Defining Soccer Playing Styles through a Data-Driven Approach

Undergraduate Thesis Report

Author: Mohammad Mustafa Arif

Supervisor: Professor Timothy Chan

April 16th, 2021

B.A.Sc. Thesis

Division of Engineering Science
UNIVERSITY OF TORONTO

Page intentionally left blank

Defining Soccer Playing Styles through a Data-Driven Approach

Author: Mohammad Mustafa Arif
Supervisor: Professor Timothy Chan

Abstract

This project exploits data from soccer event logs to extract features via graph theory, and characterizes team performances across multiple dimensions to encapsulate different aspects of the game. The research proceeds further to summarise these multi-dimensional performance representations for each team across Europe's top five domestic leagues during the 2017/18 season, and clusters teams in vector space using the features derived. To do so, hierarchical clustering with Ward's method is used to arrive at four distinct playing style clusters. A significant performance discrepancy across playing styles is found, and it is noted that teams in high performing clusters, on average, display higher possession, higher variability in passing across players and pitch zones, more indirect passing behavior (a high advance ratio), and a high defensive line compared to teams in the lower performing clusters. Analysis is also conducted to demonstrate that playing styles are not simply determined by a team's skill level, and are indeed a reflection of team identity beyond skill/ability to dominate the league. The final outcome of the research is thus a characterization of 98 teams into four playing styles spanned by custom-defined features defining different aspects of the sport. This research contributes a novel method defining team playing styles, as it generalizes network science approaches from prior work, combining them with an unsupervised clustering approach for general playing style extraction. Potential use cases for such playing style characterization include decision support for tacticians formulating game plans for upcoming games against opponents that the team does not have much direct experience with. In the larger context, the project demonstrates the value of data driven analytics in understanding the game of soccer.

Acknowledgements

First of all, I would like to express my sincere gratitude towards my supervisor, Professor Timothy Chan, for giving me the opportunity to formulate my undergraduate thesis around a topic I share a deep passion for.

I also wish to express my sincere appreciation towards my mentors for this project, Binghao Zhang and Craig Fernandes, for their regular feedback, hands-on advice, and willingness to help out in a multitude of ways over the past 10 months. Both Binghao and Craig have provided me with incredible support through this journey, and my time working under their guidance will forever be remembered and appreciated.

I would further like to extend special gratitude to Professor Scott Sanner, who inspired me towards decision support systems when I took his course, and who, on numerous occasions, has been openly willing to share his knowledge and insights with me during his office ours, helping me immensely in proceeding with my research.

In addition, I would like to thank my parents, who continue to support me in my academic and personal ventures, and consistently go out of their way to ensure I can commit myself to my work with great diligence. And finally, I would want to express appreciation towards my friends Saad Jameel, Anup Deb, Bill Sun, Kevin Zhang, Yesheng Wang, Morris Chen, and Andy Zhou, for the memorable undergraduate years we spent together.

Contents

1 Introduction

Soccer is the most popular sport in the world. Today, more than 240 million people play soccer regularly; a combined 3.5 billion viewers tuned in to the 2018 FIFA World Cup broadcast [1][2]. The historical popularity of the game drove early interest in the use of analytics to study it. As early as the 1950s, Charles Reep, an English soccer analyst, began documenting soccer statistics by hand, and indirectly laid the foundation for the long-ball movement [3][4]. Today, analytics in soccer has largely been commercialized, with vendors like Opta, a British sports analytics company, providing data to clubs and coaches to study the game in depth [5]. The field of analytics has also rapidly evolved and teams have begun using data-driven approaches for tactical decision making. With increased data collection, novel methodologies in statistics, computer science, and network theory have been employed to study teams and exploit strengths and weaknesses. For example, recently, Liverpool FC, a first division English soccer club, announced the development of "pitch control", a probabilistic model used to quantify space creation during matches in real time [6].

Despite the evolution of soccer analytics, much of the wealth of available data is exploited to a limited extent. Traditional methods involve utilizing high level match statistics such as possession percentages, numbers of shots, passes, crosses, fouls, and offsides. However, these methods are prone to oversimplifying 90+ minute interactions between two teams to a set of global statistics. An increasingly popular approach to add nuance to these methods is the use of expected goals (xG), a metric to measure the quality of shots taken by their probability of resulting in a goal [7]. In addition, modern approaches in literature have attempted to study team passing profiles through the distribution of passes amongst the players. However, these approaches are also prone to oversimplification; they do not distinguish between different types of passes or the distribution of passes across different regions of the pitch. Furthermore, they do not explore defensive activity with respect to team performance.

The general aim of this project is to demonstrate the use of data to understand team playing styles. Given the inherent shortcomings in prior work as discussed briefly above, and in more detail in the Literature Review chapter, the first objective is to expand on existing literature characterizing team passing distributions and fill in the existing gaps by:

1. Using network science to derive deeper passing metrics capturing player interactions for each performance.

2. Incorporating defensive metrics: the distribution of tackles, interceptions, and fouls to represent team performances.

These metrics span the feature space to numerically describe each performance, so that playing style clustering in vector space becomes feasible. The features for this stage of the project are derived from the Wyscout spatio-temporal soccer event logs, a public dataset with reports of soccer match events marked by time, spatial coordinates, players involved, and a standardized taxonomy of event tags [8]. The dataset covers Europe's top five domestic leagues for the 2017/2018 season, as well as international fixtures in the 2016 Euros and the 2018 World Cup. Some features are engineered though direct manipulation of the logs, while others require a deeper analysis through the construction of passing networks across players and pitch-zones. After engineering the features describing each performance, feature importance is proxied via mutual information with respect to some measure of performance.

The second phase of the project involves aggregating performance metrics for the entire season to segment teams into different playing styles. Clustering methods such as K-Means and Hierarchical Clustering are explored. The cluster assignments are then analysed from a fundamental perspective to arrive at intuitive interpretations of the established playing styles. Finally, case studies are conducted to find commonalities in playing styles amongst the top tier teams, determining what playing styles overperform/underperform with respect to others, and investigating the extent to which playing style is determined by skill.

Thus, the final outcome of this project is a classification of teams in terms of a feature space spanned by custom-defined metrics, and an analysis of the resulting playing style clusters. This study provides insights into what qualities some of the top performing teams share, and what separates these top tier teams form their mid-table and low-table counterparts. It is also found that there is not just a single playing style all the top tier teams fall under, but there are certain playing styles that are only exhibited by a subset of top tier teams. As such, this study provides a descriptive analysis on soccer performance, which can be used by coaches and tacticians to get insights for their teams or make note of the strengths/weaknesses of opposing teams. Therefore, this research adds value to the game of soccer by providing decision support, potentially narrowing the focus of human expertise. To illustrate a use case, consider a coach preparing tactically for an important upcoming game against an opponent the team has not faced in a long time period[1]. Using the playing style classification derived in this project, the coach can

[1]Such a case often comes up in the UEFA Champions League, where teams from different domestic

identify the playing style of the opponent, and study footage of his team against other recent opponents that share that playing style. This would allow the coach to focus on specific games to investigate what worked well and what failed against that style of opponent. Note that in this intended use case, the role of this research project is to provide decision support by potentially narrowing the focus of coaching staff to a subset of games from a plethora of recent footage. The project does not aim to fully automate tactical decision making and remove the role of coach philosophy/intuition. Rather, the two are complementary.

At a higher level, the project also demonstrates the value of data driven analytics in understanding the game of soccer. Despite studies finding a strong role of randomness in soccer game outcomes [9], the simple features used in this project provided enough information to segregate (in general) top performers from others. This is a promising result, as there is reason to believe that more complex analysis/feature engineering may provide more nuanced interpretations of team playing styles.

2 Literature Review

Analytics in soccer has been used on a wide array of data types including video footage, logs, and GPS tracking data. Human annotated soccer-logs capturing all events within a match have also commonly been used to study many aspects of the game [8]. In particular, work has been done on memorable/historic squads to conduct statistical performance attributions, as well as squads that achieved highly unexpected season outcomes. There have also been notable research attempts to derive embedding style representations for teams as well as individual players. While this project is framed at segmenting teams into characteristic playing styles, the following literature survey explores all the aforementioned sub-domains, which include analyses on teams as well as individual players. This is because much of the methodology discussed is reapplied in the context of this project. Note that all of the methods employed in the following literature survey can be implemented using data from soccer logs[2].

One research area in soccer analytics is to use data driven strategies to rank players. This entails answering questions such as "what makes a good player good and vice-versa?". A popular method to quantify player quality using statistics is to use information from in-game events for extracting information. For example, Pappalardo et al.

leagues contend to be champions of Europe.

[2]The dataset used for this project is a soccer-log dataset. See the 'Data' section for more details.

[10] use soccer-logs to design and implement PlayeRank, a role-aware, multi-dimensional evaluation framework for individual soccer players, tuned through team performance outcomes. They validate the performance by testing the algorithm against ground truth ratings from professional soccer scouts, achieving a significant level of agreement. They also demonstrate the use of PlayerRank to distinguish top players from others, and measure player versatility. As such, Pappalardo et al. addressed issues with the time's simple data-driven player performance metrics by:

1. Designing a "role" aware system: The evaluation function is dynamic with respect to a player's role on a team.
2. Evaluating performance across many dimensions instead of reporting a scalar score combining all aspects of player performance.

Thus, this method successfully bypasses the problem of oversimplification, where each aspect of a player's performance (including shots, passes, work-rate) are combined into an overall rating, by deriving a multidimensional, interpretable rating for each player. However, a major drawback of this method is its limited scope at individual player level. Analysing squads of 11 players (with possible substitutions and injuries) becomes an increasingly challenging task with this framework. In addition, the input features for PlayeRank are solely derived from individual player activity and they fail to measure interactions between players or "team chemistry". Nevertheless, the use of unsupervised clustering with the K-Means algorithm on spatial features for role detection is a robust method. It is generalized, with the help of involved feature engineering, to segment teams into playing styles in the context of this project.

Since there are intrinsic drawbacks to analysing players individually rather than analysing squads as dynamic systems with unique agents interacting with one another, extensive work has been done to study soccer squads at a higher level of abstraction. These approaches build on a foundation of metrics that capture interaction amongst players as a proxy for "team chemistry". One example of such high level analysis comes from Cintia et al. [11], who extract five separate team metrics summarizing passing behaviour amongst players and across zones of the pitch, eventually combining them via a weighted harmonic mean into a single indicator, the 'H-Indicator'. They report a significant correlation between the H-Indicator and performance measures, and explain game outcomes based on the H-Indicator difference across the teams. Further, they demonstrate out-of-sample predictive power in the H-Indicator by utilizing various classifiers to predict match outcomes through a season based exclusively on past H-Indicator values

and find high agreement between real and simulated rankings. The study successfully managed to quantify passing behaviour across teams, but one limitation is that only five simple statistics are used to describe passing: mean and variance of passes across players, mean and variance of passes across zones, and total passing volume. While the demonstrated descriptive and predictive power in the H-Indicator implies that these five building blocks indeed capture significant information about team passing behavior, their simplicity presents a requirement for more expressive features to characterize passing as well as other aspects of the game such as defense.

The study of soccer teams as dynamic systems with unique agents has also been approached via network science or graph theory. Originally proposed by Gould and Gatrell in 1979 [12], the idea of constructing passing networks from team performances gained popularity after Duch et al. demonstrated in 2010 [13] that player flow centrality in passing networks, a measure of the frequency of the player being involved in paths resulting in a shot, can be used to quantify the contribution of individual players in team performances. More recently, Clemente et al. [14] employed the use of network science to study team performances in the 2014 World Cup and demonstrated to a statistically significant degree that large connectivity between teammates is associated with better overall team performance. In their 2019 article, Buldu et al. [15] also utilized metrics engineered using network science to compare Pep Guardiola's FC Barcelona team from the 2009/2010 season, considered one of history's best soccer teams, with the rest of the teams in the Spanish first division. In particular, they computed[3]:

- *Clustering Coefficients*: A measure of local robustness in a network.
- *Algebraic Connectivity*: A measure of network integration or fault tolerance.
- *All-Pairs Average Topological Shortest Paths*: A measure of the directness of ball movement across teammates.
- *Advance Ratio*: A measure of passing 'directness' during a game.

Across all these metrics, they found statistically significant differences between FC Barcelona and the rest of the teams in the Spanish league. Thus, Buldu et al. [15] defined a network science framework for studying the playing style for a specific team against its rivals. For this project, their methodology can certainly be generalized by using the graph theory framework for all teams and constructing numerical representations of performances that can be clustered in vector space.

[3]Intuitive definitions are listed here. For precise mathematical definitions, please refer to the 'Methods' section.

Each of the works discussed approached soccer analysis through a statistical framework exploiting soccer-logs data-sets. However, they do not apply their methods towards generalized team playing style identification. Furthermore, relatively little exploration is done beyond passing, especially for defensive attributes. This project will therefore attempt to expand on methodology described above to fill in these gaps. Specifically, methods to characterise different aspects of the game will be taken from these studies, a separate analysis for defensive attributes will be conducted, and each team will be clustered in vector space for to derive team playing styles.

Finally, note that the use of soccer-logs is just one of many different approaches for utilizing analytics in soccer. Many other types of datasets have been used as well. There have been notable to exploit video footage, GPS tracking, and player physiological signals [16][17][18]. These approaches, however, are fundamentally different from those used in this project as the data they exploit is of an entirely different type. Thus, they are not discussed comprehensively here.

3 Data

For this project, the data being used is sourced from a public, spatio-temporal dataset of soccer-logs spanning Europe's top 5 domestic leagues for the 2017/2018 season: Spanish first division, Italian first division, English first division, German first division, and French first division [8]. In addition, international games from the 2018 World Cup and the 2016 European Cup are also covered. The data is provided by Wyscout, a leading company in the soccer industry, released under the CC BY 4.0 License, and is available on figshare[4]. The data is collected through a 3 step process.

1. Expert video analysts set team formations at the beginning of each game. This includes mapping the on-field players to their positions as well as listing the available players on the bench.

2. For each touch on the ball, the analysts, using a propriety tagger software, select one player, and create a corresponding event on the timeline. The event description involves specifying an event type and sub-type, along with the spatial coordinates, and other special tags to specify additional attributes.

[4]The dataset is publicly posted at https://figshare.com/collections/Soccer_match_event_dataset/4415000

3. The logs are quality controlled, both algorithmically and through manual cross comparisons.

The exact taxonomy of the events, sub-events, and tags can be found summarised in Table 1[5]. Altogether, the dataset covers 1941 games, 3,252,294 events, and 4,299 players.

While the dataset was provided in a standard JSON format, it comprised of files with large memory footprints. Therefore it was not feasible to open and explore the dataset on local PCs, and consequently, cloud services were used to clean/preprocess the data and filter out fields and tables not required for the project. For this task, the Standard D8s v3 virtual machine from Azure was chosen as it provided a large enough RAM of 32 GiB to accommodate the dataset. In addition to filtering out unnecessary information, the original dataset was decomposed into smaller files, and translated into a more user-friendly, tabular format. The process is summarised in Figure 1.

Table 1: Dataset Event Taxonomy

Event	**Sub-Events**	**Tags**
pass	cross, simple pass	accurate, not accurate, key pass, opportunity, assist, goal
foul	-	no card, yellow, red, 2nd yellow
shot	-	accurate, not accurate, block, opportunity, assist, goal
duel	air duel, dribbles, tackles, ground loose ball	accurate, not accurate
free kick	corner, shot, goal kick, throw in, penalty, simple kick	accurate, not accurate, key pass, opportunity, assist, goal
offside touch	acceleration, clearance, simple touch	counterattack, dangerous ball lost, missed ball, interception, opportunity, assist, goal

4 Methodology & Results

This chapter will summarize and justify the methodology, implementation challenges, and experimental design, keeping in mind the overarching goal to eventually arrive at a classification of distinct playing styles for soccer teams. The results will also be presented, while detailed discussion on the results will follow in the next chapter. In broad-strokes, the methodology can be broken down into four categories:

[5]Table taken from [8].

1. Data Preprocessing
2. Exploratory Data Analysis
3. Feature Engineering
4. Unsupervised Clustering

The following sections will elaborate more on work done and insights obtained from thees methods.

