by the end of the season while the last place team had 22±5. These totals correspond to ~82% and ~20% of the total possible points available, respectively.

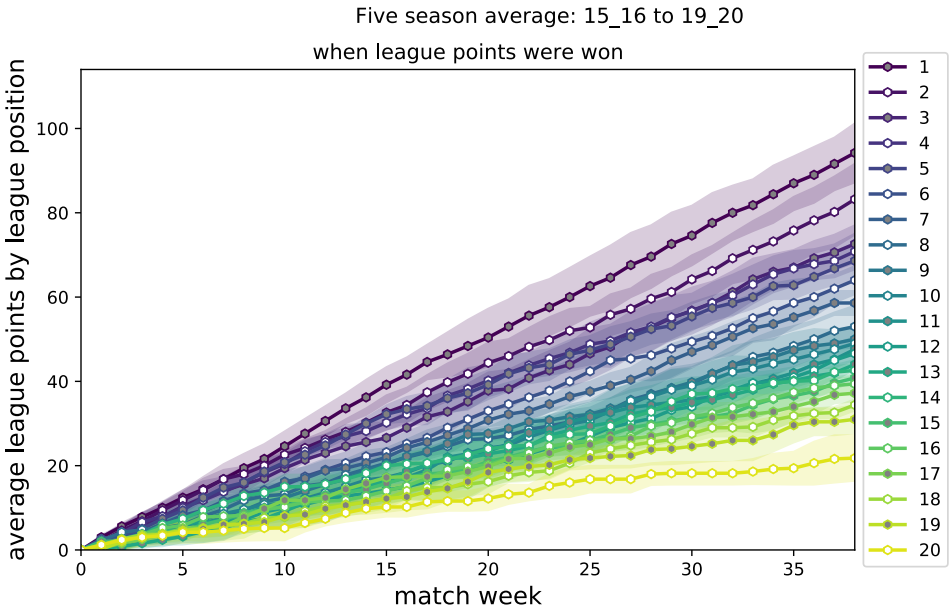

Figure 7 Average weekly league points of EPL teams that ended the season in 1st, 2nd, 3rd, etc. league position according to the legend at right, over the past five seasons. Shading indicates ± standard deviation.

The league points for all 20 EPL teams are plotted at each match week for the 2019-2020 season (**Figure 8**). Aside from the two best performers, Liverpool FC and Manchester City FC, and the worst performer, Norwich City, league points were very close amongst most teams. Thus, stretches of good or bad performance for several games in a row is shown to have a large impact on the outcome of a season for teams only separated by a few league points.

