NOVA IMS Information Management School

# MEGI

Master Degree Program in

**Statistics and Information Management**

# Exploratory research about playing styles and performance patterns in football teams

Gonçalo Ribeiro Ramos

Dissertation

presented as partial requirement for obtaining the Master Degree Program in **Statistics and Information Management**

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# EXPLORATORY RESEARCH ABOUT PLAYING STYLES AND PERFORMANCE PATTERNS IN FOOTBALL TEAMS

By

Gonçalo Ribeiro Ramos

Master Thesis presented as partial requirement for obtaining the Master's degree in Statistics and Information Management with a specialization in Marketing Research and CRM

**Advisors:** Vítor Duarte dos Santos, NOVA IMS

João António Araújo Barroso, Sport Lisboa e Benfica

Catarina Paisana Pires Costa das Neves, NOVA IMS

November 2022

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 16th of November 2022

# Acknowledgements

Quero começar por agradecer aos meus orientadores, o Prof. Vítor Santos, ao João Barroso e à Prof. Catarina Neves, pelo apoio incansável e disponibilidade ao longo de todo o desenvolvimento do projeto.

Ao Sport Lisboa e Benfica pela oportunidade e pela disponibilidade ao longo de todo o processo.

À minha família, nomeadamente aos pais e ao Afonso pelas palavras de conforto e força ao longo dos últimos meses.

À Mariana, pelo apoio e suporte constante que permitiu que acreditasse em mim e no projeto.

Aos meus amigos, pela ajuda e apoio.

Ao meu avô, esta é por ti.

# Abstract

The sports industry has grown exponentially over the years and is becoming an increasingly attractive market due to high levels of investment. In this way, sports institutions seek to obtain a competitive advantage both outside and inside the playing field. One of the strategic approaches to obtain this competitive advantage is through the study of performance data collected from athletes and teams. In this study, through exploratory research, a performance analysis model is developed where the main objective is the identification of various styles of play and team-level performance patterns regarding technical aspects of the games across the Bundesliga during the 2020/2021 season. The analysis model consisted of unsupervised methods with the application of a Factor Analysis that allowed the characterization of teams in terms of playing styles and the identification of various performance patterns through the application of clustering algorithms.



**Keywords:** Football; Sports Analytics; Match Statistics; Event Data; Performance Patterns; Playing Styles.

# INDEX

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **BMG** | Borussia M'gladbach |
| **BSC** | Hertha BSC |
| **BVB** | Borussia Dortmund |
| **DSC** | Arminia Bielefeld |
| **FCA** | Augsburg |
| **FCB** | Bayern München |
| **FCU** | Union Berlin |
| **KMO** | Kaiser-Meyer-Olkin |
| **KOE** | Köln |
| **LDA** | Linear Discriminant Analysis |
| **LEV** | Bayer Leverkusen |
| **MO5** | Mainz 05 |
| **MSA** | Measure of Sampling Adequacy |
| **PCA** | Principal Components Analysis |
| **RBL** | RB Leipzig |
| **SCF** | Freiburg |
| **SGE** | Eintracht Frankfurt |
| **SO4** | Schalke 04 |
| **SVW** | Werder Bremen |
| **TSG** | Hoffenheim |
| **VFB** | Stuttgart |
| **WOB** | Wolfsburg |

# 1 INTRODUCTION

## 1.1 CONTEXT

As everyone knows, sports are an essential endeavor and represent a considerable part of many people's lives. Even people not connected with the sports environment and professional sports could feel sports as a way of escaping their daily problems. Currently, they can put aside all their dilemmas and situations that cause them more anxiety (Brymer & Schweitzer, 2013). Sport has enormous importance in people's mental and physical health, and some people find their balance in practicing some sport to exercise, improve their health and lifestyle, or keep track of their favorite team (Martinez Arastey Guillermo, 2013)). But at the end of the day, it's all about the passion for sports.

Watching and keeping track of professional sports is a significant activity shared by young and adult individuals (Morgulev et al., 2018). This fact often leads sports fans to question and focus on details related to athletes' performance or coach decisions trying to justify the results or harmful situations negatively. These so-called "details" can consist of reflections on the coach's decisions or comparing players' metrics to predict outcomes of games and the final ranks of individuals and teams playing in competitions. And that's here when passion comes in and is when love leads to money.

Major Sports involve extensive media coverage, which enhances the spread of these high-competition sports—promoting sport and attracting potential investors who invest billions of dollars in the sports industry. The sports media coverage goes around the advertising and broadcasting licenses representing a considerable part of a sports organization's profitability (Guillermo Martinez Arastey, 2018).

The rise of lucrative financial opportunities in most major sports due to the growth of revenue from broadcasting deals and the increase in the offer of streaming platforms have taken the process of preparing high-performance athletes to another level (Morgulev et al., 2018). The world of high competition is currently facing a revolution in professionalization and research in most high-competition sports. The records and differences between the results obtained by athletes have been less expressive, where trivial details could make a

significant difference in getting the desired results. This fact was verified when analyzing the evolution of the results obtained by Olympic athletes over the years. The margin of these results has been decreasing (Guillermo Martinez Arastey, 2018), which is why athletes, coaches, and federations have reformulated and focused their entire training process based on details (data) that were previously not relevant.

## 1.2 MOTIVATION

Despite recognizing the importance of high-performance data analysis, it is possible to identify a gap in structuring the data analysis architecture where we could find it during data collection and in the a posteriori estimation.

This project focuses on data-driven analysis for quantitative behaviors in team sports. It introduces two main approaches: (1) structuring an architecture for analyzing data collected from targeted teams from a chosen football league and (2) collecting and analyzing teams' data. The first approach involves structuring the various phases of analysis and collecting data used through the performance recorded by high competition teams to allow for a much more focused and simplified analysis. All data obtained are considered relevant, but the fundamental purpose of structuring an analysis architecture is to select/define which features are indicated when performing a given analysis.

The second approach is to apply that architecture to a specific data, and therefore, contribute with a detailed analysis of a particular feature to support coaches' decision.

## 1.3 OBJECTIVES

The objective of the exploratory study is to identify various styles of play performed by teams of a specific football league throughout a pre-defined season and, later, to identify the various performance profiles obtained through the games played by these same teams. In addition, two required methods will be used to develop the project. The application of factor analysis will allow the identification of the styles of play practiced by the teams and, in addition several profiling techniques were performed where some clustering algorithms were proposed on resulting defined data from games played during a specific time window.

1- Previews context around the research topic;
2- Find a proper resource database;
3- Analyze the goodness and fidelity of the obtained data;
4- Definition of the analysis methodology;
5- Analyze the dataset through the analysis proposed;
6- Results interpretation;
7- Present study limitations and future recommendations;

## 1.4 STUDY IMPORTANCE AND RELEVANCE

Consequently, allied to the evolution of new technologies, sports have become more scientific, and sports organizations are increasingly looking for methods to support coaches. Sports performance has been widely discussed and studied, is characterized by complex and dynamic situations that produce a large volume of quantitative and qualitative information (Chen et al., 2010). Derived from the high volume of data, coaches look for ways to process all this information in order to migrate from a subjective assessment of performance to a more objective appraisal (Kirkbride, 2013). This support can be done by monitoring technological devices where it may be possible to identify critical areas through the outputs obtained by high-performance athletes. The essential areas may consist of detailed analyzes of athletes' performance, where coaches and players can identify why and how performance can improve and to make decisions about training to enhance performance (O'Donoghue & Mayes, 2013).

This study will contribute to the optimization and development of data analysis from high-performance teams. The small details represent and contain a vast information potential that is still little explored today. These details are analyzed and scrutinized to contribute to these athletes' exponential evolution and confer important information for the case study. Furthermore, this methodology will entrust coaches with a work tool to identify and analyze some aspects of the teams to enhance their performance. As Keisuki Fuji (2021) said, most data-driven models have non-linear structures and high predictive performances, but it is sometimes hard to interpret them. So, it's crucial to find a way to reduce interpretation errors and build on this process.

In addition to improving technological tools to find the necessary inputs for analyzes, it is crucial to fine-line the entire interpretation process. Deal with the outputs resulting from the studies and then work them in a relevant way. The more tools coaches have at their disposal, the more likely their teams will achieve their goals. Which could mean significant returns on a financial level in case of success in big competitions or home leagues. So, this research can contribute to an improvement in sports analysis and sports environment.

# 2 LITERATURE REVIEW

The importance of this chapter aims to expose the main aspects that are determinants for the evaluation of sports performance. This section will discuss and justify the type of data collected and later, discuss about the various statistical approaches used to carry out the project.

The theme "performance" has already been discussed and deepened in several areas over the last decades. The first record refers to the year 1912, when Fullerton developed the first-hand Notation system for baseball, allowing sports teams to apply the same practice later (M. Hughes & Franks, 2015).

The discussion around the concept of "Performance" aims to identify the reasons that lead individuals and groups to achieve said success and verify which common aspects are determinants to success. According to (Hodges et al., 2012), the methodologies that have been applied to study performance at work and sport are analyzes based on comparisons between cognitive aspects, decision making and actions well-being successes, and less on studies of successful individuals and groups. Performance analysis systems have been developed to fit several purposes: to identify physiological parameters that characterize different team sports, create game models, and identify patterns of play of successful and unsuccessful teams (M. D. Hughes & Bartlett, 2010). In terms of sports performance, more specifically with football, performance translates into various stimuli and physical interactions, techniques, tactical actions, and even movements from all competing players (Bangsbo, 1994; Bradley et al., 2014). Performance analysis is the area of science that focuses all of its analysis on the effective sports performance of athletes and teams rather than self-reports by athletes or activities undertaken in laboratory settings (O'Donoghue, 2010). According to McGary (2009), advanced knowledge of game behaviour is achieved through scientific analysis of sports performance to improve future results. Research involving the analysis of sports performance in training or competition can be considered performance analysis. Various metrics are used to perform measurements through some attributes, for example, heart rate response or blood rate accumulation, which ensure that

these fall under the theme of performance analysis of sports (M. D. Hughes & Bartlett, 2010). The appropriate choice of specific performance indicators is a crucial factor at the time of the Performance Analysis, as it allows the identification of good and bad performances of various individuals or team members through the evaluation of their ratios. As a result, it is possible to compare the performance between individuals and even between teams (M. Hughes & Franks, 2015).

The discussion and application of performance indicators have assumed a more relevant role in sports science, where these same performance indicators are applied to identify, characterize, and enhance the training methodologies (M. D. Hughes & Bartlett, 2010) and contribute to improvements for coaches and athletes (Carling et al., 2009). More specifically, in football, the analysis provides data inputs during games by time-motion, notational analysis (M. Hughes & Franks, 2015) that allows the identification of key performance indicators and allows the application of modelling methods that will later be useful as support to define playing style and can even help in the re-design of training exercises (Manuel Clemente et al., 2018).

Many variants dictate the performance of high competition athletes. As Gai D. (2019) indicated after their study "Physical, Technical and Tactical Performance Analysis of technical, tactical and physical variables collected teams from Chinese football super League", the limited number of physical-related differences between teams means that the analysis of football parameters is not just about physical aspects. It is a set of interactions between performance indicators (physical, tactical and technical). However, the inclusion of all variables in the same analysis may not be the most indicated due to the different types of data collected from each theme, contributing to a distortion of the results obtained later.

Currently, most studies based on the observation of sports performance in football focus primarily on the comparison between physical attributes and a few technical performance attributes. But few studies have investigated the technical performance attributes of professional football athletes (Rampinini et al., 2009; Russell & Kingsley, 2011). Technical actions can provide better predictors of football success and be more accurate than physical

attributes (Bradley et al., 2014; Bush et al., 2015; Castellano et al., 2012; Lago-Peñas et al., 2010, 2011; Rampinini et al., 2009). Clemente et al. (2016) emphasized that very few studies focus on analyzing and discussing game patterns and technical performance indicators necessary for success.

This way, deepening the study of technical actions could determine success in various sports (Di Salvo et al., 2007). In 2019, Gai D, in his research, stated that the analysis of technical performance indicators confers a core help to physical trainers, coaches and performance analysts in improving performance through the team patterns previously defined based on these same technical performance indicators.

Currently, it is notorious that the style of play of the teams developed over the years has undergone evolutionary changes. For that same reason, football ceased to depend only on tactical and physical aspects and technical elements that end up working as differentiating aspects.

The data relating to technical actions can be used in two moments: preparation of the pre-competition and post-competition athletes. In the pre-preparation process, trainers and analysts use data from technical actions to change and improve the training methodology. The analysis thus allows scrutinizing the various techniques developed in order to be able to specify some preponderant details for the development of the work method of the teams (Russell & Kingsley, 2011). To reveal new trends in football performance, the development of technical performance profiles is an essential task in order to contribute to improving task representativeness in practice sessions and during the process of selecting the most appropriate players for each match scenario (Liu et al., 2013).

Football has evolved a lot at a tactical, technical, and physical level over the years and it is clear that football increasingly requires a more significant physical demand. This aspect can be explained by the investment in different training approaches developed by coaches in each game preparation. Due to the importance given to the training methodologies, the players reveal a better preparation for the physical demands inherent to the sports season, which raises the level of play of each team (Bangsbo et al., 2006). However, in addition to

being physically demanding, players must have a high level of technical aspects to stand out from the crowd and play at a highly competitive level.

Accordingly, several technical researchers present variables in common in their analyzes where they have been developing and discussing, considering the results obtained. The variables from these studies tend to have more significance and are helpful for studying playing style patterns. However, as Kuhn et al. (2005) stated the styles of play appear to have changed less than what was expected because performance indicators are influenced by playing styles in elite football and identifying specific key indicators or applying different exploratory methods to deepen knowledge about the theme "playing styles", may provide information for coaches and could help them to reprioritize their training and game approaches.

Those technical variables are extremely useful because they allow the collective and specific understanding of a particular team and could provide a better understanding of teams' performance profiles (Rein & Memmert, 2016). In this way, the trainers can focus their attention on better developing the variable that has more statistical relevance (Gai, 2019). As such, over the last few years, some authors, such as Castellano et al., (2012), Marcelino et al., (2011), Taylor et al., (2008) have been referring to similar technical variables throughout their studies that describe explanatory variables to performance profiles. Namely, variables include ball possession, dribbling, shots on goal, crosses, passes, successful passes, dribbling with success, or corner kicks. For example, several studies evidence ball possession as an important key indicator, where teams that promote their style of play based on this key indicator, are more successful throughout their sporting season compared to other styles of play. Churchill et al. (2005) stated that through the analysis of the performances obtained by teams in some competitions, the teams that kept the ball for longer ended up creating offensive moments of great danger for the opponents and, consequently, having more opportunities to be successful concerning the result. end of the game. This fact was observed by Gómez et al. (2012), after analyzing the main Spanish football league "La Liga", where he identified that one of the key elements for teams to have more opportunities to be more successful would be through the enhancement of the

“Ball possession” style of play. The teams that started their game with the recovery of ball possession in the defensive half, organized their transitions to the offensive midfield through long ball possessions and penetrative passes into the decision area, significantly increasing the number of shots and consequently the possibility of making a goal. Fernandez-Navarro et al. (2016) specified in the article “Attacking and defensive styles of play in football: analysis of Spanish and English elite teams” that both La Liga and Premier League had very similar playing styles suggesting that teams that explored ball possession work were more successful sporty. Another interesting finding is obtained by interpreting the results obtained by Liu et al. (2016) that events such as tackles, aerial duels could represent a less controlled style of play. At a practical level, this may reflect that teams, by basing their style of play on previously mentioned events, may lead to more unstable performances and represent a high level of volatility in terms of the outcome of each game.

The theme of performance indicators associated with styles of play differs among analysts depending on the approach and interpretation that each one has. There is no consensus on which performance indicators are crucial for certain styles of play. For example, Hughes and Franks (2007) stated that low passing sequences are key performance indicators that describe the style of play, direct play. In contrast, Tenga and Larsen (2017) state that the key performance indicators that represent the “Direct Play” style of play are counter-attacks, attacks with at least one long ball and attacks with a maximum of two passes. Additionally, Fernandez-Navarro et al. (2016) state that the performance indicators for the different styles of play are not defined and should be included in the study, other performance indicators may help in the definition of a certain type of style of play.

Manuel Clemente et al. (2018) indicated that the game taking place outside or at home (Match Location) of one of the teams was a differentiating factor in the value obtained in the technical variables. That is, researchers realized that with the changes in the value of the situational variables, the remaining variables would eventually undergo modifications in different scenarios. Thus, researchers have tried to understand and test the influence of situational variables on other variables, namely technical variables (Gai, 2019). Also, previous studies highlight that the variables of passing, dribbling, and shooting performance

can reduce during and immediately after a simulated football match (Russell & Kingsley, 2011).

The study of key performance indicators in football involves the analysis of the interactions that occur between the defence and the attack, and these interactions are difficult to interpret at the performance level without considering the interactions of opposing teams (Fernandez-Navarro et al., 2016). However, the deepening and understanding of the diversity of a specific league will help coaches and technical teams to develop the key performance indicators identified to work team performances and players for optimal success within the league (Gai, 2019).

# 3 METHODOLOGY

To achieve the goal of this study, a process model with Six phases that describe the data Science life cycle will be used to define the various phases of a data science project. The process model will be the CRISP-DM - Cross Industry Standard Process for data mining that is one of the most conventional knowledge discovery processes developed by a European funded consortium (Martınez-Plumed, 2019).

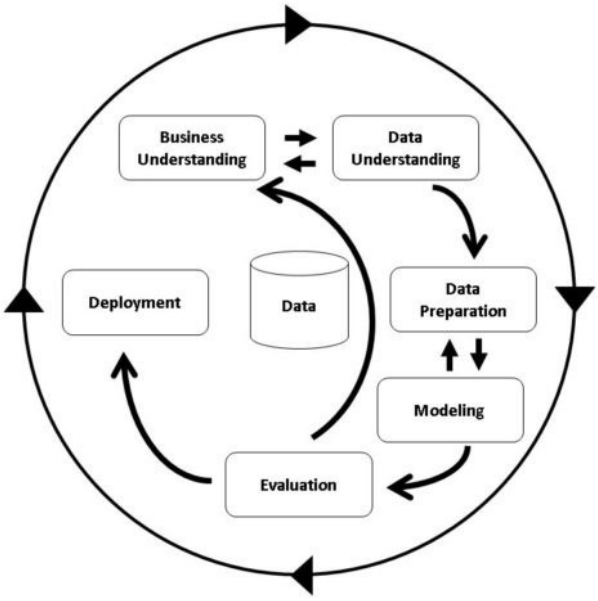



***Figure 1 – CRISP_DM project phases***
***(Source: https://www.datenbankenverstehen.de/lexikon/crisp-dm/)***

The application of this model followed a methodology to respond to and fulfill a series of steps inherent to the model. This research approach focusses on unsupervised Machine Learning (ML) techniques to combine descriptive and predictive models. The methodological approach was divided into three phases: (1) data preparation, (2) data selection, and (3) unsupervised ML. Initially, the first step in developing a project was deepening and understanding the surrounding theme around the object of study. This phase is given the context and presents the importance of this project, where it will serve to produce a project plan and define data mining goals. Subsequently, evaluating the necessary material for developing the object under study and its typology is of paramount

importance. The perception of how all the previous variables were collected and extracted from the dataset was highly relevant for further analysis. This step involves the identification of relationships and data quality verification of the data through the exploratory study.

In the data preparation phase, a series of procedures involved the treatment and cleaning of the dataset were done to prepare the dataset for further analysis, which involved the identification of the missing values and outliers, the evaluation of the typology of each variable's output and discuss the possibility of applying a standardization process to improve the quality of analysis output that used to be influenced when data are expressed in different scales.

After the conclusion of the data preparation, the second phase (data selection) performed the feature selection to improve the dataset performance and eliminate data redundancy. To improve the performance of the dataset, aggregations of the variables were performed by their respective summations, where it was possible to subsequently obtain the total of events (variables) that occurred in each game during the previously defined season. Additionally, some variables did not justify their use individually and were aggregated into formulas to constitute a more relevant variable. Also, an analysis of extreme outliers was applied to treat the noise and errors for every set of features. Having in mind that this outlier treatment could never affect the entire dataset because it could risk losing valuable information for other features.

