Department of Computer and Information Science - Linköping University

&

Mathematical Engineering and Data Science - Polytech Clermont

Internship Report

Modeling Team Playing Styles Based on Match Location

Home vs Away Performance Analysis

Author :
Ayman Zejli

Supervisor :
Patrick Lambrix

Jury :
Christophe de Vaulx
François Bouchon

Academic Year : 2024-2025

Acknowledgments

First and foremost, I wish to express my profound gratitude to **Mr. Lambrix** for his exceptional supervision, unwavering patience, and steadfast support throughout the duration of this research project. From the initial stages of this investigation, Mr. Lambrix demonstrated remarkable availability and attention to my research proposals, inquiries, and methodological challenges. We maintained a schedule of regular meetings during the internship period, during which he consistently provided perspicacious feedback and constructive recommendations that proved instrumental in maintaining the project's trajectory and ensuring continuous progress. His supervisory approach skillfully balanced guidance with academic freedom, fostering an environment conducive to intellectual exchange while maintaining rigorous academic standards. I particularly appreciated his professional expertise, prompt response, and confidence he placed in my work throughout all phases of this research endeavor. The mentorship of Mr. Lambrix constituted a fundamental contribution to both the conceptual development and the scholarly quality of this work. His support was indispensable in shaping the research methodology and analytical framework presented in this report.

I would like to express my sincere appreciation to **Linköping University**, the **Department of Computer and Information Science**, and all the staff who welcomed me warmly and supported my academic journey during this internship abroad. I would also like to thank him for the *Sport Analytics* course, which gave me a broader perspective on different sports and inspired several ideas that I applied in this research. I also gratefully acknowledge all those who contributed indirectly to this work through insightful discussions, technical advice, or their presence when needed.

Finally, I extend my warmest thanks to **Mr. de Vaulx**, my academic tutor, for his regular follow-ups during my internship and for his role in facilitating all administrative aspects of this international academic experience.

Contents

List of Figures

List of Tables

Introduction

Context and Motivation

Football has evolved into a highly data-driven sport, where understanding team behavior and tactical trends is increasingly supported by advanced analytics. In this context, **sports performance modeling** and **tactical analysis** have become essential tools for both clubs and analysts seeking to gain a competitive edge. The ability to translate raw match data into meaningful insights allows coaches, scouts, and decision-makers to optimize strategies, adapt to opponents, and evaluate performance more objectively.

One key contextual factor that has consistently influenced match outcomes is **match location** whether a team plays at home or away. The concept of *home advantage* is well documented in both academic research and mainstream football discussions. It typically encompasses a mix of psychological, physical, and situational elements such as fan support, travel fatigue, and familiarity with the playing environment. However, despite its importance, there remains a gap in understanding exactly *how* and *to what extent* this context shapes a team's tactical identity and in-game behavior throughout an entire season.

In recent years, the growth of publicly available football data and the development of machine learning tools have opened up new possibilities to study these questions with more precision. Yet, many existing studies focus primarily on outcome-based metrics (like goals or win probability) rather than exploring deeper tactical shifts that may occur depending on the match context. This project addresses that gap by proposing a systematic approach to modeling and analyzing how teams adapt their playing style in response to different situational conditions, with a particular focus on the contrast between playing at home and playing away.

Project Overview

This project was carried out within the framework of an internship, in collaboration with **Linköping University** (LiU), and focused on the **modeling of team playing styles based on match location**, using real match data from the 2022–2023 English Premier League season. The objective was to identify and understand the tactical differences exhibited by teams depending on whether they play at home or away, through a combination of `data science`, `performance analysis`, and `machine learning` techniques. The project offers a structured example of how contextualized match data can be turned into tactical insights, illustrating how data-driven methods help reveal meaningful variations in team behavior across different match conditions. This approach not only supports deeper academic research but also has practical implications for *coaches*, *video analysts*, *practitioners* who seek to develop models for team structure and playing styles based on data-driven definitions of role types and `player behaviors`, aiming to create insights that are both clear and actionable in specific game contexts to optimize `performance`.

Methodological Approach

The approach adopted was both technical and analytical. It involved the implementation of a complete `ETL (Extract, Transform, Load)` pipeline, followed by the design of a methodology combining `ELO ratings`, `clustering`, and `classification models` to analyze performance patterns. The project's objectives were to create a clear and useful way to analyze how teams play differently at home and away, to find groups of similar matches, and to understand which factors influence team tactics the most. To achieve this, several advanced methods were employed: the `ELO rating system` was adapted as a dynamic measure of team strength and context; unsupervised learning via `PCA` and `KMeans clustering` allowed for the discovery of natural groupings within matches; and supervised machine learning models, notably `Random Forests`, provided interpretable classification and feature importance insights. Visualization tools such as `Matplotlib` or `Streamlit` were also used to support the exploratory and explanatory phases of the analysis. This work adds to the field of `sports analytics` by combining the idea of team strength with machine learning to better understand football tactics. It shows how classic methods like `ELO ratings` can work well together with modern data analysis to give deeper insights about how teams adapt their play depending on the situation.

Structure of the Report

The structure of this report is organized to follow these ideas. It begins by describing the data preparation, preprocessing and the reason for focusing on the 2022–2023 Premier League season. It then explores the modeling of home and away tactical behaviors, before introducing the use of ELO ratings as a contextual variable. Subsequent chapters delve into clustering and classification methodologies applied to football matches, highlighting key tactical patterns and providing clear explanations of their significance. Finally, the report concludes by synthesizing the findings and discussing potential improvements for future research and applications in football data analytics.

1. Discovering Linköping University

A Research-Driven and Interdisciplinary Institution

Linköping University (LiU) [1] is a leading Swedish public university, founded in **1975**, and widely recognized for its forward-thinking educational model and strong focus on interdisciplinary collaboration. Located in southern Sweden, LiU has grown to become one of the most innovative academic institutions in Scandinavia.

The university operates across **four campuses** in three cities, each with its own academic specializations:

- **Campus Valla** (Linköping) – the largest and main campus, home to most departments in engineering, sciences, and teacher education.
- **Campus US** (Linköping) – focused on medicine and health sciences, located next to the Linköping University Hospital.
- **Campus Norrköping** – offering programs in media technology, logistics, social sciences, and environmental engineering.
- **Campus Lidingö** (Stockholm) – dedicated to the Carl Malmsten Furniture Studies and specialized design education.

Figure 1 below provides an overview of the geographical locations of the university's campuses across Sweden.

Figure 1: Geographical Locations of Linköping University's Four Campuses

What sets LiU apart is its unique organizational model. Instead of traditional faculty hierarchies, LiU is composed of multidisciplinary departments that report directly to the Vice-Chancellor.

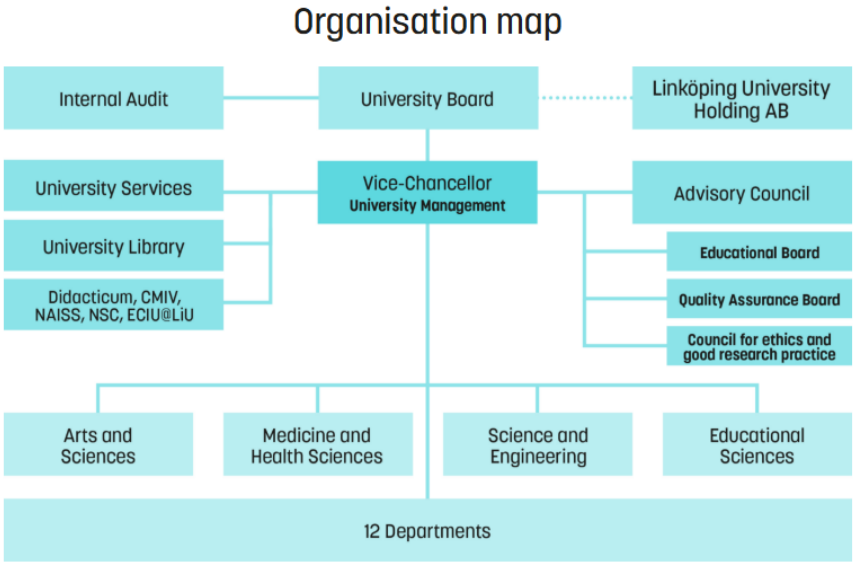

Figure 2: Organizational Structure of Linköping University

This governance model facilitates cross-disciplinary research and encourages innovative approaches to complex societal challenges. The overall structure is illustrated in the Figure 2 above.

The internship took place at **Campus Valla** [2] in Linköping, which is the heart of the university's science and engineering activities. The campus features modern educational facilities, student spaces, and a vibrant academic environment. Figure 3 presents a visual of the campus where the internship was conducted.

Figure 3: Campus Valla – Main Science and Technology Campus of Linköping University

About the Host Department: IDA

My internship took place at the **Department of Computer and Information Science (IDA)** at Linköping University, one of the largest and most prestigious computer science departments in Northern Europe. Founded in 1983, IDA hosts around 300 staff and is active in both teaching and advanced research.

IDA's research is structured into four main areas:

- **Artificial Intelligence (AI)**: autonomous systems, knowledge representation, semantic web, machine learning, natural language processing, and robotics.
- **Cognition, Interaction, and Design**: human-computer interaction, service design, cognitive systems, and public safety technologies.
- **Software and Computer Systems**: embedded systems, cybersecurity, real-time computing, distributed systems, and programming languages.
- **Data Science**: statistics, database technologies, and scalable data analytics.

The Sports Analytics Research Group at IDA

During my internship, I joined the **Sports Analytics Research Group**, an interdisciplinary team **founded in 2017**, focused on applying data science to sports.
Since 2019, the group has offered an annual **Sports Analytics course** for students in **computer science, AI, or statistics**. I had the opportunity to attend this course, which significantly enriched my understanding by exposing me to applications of sports analytics beyond my primary focus, including analyses from various sports domains. The group also organizes the **Linköping Hockey Analytics Conference (LINHAC)**, a key event in sports data science, coordinated by **Prof. Patrick Lambrix** and **Assoc. Prof. Niklas Carlsson**, experts in ice hockey analytics. This engaging environment allowed me to apply computational methods to real sports data, further strengthening my interest in the field of sports analytics.

Sustainable Development and Social Responsibility at Linköping University

During my time at Linköping University (LiU), I engaged with and observed various real-world sustainability and social responsibility activities. These initiatives spanned three areas: student life, research practices, and institutional efforts.

Sustainability as a Student

For everyday travel, I rode a second-hand bicycle, reducing the environmental impact of purchasing new products. This choice also encouraged healthy living and cost savings.
LiU offers shared electric scooters and bikes across campus, helping students and staff choose clean transport options. These services are maintained by volunteers who repair bikes free of charge—extending lifespan and accessibility for non-technical users.
The city of Linköping continues to expand green public transport: they now operate hybrid and electric buses, an improved tram network, and better bike-to-bus connectivity. These changes are part of the city's Climate Plan 2030, aiming for zero fossil fuel usage by 2030—promoting a practical shift away from private cars.

Sustainability in the Research Project

My research project involved a lot of time working on a computer. I used my personal laptop, and the university provided me with an extra screen. This setup allowed me to work efficiently with two screens, without needing to buy new equipment. Reusing existing devices helped reduce the environmental cost of new materials.
My work also included thinking about the energy impact of using digital tools. Computers, servers, and networks use a lot of electricity. The project encouraged me to reflect on how to reduce unnecessary data processing and use digital tools in smarter ways. These ideas are very important today, because digital activity is growing everywhere. Learning to work with digital efficiency is a step toward being more sustainable, even in scientific research.

Sustainability in the Institution

A key project illustrating these efforts is the electric autonomous bus called "Ride the Future." This innovative bus was developed in partnership with `Transdev`, an international company specializing in sustainable transport solutions, and `VTI` (the Swedish National Road and Transport Research Institute), which focuses on research to improve transport systems in Sweden. In June 2025, the King of Sweden attended the first test ride on campus, showing strong support at the national level. The bus operates on renewable energy, offers on-demand service to users, and is designed to be accessible for all, including people with reduced mobility. This project highlights how advanced technology, environmental sustainability, and social inclusion can work together in practice. Additionally, LiU collaborates with local partners like `Östgötatrafiken`, the regional public transport authority for the Östergötland County where Linköping is located. Together, they run initiatives such as donating used IT equipment to community centers and implementing energy-efficient lighting in campus buildings.

These partnerships reinforce the university's commitment to sustainability by extending its impact beyond the campus and fostering community development.
Beyond these, LiU hosts dedicated research groups focused on sustainable development. For example, the *Environmental Change* researchers work within areas such as `Knowledge Politics, Communication, and Learning`. They investigate how environmental knowledge is formed, shared, and understood in education, media, research, and policy. This includes studying how 'sustainable development' as a concept is taught in schools and universities, and how learning tools—such as visualizations—help foster deeper understanding and societal transformation. These efforts show how the university promotes sustainability not only through technology but also through education and knowledge dissemination.
Another key research area at LiU is `Sustainable Supply Chains and Urban Logistics`. These studies explore how complex supply chains impact the environment and society, often beyond the immediate company involved. The university focuses on improving collaboration between different actors in supply and transport chains, creating innovative tools and methods to enhance sustainability. For example, the research on sustainable city logistics addresses the "last mile" delivery challenges, seeking to reduce emissions and improve efficiency by coordinating municipalities, property owners, retailers, and logistics providers. This systemic approach underlines LiU's commitment to tackling sustainability in practical, real-world settings.

Conclusion

My experience at LiU taught me that sustainable development is a complex and multilayered challenge requiring action at many levels. Individual choices, such as using secondhand bicycles and supporting repair services, show how personal responsibility contributes to environmental goals. At the same time, adapting research practices to be more energy-efficient helps reduce the often overlooked carbon footprint of digital work. On a larger scale, institutional initiatives like the electric autonomous bus project demonstrate how innovation can directly improve urban mobility while lowering emissions and increasing accessibility. Beyond technology, LiU's strong focus on knowledge production and communication—through research on environmental education, sustainability learning, and supply chain logistics—highlights the importance of understanding, teaching, and collaborating effectively to meet sustainability goals. Altogether, these combined efforts illustrate how a university can serve as a living laboratory and a catalyst for societal transformation. By integrating practical actions, research, and education, LiU fosters a culture of sustainability that prepares both the campus community and the wider society to face environmental challenges with awareness, responsibility, and cooperation.

Building on this environment of innovation and responsibility, my main project at LiU focused on "`Modeling Team Playing Styles Based on Match Location: Home vs Away Performance Analysis.`" This project aims to analyze how football teams adapt their tactics depending on whether they play at home or away, exploring the differences in strategy, player behavior, and overall team performance. By applying data-driven methods and statistical modeling, the study seeks to provide deeper insights into the dynamics of match location and its impact on playing styles, contributing valuable knowledge both for sports analytics and tactical decision-making in professional football.

