Sports Data Analytics: An Analysis on Football Teams

Gurpreet Singh Deol

Supervisor: Professor Yun Kuen Cheung

Submitted for the Degree of Master of Science in

Data Science and Analytics

Department of Computer Science
Royal Holloway University of London
Egham, Surrey TW20 0EX, UK

September 7, 2022

Declaration

This report has been prepared on the basis of my own work. Where other published and unpublished source materials have been used, these have been acknowledged.

Word Count:

Student Name:

Date of Submission:

Signature:

Abstract

Using data collected on football teams from the Premier League, La Liga, Serie A, Bundesliga and Ligue 1, I have analysed how football has transitioned in the last 13 years. The analysis was performed using machine learning algorithms such as Principal component analysis and K-means. The analysis is divided into different segments, focusing on a general analysis, defensive performance, passing variation, and finishing ability. Teams in general have shifted more towards a possession and passing approach rather then crosses and long balls over the last 13 years.

Each league is also analysed separately to determine the nature of the competitiveness and the variance in playing styles. The Bundesliga and Ligue 1 were determined to be the least competitive. The La Liga had two dominant teams across the decade exhibiting distinct playing styles. The Serie A had the most competitive upper league table. The Premier League has seen the rise of two new title contenders, almost imitating the La Liga.

Contents

1 Introduction

Football is a global sport and a global market, with the Premier league alone estimated to be worth around £8 billion. It can essentially be seen as a global business, and successful businesses make decisions through careful data analysis. Hence it is no surprise that data analysis has been rising in importance in football. This includes driving decisions for marketing and profits as well as analysing performance.

The focus of this project is analysing performance. Tracking data initially started to develop around the 1990s, with time, the accuracy and variety of data has only served to increase since then. Analysing the data could very well tell a different story to the performance on the pitch and provide a different point of view in analysing a players or a teams playing style. It could help clubs find certain types of players with specific traits and qualities. No matter the purpose, the data collected can be used for various purposes.

Using data from previous seasons, we can analyse a teams performance to observe how they have changed and how it has affected their success. Taking it a step further, we can analyse leagues. Is there a distinct playing style for each league? If so, how has it changed over the years? Do successful teams across different leagues share any similarity? Are teams that were successful 10 years ago still successful now? These questions fall under a general umbrella, how has football changed across the years? That is the focus and goal of this project. Which will be answered by analysing leagues and teams across the years.

The data used is collected from the top 5 leagues in Europe, Premier league, La Liga, Budesliga, Serie A and Ligue 1. However, we can only use data from the 2009/2010 season to the present, as data before then is very inconsistent. Even with this limit, for some attributes we can only analyse the data for the last 5 seasons for the same reasons. Data analysis is key part of this project, therefore it is vital to understand some underlying concepts that will be used for our task.

2 Theoretical Background

The information present in this section is based on notes from Professor Yun Kuen Cheung [3] as well as on the Mathematical methods for Physics and Engineering [6]. Data collection in sports is highly advanced now, resulting in a lot of unique attributes of data after each match. These attributes

include crosses, forward passes, tackles made in the opposition box to name a few. While this provides a large quantity of data for analysis, it also leads to an issue with the visualisation of the results. A maximum of 3 attributes can be compared at once using a standard 3D grid, which is highly limiting considering the sheer amount of attributes of data. It is also incredibly time consuming and wasteful to use every possible combination of 3 attributes. To solve this issue, we first need to determine what type of learning algorithm we will be using. There are two types, supervised learning and unsupervised learning.

2.0.1 Supervised and Unsupervised Machine Learning

Data sets can be labeled or unlabeled, supervised machine learning algorithms first build a model using a labeled training data set, this model can then be used to predict labels for other data sets. On the other hand, unsupervised machine learning algorithms are useful for unlabeled data sets, where you are trying to find the structure of the data. The goal of this project is to determine if and how football has changed by analysing data and trying to find any patterns or hidden structures in the data, therefore it is preferable to use unsupervised learning.

There are a few unsupervised machine learning algorithms that can be used, such as K-Means clustering and Principle Component Analysis(PCA). We are dealing with a large dataset with multiple attributes, therefore PCA is more appropriate, however the K-means algorithm is also used in conjunction with PCA. Principle component analysis helps perform dimension reduction, simply put, it combines as much information as possible from the data and projects it onto newly calculated dimensions. This is an oversimplified explanation, the exact process and description is explained in the following sections.

2.1 Structure of Data sets

Before focusing on how machine learning algorithms are used, it is vital to understand the structure of data sets we could encounter. A lot of real world data sets have a large amount of observations as well as a large amount of attributes (high-dimensional), such as the case of sports data sets. Some of these attributes may be redundant or irrelevant to the purpose of the task. Redundant attributes are usually strongly linearly correlated to each other, such as age and date of birth especially since age can simply be calculated from the date of birth. Irrelevant data can be seen as random noises, which

can critically affect the result of the analysis if they are significant. Another aspect to consider is boredom, which occurs when most of the values in an attribute are clustered close to the mean of the attribute, commonly measured using the variance of the attribute[3].

Considering these aspects, we want attributes that have a wider range of values, since they will be "less boring". In other words, we want to maximise the variance in our dataset, however, as stated before, data sets in the real world have a large amount of dimensions, it is a tedious task to manually go through each attribute to check for the aforementioned aspects. Therefore to calculate variance in high dimensional datasets, we need linear algebra[3].

2.2 Linear Algebra

We begin with vectors, an object with magnitude and direction, and their properties, such as norm and dot product. Consider two vectors, v and u[6].

2.2.1 Vectors and their dot product

$$v = (v_1 + v_2 + v3)u = (u_1 + u_2 + u_3) \tag{1}$$

The dot product of such vectors is calculated as[6]:

$$\langle v, u \rangle = v_1u_1 + v_2u_2 + v_3u_3 \tag{2}$$

The norm of a vector is calculated as[6]:

$$||v|| = \sqrt{\langle v, v \rangle} \tag{3}$$

2.2.2 Orthogonality[6]

The vector v is a unit vector if $||v|| = 1$. The dot product is important because it shows us if two vectors are orthogonal(Perpendicular to each other) or not. If we think of attributes as vectors, then if they are orthogonal to each other, they are also uncorrelated, which is helpful in identifying redundant data. To check if two vectors are orthogonal, they must satisfy the following condition:

$$\langle v, v \rangle = 0 \tag{4}$$

2.2.3 Projection

To understand projection, consider a set of non-zero vectors $v_1, v_2, v_3, ..., v_k$ where every pair of vectors is orthogonal. Now consider another vector x, the projection of x on the subspace spanned by $v_1, v_2, v_3, ..., v_k$ can be calculated as[3]:

$$\frac{\langle x, v_1 \rangle}{\langle v_1, v_1 \rangle}.v_1 + \frac{\langle x, v_2 \rangle}{\langle v_2, v_2 \rangle}.v_2 + \frac{\langle x, v_3 \rangle}{\langle v_3, v_3 \rangle}.v_3 + ... + \frac{\langle x, v_k \rangle}{\langle v_k, v_k \rangle}.v_k \tag{5}$$

Dissecting the first term, we divide the dot product between x and v_1 by the dot product of v_1 and itself, multiplied by the vector v_1. Projection is a key part of PCA, in layman's terms, PCA projects vector x onto some new vectors (components). However, it is important to calculate appropriate components for the projection, such that the values of the projection of x are still as close as possible to the original values of x. To achieve this, we need to use eigenvectors and eigenvalues.

2.3 Eigenvectors and Eigenvalues

Vectors can be seen as either column or row vectors, depending their shape, however they are still one dimensional. Matrices however, are two dimensional. It is necessary to understand some concepts related to matrices before learning about eigenvectors, such as the transpose as well as checking the symmetry of a matrix.

2.3.1 Transpose and Symmetry

The transpose of a $a_1 \times a_2$ matrix A is $a_2 \times a_1$, it is denoted by A'. The following equation is true for $1 \leq i \leq a_1$ and $1 \leq j \leq a_2$ [6]

$$A_{ij} = A'_{ji} \tag{6}$$

For example,

$$\begin{bmatrix} 2 & 4 & 7 \\ 3 & 1 & 9 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 4 & 1 \\ 7 & 9 \end{bmatrix}' \tag{7}$$

Symmetry only applies to a $a \times a$ square matrix, a matrix which has dimensions of equal length, when $A = A'$[6].

$$\begin{bmatrix} 3 & 7 & 2 \\ 7 & 9 & 4 \\ 2 & 4 & 5 \end{bmatrix} \tag{8}$$

Drawing a straight line from the top left corner to the bottom right corner, shows us that the matrix is symmetrical, as each number can be found in the opposite space across the line.

2.3.2 Determinant

The determinant is a number calculated from a square matrix denoted with |||[6].

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \tag{9}$$

It can be calculated with the following formula for a 3 by 3 matrix,

$$|A| = a(ei - fh) - b(di - fg) + c(dh - eg) \tag{10}$$

The process has the same basic principle for larger square matrices.

2.3.3 Covariance matrix

A covariance matrix shows us the directional relationship between two variables and shows us the variance of each variable. It can be calculated using[10],

$$cov(X, Y) = \frac{1}{N-1} \Sigma_{i=1}^{N} (X_i - \overline{x})(Y_i - \overline{y}), \tag{11}$$

where $\overline{x}$ and $\overline{y}$ are the mean of each variable. For example, consider a matrix with test scores for 5 students[4],

$$\begin{array}{ccc} Maths & English & Arts \\ \left(\begin{array}{ccc} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{array} \right) & & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \end{array}$$

The covariance matrix for this data is a $d \times d$ square matrix[4],

$$\begin{array}{c} \begin{array}{ccc} Maths & English & Arts \end{array} \\ \begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} \end{array} \begin{array}{c} Math \\ English \\ Art \end{array}$$

The diagonal numbers from the top left to the bottom right tell us the variance of each attribute, in this example, art has the most variance and is the least "boring" attribute. If a number is positive between two attributes, it suggests a positive relationship, if one attribute increases so does the other. Whereas a negative value would suggest the opposite. A zero suggests that there is no obvious relationship between two attributes. The covariance matrix can also be applied to much larger datasets with greater dimensions and it allows us to measure the variance and see the relationship between attributes.

2.3.4 Eigenvalues

An eigenvector can be defined as vector whose direction remains unchanged when you apply a linear transformation to it. For a square matrix A, u (non-zero vector) is an eigenvector of A if there exists λ, a real number, such that[6]:

$$A.u = \lambda.u \tag{12}$$

where λ is the eigenvalue of the eigenvector u. In the example below, the eigenvalue is 3 and the eigenvector is [1 1 2].

$$\begin{bmatrix} 3 & 4 & -2 \\ 1 & 4 & -1 \\ 2 & 6 & -1 \end{bmatrix} . \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 6 \end{bmatrix} = (3). \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \tag{13}$$

This however is a very simple scenario where the eigenvalue is intuitive thus making the eigenvector a simple task to calculate. In most other cases, we first need to calculate the eigenvalue and then proceed towards the eigenvector. We can calculate the eigenvalues using equation 10 and an identity matrix I, a square matrix where each diagonal entry from the top left to the bottom right is filled with 1s while every other entry is 0, It can be seen as the matrix equivalent of the number 1, since if you multiply it with a matrix A, the outcome is still A[6].

