INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data grouping by performance characteristics of athletes in Portugal's First Football League

João Carlos Pereira da Silva

Master in Data Science

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Rui Jorge Henriques Calado Lopes, Associate Professor, Iscte – Instituto Universitário de Lisboa

September, 2024

TECHNOLOGY
AND ARCHITECTURE

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics
Department of Information Science and Technology

Data grouping by performance characteristics of athletes in Portugal's First Football League

João Carlos Pereira da Silva

Master in Data Science

Supervisor:
Doctor Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
Doctor Rui Jorge Henriques Calado Lopes, Associate Professor, Iscte – Instituto Universitário de Lisboa

September, 2024

To my family, the sweetest in the world

Acknowledgment

I thank God, who, without my having done anything to deserve it, granted me the Grace of being born and raised in the bosom of a family that has always made every day of my life more pleasant and that makes all things possible. To my dear Father, João, for the prompt provision of all the resources necessary for my academic success. To my dear Mother, Ana, for the kind and patient task of motivating and brightening my journey with all her maternal love. To my noble grandparents Joaquim and Rosa, for the tenderness of their wise words and advice, as well as for the loving gestures with which they constantly shower me. To my dear grandmother Engrácia, who witnessed the beginning of this journey but was unable to hear on this earth the news of its conclusion.

To Professors Ricardo Ribeiro and Rui Lopes, for their valuable advice, diligent assistance, and ascetic patience, which helped make this project possible, as well as to the rest of the faculty of the Master's program.

To all my friends who closely followed my study process and who had the patience and interest to listen to me talk about this project, especially those from the *Ora et Labora* group, for the joy they brought to the final stage of this dissertation work.

To Our Lady of Fátima, Saint Joseph of Cupertino, and Saint Thomas Aquinas, my main patrons and protectors during this Master's program, for the downpour of Graces that flooded the realization of all my work.

To all, thank you very much for your presence during this phase of my young existence.

Abstract

The need to obtain useful information on any topic, easily and quickly, while utilizing an increasingly larger volume of data to support the accuracy and precision of that information, emerges as one of the great dilemmas of the young 21st century. The "King of Sports" could not remain untouched by this trend, with more and more people seeking Artificial Intelligence and Machine Learning as a means to collect, study, and propose improvements in the performance of athletes and football teams. In this work, which evaluates the performance data of players from Portuguese First Football League over three seasons, accounting for more than 1,000 observed participations, we see how these technologies contribute to providing insights into the evolution of a player's performance across different seasons. Using Principal Component Analysis, it was possible to identify some of the most distinct characteristics of various players within the framework of a simple Cartesian reference system, such as perceiving whether their offensive and possession-based style of play leans more toward direct actions and shots or whether they are more focused on ball carrying. Additionally, it was possible to understand the positional dynamics and their evolution over the seasons for some players, thus supporting conclusions about a team's collective performance and potential decisions to be made based on data science work.

KEYWORDS: *Artificial Intelligence*, *PCA*, *Performance*, *Cartesian Reference System*, *Portuguese First Football League*, *Data Science*

Resumo

A necessidade de obter informação útil sobre qualquer tema, de modo fácil e rápido, utilizando cada vez um maior volume de dados de modo a sustentar a veracidade e precisão dessa informação surge como um dos grandes dilemas do jovem século XXI. O "Desporto Rei" não poderia passar intocado por esta inclinação de cada vez mais Homens que procuram na Inteligência Artificial e na Aprendizagem Automática uma forma de recolher, estudar e propor melhorar os desempenhos dos atletas e das equipas de futebol. Neste trabalho, onde se avaliam os dados de desempenho dos jogadores da primeira liga de futebol portuguesa durante três épocas e contabilizando mais de 1000 participações observadas, vemos como estas tecnologias concorrem para fornecer conhecimento sobre a evolução do desempenho de um jogador ao longo das diferentes temporadas. Com recurso à realização de uma Análise das Componentes Principais foi possível identificar algumas das características mais distintas de vários jogadores no prisma de um simples referencial cartesiano, tais como a perceção de que o seu estilo de jogo ofensivo e de posse de bola é mais de ações diretas e remates ou se são mais atletas de transporte de bola, assim como perceber a dinâmica posicional e a sua evolução ao longo das épocas de alguns deles de modo a fundamentar conclusões sobre o desempenho coletivo de uma equipa e sobre potenciais decisões a tomar tendo como base o trabalho de ciência de dados.

Palavras Chave: *Inteligência Artificial*, *PCA*, *Desempenho*, *Referencial Cartesiano*, *Primeira liga de futebol portuguesa*, *Ciência de Dados*

Contents

List of Figures

List of Tables

List of Acronyms

AI: Artificial Intelligence
FMR: FactoMineR
ML: Machine Learning
PCA: Principal Component Analysis
DS: Data Science
SLR: Systematic Literature Review
CRISP-DM: Cross Industry Standard Process for Data Mining
YC: Yellow Card
RC: Red Card
PFL: Portuguese First League

CHAPTER 1

Introduction

The spirit of Man has been overwhelmed by the obsession with performance in all areas of our existence where some form of competition or judgment from others may exist. The steamy technological revolution has given Man more instruments to feed his hunger for better results in all areas of his life and football was, of course, invaded by the combination of these two realities. Today gorgeous amounts of performance data relating to football players and teams are extracted, stored and processed in order to constantly improve sporting results and build useful knowledge for evaluating the entire sports planning process of a club. What ML methods can football clubs use to identify their collective patterns in terms of sports performance and to cluster potential candidates to address their weaknesses?

To introduce the topic, lets begin by summarizing what has already been done to address it. According to Akhanli and Hennig [1] the index combination from calibrated average within-cluster dissimilarities, Pearson-Γ, entropy, Bootstab stability, and (with half the weight) separation may generally be good for balancing within-cluster homogeneity and "natural" separation as far as it occurs in the data in situations where for interpretative reasons useful clusters should have roughly the same size, Ven [2] concluded that by first assigning soccer players to different player types based on the outcomes of several cluster algorithms, the attributes distinguishing one soccer player from another are identified and subsequently, the presence of combinations of the resulting player types are found with the use of rare correlated pattern mining and D'Urso, Giovanni, and Vitale [3] found that a fuzzy clustering model for mixed data allows different types of variables, or attributes, to be taken into account.

In matters of directly looking for talented players, Bergkamp *et al.* [4] said that scouts value a multidimensional collection of attributes, but mostly account for general technical soccer attributes, so, this leads us to believe that it is necessary to analyze many performance attributes of a player, and with the continuous advancement of technologies and DS, future analyzes may require even more attributes. When looking at the effect that a player can have on a team, Aquino *et al.* [5] concluded that player prominence is affected by playing formation, which will be a care that will have to be taken into consideration in this paper. By analyzing and clustering team performance data, Bekkers and Dabadghao [6] found that unique styles for both teams and players can be found by clustering them based on their motif use, *i.e.* passing behaviors. When the interest is in detecting weaknesses or strengths of a team, looking at the performance of individual players Menéndez, Bello-Orgaz, and Camacho [7] classified players from different positions in such a simple

way as "weak", "medium" and "strong", as in the case of the defenders which gives us an excellent example of how is possible to detect the positions where a team may need to search for reinforcements and thus fill gaps that were causes of bad results in the past. Carpita, Ciavolino, and Pasca [8] tested various models in order to estimate the win probability of the home team, and they did it using player and team performance data. In case of need to compare the performance of a team before and after any change that merits this temporal separation, regardless of the type of changes made to the squad structure, we can use Drezner *et al.* [9] model in order to analyse, for example, the changes in the offensive style of playing in the team. Pappalardo *et al.* [10] concluded that excellent performances are rare and unevenly distributed, since a few top players produce most of the observed excellent performances, and with that base, it somehow makes us think about the need to adjust and tone a model that is capable of bringing out players with the characteristics we want to look for in a balanced and fair way, because only in this way will we be able to ensure that good player suggestions for certain collective needs are suggested by the ML model and can therefore be adapted to clubs that do not have the capacity to sign top players. The process of choosing performance variables to evaluate and statistical attributes to tone up the model, such as weights, will have to take into account in order to achieve accurate results, considering the Akhanli and Hennig [11] paper which concluded that clustering and mapping multivariate data are strongly affected by pre-processing decisions such as the choice of variables, transformation, standardisation, weighting.

1.1. Motivation

With the rapid digital and technological transition that is taking place all over the planet, humanity has found itself able to collect in small spaces and physical structures, such as a computer or smartphone, enormous amounts of information that if had been recorded in more traditional archival media, such as sheets of paper stored in filing cabinets and furniture, would have taken up an entire block of flats. Everything we do, from the most basic activities of our existence to the most complex and mellifluous, can be measured, or recorded using numbers, words or images. Today, companies all over the world store several gigabytes of information in digital format on their computers. These records, being a collection of discrete values that convey information, describing quantity, quality, facts, statistics, other basic units of meaning, or simply sequences of symbols that can be subsequently interpreted, are called data according to OECD [12]. This data can then be used, with the help of various ML and AI mechanisms, to obtain useful information capable of creating sustained knowledge about the performance of various organizations in their areas of operation, and thereby aim for better results in the future.

Professional football, which essentially thrives on the sporting results of its players and teams, is therefore a natural candidate to leverage this continuous growth of knowledge about these technologies that provide the ability to collect, store, and process performance data. This data can help clubs and their decision-making bodies evaluate, predict,

and find patterns in their performances, and by refining this information, improve their sporting results. This work has been ongoing for several years and has benefited from significant contributions over time, which have allowed for the continuous and uninterrupted development of knowledge on the application of DS to the study of the performance of football teams and players.

1.2. Limitations of previous works

The topic in discussion still lacks a more direct and simplified *modus operandi* to find ways to meet the aforementioned objectives in the field of performance analysis in football, despite all the highly significant contributions made by other authors. This theme still lacks an objective segmentation tool, in programming code, that allows for faster extraction of information from a performance data set in order to meet the needs posed by each of the different queries that may be formulated. The goal is to obtain figures with an objective and easily interpretable qualitative evaluation of the performances of players and teams based on different performance variables, alongside categorical variables such as the player's position, club, or the sports season in consideration. Essentially, what is needed is the creation of a modest graphic representation that summarizes the entire story of a sports season, making, so to speak, an image worth more than a thousand words that could be used to describe a particular time period in the sports context.

I propose to focus the object of this study on the Portuguese First League (PFL), with the aim of providing a more detailed analysis of our own football, which, in previous works, was not given special emphasis that would allow for a deeper understanding of our league. By concentrating the research on our championship, certain insights and patterns that might be overlooked in a more generalized and international analysis could be revealed and examined. This approach may lead to acquiring a range of knowledge about our tournament that helps better understand its characteristics and the fundamentals of its uniqueness.

1.3. Research questions and goals

The main research questions to answer are:

First: Is it possible to create a tool, using a programming language, that can quickly and easily select useful information to segment categorical and numerical performance variables of football players or a team?

Second: Can we develop a model capable of accurately evaluating, from a qualitative point of view, the performance of players and thereby selecting the best athletes for specific positions or "types" of player?

1.4. Methodology

The methodology used in the development of the project was CRISP-DM (Cross Industry Standard Process for Data Mining). This methodology is the most suitable for solving this DS problem since, recalling the initial research question, it aims to reduce the required time until the visualization of results while simultaneously decreasing the methodological

and scientific complexity of the data engineering process until the effective acquisition of useful knowledge. With large volumes of data, and addressing the second research question, it is possible through this methodology to aim for more precise results with greater statistical grounding by building a model robust to environmental changes so that it can analyze, with the same precision and multiple times over different periods, various sports seasons and different championships. The CRISP-DM methodology has the following steps: 1) Business understanding, 2) Data understanding, 3) Data preparation, 4) Modeling, 5) Evaluation, 6) Deployment. In the subsequent sections, we will discuss in more detail some of the sub-tasks of this methodology that were particularly important in this work, such as the initial data acquisition, data cleaning, data construction, model building, and the evaluation and review of the process.

1.5. Document structure

The structure of the document is organized to consistently and chronologically tell the story of the execution of this work, considering the justification of all steps, tasks, and corrections made during the development of the solution code that enabled the creation of the model. The ML tool used is described in detail, along with the aforementioned points in the previous section regarding the methodology. Additionally, there is a discussion of the results and an evaluation of the model's applicability to the current football landscape. The limitations of this work are also discussed, along with suggestions for improvements that can be applied to future work exploring this topic. Attachments will also be provided with graphical visualizations of the obtained results, as well as a glossary to aid in the interpretation of the statistical variables used in the dataframes and brief explanations of them.

CHAPTER 2

Systematic literature review

The objective of conducting a SLR on the clustering of professional football player performance data is to summarize all studies that are similar to the dissertation's goals. The dissertation developed a methodology capable of clustering players in a manner almost similar to what Pappalardo *et al.* [10], D'Urso, Giovanni, and Vitale [3], and Ven [2] accomplished. Additionally, using clustering models like those employed by Bekkers and Dabadghao [6] and Menéndez, Bello-Orgaz, and Camacho [7], the goal is to identify performance aspects that may characterize players or a team and then trying to discuss how this can improve sporting success considering the study by Carpita, Ciavolino, and Pasca [8].

2.1. Historical Perspective

In a pioneering study in 2004, after analyzing data from various editions of the Brazilian football championship, it was found to be possible to use considerable volumes of data to discover interesting and unexpected patterns in the temporal evolution of the competition [13]. In the analysis of collective results from major international competitions, such as the 2010 FIFA World Cup, these technologies made it possible to determine that the results of the last three matches of the Spanish national team indicated that the clustering coefficient of the passing network increased over time and remained high, indicating Spanish players' ball possession. This possession eventually led to victory, even as the density of the passing network decreased over time [14]. The prolonged ball possession style became a hallmark of Spain in that competition. Furthermore, looking at the performance of individual players in that competition, Menéndez, Bello-Orgaz, and Camacho [7] classified players from different positions in a simple manner as "weak," "average," and "strong," such as in the case of defenders. In line with the parameters explored in these papers, we can already introduce the possibility of creating graphics that allow us to draw conclusions about the history of a football season. In 2012, a novel approach was introduced that attempted to evaluate players using a simple score, irrespective of their playing characteristics, based on their contributions to winning performances [15]. The study of this topic must take into consideration that the clustering and mapping of multivariate data are strongly affected by preprocessing decisions, such as variable selection, transformation, standardization, and weighting [11]. The study of subsequent World Cups has added evidence that success in this competition is associated with dense passing networks, high frequency, and high clustering coefficients [16]. Additionally, the prominence of a player is affected by the team's formation [5]. Considering individual player performance and talent scouting,

scouts value a multidimensional collection of attributes, but primarily account for general technical football skills [4]. They also seek to predict future performances based on present exhibitions [17].Using clustering methods, unique passing network styles of teams and players can be identified by grouping them based on passing behaviors as issued by Bekkers and Dabadghao [6]. These methods also allow us to find players with unique characteristics in this specific attribute, such as Peña and Navarro [18] issued. Considering the need to predict sporting success, it is possible to do so, in the case of the visiting team, using both player performance data and team performance data as done by Carpita, Ciavolino, and Pasca [8]. For direct evaluation of changes in team performance over time, we can use models that have been used to find differences between different teams in the same time frame. An example is the detection of significant differences between Barcelona and Manchester United in incomplete penetration dynamics in the defensive field in long passes in incomplete penetration dynamics in the defensive field and in back-passing in incomplete penetration dynamics in the offensive field as supported by Drezner *et al.* [9]. Moving away from the study of ball possession models, counter-attack methods have also been extensively studied. We now know that if a club creates a goal-scoring opportunity within the first 15 seconds after recovering the ball, the probability of scoring a goal or taking a shot is higher [19].

There's also been found evidence that excellent player performances are rare and unevenly distributed, as some top players produce most of the outstanding performances observed [10]. This investigation was able to group players according to their field positions, ranking them from the best to the worst. The application of such techniques in football still lacks more examples, but in other sports, like basketball, Kosmidis and Karlis [20] managed to cluster NBA players based on their performance using finite mixture models combined with flexible clustering methods. With the intention of estimating a player's market value, this is already achievable with an accuracy of 0.74 [21]. Although using different types of data from those we intend to work with, a technique has been developed to model athlete performance to automatically assign the tested athlete to a group of similar athletes regarding physical parameters and test development. This technique evaluates these groups concerning two performance quality indices [22]. Additionally, considering the need to link sports results with the club's financial area, a positive bidirectional relationship was found between corporate performance and sports performance, and a unilateral inverse relationship where financial performance affects sports performance, as stated by Galariotis, Germain, and Zopounidis [23]. There are several clustering methods that can be used based on player evaluation to achieve the objectives of this work. Supporting ourselves on existing literature, we know that a comparison of all these methods has already been conducted based on the behavior of various athletes as affirmed by Drachen *et al.* [24]. Based on the literature review by Ven [2], it is concluded that the most popular form of clustering in both scientific and industrial terms is K-means as previous defended by Rai and Shubha [25].

Advancing in time and using more complex methodologies, some difficulties presented at the beginning of the thematic study have been addressed. Ven [2] managed, by first assigning football players to different player stereotypes based on the results of various clustering algorithms, to identify the attributes that distinguish one football player from another. Subsequently, the presence of combinations of the resulting player types was found using rare correlated pattern mining. D'Urso, Giovanni, and Vitale [3] demonstrated that different types of variables can be used in a clustering model while remaining robust to outliers. Akhanli and Hennig [1] discussed adjusting weights to improve the balance between homogeneity and natural separation between two clusters of different players. Clustering players based on their characteristics and using performance data led Li *et al.* [26] to discover 18 different playing styles in players based on their characteristics using six performance indicators. They also compared the same player in different years. In analyzing the influence of players in their respective teams, Lionel Messi was identified as an extraordinary player due to the high number of actions per game, with a very good qualitative evaluation of these actions as found by Decroos *et al.* [27]. In the evaluation and comparison of players based on nationality and position, Gai *et al.* [28] discovered significant differences in performance and physical indicators between domestic and foreign players in the Chinese Super League.

2.2. Searching Methodology

The selection of literature for the systematic review was partly guided by the PRISMA guidelines of Moher *et al.* [29]. The search for articles was conducted in three phases:

First Phase: Keywords were used in quotation marks to find articles corresponding to the method intended for the dissertation and the type of data. At this stage, there was no defined rationale.

Second Phase: A phrase-based search was conducted without any rationale, which resulted in an excessive number of results, making it less effective than the first phase.

Third Phase: A search was performed with a defined rationale. The dimensions used in the Mendeley database search engine in this phase were as follows: ("Technique" + "Domain" + "Focus/Objective"). In the "Technique" dimension, the search focused on Clustering. In the "Domain" dimension, the search focused on the words Football and Soccer, using the "OR" operator in some cases to broaden the search due to the different terminology used to refer to the sport in Europe and North America. The "Focus/Objective" dimension concentrated on the data intended for study, specifically Team Data, Player Data, Performance, and Performance Data. Although the research rationale was not yet fully clarified during the first search phase, it still allowed for the identification of articles to be included in this review. Due to a lack of articles that met the selection criteria, the second phase involved a search without any defined rationale, which also included the use of Google Scholar. It became evident that this approach was not yielding the desired results, as the number of articles retrieved from that database was extremely high, making it impossible to review all the titles. Additionally, most of

the articles that appeared were not related to the search phrase. The first two search phases ended up being crucial in a learning-from-error approach, as they demonstrated that only with a properly defined rationale, supported by the experience of other authors who have done similar work, is it possible to more effectively and systematically find documents that align with the research objectives. This approach ensures that the results are systematic and well-founded, unlike what occurred in the initial phases. Despite the continuous improvements made to the search methods, duplicate articles were not accounted for. The selection of articles included the following steps:

Initial Screening: Articles were first reviewed by reading their abstracts.

