Football Player Role Clustering and Quantification of Role Compatibility

Ignacio Álvarez Carreiro

Student ID: 2342084

M.Sc. Data Science

Supervisor: Dr Sonia Marin

UNIVERSITY OF BIRMINGHAM

School of Computer Science

September 2025

Abstract

This project introduces a role-based framework for tactical analysis and lineup optimisation in football. Using season-long data, players are clustered into roles and evaluated through synergy matrices that measure how combinations affect team output across metrics like xG, pressures, and win rate. In addition, a lineup optimiser builds full XIs based on selected metrics and formation constraints. The system is adaptable to any season dataset and supports applications in recruitment, tactical planning, and match preparation. Future work includes match simulation, squad-based optimisation, and real-world validation.

Honour Code

I certify that this project is my own work. Code development was assisted by ChatGPT 4.0 and 5.0, mainly used to duplicate code for different statistics and debugging. Copilot AI was also used to help with report structure ideas and final editing. All outputs have been reviewed and edited to accurately reflect my work.

Contents

1.0 Introduction

Football analytics has gone through a drastic change over the past decade, moving away from traditional scouting and strict positional labels, as “centre-back,” “striker,” or “central midfielder”, towards data-driven methods. These labels are still useful for organizing teams, but they often struggle when it comes to capturing the individual tactical playing styles that define modern football. **[1]** For example, two central midfielders could vary in their defensive contributions, passing style, and area cover, yet both are labelled identically on a team sheet.

This distance between position labels and actual player behaviour has encouraged researchers to explore more nuanced frameworks for role classification. Advances in dimensional reduction and machine learning, particularly unsupervised clustering techniques, have facilitated the categorising of players into tactical roles.

However, understanding individual roles is only part of the puzzle. Effective team construction requires insight into how these roles interact, what is referred to as role compatibility. While some studies have explored player chemistry and team synergies, few have developed models that quantify how well different roles function together on the pitch.

This project takes those ideas further. Using detailed event data from StatsBomb’s Open Data, specifically the 2015/16 season across Europe’s top five leagues, we apply dimensionality reduction and clustering to redefine player roles within their traditional positions. **[2]** From there, we build compatibility matrices that measure things like win rate, minutes played together, and role co-occurrence. Finally, we bring it all together with a lineup optimization tool that selects players based on tactical fit, not just positional labels.

By combining descriptive analytics with practical decision-making tools, this project aims to support football industry decision-making with data-based methods.

2.0 Literature review

In this chapter we examine the current state of football analytics across four key domains: player roles and positional analysis, machine learning and clustering techniques, relevant data sources, and the study of complementary roles in team construction. This review situates the dissertation within a broader academic and applied context, highlighting both foundational work and emerging trends.

2.1 Relevant Data

The foundation of modern football analytics lies in the growing availability of event-level data. StatsBomb's Open Data repository is among the most widely used sources, offering detailed records of passes, tackles, shots, and defensive actions. Its structure allows for both spatial and temporal analysis, enabling researchers to construct nuanced player profiles. The normalization of performance metrics to per-90 statistics ensures comparability across players with varying match exposure, which is essential for fair evaluation. **[2]**

Complementing StatsBomb's data is Wyscout's glossary, which provides standardised definitions for complex metrics such as progressive passes, reflex saves, and defensive duels that are not explicitly defined by StatsBomb. This standardization is critical to maintain a consistency across studies and for smoothing communication between analysts, coaches, and data providers. **[3]**

Methodological insights into feature engineering are also found in academic literature. Decroos, Bransen, Van Haaren, & Davis (2018) propose a framework for ranking players using multidimensional metrics, including both raw event counts and derived statistics such as possession-adjusted actions. **[4]** This approach emphasizes the importance of selecting features that capture tactical intent rather than merely describing surface-level activity. Additionally, Pappalardo, Cintia, & Ferragina (2019) introduce network-based models that quantify interactions between players, using shared events and spatial proximity to evaluate synergy. **[5]** These relational metrics offer a new dimension to player evaluation, moving beyond individual performance toward team-level dynamics.

2.2 Player Roles and Positional Analysis

Traditional football analysis has always focused on fixed positional labels such as “central midfielder” or “left back,” as if every player behaved this way. However, recent research has challenged this idea, arguing that players within the same position label often show very different tactical behaviours and responsibilities. Hallberg (2022) examines the relevance of positional labels in modern football, using match data to demonstrate that variation within the same position. **[1]** His work suggests that data-driven approaches to classify players into roles give a much clearer and more useful picture.

Casati (2021) builds on this idea through an unsupervised learning project on GitHub, where he applies clustering algorithms to performance metrics in order to uncover role archetypes. **[6]** His approach groups players into categories such as “ball-playing defenders,” “press-resistant midfielders,” and “wide creators,” based on their statistical profiles. This method reveals tactical patterns that are often hard to find by traditional positional analysis and demonstrates the potential of machine learning to redefine how roles are understood.

Further support for behaviour-based classification comes from Gómez, Lago-Peñas, & Pollard (2023), who use statistical modelling to identify playing styles across multiple leagues. **[7]** Their study emphasizes the importance of contextual metrics, such as pass completion under pressure or defensive actions in transition, in distinguishing player roles. Unlike Casati (2021), which focuses on clustering, Gómez et al. (2023) employ supervised classification techniques to validate their role definitions against expert annotations, adding a layer of interpretability to the model. **[6] [7]**

Educational platforms like Footballizer Academy also contribute by offering examples of tactical roles within each position. **[8]** Their resources highlight the variety of responsibilities that players can have even within the same formation. This reinforces the argument that positional labels are insufficient for tactical analysis.

2.3 Machine Learning and Clustering Techniques

Machine learning has emerged as a central tool in football analytics, particularly for unveiling hidden patterns in player performance. For unsupervised learning algorithms K-Means clustering is one of the most widely used due to its simplicity and scalability. FC Python shows how K-Means can group players based on offensive metrics such as expected assists, key passes, and shot-creation actions. **[9]** The resulting clusters provide distinct attacking profiles, offering insights into tactical deployment and recruitment strategies.

Sari, Nugroho, & Prasetyo (2023) applies clustering validity indexes to assess the quality of their groupings. **[10]** Using silhouette scores and Davies-Bouldin metrics, they evaluate how well players are separated within clusters and how compact each group is. Their study highlights that clustering is not merely a descriptive tool but one that requires careful validation to ensure tactical relevance.

Dimensionality reduction techniques also play a crucial role in clustering workflows. Lopes & Machado (2021), in a study published in PMC, explore the use of UMAP (Uniform Manifold Approximation and Projection) to visualize high-dimensional player data. **[11]** UMAP preserves local structure while reducing dimensionality, making it ideal for identifying subtle differences in player behaviour. Their findings suggest that UMAP outperforms traditional methods like PCA in maintaining the integrity of tactical features during projection.

Academic work such as the Gijs Wijngaard Utrecht University Thesis applies clustering to player data with a focus on stylistic groupings. **[12]** The thesis integrates both performance metrics and contextual variables, such as opposition strength and match location, to refine its role definitions. Steve AQ further illustrates how mixed attributes can be used to define player roles. **[13]** His methodology aligns with the framework proposed by D'Urso, De Giovanni & Vitale (2023), who outline best practices for clustering with mixed data types, including normalization techniques and distance metrics. **[14]**

The Harber Institute (n.d.) also contributes to this field by applying K-Means clustering to FIFA player attributes, demonstrating how even synthetic datasets can reveal tactical tendencies when properly modelled. **[15]**

2.4 Complementary roles and team construction

While individual player analysis is essential, an effective team construction requires an understanding of how roles interact within a tactical system. Bransen & Van Haaren (2020) explore this concept through the lens of player chemistry, analysing historical pairings to identify combinations that consistently yield positive outcomes. **[16]** Their study introduces synergy metrics based on co-occurrence and match results, suggesting that certain role pairings enhance team performance beyond the sum of their parts.

The psychological dimension of team synergy is addressed by Davids and Araújo (2016) in *Frontiers in Psychology*. **[17]** Their work emphasises mutual adaptation and collective behaviour

as key components of tactical success, arguing that team dynamics cannot be reduced to individual attributes alone.

Further research by Ramos et al. (2020) at Sheffield Hallam University applies constraint-led approaches to team construction. **[18]** Their methodology focuses on designing training environments that promote tactical fit between the players. This approach has implications for lineup optimization, suggesting that compatibility should be evaluated in context rather than in isolation.

Together, these studies underscore the importance of modelling role compatibility in football analytics. They advocate for a holistic view of team construction that accounts for relational data, psychological factors, and tactical fit, moving beyond traditional metrics of individual quality.

2.5 Gaps in the literature

Despite the rapid growth of football analytics, there are still many limitations. Many clustering studies apply generic models across all players, failing to account for positional specificity. This can lead to players being misclassified, such as full backs that invert into midfield being grouped with other midfielders. The ideal approach does not just forget about the traditional positional labels, but it combines the old with the new.

Compatibility modelling is another underdeveloped area. While studies like Bransen & Van Haaren, and Pappalardo et al. (2019) introduce promising frameworks, few researchers have quantified role interactions in a way that could be useful to create line ups or build squads. **[5] [16]** Most existing models focus on individual performance, neglecting the relational dynamics that are key to a team's success. Interactive tools that operationalize compatibility metrics are also scarce. Although some platforms offer basic lineup builders, they rarely incorporate data-driven synergy models or allow users to prioritize tactical fit.

Finally, clustering models often lack contextual validation. Many studies evaluate cluster quality using internal metrics but do not test their tactical relevance in real match scenarios. Without this validation, it remains unclear whether data-driven role classifications translate into actionable insights for coaches and analysts. These gaps highlight the need for more tools that integrate positional specificity, compatibility metrics, and real-world validation into football analytics.

3.0 Data and Preprocessing

This section outlines the steps taken to prepare the player dataset used throughout the analysis. This includes the extraction of event data from StatsBomb's Open Data repository, the filtering and categorization of relevant actions, and why these metrics were chosen. **[2]**

3.1 StatsBomb Data

The StatsBomb Open Data repository is one of the most comprehensive publicly available football datasets. It provides detailed event-level information data for matches across multiple competitions, from both men's and women's football. For this project the competitions chosen were the 2015/16 Premier League, La Liga, Bundesliga, Serie A and Ligue 1 seasons. This was because they were the most recent seasons that had match information available for every game in the competition. **[2]**

Statsbomb provides metadata files for competitions, matches, teams and players. These files enable researchers to filter and group data by season, league, or team. It also allows to link individual event specific players and tactical contexts. Each match file contains a chronological sequence of events such as passes, carries or shots. These events are stored with rich metadata, including player and team identifiers, timestamps, spatial coordinates, type and subtype of event, and outcome. This allows for precise reconstruction of match flow and player behaviour.

One of the key advantages of the StatsBomb schema is its consistency and extensibility. Each type of event follows the same format, which makes it easier to compare data from different competitions. The files are stored in JSON format, a widely used data format that is easy to read and process with most programming languages. This makes it straightforward to load the data and use it in custom analysis workflows. The dataset was used to calculate player statistics, such as per-90-minite performance metrics for a fair comparison between players. This setup is both useful for individual player profiling and analysis of how different players might work together.

3.2 Building the Dataset

For each of the five leagues selected from the StatsBomb database, a separate dataset was created. **[2]** These datasets include all players who appeared in match lineups or came on as substitutes. Once the full list of players was compiled, the different performance data was added to each entry to build a complete player profile.

3.2.1 Position Data

The first attributes added to each player entry was their positional data. This was also extracted from the 'Starting XI' and 'Substitution' events within the StatsBomb dataset. **[2]** Each of these events include 'tactics' data, which contains the position assigned to every player at kick off or when entering the game as a substitute.

Throughout a season, players often appear in different positions depending on tactical decisions or squad necessities, such as covering for injuries. To account for this, the position stored for each player was the one they played most frequently throughout the season, labelled as their main position. StatsBomb provides 25 distinct position labels, including roles like 'Goalkeeper', 'Left Wing Back', 'Right Attacking Midfielder', and 'Left Center Forward'. **[2]**

To make the dataset more consistent and easier to work with, each specific position label was mapped to a broader positional category. These detailed roles were grouped into six general categories: 'Goalkeeper', 'Centerback', 'Fullback', 'Midfielder', 'Winger' and 'Striker'. This generalization helped reduce noise in the data and made it easier to compare players across teams and leagues. It also allowed for more meaningful clustering and role-based analysis later in the project, especially when evaluating player compatibility or tactical balance.

