University of Essex
Department of Mathematical Sciences

MA981: DISSERTATION

Performance Analysis of Chelsea F.C: 2021/22-2023/24

Lenore Daniel Dsouza
2309842

Supervisor: Dr Hirbod Assa & Dr Stella Hadjiantoni

September 18, 2024
Colchester

Contents

List of Figures

List of Tables

CHAPTER 1

Abstract

Performance analysis in football plays an integral role in a team's success or failure. Chelsea Football Club has been one of the top teams in the English Premier League, particularly over the past two decades. This study focuses on Chelsea's performance during the 2021/22, 2022/23, and 2023/24 seasons, a period marked by significant managerial changes. Each season saw a different manager at the helm. The reason for selecting these three seasons lies in the variability of team performance under each manager, with distinct tactical approaches and managerial philosophies influencing the results. The aim of this study is to analyze key performance metrics such as offensive and defensive actions to determine which manager had the most successful impact on Chelsea's overall performance. By employing comparative analysis and predictive modeling, this research seeks to uncover patterns that contributed to Chelsea's successes and failures under each manager, ultimately identifying the best season out of the three seasons.

A season-wise analysis was conducted for Chelsea FC's 2021/22, 2022/23, and 2023/24 seasons using visualization techniques and descriptive statistics. Comparative analysis was carried out using additional visualizations, MANOVA, correlation analysis, and clustering. The analysis identified the 2021/22 season as Chelsea's best, characterized by strong offensive and defensive metrics, including fewer goals conceded and more consistent match performances.

Predictive modeling for goals scored in the 2023/24 season was performed using Linear Regression, Decision Tree, Random Forest, and XGBoost models. Linear Regression emerged

as the best model, predicting a total of 79.12 goals for the season, closely matching the actual figure of 77 goals, with a low Mean Squared Error (MSE) of 1.077.

For predicting match outcomes (win, loss, or draw), models including Multinomial Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes were used. Multinomial Logistic Regression performed the best with an accuracy of 77.68%, outperforming other models in precision and recall, making it the most effective model for predicting match outcomes.

1.1 Introduction

Soccer, referred to as football, is arguably the most popular sport in the world.It is played and watched passionately by many nations, attracting over 2 billion spectators each year[1].Every nation have their own national leagues like Serie-A in Italy, La Liga in Spain etc. The English Premier League(EPL) which is played in England is one of the most watched and entertaining leagues in the world, consisting of 20 teams playing 38 games each every year, competing for the league title[2].

Chelsea Football Club has been one of the most prestigious and successful clubs in England and in the world, winning pretty much every trophy possible including 5 EPL titles[3]. Despite being successful, Chelsea, just like many other elite clubs goes through a rough patch in maintaining a high level of performance every now and then. To have success and keep performing at a high level consistently, it is very important to analyse and know about a teams performance.

Performance analysis(PA) has become an essential part in football and various other sports. It enables the coaching staff to have an effective strategy and therefore enhance overall performance of the team[4]. Although, PA has been used by many different clubs, a vast majority of football clubs, including some of the elite teams in the world still don't give much importance to analysing the performance and giving feedback to the players.[5]

The objective of this dissertation is to analyse the performance of Chelsea F.C over the three seasons: 2021/22, 2022/23 and 2023/24 by examining some of the key performance metrics like goals scored, possession, defensive actions, etc. The analysis aims to identify some of the key trends and factors leading to the team's success and failures.

1.2 Literature review

Performance analysis in football has grown a lot in recent years, with more focus on using data to evaluate how teams perform. This data-driven approach has changed the way football clubs make decisions and adjust their strategies[6]. However, even though there is plenty of research on general football performance, there is still a noticeable lack of studies that look closely at how individual clubs, like Chelsea FC, perform over several seasons. Most research focuses on broad performance indicators like possession, passing accuracy, and defense, but few studies explore how these factors play out over time within one club. This gap means that we often miss the details of how a club's strategies and player contributions evolve from season to season. Studying Chelsea FC over multiple seasons could provide valuable insights into what makes the team successful or what challenges they face over time. This kind of research could help the club improve and also add valuable knowledge to the field of sports analytics.

In the past few years, performance analysis has become more common in football. Early analysis mainly used basic statistics like goals scored, shots taken, and possession. But with new technology and better ways to collect data, performance analysis now includes more detailed metrics like Expected Goals (xG), Expected Assists (xA), and pass completion rates. These new metrics give a better understanding of how well a team is doing, beyond just the basic numbers.

Mackenzie et al. (2012) explored the role of performance analysis in football, emphasizing the use of objective performance indicators such as passing accuracy, possession, and spatial organization to enhance tactical awareness and improve player performance. Their review also pointed out the limitations of these metrics, suggesting that a more contextual understanding of match dynamics is necessary to fully utilize the data for player development and match strategies[7]. One of the biggest developments in football analysis has been the introduction of Expected Goals (xG). Lucey et al. (2014) argued that xG provides a more accurate picture of a team's attacking abilities by estimating how likely a shot is to result in a goal, based on factors like where the shot was taken from and how the defenders were positioned[8]. This metric has become a key part of modern football analysis, helping teams understand their attacking efficiency and make better tactical decisions.

Team performance metrics are often studied to understand what makes a team successful or not. For example, Hughes et al. (2002) found that certain key performance indicators (KPIs), like possession, pass completion rate, and strong defense, are closely linked to winning in football[9]. These KPIs are now commonly used to evaluate a team's performance over a season. For Chelsea FC, looking at these performance metrics over multiple seasons, from 2021/22 to 2023/24, can provide useful insights into how consistent and effective their tactics have been. For example, possession is an important metric for understanding Chelsea's strategy of controlling the game. Research by Jones et al. (2004) and Penos et al. (2010) showed that teams with higher possession percentages tend to be more successful[10][11]. Even though there is a lot of research on these general team performance metrics, there is still not enough research that specifically focuses on Chelsea FC's unique strategies, which shows a gap in the literature.

Another significant development in football performance analysis is the use of advanced data tools and technologies like tracking systems, machine learning, and AI-driven insights. These tools have changed how data is gathered, analyzed, and used for both match preparation and in-game strategies. For example, wearable devices and GPS tracking allow teams to monitor player movements, work rates, and physical conditions in real-time, helping coaches make precise adjustments during games and training. As Sarmento et al. (2018) pointed out, these technologies give coaches and analysts detailed insights into player performance, making it easier to optimize tactics and training[12]. For Chelsea FC, using these technologies could be key to staying competitive over a long season, especially in managing player fitness and adjusting tactics. Combining advanced analytics with traditional performance metrics could provide a more complete view of the team's performance, addressing gaps in existing research on long-term club analysis.

Predictive modeling has become an important tool in sports analytics, helping teams forecast future performance based on past data. Various machine learning models, such as linear regression, decision trees, and neural networks, have been used to predict things like match results and player performance. Constantinou and Fenton (2017) emphasize the effectiveness of Bayesian networks in predicting the evolving performance of football teams across a season, highlighting how integrating expert knowledge with data-driven approaches can improve long-term forecasts[13]. Koopman et al. (2012) introduced a dynamic bivariate Poisson model that accounts for the time-varying nature of teams' attack and defense strengths. Their model also incorporates the dependence between home and away goals, acknowledging that the progression of a match can influence team behavior and outcomes[14]. For Chelsea FC, predictive modeling can be used to predict how many goals the team might score in the upcoming season and to identify the key metrics that have the biggest impact on these predictions. For instance, Markopoulou et al. (2024) applied diverse machine learning models, including XGBoost and random forest, to accurately predict the likelihood of goal-scoring in elite football leagues, demonstrating the effectiveness of data-driven approaches for forecasting player performance based on historical data[15]. Igiri et al. (2014) developed an improved football match result prediction system within the framework of Knowledge Discovery in Databases (KDD), utilizing Artificial Neural Networks (ANN) and Logistic Regression (LR) techniques to incorporate relevant features affecting match outcomes[16].

While these advanced models hold significant potential for analyzing football data, particularly for predicting trends and future outcomes, their application to Chelsea FC's performance has not been extensively researched. This opens up opportunities for future studies to explore these models in the context of long-term team analysis. However, for this research, models like Multinomial Logistic Regression,random forest, SVM etc, were chosen due to their interpretability and effectiveness in handling the multi-class classification problem of match outcomes, given the available dataset.

CHAPTER 2

Data and Methodology

2.1 Datasets

The datasets used to analyze Chelsea FC's performance over the three consecutive English Premier League seasons—2021/22, 2022/23, and 2023/24—come from Fbref[1].[2].[3], a well-known source for football statistics. These datasets include match-level data, giving a detailed look at how the team performed across these seasons. The sample size consists of 114 matches (38 matches per season), which provides a robust dataset for statistical analysis. This analysis helps to understand Chelsea FC's overall performance, identify trends, and make predictions about future outcomes.

The match statistics cover key performance indicators (KPIs) that are crucial for evaluating Chelsea's performance in each game. These metrics include basic offensive and defensive stats like Goals For (GF) and Goals Against (GA), which show how many goals Chelsea scored and conceded. The datasets also include advanced metrics like Expected Goals (xG) and Expected Goals Against (xGA), which estimate the quality of goal-scoring chances and the challenges faced by the defense. The xG metric is especially important as it provides deeper insights into the team's attacking efficiency beyond just the goals scored.

The datasets also include possession stats like Possession Percentage (Poss), showing how much of the game Chelsea controlled, and passing stats such as Passes Completed (Cmp)

[1] 2021/22 season stats

[2] 2022/23 season stats

[3] 2023/24 season stats

and Pass Completion Percentage (Cmp%). These stats help to understand Chelsea's strategy, particularly their focus on controlling the ball and passing accurately. Defensive stats like Tackles (Tkl), Tackles Won (TklW), and Interceptions (Int) reveal the team's ability to disrupt the opponents' play and maintain defensive discipline. In-depth descriptions of the match statistics features used in this analysis can be found in the Appendix, Table A.1 and Table A.2.

2.2 Data Pre-processing

The data pre-processing process involved several critical steps to ensure that the datasets for the 2021/22, 2022/23, and 2023/24 seasons were clean, consistent, and ready for analysis. Initially, the datasets were loaded, and basic information about the data was displayed, including the structure, data types, and the presence of missing values. The datasets contained 112 columns, representing various match statistics, with 38 rows corresponding to each match in the season.

The pre-processing step involved addressing the missing values in the datasets for each of the three seasons. For the 2021/22 season, there were initially 8 missing values, while the 2022/23 season had 1 missing value, and the 2023/24 season had 4 missing values. These missing entries were handled using mean/mode imputation. To handle these, numeric columns with missing values were filled using the column mean, ensuring that the data remained statistically accurate. For non-numeric columns, the mode was used to replace missing values, preserving the integrity of categorical data. After filling the missing values, the datasets were rechecked to confirm that no missing data remained.

Next, the datasets were validated to ensure consistency and accuracy. This validation process included checking data types, identifying and counting any duplicate rows, and generating summary statistics for numerical columns to identify any anomalies. Domain-specific constraints were also checked, such as ensuring that the 'Goals' column contained no negative values, which would be invalid in the context of football matches. Additionally, the 'Date' column was validated to ensure the dates fell within a reasonable range, and, if relevant, the number of unique players was checked to confirm the integrity of player-related data.

This pre-processing ensured that the datasets were accurate, consistent, and free from errors, providing a reliable foundation for subsequent analysis.

2.3 Analysis

This analysis of Chelsea FC's performance over the 2021/22, 2022/23, and 2023/24 seasons employed a variety of advanced statistical and machine learning methods. The methodology involved gathering match-level data, performing statistical summaries, visualizing team performances, conducting multivariate analyses, and building predictive models to forecast goals and match outcomes.

2.3.1 Summary Statistics

Key performance indicators were summarized to give an overview of Chelsea's offensive and defensive strengths and weaknesses across the three seasons. This was followed by the creation of visualizations, such as pie charts, scatter plots, and violin plots, to assess relationships between metrics like possession, pass completion rates, and match outcomes. For these visualizations, the libraries/packages used included Matplotlib for creating static, animated, and interactive plots, Seaborn for statistical data visualization, and Pandas for data manipulation and analysis. These packages were integral in generating clear visual insights from the match data.

2.3.2 Comparative Analysis

Correlation Analysis

To explore the relationships between key performance metrics and match outcomes, a correlation analysis was conducted across three Chelsea FC seasons (2021/22, 2022/23, and 2023/24). The metrics included Possession (Poss), Expected Goals (xG), Goals Scored (GF), and Wins (Win). The purpose of this analysis was to assess how closely these variables are related and to determine which metrics are the most predictive of match outcomes.

A correlation heatmap was generated using the seaborn library in Python, which visualized the strength and direction of relationships between the selected variables. The heatmap included Pearson correlation coefficients, ranging from -1 (a perfect negative correlation) to 1 (a perfect positive correlation), with 0 indicating no linear relationship.

The correlation analysis primarily focused on the relationships between Goals Scored (GF) and Wins, as well as Expected Goals (xG) and Wins. These correlations were calculated to examine whether metrics like xG, which estimate the quality of chances, are as strongly linked to winning as the actual number of goals scored. Additionally, the correlation between Possession (Poss) and Wins was explored to determine the effectiveness of ball control in securing victories.

The scipy.stats package was used to compute the Pearson correlation coefficients for each metric pair, and the results were visualized in the form of a heatmap for ease of interpretation. This approach provided insights into how performance metrics impacted Chelsea FC's match outcomes and their overall importance in football analysis.

MANOVA

MANOVA (Multivariate Analysis of Variance) was employed to investigate whether performance metrics such as Goals For (GF), Goals Against (GA), and Expected Goals (xG) significantly differed across the three Chelsea FC seasons. Given that MANOVA can simultaneously analyze multiple dependent variables, it was well-suited to evaluate Chelsea's offensive and defensive performances over time. Due to violations of multivariate normality and homogeneity of variance-covariance matrices, Pillai's Trace was used as the primary test statistic, as it is more robust to assumption violations. Additional tests, including Wilks' Lambda, Roy's Largest Root, and Hotelling's Trace, were calculated to ensure the robustness of the results. Q-Q plots were generated to assess the normality of GF, GA, and xG. These plots were created using Python libraries such as matplotlib and scipy.stats. Scatter plots and correlation matrix analysis were conducted to evaluate linear relationships between the variables, and Variance Inflation Factor (VIF) was calculated to check for multicollinearity among the performance metrics. Moreover, Bartlett's test identified violations of homogeneity, further supporting the use of Pillai's Trace.

