City University of London

MSc in Data Science

Project Report

2024

Applying Data Science and Visual Analytics to Explore Football Event Sequences

Andreas Ioannides

Supervised by: Gennady Andrienko

October 2nd 2024

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.
Signed: Andreas Ioannides

Abstract

This dissertation develops a comprehensive framework for analyzing football event sequences enhancing the understanding of game dynamics and tactical strategies. Focusing on detailed event data from the UEFA European Championship match between Portugal and Iceland, the research applies advanced data science techniques, including clustering, Principal Component Analysis (PCA), and machine learning. By integrating these methodologies, the study identifies patterns in passing, shooting, duels, and fouls, providing deeper insights into player performances and team strategies. The findings reveal that successful football strategies are not solely the result of isolated events but are deeply intertwined with the dynamic sequence of interactions on the pitch. This research bridges gaps in traditional football analytics and offers actionable insights for coaches, players, and sports scientists, promoting a more nuanced understanding of football tactics.

Keywords: Football Analytics, Event Sequences, Machine Learning, Clustering, Data Visualization

Table of Contents

Chapter 1 – Introduction and Objectives

1.1 Background of the Problem

Football is celebrated worldwide for its fluidity, complexity, and strategic interplay. Unlike other sports, football is continuous and dynamic, where events occur rapidly. This constant motion and tactical complexity make football analytics particularly challenging. Traditional football analytics relies on isolated metrics such as possession percentages, shot counts, and player ratings. While useful, these metrics often provide a fragmented view of the game, failing to capture the underlying sequences and patterns that define a team's strategy and performance. For instance, a high possession rate might indicate dominance but doesn't reveal how the ball was moved, how the team structure adapted, or how specific passing sequences led to scoring opportunities.

Football events like passing sequences, shots, duels, and fouls collectively influence the flow and outcome of a match. A single pass can trigger a chain reaction, altering the defensive structure, opening space for an attack, or leading to a goal. Similarly, duels and fouls can disrupt rhythm, shift momentum, and impact the psychological aspect of the game. Understanding these interactions requires sophisticated analytical methods, beyond static metrics, to provide a holistic view of how these events shape match outcomes.

Recent advancements in data collection have significantly improved the granularity of football data. Modern datasets include detailed event data for every action on the pitch, such as passes, shots, duels, and player movements, captured with high temporal and spatial revolution. These datasets, often covering entire seasons from major leagues, offer an opportunity to apply advanced data science techniques to uncover the hidden dynamics of the game. Despite these advancements, there is a significant gap in how this data is analyzed and visualized. While there is plenty of data, the challenge lies in transforming it into actionable insights that influence decision-making.

Current analytical methods often struggle to address the data's complexity, leading to analysis that may overlook critical aspects of the game. For example, network analysis can identify key players in a passing network, but might not account for contextual factors like match location or opponent strength. Similarly, machine learning models can predict outcomes based on historical data, but may miss the nuanced interplay of sequences leading to those outcomes.

This dissertation aims to bridge the gap by developing a comprehensive framework for analyzing and visualizing football event sequences. Focusing on key aspects such as passing patterns, play sequences, duels, shots, and fouls, this research seeks to provide a deeper understanding of how these events interact to influence match outcomes. This study will uncover the underlying patterns defining successful team strategies and player performances, through clustering, network analysis, Principal Component Analysis (PCA), and machine learning.

1.2 Reasons for the Choice of Project

This dissertation is deeply rooted in the evolving field of football analytics, addressing the need for more sophisticated approaches to understanding the game. Traditional metrics, while valuable, often fail to capture football's complexity, where outcomes are shaped by sequences of interactions rather than isolated events. This research seeks to bridge this gap by focusing on the dynamic analysis of event sequences, driven by the limitations of current methods and the potential for deeper insights into football strategy and performance.

A key reason for choosing this project is to **address current gaps in football analytics**, particularly in event sequences analysis. Current methodologies provide insights into individual events, such as pass success rates or shot frequency, but often overlook the sequential and contextual event nature. For example, analyzing a pass in isolation provides limited information, but understanding it as part of a sequence reveals much more about team strategy, player roles, and scoring opportunities. This project aims to develop a framework that analyzes individual events and examines how these events interact over time to influence the game's flow and outcome.

The increasing availability of detailed football match data facilitates the **leverage of advanced data analysis techniques.** Modern datasets, capturing every on-pitch action with a highly temporal and special resolution, enable the application of clustering, Principal Component Analysis (PCA), network analysis, and machine learning. These techniques can identify patterns and relationships that traditional analysis might miss. By leveraging these methods, the project aims to provide actionable insights for those in the football industry.

Focusing on key aspects of football strategy such as passes, shots, fouls, and duels is driven by their critical roles in match outcomes. Passing is fundamental to most strategies, allowing teams to control the game, create opportunities, and disrupt opponents. Analyzing passing patterns can uncover key players, team preferences, and tactical adaptations. Similarly, analyzing shooting patterns is crucial for understanding offensive strategies and scoring efficiency, revealing factors that lead to successful goal-scoring opportunities. Fouls can significantly alter match momentum, influencing both psychological and tactical aspects. Understanding foul patterns can reveal strategic use and their impact on match dynamics. Duels are essential for gaining and maintaining possession, offering insights into individual player performance, physicality, and tactical decisions.

While this research focuses on football, its methods and insights have **broader implications for sports analytics.** The analytical framework and visualization techniques can be adapted to other sports where event sequences and player interactions are critical, such as basketball, rugby, and hockey. By introducing new tools and methods, this project has the potential to enhance decision-making across a range of sports.

Another compelling reason for undertaking this project is the opportunity to **innovate in data visualization.** Effective visualizations are crucial for transforming complex data into actionable insights. This project aims to develop visualization tools that allow users to explore events in football. By ensuring these tools are user-friendly, the research aims to bridge the gap between data science and practical football analysis, making insights accurate and easily understood and applied by practitioners.

1.3 Project Beneficiaries

This research is designed to benefit various sectors within the football community, from those involved in match play and strategy to spectators and commentators. The advanced analysis and visualization techniques developed in this dissertation are intended to serve the following key beneficiaries:

Football Coaches play a crucial role in tactical decision-making and the insights from this research are valuable to them. By understanding event sequence patterns, coaches can refine tactics, tailor training sessions to address specific weaknesses, and make real-life adjustments

during matches. This research helps identify consistent patterns of success and failure, enabling coaches to make informed decisions in the transfer market and align new signings with the team's overall strategy and goals.

Football Players can enhance individual and team performance by understanding the game's context and flow. This research allows the players to analyze how their actions contribute to broader team strategies, and adjust their behavior in response to different in-game situations. For example, studying successful passing sequences can help players anticipate movements better, leading to more efficient positioning and decision-making. Insights into duel dynamics and shooting patterns can also help players refine their techniques and strategies for improved on-field performance.

Sports Scientists focused on optimizing player performance will benefit from the detailed sequential analysis provided by this research. Scientists can understand the physical and mental demands, by examining how events unfold during a game. This can inform the development of more efficient training programs, that improve physical conditioning and tactical awareness. Additionally, identifying high-risk sequences that lead to injuries can prevent them, ensuring teams maintain peak performance throughout the season.

Football Broadcasters are integral to the football ecosystem, and the enhanced visualizations developed in this research can enrich their coverage of the game. The ability to present complex event sequences in a clear and visually appealing manner, enhances analysis and commentary, making it easier to explain key moments and decisions to a wide audience. By providing more sophisticated tools for engaging with football, the research contributes to a richer and more immersive experience for all involved.

1.4 Research Questions

This dissertation is structured around the following research questions, each addressing a specific aspect of football event sequences.

Pass Analysis: How do various passing strategies, including pass types, directions, and network centrality, influence team structure, player involvement, and overall performance?

Play Patterns: How do varying play patterns, event sequences, and player influence areas affect spatial dynamics, coherence, and success of team strategies?

PCA Analysis: How do pass features, interactions, and predictive models impact match outcomes and performance metrics?

Duels Analysis: To what extent do player duel positions, types, frequency, and temporal dynamics contribute to understanding player performances, roles, and team strategies?

Shots Analysis: What insights into offensive strategies, scoring efficiency, and team tactics can be gained from analyzing shooting patterns, and positional data, and using machine learning models to predict goal outcomes?

Fouls Analysis: What insights can be gained into match dynamics by analyzing spatial and temporal foul patterns?

1.5 Objectives and Tests for Success

The objectives of this dissertation are carefully defined to ensure focused and achievable research. Each objective contributes to a comprehensive understanding of event sequences in football, with specific tests established to measure success. These objectives align closely with the research questions and address gaps in football data analytics.

Objective: Collect and organize pass data from a publicly available dataset, ensuring accuracy and relevance for analysis.

Test for Success: Success will be determined by the accuracy and completeness of the extracted data, verified through cross-referencing with source datasets and alignment with known match events.

Objective: Calculate essential pass attributes such as pass length, angle, and direction, to analyze their impact on team structure and performance.

Test for Success: Effectiveness will be evaluated by comparing the derived metrics with existing studies, assessing their correlation with key match outcomes.

Objective: Examine how various contextual factors such as in-game events influence passing networks and team dynamics.

Test for Success: Validity will be tested through statistical methods, ensuring consistently observed effects across matches.

Objective: Develop advanced visualization tools to dynamically represent passing networks over time, allowing interactive exploration of team and player performance.

Test for Success: Success will be measured by the usability of tools, the clarity and informativeness of visualizations, and their ability to uncover insights not apparent through traditional methods.

Objective: Apply clustering and topic modeling to identify patterns in player behavior, team strategies, and event sequences, revealing latent structures within the data.

Test for Success: The effectiveness of clustering will be assessed by evaluating cluster coherence and distinctiveness, using the elbow method and perplexity scores. Clusters will be examined based on their relevance to known tactical groupings or new insights that align with expert knowledge.

Objective: Use PCA and machine learning techniques to explore interactions between passing features, and develop predictive for match outcomes and performance metrics.

Test for Success: The accuracy of predictive models will be validated against match outcomes, with insights enhancing the understanding of how specific features contribute to overall match performance.

Objective: Analyze fouls, duels, and shots to uncover patterns that provide a deeper understanding of match dynamics, player roles, and team strategies.

Test for Success: Success will be measured by aligning insights with known tactical behaviors and their predictive power in real-world football scenarios.

1.6 Methods in Broad Terms

The methodology for this dissertation is structured to address the research questions and objectives through advanced analytical techniques exploring football event sequences. Each step is designed to ensure comprehensive, accurate, and goal-aligned analysis.

Data collection: The study uses publicly available data, focusing on matches played by the Portuguese national team. The data will be precisely filtered and organized to ensure relevance, accuracy, and completeness, emphasizing consistency to facilitate meaningful and reproducible analysis.

Pass and Network Analysis: Passing networks will be constructed using network theory, analyzing team patterns. Centrality metrics, such as degree centrality, will identify key players and their influence within these networks, providing a deeper understanding of individual roles and team strategies.

PCA and Machine Learning: Principal Component Analysis (PCA) will be applied to reduce the dataset's dimensionality, focusing on key pass attributes. This will help identify significant patterns and interactions within the data, making further analysis more manageable and insightful.

Clustering Analysis: Clustering and topic modeling techniques will identify and categorize patterns in player behavior, team strategies, and event sequences. These techniques will reveal latent structures within the data, grouping similar events and distinguishing them from others.

Visualization Tools: Effective communication of findings is crucial. Advanced visualization techniques such as heatmaps, network diagrams, pitch maps, radar charts, bar charts, and line charts, will be developed, dynamically exploring and presenting the data. Visualization tools will be designed considering their usability, utilizing Python and libraries such as Seaborn and Matplotlib to create interactive and intuitive visuals.

1.7 Work Plan

The work plan for this dissertation is structured into five distinct phases, each designed to address the research objectives and ensure timely project completion.

The first phase involves finalizing the research questions guiding data collection and analysis. An initial evaluation of datasets ensures they meet the study's requirements, focusing on event-level data. This phase includes rigorous data cleaning and processing, filtering the dataset into relevant events such as passes, shots, duels, and fouls while handling missing data. Feedback loops will be integrated, allowing continuous reassessment of research questions, data quality, and initial findings. Early visualization efforts including Exploratory Data Analysis (EDA) will begin to understand the dataset's structure and guide subsequent analysis. The second phase starts with an in-depth review of existing literature, to inform the analytical framework and contextualize the research within football analytics. This phase will focus on feature engineering, developing key features from the cleaned data. Passing networks will be constructed using network theory, emphasizing centrality measures, and clustering techniques will be applied to event labels. Ongoing feedback loops will refine methods and features, ensuring alignment with research objectives and responsiveness to new insights. Visualization development will continue, with exploratory visualization testing and refinement of analytical techniques. Regular meetings with the supervisor will ensure progress remains on track and necessary adjustments are made. In the next phase, developed methods and analytical techniques will be applied to produce the final results. These findings will be analyzed in the context of the research questions, focusing on their implications for football

strategy and performance. A supervisor meeting will ensure the analysis aligns with research objectives and facilitates any final adjustments. The final phase involves compiling the findings, discussions, and conclusions into a comprehensive report. This process will emphasize clarity, coherence, and thorough documentation to present the research. A final supervisor meeting will be held, to review the draft, ensuring that all objectives have been met. Before submission, a comprehensive review will ensure that the dissertation meets the highest standards and that all aims have been successfully achieved.

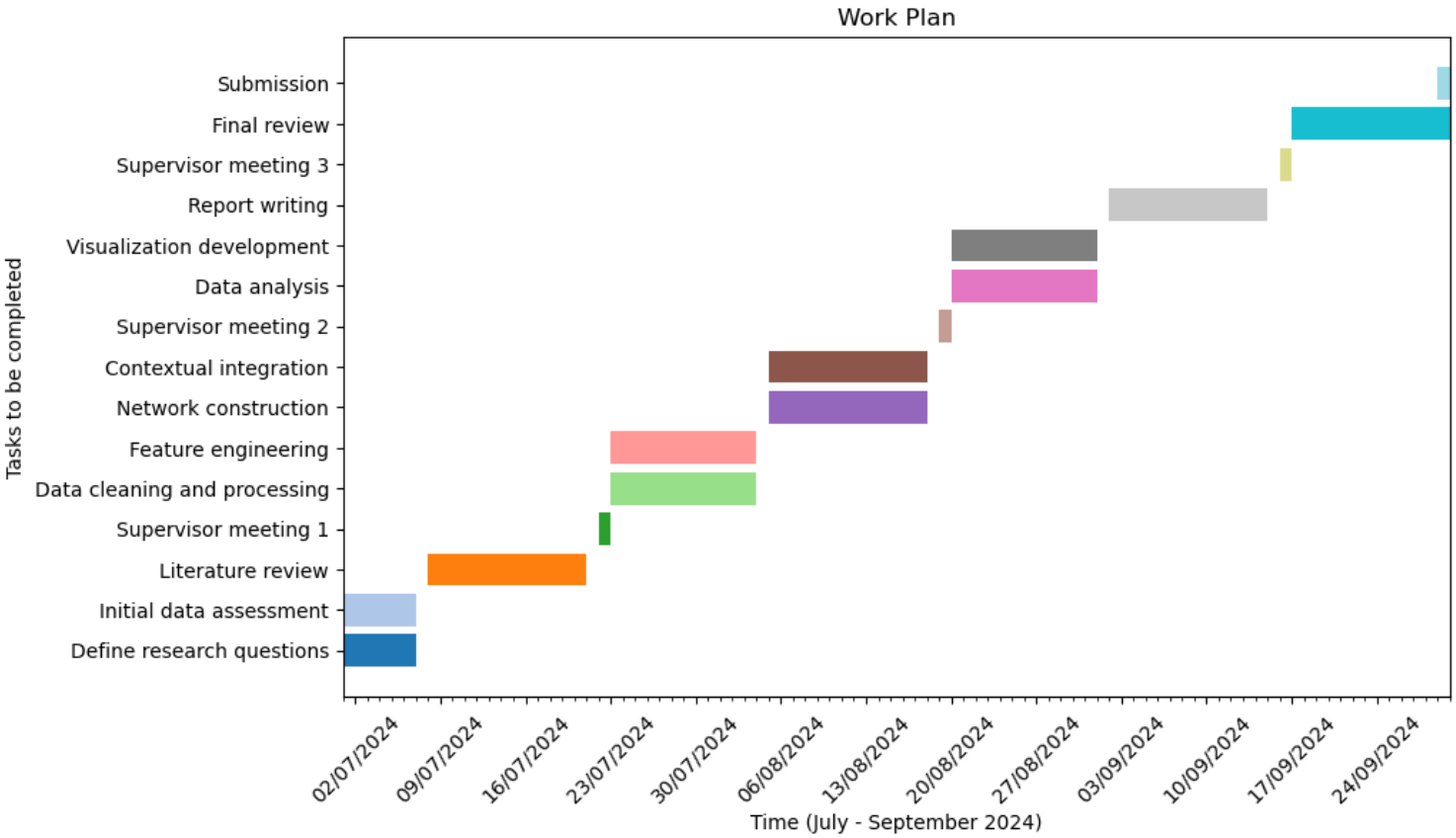


Figure 1.7(a) Work Plan of the Dissertation

1.8 Major Changes in Goals or Methods

Several adjustments were made in this project, to ensure the research remained feasible and aligned with evolving study needs.

Refinement of scope: The initial plan to analyze data from multiple seasons across various matches and teams was revised to focus on a single game for one selected team. This change, driven by time constraints and data complexity, kept the research manageable while providing meaningful insights.

Emphasis on Machine Learning for Predictive Modelling: A range of statistical techniques was planned to analyze event sequences, however, as the project progressed, greater emphasis was placed on machine learning techniques, due to early Exploratory Data Analysis (EDA), indicating that these methods could offer more powerful insights.

Incorporation of Feedback Loops: To accommodate the dynamic nature of the research, feedback loops were integrated into the work plan. These loops allowed for revisited and refined earlier stages as new findings emerged, supporting a more responsive and iterative research process.

Continuous Visualization: Visualization was integrated from the start of data cleaning and processing, continuing through the entire analysis. This ongoing effort ensured that data visualizations evolved alongside the analysis, aiding both exploratory work and the final presentation. This approach ensured that final visualizations were both insightful and commutative.

1.9 Structure of the Dissertation

The dissertation is organized as follows:

Chapter 1 – Introduction and Objectives: Introduces the research problem outlining the study's purpose and significance, presents research questions, and details objectives and methodology.

Chapter 2 – Critical Context: Reviews relevant literature and the theoretical framework in football analytics, identifying gaps this study aims to address.

Chapter 3 – Methods: Describes the data collection, analytical techniques, and visualizations used in the study.

Chapter 4 – Results: Presents the findings, including detailed analysis of passes, shots, duels, and fouls.

Chapter 5 – Discussion: Discusses the implications of the findings, compares results with objectives, and assesses the effectiveness of methods.

Chapter 6 – Evaluation, Reflections, and Conclusions: Evaluates the overall project, summarizes key conclusions, considers future work, and reflects on lessons learned and potential improvements.

Chapter 2 – Context

2.1 Introduction

Football analytics has evolved over the past two decades from statistical metrics to advanced methodologies that delve into the game's complex dynamics. Traditional metrics like possession percentages, shots on goal, and pass completion rates, provide a basic understanding of match performance but fail to capture the nuanced sequences and interactions pivotal in determining match outcomes. Football analytics has incorporated techniques such as clustering, topic modeling, Principal Component Analysis (PCA), network theory (including degree centrality), machine learning models, and sport-specific visualization techniques to overcome these limitations. This chapter reviews these methodologies, which are central to this dissertation. It examines their theoretical foundations and applications, identifies gaps in the literature, and outlines how this study contributes to the field.

2.2 State of the Art in Football Analytics

Assessing Tactical Patterns: Possession percentages traditionally indicate game control, with higher rates suggesting dominance. However, they fail to differentiate between effective attacking play and passive ball retention. Similarly, Scarf et al. identified that shots on goal do not account for shot quality, and pass completion rates may not reflect the impact of passes on advancing play or creating scoring opportunities. (Scarf, Khare and Alotaibi, 2021). Hughes & Franks challenge the traditional focus on short-passing sequences, historically linked to effective goal-scoring, emphasizing the significance of evaluating passing sequences for their strategic context and execution quality. (Hughes and Franks, 2005). Clemente et al applied network theory to map player interactions on the field, representing players as nodes and passes as edges, with edge thickness indicating pass frequency. (Clemente *et al.*, 2014). This method helps identify key players and influential passing patterns, showing that midfielders often play crucial roles in maintaining possession and linking defense with attack. Successful teams display more connected and balanced networks, involving multiple players rather than a few individuals. Additionally, metrics like team width and height, team compactness, and packing rate have gained relevance in tactical analysis, width and height measure spatial distribution, indicating offensive or defensive orientations. Frencken et al. showed that variations in inter-team distances can signal tactical changes, highlighting the role of spatial organization. (Frencken *et al.*, 2012). Pollard discussed that compact teams maintain tight formations, enhancing defensive solidity and effectiveness in pressing. (Pollard, 1986)The Packing Rate measures the number of opponents bypassed during a pass, revealing the strategic value of forward passes in penetrating defenses. (Araújo *et al.*, 2019). Advanced methodologies now capture the game's complex dynamics with clustering techniques categorizing player roles and team tactics. Bialkowski et al. introduced a dynamic role-based player representation, contextually analyzing player behavior and team formations, uncovering distinct play styles and adaptive strategies. (Bialkowski *et al.*, 2014). Visualization tools like pitch maps and heatmaps, are essential in translating spatiotemporal data into actionable insights. Pappalardo et al. stress the importance of these tools in understanding player movements, positioning, and overall team strategies, aiding analysts and coaches in interpreting and communicating tactical insights effectively. (Pappalardo *et al.*, 2019).

Predicting Game Results: Traditional metrics like possession percentages, shots on goal, and pass completion rates have long been key in football analysis, providing straightforward team performance measures. However, high shot numbers might reflect many low-percentage attempts, and possession statistics don't distinguish between effective attacking play and passive ball retention. Lago-Penas & Dellal explored the limitations of traditional possession metrics by examining how match status influences ball possession strategies in Spanish La Liga. They found that successful teams, such as FC

Barcelona maintained more stable possession patterns and exhibited a smaller coefficient of variation in possession time, suggesting the quality and consistency of possessions are more crucial than quantity. (Lago-Peñas and Dellal, 2010). For instance, teams often increase their possession when losing to create more goal-scoring opportunities, highlighting the need to contextualize possessions within the match scenario. Hughes & Franks analyzed passing sequences in the FIFA World Cups of 1990 and 1994, demonstrating that longer passing sequences could lead to goals per possession, particularly for successful teams. (Hughes and Franks, 2005). This suggests a shift from viewing possession as a static indicator to considering its strategic use relative to the match contexts. Advanced analytical methods like the Expected Goals model, have been developed to provide deeper insights. This model evaluates the quality of scoring opportunities, assigning a probability to each shot based on factors like distance, angle, and type of assist, offering a nuanced understanding of offensive efficiency, beyond simple shot counts. Despite their advantages, metrics like expected goals and passing networks have limitations. PCA has been applied to simplify the analysis of large and complex datasets. Lepschy et al. discuss how PCA can reduce dimensionality, focusing on the most significant variables contributing to a team's performance and facilitating a more straightforward interpretation of data. (Lepschy, Wäsche and Woll, 2018). Machine learning models are increasingly used to predict match outcomes and player performances. Bunker & Thabtah outline how these models can forecast results using historical data, demonstrating their versatility and accuracy in identifying patterns and trends difficult to detect with traditional methods. (Bunker and Thabtah, 2017). These emerging trends in football analytics reflect a broader shift towards more holistic analyses that consider spatial and temporal aspects of the game.

2.3 Theoretical Foundations of Event Sequence Analysis

Event Sequence Theory: Event sequence analysis is used in various disciplines to study patterns and processes over time. It involves analyzing the sequence and context of actions such as passes, dribbles, and tackles, leading to significant outcomes like goals and defensive stops. The core premise is that the order and interplay of these events provide deep insights into team strategies, decision-making processes, and game flow. A pivotal study by Link et al. introduced the concept of “dangerousity”, a metric that quantifies the potential threat posed by a sequence of actions during a match, based on player and ball positions. (Link, Lang and Seidenschwarz, 2016). Leveraging spatiotemporal tracking data, the study highlights how the continuous flow of play, rather than isolated events, generates goal-scoring opportunities. It shows that sequences like well-timed passes and effective player positioning can enhance the likelihood of scoring, emphasizing football's dynamic nature. Their analysis stresses the importance of considering spatial and temporal dimensions in football analytics. Link et al. examined how events unfold concerning one another, deepening our understanding of the mechanism behind successful plays. (Link, Lang and Seidenschwarz, 2016). This theoretical foundation is crucial to the methodologies used in this research viewing football as a dynamic and interconnected sequence of actions. This dissertation aims to uncover how these sequences, analyzed in context, contribute to team performance and match outcomes.

Analytical Methods: In football analytics, advanced methods like clustering, topic modeling, PCA, and network theory are key to uncovering deeper insights into game dynamics and team strategies. Clustering helps group similar events, allowing analysts to identify common gameplay patterns. For instance, Stefano et al. used decision tree classifiers with clustering to predict football match outcomes, revealing tactical patterns and strategic behaviors not evident from raw data. (Stefano *et al.*, 2020). This approach highlights team behaviors, pinpointing formations and playing styles that consistently lead to success. Topic modeling traditionally used in text analysis, has been adapted to uncover underlying patterns in football match data. Topic modeling adapted from text analysis uncovers underlying patterns in football data. Decroos et al. developed a technique to assess a player’s actions by identifying recurring sequences

that significantly enhance team performance. This method elucidates strategic themes underlying successful plays, providing insights into the tactical decisions of players and coaches. PCA simplifies data complexity by isolating key variables. (Decroos *et al.*, 2018). Moura et al. applied PCA to positional data, facilitating the analysis of team dynamics and clarifying how various factors influence match outcomes. (Moura *et al.*, 2013). Network theory, particularly degree centrality, is crucial in analyzing passing networks within teams. Clemente et al. used network metrics to evaluate a player's role in a team's passing structure. By assessing player centrality in these networks, the study identifies key players essential to a team's strategy, understanding how ball movement and player positioning drive team success. (Clemente *et al.*, 2014).

Modeling techniques: Machine learning models are increasingly essential in sports analysis, providing advanced methods to predict match outcomes and analyze complex event sequences. For instance, Routley & Schulte introduced a Markov game model to evaluate player actions in ice hockey, focusing on the context and long-term impact of actions. They employed Q-learning to quantify how player actions influence game outcomes, demonstrating how the sequence and timing of events can critically affect results. (Routley and Schulte, 2015). This model highlights the importance of including temporal and contextual factors in assessing a player's performance, a concept also crucial in football analytics. Similarly, Liu et al. developed a machine-learning framework that uses temporal relational models to analyze player performance in team sports. This study models player interactions over time, providing insights into how these interactions evolve during a match and affect overall team success. (Liu and Schulte, 2018). By integrating machine learning with analytical methods such as clustering and PCA, these models offer a more comprehensive analysis of event sequences, yielding deeper insights into player performance and team strategies.

Episode Segmentation Approach: In this study, segmenting football match events is critical for analyzing game dynamics. The approach, which involves segmenting based on possession changes, stoppages, and significant events is supported by various methodologies highlighted in the literature. Narayanan et al. align with the research's focus on team-specific possession tracking and the dynamic nature of events. Their emphasis on considering the spatiotemporal interactions during game phases supports segmenting episodes based on key possession and stoppage events, ensuring that only meaningful challenges are captured. (Narayanan, Kosmidis and Dellaportas, 2021). The work by Bialkowski et al. closely aligns with this dissertation's segmentation logic, where the focus is on identifying key moments impacting the game flow. Their study emphasizes detecting significant events to capture shifts within a match, correlating directly with how this research handles possession changes and key events. (Bialkowski *et al.*, 2014). This method is corroborated by researching event detection and classification, such as the study by Wei et al. It shows the importance of tracking possession and significant events to extract meaningful insights from match data. (Wei *et al.*, 2013). This aligns with the research's approach, considering significant events and stoppages as the start of new episodes. Including a ping-pong, detection mechanism enhances the accuracy of the segmentation by ensuring that only substantive shifts in possession are included. This concept is reinforced by the approaches discussed in the literature, particularly in how these studies emphasize the importance of considering the significance of each event. Andrienko et al. incorporate tolerance episodes, addressing potential data imperfections such as the imbalance between manually annotated events and automatically recorded positions. Their approach to dynamically aggregating data and identifying episodes based on specific events, such as possession transitions, enhances the reliability of episode segmentation and event detection. (Andrienko *et al.*, 2021). "Stats perform" described in their possession framework, a conceptual alignment with this study's segmentation approach, particularly in how it structures sequences and possessions around key game events. "Stats performs" recognition that not all events belong to a sequence or possession, echoes

the necessity of a ping-pong detection mechanism, which prevents rapid and trivial possession changes from creating new episodes. (*Introducing a Possessions Framework*, no date). Vidal-Codina et al. share similarities with this segmentation method, particularly in rule-based algorithms for event detection and emphasis on accurate classification of passes and shots. (Vidal-Codina *et al.*, 2022).Their attention to validating event detection methods across multiple datasets resonates with the need for robust episode segmentation that accounts for synchronization issues, a concern also addressed in this study through a tolerance adjustment mechanism.

2.4 Comparative Analysis of Existing Football Event Sequence Models

Key Studies and Models: Several significant studies have advanced event-sequence analysis in football, each contributing unique methodologies illuminating the game's tactical dimensions. Kempe et al. developed indices like the Index of Offensive Behavior (IOB) and Index of Game Control (IGC) to assess tactical behaviors in elite football. These indices combine multiple performance indicators to distinguish between possession-based and direct play strategies, showing that successful teams often display superior game control and offensive behaviors. (Kempe *et al.*, 2014). This study highlights the importance of integrating multiple indicators to capture the complexity of tactical behaviors, moving beyond isolated metrics. Coutinho et al. used clustering techniques to analyze ball possession, emphasizing the role of temporal data in understanding tactical dynamics. (Coutinho *et al.*, 2022). Their research offers a nuanced view of how different ball possession sequences affect team strategies and outcomes, enhancing our understanding of possession as a tactical element. Lang et al. employed machine-learning approaches to predict in-game status, examining how different situational variables impact match outcomes. Their study combines K-means clustering and predictive models to categorize match situations and predict future events based on historical data, underscoring the potential of machine learning to enhance tactical decision-making. (Lang *et al.*, 2022). These studies provide a robust foundation for understanding football tactics through event-sequence analysis. Each study focuses on specific game aspects, like ball possession and predictive modeling, however, they don't fully integrate these approaches into a holistic framework. This dissertation aims to bridge this gap by combining these methodologies, offering a comprehensive analysis, and capturing the dynamically interconnected nature of football events.

Comparative Review: Critical factors such as data requirements, computational complexity, and interpretability stand out when comparing methodologies in football analytics. Qu et al. demonstrated the effectiveness of Principal Component Analysis (PCA) in reducing the dimensionality of large datasets, simplifying the analysis of complex-multidimensional football data. However, they caution that while PCA effectively distills data to its most significant components, it may obscure crucial nuances, particularly interaction variables critical to tactical decisions. (Qu *et al.*, 2002). Conversely, Inan explores the application of statistical methods to address the home-field advantage in European football, emphasizing the importance of context-specific factors like crowd support and density, which significantly influence match outcomes. (Inan, 2020). While insightful for understanding socio-psychological aspects of football, this approach demands extensive and detailed data collection, which can be resource-intensive and not always feasible. This comparison reveals the strengths and limitations of each methodology, PCA is valuable for data reduction but may oversimplify relationships between variables. In contrast, statistical models provide insights into aspects like home-field advantage, requiring more data and greater computational effort. This research aims to investigate the strengths of PCA, statistical models, and other approaches like machine learning to develop a comprehensive and nuanced model for analyzing event sequences in football.

2.5 Legal, Ethical, and Societal Context

Data Privacy: The increasing use of detailed player data in football analytics raises significant concerns about data privacy, especially under regulations like the General Data Protection Regulation (GDPR) in Europe. Tataru & Tataru analyze GDPR's impact on the sports industry, stressing the necessity for strict compliance with data protection laws when handling sensitive athlete information, such as health and biometric data. They highlight the need for the sports industry to implement robust data protection measures, including anonymization techniques and obtaining explicit consent for data use, particularly when the data could impact the athlete's human rights. (Tataru and Tataru, 2020). In response, this research will adhere to stringent privacy protections to maintain research integrity and build trust in data analytics within football and other sports.

Ethical Considerations: Ethical considerations are essential in football analytics, extending beyond legal obligations. Vermeulen & Sarma identify several ethical challenges associated with using big data in sports, including concerns over data validity, security, and athletes' autonomy. (Vermeulen and Sarma, 2018). Predictive models in scouting and player transfers can significantly influence a player's career, raising fairness issues, transparency, and algorithmic bias. For instance, models based on historical data may unintentionally reinforce existing biases, leading to unfair player assessments. This research will incorporate rigorous checks to prevent perpetuating or introducing biases, through implementing fair algorithms and scrutiny of data sources for potential biases. Additionally, integrating analytics into coaching decisions should balance data-driven insights with the sport's human elements. Vermeulen & Sarma stress that while data offers valuable insights, it should not override human judgment or the enjoyment of the sport, ensuring athletes remain central to decision-making processes. (Vermeulen and Sarma, 2018). This research aims to promote a more transparent application of analytics in football by addressing these ethical considerations.

2.6 Knowledge Gaps and Challenges

Gaps in Literature: Despite advancements in applying clustering, PCA, network theory, and machine learning to football analytics, significant gaps remain. Borrie et al. point out that while clustering is widely used to categorize player roles and tactics, research on how these clusters interact dynamically during matches is lacking. They advocate for more advanced methods to capture the temporal sequences and complex interactions that characterize football. (Borrie, Jonsson and Magnusson, 2002). Similarly, Mchale et al. note that PCA effectively reduces data complexity, but often neglects the sequential and contextual nature of football events, leading to oversimplifications. (Mchale, Scarf and Folker, 2012). Duch et al. highlights the importance of using degree centrality within network theory to quantify individual player performance and influence within the team. (Duch, Waitzman and Amaral, 2010). However, they recognize that this approach might not fully account for the evolving match context, such as changes in player positioning or tactical adjustments over time. Robertson et al. emphasize the need for comprehensive models that integrate multiple analytical methods, such as combining network theory with machine learning, to grasp the complex interplay of factors that determine match outcomes. (Robertson, Back and Bartlett, 2015). This dissertation aims to bridge these gaps by developing a framework that uses these methods in isolation and integrates them to provide a more holistic analysis of football event sequences. It seeks to enhance the understanding of dynamic interactions on the field and deliver more accurate and actionable insights into player and team performance.

Practical Challenges: Implementing advanced analytics in football comes with several challenges. One major issue is data quality. Fathima et al. stated that football data can often be incomplete or inconsistent, especially when collected from different sources or under varying conditions. (Fathima S J, Sumathi, and

Sumanth, 2018). These inconsistencies pose significant challenges to accurately applying analytical methods like clustering and PCA, which rely on high-quality data to produce reliable results. Additionally, the computational complexity of models such as machine learning algorithms can hinder their widespread use in football analytics. Mackenzie & Cushion discuss how these models require substantial computational resources to process and analyze the large datasets that modern football data analytics demands. (Mackenzie and Cushion, 2012). Furthermore, there is often resistance within the sport to adopting data-driven approaches. Coaches and players may be skeptical of the value of analytics, preferring to rely on their experience and intuition. This dissertation addresses these challenges by ensuring rigorous data preprocessing and validation, employing computationally efficient methods, and carefully considering the practical implications of integrating analytics into real-world football contexts.

Chapter 3 – Methods

3.1 Introduction

This chapter outlines the methods used to analyze football match data, addressing the research questions from Chapter 1. It covers the process from data collection and preprocessing to applying sophisticated analytical techniques such as episode segmentation, network analysis, clustering, Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), topic modeling, and machine learning.

Football's continuous flow and dynamic interactions require a systematic approach beyond traditional metrics. This research employs a multi-faceted strategy using recent data science advances to examine the game at a granular level, uncovering patterns and relationships that impact team performance, player roles, and match outcomes.

The analysis focuses on Portugal's opening group-stage match against Iceland, part of the UEFA European Championship (commonly known as the Euros). A key step was segmenting matches into distinct episodes, breaking the game's flow into analytically meaningful units representing specific play phases. By identifying pivotal events like possession changes, duels, passes, fouls, and other stoppages, the game was divided into approximately 50 episodes, facilitating a detailed examination of how different play phases affect team strategy and performance. These episodes form the basis for further analysis, including passing networks, clustering player behaviors, and machine learning models to predict match outcomes.

The data underwent preprocessing to ensure suitability for advanced analysis. This involved cleaning, feature extraction, and recalculating attributes like episode duration and event frequency, essential for techniques like network analysis, which explores player relationships and centrality within passing networks.

This chapter will detail the specific analytical techniques used, including constructing and analyzing passing networks to understand player centrality and team performance, clustering and topic modeling to categorize player behaviors and tactics, PCA for data dimensionality reduction, and machine learning for predicting match outcomes and analyzing passing features. The development of visualization tools will also be discussed, highlighting their role in presenting findings clearly and accessible.

3.2 Data Collection

This section outlines the data collection and preprocessing steps for analyzing football match data in this study. The research utilizes the publicly available 'Soccer match event dataset' on Fig Share, one of the largest collections of football data, providing detailed records of significant events from matches across various European leagues and International competitions. (Pappalardo *et al.*, 2019).

Data Sources: The primary dataset is the 'Soccer match event dataset', compiled and released by researchers from the University of Pisa, Italy. It includes all spatiotemporal events from seven major football competitions, including the FIFA World Cup 2018, and UEFA Euro 2016. With over 3 million time-stamped events, the dataset offers comprehensive information on event locations, players involved, and outcomes. (Pappalardo *et al.*, 2019).

3.3 Data Preprocessing

Given the complexity and size of the dataset, several preprocessing steps were necessary to prepare the data for analysis, including data cleaning, feature extraction, normalization, segmentation of the game into meaningful episodes, and Exploratory Data Analysis (EDA).

Data Cleaning: While the data was largely complete, some missing data, particularly in player tracking coordinates and event details, required attention. Missing values were imputed using techniques appropriate to the data type. Considering time series data, such as player positions, linear interpolation ensured continuity. Outliers, identified using z-scores and inter-quartile ranges (IQR), were scrutinized and corrected or excluded based on contextual verification to avoid skewing results. Normalization and transformation steps, such as min-max scaling for event attributes like pass length and logarithmic transformations for player velocity, were applied to enhance comparability across features and improve interpretability in PCA and machine learning.

Feature Extraction: Key features such as pass type, pass length, pass direction, and the positions of the passer and the receiver, were extracted. These features were crucial for constructing passing networks and analyzing player centrality and team performance. From player tracking data, additional metrics such as average position, movement patterns, and distances covered were derived to highlight player roles and areas of influence. Furthermore, sequences of events leading to key outcomes, such as goals, were meticulously extracted to analyze the order and timing of passes and other actions, providing insights into the flow of play and team strategies.

3.4 Data Segmentation

Segmentation of match data into episodes was an essential preprocessing step to breaking down the continuous flow of the game into distinct, analyzable segments, each representing a unique phase of play.

Event-Based Segmentation: Segmentation was designed to transform the continuous flow of play into distinct episodes, providing detailed insights into match dynamics. Significant events were categorized into three primary groups: possession-related events, stoppages, and key events. Possession-related events, such as duels, interceptions, clearances, tackles, save attempts, shots, and head passes, were pivotal in determining the game's rhythm and were key markers for segmentation. Stoppages, including fouls, offsides, goal kicks, throw-ins, free kicks, and penalties, reset the game's momentum and transitioned play between episodes. Severe stoppages like violent, and late card fouls, were also considered. Significant events that impacted match outcomes like goals, penalties, free kicks, and corners, served as important reference points for segmentation.

Episode Segmentation Process: The goal was to divide each match into approximately 160 episodes, corresponding to about 80-90 possessions per team, for detailed analysis while maintaining consistency in the number of segments. Matches were organized by match ID and processed individually, with events listed chronologically. Episodes were formed by sequentially grouping events until a predefined event limit was reached, adjusted based on total events per match. Key events like goals and fouls were contained within a single segment for coherence. Any remaining events were included in the last segment, ensuring no data loss and a comprehensive representation of each match phase.

Recalculation of Episode Attributes: After segmentation, key attributes like episode duration and frequency were recalculated to reflect each segment's characteristics. Episode duration was determined by the difference between the timestamps of the first and last event, offering insights into the temporal dynamics of each play phase. Event frequencies within each episode were recounted and stored as episode attributes. These recalculations were crucial for understanding the nature of each episode, particularly

when analyzing the intensity and flow of play. The data was standardized using a Standard Scaler to ensure equal contribution of attributes like event frequency and episode duration during clustering.

Clustering of Episodes: The final step involved applying K-means clustering to group similar episodes. The standardized attributes of each episode were used as input features for the K-means algorithm, which partitioned the episodes into distinct clusters based on similarity. Each episode received a cluster label indicating the type of phase it represented. These labels were integrated back into the dataset, facilitating the analysis of patterns within and across clusters. Clustering was essential for identifying and analyzing tactical patterns and strategies employed during match phases. Grouping episodes with similar characteristics allowed a focused analysis of the tactics and strategies that shape match dynamics.

3.5 Methodologies for Answering Research Questions

This section outlines the methodologies employed to address the research questions posed in this study, utilizing Python. Building on the data preprocessing and segmentation process, the focus is on analytical techniques tailored to each research question, leveraging network analysis, spatiotemporal modeling, and machine learning to extract insights from the dynamics of football matches.

Research Question: How do various passing strategies, including pass types, directions, and network centrality, influence team structure, player involvement, and overall performance?

Rationale and Justification: Network analysis and clustering were chosen to explore intricate relationships and patterns in player interactions that basic statistical methods might overlook. These techniques visualize and quantify key players' roles within the network, influencing team dynamics. This method offers deeper relational insight compared to standard statistical analysis, which is less effective in understanding complex team dynamics. PCA was used to handle high dimensionality, and it was chosen over simpler techniques like factor analysis because it preserves more variance without assuming underlying distributions.

Data Cleaning and Preprocessing: Passes with invalid coordinates (0,0), likely from data collection errors, were removed using a custom function. Only valid passes were retained. Selected features for clustering, including spatial and temporal metrics, were standardized using a Standard Scaler to achieve equal weighting during clustering, as K-means rely on Euclidean distances. This ensured that no single feature dominated the analysis. Missing values were filled with 0s, enabling smooth processing of the dataset.

Feature Selection for Clustering: A comprehensive set of features captured various aspects of passes, including the timestamp, the distance covered, vertical and horizontal displacements, pass angle, forward movement, spread of play, progression down the pitch, and ball movement speed. These features provided a multi-dimensional view of each pass, which is crucial for analyzing structure and dynamics.

Clustering to Identify Pass Patterns: K-means clustering grouped similar patterns. Using Figure 3.5(1a), the Elbow Method, plotting the Sum of Squared Errors (SSE) against various k values, identified 4 clusters as optimal, balancing SSE minimization and interpretability. Silhouette analysis (Figure 3.5(1c)) validated clustering results and confirmed well-defined clusters for long passes and wide play switches. Considering passes across the pitch, K-means assigned each pass to a cluster, and medoids represented typical behaviors within each cluster, offering a clear depiction of tactical patterns. A summary of the clustering results can be seen in Figure 3.5(1a).

Cluster	Pass Type	Description	Tactical Insight
Cluster 1	Long Passes	Mainly long passes from defense to midfield	Quick transitions aimed at bypassing midfield opposition
Cluster 2	Short Lateral Passes	Short lateral passes in the midfield	Possession maintenance and slow buildup
Cluster 3	Forward Passes	Direct forward passes into attacking zones	High-risk, high-reward strategy for penetrating defenses

Figure 3.5(1a) Summary of Clustering Results for Passing Patterns

Visualizations and Summary of Clusters: Medoid passes were plotted to represent typical passes within each cluster, distinguishing long, short, lateral, or forward passes. Statistical summaries provided average pass lengths, enabling quantitative comparisons. Spatial distributions of passes within clusters were also visualized, emphasizing tactical importance.

Network Centrality and Player Involvement: Players' average pitch positions were calculated to visualize positioning and involvement in passing sequences. A passing network was constructed, with nodes representing players and edges representing passes. Edge thickness indicated pass frequency, highlighting key connections. Degree centrality quantified each player's involvement, indicating their importance to the team's structure. A network graph visually emphasized players with the highest centrality.

Pass Direction and Type Analysis: Passes were classified forward, lateral, or backward, based on their angle, providing insights into ball progression and possession maintenance. Bar plots displayed the distribution and accuracy of each pass type, offering insights into the effectiveness of various strategies. Cross-pitch passes were analyzed for their tactical significance in shifting play direction and disrupting defensive structures.

Temporal Patterns in Passing: The game was divided into minutes analyzing temporal passing patterns, and comparing average pass lengths between the first and second halves to observe strategic shifts or fatigue effects. A line graph visualized the average pass length for each minute. A summary table of temporal passing patterns can be seen in Figure 3.5(1b).

Time Period	Average Pass Length	Pass Type Frequency	Tactical Adjustments Noted
First Half	25 meters	60% short, 30% long	Controlled possession, focus on maintaining midfield control
Second Half	30 meters	40% short, 50% long	More direct play, increased risk to break through opposition

Figure 3.5(1b) Summary of Temporal Passing Patterns and Tactical Adjustments

Heatmaps and Performance Metrics: Hexbin heatmaps provided a spatial understanding of player involvement, showing pass density across pitch areas and highlighting zone influence. A bar plot summarized key performance metrics such as pass length, vertical change, horizontal spread, and speed of play, quantifying team performance across different clusters and passing strategies.

