17-18 EPL Teams Attack Sequence Pattern

Junseok Yang
Department of Statistics
University of Illinois Urbana-Champaign
jyang247@illinois.edu

Abstract

This project investigates whether English Premier League (EPL) teams from the 2017–18 season exhibit distinctive attacking sequence patterns, with a particular focus on sequences leading to shot opportunities. To capture both spatial and contextual aspects of play, the data were transformed into distance matrices using the Fréchet distance for spatial trajectories and Gower distance for categorical and numerical features. These representations were combined to model a latent variable termed "attack pattern", defined through unsupervised clustering via K-Medoids. A statistical test of independence using the Chi-squared test was conducted to assess the association between teams and their corresponding attack patterns. Lastly, further analysis was conducted focusing on Big 6 clubs and Leicester City, which stand out among other EPL teams for generating the highest number of goals, potentially suggesting a distinct tactical identity in their scoring patterns.

1 Introduction

Football (or soccer) is the most widely played and watched sport in the world, with 211 countries affiliated with the global organization FIFA. Among the various leagues around the world, the European football leagues are often considered the most competitive, both in terms of quality and popularity. According to Opta's global rankings, the English Premier League (EPL) consistently ranks as one of the strongest leagues in world football, reflecting its global fanbase, market size, and on-field performance (Analyst, 2024).

One of the reasons for the EPL's enduring popularity is its dynamic and often unpredictable match outcomes. Matches where underdog teams defeat top-ranked opponents are not uncommon, contributing to the league's excitement and competitiveness. While many factors can influence such outcomes—such as player form, injuries, and refereeing decisions—this study focuses on tactical differences, particularly attacking strategies, as a major contributing factor. In modern football, innovative tactics such as the use of inverted fullbacks, positional play involving triangular and square passing structures, goalkeeper participation in buildup phases, and defensive tactics like extreme low blocks (often referred to as "parking the bus") have reshaped the way teams construct goal-scoring opportunities (Carling et al., 2005).

This project aims to analyze these strategic dimensions quantitatively by focusing on "attack sequences"—defined as two previous sequences of passes or events that connect to a shot attempt. By studying the structure of these sequences across all teams in the 2017–18 EPL season, the project seeks to identify whether teams exhibit distinctive attacking patterns. Leveraging the open-access event-based soccer dataset curated by Pappalardo et al. (Pappalardo et al., 2019), this analysis combines spatial and categorical information to define a latent variable termed *attack pattern*, which is then explored through unsupervised clustering and statistical testing.

2 Data & Pre-processing

2.1 Data

The dataset used in this project is derived from the open-access event-based soccer data introduced by Pappalardo et al. (Pappalardo et al., 2019), originally sourced from WyScout, a leading sports data provider. This dataset contains spatio-temporal event logs from several top-tier football leagues and international competitions. For the purpose of this project, we focus exclusively on matches from the 2017-18 season of the English Premier League (EPL), extracting all relevant event-level data including timestamps, event types, player/team information, and spatial coordinates. Additional details

about the data structure and collection methodology are available in the dataset's official record, (Section 9, see here) (Pappalardo et al., 2019).

2.2 Attack Sequence

To answer our central research question—**Do EPL teams exhibit their own unique attack sequence patterns?**—we reconstruct the data so that each unit of observation corresponds to an **attack sequence**. Specifically, we define an attack sequence as the two events immediately preceding any shot classified as a "goal-scoring opportunity" (as tagged in the dataset), together with the shot itself. This results in a sequence of three temporally ordered events, which we compress into a single observation by combining features such as spatial coordinates, event type, and time.

The motivation behind this construction is that while goals and assists are typically emphasized in performance analysis, the events leading up to these final actions are often undervalued. In particular, long or incisive passes that initiate promising movements can be tactically decisive. For instance, consider the following stylized sequence:

Long pass (20-30m) to penalty box (Defender)
→ Short assist pass (Striker)
→ Shot / Goal (Midfielder)

Figure 1: Example of a stylized three-event attack sequence

In this example, it could be argued that the initiating pass from the defender plays a crucial role in the final shot or assist. While soccer is inherently continuous and fluid, isolating these three-event windows allows us to extract interpretable and comparable attack patterns across teams. These sequences serve as the foundation for downstream modeling and clustering.

Each processed observation encodes this three-event sequence using a mixture of spatial, temporal, categorical, and contextual features, which are later embedded into a distance space for clustering analysis.

Adding New Features: Progression and Time Duration

In addition to extracting spatial and event-type information for each attack sequence, we engineered two features that capture important tactical dynamics: *progression* and *time duration* between events.

Progression Distance and Ratio Advancing the ball toward the opponent's goal is a key factor in increasing the likelihood of creating high-quality chances. For this reason, we introduce two spatial metrics: **progression distance** and **progression ratio**.

The progression distance measures the horizontal advancement between consecutive events, computed as the difference in x-coordinates:

$$\text{Progression Distance} = x_{i+1} - x_i$$

This value can be negative, indicating a backpass or a ball played away from the opponent goal side. To normalize progression with respect to overall movement, we compute the **progression ratio**, defined as:

$$\text{Progression Ratio} = \frac{x_{i+1} - x_i}{\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}}$$

This ratio captures the extent to which a pass contributes to forward progression relative to its total travel distance.

Time Duration We also calculate the time elapsed between consecutive events in a sequence. Rapid transitions are a hallmark of effective attacking teams, allowing them to catch defenses off guard. By incorporating **event duration** features, we can explore whether quicker sequences correlate with more threatening attacks. This reflects recent observations in football match analysis suggesting that shorter time intervals between ball movements can increase offensive effectiveness (Sarmento et al., 2018).

These engineered features, combined with the spatial and categorical characteristics of each event, form a rich representation of attacking behavior that is used in the subsequent clustering and statistical analysis.

2.3 Distance Matrix

To apply clustering algorithms to sequences of football events in later analysis, it is necessary to convert each observation into a pairwise distance representation. This transformation allows us to capture the similarity or dissimilarity between attack sequences and group them based on underlying patterns. Because the attack sequences in this project

are composed of both spatial trajectories and structured features (e.g., event types, time durations, and outcome labels), we employ two complementary distance metrics: Fréchet distance for spatial similarity and Gower distance for mixed-feature similarity.

Fréchet Distance The Fréchet distance is a well-established metric for comparing curves or trajectories, taking into account the location and ordering of points along the paths. Unlike simpler pointwise distances, the Fréchet distance considers the overall shape and directionality of movement, making it particularly well-suited for comparing spatial sequences such as ball progressions in football. Intuitively, it can be described as the minimal leash length needed to connect two entities (e.g., a dog and its owner) walking along two separate curves without backtracking (Alt and Godau, 1995).

Gower Distance To compare structured, non-spatial features that include a mixture of numerical and categorical variables, we use the Gower distance (Gower, 1971). This metric computes pairwise dissimilarities by normalizing numerical variables (e.g., progression distance, time duration) and treating categorical variables (e.g., event type, presence of opponent involved, shot accuracy) using a simple matching criterion. The Gower distance is widely used in applications that require clustering over heterogeneous data types.