K-Means

K-Means clustering was utilized to group match data from Chelsea FC's 2021/22, 2022/23, and 2023/24 seasons based on key performance metrics such as Expected Goals (xG), Possession, Tackles, Shots, and Interceptions. The K-Means algorithm partitions the data into K distinct clusters, with K=3 selected to reflect the three seasons analyzed. This approach helps identify patterns in team performance across seasons, highlighting similarities and differences in playing style or performance consistency. The clustering process began by normalizing the selected metrics using StandardScaler to ensure that each feature contributed equally to the clustering process. Following normalization, the K-Means algorithm was applied to group the matches, with the objective of minimizing the sum of squared distances between each data point and its corresponding cluster centroid.

Additionally, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the data, allowing for a 2D visual representation of the clusters. This visual representation aids in interpreting the results by showing how the matches from different seasons were grouped into distinct clusters based on performance metrics.

Predictive Modelling

Two predictive modeling tasks were carried out. First, Linear Regression, Decision Tree, Random Forest, and XGBoost models were developed to predict the number of goals Chelsea would score in 2023/24. Metrics such as xG, possession, and SoT were selected as key features. Cross-validation (10-fold) was used to evaluate model performance, and the best-performing model, Linear Regression, was identified based on Mean Squared Error (MSE) and R-squared values.

For match outcome prediction, Support Vector Machines (SVM), Multinomial Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes models were used. Random Forest was applied for feature selection, identifying key predictors like goals per shot (G/Sh) and expected goals (xG). Multinomial Logistic Regression outperformed other models in terms of accuracy and F1-score during cross-validation.

CHAPTER 3

2021/22 Season

Chelsea's 2021/22 season starts with high hopes and expectation, after winning the highly prestigious UEFA Champions League trophy in the previous season(2020/21) under the leadership of Thomas Tuchel. With Tuchel continuing as manager, the team aimed to build on their European success and push for domestic dominance in the Premier League. This chapter provides an in-depth analysis of Chelsea's team performances throughout the 2021/22 Premier League campaign, leveraging detailed match statistics. Key performance indicators, tactical setups, and match-by-match assessments will be examined to evaluate how Tuchel's side performed throughout the season.

3.1 Summary Statistics

The key metrics used to analyze Chelsea FC's performance in the 2021-22 season provide a clear picture of the team's strengths and weaknesses in table 3.1. Offensively, Chelsea averaged 2 goals per match, with an expected goals (xG) value of 1.66, showing they created quality chances. They took about 15 shots per game, with a third of these on target, reflecting solid shooting accuracy. Chelsea also dominated possession with 61.76% on average, and maintained a high pass completion rate of 84.48%, highlighting their focus on controlling the game.

Table 3.1: Summary Statistics for Match Stats 2021-22

Metric	Count	Mean	Std Dev
GF	38	2.00	1.51
GA	38	0.87	1.04
xG	38	1.66	0.98
xGA	38	0.87	0.61
Poss	38	61.76	9.91
Sh	38	15.34	6.16
SoT	38	5.26	3.04
SoT%	38	33.52	13.35
Cmp	38	557.63	125.05
Att	38	654.95	121.15
Cmp%	38	84.48	4.51
Tkl	38	16.34	5.55
TklW	38	8.66	2.06
Sh_bl	38	2.53	2.23
Int	38	9.13	4.31
Err	38	0.39	0.68

Defensively, Chelsea conceded less than 1 goal per match on average, with their defense performing well in line with expected goals against (xGA). They were strong in tackling, winning over half of their attempts, and frequently blocked shots and intercepted passes to disrupt the opposition. The team also showed discipline, with very few errors leading to goals. Overall, the summary statistics for the season indicate that Chelsea was effective both in attack and defense, with strong possession and passing strategies that helped them control matches and minimize mistakes.

3.1.1 Team Performance Visualisations

As shown in figure 3.1, the pie chart depicts the match results distribution for Chelsea FC in the 2021/22 season illustrates a commendable performance. With a win rate of 55.3%, Chelsea demonstrated their dominance on the field, consistently securing victories in more than half of their matches. The draw rate of 28.9% indicates that while the team often managed to avoid defeat, there were numerous instances where matches ended in stalemates, reflecting opportunities for improvement in converting draws into wins. The loss rate of 15.8% highlights the team's resilience, maintaining a relatively low number of defeats throughout the season.

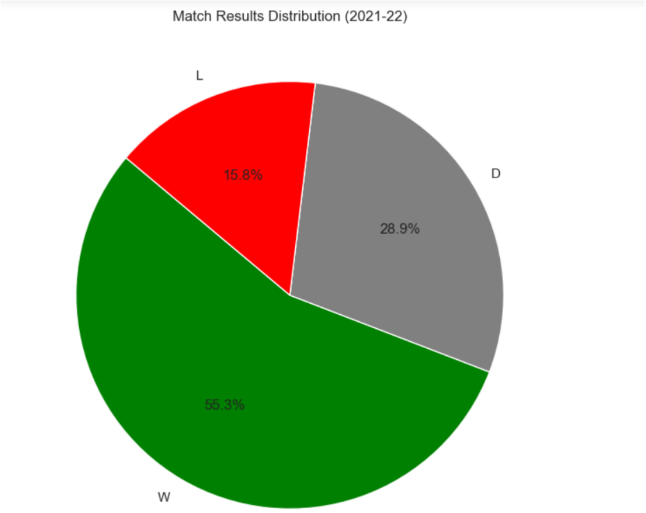

Figure 3.1: Pie Chart of Match Results

The violin plot in the figure 3.2 for the 2021/22 season shows the distribution of Expected Goals (xG), Goals For (GF), and Goals Against (GA) for Chelsea FC. The xG distribution is relatively balanced, with most values around 2, indicating that Chelsea consistently created quality chances in most matches. GF has a wider range, showing a few high-scoring matches where Chelsea scored over 6 goals, but the majority of games saw 1 to 3 goals scored, reflecting some inconsistency in converting chances. On the defensive side, GA has a more constrained distribution, with most games resulting in less than 2 goals conceded, highlighting a solid defense throughout the season.

The comparison between xG and GF shows that Chelsea's actual scoring was generally aligned with the quality of chances they created, while their defense remained strong, conceding fewer goals on average. This indicates a well-balanced performance in both attack and defense during the 2021/22 season.

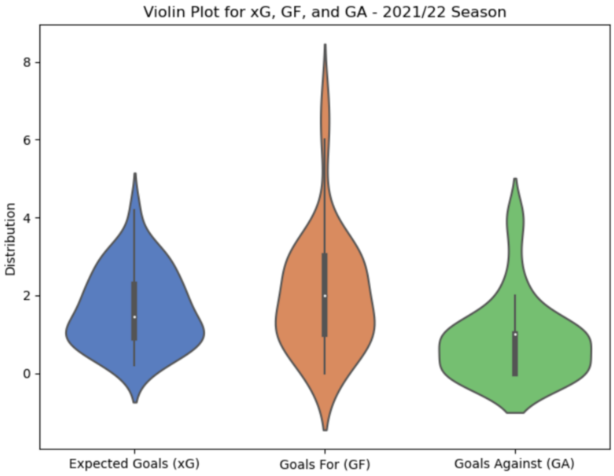

Figure 3.2: Violin Plot for xG, GF, and GA - 2021/22 Season

The scatter plot(figure 3.3) for Chelsea FC's 2021/22 season shows how possession and pass completion rates relate to match outcomes. When Chelsea had high possession (over 60%) and high pass completion (above 85%), they usually won the game, as shown by the blue circles in the top-right area. This means that when Chelsea controlled the ball well and passed accurately, they were more likely to secure a victory. Draws, represented by orange crosses, mostly happened when Chelsea had decent possession (between 50% and 70%) but didn't reach the highest pass accuracy. These games were competitive, but Chelsea couldn't turn their control into wins. Losses, marked by green squares, occurred when both possession

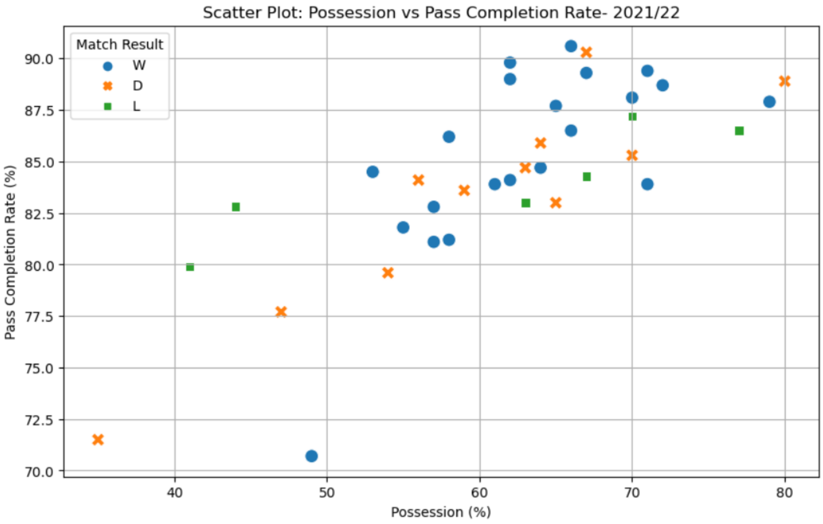

Figure 3.3: Possession vs Pass Completion Rate vs Wins

and pass completion were lower. This indicates that when Chelsea struggled to keep the ball or pass effectively, they were more likely to lose.In short, good possession and passing accuracy were key to Chelsea's success in 2021/22.

The stacked bar chart in Figure 3.4 presents the number of tackles and interceptions made by Chelsea FC in each match during the 2021/22 season. This chart highlights Chelsea's consistent defensive efforts, with regular high counts of tackles and interceptions. The number of tackles per match varies, with some matches seeing over 20 tackles, indicating periods of intense defensive efforts. Interceptions, while generally lower than tackles, remain crucial in Chelsea's defensive strategy, reflecting their ability to read the game and disrupt the opposition's passing. Matches such as numbers 2 and 37, with high counts of both tackles and interceptions, showcase Chelsea's robust and disciplined defensive performance throughout the season.

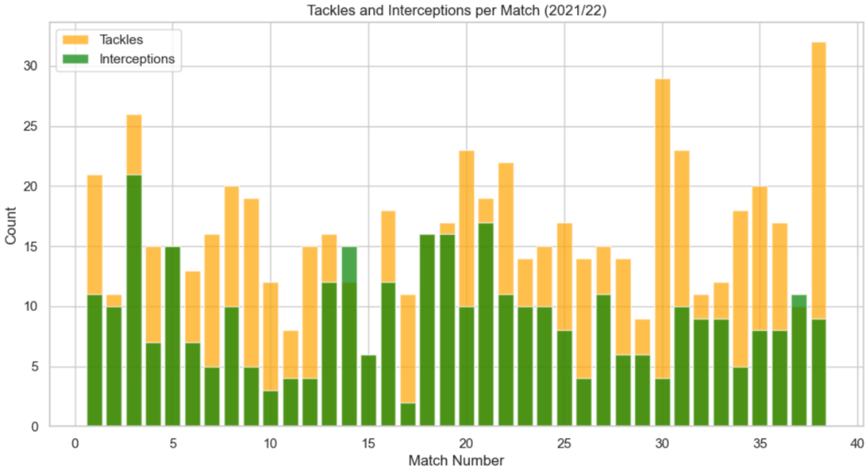

Figure 3.4: Tackles and Interceptions per Match (2021/22)

3.2 Tactical Analysis

The formation analysis as shown in table 3.2, highlights Chelsea FC's performance across four key formations during the season. The 3-4-3 formation appears to be the most commonly used and most successful, with a 54.8% win rate. Teams using this formation scored an average of 2.23 goals per match while conceding only 0.87 goals, indicating a solid balance between offense and defense. The expected goals (xG) of 1.85 further support the idea that this formation allowed Chelsea to consistently create quality scoring opportunities. In contrast, the 3-5-2 formation, used in only 2 games, had a 50% win rate, but the average goals scored and conceded were both just 0.5, suggesting a more defensive approach. The xG of 0.25 shows fewer chances created.

Formation	Win Rate	Avg Goals Scored (GF)	Avg Goals Conceded (GA)	Avg Expected Goals (xG)	Games Played
3-4-3	0.55	2.23	0.87	1.85	31
3-5-2	0.50	0.50	0.50	0.25	2
4-2-2-2	0.00	1.00	1.00	0.70	1
4-3-3	0.75	1.25	1.00	1.13	4

Table 3.2: Performance of Chelsea FC with Different Formations

The 4-3-3 formation saw a high 75% win rate but only produced an average of 1.25 goals per match, conceding 1.00 goals. This suggests that while successful, the team was still vulnerable defensively. Lastly, the 4-2-2-2 formation was used once, resulting in no wins, with 1 goal scored and 1 conceded, with an xG of 0.7, which signals less offensive efficiency.

3.3 Conclusion

The 2021/22 season was a strong campaign for Chelsea FC under Thomas Tuchel, with the team exhibiting solid performances both offensively and defensively. Averaging 2 goals per match and maintaining a high possession rate, Chelsea demonstrated consistent control of games, especially when using the 3-4-3 formation. Defensively, they conceded less than one goal per match, showcasing their tactical discipline. However, challenges remained in converting draws into wins, and the team faced tougher competition from top-tier opponents. Overall, Chelsea's tactical balance and disciplined defense helped them maintain a competitive edge.

Chapter 4

2022/23 Season

Chelsea's 2022/23 season began with hopes of building on their 3rd-place finish in the previous campaign. Graham Potter took charge mid-season, replacing Thomas Tuchel, and was tasked with steering the team towards stability and success. The new manager came in with fresh ideas, replacing the former coach to strengthen both the team's attack and defense. In this chapter, we will analyze Chelsea's team and individual player performances throughout the 2022/23 Premier League campaign, using match statistics, focusing on key metrics and the impact of Potter's tactical approach.

4.1 Summary Statistics

In the 2022/23 season, Chelsea's performance as shown in table 4.1 saw a clear decline compared to the previous season, 2021/22.

Offensively, the team struggled much more in 2022/23. They averaged only 1 goal per match, which is half of what they managed in 2021/22, where they scored an average of 2 goals per game. This indicates that Chelsea found it much harder to create and finish scoring opportunities. The expected goals (xG), which measures the quality of chances created, also dropped from 1.66 to 1.31, further showing that they weren't as effective in attack.

Defensively, Chelsea conceded more goals in 2022/23, with an average of 1.24 goals against them per match, compared to just 0.87 in the previous season. This suggests that their defense wasn't as strong, allowing more goals than before. Their possession of the ball also slightly decreased, from 61.76% to 58.66%, which indicates they weren't controlling the games as well as they did in 2021/22. Additionally, their passing accuracy dropped a little, from 84.48% to 83.49%, showing a small decline in their ability to maintain possession through accurate passing.