At an early stage, aggregation and correlation analysis methods were applied to prepare the entire dataset for further analysis. A Correlation Analysis becomes helpful in exploring the association between variables and identifying the level of multicollinearity and mediating status of independent variables in a model (Senthilnathan Samithambe, 2019).

In the third stage of methodology (3), the development of the entire process is based on two statistical themes: Factor Analysis and Cluster Analysis.

Factor Analysis

The Factor Analysis was fundamentally aimed to reduce the size of the original data contributing to the simplification of the data, such as reducing the number of variables in regression models. Factor Analysis consists in a multivariate technique that analyzes underlying patterns of complex and multidimensional phenomena (Hair, 2011) emphasizing the existence of an underlying correlation structure relative to the date used in the analysis in which it will be possible to calculate the number of factors. After checking the possibility of running this analysis on the dataset, the next step will be to choose the least number of factors that will explain the correlation between variables and preserve a satisfactory amount of information from the original data. The interpretation and selection of the ideal number of factors are achieved by considering several indicators such as factors loadings and eigenvalues, i.e., correlation with original (contributing) variables (Cruz-Jesus et al., 2016).

Cluster analysis

The Cluster Analysis will be the theme where we will base all assumptions and draw conclusions from the results obtained from the various clustering profiling methods. The main objective of these methods is to classify different objects into groups in a way that the similarity between two objects is maximal if they belong to the same group and minimal otherwise. While factor analysis is considered to be a dimensional reduction technique, cluster analysis is also a reduction method but applies over observations Through these methods, it was possible to verify the patterns obtained as an output and thus interpret them to identify the differentiating aspects between teams and games from the same team, and consider the results obtained in the games held over a defined season. To this end, two utterly different clustering approaches were adopted to obtain greater validation of the results obtained. Both were compared and analyzed.

After the conduction of the literature review with several authors to evaluate the most appropriate techniques and methods to apply, the (1) K-means method and the (2) SOM (self-organizing maps) were the chosen ones. For the first one, (1) K-means was selected because is one of the most know cluster methods ever used and applied for cluster techniques. The unsupervised method K-means has the main objective function of sum the squared Euclidean distance between each data point and find their nearest cluster center (Bação et al., 2005). The second method was SOM where the main objective is to map the data patterns onto a n-dimensional grid of neurons and units. SOM or Self-Organizing Map is a type of neural network used all over the past three decades through cluster analysis but comparing to K-means method, can be less prone to local optima than k-means (de Bodt et al., 1999).

The (1) K-Means Method aims to put data points with similar characteristics in the same cluster and separate data points with different characteristics into different clusters. All is done by minimizing the intra-cluster variance because minimizing the *SSW* (within-cluster sums of squares) will necessarily maximize the *SSB* (Between-cluster sums of squares). During the application of this method, it was decided to choose a *k* centroid based on several criteria. The *k* can be assigned to the centroids. Each centroid is a data point representing a cluster's center, meaning that all data points around each cluster will be assigned to the nearest centroid. The quality of the cluster assignments is determined by evaluating the sum of squared errors that will directly influence the sum of the squared distances (WCSS). Based on these indications, the results obtained from the inertia and *WCSS* values were evaluated for the choice of *k* clusters. Which as possible to have a visual confirmation through the "Elbow Method" Graphic.

| Basic K-Means algorithm |
| --- |
| 1: Select $k$ points as initial centroids. |
| 2: **repeat**. |
| 3: Form $k$ clusters by assigning each point to its closest centroid. |
| 4: Recompute the centroid of each cluster. |
| 5: **until** Centroids do not change. |

***Figure 1 - K-means algorithm (Source: Steinbach and Kumar 2006)***

The second applied method was (2) SOM is an unsupervised learning computational method belonging to the field of artificial neural networks (Haykin, 1999). It is often used to group sets of data observations according to their similarity in exploratory analysis. As k-means, the idea is to maximize the intra-cluster distances and minimize inter-cluster distances. SOM can dimensionally organize complex data into clusters according to their relations.

The SOM's algorithm is considered simple, where the method requires only the input parameters being ideal for problems whose patterns are unknown and indeterminate. Its structure comprises a single-layer linear 2D grid of neurons instead of a series of layers. All the nodes on this grid are connected directly to the input vector but not to one another, meaning the nodes do not know the neighbors' values and only update the weight of their connections as a function of the given inputs. All the process stops when the weighted average over the Euclidean norms of the difference between the input vector and the corresponding best matching unit. The final results consist of a descriptive model that considers how the input space is structured and projects it into a lower-dimensional space.

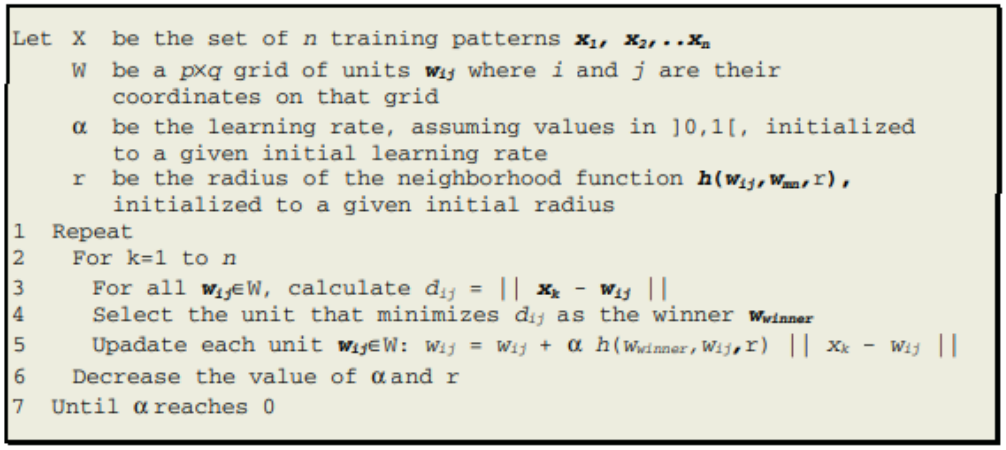



***Figure 2 - Basic SOM training algorithm (Source: Bação et al., 2005 )***

For the elaboration of this project, the programs Microsoft SQL Server Management Studio application and Spyder were used to perform all the treatment and processing of the dataset and the development of the exploratory data analysis where unsupervised methods were performed. Also, the Power BI tool was handy for better understanding and interpreting the output provided by the previously applied analysis.

# 4 DATA ANALYSIS

## 4.1 SAMPLE AND SUBJECT

Match Statistics of *306* games from the Bundesliga for the 2020/2021 season, obtained from Benfica's Sports Data Science Department through external databases provided by Stats Perform (a sports-related data and content company). Regarding the provided data, there are two types of game data types: tracking (data measured in real-time of the players and the ball) and event (location of the event and associated players in the defined period).

The tracking data is all data measured in real-time by players and the ball that provides analysts information about the spatial location of players, referees, and ball. This type of data is obtained through a system of optical tracking and broadcast systems[1].

The event data describes specific events in a game; in other words, every move that occurs on the pitch is recorded and considered event data. Usually, these data are manually recorded through broadcasting and provide the details of players' location with the ball during the specific event (i.e., pass, shot, take on).

The analysis involves *181* game event data/KPIs (Key Performance Indicators) previously analyzed and equated. All the game event variables are considered as technical performance variables, and they were gathered and went through their data-collecting methods previously validated. Later, they will be used in the analyzes with the final aim of responding to the object under study, which will identify patterns of each team in terms of the style of play and performance.

[1] camera systems with computer vision models that are installed in the stadium

Due to the abundance of KPIs in the database provided, a filter was applied to reduce the number of KPIs under study. In this way, *181* KPIs were reduced to *48* KPIs, and the respective grouping of variable predictors in the following themes:

- Variables related to defending;
- Variables related to organizing and passing;
- Variables related to attacking.

Throughout the citations and based upon prior studies, all these variables were analyzed and considered as valid performance indicators of match technical variables in football (Castellano et al., 2012; Lago-Peñas et al., 2010, 2011; Liu et al., 2013, 2015).

| Groups | Variable ID | Modified Variables |
|---|---|---|
| **Variables related to defending** | **V1** | Clearance |
| | **V14A** | Defensive Duels |
| | **V14B** | Aerial Defensive Duels |
| | **V12** | Foul |
| | **V2** | Interception |
| | **V3** | Recovery |
| | **V17B** | Save_1vs1 |
| | **V17A** | Save |
| | **V4** | Tackle |
| | **V18A** | Goalkeeper Smother |
| | **V18C** | Goalkeeper Short and Medium Pass |

| | | |
|---|---|---|
| | **V18D** | Goalkeeper Long Hand pass |
| | **V18B** | Goalkeeper Kick Hands |
| **Variables related to organising and passing** | **V5B** | Ball Conduction 20 meters |
| | **V5D** | Ball Conduction 20 meters2 |
| | **V5A** | Ball Conduction 5 meters |
| | **V5C** | Ball Conduction 5 meters2 |
| | **V15C** | Area Entries |
| | **V6E** | Construction Delayed Pass |
| | **V6F** | Preparation Delayed Pass |
| | **V6A** | Medium Pass |
| | **V6B** | Short Pass |
| | **V6C** | Long Pass |
| | **V6J** | Flank Variation Pass |
| | **V6G** | Key and Assist Pass |
| | **V6K** | Vertical Pass |
| | **V6M** | Construction Decision Vertical Pass |
| | **V6L** | Construction Preparation Vertical Pass |
| | **V6O** | Preparation Decision Vertical Pass |
| | **V16** | Game Center Variation Reception |
| | **V13A** | TakeOn |
| | **V13C** | Take On Decision |

| | | |
|---|---|---|
| | **V9** | Assist |
| | **V11** | Disposs |
| | **V14E** | Offensive Duel CAM |
| | **V14C** | Offensive Duel |
| | **V14D** | Aeriel Offensive Duel |
| **Variables related with attacking** | **V10** | Goal |
| | **V8A** | Total Shots |
| | **V8B** | Exterior Shot |
| | **V7** | Crosses |
| | **V6H** | Pass Crosses Area |
| | **V19** | Touch Area |
| | **V13B** | Take On CAM |
| | **V6I** | CAM Passes |
| | **V15A** | Entries CAM |
| | **V15B** | Entries Decision |
| | **V6D** | Decision Delayed Pass |

***Table 1 - Selected match performance statistics***

## 4.2 DATA PREPARATION

This section verifies the importance of data preparation for data mining and its application in this project. Zhang et al. (2003) mentioned that real-world data could cause pattern distortions due to missing attribute values, containing errors or outliers, or containing discrepancies in codes or names.

Initially, the question was raised about the approach adopted regarding using original variables or the standardization of the variables under study. The standardization process is usually used when the original variables are expressed on different scales and as our dataset is composed by variables with different scales, the standardization method was applied through the analysis. Although the majority of scales were numeric, the units were different, implying completely different variances (due to scale) and miscalculated distances between observations if original data was used.

At this stage, the *181* KPIs from the original data were analyzed and reduced to reduce the volume of the dataset. After grouping the chosen predictor variables by theme, the data processing process reduced the number of records obtained as output. The strategy adopted was the aggregation of each KPI by their sums, which can significantly improve the efficiency of data mining (Zhang et al., 2003).

The number of records/events obtained was significantly reduced, where the sample size under study was transformed into only 81 records. The records are classified as the set of information taken from the analyzed games that took place throughout the season and in a previously defined league, which will be filtered so that it is possible to study in detail the information collected from each KPI, grouped by teams that participated in that match.

In this phase, after the data cleaning process, where it was verified if there were missing values or "nulls," the *181* KPIs from the original data were analyzed to reduce the volume of the dataset. As previously explained, the aggregation of the KPIs was done by their sums,

which made it possible to reduce the *181* KPIs to just *48*. The formulas chosen to apply the aggregation strategy are presented below:

| **Aggregation Formulas** |
|---|
| **Medium Passes** = (Correct Passes + Wrong Passes) - ((Correct Short Passes + Wrong Short Passes) + (Correct Long Passes + Wrong Long Passes)) |
| **Short Passes** = Correct Short Passes + Wrong Short Passes |
| **Long Passes** = Correct Long Passes + Wrong Long Passes |
| **Decision Delayed Passes** = (Correct Delayed Passe + Wrong Delayed Passe) - ((Correct Construction Delayed Passes + Wrong Construction Delayed Passes) + (Correct Decision Delayed Passes + Wrong Decision Delayed Passes)) |
| **Construction Delayed Passes** = Correct Construction Delayed Passes + Wrong Construction Delayed Passes |
| **Preparation Delayed Passes** = Correct Preparation Delayed Passes + Wrong Preparation Delayed Passes |
| **Passes CAM** = Correct Passes CAM + Wrong Passes CAM |
| **Total Shots** = Shots Net + Shots Out + Shots Block |
| **Take On CAM** = Correct Take On CAM + Wrong Take On CAM |
| **Take On Decision** = Correct Take On Decision + Wrong Take On Decision |

| **Aggregation Formulas – cont.** |
|---|
| **Aerial Offensive Duel** = Correct Aerial Offensive Duel + Wrong Aerial Offensive Duel |
| **Goalkeeper Short and Medium Pass** = (Correct Goalkeeper Short Pass + Wrong Goalkeeper Short Pass) + (Correct Goalkeeper Medium Pass + Wrong Goalkeeper Medium Pass) |
| **Goalkeeper Long Hand Pass** = Correct Goalkeeper Long Hand Pass + Wrong Goalkeeper Long Hand Pass |

***Table 2 - Aggregation Formulas***

## 4.3 DATA SELECTION

This section describes below the variables resulting from previous processes, totalling forty-one technical performance indicators evaluated according to later methods. The operational definitions were collected from the Stats Perform website, simultaneously with inputs collected from the Sports Data Science Department of Sport Lisboa e Benfica.

| **Variable ID** | **Variables** | **Description** |
|---|---|---|
| **V1** | Clearance | Action by a defending player temporarily removes the attacking threat on their goal/that effectively alleviates pressure on their goal. |
| **V2** | Interception | Preventing an opponent's pass from reaching their teammates. |
| **V3** | Recovery | This is where a player recovers the ball in a situation where neither team has possession, or the ball has been played directly to him by an opponent, thus securing possession for their team. |

| Variable ID | Variables | Description |
|---|---|---|
| **V4** | Tackle | A tackle is defined as where a player connects with the ball in a ground challenge where he successfully takes the ball away from the player in possession. |
| **V5A** | Ball Conduction 5 meters | The action of dominating and moving with the ball at ground level, through a succession of touches with any part of the foot. |
| **V5B** | Ball Conduction 20 meters | The action of dominating and moving with the ball at ground level, through a succession of touches with any part of the foot. |
| **V5C** | Ball Conduction 5 meters2 | The action of dominating and moving with the ball at ground level, through a succession of touches with any part of the foot. Without any other interruption event. |
| **V5D** | Ball Conduction 20 meters2 | The action of dominating and moving with the ball at ground level, through a succession of touches with any part of the foot. Without any other interruption event. |
| **V6A** | Medium Pass | Any intentionally played the ball from one player to another. |
| **V6B** | Short Pass | Any intentionally played the ball from one player to another. |
| **V6C** | Long Pass | Any intentionally played the ball from one player to another. |
| **V6D** | Decision Delayed Pass | Passes made to the rear in the Decision zone. |
| **V6E** | Construction Delayed Pass | Passes made to the rear in the Construction zone. |
| **V6F** | Preparation Delayed Pass | Passes made to the rear in the Preparation zone. |
| **V6G** | Key and Assist Pass | The final pass or pass-cum-shot leads to the recipient of the ball having an attempt at goal without scoring. |
| **V6H** | Pass Crosses Area | Any intentionally played the ball from a wide position intending to reach a teammate in a specific area in front of the goal, including passes and crosses. |
| **V6I** | Passes CAM | Any intentionally played the ball from one player to another at the center attacking. |

| Variable ID | Variables | Description |
|---|---|---|
| **V6J** | Flank Variation Pass | Any intentionally played the ball from one player to the opposite flank to the player who makes the pass. |
| **V6K** | Vertical Pass | Any intentionally played the ball from one player to another done vertically. |
| **V6L** | Construction Preparation Vertical Pass | Any intentionally played the ball from one player to another done vertically that starts in the construction zone and ends in the preparation zone. |
| **V6M** | Construction Decision Vertical Pass | Any intentionally played the ball from one player to another done vertically that starts in the construction zone and ends in the decision zone. |
| **V6O** | Preparation Decision Vertical Pass | Any intentionally played the ball from one player to another done vertically that starts in the preparation zone and ends in the decision zone. |
| **V7** | Crosses | Any intentionally played the ball from a wide position intending to reach a teammate in a specific area in front of the goal. |
| **V8A** | Total Shots | A ball kicked or headed by a player at the opponent's net in an attempt to score a goal. |
| **V8B** | Exterior Shot | A ball kicked or headed by a player at the opponent's net in an attempt to score a goal done outside the area. |
| **V9** | Assist | A pass/cross that is instrumental in creating a goal-scoring opportunity, for example, a corner or free-kick to a player who then assists an attempt, a chance-creating through ball or cross into a dangerous position. |
| **V10** | Goal | The whole of the ball passes over the goal line, between the goalposts, and under the crossbar, provided that no offense has been committed by the team scoring the goal. |
| **V11** | Disposs | Possessions are defined as one or more sequences belonging to the same team in a row. A possession is ended by the opposition gaining control of the ball. |
| **V12** | Foul | Any infringement that is penalized as foul play by a referee. |

| Variable ID | Variables | Description |
|---|---|---|
| **V13A** | TakeOn | This is an attempt by a player to beat an opponent when they have possession of the ball. |
| **V13B** | Take On CAM | This is an attempt by a player to beat an opponent when they have possession of the ball in the offensive midfield zone. |
| **V13C** | Take On Decision | This is an attempt by a player to beat an opponent when they have possession of the ball in the decision zone. |
| **V14A** | Defensive Duels | A player in controlled possession of the ball (below elbow height) attempts to pass an opponent trying to dispossess the player in possession. |
| **V14B** | Aerial Defensive Duels | A player in controlled possession of the ball (below elbow height) attempts to pass an opponent, who in turn, is trying to dispossess the player in control in the air. |
| **V14C** | Offensive Duel | A player in controlled possession of the ball (below elbow height) attempts to pass an opponent trying to dispossess the player in possession. |
| **V14D** | Aeriel Offensive Duel | This is where two players challenge in the air against each other |
| **V14E** | Offensive Duel CAM | A player in controlled possession of the ball (below elbow height) attempts to pass an opponent, who in turn, is trying to dispossess the player in control in the offensive center zone. |
| **V15A** | Entries CAM | Passes made or entries the at the center attacking zone. |
| **V15B** | Entries Decision | Passes made or entries to these the decision zone. |
| **V15C** | Area Entries | Passes made or entries inside the area. |
| **V16** | Game Center Variation Reception | A sum of all events where a player receives the ball from a center variation pass. |
| **V17A** | Save | A goalkeeper prevents the ball from entering the goal with any part of his body when facing an intentional attempt from an opposition player. |

| Variable ID | Variables | Description |
|---|---|---|
| **V17B** | Save_1vs1 | A goalkeeper preventing the ball from entering the goal with any part of his body when facing an intentional attempt from an opposition player, including a shot done by the opposition player. |
| **V18A** | Goalkeeper Smother | A goalkeeper who comes out and claims the ball at the feet of a forward gets a smother, similar to a tackle. However, the keeper must hold onto the ball to award a00 smother. |
| **V18B** | Goalkeeper Kick Hands | A goalkeeper prevents the ball from entering the goal with any part of his body when facing an opposition player's intentional attempt, including putting the ball in position and shooting after. |
| **V18C** | Goalkeeper Short and Medium Pass | Any intentional hand pass from the goalkeeper to another with short and medium-range. |
| **V18D** | Goalkeeper Long pass | Any intentional hand pass from the goalkeeper to another with long-range. |
| **V19** | Touch Area | A sum of all events where a player touches the ball excludes things like Aerial lost or Challenge lost. Every area touch is recorded. |

***Table 3 - Variables description***

## 4.4 OUTLIERS TREATMENT

After the end of the previous processes, the time has come to proceed to the analysis of the results of descriptive statistics of the original variable. It is necessary to assess the need to remove the outliers in the original dataset. To this end, the following results presented in table 3 were considered.

