Premier League Analysis Portfolio

Adithyan M

# Premier League Analysis Portfolio

This portfolio showcases various analyses and visualizations based on the Premier League dataset.

# Introduction

Welcome to my Premier League Analysis Portfolio. As a passionate data analyst and football enthusiast, I am excited to delve into the rich tapestry of Premier League data using advanced analytical tools in R. This portfolio aims to illuminate the intricacies of team performances, player contributions, and match outcomes across various seasons, providing valuable insights that resonate with both seasoned analysts and avid football fans.

# Objectives and Goals

My primary goal with this portfolio is to uncover hidden patterns within the vast dataset of Premier League matches. By employing sophisticated data analysis techniques, I seek to not only highlight trends in team performance but also predict future outcomes with accuracy. Through this journey, I hope to contribute to the broader understanding of sports analytics and showcase the power of data-driven decision-making in football management.

# Dataset Overview

The dataset used in this analysis is a comprehensive collection of Premier League match details, covering essential information such as match dates, team line-ups, goals scored by home and away teams, and match outcomes (win, loss, draw). This dataset enables a detailed exploration of team dynamics over multiple seasons, facilitating insights into seasonal performance trends and the influence of key variables on match results.

By leveraging advanced data manipulation and visualization techniques in R, I aim to uncover meaningful patterns in team performances, assess the consistency of goal-scoring trends, and analyse the impact of match venues (home vs. away) on team outcomes. This exploration not only enhances our understanding of football analytics but also demonstrates the power of data-driven approaches in sports analysis.

# Methodology

I will employ a robust methodology that blends exploratory data analysis with advanced statistical modeling techniques in R. This approach will enable me to uncover nuanced insights and visualize complex patterns effectively. From time series analysis to regression modeling, each method will be carefully chosen to extract meaningful conclusions from the data.

# Audience and Impact

This portfolio is crafted to inform and inspire fellow football enthusiasts and analysts alike. Whether you’re a coach seeking to refine team strategies, a scout on the lookout for talent insights, or a dedicated fan eager to deepen your understanding of your team’s performance, this analysis aims to deliver actionable insights that extend far beyond the final whistle

# Structure

In structuring this portfolio, I aim to craft a cohesive narrative that begins with foundational analyses of team performances and progresses systematically towards predictive modeling of match outcomes. Each section is meticulously organized to build upon the last, offering a comprehensive view of Premier League dynamics from season to season

Through this structured approach, readers will embark on a journey of discovery, gaining insights into the intricacies of team dynamics, player contributions, and the factors influencing match results. By merging the realms of data analytics and football, this portfolio aims to showcase how these disciplines converge to unearth new perspectives and drive informed decisions in the realm of sports.

## Load Necessary Packages and Data

library(dplyr)  
library(lubridate)  
library(ggplot2)  
library(tidyr)  
# Read the Premier League dataset  
pl\_data <- read.csv("C:/Users/91730/OneDrive/premier-league-matches.csv")

# Data Structure and Summary

# Display dataset structure and summary  
str(pl\_data)

## 'data.frame': 12026 obs. of 8 variables:  
## $ Season\_End\_Year : int 1993 1993 1993 1993 1993 1993 1993 1993 1993 1993 ...  
## $ Week\_number\_of\_match: int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Date : chr "15-08-1992" "15-08-1992" "15-08-1992" "15-08-1992" ...  
## $ Home : chr "Coventry City" "Leeds United" "Sheffield Utd" "Crystal Palace" ...  
## $ HomeGoals : int 2 2 2 3 2 1 1 0 1 1 ...  
## $ AwayGoals : int 1 1 1 3 4 1 1 0 1 0 ...  
## $ Away : chr "Middlesbrough" "Wimbledon" "Manchester Utd" "Blackburn" ...  
## $ FTR : chr "H" "H" "H" "D" ...

summary(pl\_data)