Table 4.1: Summary Statistics for Match Stats 2022-23

Metric	Count	Mean	Std Dev
GF	38	1.00	0.96
GA	38	1.24	1.08
xG	38	1.31	0.55
Poss	38	58.66	9.25
Sh	38	12.66	4.88
SoT	38	3.97	2.01
SoT%	38	31.98	12.44
Cmp	38	503.87	90.49
Att	38	600.95	90.37
Cmp%	38	83.49	3.70
Tkl	38	19.47	5.71
TklW	38	11.71	3.59
Shots_bl	38	2.84	1.73
Int	38	8.97	4.09
Err	38	0.55	0.80

When it comes to shooting, Chelsea's accuracy took a hit. They had fewer shots on target per game, dropping from 5.26 in 2021/22 to 3.97 in 2022/23. The percentage of their shots that were on target also decreased slightly, from 33.52% to 31.98%. While their tackling improved, with more tackles per game, the number of errors that led to goals increased, which shows us a clear picture of their defensive struggles.

The 2022/23 season was a tougher one for Chelsea, with noticeable declines in both attacking and defensive performance compared to the 2021/22 season. They scored fewer goals, conceded more, and generally struggled to maintain the same level of control and accuracy in their play.

4.1.1 Team Performance Visualisation

The pie chart shown below in figure 4.1 for Chelsea FC's 2022/23 season shows that the team won 28.9% of their games, drew 28.9%, and lost 42.1%. This means Chelsea struggled throughout the season, losing a large portion of their matches. In the previous season, visualised in section 3.1.1 2021/22, Chelsea performed much better, winning 55.3% of their games and losing only 15.8%. The big increase in losses from 15.8% in 2021/22 to 42.1% in 2022/23 shows a clear decline in their performance. Although the draw percentage stayed the same across both seasons, the significant drop in wins and rise in losses in 2022/23 indicates that Chelsea had a much tougher time turning games into victories. This comparison highlights how much more difficult the 2022/23 season was for Chelsea compared to the stronger performance they had in 2021/22.

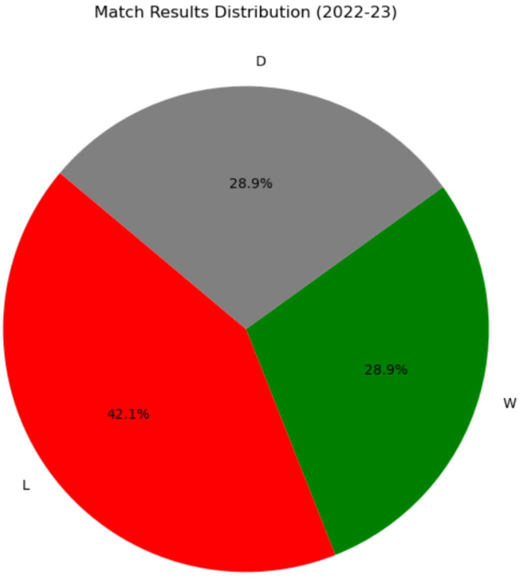

Figure 4.1: Pie Chart of the results

The density plot as shown in the figure 4.2 comparing Chelsea FC's Goals Scored (GF) and Goals Against (GA) during the 2022/23 season shows some important insights. Both curves have a similar shape, suggesting that in most matches, Chelsea scored and conceded around 1 to 2 goals. However, the Goals Scored curve is slightly shifted to the left compared to the Goals Against curve. This means Chelsea often scored fewer goals than they conceded in several matches. The peak of the Goals Against curve is slightly higher than the Goals Scored curve, particularly around the 1-2 goals mark. This indicates that Chelsea conceded 1

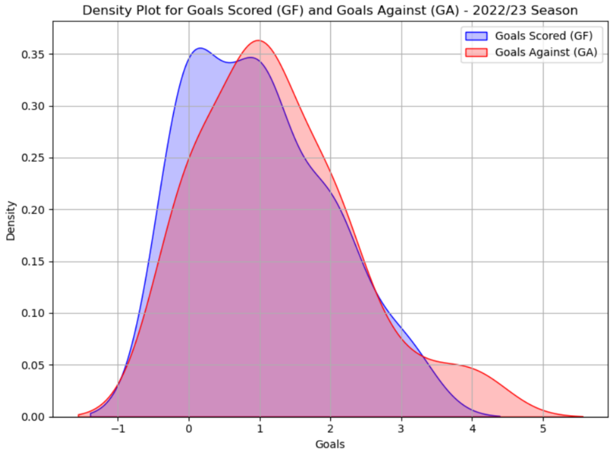

Figure 4.2: Density Plot for GF and GA

to 2 goals more frequently than they scored, which made the season more challenging for them. Additionally, the Goals Against curve extends further to the right, showing that there were instances where Chelsea conceded 4-5 goals in a match, a much rarer occurrence for their goal-scoring.

Comparatively, during the 2021/22 season, Chelsea performed better. They scored more goals and conceded fewer. The shift in these curves between the two seasons highlights the decline in both attacking and defensive performance in 2022/23, leading to tougher results for the team.

The violin plot in the figure 4.3 shows how Chelsea FC's Expected Goals (xG) and Actual Goals (GF) were distributed during the 2022/23 season. The xG distribution is spread out,

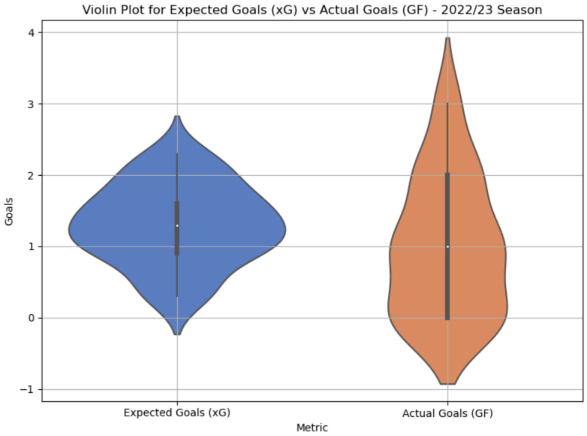

Figure 4.3: Violin Plot for xG and GF

mostly between 0.5 and 2.5 goals, indicating that Chelsea created decent scoring opportunities in many matches. However, the GF distribution is narrower, with most actual goals falling between 0 and 2 goals per match. This suggests that Chelsea often scored fewer goals than expected, underperforming in terms of converting their chances. In comparison to the 2021/22 season, where the actual goals Chelsea scored(GF) would have been more in line with xG, showing better goal conversion, the 2022/23 season highlights a decline. The gap between xG and GF in 2022/23 indicates that Chelsea struggled to finish their chances effectively. This drop in goal-scoring efficiency was a significant factor in the team's challenges during the 2022/23 season.

The scatter plot in figure 4.4 for the 2022/23 season shows how Chelsea FC's possession and pass completion rates relate to their match outcomes.

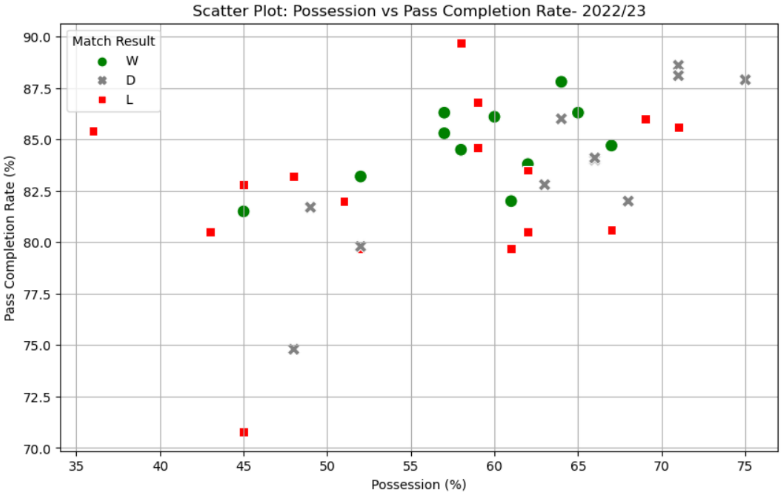

Figure 4.4: Scatter Plot Possession vs Pass Completition Rate vs Wins-2022/23

Unlike the 2021/22 season, where higher possession and passing accuracy often led to wins, the 2022/23 season shows more mixed results. In 2022/23, Chelsea's wins (green circles) are spread out across different levels of possession and pass accuracy, meaning that even when they controlled the ball well, it didn't always lead to victory. Draws (gray crosses) and losses (red squares) also appear in games where Chelsea had decent possession and passing, highlighting their struggles to turn control of the game into positive results.

Compared to the previous season, this scatter plot shows that Chelsea's possession and pass completion did not translate into success as reliably in 2022/23, indicating a drop in performance and consistency. The inconsistency in match outcomes, even with good possession and pass rates, marks a clear challenge for Chelsea during the 2022/23 season.

In the 2022/23 season, Chelsea FC's defensive performance, as measured by tackles and interceptions per match, showed significant variability. As seen in figure 4.5 Tackles remained consistently higher than interceptions, with some matches seeing over 30 tackles, but interceptions fluctuated more than in the previous season, indicating a less consistent defensive effort. This inconsistency in interceptions suggests that Chelsea's defense may have been less structured, relying more on reactive play rather than anticipatory actions.

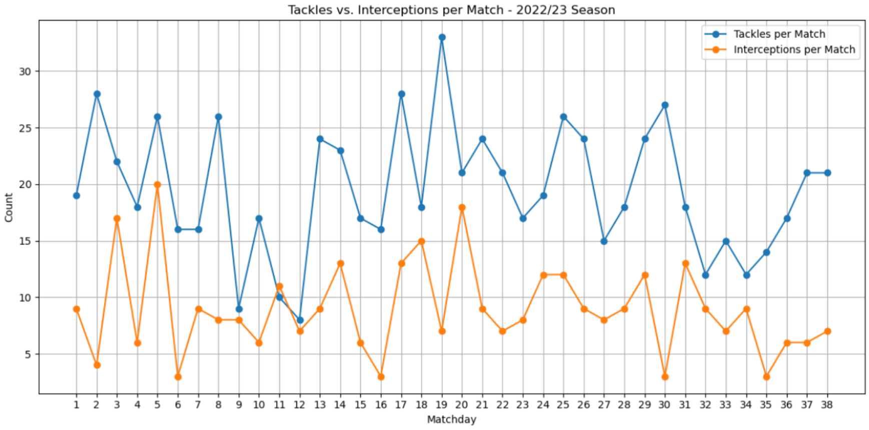

Figure 4.5: Tackles vs Interceptions Per Match 2022/23

Comparatively, in the 2021/22 season, while tackles also outnumbered interceptions, both metrics showed a more consistent pattern across matches. The higher and more stable interception counts in the 2021/22 season reflect a stronger, more coherent defensive strategy, where Chelsea's defense was more effective in maintaining its structure and disrupting the opposition's play. The drop in consistency in the 2022/23 season highlights a potential area of decline in Chelsea's defensive organization compared to the previous season.

4.2 Tactical Analysis

The formation analysis of Chelsea FC for the 2022/23 season as seen in table 4.2, highlights varying levels of effectiveness across different tactical setups. The most frequently used formation, the 3-4-3, had a moderate win rate of 36.36%, with an average of 1 goal scored and 1.18 goals conceded per match. Despite being commonly used, its win rate suggests that it was not particularly effective at securing victories. The 3-5-2 formation, used in 7 games, struggled with a low win rate of 14.29%, scoring just 0.71 goals per game while conceding 1.29. This indicates challenges in both attack and defense, making it one of the less effective formations.

Formation	Win Rate	Avg Goals Scored (GF)	Avg Goals Conceded (GA)	Avg Expected Goals (xG)	Games Played
3-4-3	0.36	1.00	1.18	1.42	11
3-5-2	0.14	0.71	1.29	1.40	7
4-2-3-1	0.22	0.78	1.00	1.10	9
4-3-3	0.25	1.25	1.50	1.29	8
4-4-2	0.67	1.67	1.33	1.33	3

Table 4.2: Performance of Chelsea FC with Different Formations

The 4-2-3-1 and 4-3-3 formations both had low win rates, with 22.22% and 25%, respectively, indicating limited success. In both formations, Chelsea conceded more goals than they scored, further highlighting tactical inefficiencies. The 4-4-2 formation, though used in only 3 matches, had the highest win rate at 66.67%, with an average of 1.67 goals scored and 1.33 conceded. This suggests that it was relatively more balanced and successful when deployed.

4.3 Coclusion

The 2022/23 season was a challenging one for Chelsea FC, marked by a clear decline in both offensive and defensive performance compared to the previous year. Under Graham Potter's leadership, the team struggled to maintain consistency, scoring only 1 goal per match on average while conceding 1.24 goals. The decline in expected goals (xG) and possession further highlighted their inability to dominate games. Despite trying various tactical formations, none proved particularly effective, with the 3-4-3 formation only achieving a moderate 36.36% win rate. The team's defensive performance also lacked stability, with interceptions and tackles showing significant variability across matches. Overall, Chelsea's inability to finish scoring opportunities and maintain defensive structure contributed to their disappointing results, demonstrating the need for stronger tactical coherence.

CHAPTER 5

2023/24 Season

Chelsea entered the 2023/24 season aiming to rebound from a disappointing 12th-place finish the previous year. With new manager Mauricio Pochettino at the helm, the club sought a revitalized approach, combining tactical changes and fresh signings to reestablish themselves as Premier League contenders. This section focuses on analyzing Chelsea's team performance throughout the season using visualization techniques and descriptive statistics. By examining key metrics such as goals scored, possession, and defensive actions, the analysis will assess the effectiveness of Pochettino's tactical strategies and the team's overall improvement compared to the challenges of the previous season.

5.1 Summary Statistics

The summary statistics for Chelsea FC's 2023/24 season as shown in table 5.1, provide a detailed overview of the team's performance across key metrics such as goals scored (GF), goals conceded (GA), expected goals (xG), possession percentage (Poss), and various other offensive and defensive indicators. On average, Chelsea scored 2.03 goals per match, showing strong offensive capability. However, the team also conceded an average of 1.66 goals per game, indicating some defensive vulnerabilities. The expected goals (xG) value of 1.96 per match suggests that Chelsea created high-quality chances consistently throughout the season. This aligns well with their actual goal-scoring average, indicating that the team performed in line with expectations in terms of finishing.