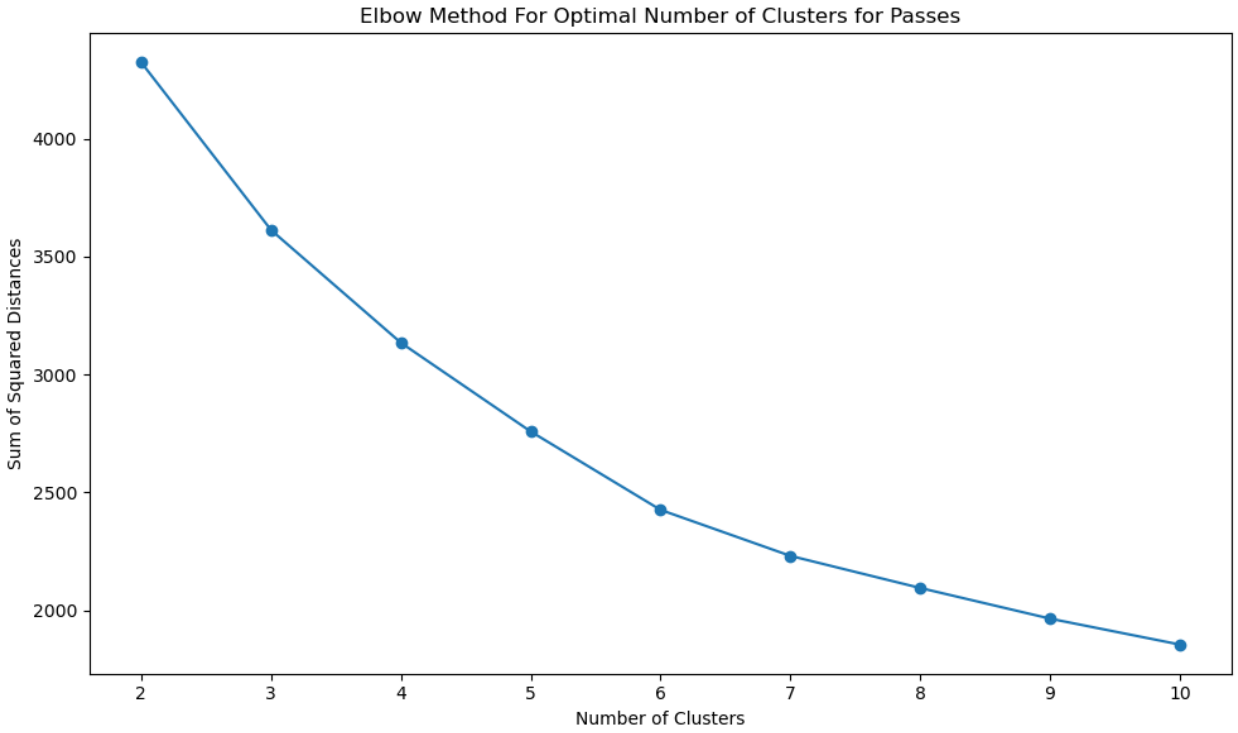


Figure 3.5(1c) Elbow Curve Showing the Optimal Number of Clusters of Passes

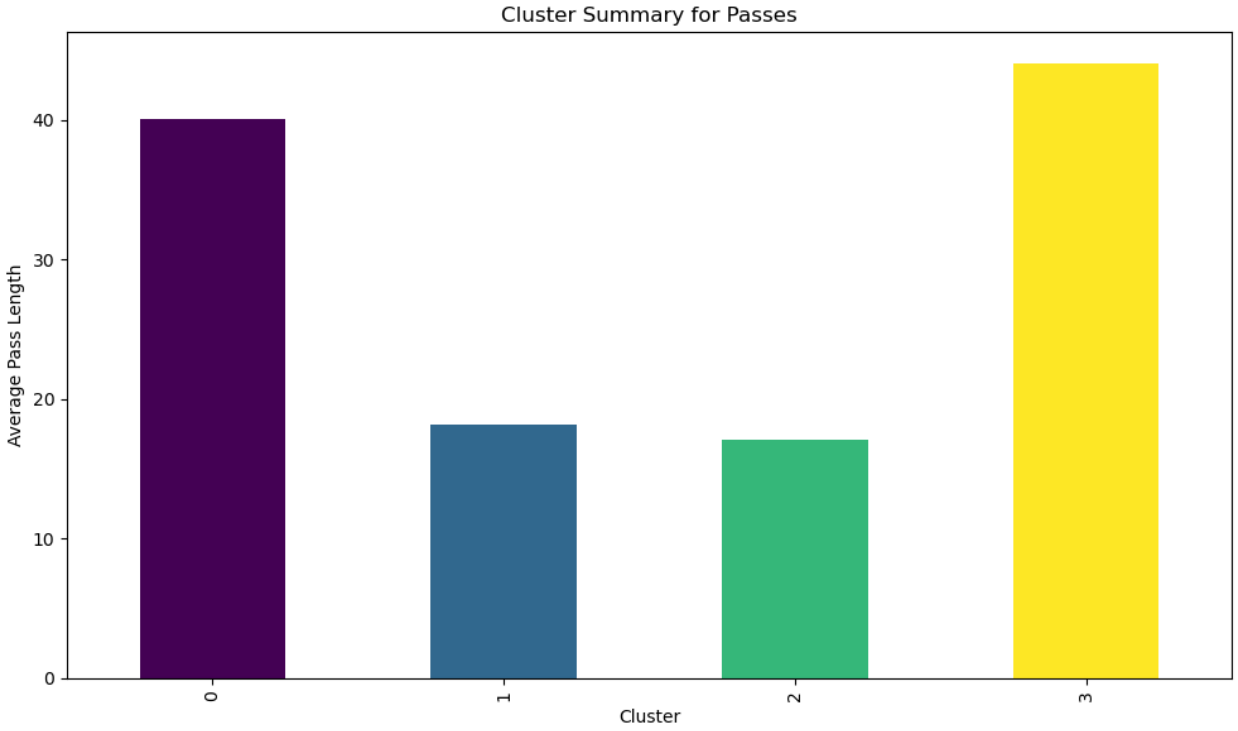


Figure 3.5(1d) Bar Plot Showing a Cluster Summary of Passes

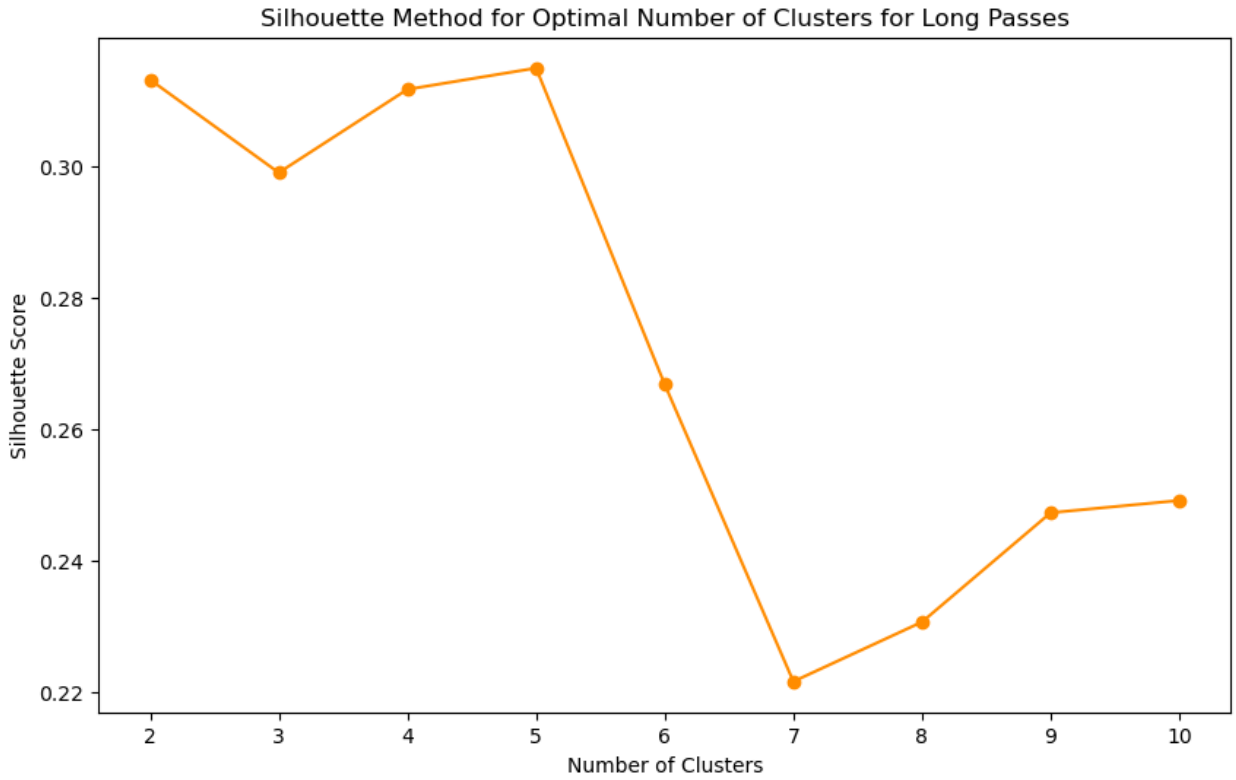


Figure 3.5(1e) Silhouette Plot Showing Optimal Number of Clusters of Long Passes

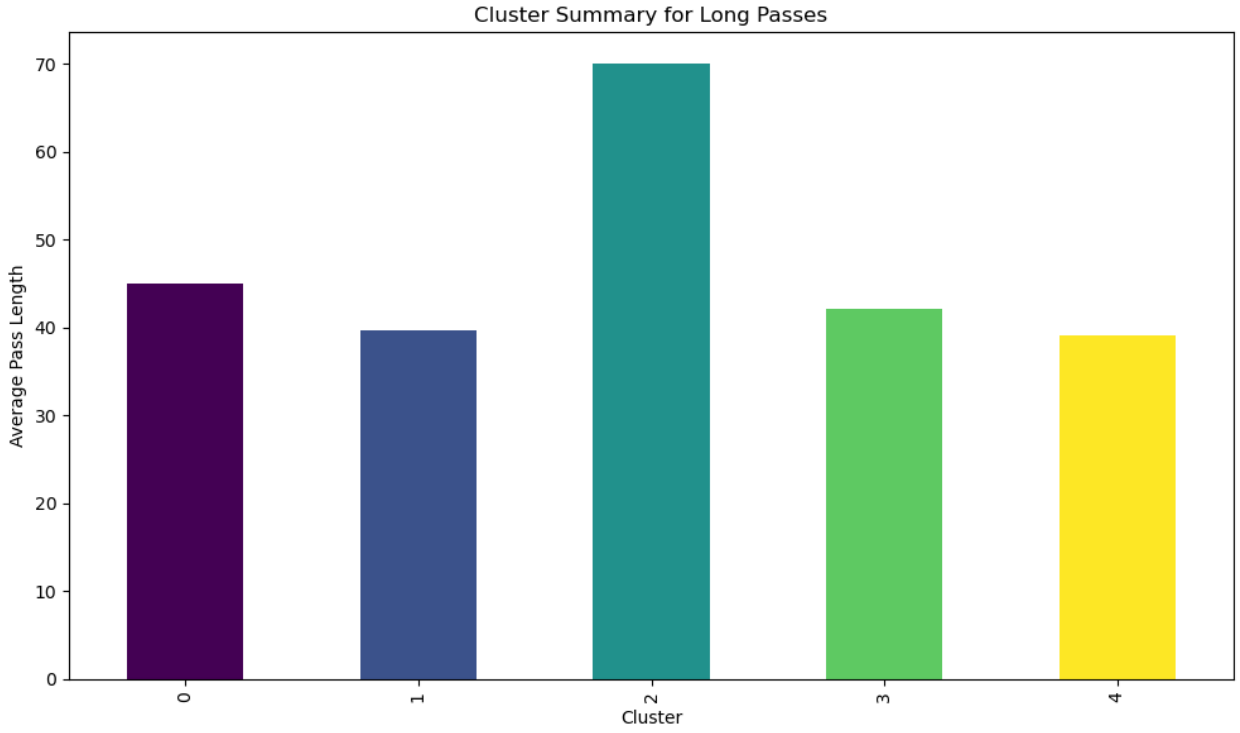


Figure 3.5(1f) Bar Plot Showing a Cluster Summary of Long Passes

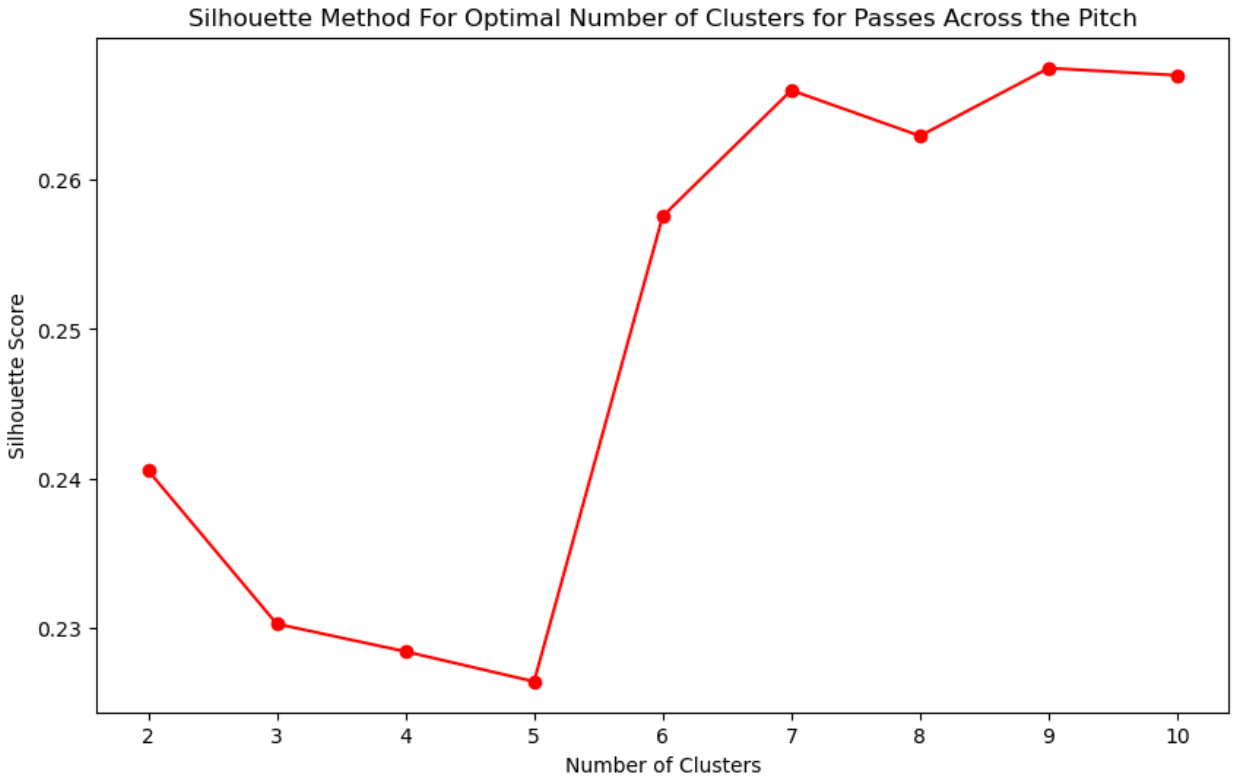


Figure 3.5(1f) Silhouette Plot Showing Optimal Number of Clusters of Passes Across the Pitch

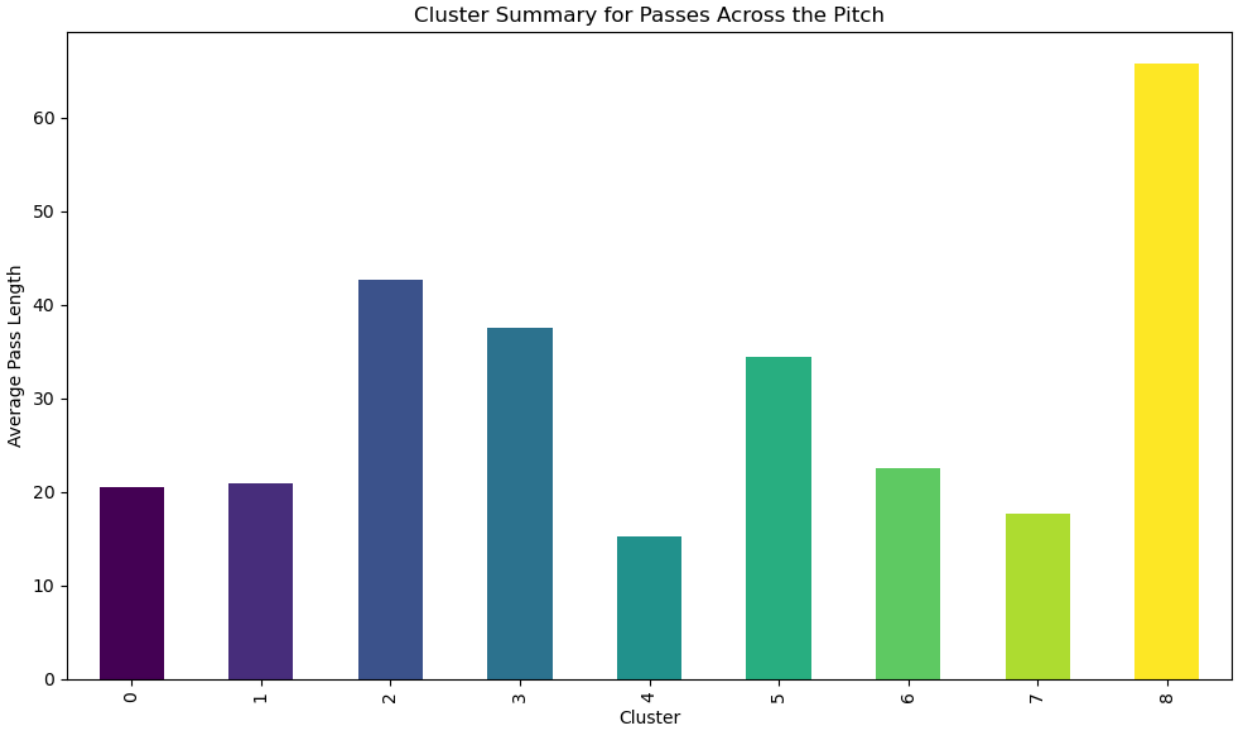


Figure 3.5(1g) Bar Plot Showing a Cluster Summary of Passes Across the Pitch

Research Question: How do varying play patterns, event sequences, and player influence areas affect the spatial dynamics, coherence, and success of team strategies in football?

Rationale and Justification: The combination of Latent Dirichlet Allocation (LDA) and transition matrices offers a structured approach to deciphering complex play patterns and sequences which is crucial for understanding strategic executions in football. This method is preferred over direct sequence analysis without topic modeling because LDA provides a nuanced understanding of underlying themes across multiple games, and transition matrices depict event flow probabilities, emphasizing connections between events over simple chronological analysis.

Data Preprocessing for Topic Modeling: Events were transformed into strings by combining relevant features such as pass length, vertical advancement, horizontal advancement, and horizontal spread. A Count Vectorizer converted these sequences into a document-term matrix (DTM), enabling the LDA algorithm to uncover latent topics in the event data. LDA identified distinct tactical approaches like building from the back, fast counter-attacks, and wide plays. As seen from Figure 3.5(2a) perplexity and coherence scores ensured the optimal number of topics, where 3 topics were identified. Each was analyzed to determine its success rate, evaluating whether episodes dominated by a particular topic led to successful outcomes, such as advancing the ball into dangerous areas or creating shots.

Event Sequences (Transition Matrix): A transition matrix captures the probabilities of transitioning from one event to another, such as how frequently short passes lead dribbles or tackles to clearances. This matrix, visualized with a heatmap, illustrated event transition probabilities, offering insights into how different strategies unfold during a match. It provided a clear picture of tactical coherence and fluidity in team play.

Player Influence Areas and Spatial Dynamics: Voronoi diagrams were used to visualize players' influence areas on the pitch, revealing the spatial organization of the team and variations in play patterns. These diagrams highlighted spatial dominance and player distribution, demonstrating shifts in player roles and their connection to tactical strategy. A passing network was also constructed, with nodes representing players and edges representing passes. Edge thickness reflected pass frequency, while degree centrality quantified each player's importance to team buildup and execution. This network provided insights into how teams maintain possession and create scoring opportunities.

Temporal Dynamics and Topic Usage Over Time: The analysis tracked the evolution of team strategies by examining the frequency of play patterns throughout the match. This highlighted tactical shifts and Portugal's adaptability regarding the game state, opposition, or adjustments. This temporal analysis sheds light on the timing of strategic changes and the decisions made by coaches or team leaders.

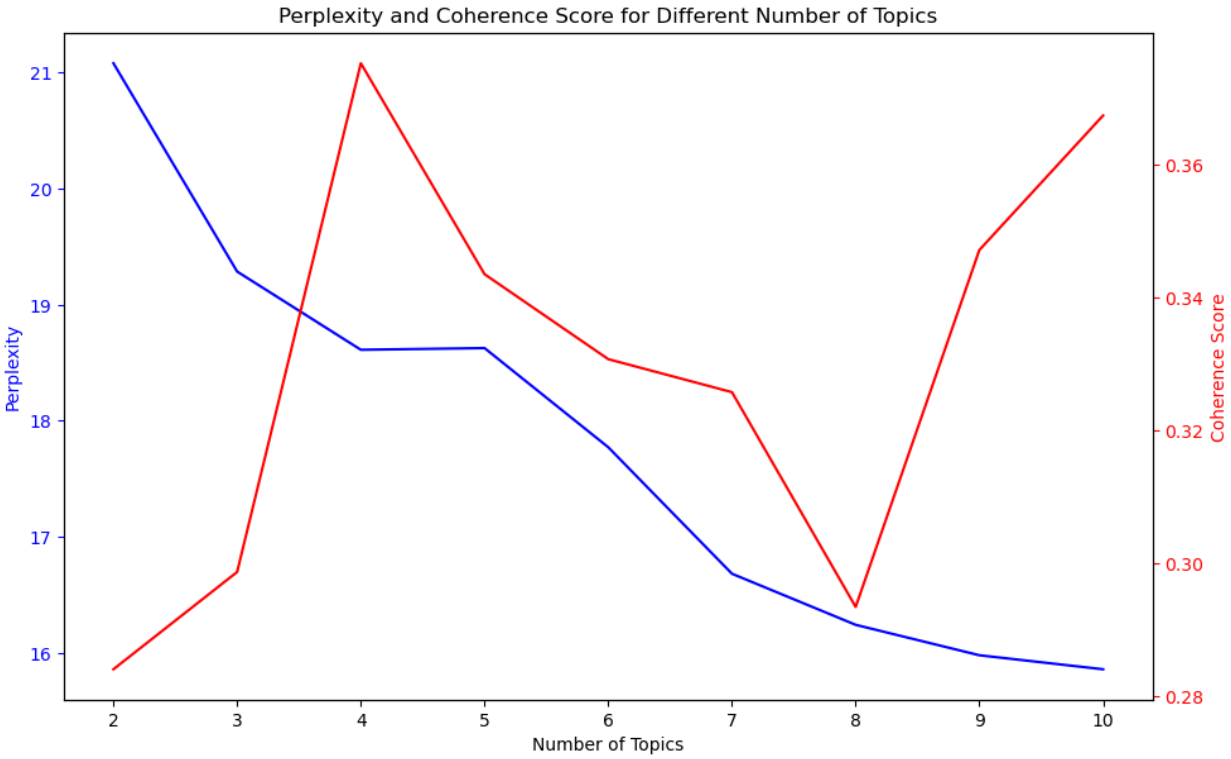


Figure 3.5(2a) Perplexity and Coherence Line Graph of Optimal Number of Topics

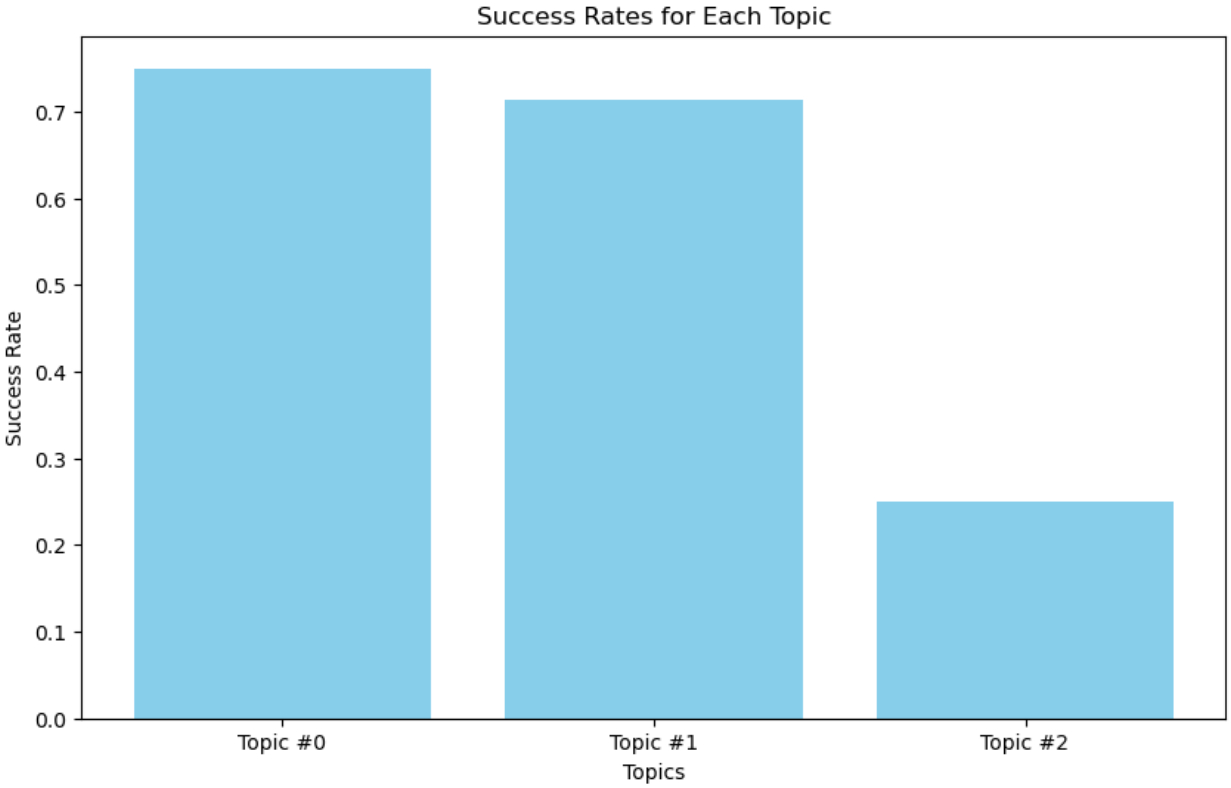


Figure 3.5(2b) Bar Plot of Success Rates of Each Topic

Research Question: How do pass features, interactions, and predictive models impact match outcomes and performance metrics?

Rationale and Justification: Machine learning models, specifically Logistic Regression and Random Forest, were applied to assess and predict the impact of passing features on match outcomes. These models outperform traditional statistical methods by effectively handling nonlinear relationships and interactions between multiple pass features. Feature selection was used to include only the most relevant predictors, optimizing accuracy and interpretability.

Data Collection and Preprocessing: The dataset included key features such as pass length, vertical and horizontal changes, vertical advancement rate, and horizontal speed rate, which are crucial for analyzing passing behaviors and their influence on match results. These features provide insights into tactical aspects like ball advancement and play spreading. Missing data was addressed by filling NaN values with 0s to ensure dataset completeness, preserving model integrity and predictive accuracy.

Standardization and Dimensionality Reduction: Following preprocessing, Principal Component Analysis (PCA) was employed to reduce the number of features for clearer visualization and analysis. PCA simplified the multi-dimensional data into 2 principal components, enabling visualization of passing characteristic interactions over time and facilitating the clustering of similar passing patterns. A Standard Scaler was used to standardize the data, ensuring all features contributed equally and preventing larger-scale features from dominating and skewing PCA results.

Temporal Analysis of Passing Behavior: By analyzing the match in 5-minute intervals, observations were made on how passing dynamics evolved throughout each half, reflecting tactical adjustments or responses to opposition play. Segmenting data into first and second halves highlighted shifts in strategy or momentum. Scatter plots of PCA components visualized passing behavior over time, revealing trends like increased long passes later in the game, or more controlled possession early on.

Clustering with DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was chosen over K-means as it detects clusters of varying densities and is resistant to noise, making it well-suited for datasets with potential outliers. Visualizing these clusters helped interpret patterns, such as clusters representing short, controlled passes, or long directed balls. This clustering offered insights into tactical decisions during different phases of play. The DBSCAN scatter plot is evident in Figure 3.5(3a).

Predictive Modeling of Pass Accuracy: Pass accuracy is critical for maintaining possession and successful build-up play. Predicting pass accuracy based on pass features provides insights into the attributes that control game flow. Since the dataset exhibits class imbalance due to more accurate passes, oversampling was used to balance the data and improve model performance. Logistic Regression, valued for its simplicity and interpretability, identified key features linked to pass accuracy, while Random Forest, captured complex interactions between features. Both models were evaluated using accuracy scores and confusion matrices, demonstrating their effectiveness in predicting pass accuracy and informing tactical decisions.

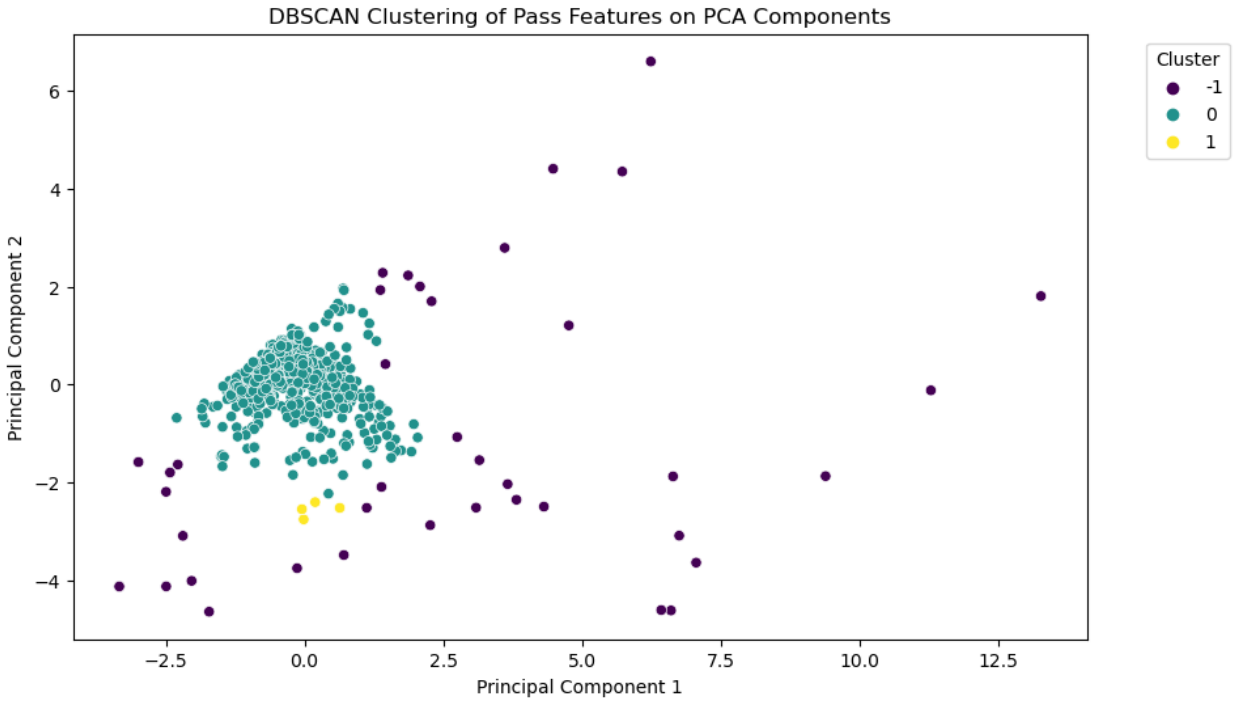


Figure 3.5(3a) Scatter Plot of DBSCAN Clustering on Pass Features on PCA Components

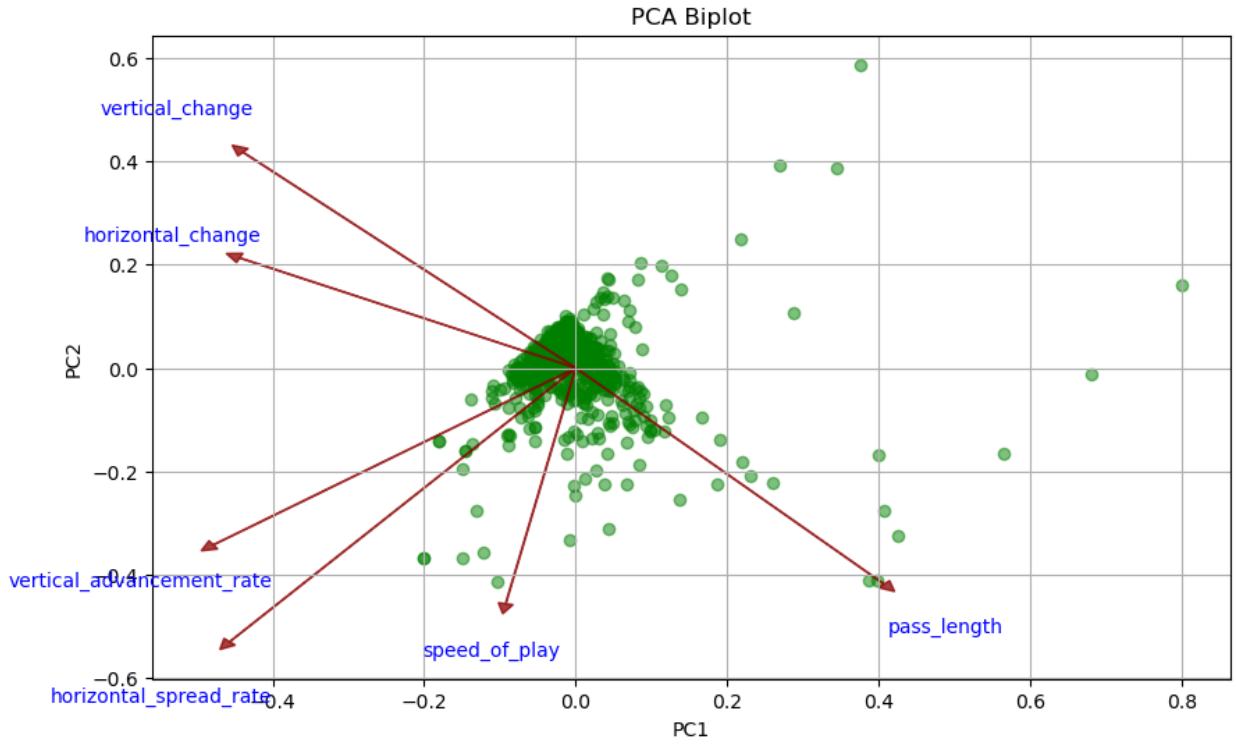


Figure 3.5(3b) Biplot of PCA Components of Pass Features

Research Question: To what extent do player duel positions, types, frequency, and temporal dynamics contribute to understanding player performances, roles, and team strategies?

Rationale and Justification: Agglomerative clustering and PCA were employed to analyze duels, offering deep insights into player behaviors and interactions. This approach preferred over simple aggregation or mean comparisons, uncovers natural groupings and patterns crucial for tactical analysis. PCA reduces the complexity of duel data while retaining its most informative aspects, proving more effective than singular value decomposition (SVD) for non-symmetric data.

Data Preprocessing and Feature Engineering: The dataset included key features like duel duration, start and end positions, and distance covered, alongside qualitative tags indicating duel outcomes such as 'Won' and 'Lost'. Missing values were filled with 0s to preserve data integrity, as they typically signify absent information rather than data errors. Duel outcome tags were one-hot encoded into binary columns, capturing their influence. A Standard Scaler was used to standardize features, ensuring that each contributed equally to the clustering algorithm, for distance-based methods like Agglomerative Clustering.

Hierarchical Clustering: Agglomerative Clustering was applied to the scaled data to reveal natural groupings of duels. Using Ward Linkage didn't require a predefined number of clusters and minimized variance between merged clusters, producing compact clusters. A dendrogram visualized the hierarchical structure, aiding in identifying optimal cluster cutoff points. The Elbow Method indicated that 3 clusters best represented the duel characteristics, such as differences in distance covered or duel outcomes. The elbow curve is shown in Figure 3.5(4a) and the dendrogram in Figure 3.5(4b). A summary table of duel clustering can also be seen in Figure 4.5(4a).

Cluster	Duel Type	Description	Tactical Insight
Cluster 1	Defensive Duels	Occurred near the defensive third, mostly interceptions	Key to preventing opposition from advancing
Cluster 2	Midfield Duels	Occurred in the middle of the pitch, involving tackles	Important for maintaining midfield control
Cluster 3	Attacking Duels	Took place in the attacking third, involved forward challenges	Key to creating scoring opportunities

Figure 4.5(4a) Summary Table of Duel Types and Tactical Insights

Dimensionality Reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE): t-SNE projected high-dimensional features into a two-dimensional space, emphasizing the preservation of local data structures. Clusters were plotted in a t-SNE space, allowing interpretation of separations and overlaps between duel types, which validated the clustering results. While t-SNE doesn't capture overall data variance as PCA does, it reveals complex, non-linear relationships. Scatterplots in the t-SNE space visualized the separation between clusters, revealing distinct patterns in duel types. Heatmaps displayed the spatial distribution of duels, highlighting tactical tendencies such as offensive versus defensive duels. The t-SNE clustering scatterplot is shown in Figure 3.4(4c).

Heatmaps of Duel Performance: Heatmaps were used to visualize duel outcomes, highlighting players who excelled in specific duels and informing their tactical roles. Another heatmap showcased the average duel duration per player, identifying those involved in prolonged duels and their success rates. A heatmap of the average distance covered during duels revealed the most mobile players, offering insights into who covered more ground during engagements. This comprehensive view of duel performances enriched the tactical understanding of individual and team contributions in various match contexts.