Inserting the identity matrix into equation 12, with the assumption that u is non zero, and doing some rearranging,

$$\begin{aligned} Au &= \lambda Iu \\ Au - \lambda Iu &= 0] \\ |A &= \lambda I| = 0, \end{aligned} \tag{14}$$

which is just the determinant[6]. We solve this to find the eigenvalues, For example,

$$\begin{aligned} |\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}| &= 0 \\ |\begin{bmatrix} -6-\lambda & 3 \\ 4 & 5-\lambda \end{bmatrix}| &= 0 \\ (-6-\lambda)(5-\lambda) - (3 \times 4) &= 0 \\ \lambda^2 + \lambda - 42 &= 0 \\ \lambda &= -7, 6 \end{aligned} \tag{15}$$

After finding the determinant and solving the quadratic equation, we can find the eigenvalues, -7 and 6.

2.3.5 Eigenvectors

The next step involves equation 12 again, we can calculate the first eigenvector using the first eigenvalue of 6,

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 6 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\begin{aligned} -6x + 3y &= 6x \\ 4x + 5y &= 6y \\ \\ -12x + 3y &= 0 \\ 4x - y &= 0 \end{aligned} \tag{16}$$

$$\begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

The eigenvector is [1 4], repeating this process but with -7 gives us the second eigenvector. This is a simple explanation on how to calculate eigenvalues and eigenvectors. A few properties about eigenvectors we need to understand going forward are, that any $b \times b$ square matrix has b pairwise orthogonal eigenvectors and that each eigenvector has a corresponding eigenvalue[6].

2.4 Principle Component Analysis

With the basic underlying theories explained, we can now focus on how PCA works. The process of PCA is divided into multiple steps, which includes two processing steps known as centering and normalization. We then need to calculate unit vectors (Principal components) that maximise the variance in the dataset.

2.4.1 Centering

Centering is performed on the data, which ensures that the centroid of a set of observations is a zero vector. The centroid is calculated as the average of the observations[3]:

$$x_1 = \begin{bmatrix}1\\2\end{bmatrix}, x_2 = \begin{bmatrix}2\\-4\end{bmatrix}, x_3 = \begin{bmatrix}1\\-1\end{bmatrix}, centroid = \begin{bmatrix}\frac{1+2+1}{3}\\\frac{2-4-1}{3}\end{bmatrix} = \begin{bmatrix}0\\-1\end{bmatrix} \tag{17}$$

We then subtract the centroid from each observation to obtain the centered observations.

$$x_1 = \begin{bmatrix}1\\1\end{bmatrix}, x_2 = \begin{bmatrix}2\\-5\end{bmatrix}, x_3 = \begin{bmatrix}1\\-2\end{bmatrix} \tag{18}$$

2.4.2 Centering and Variance

The variance cam be calculated from the centered dataset, by finding the mean of the squares of each dimension[10].

$$\frac{1}{N}\Sigma_{i=1}^{N} x_i^2 \tag{19}$$

Using the previous centered dataset:

$$\frac{1}{3} \times [1^2 + 2^2 + 1^2] = 2\frac{1}{3} \times [1^2 + (-5)^2 + (-2)^2] = 10 \tag{20}$$

The variance along the first dimension is 2 and it is 10 for the second dimension. Centering is important as it ensures the resulting components of PCA only use the variance within the dataset and not the overall mean of the the dataset as a key variable .

2.4.3 Normalization

Since our dataset contains various attributes, they all have different ranges as well. The dataset contains the total amount of goals scored by a team every season as well as the goals scored per game attributes, the range for goals per game would be somewhere between 0 and 4 depending on the season, whereas for total goals, it could easily be around 50. If we use our dataset in this state, the total goals per season will have a significant impact on the calculation of the components, compared to the goals per game data. Therefore, it is important to normalise our data to ensure that all of the variables have the same standard deviation and that they have the same weight for PCA.

One method for normalization is to divide the values by the standard deviation, which is the square root of the variance[3]. The following values are the normalized values from the previous example.

$$x_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{10} \end{bmatrix}, x_2 = \begin{bmatrix} 2/\sqrt{2} \\ -5/\sqrt{10} \end{bmatrix}, x_3 = \begin{bmatrix} 1/\sqrt{2} \\ -2/\sqrt{10} \end{bmatrix} \tag{21}$$

Normalization can be skipped if all of the attributes are to the same order of magnitude.

2.4.4 Principal Components

One of the main goals of PCA is to perform dimension reduction, which essentially means transferring as much information as possible from the original dimensions onto some new components. These components are the principal components and each component holds a different amount of the original information.

Calculating principal components involves a few of the previously mentioned steps, since they were all explained individually, I will describe the process for calculating principle components. Firstly, using the centered and normalised data with k attributes, we build a k by k covariance matrix, which shows us the variance in the data. We then calculate the eigenvalues and hence the eigenvectors of the covariance matrix.The eigenvectors

are the principal components that we are after and the eigenvalues tell us the amount of information present in the component. We want to preserve as much of the original information as possible, hence we desire the eigenvectors with the largest eigenvalues. Therefore we sort the eigenvectors in descending order, with the eigenvector with the largest eigenvalue at the top and select the top n desired eigenvectors. The largest eigenvalue eigenvector is known as the first principal component, with successive principal components named in a similar manner.

The final step involves projecting our data onto the selected principal components, which could be done using the projection formula defined earlier, since all the eigenvectors are pairwise orthogonal. Therefore allowing us to see multi dimensional data in a plot in 2 or 3 dimensions. This would normally be a very tedious and time consuming process but fortunately it can be done with a few simple steps in Python or other programming languages. For this project, all of the steps were performed in Python.

2.4.5 Variance Plot

A variance plot just shows us how much information the principal components represent, it helps us choose an appropriate value for n, the number of eigenvectors. A good rule of thumb is to select components that preserve around 75% of the information. An example of a variance plot is shown in Figure 2.

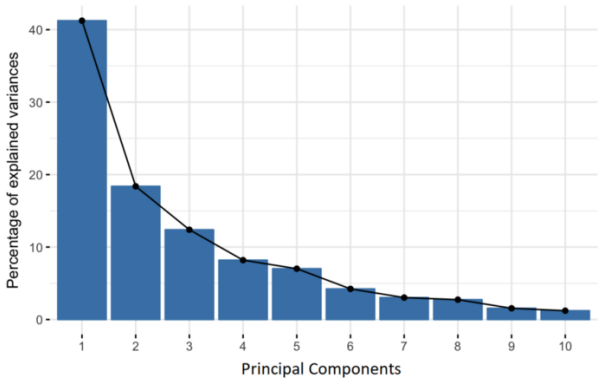


Figure 1: A simple variance plot showing the amount of information retained with principal components[5]

The percentage of information represented by each principal component

decreases sharply from the first to the tenth principal component. In this case it is enough to choose the first 3 principal components as they explain about 75% of the information.

2.4.6 Understanding a PCA Plot

To understand a PCA plot, we need to know how the original attributes relate to the principal components. Figure 3 shows a simple example of PCA plot, with both the projected data and the original attributes. The MPG attribute is negatively correlated with the both the first and second principal component, which means that the further to the bottom left a car is, the more efficient it is. whereas cars in the top left have much better acceleration. The projected variables show us how each observation performs relative to each variable. With a large enough dataset, they may also form groups or clusters, observations with the same labels, which is what we are after,,although our data is originally unlabeled.

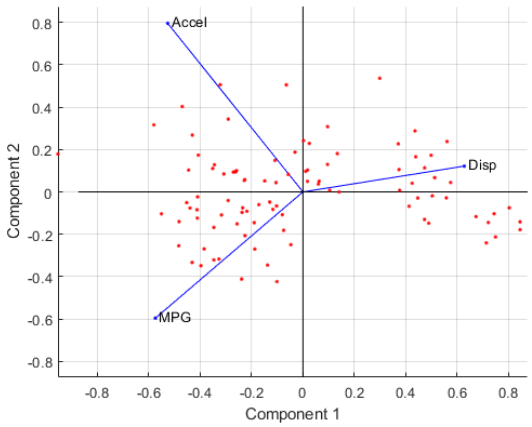


Figure 2: A simple PCA plot of data on cars with attributes Miles per gallon, displacement and acceleration[9]

We can workaround this by using some existing attributes as the final labels in the output. For example, once PCA has been performed and the output is some graph with multiple points representing each observation, we can use attributes such as the season or the league as labels. Using a unique colour for each league, we can see if each league forms their own clusters, or if teams in different seasons played differently. On the other hand, using the k-means algorithm, we can create new labels, hence new clusters for the

PCA plot.

2.5 K-Means Algorithm

For a dataset with x and y variables, after centering and normalisation if necessary, we can create n clusters after setting some simple initial conditions. Such as the number of clusters to be made and the number of iterations(number of times the algorithm should repeat) for the algorithm, the higher the iteration value, the more accurate the clusters are. However it is necessary to set a loop breaking rule to stop the algorithm once the values for the clusters are no longer changing. The algorithm first creates n random centroids, these are the center of each cluster. For each observation, the distance between it and each centroid is calculated, this observation is then assigned to that cluster. The distance can be calculated in multiple ways, such as Euclidean, used in this report, or Manhattan distance. After this first step, new centroids are calculated using the average of the observations assigned to that cluster. The process repeats for the number of iterations or when values converge and no longer change.

The K-means algorithm only works for observations with a maximum of 3 attributes, hence unless the original plot is in a maximum of 3 dimensions, it is more appropriate to use PCA. However, the whole purpose of PCA is dimension reduction, making the output a 2 or 3 dimensional plot. Therefore, the K-means algorithm can be applied to the output of PCA.

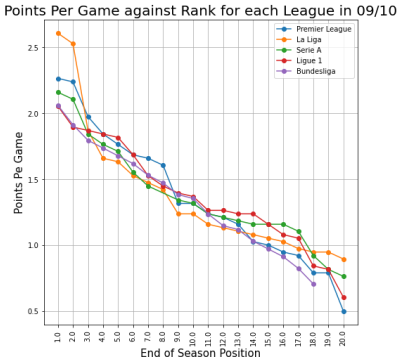
3 Analysis

The data used for analysis was collected manually and includes 97 attributes, while data for some attributes is missing for seasons before 2017/2018. The attributes were combined into groups to analyse different aspects of the sport. For example defensive attributed like tackles and clearances were used together where as passing attributes like through balls and short passes were grouped together, the code can be found in the appendix. The analysis is divided into multiple sections, depending on the leagues and teams.