Combining Distances To integrate both spatial and structured similarities into a unified framework, we normalize each distance matrix with MinMax scaling and combine them with equal weighting:

$$D_{\text{final}} = w_{\text{Fréchet}} \cdot D_{\text{Normalized Fréchet}} + w_{\text{Gower}} \cdot D_{\text{Normalized Gower}}$$

In this project, we assign equal importance to both components:

$$D_{\text{final}} = 0.5 \cdot D_{\text{Normalized Fréchet}} + 0.5 \cdot D_{\text{Normalized Gower}}$$

This final distance matrix D_{final} serves as the basis for the clustering analysis in the next stage of the project.

3 Clustering

3.1 K-Medoids

Clustering is a form of unsupervised learning, which aims to uncover hidden structure in unlabeled data by grouping similar observations together. Unlike supervised learning, where labels or outcomes guide the algorithm, unsupervised methods learn patterns purely from the input features (Hastie et al., 2009).

In this project, we apply clustering to identify a latent variable "attack patterns" among EPL teams based on their shot event sequences. Since there is no predefined label for what constitutes a tactical style or attacking identity, unsupervised clustering provides a natural framework for detecting recurring patterns without supervision. The resulting clusters serve as candidate representations of different strategic approaches to chance creation.

To perform clustering on our custom distance matrix discussed in the previous section, we use the K-Medoids algorithm. K-Medoids is a partitional clustering method that, like K-Means, assigns each observation to the nearest cluster center. However, unlike K-Means, which minimizes squared Euclidean distances around a mean, K-Medoids minimizes the sum of dissimilarities around a representative observation called the "medoid". This makes K-Medoids particularly robust to noise and outliers and allows it to work with arbitrary dissimilarity measures, including precomputed distance matrices such as those used in this study (Kaufman and Rousseeuw, 2009).

The number of clusters is a hyperparameter that we assess using model selection criteria, discussed in the next section.

3.2 Evaluation & Optimal K Selection

Selecting the optimal number of clusters is a fundamental step in clustering analysis. Since clustering is an unsupervised task, we rely on internal validation metrics that assess the quality of the resulting cluster assignments based on compactness and separation (Kodinariya and Makwana, 2013). Two commonly used tools for this purpose are the Elbow Method and the Silhouette Score.

Elbow Method The Elbow Method evaluates the total within-cluster dissimilarity (often referred to as "cost" or "inertia") across various values of K. As the number of clusters increases, the inertia naturally decreases. However, a sharp change in the

rate of decrease—forming an "elbow"—suggests a point where adding more clusters yields diminishing returns. This inflection point is interpreted as a candidate for the optimal K (Kodinariya and Makwana, 2013).

Silhouette Score The Silhouette Score quantifies how similar an observation is to its own cluster compared to other clusters. It ranges from -1 to 1, with higher values indicating more coherent and well-separated clusters. The average silhouette score across all data points provides a measure of clustering quality, with peaks suggesting better-defined clustering structures (Kaufman and Rousseeuw, 2009).

We evaluated the clustering results for $K \in [2, 20]$ using both the elbow method and average silhouette scores.

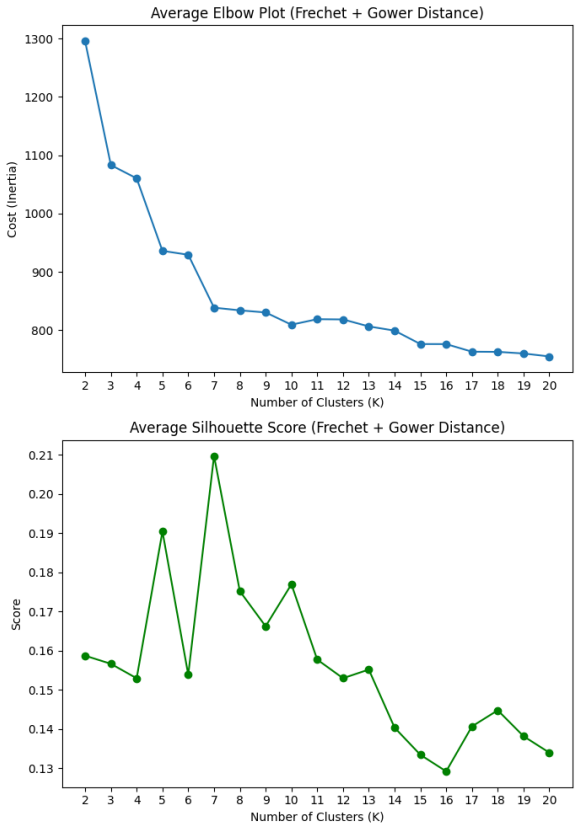


Figure 2: Average Elbow & Silhouette Score Plot

From the figure 2, we observed a prominent "elbow" at $K = 7$, where the inertia curve levels off following a steep drop from $K = 6$. Simultaneously, the silhouette score shows a sharp increase at $K = 7$, reaching its highest value (approximately 0.21), before declining significantly at $K = 8$. This alignment of both metrics supports $K = 7$ as the primary candidate for capturing stable and meaningful cluster structures in the customized distance matrix.

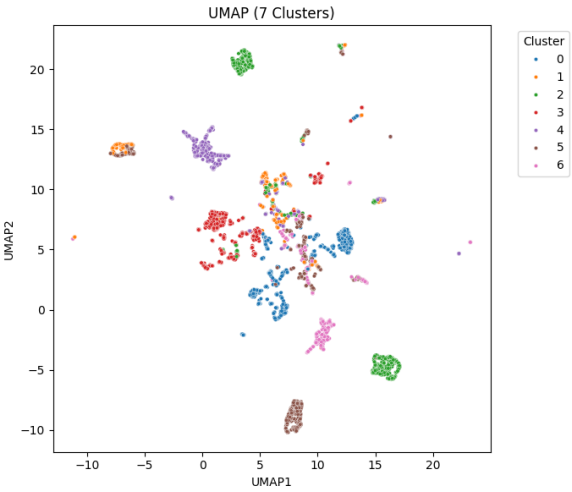


Figure 3: UMAP of K-Medoids (K = 7)

UMAP To visually assess the clustering structure in a lower-dimensional space, we applied Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique that preserves both local and global structure in high-dimensional data (McInnes et al., 2018). While some central regions in the 2D projection show overlap among cluster labels likely due to the compressed nature of the space, we observed that several well-separated clusters (blobs) were consistently and exclusively assigned to a single cluster. This provides qualitative support that the clustering algorithm, particularly at $K = 7$, captured relatively meaningful patterns and maintained coherence within these subgroups.

4 EPL Team and Attack Pattern

4.1 Translating Research Question into Mathematical Expression

After defining the latent feature "attack pattern" via clustering, we return to the original research question: **Do EPL teams exhibit their own unique attack sequence patterns?** To examine this question statistically, we define two categorical variables.

