English Premier League: Exploratory Data Analysis Report

Khoi Van

September 2, 2025

1 Background and Context

The English Premier League (EPL) provides an unusually rich, consistent record for examining scoring patterns and home–away dynamics. In its modern format, 20 clubs play each other home and away (38 matches per club), producing 380 fixtures per season—a volume that supports season-to-season comparisons of averages, distributions, and outcome rates. The league typically runs from August to May, with fixture design explicitly balancing home/away ties across clubs, which is ideal for the kinds of seasonal summaries and contrasts this project presents (Premier League, 2025).

A central theme in soccer analytics is *home advantage*, the tendency for home teams to outperform away teams. This effect is documented across decades of research, and recent "ghost-game" natural experiments during COVID-19 strengthened evidence that crowd presence is a meaningful driver: when stadiums were empty, measured home advantage generally fell. Systematic and cross-league analyses report noticeable reductions of the home edge in those conditions, highlighting how contextual factors can shape outcomes beyond team quality alone (Almeida & Leite, 2021; Leitner et al., 2022; McCarrick et al., 2021).

This project turns raw match results into a question-driven narrative by asking: how average goals per match change by season; what the distribution of total goals looks like; how the proportions of home/away/draw results vary over time; whether a home scoring edge persists when comparing average home vs. away goals (and their difference); and whether these advantages differ between top-five and bottom-five clubs. Framing the study around these five questions ensures that each plot contributes evidence toward a specific claim that tests whether impressions from a few memorable games match the league-wide evidence. Moreover, this analysis is situated in real-world context where competition rules and technologies evolve. For example, Video Assistant Referees (VAR) were introduced in the Premier League from the 2019-20 season (Premier League, 2019), a change that could alter aspects of officiating and may coincide with shifts in measured outcome or scoring patterns. While the present work is descriptive rather than causal, acknowledging such milestones clarifies why certain time-series features or breakpoints might appear and sets up sensible extensions for future modeling.

2 Data and Methodology

To build the dataset for this analysis, we gather all Premier League matches from the 2010-11 to 2024-25 seasons via Football-Data.co.uk (n.d.), add a `Season` variable to each match as a label to summarize trends over time, and then concatenate them into a single dataset. The core

variables used are the number of home and away goals in each match and the final match result (home win, away win, or draw).

To see how average goals per match change by season, we compute the average number of goals per match by season and use a line plot to visualize the fluctuations and also compare the figures with the overall average. As for the distribution of total goals in a match, we use a bar plot to show typical scorelines and spot outliers. We then calculate the probability of home win, away win and draw matches for each season and create plots to compare the probability of these three scenarios together and their trends over years relative to their corresponding overall average. We then compare average home and away goals and plot their difference over seasons to see whether a home scoring edge happens, and compare the home advantages between top-tier and bottom-tier teams by merging match data with teams' standing data to filter matches by team standings and home/away states of the teams so that we can evaluate win probability for each scenario.

3 Results

3.1 Average goals per match by season

From Figure 1, we can see that in the early 2010s, the number of goals per match stays close to the overall average, dips to a trough in the 2014-15 season, and then oscillates near the mean until a clear upswing after the COVID-19 pandemic. The jump in season 2023-24 was striking, well above the overall average, and is followed by a partial reversion to the mean in the following season. One plausible reason for this shocking increase is that there is more added time for matches in this season, leading to more goals being scored and higher goals per match on average. After the VAR technology was introduced in 2019 (Premier League, 2019), the average number of goals sees a noticeable decrease, possibly due to the fact that VAR was stricter in deciding goals and led to fewer goals being counted, but then soared in the following seasons as we have mentioned above.

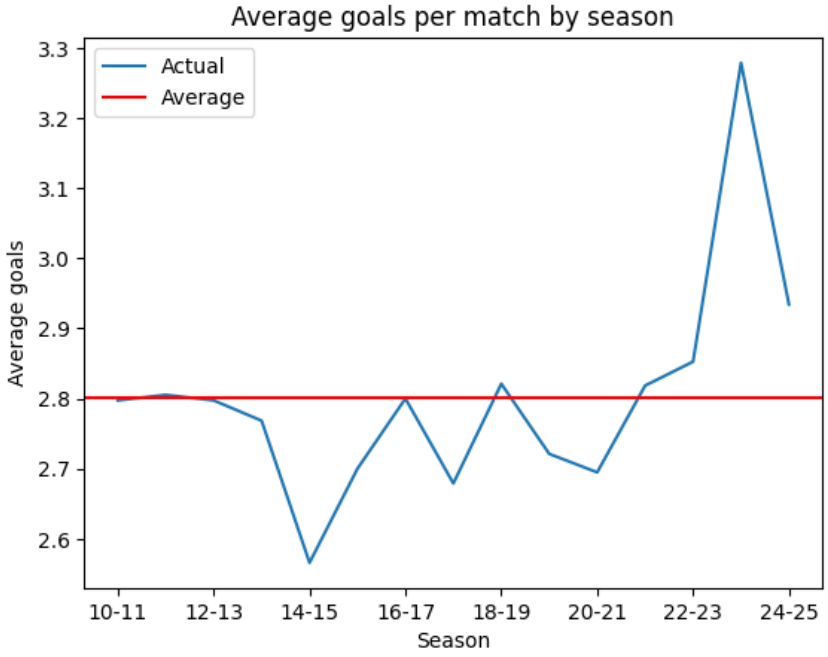

Figure 1: Average goals per match by season.