## Season\_End\_Year Week\_number\_of\_match Date Home   
## Min. :1993 Min. : 1.00 Length:12026 Length:12026   
## 1st Qu.:2000 1st Qu.:10.00 Class :character Class :character   
## Median :2008 Median :20.00 Mode :character Mode :character   
## Mean :2008 Mean :19.73   
## 3rd Qu.:2016 3rd Qu.:29.00   
## Max. :2023 Max. :42.00   
## HomeGoals AwayGoals Away FTR   
## Min. :0.000 Min. :0.000 Length:12026 Length:12026   
## 1st Qu.:1.000 1st Qu.:0.000 Class :character Class :character   
## Median :1.000 Median :1.000 Mode :character Mode :character   
## Mean :1.524 Mean :1.143   
## 3rd Qu.:2.000 3rd Qu.:2.000   
## Max. :9.000 Max. :9.000

# Convert Date to Date Format

pl\_data$Date <- as.Date(pl\_data$Date)

# Subset Data

pl\_subset <- pl\_data %>%   
 select(Season\_End\_Year, Week\_number\_of\_match, Date, Home, HomeGoals, AwayGoals, Away, FTR)  
  
# View the first few rows of the subsetted data  
head(pl\_subset)

## Season\_End\_Year Week\_number\_of\_match Date Home HomeGoals  
## 1 1993 1 0015-08-19 Coventry City 2  
## 2 1993 1 0015-08-19 Leeds United 2  
## 3 1993 1 0015-08-19 Sheffield Utd 2  
## 4 1993 1 0015-08-19 Crystal Palace 3  
## 5 1993 1 0015-08-19 Arsenal 2  
## 6 1993 1 0015-08-19 Ipswich Town 1  
## AwayGoals Away FTR  
## 1 1 Middlesbrough H  
## 2 1 Wimbledon H  
## 3 1 Manchester Utd H  
## 4 3 Blackburn D  
## 5 4 Norwich City A  
## 6 1 Aston Villa D

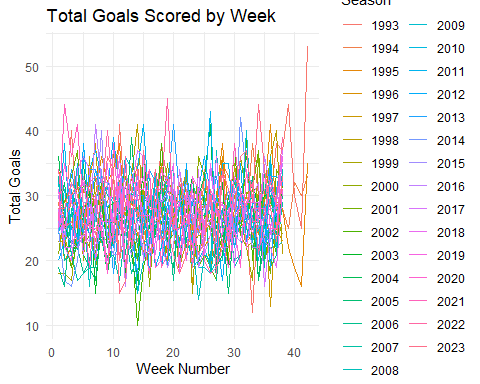
# Convert Date to Week format  
pl\_subset$Week <- week(pl\_subset$Date)

# Goals Aggregation by Week

# Aggregate data by Week and calculate total goals scored by Home and Away teams  
goals\_aggregated <- pl\_subset %>%  
 group\_by(Season\_End\_Year, Week\_number\_of\_match) %>%  
 summarise(HomeGoals = sum(HomeGoals), AwayGoals = sum(AwayGoals))

## `summarise()` has grouped output by 'Season\_End\_Year'. You can override using  
## the `.groups` argument.

# Plotting the time series of goals scored  
ggplot(goals\_aggregated, aes(x = Week\_number\_of\_match, y = HomeGoals + AwayGoals, color = as.factor(Season\_End\_Year))) +  
 geom\_line() +  
 labs(x = "Week Number", y = "Total Goals", color = "Season") +  
 ggtitle("Total Goals Scored by Week") +  
 theme\_minimal()



# Total Goals Scored by Liverpool FC per Season

# Calculate total goals scored per match for Home and Away teams  
pl\_subset <- pl\_subset %>%  
 mutate(TotalGoals = HomeGoals + AwayGoals)  
summary(pl\_subset$TotalGoals)

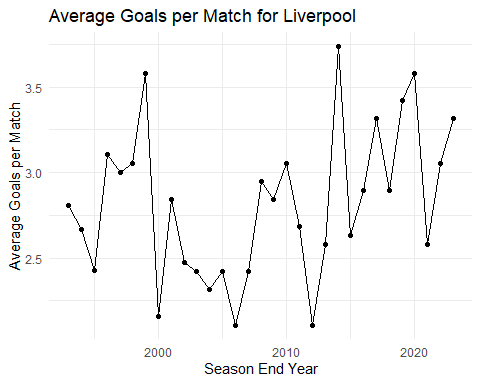
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 2.667 4.000 11.000

# Calculate average goals per match for each team  
team\_performance <- pl\_subset %>%  
 group\_by(Season\_End\_Year, Team = Home) %>%  
 summarise(AvgGoals = mean(TotalGoals))