Assume A = Attack sequence pattern (cluster label) and T = EPL team. We then translate the research question into a probabilistic statement:

$$P(A = a \mid T = t) = P(A = a) \quad \text{for all } a, t?$$

This formulation asks whether the distribution of attack patterns is independent of the team iden-

tity—that is, whether the likelihood of observing a specific attack pattern remains the same regardless of which team generated it.

4.2 Mathematical Expression to Statistical Test

The mathematical question above can be interpreted as a test for statistical independence between two categorical variables: EPL team and attack pattern. To assess this, we first construct a contingency table showing the frequency distribution of cluster assignments across teams. Then, we apply two widely used tests for independence: the **Chi-squared test of independence** and **Fisher's exact test** (Agresti, 2002; Kaufman and Rousseeuw, 2009).

H_0 : There is no association between EPL team and attack pattern (Independent).

H_A : There is an association between EPL team and attack pattern.

The results from both tests indicate strong evidence against the null hypothesis. The Chi-squared test yielded a p-value of 3.10×10^{-7}, while the Fisher's exact test produced a p-value of 0.0001. Although Fisher's exact test is known to be conservative for large contingency tables, both results support the conclusion that there is a statistically significant association between EPL teams and their attack patterns when using a conventional significance level of $\alpha = 0.05$

This suggests that the latent cluster assignments derived from the shot sequence data meaningfully differentiate teams, providing statistical support for the presence of unique attacking styles in the EPL (See **Appendix 1** for more detail).

5 Case Study 1: Big 6 Attack Sequence

In the English Premier League (EPL), the term "Big 6" traditionally refers to six historically dominant clubs: **Arsenal**, **Chelsea**, **Liverpool**, **Manchester City**, **Manchester United**, and **Tottenham Hotspur**. All six clubs exhibited remarkably similar attack sequence clustering distributions. Specifically, **Cluster 0** was the most common, followed by **Cluster 3** across all teams (**Appendix 2**). This suggests some tactical convergence in how top teams generate opportunity shots.

Cluster 0 sequences were characterized by relatively linear trajectories, often beginning near the mid-left side of the pitch (from the attacking team's perspective) and ending just outside the penalty area. These sequences frequently featured two consecutive simple passes (e.g., **Simple pass - Simple pass**) and ended with an **accurate shot**. This structure appears to reflect structured buildup play leading to high-quality chances (**Appendix 2**).

In contrast, **Cluster 3** sequences displayed a flattened "hat-shaped" trajectory closer to the left touchline. All final shots in this cluster were inaccurate, with a 0% goal conversion rate, suggesting a lower-quality outcome. Notably, the most frequent transition within this cluster was also **Simple pass - Simple pass** (**Appendix 3**).

5.1 Team-Specific Attack Patterns and Performance

Although these clubs share similar dominant cluster types, their specific attack patterns and success metrics vary considerably (**Appendix 3**):

Arsenal Arsenal frequently relied on the **Simple pass – Smart pass** combination, which accounted for a goal conversion rate of **28%** and shot accuracy of **54%**. Another prominent pattern was **Touch – Simple pass**, which proved more effective, resulting in a goal conversion rate of **43%** and shot accuracy of **71%**.

Chelsea Chelsea's most common sequence was **Simple pass – High pass**, associated with a goal conversion rate of **29%** and shot accuracy of **71%**. However, other patterns attempted by Chelsea showed relatively low or even zero goal conversion, suggesting a more limited set of efficient build-up options.

Liverpool Liverpool frequently employed two successful combinations: **Simple pass – Cross** and **Simple pass – Smart pass**. The former yielded a goal conversion rate of **25%** and shot accuracy of **46%**, while the latter performed better with a conversion rate of **40%** and accuracy of **67%**. Overall, Liverpool demonstrated above-average efficiency across these patterns.

Manchester City Manchester City displayed the most diverse and balanced attack pattern portfolio among the Big 6. Three key combinations stood out: **Simple pass – Cross** (conversion rate: **46%**, accuracy: **69%**), **Simple pass – High pass** (conversion rate: **42%**, accuracy: **67%**), and **Smart pass**

– Simple pass (conversion rate: **57%**, accuracy: **71%**). These figures reflect Manchester City's ability to vary their approach while maintaining high attacking efficiency.

Manchester United In contrast, Manchester United demonstrated the least variety in attack sequence types. Their dominant pattern was **Simple pass – Smart pass**, which resulted in a goal conversion rate of **29%** and shot accuracy of **71%**. Despite the limited range, this pattern was executed with moderate effectiveness.

Tottenham Hotspur Tottenham Hotspur's attack sequences suggest a reliance on physical advantages such as speed and aerial ability. The pattern **Simple pass – Cross** had a goal conversion rate of **33%** but relatively low shot accuracy at **38%**, which headers might be done in this sequence. In contrast, the **Acceleration – Smart pass** sequence achieved a high conversion rate of **50%** and shot accuracy of **75%**, highlighting its strategic value in Tottenham's approach.

Overall, while the Big 6 clubs share common structural patterns in their most frequent attack sequences, variations in pass types, trajectory shapes, and shot efficiency provide insight into their unique tactical identities.

5.2 Progression & Event Duration

To better understand how the Big 6 clubs build up their attacks, we compared their average progression distances, progression ratios, and event durations across Clusters 0 and 3. These numerical features can potentially offer insight into the directness and tempo of the shot-creating sequences (**Appendix 4** and **5**).

Overall, both clusters demonstrated comparable averages and standard deviations in progression and timing metrics. This similarity aligns with the dominant pattern identified in both clusters—namely, the frequent use of **Simple pass – Simple pass**, a combination often associated with controlled buildup and moderate tempo.

In Cluster 0, Tottenham Hotspur, Liverpool, and Manchester City exhibited the highest horizontal progression between the first and second events, with average distances close to or exceeding **10 meters** and progression ratios above **40%**. These figures suggest a strong emphasis on forward movement early in the sequence. In contrast, Arsenal, Chelsea, and Manchester United showed more modest progression during this phase. Interestingly, while Liverpool and Chelsea maintained a high level of progression in the final link between the second event and the shot (with distances around **11 meters** and ratios near **40%**), Manchester City's progression in this phase dropped considerably, with an average distance of only **6.75 meters** and a ratio of **24%**. This may suggest that City tended to penetrate deeper earlier in the sequence and relied on shorter passes to set up shots closer to goal.

Cluster 3 exhibited generally lower progression values across all teams. Average progression distances and ratios between the first and second events were lower than those in Cluster 0, with the exception of Manchester United, Manchester City, and Arsenal, who maintained relatively high movement in this segment (around **9 meters**). Chelsea and Liverpool had the shortest progressions during this phase. Between the second event and the final shot, Liverpool and Tottenham stood out with longer progression distances (over **10 meters**) and moderately higher ratios. This may reflect the influence of crosses as second sub-events in these teams' sequences—a pattern consistent with their tactical reliance on wide play or physical aerial presence.