In-Depth Review: Those articles that passed the initial screening were then read beyond just the title and abstract.

Table 2.1 presents the selection and rejection criteria used during these two filtering phases.

TABLE 2.1. Selection and rejection criteria

Selection Criteria	Rejection Criteria
The terms used in the search are mentioned in the title, abstract, and body of the article	The reference is not published or associated with any evidently academic institution or journal
One of the article's elements is to provide mechanisms for evaluating player and/or team performance	Articles that work with simulated data
One of the article's objectives is to relate the clustering of players to the team	Articles exclusively focused on athletes' physical data, such as weight, height, or heart rate measurement
The terms highlighted in the search results are associated with football championships	Articles that address individual sports disciplines
	Articles exclusively focused on training session programming
	Articles that do not discuss performance data
	Articles exclusively focused on studying athletes' psychology
	Articles exclusively focused on the analysis of non-sporting objectives (e.g., gender equality or abuse in sports)

The choice of selection criteria is based on the objective and natural need to gather, study, and summarize the greatest amount of relevant work, similar to what is intended for the master's dissertation, and that has been conducted to date. The first selection criterion naturally arises as a response to the need to restrict the articles to those that align with those desired to be applied, the sport to be studied, and the type of data that is the focus of the study. The second and third criteria arise out of the necessity to select articles that provide methodologies capable of serving as examples for what we

intend to do. The fourth criterion serves as a filtering mechanism due to the vast array of competition types and sports disciplines studied, as articles found in search engines often focused on data obtained from training and sports psychology studies.

The rejection criteria, in turn, primarily arise as a necessity to narrow down the articles in the literature review to sets of knowledge that align with the proposed objectives, thus rejecting documents that do not meet the dissertation's focus. The first rejection criterion is almost obvious, given that the literature review aims to support a master's dissertation, which must be academic in nature and therefore should be supported by references of that same academic character. The second and third criteria aim to make the nature of the data as similar as possible to those intended for use in the dissertation. The fourth criterion helps to eliminate issues related to the difference between comparing individual datasets that produce a collective result and other similar datasets. The subsequent rejection criteria are explained by the need to eliminate studies that aim to use data to find patterns different from those sought, e.g., for sports psychology studies, highly complex tactical and technical training issues in football, or gender equality issues.

Figure 2.1 will show us the summary of article selection.

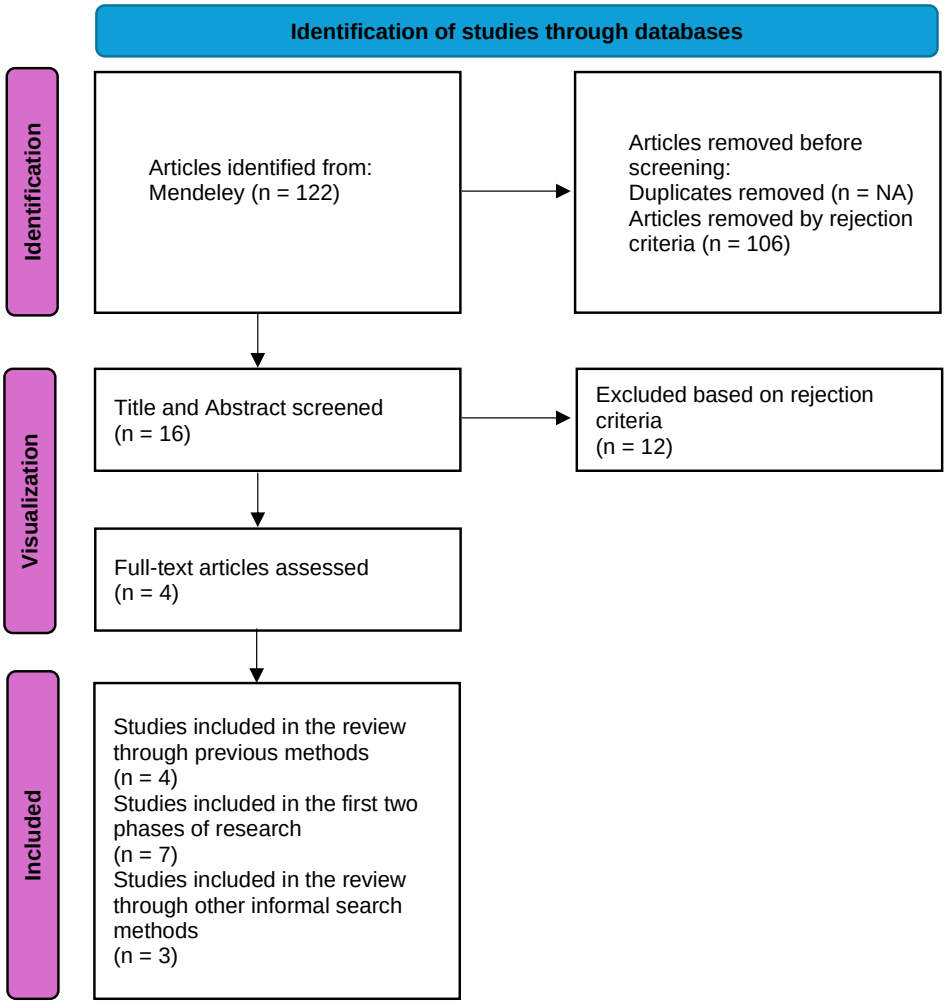

FIGURE 2.1. PRISMA process summary

The other informal search methods that led to the discovery of 3 articles included finding them through the references of other articles that were reviewed.

The exclusion criteria used to remove articles were mainly numbers 3 and 6 from Table 2.1, without counting the number of articles for each criterion. The choice to exclusively use Mendeley and not other more traditional databases like Scopus and B-on was due to the fact that terms like "Soccer" and "Football" are very broad and diverse, leading to a large number of articles in larger and more diverse databases at a time when the rationale was not yet clearly defined. As a result, the focus was eventually dedicated entirely to Mendeley, as it had already proven effective in locating highly relevant articles for this literature review.

In the descriptive analysis of the studies, we see that all the included articles are from after 2010, and the overwhelming majority were published after 2015. There is also a predominance of European authors and the use of clustering methods to find patterns. Additionally, some studies use more than one method to extract useful information from the raw data that was analyzed in order to determine which method might offer better accuracy and reliability.

The study conducted by Bergkamp *et al.* [4] introduces us to a range of traditional methods used by scouts to discover future talents, which helps us compare these methods with the ML mechanisms we use to attempt this task. However, this study focuses more on predicting future player performances based on their current techniques during adolescence. In the study of international competitions like the FIFA World Cup, we observe how the capacity for data collection has increased from 2010 with Menéndez, Bello-Orgaz, and Camacho [7] to the 2018 World Cup in Russia with Aquino *et al.* [5] that collected and stored more complex and voluminous data—such as the success and destination of all pass attempts made by a team in various football matches—more sophisticated information can be analyzed beyond the traditional performance records that mostly focused on goals, assists, saves, and disciplinary actions. For example, Drezner *et al.* [9] were able to analyze games from a season where such data collection was not yet possible and provide a fascinating profile of the differences in playing styles between Manchester United and Barcelona in 2009. Based on this study, it would be possible to compare, on many levels, teams such as Sporting CP that won the 1964 Cup Winners' Cup with the same club's team that recorded its worst-ever performance in the Portuguese Championship in 2013 by finishing in 7th place. Decroos *et al.* [27] offers an interesting model for identifying players with very specific characteristics, but the reliance on data from unconventional sources could limit future work aimed at developing and testing the model for other purposes, especially if access to such a wide range of information proves difficult. Ven [2] and Bekkers and Dabadghao [6], by creating various clusters for different player positions, place themselves in a more advantageous position compared to Li *et al.* [26] This approach provides more concrete and valuable information for the dissertation's objectives as it addresses the need to segment players by position styles while considering

a multitude of characteristics. These characteristics can later take into account the work of Carpita, Ciavolino, and Pasca [8] to predict whether integrating certain players into a team, based on the studies by Aquino *et al.* [5] and Drezner *et al.* [9], could improve sports results. However, the work of Gai *et al.* [28], when compared to similar studies, may be less relevant due to the discrepancy in physical indices and quality between the Chinese league and the major European leagues. This difference can affect the preprocessing results, as analyzed by Akhanli and Hennig [11], although their work does not directly address the allocation of these articles. Pappalardo *et al.* [10] stands out as a more robust example for understanding how to better categorize players. Akhanli and Hennig [1] and D'Urso, Giovanni, and Vitale [3] finally demonstrate effective methods to avoid some of the main pitfalls in clustering football player performance data. Nonetheless, Decroos *et al.* [27] illustrates that it is possible to identify truly exceptional players who do not constitute an outlier due to unjustifiable reasons, such as having participated in only a few minutes of play and achieving a significant number of successful actions during those minutes, thereby strongly influencing an average.

Another goal of the dissertation is to combine a set of variables that can provide a broad and clear understanding of an effective scout player request tool, relying primarily on the methodologies proven effective by Akhanli and Hennig [11] and D'Urso, Giovanni, and Vitale [3], and adopting a clustering organization approach similar to that of Ven [2].

2.3. Future Work Based on the Literature Review

The ability to gather large volumes of information about player and team performance in digital formats has paved the way for the application of advanced data analysis techniques and AI. In this context, the systematic review of the literature on clustering professional football players' performance data emerges as a valuable tool for understanding the progress, challenges, and potential advancements in this field. Over the years, pioneering studies have explored the application of clustering techniques in performance data, revealing interesting and unexpected patterns in football competitions.

The systematic review reveals an increasing integration of data and more complex methods over time. Recent studies explore the association between different player stereotypes, identifying distinctive attributes such as expected goals, expected assists, and even discovering which passing connections between specific players most effectively translate into sporting success for a team. In future work, it would be interesting to consider how this information could be used to suggest a concrete strategy for a team that is noticeably inferior to its opponent, potentially increasing its chances of winning a match by studying the most successful passing combinations. The idea of creating a concrete decision-support model for sports managers at clubs—particularly one that can automatically detect where a team is underperforming in terms of positions and performance attributes, and then suggest a group of players with the necessary characteristics to fill those gaps in alignment with the club's financial capacity—emerges as a philosophy that has yet to be applied. This model could assist in player grouping, result estimation, and qualitative evaluation

of individual and collective performances. The main limitations of this potential future work include the challenge of combining economic data with specific data oriented towards sporting results, as well as the inherent difficulty in predicting certain collective performances. This is especially true when considering the impact of a player transferring from one club to another to fill a gap, with a prediction accuracy exceeding 60%. Even without considering this nuance, it is difficult to make precise predictions, as the vast majority of result estimation models struggle when it comes to close games or highly competitive tournaments.

In conclusion, although significant progress has been made in grouping players based on their performance data or characterizing specific attributes, there is still a need to develop a method specifically focused and oriented towards supporting decision-making in player purchases. Such a method should be capable of enhancing the basis for decision-making, thereby reducing the risk of making purchases that could prove detrimental to a club. Nevertheless, the proposed method goal isn't to combine performance variables with economic variables, such as a player's market value, within the same study. Additionally, it shouldn't be able to incorporate the idea of predicting a player's sporting success at a club they are not currently part of, and how this could potentially improve the team's collective results, with the course of the investigation leading to the discovery of another type of useful information.

Study	Methods	Data Source	Data Type	Data Volume	Main Annotation
D'urso et al (2022)	Robust fuzzy clustering model	Whoscored	Performance and characterization data	397 players from Serie A	Discrepant values neutralized
Akhanli & Henning (2023)	Clustering	Whoscored	Performance data	1501 football players from 8 leagues and 107 variables	Balancing and homogeneity between two different clusters
Van de Ven (2018)	Clustering (K-means; Sparse PCA and Archetypal Analysis); Correlated Pattern Mining and Result Estimation	Kaggle	Performance data	10,000 players, 25,000 matches in 11 countries	Identification of attributes that distinguish players
Akhanli & Henning (2017)	Clustering and multidimensional scaling	Whoscored	Performance data	3152 players and 107 variables	Clustering strongly influenced by data preprocessing
Pappalardo et al (2019)	Clustering	Whoscored	Performance data	64 seasons and more than 31 million events during matches	Create a qualitative scale of players
Menéndez et al (2013)	Clustering	FIFA	Performance data	75 variables of players from the 2010 World Cup	Create a qualitative scale of teams/national teams
Aquino et al (2019)	Linear regression; Pearson correlation and Cohen's d	FIFA	Performance data	988 observations of players from the 2018 World Cup	Study of the preponderance of a player on the team
Carpita et al (2019)	Binomial logistic regression (Random Forest; Neural Networks; K-means and Naive Bayes)	Kaggle	Characterization data	10 European leagues over 7 sports seasons	Estimate the probability of the home team winning the game
Bekkers & Dabadghao (2019)	Clustering	Squawka	Performance data	8219 matches, 7412 players and 466 teams	Find unique playing styles for players and teams
Drezner et al (2020)	Finite state machine (FSM) software	Video collection	Performance data	4 matches from the 2008/2009 Champions League	Find important differences between teams
Bergkamp et al (2022)	Clustering	Royal Dutch Football Association	Performance and characterization data	125 questionnaires completed by scouts	Discover how scouts find talents
Gai et al (2019)	Clustering	Amisco Sports Analysis Services & Opta Sportdata	Performance data	240 matches from 16 teams in the Chinese championship	Player comparison
Li et al (2022)	Player vector structure	Whoscored	Performance data	960 matches from the Chinese championship	Find 16 different player styles with just 6 variables
Decroos et al (2019)	Action valuation structure through probability estimation	Opta, Wyscout, STATS, Second Spectrum, SciSports, and StatsBomb	Performance data	21 types of football actions	Find truly differentiated players

TABLE 2.2. SLR summary

CHAPTER 3

Materials and methods

3.1. Data understanding

The chosen data source for obtaining the necessary information for this dissertation is the FBREF website: https://fbref.com/pt/. It is a page of raw football statistics that analyzes both teams and players, as well as matches. The data is provided to FBREF by Data Sports Group, and the more advanced statistics are supplied by the OPTA group. First, It was verified whether the data source in question allowed the extraction of statistical information about players and teams via web scraping by adding the "robots.txt" command to the website's URL. On https://fbref.com/robots.txt, was confirmed that data extraction through the aforementioned process was not prohibited by the site's administrators. The seasons analysed are the 2021/2022, 2022/2023 and 2023/2024 ones of the PFL.

3.1.1. Data gathering

Using the "reports" library, the web scraping process begin through HTTP requests in order to access the web page, also establishing "headers" that allowed to provide information to the server about the device that is making the request. The request report was printed to confirm the success of the effort.

Using the beautifulsoup it parse HTML content and set the URL for the PFL statistics of the selected season. With help from requests.get with a custom User-Agent header (simulates a browser) to retrieve the webpage content, it checks if the response status code is 200 (indicating successful retrieval), if successful it creates a BeautifulSoup object (soup) to parse the downloaded HTML content, and saves the HTML content to a local file to be useful for later analysis. Next it open the selected file in read mode with UTF-8 encoding (to handle potential character encoding issues) and it reads the entire content of the file into a variable named text and closes the file in order to release resources. Next, it remove duplicates if present, and it prepares them for potential further processing by suggesting filenames and printing the list of links themselves. Next it retrieve information about the number of teams listed in the downloaded HTML and potentially extracts the ID of the first team (depending on the website's URL structure) and downloads webpages from a list of URLs and saves them as individual HTML files, while also providing basic error handling and a delay to avoid overloading the server.

To effectively extract information in order to save it on the computer, the code was divided into three phases: 1) extraction of tables, 2) structuring of the data frame and 3) refinement of information.

Nine out of the twelve tables available on the website were extracted. The table related to the team's schedule and match results for the season was not extracted because it lacked practical relevance to the dissertation's objectives, as described above. The tables containing goalkeeper performance data were not extracted either. This decision was made because analyzing all players from a macro-perspective could lead to an inaccurate assessment of the goalkeeper's qualitative performance and influence. The goalkeeper's data performance is heavily influenced by factors beyond their control, such as the defenders' performance and the overall team's defensive behavior, which might expose them to constant offensive threats. Ultimately, the evaluation of these tables would not produce a pattern as relevant for assessing the model's response to the research questions due to the significant differences between this position and any other field player. The specific variables for goalkeepers require a certain level of specificity to accurately interpret their performance.

In order to structure the final table, it was used the 'merge' command, which allowed it to combine all the tables into a single dataframe for each team. Along the way, it was necessary to systematically eliminate the variables 'Pos,' 'Matches,' 'Nation,' 'Age,' and '90s' because these were common to all the tables and would overlap during the merging process. At one point, it was also necessary to remove the variable "Playing Time_90s" as it was repeated in only some cases. The concatenation function was not used for this operation because it was duplicating the rows related to the players, placing the next table in the rows below whenever there was a column with the same name in different tables, even though it might refer to a different observation.

To refine the final table, it was firstly removed the '_x' and '_y' suffixes from the column names in the DataFrame to clean up the automatically generated column names, and also the rows corresponding to goalkeepers, as it didn't make much sense to evaluate their outfield data. Additionally, it was removed the 'matches' column, as it was common to all the tables and only served to contain the link to the games related to that player on the website, and also the lowercase portion of the text from the 'nation' column to simplify its name, and was selected only the first part of the 'Pos' column to streamline the player's position, as some players had two positions listed in the corresponding cell. So it ended up selecting the first one randomly. Next, it were removed the columns containing the '%' character to eliminate all success or failure rates since the data source in question does not account for the number of attempts made for a particular action on the field. For instance, if a player succeeded in their only attempt, it would be classified as '100%', potentially leading to issues with severe outliers. Afterward, it was necessary to remove the columns related to 'minutes per substitution,' 'minutes per game started,' and 'average shot distance' because they contained numerous missing values and were not of significant interest for the study. Finally, was also removed the duplicate columns, which remained due to having different column names despite representing the same values and evaluating the same actions in the athletes. This occurred because the tables had to

be extracted in a way that considered a certain hierarchy, with two groups of columns. Not following this approach would have led to the opposite problem, where tables that were not actually duplicates would be treated as such due to having identical names but referring to different objects. Doing it this way made it easier and more automated to handle potential duplicate columns. The issue with the table names led to this problem because they were either extracted based solely on the second level of the table name hierarchy, ignoring the upper level, or the extraction had to be done by splitting the variable name with an '_'. This placed the upper class level on the left and the lower level on the right, making it faster to identify and remove duplicates. This approach helped avoid mistakenly treating columns as duplicates when they were not, particularly in the case of the 'passing' table. Finally, duplicates were removed broadly using a specific code command designed for this purpose, and specifically columns whose names ended with '_y' in order to correct the problem with the merge operation where columns with the same name but different origins are suffixed with '_x' and '_y', so that the first line of code mentioned to address this issue did not yet take into account the 'merge' that was performed in an intermediate phase of this process. In the end, the dataframes were saved for the 18 teams from the 3 seasons under study individually on my computer. This was achieved through a Python loop in the code, which automatically extracted the tables in a cyclical manner, opening each team's data one after the other.

3.2. Data transformation

During the data transformation phase, an error was detected early on in the formatting of the tables, related to the hierarchical levels of the columns. In some tables, there was a discrepancy where, in certain dataframes, the column 'MP_x' appeared as 'Playing Time MP_x' because the first hierarchical level of columns was empty in some tables, while in others, the first level 'Playing Time' extended to this variable. The problem stemmed from the data source itself. To resolve this, It was chosen to rename the 'Playing Time MP_x' columns to 'MP_x' across all cases, as the latter was more common among the columns involved, thus requiring fewer renaming processes. Next, it was addresseded the issue of missing values by first identifying where these values were located. By summing the missing values by row, was found that some players had most of their performance data missing, such as Maga from Vitória de Guimarães in the 2021/22 and 2022/23 seasons and Danny Loader from FC Porto in 2021/22. Then a list of the 18 teams was made in the code and applied more detailed transformations to the datasets using a loop, allowing for the automatic editing of the data for all teams. Continuing the process, it were inserted a '-1' in the previously identified missing values in the three columns 'Standard_G/Sh,' 'Standard_G/SoT,' and 'Expected_npxG/Sh.' These columns did not have values due to the players in question not attempting any shots. By assigning a negative value, it reflects the negative connotation associated with the absence of such offensive actions.