Through the analysis of the results, it was noticeable that there were very despair values of the remaining that differ significantly from the other data or observations. For the variables 'v6b' – Short Pass, 'v6c' – Long Pass, 'v6f' – Preparation Delayed Pass was applied a restrictive value for the maximum of each variable, so they cannot distort statistical

analyzes and influence all the assumptions around the analysis. For the applied limit, the value of *950* was considered the maximum score in the variable *'v6b'*. For the applied limit, the value of *950* was considered the maximum score in the variable 'v6b'. As these variables present very similar results scales, some scores of the previously flagged variables ('v6c', 'v6f') were also eliminated after the maximum score limit in the variable 'v6b' was applied.

This limitation reduces the initial dataset by approximately *8%*, which was considered quite acceptable. Although outliers were recorded in several variables, the decision was only to limit the most discrepant variable because even knowing that it will reduce statistical significance, it will not be reasonable to remove valuable information that is part of the study area implies to produce a better fitting model. In addition, after performing the factor analysis, the value of the Rotated Factor Pattern was more expressive without outliers, meaning better interpretability.

| Descriptive Analytics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Variables** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| ***v1*** | 612,00 | 34,32 | 15,85 | 2,00 | 22,00 | 32,00 | 44,00 | 120,00 |
| ***v2*** | 612,00 | 24,76 | 9,10 | 0,00 | 18,00 | 24,00 | 30,00 | 50,00 |
| ***v3*** | 612,00 | 112,76 | 17,53 | 60,00 | 100,00 | 114,00 | 126,00 | 168,00 |
| ***v4*** | 612,00 | 30,83 | 10,01 | 6,00 | 24,00 | 30,00 | 38,00 | 64,00 |
| ***v5a*** | 612,00 | 108,54 | 44,53 | 18,00 | 74,00 | 102,00 | 140,00 | 252,00 |
| ***v5b*** | 612,00 | 10,65 | 5,99 | 0,00 | 6,00 | 10,00 | 14,00 | 30,00 |
| ***v6g*** | 612,00 | 18,58 | 8,42 | 0,00 | 12,00 | 18,00 | 24,00 | 50,00 |
| ***v6h*** | 612,00 | 56,04 | 23,20 | 4,00 | 40,00 | 54,00 | 68,00 | 142,00 |
| ***v6k*** | 612,00 | 550,58 | 103,20 | 298,00 | 472,00 | 545,00 | 618,00 | 896,00 |
| ***v7*** | 612,00 | 47,85 | 19,37 | 4,00 | 36,00 | 46,00 | 58,00 | 134,00 |
| ***v8b*** | 612,00 | 8,61 | 4,81 | 0,00 | 6,00 | 8,00 | 12,00 | 32,00 |
| ***v9*** | 612,00 | 2,13 | 2,18 | 0,00 | 0,00 | 2,00 | 4,00 | 14,00 |
| ***v10*** | 612,00 | 2,94 | 2,55 | 0,00 | 2,00 | 2,00 | 4,00 | 16,00 |
| ***v11*** | 612,00 | 17,71 | 7,47 | 0,00 | 12,00 | 18,00 | 22,00 | 44,00 |
| ***v12*** | 612,00 | 24,71 | 7,32 | 6,00 | 20,00 | 24,00 | 30,00 | 50,00 |
| ***v13a*** | 612,00 | 33,72 | 12,86 | 6,00 | 24,00 | 32,00 | 42,00 | 88,00 |
| ***v14a*** | 612,00 | 123,25 | 23,34 | 64,00 | 108,00 | 122,00 | 138,00 | 206,00 |
| ***v14c*** | 612,00 | 128,46 | 24,33 | 64,00 | 112,00 | 128,00 | 144,00 | 210,00 |
| ***v14b*** | 612,00 | 33,42 | 14,45 | 2,00 | 24,00 | 32,00 | 42,00 | 118,00 |

| Variables | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| ***v14d*** | 612,00 | 33,27 | 14,49 | 2,00 | 24,00 | 32,00 | 40,00 | 114,00 |
| ***v15a*** | 612,00 | 164,13 | 46,03 | 70,00 | 130,00 | 161,00 | 194,00 | 330,00 |
| ***v15b*** | 612,00 | 126,64 | 44,86 | 22,00 | 96,00 | 118,00 | 150,00 | 312,00 |
| ***v15c*** | 612,00 | 28,29 | 14,04 | 2,00 | 18,00 | 26,00 | 36,00 | 76,00 |
| ***v16*** | 612,00 | 5,45 | 4,12 | 0,00 | 2,00 | 4,00 | 8,00 | 22,00 |
| ***v17a*** | 612,00 | 6,07 | 3,99 | 0,00 | 4,00 | 6,00 | 8,00 | 22,00 |
| ***v17b*** | 612,00 | 0,24 | 0,73 | 0,00 | 0,00 | 0,00 | 0,00 | 4,00 |
| ***v18a*** | 612,00 | 0,11 | 0,46 | 0,00 | 0,00 | 0,00 | 0,00 | 2,00 |
| ***v18b*** | 612,00 | 2,96 | 3,62 | 0,00 | 0,00 | 2,00 | 4,00 | 20,00 |
| ***v19*** | 612,00 | 77,89 | 31,82 | 16,00 | 54,00 | 74,00 | 96,00 | 186,00 |
| ***v5c*** | 612,00 | 36,32 | 10,14 | 10,00 | 28,00 | 36,00 | 42,00 | 72,00 |
| ***v5d*** | 612,00 | 4,91 | 3,45 | 0,00 | 2,00 | 4,00 | 6,00 | 18,00 |
| ***v6b*** | 612,00 | 656,83 | 196,22 | 210,00 | 508,00 | 637,00 | 782,00 | 1424,00 |
| ***v6a*** | 612,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| ***v6c*** | 612,00 | 321,95 | 71,36 | 152,00 | 269,50 | 316,00 | 368,50 | 558,00 |
| ***v6d*** | 612,00 | 60,37 | 31,19 | 2,00 | 38,00 | 54,00 | 78,00 | 200,00 |
| ***v6e*** | 612,00 | 133,66 | 44,17 | 36,00 | 102,00 | 130,00 | 162,50 | 284,00 |
| ***v6f*** | 612,00 | 187,35 | 90,61 | 32,00 | 117,50 | 172,00 | 240,50 | 626,00 |
| ***v6i*** | 612,00 | 138,61 | 22,79 | 76,00 | 124,00 | 138,00 | 154,00 | 218,00 |
| ***v6j*** | 612,00 | 7,30 | 4,86 | 0,00 | 4,00 | 6,00 | 10,00 | 26,00 |
| ***v6l*** | 612,00 | 111,41 | 22,27 | 46,00 | 96,00 | 112,00 | 128,00 | 192,00 |
| ***v6m*** | 612,00 | 12,50 | 9,25 | 0,00 | 6,00 | 10,00 | 16,00 | 62,00 |
| ***v6o*** | 612,00 | 96,89 | 29,81 | 30,00 | 76,00 | 94,00 | 112,50 | 216,00 |
| ***v8a*** | 612,00 | 42,67 | 12,64 | 8,00 | 34,00 | 42,00 | 50,00 | 82,00 |
| ***v13b*** | 612,00 | 23,60 | 11,00 | 2,00 | 16,00 | 22,00 | 30,00 | 80,00 |
| ***v13c*** | 612,00 | 15,58 | 8,59 | 0,00 | 10,00 | 14,00 | 20,00 | 60,00 |
| ***v14e*** | 612,00 | 96,14 | 24,02 | 38,00 | 80,00 | 96,00 | 110,50 | 190,00 |
| ***v18c*** | 612,00 | 0,01 | 0,11 | 0,00 | 0,00 | 0,00 | 0,00 | 2,00 |
| ***v18d*** | 612,00 | 0,96 | 1,64 | 0,00 | 0,00 | 0,00 | 2,00 | 12,00 |

***Table 4 - Descriptive Analytics of the original dataset***

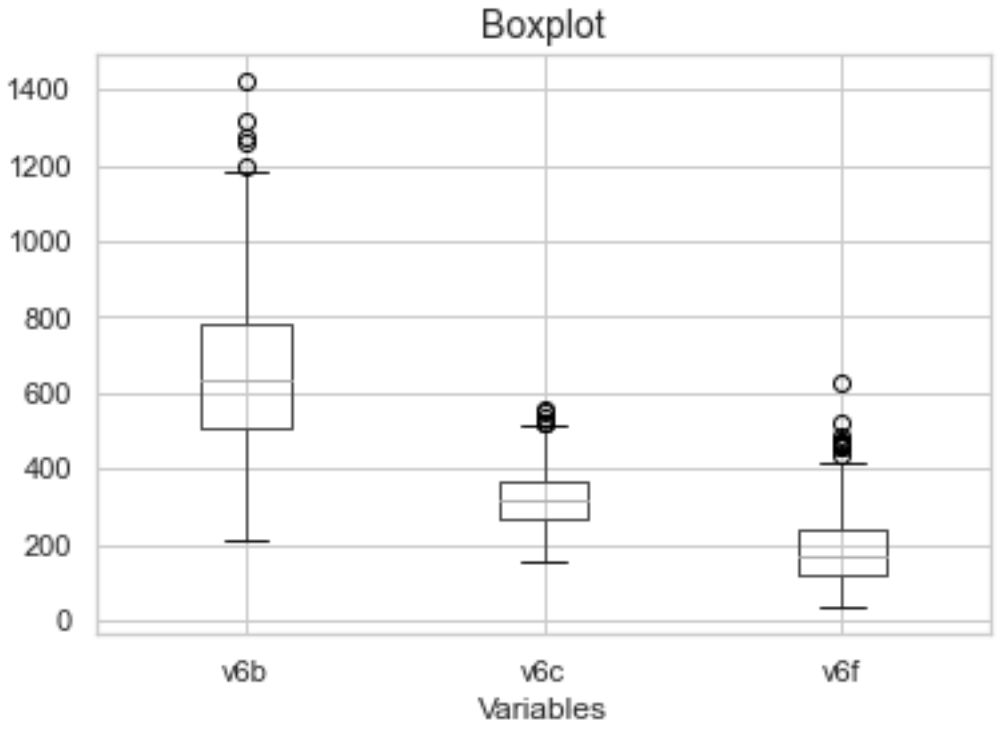



***Figure 4 - Boxplot with original dataset***

| Descriptive Analytics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Variables** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| *v1* | 564,00 | 35,65 | 15,58 | 4,00 | 24,00 | 34,00 | 46,00 | 120,00 |
| *v2* | 564,00 | 25,16 | 9,01 | 0,00 | 20,00 | 24,00 | 30,00 | 50,00 |
| *v3* | 564,00 | 112,09 | 17,63 | 60,00 | 100,00 | 112,00 | 124,00 | 168,00 |
| *v4* | 564,00 | 31,13 | 10,08 | 6,00 | 24,00 | 30,00 | 38,00 | 64,00 |
| *v5a* | 564,00 | 102,67 | 40,11 | 18,00 | 70,00 | 98,00 | 132,00 | 226,00 |
| *v5b* | 564,00 | 10,37 | 5,87 | 0,00 | 6,00 | 10,00 | 14,00 | 30,00 |
| *v6g* | 564,00 | 17,96 | 8,16 | 0,00 | 12,00 | 16,00 | 22,00 | 50,00 |
| *v6h* | 564,00 | 53,42 | 21,26 | 4,00 | 38,00 | 52,00 | 66,00 | 142,00 |
| *v6k* | 564,00 | 534,53 | 89,59 | 298,00 | 464,00 | 534,00 | 604,00 | 756,00 |
| *v7* | 564,00 | 46,26 | 18,43 | 4,00 | 34,00 | 44,00 | 56,00 | 134,00 |
| *v8b* | 564,00 | 8,48 | 4,82 | 0,00 | 4,00 | 8,00 | 12,00 | 32,00 |
| *v9* | 564,00 | 2,08 | 2,16 | 0,00 | 0,00 | 2,00 | 4,00 | 14,00 |
| *v10* | 564,00 | 2,88 | 2,51 | 0,00 | 2,00 | 2,00 | 4,00 | 16,00 |
| *v11* | 564,00 | 17,53 | 7,47 | 0,00 | 12,00 | 16,00 | 22,00 | 44,00 |

| Variables | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| *v12* | 564,00 | 24,88 | 7,26 | 6,00 | 20,00 | 24,00 | 30,00 | 50,00 |
| *v13a* | 564,00 | 33,28 | 12,81 | 6,00 | 24,00 | 32,00 | 40,00 | 88,00 |
| *v14a* | 564,00 | 124,90 | 23,00 | 70,00 | 110,00 | 124,00 | 138,00 | 206,00 |
| *v14c* | 564,00 | 127,96 | 24,70 | 64,00 | 112,00 | 127,00 | 142,50 | 210,00 |
| *v14b* | 564,00 | 33,85 | 14,56 | 4,00 | 24,00 | 32,00 | 42,00 | 118,00 |
| *v14d* | 564,00 | 33,66 | 14,64 | 2,00 | 24,00 | 32,00 | 42,00 | 114,00 |
| *v15a* | 564,00 | 157,32 | 40,09 | 70,00 | 126,00 | 156,00 | 186,00 | 290,00 |
| *v15b* | 564,00 | 120,12 | 38,79 | 22,00 | 94,00 | 116,00 | 142,00 | 290,00 |
| *v15c* | 564,00 | 26,97 | 13,20 | 2,00 | 18,00 | 26,00 | 34,00 | 76,00 |
| *v16* | 564,00 | 5,33 | 4,06 | 0,00 | 2,00 | 4,00 | 8,00 | 22,00 |
| *v17a* | 564,00 | 6,30 | 3,97 | 0,00 | 4,00 | 6,00 | 8,00 | 22,00 |
| *v17b* | 564,00 | 0,22 | 0,68 | 0,00 | 0,00 | 0,00 | 0,00 | 4,00 |
| *v18a* | 564,00 | 0,11 | 0,46 | 0,00 | 0,00 | 0,00 | 0,00 | 2,00 |
| *v18b* | 564,00 | 3,16 | 3,69 | 0,00 | 0,00 | 2,00 | 4,00 | 20,00 |
| *v19* | 564,00 | 74,81 | 29,95 | 16,00 | 54,00 | 70,00 | 92,00 | 180,00 |
| *v5c* | 564,00 | 36,41 | 10,22 | 10,00 | 28,00 | 36,00 | 42,00 | 72,00 |
| *v5d* | 564,00 | 4,90 | 3,49 | 0,00 | 2,00 | 4,00 | 6,00 | 18,00 |
| *v6b* | 564,00 | 623,10 | 162,31 | 210,00 | 496,00 | 612,00 | 756,50 | 950,00 |
| *v6a* | 564,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| *v6c* | 564,00 | 316,16 | 68,28 | 152,00 | 265,50 | 310,00 | 360,00 | 558,00 |
| *v6d* | 564,00 | 55,63 | 25,74 | 2,00 | 38,00 | 50,00 | 72,00 | 142,00 |
| *v6e* | 564,00 | 133,38 | 43,67 | 36,00 | 102,00 | 130,00 | 162,00 | 284,00 |
| *v6f* | 564,00 | 171,66 | 72,92 | 32,00 | 114,00 | 164,00 | 222,00 | 458,00 |
| *v6i* | 564,00 | 137,33 | 22,05 | 76,00 | 122,00 | 137,00 | 152,00 | 200,00 |
| *v6j* | 564,00 | 7,18 | 4,79 | 0,00 | 4,00 | 6,00 | 10,00 | 26,00 |
| *v6l* | 564,00 | 112,99 | 21,22 | 46,00 | 98,00 | 112,00 | 128,00 | 192,00 |
| *v6m* | 564,00 | 13,12 | 9,32 | 0,00 | 8,00 | 10,00 | 18,00 | 62,00 |
| *v6o* | 564,00 | 92,87 | 26,38 | 30,00 | 74,00 | 90,00 | 106,00 | 182,00 |
| *v8a* | 564,00 | 42,23 | 12,61 | 8,00 | 34,00 | 42,00 | 50,00 | 82,00 |
| *v13b* | 564,00 | 22,94 | 10,76 | 2,00 | 16,00 | 22,00 | 28,00 | 80,00 |
| *v13c* | 564,00 | 14,90 | 8,22 | 0,00 | 8,00 | 14,00 | 20,00 | 60,00 |
| *v14e* | 564,00 | 95,05 | 24,07 | 38,00 | 78,00 | 94,00 | 110,00 | 190,00 |
| *v18c* | 564,00 | 0,00 | 0,08 | 0,00 | 0,00 | 0,00 | 0,00 | 2,00 |
| *v18d* | 564,00 | 1,01 | 1,68 | 0,00 | 0,00 | 0,00 | 2,00 | 12,00 |

***Table 5 - Descriptive Analytics of the dataset after outliers' removal***

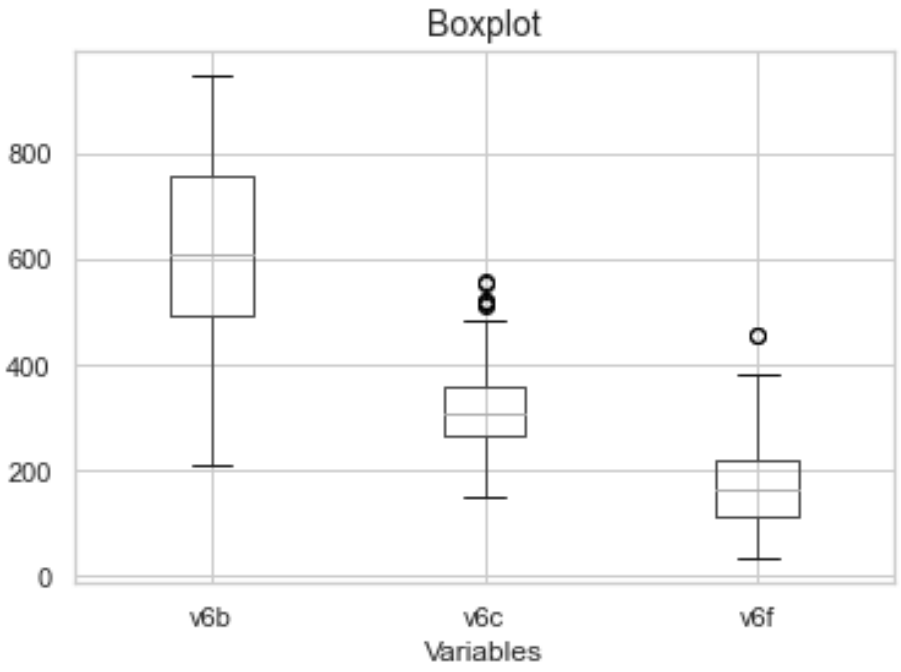



***Figure 5 - Boxplot after outliers' removal***

In addition, if the removed data is specifically analyzed, we conclude that *48* records were removed from the original dataset corresponding to the scores obtained by nine teams belonging to the Bundesliga. The table below shows the frequency obtained:

| **Outliers Frequency** | |
|---|---|
| **Teams** | **Count of Games per team** |
| Bayer Leverkusen | 6 |
| Bayern München | 6 |
| Borussia Dortmund | 11 |
| Borussia M'gladbach | 5 |
| Eintracht Frankfurt | 4 |
| Freiburg | 1 |
| Hoffenheim | 1 |
| RB Leipzig | 11 |
| Stuttgart | 3 |

***Table 6 - Outliers' Frequency***

In conclusion, it was decided not to proceed with the simultaneous analysis of the outliers due to the possibility of being necessary to include more clusters than the intended ones due to the discrepancy of the scores that led to considering the performances of the teams represented in the table above as outliers.