Table 5.1: Summary Statistics for Match Stats 2023-24

Metric	Count	Mean	Std Dev
GF	38	2.03	1.46
GA	38	1.66	1.32
xG	38	1.96	0.92
Poss	38	58.61	12.53
Sh	38	14.13	5.79
SoT	38	5.11	2.72
SoT%	38	38.34	16.50
Cmp	38	527.34	135.08
Att	38	614.84	140.87
Cmp%	38	85.17	3.71
Tkl	38	17.45	6.17
TklW	38	9.89	3.96
Shots_bl	38	4.37	2.76
Int	38	7.84	3.01
Err	38	0.63	1.00

Chelsea had an average possession of 58.61%, suggesting that they maintained control of the ball for a significant portion of each match. Their average number of shots per game was 14.13, with 5.11 of those being on target, leading to a shot on target percentage (SoT%) of 38.34%. This indicates that Chelsea was fairly effective in getting shots on target, though there is room for improvement in converting these opportunities into goals. The team's passing was also strong, with an average pass completion rate (Cmp%) of 85.17%, reflecting solid ball retention and build-up play. Defensively, Chelsea averaged 17.45 tackles per game and won 9.89 of them, showing a decent level of defensive pressure. They also blocked an average of 4.37 shots per game, but the number of interceptions was slightly lower at 7.84, suggesting that while they applied defensive pressure, there may have been moments where they were less proactive in disrupting the opposition's play. Overall, the statistics reveal a well-rounded but slightly imbalanced team, with strong offensive efforts but some defensive weaknesses, particularly in goals conceded. The team's ability to maintain possession and create chances is a positive takeaway, although improving defensive stability could be key to future success.

5.2 Team Performance Visualisation

The pie chart in figure 5.1 for Chelsea FC's 2023/24 season shows a marked improvement in performance compared to the previous season. The team secured wins in 47.4% of their matches, which is a significant increase from the 28.9% win rate in the 2022/23 season. Losses also decreased from 42.1% to 28.9%, suggesting that the team became more competitive and resilient in tight games. The percentage of draws remained similar, with 23.7% in 2023/24 compared to 28.9% in the previous season, indicating that while Chelsea managed to reduce losses, converting more drawn games into wins is still an area for improvement. Overall, the 2023/24 season reflects positive development for the team, with a stronger ability to secure victories compared to the struggles seen in 2022/23. This improvement in outcomes highlights better tactical implementation and possibly a stronger squad depth, which could be key to future success if the trend continues.

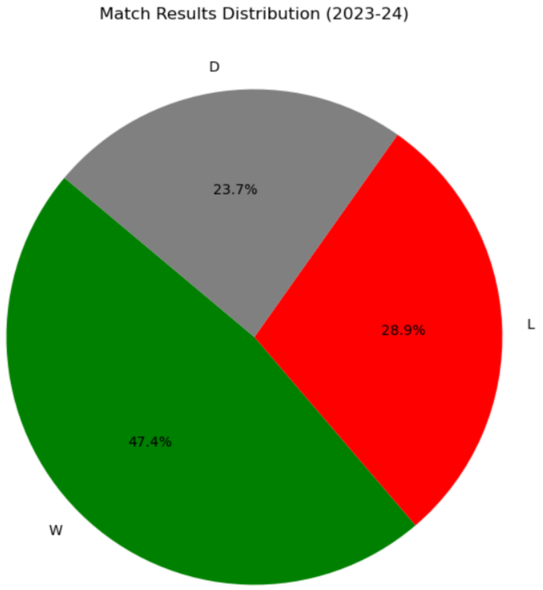

Figure 5.1: Match results 2023-24

The density plot in figure 5.2 for Chelsea FC's 2023/24 season compares Goals Scored (GF) and Goals Against (GA). The blue line represents the distribution of goals scored, while the red line represents goals conceded. Both curves peak around 1 to 2 goals, but the GA curve is slightly higher, showing that Chelsea conceded 1 to 2 goals more frequently than they scored. The GF curve tails off more slowly, indicating that while higher goal counts were rare, they did happen in a few matches. On the other hand, the GA curve shows a quicker drop-off, meaning Chelsea usually conceded fewer than 3 goals, but there were instances where they let in up to 5 goals. When compared to the 2022/23 season, the pattern is similar.

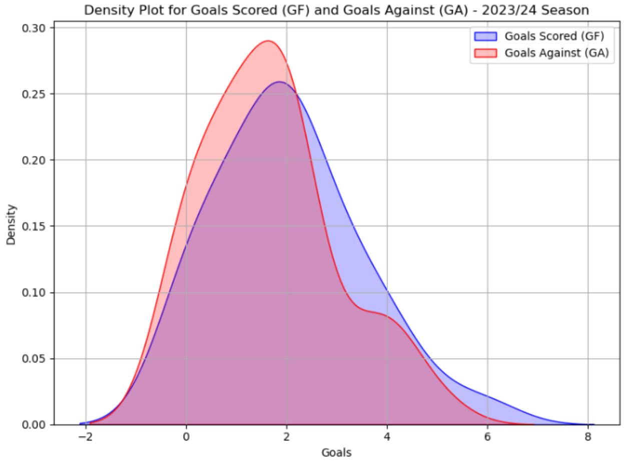

Figure 5.2: Density Plot GF vs GA

Chelsea's GF and GA distributions overlapped in both seasons, but the higher density of goals against in 2023/24 shows a recurring issue with conceding. This season's performance indicates that Chelsea's defense has not improved significantly, but their attacking output has slightly improved.

The box plot in the figure 5.3shows the distribution of Goals For (GF) and Expected Goals (xG) for Chelsea FC in the 2023/24 season. The median GF is around 2, with the interquartile range (IQR) indicating that most of the goals scored were between 1 and 3.

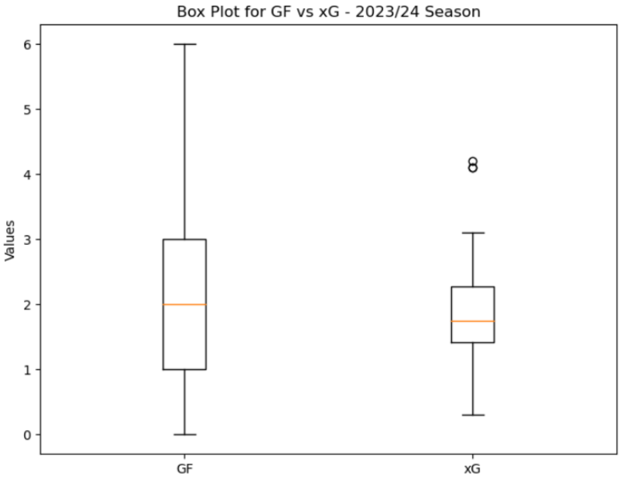

Figure 5.3: Box Plot xG vs GF

There is a wider spread for GF, with a maximum value of 6 goals and a minimum of

0. In contrast, xG, which measures the quality of scoring chances, has a narrower range. The median xG is slightly below 2, and the majority of the values lie between 1.5 and 2.5, with one noticeable outlier at around 4. This suggests that while Chelsea created some high-quality chances (as indicated by the xG), they did not always convert them into goals, shown by the wider variability in GF. When compared to the 2022/23 season, there is a similar pattern, where xG was more tightly clustered, indicating consistent creation of chances, but GF showed more variability, reflecting inconsistency in converting those chances into goals.

The scatter plot for Chelsea FC's 2023/24 season in figure 5.4 illustrates the relationship between possession percentage and pass completion rate, with match outcomes indicated by different markers. Green squares represent wins, red crosses signify losses, and grey circles indicate draws. A notable observation is that Chelsea achieved higher pass completion rates (85% and above) in many of their victories, especially in matches with possession above 60%. However, the team also lost several matches despite maintaining high possession and pass accuracy, demonstrating that possession and passing alone were not sufficient for success.

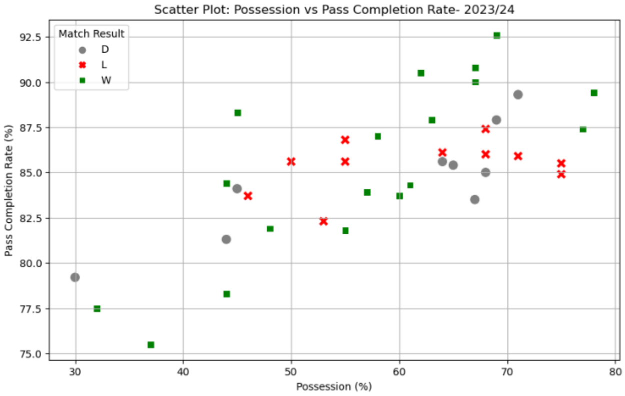

Figure 5.4: Scatter Plot Possession vs Pass Completion Rate

Compared to the 2022/23 season, the pattern remains somewhat consistent. In both seasons, Chelsea achieved better results when they controlled the game through higher possession and passing accuracy, yet there were still losses and draws even with decent ball control. This suggests that while possession and pass completion are key, other factors such as finishing and defensive solidity were also critical areas where Chelsea fell short in both seasons.

The plot in figure 5.5 shows the tackles versus interceptions per match for the 2023/24 season, offering insights into Chelsea's defensive consistency. Tackles per match are consistently higher than interceptions, peaking around matchdays 6, 18, and 28 with over 25 tackles. Interceptions, however, remain more fluctuating, with a peak around 10-12 interceptions per match. This suggests that while Chelsea maintained a strong defensive tackling game, their interceptions were less stable throughout the season. When comparing this to the 2022/23 season, as seen in the previous analysis, the trend was similar in terms of tackles being higher than interceptions, but interceptions in 2022/23 showed even more variability and lower counts overall. This suggests a slight improvement in interceptions for 2023/24, although it still doesn't match the stability and consistency seen in 2021/22, where both tackles and interceptions were more evenly distributed, indicating a more coherent defensive approach that Chelsea has yet to fully regain.

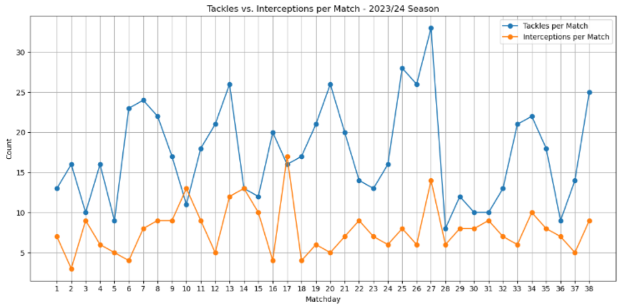

Figure 5.5: Tackles vs Interception- 2023/24

5.3 Tactical Analysis

The analysis of different formations used by Chelsea FC, which is shown in table 5.2, during the season reveals interesting insights. The "4-3-3" formation had the highest win rate (100%) with an average of 3 goals scored and only 0.5 goals conceded, although it was used only in two games, limiting the conclusions we can draw from it. The "4-2-3-1" formation was used the most frequently, with 31 games played and a win rate of 48.39%.

Formation	Win Rate	Avg Goals Scored (GF)	Avg Goals Conceded (GA)	Avg Expected Goals (xG)	Games Played
3-4-3	0.25	1.25	1.25	2.10	4
3-5-2	0.00	2.00	2.00	1.80	1
4-2-3-1	0.48	2.06	1.77	1.96	31
4-3-3	1.00	3.00	0.50	1.80	2

Table 5.2: Performance of Chelsea FC with Different Formations

Chelsea scored an average of 2.06 goals per game while conceding 1.77 goals, suggesting this formation struck a balance between offense and defense, though it wasn't the most defensively solid. The "3-5-2" formation had a disappointing performance, with no wins and 2 goals conceded per game, despite creating 1.8 expected goals. The "3-4-3" formation, while used in only four games, had a lower win rate (25%), showing inconsistencies in attack and defense, with 1.25 goals both scored and conceded.

5.4 Conclusion

In the 2023/24 season, Chelsea showed notable improvements compared to the previous year. The team averaged 2.03 goals per match, a considerable boost from the 1.00 average in 2022/23. This increase in offensive output indicates a stronger attacking performance. Defensively, however, challenges remained, with an average of 1.66 goals conceded per game, showing only a slight improvement from the 1.24 in 2022/23. Despite this, the team managed to secure wins in 47.4% of their matches, a significant rise from the 28.9% win rate in the previous season. The improvement in results can be attributed to more effective finishing and better overall team coordination.

In comparison, Chelsea's 2022/23 season was marred by struggles, both offensively and defensively. The team averaged fewer goals, conceded more, and had a higher loss rate. The 2023/24 season reflects positive steps forward, with increased scoring capabilities and a more competitive overall performance, though defensive vulnerabilities persist and will need further attention to secure long-term success.

CHAPTER 6

Comparative Analysis

6.1 Performance Comparison Across All Three Seasons

The bar chart in figure 6.1 provides a comparison of wins, losses, and draws across three seasons (2021/22, 2022/23, and 2023/24). In the 2021/22 season, the team secured the highest number of wins (over 20) with relatively fewer losses and draws. This shows that 2021/22 was a strong performance year.

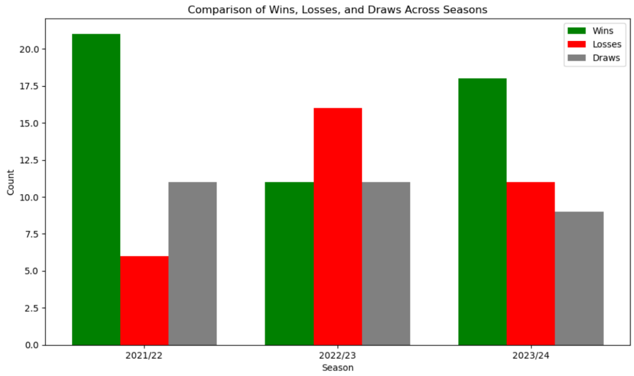

Figure 6.1: Histogram of Match Results Across All Three Seasons

The 2022/23 season displays a sharp contrast, with more losses (about 15) and fewer wins, suggesting a weaker performance. The 2023/24 season shows an improvement compared to 2022/23, with a notable rise in wins and fewer losses, although not as dominant as in 2021/22. This chart highlights the inconsistency in the team's performance, with the 2021/22 season emerging as the best among the three.

This radar chart in 6.2 illustrates Chelsea's defensive performance over the 2021/22, 2022/23, and 2023/24 seasons, focusing on key metrics like Tackles, Interceptions, Errors Leading to Goals, and Goals Against (GA).

