Analysis of Premier League Player Statistics for the 2023-24 Season

Team Members	
Kavit Mehta	Muhammad Aflah
Vignesh Gopal Rajendran	Rohith Gowda Ranganatha

Updates after Project Assignment 4:-

1. We received good feedback from our classmate, Wyl Schuth, who challenged the completeness of our dataset when using it to answer our Research Question 4: “How can we categorize teams based on their total goals scored and conceded?”. This research is supposed to be exploratory, and through our observations, we knew our dataset only contained data about the total goals scored by a team, not the ones conceded. But very recently, as of this last hour (May 8, 11:00 pm), we discovered another dataset on Kaggle named “**pl_table_2023_24” (.csv)** that contains data about the goals conceded by each team. Using that, we expanded our existing dataset and created derivative columns. We request that Professor Hutchins not deduct grades for late submission, as we are near the deadline and have put in a lot of effort.

1. Background

The Premier League is one of the most popular football leagues in the world, and analyzing player statistics provides a structured way to understand football performance. This dataset includes information on 580 players across 20 teams, capturing key performance metrics from the 2023-24 season. It contains both categorical data (such as player names, teams, and nationalities) and quantitative data (such as goals, assists, and expected goals), providing insights into player performance, team strategies, and league-wide trends. Additionally, this dataset is useful for practicing data cleaning and analysis techniques, as it offers a variety of data types and opportunities for visualization.

With football becoming more data-driven, we wanted to explore how stats like expected goals (xG), assists, and player positions reflect both team strategy and individual impact. We started by asking: “**Which teams had the highest expected goals, and how does that compare with their actual goals scored?”** This helps us see which teams are finishing their chances well and which ones aren't. Then, we looked at “**How player goals and assists differ across positions like defenders, midfielders, and forwards?”** Next, we identified **“Who are the top 10 players with the highest goal contributions?”** to highlight standout performers in the league. And lastly, we grouped teams into four categories based on how many goals they scored and conceded to better understand the balance between attack and defense. These questions help us explore trends on how individual performance (like scoring or assisting) fits into the bigger picture of team success and playing style.

2. Data cleaning & preparation

The data used for this project was obtained from Kaggle and consists of two datasets. The first contains individual player statistics for the 2023–24 Premier League season, including information such as player names, teams, positions, minutes played, goals, assists, and advanced metrics like expected goals (xG). The second dataset includes the final league table, providing each team's name, points, wins, draws, losses, and a string attribute representing both goals scored and conceded (e.g., "60-45").

To begin the analysis, we merged these two datasets using the team name as the key. During this step, we encountered missing values because three team names did not match exactly between the two files: "Brighton" vs. "Brighton & Hove Albion", "Bournemouth" vs. "AFC Bournemouth", and "Wolverhampton" vs. "Wolverhampton Wanderers". We manually mapped and corrected these names to complete the merge successfully. After the merge, we confirmed there were no remaining missing or duplicate entries that would affect our analysis.

We also created several new data fields to better support our research questions. First, we split the “scoresStr” column (which stored goals scored and goals conceded as a single string, e.g., "66-42") into two separate numeric fields: ‘Team_Goals_For’ and ‘Team_Goals_Against’. This allowed us to clearly compare offensive and defensive outputs across teams. Second, we created a derivative field called ‘**Player_Pos**’, which takes only the first-listed position from players with multiple positions (e.g., "MF, FW" becomes "MF"). This was necessary to assign each player to a single group for position-based comparisons. It was an effective way of overcoming multivariate field errors.

3. Research Questions (RQ)

3.1 RQ 1: Which teams had the highest average expected goals (xG), and how does that compare with their actual goals scored (Gls)?

This question explores whether teams that generate high-quality chances (as measured by expected goals, or xG) are actually able to convert those chances into goals. To answer it, we grouped the data by team and calculated the average xG and average actual goals scored (Gls) per player for each team.

The results showed that teams like Manchester City, Arsenal, and Liverpool were among the top performers in both average xG and actual goals scored, suggesting strong efficiency in converting chances. However, there were also teams where the gap between xG and actual goals was noticeable, indicating either poor finishing or overperformance relative to the quality of chances.