## `summarise()` has grouped output by 'Season\_End\_Year'. You can override using  
## the `.groups` argument.

# Average Goals per Match for Liverpool

# Plotting the average goals scored per match for Liverpool  
team <- "Liverpool"   
team\_performance\_filtered <- team\_performance %>%  
 filter(Team == team)  
  
# Check if there's data for the specified team  
if (nrow(team\_performance\_filtered) > 0) {  
 ggplot(team\_performance\_filtered, aes(x = Season\_End\_Year, y = AvgGoals)) +  
 geom\_line() +  
 geom\_point() +  
 labs(x = "Season End Year", y = "Average Goals per Match", title = paste("Average Goals per Match for", team)) +  
 theme\_minimal()  
} else {  
 print(paste("No data found for", team))  
}

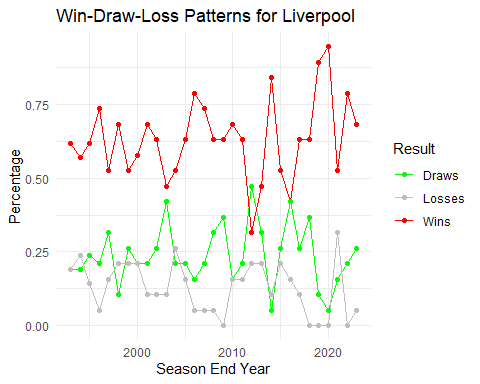


# Win-Draw-Loss Patterns for Liverpool FC

# Calculate win, draw, loss percentages for each team  
team\_results <- pl\_subset %>%  
 mutate(Result = case\_when(  
 HomeGoals > AwayGoals ~ "Win",  
 HomeGoals < AwayGoals ~ "Loss",  
 TRUE ~ "Draw"  
 )) %>%  
 group\_by(Season\_End\_Year, Team = Home) %>%  
 summarise(  
 Wins = sum(Result == "Win") / n(),  
 Draws = sum(Result == "Draw") / n(),  
 Losses = sum(Result == "Loss") / n()  
 )

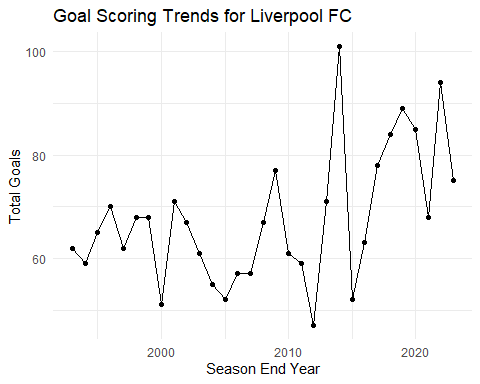
## `summarise()` has grouped output by 'Season\_End\_Year'. You can override using  
## the `.groups` argument.

team <- "Liverpool" # Replace with the team you want to analyze  
team\_results\_filtered <- team\_results %>%  
 filter(Team == team) %>%  
 pivot\_longer(cols = c(Wins, Draws, Losses), names\_to = "Result", values\_to = "Percentage")  
  
ggplot(team\_results\_filtered, aes(x = Season\_End\_Year, y = Percentage, color = Result)) +  
 geom\_line() +  
 geom\_point() +  
 labs(x = "Season End Year", y = "Percentage", color = "Result", title = paste("Win-Draw-Loss Patterns for", team)) +  
 scale\_color\_manual(values = c("green", "gray", "red")) +  
 theme\_minimal()



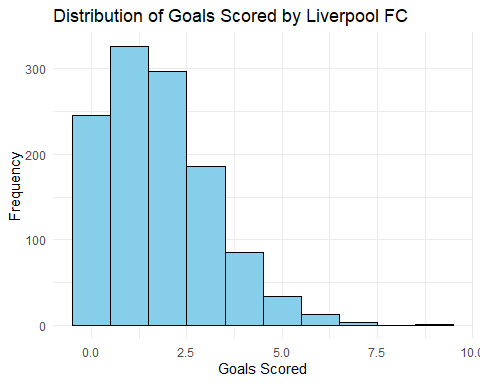
# Goal Scoring Trends for Liverpool FC

# Calculate total goals scored by Liverpool FC per season  
goals\_liverpool <- pl\_subset %>%  
 filter(Home == "Liverpool" | Away == "Liverpool") %>%  
 group\_by(Season\_End\_Year) %>%  
 summarise(TotalGoals = sum(HomeGoals[Home == "Liverpool"]) + sum(AwayGoals[Away == "Liverpool"]))  
  
