Premier League Dynamics (2019-2023): Manchester City Dominance, Arsenal Rise, Notable Trends.

Authors: Taahaa Dawe, Dhruv Gupta, Ghizlane Rehioui

Abstract

In this research, an Exploratory Data Analysis (EDA) is undertaken to unravel the performance of the Top 5 English Premier League (EPL) teams across the seasons 2019-2023. Focusing on diverse performance metrics, the primary research question explores the average goals scored per season, providing an overall perspective on the Top 5 teams' performance. The investigation extends to the percentage of wins, differentiating from goal-centric metrics to assess overall match outcomes. The analysis further explores variations in performance between home and away matches, shedding light on potential areas for team improvement. Notably, while the statistical analysis supports the rejection of the null hypothesis for the average number of goals scored by Arsenal, indicating a significant difference, it almost fails to do so for possession differences between 2022 and 2019 with a p-value close to a significance level of 0.05. This research not only addresses critical questions but also underscores the importance of varied metrics in understanding the nuanced narratives of football team performance.

Table of Contents

1. Introduction

In the vibrant landscape of contemporary football, understanding the intricate dynamics of team performance has become a crucial endeavor. As the heartbeat of global sports culture, the English Premier League (EPL) serves as an arena where strategy, skill, and sheer passion converge to create riveting narratives on the field. In this pursuit, this research aims to unravel the compelling stories hidden within the data of the Top 5 EPL teams spanning the seasons from 2019 to 2023.

The focal point of the investigation lies in the average goals scored per season, a metric that not only encapsulates the essence of team proficiency but also serves as a fundamental gauge of success. Why does this matter? Football enthusiasts, analysts, and even casual viewers are captivated by the dynamic interplay of goals, victories, and the strategies employed by their favorite teams. The quest to answer the research question, "What are the average goals scored per season for the Top 5 EPL teams, and how do those teams rank against each other?", or more generally the question of team performance, aims to provide a comprehensive understanding of the league's competitive landscape.

It is essential to recognize the evolution of football analytics and the profound impact of data-driven insights on the game. Traditional points tables have given way to a nuanced exploration of team performance metrics and shooting statistics. This work builds upon the transformative trend, offering a fresh perspective on team dynamics, player proficiency, and strategic variations that have shaped Premier League football from 2019 to 2023.

This paper is organized as follows: a data overview including how the data was collected, some potential sources of bias, the features used in the analysis, and the main data pre-processing steps, followed by the methods and rationale of the analysis, the results of the distribution and correlation analysis, exploratory analysis of the Top 5 teams, and lastly the two hypothesis tests that concern the Arsenal team.

2. Data

2.1 Data collection

The data utilized for this analysis was scraped from the FBRef website, a reputable source for football statistics and history. The focus of this project was on gathering and analyzing individual team match data from the Premier League, spanning from the 2019 season to the current season 2023. Crucial shooting statistics were also extracted for each team in every match from the shooting table. This comprehensive dataset allows for a detailed exploration of team performance metrics and shooting proficiency within the Premier League during the specified timeframe. The link to the code is available in Appendix 2.

2.2 Potential sources of bias in the data

Biases from web-scraped data

Web scraping may be a great way to quickly access real-time data. However, it is possible to encounter issues such as the untimely updates of important information, which could potentially lead to misinterpretations. Another possible source of bias would be to ask if the period between 2019-2023 is truly representative of the teams' performance or success.

Biases related to Football as a game

Football is a highly unpredictable game. Due to the low-scoring nature of the sport, a 'less qualified' team can easily win against a much more talented team. It is also easier to catch up in terms of points. Additionally, a simple mistake from a defender or keeper may be an easy opportunity for the opponent to score and win a match, especially during extra time.

In addition to the low-scoring nature of Football, referee bias is a popular theory among Football fans, hypothesizing that referees play a huge role in the match, either favoring or against a team. Some fans strongly believe that some referees turn a blind eye to some penalties, which could have made some teams easily qualify given that the opponent's team does not have a good goalkeeper. Other biases include seasonal bias or team/player bias related to the scouts or players involved in the match.

Choice of metrics (Measurement Bias)

The choice of the metrics, i.e. the domain or industry knowledge, that drive this analysis or conclusions is crucial. In other words, the standardized metrics or parameters used in Football Analytics should be respected, and using other metrics may lead to misinterpretations.

Omitted variable bias

In this analysis, omitted variable bias could be encountered, where some critical attributes that influence the outcome may be unintentionally disregarded.

2.3 Features used in the analysis

Upon saving the scraped data as a CSV file, the dataset is composed of 3138 rows and 26 columns. Not all the columns will be used to achieve the analysis goals of this paper. The main features of interest are summarized in the table below:

Column Name	**Column Description**
date	Date of the match
venue	Where the game is played (Away or Home)
result	W for Win, D for Draw, and L for Loss
goals_scored	Number of goals scored
goals_conceded	Number of goals conceded
opponent	Opponent team name
possession	Duration for which the team had the ball possession
shots	Shots taken by the team
shots_on_target	Shots on target by the team
season	Season year
team	Team

Table 2.3. Main features of interest in the scraped dataset.

The list of all 26 columns can be found in Appendix 1. The next section will deal with the pre-processing of the dataset.

2.4 Data Pre-Processing

Most of the data cleaning was done before saving the scraped data in a CSV file. This section takes another look at the scraped data for research-specific cleaning and pre-processing. At first glance, the data looks mostly clean with only one column *avg_shot_distance* having one missing value and the *attendance* column with some missing values. The column *Unnamed: 0* can be dropped as it has no relevance to the research.

Other data-cleaning steps are necessary before starting the exploratory analysis. This includes checking the names of the teams and opponents to make sure they are both proper for future analyses and comparisons.

There are 26 teams in the dataset, which is difficult to analyze. The focus will be given to the Top 5 best teams in the EPL. The choice of the best teams is done based on the total goals scored over the 5 seasons of 2019-2023. The top 5 teams in terms of highest total goals scored (*goals_scored*) over the period between 2019 and 2023 are as follows, from highest to lowest goals scored:

Rank	Top 5 Teams	Total Goals Scored (2019-2023)
1	Manchester City	392
2	Liverpool	334
3	Tottenham Hotspur	281
4	Arsenal	269
5	Manchester United	260

Table 2.4. Top 5 EPL teams based on total goals scored 2019-2023.

And with this, the rest of the analysis will focus on these Top 5 teams. New data frames will be generated from the original one as needed.

3. Methods and Analyses

This section summarizes the main methods along with the addressed research questions. This also includes the necessary changes to the dataset, distribution comparison, correlation analysis, exploratory data analysis, and hypothesis testing. The results section details the outputs and insights extracted from each of those methods. It is important to note that no extra data preprocessing was done until the EDA, where the dataset was filtered on the Top 5 Teams identified at the end of the Data Pre-Processing section.

3.1 Distribution and Correlation Analysis

Looking at all the rows of the dataset only eight variables of interest. Those are *season*, *venue*, *result*, *shots*, *shots_on_target*, *possession*, *goals_scored*, and *goals_conceded*.

3.2 Methods and Rationale of Exploratory Data Analysis

The Exploratory Data Analysis aimed to answer questions relating to the performance of the Top 5 teams based on the total goals scored over the five seasons 2019-2023 while making sure that the performance metrics are varied.

Rather than reviewing the performance in terms of total goals scored over the whole five seasons, the first question to answer in EDA aims to look at averages. This is because looking at totals per season would not be useful since the 2023 season is not complete yet. This will give an overall idea of the performance of the Top 5 EPL Teams. Other performance metrics besides the number of goals could be the expected number of goals (xG), the percentage terms, or the number of wins relative to the number of matches played over all 5 seasons (2019-2023). The latter metric is referred to as a percentage of wins. The win percentage disregards the number of goals scored and focuses on the outcome of the match. The main question it answers is how many matches a team won out of those they played.