4.1 Data Preprocessing

The soccer-logs dataset used in this project was released under the CC BY 4.0 License, and is posted on figshare[6]. While the dataset was provided in a standard JSON format, it comprised of files with large memory footprints. Therefore it was not feasible to open and explore the dataset on local PCs, and consequently, cloud services were used to clean the data and filter out fields and tables not required for the project. For this task, the Standard D8s v3 virtual machine from Azure was chosen. This virtual machine was selected because it was the least expensive resource that provided a large enough RAM of 32 GiB to accommodate the dataset and provided an integrated data science environment. In addition to filtering out unnecessary information, the original dataset was decomposed into smaller files, and translated into a more user-friendly, tabular format, as summarised in Figure 1.

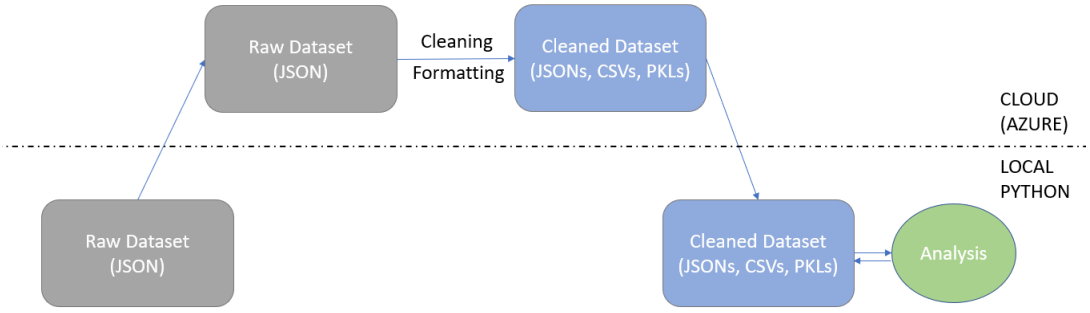


Figure 1: Data Prepossessing

The information removed at the preprocessing stage included details regarding referees and coaches, as well as events pertaining to international games in the 2018 World

[6]The dataset is publicly available at https://figshare.com/collections/Soccer_match_event_dataset/4415000

Cup and the 2016 European Championship. Note that these international games were intentionally excluded from the scope of the project because the volume of data obtained per team was severely limited in comparison to the five domestic leagues for the 2017/2018 season. This is a consequence of the longer format of the domestic leagues in which each team plays 19 games, as opposed to elimination style international tournaments. The scope of this project was thus narrowed down to classify playing styles exclusively over long-form domestic leagues.

4.2 Exploratory Data Analysis

After the dataset was prepossessed and made feasible to open locally, an exploratory data analysis was conducted to gauge the information available within the dataset. First, the data coverage was verified to note the number of games, events, and players from each of the domestic leagues. The findings are summarised in Table 2:

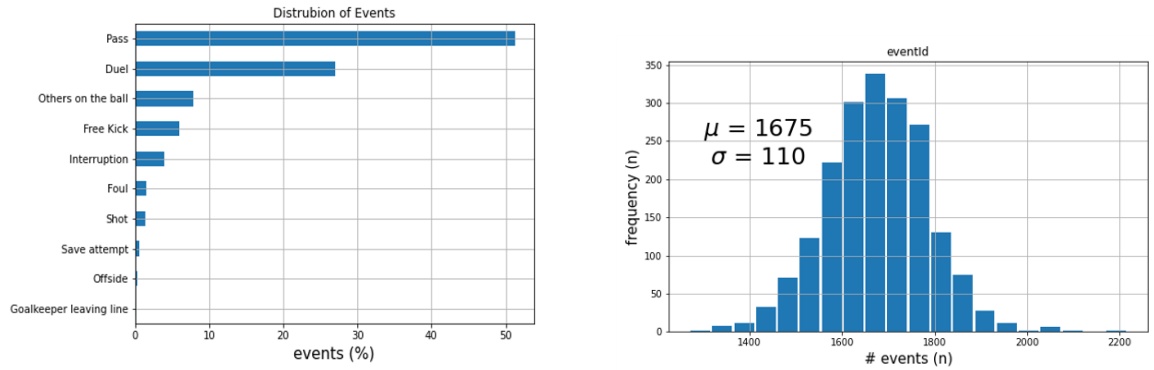
Table 2: Matches, Events, and Player Counts

League	**Matches**	**Events**	**Players**
La Liga (Spain)	380	628,659	619
Premier League (England)	380	643,150	603
Serie A (Italy)	380	647,372	686
Ligue 1 (France)	380	632,807	629
Bundesliga (Germany)	306	519,407	537

Note that with the exception of the German Bundesliga, all the leagues comprised of exactly 380 games. The reason for the discrepancy is that while all the other leagues have 20 teams, each playing 19 games a season, the Bundesliga is structured to have 18 teams, each playing 17 games per season. A breakdown of event and subs-events counts was conducted, as shown in Figure 2a. By far, the most common event was a pass, followed by a duel. The distribution of the number of events per game was analysed as shown in Figure 2b. It appears to be fairly symmetric around a mean value of 1675 events per game and a standard deviation of 110 events per game.

In the exploratory phase, the spatial aspect of the events was also studied. As noted previously, each event is marked by a set of coordinates. In the case of passes, the origin and destination points are defined separately. All coordinates are two dimensional and provided in units of f.u. (field units)[7], ranging from 0 to 100 in both the x and y

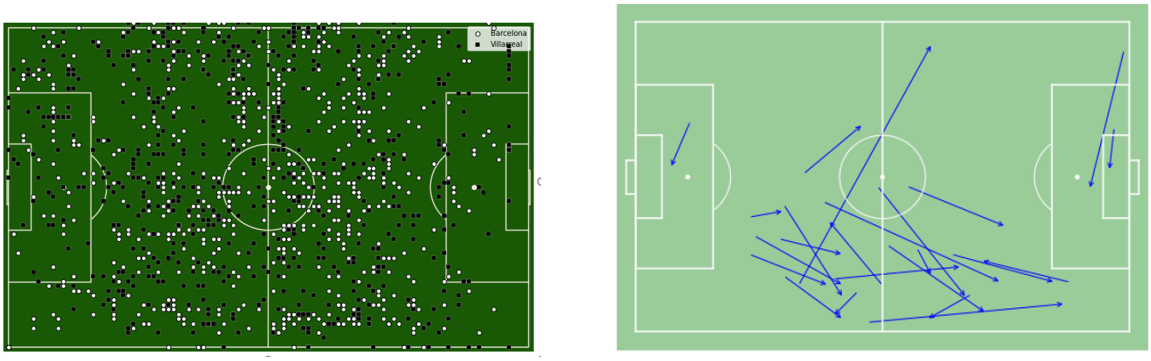
[7]The x coordinate measures nearness to the opponents goal, with $x = 100$ indicating the opponent's



(a) Event Percentages (b) Events Per Game

Figure 2: Event Breakdowns and Event Counts per Game

dimensions. Using these coordinates, visualizations for events in a game as well as passes conducted by single players were derived, as shown in Figure 3.



(a) All Events: FC Barcelona 5 - 1 Villareal (b) Passes From a Central Midfielder

Figure 3: Spatial Data Visualization with respect to the Pitch

Finally, visualizations of the temporal nature of the games were also studied. Because this project is framed at analysing team performance per game (as opposed to breaking the performance down by smaller time intervals), much of the intra-game temporal data is not thoroughly studied except for the construction of passing networks. Nevertheless, some visualizations were made to gauge how the distribution of goals vary with respect to time. As seen in Figure 4, the second half of the game is more eventful in terms of goals across all the domestic leagues.

With the data prepossessed and analysed at a high level, and frameworks for visualizations in place, the exploratory data analysis phase concluded and paved way for more involved analysis.

goal line. Similarly the y coordinate represents the nearness to the right side of the pitch. Field units conveniently provide a standardization for spatial coordinates as not all pitches share exactly the same dimensions.

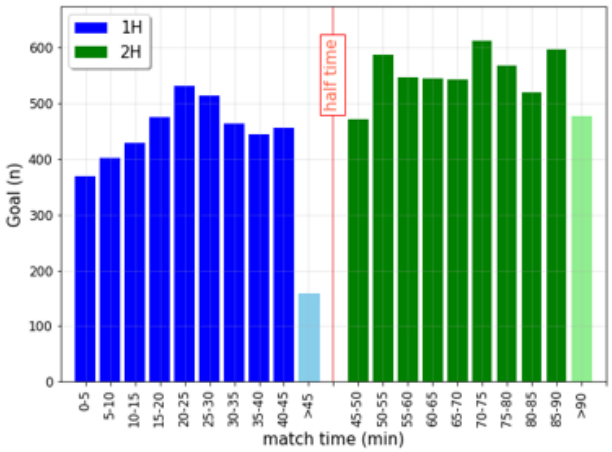


Figure 4: Distribution of Goals over Time

4.3 Feature Engineering

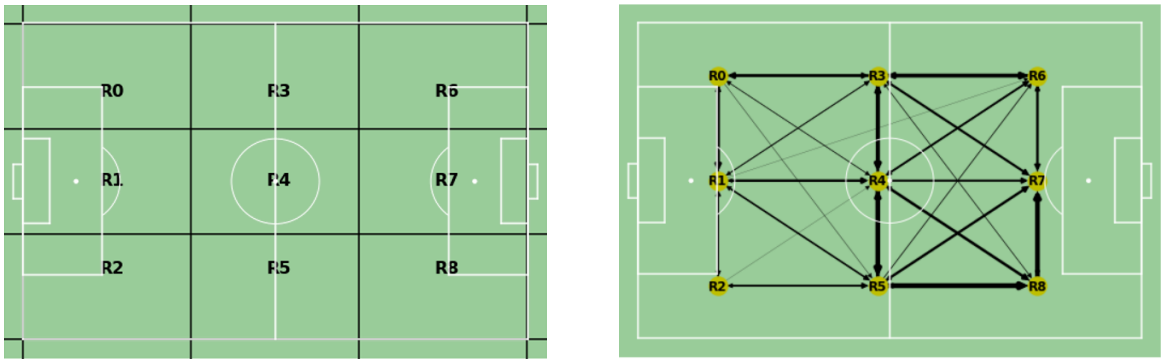
After preprocessing and exploratory data analysis, features were extracted from the dataset to characterize different dimensions of each team's performance in each match. For this, methodology from prior work was generalized and re-applied in the context of this project. Since feature engineering comprised a major component of this project, the rest of this chapter is broken down into sections by the different type of analyses conducted to extract features.

4.3.1 Zonal Networks

First, the spatial aspect of a team's passing activity was considered. The experimental work involved constructing metrics from zonal network graphs to investigate the inter-zone passing activity, and concluded with the derivation of multi-dimensional representations to describe, from a spatial lens, the passing profiles per team, per performance. Graph theory was chosen to analyse the spatial aspect of a team's passing characteristics as it provides a natural way for modelling relationships between different spatial zones on the pitch. Thus, graph theory can expose insights regarding underused and overused pitch zones, and what ball movement patterns across the pitch were most apparent. Using alternative methods to individually study each pitch zone does not provide insights stemming from the relationships between different pitch zones.

The study began by discretizing the soccer pitch evenly into 9 non-overlapping zones, as illustrated in Figure 5a. Choosing to discretize the pitch this way provided an intuitive mapping between the zones and soccer positions: defense, midfield, and forward on the

left, central, and right sides of the pitch. However, a drawback of this method is that the large zone sizes conceal the activity within the zones, as analysis was conducted at the inter-zone level, not the intra-zone level.



(a) 9 Discretized Zones for Inter-Zone Passing Analysis

(b) Team Zonal Network Graph for a Single Performance

Figure 5: Inter-Zone Analysis

After defining the zones, zonal network graphs were produced for each game, for each team. The nodes in these graphs represented the 9 zones while the edges corresponded to the inter-zone passing. The graphs were directed and weighed, with weights being set by the frequency of passes along the corresponding edge. One example of the zonal network graph can be seen in Figure 5b. Here, the boldness of the edges indicates the frequency of passes along the paths. It is visually apparent that R5-R8 was a key passing lane (by passing volume) in this performance.

Building these network graphs enabled reconstructing some of the elemental building blocks of the H-Indicator. Using the methodology from Cintia et al. [11], each zone was attributed the number of passes it "handled" by summing up the in-degree and out-degree for the corresponding node on its network graph. Then, the mean and standard deviation of the passes handled were calculated across the zones. Both these quantities were normalized by the total passing volume of the team in that performance. Equations 1-3 mathematically illustrate these definitions for team t in a certain match. The subscript i refers to one of the pitch zone index as defined in Figure 5a.

$$\forall\ i \in 0, ..., 8: \quad x_i^t = indeg_i^t + outdeg_i^t \tag{1}$$

$$\mu_{zone}^t = \frac{1}{\#passes^t} * \frac{1}{9} \sum_{i=0}^{8} x_i^t \tag{2}$$

$$\sigma^t_{zone} = \frac{1}{\#passes^t} * \sqrt{\frac{1}{9}\sum_{i=0}^{8}(x_i^t - \frac{1}{9}\sum_{i=0}^{8}x_i^t)^2} \tag{3}$$

Note that a high value of σ^t_{zone} implies the coexistence of "hot" zones with high passing activity and "cold" zones with lower passing activity for team t in the performance being analysed. In contrast, a low σ^t_{zone} indicates a more uniform distribution of passing across the zones. Macro-averaged over all performances by team t in the 2017/2018 season, σ^t_{zone} shows a positive correlation with end-of-season points. Qualitatively, the correlation seems to be directionally consistent across all 5 domestic leagues, as illustrated in Figure 6.

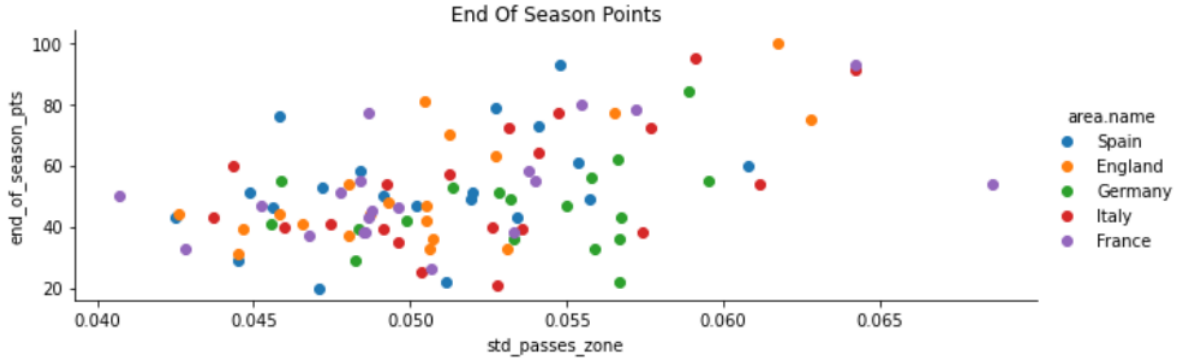


Figure 6: End of Season Points vs Macro-Averaged σ^t_{zone}

Zonal network graphs also enabled the identification of key passing lanes. For instance, from Figure 5b it is evident that R5-R8 is indeed a key passing lane (the boldness of the edge signifies the frequency of passing). This begged the question: Is it possible to segment performances through passing lane activity? If so, are there certain passing lanes that are linked to higher performance outcomes? To answer these questions, the original zonal networks were first transformed into an array of features indicating the edge weights. This yielded a 36 dimensional vector, which was then normalized by the L1 norm[8] to ensure that the weights add to 1. Equations 4-5 formally describe this process. Note that $\mathbf{w}, \mathbf{w^{raw}} \in \mathbb{R}^{36}$ are flattened vectors corresponding to interzone passing activity for all size 2 permutations of the 9 zones, and $\omega : E \to \mathbb{R}$ maps each edge on the zonal network to its corresponding edge weight[9].

$$\forall\ i, j \in 0, ..., 8\ s.t\ i \neq j:\ \mathbf{w}^{raw}_{i,j} = \omega(i, j) \tag{4}$$

[8]The L1 norm was chosen to make it such that the elements in the final vector indicate the percentage of all inter-zone passes that went through the corresponding lane

[9]Equivalently, the number of passes the team played along that edge.

$$\mathbf{w} = \frac{\mathbf{w}^{raw}}{||\mathbf{w}^{raw}||_1} \tag{5}$$

Thus, for each performance, each team had a corresponding $\mathbf{w}$ vector with the 36 elements representing the 36 passing lane intensities. The set of $\mathbf{w}$ vectors for all teams across all performances of the season was then clustered in vector space using the K-Means algorithm with the Euclidean distance metric to determine distinct passing styles from passing lane intensity [19]. Setting K=2 through the elbow method, two distinct clusters were obtained.