A key difference between the clusters also lies in **event duration**. Sequences in Cluster 3 generally took more time to unfold. Average durations between the first and second events ranged from **2.5 to 2.6 seconds** in Cluster 3, compared to **2.1 to 2.2 seconds** in Cluster 0. This temporal difference suggests that Cluster 0 sequences were executed at a faster tempo—possibly indicating more rehearsed or high-tempo attacking patterns.

6 Case Study 2: What Makes Leicester City Different?

Although Leicester City finished 9th in the 2017–18 EPL table, their offensive output stood out among all non-Big 6 teams. As shown in the league summary table (not included here), Leicester scored a total of **56 goals**, the most among all teams outside the traditional Big 6. By contrast, most other mid-table teams—such as Crystal Palace, Everton, and AFC Bournemouth—scored between **35 and 45 goals**.

What makes this even more notable is Leicester City's dominant clustering label. Unlike other teams whose attack sequences were most frequently labeled under generic or less effective clusters (e.g., Cluster 3), Leicester City had a majority

of their sequences labeled as **Cluster 5**. This suggests a fundamentally different type of attacking pattern that may explain their relatively high goal output.

6.1 Leicester City's Cluster 5 Sequences

In this section, we introduce a comparison group of three non-Big 6 teams that finished the season with relatively strong mid-table rankings—**Everton (7th, 44 goals)**, **Crystal Palace (11th, 45 goals)**, and **AFC Bournemouth (12th, 45 goals)**. These teams shared two key characteristics: their most frequent attack sequences were labeled as **Cluster 3**, and they each recorded approximately 45 goals over the season.

From the spatial trajectory visualization comparing Leicester City and the comparison group, we observe similarly structured attacking patterns—most notably, a prominent V-shaped buildup sequence (**Appendix 4**). However, when we examine the corresponding sub-event summaries and performance metrics in **Table 6**, clear differences emerge in terms of execution quality and outcomes.

Leicester City's attack sequences labeled as Cluster 5 reflect a simple but highly effective approach. Most sequences begin with **winning a ground duel in a defensive zone**, followed either by a second duel in a more advanced area or a direct cross into the final third. While structurally straightforward, these patterns were executed with notable precision: their goal conversion rates consistently exceeded **35%**, and shot accuracy remained high—up to **100%** in some cases. This combination of clarity and efficiency underlines a key tactical strength: the ability to convert fast, direct transitions into high-quality chances.

In contrast, the comparison group—whose dominant attack sequences were categorized as Cluster 3—produced strikingly ineffective results. As previously observed in the Big 6 analysis, Cluster 3 sequences were characterized by **0% goal conversion and shot accuracy**. Despite the spatial similarities to Leicester's sequences, the outcome was consistently unproductive. Their most common patterns—such as *Simple pass – Simple pass* or *Cross*—failed to translate into meaningful attacking opportunities.

This contrast highlights a critical tactical distinction. While both Leicester and the comparison teams operated in similar zones of the pitch and even followed comparable structural patterns, Leicester's superior execution in both transition timing and finishing decisively set them apart. Their ability to rapidly convert defensive actions into high-quality chances with minimal buildup likely played a major role in their significantly higher goal tally. In contrast, the comparison group relied more heavily on conservative buildup play and struggled to create or convert dangerous scoring opportunities.

6.2 Comparison of Cluster 0 Sequences Between Leicester City & Big 6

When comparing the spatial trajectory visualizations of Cluster 0 attack sequences between Leicester City and the Big 6 clubs (**Appendix 5**), we observe structural similarities but also some notable distinctions. Leicester's sequences follow a familiar path but tend to start slightly deeper—around the halfway line—and exhibit longer progression distances in both event segments. This spatial behavior is consistent with the types of sub-events involved, which include a greater proportion of **Crosses** and **High passes**, suggesting more vertical and direct play compared to the Big 6's emphasis on short, controlled passing.

While **Simple pass** events were still commonly used—bringing their patterns closer to those of the Big 6—the variety of Leicester's sub-event combinations remained relatively limited, and the overall frequency of Cluster 0 sequences was lower. Despite this, Leicester's efficiency was once again evident. According to the summary statistics (**Appendix 7**), their Cluster 0 sequences showed a notably high progression distance, with averages of **15.31 meters** from the first to second event and **9.50 meters** from the second to the shot. These were accompanied by progression ratios of **56% and 37%**, respectively—significantly higher than those observed for the Big 6 in the same cluster.

In addition to spatial advancement, Leicester's sequences also maintained strong execution quality. Their shots were accurate and frequently converted, reinforcing the idea that even with fewer attempts and less stylistic variation, the team was able to capitalize on key moments. This further supports the narrative that Leicester's attacking identity, though less complex than that of the Big 6, was highly effective in generating goals.

6.3 Key Players for Leicester City

To better understand the effectiveness of Leicester City's attacking sequences, we examine the players most and second most frequently involved in the

build-up and finishing phases of Cluster 5 and Cluster 0 sequences. The patterns of player involvement highlight not only tactical preferences, but also the individual contributions that made these attacks successful.

Cluster 5: Direct Transitions and Duels In Cluster 5, the most frequent and most effective sequence type for Leicester—**Riyad Mahrez** emerged as a key figure, heavily involved in both the second and final events of the sequence. His repeated presence at the end of fast transitional plays underscores his role as both a creator and finisher.

Christian Kabasele frequently initiated these sequences by winning ground duels in Leicester's defensive third, a key trigger for the transition. Though not a Leicester player himself, his presence in the data (likely due to a misclassification or data merge error) highlights the need for caution in interpreting raw player tables.

More relevantly, **Marc Albrighton** and **Demarai Gray** were frequently involved in the second event, typically delivering accurate crosses to Mahrez or the final shooter. Notably, the sequences involving crosses had a **67% goal conversion rate** and **100% shot accuracy**, indicating that Mahrez's ability to time runs and finish from wide service played a central role in Leicester's transition-based attack (**Appendix 8**).

Cluster 0: Structured Buildup and Involving the Back Line In Cluster 0 sequences, those that were more structured and resembled Big 6 attacking patterns—**Jamie Vardy** and **Shinji Okazaki** appeared most often in the final shot event, with **Mahrez** again involved frequently in the second pass. The Mahrez–Vardy connection, in particular, stood out, reinforcing the well-established partnership between Leicester's most creative and most clinical attackers.

Ben Chilwell and **Harry Maguire** were key players in initiating attacks, contributing often to the first event with **Simple passes** or **High passes**. Their regular presence in the buildup reflects Leicester's tendency to involve defenders—particularly full-backs and ball-playing centerbacks—in structured attacking sequences. Their involvement further reflects their role as a transitional conduit between defense and attack.

Another notable pattern was the consistent presence of **Marc Albrighton** in both clusters. In Cluster 0, he delivered several key passes in the second event (e.g., **High pass – Simple pass**) that led to goals by Okazaki and Vardy. His role as a wide playmaker bridging the buildup and finishing phases highlights his versatility and value in multiple attacking contexts (**Appendix 9**).