## 4.5 STANDARDIZATION PROCESS

With the variables chosen and data cleaned of missing values, it is necessary to ensure a few steps before proceeding to the Factor Analysis. Standardized data is the process of putting all the variables with equal importance. It is usually used when the original variables are expressed on different scales or when considering the location and variance of the variables that are irrelevant to the analysis. In this case, the variables have different scales, so it is necessary to proceed with a standardized data technique to combat the excessive production of amounts of multicollinearity. During this step, the doubt of when it is the ideal moment to apply the standardization method comes up. However, the decision ended in applying the standardization after outliers' removal because otherwise, different will end up with different standard variables. Also, variables with values of *0* were verified during this step, which would be irrelevant for the performed analyzes. As they did not represent any record, the following variables were removed from the dataset:

| **Removed Variables** | |
|---|---|
| *v6a* | Medium Pass |
| *v18c* | Goalkeeper Short and Medium Pass |
| *v18a* | Goalkeeper Smother |
| *v17b* | Save_1vs1 |

***Table 7 - Removed variables with 0 values***

## 4.6 FACTOR ANALYSIS

After completing the previous step, its crucial to evaluate the level of correlation between each variable. A Factor Analysis focuses on identifying the factors for the correlation between indicators. Moreover, to perform a Factor Analysis, it was essential to define which technique would be used. We tried different approaches, such as a PCA (Principal Components Analysis or LDA (Linear Discriminant Analysis), but the chosen one was a Factor Analysis with a Varimax rotation. The justification for choosing to perform a Varimax rotation (also called Kaiser-Varimax rotation) on this dataset is because this technique will maximize the sum of the variance of the squared loadings to clarify the relationship among factors.

Initially, the Correlation Matrix was performed to identify which pair of indicators had a high correlation and which had a lower correlation. Moreover, it is possible to observe that some variables present low values but only a few present values below *0.35*. This means that variables have moderate and vigorous values, and we have conditions to proceed with the factor analysis.

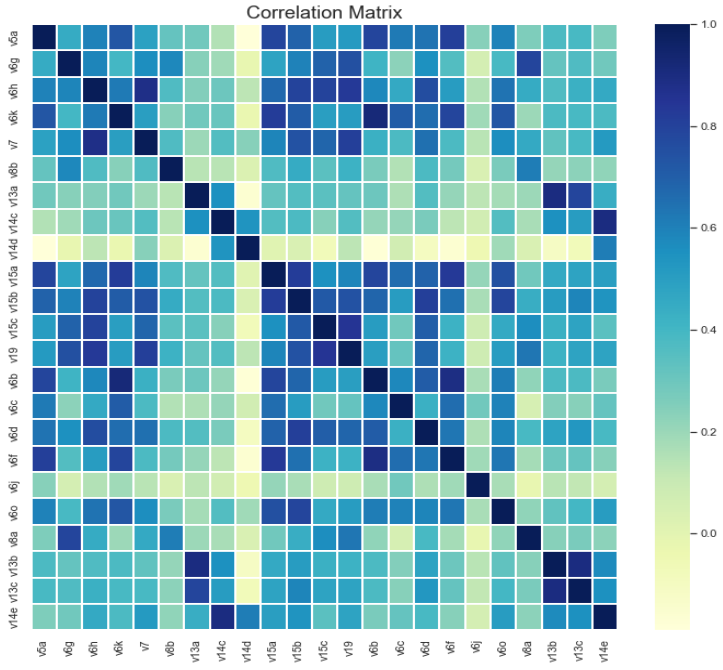



***Figure 8 - Correlation Matrix***

After that, the exploratory analysis continued verifying the sampling adequacy by applying the KMO (Kaiser-Meyer-Olkin) method. The KMO test supports our analysis by testing how suited the data is for Factor Analysis. Some insights about the correlation between variables can be assessed through the KMO test, where the sampling adequacy for each variable in the model and the complete model are measured (Cruz-Jesus et al., 2016). Based on that, it is possible to verify that the initial Overall MSA is *0.78*. However, when we performed further evaluations of the loadings and individual KMO values, it was possible to verify that many presented values below *0.5*. Even so, the analysis was conducted to understand whether removing these same variables was justified (Jolliffe, 2005).

The next step was to choose how many factors we wanted to retain analysis to facilitate understanding of the phenomenon while keeping as much information as possible. For this process, the choice will typically be made based on the values obtained through the Eigenvalues of the Correlation Matrix table, as shown below. According to the results of the eigenvalues[2] and considering the Scree Plot Method, the analysis continues with a 12-factor solution from the beginning to evaluate if each variable has representability in at least one factor.

Through the solution of twelve factors, it was possible to verify the existence of variables that did not present representation in any factor, also because these variables were not extremely relevant. Therefore, it was better to remove them and have a solution with lower number of factors. Which meant the removal of variables that did not present loadings > *0.5* and reducing the number of factors of the presented solution. Additionally, as we can see, it is not acceptable to include variables that obtained low values regarding KMO and Communalities values. To continue the analysis, it is crucial to find an ideal solution because factors will define and explain the correlations between variables. This decision is based on three criteria (Shirazi et al., 2010):

[2] Eigenvalues represent the total amount of variance that can be explained by a given factor.

- Pearson’s criteria – cumulative variance should reach 80% or more.
- Kaiser’s criteria – eigenvalues should have a value of 1 or higher (since we performed FAwith standardized data).
- Scree Plot’s criteria – analyzing the graph by looking at the elbow.

| | Eigenvalues |
|---|---|
| *0* | 12,68 |
| *1* | 3,81 |
| *2* | 3,45 |
| *3* | 2,47 |
| *4* | 2,04 |
| *5* | 1,80 |
| *6* | 1,60 |
| *7* | 1,44 |
| *8* | 1,37 |
| *9* | 1,29 |
| *10* | 1,15 |
| *11* | 1,03 |
| *12* | 0,98 |
| *13* | 0,96 |
| *14* | 0,87 |
| *15* | 0,72 |
| *16* | 0,67 |
| *17* | 0,64 |
| *18* | 0,53 |
| *19* | 0,53 |
| *20* | 0,43 |
| *21* | 0,39 |
| *22* | 0,37 |
| *23* | 0,35 |
| *24* | 0,30 |
| *25* | 0,24 |
| *26* | 0,20 |
| *27* | 0,20 |
| *28* | 0,17 |
| *29* | 0,16 |
| *30* | 0,16 |
| *31* | 0,13 |
| *32* | 0,12 |
| *33* | 0,11 |
| *34* | 0,10 |
| *35* | 0,09 |
| *36* | 0,08 |
| *37* | 0,08 |
| *38* | 0,07 |
| *39* | 0,07 |
| *40* | 0,06 |
| *41* | 0,06 |

| Eigenvalues – cont. | |
|---|---|
| *42* | 0,02 |
| *43* | 0,00 |

***Table 8 - Eigenvalues of the correlation matrix***

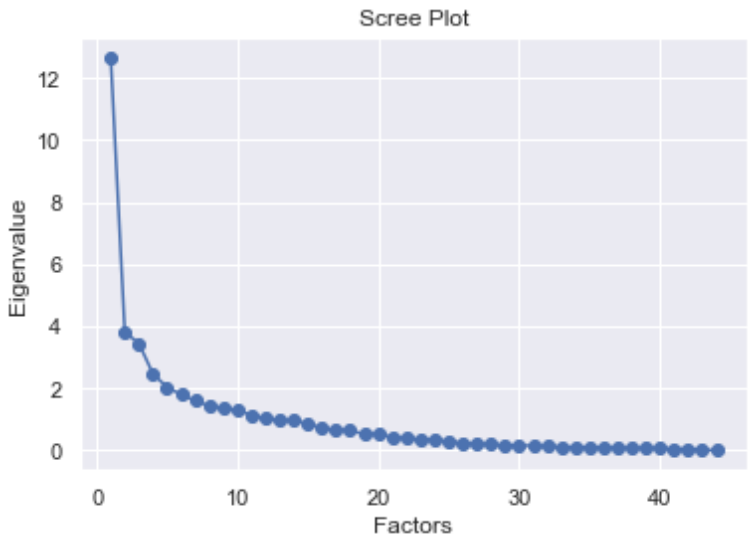



***Figure 7 - Scree Plot***

| Variables | | Factors | | | | | | | | | | | | KMO | Communalities |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | | |
| v1 | Clearance | -0,46846 | -0,13987 | -0,10987 | -0,17503 | -0,16649 | 0,22314 | -0,00007 | 0,00342 | 0,30007 | -0,03494 | -0,00952 | 0,05634 | 0,959 | 0,449 |
| v2 | Interception | -0,24793 | -0,03042 | -0,07479 | 0,00257 | -0,11337 | 0,06823 | 0,06336 | 0,03250 | -0,05091 | 0,00506 | 0,23082 | 0,04187 | 0,852 | 0,145 |
| v3 | Recovery | 0,26085 | 0,06527 | 0,06673 | 0,11858 | 0,09314 | -0,04411 | 0,08675 | -0,05800 | -0,00437 | 0,18178 | 0,46247 | 0,17436 | 0,904 | 0,457 |
| v4 | Tackle | -0,18304 | 0,05680 | -0,03901 | -0,02227 | 0,04385 | 0,08658 | 0,02882 | -0,04614 | 0,07745 | -0,06258 | 0,61020 | -0,05693 | 0,447 | 0,322 |
| v5a | Ball Conduction 5 meters | 0,75580 | 0,17939 | 0,16413 | -0,08833 | 0,17514 | -0,23647 | 0,01924 | 0,11209 | -0,10143 | -0,01195 | -0,20359 | 0,05950 | 0,953 | 0,803 |
| v5b | Ball Conduction 20 meters | 0,30945 | 0,17711 | 0,11553 | -0,18737 | 0,27847 | -0,13083 | 0,05985 | 0,11313 | -0,04198 | 0,03121 | -0,15650 | 0,20043 | 0,925 | 0,295 |
| v6g | Key and Assist Pass | 0,34524 | 0,09932 | 0,72097 | 0,05758 | 0,24464 | -0,08973 | 0,25793 | 0,00093 | -0,07954 | -0,03486 | 0,04459 | 0,11996 | 0,883 | 0,815 |
| v6h | Pass Crosses Area | 0,60642 | 0,12166 | 0,29942 | 0,11097 | 0,63976 | -0,06558 | 0,03188 | 0,05204 | -0,08484 | -0,08827 | 0,02909 | -0,00253 | 0,929 | 0,899 |
| v6k | Vertical Pass | 0,89898 | 0,14795 | 0,13294 | 0,04196 | 0,04427 | -0,13987 | 0,00620 | 0,04536 | -0,05355 | 0,23502 | 0,04847 | -0,01647 | 0,714 | 0,915 |
| v7 | Crosses | 0,47021 | 0,09893 | 0,31825 | 0,21764 | 0,65404 | -0,04833 | -0,04345 | 0,06925 | -0,05953 | -0,10627 | -0,01071 | 0,04859 | 0,916 | 0,807 |
| v8b | Exterior Shot | 0,22894 | 0,04944 | 0,57134 | 0,04564 | 0,05551 | -0,06534 | -0,06483 | -0,02852 | -0,03810 | -0,11737 | -0,07122 | 0,11447 | 0,867 | 0,434 |
| v9 | Assist | 0,04648 | 0,03916 | 0,08974 | -0,07742 | 0,01901 | -0,00788 | 0,91751 | -0,02555 | -0,05706 | -0,03457 | 0,07382 | 0,04423 | 0,595 | 0,877 |
| v10 | Goal | 0,04920 | 0,04278 | 0,10563 | -0,05904 | 0,04579 | 0,00675 | 0,92160 | -0,03964 | -0,05082 | -0,00439 | 0,02996 | 0,07893 | 0,624 | 0,875 |
| v11 | Disposs | 0,15687 | 0,16159 | 0,01222 | 0,27600 | -0,11826 | -0,10574 | -0,02808 | -0,00093 | -0,07410 | 0,05747 | 0,21009 | 0,15705 | 0,581 | 0,183 |
| v12 | Foul | -0,03440 | -0,03126 | -0,05747 | 0,12899 | -0,03216 | -0,02995 | 0,00666 | -0,02950 | 0,34035 | -0,11580 | -0,00875 | 0,02456 | 0,335 | 0,137 |
| v13a | TakeOn | 0,11001 | 0,88304 | 0,07540 | 0,09132 | 0,01896 | -0,07630 | 0,03083 | 0,04599 | -0,05357 | 0,07760 | 0,09879 | 0,20090 | 0,715 | 0,878 |
| v14a | Defensive Duels | -0,31444 | 0,03252 | -0,02507 | -0,05624 | -0,02682 | 0,09359 | -0,03766 | -0,01468 | 0,79682 | -0,01050 | 0,51159 | -0,04157 | 0,614 | 1,018 |
| v14c | Offensive Duel | 0,14547 | 0,43476 | 0,09170 | 0,80937 | 0,01961 | -0,00813 | -0,08354 | 0,01475 | 0,12928 | 0,04926 | 0,10984 | 0,13059 | 0,708 | 0,900 |
| v14b | Aerial Defensive Duels | -0,00081 | -0,07054 | 0,01041 | 0,14727 | -0,04889 | 0,14544 | -0,11209 | 0,04987 | 0,71284 | 0,16662 | -0,07774 | 0,06888 | 0,500 | 0,546 |
| v14d | Aeriel Offensive Duel | -0,07141 | -0,19393 | 0,02845 | 0,69977 | 0,14056 | 0,25928 | -0,09098 | -0,04243 | 0,33391 | 0,00949 | -0,07276 | -0,10496 | 0,736 | 0,778 |
| v15a | Entries CAM | 0,89681 | 0,18476 | 0,17526 | 0,10327 | 0,07784 | -0,11347 | 0,02325 | 0,06758 | -0,02268 | -0,06186 | -0,04064 | 0,04945 | 0,956 | 0,906 |
| v15b | Entries Decision | 0,77140 | 0,24718 | 0,28028 | 0,09066 | 0,32142 | -0,04914 | 0,05520 | 0,05514 | -0,05357 | -0,18447 | 0,02435 | 0,04309 | 0,936 | 0,882 |
| v15c | Area Entries | 0,42678 | 0,22395 | 0,41731 | 0,01434 | 0,51882 | -0,10555 | 0,21535 | 0,01231 | -0,15111 | -0,04263 | 0,15761 | 0,09107 | 0,929 | 0,791 |
| v16 | Game Center Variation Reception | 0,14524 | 0,04820 | -0,03564 | -0,00919 | 0,01831 | -0,06733 | -0,02057 | 0,94042 | -0,01417 | 0,02868 | -0,04019 | 0,02797 | 0,617 | 0,914 |
| v17a | Save | -0,38851 | 0,04914 | 0,12116 | -0,16026 | -0,09454 | 0,16599 | -0,01629 | -0,05094 | 0,03935 | -0,01495 | -0,04981 | -0,03808 | 0,722 | 0,225 |
| v18b | Goalkeeper Kick Hands | -0,28080 | -0,05859 | -0,07031 | -0,00470 | -0,06965 | 0,87113 | -0,00075 | -0,07149 | 0,04168 | -0,05796 | 0,08601 | 0,04103 | 0,849 | 0,831 |
| v19 | Touch Area | 0,44551 | 0,16466 | 0,48914 | 0,16801 | 0,56297 | -0,10646 | 0,10345 | 0,00321 | -0,11277 | -0,05401 | 0,10154 | 0,09162 | 0,936 | 0,870 |
| v5c | Ball Conduction 5 meters2 | -0,01620 | 0,21605 | 0,15065 | 0,02549 | 0,05399 | 0,12809 | 0,03709 | 0,02999 | 0,01559 | -0,03367 | 0,16489 | 0,57528 | 0,756 | 0,354 |
| v5d | Ball Conduction 20 meters2 | -0,05274 | 0,10467 | 0,07680 | 0,04800 | 0,01749 | -0,05631 | 0,07100 | -0,00751 | 0,06169 | 0,00047 | -0,06544 | 0,66704 | 0,663 | 0,169 |
| v6b | Short Pass | 0,86083 | 0,16085 | 0,16037 | -0,06880 | 0,01476 | -0,24118 | 0,05377 | 0,00689 | -0,15749 | 0,11794 | -0,02505 | -0,04447 | 0,674 | 0,879 |
| v6c | Long Pass | 0,67809 | 0,04420 | -0,04376 | 0,09812 | 0,14381 | -0,03459 | 0,00845 | 0,17808 | -0,00080 | 0,42647 | -0,14220 | -0,01248 | 0,590 | 0,722 |
| v6d | Decision Delayed Pass | 0,67325 | 0,26017 | 0,23997 | -0,00449 | 0,31974 | -0,12351 | 0,11701 | 0,05376 | -0,16477 | -0,16630 | 0,02616 | 0,00036 | 0,786 | 0,740 |
| v6e | Construction Delayed Pass | 0,18122 | 0,04368 | -0,08267 | -0,02560 | 0,03215 | -0,17969 | 0,02653 | 0,02745 | -0,16704 | 0,74028 | -0,10489 | -0,08061 | 0,319 | 0,703 |
| v6f | Preparation Delayed Pass | 0,90146 | 0,07196 | 0,10983 | -0,08408 | -0,05574 | -0,19234 | 0,06449 | 0,05215 | -0,10137 | -0,02591 | -0,12744 | -0,01332 | 0,684 | 0,888 |
| v6i | CAM Passes | 0,56592 | -0,09407 | -0,05479 | 0,27130 | -0,11435 | 0,21929 | -0,01789 | 0,01661 | 0,24813 | 0,25884 | 0,03937 | -0,05608 | 0,784 | 0,618 |
| v6j | Flank Variation Pass | 0,12577 | 0,04175 | -0,01810 | -0,01442 | 0,04401 | -0,06291 | -0,04631 | 0,95680 | 0,00734 | 0,02852 | -0,03554 | -0,00592 | 0,617 | 0,924 |
| v6l | Construction Preparation Vertical Pass | -0,14420 | -0,06435 | -0,11787 | 0,02046 | -0,17829 | 0,07971 | -0,07975 | 0,01332 | 0,14274 | 0,75147 | 0,17854 | 0,06178 | 0,704 | 0,672 |
| v6m | Construction Decision Vertical Pass | -0,36803 | -0,09714 | -0,06671 | 0,17887 | 0,02494 | 0,48999 | 0,02578 | -0,05917 | 0,14672 | 0,09922 | -0,03114 | -0,09743 | 0,831 | 0,477 |
| v6o | Preparation Decision Vertical Pass | 0,79767 | 0,09537 | 0,11907 | 0,18599 | 0,16249 | -0,01134 | -0,01248 | 0,05441 | 0,04865 | -0,10026 | -0,00256 | -0,04468 | 0,941 | 0,718 |
| v8a | Total Shots | 0,06918 | 0,11137 | 0,97168 | 0,01879 | 0,13762 | -0,01690 | 0,13399 | -0,03688 | 0,00652 | -0,03580 | -0,03656 | 0,07032 | 0,789 | 0,842 |
| v13b | Take On CAM | 0,25538 | 0,90773 | 0,08950 | 0,14046 | 0,08364 | -0,06312 | 0,02938 | 0,03808 | -0,06008 | -0,01320 | 0,01363 | 0,12046 | 0,767 | 0,924 |
| v13c | Take On Decision | 0,30585 | 0,80787 | 0,13063 | 0,13089 | 0,16632 | -0,01723 | 0,06745 | 0,02788 | -0,06784 | -0,09430 | 0,00628 | 0,08638 | 0,915 | 0,809 |
| v14e | Offensive Duel CAM | 0,31234 | 0,37514 | 0,11170 | 0,81182 | 0,16643 | 0,07168 | -0,04789 | -0,01024 | 0,13617 | -0,07700 | -0,02790 | 0,06587 | 0,780 | 0,970 |
| v18d | Goalkeeper Long Hand pass | -0,15358 | -0,01371 | -0,04600 | 0,04636 | -0,05393 | 0,66555 | -0,01264 | -0,03195 | 0,03476 | -0,06749 | 0,04736 | 0,03510 | 0,789 | 0,516 |

***Table 9 - Results of Factor Analysis with 12-factor solution***

Throughout the table below, it is possible to verify that with a solution of 12-factors, we obtain an explanation of the model in approximately *70%* through the cumulative variance, which translates into a very satisfactory value.

| | **Factors** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* |
| **Variance** | 7,248 | 3,724 | 3,454 | 2,439 | 2,300 | 2,194 | 1,972 | 1,913 | 1,639 | 1,594 | 1,226 | 1,042 |
| **Proportional Variance** | 0,165 | 0,085 | 0,079 | 0,055 | 0,052 | 0,050 | 0,045 | 0,043 | 0,036 | 0,036 | 0,028 | 0,024 |
| **Cumulative Variance** | 0,165 | 0,249 | 0,328 | 0,383 | 0,383 | 0,436 | 0,530 | 0,574 | 0,611 | 0,647 | 0,675 | 0,699 |

***Table 10 - Variance, Proportional Variance and Cumulative Variance of a 12-factor solution***

However, as stated before, a twelve-factor solution was still not easy to interpret. So as stated before, we prefer to remove non-essential variables and simplify the analysis, where we proceeded by reducing the number of factors and removing some variables over several attempts to try to obtain the best possible solution. Thus, it will be necessary to re-evaluate the importance of some variables in response to the object under study and consequently assess the need to remove some variables. In which the variables removed were the following:

| **Removed Variables** | |
|---|---|
| *v1* | Clearance |
| *v2* | Interception |
| *v3* | Recovery |
| *v4* | Tackle |
| *v5b* | Ball Conduction 20 meters |
| *v9* | Assist |
| *v10* | Goal |

| Removed Variables – cont. | |
|---|---|
| *v11* | Disposs |
| *v12* | Foul |
| *v14a* | Defensive Duels |
| *v14b* | Aerial Defensive Duels |
| *v16* | Game Center Variation Reception |
| *v17a* | Save |
| *v17b* | Save_1vs1 |
| *v18a* | Goalkeeper Smother |
| *v18b* | Goalkeeper Kick Hands |
| *v5c* | Ball Conduction 5 meters2 |
| *v5d* | Ball Conduction 20 meters2 |
| *v6e* | Construction Delayed Pass |
| *v6i* | CAM Passes |
| *v6j* | Flank Variation Pass |
| *v6l* | Construction Preparation Vertical Pass |
| *v6m* | Construction Decision Vertical Pass |
| *v18d* | Goalkeeper Long Hand pass |

***Table 11 - Removed Variables***

Over several attempts, a solution consisting of five factors was chosen, in which the following results were obtained throughout the various analyzes (table 13). Also, after some meetings with the Sports Data Science Department of Sport Lisboa e Benfica, the conclusion was reached that factors that consisted of only two variables would not be viable for consideration in a given style of play. Moreover, a solution with five factors constitutes the most viable solution for the response of the object under study.