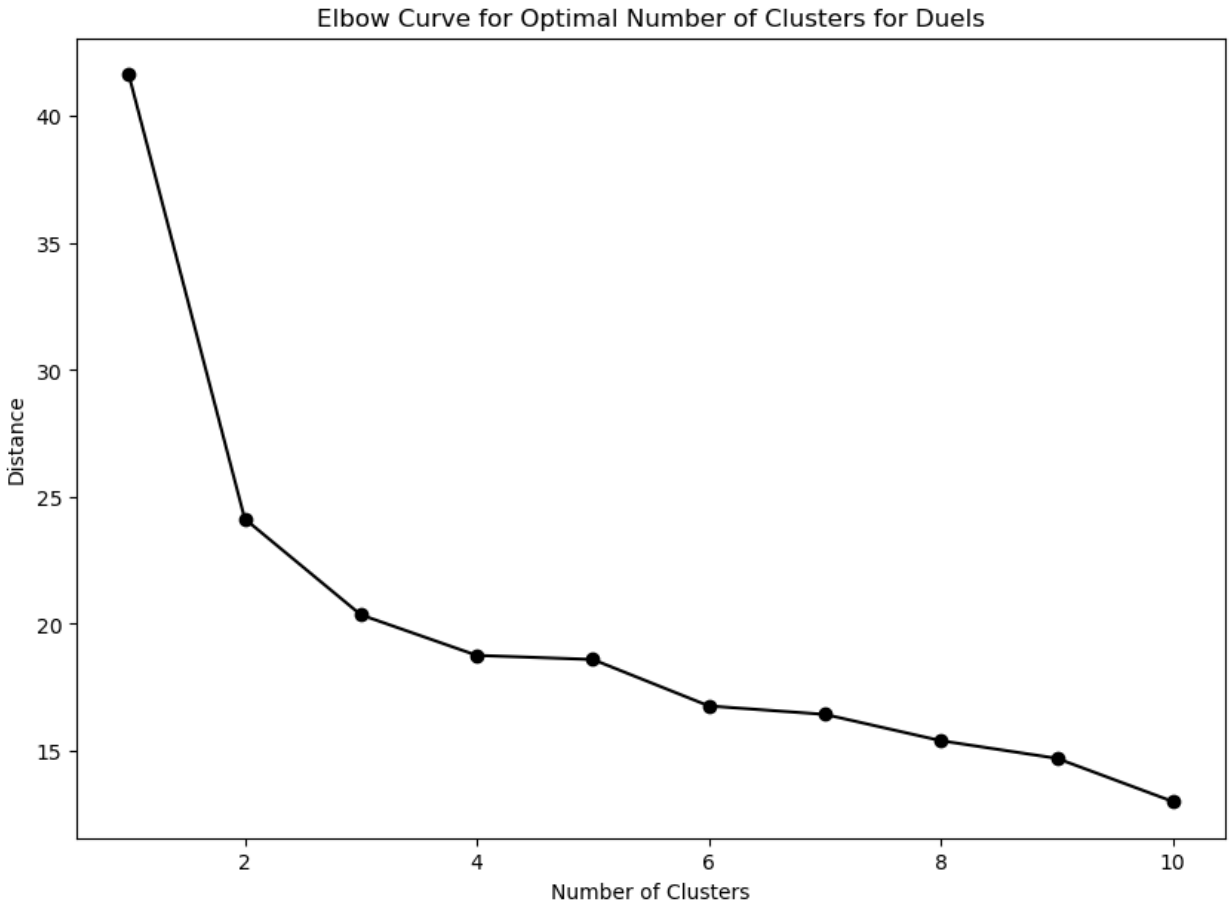


Figure 3.5(4b) Elbow Curve Showing Optimal Number of Clusters of Duels

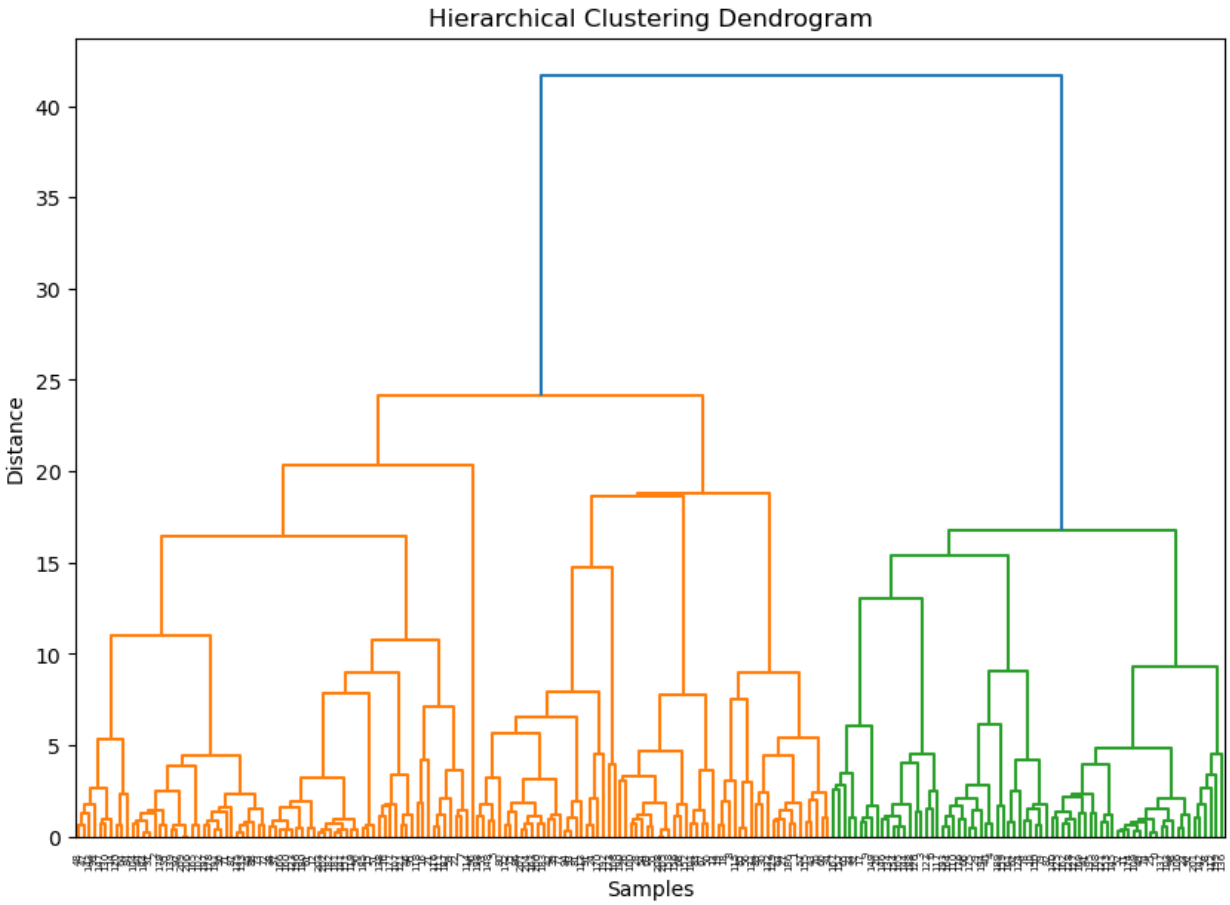


Figure 3.4(4c) Dendrogram Showing the Clustering Process of Duels

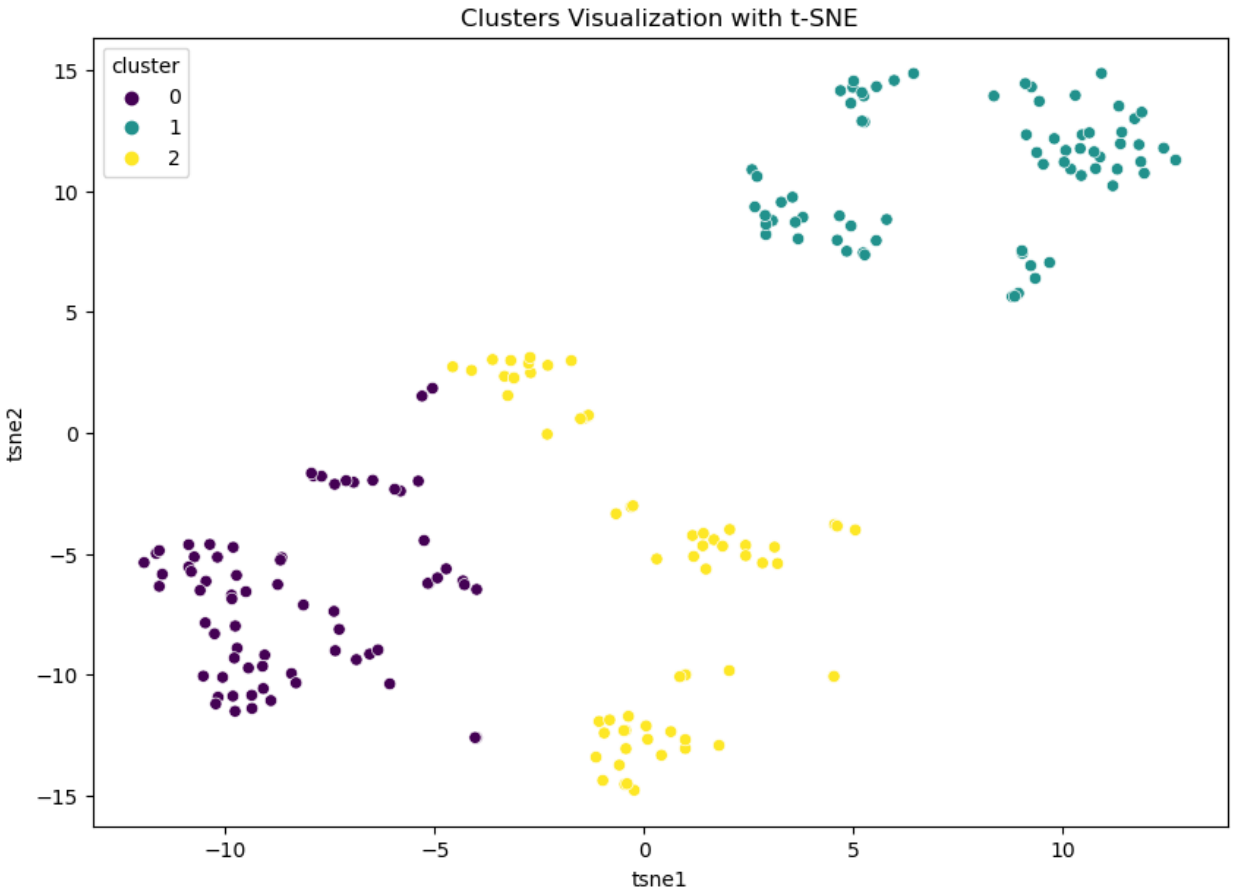


Figure 3.4(4d) Scatter Plot Showing Clustering of Duels With t-SNE

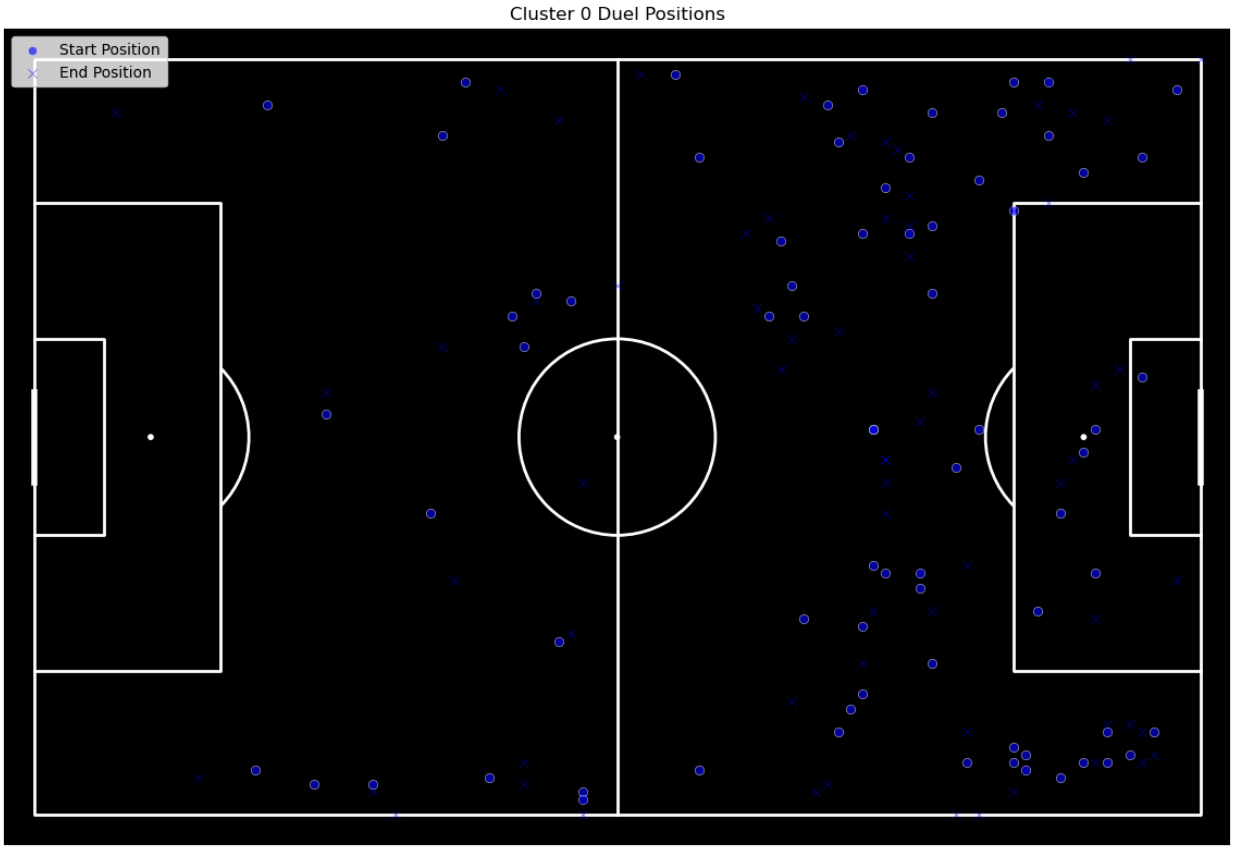


Figure 3.5(4e) Pitch Plot Showing Duel Positions of Cluster 0

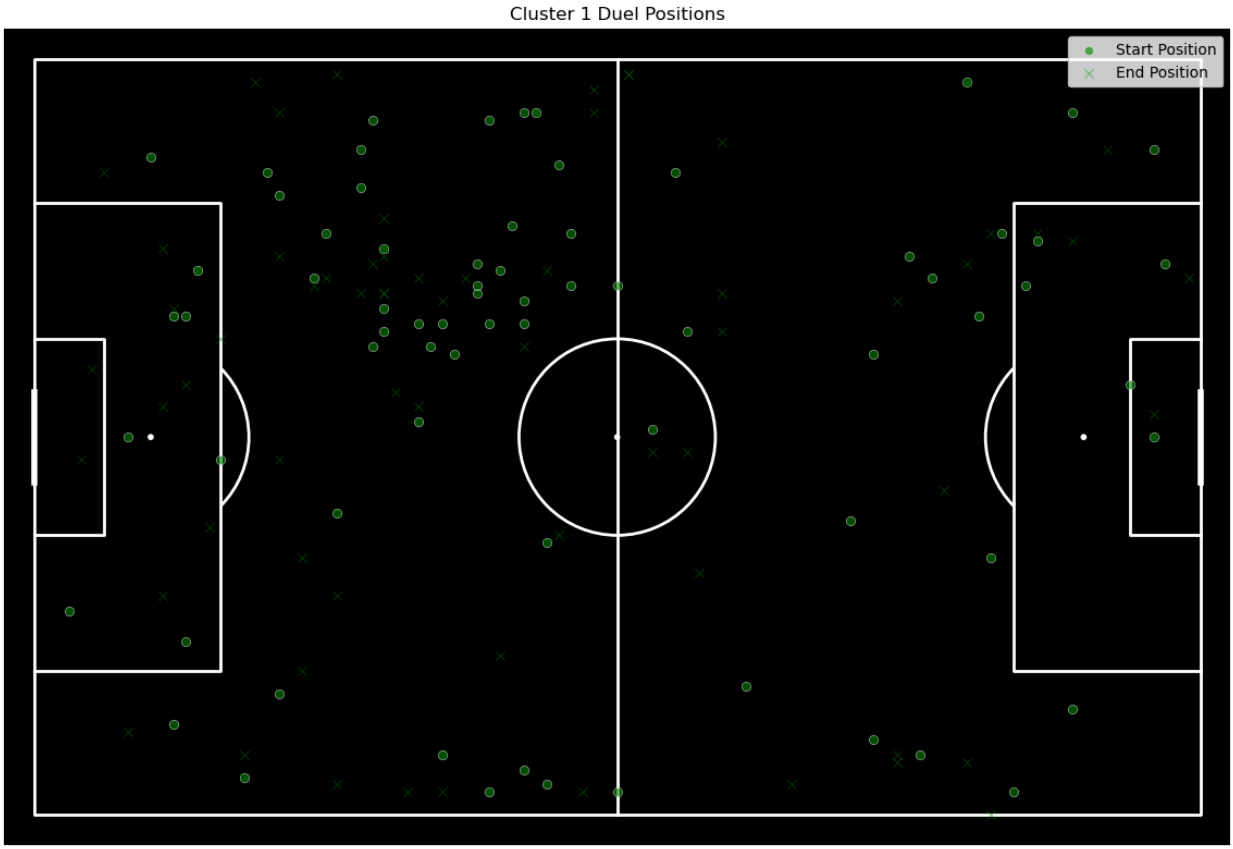


Figure 3.5(4f) Pitch Plot Showing Duel Positions of Cluster 1

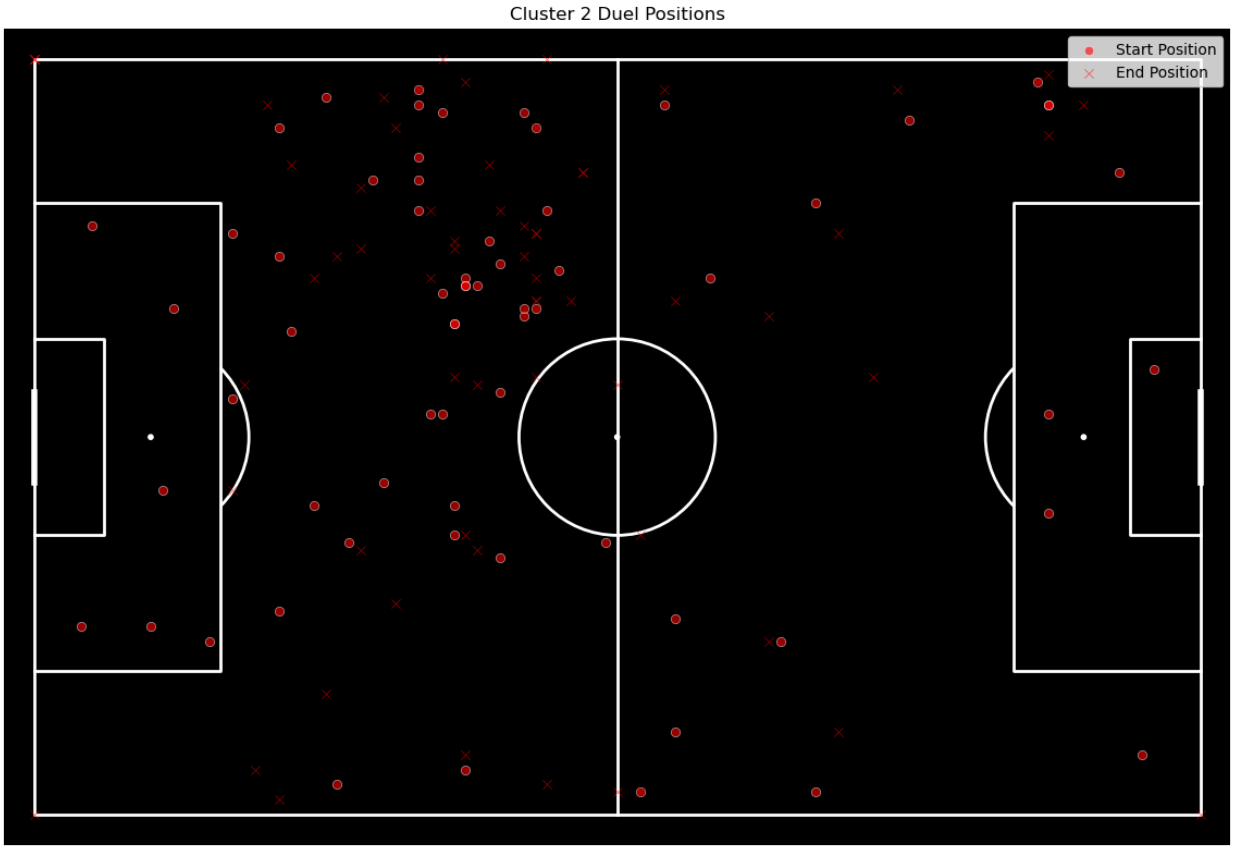


Figure 3.5(4g) Pitch Plot Showing Duel Positions of Cluster 2

Research Question: What insights into offensive strategies, scoring efficiency, and team tactics can be gained from analyzing shooting patterns, and positional data, and using machine learning models to predict goal outcomes?

Rationale and Justification: Clustering and Logistic Regression were chosen to analyze shooting patterns for their ability to classify and predict outcomes from complex input features, providing strategic insights into scoring efficiencies. This approach surpasses descriptive statistics, offering predictive power and the capacity to handle complex interactions within shooting data. Logistic Regression effectively models binary outcomes, such as scoring, and captures non-linear effects better than traditional methods.

Data Preprocessing: The dataset was enhanced with derived features representing tactical and positional shot contexts. Distance to goal was calculated using Euclidean distance from the shot's starting position to the goal center, a critical metric in expected goal models. The shot angle was computed using trigonometry to capture the player's position relative to the goal. Players were expanded into individual binary columns representing their involvement, while data tags were transformed into binary columns to include metadata for analyzing shooting outcomes.

Clustering: Key features like 'distance to goal', 'angle of shot', shot coordinates, and expected goal contributions were selected. A Standard Scaler standardized these features to ensure uniform contributions in the clustering model. In Figure 3.5(5a), the Elbow Method determined the optimal number of clusters, balancing model complexity and interpretability. K-means was applied with 4 clusters, grouping shots by position, shot quality, and tactical context. Each cluster was plotted on a

football pitch, color-coded to highlight spatial shooting patterns and strategies, such as central close-range, or long-range shots from wide angles. A summary table of Shooting patterns can be seen in Figure 3.5(5a).

Cluster	Shot Type	Distance to Goal	Shot Outcome Likelihood
Cluster 1	Close-Range Shots	<10 meters	High likelihood of scoring
Cluster 2	Mid-Range Shots	10-20 meters	Moderate likelihood of scoring
Cluster 3	Long-Range Shots	>20 meters	Low likelihood of scoring

Figure 3.5(5a) Summary of Shooting Patterns and Outcome Likelihood

Player-Level Aggregates: Player contributions were measured using total and average expected goals, providing insights into shooting efficiency based on shot position and type. The total number of shots and outcomes was tracked to assess contributions beyond goals scored. Bar plots visualized each player's expected goals, shot count, and accuracy, allowing comparison of offensive contributions. Shooting patterns in the first and second halves were analyzed for temporal trends.

Predictive Modeling: A logistic regression model was developed to predict whether a shot would result in a goal, using features like 'distance to goal', 'angle of shot', positional data, and expected goals. The model addressed class imbalance by oversampling the goal class, improving performance for predicting rare events (goals). Five-fold cross-validation assessed the model's ROC AUC score, measuring its ability to distinguish between goals and non-goals. The average AUC score indicated predictive power, and the ROC curve illustrated the model's performance across classification thresholds.

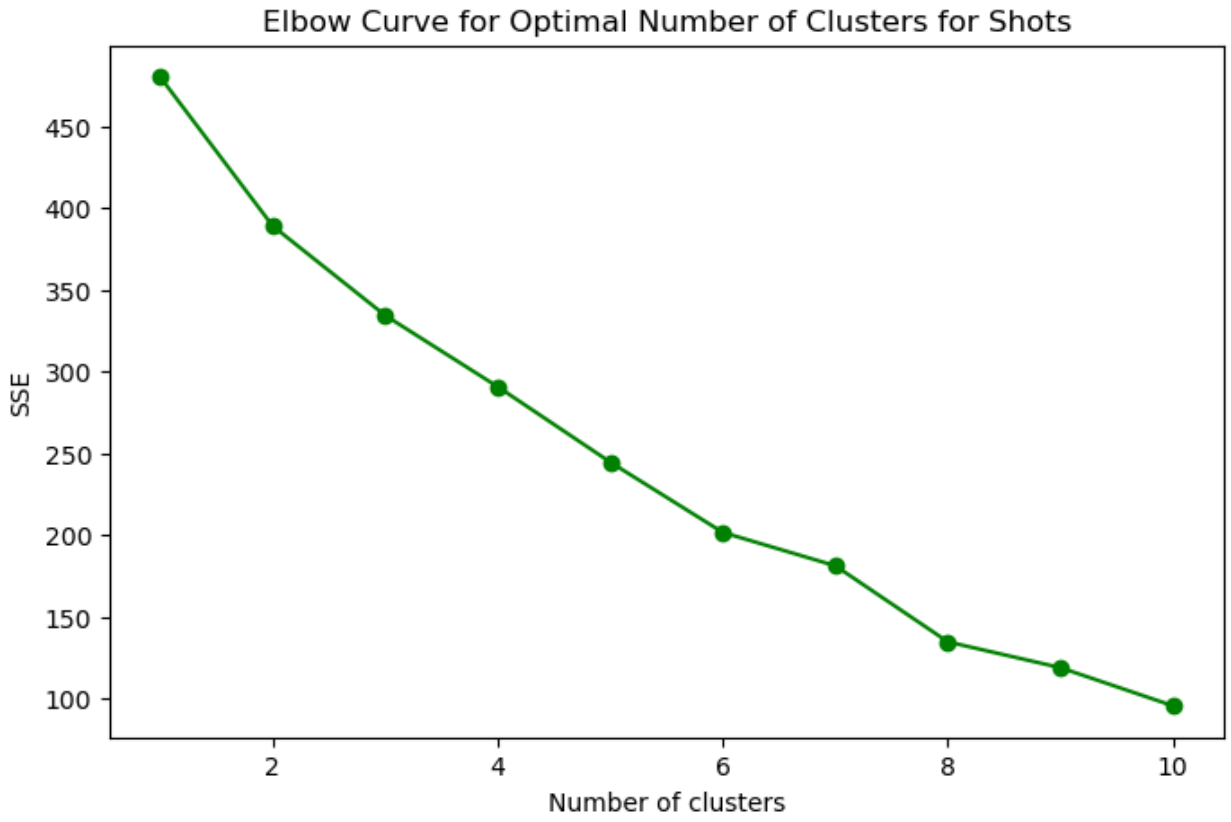


Figure 3.5(5b) Elbow Curve Showing Optimal Number of Clusters of Shots

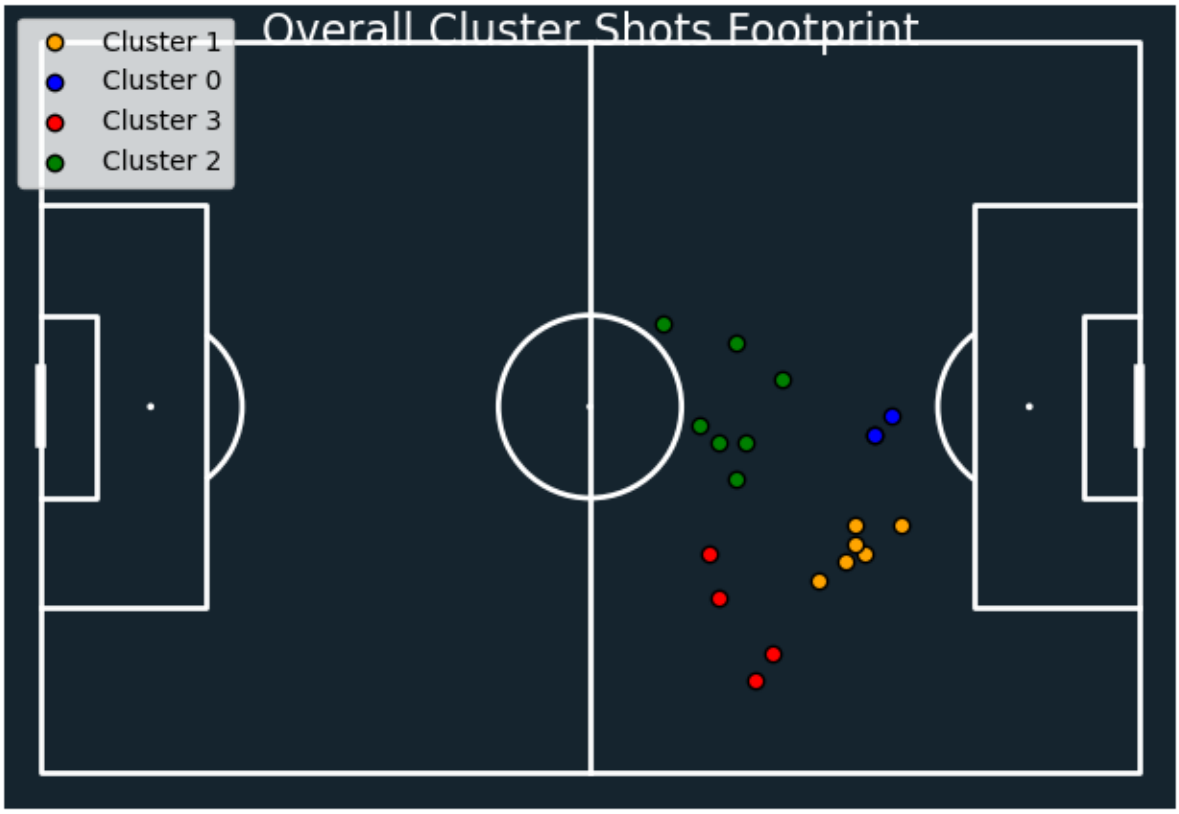


Figure 3.5(5c) Pitch Plot of Total Number of Cluster Footprints of Shots

Research Question: What insights can be gained into match dynamics by analyzing spatial and temporal foul patterns?

Rationale and Justification: The spatial and temporal analysis of fouls via clustering identifies patterns that reveal strategic or disciplinary issues within teams, which are often misled by simple observation. Clustering preferred over basic temporal tracking, provides deeper insights into the distribution and commonality of fouls, uncovering trends and anomalies critical for developing strategic responses to in-game events.

Data Preprocessing: The analysis focused on temporal and spatial variables to understand foul patterns. Temporal variables captured the specific time fouls occurred, while spatial variables represented the starting and ending coordinates of foul-related events, identifying where fouls were committed. The distance between start and end positions was calculated to highlight aggressive or spatially significant fouls. Features were standardized using a Standard Scaler to ensure uniformity for clustering algorithms like K-means and Agglomerative Clustering, which are sensitive to varying feature scales.

Clustering Approach: The Elbow Method determined that 3 clusters were optimal for the analysis. Agglomerative Clustering was then applied, grouping fouls based on spatial and temporal features. Each foul was assigned to one of the 3 clusters, considering similarity. These clusters were visualized on a football pitch, plotting the starting and ending points to identify spatial patterns. Another plot compared all clusters on a single pitch, revealing whether certain clusters were more offensively or defensively located.

Player and Team Analysis: Fouls committed by the starting XI were analyzed to identify players influencing match dynamics. Player names were extracted and linked to their foul events. A pitch plot visualized where starting XI players committed fouls, offering insights into positional tendencies and player discipline. For better understanding, the number of fouls in each half was compared, revealing shifts in aggression or defensive focus as the match progressed. A bar plot displayed foul counts from the first and second halves, highlighting temporal changes. Fouls were classified as either offensive or defensive to illuminate team tactics. Defensive fouls suggested attempts to prevent scoring opportunities, while offensive fouls indicated aggressive forward play. A classification scheme based on foul locations divided fouls into defensive or offensive zones, and their distribution was shown using a pie chart. Additionally, foul patterns across match time were analyzed, with an increase in fouls towards the end of the match potentially indicating fatigue or tactical fouling. A bar chart grouped fouls by minute, visualizing how fouling behavior evolved throughout both halves of the game.

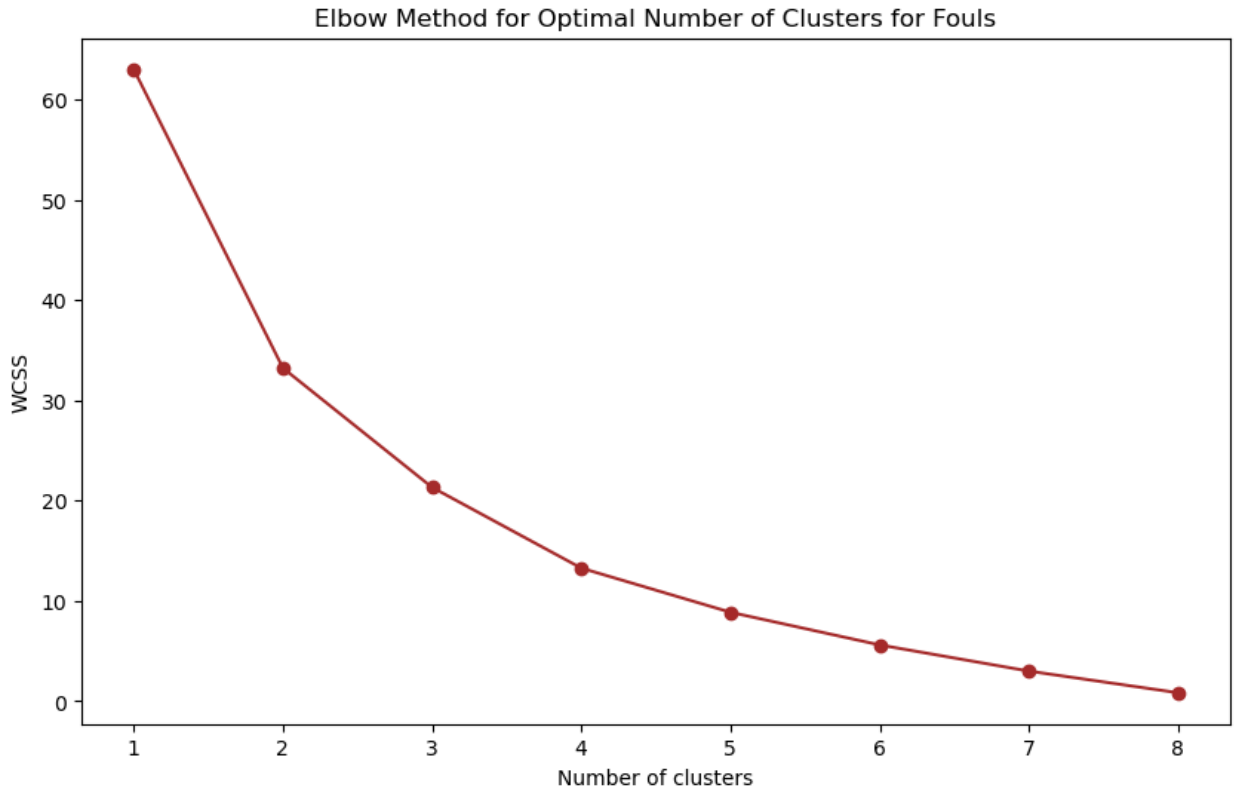


Figure 3.5(6a) Elbow Curve Showing Optimal Number of Clusters for Fouls

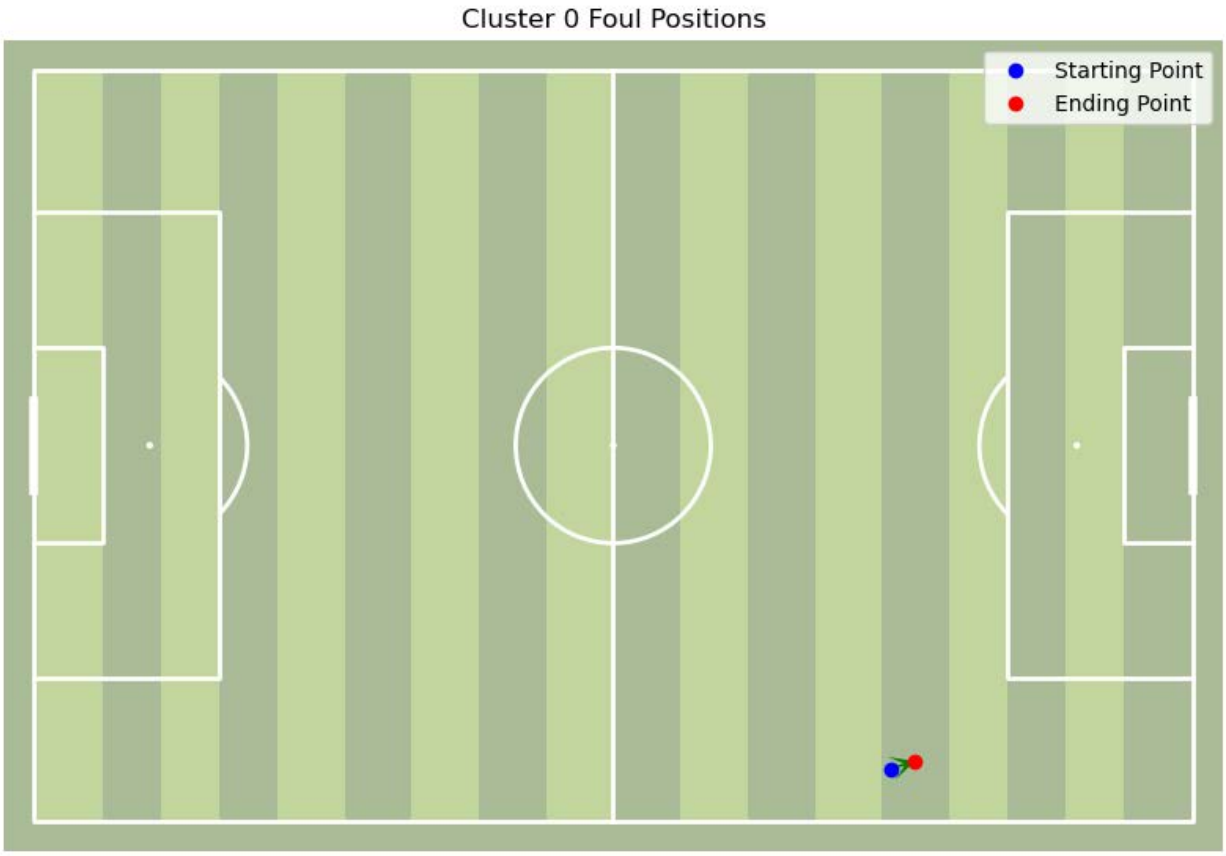


Figure 3.5 (6b) Pitch Plot showing Foul Positions of Cluster 0

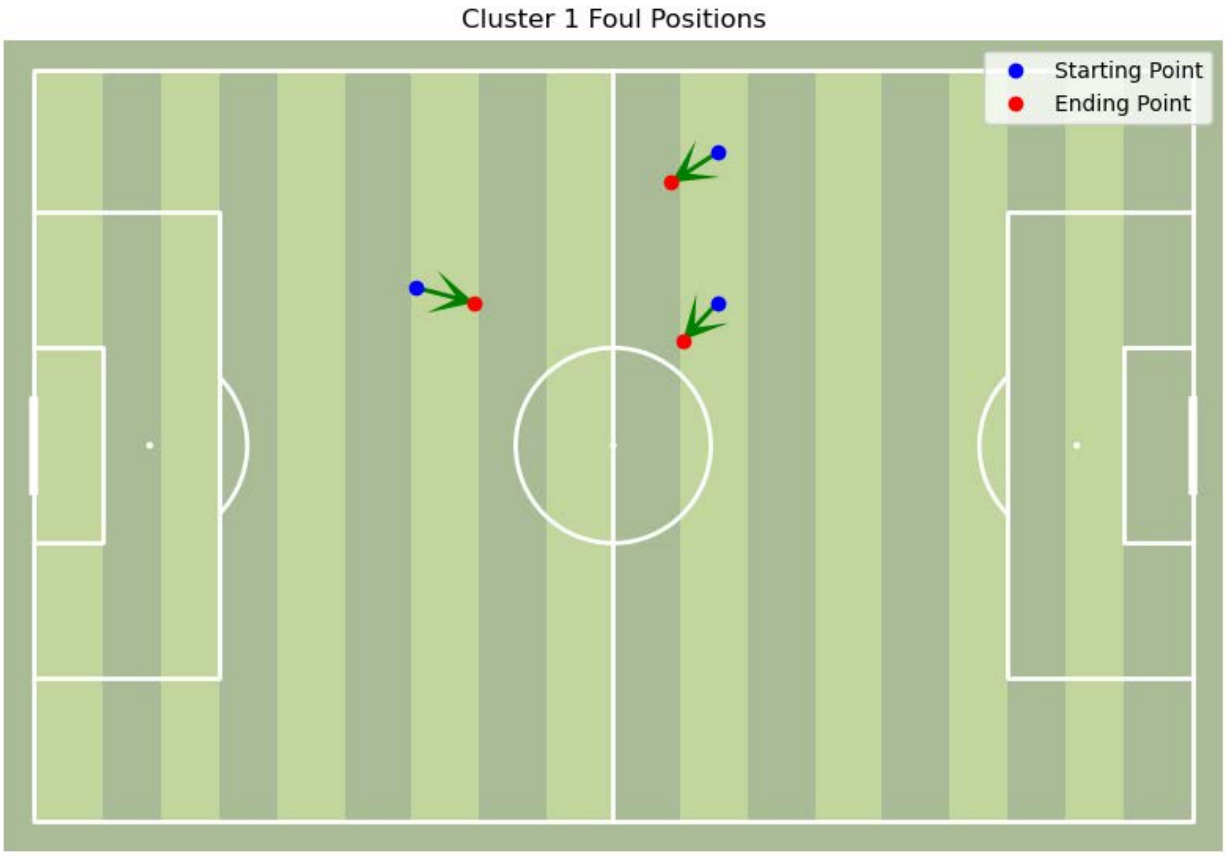


Figure 3.5 (6c) Pitch Plot showing Foul Positions of Cluster 1

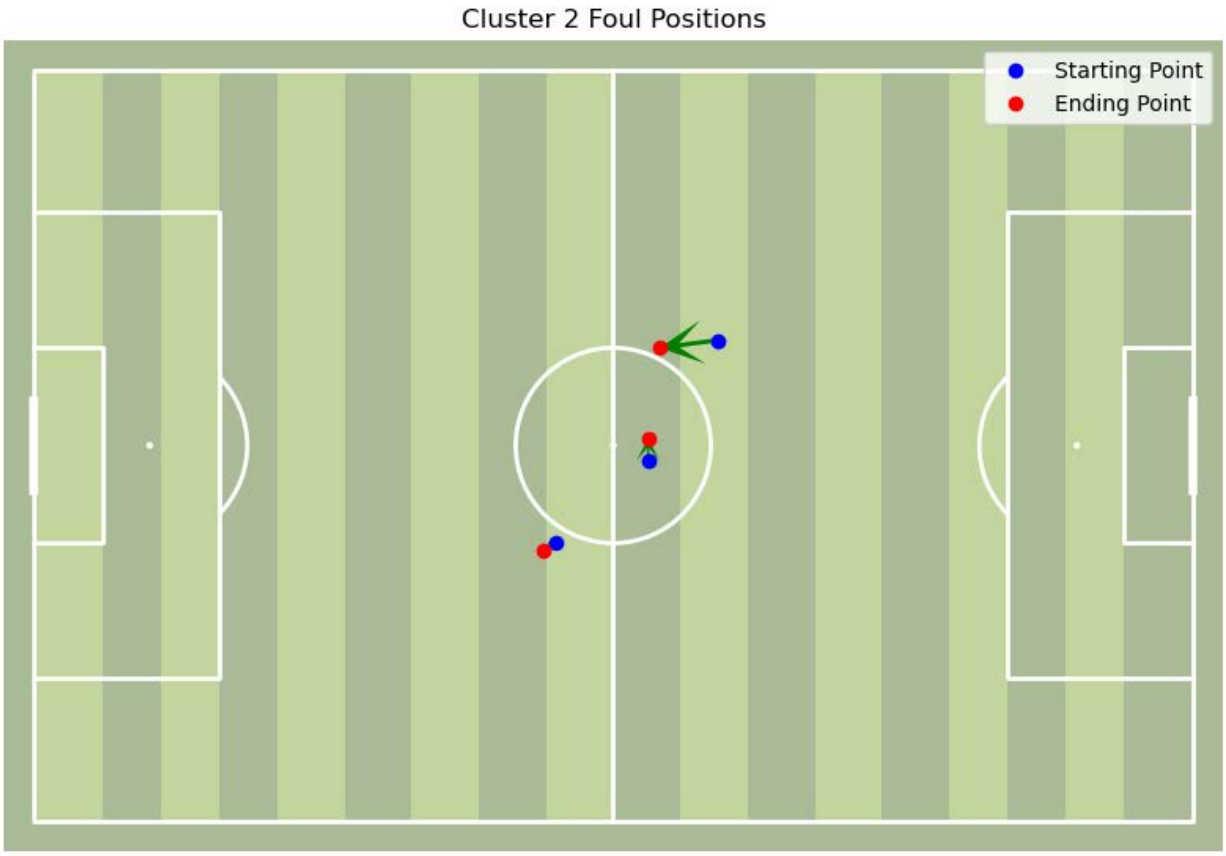


Figure 3.5 (6d) Pitch Plot showing Foul Positions of Cluster 2

Chapter 4 – Results

4.1 Introduction

This chapter presents the findings from the match data analysis, outlined in Chapter 3. It focuses on the data collected and preprocessed from the 'Soccer match event dataset' and its analysis using techniques such as network analysis, clustering, PCA, and machine learning models to address the research questions posed in Chapter 1. The results are organized around the key research objectives: understanding the impact of passing strategies on team performance and player involvement, identifying patterns in event sequences and technical strategy, and evaluating how specific football actions such as duels, shots, and fouls affect match outcomes. Techniques such as K-means clustering, network centrality metrics, and machine learning models have been applied to reveal hidden patterns and relationships within the match data. Graphical representations such as network graphs, heatmaps, and pitch plots, help visualize key findings and trends. Each section focuses on a specific research question or aspect of the game, presenting the data clearly and structured, and concluding with a summary of key insights. This chapter lays the groundwork for the deeper interpretation and discussion in Chapter 5. By the end, the reader will gain a comprehensive understanding of the outputs generated through the analytical methods and their contribution to the core research questions.

4.2 Network, Player Centrality, and Positional Dynamics Analysis

Passing Network and Average Pitch Positions for Passes

R. Carvalho to R. Guerreiro (17 passes) was the most frequent connection, showing a strong reliance on the left flank for ball possession. R. Guerreiro playing as a full-back, was a key outlet, receiving many passes from the center-back R. Carvalho. This highlights a tactical focus on the left, with R. Guerreiro advancing the ball from deep positions. R. Guerreiro to J. Moutinho (15 passes), and J. Moutinho to R. Guerreiro (14 passes) connections show Moutinho's central role in maintaining possession and linking up with Guerreiro. This connection was crucial for transitioning the ball from defense midfield, with Moutinho orchestrating play in the middle. Pepe to R. Carvalho (15 passes) indicates a defensive passing pattern, where the center backs frequently exchanged passes to retain possession and build up from the back. Pepe to Vieirinha (14 passes) represents a similar defensive link on the right side. Like Guerreiro, Vieirinha was responsible for advancing the ball down the flank, playing a comparable role to Guerreiro on the opposite side. The balanced use of wide players suggests stretching the opposition by attacking from both flanks. Average pitch positions reveal that J. Moutinho and R. Carvalho held central roles, highlighting their importance in linking defense and attack. R. Guerreiro and Vieirinha were positioned higher, reflecting their roles as wide outlets. C. Ronaldo and Nani were higher up the pitch, emphasizing their roles in receiving passes in advanced areas, although they were less involved in the build-up play.

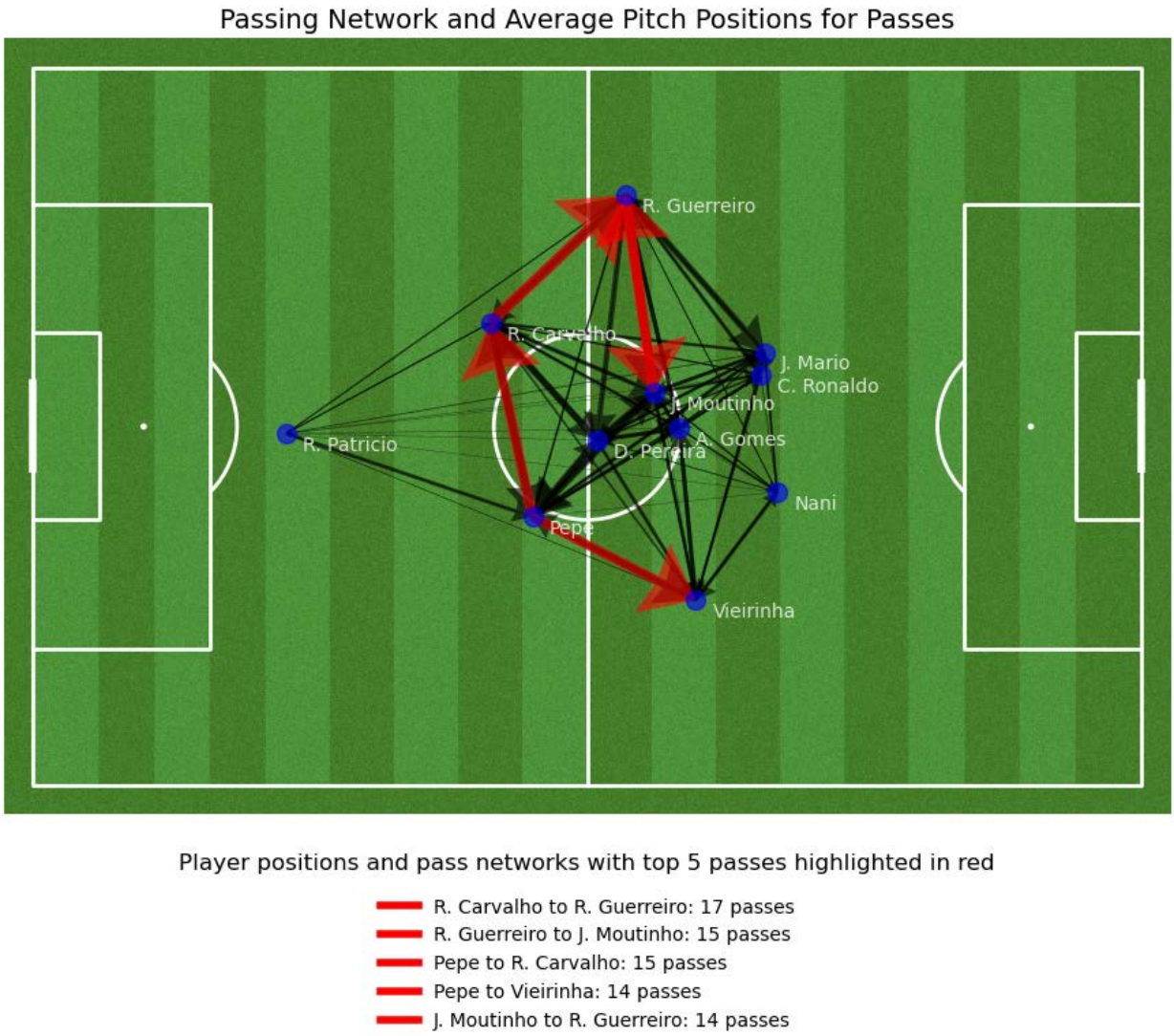


Figure 4.2(1a) Passing Network and Average Pitch Positions for Passes

Degree Centrality for Passes

J. Moutinho had the largest node in the network, indicating he was the most central player in the passing structure. His high degree, of centrality, reflects his role as the primary playmaker, responsible for distributing the ball and linking different pitch areas. Moutinho's involvement in lateral and forward passes highlights his importance in maintaining possession and driving the team forward. R. Carvalho and R. Guerreiro also had large nodes, indicating their significance in the passing network. Guerreiro was a key outlet on the left, frequently receiving passes from defenders and midfielders. Carvalho played a central role in distributing the ball within the defensive line and initiating attacks. Pepe and Vieirinha also had prominent roles, with Pepe connecting the defensive line and midfield, and Vieirinha advancing play from the right. C. Ronaldo and Nani contributed to the build-up play, receiving passes in advanced areas, though their roles focused more on finishing attacks rather than orchestrating play.

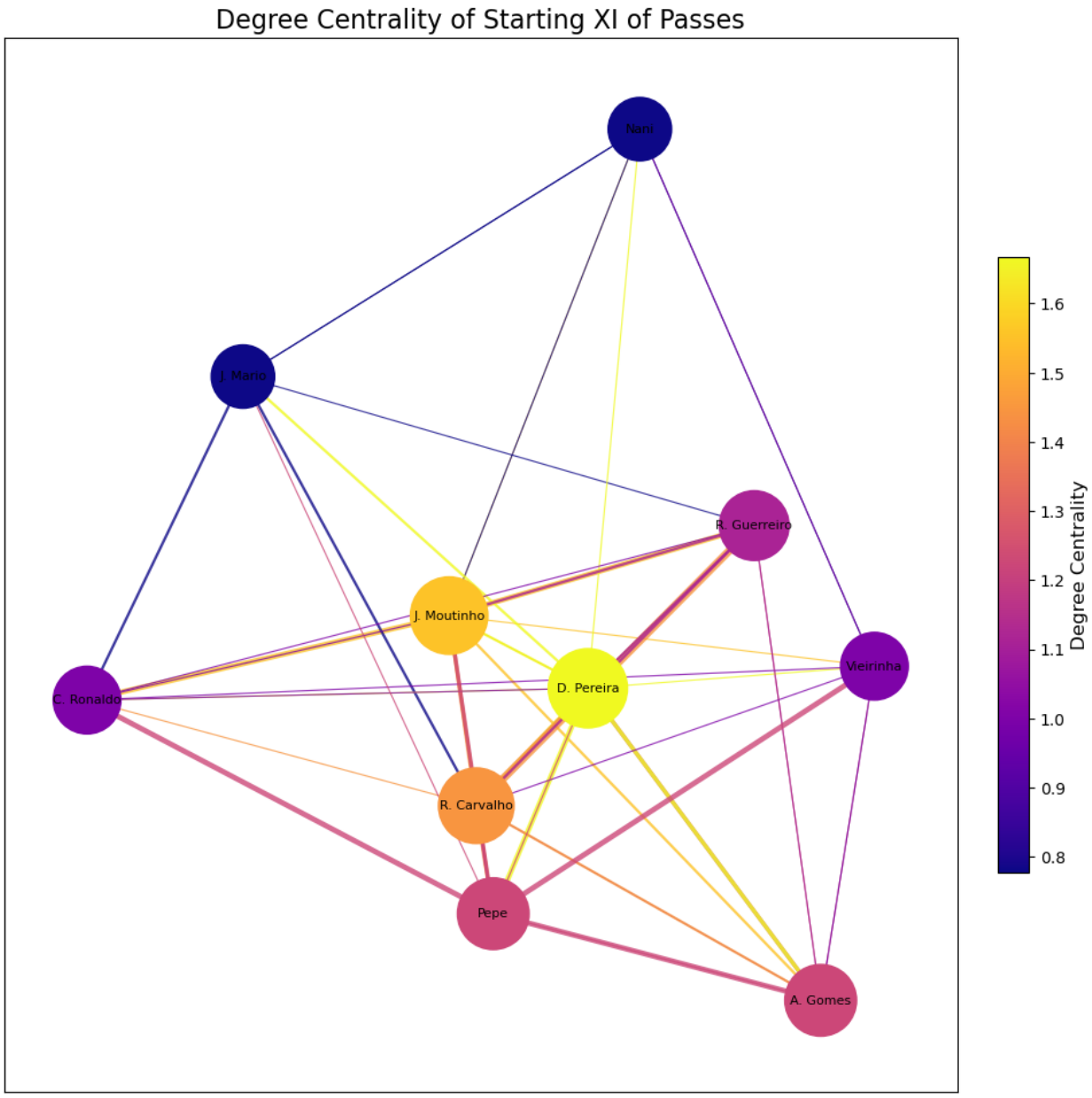


Figure 4.2(1b) Degree Centrality for Passes of Players

Passing Network and Average Pitch Positions for Long Passes

Pepe to R. Carvalho (12 passes) was the most frequent long-passing combination, showing a defensive distribution pattern between the center backs. This suggests a focus on circulating possession within the defensive line before launching forward passes to stretch opponents and create space for longer passes. R. Carvalho to Pepe (10 passes) reflects a similar dynamic in reverse, emphasizing the back-and-forth nature of play between the center backs. The connection between Pepe and Vieirinha (9 passes) suggests a more forward-oriented route. Pepe bypassed the midfield to send long passes to the right back, Vieirinha, up the pitch. This advanced play down the right flank-initiated attacks from deeper positions. R. Carvalho to R. Guerreiro (6 passes) shows a similar pattern on the left side, where Guerreiro served as an outlet for long passes. Pepe to J. Moutinho (5 passes) indicates that Pepe also looked to bypass the opposition's press by

sending long passes directly to Moutinho in midfield. The average pitch position shows defenders holding deeper positions, while full-backs were higher up the field, moving the ball forward. J. Moutinho's central position, reflects his role as a key recipient of long passes and a crucial presence in the midfield.

