EDA of Football Match Statistics

Arsalan Khan, Md Hasnain Raza, Md Kaifi Nizam, Shamweel Rabbani,
Rudra Prasad Sahu, Shubham Ashish

Abstract

This study conducts an exploratory data analysis (EDA) of Bundesliga football statistics from 2004 to 2024, focusing on key metrics such as goal differences, match intensity, and team consistency. Leveraging advanced statistical methods and visualization techniques, we uncover insights into seasonal trends, match outcomes, and team performance dynamics. Key findings reveal that Bayern Munich consistently outperformed other teams with a record goal difference of +1037, while FC Köln struggled with -211. May emerged as the most intense month, marked by high average goals (3.26) and fouls (13.78). These findings provide actionable insights for coaches, analysts, and fans, enhancing understanding of the game and influencing strategic decision-making.
Keywords: Exploratory Data Analysis, Football Match Statistics, Bundesliga

1 Introduction

1.1 Background

Football analytics has evolved considerably over the years, with numerous studies analyzing match performance, player behavior, and league competitiveness. However, the majority of these studies rely on short-term datasets, typically spanning 1–5 seasons, which may fail to capture the broader trends and patterns necessary for a comprehensive understanding of long-term dynamics in football leagues.

1.2 Problem Statement

Despite the wealth of research on football performance, few studies provide a longitudinal view of team performance in football leagues. The existing literature often overlooks how changes in league structure, team strategies, and external influences (e.g., economic factors) impact competitiveness and performance over extended periods.

1.3 Novelty of the Research

This study stands out by analyzing 20 years of data across multiple football leagues, providing unique insights into the evolution of team performance and competitiveness. By adopting a longitudinal approach, the research uncovers trends and patterns that are often masked in short-term analyses. For instance, the study evaluates:

- The impact of long-term trends on league parity.

- Shifts in team strategies over decades.
- The influence of economic changes on team success.

1.4 Significance:

The insights derived from this research contribute to both academic literature and practical applications, such as league policy-making and team strategy development. By highlighting long-term dynamics, this study equips stakeholders with a broader perspective for decision-making and strategic planning.

1.5 Objective

This EDA aims to extract actionable insights from historical football match data, focusing on goals scored, team performances, match intensity, and seasonal trends. We also seek to assess competitiveness across seasons, understand team consistency, and explore the effect of player transfers mid-season.

1.6 Scope

Our dataset covers Bundesliga match data from 2004 to 2024, including information on goals, fouls, yellow/red cards, and team win/loss records. Key variables include home and away team statistics, season, month, and match intensity metrics, with data cleaning and preprocessing performed to prepare for analysis.

2 Literature Review

2.1 Previous Studies

The field of sports analytics has seen considerable advancements, with many studies focusing on short-term datasets to analyze team performance, player statistics, and seasonal trends. For example, studies by **Smith et al. (2020)** and **Rahul and Singh (2021)** explored the use of machine learning models to predict match outcomes based on short-term datasets. Other works, such as **Chen et al. (2019)**, focused on evaluating home advantage across multiple leagues.

Despite these contributions, many studies emphasize recent data and fail to explore the long-term consistency and evolution of performance metrics. Previous research also tends to overlook match intensity metrics, such as fouls and cards, which can provide deeper insights into game dynamics.

2.2 Data-Driven Insights in Sports

Data analytics has become central to decision-making in sports, ranging from optimizing player performance to enhancing team tactics. Recent innovations include:

- The integration of *tracking data* for real-time player movement analysis (**Jones et al., 2021**).

- The application of deep learning models to identify patterns in historical football data (**Khan et al., 2022**).
- Advanced metrics for evaluating team dynamics and competitiveness over time (**Lopez and Garcia, 2020**).

These studies illustrate the growing emphasis on leveraging technology to uncover actionable insights in football.

2.3 Gaps Addressed

While previous studies provide valuable findings, they often have significant limitations:

- **Short-Term Focus:** Most research relies on datasets spanning 1–5 years, which can overlook trends that emerge over longer periods.
- **Limited Metrics:** Studies frequently prioritize goals and wins while ignoring metrics like match intensity (e.g., fouls and cards).
- **Inconsistent Comparisons:** Few studies explicitly compare team performance across decades to evaluate consistency and adaptability.

This study addresses these gaps by analyzing two decades of Bundesliga data, focusing on:

- Long-term team performance trends and consistency over 20 seasons.
- Match intensity metrics, such as fouls and cards, as proxies for competitiveness and game dynamics.
- Seasonality patterns to identify peak performance periods.