Overall, the analysis reveals a core attacking trio of **Mahrez, Vardy, and Albrighton**—each playing complementary roles across both direct and structured sequences. Mahrez acted as a hybrid creator-finisher, Vardy as a consistent final option, and Albrighton as a service provider from wide areas. In support, defenders like Chilwell and Maguire played an unexpectedly active role in shaping buildup patterns, reinforcing the team's tactical flexibility and player involvement across phases of play.

7 Discussion

7.1 Conclusion

This project aimed to explore whether teams in the English Premier League (EPL) exhibit unique patterns in their attacking sequences, particularly in the buildup to opportunity shots. Rather than focusing on individual player metrics or team statistics alone, we adopted a structural approach—defining a latent variable, *attack pattern*, based on a combination of spatial coordinates, event types, and timing between events.

We began by constructing meaningful observations from the raw match data by extracting two events prior to each shot classified as an opportunity. New features such as horizontal progression distance, progression ratio, and event duration were engineered to enhance the representation of attack sequences. These sequences were then transformed into a pairwise distance matrix using a combination of Fréchet distance (for spatial patterns) and Gower distance (for mixed-type features). The resulting distance matrix was used to define clusters via K-medoids, capturing latent attack patterns without supervision.

Statistical tests including Chi-squared and Fisher's exact test revealed a significant association between a team's identity and its attacking pattern distribution. In other words, the probability of observing a given attack pattern varied meaningfully by team—supporting the hypothesis that EPL clubs tend to rely on distinct attacking strategies.

The case studies added practical insight into these patterns. Analysis of the Big 6 clubs revealed that while their dominant sequences often belonged

to the same cluster (Cluster 0), subtle differences emerged in pass types, progression profiles, and finishing quality. For instance, Manchester City demonstrated the greatest tactical variety, while teams like Manchester United relied on more predictable sequences.

Leicester City, in contrast, stood out among non-Big 6 clubs. Despite finishing 9th in the league, they scored the most goals outside the Big 6—an outcome strongly linked to their distinct use of Cluster 5. These sequences were characterized by fast transitions initiated by ground duels and finished by high-efficiency passes or crosses, often involving Mahrez and Vardy. Compared to mid-table peers like Crystal Palace or Bournemouth—whose dominant Cluster 3 sequences yielded 0% shot accuracy—Leicester's direct but polished approach proved notably more effective.

Together, the findings demonstrate how combining spatial-temporal modeling with statistical reasoning can uncover meaningful differences in tactical behaviors across teams, offering not only academic value but also practical implications for scouting and performance analysis.

7.2 Limitation

While the project yielded meaningful insights into attacking patterns across EPL teams, several limitations should be acknowledged—primarily related to data quality and the interpretability of the clustering outcomes.

First, the spatial coordinate system in the dataset—while standardized—may lack the granularity necessary to capture subtle tactical nuances. The pitch coordinates do not adjust for variations in stadium dimensions, which can affect spacing, player behavior, and tactical structure. In addition, the positions of events are often recorded at the moment of ball contact rather than completion, which could slightly misrepresent the actual trajectory and flow of play. These limitations in spatial precision may introduce noise into the Fréchet distance calculations used to quantify trajectory similarity.

Second, while clustering offered a useful way to define latent attack patterns, the overall silhouette scores—particularly across higher values of K—were relatively low. This suggests that the structure of the data may not naturally lend itself to clear, well-separated clusters. As a result, the interpretability and reliability of some cluster assignments should be treated with caution. It is possible that certain teams' patterns lie on a continuum of styles rather than fitting into discrete groups, which could limit the effectiveness of hard clustering methods such as K-medoids.

Together, these limitations highlight that while the results provide useful directional insights, they may not fully capture the complexity of tactical behavior or the continuous nature of attacking strategy in football.

7.3 Future work

Several avenues remain for extending and enriching the current analysis. One of the most critical next steps involves the incorporation of additional contextual and spatial information—particularly that which captures player positioning beyond the immediate ball interaction. The current dataset only provides spatial coordinates for events involving the ball, leaving out the positions of other players on the pitch. This limits the ability to evaluate off-ball movement, defensive structure, and pressing patterns, all of which are essential components in modern football analysis. Future studies could benefit from datasets that include full-pitch player tracking data, enabling features such as player density, formation shape, or space creation metrics. Additional modern features such as expected goals (xG) or pressure metrics would further enhance the tactical resolution of the models.

Another direction would be to expand the focus beyond attacking sequences to include defensive analysis. For example, Burnley finished 7th in the league—outperforming several clubs with greater attacking outputs—primarily due to their strong defensive performance, conceding the fewest goals outside the Big 6. Unfortunately, the current dataset lacks sufficient defensive tracking information to study their tactics in detail. A dedicated investigation into **"What makes Burnley successful?"**, with attention to spatial compactness, dueling success, and pressing resistance, could offer a compelling contrast to the attack-focused lens used in this study. More broadly, future work could also explore the balance between offensive and defensive efficiency as factors of league success—testing whether defensive solidity is a stronger predictor of league position than scoring ability.

With richer data and more refined modeling tools, future analyses could move beyond descriptive clustering to predictive and evaluative frameworks, helping coaches, analysts, and scouts better understand the underlying mechanics of success in football.

References

Alan Agresti. 2002. *Categorical Data Analysis*, 2 edition. Wiley.

Helmut Alt and Michael Godau. 1995. Computing the fréchet distance between two polygonal curves. In *International Journal of Computational Geometry & Applications*, volume 5, pages 75–91. World Scientific.

Opta Analyst. 2024. The strongest leagues in world football: Opta power rankings. Accessed: 2025-06-04.

Chris Carling, A. Mark Williams, and Thomas Reilly. 2005. *Handbook of Soccer Match Analysis: A Systematic Approach to Improving Performance*. Routledge, London, UK.

J. C. Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edition. Springer, New York, NY.

Leonard Kaufman and Peter J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.