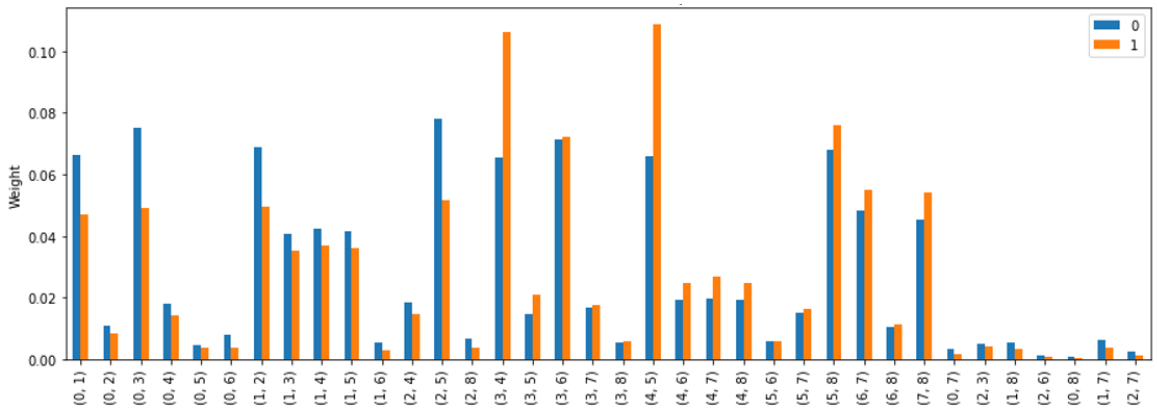
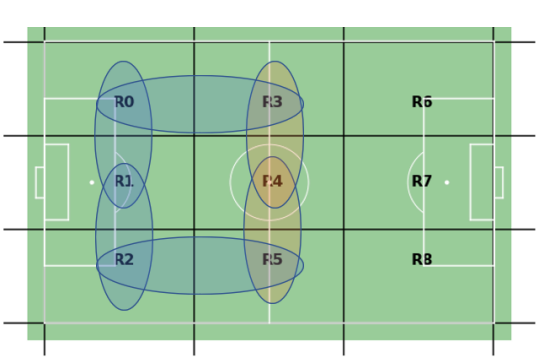


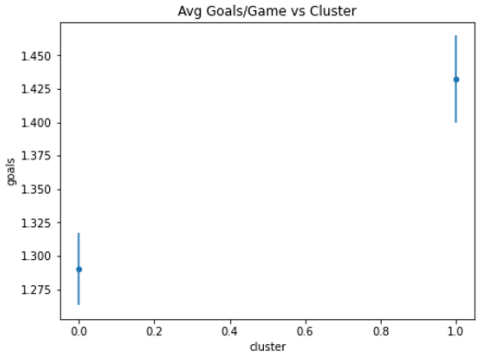
Figure 7: Centroids for the 2 Clusters from Passing Lanes

Figure 7 displays a component-wise comparison for the 2 cluster centroids with regards to the relative weight/intensity along each of the components. It is evident that R3-R4 and R4-R5 stand out as key passing lanes in cluster 1 (orange) while R0-R1, R0-R3, R1-R2, and R2-R5 stand out as key passing lanes in cluster 0 (blue). These key lanes are illustrated in terms of the pitch itself in Figure 8a. The performance discrepancy across performances belonging to the clusters is also apparent, as seen in Figure 8b. In this figure, the average goals per game, along with the standard error bars can be seen for all performances binned in cluster 0 and cluster 1 respectively. Clearly, cluster 1 is associated with more goals per game, on average, than cluster 0. From a fundamental perspective, this corresponds to one popular philosophy in soccer, 'tiki-taka', that mandates high possession and lateral movement of the ball in the midfield to exhaust the opponent.

To verify the correlation between increased lateral passing with performance, a finer grain analysis was done. Moving away from the 9-zone discretization and zonal network graphs, each pass was isolated in terms of its origin and destination coordiantes, and its advance ratio was computed. The advance ratio refers to the pass's lateral (left/right) trajectory as a ratio of its vertical (towards/away from goal) trajectory. Equation 6



(a) Key Passing Lanes for Cluster 0 (Blue) and Cluster 1 (Orange)



(b) Performance Discrepancy across Clusters

Figure 8: 2 Cluster Passing Profile Contrast

provides a mathematical definition of the mean advance ratio for team t in a given performance.

$$MeanAdvanceRatio^t = \frac{1}{\#passes_t} \sum_{i \in passes_t} \frac{|\Delta Y|_i}{|\Delta X|_i} \tag{6}$$

Here, $|\Delta Y|$ and $|\Delta X|$ correspond to the absolute value of the ball movement in the lateral(left/right) and vertical(towards/away from goal) directions respectively. Upon macro-averaging the mean advance ratio per team for all the performances by that team in the 2017/2018 season, a positive correlation with $\rho = 0.68$ was obtained with the end of season points. The relationship can be observed visually in Figure 9.

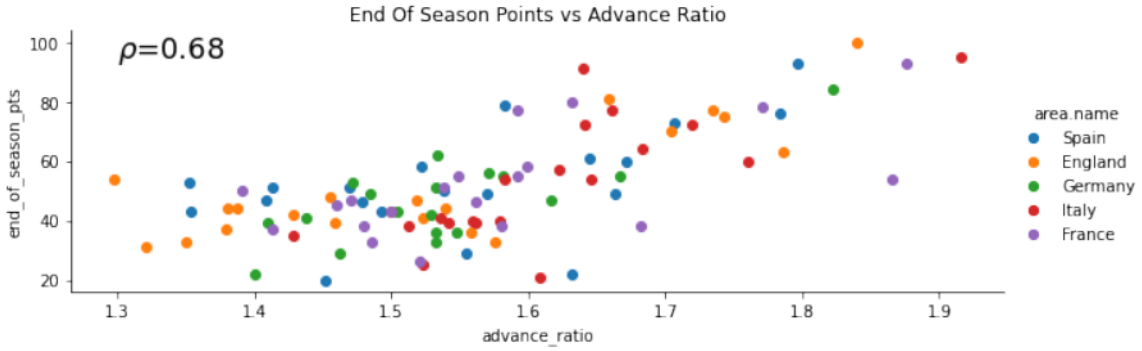


Figure 9: Macro-Averaged Mean Advance Ratio vs End Of Season Points

Thus, zonal network graph analyses yielded 3 metrics do describe passing activity in terms of spatial coordinates: μ^t_{zone}, σ^t_{zone}, and $MeanAdvanceRatio^t$. In addition, the passing lane study yielded 2 distinct clusters in terms of lane-intensities with a performance discrepancy between them. Moving forward, the 3 numerical features and the passing lane cluster (a categorical variable) were taken as individual components describ-

ing a performance, and eventually used for clustering in vector space.

4.3.2 Player Networks

Analogous to zonal networks discussed in the previous section, player passing networks were constructed by defining each player as a node, and passes amongst the players as edges. Again, graph theory was chosen to analyse passing activity amongst players because it provides metrics summarizing the relationship amongst the individual players. Such attributes can not be captured by analyzing each player individually. Note that the soccer-logs dataset [8] did not provide a player ID for the receiver of the pass. Therefore, an assumption had to be made that if a certain player i passes the ball and the next chronological event is from another player j on the same team, then it is assumed that the pass was from i to j. This assumption was necessary to construct player passing networks [10]. A possible way to avoid this assumption in future studies is to use ball-tracking and player-tracking data in conjunction with soccer-logs.

Just as with zonal networks, methodology from Cintia et al. [11] was adopted to attribute to each player the number of passes he "handled" by summing up the in-degree and out-degree for the corresponding node on its network graph. Using equations 7-10, the mean and standard deviations of passes being handled by individual players were evaluated. These definitions hold a one to one correspondence with the analogous definitions for zonal networks in equations 1-3, the only difference being that nodes corresponded to players instead of zones.

$$\forall\ i \in players(t): \quad x_i^t = indeg_i^t + outdeg_i^t \tag{7}$$

$$\mu^t_{player,unnormalized} = \frac{1}{\#players(t)} * \sum_{i \in players(t)} x_i^t \tag{8}$$

$$\mu^t_{player} = \frac{1}{\#passes^t} * \mu^t_{player,unnormalized} \tag{9}$$

$$\sigma^t_{player} = \frac{1}{\#passes^t} * \sqrt{\frac{1}{\#players(t)} \sum_{i \in players(t)} (x_i^t - \mu^t_{player,unnormalized})^2} \tag{10}$$

Again, a high value of σ^t_{player} implied the coexistence of "active" players with high passing activity and "inactive" players with lower passing activity for team t (in the

[10] To see how this assumption was used in code implementation, please refer to Appendix G.

performance being analysed). As before, investigating the player networks yielded a high correlation ($\rho = 0.56$) between the macro-averaged σ^t_{player} values for all performances of team t across the entire season, and the end-of-season points.

Player networks, however, were also analysed using algorithms from graph theory. Borrowing methodology from [15], three metrics were derived from each player network graph[11]:

The **clustering coefficient** of a node u is a measure of the network's local robustness with respect to that node. Intuitively, it counts the number of triangles centered around the node as a fraction of the number of triangles there could have been. To compute the clustering coefficient, the edge weights were not considered, because large correlations with possession were not desired[12]. The metric was derived as illustrated in Equation 11.

$$c_u = \frac{1}{(deg^{tot}(u))(deg^{tot}(u) - 1) - 2deg^{\leftrightarrow}(u)} T(u) \tag{11}$$

where $deg^{tot}(u)$ refers to the sum of in degree and out degree of node u, $deg^{\leftrightarrow}(u)$ is the reciprocal degree of node u, and $T(u)$ is the number of directed triangles through node u. More details on the intuition behind the clustering coefficient can be found in Appendix B.

The **algebraic connectivity** is a measure of integration/segregation of nodes in the network, with a value of 0 indicating complete separation amongst communities. Mathematically, it is defined as the second smallest eigenvalue of the Laplacian matrix of the network. The Laplacian matrix(L) is defined as in Equation 12:

$$L = D - A \tag{12}$$

where $D = diag(d_1, d_2, ...)$ is a diagonal degree matrix formed by the degrees of vertices d_i and A is the adjacency matrix. A high algebraic connectivity can be interpreted as the network being tolerant to faults. In the context of soccer player networks, this means that if a player is marked by an opponent's defender, or off the field (due to a red card or injury), the network as a whole is highly tolerant to it in terms of ball movement. More details on the intuition behind the clustering coefficient can be found in Appendix B.

The **average all pairs topological shortest path** is a measure of how frequently

[11] Recall that there is one player network graph per match, per team

[12] Using an unweighed definition of the clustering coefficient also ensures that this the clustering coefficient is not heavily skewed by extremely frequent 'one-two' passes between a specific pair of players.

the ball was passed between pairs of nodes on average. To compute this, the edge weight between nodes i and j was set to $\frac{1}{frequency\ of\ passes\ from\ i\ to\ j}$. A low value indicates movement of the ball between 2 random nodes, on average, was performed frequently.

All three of these metrics were computed to describe ball movement across the players. The average clustering coefficient, the algebraic connectivity, and the average all pairs topological shortest path, when macro-averaged for the entire season, yielded fairly linear trends in team performance in terms of end of season points, as illustrated in Figure 27 in Appendix A.

To summarize, player network graphs yielded 2 analogous metrics to zonal graphs: μ^t_{player} and σ^t_{player}. In addition, running graph algorithms on these networks yielded the average clustering coefficient, the algebraic connectivity, and the average all pairs topological path. These five metrics provide a way to represent the passing activity amongst players on a team across five dimensions, and they constitute components in a team's eventual performance representation in order to perform playing style clustering.

4.3.3 Defense Analysis

In addition to passing, which was explored via player and zonal network graphs, defensive aspects of games were also explored. Although a graph theory analysis was not conducted to engineer defensive features, a shallower analysis was done and the event tags from the soccer-logs dataset[8] were used to enumerate and aggregate event counts. Note that this makes the analysis more limited in scope compared to passing because it does not involve quantifying "team chemistry" amongst players in defensive roles. However, such a limitation was inevitable given the dataset because there is no information available for the movement of players off the ball. It was therefore not possible to gauge how defensive lines were structured as a whole unit, and thus individual event enumeration was adopted as the approach moving forward. Table 3 summarizes the events along with their correlation to end-of-season points when macro averaged.

Note that the fraction of slide tackles and interceptions inside the box yielded an overall negative correlation to end of season points. The next natural questions were: does this pattern continue further down the pitch? Does the position of defensive events (slide-tackles/interceptions) correlate with performance in general or is it unique to the penalty box? To answer these, the spatial X and Y coordinates were referenced again. More specifically, the centroids of slide tackles and interceptions were computed per game for each team, and the macro-averaged results analysed against the end of season per-

Table 3: Defensive Events & Correlation to End Of Season Points

Event	**Correlation To EOS Points**
Number of Yellow Cards / Game	-0.43
Dangerous Balls Lost	-0.30
Interceptions Inside Box	-0.66
Total Interceptions	-0.67
Fraction of Slide Tackles in Box	-0.48
Slide Tackles Inside Box	-0.57
Total Slide Tackles	-0.55
Fraction of Slide Tackles in Box	-0.25

formance (like all the other analyses so far). The macro-averaged X and Y centroids had correlation coefficients of 0.64 and -0.02 respectively with the end of season points. Note the correlation is only apparent in the forward/backward direction. The left/right coordinate is immaterial as far as correlation with performance is concerned. This is in line with expectations as stopping opponent attacks further down the pitch (a higher X coordinate) is less risky because set pieces become more difficult, and the likelihood of conceding a goal from a set-peace decreases. With these defensive metrics, the overall

Table 4: A Summary of All Engineered Features

Type of Analysis	**Number of Features Derived**
Exploratory Data Analysis	1
Zonal Networks	3
Passing Lane Clustering	1 cluster label or 36 individual features
Player Networks	5
Count-based Defense Metrics	10

features space comprised of 20 derived features as summarized in Table 4 [13]. Therefore, with the feature engineering concluded, each performance for each team could be characterized across these 20 dimensions as a 20 dimensional vector. The next section will describe early events to use these 20 dimensional vectors for clustering playing styles.

4.4 Unsupervised Clustering

After engineering the 20 features from the dataset describing each performance as summarised in Table 4, attempts were made to find representations for each team by going from performance features to team features. Then, clustering approaches were

[13]See Appendix C for a detailed description for each of the 20 derived metrics

used. In particular, a set of results was obtained via the following two approaches:

1. PCA Dimensionality Reduction & K-Means Clustering
2. Hierarchical Clustering: Ward's Method

The methods for these two approaches are discussed in detail in the rest of the chapter. Note that each of the methods has its pros and cons, but a major bottleneck of the PCA Dimensionality Reduction & K-Means Clustering approach is the lack of intractability in terms of the feature space. As discussed towards the end of Section 4.4.1, this lead to limitations in the analysis work, which was primarily an extension of the second approach (Hierarchical Clustering: Ward's Method). Also note that other clustering techniques such as simple K-Means, Gaussian Mixture Models, OPTICS, and Mean Shift Clustering were also explored, but they did not yield fruitful results as they were highly non-robust to initialization and some of the methods did not successfully converge. For this reason, they were dropped, and are not discussed comprehensively here.