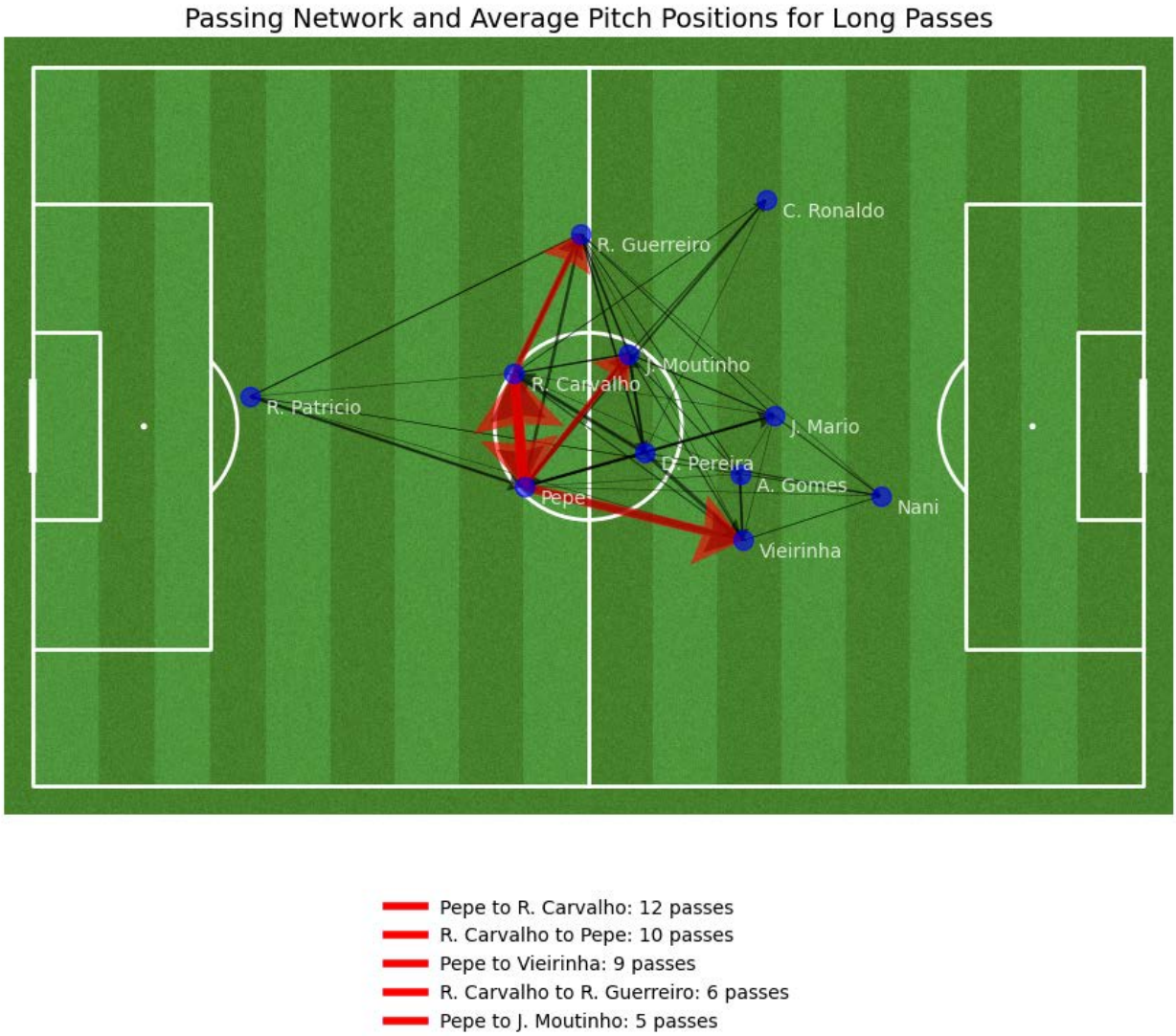


Figure 4.2 (1c) Passing Network and Average Pitch Positions for Long Passes

Degree Centrality for Long Passes

Pepe and R. Carvalho had the largest nodes, confirming their dominant roles in long-pass distribution. They were central in initiating long passes with each other and other players across the field. This reflects the team's reliance on long diagonal passes between center backs to switch play and maintain control. R. Guerreiro and Vieirinha were also frequently involved, indicating the team's interest in bypassing the midfield and targeting wide areas with long passes. J. Moutinho's relatively large node, shows his role in transitioning the team from defense to attack, controlling long balls from the backline, and redistributing them to advanced players. C. Ronaldo and Nani were less involved in distributing long passes, primarily focused on finishing attacking plays.

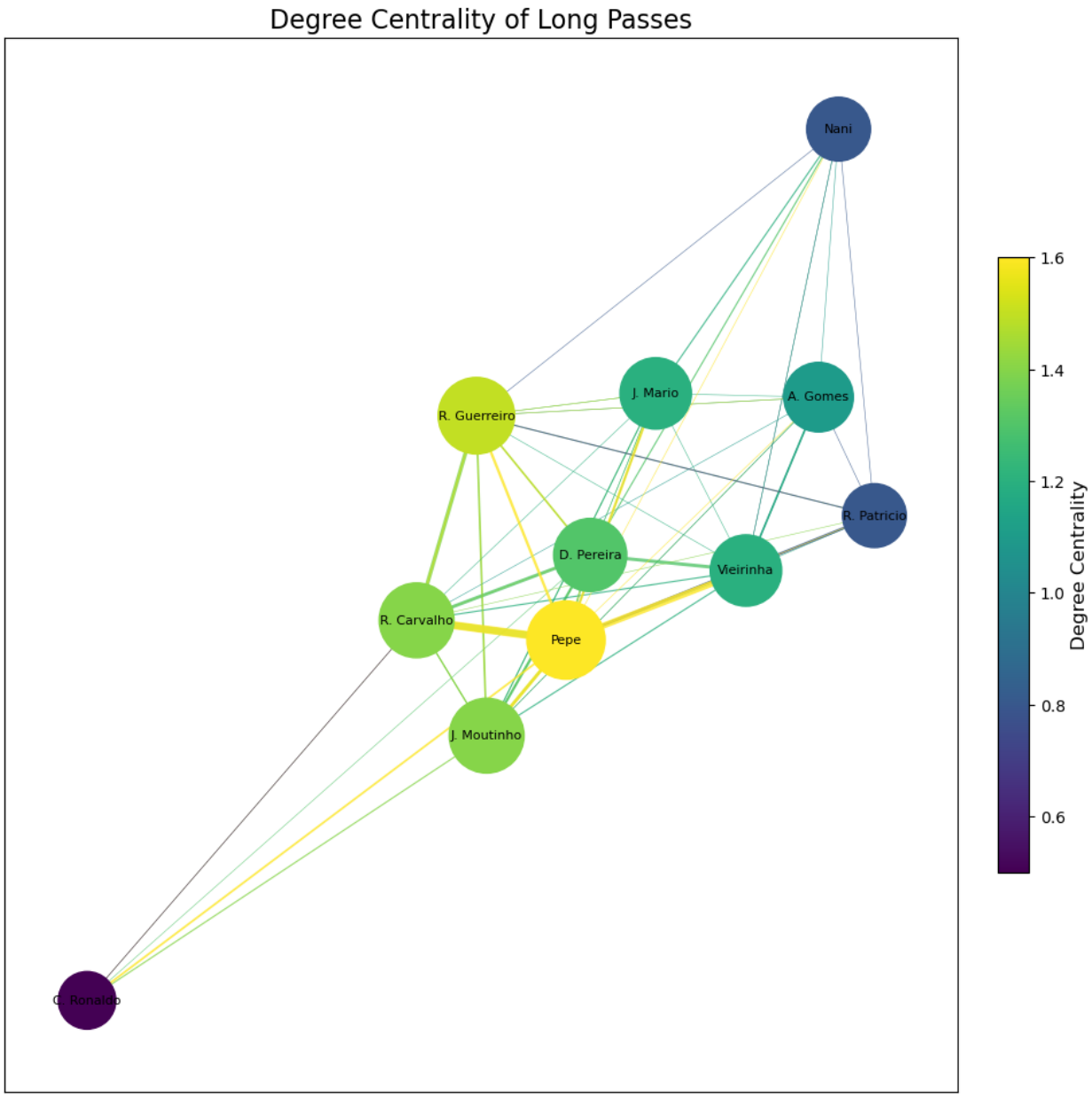


Figure 4.2(1d) Degree Centrality of Long Passes for Players

Hexbin Heatmaps for Selected Players

C. Ronaldo's heatmap reveals concentrated activity on the left side of the attacking third, with darker areas near the left side of the penalty box. This highlights his role in receiving and distributing from the left flank, often creating opportunities by taking on defenders and crossing into the box. His position in advanced areas shows his focus on contributing to attacks rather than buildup play. In contrast, J. Moutinho's heatmap shows a widespread distribution of passes across the pitch, indicating his central midfield role. Darker areas in the center of the pitch illustrate his importance in linking defense and attack, frequently dropping back to collect the ball and initiate plays. His versatility and central role in possession are clear from his activity spread across the pitch. R. Guerreiro's heatmap shows heavy involvement along the flank, with darker areas near the touchline. He was responsible for receiving wide-

positioned passes, advancing the ball forward, and linking up with midfielders and attackers. His significant activity near the opponent's box highlights his role in supporting attacking moves, often overlapping with forwards to create chances.

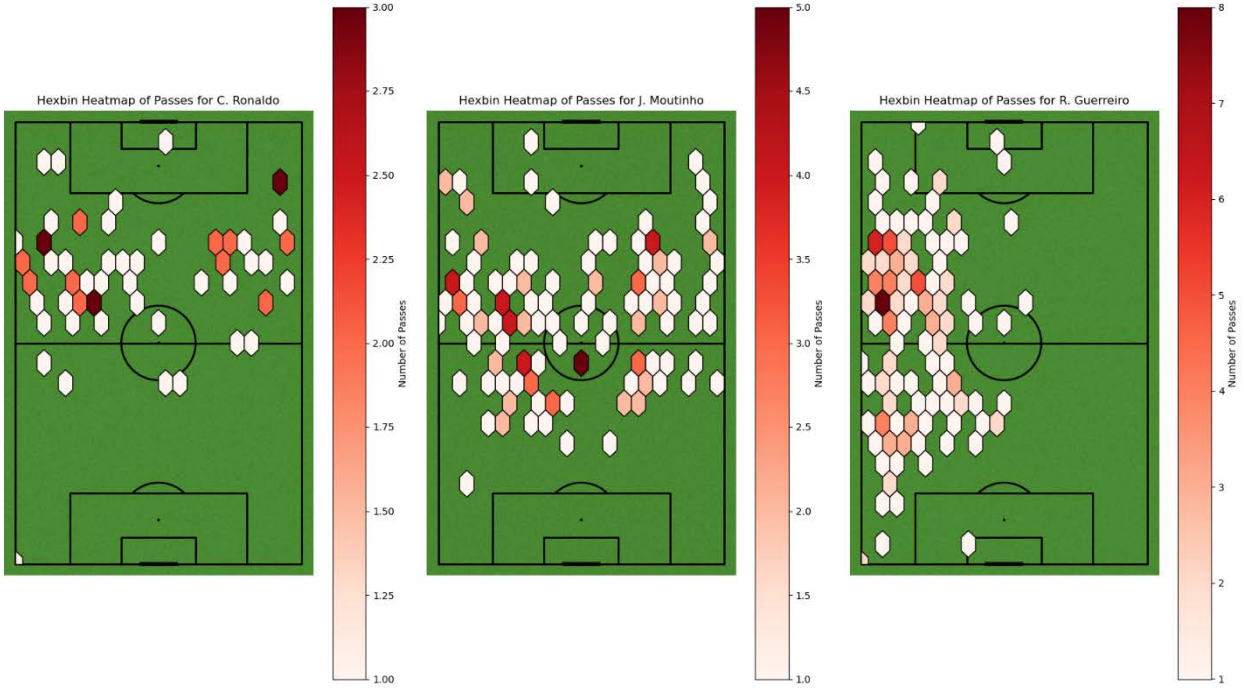


Figure 4.2(1e) Hebxin Heatmaps of Passes for Selected Players

Average Performance Metrics and Distribution of Pass Lengths in Match Analysis

Pass length was an important metric, with an average value higher than other metrics, reflecting the team's frequent use of long passes. This aligns with the earlier observations of long-pass distributions by Pepe and R. Carvalho. Possession duration was also high, focusing on ball retention and patient buildup play. Portugal prioritized controlling the game through sustained possession, gradually progressing the ball with shorter passing sequences interspersed with longer ones. Moderate values of vertical amplitude and horizontal spread suggest a balanced use of the pitch, with Portugal using vertical and horizontal movement to create space and drive into attacking areas. A relatively low average speed of play shows that Portugal emphasized controlled, precise possession rather than rapid transitions, aligning with a possession-orientated strategy. The distribution of pass lengths was right-skewed, indicating a predominance of shorter passes, with a peak in the 20–30-meter range. This supports the team's possession-based style, where shorter passes are used to maintain control. However, the longer tail of the distribution, with fewer but notable instances of passes up to 120 meters, suggests the occasional use of long balls to bypass the midfield and reach advanced players. The mix of short and long passes demonstrates a balanced approach, with shorter passes controlling the tempo and longer passes used to exploit space when needed.

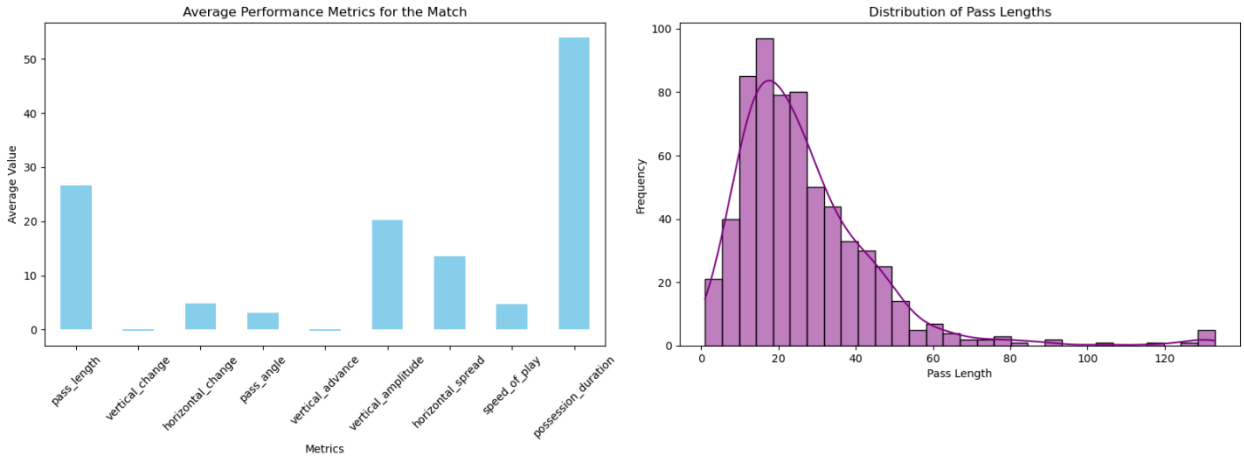


Figure 4.2 (1f) Bar Plots of Pass Length Distribution and Key Match Performance Metrics

Pass Accuracy and Dynamic Time Writing (DTW) Pass Analysis

First-half pass accuracy showed fluctuations, particularly in the early stages, possibly reflecting struggles to maintain possession under pressure, or during transitions. However, pass accuracy stabilizes as the half progresses. Accuracy stabilized as the second half progressed, indicating the team found its rhythm later. Pass accuracy was more consistent in the second half, though there were still fluctuations, indicating a drop early in the half, likely due to increased opposition pressure or tactical adjustments. Overall, pass accuracy was higher in the second half, suggesting strategic adjustments after halftime. The DTW Wrapping Path compares pass accuracy between two halves, showing how their accuracy patterns align over time. Deviations from the diagonal indicate moments where accuracy differed between halves. While the halves were generally similar, there were notable differences in certain phases, reflecting changes in dynamics or tactics during the match.

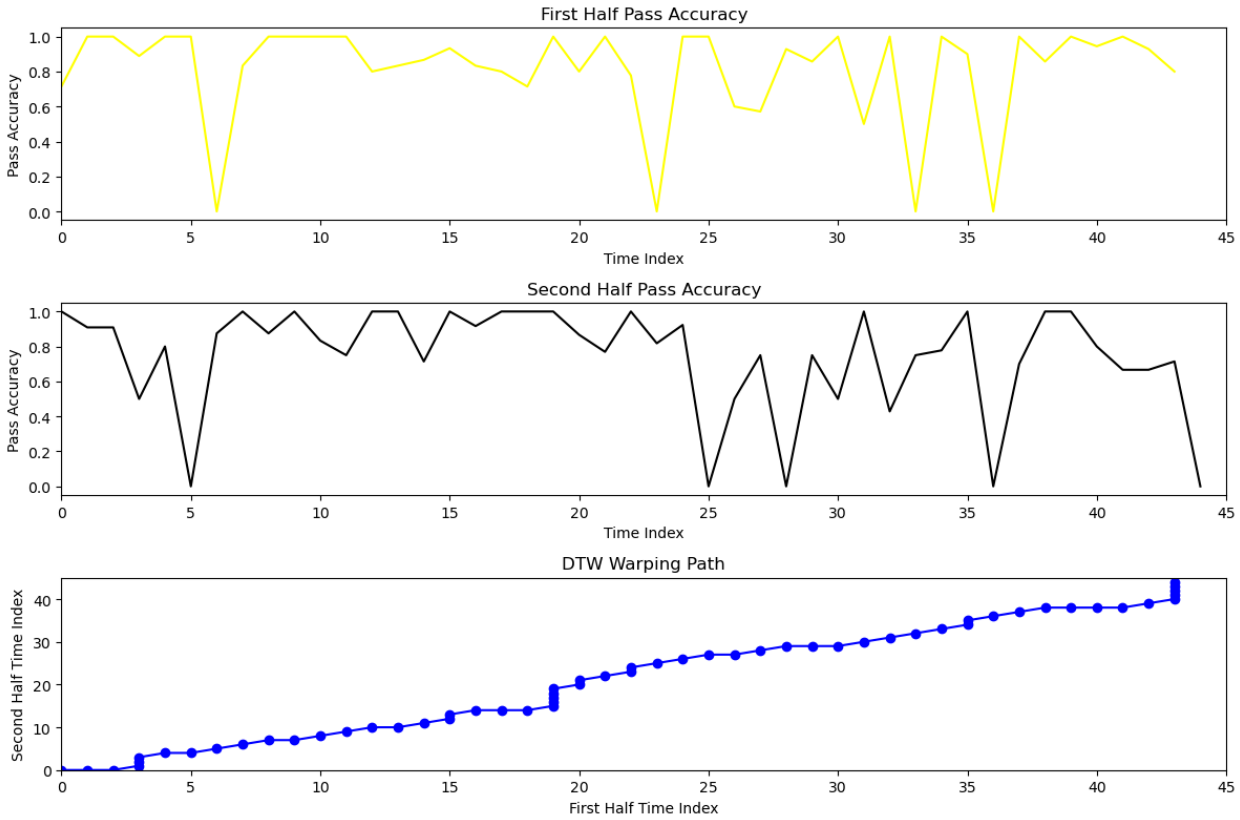


Figure 4.2(1g) Line Graphs of Pass Accuracy and a Pass Dynamic Time Wrapping

4.3 Clustering of Passing Patterns, Sequences, and Positional Analysis

Medoid Episode for Passes

Clustering of the medoid episode for general passing sequences shows a series of short to mid-range passes, primarily originating from central, or slightly leftward positions. The passes involve central midfielders, starting just beyond the halfway line, indicating they are part of build-up play or transitions to attack. The endpoints are more spread, suggesting the team distributed the ball widely, progressing towards the flanks, or into the attacking third. This visualization captures the team's passing tendencies, focusing on controlling central areas, before expanding play to the wings.

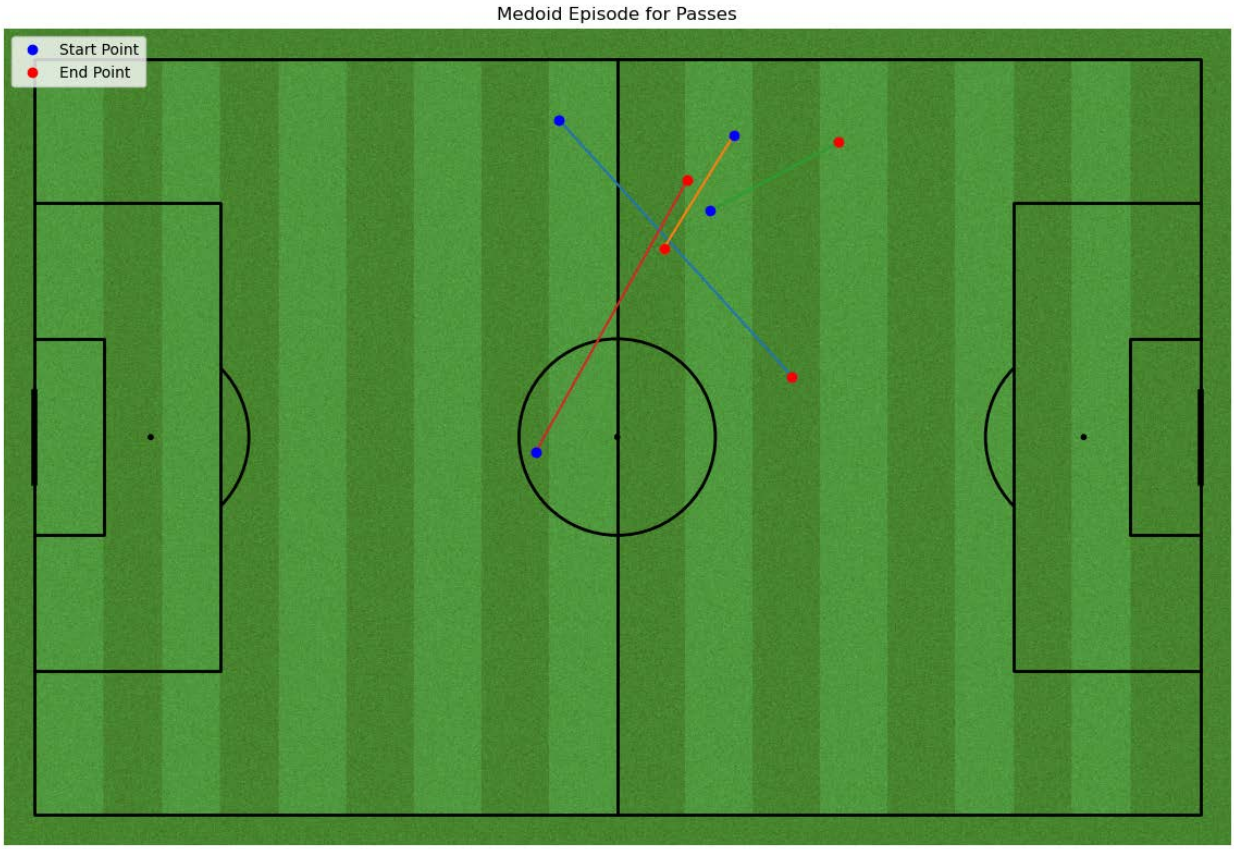


Figure 4.3(1a) Pitch Plot for Medoid Episode for Passes

Medoid Episode for Long Passes

The medoid for long passes shows they originate from much deeper positions, such as the defensive third, advancing the ball over greater distances. These passes are more direct, with longer lines stretching across the pitch. Key players, including defenders and midfielders, initiate these long passes, which typically end near the opposing goal. The medoid highlights a more aggressive strategy, bypassing the midfield to reach players in dangerous positions. Long passes are often vertical or diagonal, switching play from one side to the other, showing Portugal's ability to use long balls to stretch the opposition's defense and exploit spaces.

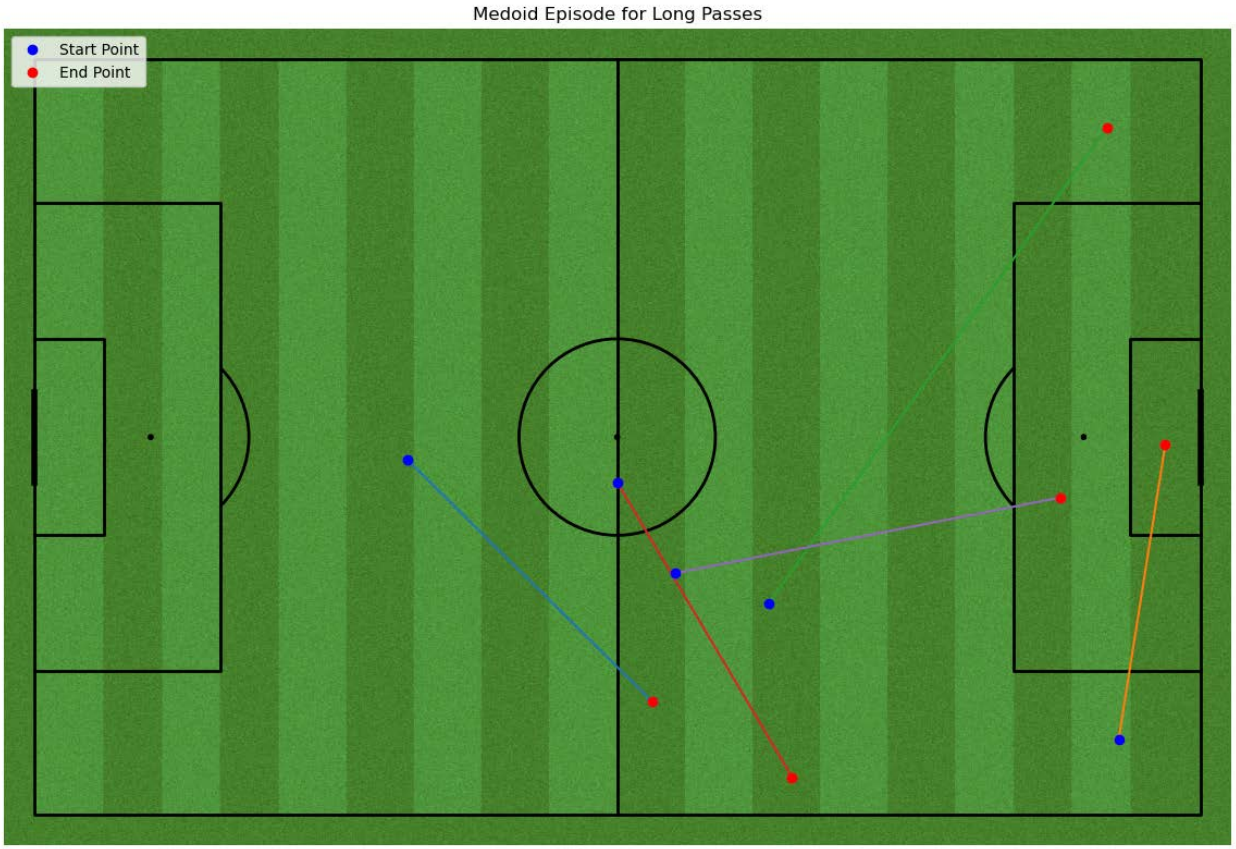


Figure 4.3(1b) Pitch Plot of Medoid Episode for Long Passes

Medoid Episode for Passes Across the Pitch

Passes across the pitch were primarily horizontal or diagonal, initiated from various points of the field. The starting points are evenly distributed across both halves, indicating these passes were used in multiple phases of play. Endpoints, often in the attacking third or on the flanks, suggest these passes were intended to switch play or move the ball across wings to create space. Diagonal passes stretched the opposition horizontally, opening gaps in the defense.

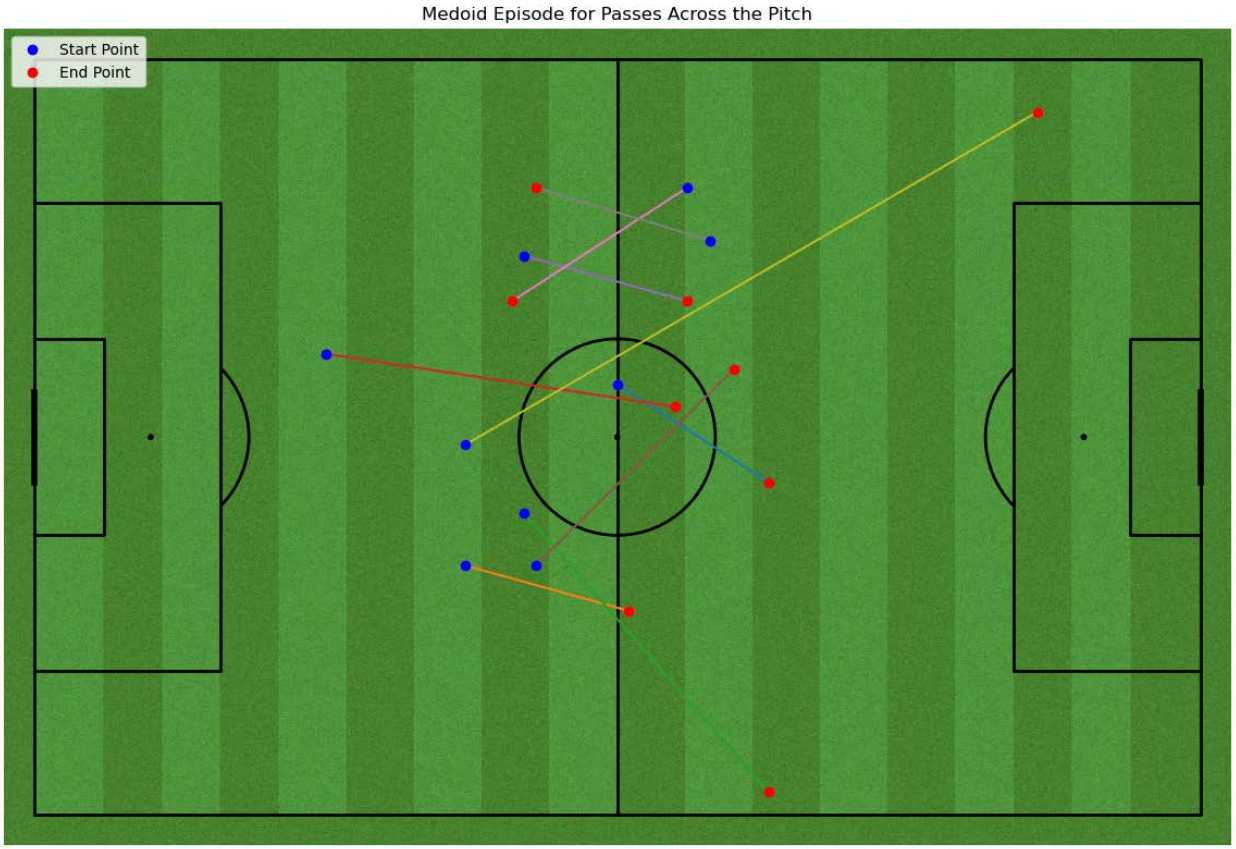


Figure 4.3(1c) Pitch Plot for Medoid Episode of Passes Across the Pitch

Passing Network and Average Pitch Positions for Passes Across the Pitch

R. Carvalho to R. Guerreiro (6 passes) was the most frequent lateral passing combination, with Carvalho playing lateral balls to Guerreiro on the left flank. This tactic stretched play horizontally, typical in build-up phases to open space and destabilize the opposition. Pepe to Vieirinha (4 passes) shows Pepe switching play to the right side, connecting with Vieirinha, positioned higher up the field. Pepe to C. Ronaldo (4 passes) reflects Ronaldo's involvement in lateral ball circulation, indicating his participation beyond attacking positions. D. Pereira to Pepe (4 passes) shows how midfielders engaged in lateral passing to help the defenders retain possession and switch play. Average pitch positions concentrated in the central midfield illustrate lateral passing's role in recycling possession and resetting play, emphasizing the full-backs' roles in advancing play from lateral balls.

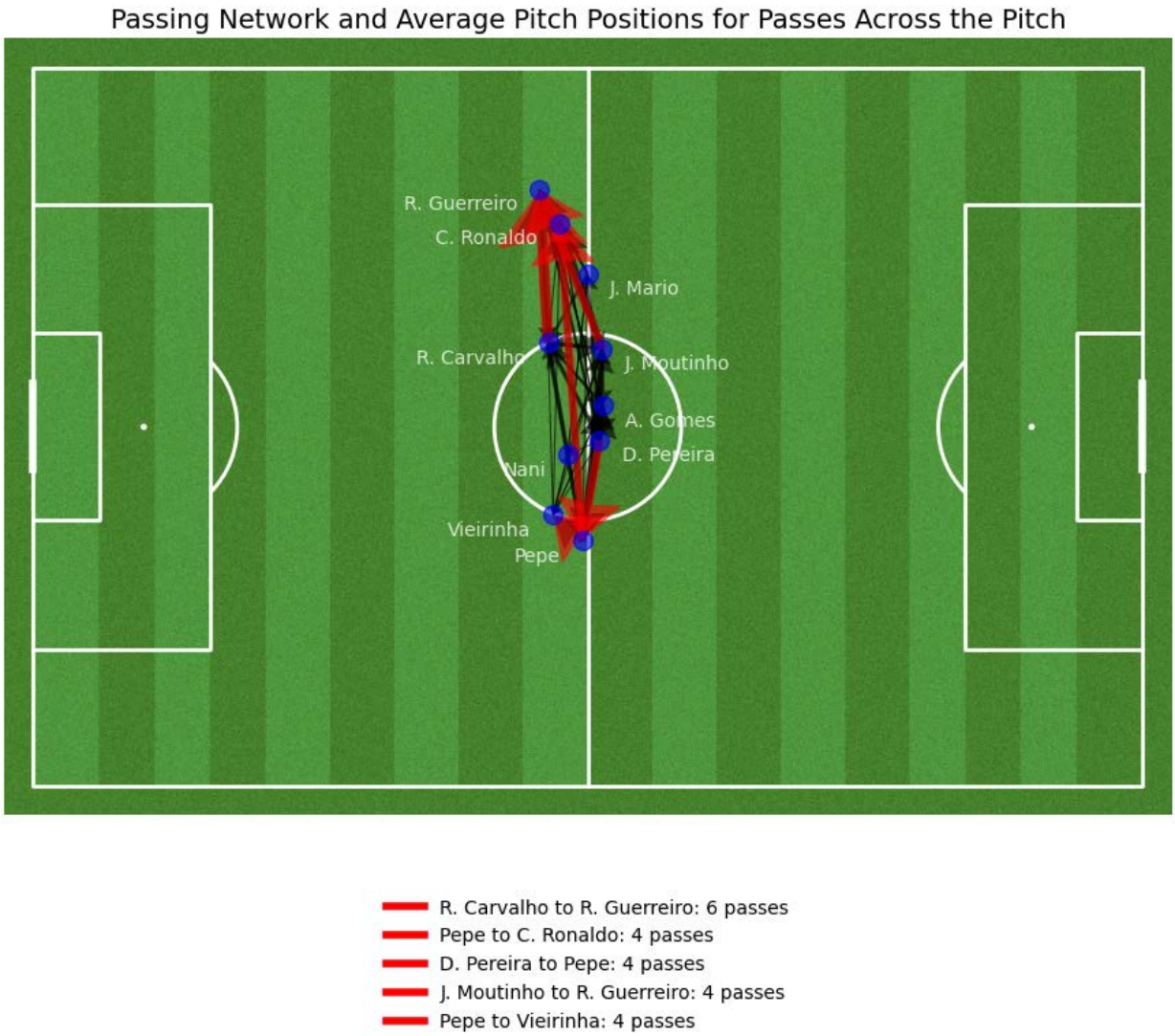


Figure 4.3 (1d) Passing Network and Average Pitch Positions for Passes Across the Pitch

Degree Centrality for Passes Across the Pitch

J. Moutinho had a large node in the network, reflecting his central role in lateral passing, maintaining control, and dictating the tempo. D. Pereira's involvement shows his importance in helping the team retain possession and switch play, especially as a defensive midfielder. Full-backs, with high degrees of centrality, were key in switching play from one side to the other. C. Ronaldo's frequent involvement in ball circulation, especially higher up the pitch, allowed him to receive lateral passes in attacking areas, facilitating play on the left flank.

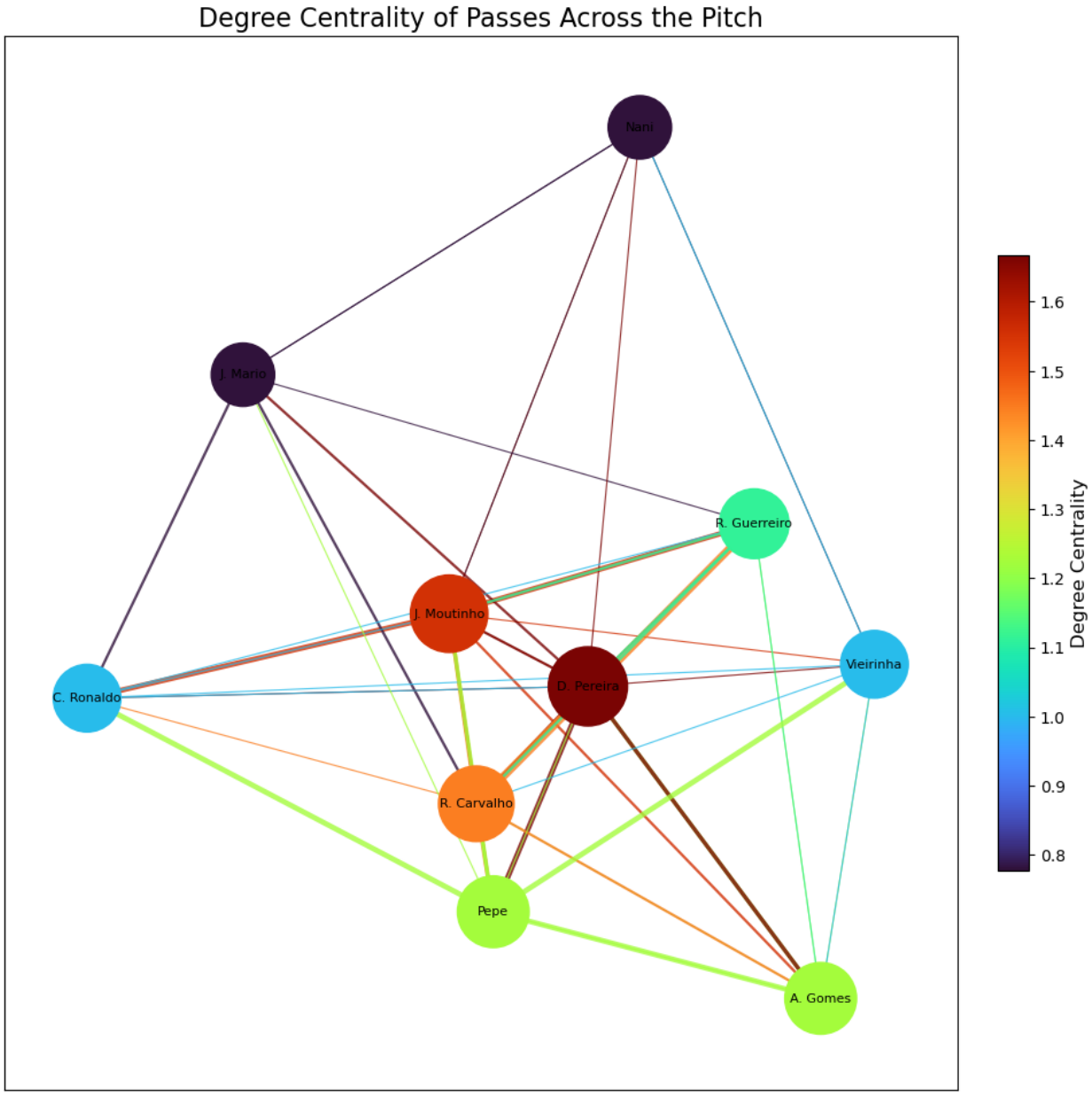


Figure 4.3 (1e) Degree Centrality of Passes Across the Pitch for Players

Pass Accuracy and Count by Pass Type

Hand passes had the highest accuracy as expected, since they are used by goalkeepers in low-pressure situations. Simple passes were also highly accurate, suggesting supporting a possession-based strategy that relied on short, safe passes. High passes and head passes showed moderate accuracy, due to their reliance on positioning and aerial duels. Nonetheless, the team performed well in aerial duels when needed. Crosses had the lowest accuracy since they are often contested by defenders, and face challenges due to the high number of players in the box. Launches also had low accuracy, reflecting their higher risk, as they are often used to switch play quickly but are prone to interception. Simple passes were the most common, with around 500 instances, indicating a preference for maintaining possession. Crosses, high passes, and head passes were less frequent but evenly distributed, suggesting they were used to break

down defenses in specific situations. Launches and hand passes were the least common, as launches are risky and hand passes are limited to goalkeepers. Smart passes, designed to break defensive lines, occurred regularly but were less common than simple passes, indicating a more creative passing strategy used selectively.

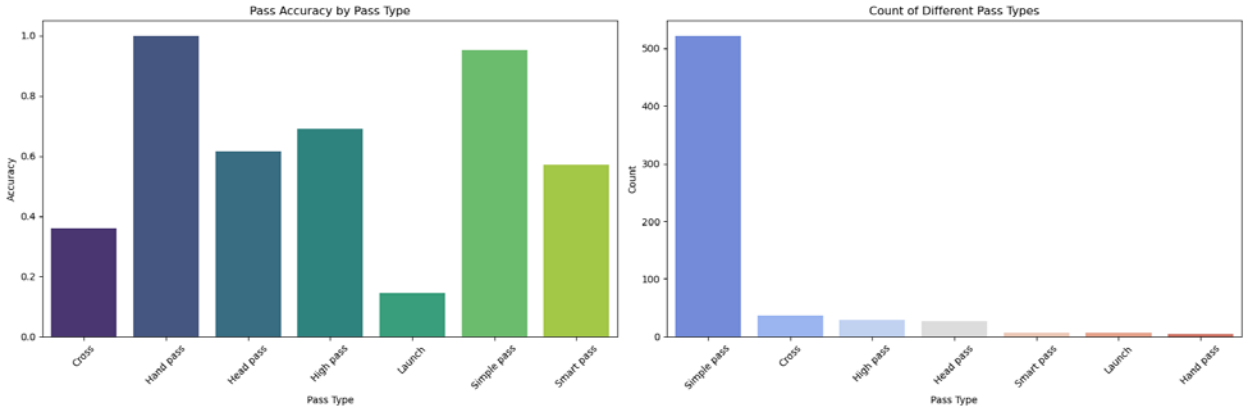


Figure 4.3(1f) Bar Plots of Pass Count and Accuracy

Pass Direction Distribution

Most passes were lateral, reflecting the team's focus on horizontal possession to stretch the opponent's defensive and recycle play. Forward passes, though less common, indicated careful progression when opportunities arose. Backward passes were the least common, showing the team's intent to move the ball forward or laterally, rather than retreat. Long passes were mostly lateral, used to switch play across the pitch rather than directly attacking. Forward long passes were rare, indicating Portugal used long passes to bypass the midfield sparingly. Backward long passes, the least frequent, further emphasizing the team's forward momentum. Considering across-the-pitch passes, the distribution was similar to the overall passing pattern, with most being lateral passes. Forward passes across the pitch were more frequent than long passes, indicating they occasionally served to advance play into dangerous areas. Backward passes across the pitch were also common, suggesting that when Portugal switched play, sometimes they chose to reset and maintain possession.

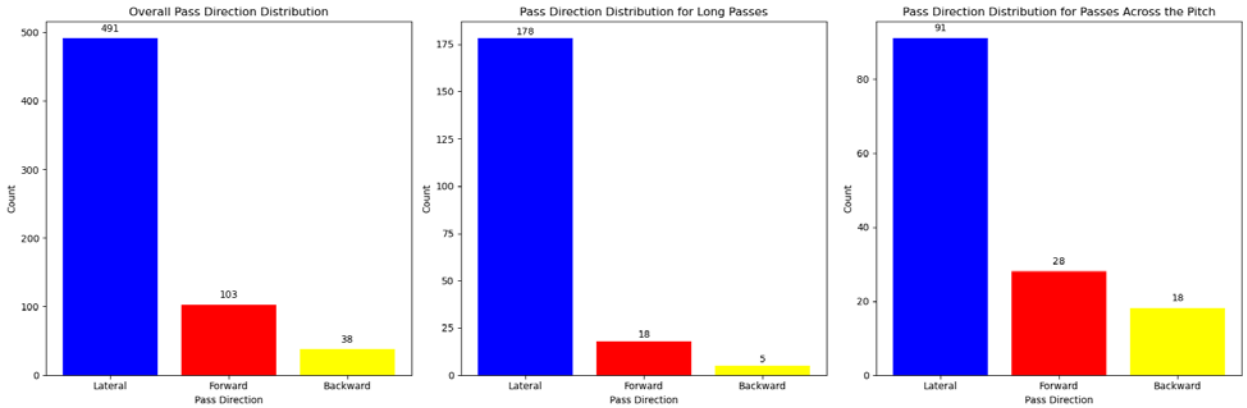


Figure 4.3(1g) Bar Plots of the Distribution of Pass Directions

Average Pass Length over Time

In the first half, pass length fluctuated, with significant dips and peaks, indicating alternating between shorter, controlled sequences, and longer direct passes. This variability suggests tactical adjustments to opposition pressing and defensive strategies. The second half was more consistent, particularly after the 70th minute, when the team relied more on longer passes, possibly to counteract fatigue or break through the defense. Peaks in pass lengths occurred at key moments, such as the 10th, 30th, 50th, and 70th minutes, indicating attempts to transition play with direct passes.

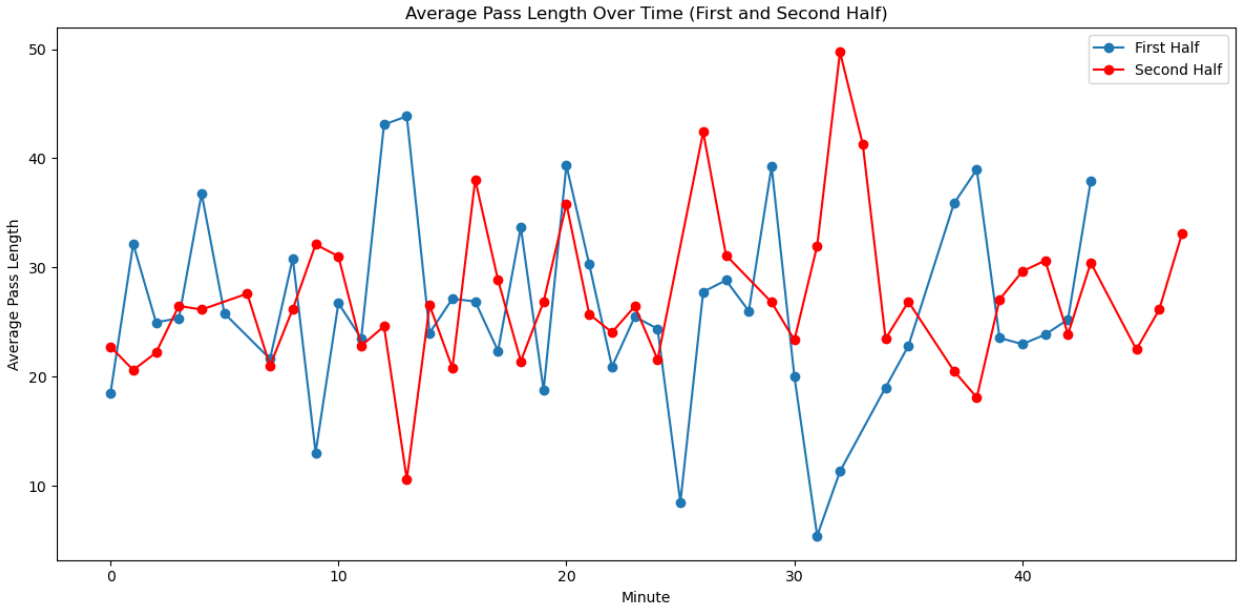


Figure 4.3(1h) Line Graph of Average Pass Length for Each Half

Rolling Average of Pass Accuracy Over Time

Pass accuracy started high in the first half, with moments close to 100% accuracy, suggesting strong possession and minimal pressure from the opponent early on. However, dips occurred after the 10th and around the 20th minute, likely due to increased pressure or riskier passing decisions. Accuracy recovered later in the half, showing improved ball retention. The second half followed a similar initial pattern but saw more fluctuations in accuracy as the game progressed. Accuracy didn't stabilize as it had in the first half, indicating fatigue, or tactical adjustments that introduced more risk in the search for scoring opportunities.