3.2 Matches distribution by total number of goals

Figure 2 illustrates a right-skewed histogram of match distribution by number of goals. Most matches cluster around 2-3 total goals, with far fewer matches ending in a 0-0 result and a quickly thinning right tail for matches with 5 goals or more. That shape says that typical EPL scorelines live in the range between 2 and 3 goals (for example, 1-1, 2-0, 2-1, 3-0), while matches with 5 goals or more are rare. This pattern is exactly what classic soccer-scoring models predict: goals arriving as approximately Poisson events yields a concentrated center with a long, light tail, which is consistent with frameworks done by Dixon and Coles (1997) that have been used for decades to model match scores. Combined with the average goals per match plot, this bar plot shows that a higher number of goals on average, such as in season 2023-24, can be attributed to either more mid-range scoring matches with 2-4 goals, or from a fatter right tail of the match distribution by number of goals plot, where a slight increase in the number of matches with 5 goals or more can have greater impact on the average.

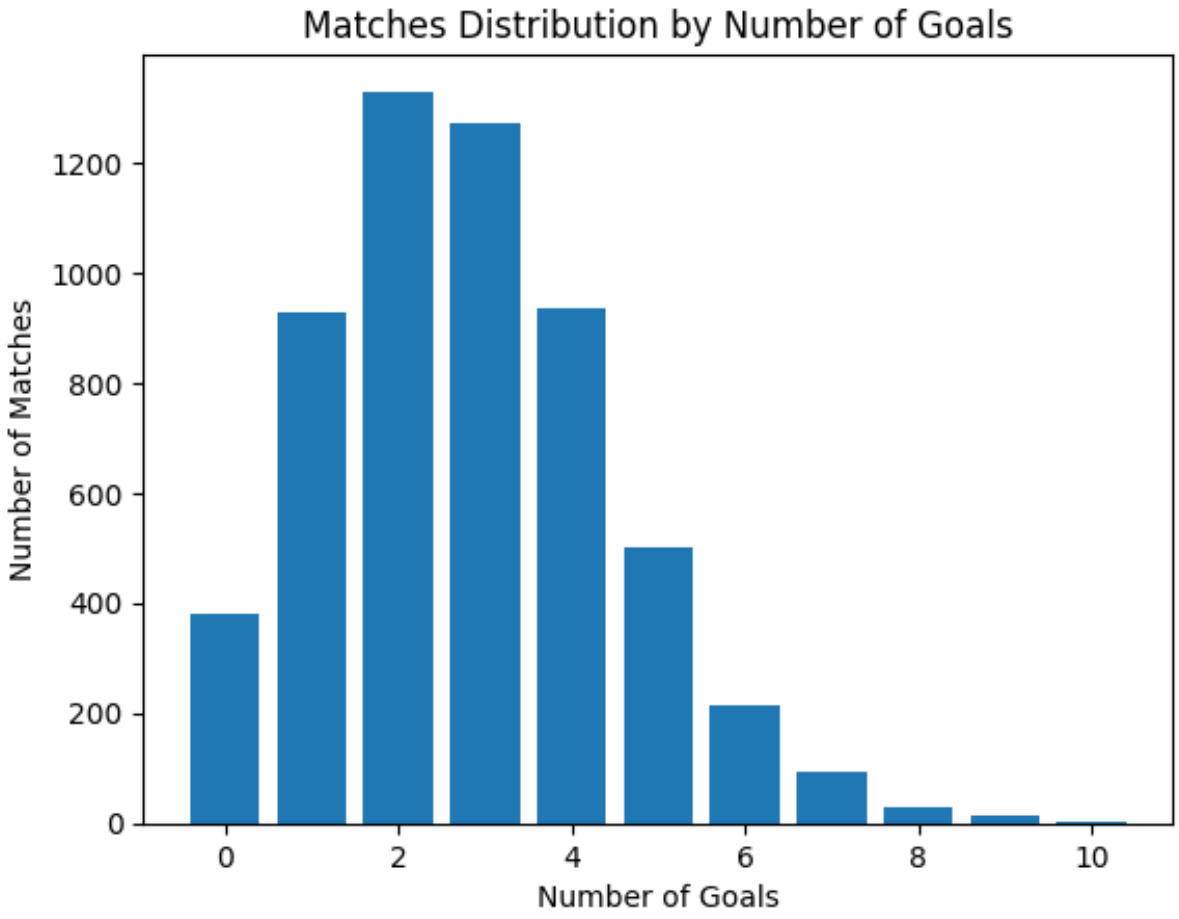

Figure 2: Distribution of total goals per match.