# Plotting goal scoring trends for Liverpool FC  
ggplot(goals\_liverpool, aes(x = Season\_End\_Year, y = TotalGoals)) +  
 geom\_line() +  
 geom\_point() +  
 labs(x = "Season End Year", y = "Total Goals", title = "Goal Scoring Trends for Liverpool FC") +  
 theme\_minimal()



# Distribution of Goals Scored by Liverpool FC

# Filter data for matches involving Liverpool FC  
liverpool\_goals <- pl\_subset %>%  
 filter(Home == "Liverpool" | Away == "Liverpool") %>%  
 mutate(TotalGoals = ifelse(Home == "Liverpool", HomeGoals, AwayGoals))  
  
# Plotting distribution of goals scored by Liverpool FC  
ggplot(liverpool\_goals, aes(x = TotalGoals)) +  
 geom\_histogram(binwidth = 1, fill = "skyblue", color = "black") +  
 labs(x = "Goals Scored", y = "Frequency", title = "Distribution of Goals Scored by Liverpool FC") +  
 theme\_minimal()

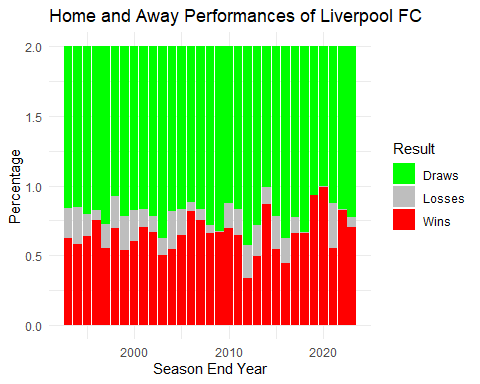


# Home and Away Performances of Liverpool FC

# Calculate home and away performance metrics for Liverpool FC  
home\_away\_performance <- pl\_subset %>%  
 mutate(Result = case\_when(  
 Home == "Liverpool" & HomeGoals > AwayGoals ~ "Win",  
 Away == "Liverpool" & AwayGoals > HomeGoals ~ "Win",  
 Home == "Liverpool" & HomeGoals < AwayGoals ~ "Loss",  
 Away == "Liverpool" & AwayGoals < HomeGoals ~ "Loss",  
 TRUE ~ "Draw"  
 )) %>%  
 group\_by(Season\_End\_Year, Venue = ifelse(Home == "Liverpool", "Home", "Away")) %>%  
 summarise(  
 Wins = sum(Result == "Win") / n(),  
 Draws = sum(Result == "Draw") / n(),  
 Losses = sum(Result == "Loss") / n()  
 )

## `summarise()` has grouped output by 'Season\_End\_Year'. You can override using  
## the `.groups` argument.

# Reshape data for stacked bar chart  
home\_away\_performance\_long <- home\_away\_performance %>%  
 pivot\_longer(cols = c(Wins, Draws, Losses), names\_to = "Result", values\_to = "Percentage")  
  
# Plotting home and away performances of Liverpool FC (stacked bar chart)  
ggplot(home\_away\_performance\_long, aes(x = Season\_End\_Year, y = Percentage, fill = Result)) +  
 geom\_bar(stat = "identity", position = "stack") +  
 labs(x = "Season End Year", y = "Percentage", fill = "Result", title = "Home and Away Performances of Liverpool FC") +  
 theme\_minimal() +  
 scale\_fill\_manual(values = c("green", "gray", "red"))



# Predictive Model for Goals Scored by Liverpool FC

# Example: Building a linear regression model to predict goals scored by Liverpool FC  
  
# Combine home and away matches into a single dataset  
liverpool\_goals <- pl\_subset %>%  
 mutate(GoalsScored = ifelse(Home == "Liverpool", HomeGoals, AwayGoals),  
 Opponent\_strength = ifelse(Home == "Liverpool", AwayGoals, HomeGoals)) # Hypothetical opponent strength example  
  