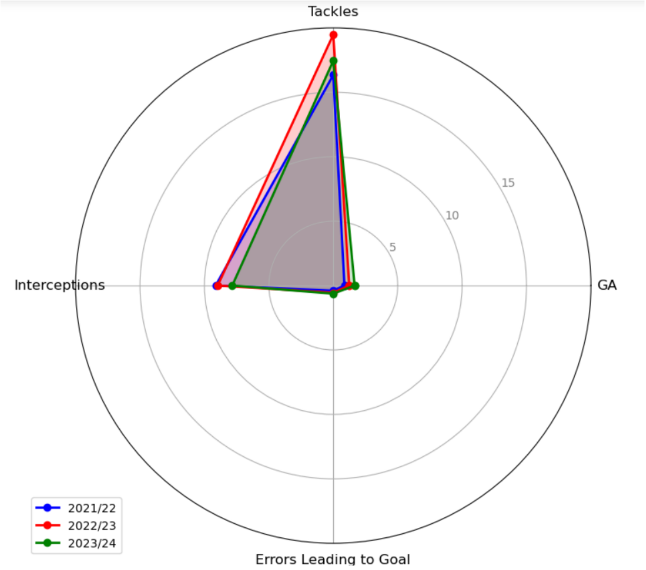

Figure 6.2: Defensive Performance-Radar Chart

The 2022/23 season (red) stands out as the worst, with the highest number of tackles and errors, leading to more goals conceded. In contrast, the 2021/22 season (blue) was the best, showing a balanced defensive approach with fewer errors and goals against, demonstrating overall stability. The 2023/24 season (green) shows a middle ground, with improvements in interceptions but still some defensive challenges.

6.1.1 Correlation Analysis

This correlation heatmap in 6.3 visualizes the relationships between Possession (Poss), Expected Goals (xG), Goals Scored (GF), and Wins (Win) for Chelsea FC across the three seasons (2021/22, 2022/23, and 2023/24). The strongest correlation is between Goals Scored (GF) and Wins (0.65), which is an expected outcome, as scoring more goals naturally increases the likelihood of winning matches. Expected Goals (xG) also shows a moderate correlation with Wins (0.34), suggesting that generating quality goal-scoring opportunities is important, though not as closely tied to winning as actual goals scored. The correlation between xG and GF (0.55) is also notable, indicating that while xG can often predict the number of goals scored, it is not a perfect indicator, especially in Chelsea's inconsistent seasons. Possession (Poss) has

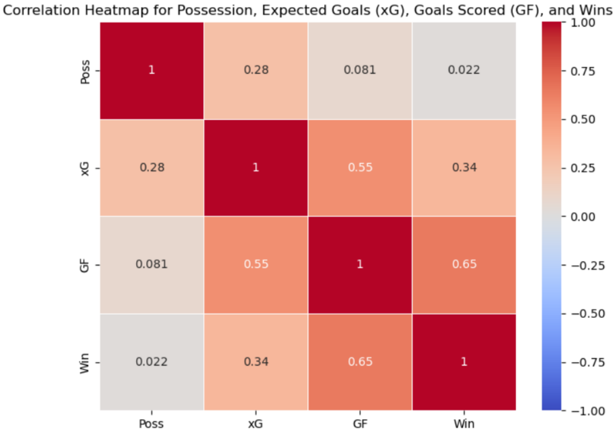

Figure 6.3: Correlation Heatmap

weak correlations with both GF (0.081) and Wins (0.022), suggesting that for Chelsea, simply controlling possession hasn't been a reliable pathway to success.

Across these three seasons, Chelsea's inconsistency in converting possession and xG into actual wins becomes evident. The 2021/22 season, which was more consistent, may have aligned better with these metrics, while the following two seasons saw more variability, explaining the weaker correlations.

6.2 MANOVA (Multivariate Analysis of Variance)

MANOVA is an extension of ANOVA that allows comparison of means across multiple dependent variables at once, rather than one[17]. It evaluates whether multiple continuous dependent variables (such as Goals Scored, Possession, xG) differ across different levels of categorical independent variables (in this case, the seasons). The MANOVA test assesses the overall effect of the independent variable on the dependent variables while accounting for correlations among them[18].

For the MANOVA analysis, the variables used were Goals For (GF), Goals Against (GA), and Expected Goals (xG). These metrics were chosen because they provide a comprehensive view of team performance, covering both offensive (GF) and defensive (GA) aspects, along with the quality of scoring opportunities (xG). The combination of these variables allows for an in-depth analysis of Chelsea FC's performance across multiple seasons, helping to identify statistically significant differences and trends. MANOVA was applied to determine if these metrics varied meaningfully across the 2021/22, 2022/23, and 2023/24 seasons.

In MANOVA, several statistical metrics are used to determine if there are significant differences between the groups (here, the three seasons) across the dependent variables[19]. The primary metrics are:

6.2.1 Wilks' Lambda Test Statistic

Wilks' Lambda is used to compare the mean vectors of p variables across g groups. The matrices are expressed as and relates to chi square distribution[20].

$$B = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T \tag{6.2.1}$$

$$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^T \tag{6.2.2}$$

Where: - B is the matrix of squares between groups, - W is the matrix of squares within groups, - $\bar{x}_i$ is the mean vector for the i-th group, and - n_i is the number of observations in the i-th group.

The Wilks' Lambda statistic is defined as:

$$\Lambda = \frac{\det(W)}{\det(B + W)} \tag{6.2.3}$$

For large samples, the Bartlett approximation is used, where the test statistic is:

$$L = -\left[(N - 1 - \frac{p+g}{2}) \ln \Lambda\right] \tag{6.2.4}$$

This follows a chi-squared distribution with degrees of freedom $p(g-1)$.

6.2.2 Hotelling's Trace Test Statistic

Hotelling's Trace is calculated from the roots of the matrix BW^{-1}, where λ_i are the eigenvalues of BW^{-1}:

$$T = \sum_{i=1}^{s} \lambda_i \tag{6.2.5}$$

Where $s = \min(g-1, p)$. A large value of T indicates a significant difference between the mean vectors.

6.2.3 Pillai's Trace Test Statistic

Pillai's Trace is defined as:

$$V = \sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} \tag{6.2.6}$$

Where λ_i are the eigenvalues of BW^{-1}.

6.2.4 Roy's Largest Root Test Statistic

Roy's Largest Root focuses on the largest eigenvalue λ_{max} of BW^{-1}:

$$\lambda_{\text{max}} = \max(\lambda_1, \lambda_2, \ldots, \lambda_s) \tag{6.2.7}$$

[21]

6.2.5 Assumptions of MANOVA

Multivariate Analysis of Variance (MANOVA) relies on several assumptions to ensure that its results are valid and interpretable. These assumptions include:

Independence of Observation

The assumption of independence requires that each observation(in this case, performance metrics like xG, GF, and GA)in the dataset must be independent of others. This means that the outcome of one match or one set of observations should not affect another. This is satisfied in the analysis of Chelsea FC's performance metrics across three seasons, as each match is treated as an independent event. Since one match result or performance does not affect subsequent matches in a structured football season, this assumption holds true without needing a formal test. The independence of matches across seasons ensures that the outcomes are distinct, allowing for reliable statistical analysis across seasons.

Multivariate Normality

The assumption states that The dependent variables (e.g., Goals For, Goals Against, and Expected Goals) should follow a multivariate normal distribution in each group (season).[22]

The Q-Q plots in figure 6.4 indicate that the assumption of multivariate normality is violated for Goals For (GF) and Goals Against (GA) across all three seasons, as significant deviations from the normal distribution line are observed, particularly at the tails. However, Expected Goals (xG) shows a distribution closer to normal, especially in the 2021/22 and 2022/23 seasons, with slight deviations for 2023/24. Since the assumption of multivariate normality is not satisfied, Pillai's Trace was used for the MANOVA analysis, as it is more robust in cases where assumptions are violated, ensuring more reliable statistical interpretations.

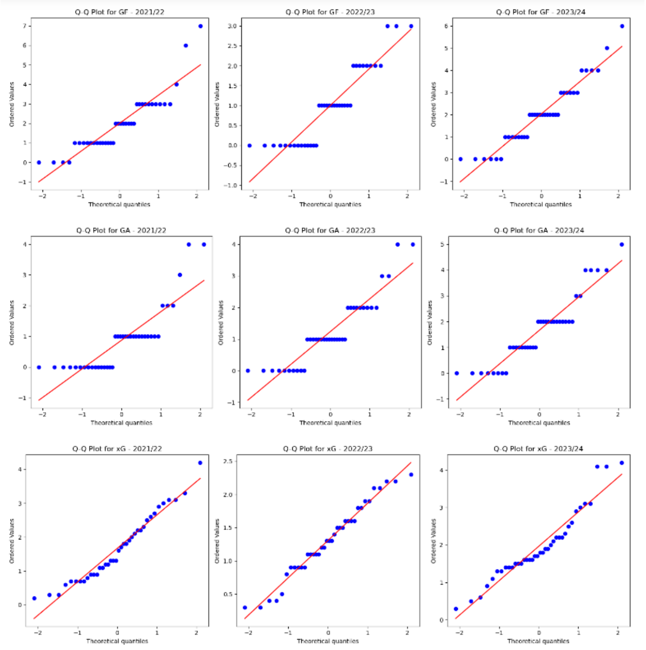

Figure 6.4: Q-Q Plots to test Multivariate Normality

Homogeneity of Variance-Covariance Matrices

The assumption of homogeneity of variance-covariance matrices in MANOVA requires that the variance-covariance matrices of the dependent variables are equal across the groups of the independent variable (in this case, across the different seasons).[22] Bartlett's test results in table 6.1 indicate a violation of the assumption of homogeneity of variances for both xG (p-value = 0.00016) and GF (p-value = 0.00478), suggesting that their variances differ significantly across the three seasons. However, for GA, the p-value is 0.978, meaning the assumption is satisfied, and variances are similar across the seasons.Since the assumption of homogeneity is violated for xG and GF, Pillai's trace is again applied for the MANOVA test interpretation. As a robust statistic, Pillai's trace effectively handles violations of both normality and homogeneity, making it the most suitable choice for this analysis.

Metric	Statistic	p-value
xG	17.46	0.00
GF	10.69	0.00
GA	0.04	0.98

Table 6.1: Bartlett's Test Results for Homogeneity of Variance-Covariance Matrices

Linear Relationships Among Dependent Variables

MANOVA assumes that the dependent variables (such as GF, GA, and xG) have linear relationships with each other within each group[23].

The scatterplot matrix in figure 6.5 and correlation matrix in table 6.2 provide insight into the linear relationships between Goals For (GF), Goals Against (GA), and Expected Goals (xG) across the three seasons. In the correlation matrix, GF and xG exhibit a moderately positive correlation of 0.55, which suggests that as xG increases, GF also tends to increase. GA has weak negative correlations with both GF (-0.15) and xG (-0.14), implying that these relationships are not strongly linear.

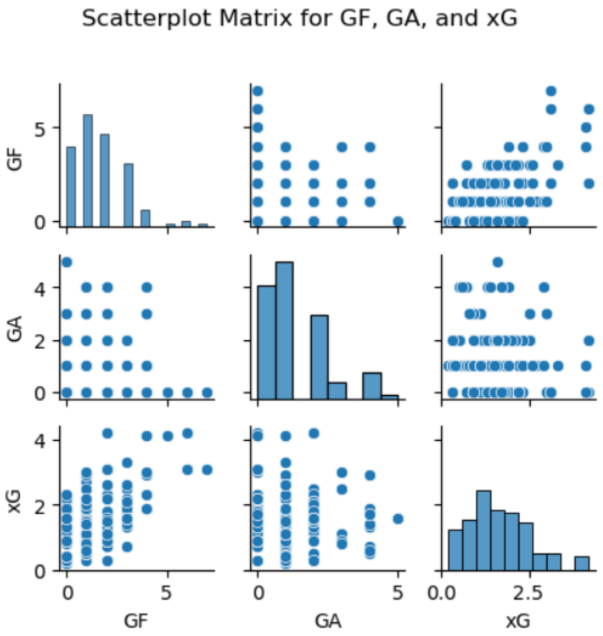

Figure 6.5: Scatter Plot matrix to test linear relationship

The assumption of linear relationships between the dependent variables for MANOVA is partially satisfied for GF and xG due to their moderate correlation. However, the weak

Table 6.2: Correlation Matrix for GF, GA, and xG

	GF	GA	xG
GF	1.00	-0.15	0.55
GA	-0.15	1.00	-0.14
xG	0.55	-0.14	1.00

correlations involving GA may indicate that the relationship is not strictly linear for all variables. Nonetheless, MANOVA can still proceed, given the presence of some linearity between key variables like GF and xG.

No Multi-collinearity

The assumption says that dependent variables should not be highly correlated (i.e., no mult-icollinearity)[23].

Table 6.3: Variance Inflation Factor (VIF) for GF, GA, and xG

Feature	**VIF**
const	6.46
GF	1.45
GA	1.03
xG	1.45

The VIF (Variance Inflation Factor) values for the dependent variables—GF (1.45), GA (1.03), and xG (1.45)—indicate that multicollinearity is not a concern in this dataset, as all values are well below the threshold of 5. This confirms that there is no significant correlation between the variables that would affect the validity of the MANOVA analysis. As seen earlier in the correlation matrix in table 6.2, the variables are not highly correlated, and this is further supported by the VIF results in table 6.3. Therefore, the assumption of no multicollinearity is satisfied in this analysis.

6.2.6 Interpretation of MANOVA Results

The first column in the table 6.4 represents the "Effect," which refers to the factors being tested in the multivariate linear model. The Intercept shows the overall mean or baseline value of the dependent variables (e.g., goals scored, possession) without considering other factors. It indicates the behavior of the response variables when all other factors are set to their baseline. The Season represents the independent variable being analyzed, assessing how the performance metrics (like goals scored and possession) differ across the 2021/22, 2022/23, and 2023/24 seasons. Each test (e.g., Wilks' lambda) determines whether the variations across seasons are statistically significant.

Effect	Test	Value	Num DF	Den DF	F Value	Pr > F
Intercept	Wilks' lambda	0.34	3.00	109.00	71.42	0.0000
Intercept	Pillai's trace	0.66	3.00	109.00	71.42	0.0000
Intercept	Hotelling-Lawley trace	1.97	3.00	109.00	71.42	0.0000
Intercept	Roy's greatest root	1.97	3.00	109.00	71.42	0.0000
Season(21/22,22/23 and 23/24)	Wilks' lambda	0.77	6.00	218.00	5.05	0.0001
Season(21/22,22/23 and 23/24)	Pillai's trace	0.24	6.00	220.00	5.05	0.0001
Season(21/22,22/23 and 23/24)	Hotelling-Lawley trace	0.28	6.00	143.57	5.07	0.0001
Season(21/22,22/23 and 23/24)	Roy's greatest root	0.19	3.00	110.00	7.07	0.0002

Table 6.4: Multivariate Linear Model Results

The results from the Multivariate Analysis of Variance (MANOVA) provide a detailed statistical comparison of Chelsea FC's performance across three seasons (2021/22, 2022/23, and 2023/24). The Wilks' lambda value of 0.7709 for the "Season" effect indicates that there is a significant difference across the seasons in the multivariate context of goals for (GF), goals against (GA), and expected goals (xG). The associated F-value of 5.0468 and p-value of 0.0001 confirm that these differences are statistically significant. Additionally, Pillai's trace (0.2420) and Hotelling-Lawley trace (0.2803) further support these findings, indicating that performance metrics shifted noticeably between the seasons.