3.3 Home, away, and draw percentages by season

From Figure 3, we see that home wins percentage hover around an approximate 45% baseline for much of the sample, with routine ups and downs. Two seasons stand out, which are a noticeable dip in season 2020-21, when the percentage falls much below the overall average, and a rebound by season 2022-23, before easing again afterwards. Comparing with the away win and draw percentages, the home win one usually has the highest probability out of the three outcomes,

but not always. The big fall in season 2020-21 lines up perfectly with the "ghost-game" period that Leitner et al. (2022) documents and finds out that playing without crowds reduced home advantage, which is reflected by the plot showing a great fall in home wins after COVID-19 break out, followed by a partial reversion when fans returned in later seasons.

In Figure 4, away wins typically sit near the approximate 31% overall average line, but spikes in the season 2020-21, which perfectly mirrors the dim in home win probability, then settles back toward its longer-run level in the subsequent seasons. In the same way as the home win probability shock in season 2020-21, the spike in away win percentage in season 2020-21 is consistent with Leitner et al. (2022)'s findings that illustrate higher away success in fan-free stadiums, which even reversed the home edge when we compare the home and away win trends if we look at Figure 6.

Figure 5 shows that draws wildly fluctuate around the overall average of approximately 24%, which is less than home win and away win probabilities if we compare them together in Figure 6, with clear dips followed by small rebounds back to the overall average. This huge variability in draw percentage helps balance the total probability when home or away win rates swing. For example, in season 2020-21, when the away wins surged, draws stayed relative lower than average, which is consistent with a temporary shift toward more decisive results in that season.

When all these trends are plotted together in Figure 6, the home wins line is usually on top, away wins below, and draw line being the lowest, which is a typical home-edge ordering. The only exception is the season right after COVID-19 outbreak (2020-21), when the away line overtakes the home line likely due to matches with no fans coming to the stadiums, making the reversal immediately visible. In later seasons, the ordering goes back to normal.

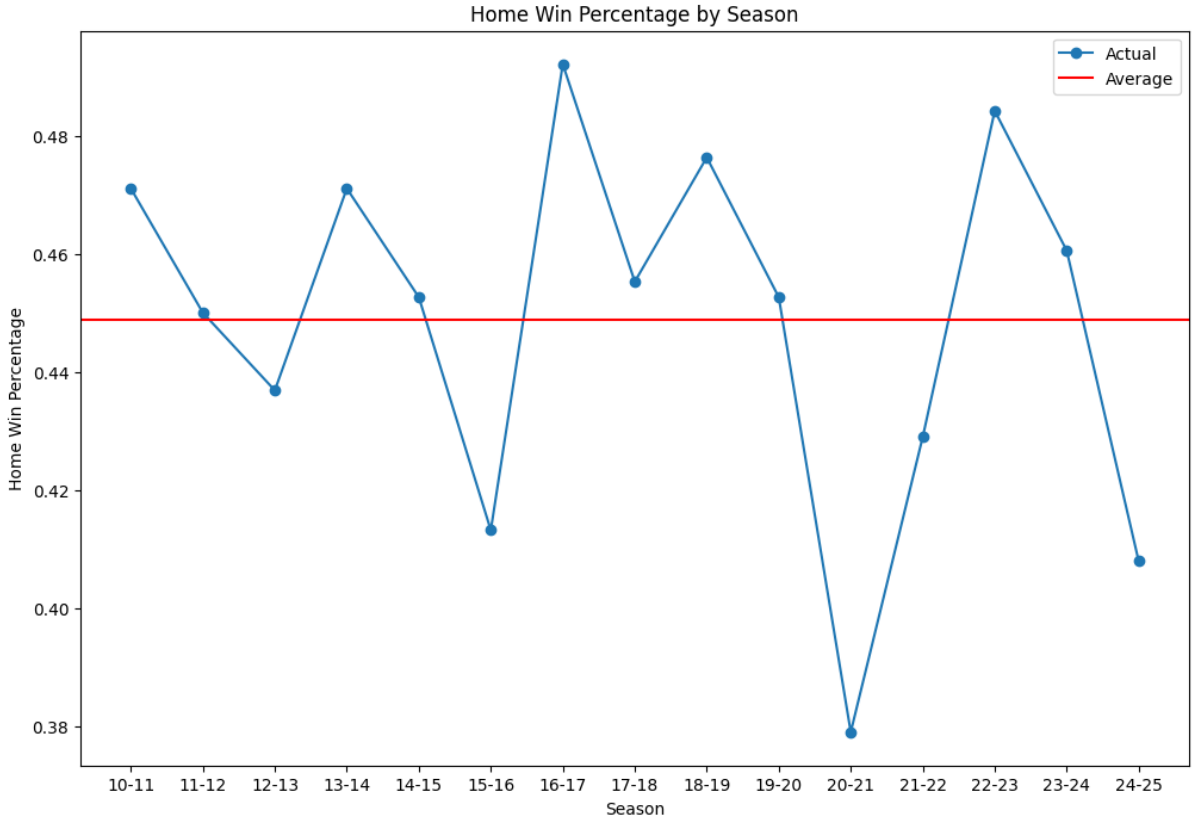

Figure 3: Home win percentage by season.