Afterward, it was decided to remove the column containing the players' nationalities, as it was not relevant to the study's objectives and was difficult to quantify. Subsequently,

it was created a new row at the end of the table that contained a summary of all the team's data, allowing for the comparison of various teams. It began by summing the various variables and in the first two columns, which are categorical variables and therefore cannot be summed, It was filled the null values left by the previous summation. In the first column, it was written the term 'Referência' to indicate that it represents a global pattern for the entire team. In the second space, it was written the name of the respective club, making it easier to interpret the table as a whole, particularly this newly created summary row.

Next, It was altered the results of some variables, as summing them didn't make sense. In these cases, applying other types of metrics would be more valuable for gaining insights from the data. For the 'age' variable, it was chosen to calculate an average to gain a more standardized understanding of this variable for the team as a whole, as the average age provides a more intuitive sense of whether a team is aging or rejuvenating. For the 'MP_x' variable, it was decided to use the maximum value among the respective players to better assess the consistency in player utilization by the team. This approach, combined with other similarly significant variables, avoids the issue where summing the values would only indicate which team benefited from longer stoppage times during games—a detail not particularly relevant to the study's objectives. The same reasoning was applied to the variables 'Playing Time_Mn/MP' and 'Playing Time_Starts,' so that, by analyzing these variables together, the precision of insights regarding the team's consistency in utilizing its players could be improved. For the 'Playing Time_Min_x' variable, it was used the average, given that the range of values is significantly larger. For the 'Playing Time_90s' variable, It was also chosen the average, since this variable is derived from the first and provides a higher level of specificity regarding the consistency of player usage. This makes it more interesting from a research perspective to obtain the average, especially since the 'MP_x' variable is very similar, and we've already used the maximum value for that one. For the 'Team Success_PPM' variable, it was also calculated the average because this variable is derived from a ratio. Taking the average is more appropriate to mitigate the impact of outliers that could arise from using the maximum value, especially in cases where both the numerator and the denominator are very low. This issue could occur when the initial statistic is influenced by the limited and possibly fortunate or unfortunate usage of a player over a certain period. For the variables 'Team Success_onG,' 'Team Success_onGA,' 'Team Success_+/-,' 'Team Success (xG)_onxG,' 'Team Success (xG)_onxGA,' and 'Team Success (xG)_xG+/-,' it was chosen to use the maximum value. For the first three, this approach provides greater clarity regarding the goals scored and conceded by the team, in conjunction with the participation of the team's most utilized player. This allows for a contrast with the general team data without considering that nuance. For the remaining variables, the decision was primarily driven by the need to standardize and simplify the data in this row, as wasn't found a clear

or intuitive practical interest in making these variables a significant factor in the overall team evaluation.

For the variables where the summation metric was retained, it was necessary to use conditional sums in some cases to avoid data distortion caused by outliers.For the variables 'Standard_Sh/90,' 'Standard_SoT/90,' 'Per 90 Minutes_Gls,' 'Per 90 Minutes_Ast,' 'Per 90 Minutes_G+A,' 'Per 90 Minutes_G-PK,' 'Per 90 Minutes_G+A-PK,' 'Per 90 Minutes_xG,' 'Per 90 Minutes_xAG,' 'Per 90 Minutes_xG+xAG,' 'Per 90 Minutes_npxG,' 'Per 90 Minutes_npxG+xAG,' 'SCA_SCA90,' 'GCA_GCA90,' 'Team Success_+/-90,' 'Team Success_On-Off,' 'Team Success (xG)_xG+/-90,' and 'Team Success (xG)_On-Off,' was used a conditional sum that only includes players whose ratio of minutes played per 90 is greater than 8.5. This decision was made because these variables are themselves the result of a ratio that uses 90 as the divisor, so this threshold was selected as the condition for summation to maintain data consistency. The advantage of this approach is that it excludes players who frequently enter games late, even if they play a large number of matches, as they are more likely to generate outliers in these variables. The reference value of 8.5 was chosen because it represents a quarter of 34, the total number of matches in the league. In summary, this method allows for the inclusion of players who joined or left the team during the winter transfer window if they had a reasonable level of participation, while excluding those who, despite being on the roster all season, contributed in a way that might be undervalued by these statistics and could disproportionately influence the team's overall performance in these variables. Choosing a quarter of the matches allows for consideration of mid-season transfers, which would not be possible if, for example, half of the games were used as the threshold. Conversely, selecting a fifth or less might have included players prone to generating outliers. For the variables 'Standard_G/Sh,' 'Standard_G/SoT,' and 'Expected_npxG/Sh,' was applied the same rationale. Since these variables are ratios based on shots taken and shots on target, where these values are used as divisors, it was set a condition that the player must have attempted more than 3 shots for 'Standard_G/Sh' and 'Expected_npxG/Sh' and more than 3 shots on target for 'Standard_G/SoT.' This was done to avoid the possibility of a player scoring a goal from a single shot, which could create a severe outlier and distort the reference norm for the team's overall data. All other variables were left with the complete summation of the values from the respective players on each team. This approach was appropriate because these variables represented statistics where the cumulative total reflected the overall objective actions performed by the entire team, making it the best summary for these metrics at the collective level.

After a new analysis of the value and meaning of each column, it was realized that some columns recorded actions that were not necessarily beneficial for evaluating the performance of the player and the team. Therefore, it was assigned negative values to the columns 'Performance_CrdY_x,' 'Performance_2CrdY,' 'Performance_CrdR_x,' 'Outcomes_Off,' 'Outcomes_Blocks,' 'Challenges_Lost,' 'Err,' 'Carries_Mis,' 'Take.Ons_Tkld,'

'Carries_Dis,' 'Performance_PKcon,' 'Performance_OG,' 'Performance_Fls,' 'Performance_Off,' and 'Aerial.Duels_Lost,' as the accumulation of these types of actions by players and teams indicated a negative aspect of their performance. Displaying these values as negatives makes it easier to understand the overall performance evaluation.

After completing the general data transformation for each of the dataframes representing the 18 teams from the three editions of the Portuguese championship, it was decided to create a summary dataframe for each of these three editions. In each of these dataframes, all the data from every team and player who participated in that particular edition were included. It was made by concatenating the 18 team tables vertically, ensuring that they were aligned sequentially. Using the same method, it was also created a single 'super' dataframe that contained the accumulated information from all three seasons. To do this, it was added a new column to each of the preceding dataframes to indicate which season the information in each row belonged to.

3.2.1. Exploratory Data Analysis

The exploratory data analysis revealed some interesting insights about the players and teams. It is notable, in figures B.1, B.2 and B.3 how the Famalicão team consistently maintains a young team identity across these three seasons, while teams like FC Porto and Arouca seem to lean more towards experience. In terms of top scorers, plot B.4 Viktor Gyokeres stands out clearly from the other players of the 2023/2024 season as well as from top scorers of other seasons. Also noteworthy is the performance of Taremi, who, with records from two distinct seasons, managed to appear in the TOP10 of the top scorers over these three years. When combining goals with assists, in plot B.5 Gyokeres again stands out from all others, with Taremi making two appearances in the Top5 with performances from two different seasons. Regarding expected goals by teams, plot B.6 SL Benfica appears in the Top5 with records from all three evaluated seasons. FC Porto occupies the other two positions in 21/22 and 22/23, while Vitória de Guimarães in 21/22 is the only team outside the consistent Top4 of the league standings over the three editions to surpass the 50 expected goals mark in a season.

3.3. Principal component analysis

3.3.1. Theory and foundation

The Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data analysis that consists in transforming a dataset with potentially correlated variables into a set of values of uncorrelated variables called principal components. These components are linear combinations of the original variables, ordered such that the first few retain most of the variation present in the original dataset. By focusing on these principal components, PCA simplifies the complexity of high-dimensional data while preserving as much variance as possible [30]. The percentage of explained variance is a key concept in PCA, providing a measure of how much information (or variance) from the original dataset is captured by each principal component. When PCA is performed,

each principal component has an associated eigenvalue, which represents the amount of variance captured by that component [30].

The explication of variance percentage enables the understanding of which variables contribute most to the variance in the data and can guide decisions on which features are most relevant for further analysis or predictive modeling [31]. Considering the use of the FMR library, which allows for performing PCA and by examining the explained variance, make informed decisions about how many principal components to retain for further analysis or modeling, FMR simplifies the process of conducting PCA and interpreting the results, making it accessible even for those who are new to multivariate analysis. It is quick and, in line with the research questions, allows for obtaining easily interpretable graphs that provide concise and objective information about the subjects under evaluation. This tool enables a broad, valuable overview of the characteristics of some players from an offensive or defensive perspective and also draws important conclusions about the collective performance of certain teams throughout the season. It also facilitates comparisons between seasons, both for players and for teams. The generated graphs can also, to some extent, be easily interpreted by individuals who are not very familiar with ML and DS.

3.3.2. Application

Initially, PCA was performed using the FMR library for each of the three investigated seasons, considering only numerical variables to prevent categorical variables from causing any constraints in the data formation. Only players were included in this analysis, and age was also excluded as it does not assess any performance attributes of the players. For the 2023/2024 season, the explanation of the first two dimensions accounts for 53.22% (38.63% + 14.59%, respectively). For the 2022/2023 season, it reaches 54.46% (40.15% + 14.31%, respectively). For the 2021/2022 season, it achieves 54.03% (39.24% + 14.79%). These results suggest that caution is needed when drawing conclusions from these graphs, as with the variance explanation hovering around 54%, there is not enough support to assert with high accuracy an objective truth about the patterns under study. Figures B.7, B.8 and B.9 show that results.

By delving deeper into the study for potential desired characteristics, for the 23/24 season, it is possible to focus solely on variables with offensive characteristics. In this B.10 chart it shows that on the top of the graph are the players with actions in set pieces and effective passing on offensive area. On the bottom it shows the ones who had a more relevant role in conducting the ball on offensive areas of the pitch and doing shots and goals. Gyokeres [nº 86] emerges as absolutely standout. In this new graph B.11 that measures expected offensive actions, it shows Gyokeres [nº 86] emerging as an exceptional creator of danger for opponents, with a very high rate of converting created opportunities, alongside Paulinho [nº 99]. On the other hand, at the lower part of the graph, it shows players who have squandered the most scoring opportunities, with particular emphasis

on Rafa Silva [nº4] for his missed goal chances and his role as a player more focused on providing assists.

In terms of participation, the B.12 shows a PCA only for the playing time variables of 2022/2023 season, which are 8 in total. On that figure the players who are more to the right have played much more time, and players who are more to the top have been substituted several times. On the other hand, players who are in the bottom of the chart and marked on the left are the ones who almost never played nor have been called to be on the bench. The ones who are in the bottom and in the right side are the ones who played several minutes but almost never had been substituted. On the other hand the ones who are in the top and on the left side are the ones who played less minutes, but have been called for the bench or played after being putted on the pitch from the bench several times and Peter Musa [nº18] is an example of that situation. Navarro [nº 316] who was at Gil Vicente that season, is an example of a player who played a lot of continuous minutes, since he played all the matches as a starter. The two first dimensions explain more than 80% of the variance, which means that we can securely make some well-founded conclusions.

Considering passing issues, in this graph B.13 it shows that at the top are the players with the greatest ability to interact with the ball and carry out useful actions in the offensive part of the field, and, on the other hand, it shows that at the bottom of the graph are players with greater ability to complete passes from behind on the field. It shows how, in the 2022/2023 season, Grimaldo [nº1] was exceptional in his ability to handle the ball in advanced positions and make accurate passing decisions in those areas. It is also noteworthy that, in the same season, several Benfica players stand out both at the top and bottom of the graph, highlighting how the club's diversity and ball possession power that season may have been decisive for their performance success or may reflect the identity of coach Roger Schmidt. Again, here, a lot of the variance is explained.

Regarding defensive matters, on this representation B.14 it shows that in the lower part are the players with the greatest capacity to resist opposing team high pressure, due to the way they present high levels of defensive actions in the most backward lines of the field, and on the other hand, in the upper part it presents the players with greater ability to recover the ball and carry out defensive actions in the team's offensive zone. Otávio [nº34] from Porto, who was the champion in this season of 2021/2022, excelled in recovering balls in the attacking midfield, which may have contributed to the club setting a record for points that year. João Mário Lopes [nº39] and Vitinha [nº37] also stood out in this season, suggesting that the influence of coach Sérgio Conceição may have also played a role in this statistical data.

Trying to map the ball possession, on this map B.15, it demonstrates on the top the players players characterized by having more secure possession of the ball on less advanced terrain, on the other hand, at the top of the graph it shows players characterized by taking more risky actions with the ball in more advanced terrain. In 2023/2024 Gyokeres [nº86]

again surges very well on this last matter, whyle At the same time, we can observe Gonçalo Inácio's [nº88] prominent position at the top of the graph as a reliable anchor for ball retention in the more defensive areas of his team. In this same quadrant, the frequent presence of central defenders from the national champions Sporting CP is evident, highlighting one of the well-known aspects of coach Rúben Amorim's tactical identity, which involves playing out from the back with three central defenders. This graph explains nearly 90% of the variance and thus allows for more insightful conclusions.

In this last interesting figure B.16, its shown on one hand, further to the right, a first squad of players who received no RC, where those who appear highest are those who make the most fouls and receive the most YC, never being sent-off and on the other hand, at the bottom there are players who commit almost no fouls and receive no disciplinary sanctions. In the remaining clouds of players we see a similar pattern, with the further to the left is their cloud, the more penalized with disciplinary sanctions the player tends to be, although the balance on the vertical axis remains the same. In the lefttmost part and above the graph are the players who are subject to severe disciplinary penalties more frequently, even without committing so many direct infractions. Those who receive more YC, and not red ones, without committing many fouls may eventually represent athletes who are sanctioned for protests or non-compliance with regulations in relation to jerseys use. In 2021/2022 Otávio [nº34], from FC Porto appears as an example of that.

The overall chart B.17, which combines data from all three seasons, primarily highlights Viktor Gyökeres [nº1148] as an outstanding offensive player and Alex Grimaldo [nº531] in the 2022/2023 season as a key player in both offensive and defensive aspects. Pedro Gonçalves , Rafa Silva, and Ricardo Horta consistently emerge as standout offensive players across all three seasons, being the most consistently impressive and distinguished offensive players in these editions of the championship. Other examples, such as Grimaldo and Otávio, only played in two of the seasons but also stood out during those periods. Taremi excelled in two seasons but experienced a sharp drop in performance in another

The chart B.18 shows the density and influence of all the variables used in the PCA, in conjunction with the values of PC1 and PC2. The tables B.1, B.2, B.3 show the objective influence of those variables in 2022/2023 season.

The evaluation of these three seasons in this context has the particularity that the coaches of the so-called 'Big Three' were the same throughout, except for SL Benfica, which experienced a coaching crisis in the 2021/2022 season.

3.3.3. Limitations of This Application

This assessment had the following limitations:

Graph Visualization: The dense data clouds in the graph make it difficult to quickly and intuitively derive insights about certain players, as their positions on the map are not easily distinguishable. The origin of the graphics also made their placement in this document extremely problematic, as they are arranged in a disorganized manner, making it more difficult to read and understand the work.

Variance Explanation: The more general graphs that cover the total of all variables for each season, and especially the graph that combines data from all three seasons (explaining 52.91% of the variance), provide a limited explanation of variance. This limitation restricts the ability to draw well-founded conclusions from some of the interpretations previously mentioned.

Need for Detailed Legends: The graphs require detailed legends to be understood by individuals who are not familiar with the field. However, this need does not compromise the simplicity and objectivity of the graphs.

CHAPTER 4

Results

4.1. Pattern analysis

Despite the insufficient explanation of variance, by handling specific objectives and queries with the data obtained through these PCAs, it is possible to cautiously recognize patterns that align with the intuitive observations of football spectators.

4.1.1. Individual patterns

Looking at players who completed all three seasons and adjusting the code to provide the desired queries, it is possible to see that, across all variables, players like Pepe, Ricardo Esgaio, and Rafa Silva maintained a consistent position on the graph throughout the three different seasons (Figure B.19). Although there were some fluctuations due to variations in their performances and minutes played, it is clear that these players generally occupied the same role within their respective teams over these three years, with their individual characteristics remaining consistent. Despite the low variance explanation, considering their positions on the graph, it would be difficult to dismiss the idea that the representation indeed refers to the same player in their respective cases.

Looking at the case of a particular player, Hidemassa Morita, who moved in 2023 from a relegation-fighting club, Santa Clara, to Sporting, its observed a notable shift in his position on the graph from a more defensive role (lower on the map) to a more offensive role. This change can be explained by the fact that, at Sporting, the player has a higher number of offensive actions compared to what he had at his previous club as we see on Figure B.20.

Examining a case where different players within the same team and under the same coach occupy similar positions and roles over three seasons, it shows that the performances and positions of João Palhinha in 2021/2022 and Morten Hjulmand in 2023/2024 at Sporting CP are extremely similar. In contrast, Manuel Ugarte showed some instability, shifting from a much more offensive pattern than the other two in 2021/2022 to a significantly more defensive role in 2022/2023 as shown in graph

The data collected and organized here show that, as exemplified by these demonstrated queries, it is possible to create a comprehensive *syllabus* of patterns to address the most detailed needs of the requester.

4.1.2. Collective patterns

Applying these same queries to study the teams, we find that the further right and higher a team is positioned, the greater the likelihood of one of the so-called "Big Three" winning

the championship. By examining three consecutive seasons where each of these teams won at least once, an interesting pattern emerges. Benfica, despite a change in coach, has maintained a more consistent performance pattern. Sporting, on the other hand, displayed a highly unstable pattern, raising questions about whether the points on the graph refer to the same team, given that they had the same coach and a relatively stable squad over these three years, with no significant changes in the core team. With Benfica maintaining a stable behavioral pattern, it is observed that Sporting and Porto won championships in seasons where they achieved exceptionally high performance. Additionally, SC Braga made notable attempts to break into the "cluster of the big teams" in at least two seasons. On the other hand, the further down and to the left a team is positioned, the greater the risk of relegation to a lower division. Figure B.22 shows that situation.

CHAPTER 5

Conclusions and recommendations for future work

5.1. Achievements

The first research question has been answered. Using RStudio, a data manipulation tool was created that provides direct, quick, and concrete answers through simple Cartesian model-based graphs. This tool partially addresses questions about the individual and collective performance of players and teams. The model can quickly reveal insights such as a player's or team's rise or decline in the PFL over the years, explore potential relationships between age and qualitative performance, assess the impact of YC on team performance, and concisely distinguish different field positions, as was predominantly done in the previously mentioned SLR.

Regarding the second research question, it is superficially addressed, as a complete answer would require a more in-depth investigation of this algorithm, specifically a greater investment of time in learning through systematic use of this tool to build solid knowledge based on experience. The tool itself does allow for answering the question, but to achieve a thorough understanding, a significantly larger number of experiments and further specification of a series of queries are needed to obtain more incisive insights, especially in areas where higher variance explanation levels can be achieved with a focus on particular characteristics. From the perspective of teams, the study of concrete and critical weaknesses in their performance remains unexplored, leaving the general idea mentioned in subsection 4.1.2.