In terms of the KMO test, a significant improvement was obtained regarding the Overall MSA, which, after removing the variables, was established at approximately *0.87*— concluding that the dataset now has better conditions to provide a satisfactory level of interpretability. Additionally, returning to the re-evaluation of the previously stated criteria (Pearson's, Kaiser's, Scree Plot's criteria), the values now obtained are acceptable. Therefore, the analysis proceeded in a planned way (Sharma, 1996).

| | Eigenvalues |
|---|---|
| *1* | 10,918 |
| *2* | 2,69 |
| *3* | 2,28 |
| *4* | 1,914 |
| *5* | 0,963 |
| *6* | 0,947 |
| *7* | 0,516 |
| *8* | 0,434 |
| *9* | 0,387 |
| *10* | 0,303 |
| *11* | 0,25 |
| *12* | 0,22 |
| *13* | 0,197 |
| *14* | 0,18 |
| *15* | 0,177 |
| *16* | 0,136 |
| *17* | 0,116 |
| *18* | 0,103 |
| *19* | 0,087 |
| *20* | 0,074 |
| *21* | 0,067 |
| *22* | 0,023 |
| *23* | 0,021 |

***Table 12 - Eigenvalues of the correlation matrix***

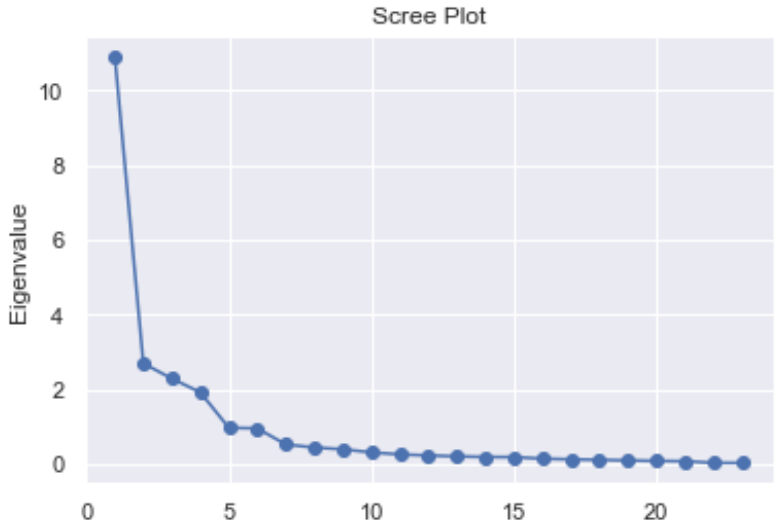



***Figure 8 - Scree Plot***

Through the loadings obtained by the Varimax rotation, we concluded that the variables are well distributed for both factors, and the loadings of each variable are visibly disseminated by each factor. At this stage, it is possible to verify that all variables found interpretability on each factor, and each variable presents acceptable KMO and Communalities values (KMO values > *0.5* & Communalities values > *0.5*). Also, as we can see in the table below, with a 5-factor solution, we obtain representability of around 80%, which means that compared to a 12-factor solution, we will have better representability, considering the cumulative variance presented in the table below.

| | **Factors** | | | | |
|---|---|---|---|---|---|
| | *1* | *2* | *3* | *4* | *5* |
| **Variance** | 6,028 | 3,535 | 3,278 | 2,568 | 2,309 |
| **Proportional Variance** | 0,274 | 0,161 | 0,149 | 0,117 | 0,105 |
| **Cumulative Variance** | 0,274 | 0,435 | 0,584 | 0,700 | 0,805 |

***Table 13 - Variance, Proportional Variance and Cumulative Variance of a 5-factor solution***

| Variables | | Factors | | | | | KMO | Communalities |
|---|---|---|---|---|---|---|---|---|
| | | *A* | *B* | *C* | *D* | *E* | | |
| v5a | Ball Conduction 5 meters | 0,78437 | 0,24215 | 0,17137 | 0,19615 | -0,09405 | 0,952 | 0,751 |
| v6g | Key and Assist Pass | 0,23342 | 0,39330 | 0,13830 | 0,76013 | 0,00545 | 0,920 | 0,806 |
| v6h | Pass Crosses Area | 0,44253 | 0,80146 | 0,11173 | 0,22690 | 0,15015 | 0,922 | 0,925 |
| v6k | Vertical Pass | 0,86627 | 0,23227 | 0,15941 | 0,09158 | 0,08403 | 0,803 | 0,845 |
| v7 | Crosses | 0,31933 | 0,74057 | 0,08342 | 0,26270 | 0,27377 | 0,910 | 0,801 |
| v8b | Exterior Shot | 0,19307 | 0,13287 | 0,06373 | 0,62578 | 0,06289 | 0,904 | 0,455 |
| v13a | TakeOn | 0,12161 | 0,06564 | 0,91253 | 0,08479 | 0,03830 | 0,705 | 0,860 |
| v14c | Offensive Duel | 0,13523 | 0,08213 | 0,49000 | 0,07409 | 0,74657 | 0,702 | 0,828 |
| v14d | Aeriel Offensive Duel | -0,10137 | 0,06982 | -0,20147 | -0,00089 | 0,85877 | 0,682 | 0,793 |
| v15a | Entries CAM | 0,82386 | 0,27745 | 0,20899 | 0,20789 | 0,15206 | 0,949 | 0,866 |
| v15b | Entries Decision | 0,61545 | 0,53489 | 0,26840 | 0,29796 | 0,15634 | 0,943 | 0,850 |
| v15c | Area Entries | 0,28313 | 0,70777 | 0,25265 | 0,37819 | -0,04460 | 0,916 | 0,790 |
| v19 | Touch Area | 0,28271 | 0,71892 | 0,20049 | 0,45573 | 0,15442 | 0,936 | 0,868 |
| v6b | Short Pass | 0,86665 | 0,19570 | 0,18215 | 0,14543 | -0,08971 | 0,780 | 0,852 |
| v6c | Long Pass | 0,71745 | 0,12987 | 0,04402 | -0,01554 | 0,14777 | 0,802 | 0,556 |
| v6d | Decision Delayed Pass | 0,55229 | 0,52653 | 0,27965 | 0,25185 | -0,01549 | 0,948 | 0,724 |
| v6f | Preparation Delayed Pass | 0,91064 | 0,09914 | 0,09274 | 0,14459 | -0,07520 | 0,836 | 0,874 |
| v6o | Preparation Decision Vertical Pass | 0,67563 | 0,33478 | 0,09252 | 0,10751 | 0,28716 | 0,912 | 0,671 |
| v8a | Total Shots | 0,02307 | 0,26804 | 0,09618 | 0,88076 | 0,04205 | 0,878 | 0,859 |
| v13b | Take On CAM | 0,21627 | 0,15667 | 0,93345 | 0,10713 | 0,10056 | 0,743 | 0,964 |
| v13c | Take On Decision | 0,22623 | 0,25615 | 0,81475 | 0,13637 | 0,11087 | 0,920 | 0,811 |
| v14e | Offensive Duel CAM | 0,23758 | 0,21589 | 0,41528 | 0,11106 | 0,82405 | 0,737 | 0,967 |

***Table 14 - Results of Factor Analysis with Varimax Rotations of a 5-factor solution***

Once the number of factors to retain was chosen, the time has come to identify which latent dimensions are responsible for intercorrelations among the indicators through the visualization and comparison of the loadings obtained in each factor by each variable. The previous rotation provides essential insights to understand the factor better and, consequently, the factor pattern.

Identifying and visualizing each variable group by factors leads us to think about the fittest name to label the factors. Thus, the best approach to achieve this objective was to classify each factor according to the set of variables obtained by each one, where every single event

provides an impression of the way a team plays. As a result, each factor will correspond to a specific style of play that will translate into the strategy adopted by each team in a game.

Through the variables obtained, the factors were identified with the labels:

A- Open Play;
B- Sustained Threat;
C- Take On;
D- Chances;
E- Duels;

The first factor is composed of the variables “Ball Conduction 5 meters”, “Vertical Pass”, “Entries CAM”, “Short Pass”, “Long Pass”, “Decision Delayed Pass”, “Preparation Delayed Pass”, “Preparation Decision Vertical Pass”. The “Open Play” Factor refers to any phase in the match where the ball is passed or kicked between teammates and both teams contesting for the ball. Subsequently, the teams with low values of open play will have the minor possession and worst results at in-possession events (such as passes, dribbles, shots, and crosses).

The second factor is composed of the variables “Pass Crosses Area”, “Crosses”, “Entries Decision”, “Area Entries”, and “Touch Area”. The Factor “Sustained Threat” focuses on possessions in the attacking third of the pitch. This factor is typically characterized by events in offensive zones of the playing field where teams are looking for opportunities to attack and strike.

The third factor is composed of the variables “Take On”, “Take On CAM” and “Take On Decision”. This factor can later help to identify if a team defines its style of play due to the individualities of its players or if it gives more importance to the work developed by the collective. If a team has high values in “Take On,” it may mean that its players are looking for a 1vs1 dispute instead of the collective to achieve the goal and, consequently, the

victory. In more physical teams with less tactical qualities, the out-of-possession events (pressures, tackles, interceptions, duels) present higher values.

The fourth factor is composed of the variables “Key and Assist Pass”, “Exterior Shot,” and “Total Shot”. That is, variables typically translate the teams’ insistence and the number of attempts a particular team try to achieve the goal.

The last and fifth factor is characterized by the variables “Offensive Duel”, “Aerial Offensive Duel”, and “Offensive Duel CAM”, where these events translate the contest actions between two players of opposing sides in the field. Teams with high scores are characterized by teams focusing their game strategy on out-of-possession events instead of focusing their strategy on tactical aspects. Usually, teams with this kind of pattern values do not have as much quality in running play and consequently low values in in-possession events. Teams with lower values support their game strategy with more technical approaches that expose their players to less high physical demand.

| A - Open Play | | B - Sustained Threat | | C - Take ON | | D - Chances | | E - Duels | |
|---|---|---|---|---|---|---|---|---|---|
| *v5a* | Ball Conduction 5 meters | *v6h* | Pass Crosses Area | *v13a* | TakeOn | *v6g* | Key and Assist Pass | *v14c* | Offensive Duel |
| *v6k* | Vertical Pass | *v7* | Crosses | *v13b* | Take On CAM | *v8b* | Exterior Shot | *v14d* | Aeriel Offensive Duel |
| *v15a* | Entries CAM | *v15b* | Entries Decision | *v13c* | Take On Decision | *v8a* | Total Shots | *v14e* | Offensive Duel CAM |
| *v15b* | Entries Decision | *v15c* | Area Entries | | | | | | |
| *v6b* | Short Pass | *v19* | Touch Area | | | | | | |
| *v6c* | Long Pass | *v6d* | Decision Delayed Pass | | | | | | |
| *v6d* | Decision Delayed Pass | | | | | | | | |
| *v6f* | Preparation Delayed Pass | | | | | | | | |
| *v6o* | Preparation Decision Vertical Pass | | | | | | | | |

***Table 15 - Labeled Factors with each variable***

## 4.7 CLUSTER ANALYSIS

After concluding the factor analysis – in which was found five latent dimensions – a cluster analysis was conducted to determine whether the observations were clustered and aggregated according to their characteristics. The main goal was to group the teams by similarity criteria using factor scores previously obtained where the observations of each cluster should be as homogeneous as possible, while between clusters, the observations should be as heterogeneous. The clusters analysis involved two main methods, hierarchical and non-hierarchical methods.

### Hierarchical Method

Hierarchical clustering is an unsupervised machine learning technique where the algorithm evaluates and assigns each data point to a specific cluster by measuring the dissimilarities between data (Murtagh & Contreras, 2012). During the analysis, agglomerative clustering was used where the clusters were joined together with the shortest distance between, and the process was repeated until one large cluster was formed containing all data points. A Euclidean distance and the Ward linkage method were used to minimize cluster variance.

With this, it is possible to visualize the solution of the hierarchical method of the cluster analysis (fig. 7) based on factor scores, where the vertical axis measures the distance of the sum of squares between clusters and the horizontal axis represents each game that was played during de 2020-2021 season.

At this stage, it is crucial to determine and decide what will be the number of k clusters to consider in the non-hierarchical methods subsequently applied. Ward's method showed higher levels of the R-Squared across all clusters' possibilities, whereas we can see in the dendrogram below that the chosen solution will be four clusters. Through this solution, the centroids of each of the four clusters were generated by Ward's method after the initial seeds were used on the further non-hierarchical methods performed – K-means and SOM.

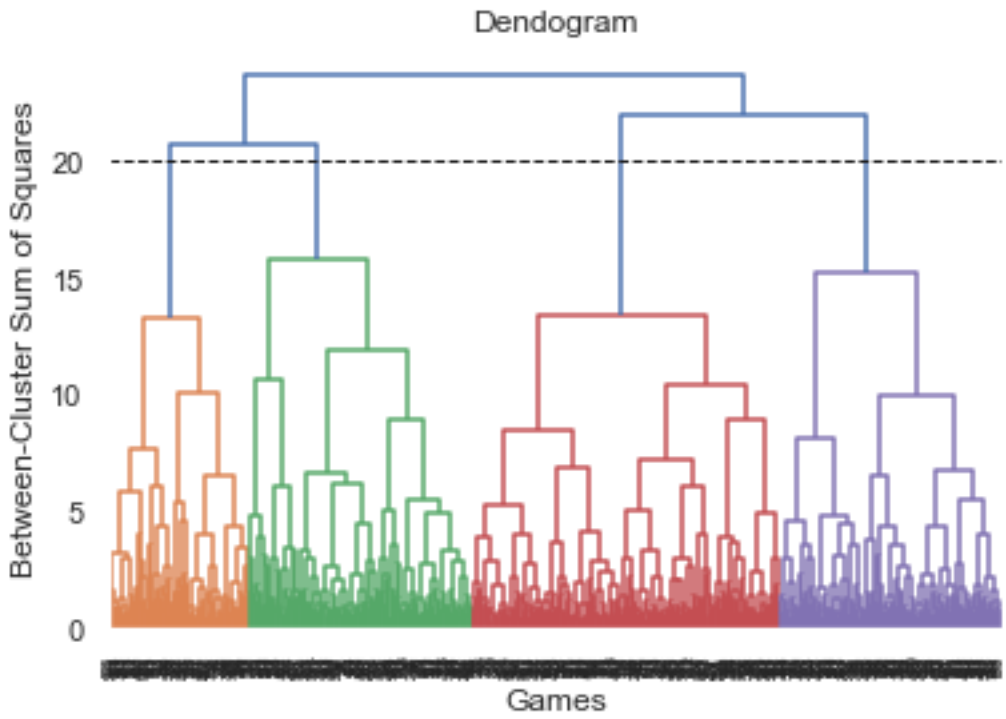



***Figure 9 - Wards' Dendrogram***

Non-Hierarchical Method - K-means

Now that a 4-cluster Solution has been chosen, it is time to perform K-Means, an unsupervised machine Learning technique. The K-Means cluster works on unlabelled data set where the main objective is to assign points to the closest centroid and then recalculate the centroids to repeat through this clustering approach (Murtagh & Contreras, 2012). The K-means algorithm started by placing centroids in a random location and then calculating the distance (Euclidean Distance) between the centroids and each point. After this process, it will choose the cluster with the minimum distance and recalculate the centroids by taking all vectors and averaging them. All processes will be repeated until there is no change between cluster constitutions. The evaluation of this method was based on the metric "WCSS" and where the visualization was obtained through the Elbow Method. In the Elbow method, the cluster number will differ between 1 to 15, where each k is calculated as the WCSS value. As the number of k increases, the WCSS value will start to decrease.

| Nr Clusters | WCSS Values |
|---|---|
| 1 | 3310.916864202699 |
| 2 | 2507.5174637035843 |
| 3 | 2089.866889304554 |
| 4 | 1765.9682324849985 |
| 5 | 1660.2670383474065 |
| 6 | 1551.733444230084 |
| 7 | 1467.7083489149345 |
| 8 | 1400.6122948351478 |
| 9 | 1327.1657052181977 |
| 10 | 1276.8554644645842 |
| 11 | 1218.3584523866086 |
| 12 | 1169.7029619350685 |
| 13 | 1154.08135633165 |
| 14 | 1116.359994896718 |

***Table 16 - WCSS***

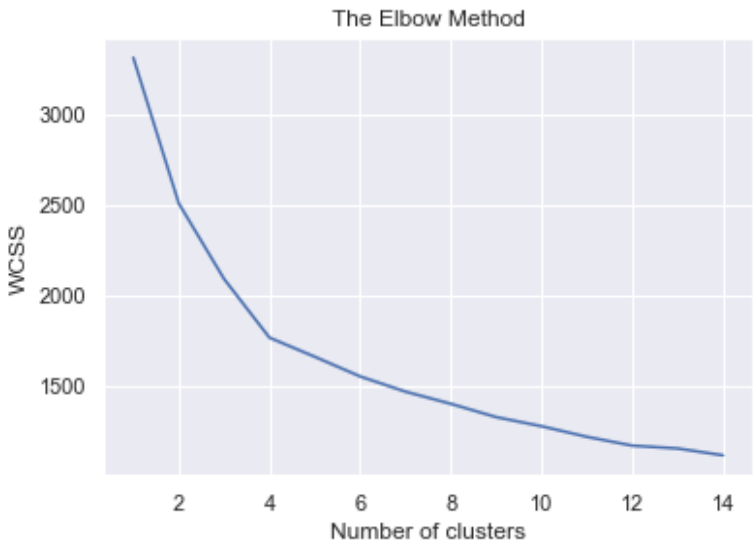



***Figure 10 - The Elbow Method***

With the application of K-means, it was possible to cluster each game in the 2020-2021 season by the 4-clusters previously chosen. In this way, based on each factor's weight on each one's formation, it was possible to label each factor through its performance in the data obtained. The performance was evaluated through the average of the factors obtained in each cluster based on the score of each game. With this approach, each cluster was intended to be labelled and later characterized. Additionally, below, we verify the total frequency of games that were considered by each cluster by applying the K-means method.

| K-Means | |
|---|---|
| **Cluster** | **Frequency** |
| 1 | 142 |
| 2 | 127 |
| 3 | 192 |
| 4 | 103 |

***Table 17 - Frequency with K-means method***

K-means

| Cluster | Open Play | Sustained Threat | Take On | Tries | Duels |
|---|---|---|---|---|---|
| 1 | 0,7434 | 0,7530 | -0,1492 | 0,5113 | -0,5344 |
| 2 | 0,2996 | -0,4001 | 1,1814 | -0,1475 | 0,3266 |
| 3 | -0,5217 | -0,3809 | -0,3735 | -0,3760 | -0,5190 |
| 4 | -0,4219 | 0,1653 | -0,5547 | 0,1778 | 1,3015 |

***Table 18 - Factor scores per cluster obtained by K-means method***

Non-Hierarchical Method - SOM

Subsequently, to complement the analysis under study and prove the results obtained previously, another unsupervised method was applied, the SOM (Self-Organizing Maps).