2. Structuring the 2022–2023 Premier League Season for Tactical Analysis

2.1 Why Focus on the 2022–2023 Premier League Season?

To conduct a detailed, tactical, and structured analysis, this work focuses on a single league and season: the English Premier League (EPL) 2022–2023. The EPL stood out as the ideal candidate due to its global competitiveness, tactical diversity, and the accessibility of high-quality match data. These qualities make it an ideal environment for investigating team behavior and performance dynamics over time (cf. [3]).

We selected the 2022–2023 season because it is both recent and complete, providing detailed information for all 38 matchweeks. The dataset used for this project was sourced from **Kaggle** [4], covering three seasons (2021–2024), with a specific focus on the 2022–2023 campaign.

The data offers a dual perspective. On the one hand, it includes detailed match-level statistics such as expected goals, score, touches, interceptions, attendance... – all essential for understanding tactical results. On the other hand, it integrates player-level information such as goals scored, assists, minutes played and positions. This dual structure allows for both tactical and player-centric analysis, making it particularly useful for modeling team behaviors over time.

To get an initial overview of the dataset's structure and available data, we look at how a match is recorded for the first matchweek of the 2022–2023 season. This overview shows key information such as teams, venue, date, and match result, before including any tactical or statistical metrics (e.g., xG or possession). As the dataset includes many columns, only the first few are shown in Figure 4 as a representative sample.

Wk	Date	Time	Home	xG_x	Score_x	Score_y	xG_y	Away	Attendance	Venue	Referee
1.0	05/08/2022	20:00:00	Crystal Palace	1.2	0	2	1.0	Arsenal	25286.0	Selhurst Park	Anthony Taylor
1.0	06/08/2022	12:30:00	Fulham	1.2	2	2	1.2	Liverpool	22207.0	Craven Cottage	Andy Madley
1.0	06/08/2022	15:00:00	Tottenham	1.5	4	1	0.5	Southampton	61732.0	Tottenham Hotspur Stadium	Andre Marriner
1.0	06/08/2022	15:00:00	Newcastle Utd	1.7	2	0	0.3	Nott'ham Forest	52245.0	St James' Park	Simon Hooper
1.0	06/08/2022	15:00:00	Leeds United	0.8	2	1	1.3	Wolves	36347.0	Elland Road	Robert Jones
1.0	06/08/2022	15:00:00	Bournemouth	0.6	2	0	0.7	Aston Villa	11013.0	Vitality Stadium	Peter Bankes
1.0	06/08/2022	17:30:00	Everton	0.7	0	1	1.5	Chelsea	39254.0	Goodison Park	Craig Pawson
1.0	07/08/2022	14:00:00	Leicester City	0.6	2	2	0.8	Brentford	31794.0	King Power Stadium	Jarred Gillett
1.0	07/08/2022	14:00:00	Manchester Utd	1.4	1	2	1.5	Brighton	73711.0	Old Trafford	Paul Tierney
1.0	07/08/2022	16:30:00	West Ham	0.5	0	2	2.2	Manchester City	62443.0	London Stadium	Michael Oliver

Figure 4: Dataset preview for Week 1 of the 2022–2023 Premier League season excluding advanced metrics

2.2 Preparing the Data : Splitting Home and Away Matches

To analyze team behaviors based on match location, we split all 380 matches of the 2022–2023 Premier League season into two distinct datasets: one for **home** matches and another for **away** matches. This separation is essential for isolating the tactical profiles of teams depending on whether they played at home or away. To do this, we created two subsets of the original dataset: one containing only the statistics for the home teams, and another for the away teams. Then, we aggregated the data by matchweek (`Wk`) and by team name using a `groupby` operation, which computes the mean performance per team for each round of the season. This allowed us to structure the data in a "per team per week" format, making it easier to track team behavior over time. As a result of this grouping process, we obtained two clean and balanced datasets that reflect team-level performances over the season, separately for home and away conditions. This structure serves as the analytical foundation for our upcoming week-by-week comparisons and the home-vs-away tactical modeling.

2.3 Matchweek-Based Comparative Analysis

Building on the structured dataset introduced earlier, we carried out a matchweek-based comparative analysis across the 38 weeks of the 2022–2023 Premier League season. The objective was to examine how teams performed from one matchweek to the next using key descriptive indicators. For every one of the 38 matchweeks, we generated a summary capturing statistics for all 10 games, with a specific focus on both the scorelines and the match context. Each game was presented with the two teams involved, the final score, and a detailed breakdown of technical and tactical performance metrics.

The following list provides an overview of the key metrics we analyzed, grouped by category:

- Score
- Expected Goals (xG)
- Fouls, Corners, Crosses
- Touches, Interceptions
- Tackles, Aerial Duels Won
- Offsides

This week-by-week view enabled us to track evolving team behavior across the season, compare home vs. away performance, and identify early tactical patterns. The first three matchweeks were analyzed in greater detail, serving as a pilot study to understand early-season dynamics, helping us understand how the season began and how to generalize the approach to the remaining games.

Figure 5 illustrates the structure used to display the data for Matchweek 1, presenting team names, final match results, and key contextual statistics—such as possession or expected goals—in a compact, clear, and visually accessible layout that facilitates quick interpretation and comparison. This visualization was implemented in Python using libraries suitable for dashboard and data display [5].

🏆 Premier League - Home vs Away Stats - Week 1 🏆

N°	Stadium	Home	Away	Referee	Score	Touches	Aerials Won	Interceptions	Tackles	Fouls	Offsides	Long Balls	Goal Kicks
1	Selhurst Park	Crystal Palace	Arsenal	Anthony Taylor	0 - 2	726 vs 599	10 vs 14	8 vs 9	18 vs 29	16 vs 11	1 vs 2	83 vs 59	4 vs 2
2	Craven Cottage	Fulham	Liverpool	Andy Madley	2 - 2	474 vs 784	23 vs 13	10 vs 10	24 vs 11	7 vs 9	4 vs 4	77 vs 94	8 vs 5
3	Tottenham Hotspur Stadium	Tottenham	Southampton	Andre Marriner	4 - 1	709 vs 554	13 vs 11	3 vs 13	24 vs 14	11 vs 6	2 vs 0	85 vs 61	5 vs 4
4	St James' Park	Newcastle Utd	Nott'ham Forest	Simon Hooper	2 - 0	634 vs 475	12 vs 16	11 vs 10	19 vs 15	9 vs 14	2 vs 0	64 vs 70	2 vs 12
5	Elland Road	Leeds United	Wolves	Robert Jones	2 - 1	515 vs 720	9 vs 7	14 vs 14	21 vs 16	13 vs 9	0 vs 1	60 vs 72	9 vs 9
6	Vitality Stadium	Bournemouth	Aston Villa	Peter Bankes	2 - 0	462 vs 721	16 vs 16	13 vs 8	20 vs 10	18 vs 16	1 vs 4	68 vs 70	12 vs 3
7	Goodison Park	Everton	Chelsea	Craig Pawson	0 - 1	495 vs 730	12 vs 22	8 vs 9	24 vs 19	14 vs 11	0 vs 2	71 vs 46	5 vs 7
8	King Power Stadium	Leicester City	Brentford	Jarred Gillett	2 - 2	725 vs 580	14 vs 15	10 vs 8	13 vs 10	6 vs 5	2 vs 2	42 vs 70	3 vs 7
9	Old Trafford	Manchester Utd	Brighton	Paul Tierney	1 - 2	708 vs 483	11 vs 9	15 vs 12	12 vs 14	7 vs 12	3 vs 1	74 vs 78	3 vs 11
10	London Stadium	West Ham	Manchester City	Michael Oliver	0 - 2	377 vs 936	10 vs 11	11 vs 3	14 vs 6	8 vs 4	4 vs 1	66 vs 64	9 vs 5

Figure 5: Sample display of Matchweek 1: home and away match data

Following the detailed analysis of the early matchweeks, this structured approach was extended to encompass all 38 rounds of the Premier League season, enabling the construction of a comprehensive and consistent comparative overview of team performances across the entire season.

2.4 Interactive HTML Visualization: A Season Summary

In order to make the matchweek analysis more accessible and visually intuitive, we developed an interactive HTML dashboard within a Jupyter Notebook environment [6]. This visualization provided an overview of the season with:

- Match-by-match results and statistics (score, xG, fouls, possession, etc.)
- Matchweek selectors to navigate through all 38 rounds
- Final league table summarizing the standings after the full season
- Highlight of the top 3 teams (podium) and the title-winning club (Figure 7)

Figures 6 and Table 2.1 illustrate the initial views of the matchweek dashboard alongside the final league standings, offering a comprehensive look at both the dynamics and key events of the early season as well as the conclusive outcomes at the end of the campaign. These results have been cross-verified with established football data sources such as `Transfermarkt` and `FBref` to ensure accuracy and reliability [7].

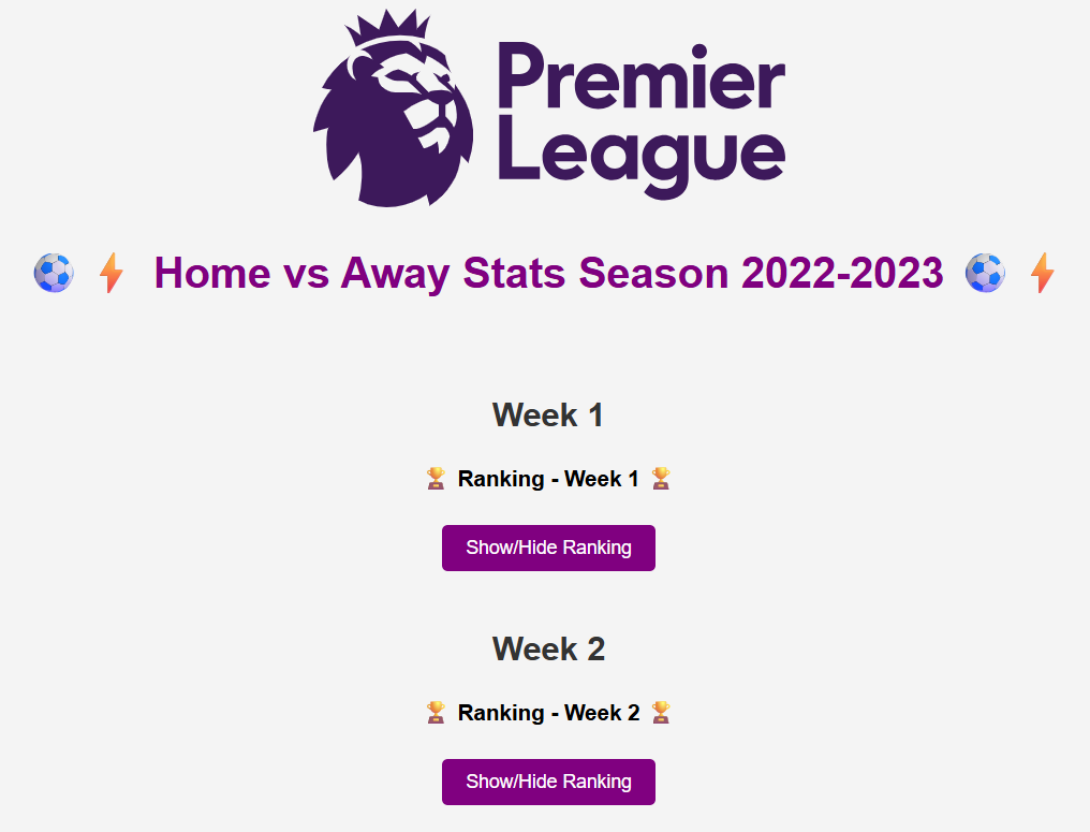

Figure 6: Interactive HTML dashboard interface: early matchweeks exploration

Figure 7: Top 3 teams podium and Premier League 2022–2023 champion celebration

Beyond the interactive navigation through matchweeks, the dashboard also included a comprehensive summary of the 2022–2023 season through the final league table. This season overview, illustrated in Table 2.1, highlights Manchester City's title-winning campaign, with Arsenal and Manchester United completing the podium. This dynamic dashboard interface (Figure 6) enables match-by-match exploration, giving users the ability to navigate through individual matchweeks and view detailed statistics.

Table 2.1: Final Premier League standings for the 2022–2023 season

Rank	Team	Pts	P	W	D	L	GD	GF	GA
1	Manchester City	89	38	28	5	5	61	94	33
2	Arsenal	84	38	26	6	6	45	88	43
3	Manchester Utd	75	38	23	6	9	15	58	43
4	Newcastle Utd	71	38	19	14	5	35	68	33
5	Liverpool	67	38	19	10	9	28	75	47
6	Brighton	62	38	18	8	12	19	72	53
7	Aston Villa	61	38	18	7	13	5	51	46
8	Tottenham	60	38	18	6	14	7	70	63
9	Brentford	59	38	15	14	9	12	58	46
10	Fulham	52	38	15	7	16	2	55	53
11	Crystal Palace	45	38	11	12	15	-9	40	49
12	Chelsea	44	38	11	11	16	-9	38	47
13	Wolves	41	38	11	8	19	-27	31	58
14	West Ham	40	38	11	7	20	-13	42	55
15	Bournemouth	39	38	11	6	21	-34	37	71
16	Nott'ham Forest	38	38	9	11	18	-30	38	68
17	Everton	36	38	8	12	18	-23	34	57
18	Leicester City	34	38	9	7	22	-17	51	68
19	Leeds United	31	38	7	10	21	-30	48	78
20	Southampton	25	38	6	7	25	-37	36	73

Note: **Pts** = Points, **P** = Matches Played, **W** = Wins, **D** = Draws, **L** = Losses, **GD** = Goal Difference, **GF** = Goals For, **GA** = Goals Against.

At the end of the season overview, Figure 7 shows the Premier League 2022–2023 podium, with Manchester City as champions, followed by Arsenal and Manchester United. This visualization highlights the best teams and reflects their steady performance and strong tactics throughout the 38 matchweeks. Together with the final ranking table (Table 2.1), these visuals give a clear and easy-to-understand summary of how the teams performed, finishing with a clear picture of the top clubs of the season. The presented information is in accordance with official season data provided by `FotMob`[8].

2.5 Summary and Transition

This chapter provided a clear overview of the Premier League 2022–2023 season by organizing and visualizing match data to compare home and away performances. Through detailed matchweek analysis and interactive views, we identified key trends in team form and important moments that influenced the final standings.

How do teams' playing styles differ between home and away matches? Are there identifiable tactical adjustments related to the venue? Do some teams consistently perform better at home while encountering challenges in away games, or vice versa? Can we measure differences in their approach, effort, or overall effectiveness depending on the match location?

The next chapter focuses on modeling these home and away differences through statistical and machine learning techniques, aiming to identify each team's distinctive playing style depending on match location. This analysis goes beyond simple results and scores to uncover how home advantage influences tactical, structural, and statistical aspects of the game.