Based on this analysis, we conclude that while high xG generally correlates with more goals, the conversion efficiency varies across teams, highlighting the importance of finishing skill, tactical systems, and possibly luck.

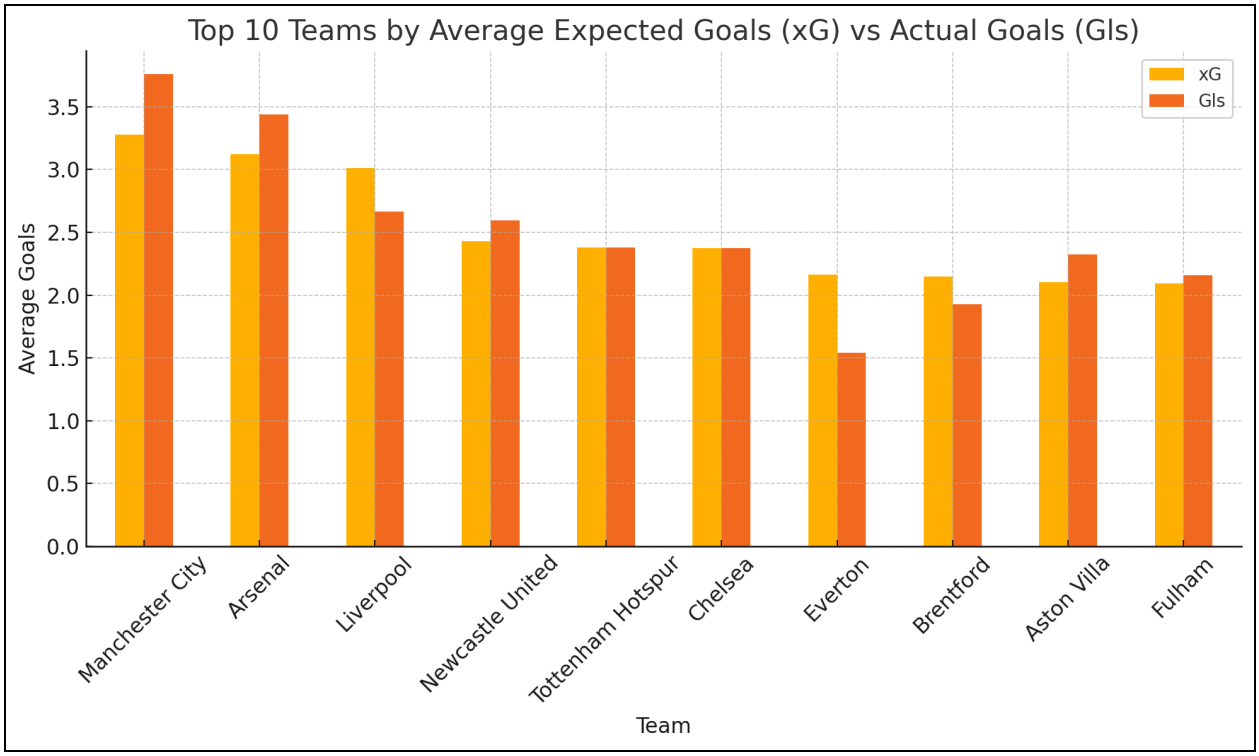

Fig. 1: Bar Chart Comparison of xG vs Gls

3.2 Research Question 2: How does player contribution (goals + assists) differ across positions (DF, MF, FW)?

This question investigates whether certain positions contribute more to attacking output than others. We focused on the total number of goals and assists (G+A) by each player, grouped by their primary position: defender (DF), midfielder (MF), or forward (FW).

The data showed a clear trend: forwards contributed the highest total G+A, followed by midfielders, with defenders contributing the least. To confirm whether the difference between midfielders and defenders was statistically significant, we ran a one-tailed t-test.

Null Hypothesis (H0): Midfielders and defenders have equal mean total G+A.
Alternative Hypothesis (H1): Midfielders have higher mean total G+A than defenders.
t-statistic: 4.29 | **p-value:** 0.0000132

Since the p-value is far below the 0.05 threshold, we reject the null hypothesis and conclude that midfielders contribute significantly more to total goals and assists than defenders.