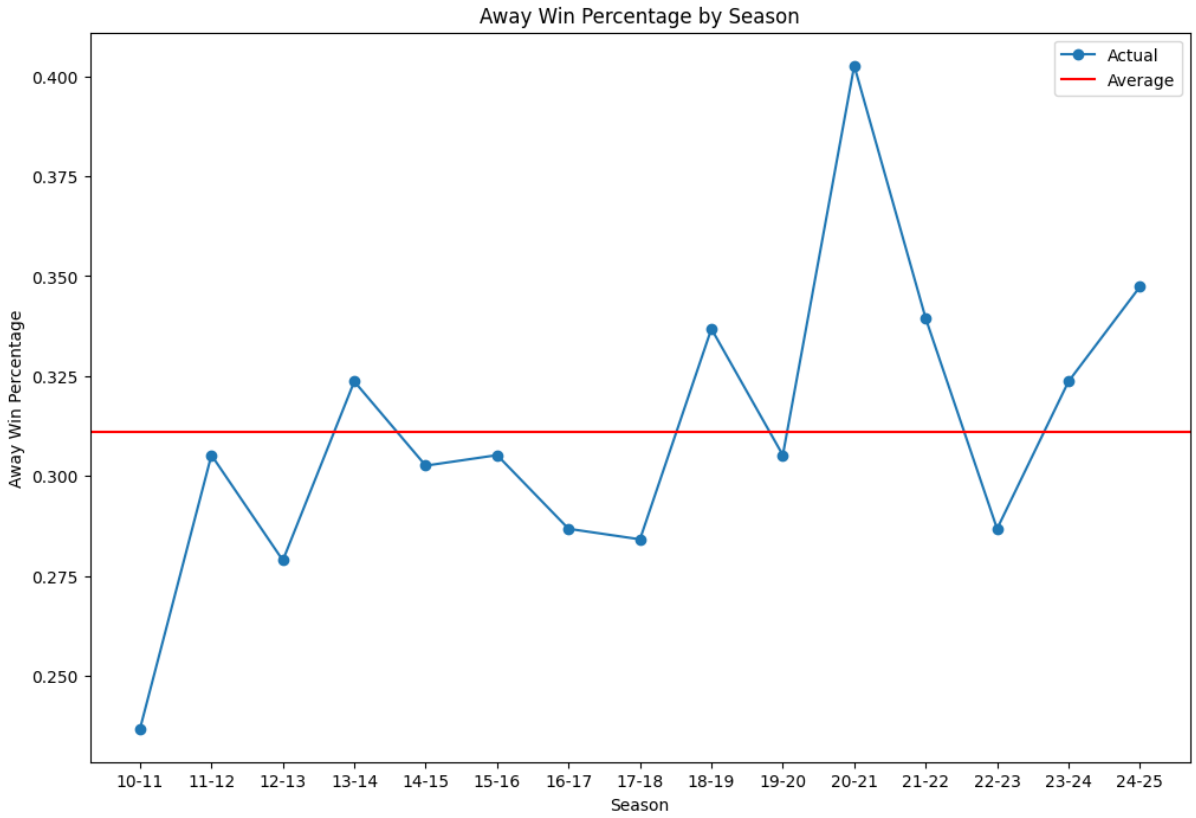

Figure 4: Away win percentage by season.