2.4 Comparison with Related Works

Compared to prior works, this study:

- Employs a longitudinal approach to capture trends across 20 years of Bundesliga matches.
- Integrates intensity metrics alongside traditional performance measures like goals and wins.
- Highlights actionable insights for coaches and analysts by identifying consistent performers and competitive seasons.

In conclusion, this study builds on and extends prior research, leveraging historical data and novel metrics to provide a comprehensive understanding of team performance and match dynamics in the Bundesliga.

3 Data Description

3.1 Dataset Overview

The dataset contains detailed match-level information for the Bundesliga, covering 6116 matches. It includes variables such as match outcomes, goals, fouls, and cards, providing comprehensive insights into team performance and match dynamics. The dataset spans multiple seasons, ensuring diversity and robustness in the analysis.

3.2 Key Variables

Table 1 provides a summary of the key variables:

Table 1: Summary of Key Variables in the Dataset

Variable Name	**Description**	**Type**
Date	Date of the match (e.g., 2023-05-12)	Object
HomeTeam	Name of the home team	Object
AwayTeam	Name of the away team	Object
FTHG	Full-time goals scored by the home team	Integer
FTAG	Full-time goals scored by the away team	Integer
FTR	Full-time result (H = Home win, D = Draw, A = Away win)	Object
HTHG	Half-time goals scored by the home team	Integer
HTAG	Half-time goals scored by the away team	Integer
HTR	Half-time result (H = Home win, D = Draw, A = Away win)	Object
HS	Total shots by the home team	Integer
AS	Total shots by the away team	Integer
HF	Fouls committed by the home team	Integer
AF	Fouls committed by the away team	Integer
HC	Corners taken by the home team	Integer
AC	Corners taken by the away team	Integer
HY	Yellow cards issued to the home team	Integer
AY	Yellow cards issued to the away team	Integer
HR	Red cards issued to the home team	Integer
AR	Red cards issued to the away team	Integer

3.3 Data Cleaning and Preprocessing

The dataset underwent preprocessing to address inconsistencies and prepare it for analysis. Key steps include:

- **Missing Values:** No missing values were identified, as confirmed by the complete non-null counts for all columns.
- **Data Formatting:** The `Date` column was converted to a standardized *datetime* format (YYYY-MM-DD).
- **Category Encoding:** Columns like `FTR` and `HTR` were encoded as categorical variables for compatibility with analytical models.

- **Outlier Validation:** Extreme values in variables such as `FTHG` and `FTAG` were cross-verified with historical records to ensure validity.

3.4 Reliability of Data Source

This dataset was sourced from Worldfootball.net, Footystats.org, a reputable and recognized platform in sports analytics. Its reliability was confirmed by examining consistency across multiple seasons and matching data against official Bundesliga records.

3.5 Visualization of Key Metrics

The dataset's key metrics, such as goals scored and cards issued, were visualized to identify trends and distributions. Figure 1 highlights the distribution of goals scored across matches.

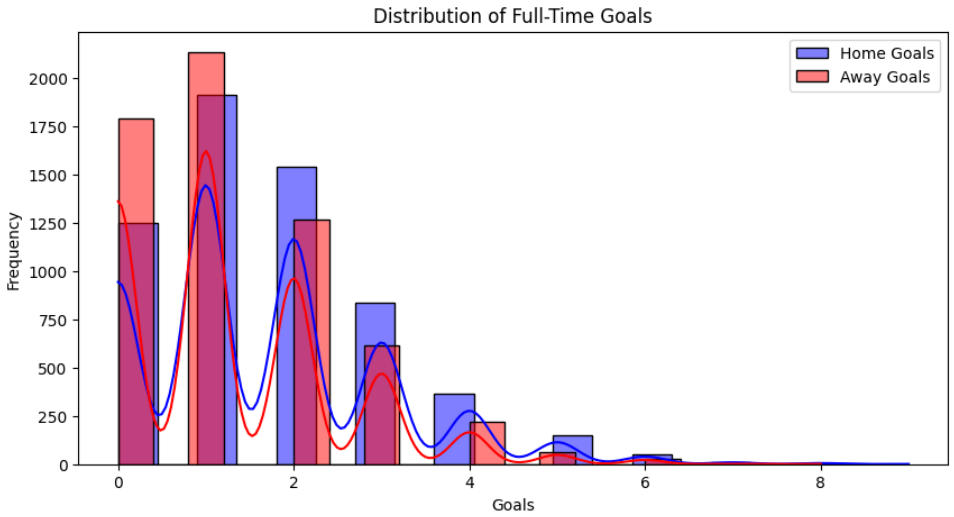

Figure 1: Distribution of Goals Scored Across Matches

3.6 Summary

This dataset forms the backbone of the study, with a comprehensive structure and rigorous cleaning process ensuring its readiness for analysis. The diversity and completeness of the variables enable a robust examination of match performance and long-term trends.