4.4.1 PCA Dimensionality Reduction & K-Means Clustering

This approach began with taking the 20 features for each performance (expressed as vector $p \in \mathbb{R}^{20}$) and compressing them to a set of features for each team. This was done by taking the mean and variance of the feature for a particular team across all games in the season. As a result, each team's representation corresponded to a 40 dimensional vector[14], $t_i \in \mathbb{R}^{36}$, where the subscript i denotes the team. The mathematical formulation of this step can be found in the following equations. In particular, for every team i in the set of all teams:

$$\mu_i = \frac{1}{|performances(i)|} \sum_{j \in performances(i)} p_j \tag{13}$$

$$\sigma_i^2 = \frac{1}{|performances(i)|} \sum_{j \in performances(i)} (p_i - \mu_i)^2 \tag{14}$$

$$t_i = [\mu_i;\ \sigma_i^2] \tag{15}$$

where $performances(i)$ is the set of all performances of team i throughout the 2017/18 season, and ';' denotes vector concatenation. The 40 dimensional team repre-

[14] 20 from the mean across performances and 20 from the variance.

sentations were then standardized via the cross sectional Z-Score normalization across all teams, as illustrated in Equations 16-18.

$$\mu_{teams} = \frac{1}{|teams|} \sum_{i \in teams} t_i \tag{16}$$

$$\sigma_{teams} = \sqrt{\frac{1}{|teams|} \sum_{i \in teams} (t_i - \mu_{teams})^2} \tag{17}$$

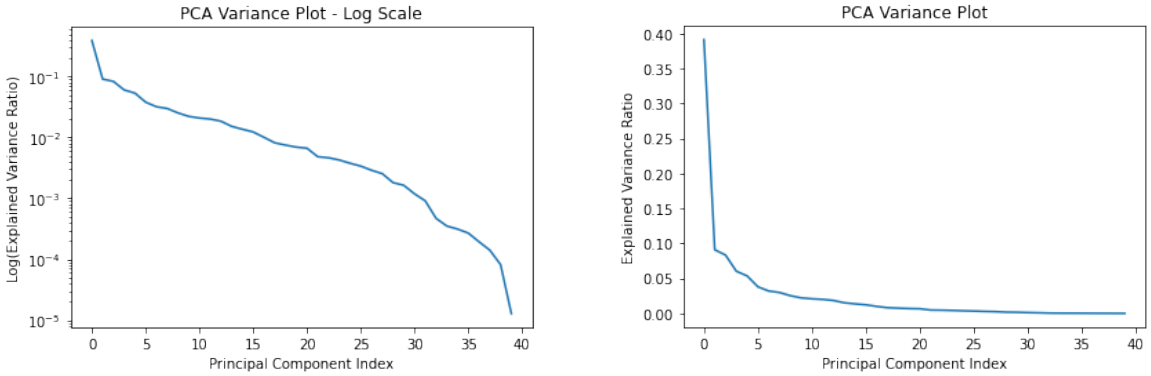
$$z_i = \frac{t_i - \mu_{teams}}{\sigma_{teams}} \tag{18}$$

Note that square and the division operations in Equations 17 and 18 are element-wise, resulting in an independent Z-Score normalization for each of the features. After Z-score normalization, the features were decomposed via singular value decomposition to perform principal component analysis. In particular, the explained variance was analysed with respect to each principal component, as shown in Figures 10a and 10b. The plots indeed suggest that many of the principal components carry little information, as much of the original feature space has correlated features. In light of this, the top 15 principal components were selected, and the rest were not considered moving forward. This decision reduced the feature space from 40 dimensions to 15 dimensions.

Taking the 15 principal components, and re-standardizing them each dimension with a Z-Score normalization using the same steps as before (Equations 16-18), clustering was attempted via the K-Means clustering algorithm, using euclidean distance as a proximity measure. Setting K=3, 3 distinct clusters were obtained. A statistically significant performance discrepancy across the three clusters was also found. The discrepancy in goals scored, goals conceded, and end of season points (normalized[15]) can be observed in Figure 12.

Note that since the clustering is occurring in the principal component space rather than the original feature space, interpretability is highly compromised. Looking at Figure 11, it is very difficult to decipher what it means for cluster 0 to be very high in component 0, because principal component 0 itself is a linear combination of the 40 original features. The same holds true for the remaining 14 principal components. Therefore, despite obtaining an orthogonal feature space, reduced dimensionality, and a segmentation by playing style that translated to a statistically significant performance discrepancy, work

[15]Normalization by games played is required for a sound comparison because the German Bundesliga is shorter than the other leagues, and teams play fewer games, as discussed previously in Section 4.2.



(a) Explained Variance Semilog Plot (b) Explained Variance Plot

Figure 10: 2 Cluster Passing Profile Contrast

Figure 11: Centroids for the 3 Clusters From PCA and K-Means

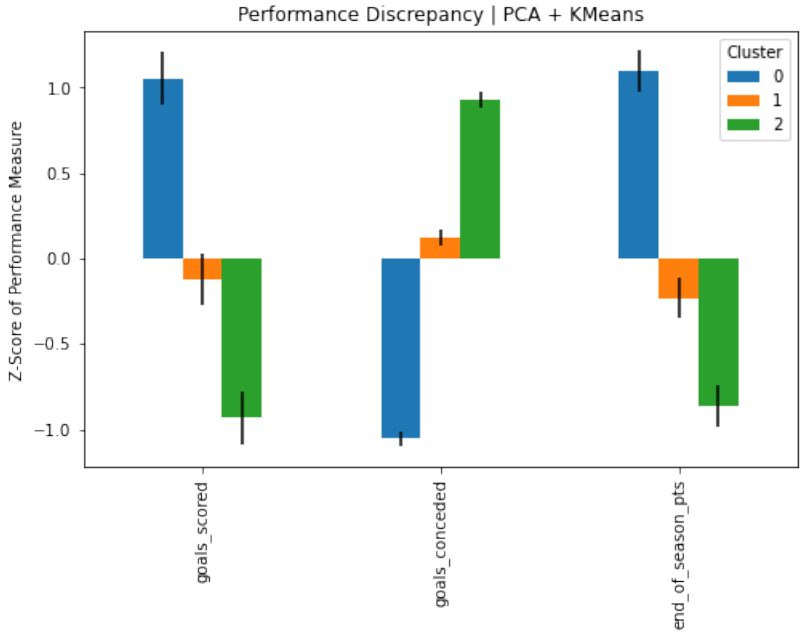


Figure 12: Performance Discrepancy Across the 3 Clusters From PCA and K-Means

in this direction was not pursued further, as the lack of playing style interpretability severely limits the insights that can be made from this work. A more detailed discussion on this limitation is presented in Chapter 6: Limitations & Next Steps.

4.4.2 Hierarchical Clustering: Ward's Method

Since the use of PCA for dimensionality reduction fundamentally compromises feature interpretability, as explained in Section 4.4.1, an alternative approach was adopted where the original 40 features, as derived by Equations 13-15, were not compressed. Instead, the original features were preprocessed, scaled by the amount of information they possess, and then used for hierarchical clustering via Ward's Method. Hierarchical clustering was chosen as opposed to K-Means as it provides a more interpretable illustration of the clustering process without presetting the number of clusters. The rest of this section expands on the methodology used.

As before, the 40 features were normalized cross sectionally, as indicated by Equations 16-18. Then, the amount of information in each of the features with regards to performance was estimated via the mutual information between the feature itself and the team's performance quartile[16]. Intuitively, mutual information measures the 'amount of information' that can be obtained about one random variable by observing another random variable. The mathematical definition for mutual information (denoted by $I(X;Y)$) between random variables X and Y can be found in Equation 19, where $H(X)$ denotes the entropy of random variable X. To estimate this between for a continuous X and discrete Y, the esitmator from Brian C. Ross[20] was used.

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) \tag{19}$$

The mutual information values derived can be seen in Figure 13. These values have been normalized to sum to 1. For a detailed plot with the exact X features described, please refer to Appendix D, Figure 30. It can be seen from Figure 13 that some features have high mutual information with performance, whereas a subset of 7 features out of the original 40 contain negligible amount of mutual information with performance.

These mutual information scores were used to scale the features so that the standard deviation of the features across all teams becomes proportional to the amount of mutual

[16]The performance quratile for each team was calculated via its end-of-season points.

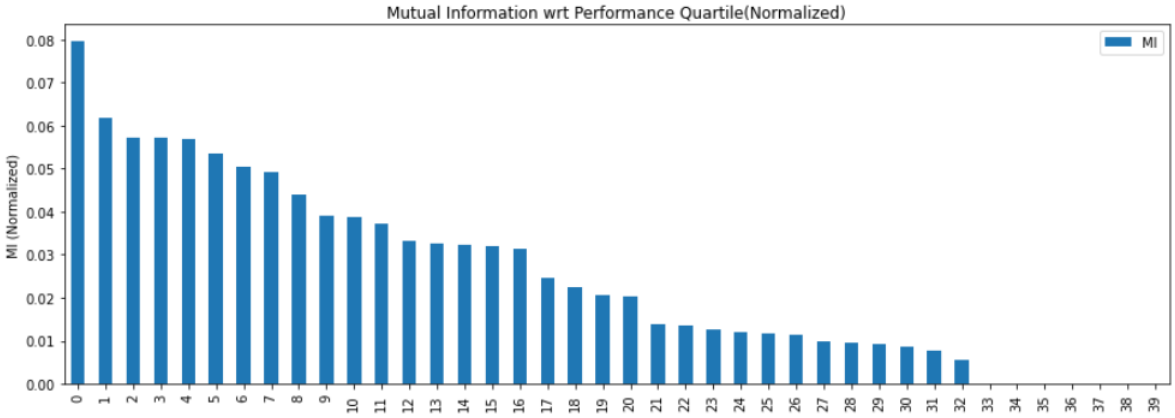


Figure 13: Normalized Mutual Information between Features & Performance Quartile

information between the features and performance. This was a necessary step in the feature processing as it is a way to design the problem such that distance between two teams in features with high mutual information is penalized more than similar distance between the two teams in features with low mutual information during the clustering process.

The next step involved taking these scaled features and clustering them via bottom up hierarchical clustering. Bottom-up clustering starts with each team belonging to it's own cluster, and merges clusters recursively until a single global cluster is obtained. As a merging criterion, Ward's method of bottom up clustering [21] was used. This method merges clusters at each step by minimizing the total internal variance within all the clusters. Thus, it merges cluster X with cluster Y if and only if merging them results in the minimal increase in the total internal variance of all clusters (equivalently, the error sum of squares or ESS) as compared to any other cluster pair. The process (for a single step in the algorithm) is mathematically described in Equations 20-22.

$$ESS(X) := \sum_{i=1}^{N_x} ||x_i - \frac{1}{N_x} \sum_{j=1}^{N_x} x_j||_2^2 \tag{20}$$

$$D(X, Y) := ESS(X \cup Y) - [ESS(X) + ESS(Y)] \tag{21}$$

where X and Y are existing clusters at some step of the algorithm, and x_i represents a single vector (corresponding to a team) in cluster X. Each iteration[17], merging occurs

[17]For the first iteration, teams are simply merged to the team that is closest by squared euclidean distance.

via minimizing $D(X, Y)$:

$$C_1, C_2 = argmin_{X,Y} D(X, Y) \tag{22}$$

C_1, C_2 are merged together, and the algorithm proceeds recursively until a single global cluster is formed. The algorithm was applied to the cluster teams; the process is illustrated in the dendrogram in Figure 14.

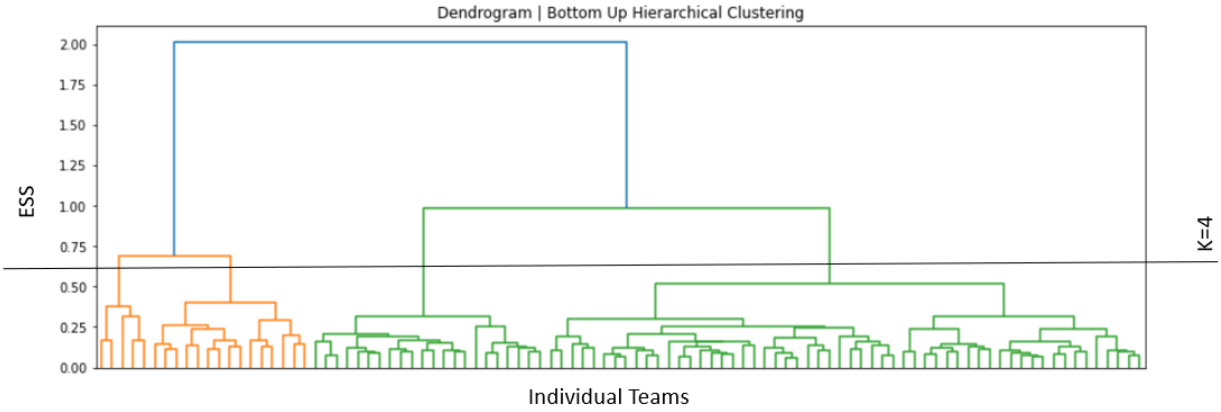


Figure 14: Dendrogram for Team Clustering with Ward's Method

One of the reasons Hierarchical Clustering with Ward's Method was chosen as a clustering method over other alternatives such as K-Means is that the algorithm does not require presetting K, the number of clusters, apriori. This allows the slicing of the denrogram in any manner to obtain the desired number of clusters after analysing the entore clustering procedure. For the purpose of team playing style classification, K=4 was chosen and the dendrogram was sliced via a horizontal line as shown in Figure 14. This value of K was chosen as it was the most granular manner to slice the dendrogram without the clusters getting too small[18].

Figure 15 shows the centroids for the four clusters obtained. For a detailed plot with exact descriptions of each of the 40 x-axis features, please refer to Appendix D, Figure 31. Note that this centroid profile reveals correlated features unlike those obtained in Section 4.4.1. The distribution of the cluster centroid vectors is on a spectrum between two extremes. This is because this method does not enforce orthogonality in the feature space (unlike PCA). This is a source of an inherent limitation in the project as the original features are required to make insights, but using the the original features also introduces such correlations. More discussion follows in Chapter 6: Limitations & Next Steps.

As before, a notable performance discrepancy across the 4 clusters was also observed,

[18]Already with K=4, a cluster with only 4 teams is obtained.

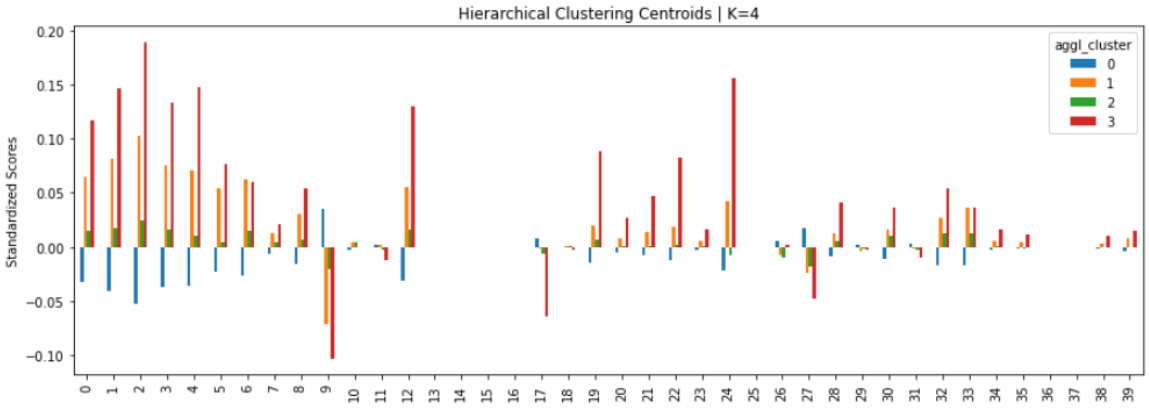


Figure 15: Centorids for the 4 Clusters Obtained

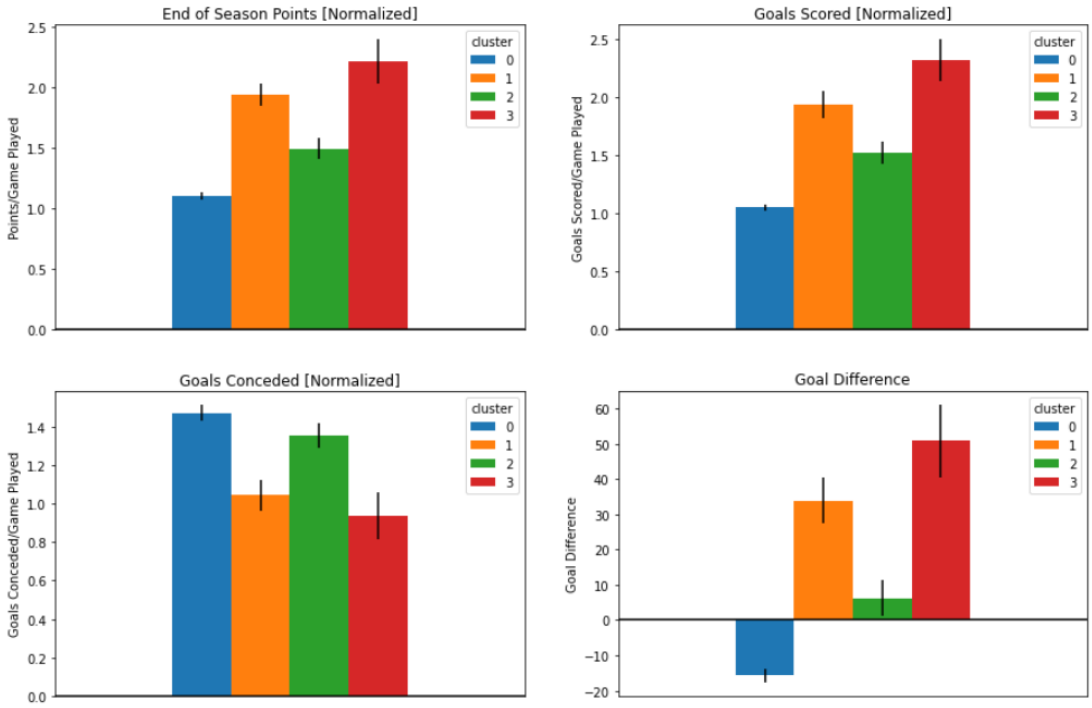


Figure 16: Performance Discrepancy Across the 4 Clusters Obtained

as shown in Figure 16. Note that the end of season points, the number of goals scored, and the number of goals conceded are all normalized by the number of games played. This is required for a balanced comparison because the German Bundesliga is slightly shorter than the other domestic leagues, as explained in Section 4.2.