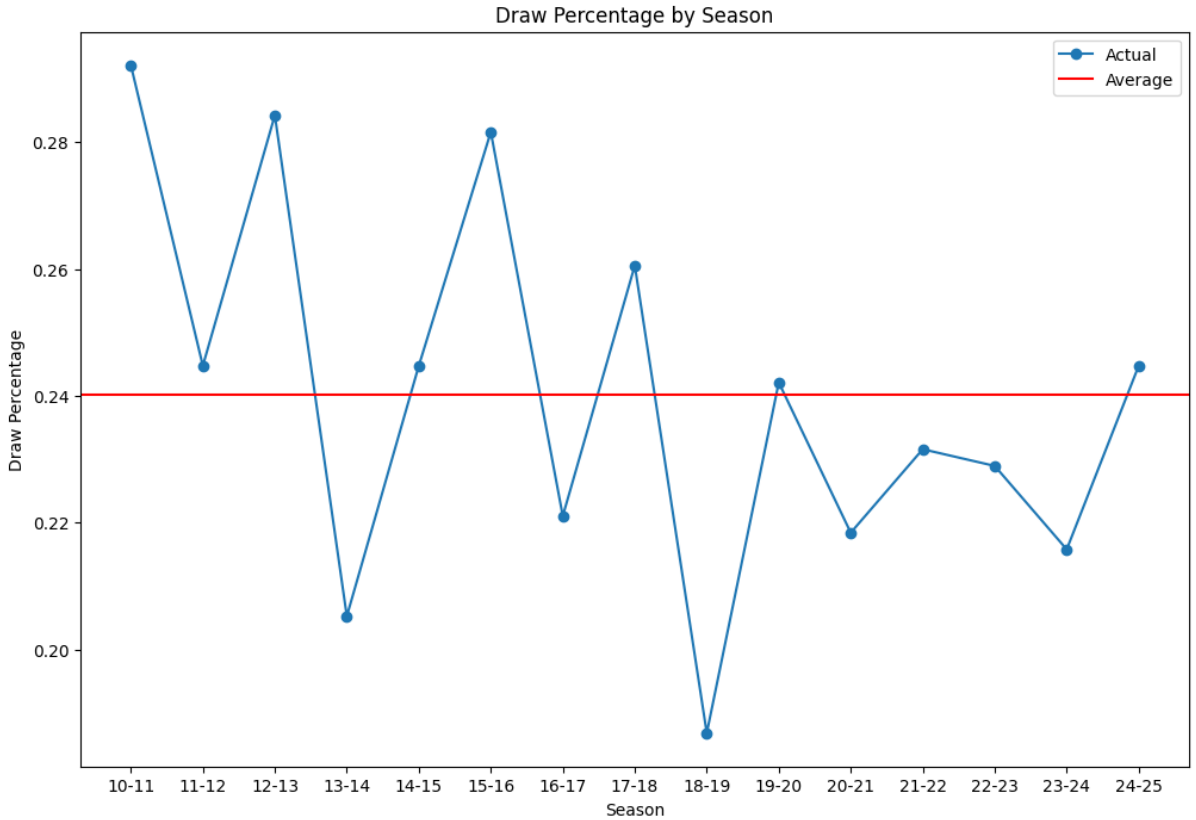

Figure 5: Draw percentage by season.

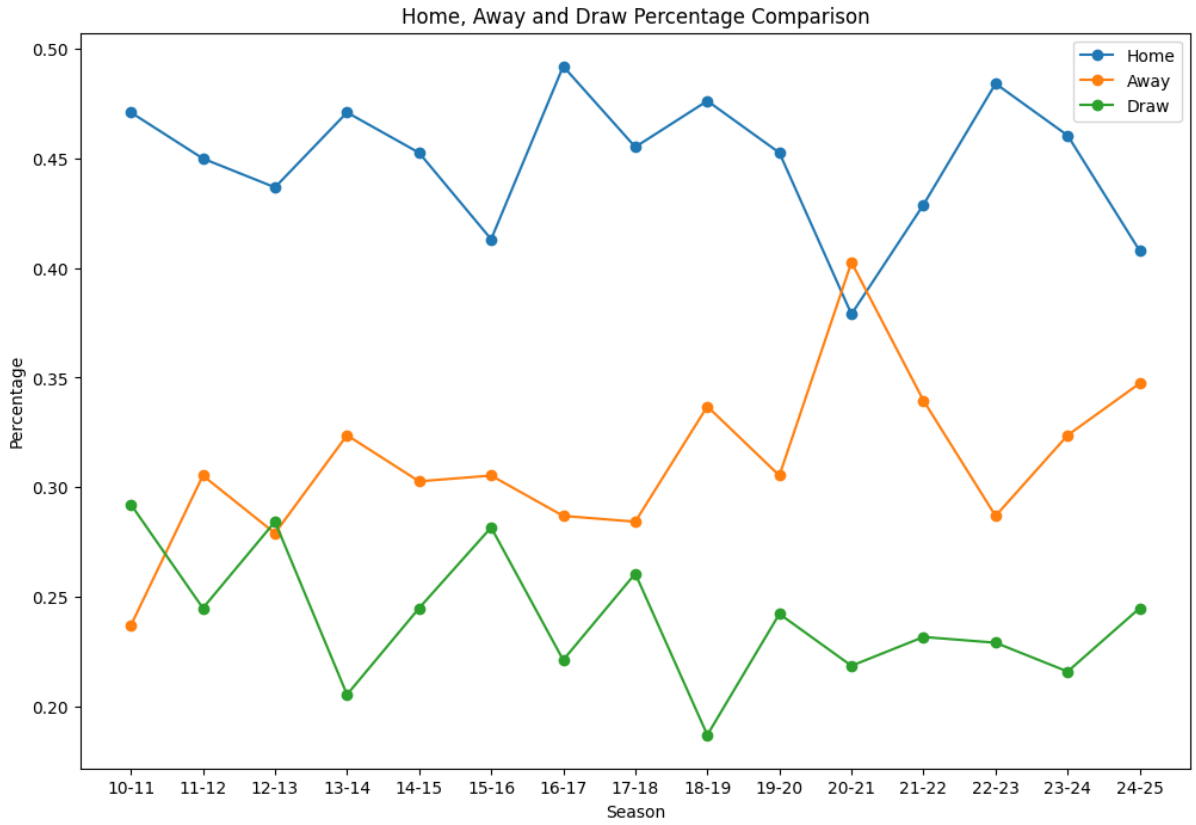

Figure 6: Home, away, and draw percentages together by season.