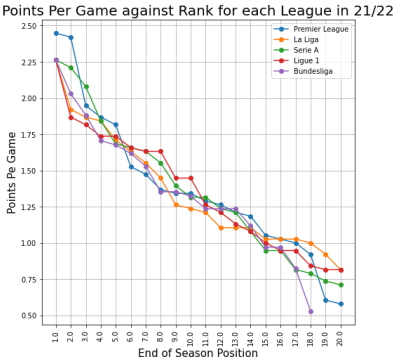
3.1 Football over 13 Years

3.1.1 Points Per Game

To begin with, we can analyse the points per game for each team in 2009/2010 and in 2021/2022. Points per game can show us if there are



(a) A plot of points per game for each team in the top 5 leagues for the 09/10 season.



(b) A plot of points per game for each team in the top 5 leagues for the 09/10 season.

Figure 3: Comparison between the 09/10 and 21/22 season.

any dominant teams or any group of teams that contest for certain positions. Any large gap between two consecutive teams (such as the first and second placed teams) suggests a difference in quality and skill between the two sides. A side by side analysis is shown in figure 3.

In the 09/10 season, there is sharp fall off (0.7 difference) for the third ranking team in La Liga, relative the first and second teams. This shows that the two giants of Spanish football, Barcelona and Real Madrid comfortably occupied the first two spots in the league. While there is also a similar drop for the third ranking team in the Premier League and Serie A (0.3 for both), it is not as exaggerated as the La Liga. An interesting observation is that the first and fifth ranking teams in Ligue 1 are only separated by about 0.3 points per game, meaning there is no distinct dominant team in the league and that first place is highly contested for. For the same positional difference, the points difference is 1 for the La Liga and 0.5 for both Serie A and the Premier League, highlighting that the fact the La Liga is mostly a two horse race. The Bundesliga on the other hand follows an almost linear pattern and is the most balanced out of all the leagues.

Comparing this to the 2021/2022 season, the script has slightly changed. The Premier League had two dominant teams who competed for the title, Manchester City and Liverpool, with a large fall for the third position. The Ligue 1 and La Liga winners actually had a relatively large difference compared to their runner up teams, showing that they dominated the league last season. The most contested league was the Serie A, with the top 3 teams

only being separated by a mere 0.15 points per game. even from this simple analysis we can get a glimpse into how the competitiveness of the leagues has changed over 13 years, we can further analyse these changes by using PCA and conducting a general analysis, with attributes relating to defending, passing and shooting to paint an overall picture of football across the last 13 years.

3.1.2 Football in the Top 5 Leagues

Figure 4 shows a plot of variance, the first 6 principal components combined explain about 80% of the information present in the data, which is ideal, however the first two components that are plotted only show about 55% of the information.

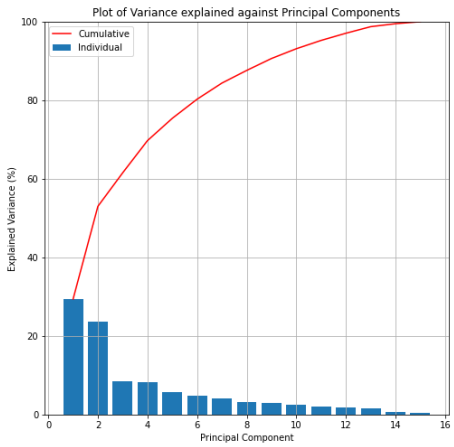


Figure 4: A plot of the explained variance and the number of principal components.

The attributes chosen, and their correlation with the principal components can be seen in figure 5. Short passes and pass success are both positively correlated with the first principal component, where as defensive attributes such as tackles and fouls correlate more positively with the second principal component.

The projected data across all 13 seasons between 2009 and 2022 can be seen in figure 6. It is interesting to see that each season forms distinct clusters, with the oldest season at the top and each successive season below it. A lot of defensive attributes correlate positively with the second principal component, and point towards the top left as seen in figure 5. That is also

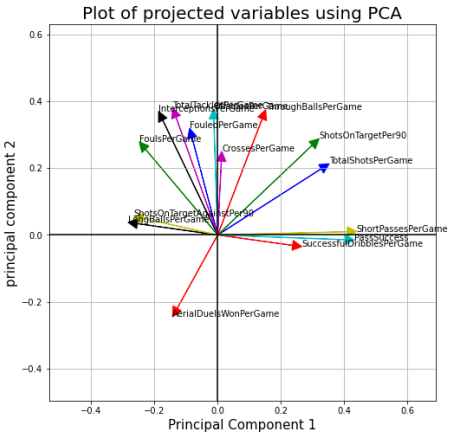


Figure 5: A plot of the projected variables and their correlation with the principal components.

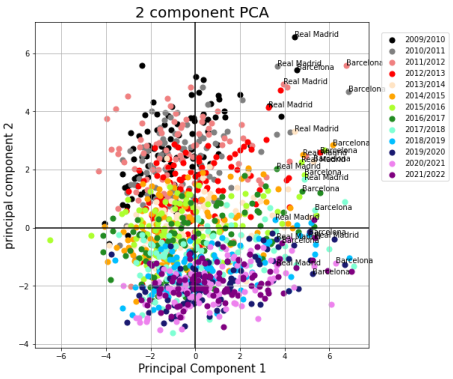


Figure 6: A PCA plot, showing projected data with a few labeled attributes, using the first two principal components calculated using various attributes from data between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

where most of the teams before 2013 are, showing us that teams were more defensive back then relative to now, since each successive season has drifted away towards the bottom right, the opposite direction.

Short passes and pass success both correlate positively with the first principal component, and it is no surprise to see extremely successful teams like Barcelona and Real Madrid far to the right of the rest of the teams. Each season, they both form a distinct group, however, other teams can be

seen gravitating towards them with each season, especially after 2016. This could show the general style of football changing, from a more defensive to aggressive style across this period of time.

This divide is further highlighted in figure 7, which shows the teams before and after the 2015/2016 season.

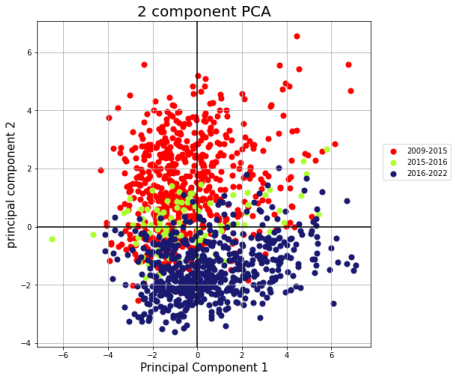


Figure 7: A PCA plot, showing projected data, using the first two principal components calculated using various attributes from data between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

3.1.3 Defensive

Since we know, from the general analysis, that teams have become less defensive over the years, we can analyse this in more detail by using defensive attributes only, such as tackles, fouls and aerials won per game. As seen in figure 8, most attributes negatively correlate with the first principal, suggesting that teams further to the left are more defensive. Blocked shots, red cards and shots against per game all point towards the bottom left, suggesting teams received more shots and blocked that much more. Teams placed towards the left generally make more tackles and interceptions per game, where as the higher they are along the y axis, the more times their players get dribbled past.

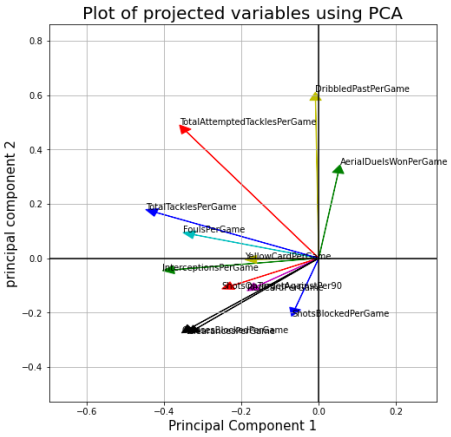


Figure 8: A plot of the projected variables and their correlation with the principal components for defensive attributes.

The teams from 2009/2010 and 2010/2011 are in the bottom left of the plot in figure 9, where as newer seasons are towards the top right, further highlighting the change in defensive tactics. The lack of fouls per game in newer seasons suggests that teams now have a less aggressive style of defending and that the general quality of defending is better. As fouls are more likely to occur when a tackle or a challenge is made with a lack of precision. There is another angle to consider when explaining this change, which will be explained when discussing passing attributes.

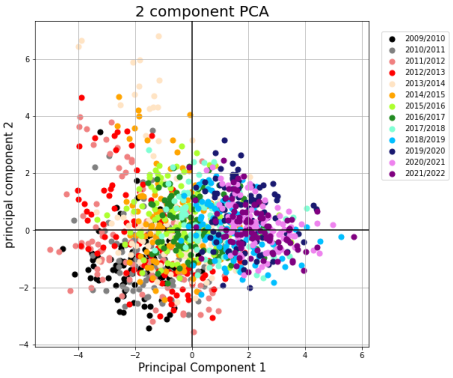


Figure 9: A PCA plot, showing projected data with a few labeled attributes, using the first two principal components calculated using various defensive attributes from data between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

Dribbled past correlates positively with the second principal component, therefore teams placed higher up tend to get dribbled past more then others. Newer seasons are relatively higher on the plot, suggesting that teams now have more aggressive players who like to dribble more towards opponents. This observation makes sense as a lot of successful teams rely on very skilled wingers. For example Liverpool and Manchester City, who have Mo Salah and Ryiad Mahrez respectively, players who thrive on beating defenders with their dribbling and pace.

3.1.4 Passing

While tactics may differ from team to team, passing is an integral part of the sport. However, there are many types of passes which will depend on the teams playing style. Some teams prefer to play short passes to build up play, while others prefer playing long balls. Using data relating to different types, we can see if there are any distinct groups. The selected attributes are shown in figure 10.

Long passes, crosses and through balls correlate positively with the second principal component, therefore teams placed further up like to get the ball forward from the back quicker, Whereas short passes and possession correlate positively with the first principal component, teams placed further to the right like to build up play through short passes and possession. Teams in the lower right have players who like to dribble through the opposition.

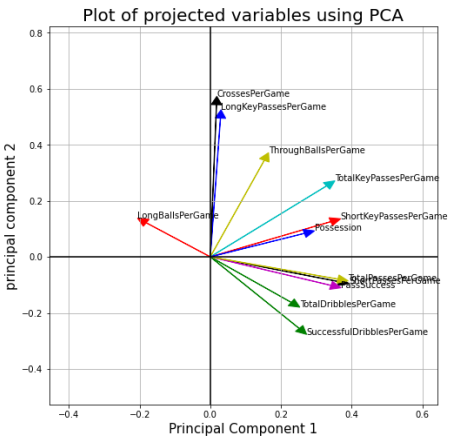


Figure 10: A plot of the projected variables and their correlation with the principal components for passing attributes.