In cases where a player's position couldn't be determined, they were classified under "Unknown". These players would eventually be taken off the data by either the minutes filter, as they usually did not have a position due to lack of game time, or manually.

3.2.2 Minutes Data

To calculate how many minutes each player spent on the pitch, match event data was parsed to track when players entered and exited each game. This included both starting players and substitutes. For each match, the exact time of entry and exit was recorded in seconds, and any added time at the first half was also added to players that remained on the pitch through that period. Players who stayed on from kick-off until the final whistle had their minutes calculated based on the last event timestamp in the match.

Additional logic was needed to handle other type of absences in matches beyond substitutions. Red cards were recorded as a player's final timestamp, and any time a player was off the pitch getting medical attention was subtracted to their overall time on the pitch. All playing time across the matches in the season was aggregated and converted to minutes.

To ensure the dataset reflected players with meaningful contributions, a minimum threshold of 810 minutes played was applied. This cutoff is consistent with standards used in football analytics, where thresholds between 810 and 900 minutes are used. This corresponds to a minimum of 10 matches, allowing for a bit of leniency to account for substitutions or injuries in those matches. Players that did not meet this requisite were excluded from further analysis.

3.2.3 Passing Data

Passing metrics were extracted from StatsBomb's event data, checking for every pass event made and received by each player across the season. **[2]** Most pass events were categorized based on the StatsBomb classification of passes. However, long passes and progressive passes were adapted from definitions from Wyscout's glossary. **[3]** All the types of passes were normalised per 90 minutes to make comparisons fair across players with different minutes played. Each pass was also evaluated for completion by checking whether the event had an outcome attached (for example it being intercepted). If no outcome was present, it meant the pass had been completed. The accuracy of each type of pass the player performed was also added to the dataset.

Total Passes & Completion Rate

This is the basic count of how many passes a player attempted, and how many of those were completed. It's a good starting point for understanding the involvement and reliability in possession. Players with high volume and high accuracy are usually central to their team's buildup. Another example of this data being useful would be players with low volume and high accuracy being players that take safer passes and are more conservative.

Progressive Passes

These are passes that move the ball significantly forward toward the opponent's goal. The definition used here comes from Wyscout, with distance thresholds depending on where the pass starts. **[3]** Progressive passes are important because they break lines and help teams advance into dangerous areas. They are a key indicator of a player's ability to move play forward.

Long Passes

Also defined using Wyscout's criteria, long passes are those that cover a lot of ground, either driven across the pitch or lofted over defenders. **[3]** These are ground passes over 45 meters and high passes over 25 meters. Long passes are useful to identify players who initiate transitions or bypass pressure.

Crosses

Crosses are wide deliveries into the box, usually from the flanks. They are tracked as a separate type of pass as they are high-risk, high-reward actions. Completed crosses can lead directly to shots or goals but also tend to have lower accuracy due to defensive pressure and crowded areas.

Key Passes & Assists

Key passes are those that directly lead to a shot, while assists lead to a goal. These are the most direct indicators of chance creation. Tracking them helps identify players who consistently set up scoring opportunities.

Passes Received

This metric counts how often a player received the ball from a teammate. It is useful to understand how players are free, trusted and part of the passing network. High passes received with low passes performed could be players that commit riskier actions on the ball or are the player the team wants to find the most to make things happen.

3.2.4 Shot Data

Shooting statistics were extracted from the shot type events in the StatsBomb dataset. **[2]** Each shot was categorised based on its outcome, body part used, location on the pitch or whether it was taken first-time. All metrics were normalised per 90 minutes, and shot accuracy was calculated by dividing number of shots on target by total shots taken.

Total Shots & Accuracy

This is the basic count of how many shots a player attempted, and how many shots were on target. It is a foundational metric for understanding a player's shooting volume and efficiency. Players with high values on both are goal threats, or players with low volume and high accuracy would be players that might wait for very clear chances to shoot.

Goals & Expected Goals

Goals are the most direct output of shooting actions, while expected goals (xG) provides information on the quality of the shots taken. Comparing goals to xG can highlight overperformance (clinical finishing) or underperformance (poor conversion). Players with high xG per 90 are consistently getting into dangerous positions, while those with high goals but low xG may be scoring from low-probability chances or shooting way less.

First-Time Shots & Headers

These are shots taken without a prior touch, often requiring quick decision-making and technique. First-time shots are useful for identifying players who are comfortable finishing under pressure or in fast-paced situations, such as rebounds or one-touch finishes from crosses. These tend to be players strikers that remain near the box or defenders that crowd the box during set pieces. This is similar to headers, which are shots players take with their head. Headers also give information on players that are aerial threats.

Shots Outside the Box

Shots taken from outside the penalty area (or box) were also separated into their own metric. Players who frequently shoot from distance tend to be players that perform late runs to receive back passes or rebounds or that have the power and accuracy to be dangerous from further away.

3.2.5 Dribbling Data

Dribbling metrics were extracted from StatsBomb's event data by identifying every dribble, carry and foul won event performed by each player across the season. **[2]** Each dribble was evaluated for success based on its outcome, to also calculate dribble success rate. Carries were assessed for whether they qualified as progressive runs using a custom spatial rule based on x-coordinate movement. Again, all metrics were normalised per 90 minutes.

Dribbles & Success Rate

This is the basic count of how many times a player attempted to beat an opponent on the ball, and how often they succeeded. Players with high volume and high success rate are typically confident in 1 on 1 situations and capable of overcoming defensive structures. On the other hand, if a player has low volume but high success it might indicate a more conservative style where they only dribble when needed.

Progressive Runs

Progressive runs were carries that moved the ball significantly forward towards the opponent's goal. The thresholds used were adapted from Wyscout's glossary, with different distance requirements based on the starting position. **[3]** These runs are important because they help break lines and advance play without relying on passing.

Fouls Won

Fouls won were tracked as a measure of how often players draw defensive pressure or force opponents mistakes. Players who win a lot of fouls tend to be tactically smart and hard to dispossess.

3.2.6 Defensive Data

Defensive metrics were compiled from StatsBomb's event-level data by aggregating key actions such as pressures, blocks, interceptions, and duels. **[2]** These events were mapped to individual players and normalised per 90 minutes. The dataset captures both active and passive defensive actions, as well as player discipline.

Clearances, Interceptions, Recoveries & Blocks

These defensive actions reflect the player's ability to disrupt opposition play and regain possession. Clearances remove the ball from danger zones, interceptions cut off passing lanes, recoveries regain control for the team, and blocks prevent shots. They help quantify individual defensive ability.

Pressures & Dribbled Past

Pressures indicate how often a player attempts to close down an opponent in possession and being dribbled past how often this player is beaten in 1 on 1 situations. High pressure volume with low dribbled past rate indicates effective pressing.

Fouls, Discipline & Physicality

Foulds committed, yellow cards and red cards were tracked to assess defensive discipline. A high volume of fouls without a high volume of cards suggests the player commits tactical fouls. However, if the card accumulation is high (especially red cards) it suggests the player arrives late to challenges or might play a bit too aggressively. When it comes to physicality metrics, duels and 50/50s were stored, as well as the success rate. This helps identify players that are duel winners, or that might avoid physical battles.

3.2.7 Goalkeeper Data

Goalkeeper metrics are different as they only apply to goalkeepers, they were also extracted from the StatsBomb's event data and normalised per 90 minutes. **[2]** Each event was categorised

based in its type and outcome. There are shot stopping, ball handling and sweeper keeper actions.

Shot-Stopping: Shots Faced, Saves & Goals Conceded

Shots faced only included shots on target. Saves correspond to the successful actions, whereas goals conceded were the failed actions. These metrics help analyse a goalkeeper's shot stopping ability, which is their core responsibility, to avoid goals.

Ball Handling: Smothers, Collections & Punches

These interventions reflect how goalkeepers deal with loose balls or crosses. Smothers involve diving on the ball to prevent further play. Collections and punches are the solutions to aerial balls or crosses, collections being catches and punches acting as clearances. These goalkeeping actions help identify more proactive keepers that try and prevent shot situations.

Sweeper Actions: Claims and Clears

Sweeper keeper actions were defined as goalkeeper interventions outside the box, either claiming the ball or clearing it. These are important to understand which goalkeepers have the ability to leave their box and act as an additional defender.

3.2.8 Involvement Data

Player involvement is quantified by a general count of ball touches, which include positive offensive and defensive actions, and average x and y position on the pitch. These stats were taken from the StatsBomb's event data. These metrics give a general idea on involvement of the player and how wide or high up these players play. In order to compare players that play similar roles but on different sides of the pitch, the y-positions (that corresponded to the width of the pitch) of the players was mirrored across the middle of the pitch. This meant that, for example, a left fullback and right fullback with the same tactical role would get classified together, instead of the clustering algorithm splitting them into right and left sided players.

4.0 Role clustering

Section 4 explains the process of clustering players into roles with the chosen high-dimensional data. The objective was to classify players into tactical roles based on their starting positions and their performance on the pitch. This involved assessing different clustering algorithms as well as dimensional reduction techniques, as well as validating the clusters were meaningful. Each final cluster was then assigned a role label using the 'average player' profile for each group. These labels were based on the Footballizer Academy glossary but serve primarily as nominal tags for reference in later analysis. **[8]**

4.1 Clustering Algorithm

There were two clustering algorithms considered for this project, Hierarchical clustering and K-Means clustering. Hierarchical clustering builds a tree-like structure by merging or splitting clusters based on distance. Although it can reveal nested relationships between the data, it struggled to fit in with what the project needed. This was due to the lack of an 'average player' data point for each cluster, the computational cost with larger datasets and the lack of flexibility of the clustering

K-Means clustering is a partitioning algorithm that assigns data points to a 'k' number of non-overlapping clusters of similar data. K-Means does not just offer that central point for each cluster and handles larger datasets faster, but it also offers a higher flexibility with the clusters. With K-Means the number of clusters can be modified and validated using metrics like silhouette score, as well as just manually selecting the number of clusters if expecting a specific amount of roles. All of this, together with the compatibility with various dimensional reduction techniques, shows why K-Means is regarded as one of the most useful clustering methods in the football analysis industry.

4.2 Dimensionality Reduction

In order to be able to cluster the high-dimensional player data into meaningful roles, dimensional reduction methods had to be implemented. The techniques evaluated were Principal Component Analysis (PCA), t-distributed Stochastic Neighbour Embedding (t-SNE), PCA followed by t-SNE, and Uniform Manifold Approximation and Projection (UMAP). The testing was done by finding the 'k' number of cluster that achieved the highest silhouette score, and then comparing those silhouette scores, as well as the projections themselves and division of players.

PCA reduces dimensionality by projecting data onto orthogonal components that capture the greatest variance . It was initially tested as it is fast, simple and captures variance. However, its

dependence on linear combinations made it struggle to create meaningful clusters when joined with the K-Means algorithm. They often only separated the data into 2 clusters, with a lack of shape and not enough separation between clusters. PCA also achieved the lowest silhouette scores overall. These limitations are shown in the PCA projection of left back data (Figure A1, Appendix A), where the lack of distance and shape between clusters is evident.

t-SNE preserves local structure and can uncover non-linear relationships in the data. The projections produced were clearer than PCA, in number, shape and separation of clusters (Figure A2, Appendix A), as well as having achieved a higher silhouette score. However, t-SNE is computationally expensive and lacks consistency. These limitations that it less suitable than other methods. The third method tested was a combination of PCA and t-SNE. This reduces the runtime of t-SNE and slightly improves the performance of PCA (Figure A3, Appendix A). However, it still kept similar issues with the inconsistency and although the silhouette score got a significant increase, it also struggled to have more than 2 clusters.

UMAP emerged as the most effective method. It preserves both local and global structure, ideal for finding tactical roles with football data. UMAP is also fast, scalable and consistent, obtaining similar results to t-SNE but with less computational cost. The UMAP projections proved to be the most useful, with clear separation and shapes in the clusters, while still keeping the data points close enough to work with different number of clusters (Figure A4, Appendix A).

4.3 Results

This subsection presents the final clustering results for each positional group. For every position, the dimensionality-reduced projections are shown, followed by a breakdown of the clusters formed, the average performance data within each cluster, and the role labels assigned. Since different positions rely on different performance metrics, the features used for each group are also outlined.

4.3.1 Goalkeepers

For goalkeepers, feature selection focused on three key groupings: goalkeeping actions, positional tendencies, and distribution metrics. Shot-stopping and ball-handling data captured the core defensive characteristics of each goalkeeper, while sweeper keeper metrics provided insight into their proactiveness and willingness to leave their goal. Average position data supported this by showing how high up the pitch the keeper would act.