Thus, by feature scaling with mutual information, and hierarchical clustering with Ward's method, 4 clusters were obtained with a noticeable performance discrepancy between them across four metrics: end of season points, goals scored, goals conceded, and goal difference. The next chapter will discuss the details regarding the constituent teams in each of these 4 clusters.

5 Discussion & Analysis

The immediate results that followed from Section 4.4.2 resulted in 4 disjoint clusters of soccer teams, with performance discrepancies and a comparison of cluster centroids presented in Figures 16 and 15 respectively. This section will discus findings from these cluster assignments and dive deeper into the exact teams within each cluster. To begin, Figure 17 below shows sample cluster assignments along with cluster sizes for each of the four clusters obtained. A complete breakdown of each cluster by constituent teams is presented in Appendix E.

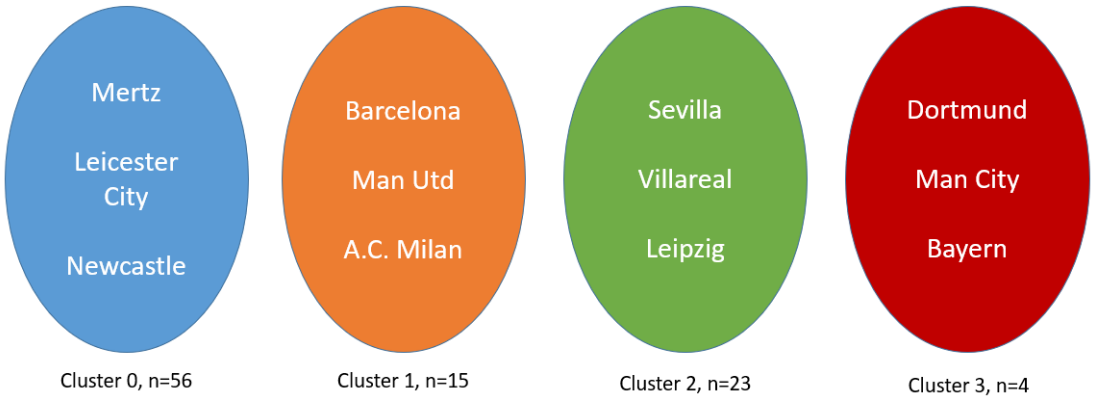


Figure 17: Performance Discrepancy Across the 3 Clusters From PCA and K-Means

From Figure 16, it can be observed that teams in clusters 1 and 3 show better performance than those in clusters 0 and 2 across all four metrics: end of season points, goals scored, goals conceded, and goal difference. Therefore, the first natural question is, how are clusters 1 and 3 different from a cluster showing low performance (cluster 0)? Furthermore, Figure 13 reveals that the centroids for clusters 1 and 3 are directionally correlated, which begs the question: are clusters 1 and 3 really different playing styles

or are we simply grouping by team skill level? The following two case studies attempt answering these questions.

5.1 Case Study: Clusters 1 & 3 vs Cluster 0

5.1.1 Differences in Feature Space

From Figure 18, it can be observed that clusters 1 and 3 are directionally correlated; cluster 3 seems to display similar characteristics to cluster 1 but in more extreme magnitudes. Conversely, cluster 0 appears to be the opposite extreme. Note that 'mean' and 'var' in Figure 18 refer to the mean and variance of the 40 features for each team used for summarizing performance metrics to team metrics, as described in Equations 13-15. Furthermore, 'w' refers to an estimate of team possession approximated by the proportion of passes a team plays, as indicated in Appendix C.

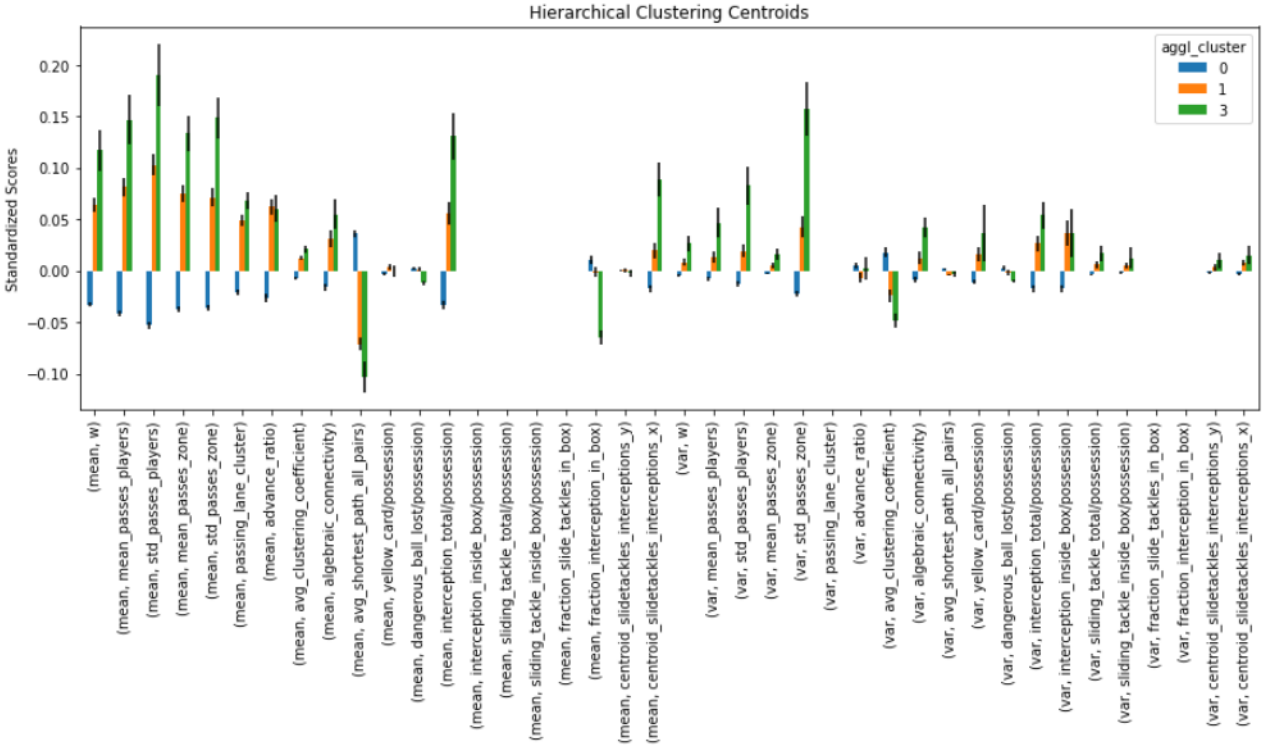


Figure 18: Clusters 1 & 3 vs Cluster 0 in Feature Space

In particular, clusters 1 and 3 both display high average possession, and consequently a high number of passes handled across players and zones. The standard deviation of the number of passes handled by each zone (and player) is also higher, on average, than cluster 0. These metrics indicate high passing activity in both clusters 1 and 3 as opposed to cluster 0, along with a higher affinity for the coexistence of hot/cold zones as well as

active/passive players. Given a low topological average shortest path length between all pairs, teams in clusters 1 and 3 distribute the ball amongst the players on the field in a comparatively more direct manner.

The discrepancy in algebraic connectivity between clusters 1 & 3 and cluster 0 is also significant. The higher average algebraic connectivity in clusters 1 & 3 indicates that the passing activity in these clusters is globally more robust to faults in the player network nodes [19]. Fundamentally, this implies that ball movement across players is not as severely bottle-necked in the event that a player (or multiple players) is marked by the opposing defenders, injured, or sent off due to a red card, because the passing network is highly integrated and remains connected even under node removal.

Note that the standardized score for the mean passing profile is also significantly higher in clusters 1 & 3 as opposed to cluster 0. This implies that on average, team performances in clusters 1 & 3 were classified in passing profile 1, as defined in Figure 8 in Section 4.3.1, more often[20] then teams in cluster 0. Therefore, teams in clusters 1 & 3 show increased lateral passing activity in the midfield (R3-R4-R5 in Figure 8) while teams in cluster 0 had higher defensive passing activity and a higher proportion of vertical passes along the defensive flanks (R0-R3 and R2-R5 in Figure 8).

This difference in the spatial aspect of passing activity is also captured by the advance ratio, which is a measure of passing directness (as derived in Section 4.3.1, Equation 6). Clusters 1 & 3 show a noticeably high average advance ratio compared to cluster 0, which indicates that passes, on average, spanned the lateral(left/right) direction significantly more than the vertical(towards/away from goal) direction in clusters 1 and 3, as compared to cluster 0. Finally, given the high X-centroids of slide tackles and interceptions in clusters 1 & 3, it can also be inferred that teams in these clusters played a higher defensive line comparatively.

To summarize, from the centroids in the feature space, it can be inferred that teams in clusters 1 & 3 displayed significantly higher possession than those in cluster 0, which translated to a higher affinity for the coexistence of hot/cold zones and active/passive players. They displayed higher global robustness in the player passing networks, implying greater resilience to faulty nodes in the passing network, while also demonstrating significantly higher lateral passing activity as opposed to direct, vertical passing. Finally, the spatial aspects of their defensive activity implied that on average, they played their

[19]For an intuitive explanation of how this 'robustness' is captured by the algebraic connectivity, please refer to Appendix B.

[20]The passing profile is encoded to and binary indicator. Thus, a higher mean value implies performances classified as '1' were more frequent.

defensive lines much higher up the pitch, corresponding to (attempted) turnovers much further away from their own goals (much closer to opponents' goals).

5.1.2 Anomalous Performance: FC Schalke

As illustrated previously, Figure 16 shows the performance discrepancy between teams in all four clusters. It is evident from all four metrics that teams in clusters 1 & 3 show superior performance. Probing further into the constituent teams within the clusters (see Appendix E for the full list), it is found that all 4 teams in cluster 3 and 10/14 teams in cluster 1 are top-5 teams by end of season table standings. Cluster 0, however, just has one team ranked in the top-5: FC Schalke. The rest of this subsection will attempt to answer why Schalke shows such anomalous performance, while the next case study will dive deeper on the nuances across clusters 1 and 3.

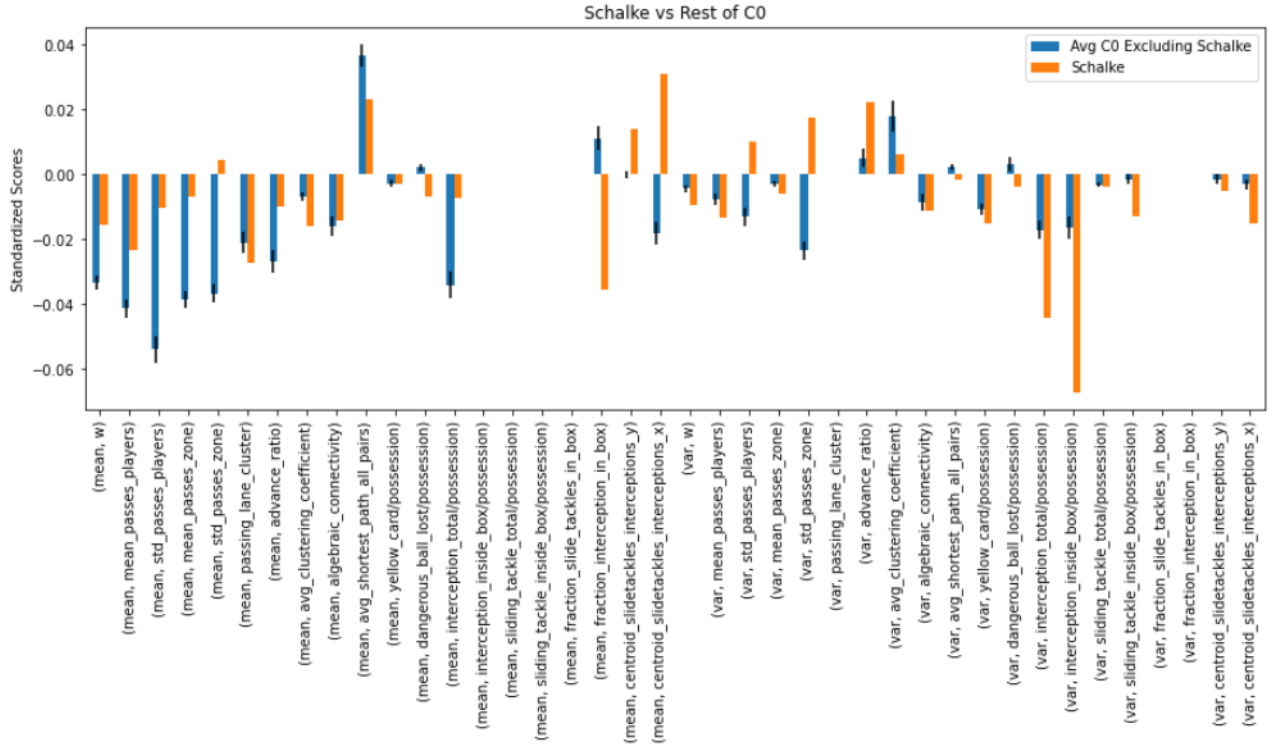


Figure 19: Schalke vs Rest of Cluster 0

The differences between FC Schalke and the rest of cluster 0 in the feature space are illustrated in Figure 19. It is immediately apparent that while Schalke did end up being classified as a cluster 0 team according to Ward's hierarchical clustering criterion, it is definitely on the extreme end of the cluster as it differs from the rest in a significant manner. Specifically, metrics stemming from passing possession such as number of passes played (w), the number of passes handled per zone and player, and the coexistence of

hot/cold zones and active/passive players are not as low for FC Shalke as they are for the rest of the teams in cluster 0. Also unlike the rest cluster 0, the fraction of slide tackles and interceptions within the box is extraordinarily low, which is indicative of Schalke forcing turnovers further down the pitch. This is reaffirmed by Schalke's high X-centroid of slide tackles and interceptions - an indication that the defensive line is set much higher up the field.

It is also interesting to note that while Shalke ended up being classified in passing profile 1 (as defined in Figure 8 in Section 4.3.1) almost as often as the rest of the teams in cluster 0, they displayed a significantly higher advance ratio. Thus, even though their passing activity was not concentrated amongst the lateral midfield zones (as opposed to defensive flanks) any more than the rest of the cluster, the average pass Schalke executed, irrespective of pitch zone, did have a comparatively larger lateral trajectory than a vertical trajectory. Finally, note that Schalke's game to game variance for the number of total interceptions and interceptions within the box is lower in comparison to the rest of cluster 0, implying a higher consistency in defensive aggressiveness. As a reminder, all the defensive metrics (including the two mentioned here) were normalized by opportunity[21].