5.2. Limitations and Potential Errors

As mentioned in point 3.3.3., the low explained variance in higher-dimensional PCA, that is, those involving more variables, necessitates caution to avoid making hasty judgments. It is important to explore in more detail which variables contribute the most to moving points on the Cartesian plane and which contribute the least. Understanding whether certain variables disproportionately shift points further to the left or right, or up or down, is crucial. The necessary graphical representation to achieve this is challenging to visualize, and it might be necessary to reduce column names to simplify the task, or alternatively, to examine the quotients of each column one by one, which would be a more time-consuming process. It should be noted that while reducing this information to a simple Cartesian reference might allow for intuitive interpretation if there is an appropriate legend, those making queries in the code need to have some advanced programming knowledge to understand what is being requested.

The work is also heavily focused on issues that are most relevant to the so-called "three big" clubs in Portuguese football and ends up neglecting the analysis of other clubs. To make the document more engaging and appealing to potential readers, and to make the best use of the available page limit, the focus has been more on Sporting, Porto, and Benfica.

At a later stage of the work, it was discovered that there was an error in the data transformation phase regarding the summarizing of the "Playing Time_Mn/MP" column, where the "maximum" parameter should have been used instead of the "mean" parameter. However, this is not expected to have produced any significant difference in the final results.

5.3. Ethical and Human Concerns

With the volatility of technological trends emerging globally, the obsession with measuring work performance through increasingly sophisticated and detailed AI and ML models may lead to a general loss of awareness regarding the humanity of others, creating an unrealistic sense of demand. With so many numbers, types of statistical variables, graphs, dashboards, and spreadsheets used to evaluate others' performance, and the growing detachment between people due to the use of more digital communication tools, there is a risk of developing a disordered view of the real capabilities of our peers. Football is no exception to this. These tools should be used with a full awareness that we are evaluating the performance of individuals on a football field, and to avoid dulling our senses, we should always remember to label our graphs with the names of real people in order to maintain some sense of empathy, by realising that we are talking about a real human being.

5.4. Recommendations for Future Work

The continuation of this work should, as previously introduced in this chapter, place greater emphasis on the remaining clubs. A deeper exploration of this model should be capable of producing insights and useful information about more objective causes and critical variables that can distinguish a championship-winning team from a podium-finishing team, identify the crucial variables that determine whether a club gets relegated, or better understand the relationship between an individual player's performance and the collective success or failure of a team. Additionally, it is essential to explore PCA models focused on different types of characteristics, as these often provide a greater explanation of variance. At the same time, it is important to identify which variables can be excluded from larger PCA models to improve the explanation without losing significant information.

Furthermore, data engineering mechanisms should be explored to modify certain columns or even add others that could enhance the understanding of the issues raised. One idea could involve trying to combine success rates, which were set aside at the beginning of the project, with the variables already studied, and examining how this affects the understanding of each sports season's narrative. Finally, a more in-depth study of the

influence of each of the over 130 variables analyzed in creating those Cartesian reference points is needed. This will help identify which columns are most critical in defining and understanding certain sporting outcomes for teams in the future, while also determining which ones most effectively support the direct assessment of the strengths and weaknesses of a particular athlete, where such conclusions are supported by a more significant and objective explanation of the patterns created through the data.

References

[1] S. E. Akhanli and C. Hennig, "Clustering of football players based on performance data and aggregated clustering validity indexes," *Journal of Quantitative Analysis in Sports*, vol. 19, 2 2023.

[2] E. van de Ven, "Clustering soccer players to find the drivers of soccer team performance," 2018.

[3] P. D'Urso, L. D. Giovanni, and V. Vitale, "A robust method for clustering football players with mixed attributes," *Annals of Operations Research*, vol. 325, pp. 9–36, 1 2018.

[4] T. L. Bergkamp, W. G. P. Frencken, A. S. M. Niessen, R. R. Meijer, and R. J. R. den Hartigh, "How soccer scouts identify talented players," *European Journal of Sport Science*, vol. 22, pp. 994–1004, 7 2022.

[5] R. Aquino, J. C. Machado, F. M. Clemente, G. Praça, L. G. C. Gonçalves, B. Melli-Neto, J. V. S.Ferrari, L. H. P. Vieira, E. F. Puggina, and C. Carling, "Comparisons of ball possession, match runningperformance, player prominence and team-network properties according to match outcomeand playing formation during the 2018 fifa worldcup," *International Journal of Performance Analysis in Sport*, vol. 19, pp. 1026–1037, 6 2019.

[6] J. Bekkers and S. Dabadghao, "Flow motifs in soccer: What can passing behavior tell us?" *Journal of Sports Analytics*, vol. 5, pp. 299–311, 4 2019.

[7] H. Menéndez, G. Bello-Orgaz, and D. Camacho, "Extracting behavioural models from 2010 fifa world cup," *Journal of Systems Science and Complexity*, vol. 26, pp. 43–61, 1 2013.

[8] M. Carpita, E. Ciavolino, and P. Pasca, "Exploring and modelling team performances of thekaggle european soccer database," *Statistical Modelling*, vol. 19, pp. 1–28, 1 2019.

[9] R. Drezner, L. Lamas, C. Farias, J. Barrera, and L. Dantas, "A method for classifying and evaluating the efficiency of offensive playing styles in soccer," *Journal of Physical Education and Sport*, vol. 20, pp. 1284–1294, 3 2020.

[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 Transactions on Intelligent Systems and Technology*, vol. 10, pp. 1–27, 5 2019.

[11] S. E. Akhanli and C. Hennig, "Some issues in distance construction for football players performance data," *Archives of Data Science*, vol. 2, 1 2017.

[12] OECD, *OECD Glossary of Statistical Terms*. OECD Publishing, 2008.

[13] R. Onody and P. De Castro, "Complex network study of brazilian soccer players," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 70, p. 037 103, Oct. 2004. DOI: 10.1103/PhysRevE.70.037103.

[14] C. Cotta, A. M. Mora, C. M. Molina, and J. J. M. Guervús, "FIFA world cup 2010: A network analysis of the champion team play," *CoRR*, vol. abs/1108.0261, 2011. arXiv: 1108.0261. [Online]. Available: http://arxiv.org/abs/1108.0261.

[15] I. Mchale, P. Scarf, and D. Folker, "On the development of a soccer player performance rating system for the english premier league," *Interfaces*, vol. 42, pp. 339–351, Aug. 2012. DOI: 10.2307/23254864.

[16] F. Clemente, F. Martins, D. P. Wong, D. Kalamaras, and R. Mendes, "Midfielder as the prominent participant in the building attack: A network analysis of national teams in fifa world cup 2014," *International Journal of Performance Analysis in Sport*, vol. 15, pp. 704–722, Aug. 2015. DOI: 10.1080/24748668.2015.11868825.

[17] P. Larkin, D. Marchant, A. Syder, and D. Farrow, "An eye for talent: The recruiters' role in the australian football talent pathway," *PLoS ONE*, vol. 15, Nov. 2020. DOI: 10.1371/journal.pone.0241307.

[18] J. Peña and R. Navarro, "Who can replace xavi? a passing motif analysis of football players," *CoRR*, vol. abs/1506.07768, 2015. arXiv: 1506.07768. [Online]. Available: http://arxiv.org/abs/1506.07768.

[19] J. Hobbs, P. Power, L. Sha, H. Ruiz, and P. Lucey, "Quantifying the value of transitions in soccer via spatiotemporal trajectory clustering," in *Proceedings of the 12th annual MIT Sloan Sports Analytics Conference*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:221471459.

[20] I. Kosmidis and D. Karlis, "Model-based clustering using copulas with applications," *Stat. Comput.*, vol. 26, no. 5, pp. 1079–1099, 2016. DOI: 10.1007/S11222-015-9590-5. [Online]. Available: https://doi.org/10.1007/s11222-015-9590-5.

[21] I. Behravan and S. M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," *Soft Comput.*, vol. 25, no. 3, pp. 2499–2511, 2021. DOI: 10.1007/S00500-020-05319-3. [Online]. Available: https://doi.org/10.1007/s00500-020-05319-3.

[22] L. Yingying, S. Chiusano, and V. D'Elia, "Modeling athlete performance using clustering techniques," in *International Symposium on Electronic Commerce and Security*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:6867082.

[23] E. Galariotis, C. Germain, and C. Zopounidis, "A combined methodology for the concurrent evaluation of the business, financial and sports performance of football clubs: The case of france," *Ann. Oper. Res.*, vol. 266, no. 1-2, pp. 589–612, 2018. DOI: 10.1007/S10479-017-2631-Z. [Online]. Available: https://doi.org/10.1007/s10479-017-2631-z.

[24] A. Drachen, C. Thurau, R. Sifa, and C. Bauckhage, "A comparison of methods for player clustering via behavioral telemetry," *CoRR*, vol. abs/1407.3950, 2014. arXiv: 1407.3950. [Online]. Available: http://arxiv.org/abs/1407.3950.

[25] P. Rai and S. Shubha, "A survey of clustering techniques," *International Journal of Computer Applications*, vol. 7, Oct. 2010. DOI: 10.5120/1326-1808.

[26] Y. Li, S. Zong, Y. Shen, Z. Pu, M. Á. Gómez, and Y. Cui, "Characterizing player's playing styles based on player vectors for each playing position in the chinese football super league," *CoRR*, vol. abs/2205.02731, 2022. DOI: 10.48550/ARXIV.2205.02731. arXiv: 2205.02731. [Online]. Available: https://doi.org/10.48550/arXiv.2205.02731.

[27] T. Decroos, L. Bransen, J. V. Haaren, and J. Davis, "Actions speak louder than goals: Valuing player actions in soccer," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis, Eds., ACM, 2019, pp. 1851–1861. DOI: 10.1145/3292500.3330758. [Online]. Available: https://doi.org/10.1145/3292500.3330758.

[28] Y. Gai, A. Leicht, C. Peñas, and M. Ruano, "Physical and technical differences between domestic and foreign soccer players according to playing positions in the china super league," *Research in Sports Medicine*, vol. 27, Oct. 2018. DOI: 10.1080/15438627.2018.1540005.

[29] D. Moher, L. Shamseer, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. Stewart, D. Altman, A. Booth, A. Chan, S. Chang, T. Clifford, K. Dickersin, P. Gøtzsche, J. Grimshaw, T. Groves, M. Helfand, J. Higgins, and E. Whitlock, "Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement," *Systematic reviews*, Jan. 2015.

[30] I. Jolliffe, "Principal component analysis," in International Encyclopedia of Statistical Science., Jan. 2011, pp. 1094–1096, ISBN: 978-3-642-04897-5. DOI: 10.1007/978-3-642-04898-2_455.

[31] J. Shlens, "A tutorial on principal component analysis," *ArXiv*, vol. abs/1404.1100, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:2051212.

APPENDIX A

General data glossary

Player: Name of the player.
Pos: Position most commonly played by the player GK - Goalkeepers DF - Defenders MF - Midfielders FW - Forwards
Age: Age at season start
MP_x: Matches played
Playing Time_Starts: Game or games started by player
Playing Time_Min_x: Minutes played
Playing Time_90s: Minutes played divided by 90
Performance_Gls: Goals
Performance_Ast: Assists
Performance_G+A: Goals + Assists
Performance_G-PK: Non-Penalty Goals
Performance_PK: Penalty Kicks Made
Performance_PKatt: Penalty Kicks Attempted
Performance_CrdY_x: Yellow Cards
Performance_CrdR_x: Red Cards
Expected_xG_x: Expected Goals
Expected_npxG_x: Non-Penalty Expected Goals
Expected_xAG: Expected Assisted Goals
Expected_npxG+xAG: Non-Penalty Expected Goals plus Assisted Goals
Progression_PrgC: Progressive Carries
Progression_PrgP: Progressive Passes
Progression_PrgR: Progressive Passes Received
Per 90 Minutes_Gls: Goals Scored per 90 minutes
Per 90 Minutes_Ast: Assists per 90 minutes
Per 90 Minutes_G+A: Goals and Assists per 90 minutes
Per 90 Minutes_G-PK: Non-Penalty Goals/90
Per 90 Minutes_G+A-PK: Non-Penalty Goals + Assists/90
Per 90 Minutes_xG: Expected Goals per 90 minutes
Per 90 Minutes_xAG: Expected Assisted Goals per 90 minutes
Per 90 Minutes_xG+xAG: Expected Goals plus Assisted Goals per 90 minutes
Per 90 Minutes_npxG: Non-Penalty Expected Goals per 90 minutes
Per 90 Minutes_npxG+xAG: npxG + xAG
Standard_Sh: Total shots

Standard_SoT: Total shots on target
Standard_Sh/90: Total shots per 90 minutes
Standard_SoT/90: Total shots on target per 90 minutes
Standard_G/Sh: Goals per shot
Standard_G/SoT: Goals per shot on target
Standard_FK: Shots from free kicks
Expected_npxG/Sh: Non-Penalty Expected Goals per shot
Expected_G-xG: Goals minus Expected Goals
Expected_np:G-xG: Non-Penalty Goals minus Non-Penalty Expected Goals
Total_Cmp: Passes Completed
Total_Att: Passes Attempted
Total_TotDist: Total distance, in yards, that completed passes have traveled in any direction
Total_PrgDist: Total distance, in yards, that completed passes have traveled towards the opponent's goal
Short_Cmp: Passes completed between 5 and 15 yards
Short_Att: Passes attempted between 5 and 15 yards
Medium_Cmp: Passes completed between 15 and 30 yards
Medium_Att: Passes attempted between 15 and 30 yards
Long_Cmp: Passes completed longer than 30 yards
Long_Att: Passes attempted longer than 30 yards
Expected_xA: Expected Assists
Expected_A-xAG: Assists minus Expected Goals Assisted
KP: Key passes
1/3: Completed passes that enter the 1/3 of the pitch closest to the goal
PPA: Completed passes into the 18-yard box
CrsPA: Crosses into Penalty Area
Pass Types_Live: Live-ball Passes
Pass Types_Dead: Dead-ball Passes
Pass Types_FK: Passes attempted from free kicks
Pass Types_TB: Completed pass sent between back defenders into open space
Pass Types_Sw: Passes that travel more than 40 yards of the width of the pitch
Pass Types_Crs: Crosses
Pass Types_TI: Throw-ins Taken
Pass Types_CK: Corner Kicks
Corner Kicks_In: Inswinging Corner Kicks
Corner Kicks_Out: Outswinging Corner Kicks
Corner Kicks_Str: Straight Corner Kicks
Outcomes_Off: Offsides
Outcomes_Blocks: Passes blocked by the opponent who was standing in the path

SCA_SCA: Shot-Creating Actions
SCA_SCA90: Shot-Creating Actions per 90 minutes
SCA Types_PassLive: Completed live-ball passes that lead to a shot attempt
SCA Types_PassDead: Completed dead-ball passes that lead to a shot attempt.
SCA Types_TO: Successful take-ons that lead to a shot attempt
SCA Types_Sh: Shots that lead to another shot attempt
SCA Types_Fld: Fouls drawn that lead to a shot attempt
SCA Types_Def: Defensive actions that lead to a shot attempt
GCA_GCA: Goal-Creating Actions
GCA_GCA90: Goal-Creating Actions per 90 minutes
GCA Types_PassLive: Completed live-ball passes that lead to a goal
GCA Types_PassDead: Completed dead-ball passes that lead to a goal.
GCA Types_TO: Successful take-ons that lead to a goal
GCA Types_Sh: Shots that lead to another goal-scoring shot
GCA Types_Fld: Fouls drawn that lead to a goal
GCA Types_Def: Defensive actions that lead to a goal
Tackles_Tkl: Number of players tackled
Tackles_TklW: Tackles in which the tackler's team won possession of the ball
Tackles_Def 3rd: Tackles in defensive 1/3
Tackles_Mid 3rd: Tackles in middle 1/3
Tackles_Att 3rd: Tackles in attacking 1/3
Challenges_Tkl: Number of dribblers tackled
Challenges_Att: Number of unsuccessful challenges plus number of dribblers tackled
Challenges_Lost: Number of unsucessful attempts to challenge a dribbling player
Blocks_Blocks: Number of times blocking the ball by standing in its path
Blocks_Sh: Number of times blocking a shot by standing in its path
Blocks_Pass: Number of times blocking a pass by standing in its path
Tkl+Int: Number of players tackled plus number of interceptions
Clr: Clearances
Err: Mistakes leading to an opponent's shot
Touches_Touches: Number of times a player touched the ball
Touches_Def Pen: Touches in defensive penalty area
Touches_Def 3rd: Touches in defensive 1/3
Touches_Mid 3rd: Touches in middle 1/3
Touches_Att 3rd: Touches in attacking 1/3
Touches_Att Pen: Touches in attacking penalty area
Touches_Live: Live-ball touches
Take-Ons_Att: Number of attempts to take on defenders while dribbling
Take-Ons_Succ: Number of defenders taken on successfully, by dribbling past them
Take-Ons_Tkld: Number of times tackled by a defender during a take-on attempt

Carries_Carries: Number of times the player controlled the ball with their feet

Carries_TotDist: Total distance, in yards, a player moved the ball while controlling it with their feet, in any direction

Carries_PrgDist: Total distance, in yards, a player moved the ball while controlling it with their feet towards the opponent's goal

Carries_1/3: Carries that enter the 1/3 of the pitch closest to the goal

Carries_CPA: Carries into the 18-yard box

Carries_Mis: Number of times a player failed when attempting to gain control of a ball

Carries_Dis: Number of times a player loses control of the ball after being tackled by an opposing player. Does not include attempted take-ons

Receiving_Rec: Number of times a player successfully received a pass

Playing Time_Mn/MP: Minutes Per Match Played

Starts_Compl: Complete matches played

Subs_Subs: Games as sub

Subs_unSub: Games as an unused substitute

Team Success_PPM: Points per Match

Team Success_onG: Goals scored by team while on pitch

Team Success_onGA: Goals allowed by team while on pitch

Team Success_+/-: Goals scored minus goals allowed by the team while the player was on the pitch.

Team Success_+/-90: Goals scored minus goals allowed by the team while the player was on the pitch per 90 minutes played.

Team Success_On-Off: Net goals per 90 minutes by the team while the player was on the pitch minus net goals allowed per 90 minutes by the team while the player was off the pitch.

Team Success (xG)_onxG: Expected goals by team while on pitch

Team Success (xG)_onxGA: Expected goals allowed by team while on pitch

Team Success (xG)_xG+/-: Expected goals scored minus expected goals allowed by the team while the player was on the pitch.

Team Success (xG)_xG+/-90: Expected goals scored minus expected goals allowed by the team while the player was on the pitch per 90 minutes played.

