Exploratory Data Analysis of Premier League Match Statistics

Prafful Raj Thapa

***Abstract*—This paper presents an Exploratory Data Analysis (EDA) of English Premier League match statistics spanning from 1993 to 2024. The dataset comprises 24 features capturing various aspects of match outcomes and team performances. Through a series of insightful visualizations and statistical summaries, the analysis uncovers key patterns and trends, such as home vs. away performance advantages and correlations among critical features. Through a series of insightful visualizations and statistical summaries, the analysis uncovers key patterns and trends, such as home vs. away performance advantages and correlations among critical features.**

***Impact Statement*—This analysis contributes to the growing field of sports analytics by offering a detailed examination of historical Premier League data. The findings can aid coaches, analysts, and sports strategists in making data-driven decisions. Additionally, this project provides a foundation for predictive models and highlights how statistical insights can enhance the understanding of game outcomes, ultimately pushing the boundary of how data is used in competitive sports environments.**

***Index Terms*—Exploratory Data Analysis (EDA)**

I. Introduction

THE Premier League is the top tier of English football and is widely regarded as one of the most competitive and globally followed football leagues in the world. Since its inception in 1992, it has captured the attention of millions of fans and analysts alike, offering a rich source of data for performance evaluation, tactical assessment, and trend analysis.

Exploratory Data Analysis (EDA) serves as a crucial first step in understanding such complex and large-scale datasets. By systematically examining historical match statistics from 1993 to 2024, this project aims to uncover hidden patterns, evaluate performance metrics, and identify influential features that contribute to match outcomes. The dataset under analysis contains 24 features covering various match statistics, including goals scored, fouls committed, corners taken, and results categorized by home and away teams. This allows for a nuanced comparison of home versus away performance, as well as an investigation into the interplay between different match factors through correlation analysis and visualizations.

II. Dataset Description

The dataset used in this project encompasses detailed statistics from English Premier League matches spanning from the 1993 season up to 2024. It comprises approximately 12,153 rows and 25 columns, making it an ideal resource for analyzing the historical evolution and dynamics of the highest level of English football. Each row in the dataset represents a single match, and the columns capture a wide range of features, from basic information such as match date and teams involved to in-depth performance metrics. The features includes:

1) **Date:** The calender date on which the match was played.
2) **Season:** The football season during which the match took place, typically spanning two years(e.g. 2023/24).
3) **Home Team:** The team playing at their home stadium.
4) **Away Team:** The team playing away from home.
5) **Half Time Home(HTH) Goals:** Total goals scored by the home team at half-time.
6) **Half Time Away(HTA) Goals:** Total goals scored by the away team at half-time.
7) **Half Time(HT) Result:** Outcome at half-time, represented as 'H', 'A', or 'D'.
8) **Full Time Home(FTH) Goals:** Total goals scored by the home team at the end of the match.
9) **Full Time Away(FTA) Goals:** Total goals scored by the away team at the end of the match.
10) **Full Time(FT) Result:** Outcome of the match, similary represented as 'H' (home win), 'A' (away win), or 'D' (draw).
11) **Referee:** Name of the match official.
12) **H Shots:** Number of total shots attempted by the home team.
13) **A Shots:** Number of total shots attempted by the away team.
14) **H SOT:** Shots by the home team that were on goal/target.
15) **A SOT:** Shots by the away team that were on goal/target.
16) **H Fouls:** Total fouls committed by the home team.
17) **A Fouls:** Total fouls committed by the away team.
18) **H Corners:** Corner kicks awarded to the home team.
19) **A Corners:** Corner kicks awarded to the away team.
20) **H Yellow:** Number of yellow cards received by home team players.
21) **A Yellow:** Number of yellow cards received by away team players.
22) **H Red:** Number of red cards received by home team players.
23) **A Red:** Number of red cards received by away team players.
24) **Display Order:** A numerical index used for sorting or organizing the match records.
25) **League** The competition in which the match was played (e.g., Premier League).

III. Methodology

The dataset utilized in this analysis was sourced from Kaggle.com, a popular platform for publicly available datasets and data science competitions. The methodology followed a structured Exploratory Data Analysis (EDA) pipeline to derive meaningful insights and ensure data quality.

