



2021-2022 Premier League Players Analysis

PROJECT REPORT SUBMITTED
IN FULFILMENT OF THE REQUIREMENTS FOR THE COURSE
STAT 250 – APPLIED STATISTICS

DEPARTMENT OF STATISTICS OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY
Samet Kenar
Bilgehan Aydoğdu
Beyza Demet

JUNE, 2023

Abstract

The Premier League is widely regarded as one of the world's most competitive and thrilling football leagues, attracting a global audience. The 2021-2022 season was no exception, featuring intense battles between top clubs, remarkable individual performances, and surprising upsets. In this project, the performance of the premier league teams is analyzed using the player's statistics that affect the league games and the 2021-22 season.








There are many factors that can affect the game, such as a player's age, position, goals, assists, expected goals, red cards, and playing minutes. These traits are examined to determine whether there is any pattern at all. Using various statistical measurements and viewpoints such as comparison of means, proportions, the mean significance by z-test, simple-multiple linear regression, and variance analysis, these traits are examined to determine the statistical significance of team performance. The findings show some significant relationships between team performance and players' statistics.

Introduction

The Premier League contains 20 teams that compete in a round-robin format, playing each other twice (home and away) over the course of a season. This results in a total of 38 matches per team. For the standings, teams earn three points for a win, one point for a draw, and zero points for a loss. The points accumulated throughout the season determine the team's position in the league table. To win a match in the Premier League, a team must score more goals than their opponent within the allotted 90 minutes of regular play time.

Throughout the 2021-2022 campaign, the Premier League showcased intense competition among twenty top-tier English clubs, each competing for victory and trying to maintain their position in the league table. The campaign was contentious until its last week and went head-to-head between Manchester City and Liverpool until the last week. In the last week, it was Manchester City that declared its championship.

Beyond the excitement of the matches, there is a wealth of statistical information that offers important insights into team and player performances. For all the teams which compete in premier league, performance on the field depends on the players' influence on the match. To measure these influence on team performance some of the statistical methods was used such as z-test, t-test, proportions, ANOVA, simple and multiple linear regression.

PREMIER LEAGUE			PTS
1		Man City	93
2		Liverpool	92
3		Chelsea	74
4		Tottenham	71
5		Arsenal	69
6		Man Utd	58
7		West Ham	56
8		Leicester	52
9		Brighton	51
10		Wolverhampton	51
11		Newcastle	49
12		Crystal Palace	48
13		Brentford	46
14		Aston Villa	45
15		Southampton	40
16		Everton	39
17		Leeds United	38
18		Burnley	35
19		Watford	23
20		Norwich	22

(NBC Sports, 2022)

Data Description

The data set, “Football Players Stats (Premier League 2021-2022)” is utilized in the project. This dataset can be reached from Kaggle.

The dataset has 678 observations and 30 variables, which include 9 discrete, 15 continuous and 4 categorical variables. Statistics such as the teams, matches played, positions, expected goals, and yellow and red cards can be accessed using the data set. Appendix Table-1 provides a complete description of each data variable with definitions. In this research, subsets were formed from this dataset for the research questions, and statistical methodologies and techniques were used. For example, a comment was made by comparing the average expected goal of champion team players and the rest of them.

Research Questions

After observing and understanding the data, seven research questions were formed to acquire better results and solutions. The research questions are given in Appendix Table-2.

We planned to reach the goal of our project by following our research questions, described in the next part.

Aim of the Study

This study is conducted to observe whether the relationship between the individual performance of the players and the overall performance of the team is statistically significant. By conducting this study, if there is a statistical significance, then future studies can be conducted to enhance the performance between players and team.

Methodology/Analysis

Seven different statistical methods used to reach conclusions and make inferences about the research questions in this study.

The first method is one-sample hypothesis testing to estimate the mean. It is utilized in our first research question in Appendix Table-2, which tests whether the average expected goal of champion team players, Manchester City Players, is greater than the other team players in the league. At first, the necessary arrangements and tests related to the claim are carried out using R. Although the population standard deviation is not given, according to Central Limit Theorem (as the sample size increases, sampling distribution follows a normal distribution) indicates, the z-test is used to test our hypothesis. The critical value found in the z-table and test value can be compared, the null hypothesis can be rejected or supported.