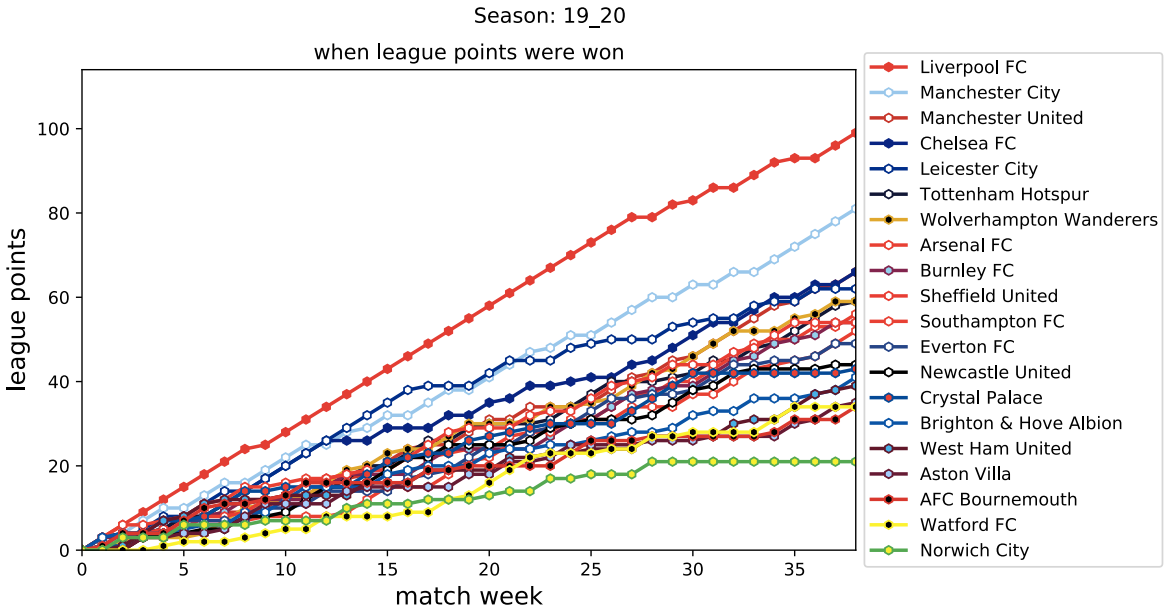

Figure 8 League points for each EPL team as accrued by week. League points are the primary metric by which league position is determined at the end of the season. A victory is worth three league points, a draw is worth one, and a loss is worth zero. The maximum possible number of league points that a team can win is 114.

How does goal differential develop throughout a season?

If multiple teams have obtained the same number of league points, the first tie breaker for league position is goal difference, defined as the number of goals scored by a club minus the number of goals they have conceded.

The weekly average goal difference for teams grouped based on their final league positions over the past five seasons is plotted below (**Figure 9**). Unsurprisingly, the teams that do well in terms of league position also do very well in terms goal difference. Teams that finish in the top half of the table typically end the season with a positive goal difference, while the opposite is true for teams at the bottom of the table. However, the magnitude of the goal difference is larger for teams that end up at the top of the table than for those that finish towards the bottom. For example, the team that goes on to win the league has scored 94±7.

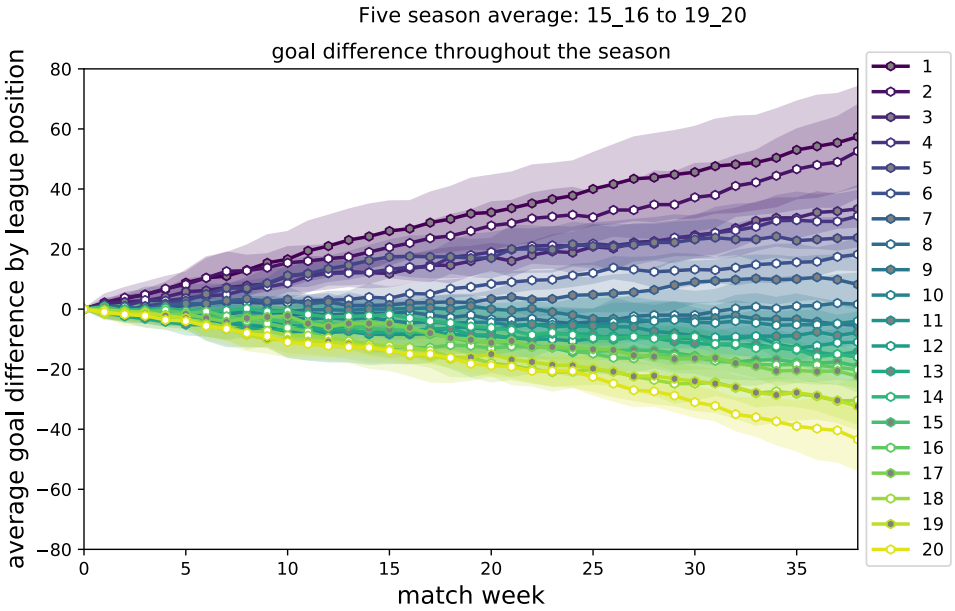

Figure 9 Average weekly goal difference of EPL teams that ended the season in 1st, 2nd, 3rd, etc. league position according to the legend at right, over the past five seasons. Shading indicates ± standard deviation.

Plotting goal difference as it developed throughout the 2019-2020 season, it is clear that most teams scored and conceded a roughly equal number of goals (**Figure 10**). This observation is consistent with the result reported above that matches are typically decided by only a few goals.