3.4 Home vs. away scoring and their difference

From Figure 7, we can see that across most seasons, home teams score more goals on average than away teams. The gap narrows sharply in season 2020-21 because home goals sharply fall while away goals surge, then the gap re-opens from 2021-22 and is especially large in the following season. Both home and away averages climb in season 2023-24, which aligns with the spike in average number of goals per match in Figure 1, before lowering the gap in the most recent season in our analysis. The convergence of two lines is consistent with the "ghost-games" period, where the crowds attending the games are either reduced or absent. The high-scoring 2023-24 season that we see in both Figure 1 and Figure 7 coincides with a new added-time guidance that increased effective playing time (Premier League, 2023).

The plot in Figure 8 demonstrates that the difference between home and away goals is positive in almost every year, hovering around 0.3 to 0.4 goals on an average match. However, it drops to almost zero in the season right after the COVID-19 outbreak, then rises again in the next two seasons before moderating like the pre pandemic. This plot reveals that the home winning edge actually exists, in which in each match, the home team scores around 0.3 to 0.4 more goals on average than the away team, yet also highlights the exceptional season 2020-21 when it nearly vanished, which is consistent with evidence that empty stadium conditions reduced home advantage in soccer.

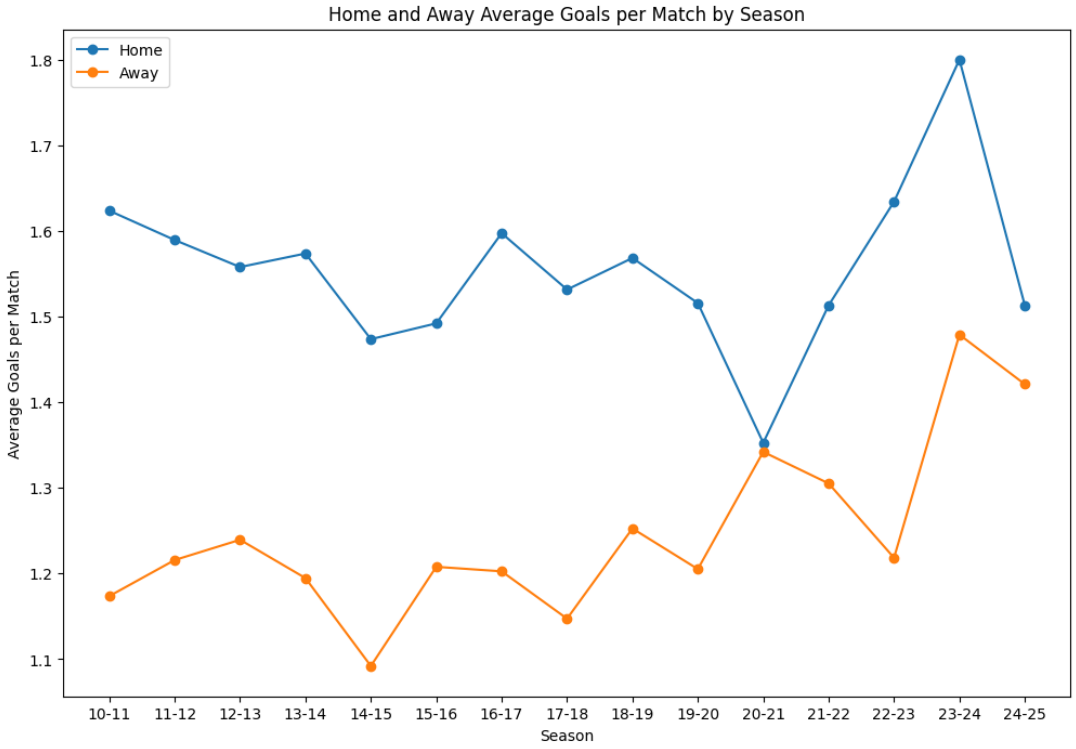

Figure 7: Average home vs. away goals by season.