The second method is two-sample hypothesis testing to compare the two population means. In this regard, the claim of whether the average expected assists of Manchester City (champion team) and Liverpool (2nd team in the league) are equivalent is investigated. Similarly, the needful adjustments and tests related to the claim are performed via R. The Central Limit Theorem is also used due to the sample size of the groups. The null hypothesis can be rejected or supported by comparing the test score and critical value.

The third method is one-sample hypothesis testing for a population proportion to infer from Premier League to all European Football Leagues. We wish to claim whether more than 50% of European Football League Players' number of minutes played in the 2021-2022 season is greater than 1000. In this respect, the R-Package named dplyr is used to subset and filter the data set. Since the sample size is appropriate for applying the Central Limit Theorem, the critical value obtained by the z-table is compared with the test statistic. This comparison makes the decision rule.

The fourth method is two-sample hypothesis testing for two population proportions to make inferences about understanding the approach of the first and last 10 teams in the league to the defensive position. In this respect, the claim is that the first and last 10 teams' defensive player proportions are compared. The R-Packages named dplyr, magrittr and ggplot2 are used for visualizing, subsetting and calculating the proportions. Therefore, the sample size is suitable for implementing the Central Limit Theorem; the critical value is determined by the z- table

and compared with the test statistic. The decision rule is made by observing the difference between them.

The fifth method is simple linear regression. This method was used to determine the relationship between goals scored and expected goals. Due to the game's dynamics, goalkeepers do not contribute to goals. So, for the analysis, the players who have an influence on goals or expected goals, such as defenders, midfielders and forwards, are selected from the dataset by using R. In this simple linear regression model, the dependent variable was chosen goals scored and the independent variable was chosen expected goals. Then, the simple linear regression model is conducted between these two variables. The assumptions are checked using the QQ Plot, Residuals and Fitted Plot, and Scale-Location. By considering the p-value and level of significance, the significance of the model is measured, and interpretations are made.

The sixth method is multiple linear regression. This method is applied to determine the factors that affect playing minutes. The playing minutes for players are decided by their managers, but managers make their decisions according to the player's effectiveness in the game. For this reason, the multiple linear regression model measures the relationship between minutes played and players' age, non-penalty expected goals and expected assists. As in the linear regression method, only the players who have an effect on goals and assists were selected from the dataset. In this multiple linear regression model, the dependent variable is minutes played, and the independent variables are age, non-penalty expected goals and expected assists. After that, the multiple linear regression model between these variables is done using R.

The seventh method is a one-way ANOVA to test whether the means of getting yellow cards differ according to positions. One-way ANOVA has three assumptions, normality, equality of variances and independence, which must be considered, and these assumptions are also considered by utilizing the QQ Plot and testing the equality of variances.

Results and Findings

In the first question (please refer to Appendix Table-2), the one sample hypothesis testing is applied. The alternative hypothesis is indicated as the average expected goal of champion team players is greater than the rest. The rest of the team players' average expected goal is found as 1.850385, and we have checked whether our null hypothesis is equivalent this value. The test statistic is 2.305909, and the critical value obtained from the z-table is 1.96. We notice that the test value is greater than the critical value obtained from the z-table having a

significance level of 0.05. By looking at the test value and the critical value, the null hypothesis can be rejected, and it can be concluded that there is enough evidence to support the alternative hypothesis that the average expected goal of champion team players is greater than the rest of the team players.

In the second question, the null hypothesis is that the average expected assists of Manchester City, the champion team, is equivalent to the average expected assists of Liverpool, the 2nd team in the league. The test statistic is 0.3193601, and the critical value from the z-table is 1.96. We fail to reject the null hypothesis because test statistic is less than the critical value. Therefore, there is not enough evidence to support the claim that the average expected assists of the champion team is not equal to the 2nd team in the league.