1) **Feature Identification** Each of the 25 features in the dataset was carefully reviewed to understand its significance and relevance to the analysis. Features were categorized based on their type (categorical or numerical) and grouped by their relationship to match events (e.g., performance metrics, disciplinary records, match results).
2) **Correlation Analysis** To identify the most influential features, a correlation matrix was computed for all numerical variables. This matrix was then visualized using a heatmap, highlighting relationships between features such as goals, shots, fouls, and cards. Strong correlations were used to guide further analysis and visualization priorities.
3) **Handling Missing Data** The dataset was examined for any missing or null values. Identified missing values were addressed through logical imputation—selecting appropriate replacement values based on domain knowledge and statistical reasoning to maintain data integrity.
4) **Feature Summarization** Summary statistics were computed for each feature to gain an initial understanding of distribution, central tendency, and variance. This step provided a foundational overview of the dataset and helped in spotting outliers or unusual patterns.
5) **Visualization and Pattern Recognition** A series of visualizations—including bar plots, line charts, and distribution plots—were created to uncover patterns across seasons, teams, and match conditions. These plots were essential for highlighting trends such as home vs. away advantages, scoring dynamics, and disciplinary behaviors over time.

This systematic approach ensured a thorough and insightful EDA, setting the stage for deeper statistical modeling and interpretation.

IV. Analysis & Result

The Exploratory Data Analysis conducted on Premier League match statistics revealed several key insights into team performances, scoring trends, and match outcomes.

A. Top Performing Teams (Home & Away)

1) Home Teams: Bar-plot Figure[1] shows a frequency analysis of matches played identified the top ten teams with the highest number of appearances, at their home stadium. These teams include **Man United, Tottenham, Arsenal, Chelsea, Everton, Liverpool, Newcastle, Aston Villa, West Ham, Man City**. Their frequent participation across multiple seasons indicate not only conistent performance but also historical dominance in the leagues.

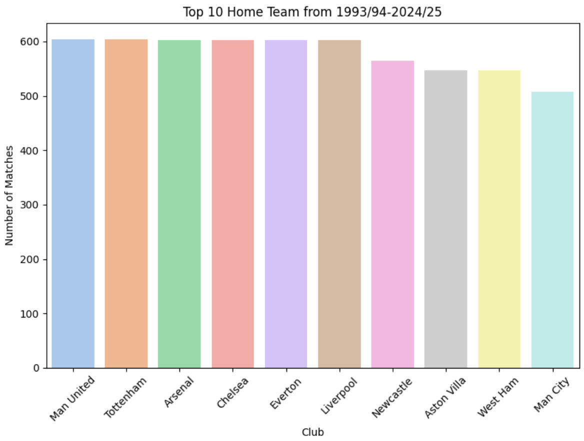

Fig. 1. Top 10 Home Teams

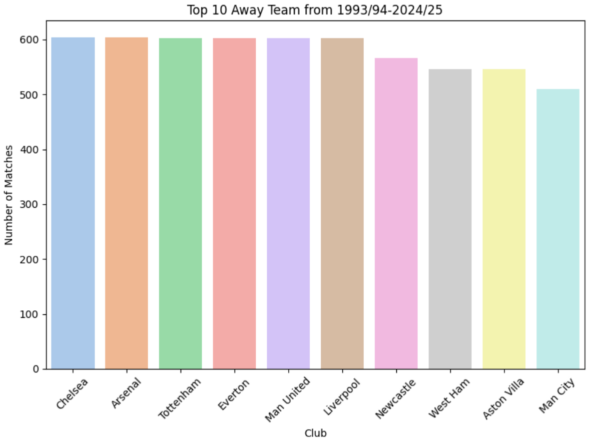

Fig. 2. Top 10 Away Teams

2) Away Teams: Bar-plot Figure[2] shows a frequency analysis of matches played identified the top ten teams with the highest number of appearances, at their opponent team stadium. These teams include **Chelsea, Arsenal, Tottenham, Everton, Man United, Liverpool, Newcastle, West Ham, Aston Villa, Man City**. Their frequent participation across multiple seasons indicate not only consistent performance but also historical dominance in the leagues.