Another way to evaluate a team's performance is the difference in performance between Home and Away matches to compare how the team performs in both cases, leading to getting an idea on which venue creates better team performance. Shooting Skill (Precision) and Blocking Opponent Goals (Defense) are other interesting measures to consider for the assessment of the Top 5 teams' skills.

3.3 Methods and Rationale of Hypothesis Testing

In the exploratory data analysis, Manchester City emerges as the leading goal-scoring team, showcasing a significant correlation between goal shots and shots on target. Noteworthy, Tottenham Hotspur distinguishes itself as a top-performing team, particularly excelling in the percentage of successfully converted shots.

Intriguingly, a disparity arises between the reported average of 2.31 goals for Arsenal on the Manchester City official website (Cox, 2023) and the calculated sample mean, prompting further investigation. Not only that, but also despite having a competitive position among the Top 3 best EPL teams as seen in the EDA, Arsenal seems to be getting more negative publicity than Tottenham (Attwood & Attwood, 2018), leading to wondering about the validity of this claim.

Additionally, to further investigate Arsenal's performance, there is a keen interest in understanding if the possession time for Arsenal has undergone a significant increase from 2019 to 2022. To substantiate these findings, thorough statistical tests and a comprehensive analysis are deemed necessary.

Accordingly, two hypotheses are explored and tested:

Research Question 1

> Does the observed sample mean of 1.71 for the average goals scored by Arsenal, derived from a dataset, provide statistically significant evidence to reject the claim made on the Manchester City's Official Website (Cox, 2023) that the average number of goals scored by Arsenal is 2.31? Furthermore, what is the probability of obtaining sample means as extreme as 1.71 or more contradictory to the stated average of 2.31, assuming the true mean is indeed 2.31, and the sample size and variance remain constant across random samples of the same size?

Null Hypothesis (H0)

The null hypothesis assumes that the average number of goals scored by Arsenal is consistent with the claim made on Manchester City's Official Website, and there is no significant difference.

$$H_0 : \mu = 2.31$$

Alternative Hypothesis (H1)

The alternative hypothesis posits that the observed sample mean of 1.71 suggests a significant difference from the claimed average of 2.31, indicating that Arsenal's actual goal-scoring average differs.

$$H_1 : \mu \neq 2.31$$

Research Question 2

Is there compelling evidence in the dataset to support the claim that there has been a substantial increase in possession for Arsenal in 2022 compared to 2019? In other words, does the observed data provide statistical significance to reject the null hypothesis that the mean possession difference is μ_d equal to zero in favor of the alternative hypothesis that the mean possession difference μ_d is greater than zero?

Null Hypothesis (H0)

The null hypothesis posits no significant difference in possession for Arsenal between 2022 and 2019.

$$H_0 : \mu_D = 0$$

Here, μ_D represents the population mean difference in possession between the two years.

Alternative Hypothesis (HA)

The alternative hypothesis suggests that there is a significant increase in possession for Arsenal in 2022 compared to 2019.

$$H_A : \mu_d > 0$$

Here, μ_d it represents the population mean difference in possession, and the greater-than symbol indicates the direction of interest—specifically, an increase in possession.

This setup allows for a one-sided hypothesis test to determine whether there is enough evidence in the data to support the claim of a substantial rise in possession during 2022.

For a one-sample t-test:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Here:
- $\bar{x}$ is the sample mean,
- μ_0 is the hypothesized population mean (for the null hypothesis),
- s is the sample standard deviation,
- n is the sample size.

It is important to note that a new column has been generated for the possession difference, being equal to the possession in 2022 minus that of 2019. This column is the concerned feature for this hypothesis test. The t-statistic for a paired-sample t-test (such as in the case of comparing means for two different years) is:

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

Here:
- n is the sample size,
- $\bar{d}$ is the sample mean of the differences (for paired-sample t-test),
- μ_d is the hypothesized population mean difference (for the null hypothesis),
- s_d is the sample standard deviation of the differences.

4. Results

4.1 Correlation and Distribution Comparison

As previously mentioned, the analysis will consist of looking at eight variables of interest. Those are *season*, *venue*, *result*, *shots*, *shots_on_target*, *possession*, *goals_scored*, and *goals_conceded*.

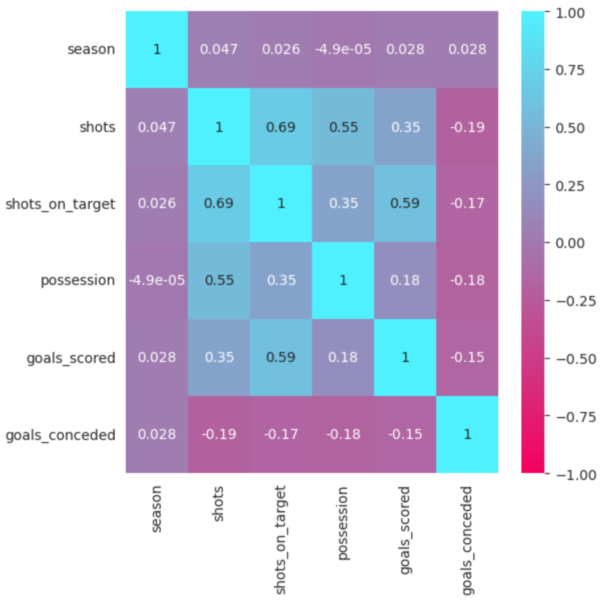


Figure 4.1.1. Correlation Matrix of the variables of interest.

The features of *possession* and *shots* are positively correlated as depicted in the Pair Plot below. This is also confirmed by the Correlation Matrix above, correlating 0.55. This makes sense because the more a team has the ball, the more chances they have at scoring. The same is found for *shots_on_target* with *shots* (0.69), *goals_scored* with *shots_on_target* (0.59), and *possession* with *shots_on_target* (0.35). Some smaller positive correlations are found for *possession* with *goals_scored* (0.18).

The *possession* with *goals_conceded*, *shots_on_target* with *goals_conceded*, *goals_conceded* with *goals_scored*, and *goals_conceded* with *shots* are negatively correlated but not strongly, with values between -0.19 and -0.15. The rest of the correlations between the remaining variables are negligible.