While it is possible to see some sort of clustering in figure 11, it is more

distinct in figure 12 which divides the data before and after the 2016/2017 season. Observing both plots, it is no surprise to see Barcelona standing out, almost like an outlier, towards the extreme right hand side. Barcelona are known for their "Tiki Taka" playing style, playing quick short passes to build up play, as well as one of the greatest players of our generation, Lionel Messi. Who is one of the best dribblers of the ball in the world, Therefore Barcelona gravitating towards both of those attributes is no surprise.

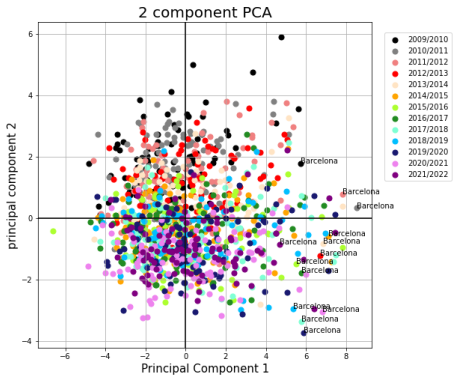


Figure 11: A PCA plot, showing projected data with a few labeled attributes, using the first two principal components calculated using various passing attributes from data collected between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

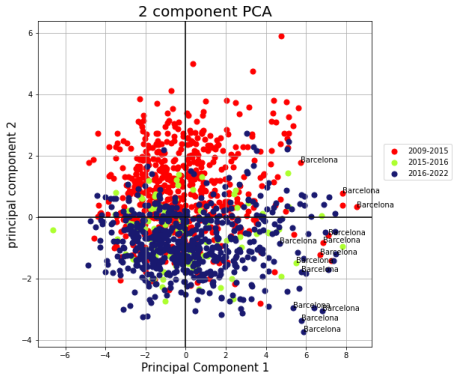


Figure 12: A PCA plot, showing projected data, using the first two principal components calculated using various passing attributes from data collected between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

However, it is still clear that teams played with more long balls and crosses before then now, while the opposite is true for short passes and dribbles. This leads back to understanding the change in defending. Fouls are more likely to occur when the ball is in the air and two players both try to gain possession of it, since the moment you have less control over your body once you are in the air. Considering the fact that teams played more crosses and long passes before, you were much more likely to be in the aforementioned state back then compared to now. Hence the chances for fouls was much higher.

Furthermore, with teams playing more short passes, it is vital for teams to maintain their defensive formation, rather then having players charge towards the opposition, since short passes are harder to intercept then others, unless you force the opposition to make mistakes with high pressure. If a gap opens up in the defense, it is very easy for current forwards, who have great pace and dribbling skills to exploit it. Hence, from this angle, it makes sense why there are less tackles in games compared to before.

3.1.5 Shooting

Scoring goals is the main aim in football, analysing how and where players score from may give us some insights into other aspects. In figure 13, attributes with ratio in their name is the number of goals scored per shot from that distance.

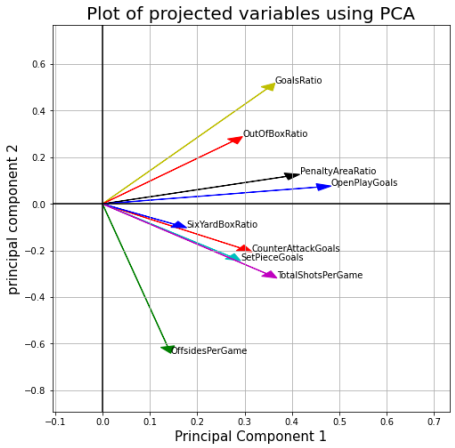


Figure 13: A plot of the projected variables and their correlation with the principal components for shooting attributes.

The goals ratio, number of goals scored per shot is positively correlated with both principal components. While most other attributes correlate positively with the first principal component, with some relationship with the second principal component.

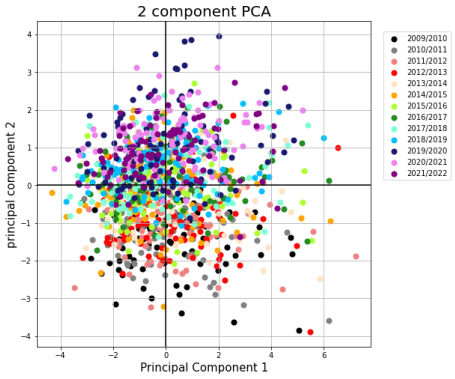


Figure 14: A PCA plot, showing projected data with a few labeled attributes, using the first two principal components calculated using various shooting attributes from data collected between 2009 and 2022. Each point is a team from the top 5 leagues that has played in the leagues.

Once again in figure 14, we see a distinction between each season, with the newer seasons at the top, showing a transition over the years. Teams score more per shot in general from various locations now then before yet take less shots per game in general. An interesting point to note is that there were more offsides per game in the older seasons, which makes sense as there were more through-balls per game in the older seasons as seen in figure 10. More through balls means that players make a run behind the defence line more often, therefore you are more likely to get caught offside if you mistime your run or if the opposition plays an offside trap.

We will now focus on analysing the top 5 leagues in Europe. Since the focus is their playing styles, the best method using the data at hand is to use passing attributes shown in figure 10. The way teams pass tells us a lot about their playing style, teams with more long passes prefer more direct football by getting the ball forward as quickly as possible, where as more short passes suggests teams prefer to slowly build up play. Since the data we have is for the last 13 seasons, we will use data from the 09/10, 15/16 and 21/22 seasons for the analysis. This shows us how teams played 13 years ago, how they play now and see if we can notice a transition in between.

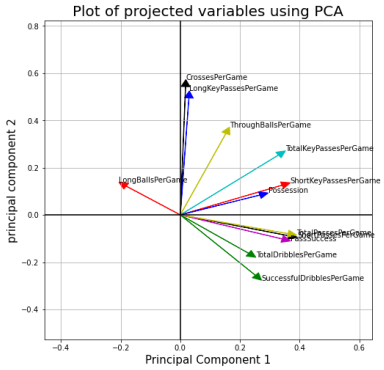
3.2 Premier League

The Premier League (PL) is the highest division of football in England, consisting of 20 Teams who play 38 matches each season. In the 90s and 00s, the league was dominated by Manchester United, with the only other contenders being Arsenal and Chelsea. However, they have had a lack of success since their manager for 25 years, Sir Alex Ferguson, retired in 2013, they have not won the PL since. Arsenal could be considered to be in a similar state, while Chelsea have had successful seasons since then, even winning the Champions League in the 2020/2021 season. Manchester City, Liverpool and Chelsea have been quite successful in the last decade. We can analyse the league over time and see how these teams have changed and how that has affected their success using the attributes shown in figure 10. The results are shown in figure 15 alongside the variables plot.

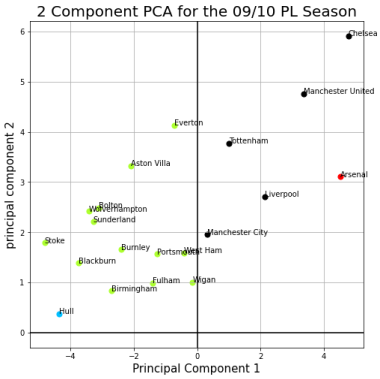
The 09/10 season in figure 15e forms a cluster high up the second principal component, which corresponds to attributes such as long passes and crosses. The other two seasons correlate more with the first principal component, which has attributed such as short passes and dribbling, The 15/16 season is between the 09/10 and 21/22 season and the teams blend in with the teams in the 21/22 season. This could highlight a transition occurring in the teams playing styles, especially considering the fact that this was the season when Jurgen Klopp joined Liverpool, one of the current most successful teams in the league.

As stated earlier, the most successful teams in the PL in the last 30 years are Manchester United, Arsenal, Chelsea, Liverpool and Manchester City[7]. In the 09/10 season in figure 15b, the standout teams are Chelsea, Manchester United and Arsenal, who formed a unique cluster. The first two are in the top right of the plot, suggesting these teams played long passes, crosses and through balls more often per game relative to the other teams, where as Arsenal adopted more of a passing and through balls approach. While the top teams standout, the rest of them are clustered in green near the bottom left. The teams also appear to be in a linear trend, from the bottom left to the top right. The decrease in the number of teams following this trend shows the increase in quality of the teams as well as the general standings at the end of the season.

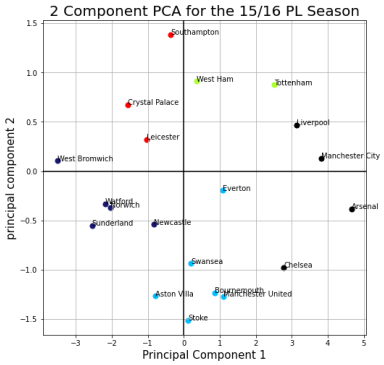
The 15/16 season shows the most diverse clusters out of the 3 seasons, with some teams playing a similar style to 09/10 and some adopting different styles. Arsenal, City and Liverpool have all adopted a passing and possession style of football, since they are placed far along the first principal component. Unexpectedly, the winners this season were Leicester City, who



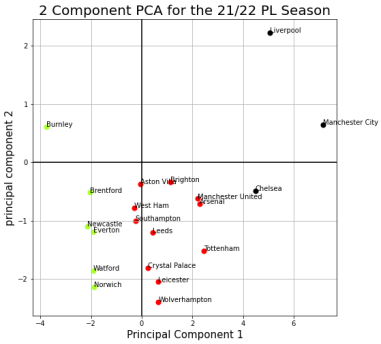
(a) Plot of the projected variables.



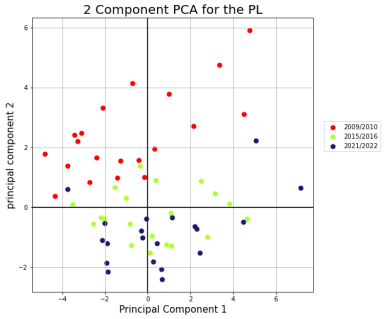
(b) Plot for the 09/10 season.



(c) Plot for the 15/16 season.



(d) Plot for the 21/22 season.



(e) A plot showing all of the data together using the seasons as the key

Figure 15: These plots show a K-Means PCA analysis of the Premier League using passing attributes. Plot (a) shows the projection of the variables. Plots (b), (c) and (d) shows the data for individual seasons, with each colour representing a cluster. Plot (e) shows all of the seasons together, with each season distinguished using the legend.