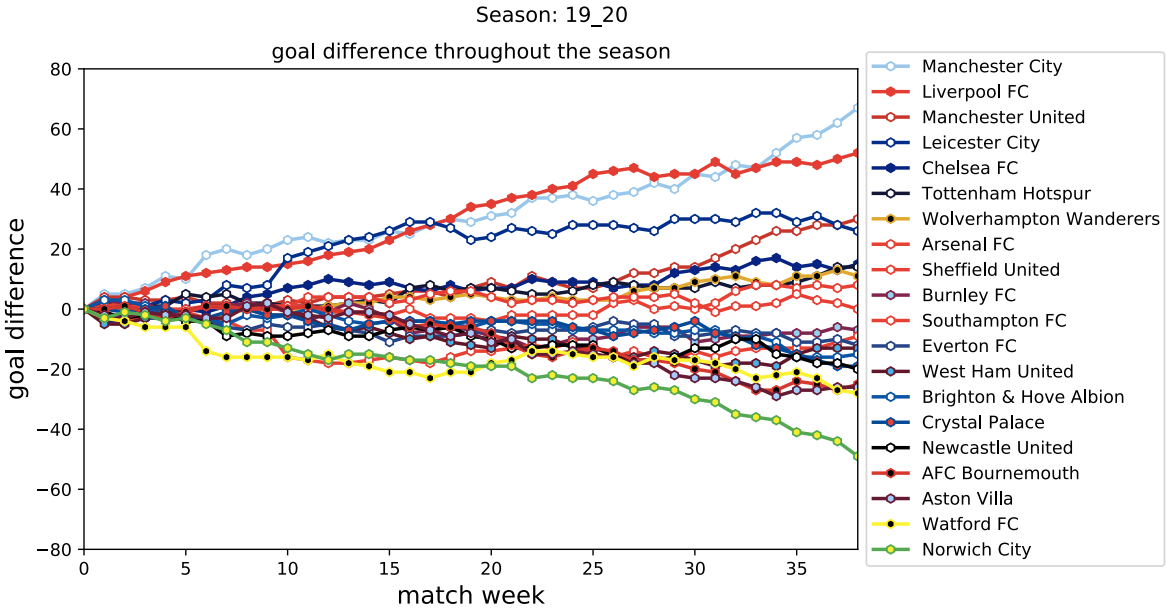

Figure 10 Goal difference for each EPL team as a function of match week. Goal difference is the secondary method by which league position is determined at the end of the season. Victories produce positive goal difference, draws will not change goal difference, and losses produce negative goal difference.

How do total goals scored develop throughout a season?

The final tiebreaker for deciding league position is the total number of goals scored. The weekly average cumulative goals total for teams grouped based on their final league positions over the past five seasons is plotted below (**Figure 11**). As expected, teams that score more goals from week to week end up finishing higher in the league standings.

#what do the slopes and the magnitudes of the values tell us? Anything (un)expected?