3. Home vs. Away: Modeling Tactical Behavior in the Premier League

3.1 Why Model Home vs. Away Performance?

While raw results and match statistics provide a first look into a football team's performance, they do not reveal the full tactical reality. Teams often adjust their style, intensity, and risk based on whether they play at home or away, influenced by factors like crowd support, travel, and psychological pressure.
This study aims to model these differences to understand:

- How tactical elements such as formations and attacking strategies differ by venue?
- Which teams change their overall approach depending on location?
- Whether top teams keep a consistent style regardless of playing home or away?

Understanding these aspects is crucial for analysts, coaches who want to simulate match scenarios, build predictive models, or develop flexible tactics.

3.2 Modeling Strategy: Dual Approach Per Team

To better understand team behaviors during the 2022–2023 Premier League season, we created two separate datasets for each team: one for their 19 home matches (`home_games`) and one for their 19 away matches (`away_games`). This separation allows us to study how match location influences team performance and tactics.

Each dataset serves as a concise tactical profile summarizing a team's season based on venue. By isolating home and away contexts, we can identify differences in performance and uncover specific tactical and behavioral trends. The datasets include a wide range of metrics, such as expected goals (xG), touches, crosses, interceptions, tackles, fouls, long balls, aerial duels, offsides, and goal kicks, covering both offensive and defensive aspects. These metrics cover both the offensive and defensive dimensions and provide a structured basis for understanding how each team has adapted or failed to adapt their style of play based on match venue.

This dual approach helps us compare how teams adjust their style when playing at home versus away and also enables comparisons between teams, revealing patterns like some clubs favoring possession at home and counterattacking away. While not intended to predict match outcomes, this method highlights consistent tactical differences linked to venue, providing a solid foundation for more advanced analyses such as clustering and machine learning, where the venue context (home or away) becomes a crucial factor in understanding team behavior.

The following sections will present visual summaries of selected teams and interpret key findings, leading into more detailed modeling of tactical behaviors.

3.3 Visualization of Tactical Profiles

Although the dual modeling approach was applied to all Premier League teams for the 2022–2023 season, presenting the full set of results would be too extensive in the context of this report. Therefore, we focus on a selected group of four teams that provide meaningful and contrasting examples of tactical behavior. These teams represent different league standings, offering a diverse range of styles, levels of success, and adaptability.

- **Manchester City** – As league champions, they offer a tactical reference point. Known for their structured and dominant 4-3-3 setup, City's consistent style makes them ideal for assessing how a top-performing team balances home and away performances.
- **Tottenham Hotspur** – A top-half team characterized by tactical inconsistency. Frequently alternating between strengths and weaknesses depending on the venue, Spurs offer a strong case of fluctuating performance. Their typical 4-2-3-1 formation serves as a base to analyze how structure and balance vary between home and away settings.
- **Everton** – A lower-table team involved in a relegation fight throughout the season. Their tactical profile reflects a predominantly reactive and defensively oriented approach, strongly influenced by the imperative to avoid losses. Operating mostly in a 4-3-3 formation, their home vs. away comparison offers valuable insight into how risk-averse strategies shift with match location.
- **Southampton** – As the league's bottom-placed team, Southampton displayed tactical instability and and struggled to maintain consistency. Mostly using a 4-2-3-1 formation, their profile helps illustrate how structural weaknesses and lack of cohesion played out differently at home versus away.

These four examples offer a representative look at how teams from different levels of the league adjusted their tactics based on match location. By comparing their home and away data through radar charts and key metrics, we gain insights into how strategy and performance varied depending on venue and competitive context.

Figures 8, 9, 10, and 11 illustrate these tactical contrasts through a combination of radar charts and bar plots. These visualizations, implemented using the Python library `Matplotlib` (cf. [9]), provide an intuitive and structured way to compare how each team adapted their playing style, intensity, and efficiency when performing at home versus away throughout the season.

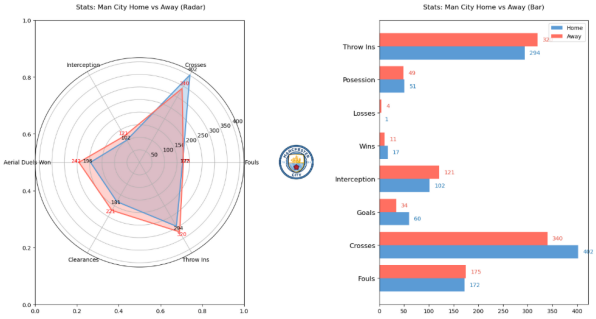

Figure 8: Manchester City – Tactical statistics comparison at home vs. away (2022–2023)

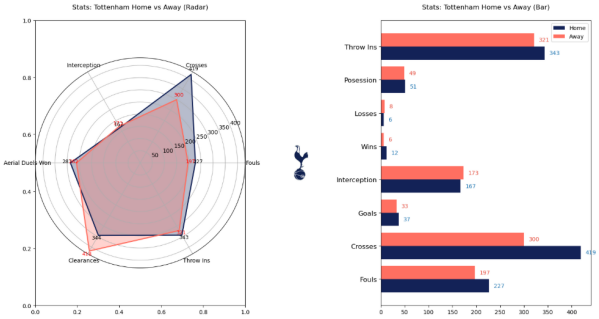

Figure 9: Tottenham Hotspur – Tactical statistics comparison at home vs. away (2022–2023)

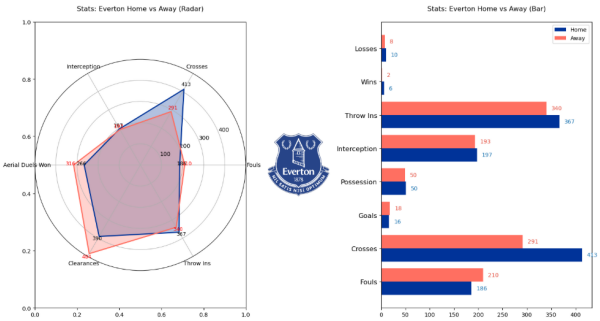

Figure 10: Everton – Tactical statistics comparison at home vs. away (2022–2023)

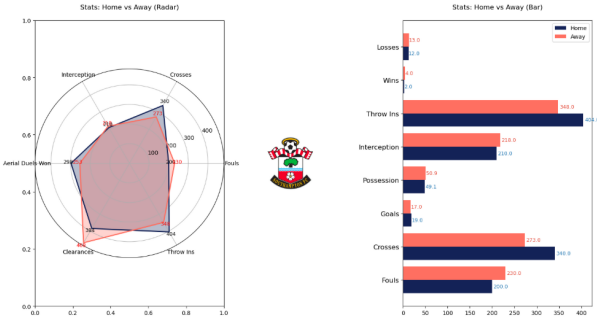

Figure 11: Southampton – Tactical statistics comparison at home vs. away (2022–2023)

3.4 Team Identity and Performance Patterns

This section presents a comparative analysis of four Premier League teams, focusing on how their tactical profiles evolved between home and away matches during the 2022–2023 season. Key metrics are summarized in the following tables, followed by a detailed interpretation for each case.

Manchester City: Home vs Away Performance (2022–2023)

As part of our tactical review, **Manchester City** emerge as a model of consistency. Under Pep Guardiola, the team displayed remarkable balance between home and away matches during the 2022–2023 season. Operating primarily in a 4-3-3 formation, they combined control, possession, and pressing intensity to maintain high performance levels at the Etihad and away from home alike.

At home, Manchester City played with excellent control and strong attack. They had over 789 touches per game and kept more than 70% possession in most matches, making the Etihad very hard for other teams. They scored 3.21 goals per game (12.5 more than expected), kept 11 clean sheets, and made the fewest tackles (just 1.6 per game), showing they stopped opponents mainly by keeping the ball and staying well positioned.

Away, Manchester City remained strong with 13 wins and 42 points. They averaged 2.05 goals per game, exceeding their xG by +4.1, and conceded just under 1 goal per match. Despite slightly less control than at home, they still showed high technical quality (almost 700 touches, 83% pass accuracy) and scored 4+ goals in 5 away games, confirming their ability to dominate even outside the Etihad.

Conclusion: Manchester City displayed elite consistency both home and away. Their home form was unmatched, and their away form remained top-tier, confirming their identity as a complete and tactically mature team. This strong performance was supported by their usual 4-3-3 formation, which allowed them to control the midfield, create many chances, and adapt well to different match situations. (see Table 3.1)

Metric	Home	Away	Diff.
Win Rate	89%	68%	-21%
Goals Scored	3.21	2.05	-1.16
Goals Conceded	0.84	0.95	+0.11
Clean Sheets	11	7	-4
xG Overperform	+12.5	+4.1	Better home
Bonus Stats			
Most Goals: Erling Haaland – 36 (8 Assists)			
Player of the Season: Erling Haaland			
Most Assists: Kevin De Bruyne – 16			

Table 3.1: Manchester City Home vs Away – Performance Comparison (2022–23)

Tottenham : Home vs Away Performance (2022–2023)

Following our analysis of Manchester City, we now turn to **Tottenham Hotspur**. Over the 2022–2023 season, Spurs exhibited noticeable contrasts in their performances depending on the venue. While their home ground offered opportunities for important wins, their away fixtures often proved more challenging, reflecting a less consistent campaign overall. Throughout the season, Tottenham primarily deployed a **4-2-3-1** formation, which balanced a solid midfield base with attacking creativity. The summary table on the right presents key statistics, which we discuss in detail below.

At home, Tottenham played with good control, focusing on keeping the ball. They had about 700 touches per game and kept possession over 60% in 12 out of 19 home matches. They won 12 games and kept 6 clean sheets. Their main goal scorer was Harry Kane, who scored the first goal in 14 home matches. However, their defense had problems, conceding 2 or more goals in 6 games. They also lost all their home games against the **"Big Six": Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and themselves**.

Away, Tottenham faced difficulties, managing only 6 wins and keeping just 2 clean sheets. Their defense conceded an average of 1.79 goals per game, including 12 goals from set pieces, highlighting some vulnerability in defense. Playing away, they relied more on direct football, using 15% more long balls and focusing on counter-attacks, despite having less possession overall. Their performance against the "Big Six" teams on the road was particularly weak, with only 1 win and 5 losses, showing struggles against the strongest opponents away from home.

Conclusion: Tottenham's 2022–2023 season showed a sharp contrast between home strength and away difficulties, with a 31% lower win rate on the road—the largest gap among the "Big Six." Their 4-2-3-1 formation relied heavily on Harry Kane as the focal point in attack, but defensive lapses, especially during transitions and set pieces, limited their success (see Table 3.2).

Metric	**Home**	**Away**	**Diff.**
Win Rate	63%	32%	-31%
Goals Scored	1.84	1.32	-0.52
Goals Conceded	1.21	1.79	+0.58
Clean Sheets	6	2	-4
xG Overperform	-2.7	-3.9	Worse away
Bonus Stats			
Most Goals: Harry Kane – 30			
Most Assists: Ivan Perišić - 8			

Table 3.2: Tottenham Home vs Away – Performance Comparison (2022–23)

Everton : Home vs Away Performance (2022–2023)

Moving on from Manchester City and Tottenham, we now look at Everton's 2022–2023 season. Unlike those teams, Everton spent most of the year near the relegation zone, showing different problems at home and away. They had trouble scoring but sometimes defended well, which helped them avoid relegation. Everton usually played with a 4-3-3 formation, like Manchester City, which influenced how they played both at home and away.

At home, Everton focused a lot on defending to stay in the game. They kept 6 clean sheets (meaning no goals allowed in 32% of their home matches) and made 18.3 tackles per game, which was the third highest in the league. This shows they worked hard to stop their opponents. For example, their 1–0 win against Arsenal at Goodison Park showed they could make it difficult for stronger teams. However, they had a hard time scoring goals, failing to score in 8 home games and scoring less than expected by 3.4 goals, which was the worst in the league.

Away from home, Everton scored slightly more goals (0.89 per game) than at home (0.79), but they conceded more often, allowing 1.63 goals per game and winning only 2 away matches. Their standout moment was a surprising 5–1 win at Brighton, but they also suffered heavy defeats like the 4–0 loss at Arsenal. They often fell behind early, conceding first in 14 out of 19 away games, showing problems with their defense. They also relied a lot on set pieces, with over half (53%) of their away goals coming from free kicks or corners.

Conclusion: Everton's season was very close to the edge. Their home results, though not strong, were important to stay in the league, while their away games showed problems in defense and a lack of flexibility in how they played. Even though they used the same 4-3-3 formation as Manchester City, Everton couldn't play it as well and didn't control games or score as much. With the fewest goals scored (32 total), wins like the one against Brighton were key to their survival. (see Table 3.3)

Metric	Home	Away	Diff.
Win Rate	32%	11%	-21%
Goals Scored	0.79	0.89	+0.10
Goals Conceded	1.16	1.63	+0.47
Clean Sheets	6	3	-3
xG Overperform	-6.85	-4.85	Worse home
Bonus Stats			
Most Goals: Dwight McNeil – 7			
Most Assists: Alex Iwobi – 7			

Table 3.3: Everton Home vs Away – Performance Comparison (2022–23)

Southampton FC: Home vs Away Performance (2022–2023)

Finally, we look at Southampton FC, the last team in our comparison. The 2022–2023 season was very hard for them, with poor results both at home and away. Unlike the other teams, Southampton did not show much difference between playing at home or away, showing bigger problems in their team and tactics. They usually played in a 4-2-3-1 formation, which did not bring enough balance or success in attack and defense.

At home, Southampton won only 3 out of 19 games at St Mary's Stadium. Even though they had 55% possession on average, they scored less than 1 goal per game (0.95). Defensively, they let in goals in 17 matches and kept just 2 clean sheets. Many of their goals (39%) came from set pieces, showing they struggled to create chances during regular play. Their best win was 2–1 against Chelsea (xG 1.1), but they also lost 0–1 to Nottingham Forest (xG 0.5), showing difficulties in games they could have won.

Away from home, things were just as bad for Southampton. They won only 3 out of 19 games and gave up 2.05 goals per match on average, one of the highest in the league. They scored 15 goals in total, but only 5 were from regular play. Their defense often broke down, like in the 4–0 loss to Manchester City. Discipline was also an issue, with 3 red cards, the most in the league. A 1–0 win at Bournemouth (xG 1.0) was one of the few good moments away from home.