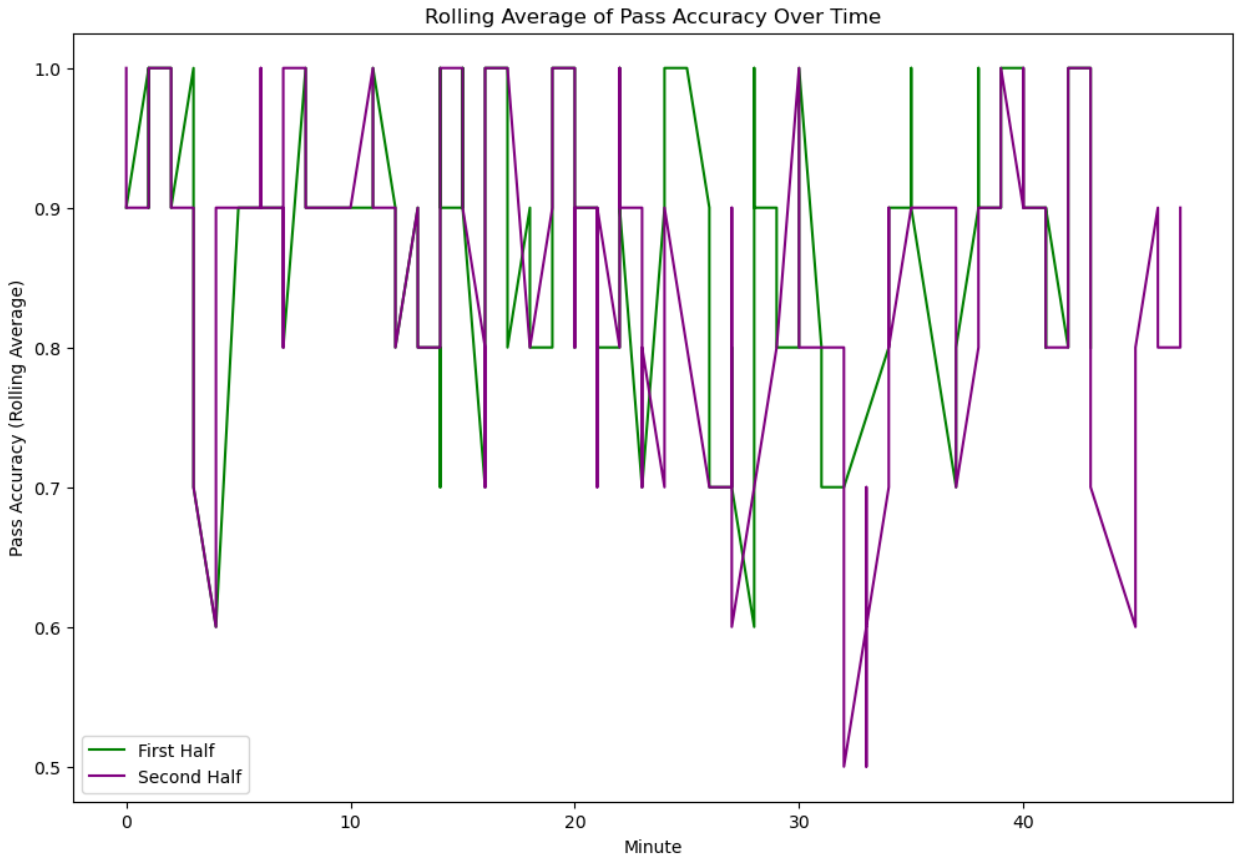


Figure 4.3(1i) Line Graph of the Rolling Average of Pass Accuracy for Each Half

Distribution of Possession Lengths

The distribution was right-skewed, with most possession sequences having fewer than 10 events. This suggests that the team often lost possession quickly or used shorter, controlled sequences before advancing or losing the ball. The most common possession length was between 5-7 events, indicating a preference for quick transitions or frequent turnovers. Longer possession sequences were rare but occurred during periods of sustained control, likely when the team was managing the game or building up play methodically. A few extremely long sequences, up to 40-50 events, indicate moments of complete possession dominance, likely used to protect a lead or break down the opposition's defense.

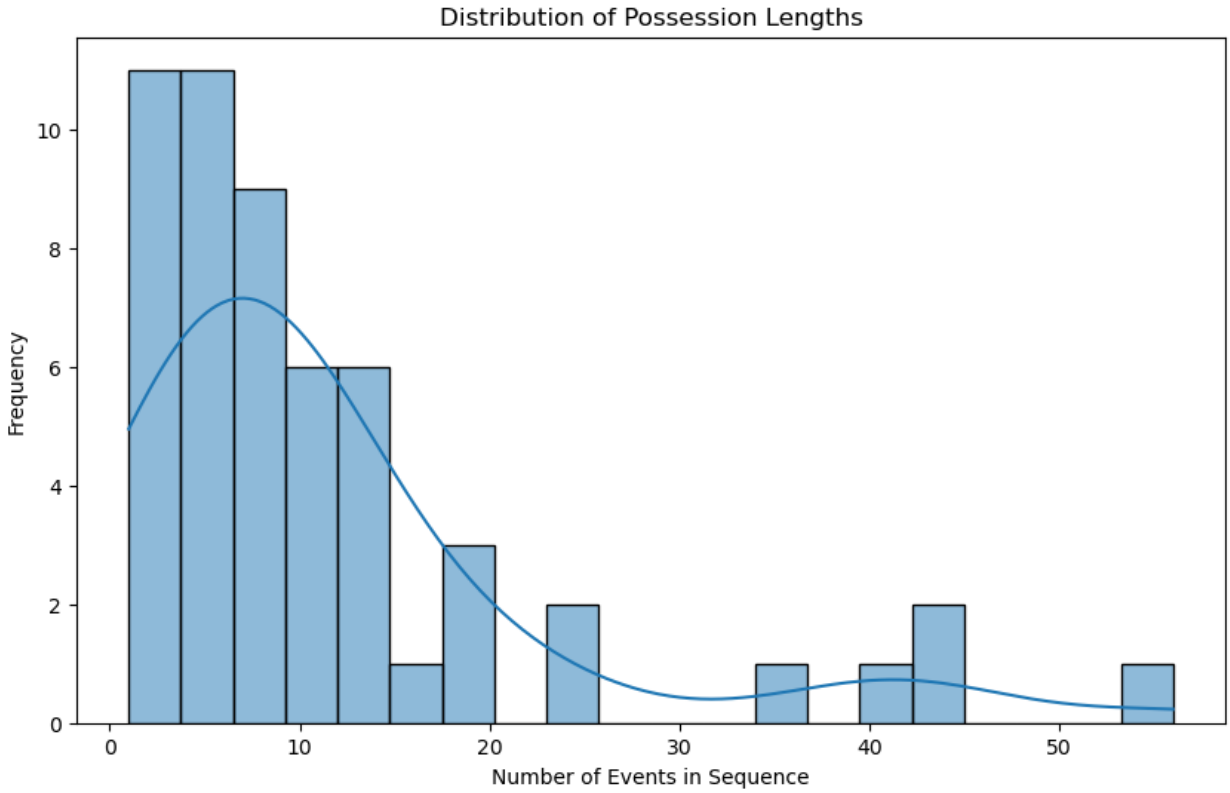


Figure 4.3(1j) Bar Plot of the Distribution of Possession Lengths

4.4 Topic Modeling and Event Sequence Dynamics Analysis

Medoid Representative Episode for Passes

The medoid visualization highlights the varied direction of passes, with Portugal utilizing multiple passing routes across the pitch, including lateral, forward, and backward movements. This variety indicates the team's comfort with recycling possession, switching play, and distributing the ball across different field zones. There is significant activity in central areas, with midfielders heavily involved in linking play and controlling possession. This reflects Portugal's strategy of dominating central areas of the pitch, for maintaining possession and controlling the tempo. Notable passes toward the flanks, particularly on the right side, suggest the team aimed to stretch the defense by switching from central to wide areas, a common tactic to create space and crossing opportunities. Many passes from wide areas target the box, showing the team's intent to create chances by playing the ball into dangerous areas. Several passes also originate from deep defensive areas, progressing toward the opposing half, indicating a clear strategy to advance play forward. Long passes toward the flanks and vertical passes through the center highlight moments of quick transitions from defense to attack. The clustering of passes in the final third near the opposition's penalty box suggests that the team frequently advanced into attacking areas, leading to goal-scoring opportunities.

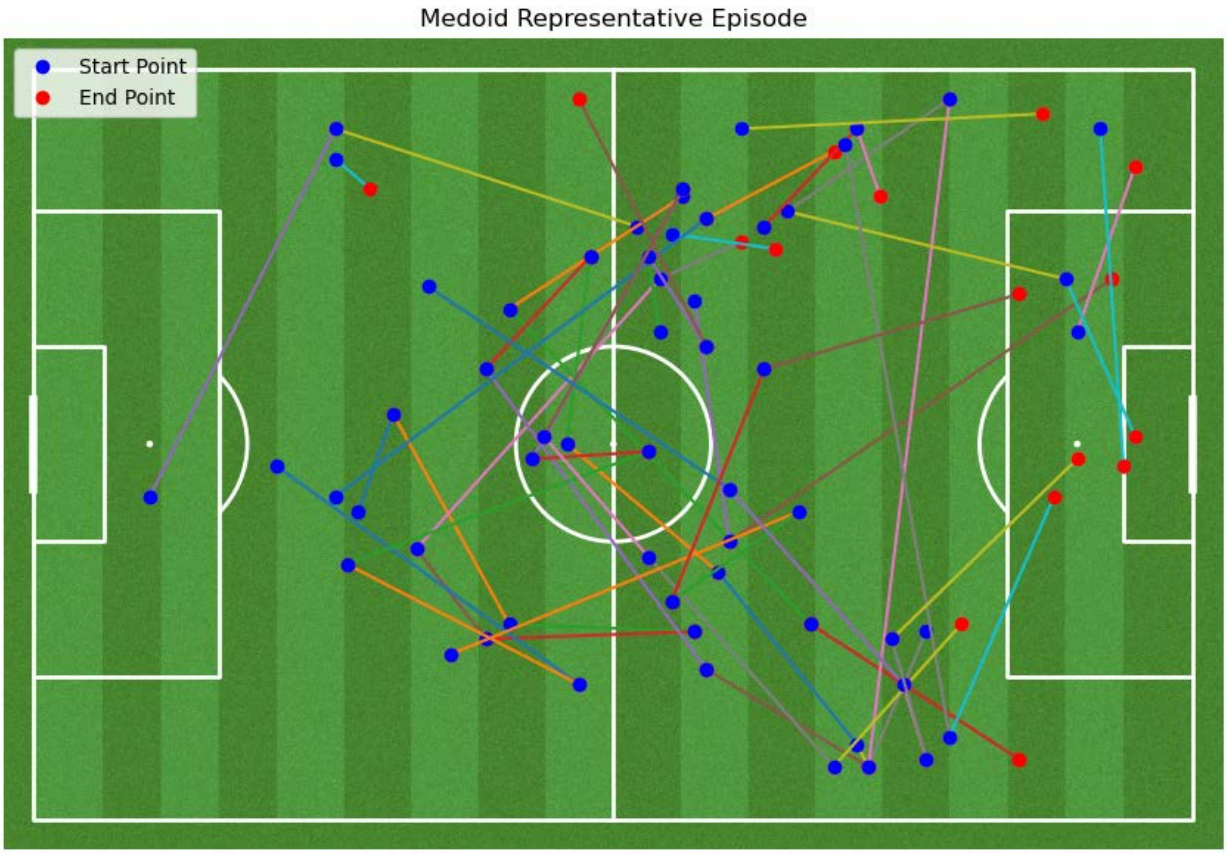


Figure 4.4(1a) Pitch Plot for Medoid Representative Topic Modeling Episode for Passes

Voronoi Diagram of Player Influence Areas

Pepe and R. Carvalho dominate the defensive zones, with Pepe covering a slightly larger area, reinforcing their key roles in the team's defensive structure. R. Guerreiro and Vieirinha control the flanks, reflecting their roles as full-backs responsible for defending and supporting attacks. J. Moutinho and D. Pereira influence the midfield areas, with Moutinho playing a more expansive role, while Pereira focuses more on defensive responsibilities. C. Ronaldo, A. Gomes, and Nani dominate the forward areas, with Ronaldo in the most advanced position, highlighting his role in creating and finishing attacking opportunities. Ronaldo's positioning suggests the team relies on him to receive the ball in high-danger areas.

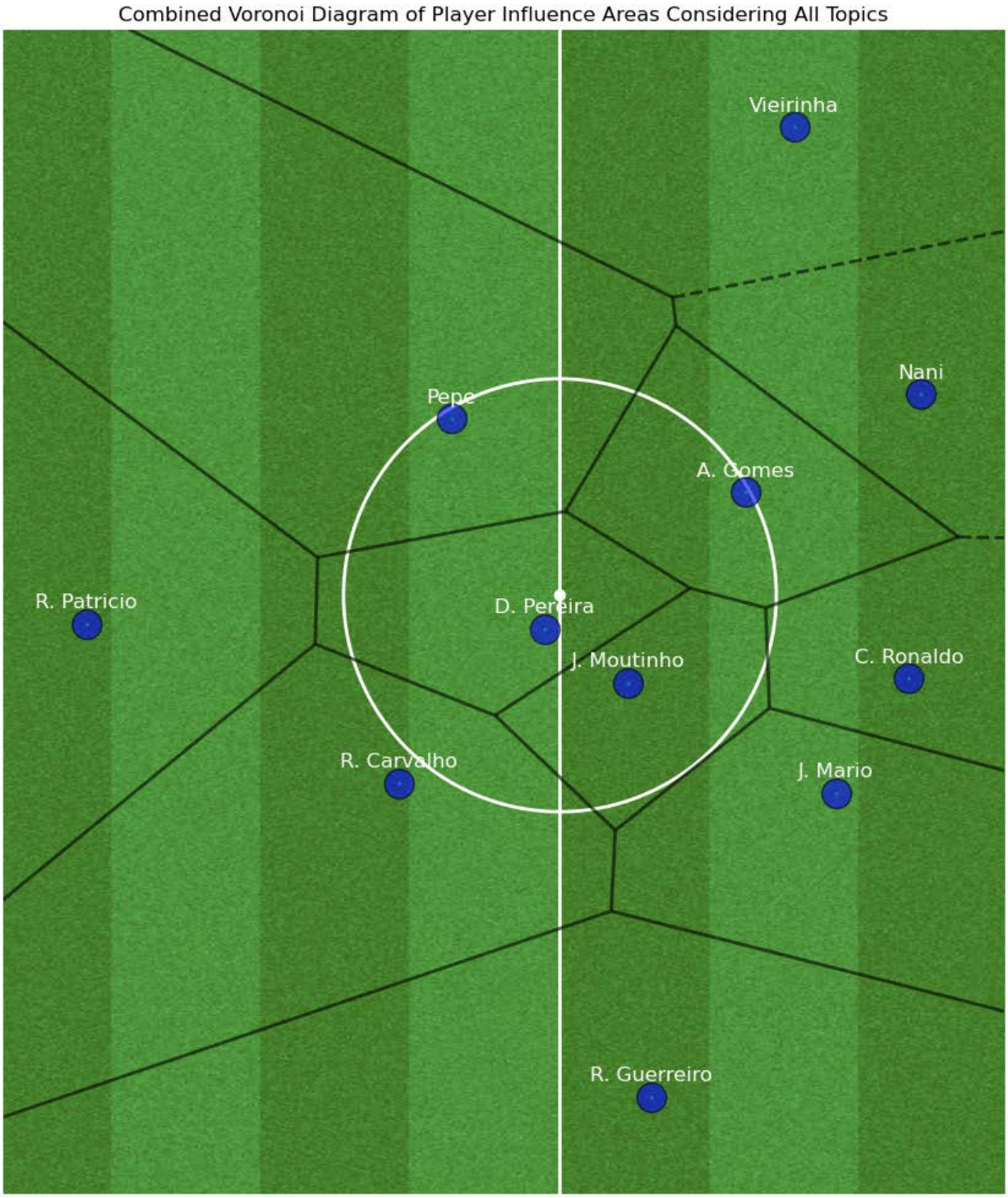


Figure 4.4 (1b) Voronoi Diagram of Player Influence Areas Considering Topic Modeling

Radar Chart for Players

C. Ronaldo's average pass length is relatively low, reflecting his role as a forward who engages in shorter passes rather than long-range distributions. His movement is moderate, aligned with his focus on receiving the ball in goal-scoring positions rather than build-up play. Ronaldo's speed of play is relatively

low, suggesting he is more deliberate in his actions, often waiting for the right moment to strike. His high pass completion rate reflects his success in short-passing situations, while his high vertical advancement rate shows his ability to push the team forward. His horizontal spread rate is low, indicating that Ronaldo rarely makes lateral passes, fitting his role in advancing vertically. J. Moutinho's average pass length is higher than Ronaldo's as he is involved in short and long passes. His high involvement highlights his role as a midfielder orchestrating play. Moutinho balances controlling possession with quicker, traditional build-up play, as reflected in his moderate speed of play. He has the highest pass completion rate, underscoring his reliability in maintaining possession. Moutinho's vertical advancement rate is moderate, as he is responsible for moving the ball forward and controlling the width by switching play. His role in controlling possession across the width of the field results in a higher horizontal spread rate than Ronaldo. R, Guerreiro's average pass length is the shortest, fitting his role as a full-back engaging in shorter passes, especially in wide areas. His significant player involvement shows his frequent connection between defense and attack, linking with midfielders and wingers. Guerreiro's high speed of play reflects his quick transitions, often moving the ball forward rapidly during counterattacks. His high pass completion rate indicates his reliability in defensive and offensive phases, ensuring possession is maintained. While his vertical advancement rate is moderate, reflecting his role in retaining possession and supporting teammates, his horizontal spread rate is the highest, emphasizing his responsibility for switching play and stretching the opposition across the field.

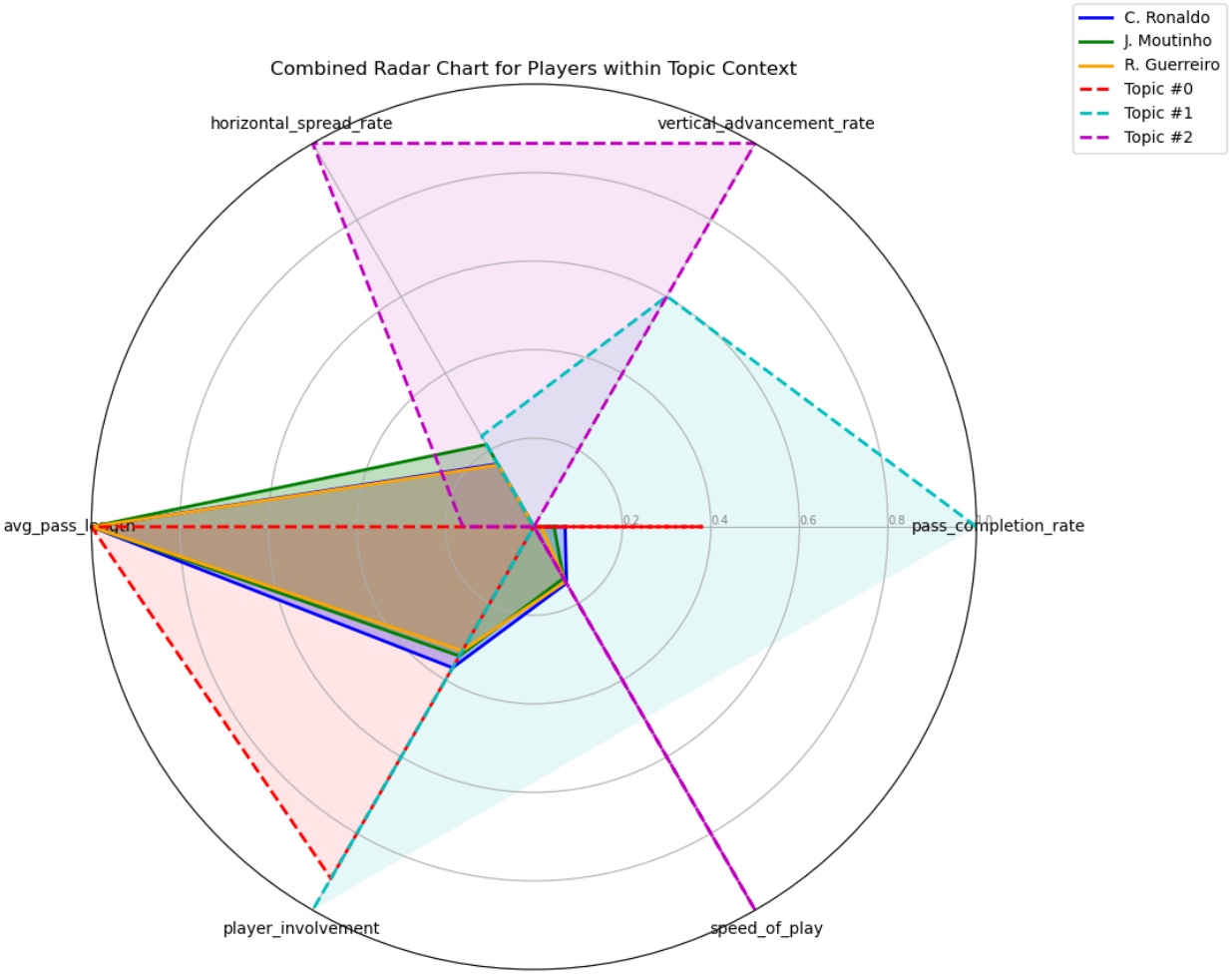


Figure 4.4(1c) Radar Chart of Players Considering Topic Modeling

Pass Flow Maps for the Starting Lineup

In the pass flow maps for topic zero, the highest pass intensity is concentrated in the center-right and central zones, indicating that Portugal primarily used these areas for building up play. Forward passing is notable in central, with upward arrows showing Portugal's focus on vertical progression through midfielders. The defensive zones show lower pass intensity, suggesting a forward-moving focus rather than circulating the ball in their half. The activity on the flanks, especially on the left, is moderate, reflecting that the team relied more on central channels than wide play. In topic one, high pass intensity is visible in both left and right midfield zones, emphasizing wide play to stretch the opposition. Lateral passing in the defensive third increases, showing the team recycled possession in their half, possibly to draw the opponent out and create space. Central midfield pass intensity is moderate, reflecting a shift toward exploiting the pitch width. Forward passes are evenly distributed across the field, indicating a more balanced passing approach. In the second topic, the right-sided zone sees increased activity, suggesting the team favored the right side for advancing play, possibly for counterattacking. More lateral passes are observed in the defensive zones, reflecting a slower build-up from the back. Portugal focused on retaining possession, to draw the opposition forward before attempting to break through. Flank usage remains moderate, with lateral and forward arrows showing balanced possession.

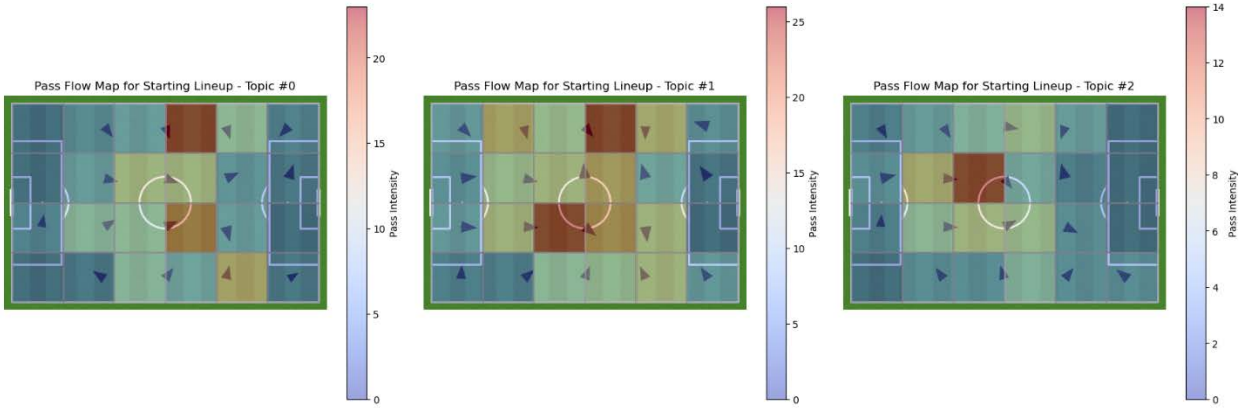


Figure 4.4(1d) Pass Flow Maps of Each Topic for Players in the Starting Lineup

Transition Matrix of Event Types

There is a 71% chance of consecutive crosses, suggesting the team often attempted multiple crosses during attacking phases, possibly after an initial clearance or missed opportunity. Simple passes follow each other 83% of the time, demonstrating a possession-based strategy with low-risk passes to maintain control. High passes follow each other 82% of the time, showing that when opting for long-range play Portugal continues this approach to cover more ground. A 20% chance exists that a head pass transitions into a high pass, indicating that aerial duels often lead to long, lofted passes to clear the ball or initiate counter-attacks. Launches and smart passes are isolated, reflecting their use in high-risk situations, to clear the ball under pressure or attempt line-breaking passes.

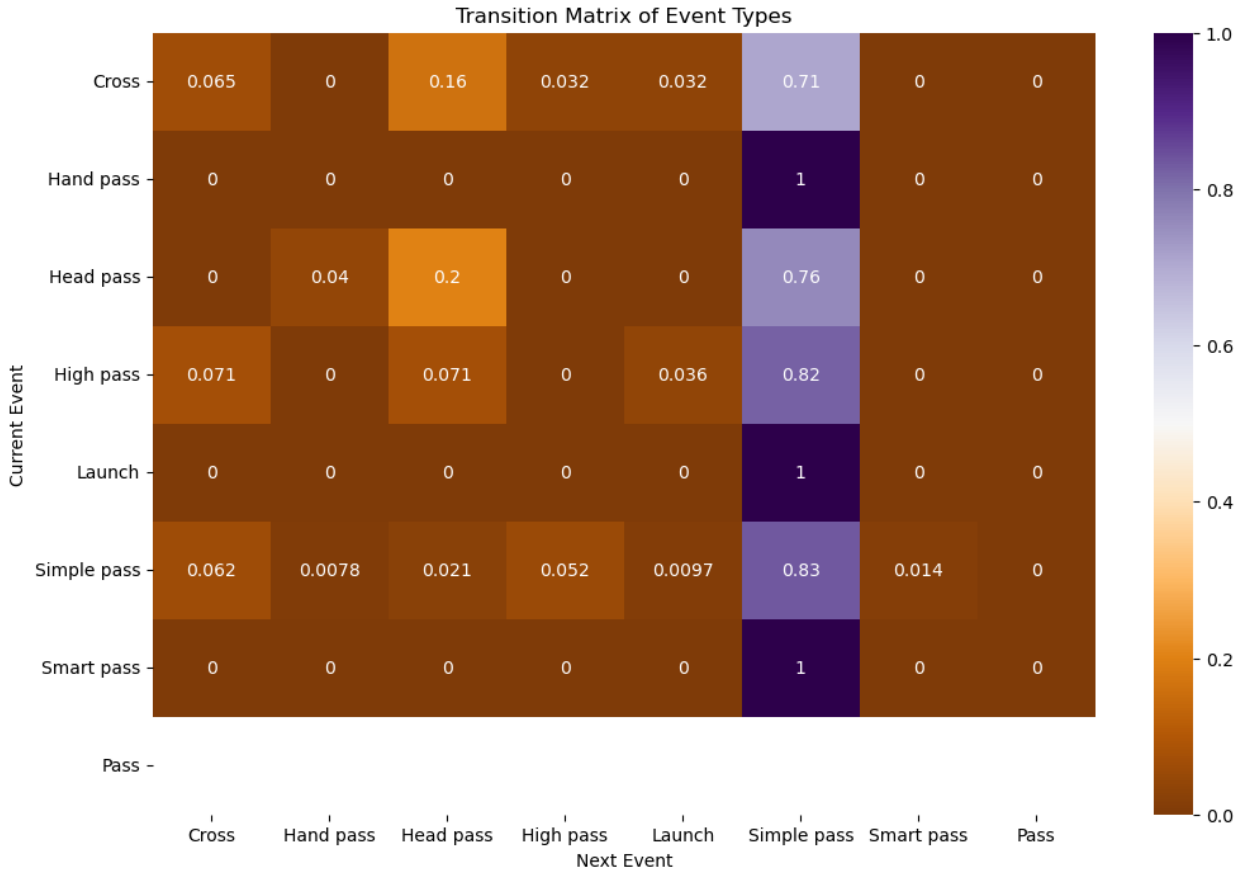


Figure 4.4(1d) Heatmap Representing a Transition Matrix of Event Types

Event Density Map

Zone X=6, Zone Y=2 (right-midfield) shows the highest event density, indicating that much of the team's play occurred here, particularly in passing sequences. This reflects a focus on building play through the right-central channel. Zone X=6, and Zone Y=7 (right flank, final third) present high density, indicating frequent offensive activity, likely related to crossing or attacking movements near the opponent's box. Zone X=4, Zone Y=6 (left central midfield) exhibits a relatively high density, reflecting the team's ability to switch between left and right flanks for balance in attack. Moderate densities in the central midfield zones suggest that Portugal maintained possession but didn't spend as much time here as on the right flank. Zone X=0, Zone Y=7, and Zone X=0, Zone Y=8 (left defensive area) present near-zero density, indicating little activity, possibly due to successful possession dominance and defensive stability.

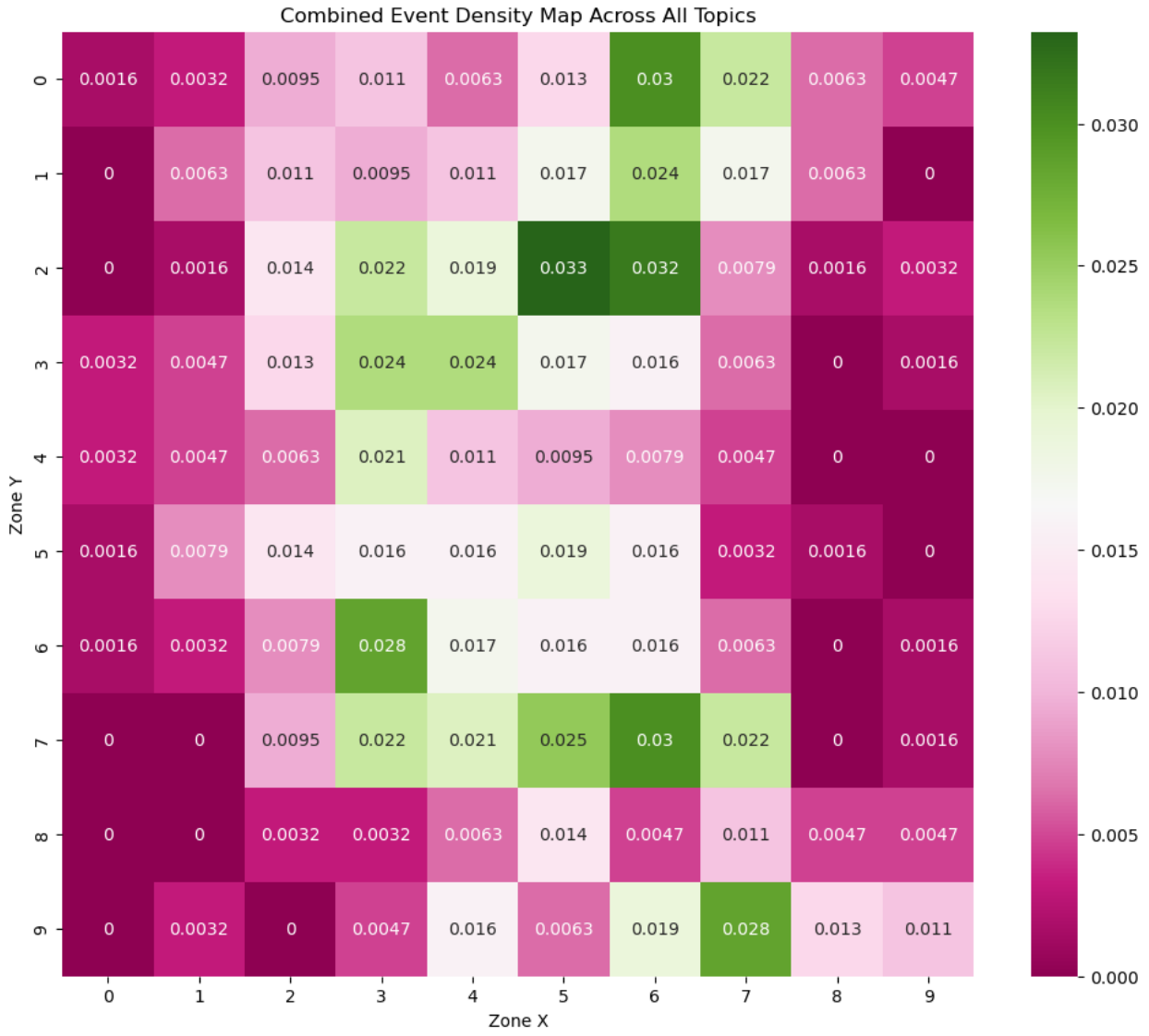


Figure 4.4(1e) Heatmap Representing Event Density Considering Topic Modeling

Passing Network and Player Involvement

In topic zero, R. Patricio, R. Carvalho, and A. Gomes are central figures in the passing network, with thick arrows showing frequent passing interactions. These players are key facilitators in maintaining possession and dictating play. Nani and C. Ronaldo are involved but appear as final third outlets rather than orchestrators. The player involvement chart shows that A. Gomes had the highest involvement count, highlighting his active role in circulating the ball. In topic one, A. Gomes remains prominent, continuing as a central facilitator. His frequent connections to Pepe, R. Carvalho, and C. Ronaldo suggest he remains key to ball circulation, though there are fever overall connections, reflecting a more focused passing strategy. Moutinho and Guerreiro also stand out, contributing to ball movement in midfield and wide areas, possibly indicating increased flank usage. A. Gomes continues to lead in involvement, but the gap between him and players like R. Carvalho and Vieirinha widens. In topic 2, Pepe, Moutinho, and Guerreiro are prominent in maintaining possession and dictating play. Pepe's significant involvement reflects a possession-focused strategy starting from the back. C. Ronaldo and Nani remain active but are

less central to the network, serving more as outlets for forward progression. J. Mario has the highest interaction count in this phase, suggesting he is the key orchestrator, acting as a primary hub for possession.

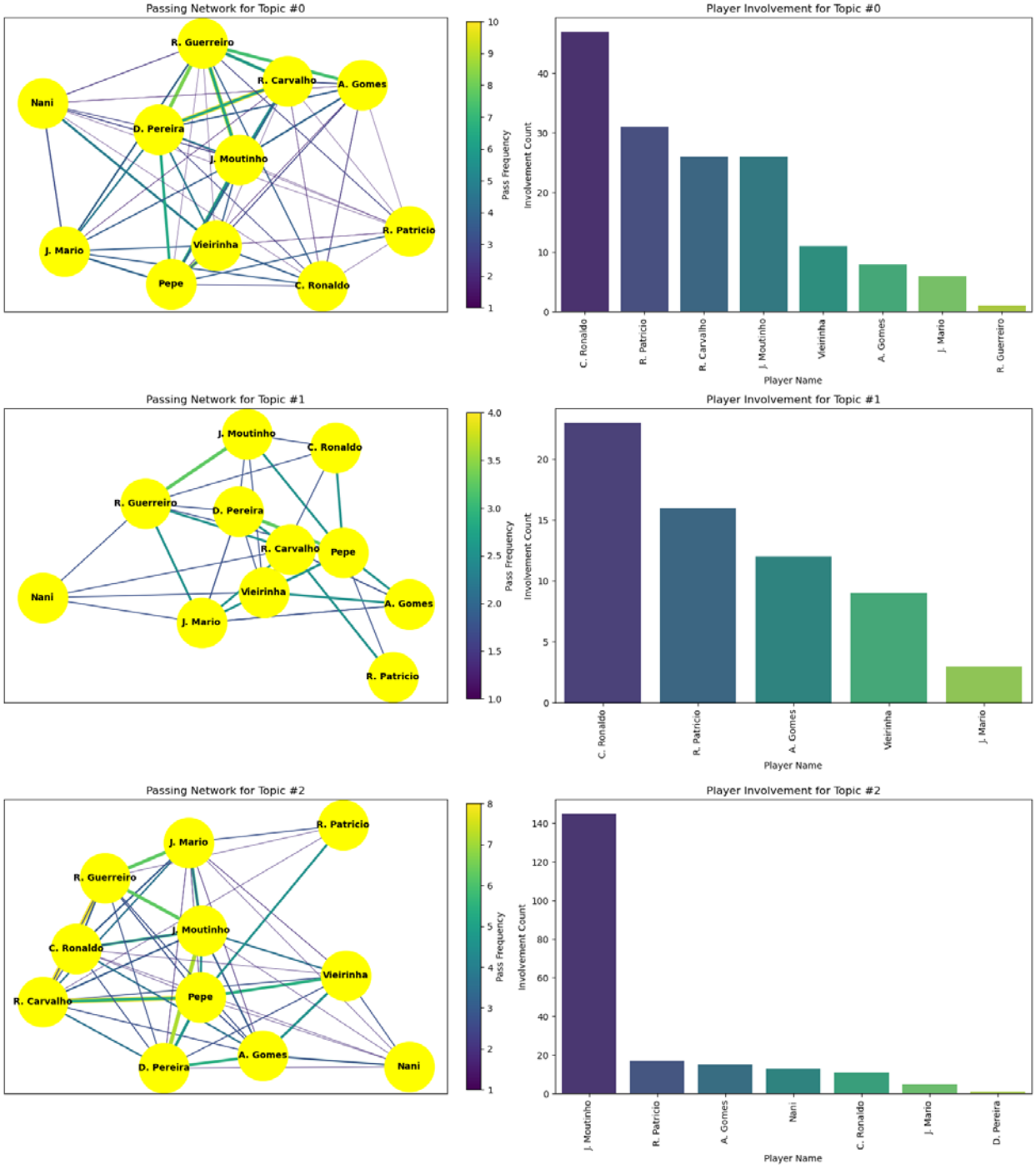


Figure 4.4(1f) Passing Network and Bar Plot of Player Involvement Considering Topic Modeling

4.5 Pass Features and Predictive Outcomes Analysis

PCA of Pass Features by Time Interval

Pass features cluster tightly in the upper left quadrant, particularly during initial intervals, in the first half. As the half progresses pass features start scattering farther along the first principal component axis, indicating increased variability in passing styles in the later stages. This dispersion suggests Portugal may have adopted more direct or varied passing strategies regarding match dynamics. The second half shows a broader spread of points across both principal component axes, indicating greater variability in pass features, particularly from the 30th minute onward. This may reflect tactical adjustments, increased opponent pressure, or fatigue, leading to more diverse passing patterns as the match enters critical phases. suggest that the team becomes less predictable or shifts its playing style in critical phases.

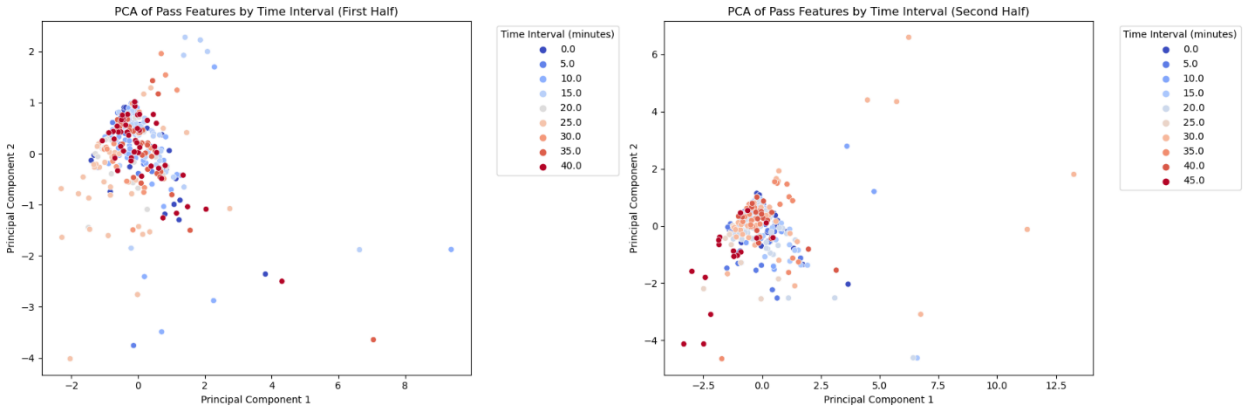
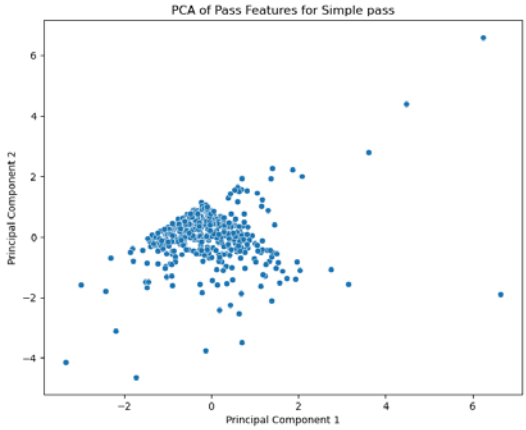
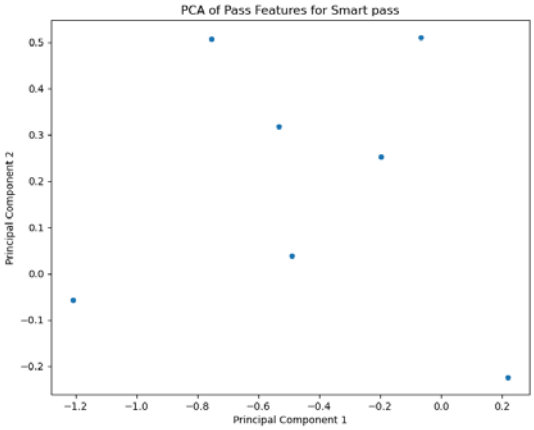


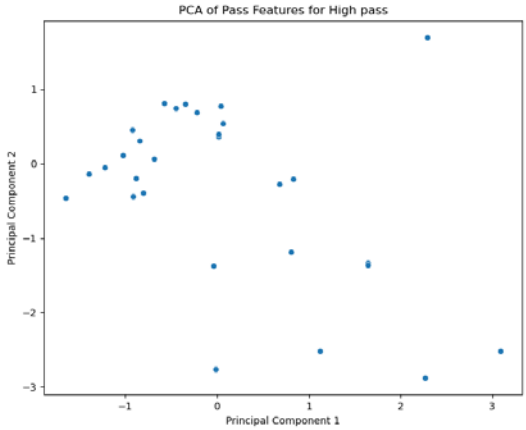
Figure 4.5(1a) PCA Scatter Plot of Pass Features Over Time

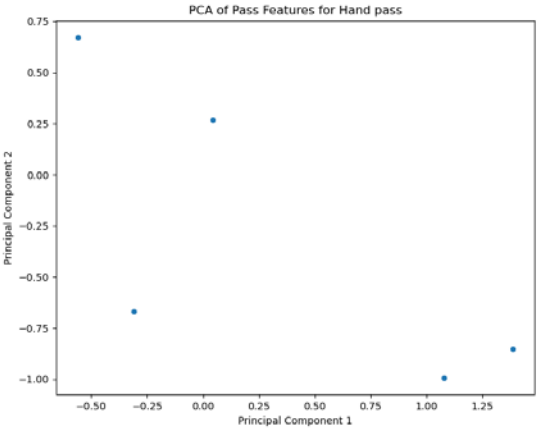
PCA of Pass Features for Different Pass Types

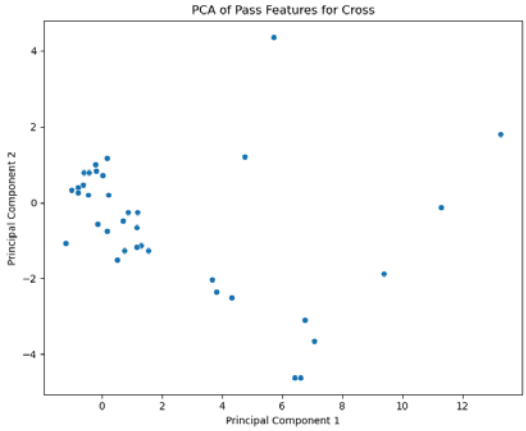
The PCA for simple passes is tightly clustered, showing low variability. This reflects the consistency of simple passes, often short and used to maintain possession in controlled situations. Smart passes, designed to break defensive lines, show more spread across the principal components, indicating greater variability due to the complexity and risk involved in executing these passes. High passes exhibit moderate dispersion, reflecting their occasional use and variability in execution, as they are often employed to bypass midfield or reach attackers quickly players. Hand passes, being rare and used in specialized situations like throw-ins, have smaller points with low variance in the PCA space, as expected. Head passes show moderate variability, similar to high passes since they are often used in contested aerial situations, adding complexity to their execution. Crosses display significant variability, with a wide spread of points, aligning with the unpredictable nature of crosses, performed under varying defensive pressure and from different field positions. Launches, long and direct passes aimed at transitioning quickly from defense to attack, also show widespread, reflecting variability in their execution considering match situations.

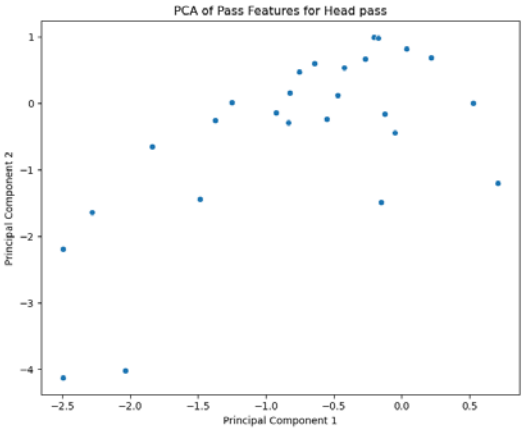

PCA of Pass Features for Simple pass
Principal Component 2
Principal Component 1


PCA of Pass Features for Smart pass
Principal Component 2
Principal Component 1


PCA of Pass Features for High pass
Principal Component 2
Principal Component 1


PCA of Pass Features for Hand pass
Principal Component 2
Principal Component 1


PCA of Pass Features for Cross
Principal Component 2
Principal Component 1


PCA of Pass Features for Head pass
Principal Component 2
Principal Component 1

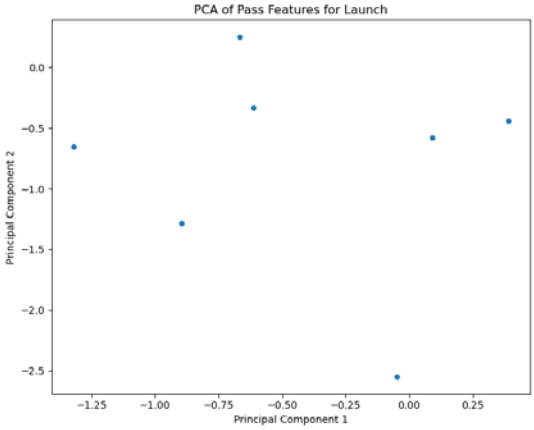

PCA of Pass Features for Launch
Principal Component 2
Principal Component 1

Figure 4.5(1b) PCA Scatter Plot of Pass Types

Pass Accuracy Prediction Based on Passing Features

In the logistic regression model, the confusion matrix shows that true positives (bottom right) are cases where the model correctly predicts an inaccurate pass, and true negatives (top left), are cases where the model correctly identifies accurate passes. False positives (top right) are instances where the model incorrectly predicts accurate passes, while false negatives (bottom left) represent cases where the model fails to identify accurate passes. Logistic regression performs reasonably well, with a 75% accuracy rate, but struggles more with false negatives, meaning it underestimates accurate passes. The random forest follows the same classification for true positives, negatives, and errors, and performs better overall. It has fewer false negatives, indicating better identification of accurate passes, but slightly more false positives. This means the model occasionally over-predicts accuracy. The random forest achieves an 89% accuracy rate, performing better than logistic regression by capturing complex interactions between pass features.