Self-Organizing Maps is an unsupervised neural network model with applicability for clustering, dimension reduction, and feature detection. Also, SOM is used by projecting higher dimensional data into lower dimensional space considering variables' similarity properties. This network consists of two layers of neurons connected by weight, where the input layer is connected to an input vector of the data set. Subsequently, the output layer will form a map consisting of a grid where several neurons are arranged. After the application of this neuronal model, we obtained results very similar to the results achieved previously with the K-means method, as can be seen below:

SOM

| Cluster | Frequency |
|---|---|
| 1 | 147 |
| 2 | 120 |
| 3 | 197 |
| 4 | 100 |

***Table 19 - Frequency with SOM method***

The pre-choice of the number of dimensions to retain was previously defined at the time of the API application due to the small dataset being analyzed. Additionally, it was also possible

to evaluate the teams' performance in each game through the average of the factors obtained in each cluster.

| SOM | | | | | |
|---|---|---|---|---|---|
| **Cluster** | **Open Play** | **Sustained Threat** | **Take On** | **Tries** | **Duels** |
| 1 | 0,8926 | 0,4729 | -0,4381 | 0,4494 | -0,7883 |
| 2 | -0,1588 | 0,9055 | 0,1475 | -0,3487 | 0,3296 |
| 3 | -0,8280 | -0,4844 | 0,1409 | -0,0936 | -0,2040 |
| 4 | 0,8100 | -0,7081 | -0,0470 | 0,2065 | 0,5881 |

***Table 20 - Factor Scores per cluster obtained by SOM method***

With the conclusion of both unsupervised methods, the time has come to interpret the results obtained by trying to label each cluster. Depending on the results, they were labelled as follows:

- Cluster 1 – Peak Performance
- Cluster 2 – Regular Performance
- Cluster 3 – Poor Performance
- Cluster 4 – Physical Performance

# 5 RESULTS AND DISCUSSION

## 5.1 FACTOR ANALYSIS INTERPRETATION

In this section, it is possible to see the comparison between teams' styles of play and performance, both directly dependent on their final season result. The main objective of this analysis is to understand Bundesliga teams better to provide several paramount insights, where other coaches and teams could use this information to have a competitive advantage. The analysis process was developed based on each style of play values obtained by each team. The strategy analysis was to perform a comparison by dividing the teams by their final season ranking score. With this, some teams were allocated to the "Top 3", "Middle 3" and "Bottom 3" based on their final ranking season result. In this way, it will be interesting to understand what types of game styles differentiated them all and which ones impacted the success of these teams. The strategy involved the implementation of a radar chart per team, where it is possible to verify the average value of each style of play and the average of each type of play of all the teams that make up this league.

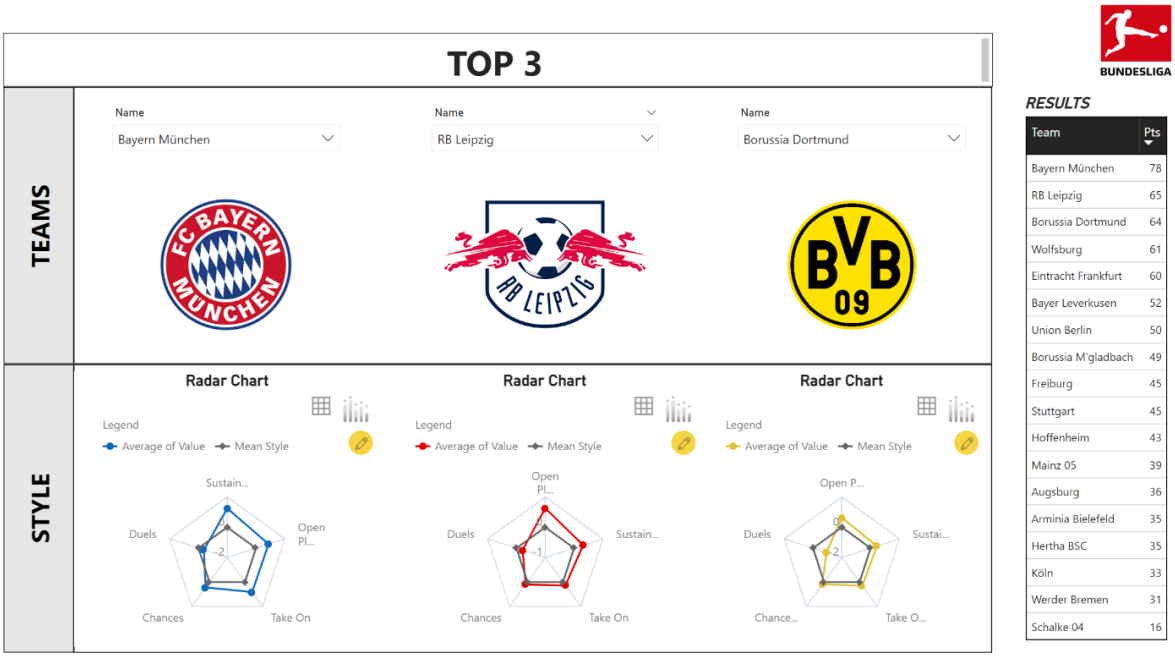
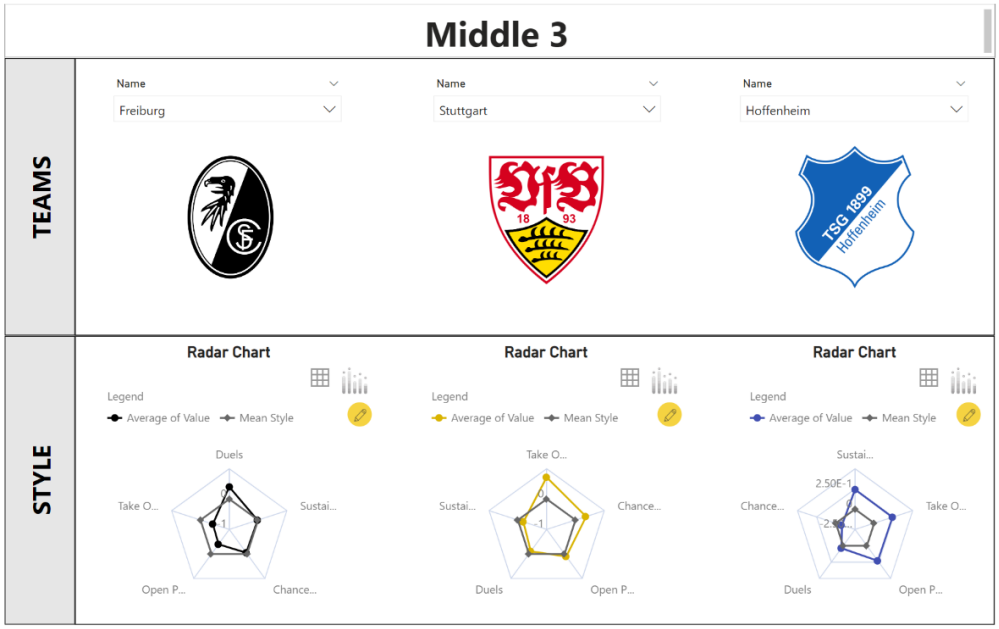



***Figure 11 - "Top 3" Dashboard***

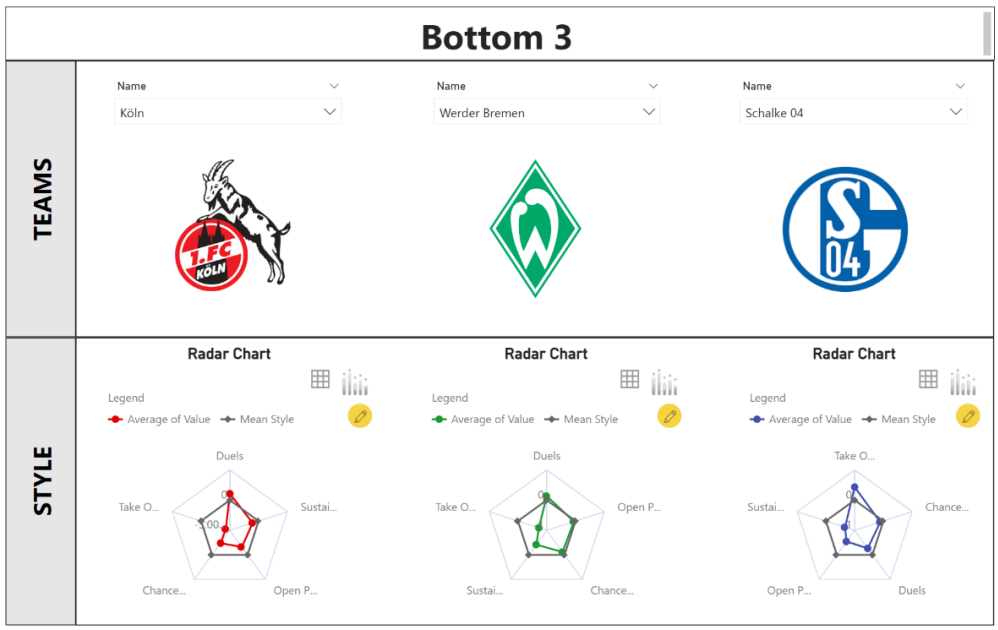



BUNDESLIGA

*RESULTS*

| Team | Pts |
|---|---|
| Bayern München | 78 |
| RB Leipzig | 65 |
| Borussia Dortmund | 64 |
| Wolfsburg | 61 |
| Eintracht Frankfurt | 60 |
| Bayer Leverkusen | 52 |
| Union Berlin | 50 |
| Borussia M'gladbach | 49 |
| Freiburg | 45 |
| Stuttgart | 45 |
| Hoffenheim | 43 |
| Mainz 05 | 39 |
| Augsburg | 36 |
| Arminia Bielefeld | 35 |
| Hertha BSC | 35 |
| Köln | 33 |
| Werder Bremen | 31 |
| Schalke 04 | 16 |

***Figure 12 - "Middle 3" Dashboard***




BUNDESLIGA

*RESULTS*

| Team | Pts |
|---|---|
| Bayern München | 78 |
| RB Leipzig | 65 |
| Borussia Dortmund | 64 |
| Wolfsburg | 61 |
| Eintracht Frankfurt | 60 |
| Bayer Leverkusen | 52 |
| Union Berlin | 50 |
| Borussia M'gladbach | 49 |
| Freiburg | 45 |
| Stuttgart | 45 |
| Hoffenheim | 43 |
| Mainz 05 | 39 |
| Augsburg | 36 |
| Arminia Bielefeld | 35 |
| Hertha BSC | 35 |
| Köln | 33 |
| Werder Bremen | 31 |
| Schalke 04 | 16 |

***Figure 13 - "Bottom 3" Dashboard***

### 5.1.1 TEAMS PLACED WITH THE HIGHEST FINAL RANKING RESULTS

**Description**

The TOP 3 is composed by FC Bayern München, RB Leipzig and Borussia Dortmund. These teams have achieved the best results in this 2020-2021 season, where these teams finished the season in the top three (Figure 11).

**Analysis**

As it is possible to verify through the dashboard in figure 14, the three teams present values of playing style higher than the average of all the teams in the league. The "Sustained Threat" and "Open Play" styles stand out as the most predominant characteristics in these three teams, which is justified by the high levels of ball possession that these teams have throughout the season and the intensity of in-possession events that develop in the “Sustained Threat” metric that increases the chances of a goal.

Other metrics highlighted are the value obtained in the “Take On” style and in the “Chances” style, in which you can show that these teams are made up of quality players who are not afraid to go 1vs1 in dribbling and end up supporting their teams to achieve the desired results thanks to the individual quality presented. The “Chances” style of play, despite being above the average of the other teams, has a low value, which means that these teams are effective at the moment of finishing the goal and do not need many attempts to obtain it.

Additionally, the average presented for the styles described above is above the average value of the other teams, constituting a distinct aspect of the performance and results obtained compared with the other teams.

Regarding the “Duels” style of play, the value obtained by the teams that make up the “Top 3” is well below the average of the other teams. Through the value of these metrics, the three teams analyzed showed that their game strategy practiced throughout this season did

not focus on non-possession events but on in-possession events, which means that they usually give more importance to the tactical aspects where teams implement someway of the play that potentiate events that show possession of the ball than to the physical duels. Typically, a strategy based on ball possession will increase the probability that a team has the opportunity to score.

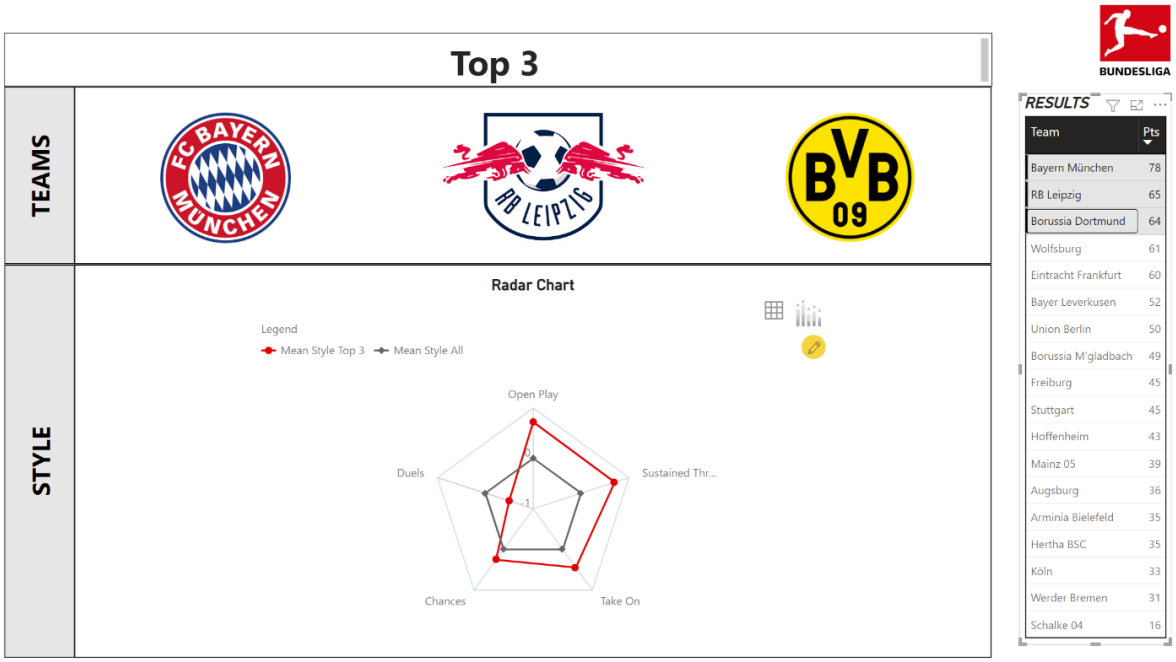



***Figura 14 - Comparison between mean style of play of "Top 3" and all Bundesliga teams***

### 5.1.2 TEAMS PLACED WITH THE AVERAGE FINAL RANKING RESULTS

**Description**

The teams allocated to the "Middle 3" were SC Freiburg, VfB Stuttgart, and TSG Hoffenheim, where they are placed in the middle of the 2020-2021 final season ranking with their respective 9th, 10th, and 11th place (figure 12).

## Analysis

Visualizing the represented dashboard, the teams in question also present values above the average playing style of all Bundesliga teams. However, the values obtained in the different game types are no longer as sound compared to the results of the "Top 3" teams. If the analysis is done individually, the three teams under analysis already have play styles whose values are lower than the average of all teams.

The "Take On" and "Duels" stand out as the playstyles that best characterize the playstyle of these three teams. Relating to the low value of "Open Play", these teams, in terms of game strategies, put more importance on the individualities and qualities of each player and not on the product inherent to the work developed by the collective. In this way, as a result, teams refuge themselves in more physical events such as duels because they could not keep possession of the ball as pretended.

In terms of the results obtained in the metrics "Sustained Threat" and "Chances", we can see that these results are correlated. These three teams present a very significant value compared to the average playing style of all teams meaning that Freiburg, Stuttgart, and Hoffenheim performed many tries throughout the season to reach the goal. However, relating to the low value obtained in the "Sustained Threat" style, these teams did not constitute a danger to their opponents. Most attempts to score (key assist pass, outside shots, total shots) were not effective enough to be considered a threat to the respective opponents.

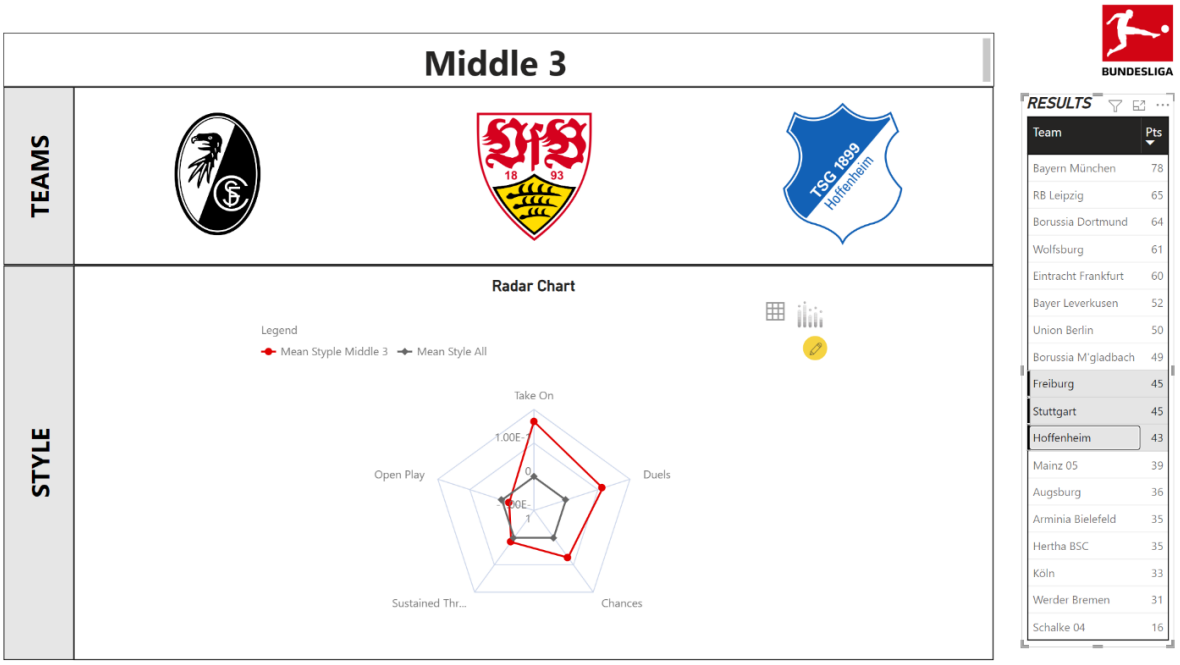



***Figure 15 - Comparison between mean style of play of "Middle 3" and all Bundesliga teams***

### 5.1.3 TEAMS PLACED WITH THE LOWEST FINAL RANKING RESULTS

**Description**

The "Bottom 3" is represented by FC Koln, SV Werder Bremen, and FC Schalke 04, which had the lowest results in the final season ranking of 2020-2021 (16th, 17th, and 18th place).

