An Empirical Analysis of Striker Performance using Data Science Techniques: A Method for Evaluating and Ranking Strikers Based on Multiple Aspects of Their Game

Strikers' Ranking

Barnsley Football Club

Written by

Marwane Hamdani

Abstract

The purpose of this report is to introduce a data-driven approach for evaluating and ranking strikers' performance based on their statistics. This methodology was developed by an individual with a passion for football analytics and with the hunger and desire to get started in the field. While not a professional expert in the field, the author of this report believes that this approach, in conjunction with the knowledge and expertise of professionals working in the club, can be used to identify and acquire top-performing strikers from the league.

Additionally, this methodology can serve as a foundation for further analysis and conditioning based on the team's playing style and specific needs. It is important to note that while data analytics can provide valuable insights, it should not be relied upon solely, and expert knowledge and judgment should be integrated into the decision-making process.

The report is structured into four sections, namely Introduction, Methodology, Practical Implementation, and Results and Conclusion. The Introduction provides an overview of the importance of data skills in football analytics and outlines the problem statement as perceived by the author. The Methodology section discusses the theoretical foundations of the methods used and explains their benefits. The Practical Implementation section describes how the methodology was applied in practice, along with any refinements made during the process. Lastly, the Results and Conclusion section presents the insights gained and summarizes the conclusions drawn from the method used.

Contents

List of Figures

1 Introduction

1.1 Data in Football

In football, goals are paramount in determining success and failure, as they are often accompanied by intense emotions. Teams that struggle to score goals often find themselves performing poorly in the league or being eliminated from competitions, despite undeserving outcomes. Therefore, it is logical to assume that finding quality goal scorers is an expensive and difficult task, particularly in today's market, where good strikers are becoming increasingly rare. Furthermore, certain strikers may not fit well with a team's playing style, or their performance may be limited to a single exceptional season. In such cases, the use of data analysis can be a valuable tool to aid in the search for effective goal scorers.

Data analysis has become an essential tool for football clubs in recent years, as it can provide valuable insights into a team's performance and the effectiveness of individual players. In the context of finding quality goal scorers, data analysis can be particularly useful as it can help identify players who are statistically proven to be effective in scoring goals.

One of the most important metrics for evaluating a striker's effectiveness is their conversion rate, which is the percentage of shots that result in goals. By analyzing a player's conversion rate, clubs can identify those who have a proven track record of scoring goals and are likely to continue doing so in the future. Furthermore, data analysis can be used to identify other key metrics such as a player's expected goals (xG) and expected assists (xA), which can provide a more nuanced view of a player's performance and potential value to a team.

Data analysis can also be used to evaluate how well a particular striker fits with a team's playing style. By analyzing a player's touch maps and heat maps, clubs can gain insights into a player's movement patterns and positioning on the pitch. This can help identify players who are likely to fit well into a team's system and contribute to the team's overall success.

In conclusion, data analysis is an important tool for football clubs in the search for effective goal scorers. By providing valuable insights into a player's performance and potential fit with a team's playing style, data analysis can help clubs identify players who are likely to contribute to the team's success on the pitch.

1.2 Problem Statement

The process of finding a top striker in football is a complex task that cannot be simplified to just one metric, such as non-penalty goals per 90. As an avid fan of football and someone who is fascinated by the evolution of strikers, I understand the importance of taking multiple parameters into consideration.

While it may be tempting to compare two top strikers, such as Harry Kane and Erling Haaland, based solely on their goal-scoring record, this approach overlooks the fact that each player has a unique style and skill set. It is important to recognize that football has evolved beyond simply defining players by their position on a lineup sheet. Instead, it is necessary to consider the spaces each player operates in and the contributions they make to their team.

For example, a striker who excels in holding up play and creating opportunities for their teammates may not have the highest goal-scoring record but can be just as valuable to their team's success as a top goal scorer. Furthermore, the playing style of a team can greatly influence the effectiveness of a striker, highlighting the importance of finding a player who is not only talented but also fits well within the team's tactics and style of play.

In conclusion, the process of finding a top striker requires a comprehensive analysis of multiple parameters, beyond just their goal-scoring record. Understanding a player's unique style and skill set, as well as how they fit within a team's tactics and style of play, is essential to identifying the most effective strikers for a team.

2 Methodology

2.1 Overall view

As previously discussed, the process of ranking strikers based on a single or even multiple features is not sufficient in defining their profile. To address this limitation, a more comprehensive approach was taken by clustering strikers based on different patterns, taking into account multiple aspects of the game.

To begin, a separation was made between different aspects of the game, such as passing features, finishing features, dribbling features, and work rate features. Each aspect was assigned a score, and then the strikers were clustered based on these scores. This method allowed for a more holistic evaluation of each striker, beyond just their goal-scoring record, and provided a more nuanced understanding of their unique style and skill set.

By clustering the strikers based on these different aspects of the game, it became possible to rank them appropriately, taking into consideration their overall contributions to their team's success. This approach offers a more nuanced and detailed analysis of the strikers, providing valuable insights for managers and coaches in their efforts to identify the most effective players for their team.

2.2 Feature separation

In my opinion, a top striker's arsenal comprises four main aspects: finishing ability, passing ability, dribbling ability, and work rate off the ball. Each of these aspects is crucial in defining a top-quality striker, and they can work effectively in almost any system, particularly those involving high pressing, energy, and intensity.

However, given the large number of features associated with each striker, dimension reduction becomes a necessity. Therefore, the UMAP algorithm was used to represent a large number of features in a smaller number of representations, with only two being necessary to obtain an overall view of the player's information in this aspect.

The UMAP algorithm enabled the dataset to be reduced without losing significant information, thereby simplifying the evaluation of each striker's unique style and skill set. This method allowed for a more efficient and precise analysis of each striker's performance and contributed to the identification of the most effective players for a particular team or playing style.

2.3 Clustering and ranking

Once the features of the four main aspects of a striker's arsenal were separated and reduced to two representations for each ability, the next step was to regroup similar strikers based on different patterns. The Gaussian mixture model proved to be an ideal tool for this task. The identified clusters of strikers were then evaluated to determine their principal characteristics.

Ranking the strikers based on these clusters enabled us to obtain a more accurate and comprehensive assessment of their overall performance. Clustering similar strikers together allowed us to

compare them based on shared traits, such as finishing ability, passing ability, dribbling ability, and work rate off the ball.

By analyzing the characteristics of each cluster, we can identify the strengths and weaknesses of each group of strikers. This allows us to better understand how they perform in different situations and what role they could play in a specific team or playing style. This approach provides a more nuanced evaluation of a striker's overall ability beyond just their goal-scoring record.

Clustering similar strikers based on different patterns and characteristics provides a more accurate and comprehensive assessment of their overall performance. The use of the Gaussian mixture model in identifying these clusters allowed us to rank the strikers based on shared traits, leading to a more nuanced evaluation of their abilities beyond just their goal-scoring record.