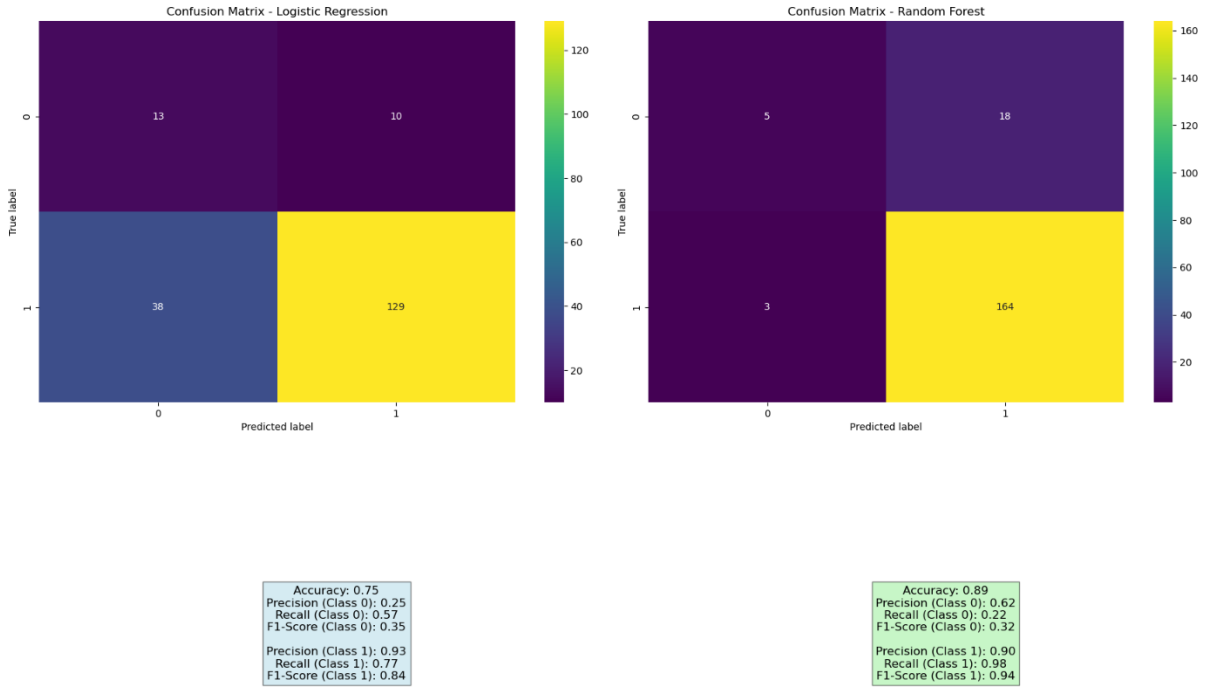


Figure 4.5(1c) Confusion Matrices and Metrics of Machine Learning Models Predicting Pass Accuracy

4.6 Duel Types and Outcomes Analysis

Average Positioning of Duel Types

The visualization of aerial duels shows the positioning and frequency of contested and successful aerial duels for each player. C. Ronaldo is positioned primarily in advanced areas, focusing on attacking aerial duels. In contrast, defenders like Pepe and R. Carvalho, are involved in numerous aerial duels in their half, as expected from their defensive responsibilities. This provides insights into the team's tactical approach, showing how key players like Ronaldo contribute offensively while defenders work to secure

aerial dominance in defense. Considering ground loose ball duels, the visualization highlights where players frequently engage to regain possession. Players like Pepe and R. Carvalho contest duels primarily in the defensive third, while midfielders like D. Pereira are active in central areas. The arrows and colors represent the players who engage in and win ground duels, offering a clear view of offensive and defensive contributions. Central midfielders like J. Moutinho play a crucial role in transitions, often regaining possession and helping Portugal move from defense to attack. Ground-attacking duels focus on players working to break through defenses. Players like Ronaldo and Moutinho are active in forward duels, with arrows illustrating their movements and challenges. This visualization shows how players in advanced positions contribute to building attacks through direct duels with defenders, adding depth to the understanding of attacking patterns and player contributions. For ground-defending duels, the visualization reveals defensive duels aimed at stopping opposing attacks. Most defensive challenges occur near the central or defensive third of the field, where players like R. Carvalho and Pepe are positioned deep, reflecting their role in contesting duels near their goal to prevent dangerous attacks. This tactical positioning during defensive actions highlights how key players contribute defensively.

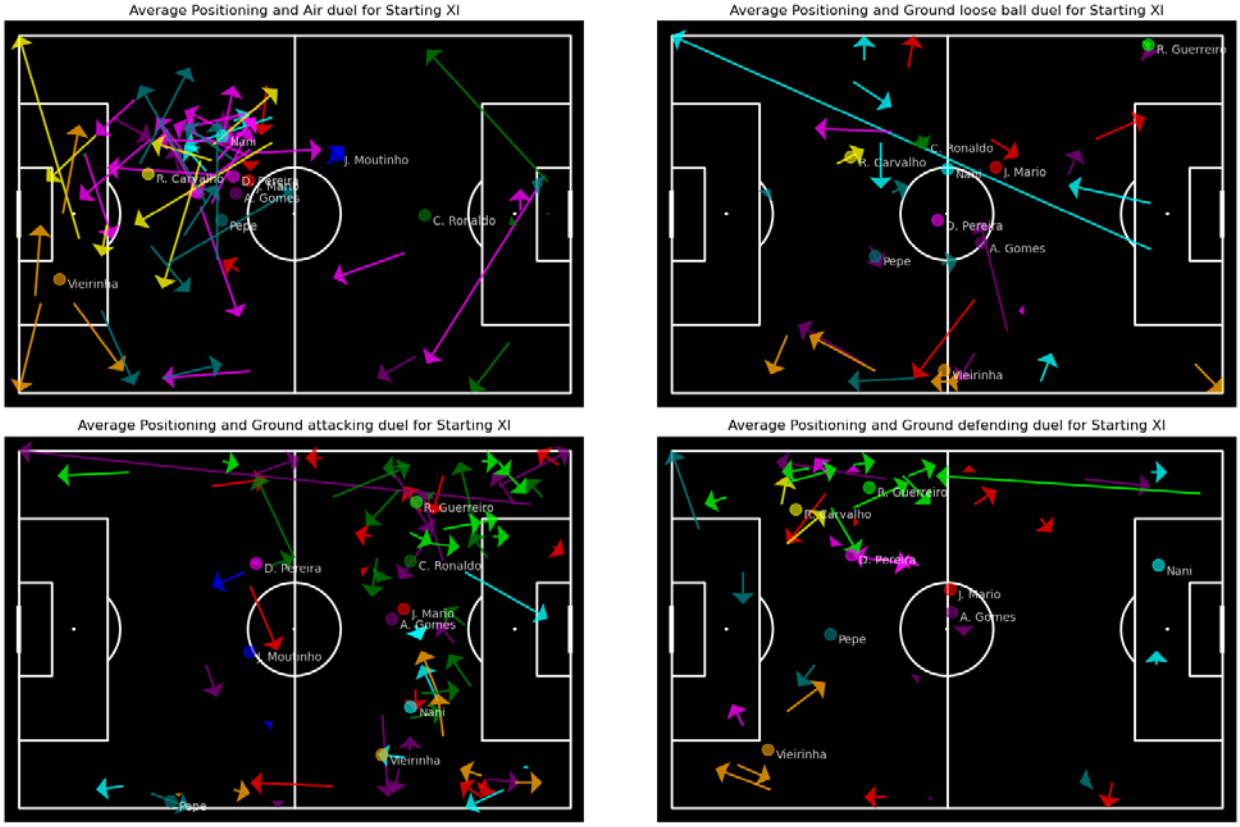


Figure 4.6(1a) Pitch Plots of Duel Types and Average Positioning of Players in the Starting Lineup

Average Positioning and Duels

The dense cluster of arrows around D. Pereira and J. Moutinho in the midfield indicates that many duels occurred in this region, consistent with the importance of midfield control for ball distribution and retrieval. Ronaldo and Nani, positioned more advanced, engage in fewer but significant duels in line with their attacking roles. Defensively, Pepe and Carvalho show arrows around the penalty area, representing their involvement in defensive and aerial duels. The diagram offers a high-level overview of where duels occur and which players are most involved, visualizing Portugal's activity and connections on the pitch.

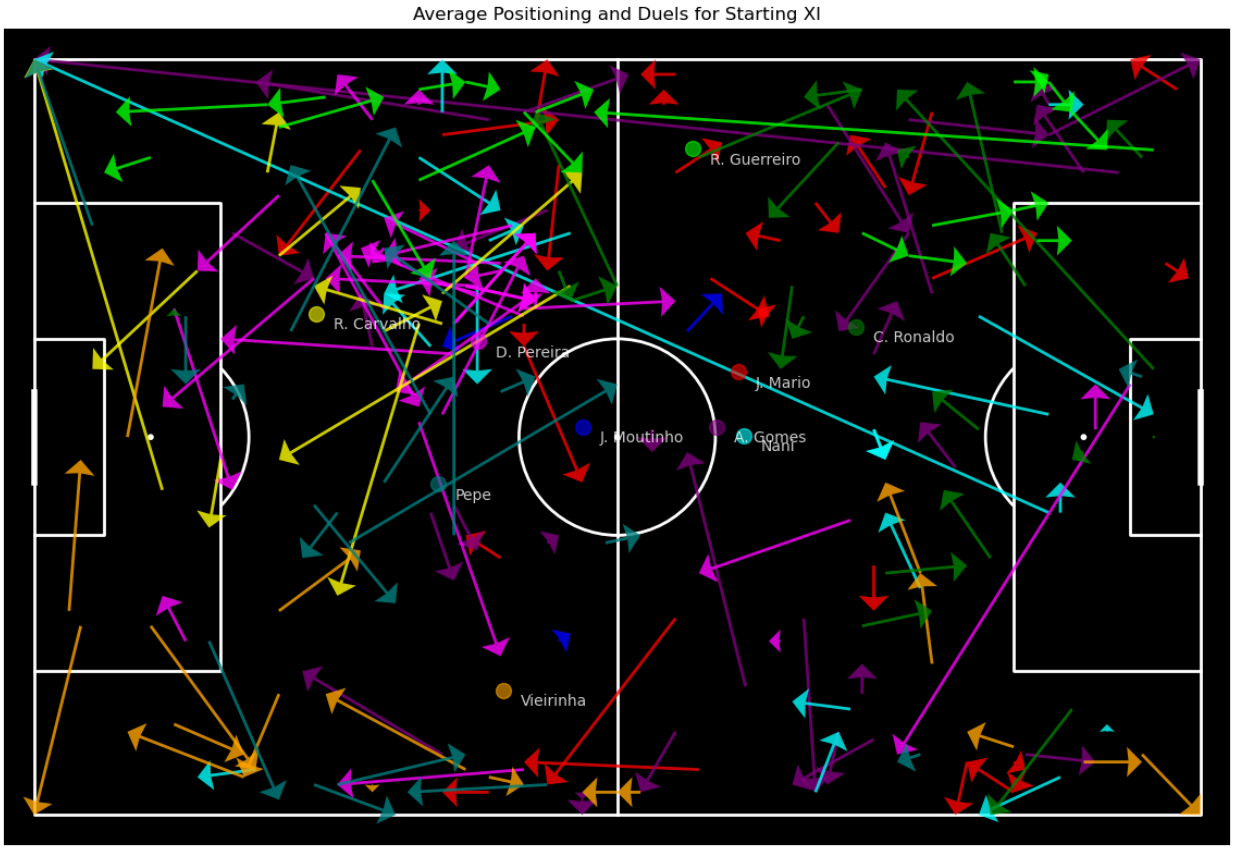


Figure 4.6(1b) Pitch Plot of all Duels and Average Positioning of Players in the Starting Lineup

Duel Positions for Individual Players

The individual player duel positions provide a more detailed view of where duels start and end for each player. J. Moutinho's duels are concentrated in the midfield, highlighting his role in transitions, ball recovery, and distribution. J. Mario and A. Gomes show a more balanced spread of duel positions across the pitch, reflecting their box-to-box play style, with many duels ending in advanced areas, indicating offensive contributions. Ronaldo's duels are mainly concentrated in the attacking third, emphasizing his role as a forward in breaking down defenses. R. Guerreiro and Vieirinha are primarily involved in duels along the flanks, consistent with their roles as wide players, supporting offensive and defensive efforts on the wings. The spread of duel positions across the pitch for individual players highlights each player's spatial responsibilities and contribution to the team's overall strategy. This detailed analysis offers a deeper understanding of tactical assignments and team dynamics.

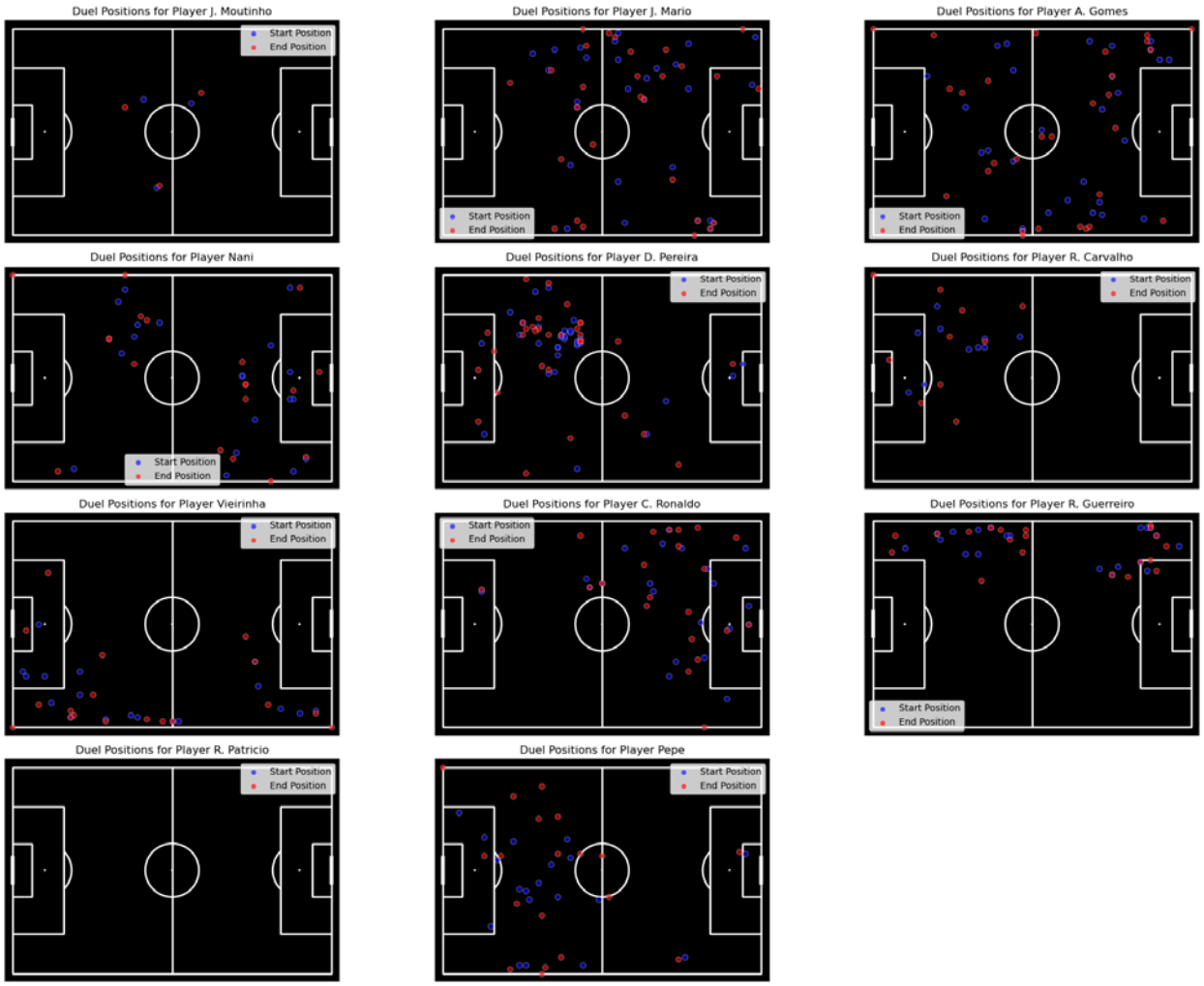


Figure 4.6(1c) Pitch Plots of Duel Positions for Individual Players

Win and Loss Rates by Duel Type

Pepe exhibits an exceptionally high win rate in air duels (90%), indicated by the intense red shading in the visualization. His dominance in aerial duels was essential for defensive stability, particularly when dealing with long balls and set-pieces. Ronaldo has a perfect win rate in ground loose ball duels, reflecting his effectiveness in recovering possession during loose ball situations and contributing to Portugal's offensive pressure. On the other hand, players like Moutinho and Ronaldo show low or non-existent win rates in ground defending duels, suggesting they were less effective in defensive scenarios. This provides insights into their positioning and roles during defensive transitions, showing a more offensive focus for these players.

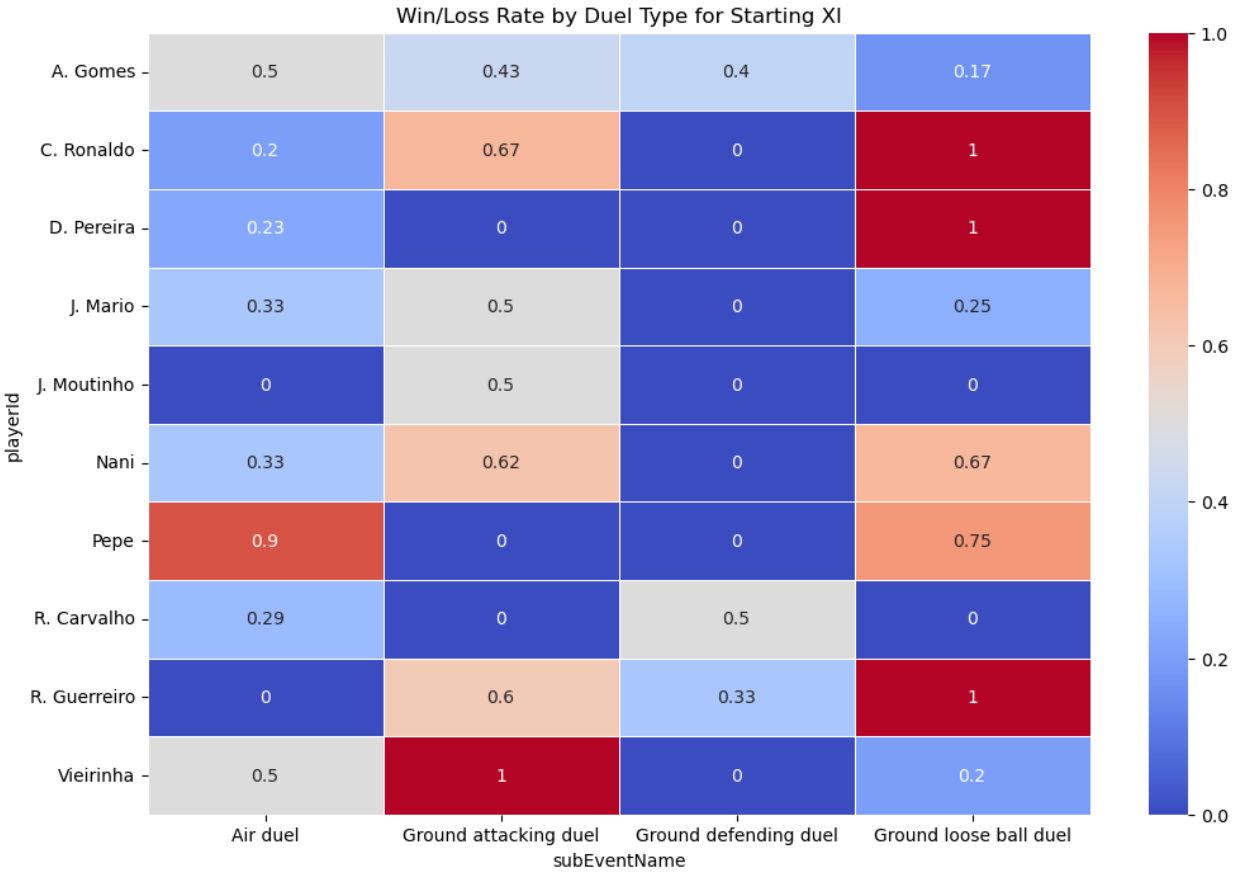


Figure 4.6 (1d) Heatmap of Win and Loss Rate by Duel Type for Players in the Starting Lineup

Average Duel Duration by Duel Type

J. Moutinho has short duel durations across all categories, indicating his ability to engage in quick, decisive actions, whether intercepting passes or quickly dispossessing opponents. Ronaldo's duels in ground-attacking situations last longer, likely due to his role in breaking down defenses and holding up play to involve teammates in the attack. D. Pereira and J. Mario have longer duel durations in ground-defending situations, reflecting their involvement in sustained defensive actions to shield the backline.

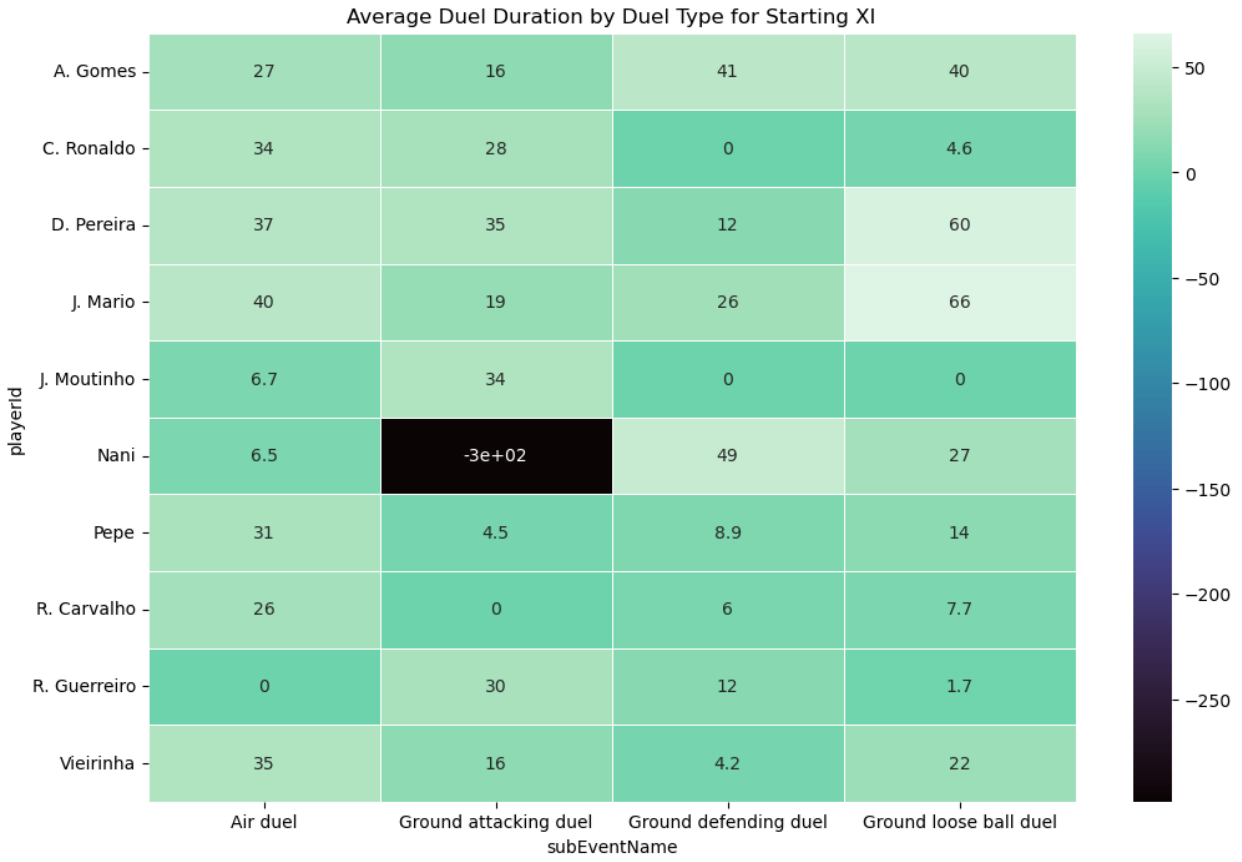


Figure 4.6(1e) Heatmap of Average Duration of Duels for Players in the Starting Lineup

Average Distance Covered in Duels by Duel Type

Pepe covers significant distance during air duels, correlating with his role as a center-back often stepping up to contest aerial balls. His defensive positioning and movement are key to his success in these situations. Nani covers considerable ground in ground loose ball duels, reflecting his role in wide areas, where he often competes for second balls or presses higher up the field. Similarly, R. Carvalho covers a significant distance in air and ground-defending duels, highlighting his contribution to protecting the backline and initiating defensive actions from deeper positions. These insights help understand the physical demands placed on players during critical match moments and inform decisions on player fitness, stamina, and positioning.

Figure 4.6(1f) Heatmap of Average Distance Covered in Duels for Players in the Starting Lineup

Duel Distribution in Clusters

Cluster 0 has the highest occurrence of ground-attacking duels, with 67 events, indicating that this cluster is associated with aggressive offensive plays, where ground challenges are common. Cluster 1, with 32 air duels, focuses on aerial battles likely occurring during goal-kicks, crosses, or long passes. Cluster 2 shows a balanced distribution of ground loose ball duels, suggesting a focus on regaining possession or contesting loose balls during the match. The analysis indicates that cluster 0 corresponds to periods of intense offensive pressure, while cluster 1 highlights moments of heightened aerial play, likely linked to set-pieces or long ball tactics.

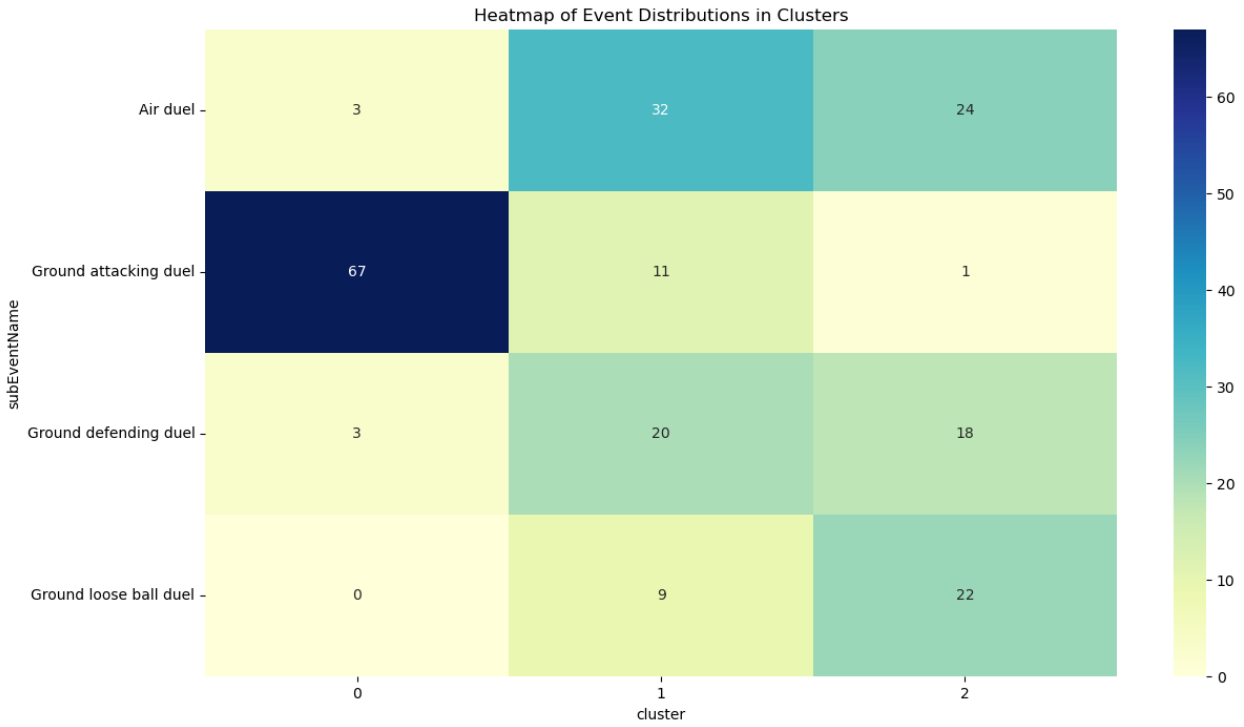


Figure 4.6(1g) Heatmap of Event Distributions in Clusters

Duel Frequency Over Time by Cluster

Duel events are spread throughout the match, with each cluster showing activity during different phases. Clusters 0 and 1 experience spikes in duel frequency around the 60th and 80th minutes, possibly corresponding to tactical shifts or moments of increased pressure. These spikes may indicate key turning points in the match, where Portugal pushed forward aggressively or adjusted their strategy, leading to more intense duels during those periods.

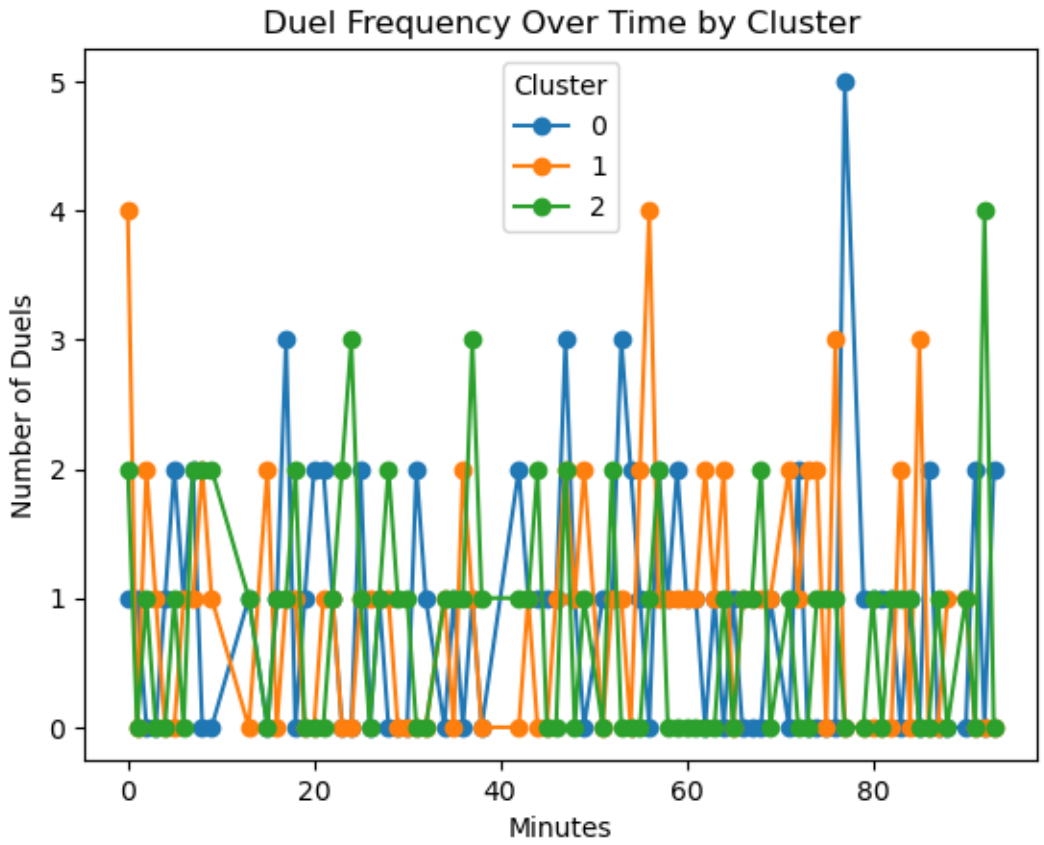


Figure 4.6(1h) Line Graph of Duel Frequency Over Time by Cluster

Cluster Transition Matrix

The diagonal elements of the matrix show the highest probabilities, indicating that once the match enters a certain cluster, it is likely to remain in that state for a while. This high transition probability is the persistence of certain match dynamics, whether ground-based or aerial-focused duels. There is moderate fluidity between Clusters 1 and 2, indicating a transition between aerial and ground-based challenges during key moments. These transitions suggest that the team frequently alternates between aerial battles and ground duels, adapting to match demands during critical phases.

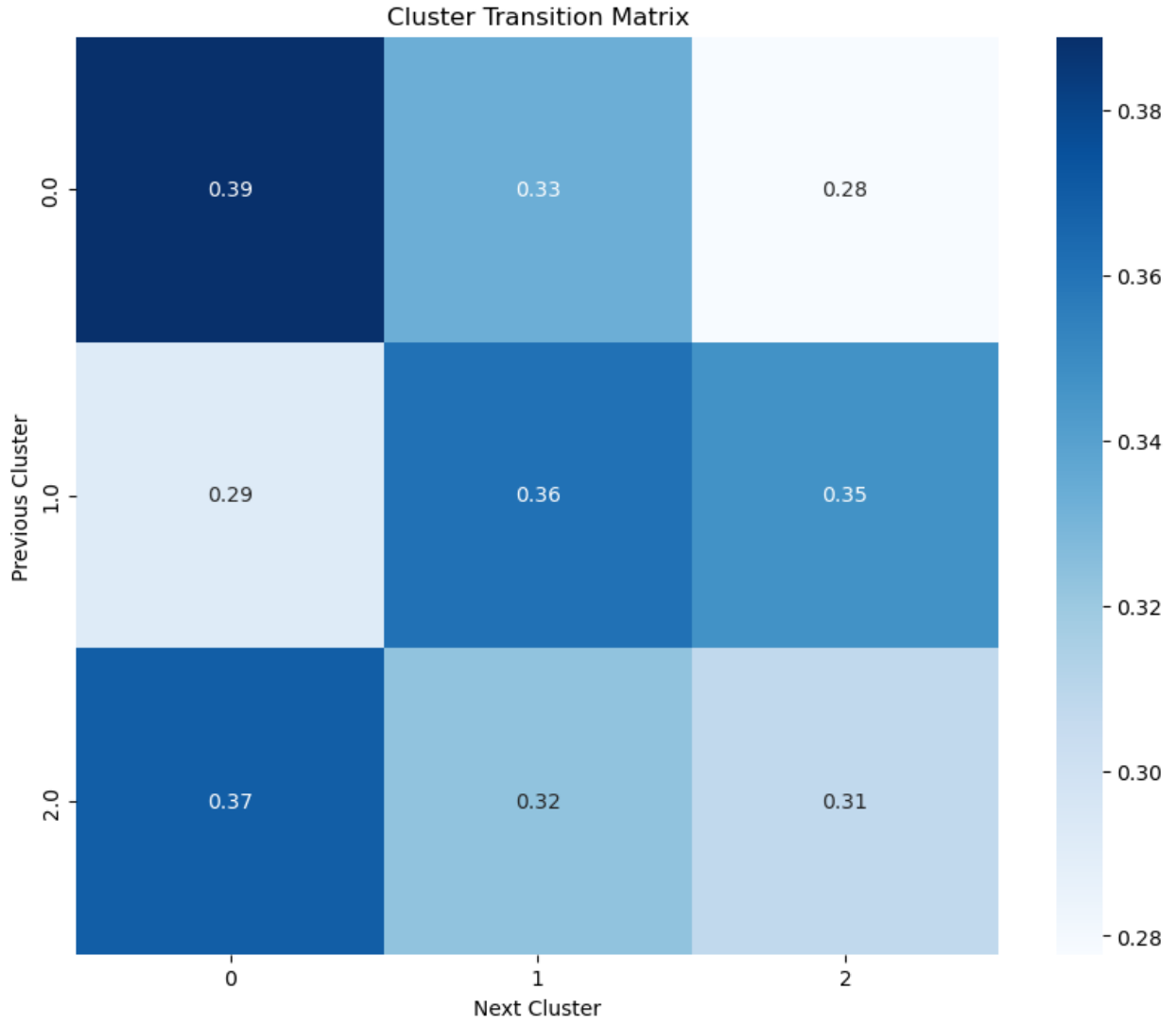


Figure 4.6 (1i) Heatmap Representing a Cluster Transition Matrix

Distribution of Duel Types and Match Period Duel Frequency

According to the bar chart, ground-attacking duels are the most frequent, emphasizing the team's focus on offensive engagements. Air duels are the second most frequent, highlighting the significance of aerial challenges in the match. Ground-defending and loose ball duels occur less frequently but are still essential for defensive play and regaining possessions. This distribution suggests that Portugal prioritizes offensive duels, with a notable reliance on direct, aerial play. The lower frequency of ground-defending duels suggests a possible area for defensive improvement. The second bar chart comparing duel frequency between the first and second halves of the match shows an increase in the number of duels in the second half, with over 110 compared to 90 in the first half. This suggests that the game becomes more intense and contested as it progresses, likely due to tactical adjustments or increased urgency from both teams. The rise in duels in the second half may indicate a more aggressive or direct approach, as Portugal seeks to secure a favorable result. This shift in intensity provides insights into how the team performs under pressure and how match dynamics involve overtime.

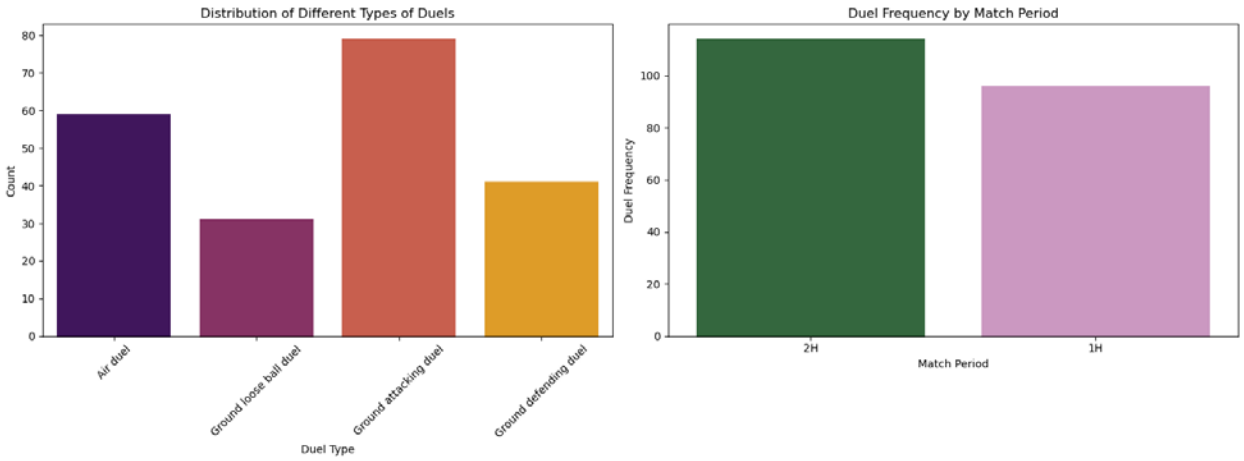


Figure 4.6(1j) Bar Plots Showing Distribution of Duel Frequency and Duel Frequency by Match Period

Due Frequency by Type and Match Period

The bar chart showing duel frequency by type and match period highlights a tactical shift as the game progresses. Ground-attacking duels increase in the second half, suggesting a more aggressive offensive approach as the match progresses. The rise in air duels reflects a greater reliance on long balls or aerial challenges, perhaps as Portugal looks to bypass the midfield or play more directly. The low frequency of ground-defending duels suggests a continued focus on offense throughout the game, with less emphasis on defensive engagements. Portugal's strength in ground-attacking duels indicates an offensive playstyle focused on winning battles in advanced areas, which might be linked to high pressing or regaining possession in dangerous zones. The lower win rate and frequency in ground-defensive duels highlight a potential area for improvement, as stronger defensive duels could help support Portugal's overall defensive performance. The increase in duel frequency in the second half, especially in ground-attacking and air duels, suggests that the game becomes more physical and contested, likely influenced by fatigue, strategic risks, or shifts in tactics as the match progresses.

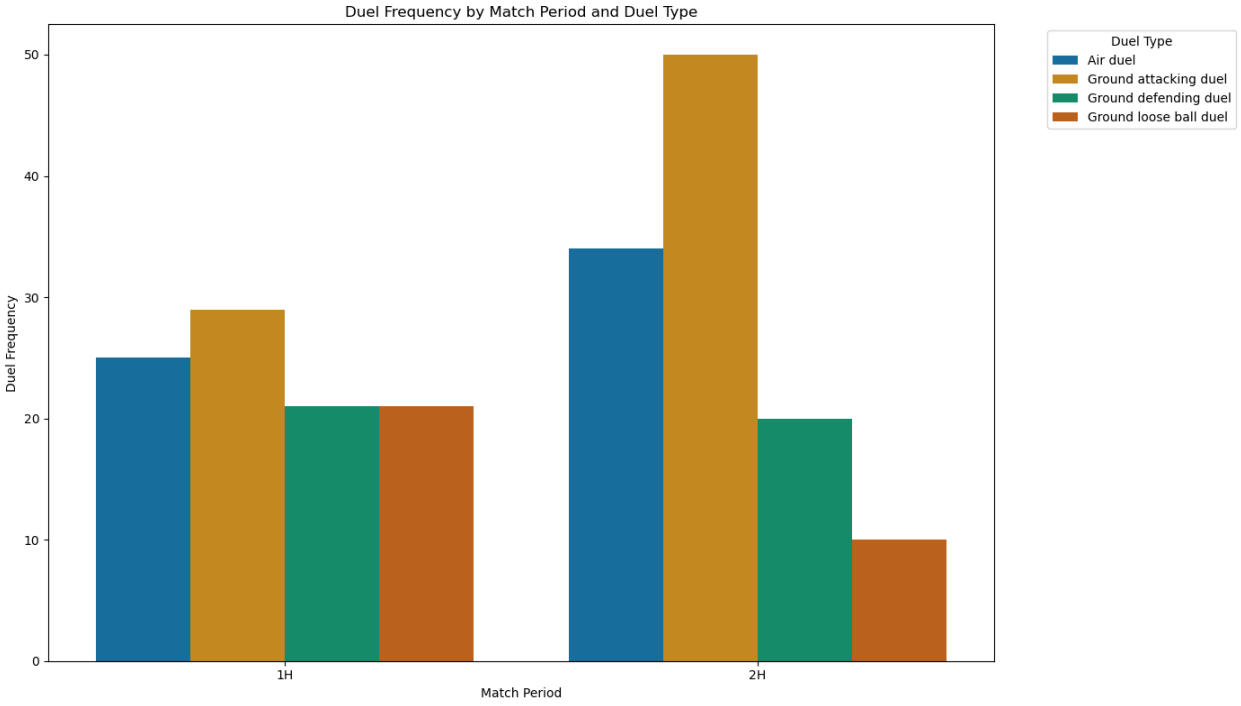


Figure 4.6(1k) Bar Plot of Duel Type and Frequency Over Time

4.7 Shooting Efficiency and Expected Goals Analysis

Total and Average Expected Goals for Each Player

C. Ronaldo leads with the highest total xG (expected goals), approximately 3, which reflects his central attacking role and consistent positioning in high-quality scoring areas. Nani follows with a total xG close to 2, showcasing his involvement in dangerous offensive plays. D. Pereira, A. Gomes, and Pepe also contribute, but with significantly lower xG values, indicating their reduced presence in goal-threatening situations compared to the forwards. D. Pereira has the highest average xG, suggesting that while his number of attempts might be low, the quality of his chances was notably high. C. Ronaldo and A. Gomes also have high average xG, reinforcing their effectiveness in positioning themselves for high-quality shots. Pepe's relatively low total xG suggests that he capitalized his limited opportunities in front of goal.

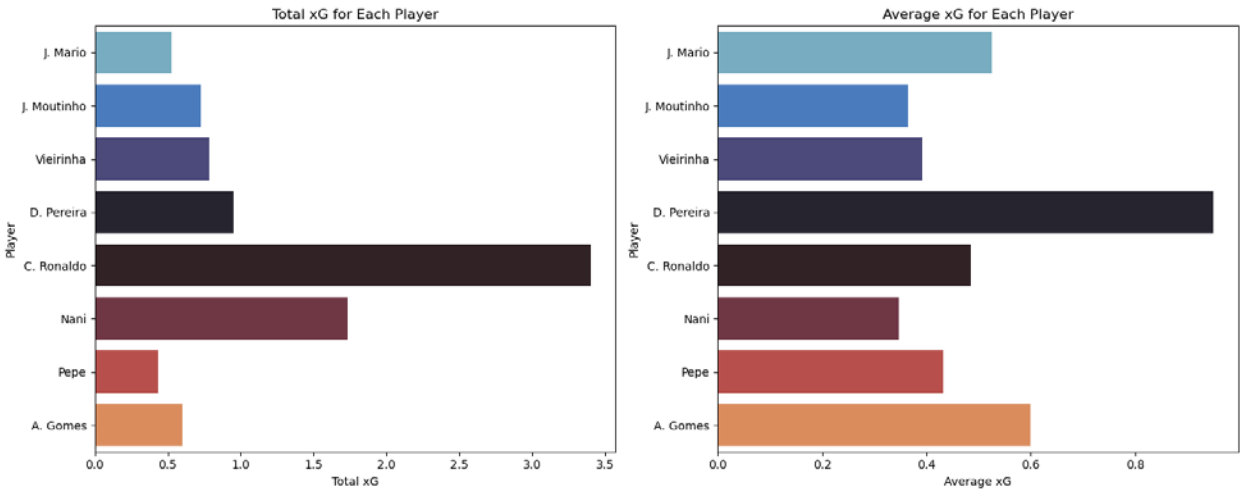


Figure 4.7(1a) Bar Plots of Total and Average Expected Goals for Each Player

Player Shot Count and Distribution of Shots by Match Period

C. Ronaldo dominates with 7 shots, emphasizing his aggressive approach and frequent attempts to score. Nani follows with 5 shots, reflecting his significant role in offensive phases. Other players like J. Moutinho, Pepe, and Vieirinha had fewer shots, underscoring their roles as more defensive or midfield-oriented. Ronaldo and Nani are central to Portugal's offense in total shots and xG, with Ronaldo being the primary attacking figure. Portugal took more shots in the first half than in the second, indicating a stronger attacking presence earlier in the game. While the distribution is relatively balanced, the slight drop in the second half could suggest tactical adjustments by the opposition or a decrease in attacking intensity. This distribution reflects a consistent ability to create shooting opportunities, with a minor decline as the game progressed.