**Analysis**

Observing the dashboard represented in figure 16, it is visible from the outset that in all five styles of play that have been scrutinized, the values of these three teams (FC Koln, SV Werder Bremen, FC Schalke 04) are below the average of the playing styles of all the other teams.

Supposing a more specific analysis is carried out, once again, we see that teams tending to have less technical quality present better results in the “Duels”, and “Take On” styles and worse results in the “Open Play” factor since they cannot have possession of the ball and consequently need to expose their players to duels in order to recover the ball. Once again, the focus on individualities overlaps with collective work, as can be seen at the "Take On" level. Due to the in-possession events not being potentiated, which will bring the teams greater possession of the ball, their players will not be so exposed to ball fights and 1vs1 duels.

Regarding the variables “Chances” and “Sustained Threat”, the inversion of the values obtained in these two metrics evidences the level of danger that these teams expose their opponents to. It is possible to see a high value in the level of "Chances" and a low value in the metric "Sustainable Threat" which means that these three teams are ineffective in the last third of the opponent's field and cannot reach opportunities to score.

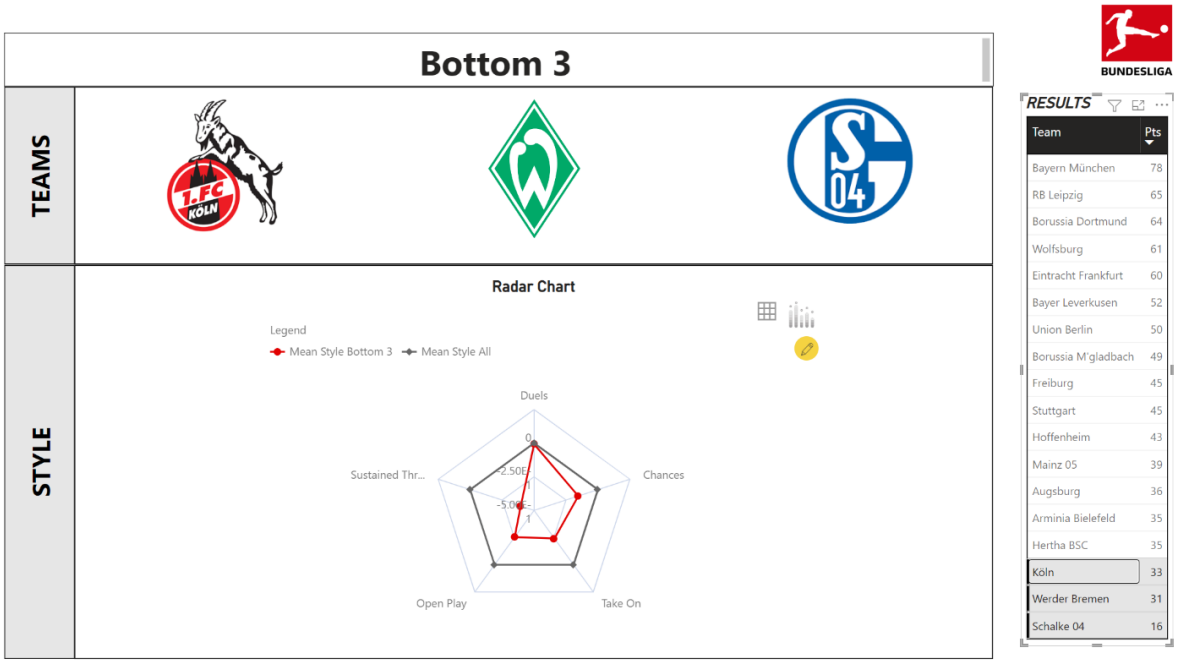



***Figure 16 - Comparison between mean style of play of “Bottom 3” and all Bundesliga teams***

## 5.2 CLUSTER ANALYSIS INTERPRETATION

The main objective of this section is to characterize better and interpret the results obtained through the application of the various clustering methods. The following topics will be discussed later: Crossing the results of both applied unsupervised machine learning techniques (K-means vs. SOM); Interpretation of the frequency obtained in each cluster by each team; Specific analysis of a Bundesliga team.

### 5.2.1 K-MEANS vs. SOM

**Description**

After both unsupervised machine learning techniques were successfully completed, it was time to compare the results of these clustering techniques to interpret and evaluate the achievements' validity and, later, the agreement between both methods. The clustering methods mentioned above made it possible to detail the various characteristics performed by the teams in each game through the grouping in each cluster. The identification of each cluster per game allowed the interpretation of the game patterns played by the teams throughout the 2020-2021 season.

**Analysis**

Accordingly, with the following table (Table 18), we verify the frequency of the distribution of teams for each cluster. Each game of the German league of the 2020-2021 season was evaluated, based on the scores resulting from each variable, and later grouped in the respective cluster according to the criteria established in the K-Means and SOM methods. When crossing the results obtained by each method, we found that the frequency of each cluster is very similar. Thus, it is possible to conclude that the results obtained after using two different clustering methods are similar. Therefore, we confirm the validity of the results.

However, the differences between the results of each clustering method can be justified because the K-Means method is considerably more susceptible to initial settings such as k value and seed. However, SOM can achieve better clustering quality when neurons in the output layers can all be used (Chen et al., 2010).

Additionally, to better interpret the results, the average of the scores of each factor was calculated according to the cluster group. Where it was possible to characterize each cluster and label each one:

- **Cluster 1** – *Peak Performance* – This cluster was labeled as "Peak Performance" for presenting significant values in the variables "Open Play", "Sustained Threat" and "Chances". Considering the factor analysis results, these are the three styles of play that characterize the "Top 3". Those teams obtained the best results in the German league during the 2020-2021 season and, consequently, the best performance.
- **Cluster 2** – *Regular Performance* – This cluster was labeled "Regular Performance" for presenting significant values in the factors "Take On" and "Duels" in both methods. However, we verified a difference between the values obtained by the different clustering methods. In terms of the K-means method, cluster 2 presents a significant value in the "Open Play" factor, and in terms of the SOM method, this significance is verified in terms of "Sustained Threat". This difference is irrelevant because the teams considered in "Middle 3" are characterized by both play styles.
- **Cluster 3 –** *Poor Performance* – Through the results obtained, this cluster presents the worst scores in each of the game styles: "Open Play", "Sustained Threat", "Take On", "Chances", and "Duels". The values obtained are considered negative and irrelevant in all game styles, both through K-means and SOM methods.
- **Cluster 4** – *Unstable Performance* – In this last cluster, we found that most of the teams that make it up are teams that are in a downhill danger zone. Teams that occupy the last places of the league table and, in the worst scenario, end up going down the division. In crossing with the factor analysis, the most significant scores of the styles of play are similar to those obtained by the teams belonging to the "Bottom 3".

K-Means

| Cluster | Frequency |
|---|---|
| 1 | 142 |
| 2 | 127 |
| 3 | 192 |
| 4 | 103 |

SOM

| Cluster | Frequency |
|---|---|
| 1 | 147 |
| 2 | 120 |
| 3 | 197 |
| 4 | 100 |

***Table 21 - Comparison of frequencies between clustering methods***

| Cluster | Open Play | Sustained Threat | Take On | Chances | Duels |
|---|---|---|---|---|---|
| 1 | 0,7434 | 0,7530 | -0,1492 | 0,5113 | -0,5344 |
| 2 | 0,2996 | -0,4001 | 1,1814 | -0,1475 | 0,3266 |
| 3 | -0,5217 | -0,3809 | -0,3735 | -0,3760 | -0,5190 |
| 4 | -0,4219 | 0,1653 | -0,5547 | 0,1778 | 1,3015 |

***Table 22 - Cluster Profiles with K-means method***

| Cluster | Open Play | Sustained Threat | Take On | Chances | Duels |
|---|---|---|---|---|---|
| 1 | 0,8926 | 0,4729 | -0,4381 | 0,4494 | -0,7883 |
| 2 | -0,1588 | 0,9055 | 0,1475 | -0,3487 | 0,3296 |
| 3 | -0,8280 | -0,4844 | 0,1409 | -0,0936 | -0,2040 |
| 4 | 0,8100 | -0,7081 | -0,0470 | 0,2065 | 0,5881 |

***Table 23 - Cluster Profiles with SOM method***

### 5.2.2 FREQUENCY PER TOP, MIDDLE AND BOTTOM

**Description**

To better understand the various clusters, it was necessary to analyze the frequency achieved by the teams throughout the 2020-2021 season. This same analysis was based on the results of the application of the K-means clustering method, where the teams comprising the "TOP 3", "Middle 3", and "Bottom 3" were evaluated. The unsupervised method chosen to continue the study was the K-means method since the results obtained between clustering methods (K-means vs. SOM) are similar and, in terms of performance, the K-means algorithm performs better than SOM (Mingoti & Lima, 2006).

**Analysis**

According to the figures presented below, regarding the analysis of the frequency of the teams that make up the "TOP 3", "Middle 3" and "Bottom 3", we see that the teams that make up the "TOP 3" have lower volatility between clusters, being usually very faithful to clusters 1 and 2. Regarding the teams that make up the "Middle 3" and "Bottom 3", it was found that the variation between clusters was higher than that of the "TOP 3" teams.

As you can see in table 24, the teams belonging to "Middle 3", and "Bottom 3" presented several games for which they had poor performances (Cluster 3) and unstable performances (Cluster 4). At a practical level, the trend increases in the values obtained in cluster 3 and cluster 4 by these two groups (Middle 3 and Bottom 3) reduces the probabilities of the teams being able to score and, consequently, win the game.

The results obtained conclude that the teams that present greater volatility and many changes at the level of clusters are less consistent teams that consequently show worse performances and may affect their success for winning games.

**TOP 3**

| Team | Frequency |
|---|---|
| **Bayern München** | |
| 1 | 19 |
| 2 | 7 |
| 3 | 1 |
| 4 | 1 |
| **Borussia Dortmund** | |
| 1 | 18 |
| 2 | 2 |
| 3 | 3 |
| **RB Leipzig** | |
| 1 | 13 |
| 2 | 8 |
| 3 | 1 |
| 4 | 1 |

**Middle 3**

| Team | Frequency |
|---|---|
| **Freiburg** | |
| 1 | 3 |
| 2 | 2 |
| 3 | 14 |
| 4 | 14 |
| **Hoffenheim** | |
| 1 | 12 |
| 2 | 11 |
| 3 | 6 |
| 4 | 4 |
| **Stuttgart** | |
| 1 | 7 |
| 2 | 15 |
| 3 | 8 |
| 4 | 1 |

**Bottom 3**

| Name | Frequency |
|---|---|
| **Köln** | |
| 1 | 5 |
| 2 | 2 |
| 3 | 18 |
| 4 | 9 |
| **Schalke 04** | |
| 2 | 13 |
| 3 | 18 |
| 4 | 3 |
| **Werder Bremen** | |
| 1 | 2 |
| 2 | 3 |
| 3 | 21 |
| 4 | 8 |

***Table 24 - Frequencies between teams with the K-means method***

Through the scatter plot below, we can see a more visual representation of the results of the frequencies presented in figure 17. In the first representation of the dispersion of the frequency of the games played by the teams that make up the “TOP 3”, we verify the existence of many games that were considered as Cluster 1 – Peak Performance and, in the remaining graphs, a decrease in representations of this same cluster. Analyzing the remaining “less good” clusters (Cluster 3 and Cluster 4), we verified the existence of more games considered as poor and unstable performances in the representations related to “Bottom 3” and “Middle 3”.

Regarding cluster 2, we observed a greater incidence in the graphic representing the teams that constituted the "Middle 3", where most of the games played were considered regular performances.

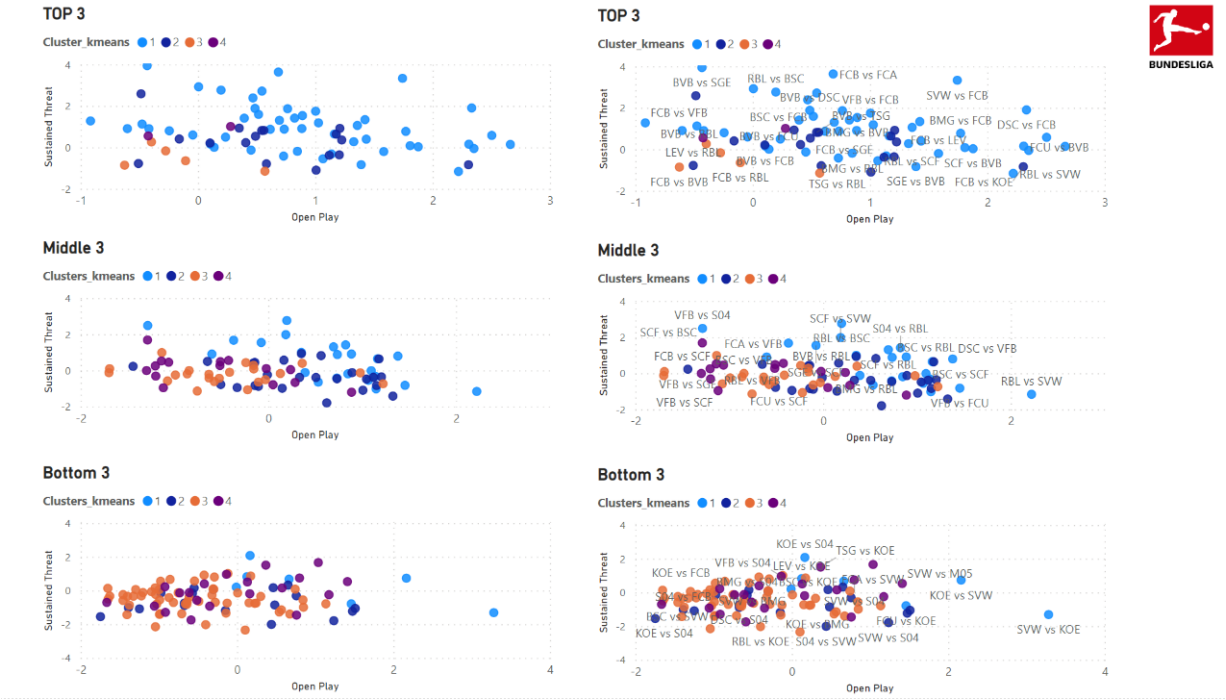



***Figure 17 - Dispersion Graphs***

## 5.3 SPECIFIC TEAM ANALYSIS

**Description**

In this section, the performance developed by Hoffenheim is explicitly analyzed in terms of the performance of each game played in the German league in the 2020-2021 season. Hoffenheim will be scrutinized for its performance, and the style of play practiced and adopted throughout the sports season. The primary purpose of this analysis will be to interpret the various patterns developed by the teams and, later, draw inferences from the performances obtained in each game.

## Analysis

As we can see in the figure below, a dashboard was built that presents the various styles of play adopted throughout the season by the constituent teams of Bundesliga this season; the frequency of each cluster, where the various performances performed by Hoffenheim through its games are identified; and an inclusive table of all games played during the 2020-2021 season, which contains information regarding the cluster belonging to it and the respective final result of the game. The addition of information on the final result of each game is justified because it is not coherent a priori to consider that a specific game identified as cluster 3 – Poor Performance, would have been a game in which the team was defeated. As will be shown later, teams that achieved Poor or Unstable Performance did not always end up losing the game.

The analysis of the Hoffenheim season was based on the division of the sports season into two phases: 1st phase (18/09/2020 to 01/01/2021) and 2nd phase (01/01/2021 to 22/05/2021).

The purpose of dividing the sports season into two phases was to identify the various changes that potentiated the change in the style of play practiced in the 1st phase of the season and which aspects were decisive for this change. Additionally, it will also be essential to investigate the variations between clusters between the two phases of the season. Starting the analysis by the overview of the 2020-2021 season of Hoffenheim, it appears that the team practiced a style of play with greater incidence in the styles "Sustained Threat," "Open Play," and "Take On." This fact is quite curious because it shows significant similarity with the characteristics developed by the teams that make up the “TOP 3”. However, focusing on the analysis in more detail, we identified that the “Chances” style of play has a lower value than the average value obtained by all Bundesliga teams. Thus, we conclude that Hoffenheim presented reasonable game rates through the high levels of ball possession and the intensity developed in in-possession events. However, looking at the ratio of the “Chances” metric, we can see that it was a team with few scoring opportunities and few attempts (i.e., Total Shots, Exterior Shots, Key Assist Pass).

Moving on to the frequency analysis, we found that in the complete season, Hoffenheim had twelve games considered as Cluster 1 - Peak Performance, eleven games evaluated as Cluster 2 - Regular Performance, six games marked as Cluster 3 - Poor Performances, and four games as Cluster 4 – Unstable Performances.

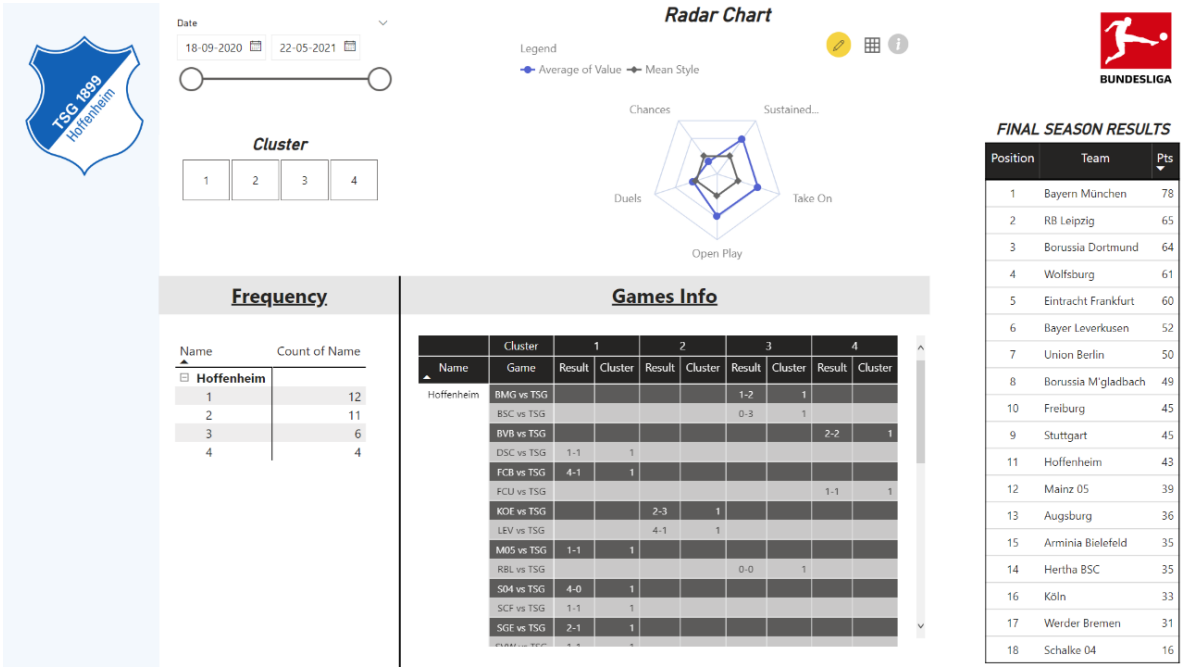



***Figure 18 - TSG 1899 Hoffenheim's general season dashboard***

Continuing to analyze the first half of the season, we can see in figure 19 that Hoffenheim played football in which the styles "Take On," "Open Play," and Sustained Threat" were the most evident. However, contrary to the entire season radar chart, in this first half of the season, the team placed more importance on individual aspects and the quality of each player than on aspects that promote ball possession and in-possession game events. Thus, Hoffenheim presented only four games in which they were considered as "Peak Performance" and five games as "Regular Performance". Four of the nine games played in this first half of the season were considered Poor and Unstable performances.