Conclusion: Southampton had a very tough season, performing poorly both at home and away. With only 25 points, they finished last—the first team to do so since 2007. Scoring just 33 goals and conceding in almost every match showed major problems, made worse by four coaching changes. Their usual 4-2-3-1 formation did not bring stability, and despite a few strong games, it wasn't enough to avoid relegation. (see Table 3.4)

Metric	Home	Away	Diff.
Win Rate	16%	16%	0%
Goals Scored	0.95	0.79	-0.16
Goals Conceded	1.79	2.05	+0.26
Clean Sheets	2	2	0
xG Overperform	-3.1	-4.8	Worse home
Bonus Stats			
Most Goals: James Ward-Prowse – 9			
Most Assists: James Ward-Prowse – 4			
Used 4 different coaches during the season			
4–4 draw vs Liverpool was their only 4+ goal game			

Table 3.4: Southampton Home vs Away – Performance Comparison (2022–23)

3.5 Why These Teams Were Chosen

These four teams were selected to show different ways of playing and different levels of success in the Premier League during the 2022–2023 season:

- **Manchester City** showed top-level consistency and control, with strong possession and high performance both at home and away. Their organized style and smart adjustments made them a reference point for other teams.
- **Tottenham** showed a big difference between home and away games. At home, they kept the ball well and leaned on their main striker, but their defense stayed shaky. Away, they played more direct and on the counter, which led to more goals conceded and fewer wins.
- **Everton** had problems both at home and away, but for different reasons. At home, they defended well with several clean sheets, but struggled to score. Away, their defense was weaker, they had less of the ball, and most goals came from set pieces.
- **Southampton** performed poorly in nearly every game, whether home or away. They often conceded goals, scored very little, and picked up several red cards. Their weak tactics and lack of efficiency led to relegation.

Together, these teams illustrate show how playing at home or away can change the way a team plays and how well they perform. From top teams with strong and steady tactics to those that struggled all season, they help us understand different styles and challenges in the Premier League.

3.6 Summary and Transition

The home and away analysis of these four teams shows how match location can strongly influence a team's playing style, results, and overall performance. From Manchester City's steady and dominant control to Southampton's ongoing struggles, the data (see Figures 8 to 11) highlight how teams adapt—or fail to adapt—depending on where they play. These differences reflect tactical choices such as possession, defensive organization, and dependence on key players.

Understanding these patterns is an important step before moving to a more data-driven evaluation of team strength that changes throughout the season. To do this, we will now introduce the **ELO rating system**, a method that captures team form over time. This system will help us model not only match outcomes but also how tactics and team strength evolve week by week in the Premier League.

4. ELO as a Tactical Contextualizer in Football Analytics

4.1 ELO: Origins and Applications in Football

The **ELO rating system**, originally conceived by Hungarian-American physicist Arpad Elo, was first developed to evaluate the relative skill levels of chess players. Its mathematical rigor and adaptability have since led to its application across a wide range of competitive domains, including basketball, ice-hockey, and most notably, football. Its strength lies in its ability to adjust team ratings dynamically after each match, based on performance and expectations. Unlike traditional point systems, ELO accounts for the strength of the opponent, the result, and the match context (such as goal difference or home advantage).

From Chess to Football: A Historical Perspective

In 1997, Bob Runyan adapted the ELO system for international football, publishing an open-source version that is now maintained by Kirill Bulygin on the *World Football Elo Ratings* website. This adaptation marked a shift by providing a more dynamic alternative to traditional ranking systems, which were often based on fixed points or simple win-loss records. Studies have shown that ELO ratings outperform earlier FIFA ranking models in predicting match outcomes [10].
The ELO system assigns each team a rating that reflects its skill relative to other teams, dynamically updating this rating after each match based on the actual result compared to the expected outcome. It also takes into account factors such as goal difference, match importance, and home advantage. This system is asymmetric: teams earn more points for beating higher-rated opponents and face harsher penalties when losing to weaker teams. Draws are also evaluated in context, being rewarded when achieved against stronger teams and penalized otherwise.

Football-Specific Enhancements

To better fit football, the ELO system adds a few key factors: a bonus for playing at home (usually +100 points), bigger rating changes when the score difference is large, and more weight given to important matches like qualifiers. These help the ratings reflect not just who won, but also how strong the performance was under different pressures. After about 30 official games, a team's ELO rating usually shows its true level, but useful insights can come earlier. ELO is important for tactics because it gives a continuous and time-based measure of team strength, unlike fixed rankings. It helps us understand if poor results come from playing away or facing a strong opponent, and whether changes in form relate to tactics or just tough schedules. Using ELO in analysis adds detail and context beyond just looking at home vs away performance.

Implementation Scope

Although there is no single official ELO system for football, several well-known versions are used today. Since 2018, FIFA adopted a modified ELO-based formula for its men's rankings, replacing the older points-based system. The main differences between FIFA's version and the *World Football Elo Ratings* lie in how they treat factors such as goal difference and match importance. In the next section, we present the mathematical formula we use to calculate ELO ratings and explain how these scores help guide clustering, dimensionality reduction, and classification tasks throughout our analysis

4.2 Mathematical Formulation of ELO in Football

In our project, we adopt a variant of the ELO system inspired by FIFA's official adaptation. This version refines the classical formula by incorporating additional factors such as goal difference and match importance. Each team's rating is initialized at 1500 at the start of the season and evolves match by match according to the following rules.

General Formula

The updated rating for a team after a match is given by:

$$R_n = R_o + P \tag{4.1}$$

Where:

- R_n is the new ELO rating after the match.
- R_o is the previous ELO rating.
- P is the number of points gained or lost from the match, rounded to the nearest integer.

The value of P is computed as:

$$P = K \cdot G \cdot (W - W_e) \tag{4.2}$$

Each component plays a specific role:

- K is the weighting constant based on the match's importance.
- G is a goal-difference factor that adjusts the score change depending on how dominant the victory was.
- W is the actual match result: 1 for a win, 0.5 for a draw, and 0 for a loss.
- W_e is the expected result based on the relative ratings of the two teams.

Expected Result

The expected outcome of a match is calculated using the following logistic function:

$$W_e = \frac{1}{1 + 10^{-dr/400}} \tag{4.3}$$

Where:

- $dr = R_{\text{team}} - R_{\text{opponent}} + H$
- R_{team} is the **ELO rating** of the team before the match. It reflects the team's strength based on past performance and match results.
- R_{opponent} is the **ELO rating** of the opposing team before the match.
- H is a constant home advantage bonus, typically set at 100 points.

Thus, the higher the rating difference in favor of a team (after accounting for home advantage), the higher the expectation of winning. For example, when $dr = 0$, $W_e = 0.5$, and when $dr = 120$, $W_e \approx 0.66$.

Match Importance Factor K

The constant K adjusts the sensitivity of the system depending on the tournament or competition type. Typical values are:

Match Type	**K Value**
World Cup, Olympic Games (1908–1980)	60
Continental/Intercontinental Tournaments	50
Major Qualifiers and Finals	40
Other Competitive Matches	30
Friendly Matches	20

Table 4.1: FIFA ELO K constants by match type

Goal Difference Factor G

The goal difference in a match affects the magnitude of the rating change. The value of G is defined as follows:

- $G = 1$ if the match is drawn or won by only one goal.
- $G = 1.5$ if the match is won by two goals.
- $G = \frac{11+N}{8}$ if the match is won by 3 or more goals, where N is the goal difference.

Example values:

Goal Difference	0	1	2	3	4	5	6	7	8	9	10
G Value	1	1	1.5	1.75	1.875	2	2.125	2.25	2.375	2.5	2.625

Table 4.2: Goal Difference Factor G

Match Result W and Special Cases

The result W reflects the outcome of the match from the perspective of the team:

- Win: $W = 1$
- Draw: $W = 0.5$
- Loss: $W = 0$

In cases where the match is decided by penalties, FIFA considers it a draw ($W = 0.5$), regardless of which team wins the shootout. If extra time is played, the result of extra time is taken as the final result.

In brief, this ELO method helps capture important details of each football match—not just who won, but how convincing the win was, how important the match was, and whether it was home or away. It provides a clear and dynamic way to measure a team's strength and follow their tactical changes over the season. We use these ELO ratings for every match as key inputs in our clustering, classification, and regression analyses.

4.3 Using ELO in Our Analysis

In this project, we used a custom ELO system for the 2022–2023 Premier League season where teams began at 1500 and ratings changed after each match based on goals and home advantage. Unlike fixed points in the official table, our ELO adjusts dynamically with results and opponent strength, giving a clearer and more detailed picture of team performance and form changes during the season. Below, we compare the final ELO rankings with the official league standings to highlight differences and insights from the dynamic rating system.

Rank	Team	Final ELO
1	Manchester City	1769
2	Arsenal	1667
3	Liverpool	1643
4	Manchester Utd	1637
5	Newcastle Utd	1609
6	Brentford	1605
7	Aston Villa	1604
8	Brighton	1544
9	Tottenham	1499
10	Fulham	1492
11	Crystal Palace	1472
12	West Ham	1456
13	Nott'ham Forest	1426
14	Wolves	1423
15	Everton	1416
16	Chelsea	1415
17	Leicester City	1399
18	Bournemouth	1355
19	Leeds United	1285
20	Southampton	1284

Figure 12: Final ELO Ranking of the 2022–2023 Premier League Season

PL Rank	Team	Points	ELO Rank	Final ELO	Δ Rank
1	Manchester City	89	1	1769	0
2	Arsenal	84	2	1667	0
3	Manchester Utd	75	4	1637	-1 ▼
4	Newcastle Utd	71	5	1609	-1 ▼
5	Liverpool	67	3	1643	+2 ▲
6	Brighton	62	8	1544	-2 ▼
7	Aston Villa	61	7	1604	0
8	Tottenham	60	9	1499	-1 ▼
9	Brentford	59	6	1605	+3 ▲
10	Fulham	52	10	1492	0
11	Crystal Palace	45	11	1472	0
12	Chelsea	44	16	1415	-4 ▼
13	Wolves	41	14	1423	-1 ▼
14	West Ham	40	12	1456	+2 ▲
15	Bournemouth	39	18	1355	-3 ▼
16	Nott'ham Forest	38	13	1426	+3 ▲
17	Everton	36	15	1416	+2 ▲
18	Leicester City	34	17	1399	+1 ▲
19	Leeds United	31	19	1285	0
20	Southampton	25	20	1284	0

Figure 13: Comparison Between Official Points Ranking and ELO Ranking

The ELO model shows important differences from the official Premier League table. One key limitation is that the model starts every team with an initial ELO of 1500, which assumes equal strength at the season's start. This simplification overlooks historical team quality differences. On the other hand, using previous season ELOs is complicated since teams change between seasons, with new teams promoted and others relegated, making it unclear how to carry over ratings fairly. On the one hand, Liverpool and Brentford rank higher in ELO because they had consistent performances against strong teams. Nottingham Forest is also underrated by the official table. On the other hand, Chelsea, Brighton, and Bournemouth rank lower in ELO, suggesting their points may not fully reflect their true strength. On the other hand, the model matches the official ranking for teams such as Manchester City, Arsenal, Leeds, and Southampton, showing it works well at both top and bottom. While ELO does not replace league points, it gives a clearer view of team performance by considering opponent strength and match context.

Case Study: Manchester City's ELO Evolution

To better illustrate the value of ELO in tracking performance trends, we present the full-season evolution of Manchester City's ELO score. Starting at 1500, City's rating progressed steadily across the 38 matchweeks, finishing at 1769. This evolution reflects their sustained dominance and consistent performance against both top-tier and mid-table teams. This time-based perspective clearly highlights key moments in the season of Manchester City. The team followed a steady and impressive upward trend in their ELO rating, showing strong form and consistent performances. There were brief slowdowns during periods with many matches close together, but their rating increased significantly after crucial victories, particularly against strong opponents like Arsenal and Liverpool. Notably, Manchester City never fell below the starting threshold of 1500 ELO points, which underlines their dominance and stability across the entire campaign. Their ELO reached a season-high of 1795 points in Matchweek 37 after a hard-fought 1–0 home win against Chelsea. Although they ended the season with a narrow defeat away to Brentford, their final rating remained high at 1769 points. These developments are illustrated in the two figures below: Figure 14 shows the full progression of their ELO rating across all matchweeks, while Figure 15 presents a summary of key ELO metrics.

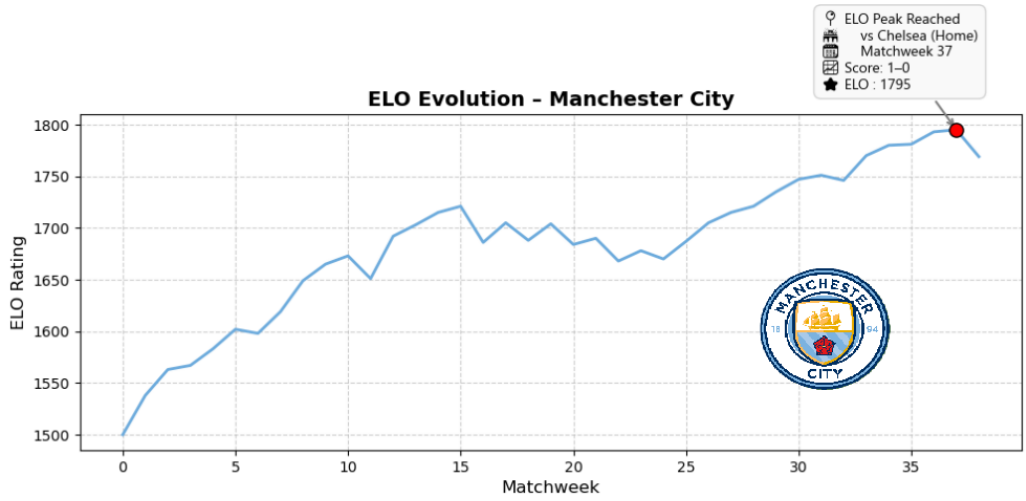

Figure 14: Manchester City — ELO Progression (Matchweek 1–38)

Highlighted row (Wk 37) corresponds to Manchester City's peak ELO.