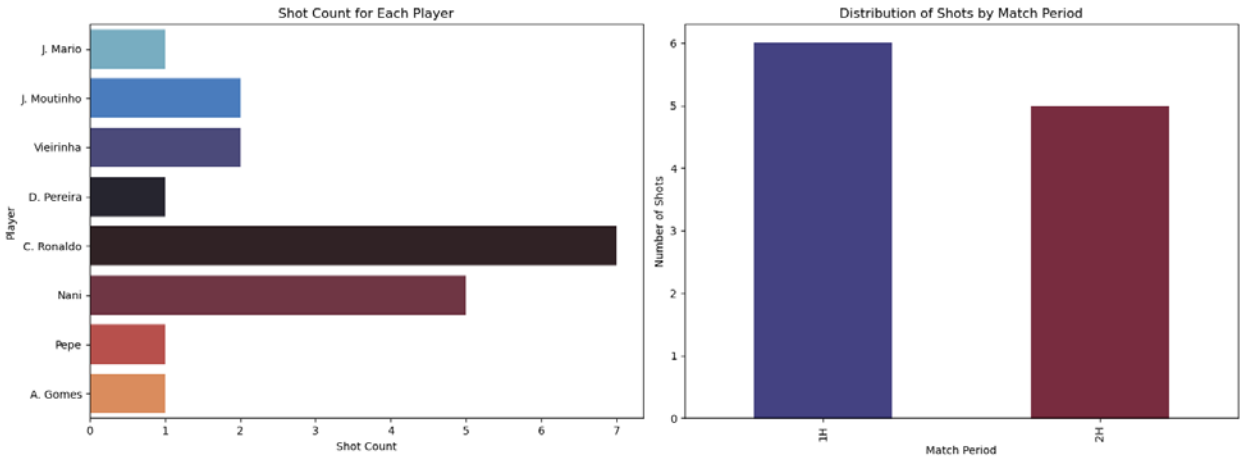
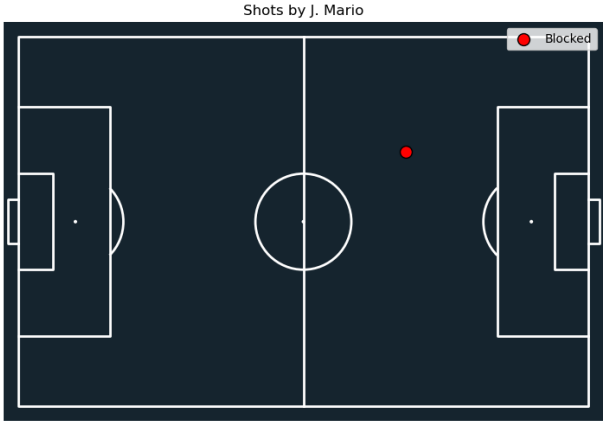


Figure 4.7(1b) Bar Plots of Player Shot Count and Distribution of Shots by Match Period

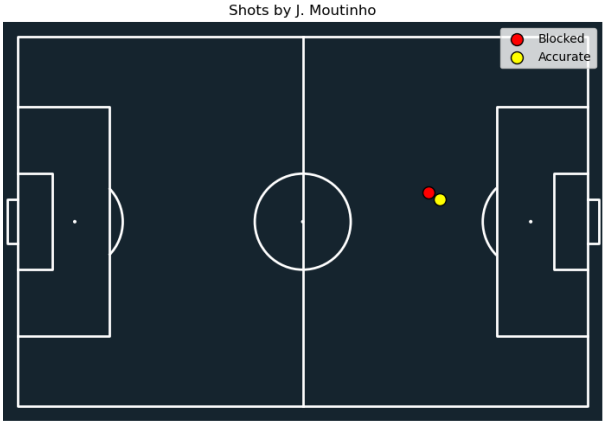
Shot Map for Each Player

J. Mario had one blocked shot from a central position outside the box, suggesting that defensive pressure limited his attempts. J. Moutinho's shot map shows 2 attempts, 1 blocked and 1 on target, both taken from just outside the penalty area. This highlights his willingness to take long-range shots from central positions, with 1 shot finding the target. Vieirinha had 1 accurate shot from a deeper midfield position, reflecting a more conservative approach, preferring precision shots from longer distances. D. Pereira took a blocked shot from just outside the box, showing his tendency to shoot from medium range while facing defensive pressure. C. Ronaldo is heavily involved in the team's shooting efforts, with multiple attempts from inside and outside the box. His map displays 1 accurate shot, 1 blocked shot, and 2 non-accurate shots, reflecting his versatility and role as Portugal's primary goal-scoring threat. Nani's shot map is clustered inside the box, with 1 goal, 1 accurate shot, and several accurate attempts, highlighting his positioning as a forward and success from close range. Pepe had 1 non-accurate shot from the right edge of the box, likely taken during a set piece or rare offensive move from his defensive role. A. Gomes had 1 accurate, and 1 blocked shot from just outside the penalty area, showing his preference for medium-range attempts though with mixed success.

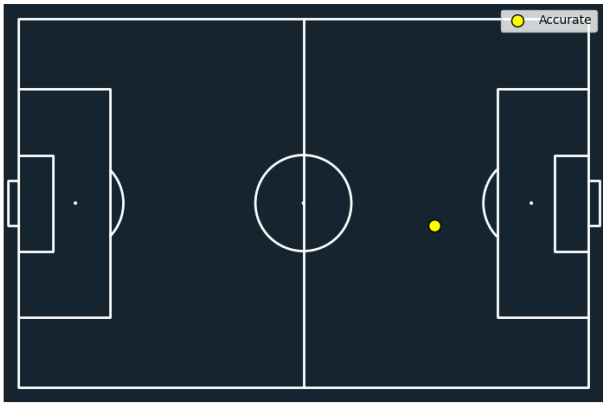
Shots by J. Mario



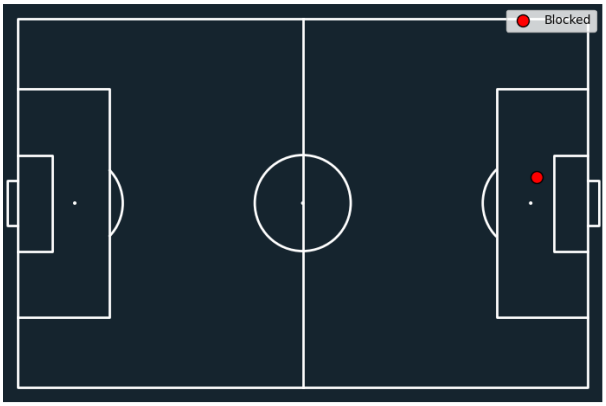
Shots by J. Moutinho



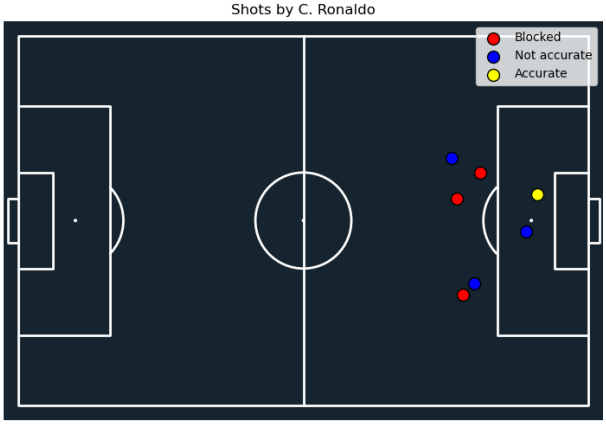
Shots by Vieirinha



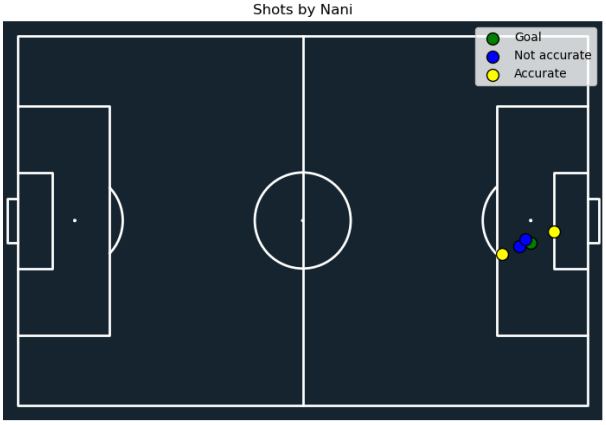
Shots by D. Pereira



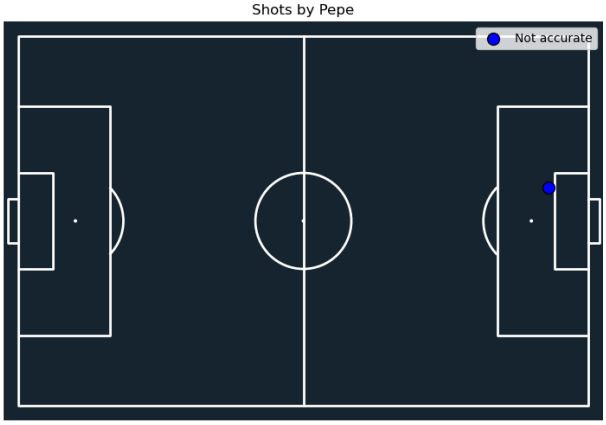
Shots by C. Ronaldo



Shots by Nani



Shots by Pepe



Shots by A. Gomes

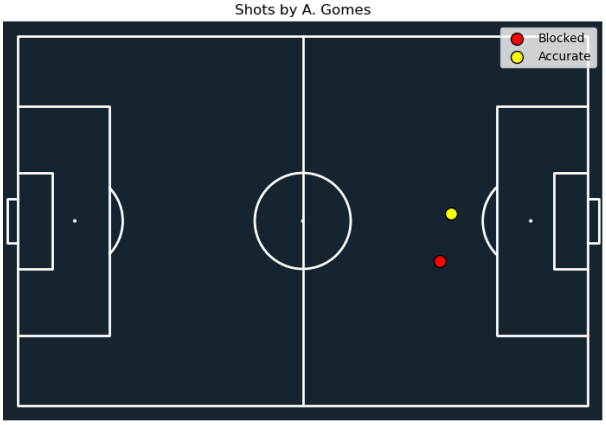


Figure 4.7 (1c) Pitch Plots of Shots for Each Player

Combined Shot Map

The combined shot map reveals that most attempts come from central positions in and around the penalty area, indicating a tactical emphasis on breaking through the opposition's defensive block via the center. Players like Ronaldo, Nani, and Pereira are the primary contributors, with Ronaldo and Nani significantly outshooting other players, highlighting their central role in Portugal's attack. Several blocked shots show the opposition's defensive ability to close down shooting lanes, suggesting adjustments to shot timing or selection. Nani's ability to convert 1 of his close-range attempts into a goal contrasts with other players, who struggled to hit the target despite having several chances. This disparity in finishing highlights the differences in individual shooting accuracy and the potential for tactical adjustments to improve finishing efficiency.

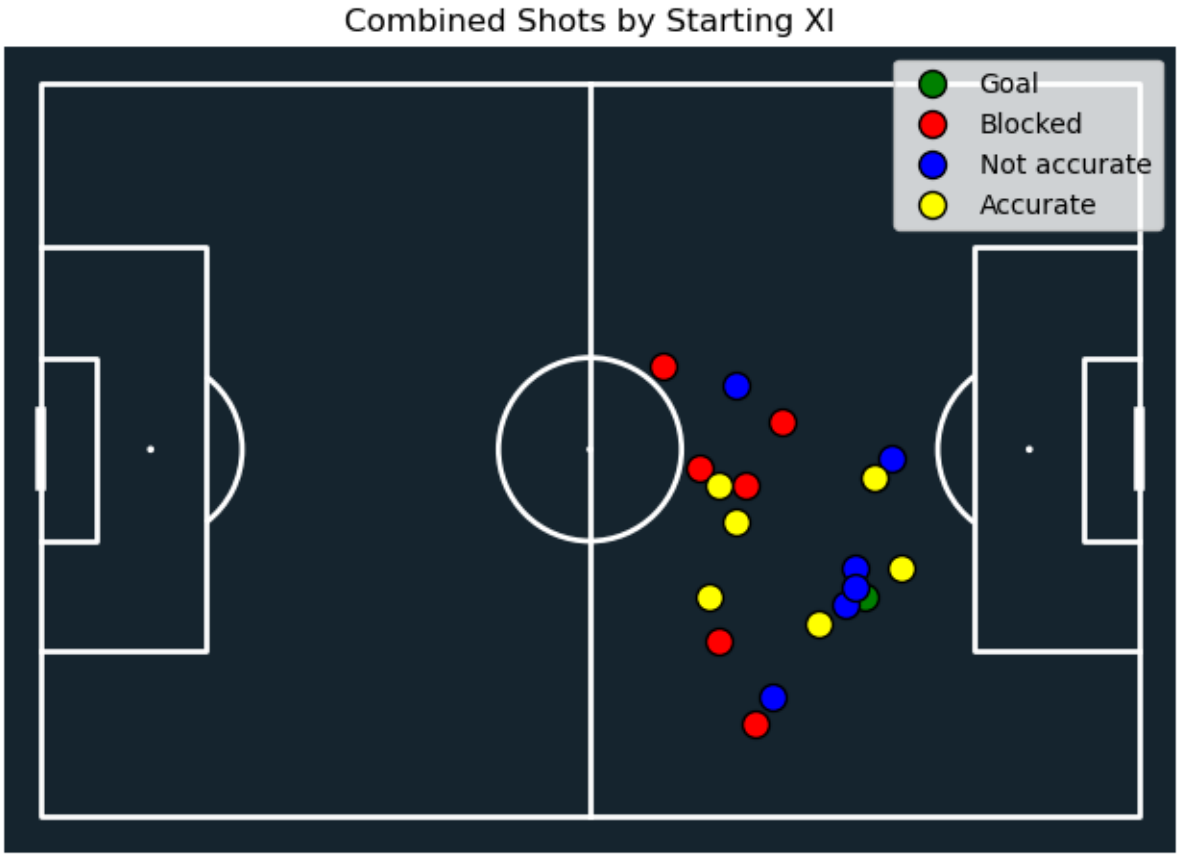


Figure 4.7(1d) Pitch Plot of Shots by All Players

Player Comparison Radar Chart

Ronaldo dominates in total xG and shot count, underscoring his pivotal role in generating high-quality chances and consistently being involved in Portugal's attacking play. This aligns with his position as the main striker, responsible for creating and converting scoring opportunities. Nani also stands out with a notable xG and shot count, reflecting his involvement as a forward in the team's offensive efforts. D. Pereira's high average xG is noteworthy, indicating that while he may not take many shots, the chances he does have are from dangerous areas and of high quality. In contrast, J. Mario, J. Moutinho, and

Vieirinha have lower xG and shot counts, consistent with their midfield or defensive roles, contributing less to direct scoring opportunities but supporting overall team play.

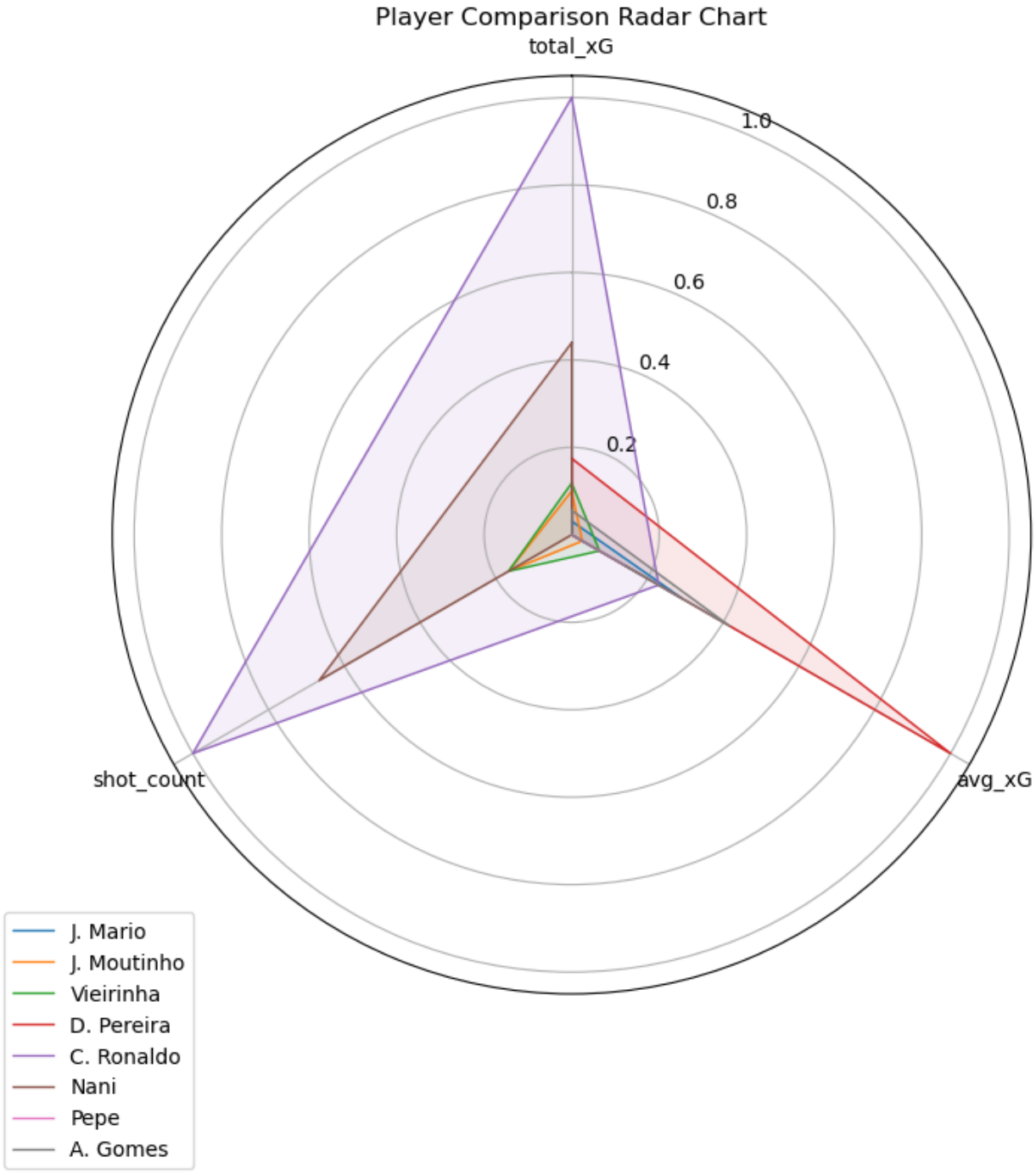
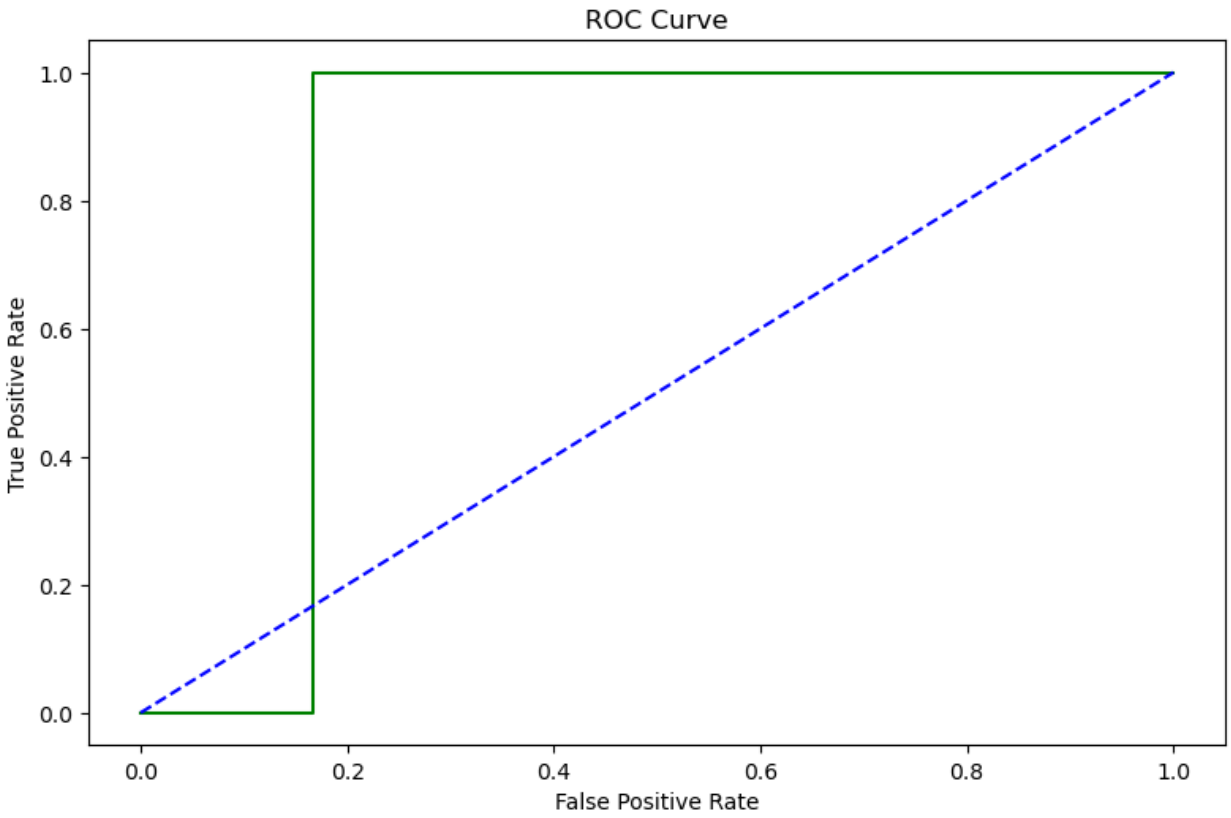


Figure 4.7 (1e) Radar Chart of Player Performance Metrics

Goal Outcome Prediction Based on Shot Characteristics

The logistic regression model used to predict goal outcomes based on shot characteristics shows strong performance, with an AUC of approximately 83% This value indicates the model's ability to distinguish

between goal and non-goal outcomes effectively. The precision score for non-goals highlights the model's reliability in not predicting goals when they don't occur, while the recall score reflects how well the model identifies actual goals. The accuracy score further supports the model's effectiveness, and cross-validation was used to ensure robustness by iterating through different subsets of the data. the mean ROC AUC score from 5-fold cross-validation demonstrates consistent performance across different data splits, validating the model's predictive power.



Precision (Class 0): 1.00
Recall (Class 0): 0.67
F1-Score (Class 0): 0.80

Precision (Class 1): 0.75
Recall (Class 1): 1.00
F1-Score (Class 1): 0.86

AUC-ROC: 0.83

Figure 4.7(1f) ROC Curve and Metrics for Machine Learning Model Predicting Shot Outcomes

4.8 Foul Patterns and Match Dynamics Analysis

Foul Locations by Players

The foul location map, color-coded by each player, reveals insights into the areas where fouls were committed. Players like D. Pereira, Pepe, and C. Ronaldo committed fouls primarily in the central and defensive zones, suggesting these fouls were tactical, likely aimed at disrupting the opposition's play or preventing dangerous transitions. Nani's fouls, located to the attacking third, reflect more aggressive play or attempts to press high up the pitch. This visualization outlines defensive and aggressive tendencies based on their foul locations across the pitch.

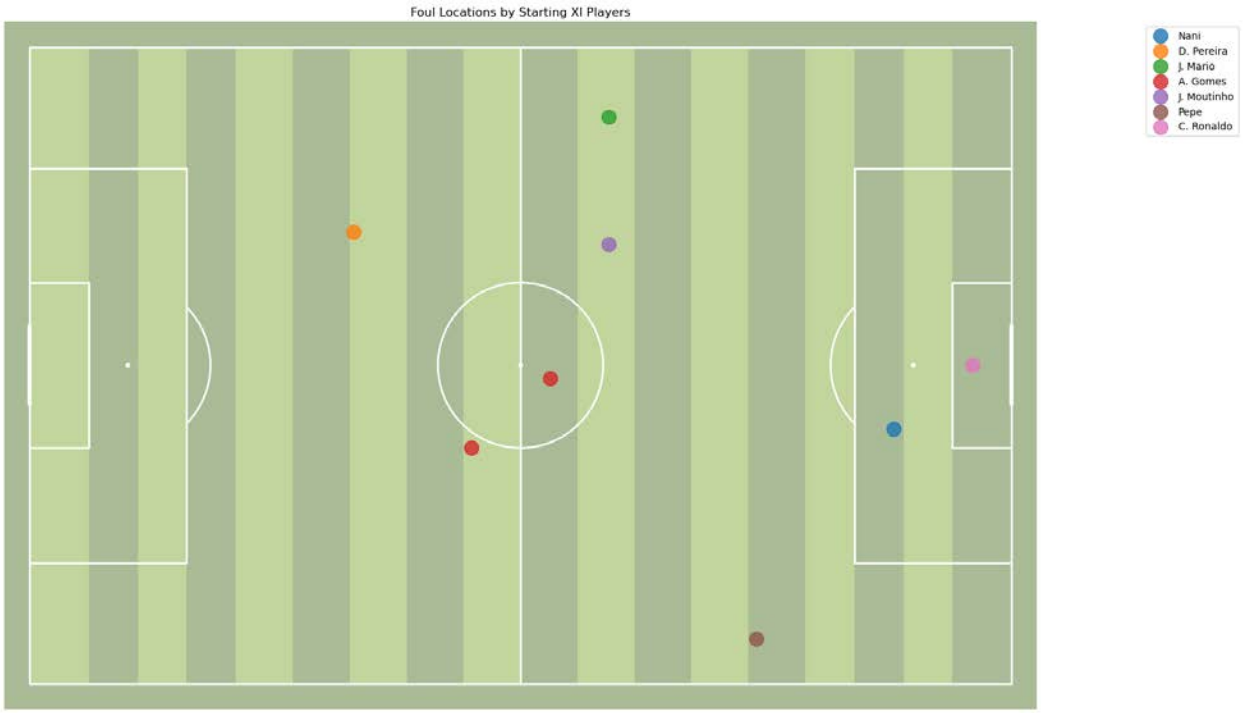


Figure 4.8(1a) Pitch Plot of Foul Locations by Players in the Starting Lineup

Distribution and Comparison of Fouls Between Halves

The first visualization highlights the distribution of fouls between the first and second halves. In the first half, fouls primarily occurred between the 10th and 25th minutes, suggesting a period of early intensity or aggressive pressure. In contrast, fouls in the second half, are more evenly spread, indicating that Portugal maintained physical engagement throughout the match but at a more controlled pace. This shift in distribution may reflect changes in match tempo or adoption of the opponent's tactics. The slight decrease in fouls from the first to the second half could suggest a more disciplined approach as the game progressed, possibly reflecting a strategic adjustment or response to the evolving game state. The second visualization compares the total number of fouls between the halves. A small reduction in fouls in the second half implies that Portugal may have adopted a more cautious or controlled strategy as the match progressed.

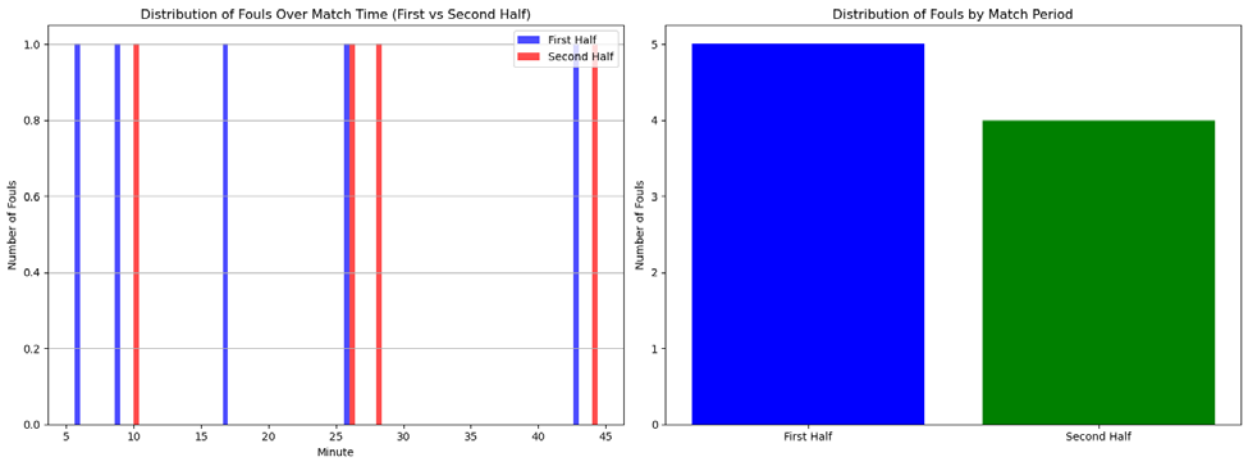


Figure 4.8(1b) Bar Plots Representing the Distribution and Comparison of Fouls Over Time

Total Fouls Committed by Players

The bar chart breaks down the total number of fouls committed by individual players. A. Gome leads with the most fouls (2), while C. Ronaldo, D. Pereira, J. Mario, J. Moutinho, Nani, and Pepe each contributed with 1 foul. This even distribution suggests that no single player was responsible for a disproportionate number of fouls. The balanced spread of fouls among players points to a collective tactical approach, likely aimed at disrupting play across different pitch areas. This indicates that fouling was part of a broader team strategy, rather than concentrated in the actions of a particular individual.

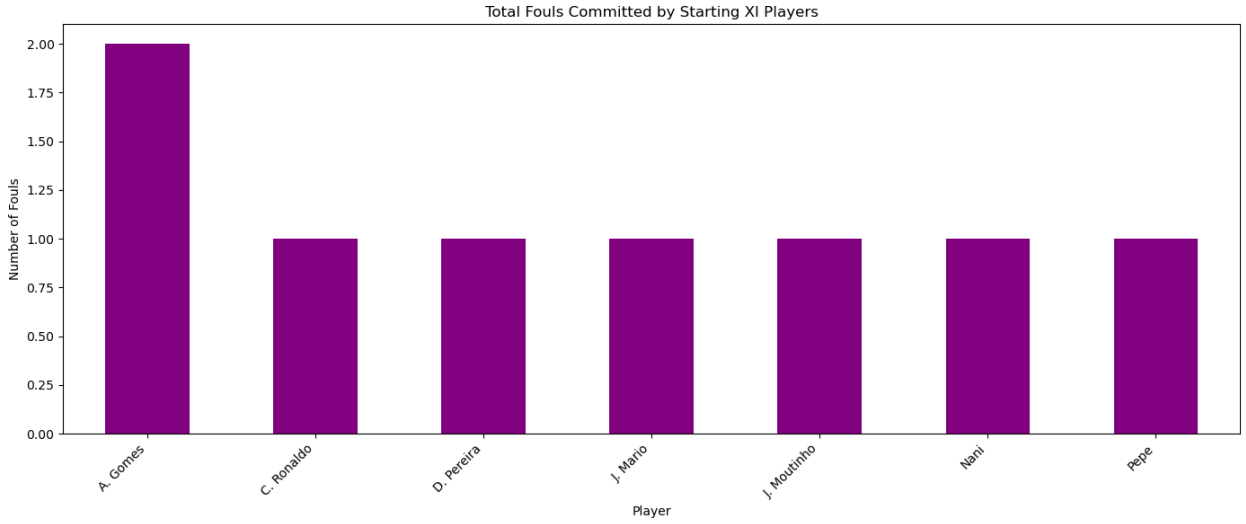


Figure 4.8(1c) Bar Plot of Total Fouls Committed by Players in the Starting Lineup

Distribution of Fouls by Zone

The pie chart illustrates the fouls committed in offensive versus defensive zones. Most fouls occurred in the defensive zone, reflecting Portugal's focus on containing the opposition's attacks and preventing counter-attacks in the defensive third. The proportion of offensive fouls suggests an aggressive pressing strategy higher up the pitch, where Portugal sought to regain possession or disrupt the opposition's

buildup. The emphasis on defensive fouls highlights the use of tactical fouls to break up play and control the game, particularly in the midfield and defensive zones, to prevent dangerous situations from developing.

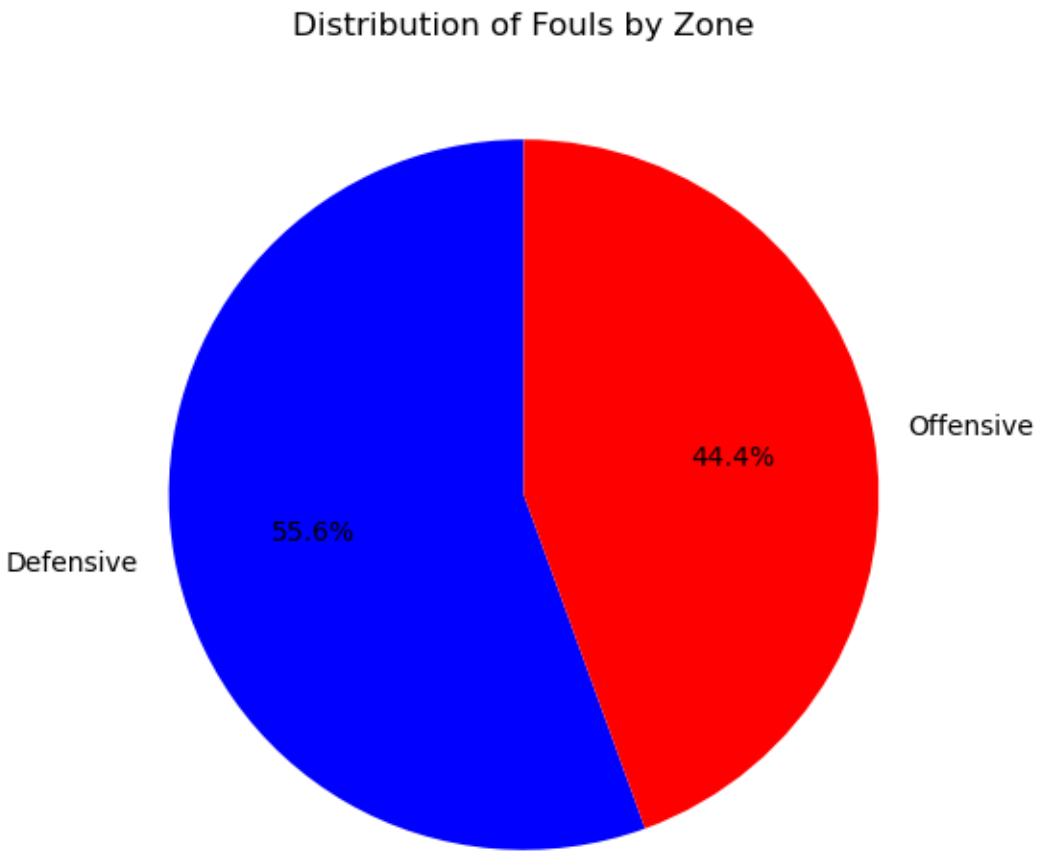


Figure 4.8(1d) Pie Chart of Foul Distribution by Zone

Chapter 5 – Discussion

5.1 Introduction

The chapter interprets and discusses the findings from Chapter 4 relative to the research objectives and hypotheses from Chapter 1. It examines how well the results address the key research questions concerning the impact of passing strategies, event sequences, and shooting behaviors on team performance. The discussion also compares these results with theoretical and empirical work from Chapter 2, providing deeper insights into tactical and performance implications. Additionally, the chapter assesses the confidence, validity, and generalizability findings, highlighting practical implications for football coaches, analysts, and researchers. Finally, it outlines suggestions for future research to address gaps and explore new opportunities in football performance analysis.

5.2 Comparison of Results with Research Questions

Key Findings

The analysis of passing strategies in Chapter 4 revealed how short and long passes influenced team dynamics. Short passes, particularly in midfield by players like J. Moutinho and R. Guerreiro, were crucial for maintaining possession and controlling the game's tempo. These players acted as key nodes in the passing network, facilitating transitions from defense to attack. Long passes, typically from defenders like Pepe and R. Carvalho, shifted play across the pitch, stretching the opposition's defense and enabling wide players like Vieirinha and R. Guerreiro to advance higher up the pitch, contributing to the team's adaptability. A summary table of key insights on passing strategies can be seen in Figure 5.2(1a).

Objective	Aspect	Key Insights
Passing Strategies and Team Performance	Short Passes in Midfield	Maintains possession and controls the game's tempo, with players like J. Moutinho and R. Guerreiro acting as key nodes
	Long Passes from Defenders	Stretches the opposition defense, enabling wide players to advance higher up the pitch
	Lateral Passes between Defenders	Helps retain possession in defensive areas and creates opportunities to push forward
	Wide Players Contribution	Contributes to offensive transitions, stretching the field and advancing play

Figure 5.2(1a) Summary of Key Insights on Passing Strategies and Team Performance

Alignment with Initial Hypothesis

The findings support the hypothesis that lateral passes between defenders and short passes in midfield contribute to ball retention and offensive play creation. For example, frequent lateral passes between R. Carvalho and Pepe allowed Portugal to retain possession in defensive areas, waiting for opportunities to push forward. Long diagonal passes helped stretch the opposition and create space for wide players like R. Guerreiro and Vieirinha. The network analysis confirmed that Moutinho and Guerreiro's high centrality in the passing network played a key role in Portugal's possession-based strategy.

Literature Comparison

These findings align with Hughes & Franks (2005), who highlighted the importance of midfield coordination in successful teams, as seen in Moutinho's pivotal role. Clemente et al. (2015) also emphasized network centrality in controlling the game, as shown in Moutinho's short-passing game. The study expands on these works by showing how long passes defenders like Pepe and R. Carvalho to the flanks disrupted defensive formations, and how wide players like R. Guerreiro and Vieirinha contributed to offensive transitions the study also supports Hughes and Franks (2005), model of teams using short, lateral passes in defense to maintain possession before launching attacks.

Key Insights

The findings align with existing literature and provide new insights into the evolving role of wide players like R. Guerreiro and Vieirinha in stretching the field and advancing play. While central midfielders were traditionally seen as pivotal, this study highlights the growing importance of wide players in maintaining defensive stability and contributing to offensive transitions, adding depth to the understanding of network centrality and player positioning in modern football.

Conclusion

In conclusion, Portugal's passing strategies, through short passes in midfield and long diagonal passes from defenders, were integral to maintaining team structure and performance. The results confirm the original hypothesis that passing patterns help retain possession, control the game, and create goal-scoring opportunities. These studies align with previous findings from Hughes & Franks and Clemente et al. while offering new insights into the role of wide players in the passing network and their contribution to possession retention and tactical flexibility.

Key Findings

The analysis of event sequences and tactical patterns explored how passing sequences duels, and fouls, influenced match dynamics and team strategies. The study segmented the match into distinct episodes, critical periods were identified where tactical patterns emerged. For example, clusters of dueling activity in central areas often resulted in possession turnovers, leading to rapid counter-attacking opportunities. These patterns highlighted the tactical importance of winning duels in key areas of the pitch. Sustained possession changes led to short-passing sequences to maintain control and build toward scoring opportunities, while counter-attacking phases involved longer, direct passes to exploit space left by the opposing defense. This segmentation illustrated how teams adapted strategies based on evolving match scenarios, with specific tactical responses required depending on game momentum and opposition positioning. Figure 5.2(1b) shows a summary table of event sequences.

Objective	Event Sequence	Tactical Outcome
Event Sequences and Tactical Patterns	Short Passing Sequences	Maintains possession, builds toward goal opportunities
	Long Passing Sequences	Exploits wide areas, stretches opposition defense
	Duels in Central Areas	Leads to turnovers, creates counter-attacking chances
	Fouls in Wide Areas	Disrupts attacking play, allows defensive reorganization

Figure 5.2(1b) Summary of Event Sequences and Their Tactical Outcomes

Clustering of Tactical Patterns

Clustering methods revealed distinct tactical patterns, such as high-risk, high-reward sequences following duels or interceptions. These rapid, direct passing sequences moved the ball into attacking areas while the opposing team was out of position. Clustering also exposed defensive patterns, where teams committed fouls in wide areas to disrupt attacks and regain defensive shape. This demonstrated how fouls were strategically used to slow down play and reorganize, especially after an opponent's fast break, showing the importance of fouls as a defensive tactic.

Literature Comparison

The findings align with Pappalardo et al. (2019), who focused on event sequences to predict match outcomes. However, this study expands on their work by emphasizing the role of duels and fouls, which have received less attention in previous research. While Pappalardo highlighted the importance of pass sequences, this analysis shows how duels can significantly influence momentum shifts and how fouls serve as tactical resets, offering a more detailed understanding of game dynamics and illustrating how defensive actions can actively disrupt an opponent's momentum.

Conclusion

In conclusion, tactical success is shaped by passing sequences and by duels, fouls, and transitions. By segmenting matches into tactical phases, this study provided new insights into how teams adapt their strategies around key defensive and offensive events. These findings expand on prior research, such as Pappalardo et al. (2019), contributing to a broader understanding of how teams react to pivotal moments in a match and offering fresh perspectives on event sequences in football.

Key Findings

The analysis of shooting behavior and goal outcomes in Chapter 4, used expected goals (xG) models to assess shooting efficiency and tactical decision-making in goal-scoring opportunities. The xG model performed well in predicting the likelihood of a shot resulting in a goal, factoring in elements such as shooting distance, angle, and pass type. The model's accuracy was particularly high for shots taken inside the penalty area, aligning with the understanding that closer shots have a higher probability of success. Additionally, the analysis identified patterns in shot-taking behavior, where certain players, particularly forwards like C. Ronaldo, consistently positioned themselves in high-probability zones, such as inside the box, to maximize their chances of scoring. This demonstrates the tactical importance of positioning and shot selection in offensive strategies.

Objective	Aspect	Key Insights
Shooting Behavior and Goal Outcomes	Expected Goals (xG) Model	Predicts likelihood of a goal based on factors like shooting distance, angle, and pass type
	Shooting Distance and Angle	Closer shots, especially inside the penalty area, have a higher probability of success
	Player Positioning	Forwards like C. Ronaldo consistently position themselves in high-probability zones to maximize scoring chances
	Shot Selection	Inefficiencies arise from low xG shots taken from difficult angles or long distances

Figure 5.2(1c) Summary of Key Insights on Shooting Behavior and Goal Outcomes

Prediction Accuracy and Tactical Implications

The xG model revealed key tactical insights, particularly in Portugal's approach to chance creation. The model's high predicted values for shots following crosses and through balls into the box reflect a tactical decision to create high-quality scoring opportunities by exploiting wide areas and delivering passes into dangerous zones. However, the model highlighted inefficiencies where players took low xG shots from difficult angles or long distances, rarely leading to goals. These findings suggest that focusing on high xG opportunities and avoiding speculative shots could improve overall shooting efficiency.

Literature Comparison

The xG model's results are consistent with previous research such as Lucey et al. (2015) and Caley (2014), emphasizing the importance of shooting distance and angle for scoring success. However, this analysis extends traditional xG models by incorporating additional factors such as pass type and defensive pressure, providing a more comprehensive understanding of what influences shot outcomes. Furthermore, using machine learning to predict shooting efficiency introduces an innovative layer by examining the sequence leading up to the shot, revealing how tactical decisions in build-up play, such as crossing patterns and through passes, significantly impact xG values and goal likelihood.

Conclusion

In conclusion, the xG model offered valuable insights into shooting behavior and goal-scoring efficiency, confirming that shots from closer distances with favorable pass types have higher success rates. The analysis underscores the need to improve shot selection, prioritizing high xG chances over speculative attempts. Compared to traditional xG models, additional variables and a focus on pre-shot events provide new insights into football analytics, enhancing tactical decision-making in offensive play.

5.3 Implications of the Findings

Practical Implications for Football Teams

The findings from this research offer several practical insights that can be applied to improve team strategies in football. One key insight is the significance of passing network centrality for ball control and team coordination. J. Moutinho and R. Guerreiro were central to maintaining possession and controlling the game's tempo, particularly in midfield. This suggests that coaches should focus on developing passing networks where key players, like central midfielders, are positioned to maximize their involvement in

linking play between defense and attack. Teams could enhance passing routes by encouraging short-passing sequences in midfield and using longer diagonal passes from defenders to stretch the field and exploit space on the flanks. Utilizing wide players, like fullbacks, can improve defensive stability and support offensive transitions. As seen with R. Guerreiro and Vieirinha, occupying wide spaces helps maintain team shape and creates counter-attacks. The findings on duels also provide valuable insights into how teams can adjust their tactics in high-pressure situations. The analysis showed that winning central duels often leads to rapid transitions and scoring chances. Teams can use this information to emphasize the importance of positioning in key areas, like central midfield, where winning duels can shift momentum. Players in these positions should be trained to be more aggressive in challenging for the ball and to quickly initiate fast-paced offensive plays after winning possession. Additionally, the insights on foul clustering in defensive areas highlight the tactical use of fouls to disrupt the opposition's play and allow the team to regain defensive shape. Teams can incorporate this strategy by committing tactical fouls in non-threatening areas when out of position, helping to reset the formation and recover defensively.

Theoretical Contributions

This research contributes to sports analytics by integrating machine learning models with event sequence data to analyze outcomes such as passing efficiency, duel success, and shooting behavior. Using clustering techniques to analyze passing patterns and event sequences introduces a new framework for understanding the tactical interplay between passing networks, duels, and fouls in shaping match dynamics. A key contribution is the focus on duel frequency and its impact on tactical success. While past research has mainly emphasized passing sequences as the primary driver of performance, this study highlights the role of defensive and offensive duels in determining match outcomes. This adds depth to understanding team transitions, showing how winning key duels, especially in central areas, can create rapid counter-attacking opportunities. Moreover, this research extends the literature on expected goals (xG) by incorporating pre-shot events such as pass types, positioning, and passing routes into predictive models. Traditional xG models focus on shot characteristics, however, this study expands the analysis by examining the sequences leading up to shots, offering deeper insights into how build-up play and chance creation affect scoring probabilities. The integration of sequence-based event data with predictive modeling can be applied to other areas of football, such as analyzing defensive formations or pressing tactics, providing a broader scope for performance analysis in football analytics.

Implications for Further Research

The findings from this research open multiple avenues for further studies in football analytics. One key area for future research is the development of real-time game analysis tools that can track and visualize passing networks, duels, and fouls as they occur. By incorporating real-time data, teams could adjust their strategies dynamically, responding to changes in possession patterns or duel outcomes. Future research could focus on creating real-time predictive models that assess the likelihood of winning duels or scoring goals based on current game conditions. Additionally, future studies could expand the predictive models by incorporating variables such as player fatigue, weather conditions, and opponent strength. These additional factors could improve the accuracy of the models and provide a more holistic understanding of the influences on team performance. Integrating physical metrics, such as distance covered or player velocity, could offer insights into how physical performances affect decision-making in high-pressure situations, like duels or counter-attacks. This research could be applied to other sports, such as basketball or hockey, where passing networks and event sequences similarly impact team performance. Clustering techniques and predictive models could help researchers analyze how team dynamics and player interactions affect match outcomes in different sports contexts. This cross-sport application could yield comparative insights, contributing to multi-sports analytics. Finally, further research could investigate

how substitutions and tactical changes affect match phases. For instance, analyzing the impact of substitutions on passing networks and duel performance in the final stages of a game could help teams optimize tactical adjustments in real-time, building on the concept of tactical resets identified in this study.

5.4 Confidence in Results, Validity, and Generalizability

Confidence in Results

The findings from this study are robust due to the comprehensive event data and rigorous clustering techniques used. Detailed data on passes, duels, shots, and fouls, enabling in-depth analysis of tactical patterns across different play phases. The machine learning models, especially the xG model, showed strong predictive accuracy for goal-scoring probabilities, particularly for shots taken inside the penalty area, reinforcing confidence in their ability to evaluate shooting efficiency and goal outcomes. The clustering techniques applied to event sequences such as duels and fouls successfully identified tactical phases, providing insights into how teams transitioned between offensive and defensive strategies. While the clustering approach is powerful, its effectiveness depends on data quality and may vary depending on match scenarios. However, given the overall predictive performance and identification of tactical phases, there is a high confidence in the study's findings.