On matchday 13, Hoffenheim occupied the 13th position of the league table, where any slip would put the team in a danger zone, to which they could enter the relegation zone for the lower league.

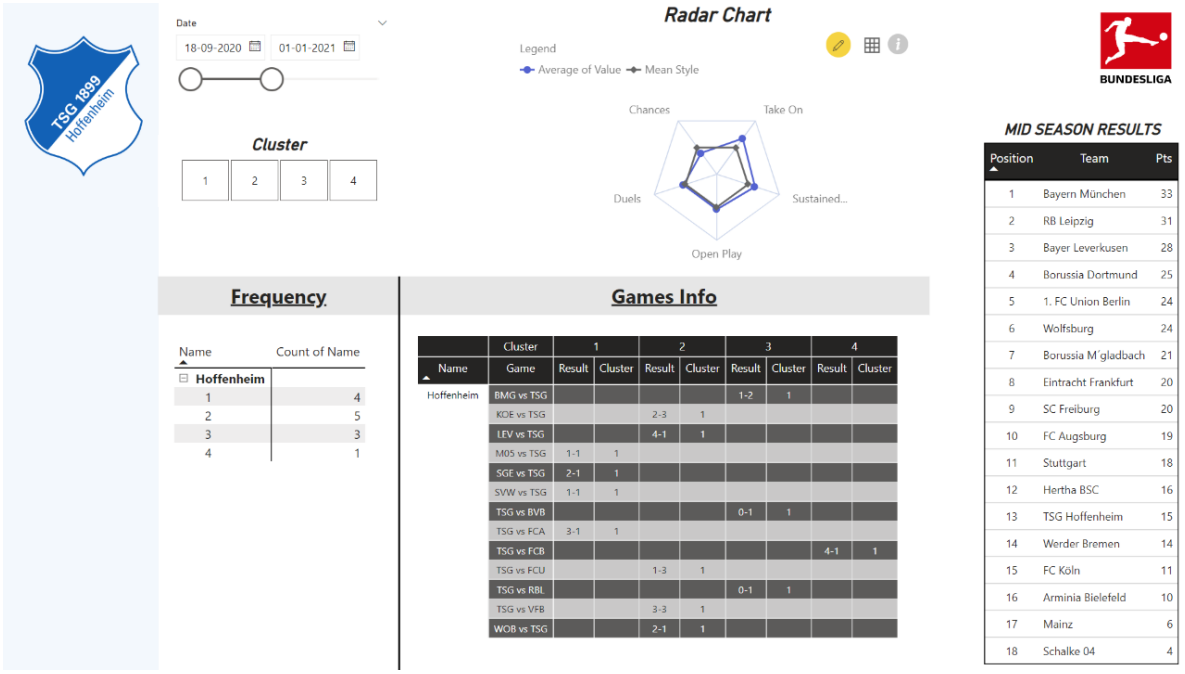



***Figure 19 - TSG 1899 Hoffenheim's first phase dashboard***

Continuing the analysis moving to the 2nd phase of the season, Hoffenheim presented significant playing style changes. In this half of the season, the team changed its play pattern with notable improvements in the metrics "Open Play" and "Sustained Threat." Both styles of play have values well above the values presented in the 1st phase of the season and considerably above the average of all teams in the German league. At this stage, Hoffenheim prioritized individual aspects (Take On) and focused its game strategy towards events that would promote the team's success (Open Play and Sustained Threat).

As a result, the "Chances" ratio also increased compared to the first half of the season, where the changes described above were reflected in the performance developed over the games. This change is reflected in the frequency shown in the dashboard below.

In the second half of the season, Hoffenheim had eight games considered as “Peak Performance” (2x the value obtained in the first half of the season) and six games considered as “Regular Performance” (a value higher than the results obtained in the 1st phase). Looking now at the frequency of games of the “worst” clusters, the team of the 20 games played only presented six games with lower-than-expected performances (Cluster 3 – Poor Performances and Cluster 4 – Unstable Performances).

Hoffenheim closes the 34th round of the 2020-2021 season with a 2-1 victory over Hertha BSC, ending the season in 11th place in the league table (two places higher than in the first phase of the season). Considering the results presented, it is evident that the changes observed in terms of playing style in the 2nd phase of the season were preponderant in the impact of the performance demonstrated by the team throughout the played games, being decisive for the success of the team and the results achieved.

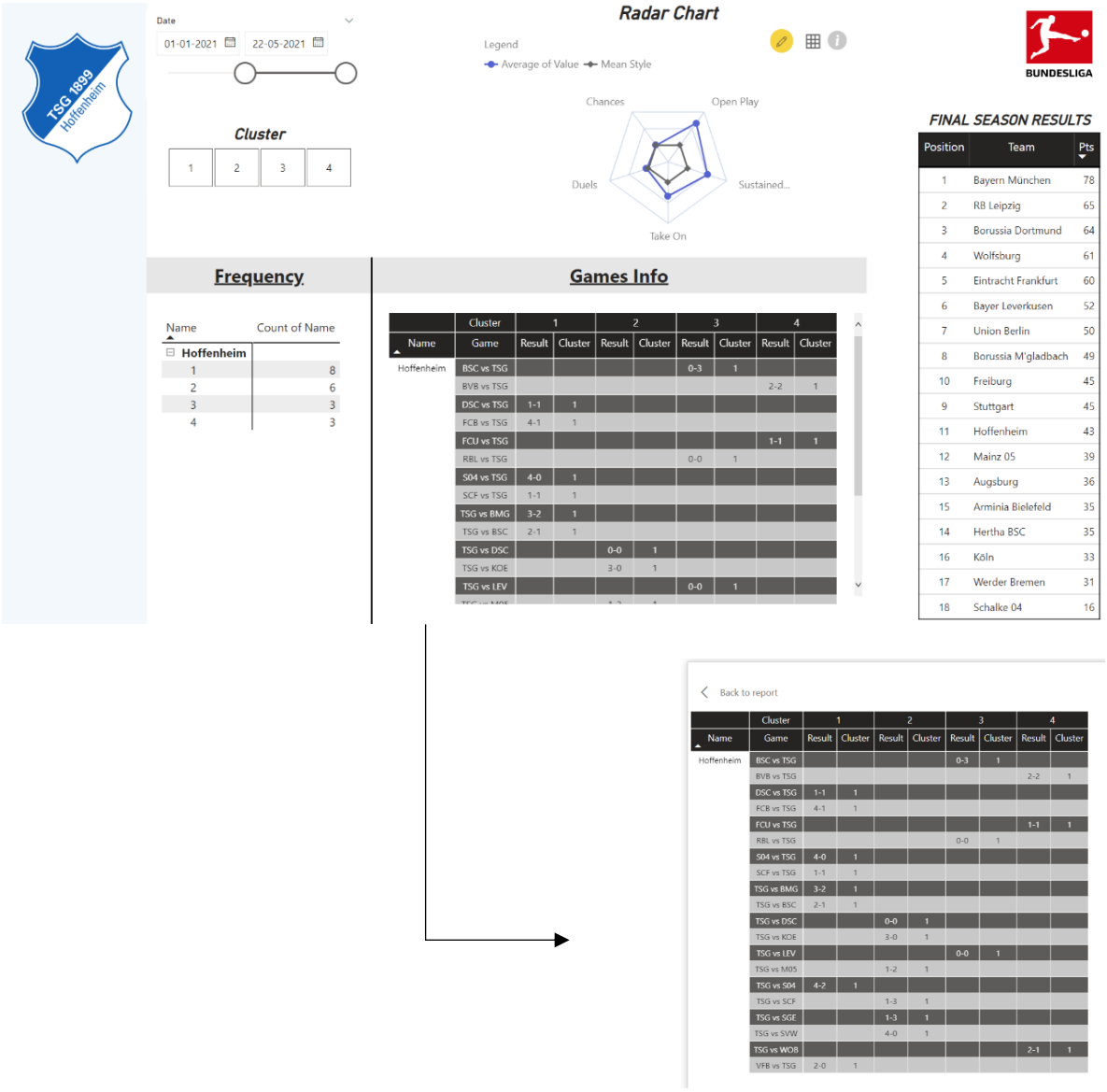



***Figure 20 - TSG 1899 Hoffenheim's' second phase dashboard***

# 6 CONCLUSION

## 6.1 SYNTHESIS OF THE DEVELOP WORK

The current research aimed to structure an analysis model to support coaches and technical teams in the process of interpreting game patterns performed by opposing teams and even by their own teams.

Through the develop model, it was possible to identify which styles were implemented throughout the 2020-2021 season of Bundesliga by the teams that constituted it through the application of a factor analysis and cluster analysis. In this way, the styles that were most prevalent in the teams were identified, evaluating their success and failure. Through the interpretation of the data collected from these teams, four performance profiles were identified: Peak performance, Regular performance, Poor performance, and Unstable performance. The study of these four clusters aimed to interpret and cross-reference the various styles of play identified, in order to explain the various patterns obtained. We conclude, based on the results obtained, that the teams that define their game by play Styles that promote "Open play", "Sustained Threat", and "Chances", are teams that enhance the achievement of good performances and consequently good results. The teams the focus their strategy in play styles such as “Take on” and “Duels” have performances considered to be worse and consequently obtain worse results.

Therefore, the research goal was achieved to identify various styles of play performed by teams of a specific football league and, later, identify the different performance profiles throughout a particular season. The results from the current study provide coaches’ the opportunity to analyze in detail the various aspects that characterize the teams in terms of style of play and performance. It will allow the identification of which aspects determine a particular type of performance and, thus, help technical teams structure their strategy and optimize coaches’ decisions to explore the identified critical aspects.

The use of factor analysis was beneficial in reducing the initial dataset to obtain a more focused and reliable analysis to identify the various play styles.

The application of cluster analysis allowed the exploration of the various existing patterns and thus identified the various performance indices developed by the teams of the main league in Germany in the 2020-2021 season.

## 6.2 FUTURE WORK AND LIMITATIONS

For further research, it will be essential to apply the location information to each event/variable to be more assertive in classifying each variable's typology. The inclusion of situational factors (event's location) of the variables allowed the analyst to identify precisely where the event happened. It allowed organizing the variables by area (Construction zone, Preparation zone, or Decision zone), which will help classify them concerning the style of play and characterize the different ball possessions developed along the defensive midfield or in the offensive midfield.

Adding the event's location can change the loadings obtained previously in Factor Analysis and, consequently, the organization of each factor because we have information about where each event occurs.

# REFERENCES


Bação, F., Lobo, V., & Painho, M. (2005). Self-organizing Maps as Substitutes for K-Means Clustering. LNCS, 3516, 476–483.

Bangsbo, J. (1994). The physiology of soccer - With special reference to intense intermittent exercise. *Acta Physiologica Scandinavica, Supplement*, *151*(619), 1–155.

Bangsbo, J., Mohr, M., & Krustrup, P. (2006). Physical and metabolic demands of training and match-play in the elite football player. Https://Doi.Org/10.1080/02640410500482529, 24(7), 665–674. https://doi.org/10.1080/02640410500482529

Bradley, P. S., Lago-Peñas, C., & Rey, E. (2014). Evaluation of the match performances of substitution players in elite soccer. *International Journal of Sports Physiology and Performance*, *9*(3), 415–424. https://doi.org/10.1123/IJSPP.2013-0304

Brymer, E., & Schweitzer, R. (2013). Extreme sports are good for your health: A phenomenological understanding of fear and anxiety in extreme sport. Journal of Health Psychology, 18(4), 477–487. https://doi.org/10.1177/1359105312446770

Bush, M., Barnes, C., Archer, D. T., Hogg, B., & Bradley, P. S. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. *Human Movement Science*, *39*, 1–11. https://doi.org/10.1016/j.humov.2014.10.003

Carling, C., Reiley, T., & Williams, A. M. (2009). Performance assessment for field sports.

Castellano, J., Casamichana, D., & Lago, C. (2012). The Use of Match Statistics that Discriminate Between Successful and Unsuccessful Soccer Teams. *Journal of Human Kinetics*, *31*, 139–147. https://doi.org/10.2478/v10078-012-0015-7

Chen, Y., Qin, B., Liu, T., Liu, Y., & Li, S. (2010). The Comparison of SOM and K-means for Text Clustering. Computer and Information Science, 3(2), 268–274. https://doi.org/10.5539/cis.v3n2p268

Churchill, S., & Hughes, M. (2005). Attacking Profiles of Successful and Unsuccessful Teams in Copa America 2001. In Science and Football V. Routledge. https://doi.org/10.4324/9780203412992-81

Churchill, S., & Hughes, M. (2005). Attacking Profiles of Successful and Unsuccessful Teams in Copa America 2001. In *Science and Football V*. Routledge. https://doi.org/10.4324/9780203412992-81

de Bodt, E., Cottrell, M., & Verleysen, M. (1999). Using the Kohonen algorithm for quick initialization of simple competitive learning algorithm. 7th European Symposium on Artificial Neural Networks. ESANN'99. Proceedings. D-Facto, Brussels, Belgium, August, 19–26.

Di Salvo, V., Baron, R., Tschan, H., Calderon Montero, F. J., Bachl, N., & Pigozzi, F. (2007). Performance characteristics according to playing position in elite soccer. *International Journal of Sports Medicine*, *28*(3), 222–227. https://doi.org/10.1055/S-2006-924294

Fernandez-Navarro, J., Fradua, L., Zubillaga, A., Ford, P. R., & McRobert, A. P. (2016). Attacking and defensive styles of play in soccer: analysis of Spanish and English elite teams. Journal of Sports Sciences, 34(24), 2195–2204. https://doi.org/10.1080/02640414.2016.1169309

Gai, D. Y. (2019). Physical, technical and tactiocal performance analysis of the Chinese football league super league. *Tesis Doctoral*.

Gómez, M. A., Gómez-Lopez, M., Lago, C., & Sampaio, J. (2012). Effects of game location and final outcome on game-related statistics in each zone of the pitch in professional football. European Journal of Sport Science, 12(5), 393–398. https://doi.org/10.1080/17461391.2011.566373

Guillermo Martinez Arastey. (2018, June 28). Gps technology in professional sports. https://www.sportperformanceanalysis.com/article/gps-in-professional-sports

Hair, J. F. (2011). Multivariate Data Analysis: An Overview. International Encyclopedia of Statistical Science, 904–907. https://doi.org/10.1007/978-3-642-04898-2_395

Haykin, S. (1999). Neural networks: a comprehensive foundation by Simon Haykin. The Knowledge Engineering Review, 13(4), 409–412. https://books.google.com/books/about/Neural_Networks.html?hl=pt-PT&id=bX4pAQAAMAAJ

Hodges, N. J., Starkes, J. L., & MacMahon, C. (2012). Expert Performance in Sport: A Cognitive Perspective. *The Cambridge Handbook of Expertise and Expert Performance*, 471–488. https://doi.org/10.1017/CBO9780511816796.027

Hughes, M. D., & Bartlett, R. M. (2010). *The use of performance indicators in performance analysis*. https://doi.org/10.1080/026404102320675602

Hughes, M., & Franks, I. M. (2015). Essentials of performance analysis in sport.

Kirkbride, A. (2013). Scoring/judging applications. Routledge Handbook of Sports Performance Analysis, 158–170. https://doi.org/10.4324/9780203806913-22

Lago-Peñas, C., Lago-Ballesteros, J., & Rey, E. (2011). Differences in performance indicators between winning and losing teams in the UEFA Champions League. *Journal of Human Kinetics*, *27*(1), 135–146. https://doi.org/10.2478/v10078-011-0011-3

Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., & Gómez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the Spanish soccer league. In *©Journal of Sports Science and Medicine* (Vol. 9). http://www.jssm.org

Liu, H., Gómez, M. A., Gonçalves, B., & Sampaio, J. (2015). Technical performance and match-to-match variation in elite football teams. *Https://Doi.Org/10.1080/02640414.2015.1117121*, *34*(6), 509–518. https://doi.org/10.1080/02640414.2015.1117121

Liu, H., Hopkins, W. G., & Gómez, M. A. (2016). Modelling relationships between match events and match outcome in elite football. *European Journal of Sport Science*, *16*(5), 516–525. https://doi.org/10.1080/17461391.2015.1042527

Liu, H., Hopkins, W., Gómez, A. M., Molinuevo, S. J., Gómez, M. A., & Molinuevo, J. S. (2013). Inter-operator reliability of live football match statistics from OPTA Sportsdata. *International Journal of Performance Analysis in Sport*, *13*(3), 803–821. https://doi.org/10.1080/24748668.2013.11868690

Manuel Clemente, F., Sarmento, H., Rabbani, A., I Van Der Linden, C. M., Kargarfard, M., & Teoldo Costa, I. (2018). *Variations of external load variables between medium-and large-sided soccer games in professional players*. https://doi.org/10.1080/15438627.2018.1511560

Marcelino, R., Mesquita, I., & Sampaio, J. (2011). Effects of quality of opposition and match status on technical and tactical performances in elite volleyball. *Journal of Sports Sciences*, *29*(7), 733–741. https://doi.org/10.1080/02640414.2011.552516

Martinez Arastey Guillermo. (2013). What is Performance Analysis in Sport? | Sport Performance Analysis. https://www.sportperformanceanalysis.com/article/what-is-performance-analysis-in-sport

Morgulev, E., Azar, O. H., & Lidor, R. (2018). Sports analytics and the big-data era. International Journal of Data Science and Analytics, 5(4), 213–222. https://doi.org/10.1007/s41060-017-0093-7

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. WIREs Data Mining Knowl Discov, 2, 86–97. https://doi.org/10.1002/widm.53

O'Donoghue, P. (2010). *Research Methods for Sports Performance Analysis*. https://books.google.pt/books?id=wW54AgAAQBAJ&printsec=frontcover&dq=Research+Methods+for+Sport+Performance+Analysis&hl=pt-PT&sa=X&redir_esc=y#v=onepage&q&f=false

O'Donoghue, P., & Mayes, A. (2013). Performance analysis, feedback and communication in coaching. Routledge Handbook of Sports Performance Analysis, 173–182. https://doi.org/10.4324/9780203806913-24

Rampinini, E., Impellizzeri, F. M., Castagna, C., Coutts, A. J., & Wisløff, U. (2009). Technical performance during soccer matches of the Italian Serie A league: Effect of fatigue and competitive level. *Journal of Science and Medicine in Sport*, *12*(1), 227–233. https://doi.org/10.1016/j.jsams.2007.10.002

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. SpringerPlus 2016 5:1, 5(1), 1–13. https://doi.org/10.1186/S40064-016-3108-2

Russell, M., & Kingsley, M. (2011). Influence of exercise on skill proficiency in soccer. *Sports Medicine*, *41*(7), 523–539. https://doi.org/10.2165/11589130-000000000-00000/FIGURES/TAB3

Senthilnathan Samithambe. (2019). Usefulness of correlation analysis. https://ssrn.com/abstract=3416918https://ssrn.com/abstract=34169182https://ssrn.com/abstract=3416918

Steinbach, M., V. Kumar, et al. (2006). Introduction to data mining.

Taylor, J., Mellalieu, S., James, N., & Shearer, D. (2008). The influence of match location, quality of opposition, and match status on technical performance in professional association football. Journal of Sports Sciences, 26(9), 885–895. https://doi.org/10.1080/02640410701836887

Tenga, A., & Larsen, Ø. (2017). Testing the Validity of Match Analysis to describe Playing Styles in Football. *Http://Dx.Doi.Org/10.1080/24748668.2003.11868280*, *3*(2), 90–102. https://doi.org/10.1080/24748668.2003.11868280

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. Applied Artificial Intelligence, 17(5–6), 375–381. https://doi.org/10.1080/713827180