3 Practical Implementation

3.1 A Simple Exploratory Data Analysis

The dataset is well-prepared for conducting data analysis, as it does not contain any missing or duplicated values, nor does it include any outliers. This creates a suitable foundation for applying data science methodologies to the dataset. However, before proceeding with any analysis, it is essential to identify the correlations between the various features. Given the large number of features present in the dataset, it is imperative to segment and categorize them efficiently for effective analysis.

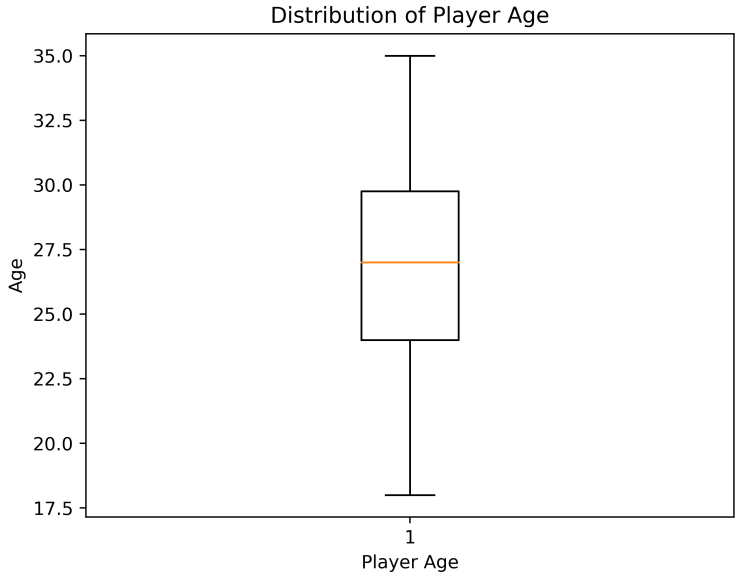

Figure 1: No outliers present in the Ages feature.

The Age feature is well distributed.

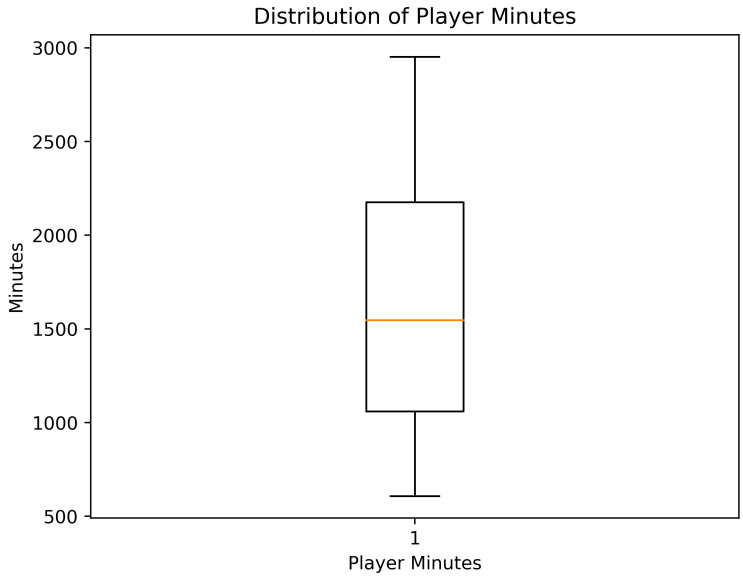

Figure 2: No outliers in the Minutes feature.

The Minutes feature is well distributed.

3.2 Feature Separation

The distinct characteristics and attributes that contribute to a striker's performance in football necessitated a structured approach to feature extraction. Specifically, the creative process involved segmenting the features into four primary stages that correspond to a striker's playing style: passing ability, finishing ability, dribbling and ball carrying, and work rate off the ball. Subsequently, we endeavored to uncover the correlations that exist between the different features across these various aspects of a striker's gameplay.

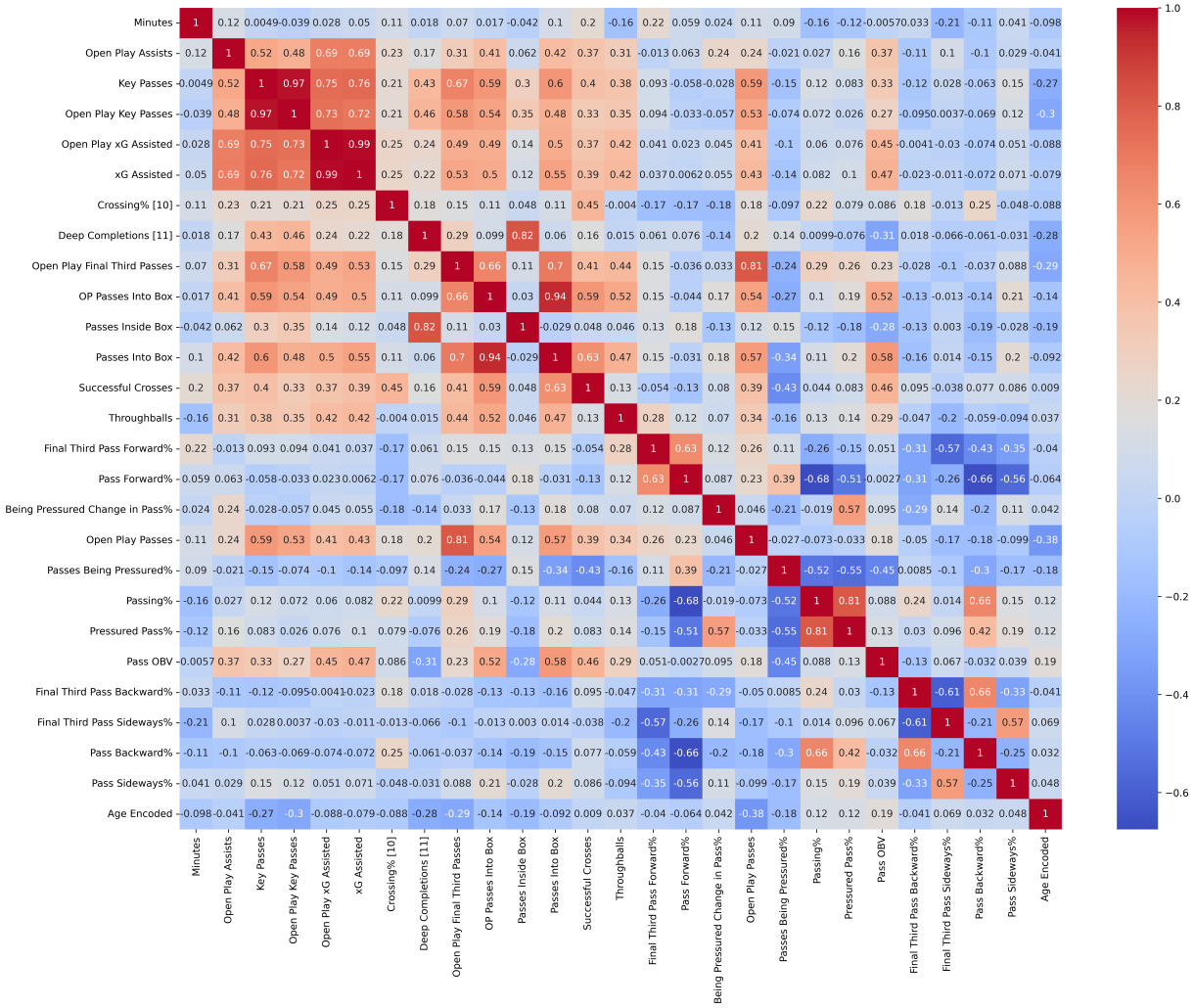

Figure 3: Correlation matrix of the passing features.