To summarize, FC Shalke's anomalous performance may potentially stem from their relatively high possession, as well as a higher variance in passing activity across players and zones. Furthermore, FC Shalke displays a significantly higher advance ratio compared to the rest of the cluster, indicating that they move the ball more frequently in a side to side manner compared to directly going for a direct approach towards the opponent's goal. Finally, FC Shalke also stands out as the team shows high consistency in defensive aggressiveness compared to the rest of the cluster. These characteristic qualities explain, to an extent, why FC Shalke shows much better performance than the rest of the teams in cluster 0,

5.2 Case Study: Is the Cluster Distribution Capturing Anything Beyond Skill?

As noted previously, clusters 1 and 3 are directionally similar, but cluster 3 has more extreme features (represented by the higher magnitudes in the cluster centroid

[21]This indicates that metrics involving enumeration of slide tackles and interceptions were normalized by the **opponent's** possession since it is a measure of the opportunity the team head to execute slide tackles and interceptions.

in Figure 15). It was also noted in Section 5.1.2 that cluster 3 has higher performing teams than cluster 1, as all 4 teams in cluster 3 are top-5 by end of season points, while only 10/14 teams in cluster 1 fall under this category. This begs the question: are the cluster centroids simply measuring skill level? Is playing style (by feature space) yielding anything beyond a team's ability to dominate over a season?

To investigate if the extreme centroids are a direct reflection of the skill of the teams in the corresponding cluster, cluster 1 was split such that the 10 teams having the top-5 ranking were isolated from the 4 teams that did not fall under this category. The two sub-clusters, denoted cluster '1+' and '1-' respectively, were compared to cluster 3 independently. Note that the lowest ranked team in the '1-' cluster is ranked 10. Plots comparing the centroids of cluster 1+, cluster 1-, and cluster 3 can be found in Figures 20-24. For a more detailed visualization for all 40 features, please refer to Appendix F.

From Figures 20-24 it is apparent that the difference between the centroids for cluster 1+ and cluster 1- is not significant, especially when intra-cluster variance is accounted for. However, they both differ significantly from cluster 3's centroid. The same pattern generalizes beyond these 5 features, as is evident from the visualization in Appendix F. This breakdown indicates that the top-5 teams in cluster 1 are not significantly different in the feature space from the rest of the cluster. Furthermore, the top-5 teams in cluster 1 and the remaining teams in the cluster compare to cluster 3 similarly. Thus, it can be concluded that playing style is not exclusively a measure of skill. The extreme feature profile displayed by cluster 3 teams is not simply because the teams in the cluster are more dominant on average, since the top-5 cluster 1 teams display a noticeably different feature distribution to cluster 3 while being at the same level of dominance. Therefore playing style, as derived by hierarchical clustering, is not purely a function of skill in the general case. However, it is still important to note that cluster 3 teams possess extreme features, and are all ranked in the top-5 while cluster 1 (with less extreme features) is more mixed.

To summarize the discussion, by splitting cluster 1 into cluster 1+ and cluster 1-, and comparing to cluster 3, it was found that:

1. Even under similar skill levels, teams in different clusters display different feature profiles (on average). This is indicative that these teams have intrinsically different playing styles, as captured by the features.

2. Splitting cluster 1 into highly skilled teams (1+) and second tier teams (1-) did not result in a statistically significant difference between the two sub-clusters, as

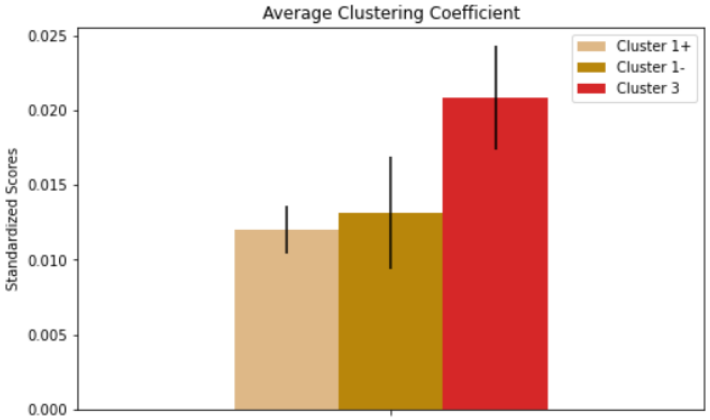


Figure 20: Average Clustering Coefficient

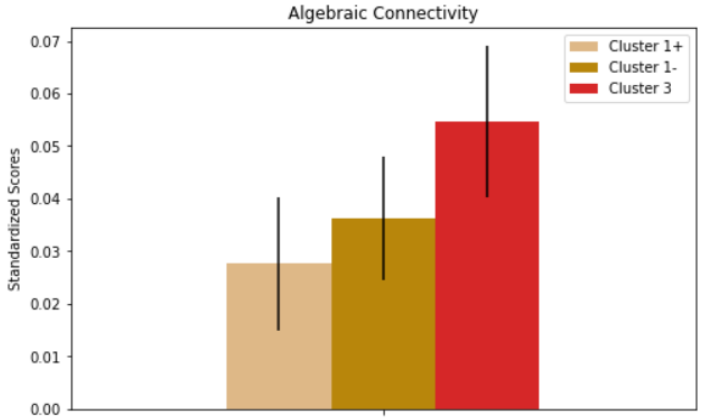


Figure 21: Algebraic Connectivity

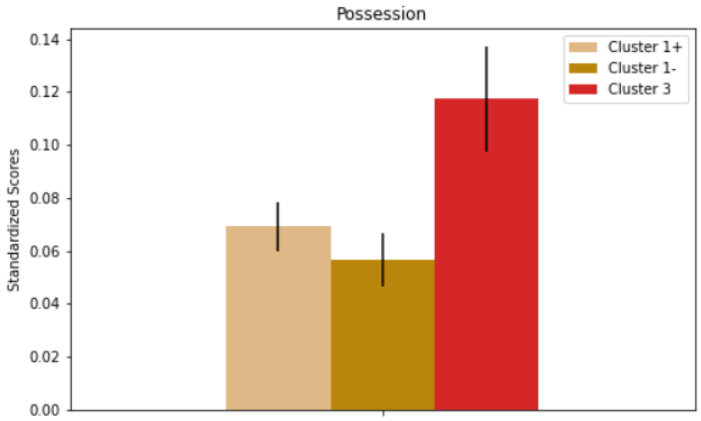


Figure 22: Possession

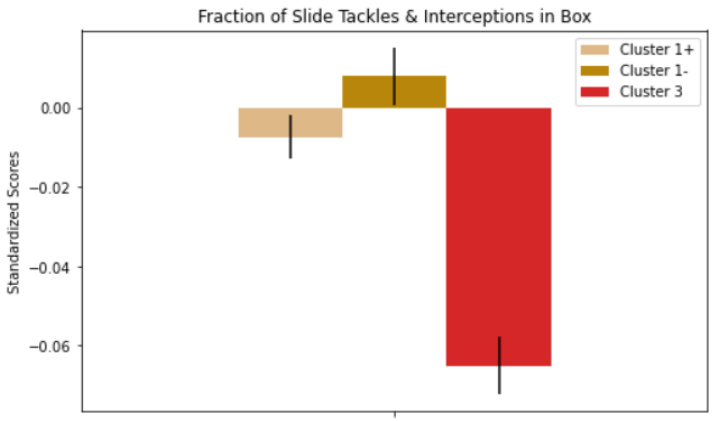


Figure 23: Fraction of Slide Tackles & Interceptions in Box

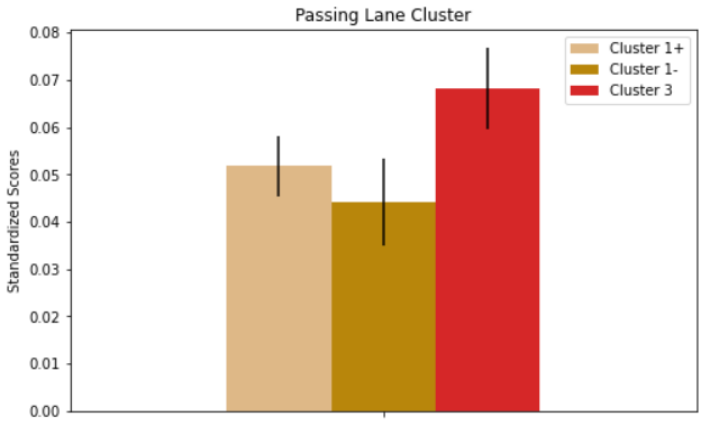


Figure 24: Passing Lane Profile

shown in Figures 20-24. This demonstrates that playing style, as captured by the engineered features, goes beyond skill level.

Both these findings are evidence for demonstrating a level of independence between pure skill/dominance and playing styles derived by hierarchical clustering. However, the correlated feature space limits the strength of the evidence to support the claim as the clustering process itself is prone to being skewed towards some features, as elaborated on in chapter 6.

6 Limitations & Next Steps

As hinted earlier, a major limitation in the framework presented is the fact that the 40 features being used for clustering and analysis are inter-correlated. This is especially evident in Figure 15, as the cluster centroids are not orthogonal. Rather, they are on a spectrum from one extreme (cluster 0) to another extreme (cluster 3). The correlated feature space indicates that the information from the 40 features is limited. Practically, this means that in the hierarchical clustering process, where each feature is given importance in proportion to its mutual information with performance, there may be skew towards certain features that are highly correlated. For instance, if two features are perfectly correlated, then only one of them should ideally be considered for calculating the euclidean distances and intra-cluster variance in Ward's algorithm, as the other feature is providing no extra information. Since the current framework does not penalize these correlations, the skewness in the clustering outcome is inevitable, and it thus forms an inherent limitation in the current framework.

A workaround to accommodate for this limitation is to use feature transformation approaches for dimensionality reduction. One such approach, principal component analysis (PCA), was discussed in Section 4.4.1, and it indeed resulted in uncorrelated cluster centroids due to the orthogonal feature space, as presented in Figure 11. However, this came at a cost of interpretability, as each principal component itself was a linear combination of all 40 of the original features. Thus, it became very difficult to conduct qualitative analysis (similar to the analysis presented in chapter 5) on what the different centroids meant for the clusters obtained, and how they translated to soccer fundamentals to describe playing styles. Work on this approach was therefore dropped. However, if the context of the research problem changes such that interpretability is not a strict requirement, using PCA for dimensionality reduction and eliminating correlated features

is a viable option to make the clustering process more robust. An example context could be the unsupervised clustering of teams where playing style insights are not required.

Since any form of feature transformation will, to an extent, compromise interpretability of the clusters, the problem must use the 40 original features if fundamental insights are required. To make the clustering procedure more robust to this inherent limitation in the feature space, more complex clustering approaches could be adopted, where the distance measure is not a simple euclidean, cosine, or manhattan distance, but rather a custom defined metric that accommodates for correlated features. If bottom up hierarchical clustering is desired, then Ward's merging criteria, as presented in Equation 22, would need to be modified to account for feature correlations. A potential next step could be to explore these avenues of research and derive novel methodologies for robust clustering in a correlated feature space. This would take the research from an applied data science realm (as it is presented in this project) to a domain of more theoretical statistics, as out of the box clustering algorithms would no longer suffice.

Another major limitation of the project is the granularity of the analysis. The current approach attempts playing style classification for teams, rather than for individual games. While the game to game variance of the features over the course of the season is used as an input feature[22], future work may attempt analysis at a more granular level. The added granularity can certainly be used to capture nuance between early season, when players come back from a summer holiday, and end season, when many squads become more vulnerable due to a lacking squad depth resulting in high fatigue. More granularity can also capture the impact of signing players in the January transfer window, as certain key players may change team playing styles entirely. Furthermore, with the added granularity, the effect of 'form' (how teams were playing leading upto a game) and injured players can also be analysed. For such analysis, playing style segmentation must be done for each performance/game, rather than each team.

Finally, note that there is an inherent limitation in the current project that stems from the nature of the metrics derived in chapter 4. All the features/metrics are computed to summarize a team's performance in a particular game, but no intra-game analysis is conducted. In particular, the temporal nature of the data available is not currently exploited to its full potential because rather than analysing how the performance of a team evolves during a game, the current framework is using summary metrics to encapsulate the whole performance. Time varying feature extraction would thus be a significant area of research that can potentially be explored as a next step. Such research could enable

[22]Game to game consistency is being used to define playing style, as illustrated in Equations 13-15.

even more nuanced analysis, as more detailed questions like those listed below can be answered:

1. How did a team's playing style differ from the first half to the second?
2. Did a team play differently once they were leading (or trailing)?
3. Did substituting player X for player Y change the team playing style?
4. How did a team react to a send-off (red card)?

While this would involve ground up feature re-engineering, such work can bring much detail to analysis and be used to break down historical performances (such as a champions league final game).

7 Conclusion

This project applied methodology from prior work to exploit soccer-logs data, generating human interpretable features characterising different aspects of the game, and used the derived metrics to cluster teams in vector space, making insights regarding their playing styles. Network science was employed to specifically study team passing across players and pitch zones, while direct log manipulation yielded metrics summarizing other aspects of team performances. Using these metrics, different clustering approaches were explored, and by the process of elimination, hierarchical clustering with Ward's method was chosen to arrive at four disjoint clusters such that each team was classified into one of the four clusters by playing style. The clusters were revealed to have significant performance discrepancy, and analysis was done to determine how high performing clusters differ from their lower performing counterparts. In short, it was found that teams in high performing clusters, on average, displayed higher possession, higher variability in passing across players/zones, more indirect passing, and a high defensive line compared to teams in the lower performing clusters. It was also demonstrated that playing styles are not simply determined by a team's skill level, and is indeed a reflection of team identity beyond skill/ability to dominate the league.

As such, this project's outcome is a categorization of 98 teams from five European domestic leagues (Spanish first division, Italian first division, English first division, German first division, and French first division) into four clusters representing their playing styles, along with a qualitative interpretation for the playing styles from a fundamental

perspective. This is an important contribution to the study of soccer analytics, as the idea of combining network science algorithms characterizing passing with unsupervised clustering to determine and analyse general team playing style is indeed novel. One potential use case for such research is for coaches and tacticians to direct their human expertise when preparing for upcoming games. For instance, if Schalke (a team from the German first divison) has an upcoming game against Liverpool (a team from the English first division) in the European Champions League, Shalke's coach could use this playing style classification to note that Liverpool is a cluster 3 team. While Schlke may not have direct prior experience against Liverpool, as it is a team from a different domestic league, the coach can note that Bayern Munich (also from the German first division) is also a cluster 3 team. Thus, the coach would be able to narrow his focus by studying footage of Schalke playing against Bayern Munich to make insights informing his tactics against Liverpool. As such, this research stands as a decision support tool to guide human expertise to a subset of otherwise overwhelming amount of data/game footage. Note that the project is not meant to replace human decision making; some intangibles and 'intuition' are crucial to inform tactics. Rather, this project is complementary to human decision making, and simply serves to narrow human expertise.

Beyond decision support, the research can also used by coaches to study their teams in greater detail in order to expose potential weaknesses. For instance, the Liverpool coach could note that Liverpool, being a cluster 3 team, is prone to play a very high defensive line. This leaves them very vulnerable to quick counterattacks. Noting this, the coach may choose to instruct his squad to remain more structured in their defensive position when playing against teams with fast attacking wingers.

Fundamentally, segmentation of soccer teams into distinct, interpretable playing styles provides a way to study teams in great depth, and also narrow human efforts, as shown by the two use cases above. The use of passing metrics from network science, along with other defensive statistics is new, and thus presents challenges and opportunities. As elaborated on in chapter 6, the main limitation with the current methodology is the correlated feature space. By deriving custom-defined clustering methods to accommodate for these correlations, the claims for differences across the derived playing styles would become stronger. Thus, there is potential for future theoretical research to arrive at such a clustering method. Revising the scope of the project in a ground-up manner to analyse intra-game playing style changes would also provide much granularity to answer more nuanced questions, as noted in chapter 6.

Even with its current limitations, the playing styles segmentation still represents a

valuable contribution to the field of soccer analytics. This thesis only managed to explain the methodology to derive team playing styles, and summarise findings from certain case studies highlighting differences between top tier teams and their lower tier counterparts, while also exploring certain anomalies. Thus, the project demonstrates the value of data driven analytics in understanding the game of soccer. Despite studies finding a strong role of randomness in soccer game outcomes [9], the simple features used in this project provided enough information to segregate (in general) top performers from others. This is a promising result, as there is reason to believe that more complex analysis/feature engineering may provide more nuanced interpretations of team playing styles.