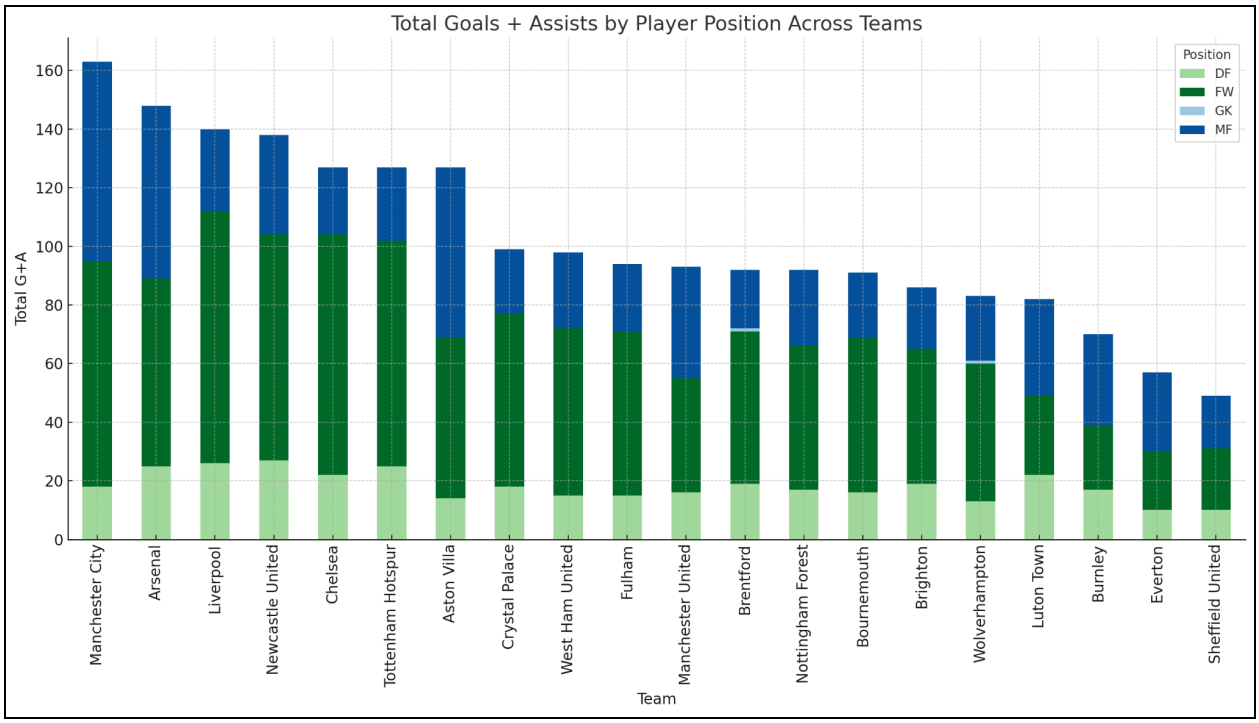

Fig. 2: Histogram Comparison of Gls and Ass

3.3 Research Question 3: Who are the top 10 players with the highest combined goals and assists?

This question highlights individual excellence by identifying the players who had the biggest attacking impact during the season. We sorted the dataset based on total G+A and selected the top 10 performers.

The resulting list featured expected names such as Erling Haaland, Mohamed Salah, and Ollie Watkins—players who consistently contributed to their teams' attacking output. These players often serve as focal points of their team's offensive strategy, and their high G+A numbers validate that role.

This analysis is useful for evaluating player value and consistency, and it aligns with how top-performing forwards and attacking midfielders typically dominate both scoring and playmaking metrics.