Team Success (xG)_On-Off: Net expected goals per 90 minutes by the team while the player was on the pitch minus net expected goals per 90 minutes by the team while the player was off the pitch

Performance_2CrdY: Second yellow Card

Performance_Fls: Fouls Committed

Performance_Fld: Fouls Drawn

Performance_Off: Offsides

Performance_Int: Interceptions

Performance_PKwon: Penalty Kicks Won

Performance_PKcon: Penalty Kicks Conceded
Performance_OG: Own Goals
Performance_Recov: Number of loose balls recovered
Aerial Duels_Won: Aerials Won
Aerial Duels_Lost: Aerials Lost

APPENDIX B

General Figures

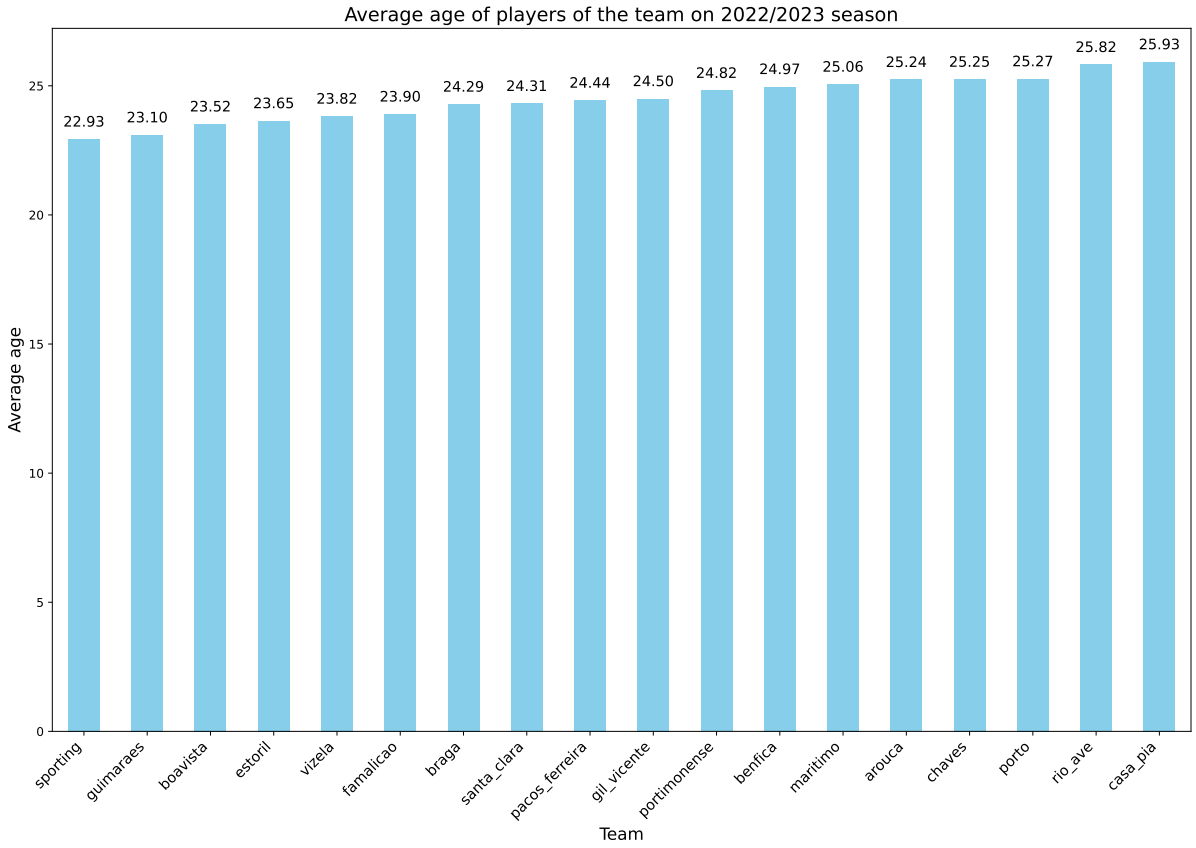

FIGURE B.1. Average age of players of the team on 2022/2023 season

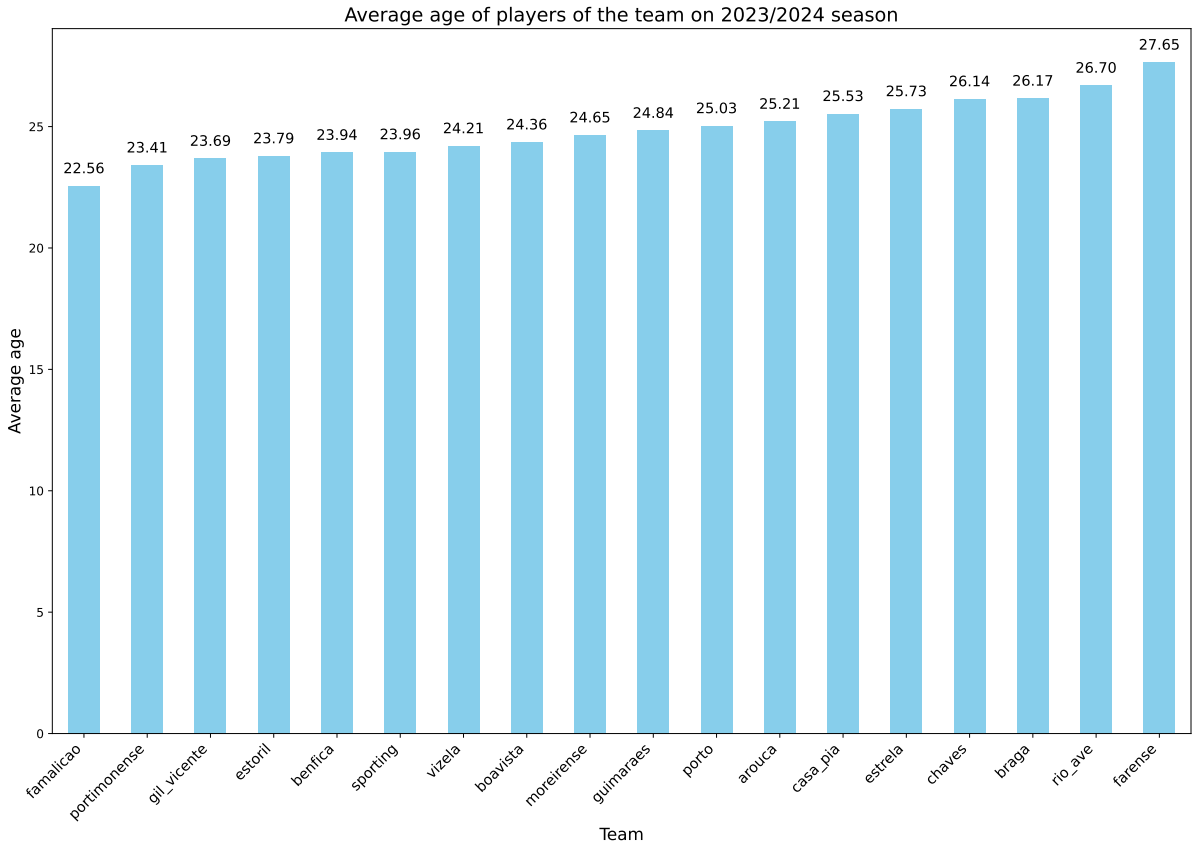

FIGURE B.2. Average age of players of the team on 2023/2024 season

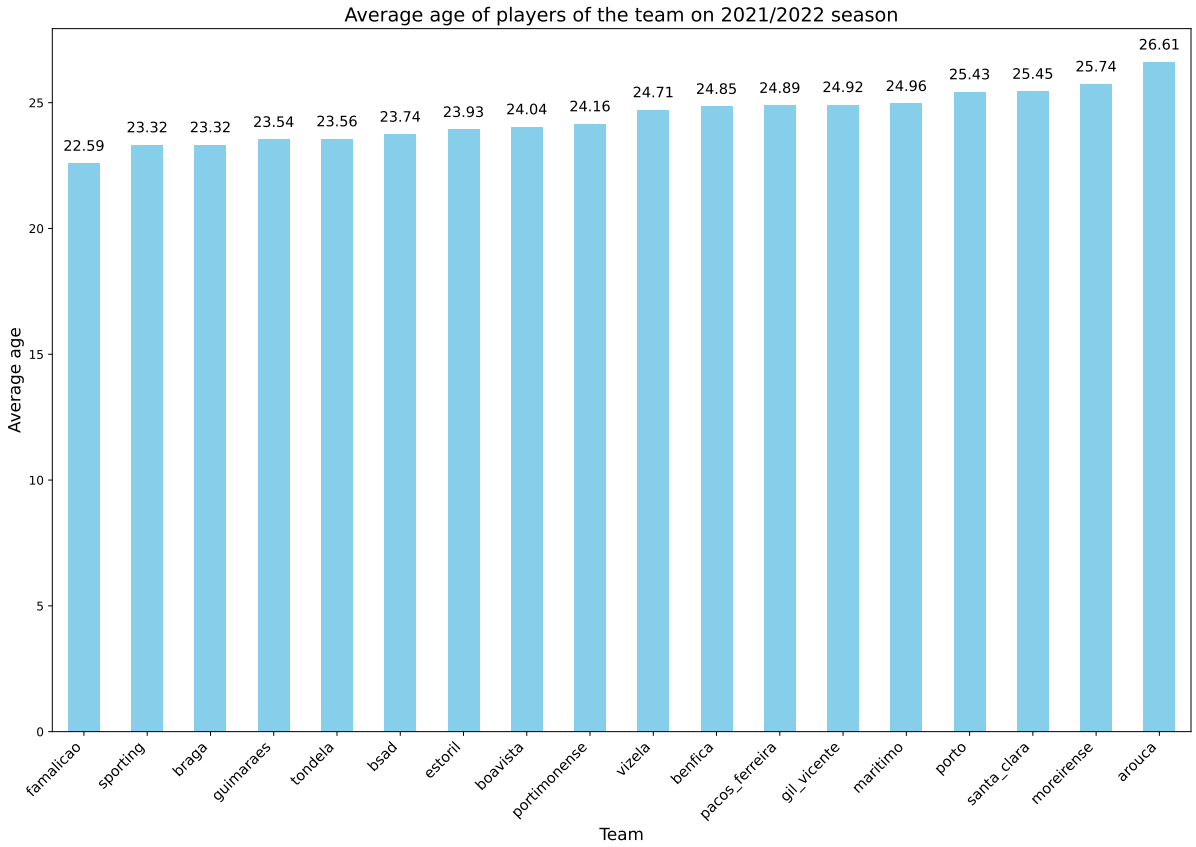

FIGURE B.3. Average age of players of the team on 2021/2022 season