Matchweek	Home	Away	Score	Home Δ	Away Δ	New Home ELO	New Away ELO
1	West Ham	Manchester City	0 - 2	-38	+38	1462	1538
2	Manchester City	Bournemouth	4 - 0	+25	-25	1563	1497
3	Newcastle Utd	Manchester City	3 - 3	-4	+4	1524	1567
4	Manchester City	Crystal Palace	4 - 2	+16	-16	1583	1477
5	Manchester City	Nott'ham Forest	6 - 0	+19	-19	1602	1442
6	Aston Villa	Manchester City	1 - 1	+4	-4	1438	1598
7	Manchester City	Tottenham	4 - 2	+21	-21	1619	1567
8	Wolves	Manchester City	0 - 3	-30	+30	1436	1649
9	Manchester City	Manchester Utd	6 - 3	+16	-16	1665	1525
10	Manchester City	Southampton	4 - 0	+8	-8	1673	1395
11	Liverpool	Manchester City	1 - 0	+22	-22	1562	1651
12	Arsenal	Manchester City	1 - 3	-41	+41	1641	1692
13	Manchester City	Brighton	3 - 1	+11	-11	1703	1523
14	Leicester City	Manchester City	0 - 1	-12	+12	1444	1715
15	Manchester City	Fulham	2 - 1	+6	-6	1721	1495
16	Manchester City	Brentford	1 - 2	-35	+35	1686	1513
17	Leeds United	Manchester City	1 - 3	-19	+19	1432	1705
18	Manchester City	Everton	1 - 1	-17	+17	1688	1394
19	Chelsea	Manchester City	0 - 1	-16	+16	1503	1704
20	Manchester Utd	Manchester City	2 - 1	+20	-20	1628	1684
21	Manchester City	Wolves	3 - 0	+6	-6	1690	1369
22	Tottenham	Manchester City	1 - 0	+22	-22	1575	1668
23	Manchester City	Aston Villa	3 - 1	+10	-10	1678	1478
24	Nott'ham Forest	Manchester City	1 - 1	+8	-8	1442	1670
25	Bournemouth	Manchester City	1 - 4	-17	+17	1358	1687
26	Manchester City	Newcastle Utd	2 - 0	+18	-18	1705	1620
27	Crystal Palace	Manchester City	0 - 1	-10	+10	1407	1715
28	Manchester City	West Ham	3 - 0	+6	-6	1721	1389
29	Manchester City	Liverpool	4 - 1	+14	-14	1735	1572
30	Southampton	Manchester City	1 - 4	-12	+12	1346	1747
31	Manchester City	Leicester City	3 - 1	+4	-4	1751	1389
32	Brighton	Manchester City	1 - 1	+5	-5	1572	1746
33	Manchester City	Arsenal	4 - 1	+24	-24	1770	1707
34	Fulham	Manchester City	1 - 2	-10	+10	1469	1780
35	Manchester City	Leeds United	2 - 1	+1	-1	1781	1315
36	Everton	Manchester City	0 - 3	-12	+12	1396	1793
37	Manchester City	Chelsea	1 - 0	+2	-2	1795	1409
38	Brentford	Manchester City	1 - 0	+26	-26	1605	1769

Figure 15: Manchester City — Season Summary Based on ELO Metrics

4.4 Summary and Transition

ELO ratings bring valuable tactical context by showing whether team behavior is influenced by match location or opponent strength. This helps improve the accuracy of our machine learning models. In the next chapter, we use this enriched data in Principal Component Analysis (PCA), KMeans clustering, and Random Forest classification to better understand and predict team playing styles.

5. Clustering Football Matches Using KMeans

5.1 Motivation and Context

In modern football analytics, understanding team performance goes far beyond simply observing results. Contextual factors such as the strength of the opponent, the match venue (home or away), and the tactical setup play a crucial role in shaping match outcomes. To address this, we previously introduced ELO ratings, a dynamic metric that reflects relative team strength over time. This allowed us to control for contextual biases and better compare performances across heterogeneous fixtures. Building on this foundation, we turned to unsupervised learning techniques to go a step further, not just evaluating team performance but uncovering latent tactical patterns inherent in the matches themselves. That is why, we sought to answer the following question:

- ***Can we identify categories of matches that share similar tactical profiles, regardless of the teams involved?***

We applied KMeans clustering to group all Premier League 2022/2023 matches based on various game statistics such as expected goals, shots, possession, and defensive actions. This method, implemented using the Python library `scikit-learn` (cf. [11]), helps us identify different tactical patterns and understand how they vary between teams and venues. It also reveals playing styles directly from the data, beyond team reputations or rankings. Additionally, we examined how playing at home or away influences these tactics, as it often affects pressing and formation choices. This analysis prepares us for more detailed and predictive modeling in later steps.

5.2 Clustering Methodology

Why KMeans?

KMeans [12] is a popular unsupervised learning method because it is simple, fast, and effective at finding groups in complex data. It works by dividing all observations into a set number of clusters, grouping each observation with the closest average (centroid) to reduce overall differences within clusters. In our project, each football match is represented by numbers summarizing important statistics from both home and away teams, such as expected goals, corners, touches, interceptions, clearances, and aerial duels. These features describe how teams played offensively and defensively. Using KMeans, we can easily group matches by similar tactical patterns, creating categories that help us analyze different styles of play and find strategic trends across teams and locations.

Determining the Optimal Number of Clusters

Before applying the KMeans algorithm, it is essential to determine the most appropriate number of clusters K. An unsuitable choice of K can lead to overfitting (too many

clusters) or underfitting (too few clusters), affecting the interpretability and relevance of the resulting tactical groupings. To address this, we employed two widely used evaluation methods: the **Elbow Method** and the **Silhouette Score**.

Elbow Method: The Elbow Method helps identify the value of K beyond which the reduction in inertia (i.e., within-cluster sum of squares) becomes marginal. By plotting inertia values across a range of K, the goal is to find the "elbow" point where the curve begins to flatten. This point gives a good balance between making the model simple and explaining the data well. As shown in Figure 16, the elbow appears clearly at $K = 2$, suggesting that dividing the dataset into two clusters captures the core structure without overly complicating the segmentation.

Silhouette Score: The Silhouette Score evaluates how well each data point fits within its assigned cluster, compared to other clusters. It ranges from -1 to 1, where higher values indicate better separation and cohesion of clusters. We computed the silhouette score for each value of K and observed a peak at $K = 2$ (see Figure 16). This reinforces the Elbow Method result and further validates our choice to proceed with two tactical clusters.

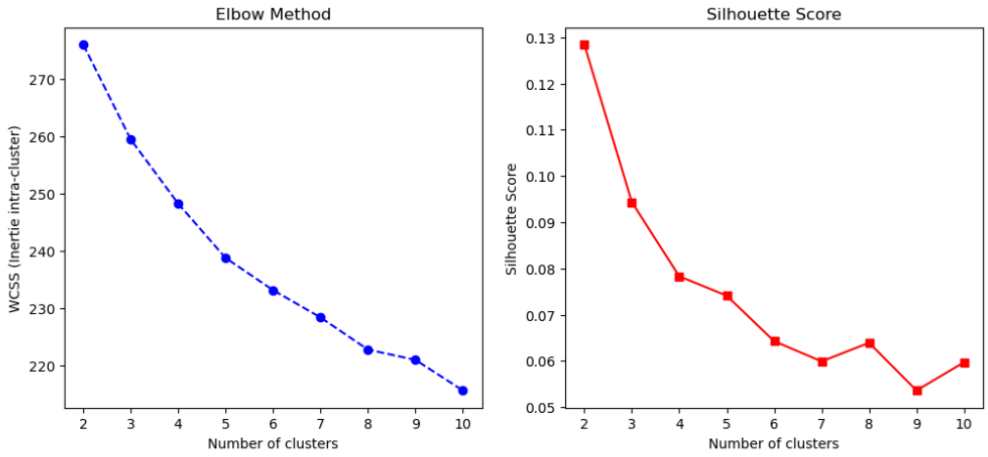

Figure 16: Evaluation of the optimal number of clusters using the Elbow Method (left) and the Silhouette Score (right).

Together, these methods suggest that $K = 2$ is the most suitable number of clusters for our dataset. This decision allows us to focus on analyzing and interpreting two distinct match profiles.

Visualizing Clusters using PCA

To better see how the KMeans clusters are arranged, we used **Principal Component Analysis (PCA)** [13], a method that reduces many variables into just two main ones

while keeping most of the important information. Each match is described by several normalized stats from both teams. PCA transforms these into two new values so we can plot the matches on a 2D graph. This helps us check if the groups found by KMeans still make sense as shown in Figure 17. The visualization was produced using the interactive plotting library Plotly [14]. Indeed, each point represents a Premier League match from the 2022/2023 season, annotated with the home and away team names (e.g., *Nottingham Forest vs Manchester City*), and colored according to its cluster assignment by the KMeans algorithm with $K = 2$. This particular match belongs to Cluster 0, which is associated with a more defensive tactical profile.

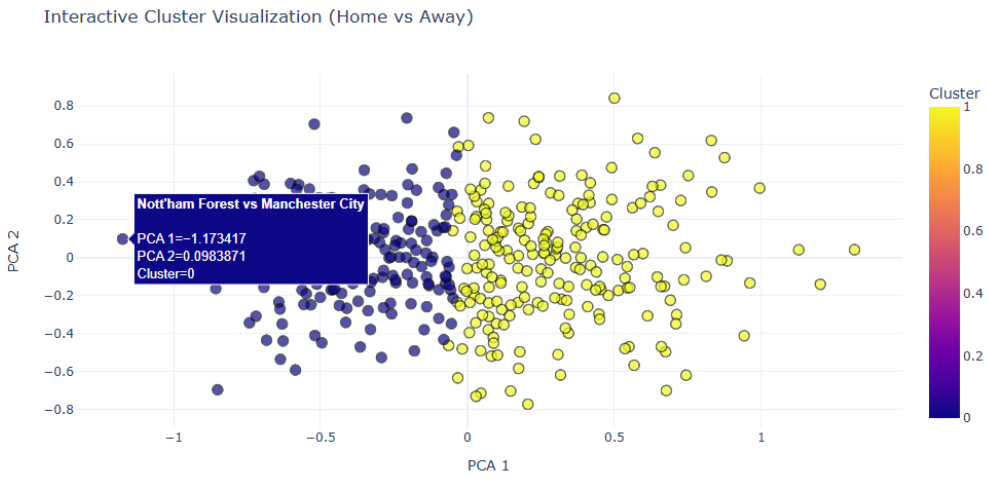

Figure 17: PCA Projection of Premier League Matches Colored by KMeans Clusters ($K = 2$).

The PCA plot shows a clear separation between the two clusters. Matches in **Cluster 0** and **Cluster 1** form distinct groups, meaning the clustering reflects real tactical differences. The labeled team names also reveal that some teams often appear in the same cluster, showing consistent playing styles during the season.

5.3 Cluster Interpretation and Tactical Profiling

Clustering Implementation and Distribution

Before interpreting the clusters, we first normalized all match statistics using standard scaling to ensure that each feature had equal weight in the clustering process. We then ran the KMeans algorithm with two clusters, as identified earlier by the Elbow and Silhouette methods. Each match was represented as a data point combining key statistics from both the home and away teams. The resulting distribution of matches across the two clusters, displayed in Figure 18, shows a fairly balanced split. This balanced partition provides a clear starting point for deeper analysis of the specific tactical traits and differences captured by each cluster.

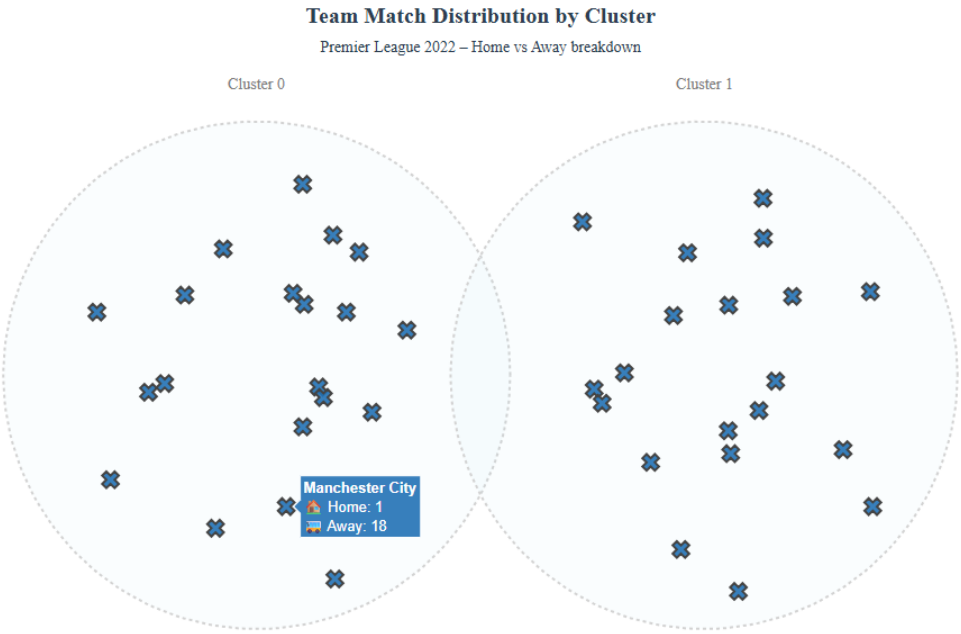

Figure 18: Distribution of matches into clusters using KMeans ($K = 2$).

5.3.1 Interpretation of Clusters Based on Average Match Statistics

The clustering process produced two clearly distinct tactical profiles, which we interpret as follows:

- **Cluster 0 – Defensive-Oriented Matches**

Matches in Cluster 0 tend to follow a reactive and defensively structured tactical approach. Teams involved in these matches generate fewer scoring chances, defend deeper, and rely on physical interventions and clearances to resist attacking pressure.

Table 5.1: Average Statistics in Cluster 0 (Defensive Matches)

Metric	Team Value	Opponent Value	Interpretation
xG (Expected Goals)	1.21	1.58	Teams concede more than they create
Corners Won	3.68	6.06	Low attacking pressure, more corners conceded
Tackles	18.1	16.3	Active defensive pressing
Interceptions	10.1	8.25	Disruptive, reactive defending
Clearances	22.5	15.9	Defensive reliance on clearances
Aerial Duels Won	13.1	14.3	Physical games with aerial challenges

We interpret Cluster 0 as representing matches where teams prioritize strong defensive play. These teams often sit deeper, use low blocks, and engage in aggressive duels combined with quick counter-attacks. The higher numbers of clearances, tackles, and interceptions reflect a reactive and physical style of play, as detailed in Table 5.1. An example of this cluster is the match between Nottingham Forest and Manchester City, where the

lower-ranked home team defends deeply against a technically superior opponent and concedes more attacking opportunities. Looking at the distribution of match results in this cluster, home teams win about 41% of the matches, away teams win roughly 35% , and approximately 24% of the games end in a draw. The teams that appear most frequently in Cluster 0 include Nottingham Forest, Leicester City, and Bournemouth.

- **Cluster 1 – Attacking-Oriented Matches**

Cluster 1 is defined by an active and intense style of play, where teams aim to control the game by keeping possession, creating many scoring opportunities, and applying strong offensive pressure.

Table 5.2: Average Statistics in Cluster 1 (Attacking Matches)

Metric	Team Value	Opponent Value	Interpretation
xG (Expected Goals)	1.94	0.96	Teams create significantly more than they concede
Corners Won	7.44	2.94	High attacking output, more set-pieces won
Tackles	15.6	17.9	Lower defensive work, more possession
Interceptions	7.6	9.89	Teams dominate play, intercept less
Clearances	13.1	26.3	Opponents are forced to clear under pressure
Aerial Duels Won	14.5	12.9	Offensive teams also win physical duels

Cluster 1 includes matches where teams adopt a proactive and dominant style of play. These games are driven by attacking intent, with teams pressing high, maintaining possession, and creating frequent chances. Defensive statistics tend to be lower—not due to weakness, but because these teams control the game and limit their opponents' ability to build attacks. This interpretation is supported by the data shown in Table 5.2. A typical example of this cluster is the match between Arsenal and Manchester United, where both teams aimed to dominate through fast-paced and offensive football. The distribution of results in Cluster 1 also reflects this assertive approach. Home teams win more often, around 55%, while away wins occur in about 23% of cases, and draws represent roughly 22%. The teams that most frequently appear in this cluster are Arsenal, Manchester City, and Liverpool, all known for their consistent attacking strategies throughout the season.