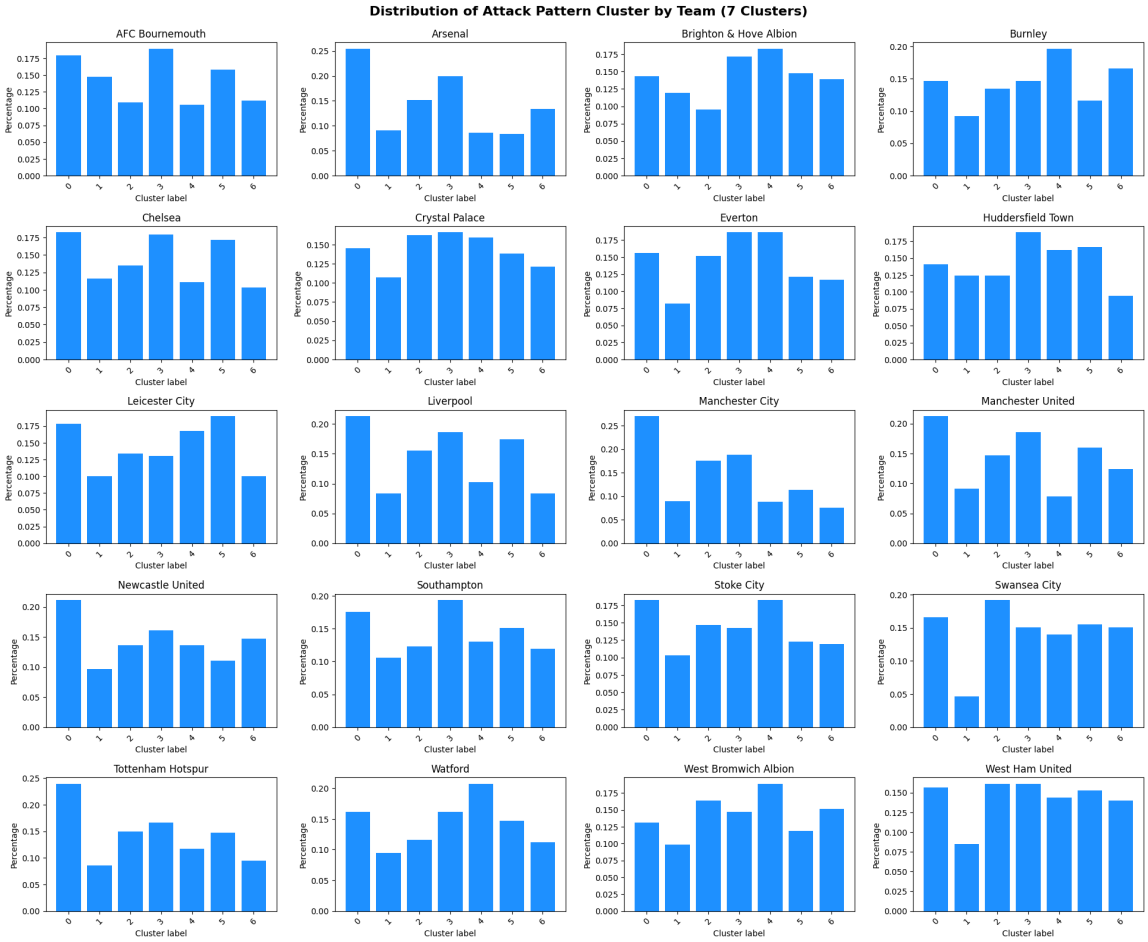
Trupti M. Kodinariya and Prashant R. Makwana. 2013. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Luca Pappalardo, Paolo Cintia, Alessandro Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. 2019. A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6:236.

Hugo Sarmento, Filipe Manuel Clemente, Liam D. Harper, Ítalo T. da Costa, Adam Owen, and António J. Figueiredo. 2018. Match analysis in football: A systematic review. *Journal of Sports Sciences*, 36(14):1533–1543.

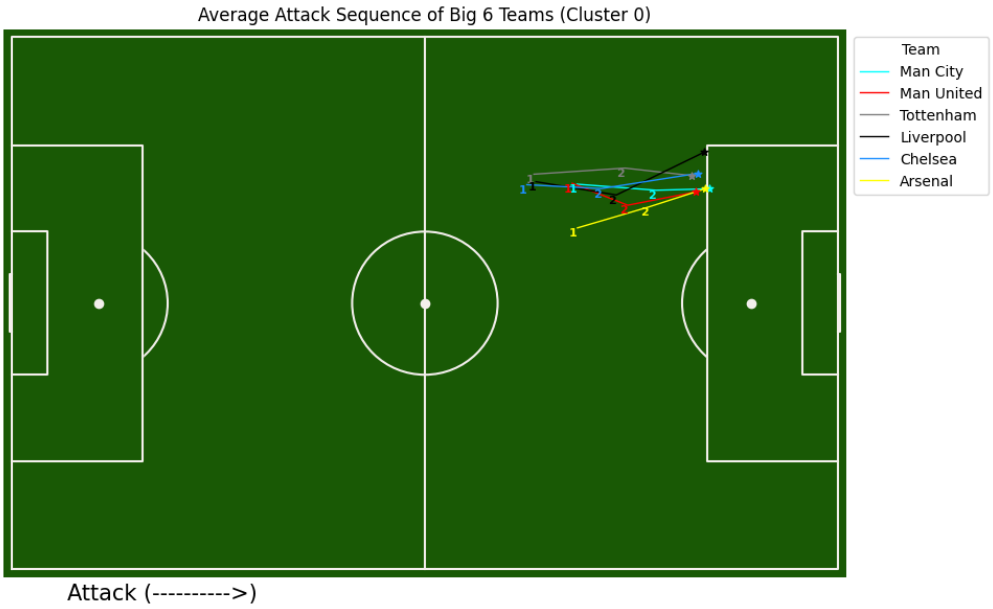
A Appendix



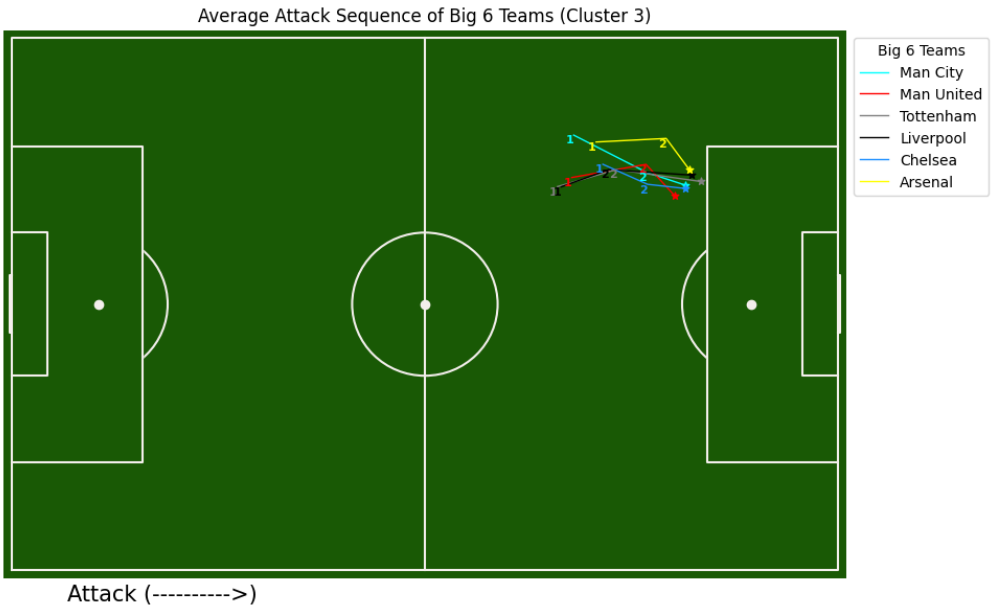
Appendix 1: Distribution of Attack Pattern by Team