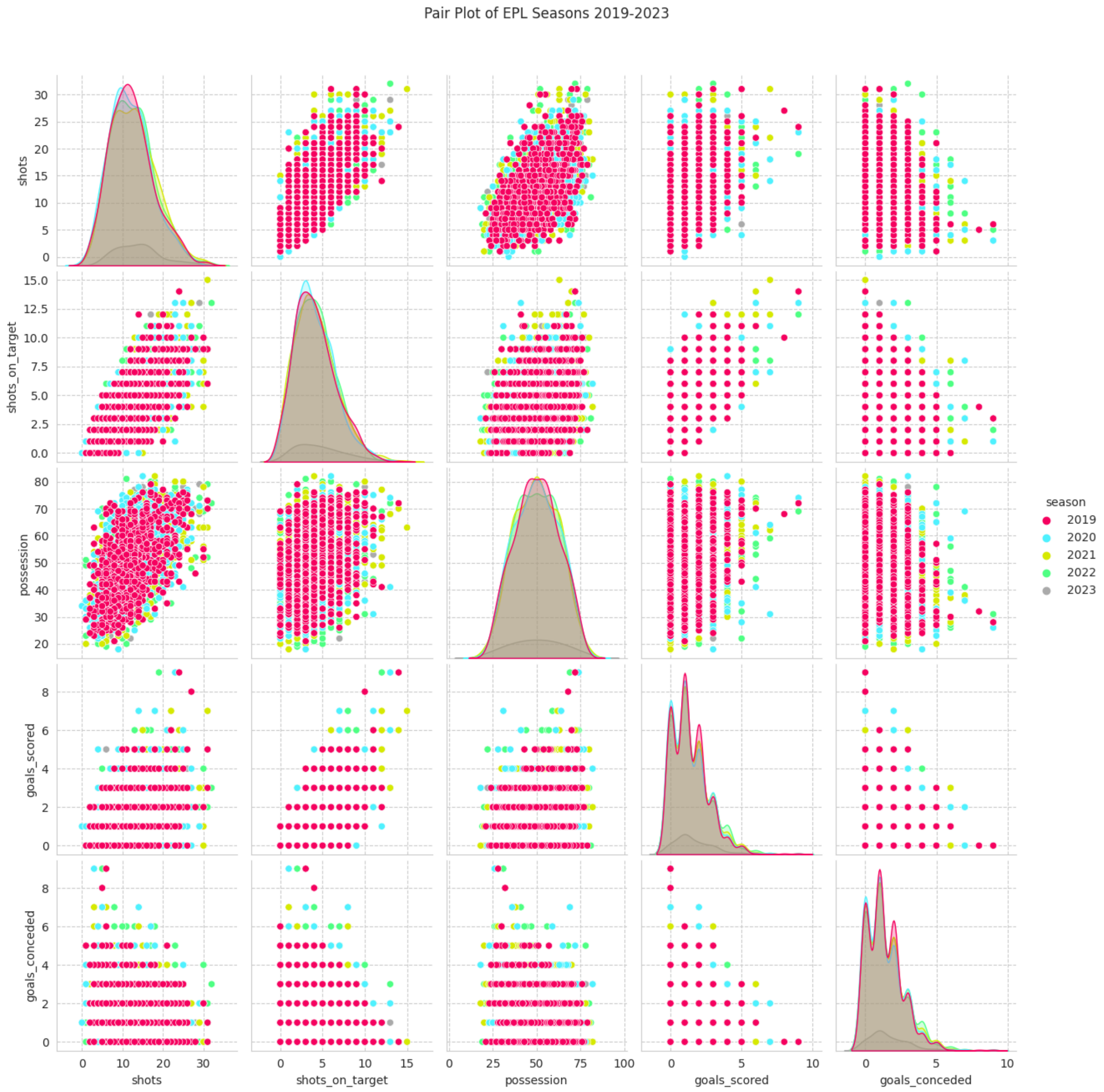


Figure 4.1.2. Pair Plot of the variables of interest.

Based on the Pair Plot, the distributions of each variable of interest for all teams is the same regardless of the season. The features *shots*, *shots_on_target*, and *possession* appear to have a distribution that is close to the normal distribution with median values of 10 shots, 2.5 shots on target, and 50 minutes respectively for every season. The features *shots* and *shots_on_target* have slightly skewed curves in the same direction. The *goals_scored* and *goals_conceded* appear to be about the same shape.

Only the year 2023 looks different and much lower than the other seasons. This is because the season 2023 is not yet finished and the data for it is not yet available.

4.2 Exploratory Data Analysis

4.2.1 Average goals scored per season for Top 5 EPL

Given that the year 2023 lacks data, it is more meaningful to plot the average scored goals to have an overall idea of the performance of the Top 5 EPL Teams.

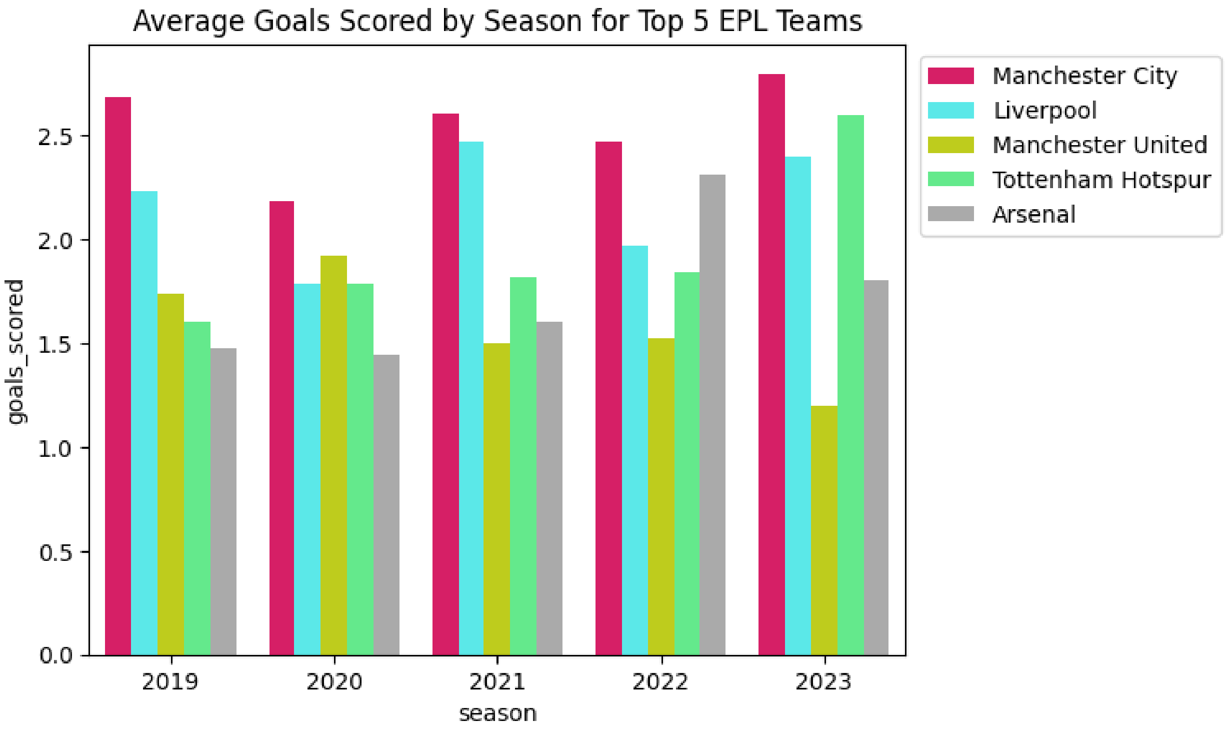


Figure 4.2.1. Average Goals Scored by Season for Top 5 EPL Teams.

Similarly to what was found in the Data Preprocessing section, Manchester City ranks first in all seasons, followed by either Liverpool (2019 and 2021), Arsenal (2022), Manchester United (2020), or Tottenham Hotspur (2023) in the second place in terms of average scored goals per season.

On average, it seems that Manchester City, compared to the other Top 5 teams, is able to maintain the highest seasonal goals scored from one year to the next. In 2023, it looks like Tottenham Hotspur is competing fiercely, ranking 2nd on average, as opposed to previous years where it was third or fourth. Liverpool is a strong competitor against Manchester City (1st team) and is ranked either 2nd or 3rd at most. The average number of goals scored by Manchester United looks like they are decreasing over time, ranking last among the Top 5 in the last three seasons.

4.2.2 Percentage of wins over all 5 seasons (2019-2023)

The win percentage disregards the number of goals scored and focuses on the outcome of the match. The main question it answers is how many matches a team won out of those they played.

Top 5 Teams	% Wins
Manchester City	73.25%
Liverpool	65.61%
Arsenal	53.50%
Manchester United	50.96%
Tottenham Hotspur	49.68%

Table 4.2.2. Percentage of wins over all 5 seasons for Top 5 Teams (2019-2023).