B. Goal Statistics (Home & Away)

1) Home Team: Bar-plot Figure[3] visualizes the distribution of **Full Time Home (FTH) Goals** scored by the home team through the league. The x-axis represents the number of goals scored by the **home team** as full time. Whereas, y-axis represents the **frequency**(or count) of matches where the home team scored that number of goals.

Statistic

- **Arithmetic Mean:** 1.531, indicating a low average value overall.
- **Standard Deviation:** 1.306, showing how spread out the values are.
- **Min:** 0.0, Lowest goal scored by home team.
- **25%:** First quartile, 1.0, 25% of goals scored by home team is $\leq$ 1.0

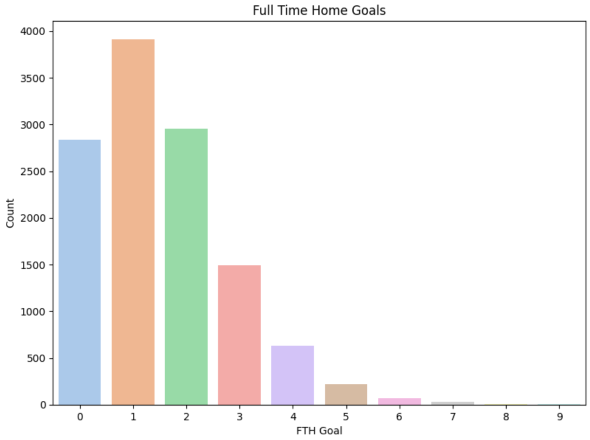

Fig. 3. Full Time Home(FTH) Goals

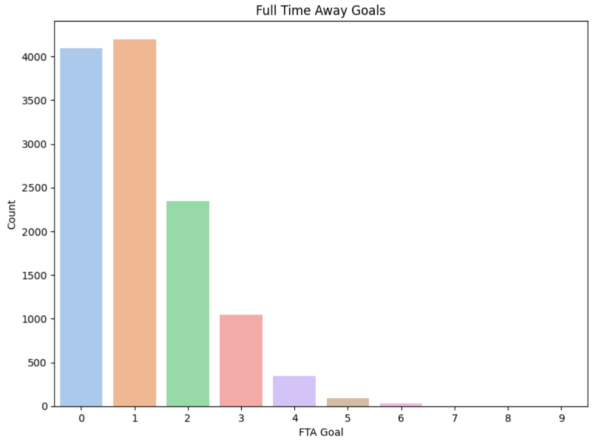

Fig. 4. Full Time Away(FTA) Goals

- **50%:** Median, 1.0, 50% of goals scored is $\leq$ 1.0
- **75%:** Third quartile, 2.0, 50% of goals scored is $\leq$ 2.0
- **Max:** 9.0, Highest goal scored by home team.

2) Away Team: Bar-plot Figure[4] visualizes the distribution of **Full Time Away (FTH) Goals** scored by the away team through the league. The x-axis represents the number of goals scored by the **away team** as full time. Whereas, y-axis represents the **frequency**(or count) of matches where the home team scored that number of goals.

Statistic

- **Arithmetic Mean:** 1.160, indicating a lower average value that home team goals mean.
- **Standard Deviation:** 1.147, showing low spread of the values that home team goals standard deviation.
- **Min:** 0.0, Lowest goal scored by away team.
- **25%:** First quartile, 0.0, 25% of goals scored by away team is = 0.0
- **50%:** Median, 1.0, 50% of goals scored is $\leq$ 1.0
- **75%:** Third quartile, 2.0, 50% of goals scored is $\leq$ 2.0
- **Max:** 9.0, Highest goal scored by away team.

To better understand the goal-scoring dynamics, **Difference Table**[I] was constructed comparing the frequency of full-time goals scored by home and away teams. This table revealed a consistent pattern of higher goal-scoring by home teams across most goal ranges, reinforcing the hypothesis of a **home advantage** in the Premier League.