Some features of the same sequence are correlated while some just cause the others.

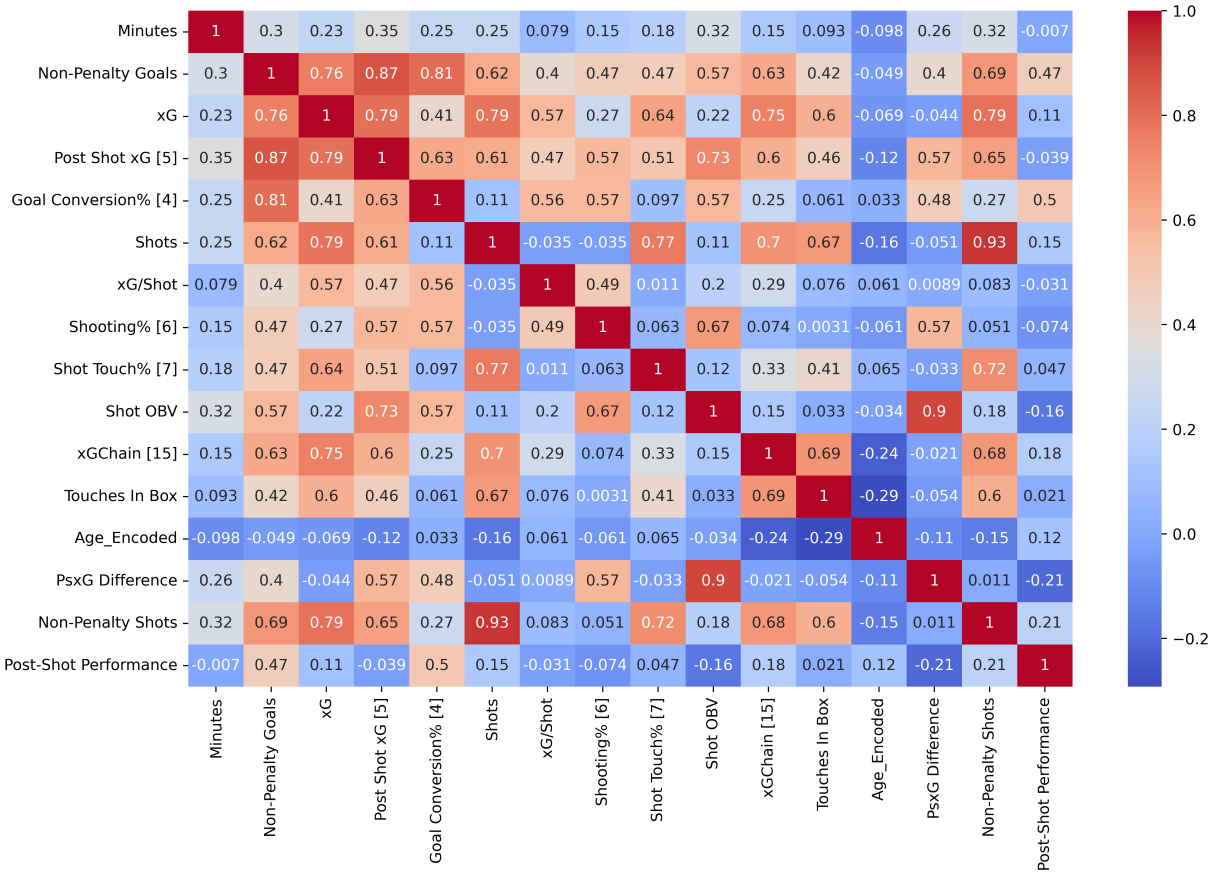

Figure 4: Correlation matrix of the finishing features.

Just like the passing features, some belong to the same sequence while some cause others.

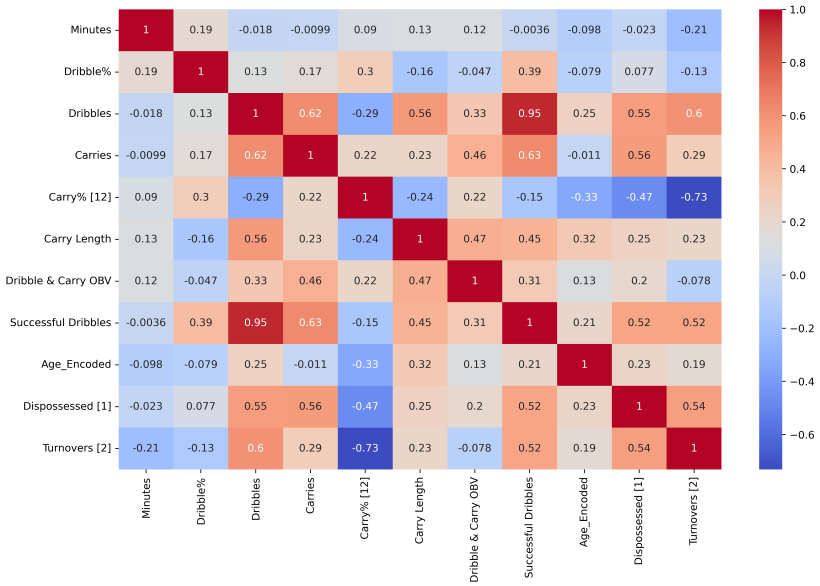

Figure 5: Correlation matrix of the dribbling features.

A high number of dribbles and carries comes with both positive and negative to the game of a striker.

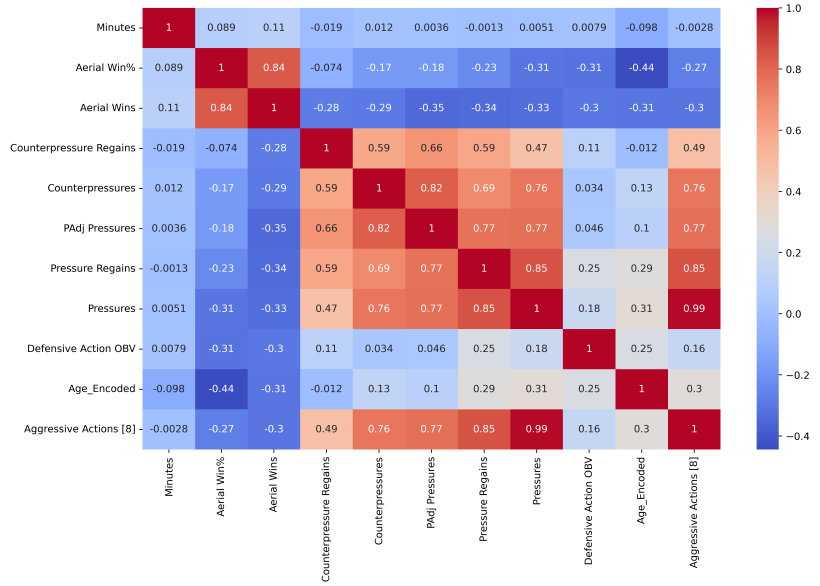

Figure 6: Correlation matrix of the work rate features.