In terms of the percentage of wins, there seem to be a few differences from the performance based on the sum of goals scored over the 5 seasons 2019-2023.

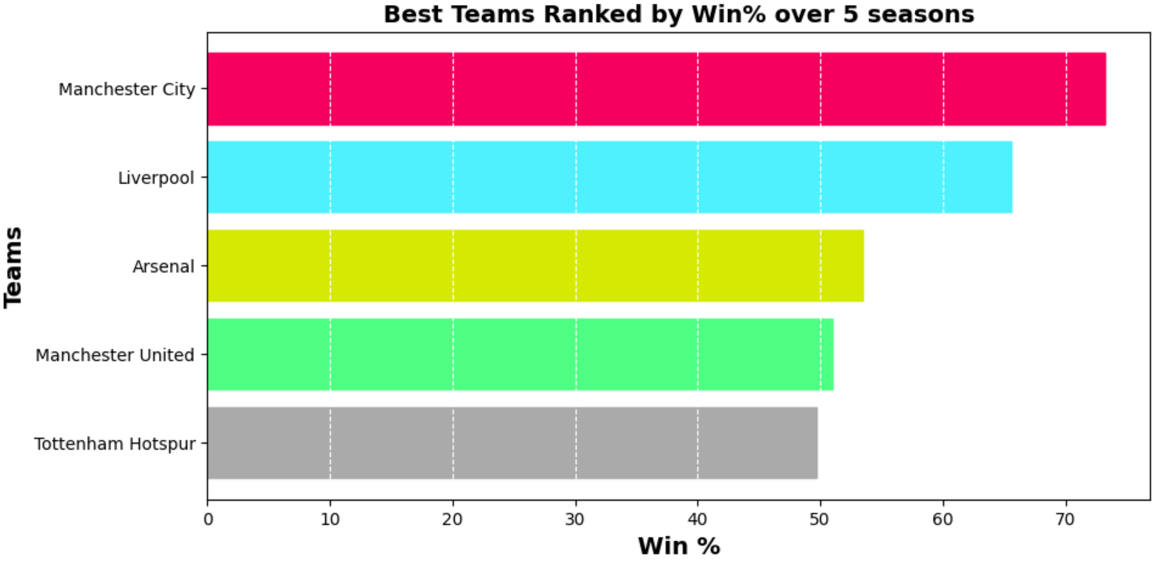


Figure 4.2.2. Percentage of wins over all 5 seasons for Top 5 Teams (2019-2023).

From the graph above, it looks like Manchester City is the team with the highest percentage of wins over the 5 seasons of 2019-2023. Liverpool comes second, followed by Arsenal, Manchester United, and finally Tottenham Hotspur.

Manchester City (73.25%) and Liverpool (65.61%) remain the first two in terms of percentage of wins as well as total goals scored between the seasons 2019 and 2023. However, Tottenham

Hotspur (49.68%) turns out to have the lowest Percentage of Wins among the Top 5 Teams despite having a higher sum of goals scored over the 5 seasons. Arsenal (53.50%) and Manchester United (50.96%) climb up one step each, having a higher win percentage than Tottenham Hotspur.

4.2.3 Any difference in performance between Home and Away matches?

Looking at the away vs. home total of goals scored allows for the comparison between how the team performs in both cases and could lead to further investigation on how to improve team conditions or where the team needs to put more effort.

Top 5 Teams	Away Goals	Home Goals
Manchester City	168	224
Liverpool	152	182
Tottenham Hotspur	131	150
Arsenal	114	155
Manchester United	109	151

Table 4.2.3. Total of Goals Scored of Top 5 for Away vs. Home over 5 seasons (2019-2023).

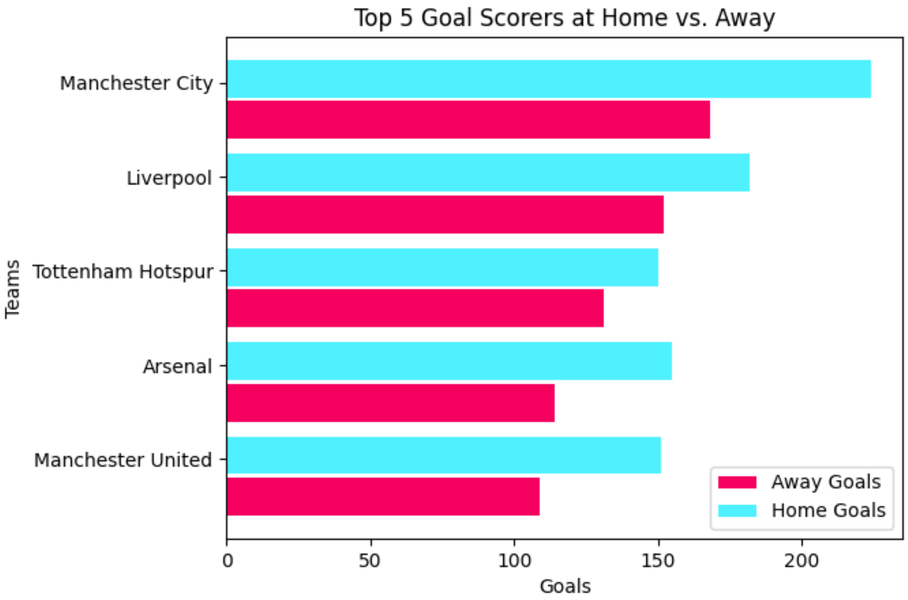


Figure 4.2.3. Bar Plot of Goals Scored of Top 5 Away vs. Home over 5 seasons (2019-2023).

In general, the goals home are higher than away goals for all teams. As seen in the table above, Manchester City and Liverpool are still maintaining the position of top two respectively in terms of home and away goals. Tottenham Hotspur scores much more than Arsenal and Manchester United only when playing at home. But in away games, the latter two score slightly higher.

4.2.4 Identifying Overperforming Teams Based on Goals Scored and Expected Goals (xG)

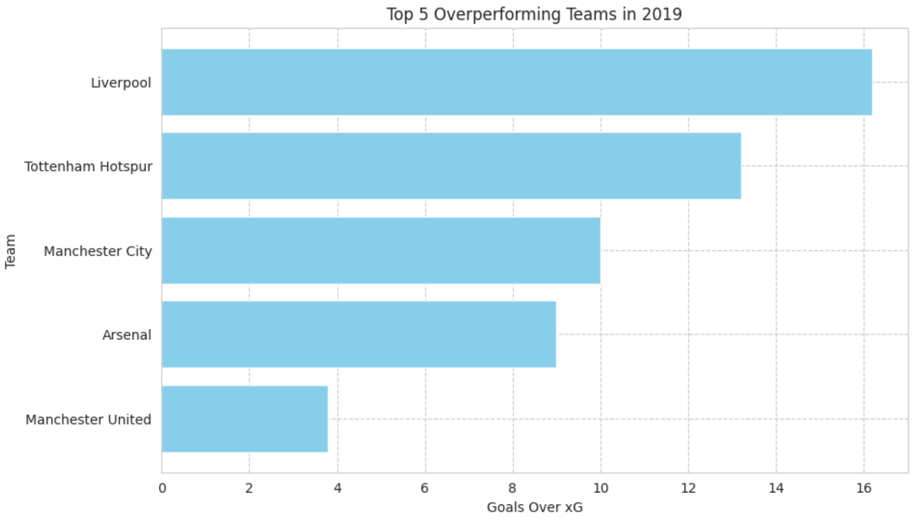
In the realm of football analytics, Expected Goals (xG) have emerged as a pivotal metric for evaluating team performance. xG quantifies the likelihood of a shot resulting in a goal, taking into account numerous factors like shot location, type, and defensive pressure. Assessing how a team performs relative to their XG provides crucial insights into their offensive efficiency.