Team	0	1	2	3	4	5	6	Total	Max	2nd Max	Proportion
AFC Bournemouth	51	42	31	54	30	45	32	285	3	0	0.189
Arsenal	101	36	60	79	34	33	53	396	0	3	0.255
Brighton	36	30	24	43	46	37	35	251	4	3	0.183
Burnley	38	24	35	38	51	30	43	259	4	6	0.197
Chelsea	69	44	51	68	42	65	39	378	0	3	0.183
Crystal Palace	42	31	47	48	46	40	35	289	3	2	0.166
Everton	36	19	35	43	43	28	27	231	3	4	0.186
Huddersfield Town	33	29	29	44	38	39	22	234	3	5	0.188
Leicester City	48	27	36	35	45	51	27	269	5	0	0.190
Liverpool	89	35	65	78	43	73	35	418	0	3	0.213
Manchester City	126	42	82	88	41	53	35	467	0	3	0.270
Manchester United	65	28	45	57	24	49	38	306	0	3	0.212
Newcastle United	59	27	38	45	38	31	41	279	0	3	0.211
Southampton	50	30	35	55	37	43	34	284	3	0	0.194
Stoke City	46	26	37	36	46	31	30	252	0	4	0.183
Swansea City	32	9	37	29	27	30	29	193	2	0	0.192
Tottenham Hotspur	86	31	54	60	42	53	34	360	0	3	0.239
Watford	46	27	33	46	59	42	32	285	4	0	0.207
West Bromwich Albion	32	24	40	36	46	29	37	244	4	2	0.189
West Ham United	37	20	38	38	34	36	33	236	2	3	0.161

Table 2: Distribution of Attack Pattern with Maximum & 2nd Maximum Patterns by Team



Appendix 2: Spatial Trajectories of Average Attack Sequences for Big 6 Clubs (Cluster 0)



Appendix 3: Spatial Trajectories of Average Attack Sequences for Big 6 Clubs (Cluster 3)

Team	First Event	Second Event	Count	Goals	Goal %	Accuracy
Arsenal	Simple pass	Simple pass	70	8	0.11	0.543
		Smart pass	18	5	0.28	0.667
		Cross	13	2	0.15	0.462
	Smart pass	Simple pass	8	1	0.12	0.375
		Cross	7	2	0.29	0.571
	Touch	Simple pass	7	3	0.43	0.714
	Ground attacking duel	Simple pass	6	2	0.33	0.833
	Simple pass	High pass	6	0	0.00	0.500
	High pass	Cross	5	2	0.40	0.400
Chelsea	Simple pass	Simple pass	36	2	0.06	0.556
		Smart pass	14	2	0.14	0.571
		Cross	13	1	0.08	0.462
		High pass	7	2	0.29	0.714
	Ground attacking duel	Simple pass	5	1	0.20	0.600
	Simple pass	Acceleration	5	0	0.00	0.600
Liverpool	Simple pass	Simple pass	41	4	0.10	0.512
		Cross	24	6	0.25	0.458
		Smart pass	15	6	0.40	0.667
		High pass	9	1	0.11	0.667
	Ground attacking duel	Cross	7	2	0.29	0.571
		Simple pass	6	2	0.33	0.500
	Simple pass	Touch	6	0	0.00	0.333
	Acceleration	Simple pass	5	2	0.40	0.600
Manchester City	Simple pass	Simple pass	64	9	0.14	0.516
		Cross	26	12	0.46	0.692
		Smart pass	21	4	0.19	0.619
		High pass	12	5	0.42	0.667
	Smart pass	Cross	11	2	0.18	0.455
	Touch	Simple pass	8	1	0.12	0.375
	High pass	Simple pass	7	1	0.14	0.429
	Smart pass	Simple pass	7	4	0.57	0.714
	Acceleration	Simple pass	6	1	0.17	0.667
	Corner	Simple pass	6	2	0.33	0.500
	Ground attacking duel	Simple pass	5	0	0.00	0.800
	Simple pass	Touch	5	0	0.00	0.200
Manchester United	Simple pass	Simple pass	42	6	0.14	0.524
		Cross	8	1	0.12	0.250
		Smart pass	7	2	0.29	0.714
	Ground attacking duel	Cross	5	2	0.40	0.600
Tottenham Hotspur	Simple pass	Simple pass	37	2	0.05	0.595
		Cross	21	7	0.33	0.381
		Smart pass	16	1	0.06	0.688
	Acceleration	Smart pass	8	4	0.50	0.750
	Ground attacking duel	Simple pass	7	1	0.14	0.571
	Simple pass	High pass	7	1	0.14	0.429
		Touch	5	0	0.00	0.400
	Smart pass	Simple pass	5	3	0.60	0.600

Table 3: Big 6 Attack Sequence Event & Stats (Combined Clusters)

Team	Count	progress_dist_12	progress_ratio_12	progress_dist_23	progress_ratio_23	event_duration_12	event_duration_23
Arsenal	101	8.93 (13.44)	0.33 (0.58)	7.1 (11.42)	0.36 (0.52)	2.16 (1.45)	1.69 (0.82)
Chelsea	69	9.36 (17.64)	0.32 (0.55)	12.14 (13.3)	0.4 (0.44)	2.21 (1.28)	2.08 (1.33)
Liverpool	89	10.11 (14.75)	0.46 (0.52)	11.07 (13.05)	0.41 (0.46)	2.23 (1.33)	1.98 (1.01)
Manchester City	126	9.94 (14.82)	0.42 (0.58)	6.75 (13.08)	0.24 (0.49)	2.4 (1.22)	1.84 (0.93)
Manchester United	65	6.98 (13.4)	0.29 (0.51)	8.52 (11.53)	0.4 (0.43)	2.17 (1.36)	1.83 (0.9)
Tottenham Hotspur	86	11.45 (15.64)	0.43 (0.53)	8.35 (11.89)	0.31 (0.46)	2.68 (1.39)	1.9 (1.03)

Table 4: Big 6 Average Progress & Event Duration of Cluster 0 (Standard Deviation in Parentheses)

Team	Count	progress_dist_12	progress_ratio_12	progress_dist_23	progress_ratio_23	event_duration_12	event_duration_23
Arsenal	79	8.80 (12.19)	0.34 (0.55)	2.92 (9.91)	0.10 (0.46)	2.38 (1.73)	1.69 (0.85)
Chelsea	68	5.62 (18.77)	0.13 (0.58)	4.81 (14.79)	0.06 (0.52)	2.49 (2.35)	1.61 (1.02)
Liverpool	78	5.99 (14.95)	0.23 (0.58)	10.40 (14.14)	0.38 (0.51)	2.77 (3.01)	1.94 (1.02)
Manchester City	88	9.20 (15.41)	0.37 (0.58)	4.86 (11.72)	0.17 (0.48)	2.54 (1.38)	1.79 (0.79)
Manchester United	57	9.33 (16.74)	0.26 (0.63)	3.56 (11.14)	0.10 (0.48)	2.56 (2.01)	1.94 (0.92)
Tottenham Hotspur	60	7.43 (10.71)	0.34 (0.57)	10.55 (14.30)	0.33 (0.43)	2.65 (1.52)	1.93 (1.03)