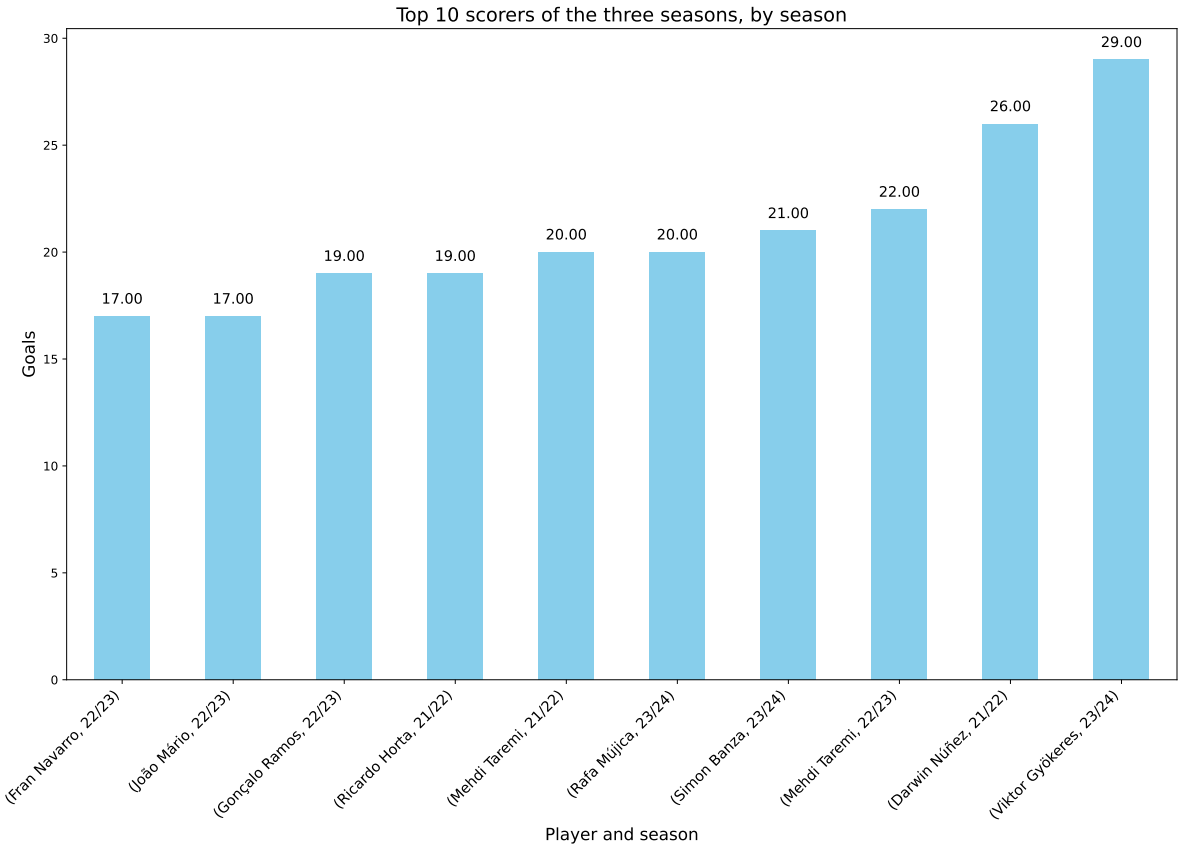

FIGURE B.4. Top 10 scorers of the three seasons, by season

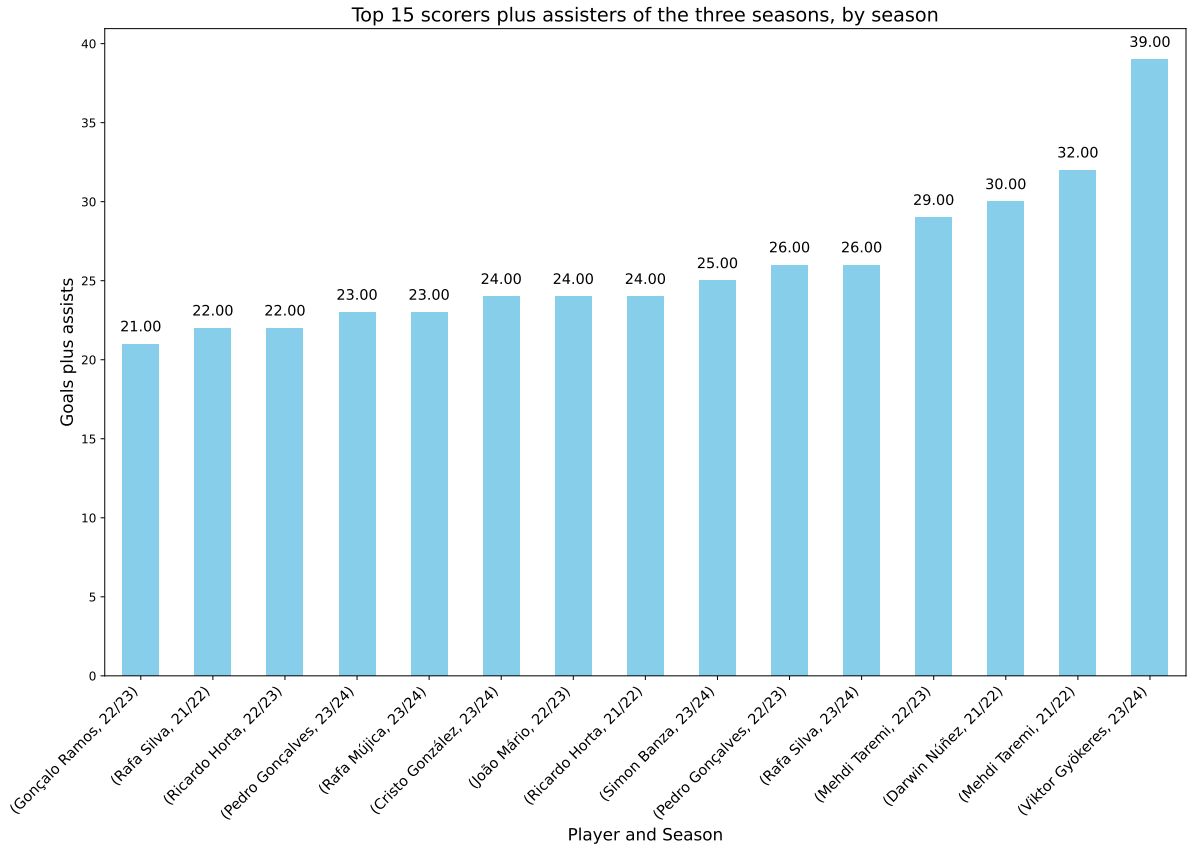

FIGURE B.5. Top 15 scorers plus assisters of the three seasons, by season

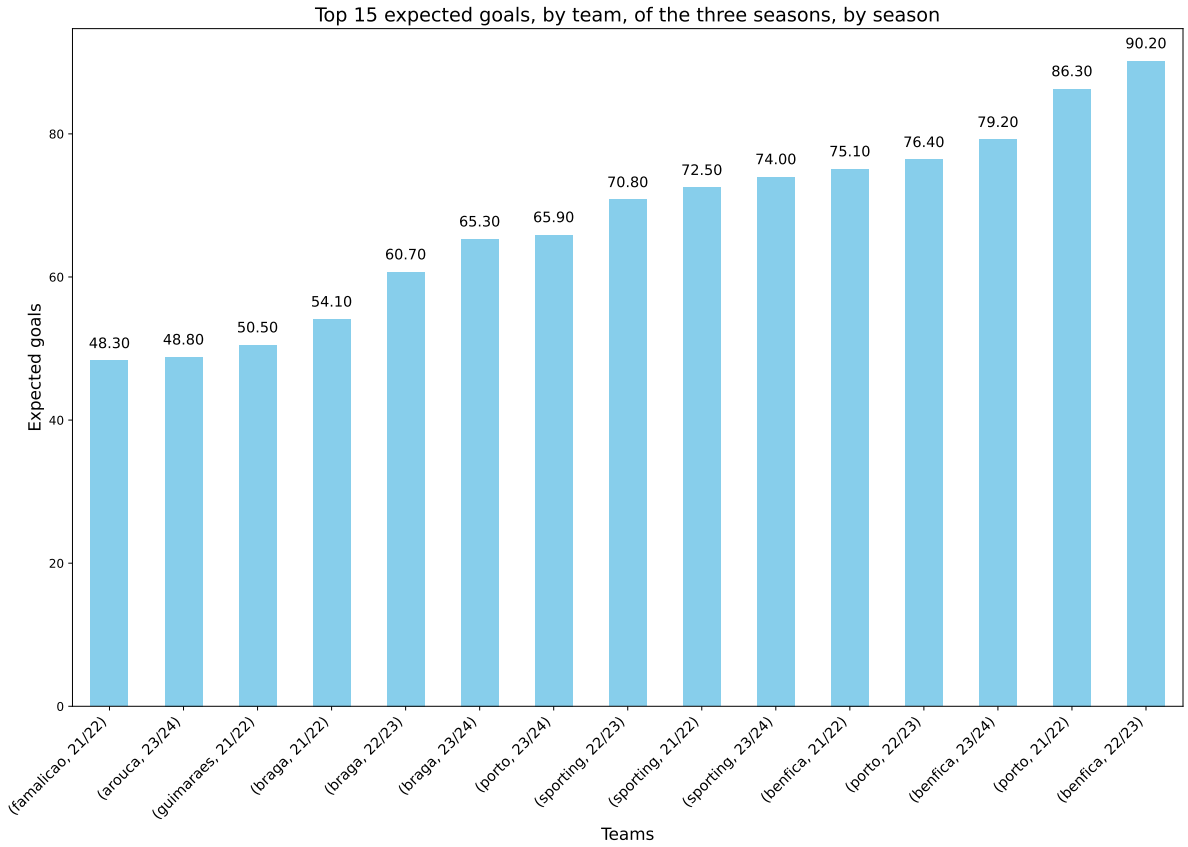

FIGURE B.6. Top 15 expected goals, by team, of the three seasons, by season

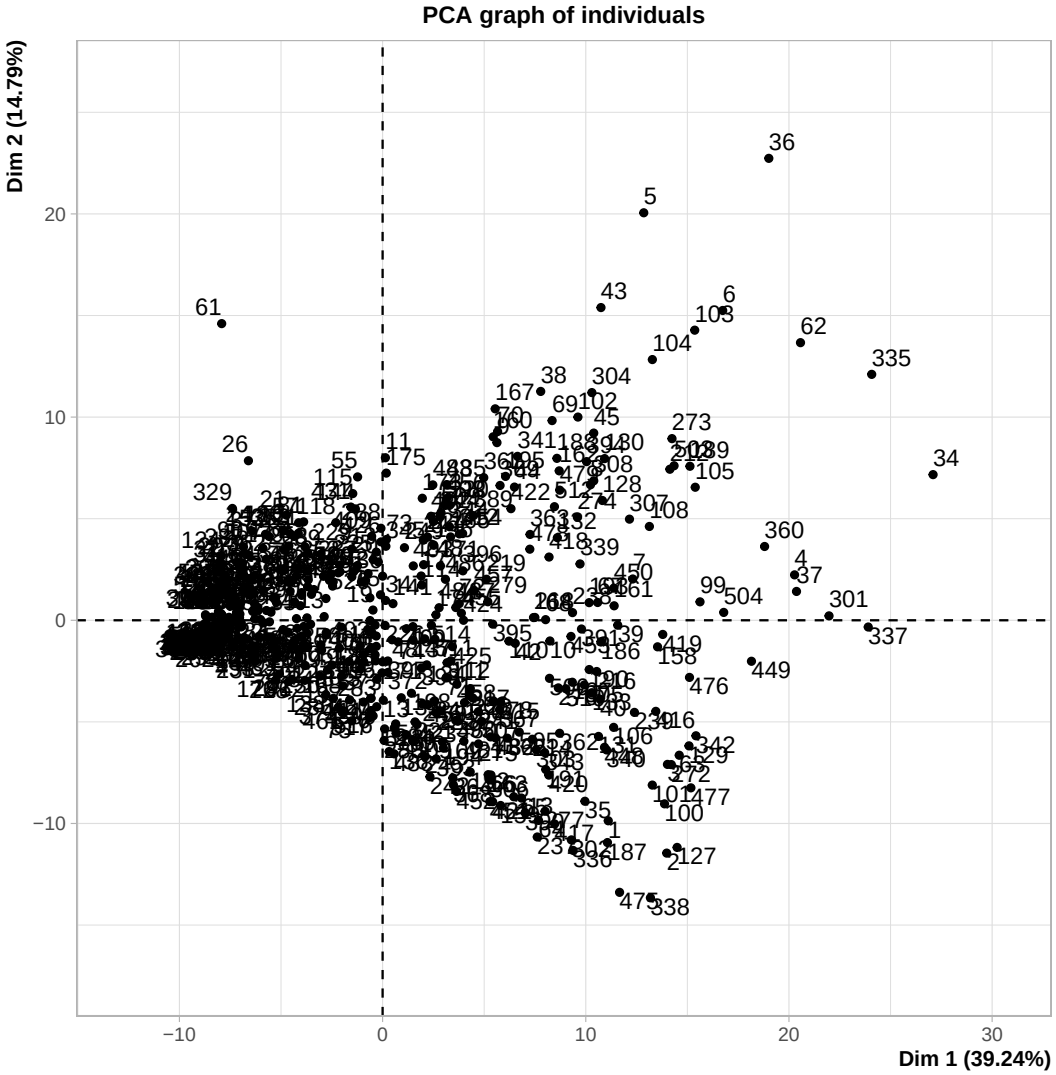

FIGURE B.7. PCA for all statistical variables of the season 2021/2022

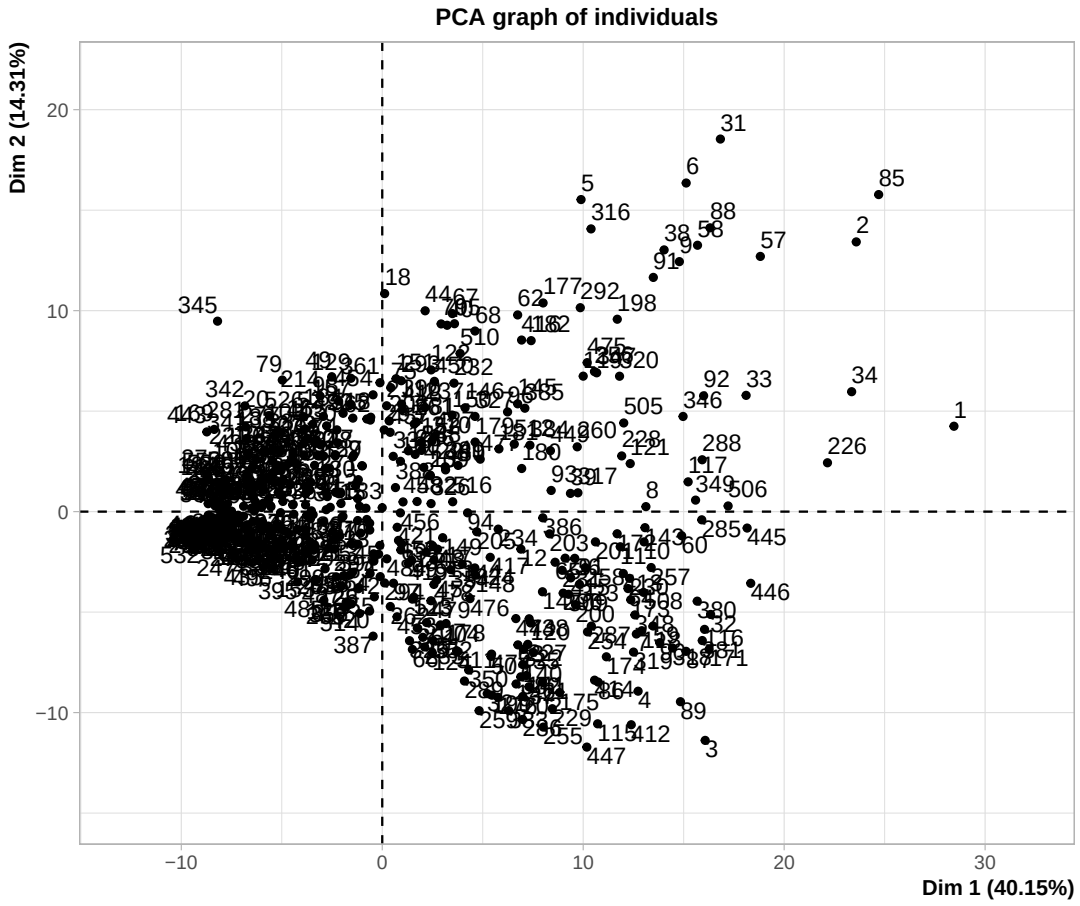

FIGURE B.8. PCA for all statistical variables of the season 2022/2023

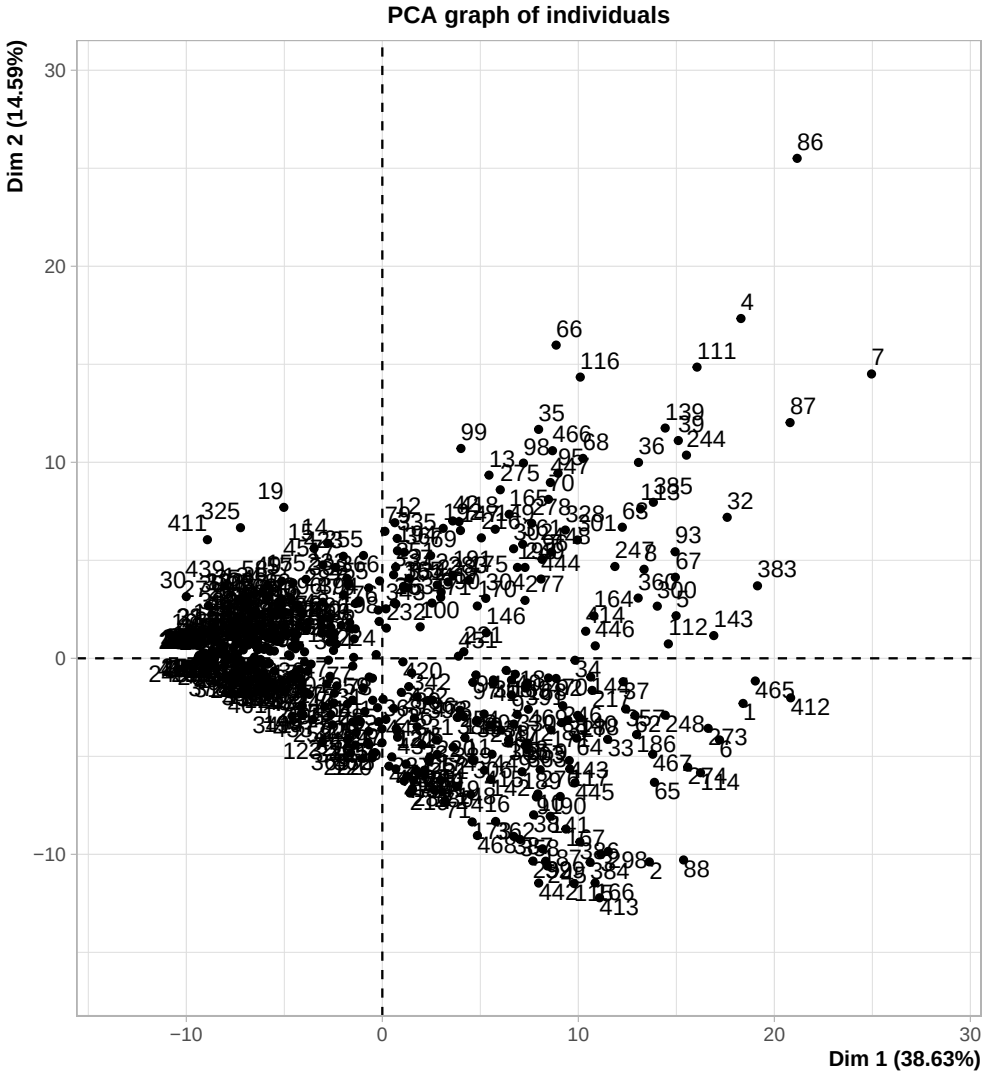

FIGURE B.9. PCA for all statistical variables of the season 2023/2024

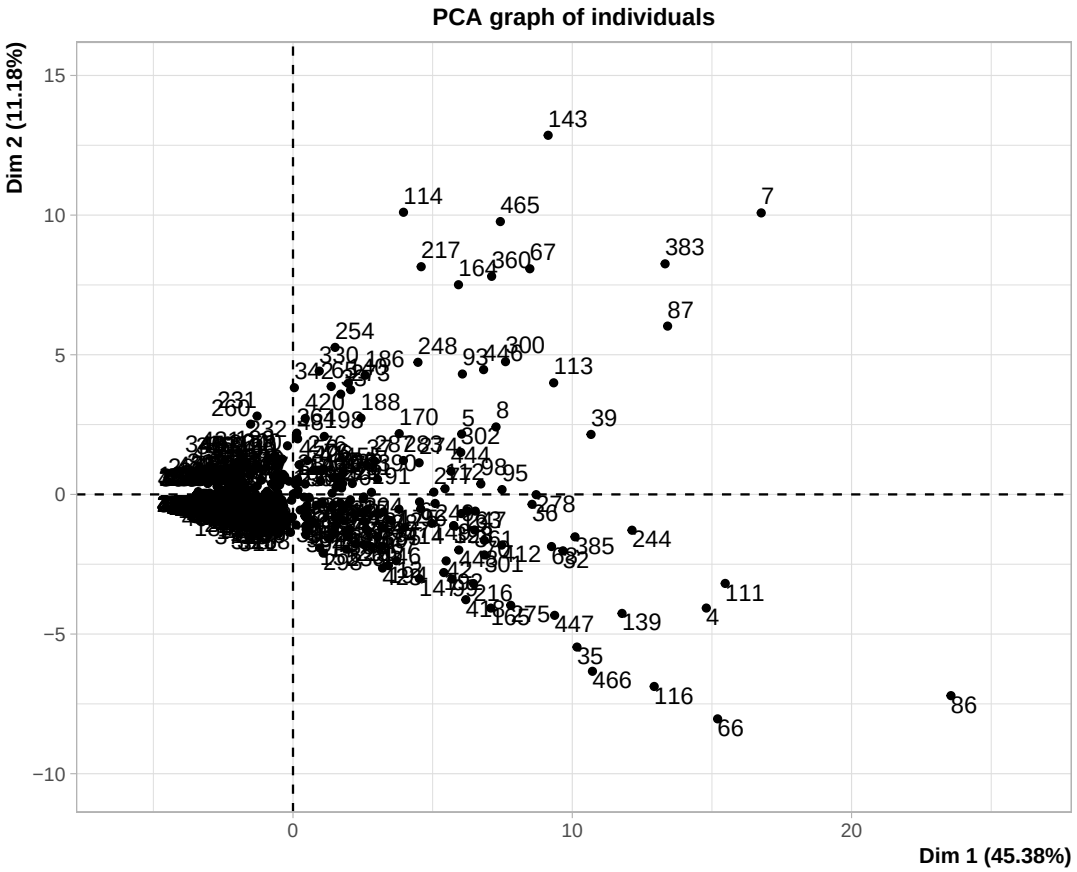

FIGURE B.10. PCA for offensive style variables of the season 2023/2024

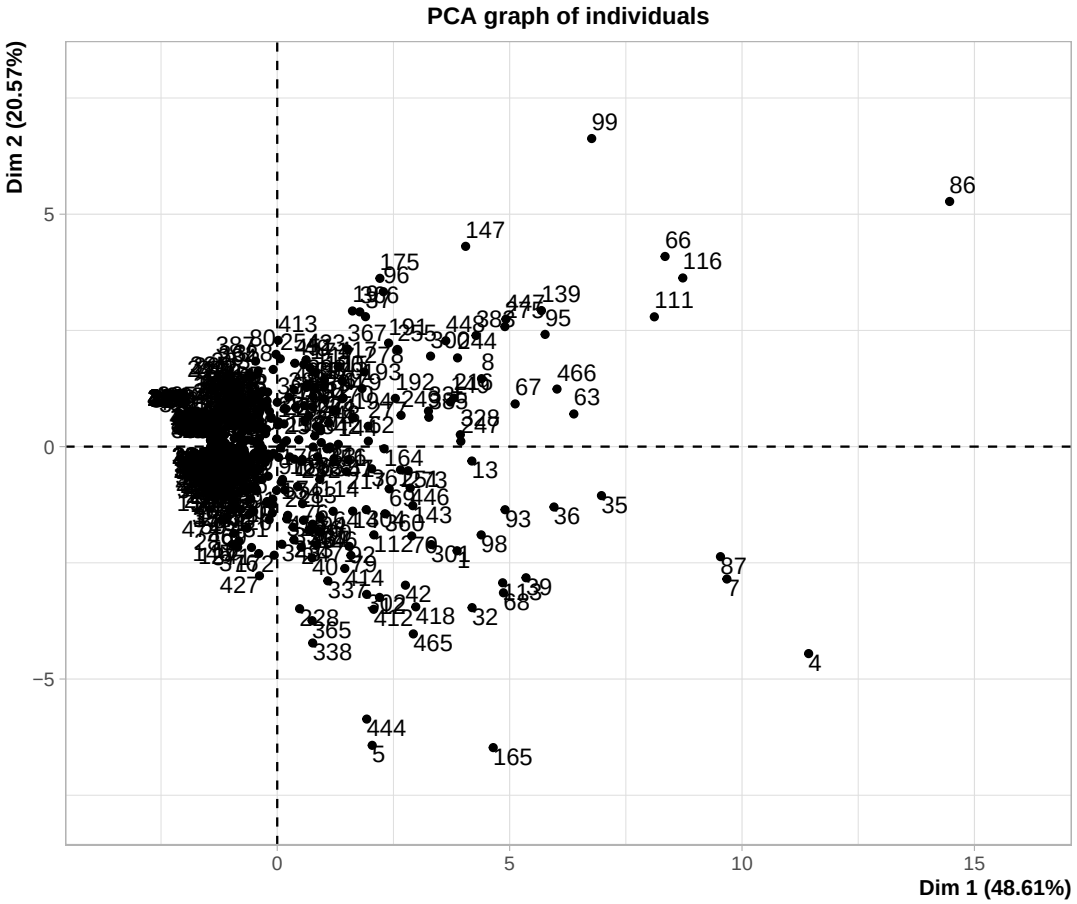

FIGURE B.11. PCA for expected offensive actions variables of the season 2023/2024