Future work should continue to adopt this methodology beyond the 2017/2018 season, and perhaps even to international teams, to better understand team playing styles. Given more data over a longer term, it would be interesting to study the evolution of team playing styles over long time horizons and determine if playing styles significantly differ between clubs playing long format leagues and international teams participating in shorter, elimination style tournaments.

Appendices

A Correlations of Raw Features with Performance

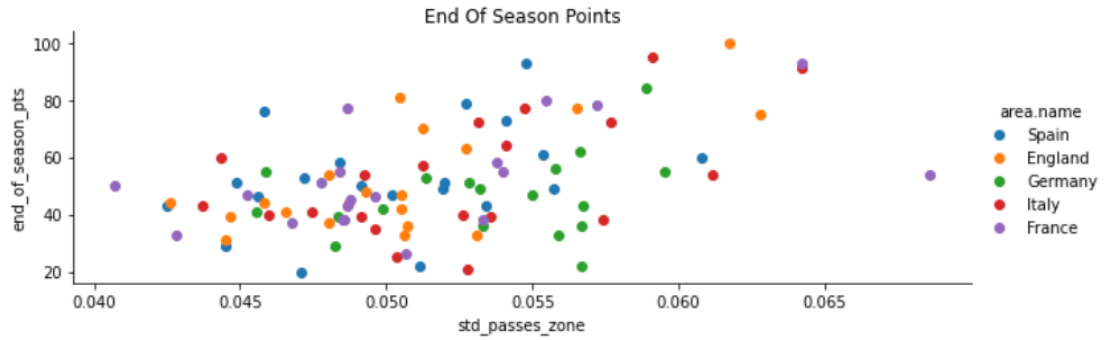


Figure 25: End of Season Points vs Macro-Averaged σ^t

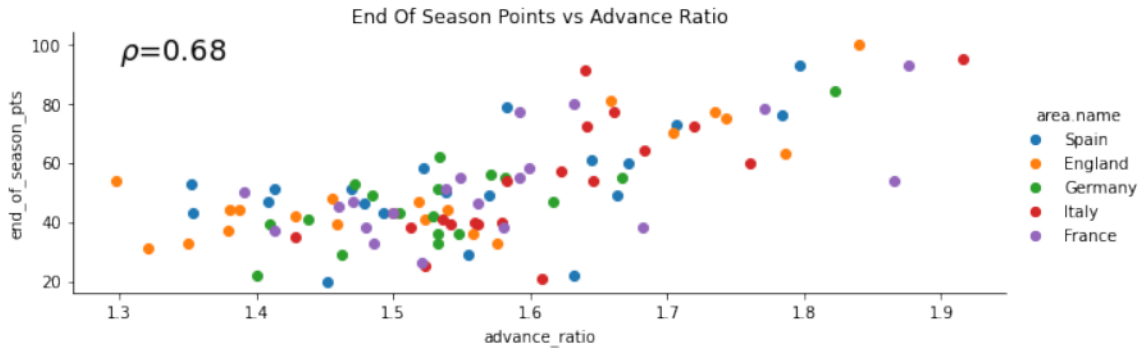
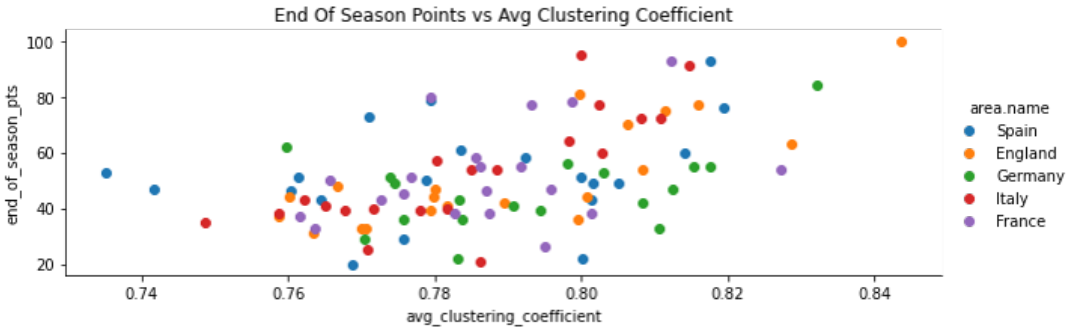
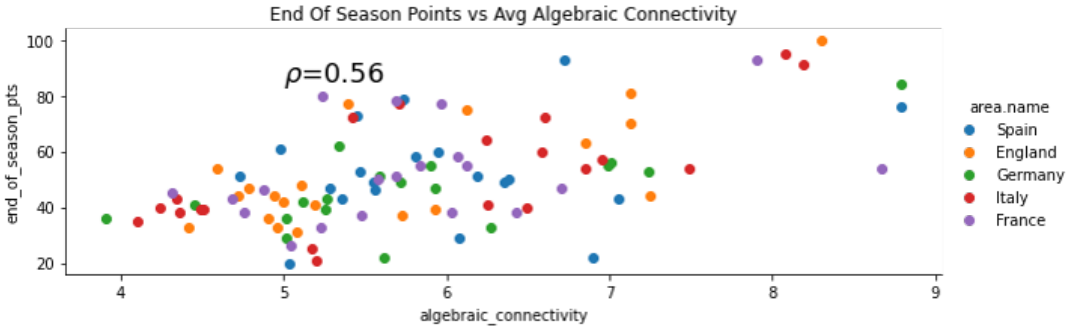


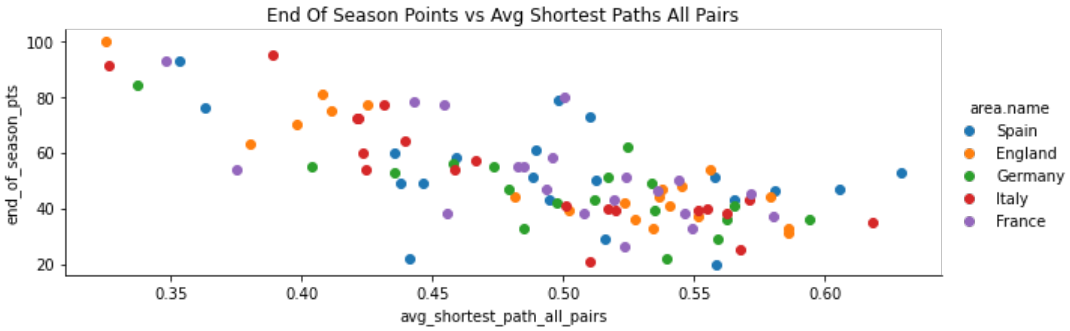
Figure 26: Macro-Averaged Mean Advance Ratio vs End Of Season Points



(a) Average Clustering Coefficient, $\rho = 0.52$



(b) Algebraic Connectivity, $\rho = 0.56$



(c) Average All Pairs Topological Shortest Paths, $\rho = -0.73$

Figure 27: Player Network Metrics vs End of Season Points

B Detailed Examples of Graph Theory Metrics

Clustering Coefficient

Intuitively, the clustering coefficient of node u count the number of triangles around node u as a ratio of the number of triangles that could have been (accounting for directionality). Equivalently, it measures how many of u's neighbours are interconnected with each other. Figure 28 below shows a contrast between a low clustering coefficient and a high clustering coefficient for the central node.

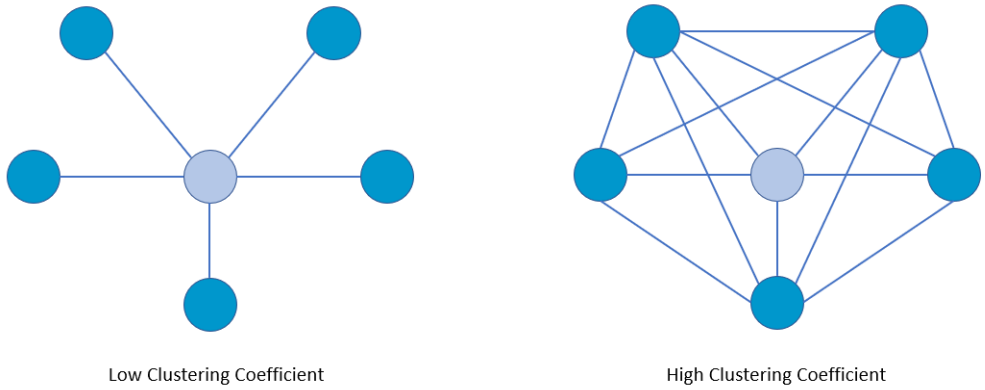
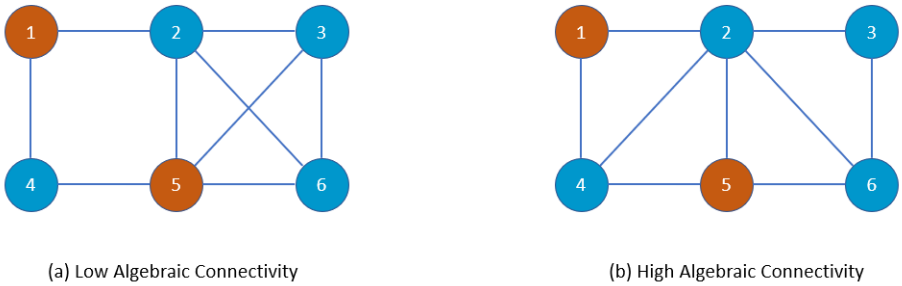


Figure 28: A Contrast Between a Low and High Local Clustering Coefficient

Algebraic Connectivity

Algebraic connectivity is a measure of a graph's integration or tolerance to failure. It is intuitively demonstrated by the graphs in Figures 29a and 29b.



(a) Low Algebraic Connectivity (b) High Algebraic Connectivity

Figure 29: A Contrast Between a Low and High Local Algebraic Connectivity

In Figure 29a, if nodes 1 and 5 (indicated in orange) are removed from the graph, the resulting graph would be completely separated as node 4 will be isolated from nodes 2, 3, and 6. In contrast, if the same nodes are removed from the graph in Figure 29b,

the resulting graph will still remain as a single connected component. The algebraic connectivity captures this intuition, and thus the algebraic connectivity for the graph in Figure 29b is higher than it is for the graph in Figure 29-a.

C List of Derived Metrics

Metric	Description
Passing Volume (w)	Number of passes (successful/unsuccessful) executed by a team in a given match.
μ^t_{zone}	Mean number of passes "handled" by each of the 9 zones
σ^t_{zone}	Standard deviation of passes "handled" across the 9 zones
$MeanAdvanceRatio^t$	The average ratio of the lateral trajectory to the vertical trajectory for all passes in a game
Passing Lane Cluster	A binary value indicating which cluster the passing lane profiles belong to according to Section 4.3
$\mathbf{w}$	A 36 dimensional vector measuring passing lane intensities
μ^t_{player}	Mean number of passes "handled" by each of the player
σ^t_{player}	Standard deviation of passes "handled" across the players on a team
Average Clustering Coefficient	A measure of a player network's local robustness
Algebraic Connectivity	A measure of a player's network's segregation and fault tolerance
Average All-Pairs Topological Shortest Path	A measure of how frequently the ball was passed between two players, on average
Number of Yellow Cards / Game	Aggregated per match, per team
Dangerous Balls Lost	Aggregated per match, per team
Interceptions Inside Box	Aggregated per match, per team
Total Interceptions	Aggregated per match, per team
Fraction of Slide Tackles in Box	Aggregated per match, per team
Slide Tackles Inside Box	Aggregated per match, per team
Total Slide Tackles	Aggregated per match, per team
Fraction of Slide Tackles in Box	Aggregated per match, per team
$\mathbf{X}$	X centroid of slide tackles and interceptions
$\mathbf{Y}$	Y centroid of slide tackles and interceptions

D Detailed Mutual Information & Cluster Centroid Plots

In the figures below, 'mean' and 'var' refer to the mean and variance of the 40 features for each team used for summarizing performance metrics to team metrics, as described in Equations 13-15. Furthermore, 'w' refers to an estimate of team possession approximated by the proportion of passes a team plays, as indicated in Appendix C.

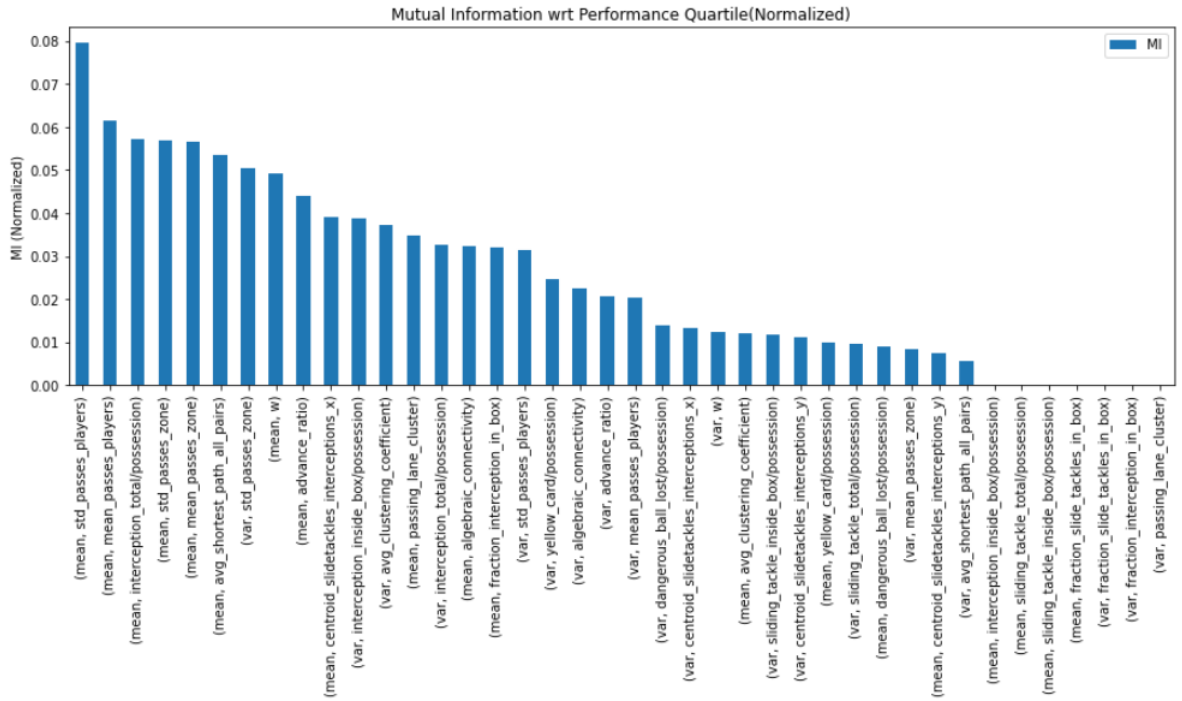


Figure 30: Normalized Mutual Information between Features & Performance Quartile

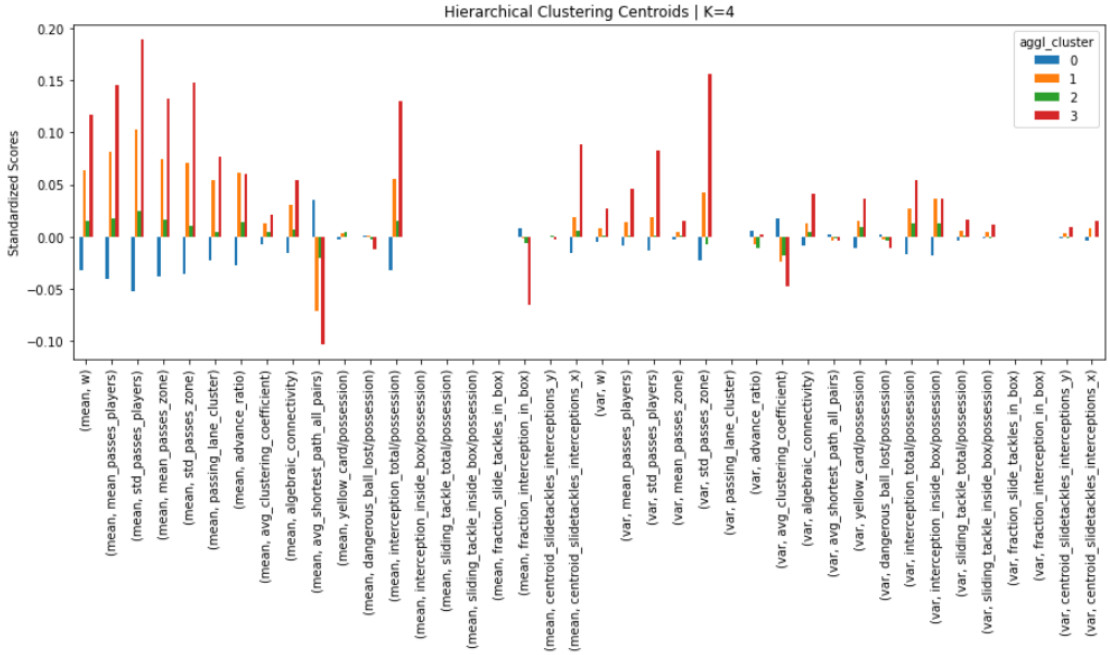


Figure 31: Centorids for the 4 Clusters Obtained

E A Breakdown of the 4 Clusters by Constituent Teams

League	Team	Cluster
England	AFC Bournemouth	0
England	Arsenal	1
England	Brighton & Hove Albion	0
England	Burnley	0
England	Chelsea	1
England	Crystal Palace	0
England	Everton	0
England	Huddersfield Town	0
England	Leicester City	0
England	Liverpool	3
England	Manchester City	3
England	Manchester United	1
England	Newcastle United	0
England	Southampton	2
England	Stoke City	0
England	Swansea City	0
England	Tottenham Hotspur	1
England	Watford	0
England	West Bromwich Albion	0
England	West Ham United	0
France	Amiens SC	0
France	Angers	0
France	Bordeaux	2
France	Caen	0
France	Dijon	0
France	France	2
France	Guingamp	0
France	Lille	2
France	Metz	0
France	Montpellier	0
France	Nantes	0
France	Nice	1
France	Olympique Lyonnais	1