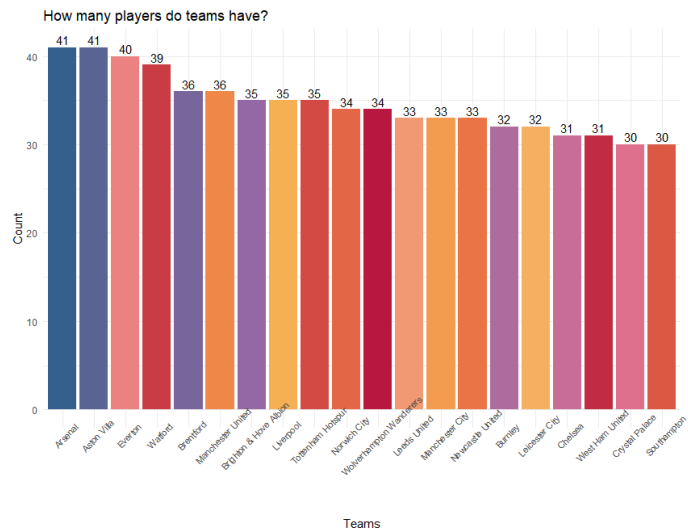


Figure-1

In one-sample hypothesis testing for population proportion, the claim that whether more than 50% of European Football League Players' number of minutes played in the 2021-2022 season is greater than 1000 or not is tested. The data shows that 322 out of 691 players played +1000 minutes in the 2021-2022 season. The test statistic is -2.150898, and the critical value

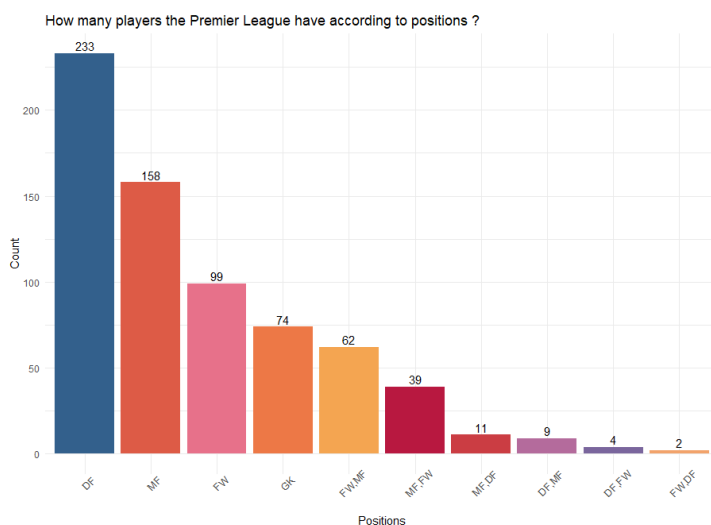


Figure-2

acquired from the z-table is 1.96. By observing the test value and the critical value, it can be stated that from the rejection region of the null hypothesis that there is enough evidence to support the claim that more than 50% of European Football League Player's number of minutes played in 2021-2022 is greater than 1000.

A two-sample hypothesis testing for population proportions is formed for our fourth question. As shown in Figure-2, the

defensive players cover the majority of the league according to position. The claim that whether the first and last 10 teams' defensive players' proportions are equivalent is tested. The test statistic is obtained as 1.020853, and the critical value, according to obtained z-table, is 1.96. By comparing them, we can state that we fail to reject the null hypothesis and say there is insufficient evidence to claim that the first and last 10 teams' defensive players' proportions are equal.

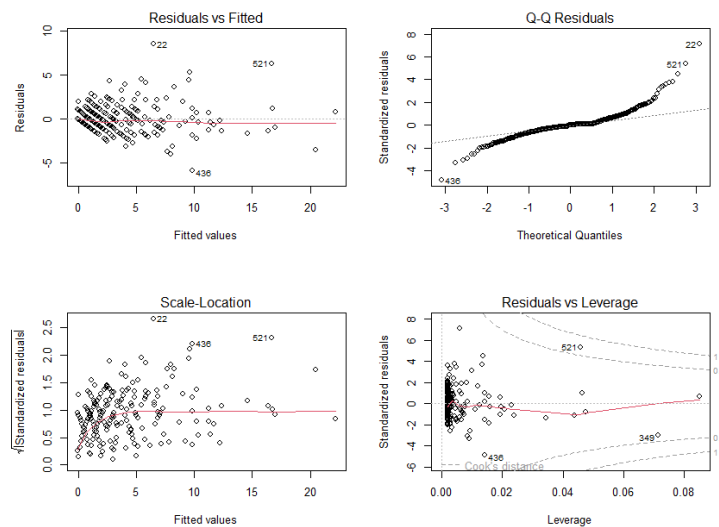


Figure-3

is, as seen in Figure-3, QQ Plot roughly follows the line, a sign of normality. The third assumption is met because, as seen in Figure-3, the residuals versus the fitted plot's red line is almost horizontal, ensures linearity. The scale-location plot in the figure met the homoscedasticity. The red line seems linear and looks like horizontal. Since all the assumptions of simple linear regression for our model (please refer to