The high Roy's greatest root value of 0.1929 and the significant F-value of 7.0742 emphasize that the differences in goals and expected goals were most substantial in certain comparisons, specifically highlighting that the 2021/22 season outperformed the others, with Chelsea scoring more goals and conceding fewer than in the subsequent seasons. This analysis suggests that the 2021/22 season was statistically Chelsea's best in terms of offensive and defensive metrics.

6.3 Comparison Using Clustering

In this clustering analysis, K-Means was applied to examine patterns in the performance metrics of Chelsea FC across three football seasons: 2021/22, 2022/23, and 2023/24. The goal of clustering is to group matches based on similar characteristics without prior knowledge of match outcomes or labels[24]. By using performance metrics such as Expected Goals (xG), Possession, Tackles, Shots on Target, and others, clustering helps uncover hidden patterns in how the team performed over time.

The variables used in clustering are key football performance metrics: Expected Goals (xG), Possession (Poss), Shots (Sh), Shots on Target (SoT), Completed Passes (Cmp), Tackles (Tkl), Shots Blocked (Shots_bl), and Interceptions (Int). These metrics reflect both offensive and defensive aspects of the team's performance across the three seasons. They provide a comprehensive view of how Chelsea FC performed in each match and allow the clustering algorithms to group similar matches together based on these features.

This combination of clustering and PCA allows us to visualize how Chelsea's performances have evolved over time and which seasons exhibit the greatest variance in key metrics.

6.3.1 K-Means Clustering

K-Means clustering is a popular unsupervised machine learning algorithm used to partition data points into K distinct clusters. Each cluster is represented by its centroid, which is the mean of the points in the cluster. The algorithm aims to minimize the sum of squared distances between each point and its assigned centroid. It operates iteratively, assigning data points to clusters based on proximity to the centroid, then updating the centroids based on the mean of the newly assigned points.

The objective function of K-Means is:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{6.3.1}$$

Where:

- J is the sum of squared distances (or the inertia). - x is a data point. - μ_i is the centroid of cluster C_i[25].

In the context of this analysis, K-Means is used to group the matches from different seasons based on their performance metrics (xG, Possession, Tackles, etc.). $K = 3$ clusters were chosen to reflect the three seasons (2021/22, 2022/23, and 2023/24).

In this analysis, PCA was used for the K-Means clustering. This allows for a clearer visual representation of the clusters identified by K-Means in a 2D space, which can be difficult when working with many variables like xG, Possession, Tackles, etc. By reducing the dimensions, we can better visualize how different matches across seasons are grouped. The PCA visualization of the K-Means clustering in figure 6.6 shows the distribution of matches from the 2021/22, 2022/23, and 2023/24 seasons across three distinct clusters. The purple cluster, which contains more matches from the 2021/22 season, indicates that this season had a distinct playing style or more consistent performance compared to the others. On the other hand, the yellow and green clusters contain a mix of matches from the 2022/23 and 2023/24 seasons, suggesting that these seasons had more varied performances, with matches spread across different clusters. The consistency of the 2021/22 season is evident from the clustering pattern, where matches are tightly grouped in the purple cluster, indicating a more stable performance throughout the season. In contrast, matches from the 2022/23 and 2023/24 seasons are more dispersed, suggesting inconsistent performances with fluctuations between high and low-performing games.

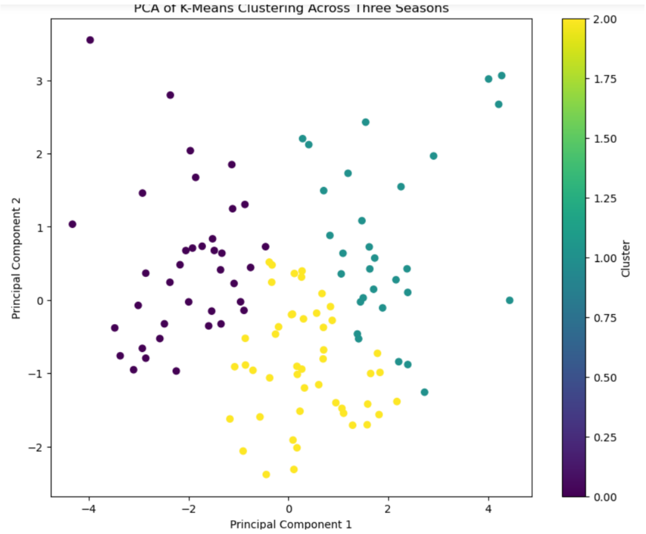

Figure 6.6: KMeans Clustering using PCA

Distinct grouping is also apparent, with the 2021/22 season showing a centralized cluster, while the 2022/23 and 2023/24 seasons display more diversity in match outcomes. This suggests that the 2021/22 season had a more predictable and structured performance, whereas the other two seasons exhibited varied playing styles and results.

In conclusion, the 2021/22 season emerges as more consistent and potentially better performing, while the 2022/23 and 2023/24 seasons show greater variability in match performance, highlighting more erratic outcomes.

CHAPTER 7

Predictive Modelling

Predictive modeling in this study is centered around forecasting match outcomes (Wins, Losses, Draws) and goals scored using historical football data from Chelsea FC's previous seasons. This approach leverages advanced machine learning techniques to analyze a variety of match metrics, including expected goals (xG), possession, shots on target, tackles, and interceptions, among others, to develop predictive models.

Football is basically a complex and dynamic sport, with multiple interdependent factors influencing the result of a match. By building predictive models, we can quantify and evaluate how these variables impact match outcomes. The goal of this study is to develop reliable models that can forecast match results and goals scored for future games based on historical performance, and to assess which models are most effective in this context.

7.1 Predicting Number of Goals for 2023/24 Season

Predictive modeling was conducted to estimate the number of goals that Chelsea FC would score during the 2023/24 season. The analysis utilized match statistics from the 2021/22 and 2022/23 seasons. To achieve accurate predictions, four different machine learning models were selected: Linear Regression, Decision Tree, Random Forest, and XGBoost. Each of these models was chosen for specific reasons, taking into account their suitability for the dataset and the goals of the analysis.

In this analysis, feature selection was based on domain knowledge, focusing on key offensive metrics that directly impact goal-scoring performance. Metrics such as Expected Goals (xG), Possession Percentage (Poss%), and Shots on Target Percentage (SoT%) were selected as they have a significant influence on predicting Goals For (GF). This approach ensures that the model remains interpretable and relevant to the objective of predicting goal outcomes, while avoiding unnecessary complexity from defensive or less relevant features. Prioritizing these offensive metrics helps maintain a clear focus on the prediction of goals scored.

7.1.1 Cross Validation Approach

To ensure that the models were reliable and could perform well on new data, cross-validation approach was used. Cross-validation is a method where the dataset is divided into several parts. In this case, the dataset was split into ten parts, known as 10-fold cross-validation. The model was trained on nine of these parts and tested on the remaining one. This process was repeated ten times, each time using a different part of the data for testing. By averaging the results from these ten rounds, we can get a more accurate and reliable estimate of how well the model will perform on unseen data. This method helps prevent the model from being too closely fitted to the specific data it was trained on, which is known as overfitting, and ensures it can generalize well to other data. The overall performance metric from this cross-validation gives us a good understanding of how the model is likely to perform in practice[26].

The process can be summarized with a formula:

$$\text{CV\_Metric} = \frac{1}{10}\sum_{i=1}^{10} \text{Metric}(\text{Model}_{\text{train}_i}, \text{Validation}_i) \tag{7.1.1}$$

Here, CV_Metric is the average result from all the tests, showing how well the model performed, and $\text{Metric}(\text{Model}_{\text{train}_i}, \text{Validation}_i)$ is the result from each individual test. This method helps ensure the model works well not just with the training data but also with new, unseen data.

7.1.2 Model Selection

Linear regression

Linear Regression is a simple yet powerful model that assumes a linear relationship between the input features—such as expected goals (xG), possession percentage (Poss), and shots on target percentage (SoT%)—and the output, which in this case is the number of goals scored. The model's equation:

$$\text{Predicted Goals} = \beta_0 + \beta_1 \times \text{xG} + \beta_2 \times \text{Poss} + \beta_3 \times \text{SoT\%} + \epsilon \tag{7.1.2}$$

- **Predicted Goals**: Estimated goals based on model variables.
- β_0: Intercept (baseline goals when other variables are zero).
- β_1: Coefficient for expected goals (xG).
- β_2: Coefficient for possession percentage (Poss).
- β_3: Coefficient for shots on target percentage (SoT%).
- ϵ: Error term accounting for unexplained variations.

allows for clear insights into how each feature contributes to the prediction. Its straightforward nature makes it particularly useful when transparency and interpretability are important, especially when understanding how each factor influences goal-scoring[27].

Decision Tree

Decision Tree, on the other hand, is a model designed to handle more complex, non-linear relationships between features and the target outcome. It works by splitting the data into branches based on the values of the input features, making it effective in situations where the relationship between features and goals scored is not straightforward. However, Decision Trees can sometimes overfit the data, performing well on training data but less effectively on new, unseen data[28].

Random Forest

To address the potential overfitting of Decision Trees, Random Forest was employed. This model improves robustness by creating multiple decision trees and averaging their predictions. The ensemble approach of Random Forest reduces the risk of overfitting and enhances accuracy, making it well-suited for complex datasets that may contain various patterns and interactions among the features[29].

XGBoost

Lastly, XGBoost was selected for its exceptional predictive power. XGBoost builds models sequentially, where each new model tries to correct the errors made by previous ones, significantly improving the overall prediction accuracy[30], making it an ideal choice for ensuring high accuracy in predicting goals based on features like xG, possession, and SoT%.

7.1.3 Evaluation Metrics

The performance of these models was evaluated using three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R 2 2). MAE measures the average difference between the predicted and actual number of goals, giving a direct indication of the accuracy of the model. The formula for MAE is given by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{7.1.3}$$

MSE measures the average of the squares of the errors, providing more weight to larger errors. It is useful for identifying models that might occasionally make large mistakes. The formula for MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7.1.4}$$

R-squared explains how much of the variability in the actual goals can be accounted for by the model's predictions. A higher R-squared value indicates a better fit of the model to the data. R-squared is calculated using the formula: [31]

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2} \tag{7.1.5}$$

The cross-validation results revealed that Linear Regression outperformed the other models, achieving the lowest MSE (1.077914) and proving to be the most reliable predictor of goals scored. Although its R² value (0.124703) was not particularly high, it was still superior to those of the other models, making it the best choice for this task.

Table 7.1: Cross Validation Results

Model	MAE	MSE	R Squared
Linear Regression	0.827788	1.077914	0.124703
Decision Tree	1.025000	1.682143	-0.643409
Random Forest	0.919964	1.319148	-0.063773
XGBoost	1.023096	1.747830	-0.392164

In contrast, the Decision Tree, Random Forest, and XGBoost models all exhibited higher MSE values and negative R² scores, indicating poor fit and less reliable predictions. Consequently, Linear Regression was selected as the best model for predicting the number of goals, as it demonstrated the most consistent and accurate performance during cross-validation. The performance metrics for each model are summarized in Table 7.1.

7.1.4 Best model

Linear Regression was used to predict how many goals Chelsea FC would score in the 2023/24 season. The model predicted that they would score 79.12 goals in total, with an average of 2.08 goals per match. In reality, Chelsea scored 77 goals, which means the prediction was very close, with only a small difference of -2.12 goals. This shows that the model did a good job estimating the total number of goals. The prediction for the average goals per match was 2.08, but the actual average was 1.01, resulting in a difference of -1.07 goals per match. While the model slightly overestimated the number of goals per match, it was still quite close to the actual numbers, showing that the prediction was generally accurate.

The analysis of the model's coefficients highlighted that expected goals (xG) emerged as the most critical factor in predicting the number of goals Chelsea FC is likely to score. This finding is substantiated by the significant coefficient value associated with xG in the Linear Regression model. The large coefficient indicates that xG has a substantial influence on the model's output, underscoring its importance as a predictor of goal-scoring outcomes.

Essentially, the higher the xG value, the greater the number of goals Chelsea is expected to score, which aligns with the broader understanding in football analytics where xG is considered a key metric for assessing the quality of goal-scoring opportunities.

While xG was the most influential feature, other factors like Possession Percentage (Poss) and Shots on Target Percentage (SoT%) also played roles in the model's predictions. However, their impact was significantly less pronounced compared to xG. Possession percentage, which reflects the control a team has over the game, and SoT%, which indicates shooting accuracy, did contribute to the model, but their coefficients were much smaller, indicating that while they have some predictive power, they are not as decisive as xG in determining the number of goals.

This relationship between the features and the predicted goals is visually represented in Figure 7.1. As seen in the figure, xG stands out as the dominant feature, clearly dwarfing the contributions of Poss and SoT%. This visualization reinforces the conclusion that xG is the most powerful predictor among the features analyzed, driving the model's predictions of Chelsea FC's goal-scoring potential.

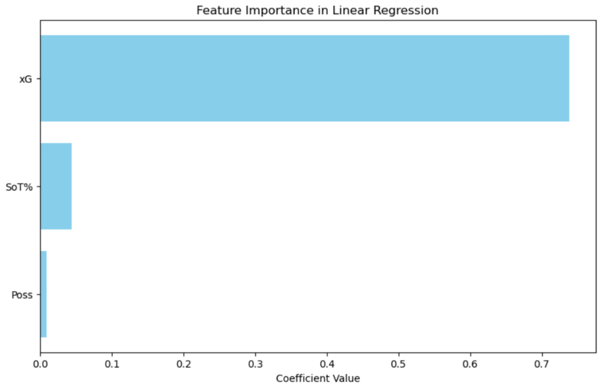

Figure 7.1: Significant Feature

7.2 Predicting Match Outcomes for 2023/24 Season

Multiple machine learning models were used to predict the match outcomes (Wins, Losses, and Draws) for Chelsea FC in the 2023/24 season, based on historical data from the 2021/22 and 2022/23 seasons. The models used include Support Vector Machines (SVM), Multinomial Logistic Regression, K-Nearest Neighbors (KNN) and Naive Bayes.

7.2.1 Support Vector Machine (SVM)

SVM aims to find the optimal hyperplane that best separates the classes (Wins, Losses, Draws) in a high-dimensional space. It is effective in high-dimensional spaces and works well when the relationship between variables is non-linear, making it ideal for football data where multiple metrics interact in complex ways.