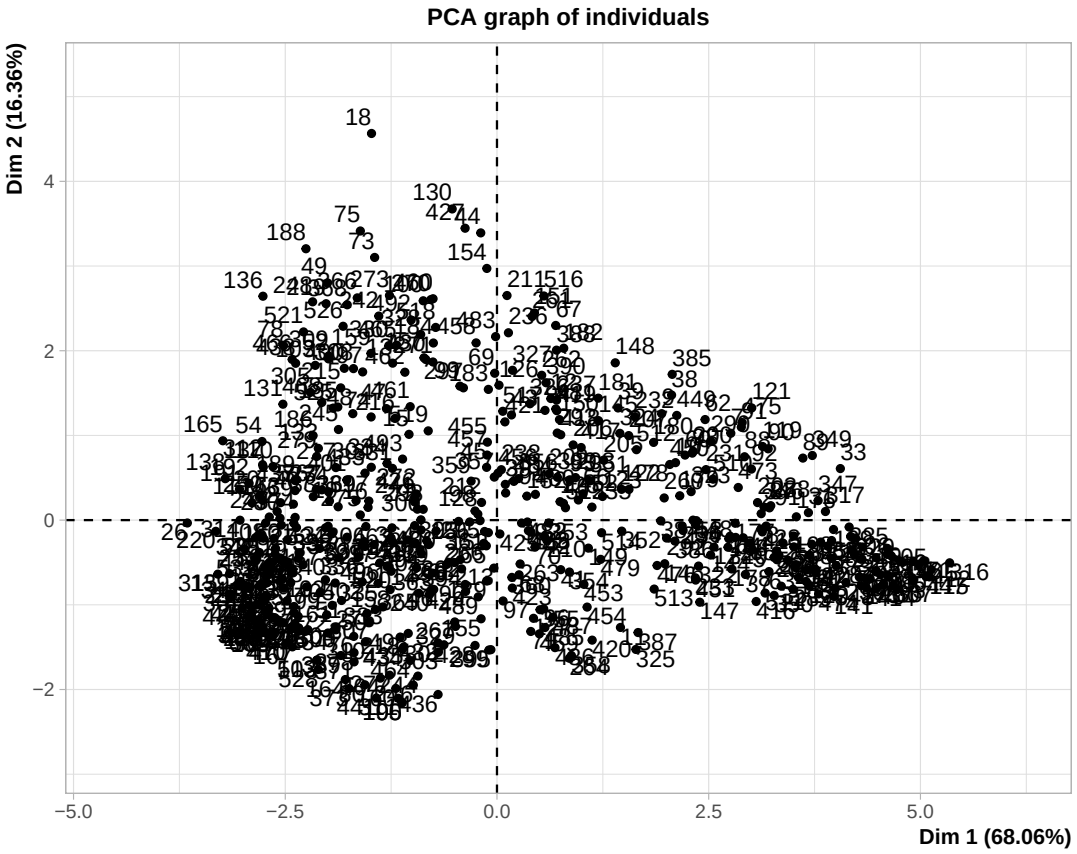

FIGURE B.12. PCA for time playing variables of the season 2022/2023

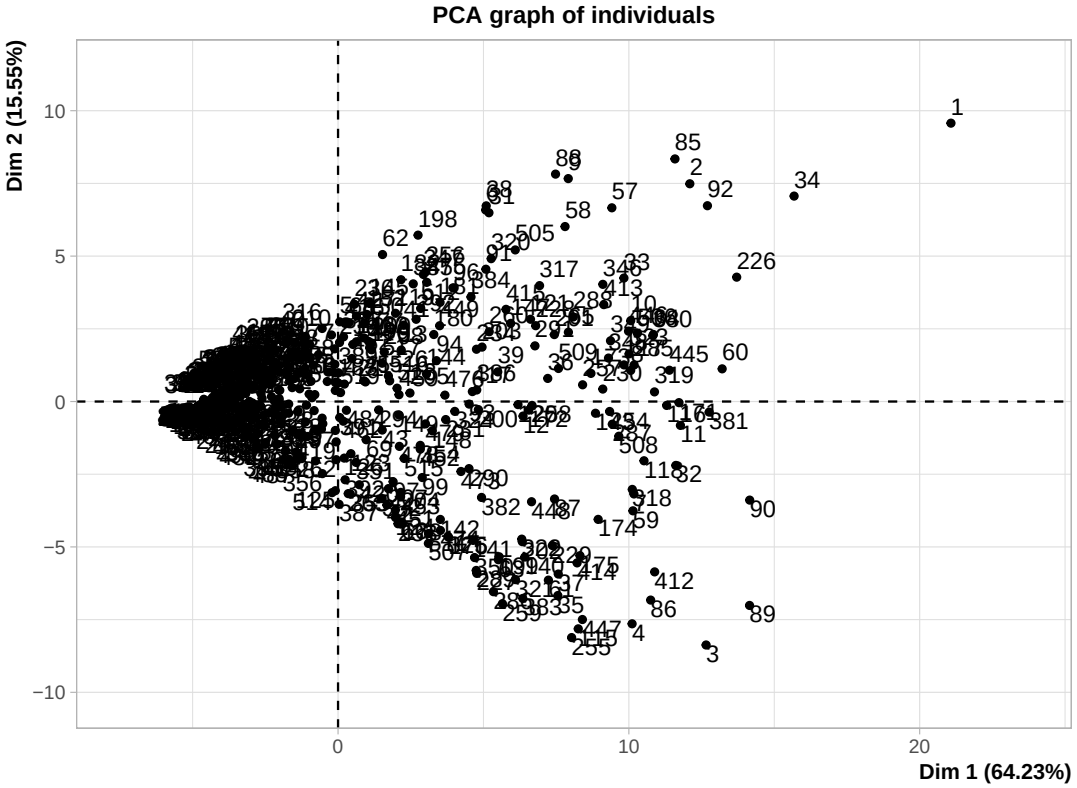

FIGURE B.13. PCA for passing variables of the season 2022/2023

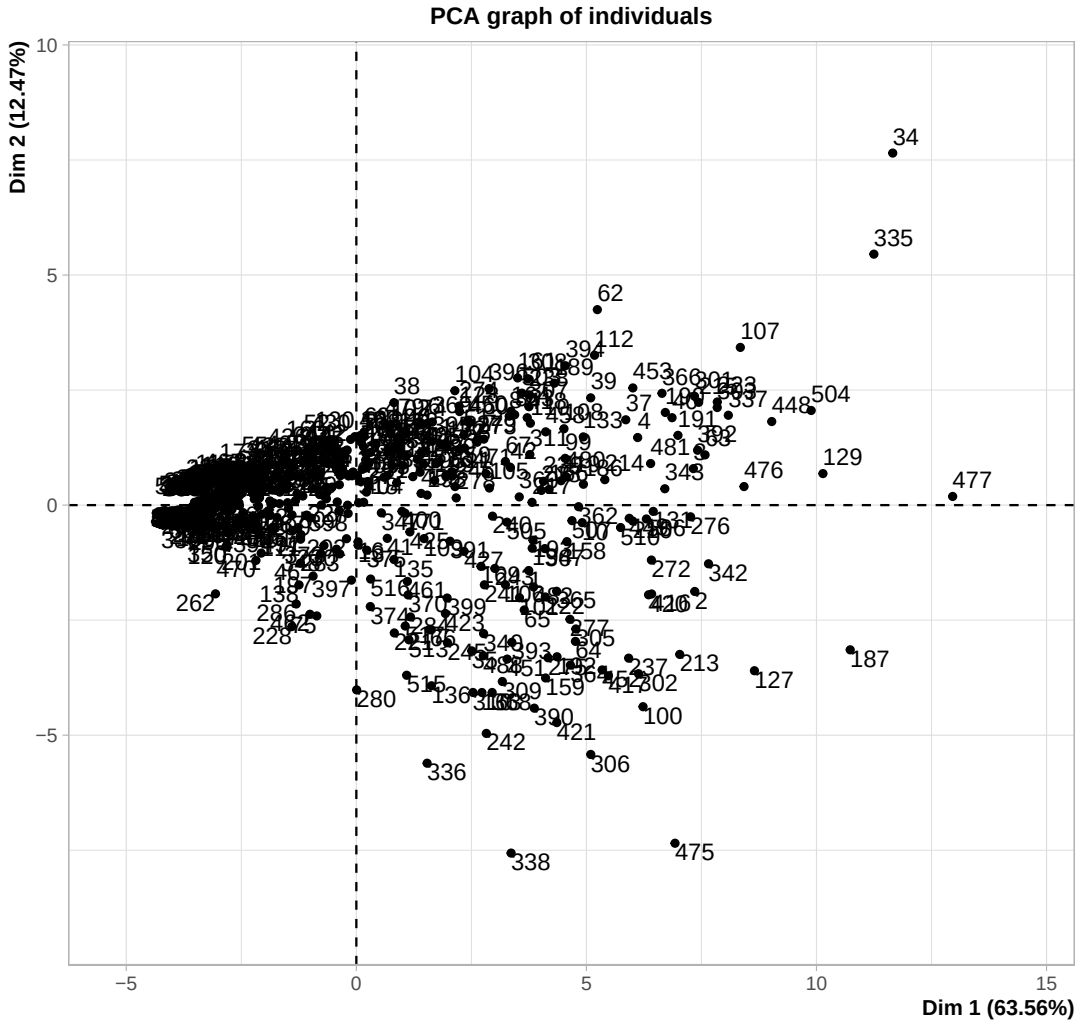

FIGURE B.14. PCA for time defensive variables of the season 2021/2022

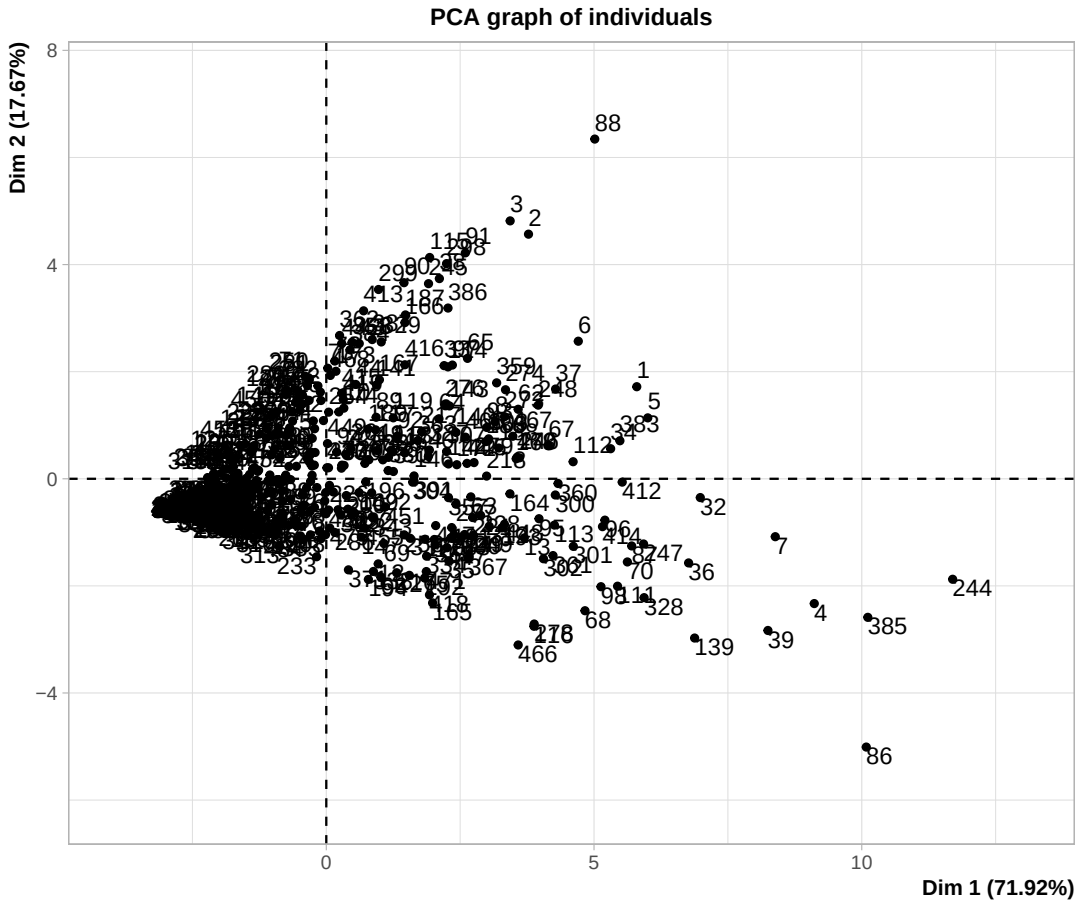

FIGURE B.15. PCA for ball possession variables of the season 2023/2024

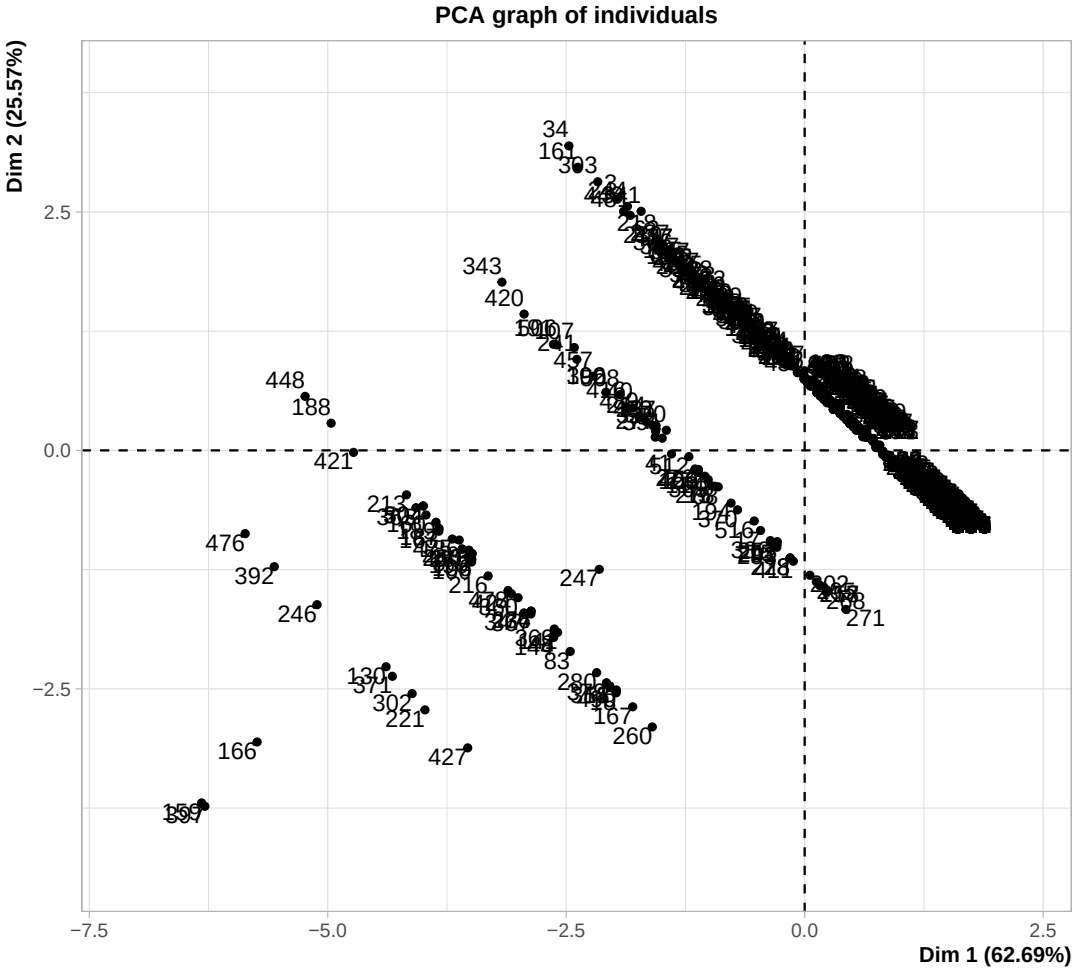

FIGURE B.16. PCA for disciplinary variables of the season 2021/2022

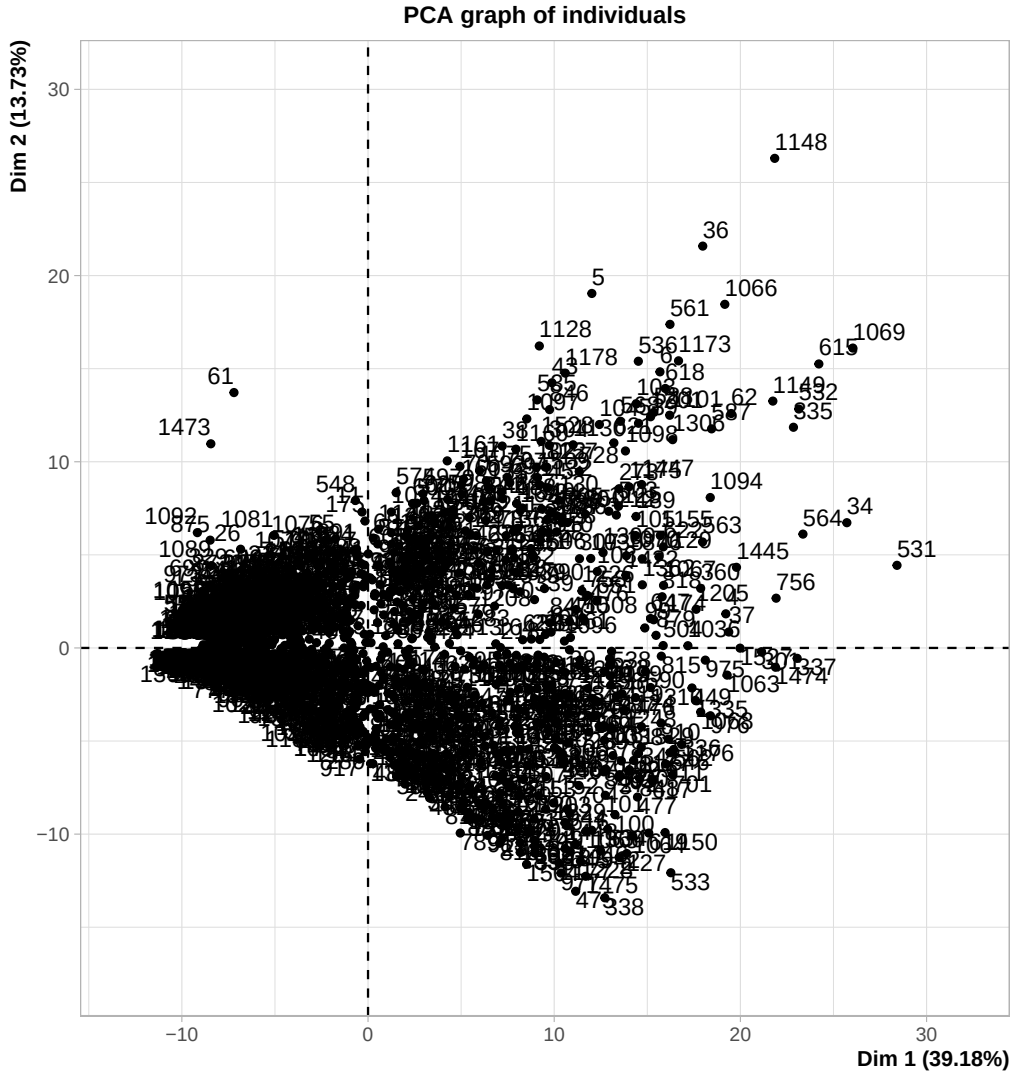

FIGURE B.17. PCA for all variables of all seasons

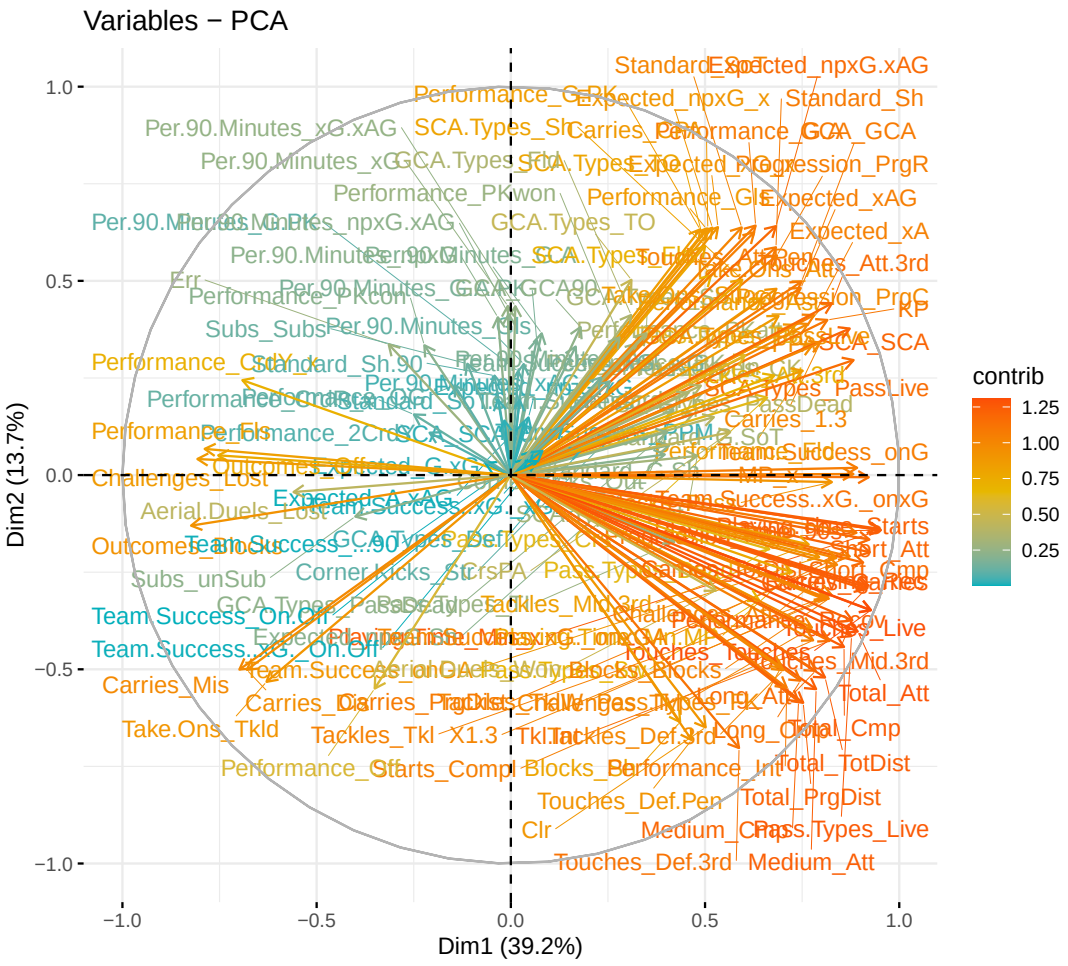

FIGURE B.18. Variables influence on PC1 and PC2

TABLE B.1. Variables influence on PCA

	Dim.1	Dim.2
MP_x	0.885271	0.009362
Playing.Time_Starts	0.942761	-0.14846
Playing.Time_Min_x	0.950045	-0.14974
Playing.Time_90s	0.950095	-0.14966
Performance_Gls	0.518262	0.629278
Performance_Ast	0.643554	0.423562
Performance_G.A	0.646477	0.628071
Performance_G.PK	0.48277	0.638311
Performance_PK	0.374003	0.285145
Performance_PKatt	0.391617	0.323686
Performance_CrdY_x	-0.69043	0.215745
Performance_CrdR_x	-0.2347	0.185858
Expected_xG_x	0.541818	0.63077
Expected_npxG_x	0.514643	0.641546
Expected_xAG	0.762201	0.436031
Expected_npxG.xAG	0.698541	0.633603
Progression_PrgC	0.778739	0.351176
Progression_PrgP	0.889894	-0.15517
Progression_PrgR	0.697977	0.529886
Per.90.Minutes_Gls	0.220189	0.624132
Per.90.Minutes_Ast	0.135519	0.344383
Per.90.Minutes_G.A	0.227626	0.621521
Per.90.Minutes_G.PK	0.170529	0.600286
Per.90.Minutes_G.A.PK	0.195454	0.603821
Per.90.Minutes_xG	0.008388	0.470041
Per.90.Minutes_xAG	0.005571	0.214838
Per.90.Minutes_xG.xAG	0.010169	0.481863
Per.90.Minutes_npxG	-0.02797	0.439998
Per.90.Minutes_npxG.xAG	-0.01503	0.460324
Standard_Sh	0.683402	0.590509
Standard_SoT	0.593647	0.647107
Standard_Sh.90	-0.01501	0.513085
Standard_SoT.90	0.000938	0.443181
Standard_G.Sh	0.404263	0.10946
Standard_G.SoT	0.477806	0.221674
Standard_FK	0.456932	0.161575
Expected_npxG.Sh	0.382444	0.089359
Expected_G.xG	-0.11388	-0.05496
Expected_np.G.xG	-0.08848	-0.00996
Total_Cmp	0.865297	-0.42463
Total_Att	0.897816	-0.38953
Total_TotDist	0.817944	-0.50277
Total_PrgDist	0.762396	-0.57081
Short_Cmp	0.91408	-0.2199
Short_Att	0.929277	-0.19585
Medium_Cmp	0.747664	-0.56369
Medium_Att	0.792117	-0.53591