5.3.2 Cluster Membership by Team and Role

This section explores how each team's matches are distributed across the two clusters, depending on whether they played at home or away. The table below (Table 5.3) helps identify tactical patterns based on team profiles and roles.

- **Defensive Teams (Cluster 0 dominant):** Nottingham Forest, Bournemouth, Leicester City, and Wolves have significantly more matches in Cluster 0, especially when playing at home. This suggests that these teams frequently adopt reactive game plans.
- **Attacking Teams (Cluster 1 dominant):** Arsenal, Manchester City, Brighton, and Liverpool appear more in Cluster 1, particularly at home, reflecting a proactive and possession-based approach.

- **Balanced Teams:** Clubs like Aston Villa, Crystal Palace, and Tottenham show a relatively even distribution, suggesting tactical flexibility depending on the opponent.
- The **home vs away role** greatly influences match clustering, confirming that tactical behavior shifts based on context

This breakdown clearly shows that cluster membership is not random, but strongly connected to each team's typical playing style and the match context—especially whether the team is playing at home or away. These patterns are summarized in the Table 5.3 below.

Table 5.3: Match distribution by cluster and role (Home vs Away) for each team.

Team	**Cluster 0 (H–A)**	**Cluster 1 (H–A)**
Arsenal	2–13	17–6
Aston Villa	7–8	12–11
Bournemouth	15–2	4–17
Brentford	10–6	9–13
Brighton	2–16	17–3
Chelsea	5–15	14–4
Crystal Palace	9–6	10–13
Everton	13–4	6–15
Fulham	9–8	10–11
Leeds United	11–6	8–13
Leicester City	14–7	5–12
Liverpool	2–15	17–4
Manchester City	1–18	18–1
Manchester Utd	8–12	11–7
Newcastle Utd	2–9	17–10
Nott'ham Forest	16–3	3–16
Southampton	10–4	9–15
Tottenham	8–6	11–13
West Ham	11–6	8–13
Wolves	15–6	4–13

Overall, clustering helps reveal key tactical patterns that go beyond team names or reputations. Instead of looking only at who is playing, we can now group matches by how the game was played. This gives us a more data-driven way to understand football styles.

Cluster 0 includes matches with a more defensive and reactive approach. In these games, teams often sit deep, defend in numbers, and create fewer chances. This type of match is common when a weaker team faces a stronger one, especially at home.

Cluster 1, on the other hand, shows games with attacking and high-pressure styles. These matches have more expected goals, more corners, and fewer defensive stats because the teams in control keep possession and dominate the game.

This method allows us to look at matches based on playing style, not just team identity. It also helps us explore how teams adjust their tactics depending on the opponent, whether they play at home or away, or how the score changes during the match. For example, even strong teams like *Manchester City* often appear in Cluster 1 when playing at home, but shift to Cluster 0 in tougher away games—showing how tactics change with context.

5.4 Conclusion and Transition

By applying KMeans clustering on match statistics from the 2022/2023 Premier League season, we identified two distinct tactical profiles. This unsupervised learning approach highlighted meaningful patterns in match dynamics, distinguishing between games dominated by defensive resilience and those characterized by offensive control. It also revealed how team roles (home vs away) influence tactical behavior, with underdog strategies often linked to the defensive cluster.

In the next chapter, we move from descriptive segmentation to predictive modeling. Using supervised machine learning, particularly Random Forests, we aim to classify matches based on their tactical style and determine which features are most influential in shaping these profiles. This approach will provide a deeper understanding of the key drivers behind the identity of matches and will offer valuable tools to anticipate tactical trends in modern football.

6. Supervised Modeling with Random Forest

6.1 Introduction and Methodology

After grouping matches with unsupervised clustering, we now move to a supervised learning approach to better understand how matches unfold. For this part, we use the Random Forest algorithm which is known for being both powerful and easy to interpret. It helps us predict match types and understand what in-game actions matter most.
In fact, `Random Forest` [15] is a machine learning algorithm that builds many decision trees using different parts of the data and different features. Each tree gives a prediction, and then the Random Forest combines all of them to make a final decision. For classification, it chooses the most common result among the trees, and for regression, it takes the average. This makes the model more reliable and less sensitive to noise than using just one decision tree. What makes this method especially useful is that it shows us which features are most important for prediction. In our case, it helps reveal which match statistics like passes, duels, or possession play the biggest role in shaping tactical styles—whether a team tries to control the game or prefers to sit back and react.

Objective. With Random Forests the goal is not to predict results but to understand which stats best explain tactical behavior and how team roles like home or away influence these patterns so we can eventually build new tactical groups based on this information. To enable this analysis, we restructured the dataset from a *match-level* to a *team-level* format. Instead of having one row per match (with both home and away data), we split each match into two rows — one for each team's performance.

Transformation of Match Data Format

Original Format				
Team	Opponent	Score_x	Score_y	Label
Arsenal	Chelsea	2	1	1

$\Longrightarrow$

Transformed Format				
Team	Opponent	Side	Score	Label
Arsenal	Chelsea	Home	2	1
Chelsea	Arsenal	Away	1	0

This change was important for two main reasons. First, it let us study each team's performance separately, making it easier to compare how teams play at home versus away. We split each match into two rows—one with features ending in `_x` for the home team and one with `_y` for the away team, so the data clearly matches the team's side. This avoids mixing home and away stats. Second, this setup allows us to train a model that can learn different patterns for home and away performances, helping us better understand tactics from both angles.

After transforming the data, we applied a consistent preprocessing step. Initially, when some stats were missing for the home team in the first half of the dataset, we filled them using the corresponding away team stats from the same match, and did the reverse for the second half to ensure completeness. Then, we rescaled all features between 0 and 1 using the Python function `MinMaxScaler` from the Scikit-learn library to keep features

comparable and avoid those with large values from having too much influence. Finally, we created two label columns, one for home performance and one for away, marking rows with home stats as "home" (`_x` features and label = 1) and those with away stats as "away" (`_y` features and label = 0) so the model can clearly distinguish team roles.

With the dataset now clearly split between home and away team performances using `_x` and `_y` features and labels, we trained a Random Forest Classifier to understand how well match statistics alone can tell if a team was playing at home or away. We used the model to check the prediction accuracy and to see which features were the most important overall for this task. To go deeper, we also **figured out the most important stat for each team in each match**. We did this by multiplying each normalized stat by its feature importance score and picking the one with the highest result. This gave us tactical insight at the match level, helping us see, for example, whether `Shot_Accuracy_x` or `Tackles_Defense_Ratio_y` was the key to a team's performance.

6.2 Model Performance and Feature Insights

The Random Forest classification model achieved solid results on the task of distinguishing home and away team performances, with an accuracy exceeding 70% on the test set. But more importantly, it helped us understand which stats matter most through its feature importance system, giving us tactical insights from two angles. First, we looked at global feature importance, where the model ranked all performance stats based on how often they helped across the whole season. This tells us which variables were most useful overall to spot home or away behavior. Then, we checked match-specific dominance, where we figured out for each team in each match which stat had the strongest impact on the model's decision. This helped us understand what stood out in each performance based on the 36 normalized metrics. Figure 19 below shows the global distribution of the most important features per match, ranked by how frequently each metric appeared as the top contributor—highlighting recurring patterns such as the presence of `Long Balls` or `Shot Accuracy` across many games.

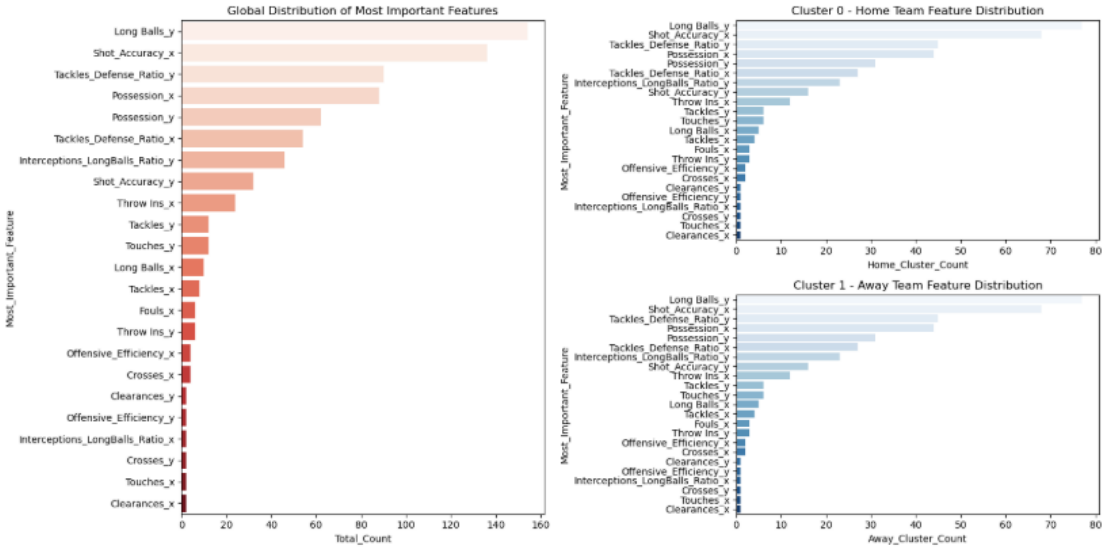

Figure 19: Most Frequent Top Features per Match (Random Forest)

Case Study : What Makes Manchester City Stand Out

To better understand how most important features translate into concrete match contexts, we highlight the case of **Manchester City**, a team known for its tactical versatility and dominance throughout the 2022/23 season. In Figure 20, each row shows one of their games with the key stat that influenced the model, whether they played home or away, the tactical style guessed by the model, and the score. This gives a simple way to understand what often makes City different, like their strong finishing or control in key moments.

	Home	Away	Most_Important_Feature	Team_in_Cluster	Cluster	Score_x	Score_y
9	West Ham	Manchester City	Possession_y	Manchester City	1	0	2
11	Manchester City	Bournemouth	Shot_Accuracy_x	Manchester City	0	4	0
28	Newcastle Utd	Manchester City	Possession_y	Manchester City	1	3	3
31	Manchester City	Crystal Palace	Possession_x	Manchester City	0	4	2
45	Manchester City	Nott'ham Forest	Possession_x	Manchester City	0	6	0
57	Aston Villa	Manchester City	Shot_Accuracy_y	Manchester City	1	1	1
62	Wolves	Manchester City	Interceptions_LongBalls_Ratio_y	Manchester City	1	0	3
74	Manchester City	Manchester Utd	Offensive_Efficiency_x	Manchester City	0	6	3
79	Manchester City	Southampton	Shot_Accuracy_x	Manchester City	0	4	0
96	Liverpool	Manchester City	Touches_y	Manchester City	1	1	0

Figure 20: Sample of Manchester City matches with most important feature per match, cluster label, and score.

Some match examples from Figure 20 help explain the tactical patterns found by the model:

- In the 4–0 home win vs Bournemouth, `Shot_Accuracy_x` was key, showing how City's strong finishing made the difference in front of goal.
- In the 3–3 away draw at Newcastle, `Possession_y` stood out, showing that even away, City kept control of the ball — a core part of their usual game style.
- In the 6–3 home win vs Manchester United, `Offensive_Efficiency_x` was most important, proving how well City turn chances into goals, often using quick passing in the final third.
- In the 3–0 away win at Wolves, `Interceptions_LongBalls_Ratio_y` was dominant, showing their ability to stop long passes and defend high, which fits their pressing system even away from home.

These examples reflect how City's typical 4-3-3 formation plays out in the stats: strong finishing and chance creation at home, possession and pressing away. The model highlights how different stats stand out depending on context, giving more depth than a simple win/loss analysis.

6.3 Tactical Interpretations and Global Observations

Using **Random Forest** allowed us to both classify team performances and reveal key tactical patterns through feature importance. Comparing home and away matches (see Figures 19 and 20) shows clear differences in style. On the one hand, at home, teams generally control the game with strong possession and precise passing, as reflected by features like `Touches_x` and `Possession_x`, which lead to better chance creation and finishing, highlighted by metrics such as `Shot_Accuracy_x` and `Offensive_Efficiency_x`. These offensive strengths, often seen in Manchester City's home matches, combine with active defensive efforts through tackles and interceptions, while long balls help to stretch or shift play. On the other hand, away teams tend to adopt a more defensive, reactive approach, with important features like `Clearances_y`, `Interceptions_y`, and `Tackles_y` signaling compact defending and pressure absorption. Offensively, they rely more on quick counter-attacks and long balls but struggle with shot accuracy and final-third efficiency. This pattern appears in matches such as City's away game at Wolves (20), where defensive transitions were crucial. Overall, these results confirm the well-known tactical differences between home and away games, quantifying them with model-driven feature importance, and offering a dual perspective that combines both predictive accuracy and deeper tactical understanding.

6.4 Conclusion and Transition

This analysis demonstrates that match identity—whether a team played at home or away—can be predicted well using match statistics, and the main features behind these predictions match real tactical behaviors like possession, finishing, defense, and long-ball use. However, simply dividing matches into home or away doesn't fully capture how teams actually play. For example, a team playing away against a top opponent might use very different tactics than against a weaker team, and home teams don't always dominate the same way, especially versus strong rivals.

To explore this further, the next section expands the analysis by using clustering and a refined Random Forest model trained on four classes that mix match location with opponent strength (like home vs strong team or away vs weaker team). To capture these differences, the next step extends the analysis by training a model on four categories that combine match location and opponent strength, using ELO ratings to measure team quality throughout the 2022-2023 Premier League season. By mixing these ELO-based contexts with clustering and feature importance, we aim to better understand the many ways teams adapt their play in this competition.

7. ELO-Informed Clustering for Tactical Match Analysis

7.1 Motivation & Contextual Refinement of Tactical Groupings

The previous binary classification approach that simply classified matches as home or away showed we can tell a match's identity from stats. But just knowing the location does not fully capture and figure out the complexity of tactics. A team's approach in a home match against a title contender differs significantly from a home match versus a relegation-threatened side. Likewise, an away performance can vary drastically depending on the caliber of the opponent. To better reflect these differences, we added team strength using ELO ratings—updated after each match to show current form. Combining location and opponent quality, we defined four match types: home vs strong, home vs weak, away vs strong, and away vs weak. In practice, to use ELO ratings in our analysis, we had to reorganize the data. At first, each match was in one row with both teams and their **ELO scores** recorded after the game. But since our model looks at each team separately, we needed to create one row per team, with the ELO scores from just before the match for both the team and its opponent.