Equation: For linear SVM, the decision boundary is found by:

$$f(x) = \mathbf{w}^T \cdot x + b \tag{7.2.1}$$

where w is the weight vector, x is the input, and b is the bias term. The goal is to maximize the margin between classes[32].

7.2.2 Multinomial Logistic Regression

This model is used when the outcome is a categorical variable with more than two classes. It models the probability of each class (Win, Loss, Draw) and is highly interpretable. It is especially useful in multi-class classification problems like football match predictions.

Equation: For class k, the probability $P(Y = k \mid x)$ is given by:

$$P(Y = k \mid x) = \frac{\exp(\beta_k^T \cdot x)}{1 + \sum_{k=1}^{K-1} \exp(\beta_k^T \cdot x)} \tag{7.2.2}$$

where β_k are the coefficients associated with class k[33].

7.2.3 K-Nearest Neighbors (KNN)

This model is non-parametric and classifies an observation based on the majority vote of its nearest neighbors in the feature space. It is simple but can struggle with noisy data or large feature sets[34].

7.2.4 Naive Bayes

The Naive Bayes classifier assumes conditional independence between features given the class, simplifying the joint probability calculation. The likelihood of a class C given a set of features $X = \{X_1, X_2, \ldots, X_n\}$ is calculated using Bayes' Theorem as:

$$P(C \mid X) = \frac{P(C) \prod_{i=1}^{n} P(X_i \mid C)}{P(X)} \tag{7.2.3}$$

Where $P(C)$ is the prior probability, and $P(X_i \mid C)$ represents the conditional probabilities of the features X_i given the class C[35].

7.2.5 Feature Selection

Feature selection was performed using a Random Forest Classifier to determine the most important features for predicting match outcomes. The model was trained using a combination of numerical and categorical features, which were preprocessed using StandardScaler for numerical data and OneHotEncoder for categorical data (such as "Venue," "Opponent," and "Formation"). The Random Forest model ranked the feature importances, allowing the top 10 most important features to be selected for further analysis.

The top 10 features selected are: G/Sh (Goals per Shot), G/SoT (Goals per Shot on Target), Save%, CS (Clean Sheets), GA (Goals Against), SoT% (Shots on Target Percentage), xG (Expected Goals), Err (Errors leading to shots or goals), xGA (Expected Goals Against), and npxG/Sh (Non-Penalty Expected Goals per Shot). These features encapsulate a combination of offensive, defensive, and overall team performance metrics, providing a comprehensive view of the factors most influential in determining match outcomes. These selected features were used in the subsequent model training and evaluation.

7.2.6 Best Model

The cross-validation results in table 7.2 indicate that Multinomial Logistic Regression performs the best across all metrics when predicting match outcomes. It achieved an accuracy of 77.68%, which is slightly higher than that of SVM at 76.79%. In terms of precision and recall, Multinomial Logistic Regression scored 0.8003 and 0.7768, respectively, reflecting a strong balance between correctly identifying wins, losses, and draws, and minimizing false positives. Its F1-score, which balances precision and recall, was 0.7610, reinforcing the model's capability to handle imbalanced classes effectively.

Table 7.2: Cross Validation Results

Model	Accuracy	Precision	Recall	F1-Score
SVM	0.7679	0.8051	0.7679	0.7485
Multinomial Logistic Regression	0.7768	0.8003	0.7768	0.7610
KNN	0.5893	0.5872	0.5893	0.5594
Naive Bayes	0.6393	0.6229	0.6393	0.5902

SVM also performed well, with a comparable accuracy of 76.79%, precision of 0.8051, and recall of 0.7679, suggesting that both models are highly competitive. However, the slight edge in performance metrics makes Multinomial Logistic Regression the preferred model. In contrast, KNN and Naive Bayes showed relatively weaker performances, with KNN achieving only 58.93% accuracy and Naive Bayes at 63.93%. Both models had lower F1-scores, with KNN at 0Z5594 and Naive Bayes at 0.5902, highlighting their struggle in this classification task.

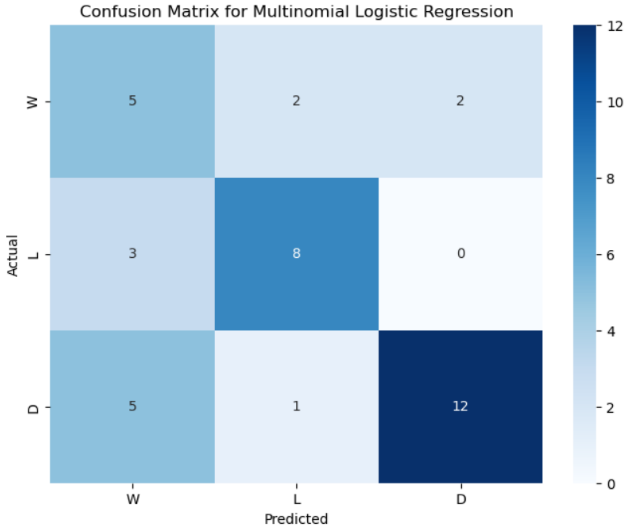

Figure 7.2: Confusion Matrix

Looking at the confusion matrix in figure 7.2 for Multinomial Logistic Regression, we see that out of 12 actual draws, the model correctly predicted 12, misclassifying only 1 as a loss. However, it misclassified some actual wins (5) as losses or draws, and some losses (3) as wins or draws. Despite these misclassifications, the model still performs well in classifying the majority of matches accurately.

CHAPTER 8

Conclusions

8.1 Discussion

This study aimed to comprehensively analyze Chelsea FC's performance over three seasons, employing a range of techniques such as data visualization, summary statistics, clustering, and MANOVA. The results confirmed that key performance indicators (KPIs) play a crucial role in evaluating team success, which aligns with the findings of Hughes et al. (2002). These tools enabled a detailed assessment of both offensive and defensive metrics, providing valuable insights into the factors that contributed to Chelsea's performance. The analysis reinforced the importance of performance indicators in the broader context of football analytics.

The correlation analysis showed that Goals For (GF) and Expected Goals (xG) had a strong positive correlation with wins, supporting Lucey et al. (2014), who emphasized that xG is a critical metric in assessing a team's offensive capabilities. This result highlights the effectiveness of xG as a predictor of success in football, confirming its value in tactical and strategic planning. Interestingly, possession was found to have the weakest correlation with wins, which is contrary to the conclusions drawn by Jones et al. (2004) and Penos et al. (2010), who argued that possession is pivotal to a team's chances of winning. While Jones et al. and Penos et al. highlighted possession as a critical factor in football success, this analysis found it to be far less impactful for Chelsea FC's performance, especially in the seasons analyzed. This contrast could be due to the context of Chelsea's playing style or specific tactical changes during the 2021/22 to 2023/24 seasons. For example, the possession-heavy approach might not have directly translated into results, suggesting that merely dominating possession is insufficient without tactical adjustments to convert possession into goal-scoring opportunities. This divergence highlights the evolving nature of football tactics, where possession may not be as crucial in achieving success as it once was, and where modern metrics like xG provide a more comprehensive understanding of team performance.

In predictive modeling, xG emerged as a significant factor in forecasting the number of goals Chelsea would score, further validating the conclusions of Lucey et al. (2014). The multinomial logistic regression model was identified as the best-performing model, with an accuracy of 77.68%. While Igiri et al. (2014) reported a higher accuracy of 93% with logistic regression, their model was limited to predicting only wins and losses, excluding draws. This distinction underscores the importance of comprehensive models that can handle multiple match outcomes for more precise predictions.

Limitations

While the study effectively analyzed Chelsea FC's performance across three seasons, several limitations could be addressed in future research. First, the features used in the analysis were primarily team-level metrics such as possession, expected goals, and shots on target. Incorporating more granular data, like player-specific metrics (e.g., individual player form, injuries, fitness levels), could enhance the predictive power of the models. Second, the analysis did not account for external factors such as opposition strength, tactical variations, or player injuries, which could significantly affect performance and predictions. Lastly, expanding the scope to include analysis of opposing teams' strengths and weaknesses could provide a more holistic view of match outcomes and improve model accuracy in predicting real-world results. By including these additional variables, future studies could offer more robust insights into both team and player performances.

Conclusion

This study found that visualization techniques and statistical methods provide essential insights into Chelsea FC's performance over the 2021/22, 2022/23, and 2023/24 seasons. Key performance indicators such as goals scored and expected goals were shown to be significant predictors of success, while possession was less impactful. Predictive modeling confirmed that expected goals are crucial for forecasting goal outcomes, with multinomial logistic regression being the most effective model for predicting match outcomes.

Among the seasons analyzed, the 2021/22 season was identified as the best, with Chelsea displaying strong offensive and defensive performances, leading to more wins. In contrast, the 2022/23 season was the worst, marked by a sharp decline in both attacking output and defensive stability, resulting in a higher loss rate and fewer wins. The 2023/24 season, though an improvement from 2022/23, still did not reach the heights of 2021/22 but demonstrated positive steps in both goal scoring and overall team performance.

APPENDIX A

Chelsea Match Statistics Features

Table A.1: Features in the Chelsea Match Statistics Dataset

Feature	**Description**	**Feature**	**Description**
Date	Date of the match	Poss	Possession percentage
Time	Time when the match was played	Attendance	Number of spectators
Round	Round or matchday number	Captain	Captain for the match
Day	Day of the week	Formation	Team formation used (e.g., 4-3-3)
Venue	Home or Away	Referee	Match referee
Result	Match result (Win, Loss, Draw)	Gls	Goals scored in the match
GF	Goals For (Goals scored by Chelsea)	Sh	Shots taken
GA	Goals Against (Goals conceded by Chelsea)	SoT	Shots on target
Opponent	Opposing team	SoT%	Percentage of shots on target
xG	Expected Goals	G/Sh	Goals per shot ratio
xGA	Expected Goals Against	G/SoT	Goals per shot on target ratio
npxG	Non-penalty expected goals	Dist	Distance from goal for shots
npxG/Sh	Non-penalty xG per shot	FK	Free kicks
G-xG	Difference between goals and expected goals	PK	Penalty kicks
np:G-xG	Non-penalty goals minus non-penalty xG	Cmp	Completed passes
Cmp	Completed passes	Att	Attempted passes
Cmp%	Pass completion percentage	TotDist	Total distance covered in passes
PrgDist	Progressive distance	Cmp.1	Completed short passes
Att.1	Attempted short passes	Cmp%.1	Short pass completion percentage
Cmp.2	Completed medium passes	Att.2	Attempted medium passes
Cmp%.2	Medium pass completion percentage	Cmp.3	Completed long passes
Att.3	Attempted long passes	Cmp%.3	Long pass completion percentage
Ast	Assists	xAG	Expected assists
xA	Expected assists (same as xAG)	KP	Key passes
1/3rd	Passes into the final third	PPA	Passes into the penalty area
CrsPA	Crosses into the penalty area	PrgP	Progressive passes
Touches	Total touches	Def Pen	Defensive touches in the penalty area
Def 3rd	Touches in the defensive third	Mid 3rd	Touches in the middle third

Table A.2: Features in the Chelsea Match Statistics Dataset Cont..

Att 3rd	Touches in the attacking third	Att Pen	Touches in the attacking penalty area
Live	Live-ball touches	AttTakeons	Attempted take-ons
Succ	Successful take-ons	Succ%	Success percentage for take-ons
Tkld	Tackles made	Tkld%	Tackle success percentage
Carries	Total carries	TotDis_poss	Total distance covered while in possession
PrgDis_poss	Progressive distance while in possession	PrgC	Progressive carries
1/3rd_poss	Carries into the final third	CPA	Carries into the penalty area
Mis	Miscontrols	Dis	Dispossessions
Rec	Passes received	PrgR	Progressive receptions
Tkl	Total tackles	TklW	Tackles won
Def 3_def	Defensive third tackles	Mid 3_def	Middle third tackles
Att 3_def	Attacking third tackles	Tkl.1	Tackle attempts
A_def	Tackles against dribblers	Tkl%	Tackle success percentage against dribblers
Lost	Ball losses	Blocks	Blocks made
Shots_bl	Shots blocked	Pass	Passes blocked
Int	Interceptions	Tkl+Int	Total tackles and interceptions
Clr	Clearances	Err	Errors leading to shots or goals
SoTA	Shots on target against (goalkeeping)	Saves	Saves made
Save%	Save percentage	CS	Clean sheets
PSxG	Post-shot expected goals (goalkeeping)	PSxG+/-	Difference between PSxG and goals allowed
Pkatt_goalkeeping	Penalty kicks faced (goalkeeping)	PKsv	Penalties saved
Cmp_goalkeeping	Completed passes (goalkeeping)	Att_goalkeeping	Attempted passes (goalkeeping)
Cmp%_goalkeeping	Pass completion percentage (goalkeeping)	Att (GK)	Goal kicks attempted
Thr	Throw attempts (goalkeeper distribution)	Launch%	Percentage of launched passes (goalkeeper distribution)
AvgLen	Average length of launched passes (goalkeeper distribution)	Opp	Opponent's attempted crosses into penalty area
Stp	Sweeper keeper actions	Stp%	Success percentage of sweeping actions
#OPA	Off-penalty area actions (goalkeeping)	AvgDist	Average distance of off-penalty actions

APPENDIX B

Code Snippets

B.1 Data Preprocessing

Reading the match stats for three seasons

```
# Reading the match stats for three seasons
matchstats_21_22 = pd.read_csv(’2021-22_matchstats.csv’)
matchstats_22_23 = pd.read_csv(’22-23_matchstats.csv’)
matchstats_23_24 = pd.read_csv(’23-24_matchstats.csv’)
```

Handling Missing Values

```
def fill_missing_values(df):
    for column in df.columns:
        if df[column].isnull().any():
            if df[column].dtype in [np.float64, np.int64]:
                mean_value = df[column].mean()  # replace numeric values
                df[column].fillna(mean_value, inplace=True)
            else:   #replace non-numeric values
                mode_value = df[column].mode()[0] columns
                df[column].fillna(mode_value, inplace=True)
    return df

# Applying missing value handling to each dataset
matchstats_21_22 = fill_missing_values(matchstats_21_22)
matchstats_22_23 = fill_missing_values(matchstats_22_23)
matchstats_23_24 = fill_missing_values(matchstats_23_24)
```

Validation

```
def validate_dataset(df, name):
    print(f"Validating {name}")
    print(df.describe())  # Checking summary statistics
    duplicates = df.duplicated().sum()
    print(f"Number of duplicate rows: {duplicates}")
```

```
# Apply validation to all datasets
datasets = {
    'matchstats_21_22': matchstats_21_22,
    'matchstats_22_23': matchstats_22_23,
    'matchstats_23_24': matchstats_23_24,
}
for name, df in datasets.items():
    validate_dataset(df, name)
```