Objective

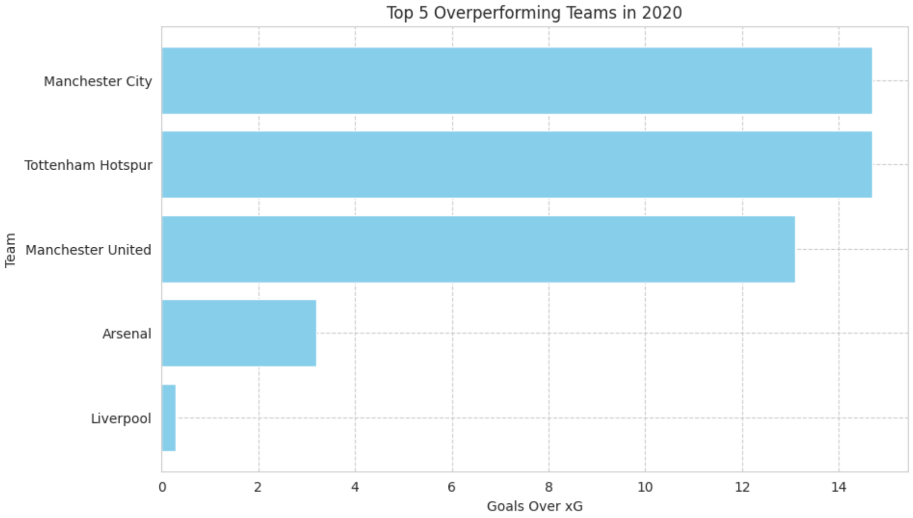
This analysis seeks to identify overperforming football teams by contrasting their Goals Scored with Expected Goals (xG). Overperforming teams consistently score more goals than their xG projections suggest. The analysis is divided into two main parts:

Part 1: Identifying Top 5 Overperforming Teams Each Year

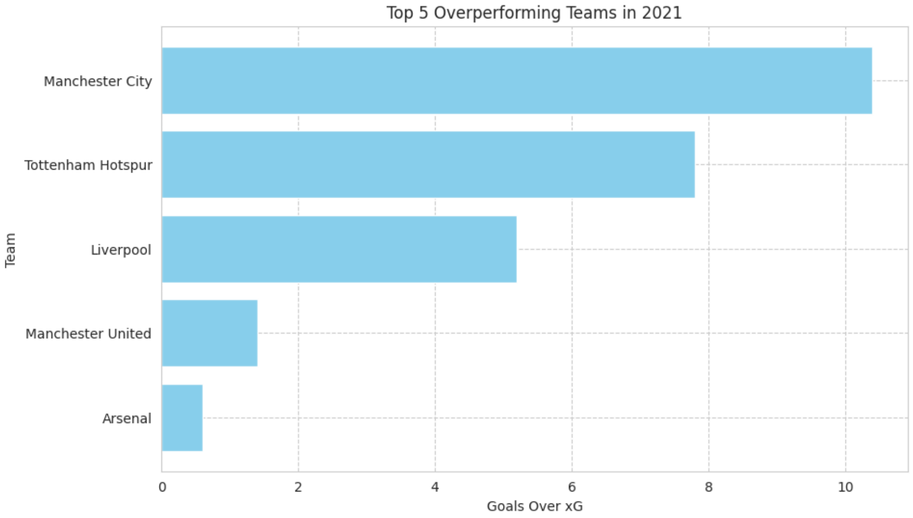
In this segment, the individual football seasons are dissected to pinpoint teams that have excelled in surpassing their XG projections. The process involves calculating Goals Scored and xG for each team within a season and identifying the top 5 teams with the highest positive differences between Goals Scored and xG.



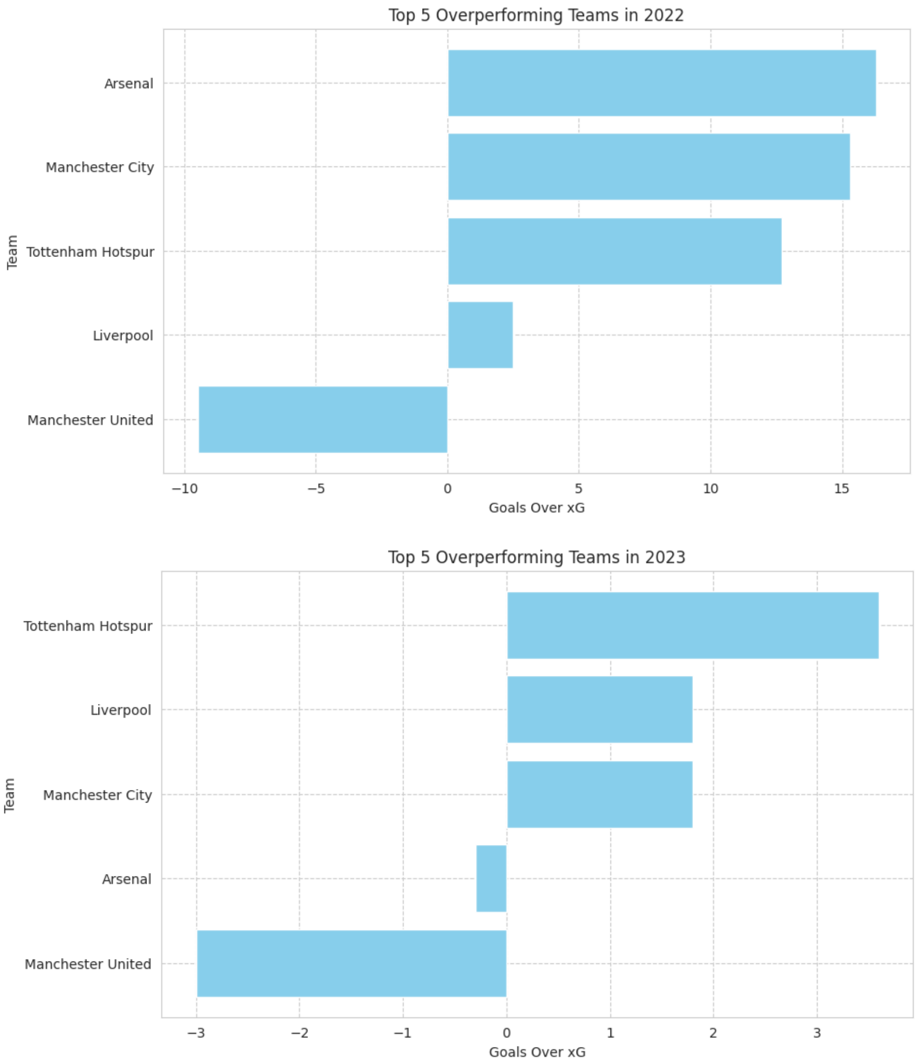
Top 5 Overperforming Teams in 2019



Top 5 Overperforming Teams in 2020



Top 5 Overperforming Teams in 2021


Top 5 Overperforming Teams in 2022

Top 5 Overperforming Teams in 2023

Figure 4.2.4.1. Top 5 Overperforming Teams Each Year for Goals over Expected Goals.

Part 2: Identifying Top 5 Overperforming Teams Combined for All Seasons

Here, data across multiple seasons is aggregated to spot teams that have consistently outperformed their xG expectations. The cumulative Goals Scored and xG are computed for each team across all seasons and ascertain the teams with the greatest collective positive differences between Goals Scored and xG.