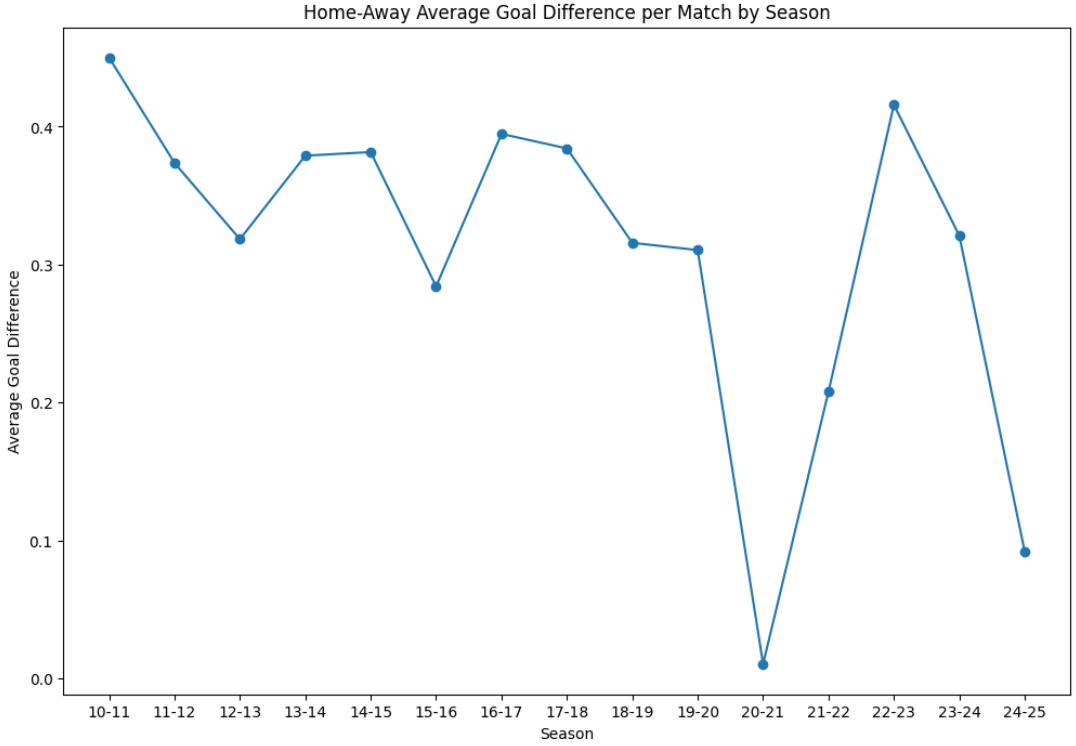

Figure 8: Home–away average goal difference by season.

3.5 Top-5 vs. Bottom-5: home/away win percentage

Figure 9 reveals that across all seasons, top 5 team's home win rates are consistently the highest, followed by top 5's away win rate, bottom 5's home win rate, and the bottom 5's away win rates are the lowest. This means that elite teams dominate at their home stadiums and also perform far better in away matches than weaker teams, even if they are played at home. However, there are two points that stand out from this plot. First, in season 2020-21, the gap compresses: home win rates for both high-tier and low-tier teams fell while their away win rates rise, since venue affected match results less in that season. Second, the gap reopens from season 2021-22 onward: the top 5 teams regain strong home form when fans come back to stadiums and maintain clearly higher away win rates than bottom 5 teams, while bottom 5 home and away win rates drop back toward their usual lower levels. This pattern lines up with the broader literature: home advantage weakened in "ghost-game" conditions during the pandemic (Leitner et al., 2022), and team ability is strongly related to the size of home/away performance gaps, where stronger teams export better quality, so their away win rates are higher even than home win rates of weaker teams (Ramchandani et al., 2021).