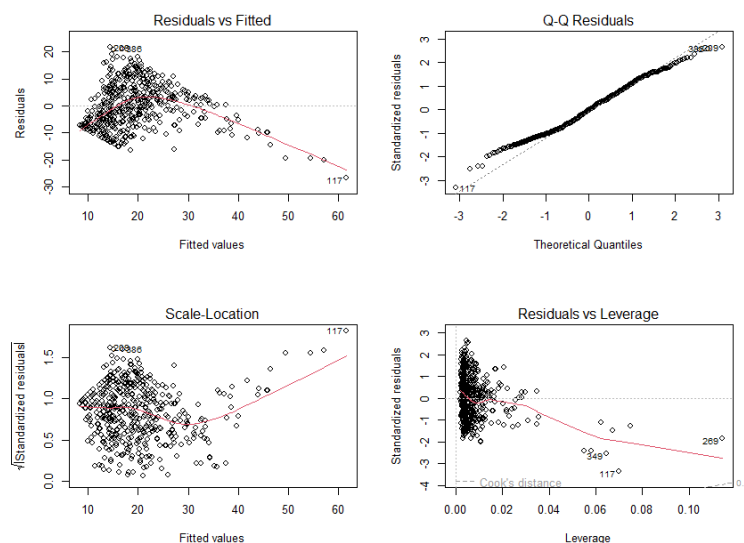


Figure-4

The fifth research question aimed to determine the relationship between goals scored and expected goals. In single linear regression, there are three assumptions. The first assumption is independence which our model satisfies because it is assumed that the data set was gathered using statistically appropriate sampling methods. The second assumption is the normality

is, as seen in Figure-3, QQ Plot roughly follows the line, a sign of normality. The third assumption is met because, as seen in Figure-3, the residuals versus the fitted plot's red line is almost horizontal, ensures linearity. The scale-location plot in the figure met the homoscedasticity. The red line seems linear and looks like horizontal. Since all the assumptions of simple linear regression for our model (please refer to Appendix Figure-6) is appropriate. The 86,95% of the variation in the goals can be explained by our model. Our p-value is $2.2e-16$, it can be observed that the p-value is less than our significance level. Thus, there is a significant relationship between goals and expected goals.

The sixth research question aimed to determine the

relationship between minutes played and players' age, non-penalty expected goals and expected assists. According Figure-4, since the QQ Plot follows an approximately perfect straight line, we can interpret it as our data is distributed normally. Secondly, by figure, the red line looks like curved. The relationship between independent and dependent variables may be linear. Thirdly, as can be seen from the scale-location, the red line is smoothly curved, which can be concluded as there can be some heteroscedasticity. Lastly, it was assumed that there observations are independent from each other. As all assumptions seems suitable, we can comment on the result. The Adjusted R-Squared is 0.462. The 46.2% of variation in the minutes is explained by our model. Our significance level is 0.05. Our p-value is found smaller than $2.2e-16$, smaller than 0.05. We can conclude that non-penalty expected goals, expected assists, and age is essential for minutes played.