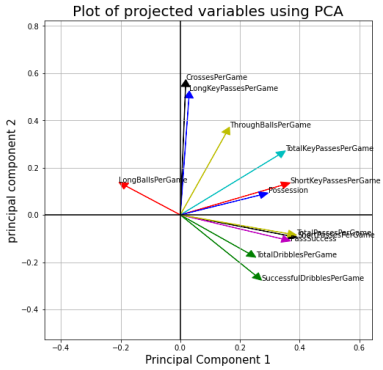
had only recently been promoted. Correlating with long passes and crosses, they played completely differently to the aforementioned teams. Manchester United however are in a position that City were in 09/10. The change in managers after Sir Alex Ferguson retired, resulted in a lack of direction for the club. Whereas other teams adapted and changed like City, United were neither playing like they were before, or adapting a new playing style, leading to one of their worst seasons in the modern history of the club.

With the most recent season, 21/22, the shear gulf in between teams in easily noticeable. It can essentially be dubbed as Manchester City and Liverpool against the rest. Both of these teams are probably the best their respective managers have managed at the clubs. No surprise seeing Pep Guardiola's City playing with short passes and possession, the same as he did when he managed Barcelona. Liverpool on the other hand performing more crosses and through balls, making full use of their wing backs (Left and right backs who play a more offensive role). Their playing style is reflective of their team structure, with City having the better midfield allowing them to maintain possession and Liverpool the better wingers giving them more opportunities to use the full width of the pitch with more crosses. The rest of the teams form two main groups highlighted in green and red, showing that this season was just a two horse race, like the 09/10 season but with different horses.

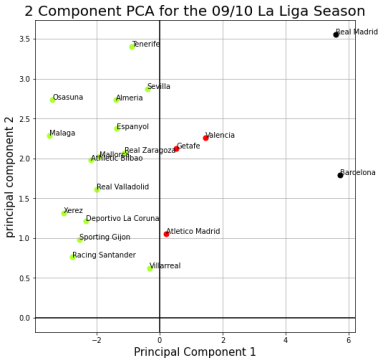
3.3 La Liga

The La Liga (LL) is the highest division of football in Spain, hosting giants such as Barcelona and Real Madrid. The analysis follows the same format as the PL. Similar to the PL, the La Liga has two teams that have been dominant each season, the only difference is that they have not changed. Apart from the 13/14 season which saw Athletico Madrid lift the trophy, the only other winners since 2004 have been Barcelona and Real Madrid[8].

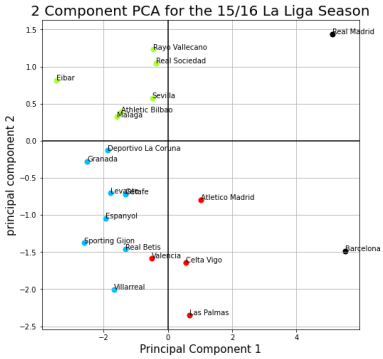
Figure 16e shows a similar transition to the PL except for the clusters for the 15/16 and 21/22 seasons, which overlap more then the PL. Across each plot in figure 16, it is clear to see the gulf between the top 2 teams and the rest. In fact, even Barcelona and Real Madrid are quite distinct from each other. Barcelona far to the right in each plot, known for playing short passes and keeping possession to build up play. Especially considering that their midfield consisted of some of the greatest midfielders with the likes of Iniesta and Xavi. Real Madrid in the top right in plots a and b, playing crosses and long balls, with talented wingers such as Ronaldo and Bale and attacking fullbacks like Marcelo



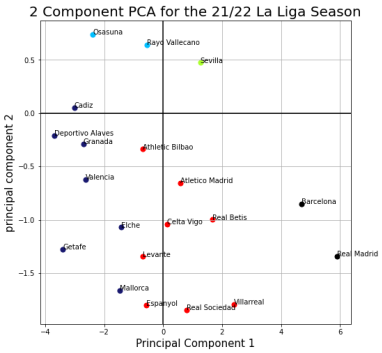
(a) Plot of the projected variables.



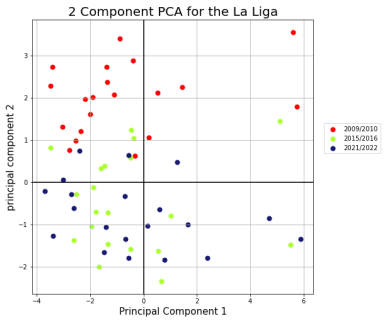
(b) Plot for the 09/10 season.



(c) Plot for the 15/16 season.



(d) Plot for the 21/22 season.



(e) A plot showing all of the data together using the seasons as the key

Figure 16: These plots show a K-Means PCA analysis of the La Liga using passing attributes. Plot (a) shows the projection of the variables. Plots (b), (c) and (d) shows the data for individual seasons, with each colour representing a cluster. Plot (e) shows all of the seasons together, with each season distinguished using the legend.

Both of these teams are very different yet also very successful, Madrid plays more direct football where as Barcelona plays with a possession approach. These two styles counter each other in a lot of ways, since playing a possession style almost requires you to have a high defensive line. Since your opponent will have to sit back to defend, your midfield is required to push up to help attack which leads to the defensive line also pushing up to prevent a counter attack in the case of possession loss. The obvious weakness of a high defensive line is the amount of space behind the defense, and the best method to exploit that space is to attack quickly with speed, a characteristic Madrid players have in abundance. The PL has had almost the exact same situation in recent seasons, with City and Liverpool as mentioned before. Liverpool's team structure is very similar to Madrid's, same with City and Barcelona.

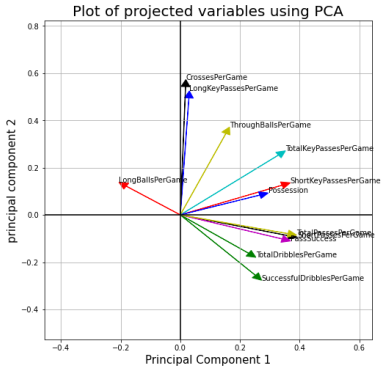
Finally, although the gap between these two teams and the rest of the league has reduced in the 21/22 season, it still exists. With Messi leaving the club at the start of the season, Barcelona were definitely missing an integral part of their team and entered a sort of a transition period. With their only real competitor gone, it is no surprise that Madrid won the La Liga and even the Champions League this season. Regardless of the season however, the lack of quality from the rest of the competitors is quite evident.

3.4 Serie A

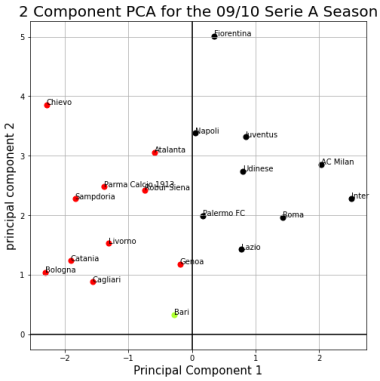
The Serie A is the highest division of football in Italy, The K-means PCA plot is shown in figure 17 alongside the variables plot. Figure 17e shows the previously witnessed pattern, however, unlike the PL and LL, the density of teams for each season is fair;y uniform.

In the 09/10 season, the teams generally form two groups, the top 10 teams in black while the rest in red. Compared to the PL and La Liga, there is no distinct gap between some teams except for Florentina, who finished 11th that season. The clustering of the top teams, such as Inter, Roma and AC Milan could highlight the fact that a lot of these teams play a similar style of football.

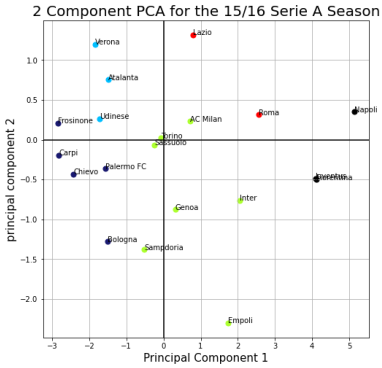
In comparison with the 09/10 PL and La Liga season, United and Chelsea are above $y = 5$ on the y axis and both Barcelona and Real Madrid are at around $x = 6$ on the x axis. Both of these leagues had teams who play a very distinct type of football, which was also very effective. Apart from Florentina, the range in both the x and y axis for the black cluster is only 3. The majority of teams in the black cluster play a balanced style, a mix of possession and direct football.



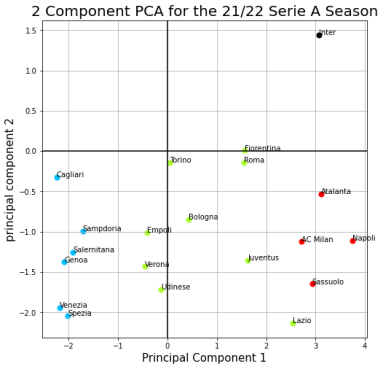
(a) Plot of the projected variables.



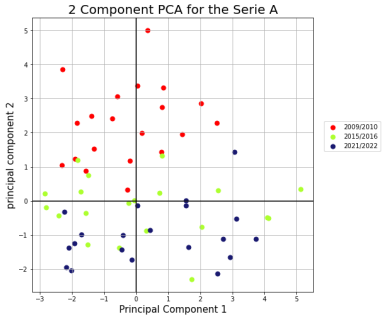
(b) Plot for the 09/10 season.



(c) Plot for the 15/16 season.



(d) Plot for the 21/22 season.



(e) A plot showing all of the data together using the seasons as the key

Figure 17: These plots show a K-Means PCA analysis of the Serie A using passing attributes. Plot (a) shows the projection of the variables. Plots (b), (c) and (d) shows the data for individual seasons, with each colour representing a cluster. Plot (e) shows all of the seasons together, with each season distinguished using the legend.

This lack of diversity would normally suggest that the league is very competitive between the top teams, since a distinct playing style may lead to an advantage over a lot of your opponents, as we have seen in the PL and La Liga. However the Serie A has only seen 3 unique winners in the last 13 seasons, with Juventus winning 9 in a row.[1] Second position varied between Milan, Inter, Roma and Napoli. The difference between first and second place was only one point in the 19/20 season. So while Juventus have been dominant in the league, the rest of the teams have been very competitive on several occasions, especially for the top 5 positions.

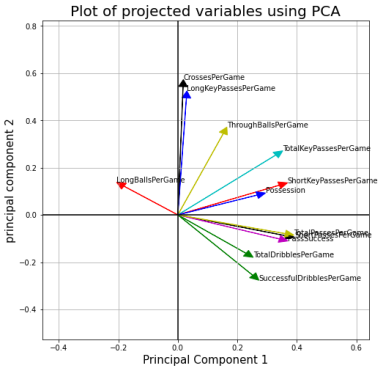
The 15/16 and 21/22 seasons show more teams further along the first principal component rather then the second, showing a trend we have seen for teams with teams adopting a possession style of football. The teams in each cluster are also very close to each other, with no large differences between clusters, a very different picture compared to the PL and La Liga, further supporting the point made earlier about competitiveness.