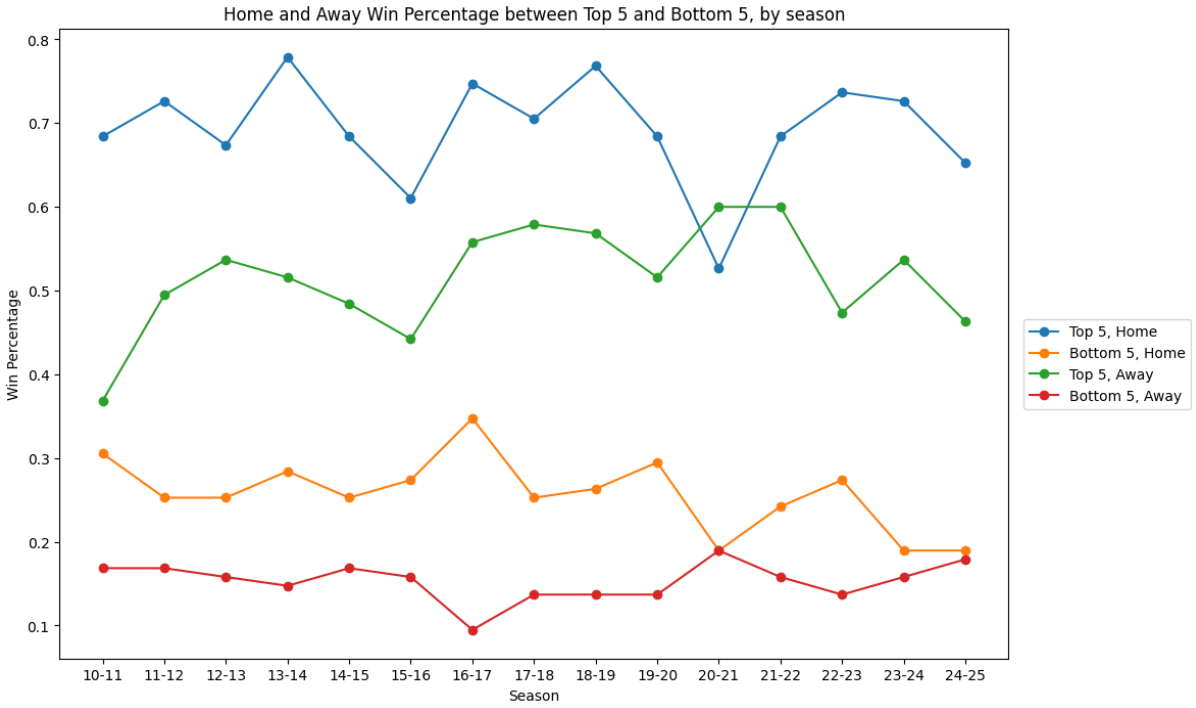

Figure 9: Home and away win percentages for top-5 vs. bottom-5 teams by season.

4 Conclusion

Together, the figures show that Premier League has a largely stable number of goals on an average match at around 2.8 for most of the 2010s, dipped in 2015, and then shifted upward in the early 2020s, pealing in 2023-24 before easing back in 2024-25. The distribution of full-time goals is right skewed: most games finish with two or three total goals, and matches with five goals or more are uncommon. By looking at the mean and the distribution side by side, we know why average number of goals move: either because games with 2-4 goals become more common or because we see a little bit more games with five goals or more.

Outcome patterns over time reinforce that story. Home wins generally lead the three scenarios in terms of probability, trailed by away wins, and draws are lowest. However, 2020-21 stands out as a compression year in which away wins rise and home wins fall relative to their long-run levels, which is consistent with the "ghost-game" period when crowd restrictions coincided with a reduced home edge. In terms of scoring, home teams usually outscore away teams by roughly 0.3-0.4 goals per match, but that margin nearly disappears in 2020-21 and then returns toward usual levels from 2021-22 onward. These two views tell the same story from different angles: home advantage is persistent but varies over time due to external factors, such as COVID-19.

By considering team quality, we can see that venue also interacts with ability. Top-five clubs post the highest home win percentages and also carry distinctly higher away win rates than bottom-five clubs. The gap between tiers narrows in 2020-21 season and re-opens afterward, indicating that stronger teams perform better even in away matches, especially once normal conditions resume. Together with the scoring and outcome trends, this suggests that league-wide shifts affect all teams but do not erase underlying performance differences.

Practically, these findings provide baselines for analysis and forecasting. For expected goals per match, a sensible prediction remains centered near a total of three goals per match, with awareness that recent seasons have more goals. For result modeling, long-run shares around 45% win-rate for home teams, 31% win-rate for away teams and 24% draw-rate are reasonable anchors, while the 2020-21 outlier reminds us that context can affect venue effects.

References

Almeida, C. H., & Leite, W. S. (2021). Professional football in times of covid-19: Did the home advantage effect disappear in european domestic leagues? *Biology of Sport*, *38*(4), 693–701. https://doi.org/10.5114/biolsport.2021.104920

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(2), 265–280. https://doi.org/10.1111/1467-9876.00065

Football-Data.co.uk. (n.d.). *England football results and betting odds data*. Retrieved August 11, 2025, from https://www.football-data.co.uk/englandm.php

Leitner, M. C., Daumann, F., Follert, F., & Richlan, F. (2022). The cauldron has cooled down: A systematic literature review on home advantage in football during the covid-19 pandemic from a socio-economic and psychological perspective. *Management Review Quarterly*, *73*(2), 605–633. https://doi.org/10.1007/s11301-021-00254-5

McCarrick, D., Bilalić, M., Neave, N., & Wolfson, S. (2021). Home advantage during the covid-19 pandemic: Analyses of european football leagues. *Psychology of Sport and Exercise*, *56*, 102013. https://doi.org/10.1016/j.psychsport.2021.102013

Premier League. (2019, November). *Clubs agree to introduce var from 2019/20 season* [Accessed 2025-08-11]. https://www.premierleague.com/en/news/912537

Premier League. (2023, August). *What's new in 2023/24: Applications of the laws of the game* [Accessed 2025-08-11]. https://www.premierleague.com/en/news/3617054

Premier League. (2025, June). *How the premier league fixture list is compiled* [Accessed 2025-08-11]. https://www.premierleague.com/en/news/4324541

Ramchandani, G., Millar, R., & Wilson, D. (2021). The relationship between team ability and home advantage in the english football league system. *German Journal of Exercise and Sport Research*, *51*(3), 354–361. https://doi.org/10.1007/s12662-021-00721-x