Validity

The methods used in this study are valid, as the results align with established football theories and patterns. The analysis of passing strategies supports the findings of Hughes & Franks (2005) and Clemente et al. (2015), who emphasized the importance of midfield coordination and network centrality in maintaining possession and dictating match dynamics. Using xG models to predict shooting efficiency was widely validated in prior research, indicating that the methods employed in this study are suitable for answering the research questions. However, there are limitations to consider. This study focused on a single match, Portugal vs Iceland, which may limit the representativeness of the findings. Match-specific factors, such as team formations, opposition strength, and tactical approaches, could influence the results, raising concerns about the broader applicability of the conclusions. Additionally, the reliance on event data, without including tracking data limits the depth of tactical insights. While the machine learning models performed well in predicting shooting outcomes, incorporating additional variables, such as defensive pressure, player fatigue, or weather conditions, could enhance predictive power. These factors could have influenced in-game decisions and performance but were not accounted for in this analysis.

Generalizability

The generalizability of the findings is somewhat restricted due to the focus on a single match between Portugal and Iceland. While the study provides valuable insights into team dynamics, passing strategies, and event sequences, it's unclear if these findings can be applied broadly to other matches, teams, or competitions. The tactical approach of Portugal in this specific match might differ significantly from how other teams operate, or how Portugal performs against different opponents with varying play styles. The results might not apply to teams with different tactical priorities, such as those that favor a direct, long-ball approach which would likely not emphasize network centrality and short-passing sequences as much as Portugal did in this game. To improve generalizability, future research should extend the analysis across multiple matches and teams, incorporating a broader dataset that includes a variety of play styles, formations, and competitive contexts. This would allow a more comprehensive understanding of how passing networks, duels, and shooting efficiency impact team performance under various conditions. Tracking data could provide more detailed insights into player interactions and tactical adaptations,

further enhancing the applicability of the results. In conclusion, while the findings of this study are robust and valid within the context of the specific match analyzed, their broader generalizability is limited. Future research should test these methods across a wider range of matches and contexts to evaluate the consistency and applicability of the results to different teams, leagues, and tactical scenarios.

Recommendations for Future Work

The accuracy and predictive power of the machine learning models used in this study can be improved through several refinements. One key enhancement would be additional player metrics, such as fitness levels, positioning, and velocity. Incorporating these variables would provide a more comprehensive view of player performance and how physical conditions influence their ability to execute tactical decisions like passes, duels, and shots. Furthermore, more advanced machine learning models like neural networks or deep learning algorithms could capture complex, nonlinear relationships between in-game events and outcomes. These models can analyze high-dimensional datasets and uncover patterns that traditional models might miss, potentially increasing the accuracy of predictions related to goal-scoring opportunities.

Broader Data Analysis

Generalizability can be improved by expanding the scope beyond a single match. Future research should analyze a broader dataset with matches from different teams, leagues, and competitions, accounting for different tactical styles and opposition strategies, providing a more comprehensive understanding of how teams use passing networks, handle duels, and create goal-scoring chances. This larger sample would also allow researchers to identify context-specific trends, such as how teams adapt passing strategies under pressure or against stronger opponents. Integrating tracking data would further enhance the analysis, capturing dynamic game elements such as player movement, speed, and positional changes. Tracking data could provide insights into off-ball actions, such as how players adjust positioning during defensive transitions or contribute to passing networks through off-ball movement. Additionally, real-time analytics could be developed to offer instant feedback to coaches and analysts during live matches.

Collaboration with Data Providers

Collaborating with data providers such as Opta, StatsBomb, or Wyscout could significantly enhance future analyses by providing access to more granular data. These providers offer detailed information on player actions, such as possession chains, defensive pressure, and progressive passes, allowing for a deeper examination of tactical patterns. Partnering with these data providers would enable proprietary metrics like expected assists (xA) or defensive coverage, offering new perspectives on how individual player actions contribute to overall team performance. Moreover, data providers are developing new models and metrics that incorporate positional and tracking data, making it easier to evaluate the effectiveness of possession strategies and the impact of player movement on creating space. Establishing partnerships with these providers would broaden the range of variables for analysis, facilitating studying more complex tactical interactions in football.

Chapter 6 – Evaluation, Reflections, and Conclusions

6.1 Evaluations of the Project as a Whole

Objectives Review

The primary objectives of this dissertation were to analyze passing strategies and their impact on team performance, explore event sequences and tactical patterns in football matches, and assess shooting behavior using advanced analytics and machine learning models. These objectives addressed gaps in football analytics by focusing on dynamic interactions rather than isolated statistics. The research effectively met these goals through network analysis, PCA, clustering, and machine learning models. Network analysis was useful in revealing the roles of key players, such as J. Moutinho and R. Guerreiro, in Portugal's possession-based strategy. It visualized passing structures providing insights into team performance. However, future research could improve this by integrating real-time tracking data to capture off-ball movement, and positional adjustments, leading to a more complete understanding of how teams manage space. Clustering techniques successfully identified tactical patterns during match phases, especially around duels and fouls. These methods offered valuable insights into game dynamics and were further enhanced by including spatial positioning and defensive formations, providing a more detailed view of tactical adjustments across match phases. The xG model offered strong predictive insights into shooting behavior, particularly in high-probability zones. However, incorporating more contextual data, such as defensive pressure during build-up play, could improve the model's accuracy and relevance. Including variables accounting for the timing and intensity of defensive actions would enable a more detailed analysis of how opposing defenses impact shot quality. Overall, the methodologies used were well-suited to the project's objectives. However, further refinements such as integrating tracking data and expanding the range of variables analyzed could have added depth and broadened the applicability of the analysis to a wider range of football scenarios. A summary table of the research objectives, methods, and key findings can be seen in Figure 6.1(1a).

Objective	Methods	Key Findings
Analyze passing strategies and their impact on team performance	Network analysis to visualize passing structures and key player roles	Short passes in midfield and long diagonal passes were essential for possession and team control
Explore event sequences and tactical patterns	Clustering & PCA to identify tactical patterns during key phases	Duels and fouls played a key role in tactical phases and momentum shifts
Assess shooting behavior using advanced analytics	Expected Goals (xG) models to assess shot efficiency and outcomes	Shooting from closer distances had higher success; shot selection and positioning were vital for goal efficiency

Figure 6.1(1a) Summary of Research Objectives, Methods, and Key Findings

Review of Literature

The literature review in Chapter 2 provided a strong theoretical foundation by examining the evolution of football analytics, focusing on passing networks, event sequence analysis, and shooting models. Studies by Hughes & Franks (2005) and Clemente et al. (2015) guided the research, particularly in understanding midfield coordination and network centrality in team performance. This dissertation also built upon Pappalardo et al. (2019), emphasizing event sequences in predicting match outcomes. While Pappalardo's

work focused on passing sequences, this study extended the analysis to include duels and fouls, demonstrating their importance in disrupting play and resetting defensive structures. These non-passing actions added depth to understanding match dynamics and event-based football analytics. The xG model analyzing shooting behavior was consistent with the work of Lucey et al. (2015) and Caley (2014), showing the relationship between shot distance, angle, and goal probability. However, this dissertation advanced the literature by incorporating pre-shot events, like pass types and defensive pressure, providing a more comprehensive view of factors influencing shot outcomes. Most existing literature on event sequences focuses on European leagues, but this study's findings could be applied more broadly. Future research could explore how non-Western leagues, such as those in South America or Asia, utilize passing networks and tactical duels, expanding the global body of football analytics and showing the versatile methods used. In conclusion, the literature review established a solid theoretical basis, and the dissertation's findings aligned with and expanded upon existing academic work in football analytics. Incorporating elements like pre-shot decision-making and broadening the analysis beyond passing sequences helps the research offer a deeper understanding of football tactics and strategies across different cultures.

Methods Used

The methods in this dissertation, including network analysis, clustering, PCA, and machine learning, were well-suited for addressing the research objectives, each contributing uniquely to understanding football performance. Network analysis effectively evaluated player importance within the passing network, identifying key players like R. Guerreiro and J. Moutinho. Metrics such as degree centrality provided quantifiable measures of player involvement, aligning with studies like Clemente et al. (2015). Future studies could enhance this by incorporating tracking data for a more dynamic understanding of player positioning off the ball, offering a more detailed view of team tactics. Clustering techniques identified tactical patterns during match phases, particularly high-risk high-reward sequences in duels and fouls. However, including spatial positioning and detailed team formation data would provide clearer insights into how teams adapt during key moments. This would reveal more about defensive structure and how they shift throughout the game. PCA was used to reduce data dimensionality, helping isolate significant features in passing strategies and shooting behavior. This enhanced result interpretability by focusing on key factors contributing to successful tactical plays. Future research could incorporate more granular data, such as player-specific attributes, for a more personalized analysis of how players impact match dynamics. Machine learning models for predicting shot outcomes using xG provided valuable insights into shooting efficiency and tactical decision-making. While the model performed well, it could be improved by adding variables like defensive pressure and player positioning during the build-up, offering more context for how defensive actions affect shot quality. In summary, the methods used effectively achieved the research objectives, though the lack of tracking data limited the depth of analysis. Including tracking data would have provided richer insights into player movement and off-ball actions, offering a comprehensive understanding of team tactics and performance.

6.2 Summary of General Conclusions

Passing Strategies and Team Performance

The analysis of Portugal's passing network revealed a strategy centered on short, controlled passes through the midfield to maintain possession, complemented by long diagonal passes from defenders to exploit spaces on the flanks. J. Moutinho and R. Guerreiro were pivotal in connecting defense to attack, highlighting the importance of network centrality in key midfield and wide players. The findings

emphasize positional play in modern football, where utilizing wide players to stretch the opposition is crucial for maintaining team structure and controlling the game's tempo.

Event Sequence and Tactical Patterns

The study uncovered tactical patterns during various match phases by clustering match events. Key moments such as duels, fouls, or possession turnovers, were identified as momentum-shifting points. Winning central duels often led to counter-attacks, while defensive fouls helped Portugal reset their shape and disrupt the opposition. This demonstrates the fluid nature of tactical phases in football, where teams must adapt to maintain control. The strategic use of fouls and duels is crucial in these adaptations.

Shooting Behavior and Goal Outcomes

The xG models showed that shots taken inside the penalty area had a significantly higher chance of success. Tactical decisions, like creating chances through wide-area crosses or forward passes into the box, were essential in generating high-probability shots. Forwards like C. Ronaldo consistently positioned themselves in high-scoring zones, underscoring the importance of shot selection and positioning. The findings suggest that improving shot selection, especially by reducing low-probability long-range efforts could increase scoring efficiency. Overall, this dissertation enhanced the understanding of football dynamics, showing how passing networks, event sequences, and shot selection impact team performance. These insights offer valuable guidance for coaches, analysts, and researchers looking to improve their tactical approaches.

6.3 Implications of the Conclusions

Practical Implications for Football Teams

A major takeaway is the importance of passing network centrality in maintaining possession and controlling match tempo. Teams can develop training strategies focusing on central players, like midfielders and full-backs, effectively linking defense and attack. By optimizing passing routes and positioning key players in critical areas, teams can improve ball retention and create more efficient attacking opportunities. The insights related to duels and fouls also guide managing high-pressure situations. Winning key duels in central areas allows teams to transition quickly into attack, exploiting the opposition's disorganization. Similarly, using fouls strategically in non-threatening areas can help slow the game down allowing teams to regain defensive shape. Findings on passing strategies and shooting behavior have broader applications beyond Portugal's approach. Teams in other leagues, such as the Premier League or La Liga, could benefit from focusing on wide players and network centrality to adopt a possession-based style. In leagues where teams frequently switch between high-pressure, and counter-attacking tactics, optimizing player roles, especially by using midfielders as key transitional players, can improve team performance. Furthermore, teams can refine their attacking strategies by concentrating on high-probability shooting zones, as shown by C. Ronaldo's positioning, and reduce reliance on low-probability long-range shots. This research provides a framework for tactical refinement that is globally applicable, particularly in managing possession and creating high-quality scoring chances across various football contexts.

Broader Strategic Implications

Beyond tactical applications, these findings contribute to a broader understanding of how data analytics can shape football strategy at both team and league levels. As more teams embrace data-driven decision-making, this research offers valuable insights into how passing and shooting analytics can be integrated into real-time match analysis. By focusing on aspects like duels, fouls, and pre-shot events, teams can

develop more adaptive strategies that respond to match conditions, allowing for greater tactical flexibility. These insights can also influence player recruitment and development, helping teams identify players who excel in key roles within the passing network or consistently position themselves in high xG zones. Expanding these techniques to other leagues and competitions can contribute to global trends in football analytics, offering a standardized framework for evaluating tactical success.

Theoretical Contributions

This dissertation contributes to sports analytics by integrating machine-learning models with event sequence data. Using clustering techniques to analyze passing patterns and tactical phases adds a new dimension to understanding how teams adjust their strategies during matches. The focus on duels highlights their often-overlooked role in shaping match outcomes. This research provides a more comprehensive view of match dynamics by extending event-based football analytics beyond passing networks to include non-passing actions like duels and fouls. The enhanced xG model, incorporating pre-shot events like pass types and positioning, adds to the literature on shooting behavior. By showing that pre-shot decision-making significantly impacts shooting efficiency, this research offers a deeper understanding of goal-scoring opportunities, influencing future developments in football analytics.

Broader Theoretical Implications

This research advances the understanding of tactical event sequences and creates a framework for integrating machine learning in real-time analysis. As football increasingly adopts data-driven approaches, this work underscores the potential of predictive models to support in-game decision-making, enabling coaches to adjust tactics based on real-time feedback. Combining machine learning with event sequence data opens new opportunities for studying other sports where event sequences and player interactions significantly impact performance. Moreover, by incorporating non-passing actions into event-sequence analysis, this research expands the theoretical framework of football analytics, suggesting that future studies could apply similar methods in contexts such as basketball or hockey, where off-ball movements and player interactions also play a crucial role in outcomes.

6.4 Proposals for Further Work

Expanding Data Sources

A key limitation of this dissertation was the reliance on event data without including tracking data. Future research could incorporate tracking data to offer deeper insights into player movement, off-ball reactions, and spatial dynamics during matches. Tracking data would provide a more comprehensive view of how players create space, contribute to build-up play, and interact with opponents. This would allow for a richer analysis of positional play and off-ball movement essential for understanding tactical decision-making. Tracking data could also reveal how spatial interactions between players impact team strategies, especially during transitional play phases. Integrating this data type would enable future studies to develop advanced models that capture the full complexity of football dynamics, making the findings more applicable to real-time tactical adjustments and overall team management strategies. A dissertation flowchart can be seen in Figure 6.4(1a).

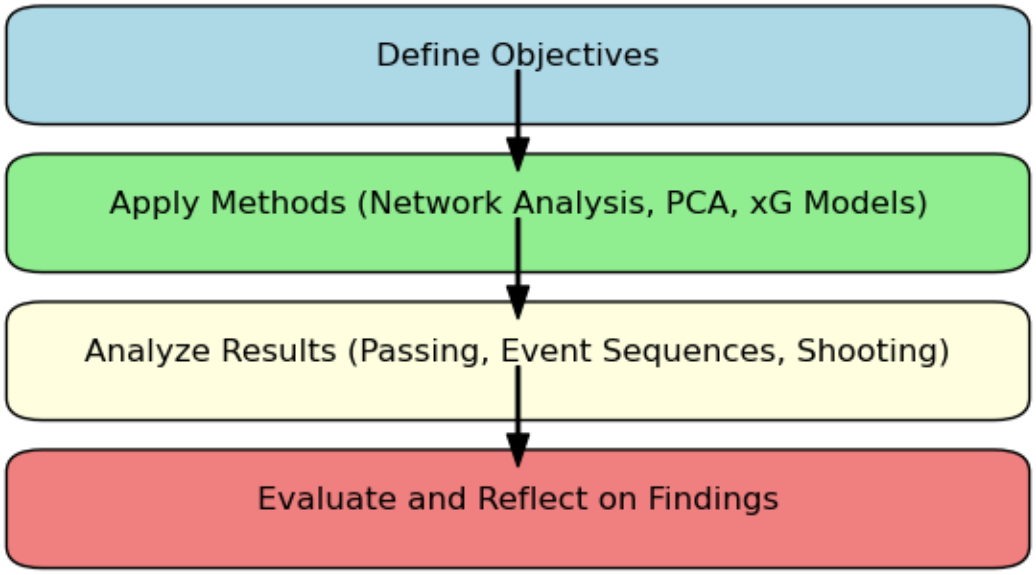


Figure 6.4(1a) Research Process Flowchart: From Objectives to Findings

Broader Data Analysis

Further research could expand the analysis to include multiple matches, teams, and competitions, offering a broader understanding of how different tactical approaches, formations, and opposition strategies affect match outcomes. Analyzing data from various leagues and competition levels could help identify trends, such as how teams adapt their passing strategies when facing stronger opposition or pressure. Expanding the dataset across different regions, including non-European leagues, would provide a more global perspective on football analytics, highlighting regional influences on tactical strategies. A larger dataset would also enhance the robustness of machine-learning models, improving their predictive accuracy of match outcomes by incorporating diverse tactical approaches. This could lead to identifying regional or competition-specific tactical patterns, contributing to a more holistic understanding of football strategies worldwide.

Real-Time Game Analysis

An additional area for exploration is the development of real-time game analysis tools that track passing networks, duels, and fouls as they occur. Real-time data would enable teams to adjust strategies mid-match responding to opposing tactical shifts. Future studies could explore how predictive models can be adapted for in-game decision-making, helping coaches and analysts optimize their tactics during high-pressure moments. By integrating real-time data with machine learning models, teams could receive instant feedback on player performance, tactical shifts, or weaknesses in the opposition's strategy. This could transform in-game coaching by providing actionable insights and enabling dynamic adjustments that give teams a competitive edge. Furthermore, real-time data could inform substitution strategies and formation changes considering ongoing match developments, allowing teams to make data-driven decisions in high-stakes situations.

Broader Strategic Implications

The potential integration of real-time game analysis tools extends beyond football by applying these methods across various sports, such as basketball, hockey, or rugby, where similar event-sequence data is available. Developing sport-specific models could enhance the generalizability of techniques used in this research, fostering real-time tactical decision-making tools across different performance-based industries.

6.5 Reflective Section

Reflecting on the research process, several key lessons emerged. One of the main challenges was the limitation of the event data, particularly in analyzing off-ball movements and player positioning. While event data provided valuable insights into on-ball actions, the absence of tracking data restricted the analysis of tactical decision-making, especially for players like J. Moutinho, whose influence extends beyond visible actions. Tracking data would have enabled a deeper understanding of off-ball contributions, such as pressing and creating space, allowing for a more nuanced analysis of team dynamics during defensive or transitional play. If I were to start this project again, I would emphasize incorporating additional variables in the machine-learning models, such as player fatigue, weather conditions, and match context. These factors can significantly impact performance and tactical decisions, offering a more holistic view of match outcomes. Furthermore, managing large datasets and machine learning models presented challenges, particularly in data processing and feature engineering. A more streamlined workflow would have saved time and allowed deeper exploration of additional variables. The project deepened my understanding of football tactics and the importance of data-driven analysis in sports. Applying advanced techniques such as network analysis, clustering, PCA, and machine learning, has offered valuable insights into how teams can optimize performance. In future research, I would prioritize integrating real-time analytics and predictive modeling to support in-game decision-making. Developing tools that track real-time data and combining predictive models, would enable dynamic tactical adjustments based on ongoing match events, providing coaches with actionable insights during matches.

6.6 Conclusions

In conclusion, this dissertation successfully addressed the research objectives and provided new insights into football tactics and performance analysis. This research has deepened the understanding of passing networks, event sequences, and shooting behavior in football, by applying network analysis, clustering, PCA, and machine learning models. The findings offer practical implications for coaches and analysts, presenting strategies to optimize passing, shooting, and defensive actions. Additionally, this research makes theoretical contributions to sports analytics, particularly through using machine learning to enhance traditional models like xG. While the study had limitations, including its reliance on event data and the focus on a single match, it opened new avenues for further exploration. Future research could benefit from integrating tracking data and real-time analysis tools to expand the scope of football analytics. Overall, this dissertation contributes valuable insights into football performance analysis and lays the groundwork for future research.

References

Andrienko, G. *et al.* (2021). 'Constructing Spaces and Times for Tactical Analysis in Football', *IEEE Transactions on Visualization and Computer Graphics*, 27(4), pp. 2280–2297. Available at: https://doi.org/10.1109/TVCG.2019.2952129.

Araújo, L. *et al.* (2019). 'An experimental analysis of deepest bottom-left-fill packing methods for additive manufacturing', *International Journal of Production Research*, 58, pp. 1–17. Available at: https://doi.org/10.1080/00207543.2019.1686187.

Bialkowski, A. *et al.* (2014). 'Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data', in *2014 IEEE International Conference on Data Mining*. *2014 IEEE International Conference on Data Mining (ICDM)*, Shenzhen, China: IEEE, pp. 725–730. Available at: https://doi.org/10.1109/ICDM.2014.133.

Borrie, A., Jonsson, G. and Magnusson, M. (2002). 'Temporal pattern analysis and its applicability in sport: An explanation and exemplar data', *Journal of sports sciences*, 20, pp. 845–52. Available at: https://doi.org/10.1080/026404102320675675.

Bunker, R. and Thabtah, F. (2017). 'A Machine Learning Framework for Sport Result Prediction', *Applied Computing and Informatics*, 15. Available at: https://doi.org/10.1016/j.aci.2017.09.005.

Clemente, F. *et al.* (2014). 'Using network metrics to investigate football team players' connections: A pilot study', *Motriz. Revista de Educação Física*, 20, pp. 262–271. Available at: https://doi.org/10.1590/S1980-65742014000300004.

Coutinho, D. *et al.* (2022). 'Clustering ball possession duration according to players' role in football small-sided games', *PLOS ONE*, 17, p. e0273460. Available at: https://doi.org/10.1371/journal.pone.0273460.

Decroos, T. *et al.* (2018). 'Actions Speak Louder Than Goals: Valuing Player Actions in Soccer'. Available at: https://doi.org/10.48550/arXiv.1802.07127.

Duch, J., Waitzman, J. and Amaral, L. (2010). 'Quantifying the Performance of Individual Players in a Team Activity', *PloS one*, 5, p. e10937. Available at: https://doi.org/10.1371/journal.pone.0010937.

Fathima S J, S., Sumathi, V.P. and Sumanth, S. (2018). 'Data analytics in football sport to identify gaps for the improvement of quality opportunities throughout world-wide teams', *International Journal of Recent Technology and Engineering*, 7, pp. 364–368.

Frencken, W. *et al.* (2012). 'Variability of inter-team distances associated with match events in elite-standard soccer', *Journal of sports sciences*, 30, pp. 1207–13. Available at: https://doi.org/10.1080/02640414.2012.703783.

Hughes, M. and Franks, I. (2005). 'Analysis of passing sequences, shots and goals in soccer', *Journal of Sports Sciences*, 23(5), pp. 509–514. Available at: https://doi.org/10.1080/02640410410001716779.

Inan, T. (2020). 'The Effect of Crowd Support on Home-Field Advantage: Evidence from European Football', *Annals of Applied Sport Science*, 8, pp. 7–16. Available at: https://doi.org/10.29252/aassjournal.806.

Introducing a Possessions Framework (no date). *Stats Perform*. Available at: https://www.statsperform.com/resource/introducing-a-possessions-framework.

Kempe, M. *et al.* (2014). 'Possession vs. Direct Play: Evaluating Tactical Behavior in Elite Soccer', *International Journal of Sport Science*, 4, pp. 35–41. Available at: https://doi.org/10.5923/s.sports.201401.05.

Lago-Peñas, C. and Dellal, A. (2010). 'Ball Possession Strategies in Elite Soccer According to the Evolution of the Match-Score: the Influence of Situational Variables', *Journal of Human Kinetics*, 25(2010), pp. 93–100. Available at: https://doi.org/10.2478/v10078-010-0036-z.

Lang, S. *et al.* (2022). 'Predicting the in-game status in soccer with machine learning using spatiotemporal player tracking data', *Scientific Reports*, 12. Available at: https://doi.org/10.1038/s41598-022-19948-1.

Lepschy, H., Wäsche, H. and Woll, A. (2018). 'How to be Successful in Football: A Systematic Review', *The Open Sports Sciences Journal*, 11, pp. 3–23. Available at: https://doi.org/10.2174/1875399X01811010003.

Link, D., Lang, S. and Seidenschwarz, P. (2016). 'Real Time Quantification of Dangerousity in Football Using Spatiotemporal Tracking Data', *PLOS ONE*, 11, p. e0168768. Available at: https://doi.org/10.1371/journal.pone.0168768.

Liu, G. and Schulte, O. (2018). *Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation*, p. 3448. Available at: https://doi.org/10.24963/ijcai.2018/478.

Mackenzie, R. and Cushion, C. (2012). 'Performance analysis in football: A critical review and implications for future research', *Journal of sports sciences*, 31. Available at: https://doi.org/10.1080/02640414.2012.746720.

Mchale, I.G., Scarf, P. and Folker, D. (2012). 'On the Development of a Soccer Player Performance Rating System for the English Premier League', *Interfaces*, 42, pp. 339–351. Available at: https://doi.org/10.2307/23254864.

Moura, F. *et al.* (2013). 'A spectral analysis of team dynamics and tactics in Brazilian football', *Journal of sports sciences*, 31, pp. 1568–1577. Available at: https://doi.org/10.1080/02640414.2013.789920.

Narayanan, S., Kosmidis, I. and Dellaportas, P. (2021). *Flexible marked spatio-temporal point processes with applications to event sequences from association football*. Available at: https://doi.org/10.48550/arXiv.2103.04647.

Pappalardo, L. *et al.* (2019). 'A public data set of spatio-temporal match events in soccer competitions', *Scientific Data*, 6. Available at: https://doi.org/10.1038/s41597-019-0247-7.

Pollard, R. (1986). 'Home advantage in soccer: A retrospective analysis', *Journal of sports sciences*, 4, pp. 237–48. Available at: https://doi.org/10.1080/02640418608732122.

Qu, Y. *et al.* (2002). *Principal Component Analysis for Dimension Reduction in Massive Distributed Data Sets*, *Knowledge and Information Systems - KAIS*.

Robertson, S., Back, N. and Bartlett, J. (2015). 'Explaining match outcome in elite Australian Rules football using team performance indicators', *Journal of Sports Sciences* [Preprint]. Available at: https://doi.org/10.1080/02640414.2015.1066026.

Routley, K. and Schulte, O. (2015). 'A Markov Game model for valuing player actions in ice Hockey', pp. 782–791.

Scarf, P., Khare, A. and Alotaibi, N. (2021). 'On skill and chance in sport', *IMA Journal of Management Mathematics*, 33. Available at: https://doi.org/10.1093/imaman/dpab026.

Stefano, E. *et al.* (2020). 'Decision Trees for the Prediction of Outcome of Soccer Games - Historical Data Analysis', *Brazilian Journal of Development*, 6, pp. 4719–4732. Available at: https://doi.org/10.34117/bjdv6n1-339.

Tataru, S.R. and Tataru, I. (2020). 'Privacy & Data Protection in Sport Industry', *SPORT AND SOCIETY* [Preprint]. Available at: https://doi.org/10.36836/2020/1/12.

Vermeulen, E. and Sarma, V. (2018). *Big data in sport analytics: applications and risks*.

Vidal-Codina, F. *et al.* (2022). 'Automatic event detection in football using tracking data', *Sports Engineering*, 25. Available at: https://doi.org/10.1007/s12283-022-00381-6.

Wei, X. *et al.* (2013). *Large-Scale Analysis of Formations in Soccer*, *2013 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2013*, p. 8. Available at: https://doi.org/10.1109/DICTA.2013.6691503.

Dataset Link: https://figshare.com/collections/Soccer_match_event_dataset/4415000/2

Appendices

Appendix A – Dissertation Proposal

Dissertation Proposal

Dynamic Patterns in Ball Possession: Analyzing Similarity and Unveiling Typical Sequences in Football

Andreas.Ioannides@city.ac.uk

Supervisor: Gennady Andrienko

1) Introduction

Purpose of work

The dissertation aims to investigate and visualize passing patterns in football matches, providing coaches and managers with data-driven insights to assess team strategies and player performance. The publicly available datasets for the top 5 European football leagues for a season, this study will employ various analytical techniques, including supervised and unsupervised learning, and diverse methods for feature extraction and pattern recognition. Analyzing football quantitatively has been complex due to its dynamic nature and limited data availability. However, recent datasets offering comprehensive time-series data at the coordinate level, and growing applications of data science techniques, present a valuable opportunity to advance football analytics.

Overall objectives:

The primary objectives of this dissertation are to develop a comprehensive framework for extracting, analyzing, and visualizing passing data in football. Specifically, the study aims to

- Extract pass data by collecting and filtering all pass events for selected teams, from the five major European football leagues over a season.
- Calculate pass attributes by deriving additional metrics such as pass length, vertical changes, horizontal changes, and pass angles.
- Perform a contextual analysis by analyzing contexts such as home and away games, the strength of opponents, and match events that affect passing networks.
- Develop visualization tools to represent passing networks and their changes over time and across different contents.

Products of the work

- Accessing the publicly available dataset of passes with enriched attributes such as pass length and angles.
- Reconstructed passing networks for selected teams, illustrating how different types of passes are distributed across the football pitch.
- Gain contextual insights by analyzing reports detailing how various contexts influence passing behavior.

- Create interactive visualization dashboards to explore passing networks for different teams and players, and their changes dynamically.

Beneficiaries

The primary beneficiaries of this research are **football analysts and coaches**, as insights from passing networks can help them understand team dynamics and devise more effective strategies. **sports scientists** can use detailed analysis to contribute to performance optimization and tactical decision-making research. **football players** can adjust strategies and improve performance by understanding the impacts of different contexts on passing networks. Enhanced visualizations can provide deeper and more engaging insights for **football enthusiasts and broadcasters**.

Project scope

The scope of this dissertation is defined to ensure that topic objectives are realistic and achievable. Key aspects include **analyzing one team from each of the 5 major European football leagues** (Premier League, La Liga, Serie A, Bundesliga, Lige 1). The analysis will cover all games played in a **single season.** The research will initially focus on the different types of **passing events** in football, considering **contextual factors** such as match location (home and away games), opponent strength, and in-game events such as (fouls, and red cards).

2) Critical Context

Literature Review

Passing Network Analysis in Football

A network analysis of a football team's passing dynamics.(Cintia, Rinzivillo and Pappalardo, 2015) emphasizes the significance of network theory in analyzing football tactics. By examining passing networks, the researcher identified key players and pass patterns that contribute to a team's overall performance. The methodology used, including constructing and analyzing passing networks, provides a foundation for this dissertation. Insights from this study will inform the construction of passing networks in this research.

Contextual influence on passing

Evaluating the performance of football teams based on passing networks consisting of different types of passes. (Zhou *et al.*, 2023) explores how contextual factors such as match location (home games and away games) and the opponent's strength, influence team performance. This research highlights the importance of considering event factors when analyzing football networks. The dissertation will incorporate match location and opponent strength to examine their impact on passing networks. Below is a network analysis of the English Premier League side Everton, which I can use to compare Everton's passing networks against West Ham United's passing networks, which I will construct while doing this research.

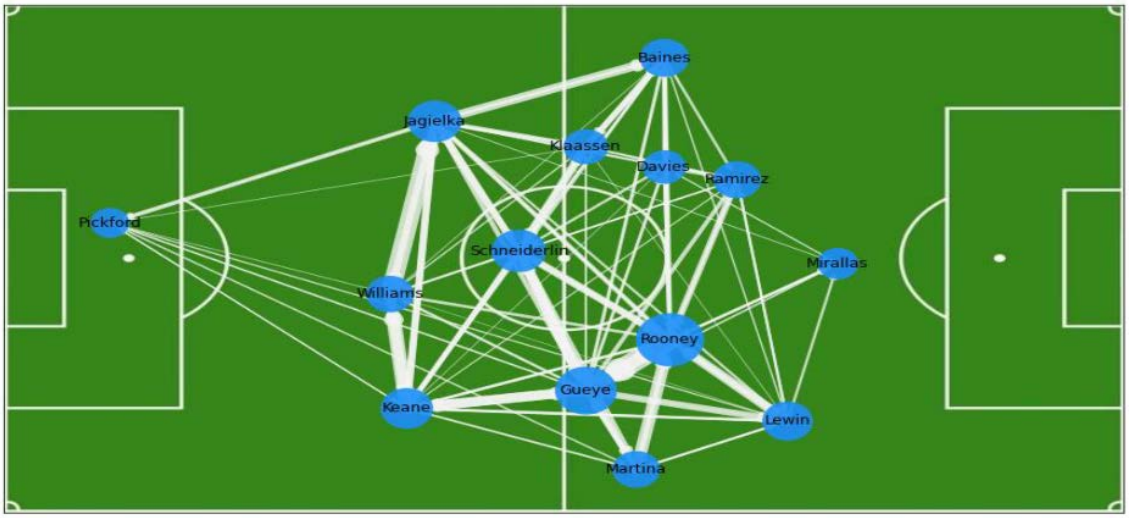


Fig.2 Passing networks of Premier League side Everton during the first round of the season.

Data-Driven Performance Analysis

The approach of big data in sports science and the overall challenges of football. (Rein and Memmert, 2016) discusses the potential and challenges of big data for tactical analysis in football. It highlights the importance of data accuracy, integration, and advanced analytical techniques to derive meaningful insights. The challenges and methodologies outlined will inform the data collection and analysis process. Ensuring data accuracy and employing advanced analytical techniques will be crucial for producing reliable results.

Statistical Methods in Sports Analysis

Analyzing different football strategies in event competitions.(James, Mellalieu and Hollely, 2002) provides methodologies for statistical analysis in sports, emphasizing the importance of robust data analysis techniques to understand team strategies. The statistical methods from this study will be adapted to analyze passing networks, employing techniques such as multivariate analysis to understand the relationship between different variables and passing behavior.

Visualization Techniques

Showcasing sports visualization techniques(Du and Yuan, 2021) highlights various visualization techniques in sports analytics, emphasizing the importance of effective visualizations in communicating and extracting complex data insights. The visualization techniques used, including heatmaps and scatter plots to construct interactive visualizations for passing networks, will help communicate the results to stakeholders.

Relevance and Application of Literature

Network theory is central to analyzing passing patterns and identifying key players. The contextual analysis work will guide the incorporation of contextual factors, ensuring a comprehensive analysis of how different situations impact passing networks. Discussions on big data challenges and methodologies will inform the data processing, and analytical techniques

used in this dissertation. Statistical approaches will be employed to perform robust data analysis, ensuring the validity and reliability of findings. Insights on effective visualization techniques will guide the creation of informative visual tools to present the analysis of results. The selection and review of these documents demonstrate a systematic approach to understanding the current state of research in football analytics. Each piece of literature has been evaluated for relevance and application to the research question. The sophisticated arguments and methodologies presented in these studies will be synthesized and applied to the task, ensuring this dissertation is grounded in high-quality academic and technical references. The selected literature represents recent advancements and ongoing discussions in sports analytics. By integrating these insights, this dissertation aims to contribute original findings to understanding passing networks in football. The relevance of the sources ensures that the research remains current and addresses the latest challenges and opportunities in the field.

3) Approaches: Methods and Tools for Design, Analysis and Evaluation

Data Collection and Preparation

The primary data source for this dissertation will be publicly available datasets from the top 5 European football leagues: Premier League, La Liga, Serie A, Bundesliga, and Ligue 1. These datasets provide detailed event-label data for each match, including player movements and actions captured at the coordinate level. This data source enables in-depth analysis of passing patterns and player interactions. To manage the project scope, the analysis will focus on one team from each league for a season. This targeted approach ensures that the volume of data remains manageable while allowing the exploration of patterns across different teams and contexts. The initial data preparation step involves data filtering to extract passing events, including essential attributes like the start and end coordinates of passes, the player initiating the passes, and the player receiving the passes. This isolate relevant data needed for subsequent analysis.

Additional attributes will be calculated such as:

- **Pass length:** Euclidean distance between start and end coordinates of passes.
- **Vertical change:** Difference in the vertical position (y-coordinate) of the ball from the start to the end of the pass.
- **Horizontal change:** Difference in the horizontal position (x-coordinate) of the ball from the start to the end of the pass.
- **Pass angles:** Pass angle relative to the horizontal axis, providing insight into the directionality of the pass.

Passing Network Construction

Using network theory, passing networks will be constructed for the selected teams, where nodes represent players, and edges represent passes between players. Each edge will be annotated with attributes like the number of passes, average pass length, and changes in ball possession. This allows for a detailed analysis of the team's passing dynamics.

To add depth to the analysis, contextual information will be integrated into the passing networks. This includes:

- **Match location:** Home and away games examining location effects on passing behavior
- **Opponent Strength:** Passes against top-tier and bottom-tier teams to understand competitive level influence.
- **Time and Score of the Game:** How match timing (1^{st} and 2^{nd} half) affects passing patterns, such as more conservative passing strategies when leading.
- **Pass Position:** Sequence within possession, (simple pass, high pass) to identify critical moments in play.
- **Player roles:** Differentiating passes by defenders, midfielders, and attackers to highlight role-specific behaviors.
- **Passes Between Players:** Granular analysis to reveal individual tendencies and interlayer dynamics.
- **Directionality:** Distinguish A→B and B→A passes to understand reciprocal relationships.
- **In-game events:** Impact of fouls, red cards, and goals on passing networks.

Defining and justifying game episodes

Episodes are defined using a combined approach that incorporates possession-based and stoppage-based criteria. An episode begins when a team gains possession and ends when they lose possession, or when a stoppage occurs such as a foul or offside. This method ensures episodes reflect continuous and contextual changes, providing comprehensive match segmentation. Possession-based episodes capture game flow, analyzing team buildup and transitions, while stoppage-based reflect momentum changes due to natural breaks.

Sequential pass analysis

Analyzing passing sequences within episodes reveals team strategies and player interactions. The first pass sets the stage, intermediate passes maintain possession and build up play, and the final pass often leads to goal-scoring opportunities or significant changes in play. Statistical analysis will compare average event times and the significance of first, intermediate, and last passes. This analysis highlights the distinct roles of passes within an episode providing insights into team dynamics and strategies.

Centrality and Network Metrics

To identify key players and understand their roles within the passing network, various centrality measures will be calculated(Korte *et al.*, 2019)

- **Degree Centrality:** Measures the number of direct connections each player has, highlighting those most involved in passing plays.
- **Betweenness Centrality:** Identifies players who act as bridges, facilitating ball movement between different parts of the team.

- **Closeness Centrality:** Assess how close a player is to all other players, indicating their ability to interact quickly with the others.
- **Eigenvector Centrality:** Identifies influential players, considering their direct connections and the influence of neighbors.

Clustering and Embedding Analysis

K-means and Hierarchical Clustering: These algorithms will identify groups of players with similar passing attributes, revealing distinct playing styles and strategies. Hierarchical clustering will be visualized with a dendrogram, showing relationships between players.

PCA, UMAP, and t-SNE Embedding: These methods reduce data dimensionality, and move data to a lower dimensionality point, to reveal hidden patterns and structures of the data.(Luz, 2017) transformed data will be used for hierarchical clustering, to identify clusters.

Similarity Metrics: Metrics like cosine similarity or the Jaccard index will quantify the resemblance between different passing networks.

Visualization Techniques

Heatmaps and Composite Heatmaps: Visualize pass origins and destinations of passes. providing a spatial understanding of team strategies and player movement composite heatmaps can offer a holistic view of passing patterns, showing where passes started and when they ended.

Player and Team-Specific Visualizations

Player and Team Heatmaps: Generate heatmaps for individual players to analyze specific passing patterns and areas of influence, highlighting tactical differences and similarities in passing behaviors.

Network Visualizations

Interactive dashboards and Flow Maps: Tools like Plotly or Dash will create dashboards allowing users to explore passing dynamically, by player, team, and match context. Flow maps will visualize ball movement across different pitch areas, showing the direction and volume of passes.

Ethical Considerations

The study will ensure that all data used is publicly available and doesn't contain any personally identifiable information. Data security measures will be implemented to verify the confidentiality and integrity of the data. The research will maintain transparency in data collection and analysis methods, ensuring that all procedures are replicable and verifiable.

Evaluation

The success and impact of this dissertation will be evaluated using a multi-faced approach focusing on data accuracy, analytical validity, usability of visualizations, and the practical applicability of the insights. We will ensure the accuracy and reliability of data by cross-

referencing the extracted passing events with data sources to verify correctness. Calculated metrics such as pass lengths and angles will be validated through sample checks and manual calculations to ensure accuracy. Potential limitations of missing or inaccurate data will be acknowledged, and assumptions made during the analysis will be explicitly stated to contextualize the findings. The analytical methods applied to this dataset will be evaluated for their effectiveness in uncovering meaningful patterns in passing data. Statistical tests will be employed to confirm the robustness of findings determining if the observed differences in passing patterns are statistically significant. Comparative analysis with existing studies will be conducted to verify that the results are consistent. Innovation will be integrated into the approach by employing advanced machine-learning techniques for clustering and embedding, allowing for the discovery of nuanced patterns in data. These methods will be designed to apply to real-world scenarios, delivering actionable insights for football coaches, analysts, and other stakeholders. Through these comprehensive and carefully considered approaches, the dissertation aims to provide robust, innovative, and practical insights into football passing patterns, contributing to academic research and practical applications in sports analytics.

4) Work Plan

Figure 3 shows the Gannt chart spanning June-September. The main tasks are under the tasks to be completed section. The bar length shows an approximate time for each task to be completed. The work plan will be adapted in case of periods of deviation from the work plan and tasks.

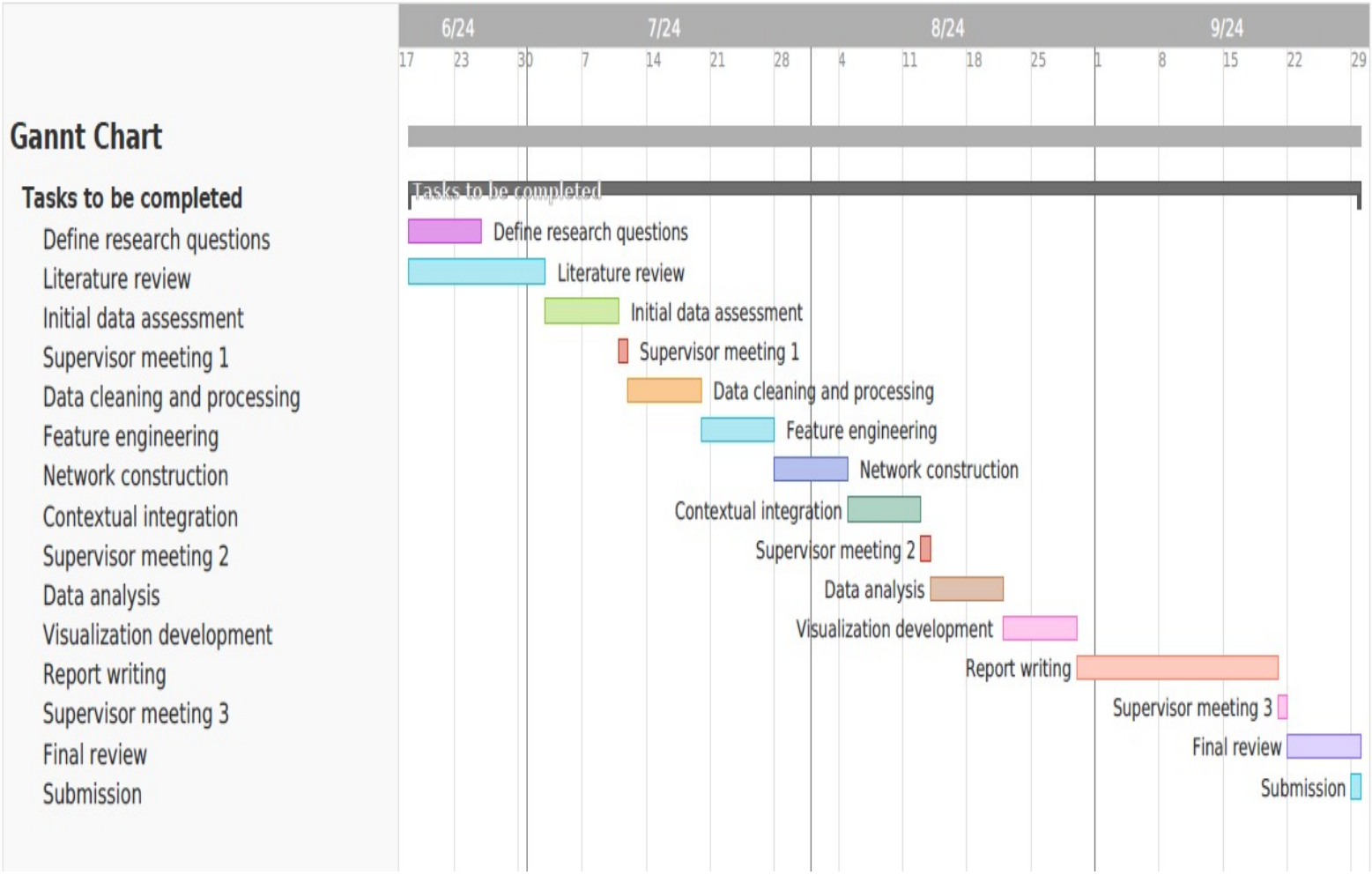


Figure 3. A realistic Gannt chart for the duration of the dissertation.

5) Risk

A risk register is essential for any long-term project. This table shows potential risks that might occur during the 12 weeks, the likelihood of the risks happening, the impact that the risks can have on the project, and possible mitigation strategies to avoid the risks occurring.

Risk Description	Likelihood	Impact	Mitigation Strategy
Data Quality Issues	Medium	High	Perform thoroug data cleaning and validation
Incomplete Contextual Information	Medium	Medium	Include additional data sources or adjust analysis scope
Technical Challenges in Network Analysis	Low	High	Use well-documented algorithms and seek guidance from supervisor
Visualization Clarity	Medium	High	Testing and feedback from supervisor to ensure effectiveness
Scope Creep	Medium	Medium	Strictly adhere to defined project scope and review progress regularly
Software or Tool Failure	Low	High	Maintain backups and have alternative tools ready to use
Data Loss	Low	High	Implament robust data backup and control version systems
Misunderstanding	Medium	Medium	Commuunicate with supervisor about research goals and methods
Time Management	Medium	High	Develop a work plan and monitor progress regularly

Table 1. Risk register documenting potential risks that can occur during the project.

References:

Cintia, P., Rinzivillo, S. and Pappalardo, L. (2015) *A network-based approach to evaluate the performance of football teams*. (Accessed: May 2024)

Du, M. and Yuan, X. (2021) 'A survey of competitive sports data visualization and visual analysis', *Journal of Visualization*, 24(1), pp. 47–67. Available at: https://doi.org/10.1007/s12650-020-00687-2. (Accessed May 2024)

James, N., Mellalieu, S. and Hollely, C. (2002) 'Analysis of strategies in soccer as a function of European and domestic competition', *International Journal of Performance Analysis in Sport*, 2, pp. 85–103. Available at: https://doi.org/10.1080/24748668.2002.11868263. (Accessed: May 2024)

Korte, F. *et al.* (2019) 'Play-by-Play Network Analysis in Football', *Frontiers in Psychology*, 10. Available at: https://doi.org/10.3389/fpsyg.2019.01738. (Accessed: May 2024)

Luz, A. (2017) 'Visualising football players in two dimensions with PCA', *Medium*, 21 December. Available at: https://datalesdatales.medium.com/visualising-football-players-in-two-dimensions-with-pca-92c7bb005ab4 (Accessed: May 2024).

Rein, R. and Memmert, D. (2016) 'Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science', *SpringerPlus*, 5(1), p. 1410. Available at: https://doi.org/10.1186/s40064-016-3108-2. (Accessed: May 2024)

Zhou, W. *et al.* (2023) 'An Improved Passing Network for Evaluating Football Team Performance', *Applied Sciences*, 13(2), p. 845. Available at: https://doi.org/10.3390/app13020845. (Accessed: May 2024)

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.
The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**

2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**