3.5 Ligue 1

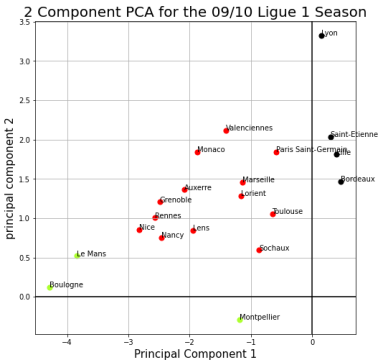
Ligue 1 is the highest division of football in France, the league has become more popular in recent years due to star players like Neymar and Messi playing in the league.

Ligue 1 is generally seen as a contender for the worst league in the top 5, especially in recent seasons due to Paris Saint-Germain (PSG). As seen in figure 18b, almost all of the teams are in the top left quadrant of the graph, with the overall range in the y axis being about 3.5 (6 for the PL) and no team past 0.5 on the x axis. The general standard of football was definitely lower relative to the rest of the leagues. While all of the previously mentioned leagues have had teams in the top left quadrant, they have also had teams on the opposite end of the spectrum. Apart from the Bundesliga, 20 teams play in every league for the final trophy, the final table can be divided into 3 groups, top, middle and bottom. Difference in quality between teams will exist, but there will be teams fighting for high positions in the middle and bottom of the table. The reason the PL is generally seen as a very competitive league is due to this reason, due to having teams of different quality, the competition is tense throughout the table. The point is that the Ligue 1 lacked any teams that could contend with perhaps even the mid table teams in the PL or Serie A.

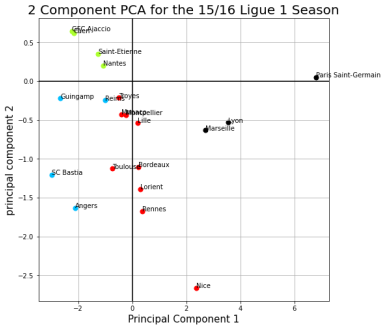
This would change in 2011 once PSG changed owners and had a tremendous increase in funding, they won the league 8 times since then[11]. From being an inconspicuous team in 09/10, they became almost an outlier in the



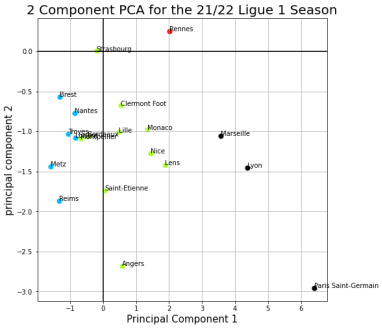
(a) Plot of the projected variables.



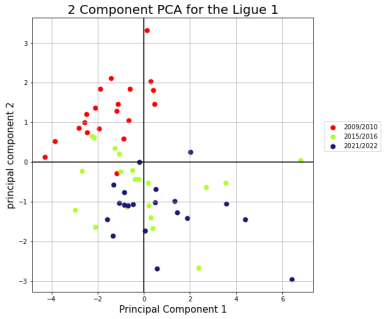
(b) Plot for the 09/10 season.



(c) Plot for the 15/16 season.



(d) Plot for the 21/22 season.



(e) A plot showing all of the data together using the seasons as the key

Figure 18: These plots show a K-Means PCA analysis of Ligue 1 using passing attributes. Plot (a) shows the projection of the variables. Plots (b), (c) and (d) shows the data for individual seasons, with each colour representing a cluster. Plot (e) shows all of the seasons together, with each season distinguished using the legend.

15/16 season, placing far along the first principal component, adopting a possession style of football, The rest of the teams forming two main groups, Lyon, Marseille and the rest. If PSG were out of the picture, Ligue 1 would be similar to the PL and LL with having two dominant teams in Lyon and Marseille.

An interesting difference can be seen between 21/22 and 15/16, although PSG is still placed relatively high along the first principal component, it is in the bottom right corner. This direction correlates to the dribbling attribute. With PSG having signed Neymar in 2017 and Messi in 2021, arguably two of the greatest dribblers of all time, it is no surprise the team to have such an increase in dribbles per game. Although PSG have been successful domestically, they are yet to attain any Champions League success.

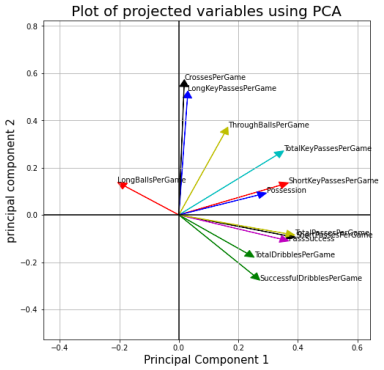
3.6 Bundesliga

The Bundesliga is the highest division of football in Germany, home to giants such as Bayern Munich and Borussia Dortmund. The K-means PCA analysis can be seen in figure 19.

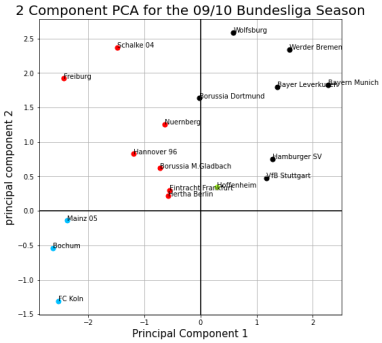
The general layout of teams in the 09/10 season is very similar to the Serie A, with a lot of the top ranking teams clustering together. The general approach is a balanced playing style since the teams are near the top right corner, apart from Dortmund. The range along the y and x axis is also very similar to the Serie A, being 4 and 5.5 respectively.

Between 2000 and 2010, the league saw 5 different champions, highlighting the competitiveness at the top of the table. This changed post 2010, where apart from Dortmund winning one season, Bayern Munich have just won their 10th league in a row[12]. Following a similar pattern to PSG, Bayern essentially becomes an outlier compared to the rest of the league. They are placed far along the x axis in both the 15/16 and 21/22 seasons, indicating to their high possession and passes per game. While Dortmund has shifted more towards the bottom right, in line with the dribbling attribute. This is not surprising since they have had talented wingers like Adnan Januzaj and Jadon Sancho. Although the Bundesliga may appear to suffer the same issue as Ligue 1, there are more mid table table teams in the Bundesliga compared to Ligue 1, but the league is still dominated by Bayern Munich.

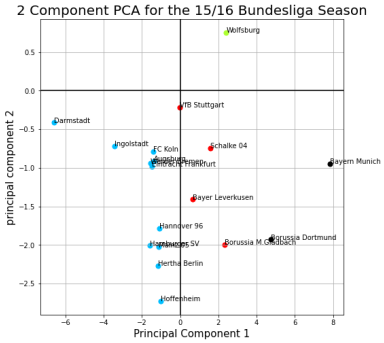
Another angle to consider, for why both PSG and Bayern Munich place so high along the first principal component, is that the general quality of opponents is so low that PSG and Bayern naturally dominate possession rather then saying they adopt such a playing style. This is especially so



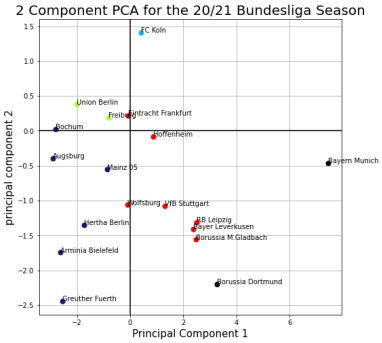
(a) Plot of the projected variables.



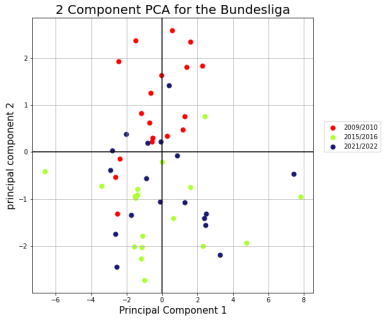
(b) Plot for the 09/10 season.



(c) Plot for the 15/16 season.



(d) Plot for the 21/22 season.



(e) A plot showing all of the data together using the seasons as the key

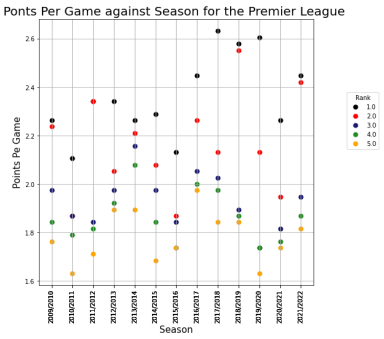
Figure 19: These plots show a K-Means PCA analysis of the Bundesliga using passing attributes. Plot (a) shows the projection of the variables. Plots (b), (c) and (d) shows the data for individual seasons, with each colour representing a cluster. Plot (e) shows all of the seasons together, with each season distinguished using the legend.

considering the gulf of quality in each league.

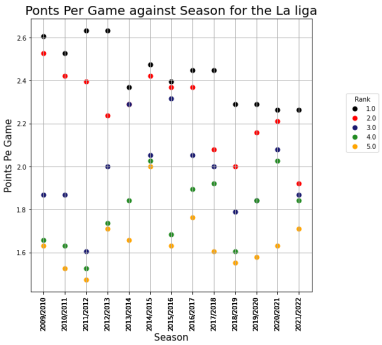
3.7 Points Analysis

The competitiveness can also be seen in figure 20. The top 2 teams in the Serie A have been relatively close to each other, apart from a few seasons such us 13/14. the average range between the first and fifth team appears to be around 0.6. For the Bundesliga, the gap between the top two teams is evident in most seasons. The gap slowly increases in 09/10, maximises in in 13/14 to 0.6 between the top two teams, the average between the first and fifth is about 0.9 across all seasons. The Ligue 1 follows a similar trend with the 11/12 and 20/21 seasons being an exception. The average range between the first and the fifth team is also about 0.9. Considering the fact the range between between the top two teams, in both Ligue 1 and the Bundesliga, is similar to the gap between the top 5 teams in Serie A shows the gulf in competitiveness in the leagues.

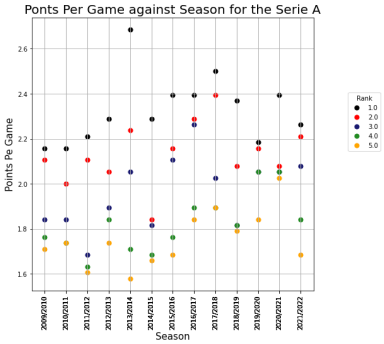
The top two teams in the La Liga have been extremely close across the years, with an average range of about 0.1, yet the average range between the first and fifth has been about 0.9, further supporting the fact that two teams have remained dominant across the years. The Premier League has had various fluctuations. Some seasons, such as 12/13, 15/16, 17/18 and 19/20 shows a clear dominant team, yet there have also been seasons where multiple teams have been close such as 13/14, or seasons with two teams challenging for the title such as 09/10, 18/19 and 21/22.