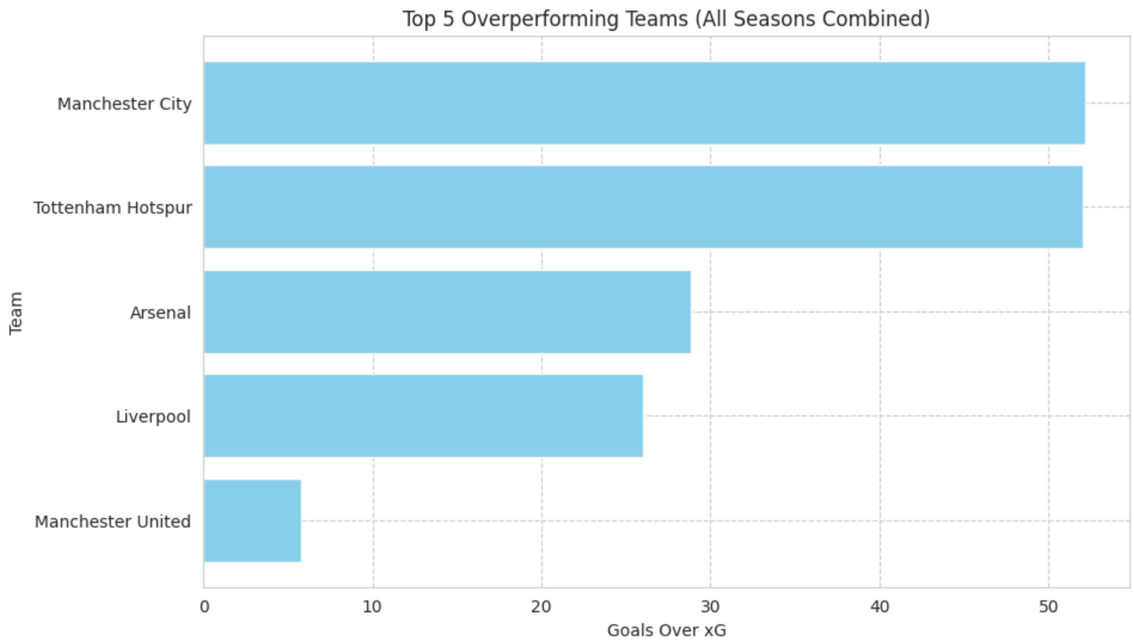


Figure 4.2.4.2. Top 5 Teams Combined for All Seasons for Goals over Expected Goals.

Both parts of the analysis incorporate visualizations in the form of bar charts. These charts showcase the top 5 teams for each season (Part 1) and the top 5 teams across all seasons (Part 2). They provide a clear representation of which teams have consistently demonstrated superior goal-scoring capabilities in relation to their xG expectations. Despite having fluctuating performance among the top 5 teams from one year to the next, Manchester City has the most goals over xG over the 5 seasons as seen in figure 4.2.4.2, which is somewhat expected as they scored the most goals.

4.2.5 Using Percentage Terms for Overperformance Analysis

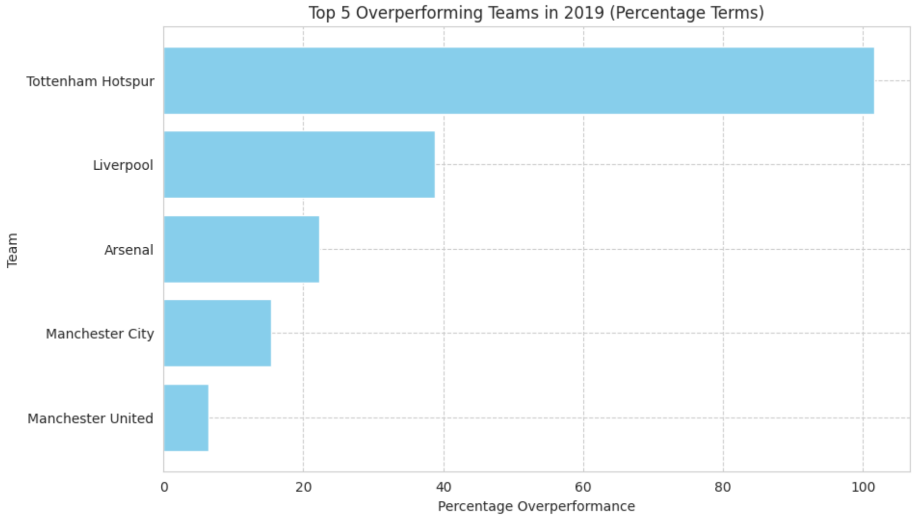
In this part, a significant adjustment is made to the way teams' overperformance relative to Expected Goals (xG) is assessed. Rather than relying solely on total overperformance values, the percentage terms are used. This shift offers a more meaningful and equitable basis for comparison across seasons and teams. The percentage terms are a better choice for comparison for the following reasons:

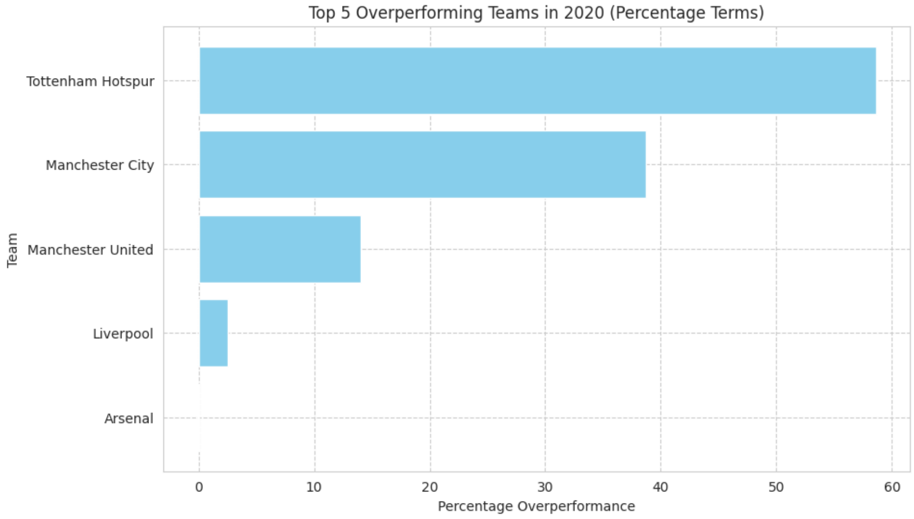
Standardized Assessment: Total overperformance values may be influenced by the volume of scoring opportunities a team encounters in a given season. Teams with more opportunities are likely to have higher total overperformance figures. By calculating percentage overperformance, the assessment is standardized, enabling fair comparisons irrespective of variations in opportunities.