On-ball passing metrics were used to assess comfort and decision-making in possession. Progressive passes and passes received reflected a goalkeeper's involvement in buildup play and their appetite for risk when distributing under pressure. Conversely, long passes and key passes highlighted more direct tendencies, the goalkeepers who skip the buildup phase to target the forward line.

After filtering the data, dimensionality reduction was performed using UMAP. A grid search across multiple hyperparameter values was done to find the best projection for clustering, using the silhouette score. K-Means clustering was applied to the reduced data, with k values ranging from 2 to 4, as more than 4 clusters for goalkeepers would be excessive and lead to overfitting. The best result was achieved with k = 3, a silhouette score of 0.412 (see Figure B1 Appendix B). From the clusters created:

- **Cluster 0 – Shot Stopper:** Higher frequency of saves and 'in goal' actions, with lower sweeper and passing metrics, with an average position closer to their goal.
- **Cluster 1 – Sweeper Keeper:** Highest passing metrics and most sweeper actions, they were also the most involved in play and had the furthest from goal average position.
- **Cluster 2 – Conservative:** Lowest number of passes and goalkeeping actions, but the highest accuracy in passes. This meant that they took low risk on-ball actions.

4.3.2 Centerbacks

The first outfield position to go through the clustering process were the centerbacks. The features used for them were passing, dribbling, positioning and defensive metrics. Passing metrics were used to assess involvement in build-up and comfort on the ball. Dribbling data helped identify defenders that advance the ball through carries and can overcome opponent pressure. Defensive data such as interceptions, blocks, clearances and recoveries were used to evaluate the core defensive abilities of the defenders in one on one actions and shot defending actions. Pressures show how proactive the defender, and when also taking position into account you can find which defenders push up to press. The other side of defensive data would be the discipline and physicality data, that assesses how many duels the centerbacks enter, how many they win and how many fouls they commit. Finally, adding to the duels data, headers (from the shot data) help evaluate the aerial presence of the player.

The clustering method used was the same as for goalkeepers, with the only change of the k range through which the code explores is extended to 3 to 7, as there is a higher volume of centerbacks and centerback roles. The iteration chose the value k = 5, which got a silhouette score of 0.443. The projection is included in Figure B2, Appendix B. From this projection, the 5 clusters were labelled, based on their average data, as:

- **Cluster 0 – Stopper:** High number of clearances, blocks, interceptions and duels with very low passing (often relying on long passes) and ball carrying metrics. Highest value

for headers per 90, hinting at a good aerial presence. Deepest average position on the pitch. Traditional, risk avoiding defender that is the team's last man.

- **Cluster 1 – Sweeper:** Average passing metrics and defensive coverage. These players show a better progression performance, duel win rate and slightly higher positioning than Stoppers. Their role focuses on being an all-round defender that can be the last man but can also go challenge for the ball on duels, with safe ball progression.
- **Cluster 2 – Enforcer:** Marked by having the highest pressing and duel stats, with similarly low passing metrics as the Stopper. Also commits the highest amount of fouls. These defenders tend to be very proactive and aggressive off the ball and commit fouls when getting caught out of position or beaten one on one.
- **Cluster 3 – Libero:** The most progressive and offensive profile. Highest average position, number of passes, carries and touches. Meaning they tend to move the ball up safely and get involved in play higher up on the pitch. They also have the highest amount of shots. They also commit a high amount of fouls, probably tactical fouls when getting caught out of position. Their one on one defensive data is weaker than the other roles.
- **Cluster 4 – Ball-Playing CB:** A balanced profile, combining average defensive metrics with high passing metrics, only beaten by the Libero. These are defenders that are comfortable on the ball and initiate the build-up, but do not push as high up as the Libero. Their average position is not as high as the Libero, but it is the one with most width, meaning they rather find space for the build-up in wider areas of the pitch.

4.3.3 Fullbacks

For fullbacks, feature selection focused on passing, crossing, dribbling, defensive actions, and positional tendencies. The metrics are similar to centerbacks, but with an added passing stat, crosses, as they are key to how the fullbacks generate chances. Headers were removed as aerial presence is not as key to the fullback position as for centerbacks.

The clustering method was the same as previously, with a k range of 2 to 7. Leading to an optimal k value of 3, with a silhouette score of 0.429 (See Figure 3B, Appendix B). This projection and their average data had the clusters as:

- **Cluster 0 – Defensive Fullback:** These fullbacks had the lowest passing involvement, as well as ball carrying and dribbles. However, they had the highest number of blocks, clearances and interceptions, with a high amount of duels and pressures. This and the average position being the closets to the goal on both axis shows these fullbacks preferred to defend in a more conservative manner without pushing to high up and keeping the defensive block compact.
- **Cluster 1 – Balanced Fullback:** These fullbacks preferred to be more proactive in their defending, with the highest amount of pressing and duels. They also had the highest

build-up passing metrics and progressive runs, but no the highest chance-crating passing metrics (crosses and key passes). These fullbacks would push up and stay close to the midfield but not overcommitting to getting too involved in attacking areas, more focused on the build-up and possession recycling.

- **Cluster 2 – Attacking Fullback:** Attacking fullbacks had the lowest defensive contributions, but the highest chance-creating metrics. They had a similar average position to Balanced Fullbacks in terms of how high up the pitch, but much wider placement. They focused more on first and final third actions and less on the build-up.

4.3.4 Wingers

Wingers were analysed using a feature set that captured their dual role as creators and finishers, similar to the fullback one but with the addition of shooting metrics. This meant not just their build-up and chance creating abilities were measured, but also their defensive commitment, attacking output and pitch positioning were considered. The shooting data included expected goal created, as well as the different types of shots taken.

For the clustering iteration, the k value range was set to be between 3 and 7, as the k value 2 gave the highest score but the clusters had little tactical relevance and had more to do with the quality of the players. This meant the final k value was 3, with a silhouette score of 0.472 (see Figure 4B, Appendix B). The three clusters were labelled:

- **Cluster 0 – Defensive Winger:** These wingers have the highest defensive data, specially pressures, and an average involvement in play. Their passes metrics focused more on build-up phases. They also had the lowest shooting metrics and ball carrying metrics, as well as the deepest average position. The Defensive Wingers tend to be more focused on disrupting opponents build up and supporting the fullbacks and midfielders in defensive and build-up duties.
- **Cluster 1 – Direct Winger:** Winger role that has the highest average position, with low involvement in the play, low touches and passes. They also have the low defensive data. Although they have high shooting metrics, they have lower dribbling and crossing data than cluster 2, with a similar crossing accuracy. These wingers tend to be very direct in attack looking for a shot or a cross whenever they can, avoiding slowing the play. This also means they depend on other players to create the chances instead of being self-reliant, thriving in counterattacks and quick transitions.
- **Cluster 2 – Creative Winger:** Creative Wingers have the highest volume of passing and dribbling metrics, especially when it comes to chance-creation metrics such as key passes or crosses. They have solid shooting and defensive metrics and the highest touches per 90, as well as the widest average position. These wingers are balanced, being competent at all aspects of play, but thrive as wide chance-creators or playmakers, needing a high involvement in the game.

4.3.5 Strikers

Strikers were analysed with the same metrics as wingers but not including accuracy of specific passes, to reduce the number of features. Shooting metrics help identify what style of shots they tend to make, where from, and how accurately. Passing and dribbling show how combinative and mobile the strikers are, and how likely they are to create chances for others or themselves. Defensive data can show if they are aggressive on the press or to disrupt opponent's build-up or if they track back and help the midfield after that first line of pressing is broken.

The k range used for the clustering iteration is from 3 to 6, giving k = 3 as the best value with a silhouette score of 0.460 (Figure 5B, Appendix B). The resulting clusters revealed three distinct striker roles:

- **Cluster 0 – Complete Forward:** This group showed the highest involvement across most offensive metrics. Highest passing stats, progressive actions, shots and expected goals. They had overall average to low defensive metrics except for pressing and interceptions. These are the most balanced and complete strikers, with a lot of mobility and that tend to not just stay around the opponent's box, but move around to help in the build-up, find space and create chances.
- **Cluster 1 – Poacher:** The most direct and goal-focused profile, with the lowest involvement. They have the lowest shot volume and very minimal passing, dribbling and defensive stats. They focus more on final-third positioning and finishing rather than all-phase contributions.
- **Cluster 2 – Target Man:** This cluster is a middle point between the previous two in terms of general involvement. Still having low passing, dribbling and more focused on physical battles for the defensive metrics (higher pressing and duel involvement). They also had an xG similar to the Complete Forwards, with shooting metrics just under them for most attributes except headers. This striker is more focused on being a physical presence on both defence and offence but isn't very involved in the build-up.

4.3.6 Midfielders

The midfielder's analysis was done slightly differently. The vast amount of players classified as midfielders, joint with the difference between different midfield positions, meant the clusters would tend to separate them on positional differences more than tactical roles. To overcome this, they were split into four different groups: central midfielders, defensive midfielders, attacking midfielders and wide midfielders.

Despite this split, all groups were clustered using the same set of features, which included all the passing, defensive, dribbling and positioning metrics, as well as only shots per 90, xG per 90

and shots outside the box per 90 from the shooting metrics. This was because other shooting metrics are only key for attackers. All four groups also went through the same iterations, with k ranges from 2 to 6.

For defensive midfielders the ideal k was 4, with a silhouette score of 0.473. The clusters can be seen on Figure 6B, Appendix B. These were:

- **Cluster 0 – Ball Winning Midfielder:** This cluster showed very high defensive intensity across all metrics, with the highest duels and pressures. They had the second lowest passing metrics, mainly focusing on long passes and crosses. In dribbling metrics, they also had high progressive runs and dribble. This role profile is of a deep defensive midfielder that progresses the ball via carries and not passes. They also aren't static in their deep position as they come out and press and get in on one on one defensive situations.
- **Cluster 1 – Positional Midfielder:** Similar defensive stats to the Ball Winning Midfielder, but slightly higher except for pressures and duels. They also play more centrally and deeper than them. This cluster also has slightly better passing metrics and slightly worse dribbling metrics to cluster 0. Positional Midfielders are more static defensive midfielders that get involved in the initial build-up but don't push very high. They also prefer to remain a bit more cautious when defending, without overcommitting.
- **Cluster 2 – Deep Playmaker:** This was the more creative group out of the defensive midfielders. They specialise in initiating attacks and moving the ball around. They had much higher passing metrics and touches than any other defensive midfielder, especially in progressive passes. Defensively, they focus on ball recoveries instead of one on one actions or duels. They have the highest average position and shot number, showing a more offensive intent than the others.
- **Cluster 3 – Anchor Man:** This cluster had the lowest involvement in possession and attacking metrics and focused strictly on defensive actions. They still had low pressures, duels and recoveries. Also having the deepest average position meant this role was more conservative and acted as the last man resort in midfield mainly focusing on defending and maintaining positional discipline, being less involved in play than other roles.

For central midfielders the k value was 3, with a silhouette score of 0.461 (Figure 7B, Appendix B). The roles were:

- **Cluster 0 – Box to Box Midfielder:** These midfielders have high involvement in both attacking and defensive phases. They had average passing metrics and excelled in progressive runs and dribbles. Their defensive data was the highest of the three with very high pressures. This role profile focused on being a high energy player that is aggressive when defending and that can be involved in the build-up but prefers driving the ball forward themselves. They also tend to drift slightly wide.

- **Cluster 1 – Holding Midfielder:** A more reserved and disciplined role, almost a bridging role between central midfielder and defensive midfielder. They had the lowest passing volume and attacking output, but with high accuracy, indicating they avoid risky passes. They had average defensive metrics, also acting conservatively in those actions.
- **Cluster 2 – Playmaker:** This cluster led in all possession and creative metrics. Highest passing volume, progressive passes, key touches and passes. As well as leading in shooting metrics. They had defensive data similar to cluster 1. This profile was more about chance creating from midfield whilst also supporting in defensive actions from slightly higher up. They progressed the ball from defence to offence with passes, dictating the tempo from central areas.