The work rate features are a bit random with lots belonging to one sequences.

The correlation between features is an important aspect to consider in data analysis, as it can affect the effectiveness of dimensionality reduction techniques such as UMAP. In our analysis of the football data, we observed that the finishing and passing features exhibit a moderate level of correlation with each other of the same aspect, which is not surprising given the logical relationship between these features. However, we also noted that the features in the dribbling and work rate categories do not exhibit significant correlation with each other, which is a positive aspect of the data as it provides a good starting point for applying dimensionality reduction techniques.

It is important to keep in mind that the level of correlation between features can affect the results of dimensionality reduction techniques. For example, if the features exhibit strong linear correlations, techniques like PCA may be more effective at reducing the dimensionality of the data. However, if the features exhibit more complex, nonlinear relationships, UMAP may be more appropriate. Therefore, it is always a good practice to explore the data and evaluate the level of correlation between features before selecting an appropriate dimensionality reduction technique.

3.3 Dimensionality reduction

Dimensionality reduction is an important step in data analysis, particularly when dealing with high-dimensional data such as the football dataset we are working with, which contains over 60 features. The aim of dimensionality reduction is to reduce the number of features in the dataset while retaining the most important information. By reducing the dimensionality of the data, we can simplify the analysis and make it easier to visualize and interpret.

UMAP (Uniform Manifold Approximation and Projection) is a popular dimensionality reduction technique that can be used to visualize high-dimensional data in two or three dimensions. UMAP is particularly useful for preserving the nonlinear structure of the data, which is often lost in other dimensionality reduction techniques like PCA.

In our analysis of the football data, we separated the features into four different aspects of the game (finishing, passing, dribbling, and work rate), and for each aspect, we applied dimensionality reduction using UMAP. This allowed us to create two-dimensional embeddings for each aspect, which we can use to visualize and analyze the data in a more simplified form. By reducing the dimensionality of the data, we were able to focus on the most important features and relationships between them, which can provide valuable insights into the players' performance.

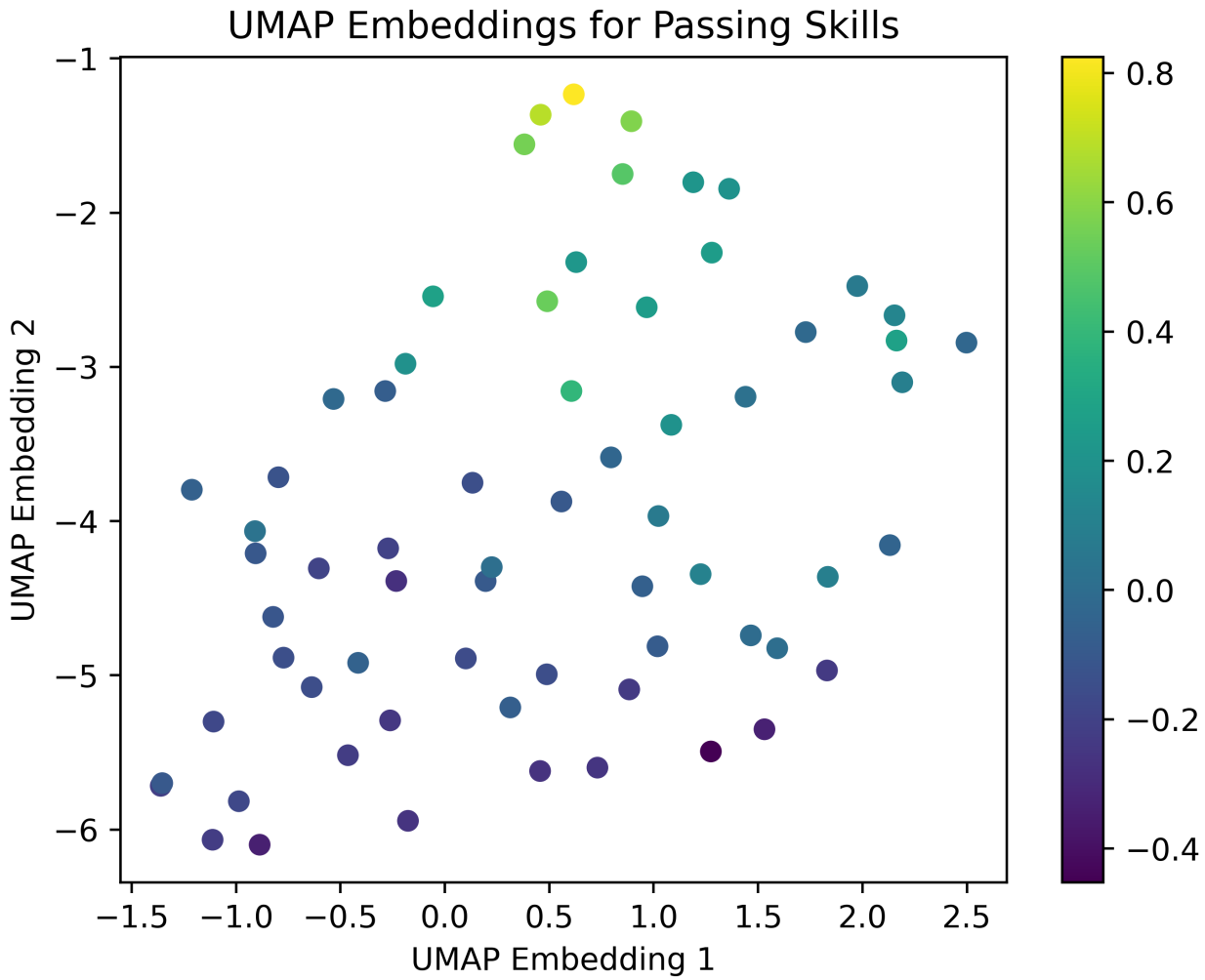

Figure 7: The UMAP Embeddings of the Passing Features.

By splitting passing features into two embeddings, we were able to effectively represent the most extensive aspect of the game. The colors in the embeddings indicate the player's overall passing score, which we will describe in detail later.