B.2 MANOVA

```
import pandas as pd
from statsmodels.multivariate.manova import MANOVA
matchstats_21_22['Season'] = '2021/22'
matchstats_22_23['Season'] = '2022/23'
matchstats_23_24['Season'] = '2023/24'

# Concatenate all seasons' data
data = pd.concat([matchstats_21_22[['GF', 'GA', 'xG', 'Season']],
                  matchstats_22_23[['GF', 'GA', 'xG', 'Season']],
                  matchstats_23_24[['GF', 'GA', 'xG', 'Season']]])

manova = MANOVA.from_formula('GF + GA + xG ~ Season', data=data) # Perform MANOVA
print(manova.mv_test())
```

B.3 Kmeans

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Combine data
matchstats_21_22['Season'] = '2021/22'
matchstats_22_23['Season'] = '2022/23'
matchstats_23_24['Season'] = '2023/24'
combined_stats = pd.concat([matchstats_21_22, matchstats_22_23, matchstats_23_24])

# Select relevant metrics for clustering
metrics = combined_stats[['xG', 'Poss', 'Sh', 'SoT', 'Cmp', 'Tkl', 'Shots_bl', 'Int']]

scaler = StandardScaler()  # Normalize the data
metrics_normalized = scaler.fit_transform(metrics)

kmeans = KMeans(n_clusters=3, random_state=42)   # Perform K-Means Clustering
combined_stats['KMeans_Cluster'] = kmeans.fit_predict(metrics_normalized)



pca = PCA(n_components=2)
principal_components = pca.fit_transform(metrics_normalized)
plt.figure(figsize=(10, 8))   # Visualize the clusters using PCA
plt.scatter(principal_components[:, 0], principal_components[:, 1], c=combined_stats['KMeans_Cluster'], cmap='viridis')
plt.title('PCA of K-Means Clustering Across Three Seasons')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.colorbar(label='Cluster')
plt.show()
```

B.4 Predictive Modelling

Predicting number of goals

```
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_score, KFold
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from xgboost import XGBRegressor  # Import XGBoost
import numpy as np
import random
# Set random seeds
np.random.seed(42)
random.seed(42)

# Features and target variable
features = ['xG', 'Poss', 'SoT%']
target = 'GF'

# Combine datasets for training
data = pd.concat([matchstats_21_22, matchstats_22_23])
X = data[features]
y = data[target]

# Prepare testing data for 2023/24 season
X_test = matchstats_23_24[features]
y_test = matchstats_23_24[target]

# Cross-validation setup
kf = KFold(n_splits=10, shuffle=True, random_state=42)

# Define models
models = {
    'Linear Regression': LinearRegression(),
    'Decision Tree': DecisionTreeRegressor(),
    'Random Forest': RandomForestRegressor(),
    'XGBoost': XGBRegressor()
}

# Evaluate models with cross-validation
results = {}
for name, model in models.items():
    cv_mae = cross_val_score(model, X, y, cv=kf,  scoring='neg_mean_absolute_error')
    cv_mse = cross_val_score(model, X, y, cv=kf, scoring='neg_mean_squared_error')
    cv_r2 = cross_val_score(model, X, y, cv=kf, scoring='r2')

    results[name] = {
        'MAE': -cv_mae.mean(),
        'MSE': -cv_mse.mean(),
        'R2': cv_r2.mean()
    }

# Convert results to DataFrame
results_df = pd.DataFrame(results).T

# Printing the cross-validation results
print("Cross-Validation Results:")
print(results_df)

# Identify the best model based on MSE
best_model_name = results_df['MSE'].idxmin()
```

```
best_model = models[best_model_name]
# Printing the best model based on MSE
print(f"\nBest Model: {best_model_name}")

# Train the best model on the full dataset
best_model.fit(X, y)

# Predict for the 2023/24 season
predictions_2023_24 = best_model.predict(X_test)

# Evaluate predictions for the 2023/24 season
mae_2023_24 = mean_absolute_error(y_test, predictions_2023_24)
mse_2023_24 = mean_squared_error(y_test, predictions_2023_24)
r2_2023_24 = r2_score(y_test, predictions_2023_24)

# Calculate total and average predicted goals
total_predicted_goals = predictions_2023_24.sum()
average_predicted_goals_per_match = predictions_2023_24.mean()

# Calculate actual goals
actual_total_goals = matchstats_23_24['GF'].sum()
num_matches = len(matchstats_23_24)
actual_avg_goals_per_match = actual_total_goals / num_matches
```

Feature Importance

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

# Train the Linear Regression model on the entire dataset
linear_regression_model = LinearRegression()
linear_regression_model.fit(X, y)  # Using the full dataset X, y

# Extract the coefficients
coefficients = linear_regression_model.coef_

# Create a DataFrame to tie features with their coefficients
feature_importance_lr = pd.DataFrame({
    'Feature': features,
    'Coefficient': coefficients
})

# Sort the features by the absolute value of their coefficients
feature_importance_lr = feature_importance_lr.sort_values(by='Coefficient', ascending=False)

# Plot the feature importance
plt.figure(figsize=(10, 6))
plt.barh(feature_importance_lr['Feature'], feature_importance_lr['Coefficient'], color='skyblue')
plt.xlabel('Coefficient Value')
plt.title('Feature Importance in Linear Regression')
plt.gca().invert_yaxis()  # Invert y-axis to show the most important feature at the top
plt.show()
```

Predicting Match Outcomes

```
import pandas as pd
from sklearn.model_selection import StratifiedKFold, GridSearchCV, cross_val_score
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score, confusion_matrix, make_scorer
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier

# Combine datasets for training
common_features=matchstats_21_22.columns.intersection(matchstats_22_23.columns).intersection(matchstats_23_24.columns)

# Exclude irrelevant columns
features_to_exclude = [’Season’, ’Date’, ’Day’, ’Captain’, ’Result’, ’Referee’, ’Time’, ’Round’, ’Attendance’]
all_features = common_features.difference(features_to_exclude)

# Prepare training and testing data
X = data[all_features]
y = data[’Result’]  # Assuming ’Result’ is W, L, or D.

# Prepare testing data for 2023/24 season
X_test = matchstats_23_24[all_features]
y_test = matchstats_23_24[’Result’]

# OneHotEncoding for categorical variables
categorical_features = [’Venue’, ’Opponent’, ’Formation’]  # Assuming these are the categorical variables
numeric_features = [col for col in all_features if col not in categorical_features]

# Define transformers
preprocessor = ColumnTransformer(
    transformers=[
        (’num’, StandardScaler(), numeric_features),
        (’cat’, OneHotEncoder(handle_unknown=’ignore’), categorical_features)
    ])

# Random Forest for feature selection
rf = RandomForestClassifier(n_estimators=100, random_state=42)
pipeline = Pipeline(steps=[
    (’preprocessor’, preprocessor),
    (’rf’, rf)
])

# Fit the pipeline to data
pipeline.fit(X, y)

# Get feature importances from the Random Forest
rf_feature_importances = pipeline.named_steps[’rf’].feature_importances_

# Combine feature names with their importances
feature_importance_dict = dict(zip(numeric_features + list(pipeline.named_steps[’preprocessor’].named_transformers_[’cat’].get_feature_names_out()), r
sorted_features = sorted(feature_importance_dict.items(), key=lambda item: item[1], reverse=True)

# Select top 10 features
top_features = [feature[0] for feature in sorted_features[:10]]
print(f"Top 10 Features: {top_features}")

# Define custom scorers for precision, recall, and f1
scorers = {
    ’accuracy’: ’accuracy’,
    ’precision’: make_scorer(precision_score, average=’weighted’),
    ’recall’: make_scorer(recall_score, average=’weighted’),
    ’f1’: make_scorer(f1_score, average=’weighted’)
}

# Use only top features for training and testing
X_top_features = X[top_features]
```

```
X_test_top_features = X_test[top_features]

# Stratified KFold Cross-Validation
skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)

# Define models
models = {
    'SVM': SVC(class_weight='balanced'),
    'Multinomial Logistic Regression': LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=1000, class_weight='balanced'),
    'KNN': KNeighborsClassifier(),
    'Naive Bayes': GaussianNB(),
}

# Evaluate each model using cross-validation for multiple metrics
results = {}
for name, model in models.items():
    print(f"Evaluating model: {name}")
    results[name] = {}
    for metric_name, scorer in scorers.items():
        if name == 'SVM':
            grid_svm.fit(X_top_features, y)
            best_svm = grid_svm.best_estimator_
            score = cross_val_score(best_svm, X_top_features, y, cv=skf, scoring=scorer).mean()
        elif name == 'KNN':
            grid_knn.fit(X_top_features, y)
            best_knn = grid_knn.best_estimator_
            score = cross_val_score(best_knn, X_top_features, y, cv=skf, scoring=scorer).mean()
        else:
            score = cross_val_score(model, X_top_features, y, cv=skf, scoring=scorer).mean()
        results[name][metric_name] = score

# Convert results to DataFrame for easy viewing
results_df = pd.DataFrame(results).T

# Display cross-validation results
print("Cross-Validation Results:")
print(results_df)

# Identify the best model based on accuracy
best_model_name = results_df['accuracy'].idxmax()

# Print the best model based on accuracy
print(f"\nBest Model: {best_model_name}")

# Train the best model on the full training dataset
best_model = models[best_model_name]
best_model.fit(X_top_features, y)

# Predict for the 2023/24 season
y_pred = best_model.predict(X_test_top_features)

# Confusion Matrix
conf_matrix = confusion_matrix(y_test, y_pred)

# Visualize Confusion Matrix
plt.figure(figsize=(8,6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['W', 'L', 'D'], yticklabels=['W', 'L', 'D'])
plt.title(f'Confusion Matrix for {best_model_name}')
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```

Bibliography

[1] Jaime Orejan. *Football/Soccer: history and tactics*. McFarland, Jefferson, North Carolina, 2011.

[2] Joel Oberstone. Comparing team performance of the english premier league, serie a, and la liga for the 2008-2009 season. *Journal of Quantitative Analysis in Sports*, 7(1), 2011.

[3] Ibad Ur Rehman. Multiple regression model for predicting premier league score of chelsea fc.

[4] Peter O'Donoghue. *An introduction to performance analysis of sport*. Routledge, 2014.

[5] Craig Wright, Chris Carling, Craig Lawlor, and David Collins. Elite football player engagement with performance analysis. *International Journal of Performance Analysis in Sport*, 16(3):1007–1032, 2016.

[6] Craig Wright, Chris Carling, and David Collins. The wider context of performance analysis and it application in the football coaching process. *International Journal of Performance Analysis in Sport*, 14(3):709–733, 2014.

[7] Rob Mackenzie and Chris Cushion. Performance analysis in football: A critical review and implications for future research. *Journal of sports sciences*, 31(6):639–676, 2013.

[8] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. 2015.

[9] Mike D Hughes and Roger M Bartlett. The use of performance indicators in performance analysis. *Journal of sports sciences*, 20(10):739–754, 2002.

[10] PD Jones, Nic James, and Stephen D Mellalieu. Possession as a performance indicator in soccer. *International Journal of Performance Analysis in Sport*, 4(1):98–102, 2004.

[11] Carlos Lago-Peñas and Alexandre Dellal. Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables. *Journal of human kinetics*, 25(2010):93–100, 2010.

[12] Hugo Sarmento, Filipe Manuel Clemente, Duarte Araújo, Keith Davids, Allistair McRobert, and António Figueiredo. What performance analysts need to know about research trends in association football (2012–2016): A systematic review. *Sports medicine*, 48:799–836, 2018.

[13] Anthony Constantinou and Norman Fenton. Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, 124:93–104, 2017.

[14] Siem Jan Koopman and Rutger Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(1):167–186, 2015.

[15] Christina Markopoulou, George Papageorgiou, and Christos Tjortjis. Diverse machine learning for forecasting goal-scoring likelihood in elite football leagues. *Machine Learning and Knowledge Extraction*, 6(3):1762–1781, 2024.

[16] Chinwe Peace Igiri and Enoch Okechukwu Nwachukwu. An improved prediction system for football a match result. *IOSR journal of Engineering*, 4(12):12–20, 2014.

[17] Harvey J Keselman, Carl J Huberty, Lisa M Lix, Stephen Olejnik, Robert A Cribbie, Barbara Donahue, Rhonda K Kowalchuk, Laureen L Lowman, Martha D Petoskey, Joanne C Keselman, et al. Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of educational research*, 68(3):350–386, 1998.

[18] Russell Warne. A primer on multivariate analysis of variance (manova) for behavioral scientists. *Practical Assessment, Research, and Evaluation*, 19(1), 2014.

[19] Sweta Patel and CD Bhavsar. Analysis of pharmacokinetic data by wilk's lambda (an important tool of manova). *International Journal of Pharmaceutical Science Invention*, 2(1):36–44, 2013.

[20] A El Ouardighi, A El Akadi, and D Aboutajdine. Feature selection on supervised classification using wilks lambda statistic. In *2007 International Symposium on Computational Intelligence and Intelligent Informatics*, pages 51–55. IEEE, 2007.

[21] Can Ateş, Özlem Kaymaz, H Emre Kale, and Mustafa Agah Tekindal. Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-i error rates. *Computational and mathematical methods in medicine*, 2019(1):2173638, 2019.

[22] Aaron French, Marcelo Macedo, John Poulsen, Tyler Waterson, and Angela Yu. Multivariate analysis of variance (manova). *San Francisco State University*, 2008.

[23] Laerd Statistics. One-way repeated measures manova in spss statistics. *Statistical tutorials and software guides*, 2018.

[24] Lior Rokach and Oded Maimon. Clustering methods. *Data mining and knowledge discovery handbook*, pages 321–352, 2005.

[25] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, pages 63–67. Ieee, 2010.

[26] Daniel Berrar et al. Cross-validation., 2019.

[27] Xiaogang Su, Xin Yan, and Chih-Ling Tsai. Linear regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(3):275–294, 2012.

[28] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[29] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[30] Adeola Ogunleye and Qing-Guo Wang. Xgboost model for chronic kidney disease diagnosis. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(6):2131–2140, 2019.

[31] Abhishek V Tatachar. Comparative assessment of regression models based on model evaluation metrics. *International Journal of Innovative Technology and Exploring Engineering*, 8(9):853–860, 2021.

[32] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1):41–51, 2018.

[33] Abdalla M El-Habil. An application on multinomial logistic regression model. *Pakistan journal of statistics and operation research*, pages 271–291, 2012.

[34] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[35] Sona Taheri and Musa Mammadov. Learning the naive bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, 23(4):787–795, 2013.