Attacking midfielders had k = 3 as their ideal values, with a silhouette score of 0.466. The clusters were:

- **Cluster 0 – Advanced Playmaker:** This group is the most creative attacking midfielder, with highest chance creating metrics and high possession metrics. They were the most aggressive on the press and managed the most ball recoveries and duels. These players remain centrally and create chances for the attackers in the team, whilst still helping, occasionally, the rest of the midfield with defensive and build-up responsibilities.
- **Cluster 1 – False 10:** A more reserved and support-oriented profile. They had the lowest passing and offensive metrics out of the 3. They are not as high up as the other two roles and their balanced defensive stats indicate they act as a bridge between central midfielders and attacking midfielders, as further forward supporting player, mainly during defensive situations. They facilitate play without dominating it.
- **Cluster 2 – Second Striker:** This cluster is the most attacking midfield, almost acting as an attacker. They lead in shots, xG, progressive runs and dribbles, while also receiving the most passes and operating further forward. They are less involved in chance creation and more involved in being the player that finishes attacks themselves. They are less involved in defensive metrics. They also operate slightly wider than other attacking midfielders, looking for the half-spaces between midfielders, wingers and strikers.

For wide midfielder the k value chosen was k=2, silhouette score of 0.568 (Figure 8B, Appendix B). The clusters were :

- **Cluster 0 – Wide Playmaker:** These wide midfielders were more involved in creative and possession-based metrics. Leading in all types of passes, dribbling and shots. In terms of defensive parameters, they lead the more proactive actions such as pressing and duels. These midfielders were closer to behaving like wingers and were more focused on moving the ball from wide midfield areas to dangerous ones.
- **Cluster 1 – Wide Supporter:** A more reserved and complementary profile. These players had lower passing volume, progressive actions, and touches. But kept a higher crossing

accuracy, probably due to only crossing in mor risk free scenarios. Defensively they had balanced data. This profile was more focused on lending a hand to the fullbacks and the more centered midfielders with maintaining possession, width and defensive cover, with a lower impact on the games.

5.0 Role compatibility analysis

This section goes through the method of evaluating how well different player roles perform together, using both interaction-based metrics and contextual solo performance indicators. The goal is to identify role pairings that elevate team performance and build optimised starting elevens based off these pairings.

The eleven matrices made were split in two groups: interaction-based matrices and individual performance matrices. The first group of matrices focuses on pair-based synergy, capturing direct interactions and team-level outcomes when two roles are on the pitch together. These include minutes played together, win ratio, passes exchanged, key passes and crosses between them, team xG generated, and xG conceded. The second group tracks individual output in context, measuring how a player's solo actions vary depending on which role is playing alongside them. Any matrix using match actions was normalised to per 90 minutes. The full matrices can be found on Appendix C.

5.1 Interaction-Based Synergy Matrices

The interaction-based matrices measure direct exchanges between players (such as passes or chances created), as well as the influence that both players have on team performance (such as win ratio or xG for). These matrices are symmetrical meaning the direction of the action is treated as bidirectional: a pass from role A to Role B is counted for both roles. This approach focuses on relational dynamics.

5.1.1 Minutes Matrix

This matrix stores how many minutes each pair of roles shared the pitch together throughout the 2015/16 season. It goes beyond simple co-occurrence, serving not only as a record of tactical deployment, but also as a foundational layer for normalizing future matrices. It reflects the structural decisions made by coaches: which combinations were most trusted to play together, and which roles formed the backbone of their systems.

The data reveals clear patterns in role usage. Defensive Fullbacks appear prominently across the top pairings, highlighting their tactical versatility and frequent use in both defensive and transitional setups. Their pairing with Ball Winning Midfielders (137,673 minutes) and Enforcers (133,497 minutes) suggests a strong emphasis on ball recovery and compact defensive shape. Similarly, combinations like Stopper & Defensive Fullback (119,289 minutes) and Defensive Fullback & Shot Stopper (116,840 minutes) reflect traditional back-line structures built around stability and low-risk distribution.

On the attacking side, the frequent pairing of Complete Forwards with Attacking Fullbacks (116,837 minutes) and Balanced Fullbacks (113,137 minutes) points to systems that rely on wide support and vertical progression. The presence of Libero & Attacking Fullback (111,311 minutes) also hints at more fluid defensive setups and transitional play.

Top 10 Pairings – Minutes Matrix

1. Ball Winning Midfielder & Defensive Fullback — 137,673 minutes
2. Enforcer & Defensive Fullback — 133,497 minutes
3. Stopper & Defensive Fullback — 119,289 minutes
4. Defensive Fullback & Complete Forward — 117,119 minutes
5. Defensive Fullback & Shot Stopper — 116,840 minutes
6. Complete Forward & Attacking Fullback — 116,837 minutes
7. Balanced Fullback & Complete Forward — 113,137 minutes
8. Enforcer & Balanced Fullback — 111,683 minutes
9. Attacking Fullback & Libero — 111,311 minutes
10. Balanced Fullback & Stopper — 107,032 minutes

5.1.2 Win Ratio Matrix

The win ratio matrix evaluates how often teams won when specific role pairings were present in the starting XI. Unlike the minutes matrix, which reflects tactical frequency, this captures outcome-based synergy, identifying combinations that consistently correlate with match success. For each match, the algorithm tracked the result (win/loss/draw) and counted how often each role pair appeared in winning lineups. Substitutes were excluded to maintain tactical consistency and avoid skewing results with reactive changes.

The matrix reveals a clear trend: Libero-based combinations dominate the top-performing pairs, appearing in 7 of the top 10. The highest win ratio was recorded for Libero & Holding Midfielder (0.795), suggesting a strong tactical backbone built around deep coverage and disciplined midfield support. Other high-performing pairs include Deep Playmaker & Playmaker (0.691) and

Libero & Creative Winger (0.650), pointing to systems that blend vertical progression with defensive fluidity.

Top 10 Pairings – Win Ratio Matrix

1. Libero & Holding Midfielder — 0.795
2. Deep Playmaker & Playmaker — 0.691
3. Playmaker & Libero — 0.667
4. Libero & Deep Playmaker — 0.665
5. Libero & Creative Winger — 0.650
6. Libero & Libero — 0.614
7. Holding Midfielder & Conservative — 0.609
8. Libero & Attacking Fullback — 0.601
9. Complete Forward & Libero — 0.589
10. Libero & Conservative — 0.589

5.1.3 Passing Matrix

This matrix measures the volume of passes exchanged between role pairings, normalized per 90 minutes using the previously computed minutes matrix. By focusing exclusively on starting players, it captures stable tactical relationships rather than reactive or short-term interactions. Each pass event is mapped to the roles of both the passer and receiver, and aggregated across all matches to build a symmetric role-to-role interaction matrix. The result reflects not just positional proximity, but also which roles are actively linking play, facilitating progression, or recycling possession. High pass volumes between roles suggest strong structural connectivity and coordinated buildup patterns

The data reveals a clear centrality of creative and deep-lying roles. The most frequent pairing was Playmaker + Deep Playmaker (30.14 passes per 90), indicating a high-volume distribution axis in midfield. Libero-based combinations also featured prominently, with Libero + Libero (24.09) and Deep Playmaker + Libero (19.87) suggesting fluidity and recycling in deeper zones. Notably, Second Striker pairings with Creative Wingers and Playmakers also ranked highly, pointing to active combination play in advanced areas.

Top 10 Pairings – Passing Matrix

1. Playmaker & Deep Playmaker — 30.14
2. Libero & Libero — 24.09
3. Deep Playmaker & Libero — 19.87
4. Sweeper & Libero — 19.46
5. Playmaker & Positional Midfielder — 19.11
6. Deep Playmaker & Second Striker — 18.41
7. Creative Winger & Playmaker — 17.50
8. Second Striker & Creative Winger — 16.90
9. Second Striker & Playmaker — 16.79
10. Libero & Ball Playing CB — 16.78

5.1.4 xG For Matrix

This matrix measures the average expected goals (xG) generated by a team when specific role pairings were present in the starting XI. Unlike other interaction-based matrices, this one captures team-level attacking output, offering insight into which combinations consistently contribute to chance creation and goal threat. For each match, total team xG was attributed to all role pairs in the starting lineup, weighted by co-occurrence.

The result is a symmetric matrix that reflects not just positional synergy, but attacking productivity. High xG values suggest that the presence of certain role combinations correlates with more dangerous attacking sequences, whether through buildup, final-third movement, or transitional play.

The data reveals a clear trend: Libero-based pairings dominate the top of the matrix, appearing in 8 of the top 10 combinations. The highest average xG was recorded for Playmaker & Deep Playmaker (1.972), highlighting the creative engine of possession-heavy systems. Libero & Playmaker (1.920) and Libero & Creative Winger (1.873) also ranked highly, suggesting that deep distributors paired with advanced creators consistently drive attacking output.

Top 10 Pairings – xG For Matrix

1. Playmaker & Deep Playmaker — 1.972 xG
2. Libero & Playmaker — 1.920 xG
3. Libero & Creative Winger — 1.873 xG
4. Holding Midfielder & Libero — 1.850 xG
5. Libero & Direct Winger — 1.811 xG
6. Libero & Libero — 1.751 xG
7. Attacking Fullback & Libero — 1.737 xG
8. Conservative & Creative Winger — 1.724 xG
9. Libero & Wide Playmaker — 1.722 xG
10. Libero & Deep Playmaker — 1.703 xG

5.1.5 xG Against Matrix

This matrix measures the average expected goals (xG) conceded by a team when specific role pairings were present in the starting XI. Unlike individual defensive metrics, this matrix captures team-level defensive resilience, attributing opponent xG to the roles deployed together. For each match, total xG conceded was assigned to all role pairs in the lineup, weighted by co-occurrence and normalized using the minutes matrix to exclude pairings with zero shared time.

The result is a symmetric matrix that reflects how certain combinations contribute to limiting chance quality. Lower values indicate pairings that consistently suppress opponent threat, whether through compact shape, pressing coordination, or transitional coverage.

The top-performing combinations reveal a consistent theme: Libero-based pairings dominate, appearing in 8 of the top 10. Holding Midfielder & Libero (0.784 xG against) leads the list, suggesting a strong defensive spine built around spatial control and disciplined midfield screening. Other standout pairings include Box to Box Midfielder & Libero (0.884), Deep Playmaker & Libero (0.888), and Libero & Conservative (0.898), all pointing to systems that balance deep distribution with defensive coverage.

Top 10 Pairings – xG Against Matrix

1. Holding Midfielder & Libero — 0.784
2. Box to Box Midfielder & Libero — 0.884
3. Deep Playmaker & Libero — 0.888
4. Libero & Libero — 0.889
5. Deep Playmaker & Playmaker — 0.894
6. Libero & Conservative — 0.898
7. Conservative & Attacking Fullback — 0.910
8. Libero & Attacking Fullback — 0.911
9. Holding Midfielder & Conservative — 0.917
10. Creative Winger & Libero — 0.920

5.1.6 Chance Creating Matrix

This matrix tracks the frequency of chance-creating actions, specifically crosses and shot assists, exchanged between role pairings in the starting XI. By isolating passes that directly lead to shots or deliver balls into dangerous areas, it captures the creative chemistry between roles. Each event is mapped to the passer and recipient's roles, and normalized per 90 minutes using the minutes matrix to ensure tactical relevance.

The result is a symmetric matrix that reflects final-third interaction quality, highlighting which combinations consistently generate scoring opportunities. High values suggest coordinated attacking patterns, positional synergy, and effective spatial exploitation.

The top pairings reveal a clear emphasis on forward-to-winger dynamics. Complete Forward & Creative Winger (1.70 per 90) and Target Man & Creative Winger (1.58) lead the chart, suggesting systems that rely on aerial service, hold-up play, and wide delivery. Combinations like Direct Winger & Complete Forward (1.48) and Second Striker & Complete Forward (1.48) reinforce the importance of vertical movement and central finishing roles. Interestingly, Creative Winger & Poacher (1.45) and Creative Winger & Advanced Playmaker (1.35) show that wide creators also link effectively with both finishers and central playmakers.

Top 10 Pairings - Chance Creating Matrix

1. Complete Forward & Creative Winger — 1.70
2. Target Man & Creative Winger — 1.58
3. Complete Forward & Complete Forward — 1.56
4. Direct Winger & Complete Forward — 1.48
5. Second Striker & Complete Forward — 1.48
6. Creative Winger & Poacher — 1.45
7. Target Man & Complete Forward — 1.43
8. Direct Winger & Direct Winger — 1.42
9. Target Man & Direct Winger — 1.41
10. Creative Winger & Advanced Playmaker — 1.35

5.2 Individual Performance Synergy Matrices

While previous matrices focused on pair-based synergy, this section shifts the lens to individual output in context . They measure how a player's actions vary depending on the role of the teammate they're paired with. These matrices track per-90 rates of key actions such as pressures, interceptions, dribbles, carries, blocks, and ball recoveries, aggregated across matches and normalized using shared minutes. The goal is to isolate indirect influence: which roles tend to unlock others, suppress activity, or reshape individual behaviour within the tactical system. By mapping these patterns, we can find what roles enhance a certain player the most, as well as which roles enhance each other mutually the most.