TABLE B.2. Variables influence on PCA

	Dim.1	Dim.2
Long_Cmp	0.75818	-0.51287
Long_Att	0.801393	-0.49485
Expected_xA	0.799401	0.368954
Expected_A.xAG	-0.14349	0.014524
KP	0.826041	0.366508
X1.3	0.815495	-0.32308
PPA	0.790326	0.283963
CrsPA	0.598188	0.060088
Pass.Types_Live	0.8823	-0.39871
Pass.Types_Dead	0.681274	-0.20183
Pass.Types_FK	0.708676	-0.50206
Pass.Types_TB	0.61098	0.230455
Pass.Types_Sw	0.662975	-0.32862
Pass.Types_Crs	0.682963	0.152592
Pass.Types_TI	0.444676	-0.18601
Pass.Types_CK	0.490323	0.221881
Corner.Kicks_In	0.457741	0.217515
Corner.Kicks_Out	0.412107	0.129643
Corner.Kicks_Str	0.23665	0.098893
Outcomes_Off	-0.75756	0.070284
Outcomes_Blocks	-0.83291	-0.10681
SCA_SCA	0.882528	0.336173
SCA_SCA90	0.00473	0.283862
SCA.Types_PassLive	0.891023	0.257063
SCA.Types_PassDead	0.524626	0.110838
SCA.Types_TO	0.484506	0.556264
SCA.Types_Sh	0.572833	0.514554
SCA.Types_Fld	0.527726	0.478167
SCA.Types_Def	0.511182	0.054433
GCA_GCA	0.744608	0.488741
GCA_GCA90	0.180979	0.434112
GCA.Types_PassLive	0.722328	0.377896
GCA.Types_PassDead	0.4035	0.056652
GCA.Types_TO	0.296659	0.484521
GCA.Types_Sh	0.319918	0.365553
GCA.Types_Fld	0.311687	0.420436
GCA.Types_Def	0.158585	0.063383
Tackles_Tkl	0.844686	-0.29109
Tackles_TklW	0.828032	-0.30338
Tackles_Def.3rd	0.727452	-0.49462
Tackles_Mid.3rd	0.78971	-0.16488
Tackles_Att.3rd	0.695765	0.26346
Challenges_Tkl	0.808714	-0.3583
Challenges_Att	0.865498	-0.20863
Challenges_Lost	-0.83346	0.03098
Blocks_Blocks	0.815268	-0.35553
Blocks_Sh	0.432494	-0.61766

TABLE B.3. Variables influence on PCA

	Dim.1	Dim.2
Blocks_Pass	0.849418	-0.07825
Tkl.Int	0.835722	-0.41203
Clr	0.488526	-0.63262
Err	-0.25398	0.320561
Touches_Touches	0.928933	-0.33965
Touches_Def.Pen	0.456196	-0.65992
Touches_Def.3rd	0.600669	-0.67684
Touches_Mid.3rd	0.880772	-0.37468
Touches_Att.3rd	0.836339	0.408528
Touches_Att.Pen	0.633315	0.632146
Touches_Live	0.928688	-0.34027
Take.Ons_Att	0.708989	0.43197
Take.Ons_Succ	0.717381	0.368138
Take.Ons_Tkld	-0.65204	-0.48289
Carries_Carries	0.922681	-0.28368
Carries_TotDist	0.928647	-0.16543
Carries_PrgDist	0.882582	-0.19731
Carries_1.3	0.821631	0.291
Carries_CPA	0.502873	0.627251
Carries_Mis	-0.69524	-0.50218
Carries_Dis	-0.68602	-0.4633
Receiving_Rec	0.923371	-0.27741
Playing.Time_Mn.MP	0.716263	-0.24936
Starts_Compl	0.772398	-0.42529
Subs_Subs	-0.10844	0.328331
Subs_unSub	-0.40182	-0.09642
Team.Success_PPM	0.129581	0.124689
Team.Success_onG	0.887706	0.022556
Team.Success_onGA	0.784243	-0.24657
Team.Success_...	0.292406	0.25869
Team.Success_...90	0.154	0.143783
Team.Success_On.Off	0.103127	0.064503
Team.Success..xG._onxG	0.924661	-0.00529
Team.Success..xG._onxGA	0.828163	-0.23958
Team.Success..xG._xG...	0.270351	0.274057
Team.Success..xG._xG...90	0.091551	0.194893
Team.Success..xG._On.Off	0.023603	0.120633
Performance_2CrdY	-0.21578	0.1217
Performance_Fls	-0.805	0.05307
Performance_Fld	0.769855	0.227654
Performance_Off	-0.32917	-0.56597
Performance_Int	0.725251	-0.53616
Performance_PKwon	0.304233	0.407715
Performance_PKcon	-0.20804	0.335705
Performance_OG	-0.14678	0.248734
Performance_Recov	0.911531	-0.31176
Aerial.Duels_Won	0.542617	-0.29947
Aerial.Duels_Lost	-0.54466	-0.06977

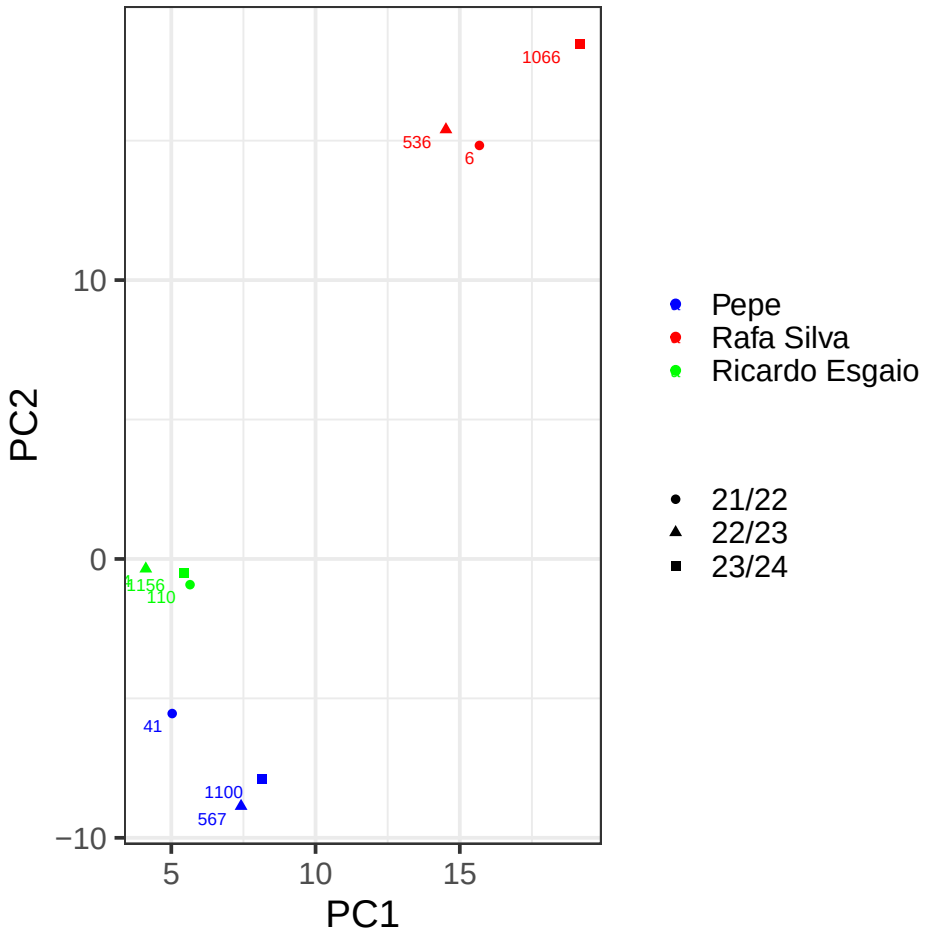

FIGURE B.19. PCA for the season evolution of Pepe, Rafa Silva and Ricardo Esgaio

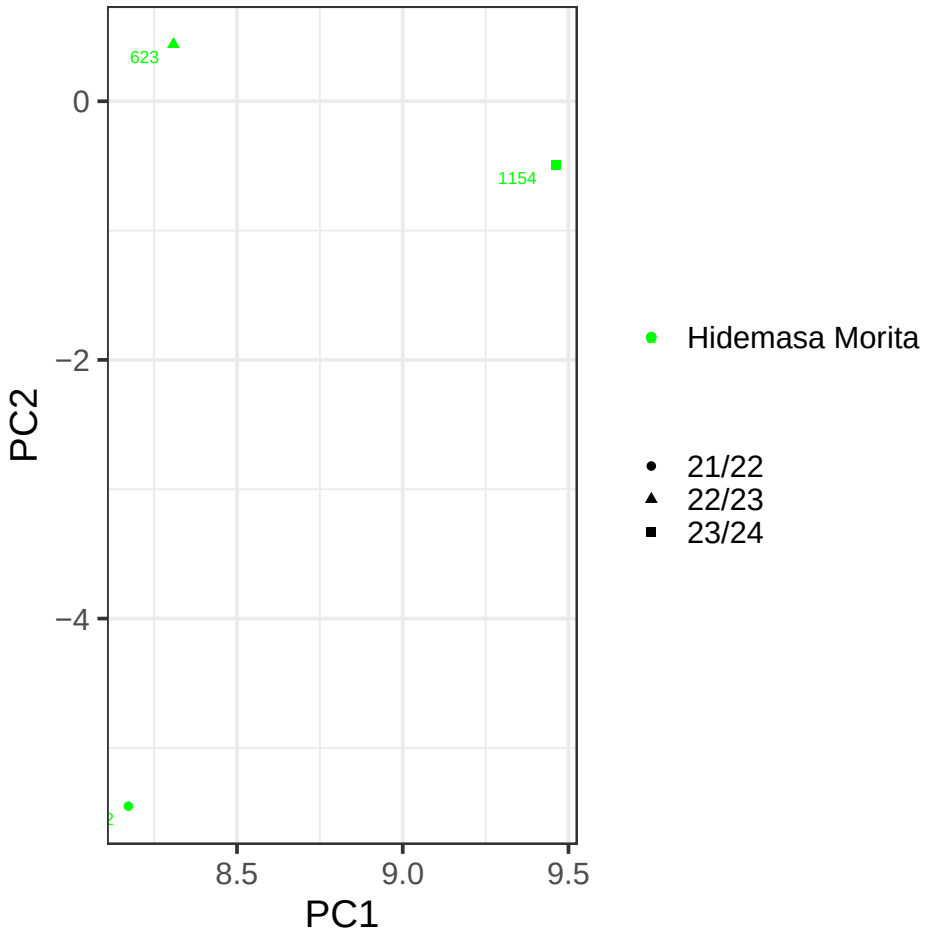

FIGURE B.20. PCA for the season evolution of Hidemassa Morita

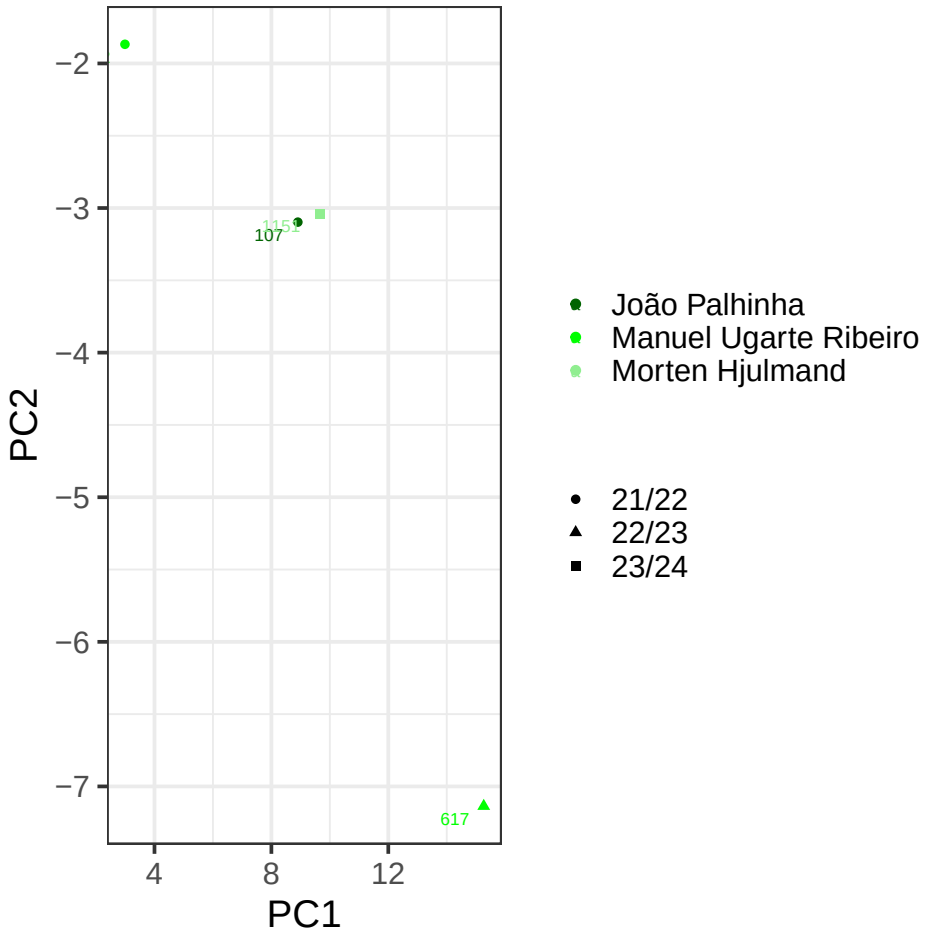

FIGURE B.21. PCA for the analysis of Sporting CP defensive midfielders

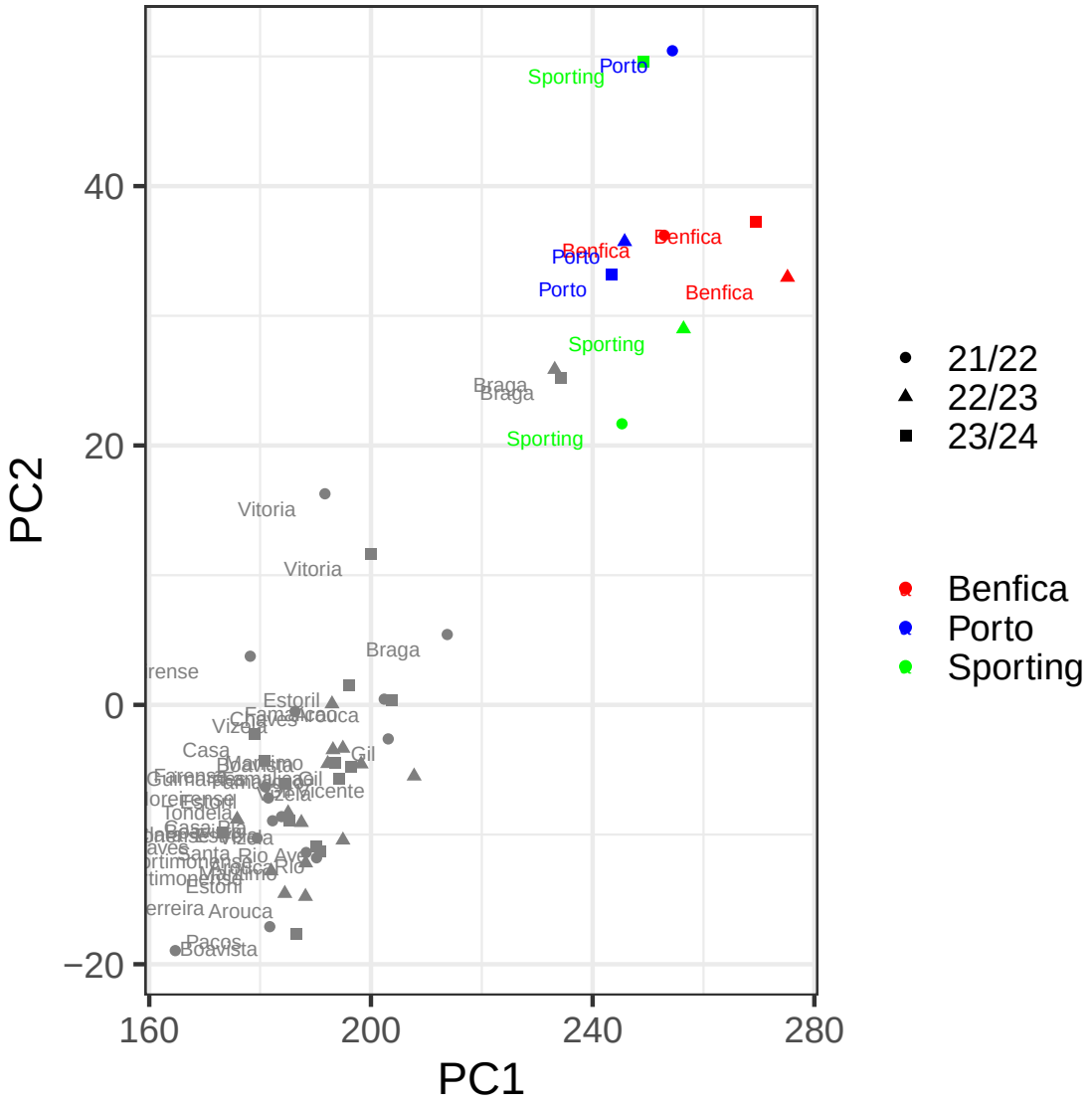

FIGURE B.22. PCA for team analysis by season