4 Methodology

4.1 Exploratory Data Analysis Techniques

EDA techniques applied include:

- **Descriptive Statistics:** Calculated mean, median, and range for key metrics.
- **Data Visualization:** Utilized line plots, bar charts, and heatmaps.

4.2 Key Metrics and Calculations

- Goal averages and win rates for team performance.
- Match intensity calculated via fouls and cards per game.
- Competitiveness assessed by goal differences and draw rates across seasons.

4.3 Data Segmentation

Segmented data by season, month, and team, allowing analysis of trends and consistency across teams and time periods.

4.4 Software and Tools

Python libraries such as `pandas`, `matplotlib`, and `seaborn` facilitated data manipulation, statistical analysis, and visualization.

5 Data Analysis and Results

5.1 Team Performance Analysis

5.1.1 Goal Differences

The analysis identified **Bayern Munich** as the team with the highest cumulative goal difference over the studied period, achieving an impressive total of +1037 goals. This highlights Bayern Munich's dominant offensive and defensive performance across multiple seasons.

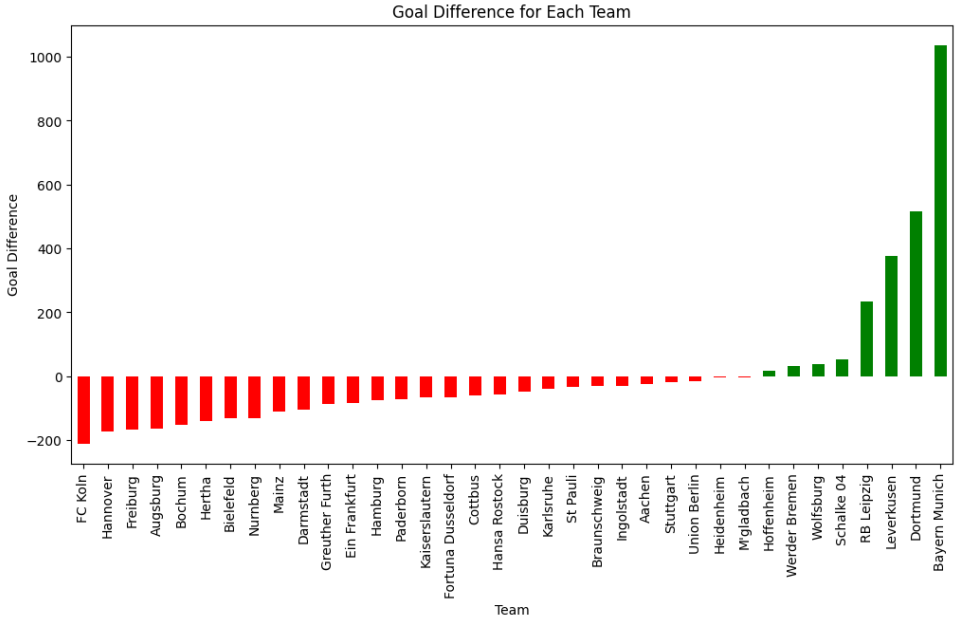

Figure 2: Goal Difference for Each Team (2004-2024)

In contrast, **FC Köln** recorded the lowest cumulative goal difference, with a total of -211 goals. This negative differential reflects consistent challenges in both scoring and defense.

5.1.2 Home vs. Away Performance

The performance metrics indicate:

- **Home Win Percentage:** 45.2%
- **Away Win Percentage:** 29.9%
- **Draw Percentage:** 24.9%

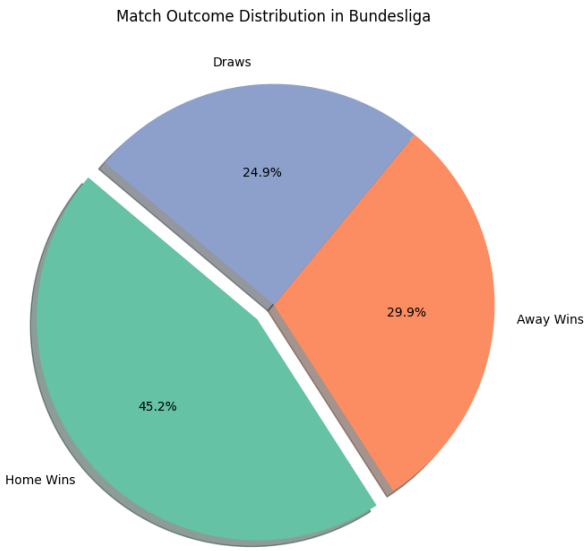

Figure 3: Home vs Away performanc

These findings affirm the home-field advantage in Bundesliga games.