5.2.1 Shots & Dribbles Matrix

This matrix tracks the combined rate of shots and dribbles performed by a player, normalized per 90 minutes, depending on which role they shared the pitch with. It highlights how certain roles enhance or supress individual attacking output, offering an insight into tactical combinations that unlock the most aggressive behaviour.

For example, when Creative Wingers are paired with a Libero, they produce their highest attacking output (averaging 8.05 actions per 90 minutes). This suggests that the Libero's deep positioning and distribution create space and support for wide attackers to take on defenders and attempt shots. In contrast, the lowest output occurs when Creative Wingers are paired with a False 10 (5.23 actions per 90), indicating possible spatial overlap or reduced service into wide areas.

Another way to analyse the data would be to look at the combined output of both roles when they share the pitch, rather than isolating the effect on just one. This approach captures roles that enhance each other instead of just one role taking advantage of the other. From this, the most productive pairing is two Creative Wingers, who together average 13.61 shots and dribbles per 90 minutes. This suggests that when both flanks are occupied by high-risk, high-reward profiles, the system becomes heavily geared toward direct attacking play. Other standout combinations include Complete Forward & Creative Winger (12.97) and Creative Winger & Wide Playmaker (12.92).

Top 10 Pairings – Shots & Dribbles Matrix

1. Creative Winger & Creative Winger: 13.61 per 90
2. Complete Forward & Creative Winger: 12.97 per 90
3. Creative Winger & Wide Playmaker: 12.92 per 90
4. Creative Winger & Direct Winger: 11.59 per 90
5. Creative Winger & Second Striker: 11.55 per 90
6. Advanced Playmaker & Creative Winger: 11.02 per 90
7. Direct Winger & Direct Winger: 10.89 per 90
8. Creative Winger & Playmaker: 10.69 per 90
9. Complete Forward & Direct Winger: 10.62 per 90
10. Advanced Playmaker & Direct Winger: 10.50 per 90

5.2.2 Carries Matrix

This matrix tracks the number of carries performed by a player per 90 minutes, depending on which role they shared the pitch with. Carries represent a player's ability to progress the ball through space, and this matrix helps identify which combinations encourage or suppress that behaviour. It reflects not just individual tendencies, but how tactical structure and role proximity influence movement and ball retention.

For Deep Playmakers, the most enabling partner is the Playmaker, with an average of 73.06 carries per 90 minutes when paired together. This suggests another creative output close to the Deep Playmaker could relieve pressure off them and leave more space to progress the ball. The

lowest output occurs when Deep Playmakers are paired with Advanced Playmakers (40.89), indicating that the distance between the defensive midfielder and the attacking midfielder might be too much and the partner does not soak up any pressure from the Deep Playmaker, or that passes are preferred to advance the ball in those scenarios.

When looking at the combined output of both roles, the most productive pairing is again Deep Playmaker & Playmaker, with a combined 137.84 carries per 90 minutes, followed by Libero & Playmaker (117.85) and Deep Playmaker & Libero (115.52). These combinations highlight systems built around deep distribution and central control, where multiple roles share responsibility for advancing the ball. The presence of Libero, Playmaker, and Deep Playmaker across most top pairings reinforces the idea that progression is most effective when distributed across flexible, technically secure roles, especially those operating in deeper or transitional zones.

Top 10 Pairings – Carries Matrix

1. Deep Playmaker & Playmaker: 137.84 per 90
2. Libero & Playmaker: 117.85 per 90
3. Deep Playmaker & Libero: 115.52 per 90
4. Playmaker & Playmaker: 110.68 per 90
5. Playmaker & Positional Midfielder: 107.86 per 90
6. Creative Winger & Libero: 107.67 per 90
7. Libero & Libero: 107.06 per 90
8. Creative Winger & Playmaker: 106.32 per 90
9. Playmaker & Second Striker: 104.17 per 90
10. Deep Playmaker & Second Striker: 103.93 per 90

5.2.3 Defensive Actions Matrix

This matrix tracks the number of defensive actions — including interceptions, clearances, and blocks — performed by a player per 90 minutes, depending on which role they shared the pitch with. It highlights how certain roles influence defensive output, offering insight into which combinations foster active defensive behaviour and structural solidity.

For Stoppers, the most enabling partner is the Playmaker, with an average of 10.12 defensive actions per 90 minutes when paired together. This may reflect systems where the Playmaker's advanced positioning leaves space behind him and defensive responsibilities to deeper roles, forcing the Stopper into more frequent interventions. Conversely, the lowest output occurs when Stoppers are paired with Liberos (6.76), suggesting that the Libero's spatial coverage and anticipation reduce the need for reactive defending from the Stopper.

For the mutual data, the most productive pairing is Enforcer & Stopper, with a combined 18.32 defensive actions per 90 minutes, followed closely by Stopper & Stopper (18.09) and Sweeper & Sweeper (18.05). The recurring presence of Enforcers, Stoppers, Sweepers, and Ball Playing Centre Backs in the top pairings reinforces the idea that defensive output scales when roles are stacked with similar intent. This also suggests that when aggressive and defensive centerback roles play together, it reduces the amount of defensive work other positions might have to do.

Top 10 Pairings – Defensive Actions Matrix

1. Enforcer & Stopper: 18.32 per 90
2. Stopper & Stopper: 18.09 per 90
3. Sweeper & Sweeper: 18.05 per 90
4. Stopper & Sweeper: 17.96 per 90
5. Enforcer & Libero: 17.88 per 90
6. Enforcer & Enforcer: 17.84 per 90
7. Ball Playing CB & Stopper: 17.62 per 90
8. Ball Playing CB & Sweeper: 17.54 per 90
9. Ball Playing CB & Enforcer: 17.53 per 90
10. Enforcer & Sweeper: 17.42 per 90

5.2.4 Ball Recoveries Matrix

This matrix tracks the number of ball recoveries performed by a player per 90 minutes, depending on which role they shared the pitch with. Ball recoveries reflect a player's ability to regain possession and reinitiate play, often in transitional or defensive phases.

For Box to Box Midfielders, the most enabling partner is the Wide Supporter, with an average of 7.40 recoveries per 90 minutes when paired together. This suggests that systems with wide structural support allow central midfielders to press and recover more aggressively. On the other end, the lowest output occurs when Box to Box Midfielders are paired with Second Strikers (5.84), indicating that more attacking setups may reduce either their defensive workload or have them act more conservatively.

The most productive pairing is Box to Box Midfielder & Deep Playmaker, with a combined 13.04 ball recoveries per 90 minutes, followed by Advanced Playmaker & Positional Midfielder (12.73) and Ball Winning Midfielder & Playmaker (12.66). These combinations reflect systems built around central control and proactive regaining, often in midfield-heavy structures. They also suggest that ball recoveries thrive in systems with layered midfield roles.

Top 10 Pairings – Combined Ball Recoveries

1. Box to Box Midfielder & Deep Playmaker: 13.04 per 90
2. Advanced Playmaker & Positional Midfielder: 12.73 per 90
3. Ball Winning Midfielder & Playmaker: 12.66 per 90
4. Advanced Playmaker & Wide Playmaker: 12.56 per 90
5. Advanced Playmaker & Deep Playmaker: 12.49 per 90
6. Box to Box Midfielder & Positional Midfielder: 12.47 per 90
7. Ball Winning Midfielder & Positional Midfielder: 12.40 per 90
8. Deep Playmaker & Positional Midfielder: 12.29 per 90
9. Advanced Playmaker & Box to Box Midfielder: 12.24 per 90
10. Ball Winning Midfielder & Wide Playmaker: 12.19 per 90

5.2.5 Pressures Matrix

This matrix tracks the number of pressures performed by a player per 90 minutes, depending on which role they shared the pitch with. Pressures reflect a player's defensive intensity, their effort to close down opponents and disrupt buildup. This matrix helps identify which combinations foster high pressing behaviour and which setups may dampen it.

For Ball Winning Midfielders, the most enabling partner is the Playmaker, with an average of 25.89 pressures per 90 minutes when paired together. This suggests that when paired with a more creative, less defensively active role, the Ball Winning Midfielder takes on greater responsibility for disrupting opposition play. On the other end, the lowest output occurs when paired with a Conservative (21.61). The most pressing pair is Ball Winning Midfielder & Box to Box Midfielder, with a combined 50.19 pressures per 90 minutes, followed by Advanced Playmaker & Box to Box Midfielder (49.50) and Ball Winning Midfielder & Playmaker (46.96). These combinations reflect systems built around midfield dynamism, where multiple roles actively engage the ball and compress space.

Top 10 Pairings – Combined Pressures

1. Ball Winning Midfielder & Box to Box Midfielder: 50.19 per 90
2. Advanced Playmaker & Box to Box Midfielder: 49.50 per 90
3. Ball Winning Midfielder & Playmaker: 46.96 per 90
4. False 10 & Wide Playmaker: 46.79 per 90
5. Box to Box Midfielder & Defensive Winger: 46.79 per 90
6. Ball Winning Midfielder & Ball Winning Midfielder: 46.52 per 90
7. Advanced Playmaker & Playmaker: 45.99 per 90
8. Ball Winning Midfielder & False 10: 45.63 per 90
9. Ball Winning Midfielder & Defensive Winger: 45.57 per 90
10. Box to Box Midfielder & Box to Box Midfielder: 44.81 per 90

5.3 Line-Up Optimiser

The Line-Up Optimiser is a tool that creates ideal starting eleven's based off the previous matrices. Users can select which matrices to use, and the weight of each of them for the synergy calculations. In order to combine the matrices, all of them were previously normalised based on their maximum value, keeping all datapoints between 0 and 1. Once the matrices are selected, the user defines a formation (e.g. 4-3-3, 5-3-2, 3-4-3...) and input roles for any number of positions. These inputs act as constraints: the optimiser will respect the locked roles and fill in the remaining positions to maximize overall synergy across the team.

The optimiser then evaluates all possible combinations of roles for the remaining positions, using the selected matrices to score each line-up. The final output includes a complete set of roles for 11 positions and a synergy score between 0 and 1. This process allows analysts, coaches, and simulation designers to explore how different role combinations affect team performance, based on actual match data. This tool could also help in identifying what roles a team needs to recruit based on what key players are already in their best starting eleven. To access the final code, go to Appendix D.

Whether the goal is to maximize win probability, minimize xG against, or build a system that prioritises possession recycling, the Line-Up Optimiser provides a data-driven foundation for tactical decision-making.

6.0 Evaluation

This project offers a flexible and data-driven approach to tactical analysis, role optimisation, and squad planning. Built on season-long performance data and role-based synergy matrices, it enables users to explore how different combinations of roles affect team output across multiple dimensions, from xG creation to defensive intensity. The following sections outline its current applications, potential future developments, and validation strategies.

6.1 Applications

This system can be used across a range of football analysis tasks, including recruitment, tactical planning, and lineup building. It adapts to different formations, role constraints, and performance priorities. For recruitment, it helps identify players who fit into a squad not just statistically but tactically, based on how their role interacts with others. If teams have a few

players that are key pieces to the team, the Line-Up Optimiser can provide roles for the other positions that can be used as recruitment suggestions. In addition, the suggestions could also be extracted by looking at which roles consistently elevate others.

For tactical analysis, it shows which formations and role combinations work best in different match scenarios; whether defending a lead, chasing a goal, or facing a high press. This allows coaches and analysts to explore setups that suit specific game states or opponent styles. The optimiser can also generate full starting XIs using selected metrics like win rate, xG, or pressures. Users can lock in certain roles and let the system fill in the rest to maximise synergy across the pitch.

6.2 Future Work

There are several areas where the framework could be expanded to increase its tactical depth and real-world utility. One direction would be modifying the synergy calculations to apply different matrices depending on pitch zones (for example, using xG creation in the final third, passing volume in midfield, and defensive actions in deeper areas). Another idea is weighting role pairings more heavily when they operate in close proximity, since local interactions often have greater tactical impact (giving a larger weight to CB-CB or CB-LB pairings in the overall synergy calculations compared to GK-ST).

The optimiser could also be constrained to a pre-built squad, selecting only from available players and their viable roles. This would make the tool directly applicable to real-world team management. Additionally, the role clusters already built could be used to train a supervised classifier that predicts optimal roles for players based on performance data and tactical context.

Finally, the system is designed to work with any season-long dataset. If new data becomes available (from more recent seasons, different leagues, or updated tracking) the optimiser and matrices can be re-run without structural changes. This ensures long-term scalability and relevance.