TABLE I
DIFFERENCE TABLE COMPARING FTH GOALS & FTA GOALS

Goals	FTH Count	FTA Count	Difference(FTH - FTA)
0	2834	4093	-1259
1	3914	4194	-280
2	2954	2344	+620
3	1496	1043	+453
4	633	346	+287
5	217	94	+123
6	27	3	+24
7	2834	4093	-1259
8	7	2	+5
9	4	1	+3

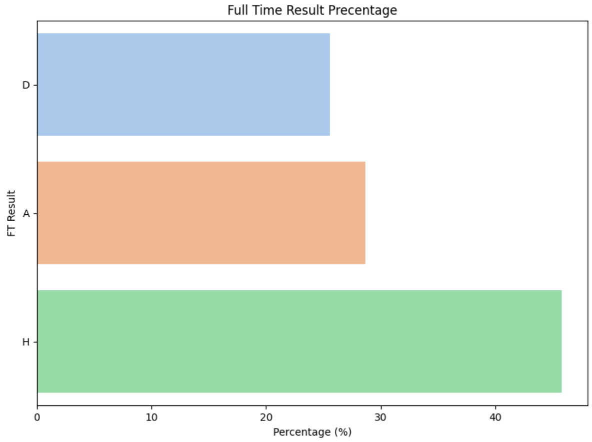

Fig. 5. Full Time Result Percentage (%)

C. Full-Time Match Results Distribution

The distribution of full-time match results figure[5] further confirms the existence of a home advantage:

- **45%** of matches resulted in a **home win**.
- **28%** of matches ended in an **away win**.
- **27%** concluded as a draw.

D. Feature Correlations

Figure[6] helps visualize the pairwise correlation coefficient between different football matches statistics. Here, each square shows the **correlation coefficient** between two variables, with 1 indicating perfect correlation, -1 with perfect negative correlation, and 0 indicating no correlation. Color ranging in the heatmap from red(highly positive) to blue(highly negative), with white(neutral) in between helps visualize the heatmap more easily.

Some interesting observations are:

1) **FTH Goals** $\leftrightarrow$ **H Shots** (0.67): More shots on target by home team leads to more home team goals.
2) **FTA Goals** $\leftrightarrow$ **A Shots** (0.68): Same logic for away team.
3) **HTH Goals** $\leftrightarrow$ **FTH Goals** (0.69): More goals at half time usually results in more at full time.

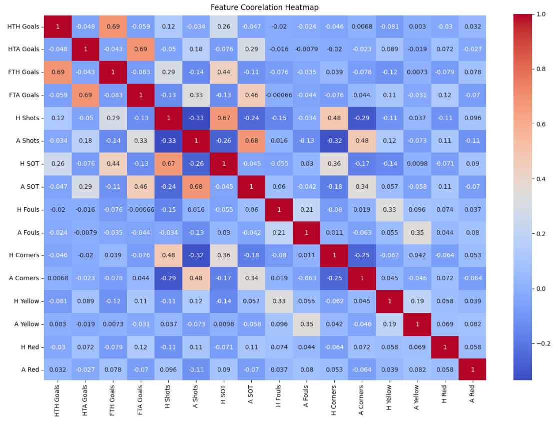

Fig. 6. Feature Correlation Heatmap

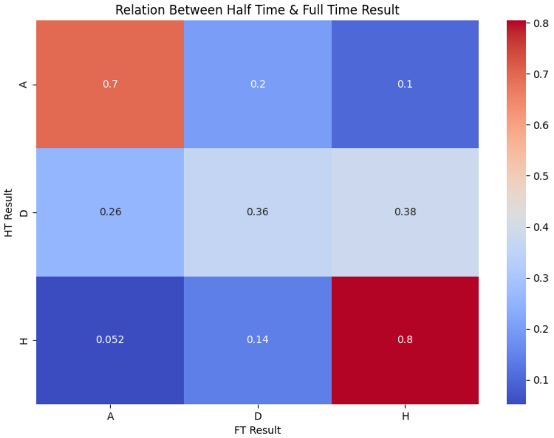

Fig. 7. Half Time(HT) & Full Time(FT) Result Relationship

4) **HTA Goals** ↔ **FTA Goals** (0.69): Similar for away team.
5) **H Shots** ↔ **H SOT** (0.67) : Expected, as more shots generally yields more shot-on-target.
6) **A Shots** ↔ **A SOT** (0.67) : Similar reasoning as for away team.
7) **H Shots/ A Shots** ↔ **H Corners/ A Corners** (0.48) : A bit connection between shooting opportunities and corners as most of the time an shot deflected by opponent team (Player/ Goal keeper) result in the team corners.