# Build a linear regression model to predict goals scored by Liverpool FC  
goal\_model <- lm(GoalsScored ~ Opponent\_strength + as.factor(Home), data = liverpool\_goals)  
  
# Summary of the regression model  
summary(goal\_model)

##   
## Call:  
## lm(formula = GoalsScored ~ Opponent\_strength + as.factor(Home),   
## data = liverpool\_goals)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.1562 -1.0179 -0.1382 0.7314 7.7310   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.886626 0.049916 17.763 < 2e-16 \*\*\*  
## Opponent\_strength -0.011220 0.008412 -1.334 0.182302   
## as.factor(Home)Aston Villa 0.239851 0.068324 3.511 0.000449 \*\*\*  
## as.factor(Home)Barnsley 0.970242 0.266726 3.638 0.000276 \*\*\*  
## as.factor(Home)Birmingham City 0.179672 0.109957 1.634 0.102282   
## as.factor(Home)Blackburn 0.225773 0.077315 2.920 0.003505 \*\*   
## as.factor(Home)Blackpool 1.078458 0.266685 4.044 5.29e-05 \*\*\*  
## as.factor(Home)Bolton 0.395071 0.086829 4.550 5.42e-06 \*\*\*  
## as.factor(Home)Bournemouth 0.672587 0.117114 5.743 9.53e-09 \*\*\*  
## as.factor(Home)Bradford City 0.653272 0.191575 3.410 0.000652 \*\*\*  
## as.factor(Home)Brentford 0.156520 0.191503 0.817 0.413760   
## as.factor(Home)Brighton 0.407953 0.117171 3.482 0.000500 \*\*\*  
## as.factor(Home)Burnley 0.408153 0.104308 3.913 9.17e-05 \*\*\*  
## as.factor(Home)Cardiff City 1.046533 0.191618 5.462 4.81e-08 \*\*\*  
## as.factor(Home)Charlton Ath 0.430695 0.104165 4.135 3.58e-05 \*\*\*  
## as.factor(Home)Chelsea -0.018051 0.066345 -0.272 0.785563   
## as.factor(Home)Coventry City 0.251550 0.098191 2.562 0.010424 \*   
## as.factor(Home)Crystal Palace 0.406990 0.084337 4.826 1.41e-06 \*\*\*  
## as.factor(Home)Derby County 0.457446 0.109977 4.159 3.21e-05 \*\*\*  
## as.factor(Home)Everton 0.251297 0.066483 3.780 0.000158 \*\*\*  
## as.factor(Home)Fulham 0.418017 0.080867 5.169 2.39e-07 \*\*\*  
## as.factor(Home)Huddersfield 0.594735 0.191785 3.101 0.001933 \*\*   
## as.factor(Home)Hull City 0.652327 0.126646 5.151 2.64e-07 \*\*\*  
## as.factor(Home)Ipswich Town 0.384686 0.123318 3.119 0.001816 \*\*   
## as.factor(Home)Leeds United 0.253665 0.081981 3.094 0.001978 \*\*   
## as.factor(Home)Leicester City 0.409186 0.079094 5.173 2.34e-07 \*\*\*  
## as.factor(Home)Liverpool 1.142499 0.067140 17.017 < 2e-16 \*\*\*  
## as.factor(Home)Manchester City 0.086195 0.069418 1.242 0.214374   
## as.factor(Home)Manchester Utd -0.126920 0.066341 -1.913 0.055752 .   
## as.factor(Home)Middlesbrough 0.365672 0.082419 4.437 9.21e-06 \*\*\*  
## as.factor(Home)Newcastle Utd 0.260254 0.068227 3.815 0.000137 \*\*\*  
## as.factor(Home)Norwich City 0.550582 0.094477 5.828 5.76e-09 \*\*\*  
## as.factor(Home)Nott'ham Forest 0.330266 0.115515 2.859 0.004256 \*\*   
## as.factor(Home)Oldham Athletic 0.631273 0.182717 3.455 0.000552 \*\*\*  
## as.factor(Home)Portsmouth 0.324385 0.109880 2.952 0.003161 \*\*   
## as.factor(Home)QPR 0.495940 0.107924 4.595 4.37e-06 \*\*\*  
## as.factor(Home)Reading 0.495771 0.158784 3.122 0.001799 \*\*   
## as.factor(Home)Sheffield Utd 0.187240 0.124401 1.505 0.132315   
## as.factor(Home)Sheffield Weds 0.313535 0.102507 3.059 0.002228 \*\*   
## as.factor(Home)Southampton 0.382334 0.071148 5.374 7.86e-08 \*\*\*  
## as.factor(Home)Stoke City 0.249072 0.095541 2.607 0.009146 \*\*   
## as.factor(Home)Sunderland 0.346170 0.081039 4.272 1.96e-05 \*\*\*  
## as.factor(Home)Swansea City 0.406670 0.109896 3.701 0.000216 \*\*\*  
## as.factor(Home)Swindon Town 1.269588 0.254153 4.995 5.96e-07 \*\*\*  
## as.factor(Home)Tottenham 0.213507 0.066384 3.216 0.001302 \*\*   
## as.factor(Home)Watford 0.679292 0.104231 6.517 7.45e-11 \*\*\*  
## as.factor(Home)West Brom 0.588313 0.086891 6.771 1.34e-11 \*\*\*  
## as.factor(Home)West Ham 0.404442 0.068959 5.865 4.61e-09 \*\*\*  
## as.factor(Home)Wigan Athletic 0.540323 0.104274 5.182 2.23e-07 \*\*\*  
## as.factor(Home)Wimbledon 0.382019 0.102549 3.725 0.000196 \*\*\*  
## as.factor(Home)Wolves 0.530006 0.099543 5.324 1.03e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.144 on 11975 degrees of freedom  
## Multiple R-squared: 0.05495, Adjusted R-squared: 0.05101   
## F-statistic: 13.93 on 50 and 11975 DF, p-value: < 2.2e-16