6.3 Validation

Validation is key to assessing the reliability and tactical relevance of the framework. While the system is grounded in match data and matrix logic, its credibility depends on both quantitative metrics and real-world feedback.

For the unsupervised role clusters, silhouette scores were used to evaluate quality. Most scores ranged between 0.4 and 0.6, indicating moderately well-separated clusters. While not perfectly distinct, these scores suggest that the roles reflect meaningful behavioural patterns rather than arbitrary groupings.

The synergy matrices are built from aggregated match data. Their reliability depends on sample size, positional consistency, and tactical context. Validation strategies include statistical sanity checks, cross-matrix consistency, and temporal robustness. The most meaningful validation, however, would come from real-world use: a football team applying the optimiser in practice and providing feedback on its tactical accuracy and usability. This kind of feedback loop would allow the system to evolve from a data-driven prototype into a trusted tactical assistant.

7.0 Conclusion

This project presents a modular, data-driven framework for tactical role analysis and lineup optimisation in football. By leveraging season-long performance data and role-based synergy matrices, it enables users to explore how different combinations of roles influence team output across key metrics such as xG, pressures, and win probability. The clustering of player behaviours into distinct roles provides a foundation for tactical reasoning, while the optimiser translates that reasoning into actionable lineups.

The system is designed to be flexible and scalable, capable of adapting to new datasets, evolving tactical trends, and different user priorities. Whether used for recruitment, match preparation, or simulation, it offers a structured way to quantify tactical fit and explore the impact of role interactions.

While the current implementation is robust, its full potential lies in future development and real-world validation. With enhancements like zone-specific weighting, squad-constrained optimisation, and supervised role prediction, the framework could evolve into a powerful tactical assistant. And with feedback from coaches, analysts, and players, it could become not just a tool for analysis, but a collaborator in decision-making.

8.0 References

[1] Kai Hallberg. (6 May 2025) *Are soccer positions still relevant? A data-driven investigation*, Medium. Available at: https://medium.com/%40khallberg10/are-soccer-positions-still-relevant-a-data-driven-investigation-3ba521f00f2e (Accessed: 2 July 2025).

[2] StatsBomb, *Open Data 360 Frames v1.0.0,* GitHub, (March 2021). Available at: https://github.com/statsbomb/open-data/blob/master/doc/Open%20Data%20360%20Frames%20v1.0.0%20(1).pdf (Accessed: 30 June 2025).

[3] Wyscout, *Wyscout Glossary*. Available at: https://dataglossary.wyscout.com/ (Accessed: 2 July 2025).

[4] Tom Decroos, Lotte Bransen, Jan Van Haaren, Jesse Davis(18 February 2018) *Actions Speak Louder Than Goals: Valuing Player Actions in Soccer,* Available at: https://doi.org/10.48550/arXiv.1802.07127 (Accessed 10 August 2025)

[5] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, Fosca Giannotti. (2019) *PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach*, ACM Transactions on Intelligent Systems and Technology, 10(5), pp. 1–27. doi:10.1145/3343172

[6] FedericoCasati (2021) *Redefining-football-players-roles-using-clustering/3. clustering.ipynb at main · Federicocasati/redefining-football-players-roles-using-clustering*, GitHub. Available at: https://github.com/FedericoCasati/Redefining-football-players-roles-using-clustering/blob/main/3.%20Clustering.ipynb (Accessed: 13 July 2025).

[7] Miguel Ángel Gómez, Carlos Lago-Peñas, Richard Pollard. (2013) *Situational Variables, Routledge Handbook of Sports Performance Analysis*. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203806913-26/situational-variables-miguel-%C3%A1ngel-g%C3%B3mez-carlos-lago-pe%C3%B1as-richard-pollard (Accessed: 20 August 2025)

[8] Footballizer.com, *Academy - footballizer*. Available at: https://www.footballizer.com/academy/ (Accessed: 10 July 2025).

[9] FC Python. (2020) Learn Python with Football, FC Python. Available at: https://fcpython.com/machine-learning/introduction-to-k-means-with-python-clustering-shot-creators-in-the-premier-league (Accessed: 10 July 2025).

[10] Akhanli, Serhat Emre and Hennig, Christian.(2023) *Clustering of football players based on performance data and aggregated clustering validity indexes.* Journal of Quantitative Analysis in Sports, vol. 19, no. 2, 2023, pp. 103-123. https://doi.org/10.1515/jqas-2022-0037

[11] Lopes AM, Tenreiro Machado JA. (23 June 2021) *Uniform Manifold Approximation and Projection Analysis of Soccer Players*. Entropy (Basel), 23(7):793. doi: 10.3390/e23070793. PMID: 34201479; PMCID: PMC8307339. (Accessed 10 July 2025)

[12] Gijs Wijngaard. *Clustering soccer players: investigating unsupervised learning on player positions*. Utrecht University. Available at: https://studenttheses.uu.nl/bitstream/handle/20.500.12932/35795/thesis.pdf?sequence=1&isAllowed=y (Accessed 30 June 2025)

[13] Steve AQ (25 March 2024) *K-Means Player Cluster Analysis*. Pitch IQ. Available at: https://steveaq.github.io/Player-Roles-Clustering/ (Accessed 30 June 2025)

[14] D'Urso, P., De Giovanni, L. & Vitale, V.(14 February 2022) *A robust method for clustering football players with mixed attributes*. Ann Oper Res **325**, 9–36 (2023). https://doi.org/10.1007/s10479-022-04558-x (Accessed 10 July 2025)

[15] Azzami, S.Y. *et al.* (2025) *Clustering and profiling analysis for FIFA football player using K-means*, Jurnal Informatika: Jurnal Pengembangan IT, 10(1), pp. 178–189. doi:10.30591/jpit.v10i1.7897. (Accessed 10 July 2025)

[16] Lotte Bransen, Jan Van Haaren. *Player Chemistry: Striving for a Perfectly Balanced Soccer Team*. SciSports. Available at: https://arxiv.org/pdf/2003.01712 (Accessed 10 July 2025)

[17] Duarte Araujo, Keith Davids (21 September 2016). *Team Synergies in Sport: Theory and Measures*. Frontiers in Psychology. Frontiers. Available at: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2016.01449/full (Accessed 5 August 2025)

[18] Ramos, A. Coutinho, P. Leitao, J.C. Cortinhas, A. Keith Davids. (2020) *The constraint-led approach to enhancing team synergies in sport - what do we currently know and how can we move forward? A systematic review and meta-analyses* , Psychology of Sport and Exercise, 50, p. 101754. doi:10.1016/j.psychsport.2020.101754. (Accessed 10 July 2025)

9.0 Appendix

9.1 Appendix A: Dimensional Reduction Mappings

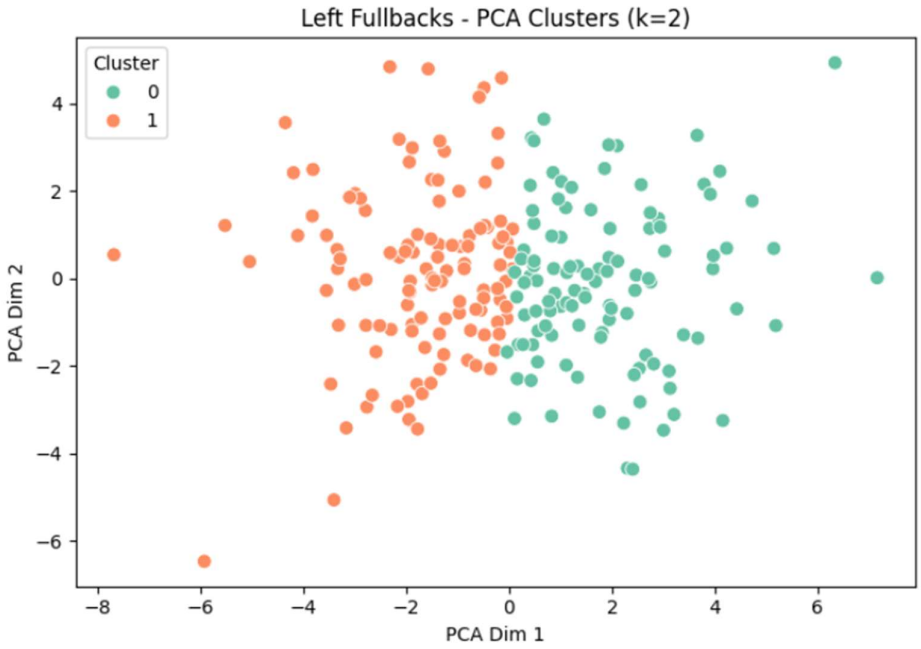

Figure A1: PCA projection of left back data, shows poor cluster separation. Silhouette Score = 0.328

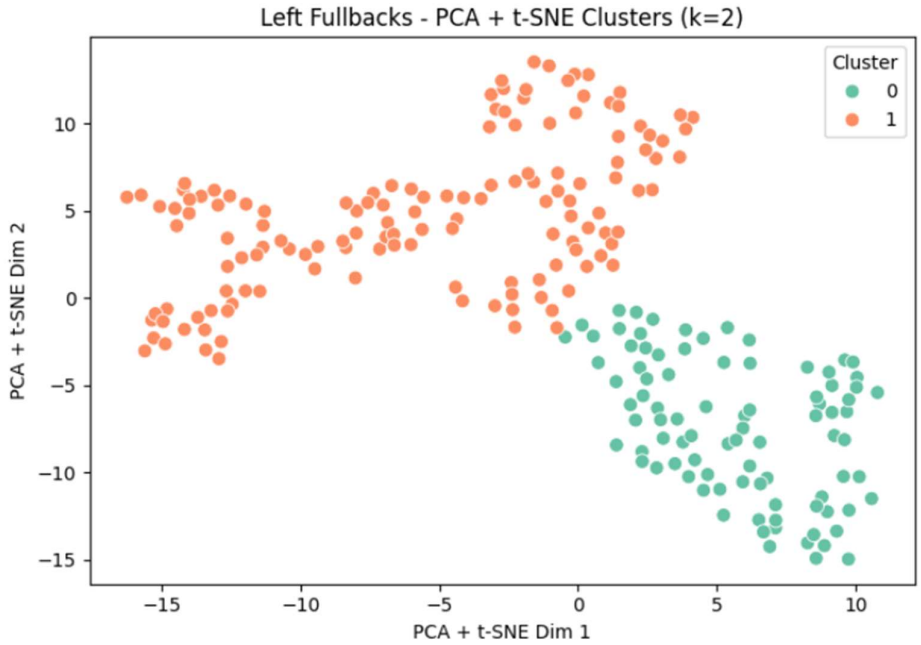

Figure A2: t-SNE projection of left back data, shows 3 distinct clusters. Silhouette Score = 0.403

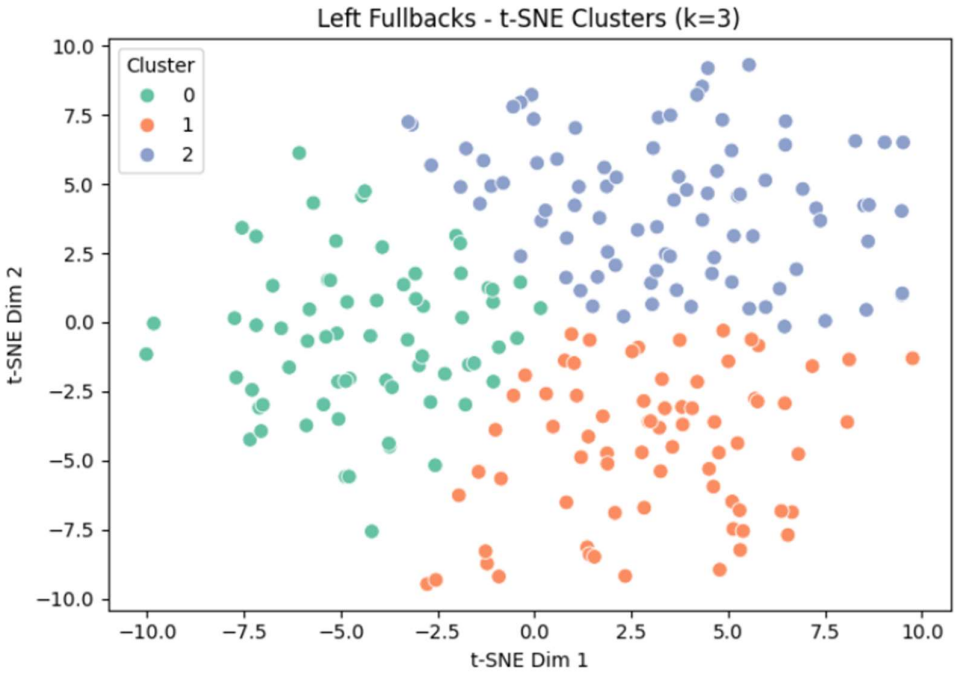

Figure A3: PCA and t-SNE combination projection of left back data, shows 2 distinct clusters. Silhouette Score = 0.532