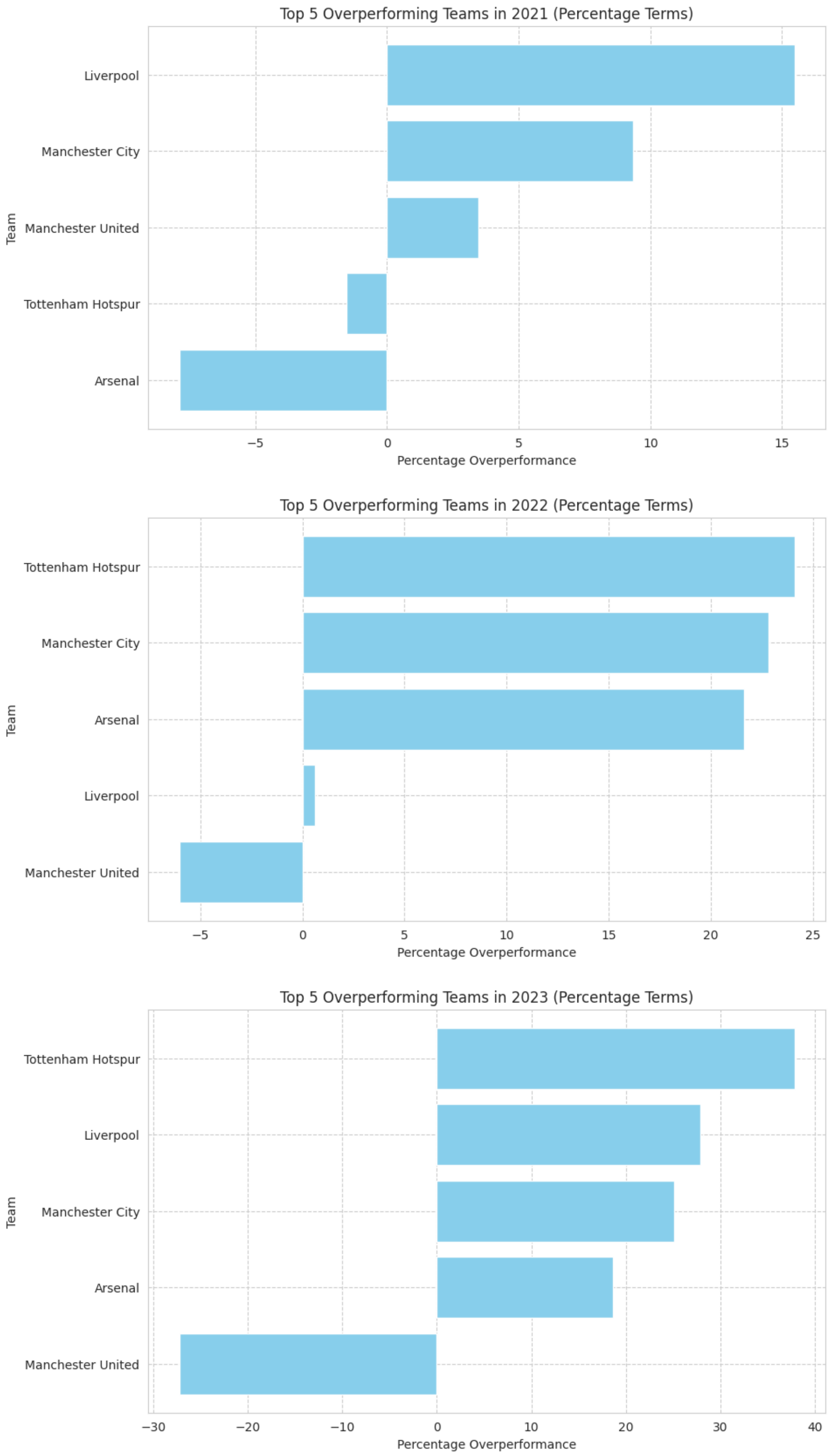
Reflecting Efficiency: Percentage overperformance provides insights into a team's efficiency in front of the goal. It gauges a team's ability to make the most of the opportunities they create, which is a key indicator of offensive prowess.

Comparative Analysis: Using percentage terms allows to evaluate teams on a level playing field, making it easier to identify consistent overperformers across different seasons. This approach helps pinpoint teams that consistently excel in converting opportunities into goals, regardless of the total number of chances.

Part 1: Identifying Top 5 Overperforming Teams Each Year (Percentage Terms)







Top 5 Overperforming Teams in 2021 (Percentage Terms)

Top 5 Overperforming Teams in 2022 (Percentage Terms)

Top 5 Overperforming Teams in 2023 (Percentage Terms)

Figure 4.2.5.2. Top 5 Overperforming Teams Each Year for Percentage Terms.

Part 2: Identifying Top 5 Overperforming Teams Combined for All Seasons (Percentage Terms)

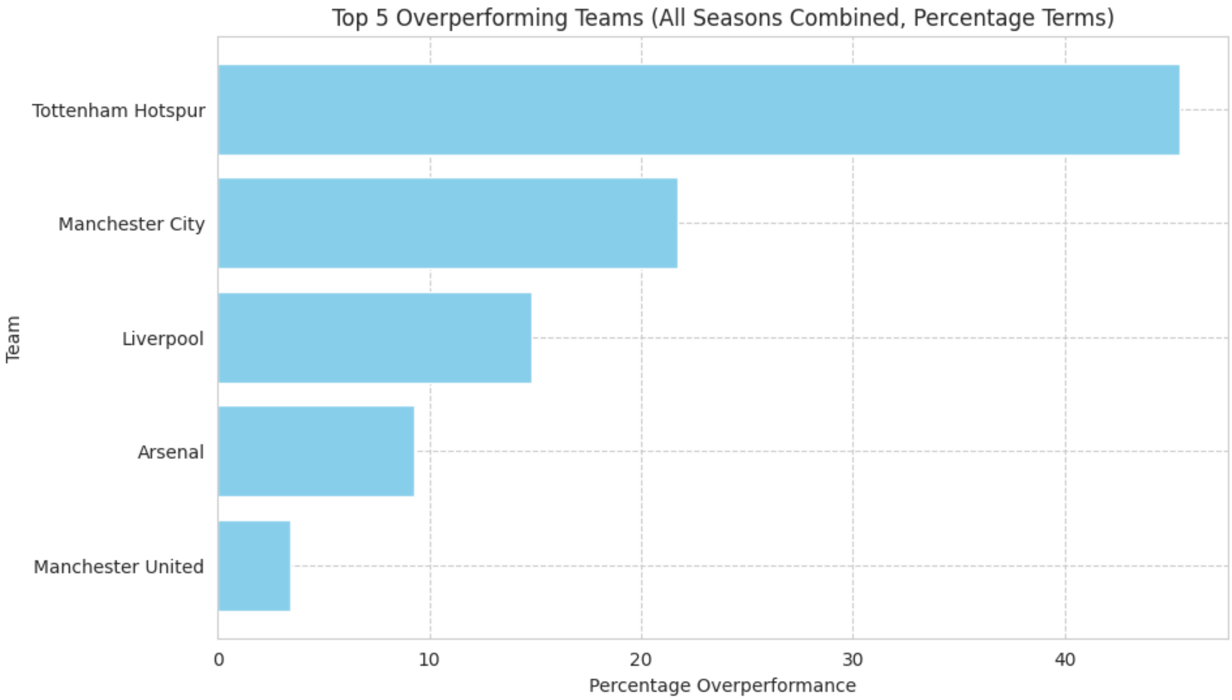


Figure 4.2.5.2. Top 5 Overperforming Teams for Percentage Terms for all Seasons.

Based on the figure above, Tottenham Hotspur makes the most of their chances as their percentage overperformance is higher than 40%. This shift to percentage terms provides a more insightful and standardized approach to evaluating teams' offensive efficiency. It allows recognition of teams that consistently excel in goal-scoring, regardless of the fluctuations in opportunities across different seasons. Ultimately, this method enhances the ability to identify patterns of success and assess team performance more comprehensively in the dynamic landscape of football analytics.

Incorporating percentage terms for overperformance analysis not only standardizes the assessment but also highlights teams' efficiency in converting opportunities into goals. It enables fair comparisons across seasons and teams, shedding light on consistent overperformers. This shift enhances the depth and accuracy of the analysis, providing a more holistic view of team performance in the realm of football analytics.

4.2.6 Shooting Skill and Precision

Besides the number of goals scored or many matches won, looking into the skills of the teams may be more meaningful, namely through computing the percentage of successful shots shown in the table below. The formula used to compute this metric is as follows:

$$\%\ Shots\ Success\ Percentage = \frac{\Sigma\ shots\_on\_target \times 100}{\Sigma\ shots}$$

Top 5 Teams	Shots Success Percentage
Tottenham Hotspur	36.98%
Manchester United	36.28%
Manchester City	34.67%
Liverpool	34.40%
Arsenal	32.97%

Table 4.2.6. Shots Success Percentage for Top 5 Teams over the 5 seasons (2019-2023).