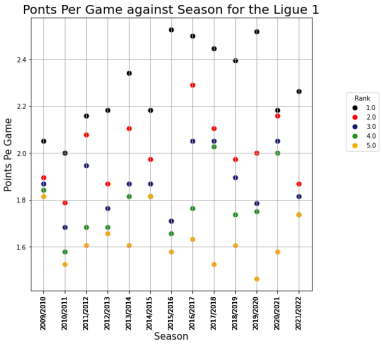
!ht (a) Plot for the Premier League.



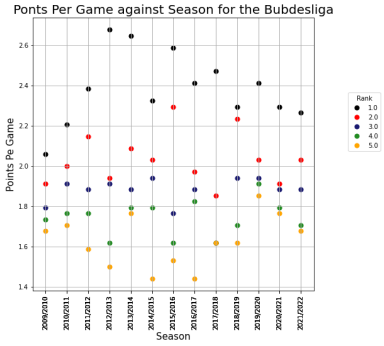
(b) Plot for the La Liga.



(c) Plot for the Serie A.



(d) Plot for the Ligue 1



(e) Plot for the Bundesliga

Figure 20: These plots show the points per game for the top 5 teams in each season in each league. The

4 Conclusion

Football has definitely changed across the last 13 years. In terms of defense, teams are now more organised compared to before. With each successive season, more successful teams, such as Manchester City, have adopted a possession style of football, while some still maintain a direct approach, such as Liverpool and Real Madrid.

Apart from the latest season, the La Liga has not seen any changes to teams that have dominated the league for over a decade, Barcelona and Real Madrid, both playing a different style of football. The Premier League has seen a change in the regular title contenders, yet the number of teams challenging for the title has not changed. The competitiveness for the title has also fluctuated across the years. The Serie A is probably the league that is generally the most competitive for the top of the table, since there is no glaring gulf that separates most teams from a one.

The Bundesliga and Ligue 1 share a similar fate, with one team essentially dominating the league for the past decades, PSG has seen an increase in funding, alongside the addition of star players, essentially creating the gulf of quality that exists in the league, with Bayern Munich mirroring PSG in the Bundesliga.

5 Professional Issues; Data Collection and Usage

For any data analysis project, data acquisition is key part of the process. Data is a very important issue in modern times, as it is a very powerful tool. Most companies use data to guide their business, such as making decisions based of sales data. Since such practise has been successful, businesses may take it a step further and use more personal information to make more focused decisions. This obviously leads to an issue with privacy and data security. Is it justified for businesses to collect, store, and use personal data from customers? What exactly is defined as 'personal' data and who has the rights to this data? While the principal of data collection may seem harmless at first, it can have enormous consequences depending on how it is used.

Most websites you visit collect data on your activity during that time, such as the things you search for or interact with the most. This information is then used to recommend products or services that you are likely to purchase. Social media websites work in a similar fashion. The posts you see depend on what type of content you consume. You will see more content

related to food if you interact with food related posts more often. While it is easy to see that these two simple processes only show you things that you prefer, that is a point that can and has been exploited. Generally, when you have to make a decision from multiple choices, you can consider the pros and cons of each and base your decision on the evidence you have. A lack of information for one choice while an abundance for another will naturally create a bias in your decision. With social media, if you are only shown information from one perspective, it limits your ability to evaluate the entire problem leading you to form a biased opinion. This simple problem has had global effects.

Over the last decade, multiple allegations, for different elections, have occurred suggesting that the voters were influenced to vote one way or the other by being shown specific posts on social media[2]. They would receive positive posts about one political group, while negative posts for others. The choice would depend on the data collected for each voter, such as the type of posts they interact with. Russia was accused of using such means to influence the 2016 general election in the United States[2], as well as during the EU referendum in the UK. With such far reaching effects, it is easy to see why data collection and usage is such an important and sensitive topic. A major point of controversy was the fact that the data had been sold to third parties, who then allegedly used it for the aforementioned purposes. This leads to questions like should personal data be collected at all, what type of information is allowed to be collected and who decides what the data can be used for?

While there are a lot of laws that were introduced to prevent or limit such activities, it is vital to ensure organisations abide by them and be held responsible for misuse. This topic is very relevant to this project as I had to collect the data I used for my analysis. The data I collected and used was performance data related to football teams, while it is collected from companies contracted by teams and leagues, it was a challenge to obtain. I had to use a few different methods to scrape the data from websites. While my purpose was for performance analysis, it is important to understand for any project related to data, that how you collect and use data requires a lot of consideration. I had to make sure I did not collect any sensitive or private information, only attributes which are already public and openly available.

6 Self assessment

Throughout my project, I have enjoyed analysing teams using techniques I have learned. As someone who enjoys partaking in various sports, it was very interesting to look at teams through data instead of just watching them, giving a unique perspective into their performance.

Analysing my process in completing my project, there were a few stages where I faced challenges and where I could have perhaps been more efficient. One of the steps I struggled with the most was data acquisition. Until I actively started to search for data to analyse, I struggled to find any sources fulfilling my needs. This proved to be a major stumbling point as it drove progress to a halt. I decided to collect data my self by using any resources I could find, this led to a long process of data collection, storage and cleaning. It was also time consuming since I had to make sure the data was consistent and accurate. Due to my lack of experience, I underestimated the difficulty of this step. Another challenge was cleaning the data, as a lot of string values were inconsistent across different sources. I had to first fix the inconsistencies and then use unique keys (the team name plus the season) for each value to match and combine data from the multiple different files they were stored in.

I understood that planning every aspect of the project is very important, even if some parts may appear to be simple, it is better to have planned for unexpected outcomes or challenges. While it may not help solve the issue, it makes the process of finding the solution efficient. Working on this project has made me interested in performing data analysis in sports, so I will try to apply what I have learned into other sports I enjoy to see what interesting observations I can find.

References

[1] Serie A. Honours list. https://www.legaseriea.it/en/serie-a/roll-of-honour, 2022.

[2] Hernan A.Makse Alexandre Bovet. Influence of fake news in twitter during the 2016 us presidential election. https://doi.org/10.1038/s41467-018-07761-2, 2019.

[3] Yun Kuen Cheung. CS5100 Data Analysis, Chapter 9: Unsupervised Learning - Exploratory Data Analysis.

[4] Akash Dubey. The mathematics behind principal component analysis. https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643, 2018.

[5] Zakaria Jaadi. A step-by-step explanation of principal component analysis (pca). https://builtin.com/data-science/step-step-explanation-principal-component-analysis, 2021.

[6] S.J.Bence K.F.Riley, M.P.Hobson. *Mathematical Methods For Physics And Engineering*. Cambridge University Press, Cambridge, third edition, 2006.

[7] Premier League. Premier league explained. https://www.premierleague.com/premier-league-explained, 2022.

[8] La Liga. Top laliga santander winners of all time. https://www.laliga.com/en-GB/news/top-laliga-santander-winners-of-all-time, 2021.

[9] Matalb. Biplot. https://www.mathworks.com/help/stats/biplot.html.

[10] Lindsay I Smith. A tutorial on principal components analysis, February 2002.

[11] Wikipedia. List of french football champions. https://en.wikipedia.org/wiki/List_of_French_football_champions, 2022.

[12] Wikipedia. List of german football champions. https://en.wikipedia.org/wiki/List_of_German_football_champions, 2022.

7 Appendix

Here is the code I used for the majority of my analysis.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
pd.options.mode.chained_assignment = None


#Inputs: 'all' or desired variables

selected_leagues =  ['Premier League']
selected_seasons = ['2009/2010','2015/2016','2021/2022']
#selected_seasons = ['2021/2022']
selected_teams = 'all'
selected_ranks = 'all'

pc = 'principal component 2'

#Type of Legend: 'Leagues', 'Year', 'Teams','2 Year' , 'KM'
type_of_plot = 'Year'

label = 'Teadm'
selective_teams = ['']

#2 Year plot?

ans = 'no'

#K-means
n = 5


#Load data
data = pd.read_csv (r'.\Complete Dataset 2.csv')
data = data[data.League != 'Champions League' ]
```

```
if ans=='yes':
    data['yearsplit'] = ''

    y1 = ['2009/2010', '2010/2011', '2011/2012', '2012/2013', '2013/2014',
    y2 = [ '2016/2017', '2017/2018', '2018/2019','2019/2020', '2020/2021',

    for i in range(0,len(data)):
        for j in range(0,6):

            if data.Season[i] == y1[j] :
                data.yearsplit[i] = '2009-2015'
            elif data.Season[i] == y2[j]:
                    data.yearsplit[i] = '2016-2022'
            elif data.Season[i] == '2015/2016':
                        data.yearsplit[i] = '2015-2016'
    unique_twoyearlabels = data.yearsplit.unique()


club_string_var = {'Key','Team','League','Season','Rank','yearsplit' }
season_standings = {'Key','Rank','Games','Wins','Draws','Losses','GoalsFor

all_key_attributes = {'Key','Team','League','Season','Rank','yearsplit','P
                      'TotalShotsPerGame',
                      'CrossesPerGame','LongBallsPerGame', 'ThroughBallsPerGame
                      'ShotsOnTargetPer90','ShotsOnTargetAgainstPer90','AerialD
                      'SuccessfulDribblesPerGame', 'OffsidesPerGame','Intercept
                      'FoulsPerGame','FouledPerGame', 'TotalTacklesPerGame'}

passing_attributes = {'Key','Team','League','Season','Rank','yearsplit','P
                      'CrossesPerGame','LongBallsPerGame', 'ThroughBallsPerGame
                      'SuccessfulDribblesPerGame','TotalDribblesPerGame','Total
                      'LongKeyPassesPerGame','ShortKeyPassesPerGame'}

defensive_attributes = {'Key','Team','League','Season','Rank','yearsplit',
                      'ShotsOnTargetAgainstPer90','AerialDuelsWonPerGame','Clea
                      'DribbledPastPerGame', 'InterceptionsPerGame',
                      'FoulsPerGame','FouledPerGame', 'TotalTacklesPerGame'}
```