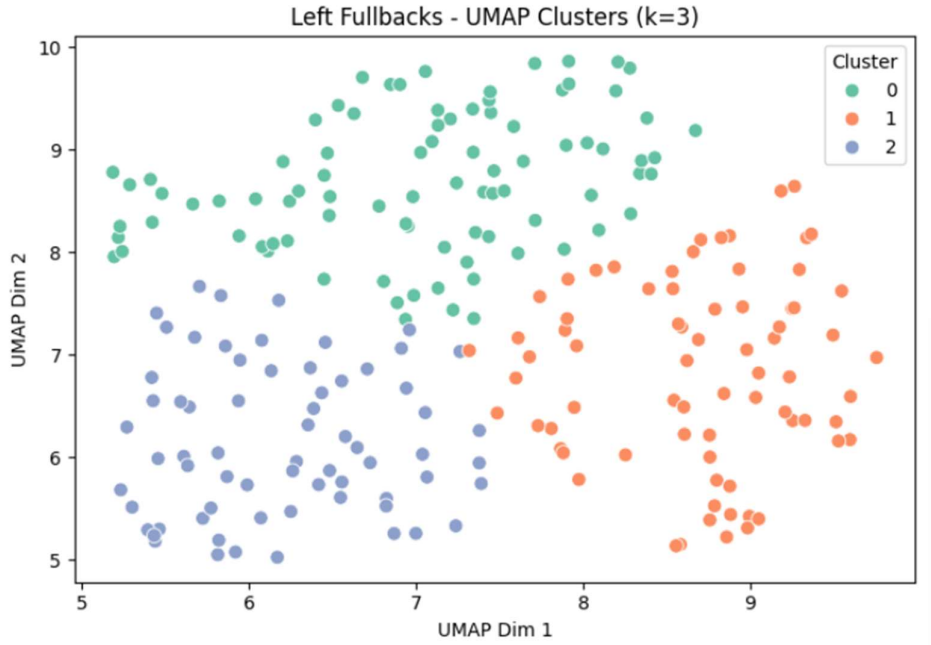

Figure 4A: UMAP projection of left back data, shows 3 distinct clusters. Silhouette Score = 0.430

9.2 Appendix B: Final Clustering Mappings

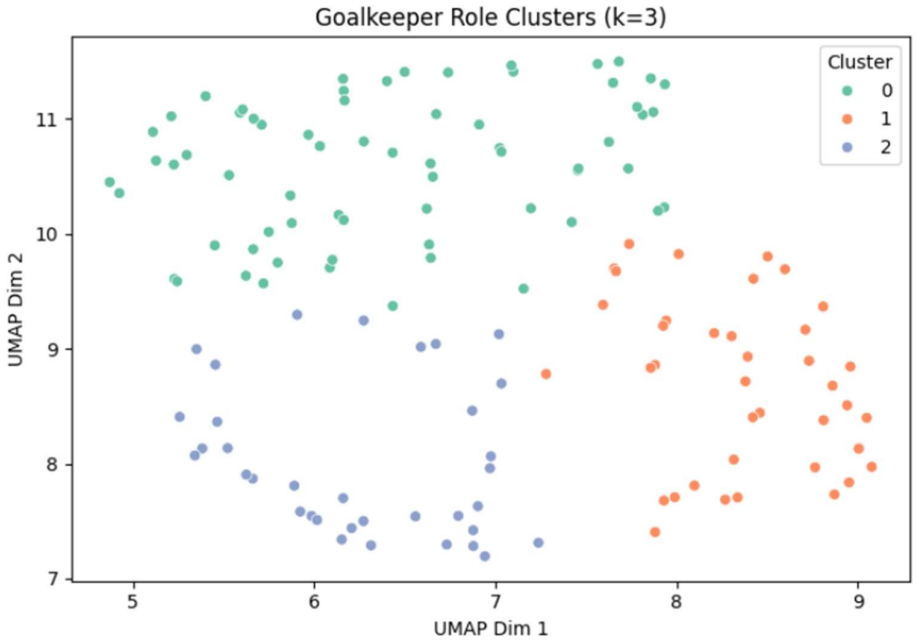

Figure 1B: Goalkeepers clustered into roles using UMAP (min_dist = 0.0, n_neighbors = 10) and K-Means (k = 3). Silhouette score = 0.459. Cluster 0: Shot Stopper, Cluster 1: Sweeper Keeper, Cluster 2: Conservative.

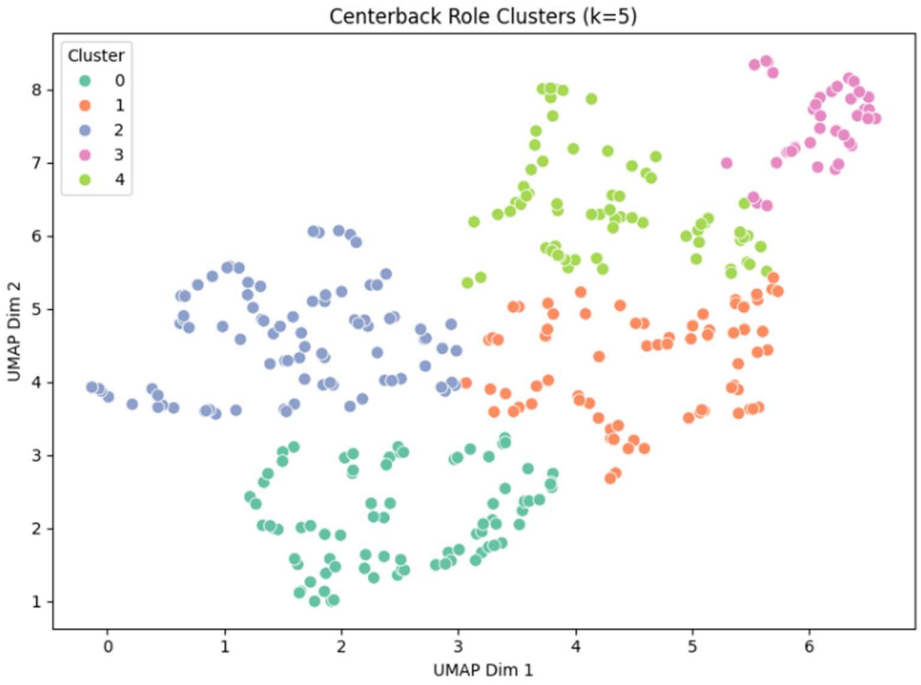

Figure 2B: Centerbacks clustered into roles using UMAP (min_dist = 0.0, n_neighbours = 5) and K-Means (k = 5). Silhouette score = 0.443. Cluster 0: Stopper, Cluster 1: Sweeper, Cluster 2: Enforcer, Cluster 3: Libero, Cluster 4: Ball Playing CB.

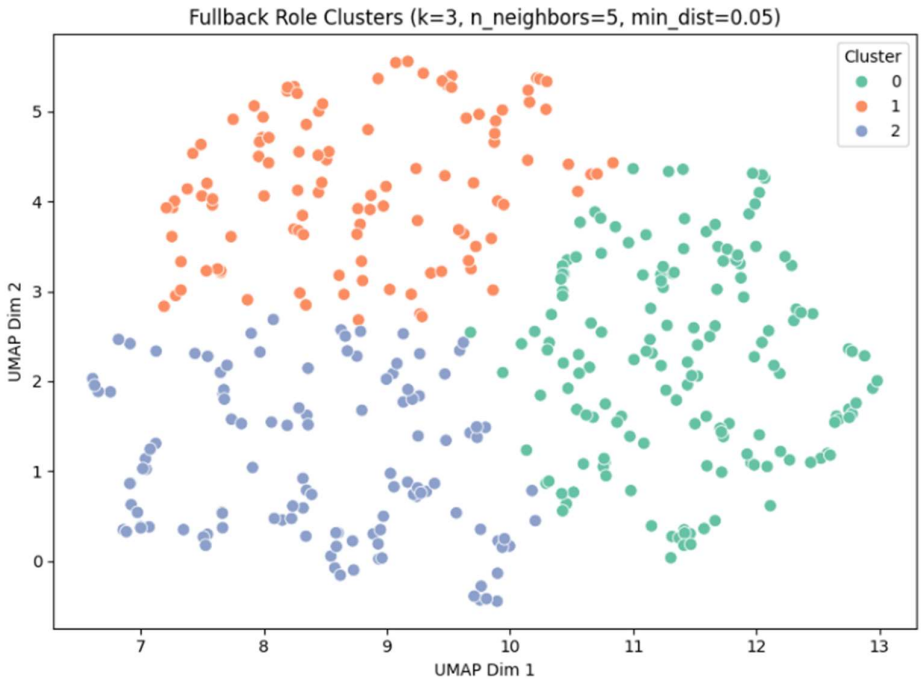

Figure 3B: Fullbacks clustered into roles using UMAP (min_dist = 0.05, n_neighbors = 5) and K-Means (k = 3). Silhouette score 0.429. Cluster 0: Defensive Fullback, Cluster 1: Balanced Fullback, Cluster 2: Attacking Fullback.

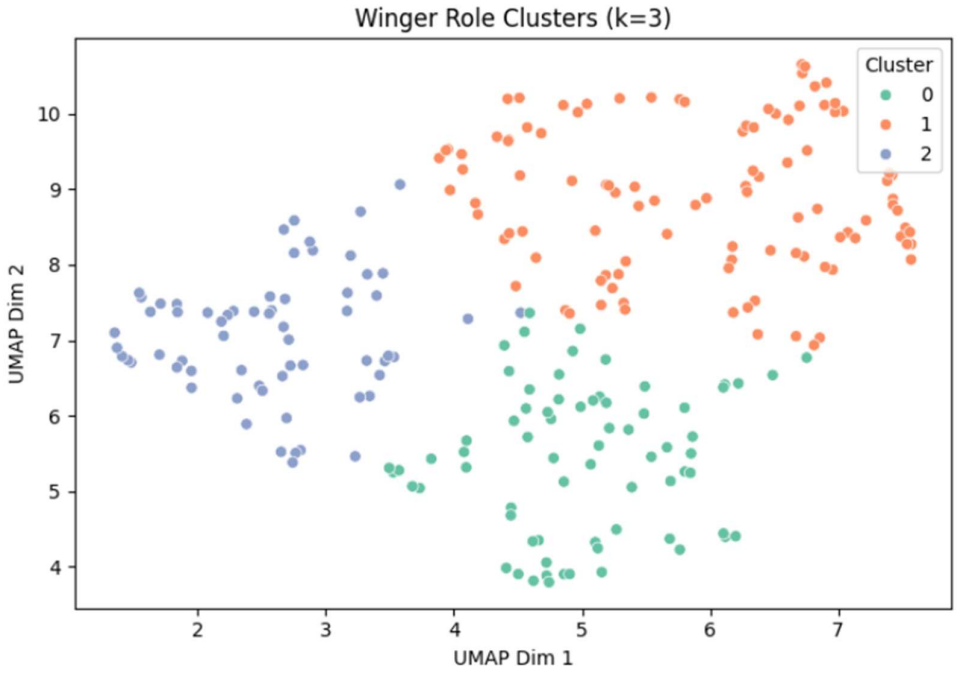

Figure 4B: Wingers clustered into roles using UMAP (min_dist = 0.05, n_neighbors = 5) and K-Means (k = 3). Silhouette score = 0.472. Cluster 0: Defensive Winger, Cluster 1: Direct Winger, Cluster 2: Creative Winger.