With these adjustments, we applied **KMeans clustering with** $K = 4$, combining usual match stats (like expected goals, defensive actions, and passes) with ELO scores. The resulting clusters reflect both tactical similarities and the context of the matches, taking into account the difficulty and balance between the teams involved. To support the analysis, we developed a `Streamlit web application` [16]. This interface allows users to explore the distribution of clusters, examine tactical profiles, and filter matches based on specific criteria such as team name, match location, or strength of the opponent. Figure 21 provides an overview of the application.

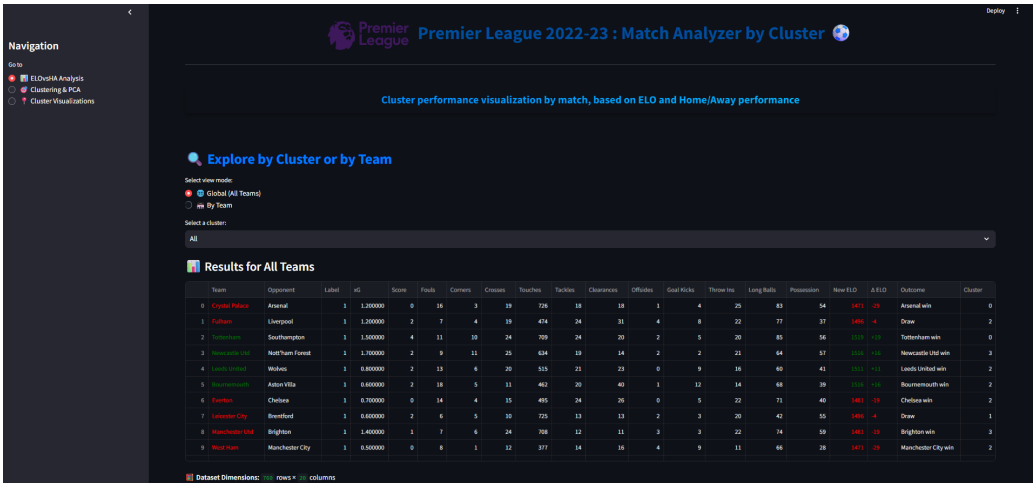

Figure 21: Homepage of the interactive Streamlit dashboard for ELO vs HA Cluster Analysis.

7.2 Model Construction and Cluster Overview

The clustering model presented in this section builds upon the structural foundation established in Chapter 5, where each team-match instance is treated as a unique observation. However, two key enhancements were introduced to better reflect the competitive and tactical context of matches: **(i) the integration of a dynamic ELO score for each team prior to kickoff**, and **(ii) the use of four clusters** ($K = 4$) rather than two, enabling a more granular partitioning of match types. By combining match location with opponent quality through ELO, the model can better distinguish different situations—like playing at home against a strong team versus away against a weaker one—giving a clearer picture of tactical adjustments. The `KMeans algorithm` was run on a set of key features such as xG, corners, crosses, possession and the ELO scores. These features were chosen because they are meaningful tactically and statistically relevant. Adding ELO helps the model not only group matches by playing style but also by the challenge level of the opponent. Following standard preprocessing (including scaling and normalization), the clustering algorithm identified four distinct groups of match behaviors. For instance, one cluster may consist primarily of high-possession home matches against weaker sides, while another may group low-possession away performances against high-ELO opponents. This setup makes it easier to analyze how teams change their tactics based on the situation. The Figure 22 below provides an overview of the four-cluster distribution across all matches throughout the season.

Figure 22: Distribution of matches into clusters using KMeans (K = 4).

Cluster-Based Distribution: Home Matches

This subsection focuses on the KMeans clustering results for **home matches** throughout the season. Each team's performance in a home game was grouped into one of four clusters, based on team's ELO rating after the match. This method captures both the on-field actions and the context in which teams made their tactical decisions. Figure 23 displays the spatial distribution of home matches across the four clusters. Each point corresponds to one match played at home, positioned based on the underlying feature space after dimensionality reduction (e.g., via PCA). The coloring reflects the assigned cluster, allowing for visual inspection of how matches are grouped.

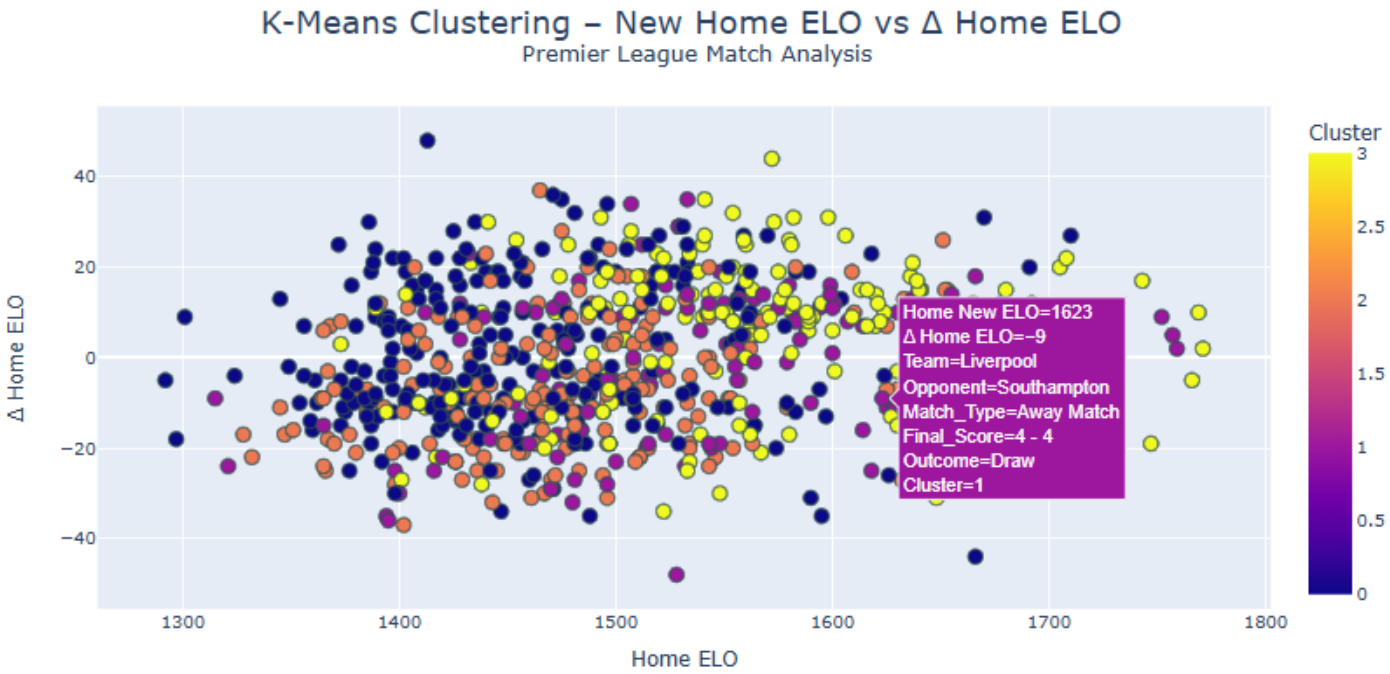

Figure 23: Clustering of all home matches using KMeans ($K = 4$) with ELO-based context.

The plot highlights clear patterns among home matches. One cluster groups games where the home team controls possession, usually against weaker. Another cluster includes matches where possession is lower, suggesting the team faced stronger opponents or played more cautiously. Using ELO scores helps distinguish between matches that may look similar in stats but differ in difficulty—for example, a 2–0 home win against a strong team versus a weaker one. The clusters also reflect how teams adjust their style based on opponent strength, showing flexibility in their tactics depending on the match context. This clustering gives a clearer picture of how teams behave at home under different conditions.

Cluster-Based Distribution: Away Matches

Following the same methodology, we applied the KMeans clustering algorithm to all **away matches**. As before, each match was assigned to one of four clusters based on a combination of key performance indicators and the ELO score of the away team after to the fixture. This helps us clearly understand how teams change their playing style when

they are playing away from home. Figure 24 illustrates the spatial distribution of home matches across the four clusters.

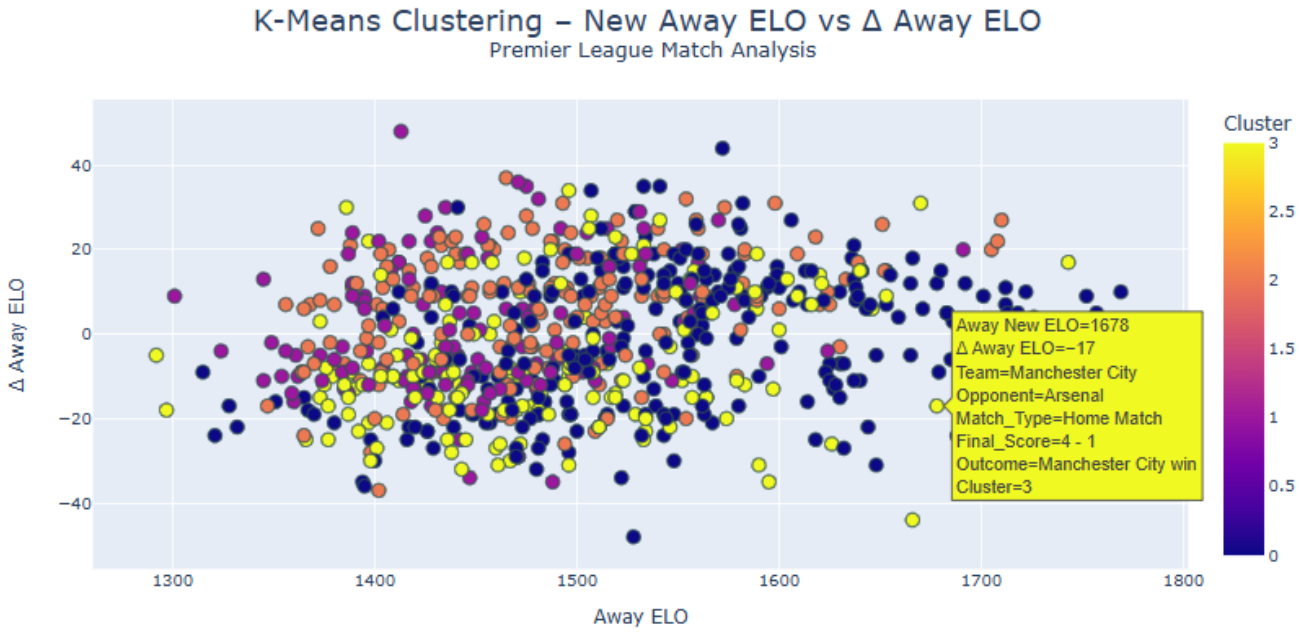

Figure 24: Clustering of all away matches using KMeans ($K = 4$). Clear tactical deviations are observed between matches of differing ELO difficulty.

To sum up, one cluster stands out with low possession, many defensive clearances, and frequent long balls, showing teams using a defensive style when facing stronger opponents with higher ELO ratings. Another cluster groups away matches where teams kept more possession and created more chances, usually when stronger teams play against weaker opponents. The ELO rating helps to clearly separate these situations, showing that similar stats can mean very different things depending on the opponent's strength. In general, the clusters reveal how teams change their tactics not just based on playing away, but also depending on how strong their opponent is, giving a clearer picture of their in-game behavior.

ELO-Based Segmentation for Interpretation

To make the analysis of match clusters more interpretable, we introduce a segmentation of matches according to post-match ELO scores. The ELO rating used for segmentation corresponds to the post-match ELO of the team playing at home or away, depending on the perspective shown in Figures 23 (Home) and 24 (Away). In other words, for home match figures, the segmentation is based solely on the home team's ELO after the match, and similarly for away matches, it is based on the away team's ELO after the match. This makes it possible to assess how a team behaves in different contexts depending on its own relative strength at that moment in the season. While the ELO thresholds are empirically defined, they are grounded in the actual distribution of ELO scores during the season and offer a practical way to stratify the data.

The three ELO groups are defined as follows:

- **Low ELO**: 1292 – 1400
- **Medium ELO**: 1400 – 1600
- **High ELO**: 1600 – 1771

These ELO groups are not intended to serve as absolute standards of team quality but practical categories to help understand different match situations. Matches with teams in the low ELO group usually mean facing stronger opponents or poor team form, while high ELO matches involve stronger teams that control the game more confidently, no matter where they play. Segmenting the data in this manner enables us to explore how the same team may behave differently when operating within a low-ELO or high-ELO context, and how cluster membership interacts with the level of competitive difficulty. This approach will be particularly useful in the upcoming sections, where we interpret the clusters one by one, using both performance statistics and ELO context to explain the dominant tactical signatures associated with each group.

7.3 Understanding the ELO-Based Clusters

ELO range: 1292 to 1413 points

In this initial ELO range, the match distribution among clusters is as follows: Cluster 0 contains 72 matches, Cluster 2 includes 37, while Clusters 1 and 3 are less represented with 8 and 7 matches respectively. The dominant clusters in this segment are clearly Cluster 0 and Cluster 2.

- **Cluster 0:** This cluster mostly includes of matches between top-tier teams (Top 6) and bottom-ranked teams. The matches cover the full range of ELO scores but are mostly between 1292 and 1340 points, which are quite low. This could mean a few things: many of these games happened early in the season when all teams started with an ELO of 1500; some matches involved teams with similar levels, whether from the Top 3 or the bottom of the table; and some games had many goals and big changes in rankings. For example, Everton's 5-1 win at home against Brighton in week 35 led to a large 48-point increase in their ELO.
- **Cluster 1:** Most matches in this cluster resulted in home losses, often with significant point losses (up to -40 points). One exception is Leeds' 1-0 win against Southampton in matchweek 26, earning a modest +10 point gain. This cluster is characterized by unfavorable outcomes for home teams.
- **Cluster 2:** Matches are again dominated by defeats for home sides, but some home wins are present. The magnitude of ELO point changes is more moderate ($\pm$20 points), such as Bournemouth's 1-0 win over Liverpool in matchweek 27 (+19 points).
- **Cluster 3:** This cluster includes only 7 matches, so it's too limited to draw clear conclusions. However, it shows similar patterns, with close games and moderate changes in points—for example, Leicester's 0-1 home loss to Southampton.

Conclusion: This ELO range includes many matches from the end of the season (weeks 25–38), often with teams facing tough challenges. Clusters 0 and 2 are the most common here, showing unpredictable results with big surprises and fights against relegation. Home teams sometimes had difficulties but also managed to win important games.

ELO range: up to 1600 points

In this broader range, match distribution across clusters is more balanced: 184 matches in Cluster 0, 156 in Cluster 2, 109 in Cluster 3, and 90 in Cluster 1. Clusters 0 and 2 remain the most populated.