4.2.7 Blocking Opponent Goals and Defense

This section focuses on exploring the percentage of goals successfully blocked or conceded by a team. This can be used as a measure of the skill of the team's goalkeeper or more generally the defense skill of the team as a whole. The defense success percentage over all seasons, shown in the table below, is computed by dividing the first column by the second one.

Top 5 Teams	Goals Conceded	Shots Attempted by Opponents	Defense Success Percent
Manchester City	129	1108	11.64%
Liverpool	152	1393	10.91%
Arsenal	182	1759	10.35%
Manchester United	190	1871	10.15%
Tottenham Hotspur	200	2080	9.62%

Table 4.2.7. Defense Success Percent of Top 5 EPL Teams (2019-2023).

Manchester City (11.64%) and Liverpool (10.91%) rank again as the top 2 teams in terms of defense. However, unlike what was previously found, Arsenal (10.35%) comes immediately after those two and followed up by Manchester United (10.15%) and finally, Tottenham (9.62%).

Another interesting way of looking at the table above is to look at the third column alone or the second alone. The *Shots Attempted by Opponents* column shows how many times opponents were attempting to score a goal. This can also be another indicator of good defense, where the lower the shots attempted by opponents, the better a team's defense is, not allowing the ball to reach the goal.

With this in mind, results similar to those of the Defense Success Percent are obtained where Manchester City had the lowest, followed by Liverpool, Arsenal, Manchester United, and lastly Tottenham Hotspur.

4.3 Hypothesis Testing

Research Question 1

Does the observed sample mean of 1.71 for the average goals scored by Arsenal, derived from a dataset, provide statistically significant evidence to reject the claim made on the Manchester City's Official Website (Cox, 2023) that the average number of goals scored by Arsenal is 2.31? Furthermore, what is the probability of obtaining sample means as extreme as 1.71 or more contradictory to the stated average of 2.31, assuming the true mean is indeed 2.31, and the sample size and variance remain constant across random samples of the same size?

Null Hypothesis (H0)

The null hypothesis H_0 assumes that the average number of goals scored by Arsenal is consistent with the claim made on Manchester City's Official Website, i.e., the mean is 2.31.

$$H_0 : \mu = 2.31$$

Alternative Hypothesis (H1)

The alternative hypothesis H_1 suggests that the observed sample mean of 1.71 indicates a significant departure from the claimed average, implying that Arsenal's actual goal-scoring average differs.

$$H_1 : \mu \neq 2.31$$

Result

The statistical analysis yielded a very small p-value (9.49×10^{-8}), providing strong evidence against the null hypothesis. The negative test statistic (-5.60) further indicates a deviation from the hypothesized mean. Additionally, the confidence interval $(1.50, 1.92)$ excludes the claimed mean of 2.31. Therefore, we reject the null hypothesis, concluding that the average number of goals scored by Arsenal, according to the dataset, significantly differs from the claimed value of 2.31.

Research Question 2

Is there compelling evidence in the dataset to support the claim that there has been a substantial increase in possession for Arsenal in 2022 compared to 2019? In other words, does the observed data provide statistical significance to reject the null hypothesis that the mean possession difference is μ_d equal to zero in favor of the alternative hypothesis that the mean possession difference μ_d is greater than zero?

Null Hypothesis (H0)

The null hypothesis H_0 assumes that there is no significant difference in possession for Arsenal between 2022 and 2019, meaning the mean possession difference μ_D is equal to zero.

$$H_0 : \mu_D = 0$$

Alternative Hypothesis (H1)

The alternative hypothesis H_1 posits that there is a significant increase in possession for Arsenal in 2022 compared to 2019, indicating that the mean possession difference μ_D is greater than zero.

$$H_A : \mu_D > 0$$

Result

Arsenal's possession (2022 vs. 2019) shows a mixed picture. The P-value (0.0321) hints at change, but the test statistic (-2.23) falls short of a significant increase. Therefore, there is insufficient evidence to support the claim of a significant increase in possession for Arsenal from 2019 to 2022.

5. Conclusions

In this research, an in-depth analysis of the English Premier League (EPL) dynamics from 2019 to 2023 was conducted, focusing on the performance of the Top 5 teams. The exploration delved into various metrics, including average goals scored, win percentages, home versus away performance, and defensive capabilities. Manchester City consistently emerged as a dominant force, leading in multiple performance indicators, while Arsenal showcased a notable rise in performance.

The results revealed a statistical difference in Arsenal's average goals scored, challenging a claim made on Manchester City's official website. However, the analysis did not support the hypothesis of a substantial increase in Arsenal's possession in 2022 compared to 2019. These findings emphasize the importance of nuanced metrics in understanding team dynamics, shedding light on both offensive and defensive aspects of performance.

The implications of the results extend beyond statistical comparisons, offering insights into the evolving landscape of football analytics. The research highlighted the significance of Expected Goals (xG), shooting efficiency, and defensive success percentages in evaluating team proficiency.

The research underscores the complexity of football dynamics and the need for a multifaceted approach to capture nuanced narratives.

To extend this research, future studies could delve deeper into the impact of individual player performance, tactical strategies, and external factors such as managerial changes on team dynamics. Additionally, incorporating advanced analytics and machine learning techniques could enhance predictive capabilities, providing a more comprehensive understanding of the intricate factors influencing Premier League team performance. Overall, this research contributes to the evolving field of football analytics and sets the stage for further exploration into the multifaceted aspects of team dynamics in elite football leagues.

6. References

Attwood, T., & Attwood, T. (2018, January 31). *Why Arsenal seems to get more negative publicity than other teams | Untold Arsenal: Supporting the club, the manager, and the team*. Untold Arsenal: Supporting the Club, the Manager, and the Team | "I Believe the Target of Anything in Life Should Be to Do It so Well That It Becomes an Art." a Wenger. https://untold-arsenal.com/archives/67017

Cox, S. (2023, May 29). *CITY'S 2022/23 PREMIER LEAGUE: STATS AND RECORDS*. Manchester City FC. https://www.mancity.com/news/mens/man-city-premier-league-stats-202223-63820961

7. Appendix

Appendix 1. Columns of the scraped data

- 'Unnamed: 0': Row number of data
- 'date': Date of the match
- 'time': Time of match
- 'comp': Competition name
- 'round': Matchweek number
- 'day': Day of the week
- 'venue': Where the game is played (Away or Home)
- 'result': W for Win, D for Draw, and L for Loss
- 'goals_scored': Number of goals scored
- 'goals_conceded': Number of goals conceded
- 'opponent': Opponent team name
- 'expected_goals': Expected goals to score
- 'expected_assists': Expected assists provided
- 'possession': Duration for which the team had the ball possession
- 'attendance': Count of audience
- 'captain': Captain of a team
- 'formation': Team formation
- 'referee': Referee name
- 'shots': Shots taken by team
- 'shots_on_target': Shots on target by team
- 'avg_shot_distance': Average shot distance
- 'freekicks': Freekicks taken
- 'penaltykicks': Penalties taken
- 'penaltykicks_opponent': Penalties given to the opponent
- 'season': Season year
- 'team': Team

Appendix 2. Links

- GitHub (Code .ipynb & HTML + Data): https://github.com/ghizzghiz/EPL---EDA-and-Hypothesis
- Video: https://drive.google.com/file/d/1ExOMMKZtLNReWVgpqHpQ3r5BhFPfxMeM/view?usp=share_link
- Slides: https://drive.google.com/file/d/1D_EWUoKtD3EiRwwHRdYO6gFC8sr6XAZK/view?usp=share_link