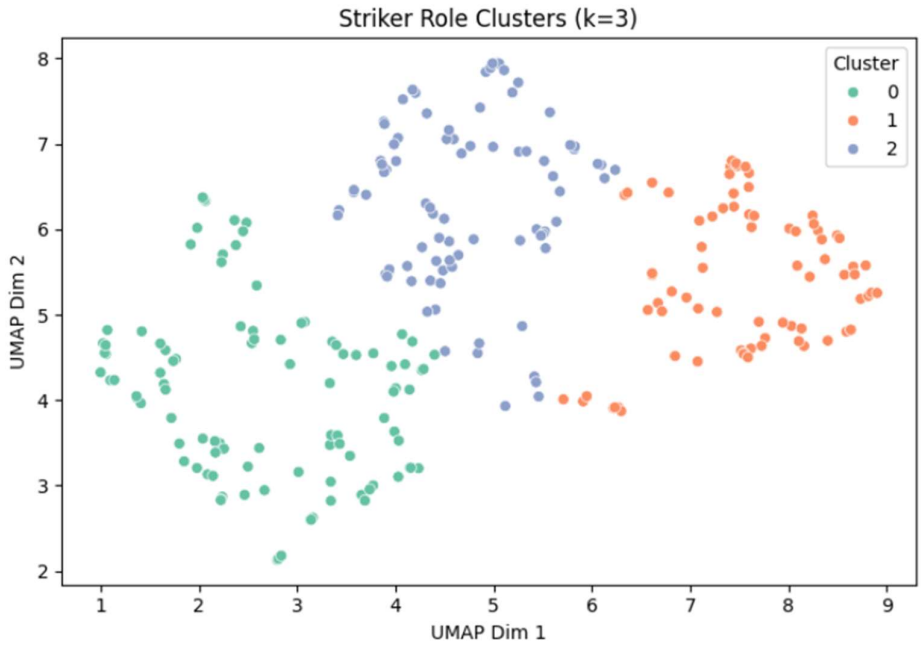

Figure 5B: Strikers clustered into roles using UMAP (min_dist = 0.0, n_neighbors = 5) and K-Means (k = 3). Silhouette score = 0.460. Cluster 0: Complete Forward, Cluster 1: Poacher, Cluster 2: Target Man.

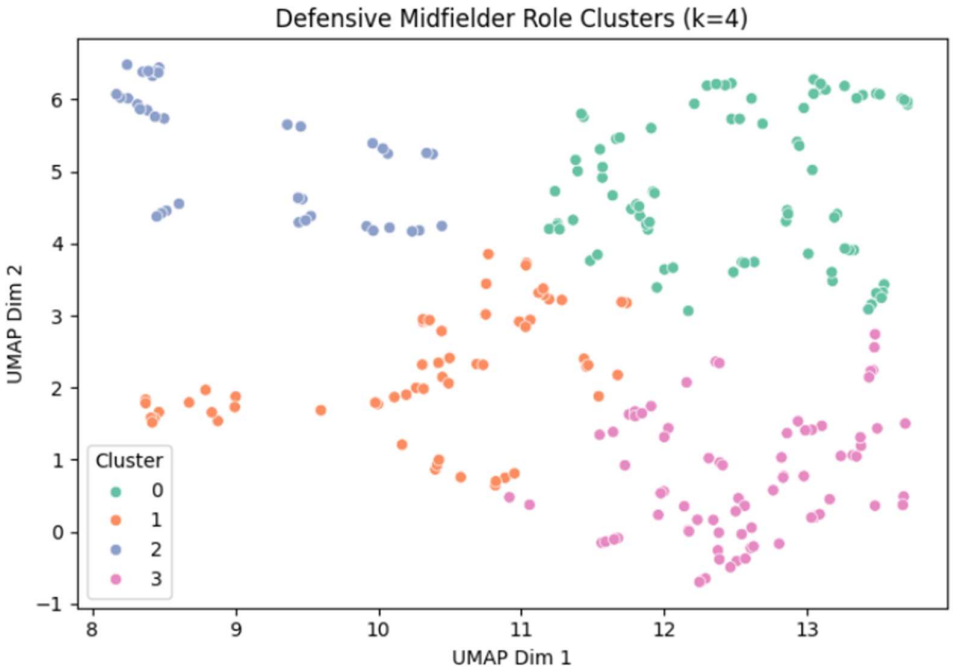

Figure 6B: Defensive midfielders clustered into roles using UMAP (min_dist = 0.0, n_neighbors = 5) and K-Means (k = 4). Silhouette score = 0.473. Cluster 0: Ball Winning Midfielder, Cluster 1: Positional Midfielder, Cluster 2: Deep Playmaker, Cluster 3: Anchor Man.

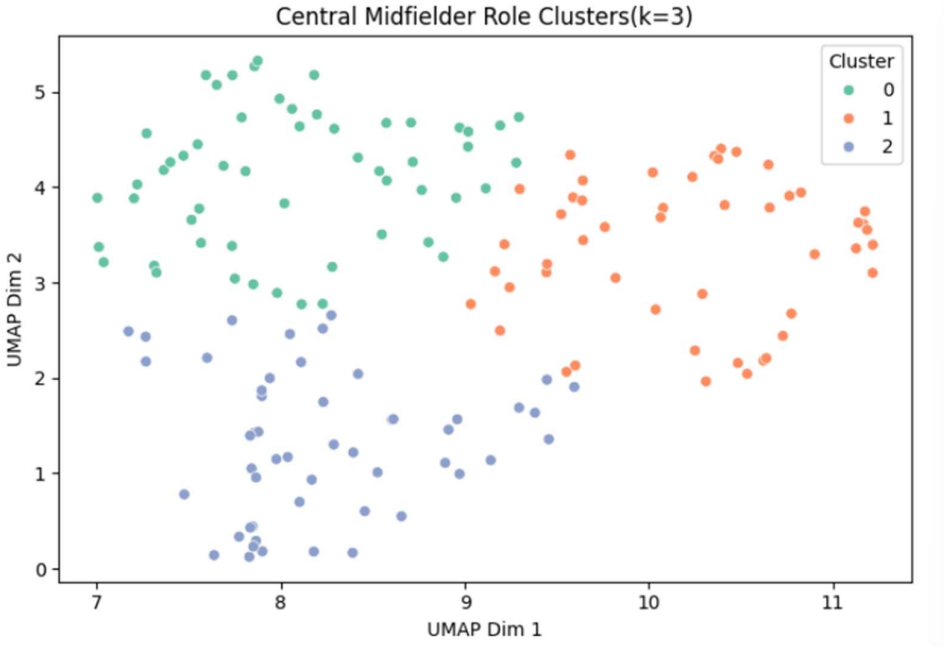

Figure 7B: Central midfielders clustered into roles using UMAP (min_dist = 0.0, n_neighbors = 15) and K-Means (k = 3). Silhouette score = 0.461. Cluster 0: Box to Box Midfielder, Cluster 1: Holding Midfielder, Cluster 2: Playmaker.

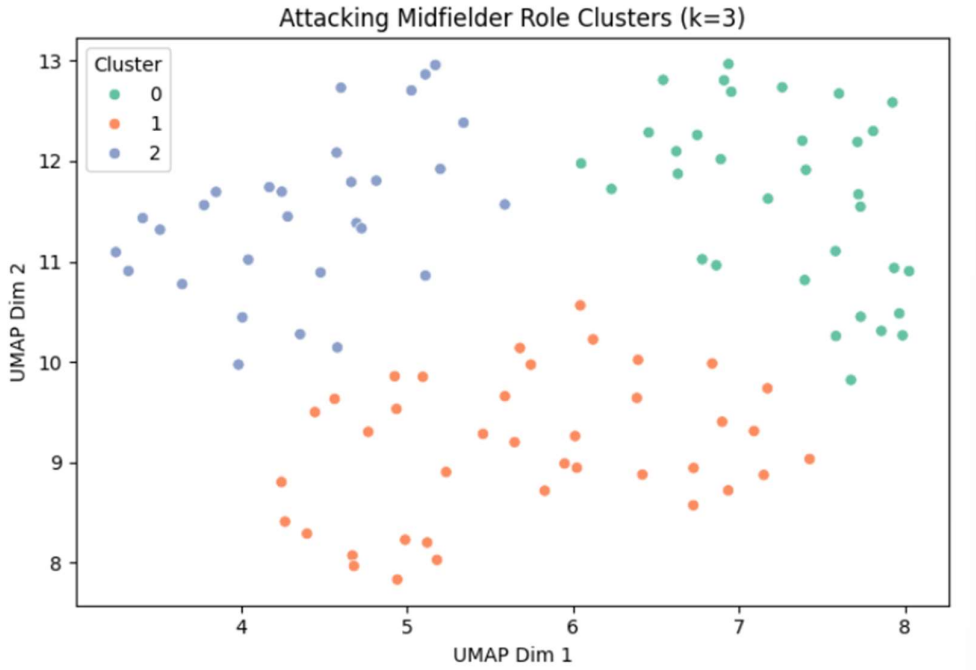

Figure 8B: Attacking midfielders clustered into roles using UMAP (min_dist = 0.1, n_neighbors = 10). Silhouette score = 0.466. Cluster 0: Advanced Playmaker, Cluster 1: False 10, Cluster 2: Second Striker.

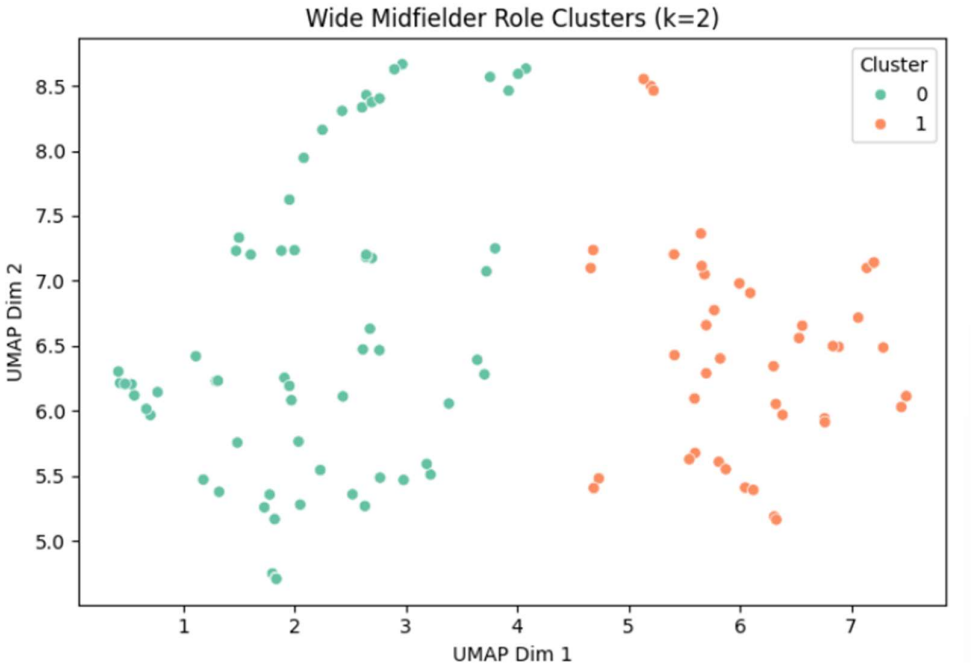

Figure 9B: Wide midfielders clustered into roles using UMAP (min_dist = 0.0, n_neighbors = 5) and K-Means (k = 2). Silhouette score = 0.568. Cluster 0: Wide Playmaker, Cluster 1: Wide Supporter.

9.3 Appendix C: Synergy Matrices

Due to size constraints, full matrices are available online:
[https://drive.google.com/drive/folders/1PZoih17T5Ro4Hafbly31tje7WmObD0hh?usp=drive_link]

9.4 Appendix D: GitLab Repository

GitLab Permalink: https://git.cs.bham.ac.uk/projects-2024-25/ixa184/-/tree/18fdd20d5e8bc39c3a0b9001a190a1c919abe77e/

Initial StatsBomb Data found at: https://github.com/statsbomb/open-data/tree/master/data

All code files were done on Jupyter Notebook, data will need to be downloaded on your device for the code to work.

Main files in the repository:

- Dataset build final: 5 different code files processing the StatsBomb data, for each league used, to create player statistics datasets.
- Cleaned_player_data.csv: Example player data set for LaLiga 2015/16
- Positions.ipynb: Each position group has its own clustering code file. Midfielders have two, 'midfielders.ipynb' is the clustering with all midfielders together, 'midfielders split.ipynb' is the final used code that first splits midfielders into more positional groups. The other position files are for 'goalkeepers', 'centerbacks', 'fullbacks', 'wingers' and 'strikers'.
- all_player_roles.csv: Final player dataset including players from the 5 leagues, with their performance stats and the final roles given to them.
- Complimentary matrices final.ipynb: The final code to create the synergy matrices.
- Player_Role_Pairings_Matrices-20250829T140030Z-1-001.zip: Zip file storing all the matrices used.
- Final_Lineup_Optimiser.ipynb: The final iteration of the line up maker code, that uses the players dataset and the matrices.