France	Olympique Marseille	2
France	PSG	1
France	Rennes	0
France	Saint-Etienne	2
France	Strasbourg	0
France	Toulouse	0
France	Troyes	0
Germany	Augsburg	0
Germany	Bayer Leverkusen	2
Germany	Bayern Mucnchen	3
Germany	Borussia Dortmund	3
Germany	Borussia M'gladbach	2
Germany	Eintracht Frankfurt	0
Germany	Freiburg	0
Germany	Hamburger SV	0
Germany	Hannover 96	0
Germany	Hertha BSC	0
Germany	Hoffenheim	2
Germany	Koln	0
Germany	Mainz 05	0
Germany	RB Leipzig	2
Germany	Schalke 04	0
Germany	Stuttgart	0
Germany	Werder Bremen	0
Germany	Wolfsburg	2
Italy	Atalanta	2
Italy	Benevento	0
Italy	Bologna	0
Italy	Cagliari	0
Italy	Chievo	0
Italy	Crotone	0
Italy	Fiorentina	2
Italy	Genoa	0
Italy	Hellas Verona	0
Italy	Internazionale	1
Italy	Juventus	1
Italy	Lazio	2

Italy	Milan	1
Italy	Napoli	3
Italy	Roma	1
Italy	SPAL	0
Italy	Sampdoria	1
Italy	Sassuolo	0
Italy	Torino	0
Italy	Udinese	0
Spain	Athletic Club	0
Spain	Atletico Madrid	2
Spain	Barcelona	1
Spain	Celta de Vigo	2
Spain	Deportivo Alaves	0
Spain	Deportivo La Coruna	0
Spain	Eibar	2
Spain	Espanyol	0
Spain	Getafe	0
Spain	Girona	0
Spain	Las Palmas	2
Spain	Legan\u00e9s	0
Spain	Levante	0
Spain	Malaga	0
Spain	Real Betis	1
Spain	Real Madrid	1
Spain	Real Sociedad	2
Spain	Sevilla	2
Spain	Valencia	2
Spain	Villarreal	2

F Detailed Comparison of Features in the Centroids of Cluster 1+, Cluster 1-, and Cluster 3

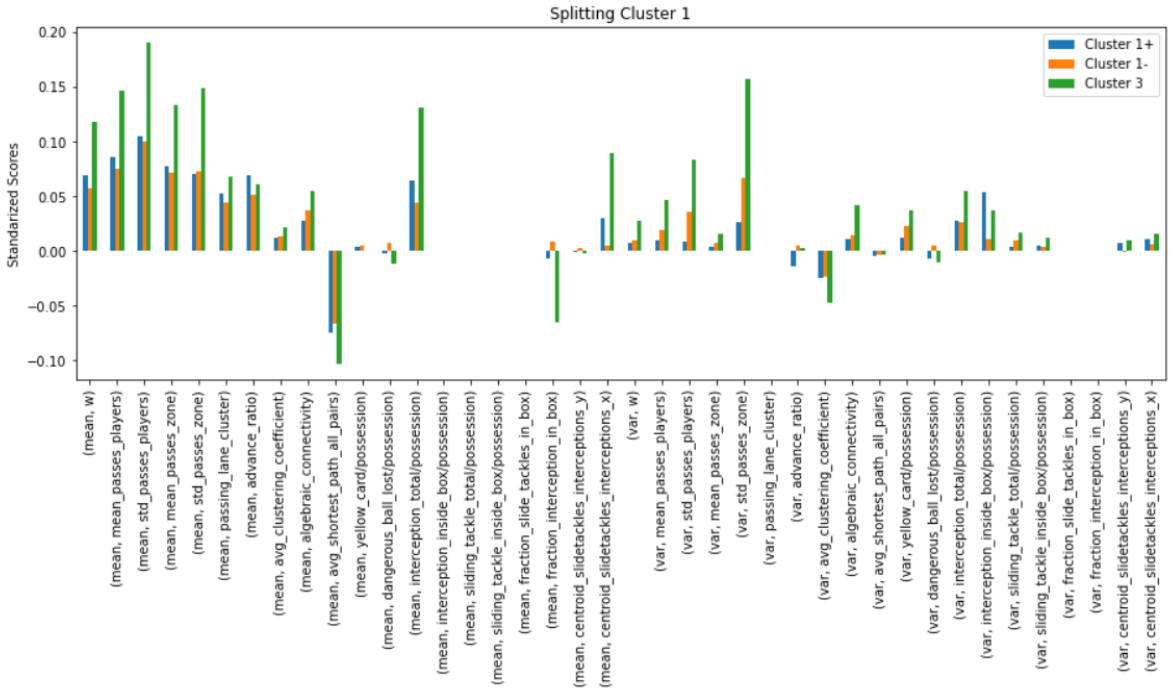


Figure 32: Splitting Cluster 1 - Isolating Top-5 Teams

G Code & Implementation Details for Player Networks

Here, the key scripts for producing player network graphs is available. As a reminder, a key assumption was made that if a certain player i passes the ball and the next chronological event is from another player j on the same team, then it is assumed that the pass was from i to j. This assumption can be found in the code below with the corresponding comment. For complete code for the project with annotated analyses, please refer to https://github.com/arifmoh2/thesis_soccer.

Player Networks

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
import networkx as nx

matches = w.index.get_level_values(0).unique(); n=len(matches)
network_stats = dict()

for i, match in enumerate(matches):
    if i % 50 == 0: print('{}/{}'.format(i, n))

    matchId = match

    #Get the time for the latest event in the first half for this match
    end_first_half = events.loc[(events['matchId']==matchId)
                                &(events['matchPeriod']=='1H'),
                                ↪  ['eventSec']].max().values[0]


    df = events_passing.loc[(events_passing['matchId']==matchId)]
    df.loc[:, 'time'] = (df['matchPeriod']=='2H')*end_first_half +
    ↪  df['eventSec']
    df = df[['playerId', 'positions', 'teamId', 'time']]
```

```
team1, team2 = df['teamId'].unique()
profile_team1 = df.loc[df['teamId'] == team1, ['playerId',
↪  'time']].sort_values(by='time').reset_index(drop=True)
profile_team2 = df.loc[df['teamId'] == team2, ['playerId',
↪  'time']].sort_values(by='time').reset_index(drop=True)

#Saving Stats as dict of dicts
network_stats[match] = dict()
network_stats[match][team1], network_stats[match][team2] = dict(),
↪  dict()


#Filter to players that have played at least 50 minutes of football
↪  to reduce outlier effects
player_time = (profile_team1.groupby('playerId').max() -
↪  profile_team1.groupby('playerId').min())
#relevant_players = player_time.loc[player_time['time']>3000].index
#profile_team1 =
↪  profile_team1.loc[profile_team1['playerId'].isin(relevant_players)]

player_time = (profile_team2.groupby('playerId').max() -
↪  profile_team2.groupby('playerId').min())
#relevant_players = player_time.loc[player_time['time']>3000].index
#profile_team2 =
↪  profile_team2.loc[profile_team2['playerId'].isin(relevant_players)]



# Naive assumption: Receiver is the player who made the next pass
↪  in this team - doesn't account for interceptions
# We will make a graph with edge (i, j) value = Number of passes
↪  from player i to player j
profile_team1['sender'] = profile_team1['playerId']
profile_team1['receiver'] =
↪  profile_team1['playerId'].shift(-1).dropna()
profile_team1 = profile_team1[['sender', 'receiver']]
```

```
profile_team1 = profile_team1.loc[profile_team1['sender'] !=
↪  profile_team1['receiver']]
profile_team1 =
↪  profile_team1.groupby(['sender','receiver']).size().to_frame('weight').reset_

profile_team2['sender'] = profile_team2['playerId']
profile_team2['receiver'] =
↪  profile_team2['playerId'].shift(-1).dropna()
profile_team2 = profile_team2[['sender', 'receiver']]
profile_team2 = profile_team2.loc[profile_team2['sender'] !=
↪  profile_team2['receiver']]
profile_team2 =
↪  profile_team2.groupby(['sender','receiver']).size().to_frame('weight').reset_

g = nx.DiGraph()
g.add_edges_from([(row['sender'], row['receiver'],
↪  {'weight':row['weight']}) for i, row in
↪  profile_team1.iterrows()])
df = pd.Series(dict(g.degree(weight='weight')),
↪  name='number_passes_invloved')
network_stats[match][team1]['avg_clustering_coefficient'] =
↪  pd.Series(nx.clustering(g)).mean()
network_stats[match][team1]['algebraic_connectivity'] =
↪  nx.algebraic_connectivity(g.to_undirected())
network_stats[match][team1]['avg_shortest_path_all_pairs'] =
↪  topological_all_pairs_shortest_paths(g)
#network_stats[match][team1]['mean_passes_players'] = df.mean()
#network_stats[match][team1]['std_passes_players'] = df.std()

g = nx.DiGraph()
g.add_edges_from([(row['sender'], row['receiver'],
↪  {'weight':row['weight']}) for i, row in
↪  profile_team2.iterrows()])
df = pd.Series(dict(g.degree(weight='weight')),
↪  name='number_passes_invloved')
network_stats[match][team2]['avg_clustering_coefficient'] =
↪  pd.Series(nx.clustering(g)).mean()
```

```
    network_stats[match][team2]['algebraic_connectivity'] =
    ↪ nx.algebraic_connectivity(g.to_undirected())
    network_stats[match][team2]['avg_shortest_path_all_pairs'] =
    ↪ topological_all_pairs_shortest_paths(g)
    #network_stats[match][team2]['mean_passes_players'] = df.mean()
    #network_stats[match][team2]['std_passes_players'] = df.std()



network_stats = pd.DataFrame.from_dict({(i,j): network_stats[i][j]
                            for i in network_stats.keys()
                            for j in network_stats[i].keys()},
                        orient='index')
```

References

[1] FIFA.com, *Who we are - news - fifa survey: Approximately 250 million footballers worldwide*. [Online]. Available: https://www.fifa.com/who-we-are/news/fifa-survey-approximately-250-million-footballers-worldwide-88048.

[2] Hermesauto, *Football: 2018 world cup watched by record 3.5 billion people, says fifa*, Dec. 2018. [Online]. Available: https://www.straitstimes.com/sport/football/football-2018-world-cup-watched-by-record-35-billion-people-fifa.

[3] C. Reep and B. Benjamin, "Skill and chance in association football," *Journal of the Royal Statistical Society. Series A (General)*, vol. 131, no. 4, pp. 581–585, 1968, ISSN: 00359238. [Online]. Available: http://www.jstor.org/stable/2343726.

[4] *Obituary: Stan cullis*, Mar. 2001. [Online]. Available: https://www.theguardian.com/news/2001/mar/01/guardianobituaries.football.

[5] J. Prince-Wright, *How opta altered the premier league, and soccer, forever - prosoccertalk: Nbc sports*, Oct. 2013. [Online]. Available: https://soccer.nbcsports.com/2013/10/02/how-opta-has-helped-alter-the-premier-league-and-soccer-forever-part-i/.

[6] J. Williams, *Liverpool are using incredible data science, and match effects are extraordinary*, Jan. 2020. [Online]. Available: https://www.liverpool.com/liverpool-fc-news/features/liverpool-transfer-news-jurgen-klopp-17569689.

[7] *Advanced metrics*. [Online]. Available: https://www.optasports.com/services/analytics/advanced-metrics/.

[8] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Scientific Data*, vol. 6, no. 1, 2019. DOI: 10.1038/s41597-019-0247-7.

[9] M. Lames and T. McGarry, "On the search for reliable performance indicators in game sports," *International Journal of Performance Analysis in Sport*, vol. 7, no. 1, pp. 62–79, 2007. DOI: 10.1080/24748668.2007.11868388. eprint: https://doi.org/10.1080/24748668.2007.11868388. [Online]. Available: https://doi.org/10.1080/24748668.2007.11868388.

[10] L. Pappalardo, P. Cintia, P. Ferragina, E. Massucco, D. Pedreschi, and F. Giannotti, "Playerank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 5, Sep. 2019, ISSN: 2157-6904. DOI: 10.1145/3343172. [Online]. Available: https://doi.org/10.1145/3343172.

[11] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi, and M. Malvaldi, "The harsh rule of the goals: Data-driven performance indicators for football teams," in *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1–10. DOI: 10.1109/DSAA.2015.7344823.

[12] P. Gould and A. Gatrell, "A structural analysis of a game: The liverpool v manchester united cup final of 1977," *Social Networks*, vol. 2, no. 3, pp. 253–273, 1979. DOI: 10.1016/0378-8733(79)90017-0.

[13] J. Duch, J. S. Waitzman, and L. A. N. Amaral, "Quantifying the performance of individual players in a team activity," *PLoS ONE*, vol. 5, no. 6, 2010. DOI: 10.1371/journal.pone.0010937.

[14] F. M. Clemente, F. M. L. Martins, D. Kalamaras, P. D. Wong, and R. S. Mendes, "General network analysis of national soccer teams in fifa world cup 2014," *International Journal of Performance Analysis in Sport*, vol. 15, no. 1, pp. 80–96, 2015. DOI: 10.1080/24748668.2015.11868778.

[15] J. Buldú, J. Busquets, I. Echegoyen, and F. Seirul Lo, "Defining a historic football team: Using network science to analyze guardiola's f.c. barcelona," *Scientific reports*, vol. 9, no. 1, p. 13 602, Sep. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-49969-2. [Online]. Available: https://europepmc.org/articles/PMC6753100.

[16] S. A. Pettersen, D. Johansen, H. Johansen, V. Berg-Johansen, V. R. Gaddam, A. Mortensen, R. Langseth, C. Griwodz, H. K. Stensland, and P. Halvorsen, "Soccer video and player position dataset," in *Proceedings of the 5th ACM Multimedia Systems Conference*, ser. MMSys '14, Singapore, Singapore: Association for Computing Machinery, 2014, pp. 18–23, ISBN: 9781450327053. DOI: 10.1145/2557642.2563677. [Online]. Available: https://doi.org/10.1145/2557642.2563677.

[17] T. Stolen, K. Chamari, C. Castagna, and U. Wisl??ff, "Physiology of soccer," *Sports Medicine*, vol. 35, no. 6, pp. 501–536, 2005. DOI: 10.2165/00007256-200535060-00004.

[18] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Identifying team style in soccer using formations learned from spatiotemporal tracking data," in *2014 IEEE international conference on data mining workshop*, IEEE, 2014, pp. 9–14.

[19] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, 2003, Biometrics, ISSN: 0031-3203. DOI: https://doi.org/10.1016/S0031-3203(02)00060-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320302000602.

[20] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, e87357, 2014.

[21] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

Page intentionally left blank