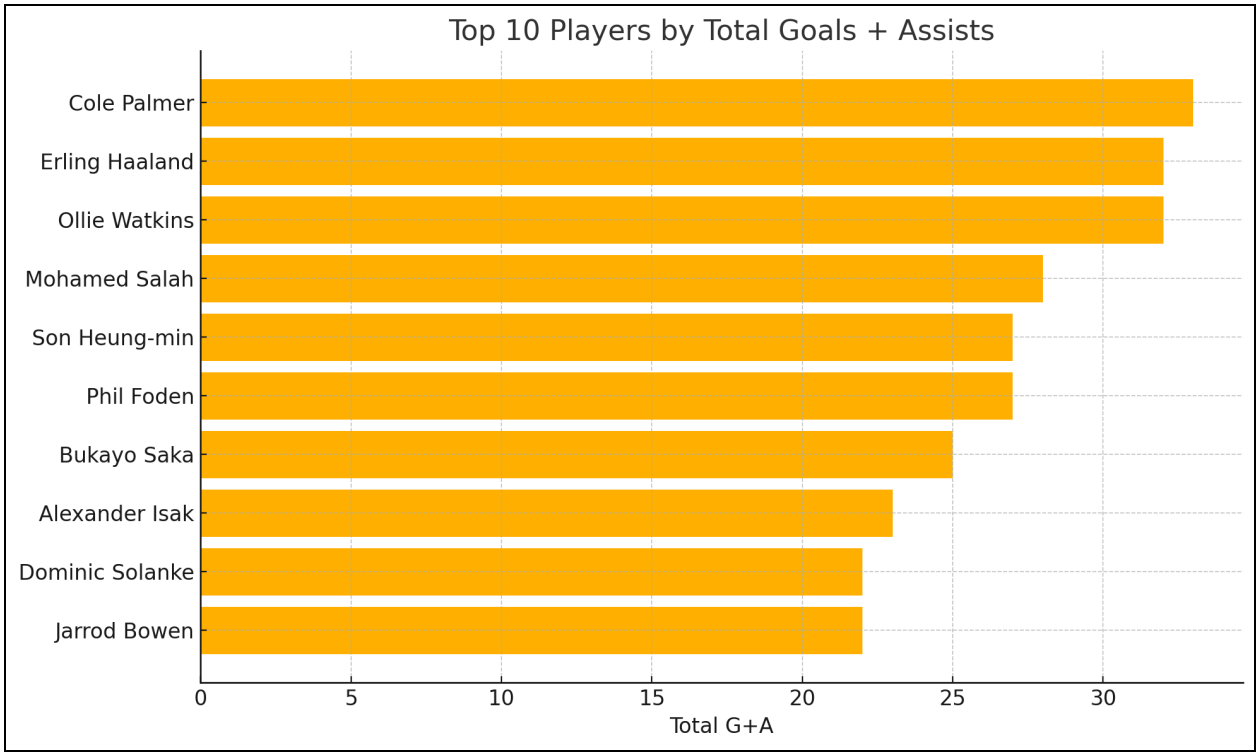

Fig. 3: Horizontal Bar Chart Comparison of players with most goal contributions

3.4 Research Question 4: How can we categorize teams based on their total goals scored and conceded?

This question explores how teams differ in style and performance by looking at both their offensive and defensive outputs. We used total goals scored (Team_Goals_For) and total goals conceded (Team_Goals_Against) for each team, and compared these values to the league averages. Based on this, each team was placed into one of four performance categories:

- **Good Attack, Good Defence**
  ('Team_Goals_For' > Avg 'Team_Goals_For' and 'Team_Goals_Against' < Avg 'Team_Goals_Against')

- **Good Attack, Bad Defence** ('Team_Goals_For' > Avg 'Team_Goals_For' and 'Team_Goals_Against' > Avg 'Team_Goals_Against')

- **Bad Attack, Good Defence** ('Team_Goals_For' < Avg 'Team_Goals_For' and 'Team_Goals_Against' < Avg 'Team_Goals_Against')

- **Bad Attack, Bad Defence** ('Team_Goals_For' < Avg 'Team_Goals_For' and Team_Goals_Against > Avg Team_Goals_Against)

To visualize this, we created a scatter plot where each team was plotted based on its goals scored and conceded. Quadrant lines were drawn using the league average values, and teams were color-coded and shape-coded according to their category, with a custom legend sorted by goal difference. This helped us clearly identify where each team falls—both tactically and in terms of overall performance. To further investigate whether teams that generate high-quality chances actually score more goals, we conducted a statistical test comparing the average goals scored (Gls) by teams with above-average xG versus those with below-average xG.