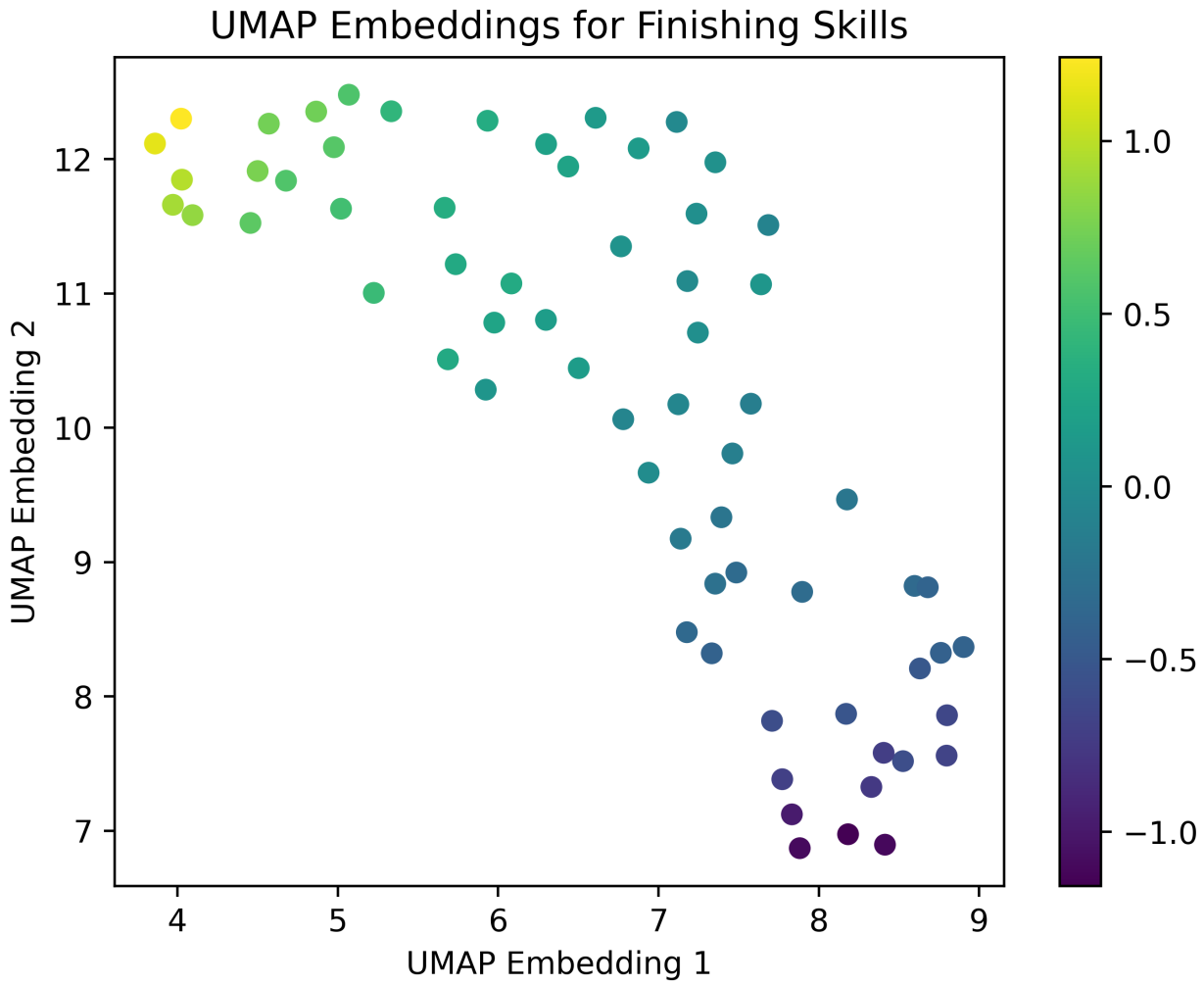

Figure 8: The UMAP Embeddings of the Finishing Features.

Similar to the passing features, we also represented all finishing features into two embeddings, and this resulted in better distribution of the finishing scores.

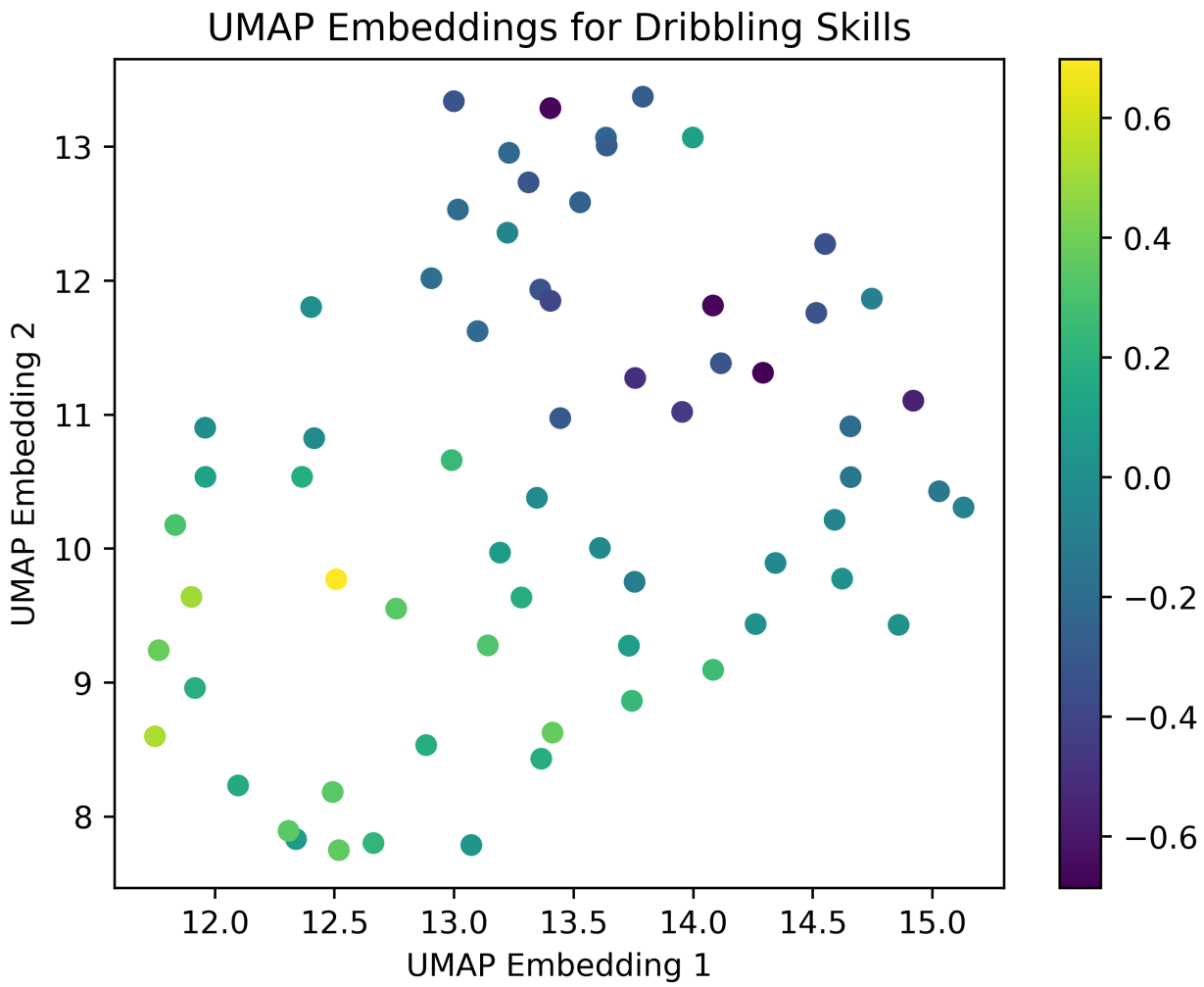

Figure 9: The UMAP Embeddings of the dribbling Features.
Reducing the dimensionality of the dribbling features, which are also randomly distributed, was performed in a similar manner.