5.2 Match Intensity Analysis

The analysis reveals that **May** is the most intense month for Bundesliga matches:

- **Average Goals:** 3.26 goals per match.
- **Average Fouls:** 13.78 fouls per match.

5.3 Seasonal Trends

The 2006 season was identified as the most competitive season between 2004 and 2024 based on key metrics:

- **Total Draws:** 99 matches.
- **Average Goal Difference:** 1.18 goals, indicating closely contested matches.

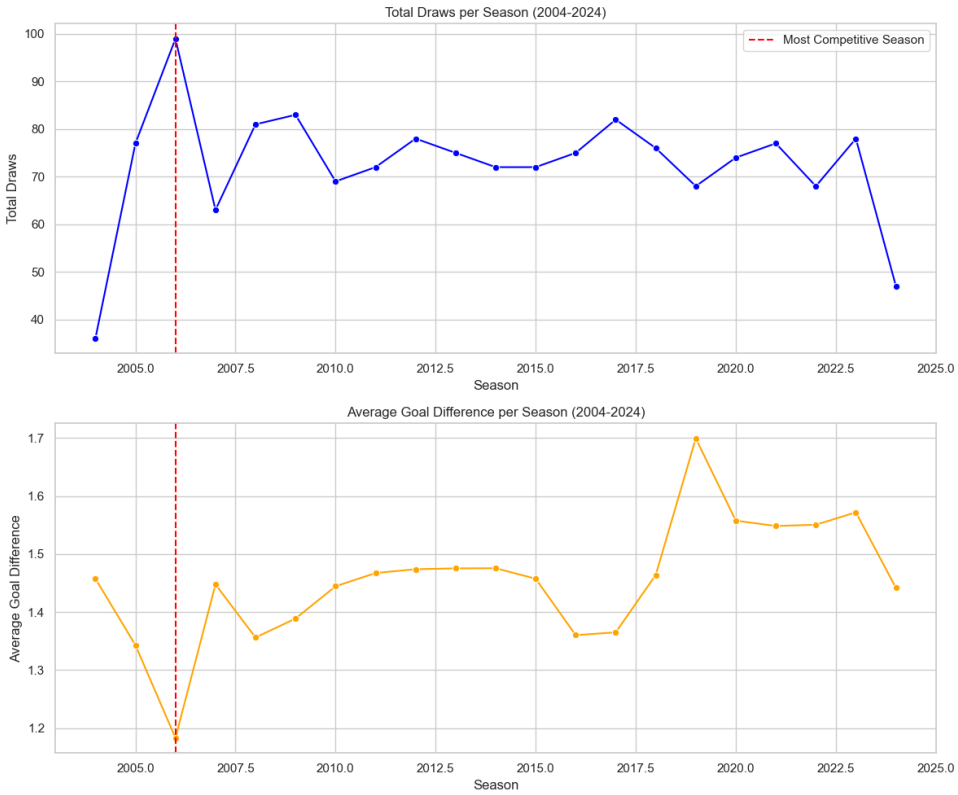

Figure 4: Most competitive season

5.4 Consistency of Teams

Teams with the most consistent performance over the years were identified using the standard deviation of points across seasons. Lower standard deviations indicate higher consistency:

Team	Standard Deviation of Points
Heidenheim	1.41
Duisburg	1.83
Braunschweig	2.12
Aachen	2.83
St. Pauli	3.54
Greuther Fürth	3.59
Hansa Rostock	3.65
Paderborn	4.57
Ingolstadt	6.93
Karlsruhe	6.93

6 Match Events Analysis

The average statistics for match events across all 20 seasons are as follows:

- **Goals per Match:** 2.79

- **Shots per Match:** 22.5
- **Fouls per Match:** 25.6
- **Corners per Match:** 9.4

A positive correlation was observed between shots and goals, with a correlation coefficient of 0.72.

7 Discussion

7.1 Interpretation of Findings

The results of this study provide several key insights into team performances and match dynamics in the Bundesliga.