Null Hypothesis (H0): Teams with high and low xG have equal average actual goals scored.
Alternative Hypothesis (H1): Teams with high xG score more goals on average than teams with low xG.
Test Used: Independent one-tailed t-test

t-statistic: 3.22 | **p-value:** 0.0036

Since the p-value is below 0.05, we reject the null hypothesis and conclude that teams with higher expected goals do, in fact, score significantly more goals on average. This supports the idea that xG is a meaningful indicator of offensive performance at the team level.

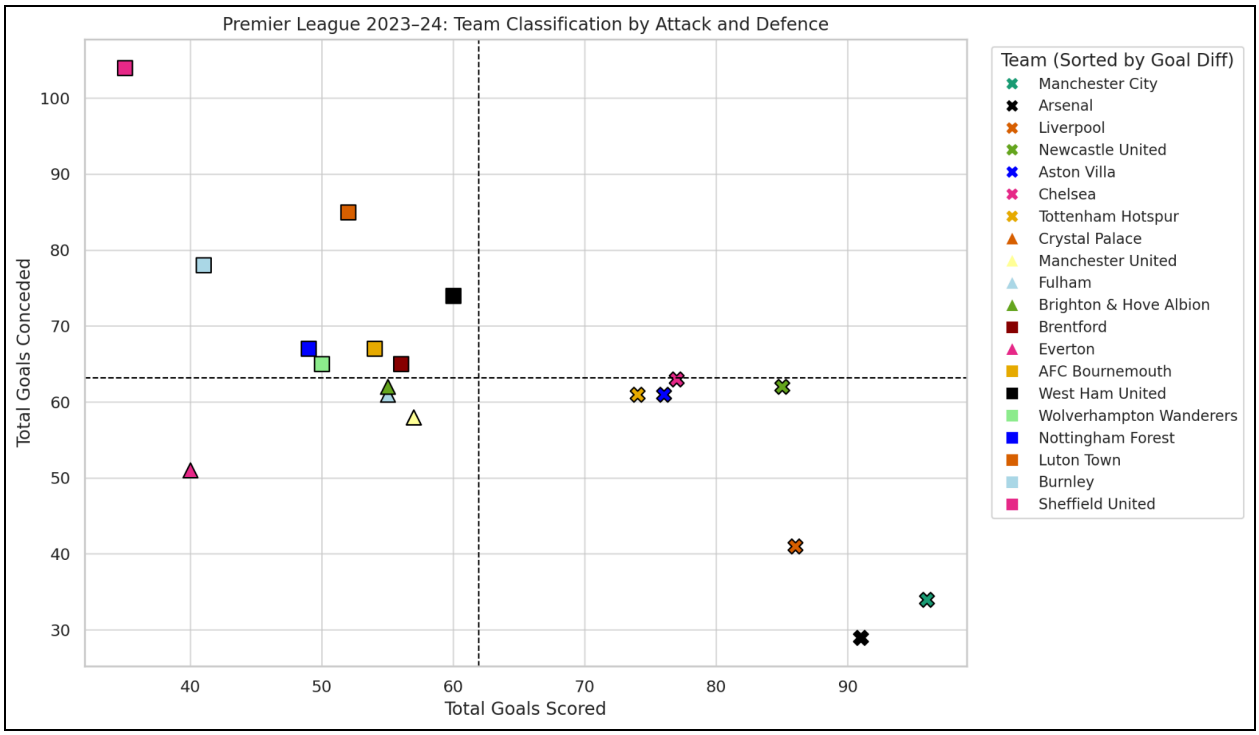

Fig. 4: Reimagined Scatterplot of goals scored vs goals conceded by each team

Conclusion:

Overall, this analysis provided a multi-level view of Premier League performance by examining both team-level efficiency and individual player contributions. We found that teams with high expected goals generally score more, but not always proportionally. Player position strongly influences goal involvement, with forwards and midfielders contributing significantly more than defenders. Lastly, the top-performing players in terms of goals and assists reflect both their skill and central role in their teams' attacking systems. Together, these findings offer valuable insights into how tactical roles and statistical efficiency intersect in modern football.