# Predicting goals scored by Liverpool FC  
# Example prediction:  
new\_data <- data.frame(Opponent\_strength = 1.5, Home = "Liverpool") # Adjust opponent strength and match location as per data  
predicted\_goals <- predict(goal\_model, newdata = new\_data)  
predicted\_goals

## 1   
## 2.012295

# Interpretation of Predictive Model Results

The results from the linear regression model provide valuable insights into predicting goals scored by Liverpool FC in the Premier League. The analysis reveals several key findings. Firstly, the coefficient for Opponent\_strength suggests that while there is a slight negative trend (-0.0112), indicating that stronger opponents may marginally reduce Liverpool’s expected goals, this relationship is not statistically significant (p-value = 0.182). This implies that opponent strength alone may not strongly influence Liverpool’s goal-scoring performance.

Secondly, the model incorporates the impact of playing at home (as.factor(Home)). Liverpool’s coefficient (1.1425) shows a significant positive effect when playing at their home ground compared to other teams. This suggests a clear advantage in goal-scoring when Liverpool plays at home, which aligns with common perceptions of home-field advantage in football.

Assessing the model’s overall performance, the Multiple R-squared (0.05495) and Adjusted R-squared (0.05101) indicate that the model explains approximately 5.5% of the variance in goals scored by Liverpool FC. While this demonstrates some predictive capability, it also highlights that other factors beyond opponent strength and home advantage play significant roles in determining match outcomes and goal-scoring patterns.

As an example, for a hypothetical scenario where the opponent’s strength is 1.5 and Liverpool plays at home, the model predicts they will score approximately 2.01 goals. This prediction showcases the model’s application in forecasting goal-scoring scenarios based on the specified variables.

In conclusion, while the model provides valuable insights into predicting Liverpool’s goal-scoring performance, it underscores the complexity of football dynamics and the multitude of factors influencing match outcomes. Future analyses could explore additional variables to enhance predictive accuracy and further refine understanding of team performance in football.

# Conclusion

This portfolio explored Premier League data using advanced R techniques to understand team performances and match outcomes. We've identified important trends in scoring, analysed home and away performances, and built models to predict future match results.

These insights help us see what makes teams successful in the Premier League. Whether you're a coach, a scout, or a fan, this analysis offers useful information. By using data analytics in football, we can make better decisions and strategies.

As we move forward, combining data analytics with football will continue to improve our understanding of the game. This approach will help us make smarter decisions on and off the field, making the sport even more exciting and strategic.