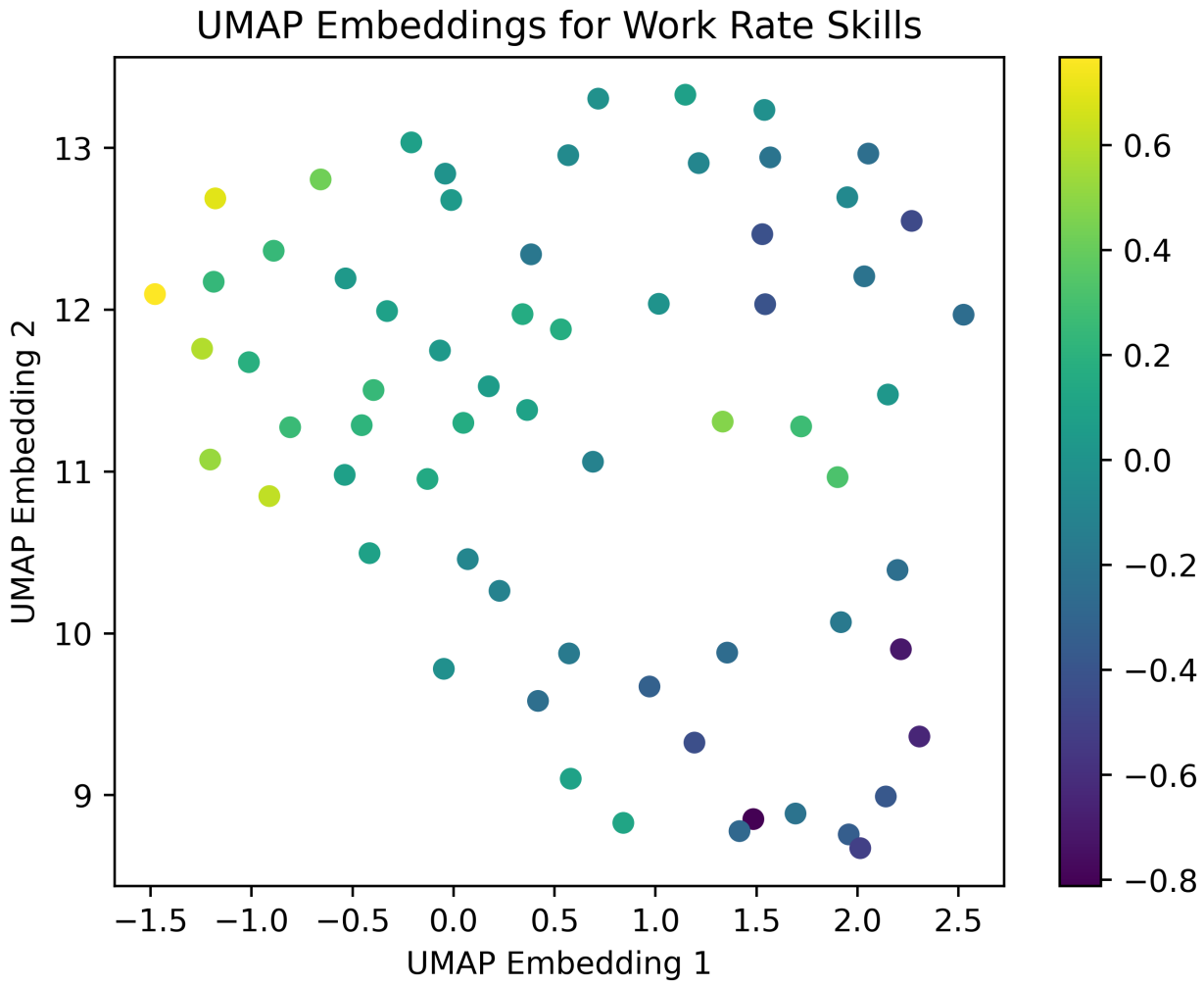

Figure 10: The UMAP Embeddings of the Work Rate Features.

To reduce the dimensionality of the work rate features, which are also random, we employed a similar technique of grouping them into two embeddings.

In the process of dimensionality reduction, we first separated the features of each aspect into positive and negative ones. Each feature was then attributed a positive or negative score, depending on whether it was indicative of good or bad performance. For example, a high number of dispossession for a striker is considered a bad characteristic, while a high dribble percentage is considered a positive one.

Some of these evaluations were straightforward, while others were more subjective. For example, a high number of sideways or backward passes should not be evaluated the same way as a high number of forward passes.

By separating the features into positive and negative ones, we were able to focus on giving an overall score to each aspect of the game. This allowed us to reduce the dimensionality of the data and create two-dimensional embeddings using UMAP, as the figures show for each aspect.

Overall, the process of dimensionality reduction was crucial to our analysis of the football dataset. By separating the features and attributing positive and negative scores, we were able to focus on the most important aspects of the game and reduce the dimensionality of the data in a meaningful way. This allowed us to create more easily interpretable visualizations and gain insights into the players' performance.

3.4 Clustering the strikers

After reducing the dimensions of each aspect to two embeddings, the next step is to group similar strikers together. To achieve this, we used the Gaussian Mixture model (GMM). The GMM is a probabilistic model that represents the distribution of data points as a mixture of several Gaussian distributions. In our case, the GMM was used to cluster the strikers into groups based on their similarity in terms of the extracted features.

The GMM model is particularly useful in cases where the underlying data distribution is complex and cannot be easily captured by a single distribution. It allows for the identification of subpopulations within a larger population, which can be useful for various applications such as anomaly detection or customer segmentation.

In our case, the GMM model was used to identify groups of strikers with similar skill sets, which can be useful for scouting or team selection purposes. By clustering the strikers based on their similarity, we can identify groups of players with similar strengths and weaknesses, which can inform decision-making in terms of team selection or player recruitment.

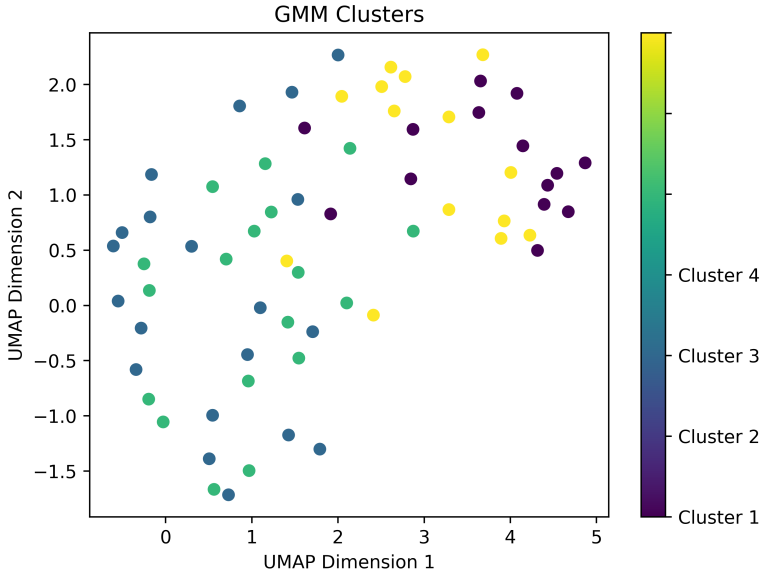

Figure 11: The Clusters obtained.