- **Cluster 0** can be divided into different groups based on ELO scores, which helps explain the variety of match outcomes within this cluster. For matches with ELO below 1475 points, we often see mid-table teams facing each other, with mixed results—like Nottingham Forest's 2-0 win against Leicester in matchweek 20, which led to a significant 36-point gain. In the mid-range between 1475 and 1550 points, matches tend to favor home victories, reflecting games involving mid-to-high ranked teams; an example is Brentford's 3-0 home win over Southampton in matchweek 25. For matches above 1550 points, the cluster mostly includes high-level games between top 10 clubs, such as Manchester United's 2-0 victory over Nottingham Forest in matchweek 17, where ELO changes were generally moderate, between 10 and 40 points. These patterns align with the cluster's performance metrics, which show how possession, expected goals, and defensive actions vary across these ELO tiers, revealing tactical differences linked to team strength and match difficulty.

- **Cluster 1** groups matches often involving well-known teams but with surprising results. For example, Newcastle's 1-0 win over Fulham led to a 35-point gain, while Brighton suffered a heavy 1-5 home loss to Everton, losing 48 points. Most matches in this cluster are closely contested, with ELO changes usually staying within ± 15 points. This reflects a mix of competitive games where the outcomes were somewhat unexpected compared to the teams usual performance levels.

- **Cluster 2** is marked by a high number of home losses, with 101 defeats compared to 51 wins, and an average ELO around 1486. This cluster mostly includes matches where weaker teams face stronger opponents, leading to unfavorable results at home. For example, Crystal Palace's impressive 5-1 away win against Leeds in matchweek 30 earned them 37 ELO points. Despite their strength, top clubs like Tottenham and Arsenal also experienced important defeats within this group. The statistics highlight a pattern of home teams struggling against higher-ranked opposition.

- **Cluster 3** mainly includes matches with ambitious teams showing strong and consistent performances at home. Out of 109 matches, 76 are home wins, like Brighton's 3-0 victory over Arsenal in matchweek 36, which earned them 44 ELO points. There are also notable away wins, such as Aston Villa's 2-0 win against Tottenham, causing a 34-point loss for Spurs. The average ELO in this cluster is 1528, indicating generally higher-quality teams. These matches often show strong performance features such as higher possession and more attacking actions, reflecting the confident and controlled playing style typical of these teams.

`Conclusion:` In this ELO range, match outcomes vary widely. Home advantage is more noticeable when teams have an ELO above 1475, but strong clubs appear in all clusters, adding unpredictability. Cluster 2 highlights how even top teams can sometimes struggle against weaker opponents.

ELO range: 1600 to 1771 points

In this highest ELO range, the number of matches is uneven across clusters: Cluster 0 has 10 matches, Cluster 1 has 38, Cluster 3 has 42, and Cluster 2 only 11. Clusters 1 and 3 are the most common, and all clusters represent games between top-level teams.

- **Cluster 0:** This cluster includes matches between top teams like Arsenal, Manchester City, and Newcastle, with an average ELO of 1657. Among the 10 games, 6 were home wins—for example, Manchester City's 3-1 victory over Arsenal in matchweek 12, which earned them +31 ELO points. There were also notable losses, such as Arsenal's 0-3 home defeat to Brighton in matchweek 36, resulting in a -44 point change. These matches typically show strong performance stats like high expected goals and possession, reflecting the quality and intensity expected at this level.
- **Cluster 1:** This cluster includes 38 matches with an average ELO of 1662, showing closely contested games with small point changes. Examples are Arsenal's 1-0 loss to Everton in matchweek 22, which led to a -24 point change, and Manchester United's 3-1 win over Fulham in matchweek 28, earning +13 points. These matches usually have balanced stats and reflect intense competition with modest shifts in team strength.
- **Cluster 2:** This smaller cluster of 11 matches includes several games involving Newcastle United, like their 3-0 win over Leicester in matchweek 37, which gained them +26 points. The results show a mix of moderate wins and losses, mostly between Top 6 teams, reflecting competitive matches with balanced performance metrics.
- This is the largest cluster at this ELO level, with 42 matches averaging an ELO around 1650. It is mostly made up of home wins, such as Brighton's 3-0 victory over Liverpool (+33 points) and Arsenal's 4-1 win against Crystal Palace (+25 points). There are also losses, like Liverpool's 1-0 defeat at Bournemouth (-18 points). The cluster shows solid performances with moderate changes in ELO ratings.

`Conclusion:` In this high ELO range, matches mainly involve elite teams competing for top positions or European spots. Cluster 3 is notable for strong and consistent home team performances. Overall, the four clusters capture different tactical patterns and competitive levels, reflected in key features like possession, shot_accuracy, and defensive actions, allowing a clear distinction between dominant, balanced, and reactive match styles.

7.4 Conclusion

The cluster analysis by ELO levels shows clear differences in team tactics and results depending on team strength and match location. Table 7.1 summarizes these patterns:

Home matches in Cluster 3 have the best average ELO gains (+17.1) and smaller losses, reflecting strong home performances by top teams. Away matches in the same cluster have the largest average losses (-20.6), highlighting the difficulty of playing away against tough opponents. In addition, Home matches in Cluster 1 also show good results (+15.0 on average), while Away matches in Cluster 0 are more unstable, with moderate wins but the highest average losses (-16.4). This confirms that context, including location and opponent strength, strongly affects team performance.

Table 7.1: Cluster-wise ELO and Δ ELO Analysis (Home vs Away)

Cluster	Match Type	Avg. ELO	Match Count	Δ ELO (Wins)	Δ ELO (Losses)
0	Away Match	1542.2	152	+10.5	-16.4
0	Home Match	1544.8	113	+15.0	-14.1
1	Away Match	1443.1	48	+13.5	-13.1
1	Home Match	1456.9	86	+16.5	-9.8
2	Away Match	1474.2	100	+12.2	-11.1
2	Home Match	1496.5	104	+15.8	-10.5
3	Away Match	1486.2	80	+12.9	**-20.6**
3	Home Match	1489.0	77	**+17.1**	-11.8

To complement this table, Figure 25 displays a `radar comparison of Home vs Away ELO per Cluster`, helping visualize the ELO levels across clusters and match types. This figure reinforces the table insights by clearly showing which contexts (e.g., Cluster 1 Home or Cluster 3 Away) are associated with higher or lower ELO performances.

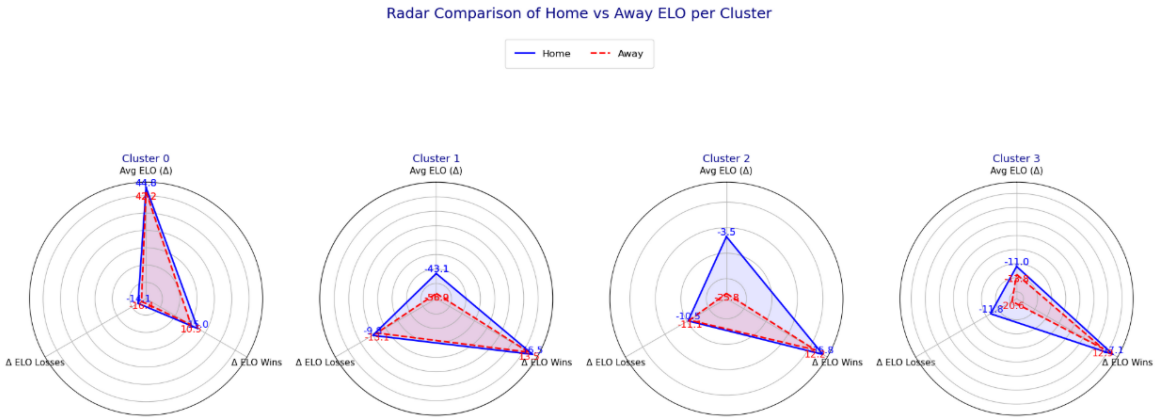

Figure 25: Radar Comparison of Home vs Away ELO per Cluster

Indeed, clusters 0 and 2 represent most Premier League matches across all ELO levels, while clusters 1 and 3, though less common, involve key games between top teams where results are critical. Home advantage plays a significant role in boosting stronger teams performance, especially in Clusters 1, 3. In contrast, away matches in clusters 0 and 3 tend to be less predictable and often lead to larger losses for weaker or mid-table teams. This highlights how team strength, tactics, and match location together shape outcomes. Therefore, including contextual features like ELO differences, venue, and opponent quality is vital for understanding performance and improving prediction models. Future enhancements could come from adding factors such as team form, injuries and fatigue from other competitions, to improve the accuracy of such models.

General Conclusion

This project, dealing with `Modeling team playing styles based on match location with home versus away performance analysis`, offered an in-depth exploration to figure out tactical trends and performance dynamics in the 2022–2023 English Premier League. Focusing on the influence of match location, we developed a structured, data-driven approach to understand how teams adapt their playing styles when competing at home versus away.

By combining open football match data with statistical modeling, `ELO rating systems` and Machine Learning techniques, we uncovered significant patterns in how team behavior shifts depending on both venue and opponent strength. The use of evolving ELO scores allowed us to contextualize each fixture's difficulty and better evaluate team performance over time. Through `Principal Component Analysis` and `KMeans clustering`, we identified distinct tactical match profiles, while `Random Forest algorithm` helped predict match identity and understand the importance of key statistical features such as possession, finishing efficiency, and defensive resilience.

These results validated our methodology, showing that match location alone is not enough to define tactical behavior, contextual elements like team strength, opposition quality, and evolving form are also crucial. Our visualizations and predictive models revealed the complexity of tactical decision-making and highlighted the non-linear nature of home advantage in modern football.

Despite these achievements, several improvements could further enrich the analysis. First, even though the 2022–2023 Premier League season offered a solid foundation, incorporating parallel competitions such as the `FA Cup`, `UEFA Champions League`, or `Europa League` based on qualification through league standings could reveal how teams adjust their tactics across different contexts. Understanding tactical adaptations due to fatigue, travel, or changes in competition priorities would make the model even more realistic. Additionally, expanding to multiple seasons would allow the study of tactical evolution over time, including coaching changes, formation shifts, and player transfers.

Another significant avenue for improvement lies in the inclusion of richer contextual features, such as formations, player roles, or in-game positional data. Notably, coupling this approach with video analysis or expert-coded match annotations could serve to validate and refine the model outputs, bridging the gap between raw data and real football intelligence. Such integrations would enhance the practical relevance of our findings for real-world applications such as scouting, tactical preparation, and performance diagnostics.

In summary, this project provided an enriching opportunity to combine theoretical knowledge and technical implementation, advancing our skills in data science, sports analytics, and performance modeling. It forms a strong foundation for future research and professional tools aimed at understanding and forecasting tactical behaviors in competitive football.

Bibliography

[1] Linköping university: Official website and research overview. https://liu.se/en, June 2025.

[2] Campus valla: Main campus of linköping university. https://liu.se/en/article/campus-valla, June 2025.

[3] Premier league : Official website. https://www.premierleague.com/. Accessed March–July 2025.

[4] Premier league dataset. https://www.kaggle.com/datasets/memocan/premier-league-games-and-player-stats-2021-2024. Accessed in March 2025.

[5] Python programming language. https://www.python.org/, June 2025.

[6] Jupyter notebook - Interactive web-based computing platform. https://jupyter.org, June 2025.

[7] Transfermarkt and fbref. https://www.transfermarkt.com, https://fbref.com. Accessed March–July 2025.

[8] Fotmob - Premier league 2022–2023 overview. https://www.fotmob.com/fr/leagues/47/overview/premier-league?season=2022-2023. Accessed March–July 2025.

[9] Matplotlib - Visualization with python. https://matplotlib.org/, June 2025.

[10] World football elo ratings. https://www.eloratings.net/about & https://en.wikipedia.org/w/index.php?title=World_Football_Elo_Ratings&oldid=1288237011. Accessed in March and April 2025.

[11] Scikit-learn - Machine Learning in Python. https://scikit-learn.org/stable/. Accessed in April 2025.

[12] Kmeans documentation. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. Accessed in April 2025.

[13] PCA documentation. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. Accessed in April-May 2025.

[14] Plotly documentation. https://plotly.com/python/. Accessed in April 2025.

[15] RandomForest documentation. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed in April 2025.

[16] Streamlit • A faster way to build and share data apps. https://streamlit.io/. Accessed in May 2025.

Résumé

Réalisé dans le cadre d'un stage académique en collaboration avec la l'université de Linköping (LiU), ce projet porte sur l'analyse des styles de jeu des équipes de football selon le lieu du match (domicile ou extérieur), avec pour terrain d'étude la saison 2022–2023 de Premier League anglaise. L'objectif principal était de comprendre comment les équipes adaptent leurs comportements tactiques en fonction du contexte géographique, à partir de données publiques issues de matchs réels. La méthodologie reposait sur un pipeline rigoureux de type `ETL` (*Extract, Transform, Load*), intégrant la collecte, le nettoyage, la structuration et le chargement des données, suivi de l'application de techniques statistiques et d'algorithmes de *machine learning*. Les principaux outils mobilisés étaient `Python`, `Pandas`, `Scikit-learn` ainsi que des bibliothèques de visualisation comme `Matplotlib` et `Plotly` ou `Streamlit`. Grâce à l'implémentation du système de notation ELO et à des algorithmes de *clustering* (PCA + KMeans) et de classification (Random Forest), le projet a permis d'identifier des profils de matchs distincts et d'évaluer les variables les plus influentes. Ce travail a permis de mettre en lumière la complexité des décisions tactiques, d'acquérir une expertise technique approfondie en traitement de données sportives, et de mener une réflexion poussée sur la contextualisation des performances en football.

Mots-clés : Football, Analyse de données, Machine learning, ELO, Clustering, Premier League, Analyse de performance sportive, Python.

Abstract

Conducted as part of an academic internship in collaboration with Linköping University (LiU), this project focuses on analyzing football team playing styles based on match location, using data from the 2022–2023 English Premier League season. The main objective was to understand how teams adjust their tactical behavior depending on whether they play at home or away, relying on publicly available match data. The methodology relied on a rigorous `ETL` (*Extract, Transform, Load*) pipeline involving data collection, cleaning, structuring, and loading, followed by the application of statistical modeling and machine learning techniques. Key tools included `Python`, `Pandas`, `Scikit-learn` and visualization libraries such as `Matplotlib` and `Plotly` or `Streamlit`. By implementing the ELO rating system and leveraging both clustering algorithms (PCA and KMeans) and classification models (Random Forest), the project identified distinct match profiles and highlighted the most influential performance indicators. This work demonstrated the complexity of tactical decision-making, provided deep technical expertise in football data processing, and offered valuable insight into the contextual dynamics shaping modern football performance.

Keywords : Football, Data analysis, Machine learning, ELO, Clustering, Premier League, Performance modeling, Python.