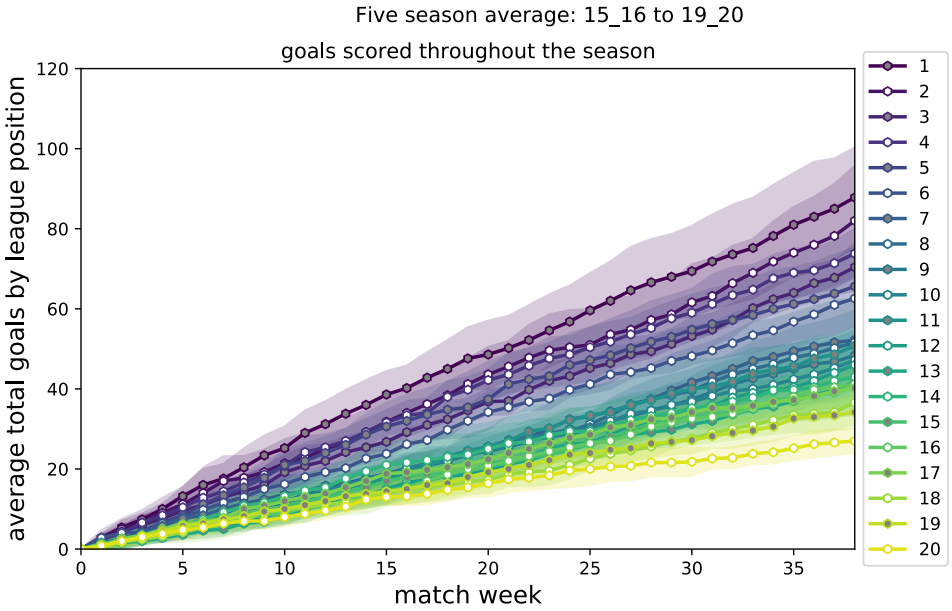

Figure 11 Average weekly total number of goals scored of EPL teams that ended the season in 1st, 2nd, 3rd, etc. league position according to the legend at right, over the past five seasons. Shading indicates ± standard deviation.

The final goal tally of each EPL club in the 2019-2020 season is plotted below by match week (**Figure 12**). Interestingly, Manchester City scored many more goals than Liverpool FC, although they finished second in terms of league position. It should not be surprising to see that scoring timely goals and seeing out games such that the maximum number of league points are secured is more vital to finishing higher up in the league than simply scoring lots of goals.

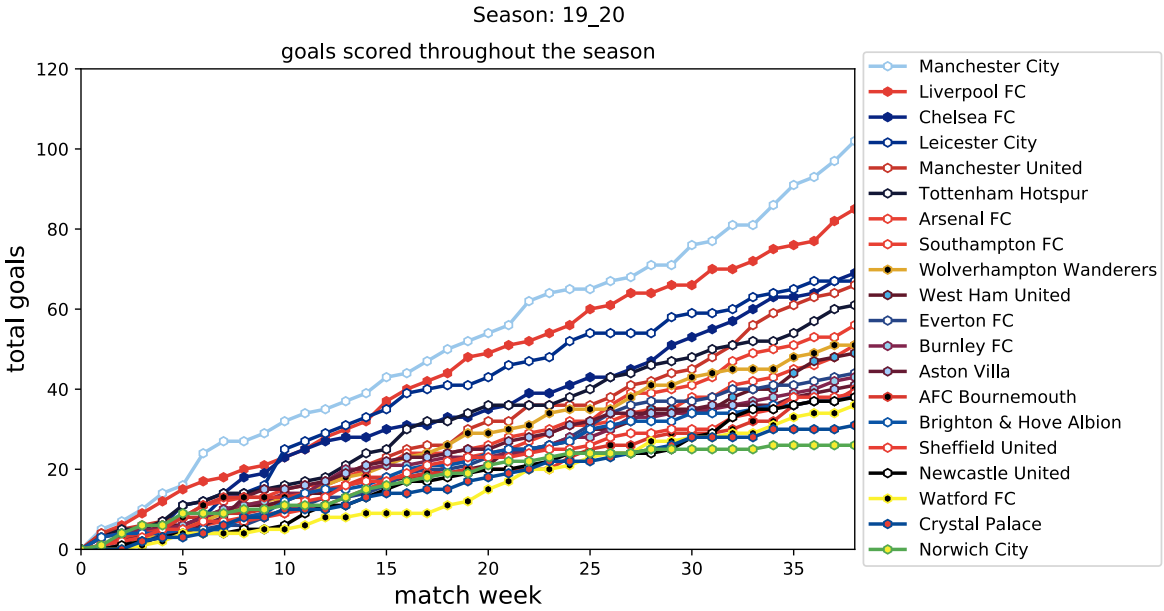

Figure 12 Total number of goals scored for each EPL team as a function of match week. Total goals scored is the tertiary, and final, method by which league position is determined at the end of the season.

Conclusions

So far, the way that goals are scored in individual matches and the components of final league position as a function of match week have been explored. # remind people of the insights you generated.

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?