After representing the aspects of the game into two embeddings each, the next step was to cluster similar strikers together using Gaussian Mixture Model (GMM). Initially, the decision was made to use four clusters that aligned with the four aspects of the game, and the small sample size seemed to justify this approach. However, upon analyzing the clusters, we found that the strikers were not grouped into specific profiles (such as the finisher or hard worker) as expected. Instead, they were grouped based on their overall performance, which was also effective in its own way.

To refine the analysis, we added two new scores, the "Overall Score," which is the sum and average of the scores of all four aspects, and the "On The Ball Score," which focuses on dribbling and passing skills to evaluate how good a striker is with their feet. This allowed us to identify more technical strikers who are not just finishers, and it made the ranking of the strikers more straightforward.

GMM was the ideal clustering algorithm in this case because it allowed us to handle the mixture of distributions and the uncertainty around the clusters, which is a common issue in unsupervised learning. By adopting GMM, we were able to identify groups of similar strikers based on their overall performance and technical skills, and this allowed us to gain valuable insights into the data.

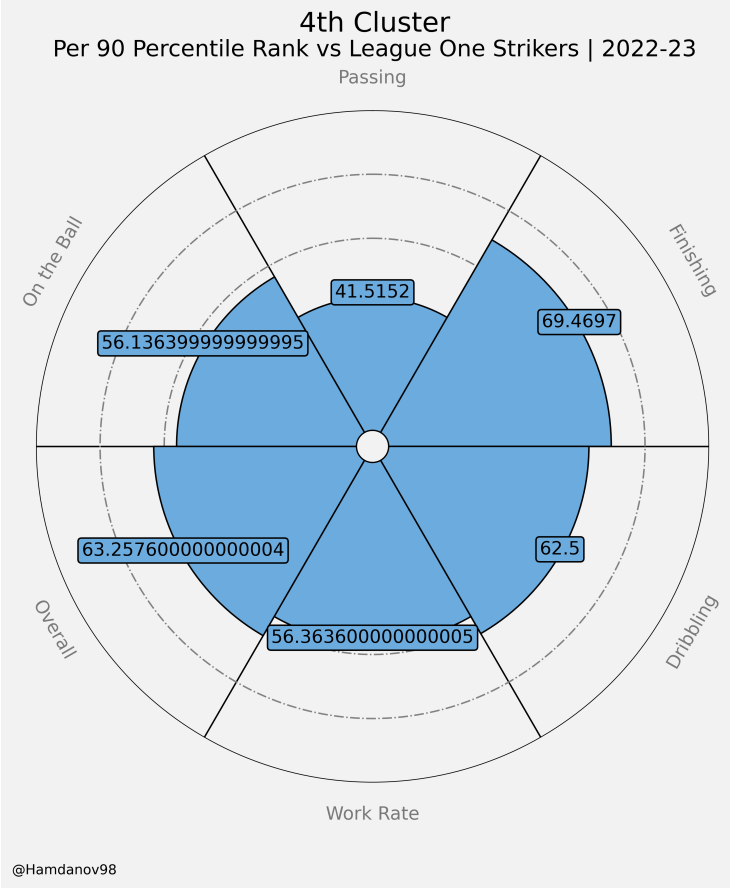

Figure 12: How an average player looks from the fourth cluster.

The most balanced of the clusters, the cluster that contains the best overall strikers.

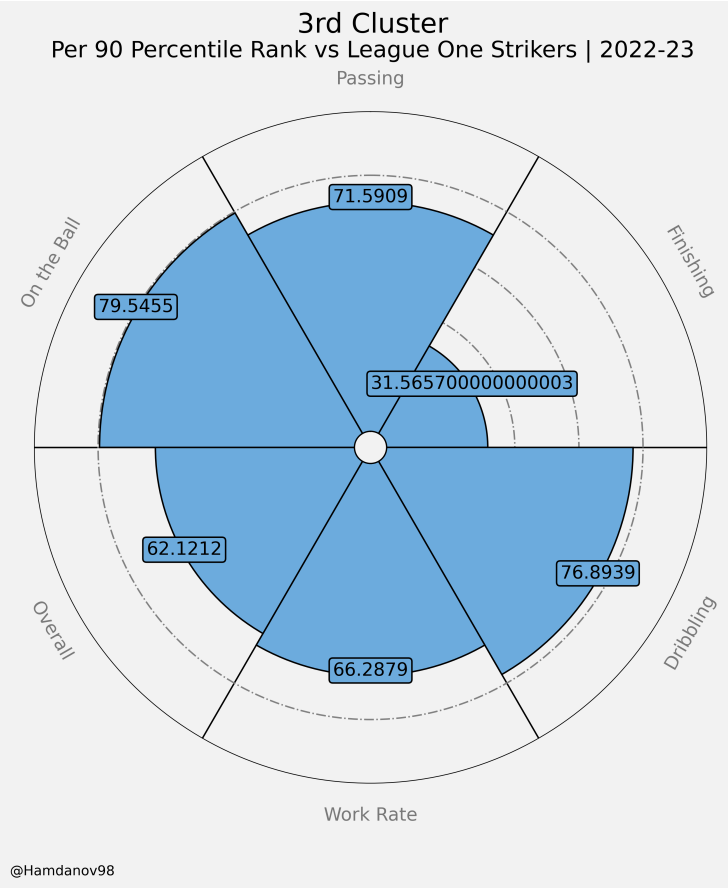

Figure 13: How an average player looks from the third cluster.
This cluster comprises the technically proficient strikers who excel at ball retention and executing accurate passes, but may fall short in terms of finishing prowess.

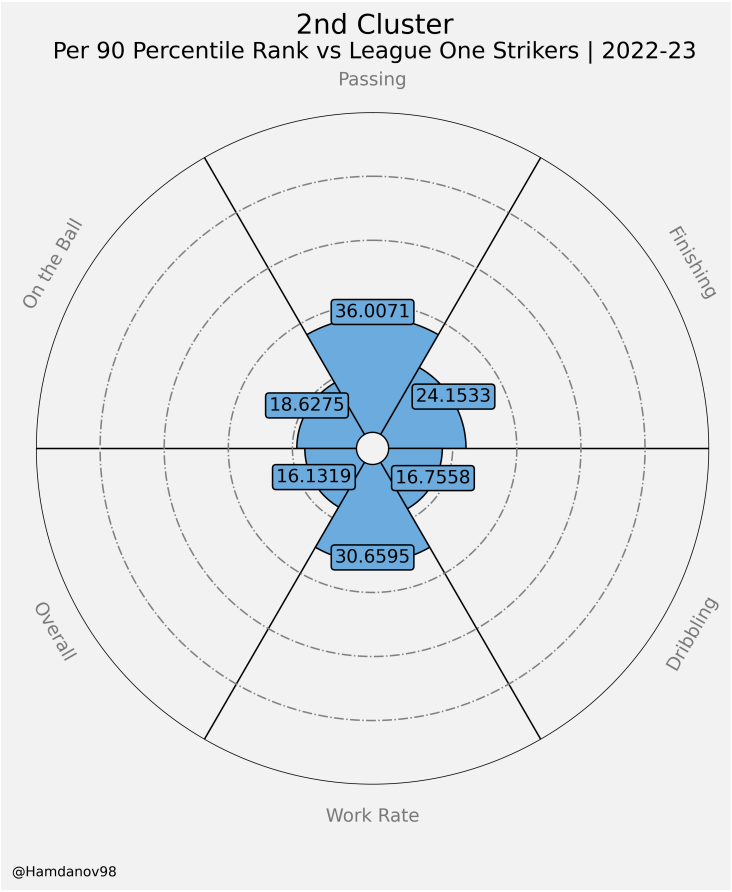

Figure 14: How an average player looks from the second cluster.