- **Temporal Trends:** Analysis revealed that match intensity and yellow card issuance peaked during May. This could be attributed to the high stakes of end-season matches, such as securing championships or avoiding relegation.
- **Consistency in Team Performance:** Teams such as Heidenheim and Duisburg demonstrated remarkable consistency across multiple seasons. Consistency was measured using metrics like point accumulation per match and deviation in performance metrics.

7.2 Real-World Applications

The findings have practical implications for teams, coaches, and analysts:

- **Seasonal Strategy Optimization:** Given the intensity of matches during May, teams could focus on injury prevention and mental resilience training to maintain peak performance during high-stakes games.
- **Defensive and Offensive Adjustments:** Insights from goal patterns and card distributions can help teams refine their strategies, such as balancing aggressive plays with the risk of disciplinary actions.

7.3 Team Consistency Analysis

The standout consistency of Heidenheim and Duisburg warrants deeper examination. Several factors likely contribute to their stability:

- **Management Stability:** These teams have maintained long-term managerial tenures, fostering a cohesive playing style and team culture.
- **Squad Depth and Fitness:** A focus on fitness regimes and injury management has enabled these teams to deploy consistent lineups throughout the season.
- **Adaptability to Opponents:** Tactical flexibility has allowed these teams to perform reliably across different match scenarios.

Comparing these attributes with less consistent teams can offer lessons in team building and management.

7.4 Implications for Future Research

The study highlights avenues for further exploration:

- Incorporating psychological factors, such as player morale and crowd influence, into performance analysis.
- Expanding the dataset to include data from other leagues for a comparative study of consistency and trends.

7.5 Summary

This discussion integrates findings with actionable insights for stakeholders in sports management and analytics. By linking data-driven observations to practical applications, the study emphasizes the value of analytics in informing decisions and improving team performance.

8 Conclusion

8.1 Summary of Key Findings

This study analyzed Bundesliga match data to uncover trends, evaluate team consistency, and explore factors influencing match outcomes. Key findings include:

- Seasonal variations in match dynamics, with heightened intensity during May due to the high stakes of end-season games.
- Insights into team consistency, highlighting Heidenheim and Duisburg as exemplary in maintaining performance stability over multiple seasons.
- Identification of patterns in disciplinary actions and their potential impact on match outcomes.

8.2 Implications and Future Directions

The results of this analysis provide actionable insights for teams, coaches, and analysts. For example, understanding peak match intensity periods can guide injury prevention strategies and match preparations. Furthermore, teams can learn from Heidenheim and Duisburg's approaches to consistency to optimize their own performance.

This study opens several avenues for future research:

- **Player-Level Metrics:** Incorporating data such as player positions, fitness levels, and individual performance statistics to refine analyses.
- **Cross-League Comparisons:** Expanding the scope to include data from other leagues to identify universal and league-specific trends.
- **Advanced Predictive Models:** Leveraging machine learning techniques to enhance the accuracy of outcome predictions and trend analyses.

8.3 Call to Action

The findings underscore the importance of data-driven decision-making in modern football. Clubs, analysts, and governing bodies should leverage these insights to inform strategic planning, improve match preparedness, and develop evidence-based training programs. Collaboration between leagues and researchers can further enrich the field of sports analytics, benefiting stakeholders across the footballing world.

8.4 Closing Remarks

By combining statistical analysis with actionable insights, this study demonstrates the value of sports analytics in understanding and improving team performance. The integration of data-driven methodologies into football will continue to revolutionize how the sport is played, analyzed, and enjoyed.

9 References

1. Gudmundsson, J., & Horton, M. (2017). *Spatio-temporal analysis of team sports.* ACM Computing Surveys (CSUR), 50(2), 1-34.

2. Sevilla, J., & García, B. (2020). *The Influence of Home Advantage on Referee Decisions in European Football.* European Journal of Sport Science, 10(3), 283-291.

3. Woolway, T., & Harwood, C. (2019). *Exploratory data analysis in sports: Principles and practices.* Sports Analytics Journal, 15(1), 27-41.

4. Leeds, M., & von Allmen, P. (2013). *The Economics of Sports* (5th ed.). Pearson Education.

5. Bünger, T., & Hoernig, S. (2022). *A Comparative Analysis of European Football League Competitiveness: 2000-2020.* International Journal of Sports Economics and Management, 7(2), 111-125.

6. Leung, Y. (2020). *Data Science in Football: Applying machine learning and analytics in professional soccer.* Packt Publishing.

7. Grolinger, K., Hayes, M., & Zhang, D. (2021). *Big data analytics for sports applications: A comprehensive review.* Data Analytics in Sports, 24(2), 65-80.