E. Half Time(HT) Result & Full Time(FT) Result Relation

Figure[7] visualize interesting relation between HT result and FT result. **Home Team** winning half-time, concludes 80% of the time in the **Home Team** winning full-time. Similarly, **Away team** winning half-time results, slightly lower 70% of the time also winning full-time. Additionally, it's relatively rare for the **Home Team** to lead at half-time but lose by full-time or vice-versa with such scenario occurring in approximately 10% of the time.

F. H Shots, A Shots vs FTH Goals, FTA Goals relation

Figure[8] highlights that there are hardly any correlation between H Shots vs A Shots, H Shots vs FTA Goals, FTH Goals vs A Shots, FTH Goals vs FTA Goals.

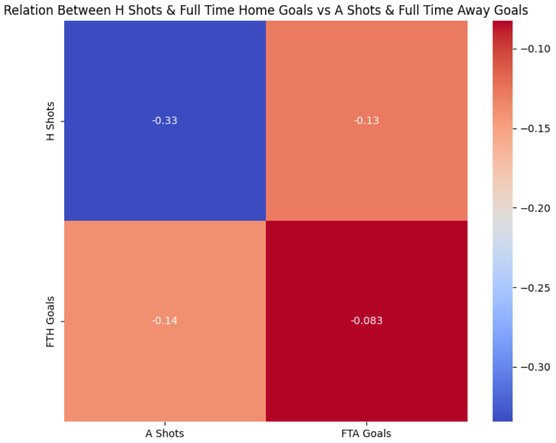

Fig. 8. Relation between H Shots, A Shots vs FTH Goals, FTA Goals

V. Discussion

Based on the analysis & result achived above, there are several discussion points that can be explored that prompt meaningful conclusions. Here are series of meaningful discussions.

- **Dominance of Legacy Clubs:**
  Teams like **Manchester United, Arsenal, Liverpool, Chelsea,** and others consistently appear at the top. This could lead to these legacy teams financial power and recruitment power among other teams.
- **Significant Home Advantage**
  The analysis **Difference Table**[I] confirms that home teams not only score more goals but also win more often. Factors that contribute to this advantage may be crowd support and psychological comfort players feel at their home stadium. Due to reduced to none travel fatigue home team perform at their best condition at their home stadium. Players familiarity with ground condition is also an factor that contribute to overall home advantage. Combining all these factors may significantly increases the chances of home team winning the game.
- **Higher Draw Rate**
  Around 27% of the matches ends in draw, which could indicate higher competitiveness on the league. Teams prioritizing not losing over wining also seem to contribute to this higher draw rate
- **Half Time Goals as a Predicate of Full Time Outcome**
  The correlations between half-time and full-time goals (HTH Goals ↔ FTH Goals: 0.69, HTA Goals ↔ FTA Goals: 0.69) suggest that teams leading at half-time are likely to maintain or extend their lead. This could prompt further exploration into momentum in matches, game management, and psychological factors influencing second-half performances.
- **Shots, Corners, Goals & Attacking Pressure**
  Strong positive correlation between shots and goals for both home(**FTH Goals** ↔ **H Shots** (0.67)) and away teams(**FTA Goals** ↔ **A Shots** (0.68)) reinforce the intuitive understanding that increasing shot volume increases probability of scoring.

The moderate correlation between shots and corners(**H Shots** ↔ **H SOT** and **A Shots** ↔ **A SOT**: both 0.67) suggests that team generating more offensive pressure often win more corners, possibly due to deflections and forced saves. This supports the idea that set-piece opportunities often arise from sustained attacking momentum.

VI. Conclusion

This Exploratory Data Analysis of Premier League match statistics from 1993 to 2024 provides valuable insights into team performances, scoring trends, and the significance of home advantage. The findings reveal a clear dominance of historically strong clubs, a consistent pattern of higher goal-scoring and win rates for home teams, and a balanced distribution of match outcomes. These insights not only highlight key dynamics of the Premier League but also lay the groundwork for further predictive modeling and deeper analysis in sports analytics.

Code Link: Premier League Google Colab

References

[1] English Premier League & Championship full dataset. (2025, January 28). Michael Panagopoulos. ‘https://www.kaggle.com/datasets/panaaaaa/english-premier-league-and-championship-full-dataset‘