The poorest cluster of strikers overall.

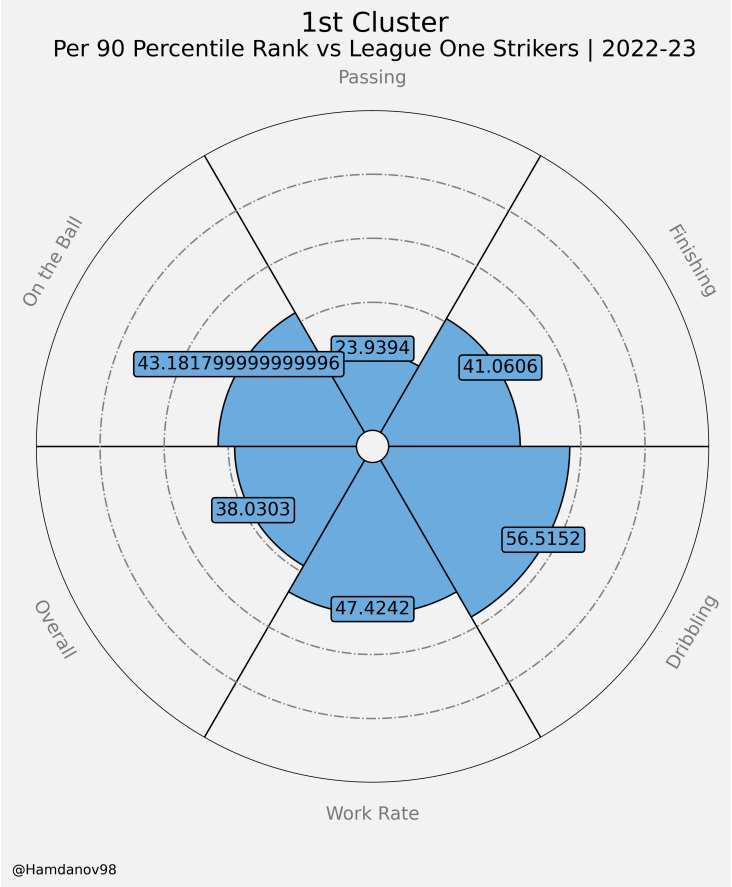

Figure 15: How an average player looks from the first cluster.

the second poorest cluster of strikers overall.

3.5 Ranking the strikers

In this step, there was confusion on how to rank the clusters, particularly who to prioritize. The third cluster had the best technical strikers, but ultimately it was decided to go with the fourth cluster based on their strong finishing feature, combined with their overall score. After ranking the clusters and reviewing the order, it became clear that this decision made the most sense. It is believed that this initial ranking can serve as a good starting point for further conditioning and analysis and an excel made based on this ranking.

4 Results and Conclusion

```
In [54]: thebest.head(5)
```

Out[54]:

	Player	Age	Overall Score	Finishing Score	Passing Score	Dribbling Score	Work Rate Score	On the Ball Score	GMM Cluster	Passing Percentile	Finishing Percentile	Dribbling Percentile	Work Rate Percentile	Overall Percentile	On the Ball Percentile
41	Player 42	29	0.518965	0.605347	0.556446	0.698381	0.215687	0.627414	4	0.954545	0.833333	1.000000	0.803030	1.000000	0.984848
15	Player 16	35	0.502542	1.215650	0.583934	0.375838	-0.165253	0.479886	4	0.969697	0.984848	0.954545	0.348485	0.984848	0.954545
10	Player 11	26	0.327904	0.499631	0.094794	0.098064	0.619128	0.096429	4	0.712121	0.803030	0.666667	0.969697	0.924242	0.742424
20	Player 21	30	0.242878	1.320941	-0.029522	-0.199404	-0.120504	-0.114463	4	0.560606	1.000000	0.333333	0.363636	0.848485	0.363636
63	Player 64	22	0.239570	0.342834	0.250600	0.116806	0.248041	0.183703	4	0.863636	0.757576	0.681818	0.848485	0.833333	0.833333

Figure 16: An overview of the top 5 strikers based on the method.

The analysis of the top strikers revealed that they mainly consisted of players in their prime or past their prime, which is a logical outcome. However, there was one young prospect who showed promise for the next three years, and a random option of a player aged 35. While this may seem surprising, advancements in sports science, fitness, and nutrition have led to players maintaining their abilities for a longer period. For instance, Inter Milan lost Lukaku, who was 28 at the time, to Chelsea in the summer of 2021 and replaced him with Dzeko, who was 35. This decision proved successful for the team, indicating the potential of older players in certain circumstances.

In [55]:
```
thebest.tail(5)
```

Out[55]:

	Player	Age	Overall Score	Finishing Score	Passing Score	Dribbling Score	Work Rate Score	On the Ball Score	GMM Cluster	Passing Percentile	Finishing Percentile	Dribbling Percentile	Work Rate Percentile	Overall Percentile	On the Ball Percentile
0	Player 1	20	-0.067648	0.105920	-0.172769	0.038429	-0.242173	-0.067170	2	0.257576	0.590909	0.606061	0.242424	0.409091	0.469697
49	Player 50	28	-0.078277	0.028708	-0.100563	0.176337	-0.417591	0.037887	2	0.393939	0.515152	0.757576	0.106061	0.378788	0.651515
21	Player 22	18	-0.083262	-0.022115	-0.258859	-0.130329	0.078257	-0.194594	2	0.090909	0.484848	0.378788	0.621212	0.363636	0.196970
45	Player 46	21	-0.118472	-0.433287	-0.337511	0.358808	-0.061897	0.010648	2	0.030303	0.227273	0.924242	0.439394	0.287879	0.590909
1	Player 2	24	-0.255050	-0.147956	-0.227142	-0.557568	-0.087533	-0.392355	2	0.181818	0.393939	0.060606	0.409091	0.196970	0.045455

Figure 17: An overview of the bottom 5 strikers based on the method.

On the other hand, the not-so-good strikers primarily consisted of younger players, which aligns with the notion of a striker being in their prime.

Overall, this method serves as a solid foundation for further analysis and conditioning based on the team's playing style and requirements. With in-depth analysis and customization, this approach has the potential to yield effective results.