```
data['OutOfBoxRatio']=data['OutOfBoxGoalsPerGame']/data['OutOfBoxShotsPerGa
data['SixYardBoxRatio']=data['SixYardGoalsPerGame']/data['SixYardBoxShotsPe
data['PenaltyAreaRatio']=data['PenaltyAreaGoalsPerGame']/data['PenaltyAreaS
data['GoalsRatio']=data['GoalsPerGame']/data['ShotsOnTargetPer90']

shooting_attributes = {'Key','Team','League','Season','Rank','yearsplit','
                       'OffsidesPerGame', 'OutOfBoxShotsPerGame', 'SixYardB
                       'GoalsPerGame', 'SixYardGoalsPerGame', 'PenaltyAreaG
                        'OpenPlayGoals', 'CounterAttackGoals', 'SetPieceGo

shooting1_attributes = {'Key','Team','League','Season','Rank','yearsplit',
                       'OffsidesPerGame', 'OutOfBoxRatio','SixYardBoxRatio
                        'OpenPlayGoals', 'CounterAttackGoals', 'SetPieceGo


defensive1_attributes = {'Key','Team','League','Season','Rank','yearsplit'
                    'ShotsOnTargetAgainstPer90','AerialDuelsWonPerGame','Clea
                    'DribbledPastPerGame', 'InterceptionsPerGame', 'YellowCar
                    'FoulsPerGame', 'TotalTacklesPerGame','ShotsBlockedPerGam

playing_attributes = {'Key','Team','League','Season','Rank','yearsplit','T
       'TouchesDefThird', 'TouchesMidThird', 'TouchesAttThird',
       'TouchesAttPen', 'LiveTouches', 'NumOfPlayersDribbledPast', 'Nutmegs
       'Controlled', 'DistMovedWithBall', 'ProgressiveDistMoved', 'ProgC',
       'ProgressiveIntoFinalThird', 'ProgressiveInto18Yard', 'Miscontrols'
       'MiscontrolsAfterTackle', 'ProgressivePassReceived'}

playing1_attributes = {'Key','Team','League','Season','Rank','yearsplit','
       'TouchesDefThird', 'TouchesMidThird', 'TouchesAttThird','TouchesAttI
       'CounterAttackGoals', 'SetPieceGoals'}


pad = {'Key', 'Team', 'League', 'Season', 'Rank','yearsplit', 'Possession'
    'SixYardGoalsPerGame', 'PenaltyAreaGoalsPerGame', 'OutOfBoxGoalsPerGame
       'OpenPlayGoals', 'CounterAttackGoals'}


gk = { 'Key','Team','League','Season','Rank','yearsplit','ShotsBlockedPerGa
       'CrossesBlockedPerGame', 'TiotalSavesPerGame', 'SixYardSavesPerGame
```

```
        'PenaltyAreaSavesPerGame', 'OutOfBoxSavesPerGame'}

key_attributes = pad

selected_data = pd.DataFrame(data, columns= key_attributes)
selected_data.fillna(0, inplace=True)
team_profile = pd.DataFrame(data, columns= club_string_var)
team_standings = pd.DataFrame(data, columns= season_standings)


#sorted_data = selected_data[selected_data.Season =='2018/2019']
sorted_data = selected_data
top_data = sorted_data.reset_index(drop=True)
top_teams_profile = pd.DataFrame(top_data, columns = club_string_var)
top_data.drop(club_string_var,axis=1, inplace=True)

#PCA Calculation

data_centered = top_data.apply(lambda x: x-x.mean())
scaler = StandardScaler()
data_centered[data_centered.columns]= scaler.fit_transform(data_centered[d
pca = PCA(n_components= len(key_attributes)-len(club_string_var))
principalComponents = pca.fit_transform(data_centered)

principalDf = pd.DataFrame(data = principalComponents[:,0:3]
             , columns = ['principal component 1', 'principal component 2'
finalDf1 = pd.concat([principalDf, top_teams_profile], axis = 1)

#K-means

kmeans = KMeans( init="random", n_clusters = n,max_iter=300,random_state=4

dat = finalDf1[['principal component 1', 'principal component 2']].values
km1 = kmeans.fit(dat)
km = pd.DataFrame(data = km1.labels_, columns = ['KMeansLabel'])
finalDf = pd.concat([finalDf1, km], axis = 1)

#Select Data to be shown

if selected_leagues == 'all':
```

```
    finalDf = finalDf
    unique_leagues = finalDf.League.unique()
else:
    finalDf = finalDf[finalDf['League'].isin(selected_leagues)]
    unique_leagues = selected_leagues

if selected_seasons == 'all':
    finalDf = finalDf
    unique_seasons = finalDf.Season.unique()
else:
    finalDf = finalDf[finalDf['Season'].isin(selected_seasons)]
    unique_seasons = selected_seasons

if selected_teams == 'all':
    finalDf = finalDf
    unique_teams = finalDf.Team.unique()
else:
    finalDf = finalDf[finalDf['Team'].isin(selected_teams)]
    unique_teams = selected_teams

if selected_ranks == 'all':
    finalDf = finalDf
    unique_ranks = finalDf.Rank.unique()
else:
    finalDf = finalDf[finalDf.Rank <= selected_ranks]
    unique_ranks = finalDf.Rank.unique()


#Plotting

fig = plt.figure(figsize = (8,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Principal Component 1', fontsize = 15)
ax.set_ylabel(pc, fontsize = 15)
ax.set_title('2 Component PCA for the Bundesliga', fontsize = 20)


fig2 = plt.figure(figsize = (8,8))
ax2 = fig2.add_subplot(1,1,1)
ax2.set_xlabel('Principal Component')
```

```
ax2.set_ylabel('Explained Variance (%)')
ax2.set_title('Plot of Variance explained against Principal Components')
ax2.set_ylim([0  , 100])
ax2.plot(range(1,len(pca.explained_variance_ratio_)+1), (pca.explained_var
ax2.bar(range(1,len(pca.explained_variance_ratio_)+1), (pca.explained_vari
ax2.legend()
ax2.grid()


fig1 = plt.figure(figsize = (8,8))
ax1 = fig1.add_subplot(1,1,1)
ax1.set_xlabel('Principal Component 1', fontsize = 15)
ax1.set_ylabel(pc, fontsize = 15)
ax1.set_title('Plot of projected variables using PCA', fontsize = 20)




if type_of_plot == 'Leagues':

    colors = ['black','red','greenyellow',
              'deepskyblue', 'midnightblue','violet']
    for league, color in zip(unique_leagues,colors):
        indicesToKeep = finalDf['League'] == league
        ax.scatter(finalDf.loc[indicesToKeep, 'principal component 1']
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
    ax.legend(unique_leagues,loc = 'lower right',bbox_to_anchor=(1.25, 0.5

elif type_of_plot == 'Year':

#     colors = ['black','grey','lightcoral','red','bisque','orange','greeny
 #              'aquamarine','deepskyblue', 'midnightblue','violet','purple

    colors = ['red','greenyellow','midnightblue', 'aquamarine','purple','o
              'violet','aquamarine','deepskyblue','grey',
              'lightcoral' ,'bisque']
    for season, color in zip(unique_seasons,colors):
        indicesToKeep = finalDf['Season'] == season
```

```
        ax.scatter(finalDf.loc[indicesToKeep, 'principal_component_1']
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
    ax.legend(unique_seasons,loc = 'lower_right',bbox_to_anchor=(1.25, 0.5

elif type_of_plot == 'Teams':

    colors = ['black','red', 'midnightblue','forestgreen','orange','greeny
              'violet','aquamarine','deepskyblue','purple','grey',
              'lightcoral' ,'bisque']
    for team, color in zip(unique_teams,colors):
        indicesToKeep = finalDf['Team'] == team
        ax.scatter(finalDf.loc[indicesToKeep, 'principal_component_1']
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
    ax.legend(unique_teams,loc = 'lower_right',bbox_to_anchor=(1.25, 0.5))

elif type_of_plot == '2Year':

    colors = ['red','greenyellow', 'midnightblue','forestgreen','orange','
              'violet','aquamarine','deepskyblue','purple','grey',
              'lightcoral' ,'bisque']
    for season, color in zip(unique_twoyearlabels,colors):
        indicesToKeep = finalDf['yearsplit'] == season
        ax.scatter(finalDf.loc[indicesToKeep, 'principal_component_1']
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
    ax.legend(unique_twoyearlabels,loc = 'lower_right',bbox_to_anchor=(1.2

elif type_of_plot == 'Rank':

    colors = ['black','red', 'midnightblue','forestgreen','orange','greeny
              'violet','aquamarine','deepskyblue','purple','grey',
              'lightcoral' ,'bisque']
    for rank, color in zip(unique_ranks,colors):
        indicesToKeep = finalDf['Rank'] == rank
        ax.scatter(finalDf.loc[indicesToKeep, 'principal_component_1']
```

```
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
    ax.legend(unique_ranks,loc = 'lower right',bbox_to_anchor=(1.25, 0.5))


elif type_of_plot == 'KM':
    colors = ['black','red','greenyellow',
              'deepskyblue', 'midnightblue','violet']
    kms = finalDf.KMeansLabel.unique()
    for km, color in zip(kms,colors):
        indicesToKeep = finalDf['KMeansLabel'] == km
        ax.scatter(finalDf.loc[indicesToKeep, 'principal component 1']
                , finalDf.loc[indicesToKeep, pc]
                , c = color
                , s = 50)
#     ax.legend(kms,loc = 'lower right',bbox_to_anchor=(1.25, 0.5))




ax.axhline(y=0, color='k')
ax.axvline(x=0, color='k')


#Plot Variance graph

pca_values=pca.components_

  #Create broken lines
minylim = min(pca_values[1,:])-0.25
maxylim = max(pca_values[1,:])+0.25
minxlim = min(pca_values[0,:])-0.25
maxxlim = max(pca_values[0,:])+0.25

ax1.set_xlim([minxlim , maxxlim ])
ax1.set_ylim([minylim , maxylim ])
ax1.axhline(y=0, color='k')
ax1.axvline(x=0, color='k')
```

```
colors = ['r', 'b', 'k','y', 'g','c','m']
if len(pca_values[0]) > 6:
        colors=colors*(int(len(pca_values[0])/6)+1)


add_string=""
for i in range(len(pca_values[0])):
        xi=pca_values[0][i]
        yi=pca_values[1][i]
        plt.arrow(0,0,
              dx=xi, dy=yi,
              head_width=0.03, head_length=0.03,
              color=colors[i], length_includes_head=True)
        add_string=f" ({round(xi,2)} {round(yi,2)})"
        plt.text(pca_values[0, i],
             pca_values[1, i] ,
             s=top_data.columns[i]  )


#Assign labels to  plot

labels = finalDf

if label == 'Team':

    annotations=labels['Team'].values
    xi=labels[ 'principal component 1'].values
    yi=labels[ pc].values

    for i in range(len(labels)):
        ax.text(xi[i],yi[i],s=annotations[i])

elif label == 'Season':

    annotations=labels['Season'].values
    xi=labels[ 'principal component 1'].values
    yi=labels[ pc].values

    for i in range(len(labels)):
```

```
        ax.text(xi[i],yi[i],s=annotations[i])

elif label == 'TS':

    annotations=labels['Key'].values
    xi=labels[ 'principal component 1'].values
    yi=labels[ pc].values

    for i in range(len(labels)):
        ax.text(xi[i],yi[i],s=annotations[i])

elif label == 'ST':

    labels = labels[labels['Team'].isin(selective_teams)]

    annotations=labels['Team'].values
    xi=labels[ 'principal component 1'].values
    yi=labels[ pc].values

    for i in range(len(labels)):
        ax.text(xi[i],yi[i],s=annotations[i])




ax.grid()
ax1.grid()
with pd.option_context('display.max_rows', None, 'display.max_columns', No
    print(finalDf)
print(pca.explained_variance_ratio_)
```