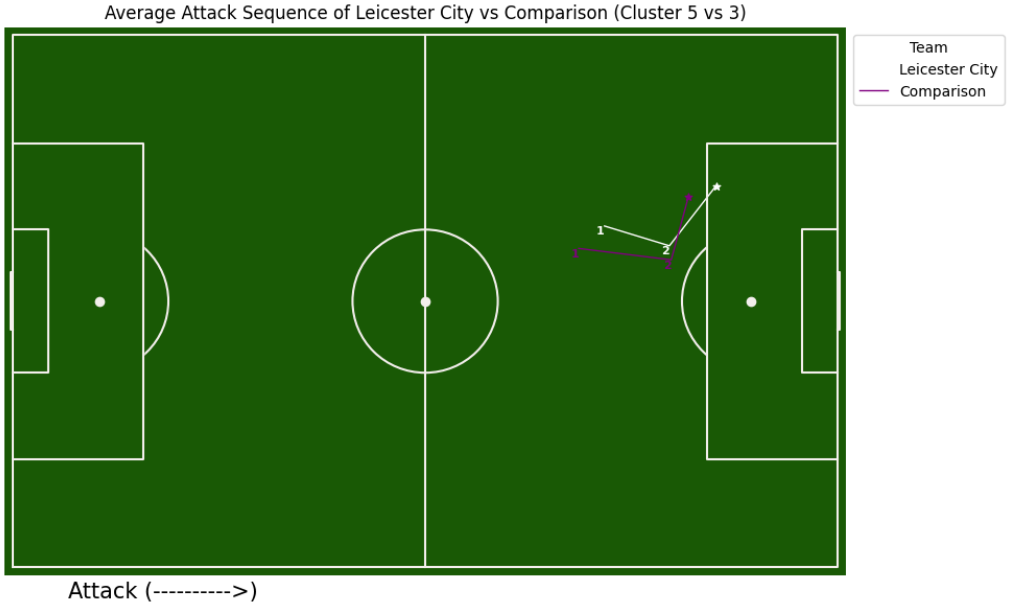
Table 5: Big 6 Average Progress & Event Duration of Cluster 3

Team	First Event	Second Event	Count	Goals	Goal %	Accuracy
Leicester City	Ground defending duel pass	Ground attacking duel	28	10	0.36	0.75
		Cross	6	4	0.67	1
Comparison	Simple pass	Simple pass	40	0	0.00	0.00
		Cross	17	0	0.00	0.00
	Ground attacking duel	Cross	9	0	0.00	0.00
		Simple pass	9	0	0.00	0.00
	Simple pass	Smart pass	7	0	0.00	0.00
	Ball out of the field	Corner	6	0	0.00	0.00
	Simple pass	High pass	6	0	0.00	0.00
	Acceleration	Simple pass	5	0	0.00	0.00

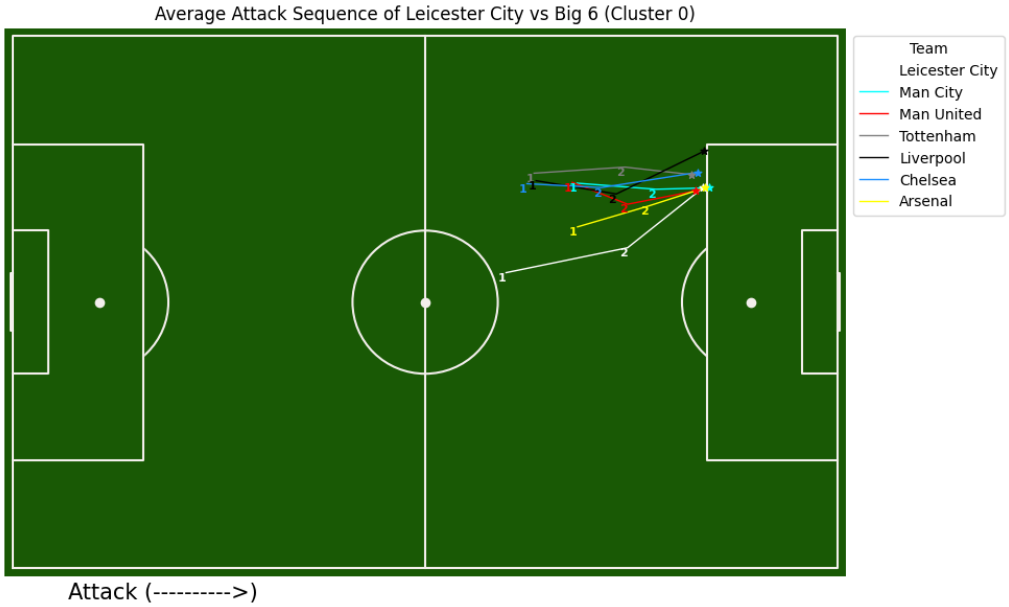
Table 6: Leicester City vs Comparison Group Attack Sequence Event & Stats

Cluster	Count	progress_dist_12	progress_ratio_12	progress_dist_23	progress_ratio_23	event_duration_12	event_duration_23
0	48	15.31 (14.09)	0.56 (0.45)	9.5 (12.13)	0.37 (0.48)	2.87 (1.46)	1.59 (0.88)
5	51	8.18 (33.78)	0.28 (0.62)	5.86 (19.04)	0.15 (0.51)	3.29 (10.68)	1.33 (0.85)

Table 7: Leicester City Average Progress & Event Duration



Appendix 4: Spatial Trajectories of Average Attack Sequences for Leicester City vs Comparison



Appendix 5: Spatial Trajectories of Average Attack Sequences for Leicester City vs Big 6

First Event	Second Event	Player 1	Player 2	Player 3
Ground defending duel	Ground attacking duel	C. Kabasele / J. Gomez	R. Mahrez	R. Mahrez
	Cross	D. Janmaat / J. Gomez	D. Gray / M. Albrighton	R. Mahrez

Table 8: Key Player Involvement in Leicester City's Cluster 5 Sequences (Most Involved / Second Most Involved)

First Event	Second Event	Player 1	Player 2	Player 3
Simple pass	Simple pass	Adrien Silva / B. Chilwell	R. Mahrez	R. Mahrez
	High pass	H. Maguire	M. Albrighton	J. Vardy
High pass	Simple pass	M. Albrighton	R. Mahrez	S. Okazaki
Simple pass	Cross	B. Chilwell / M. Albrighton	J. Vardy	S. Okazaki
	Smart pass	H. Maguire	Adrien Silva / K. Iheanacho	I. Slimani / J. Vardy
Ground attacking duel	Cross	D. Gray	D. Gray	A. King / J. Vardy
Acceleration	Simple pass	R. Mahrez	R. Mahrez	K. Iheanacho

Table 9: Key Player Involvement in Leicester City's Cluster 0 Sequences (Most Involved / Second Most Involved)