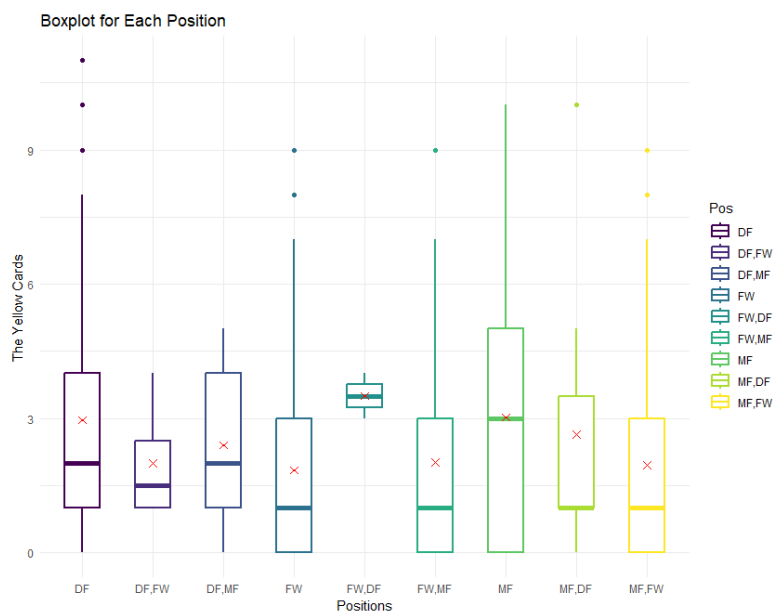


Figure-5

For our last research question, one-way ANOVA tests whether the means of getting yellow cards differ according to the positions. Assumptions were checked. Firstly, we assumed the samples making up the groups according to positions are independent. Secondly, as we can observe from the Appendix Figure-7, our observations in each group are normally

distributed because QQ Plot follows an approximately straight line. Thirdly, a boxplot is created to see if there is really a difference between the means of the groups, which can be seen in Figure-5. Finally, for equality of variances, the result is 4.063138, greater than 2. It may affect the overall effectiveness of detecting an actual difference in mean; however, this does not mean our results will be completely worthless. It is assumed that the assumptions are appropriate, which means we can construct ANOVA. The p-value is 0.0116, smaller than the usual threshold of 0.01; there is a statistical difference among the means of getting yellow cards according to positions. Moreover, it is found that the players who received the most yellow cards are the forwards and defenders.

Discussion/Conclusion

In this study, firstly, the teams with high average expected goals are more likely to win since the average expected goal of the champion team was higher than the average expected goal of the remaining teams. Secondly, it has been revealed that the expected assists by the players of the first and second teams in the league are not very important factors in the championship fight and are not statistically significant. Therefore, other factors are more critical in the difference between the two most successful teams. Thirdly, the number of minutes players play during the season depends on many factors and can vary significantly between different leagues and players. Hobbs (2022) states that the playing time of the players has a significant impact. Generally, if a football season has a league schedule of 38 games and each game lasts 90 minutes, a total of 3,420 minutes will be played. For the Premier League 2021-22 season, it is found that more than 50% of European Football League Player's number of minutes played in 2021-2022 is more significant than 1000. Fourthly, when we evaluate the distribution of positions over the defenders, we observe that the distribution of the number of players according to positions is not statistically significant. Good results may not be achieved by only defending in the league. There is a difference in the league table although there is no significant difference in defender proportion between the first and last ten teams. Thus, setting up a team plan or team by focusing on a single position may not bring successful results. Fifthly, The Premier League 2021-22 season's statistical analysis, reveals a significant linear relationship between goals scored and expected goals (xG). A strong positive correlation between goals scored, and xG suggests that teams that consistently create high-quality scoring chances tend to convert those opportunities into goals. Sixthly, it is observed that non-penalty expected goals, expected assists, and age is essential for playing minutes for players. Managers make their decisions according to what they can get from players. As clubs strive for success, understanding the relationship between minutes played and players' age, non-penalty expected goals and expected assists provide valuable insights for player development, squad management, and strategic decision-making in the Premier League. Seventhly, when the Premier League 2021-22 season is examined, it is concluded that there is a statistical difference among the means of getting yellow cards according to positions. It has been determined that the players who received the most yellow cards are the forwards and defenders. Forwards are in more frequent contact with opposing defenders in the offensive zone and have an aggressive style of play. On the other

hand, Defenders are in more physical contact with opposing forward players. These situations may explain why forwards and defenders are more likely to receive a yellow card.

References

Gowda, O. *Football Player Stats (Premier League 2021-2022)*. Kaggle.

<https://www.kaggle.com/datasets/omkargowda/football-players-stats-premier-league-20212022>

Hobbs, J. (2022, June 06). *The teenagers with the most minutes played in Europe's top*

leagues in 21/22. Stats24.com. <https://www.stats24.com/football/news-the-teenagers-with-the-most-minutes-played-in-europes-top-leagues-in-2122-234>

Appendix

Table – 1

Features	Definitions
Player	Demonstrates the player's name
Team	Demonstrates the player's played club
Nation	Demonstrates the player's nation
Pos	Demonstrates the player's position
Age	Demonstrates the player's age
MP	Demonstrates the player's matches played
Starts	Demonstrates the number of games in which the player started in the first 11
Min	Demonstrates the minutes played
90s	Demonstrates the minutes played divided by 90
Gls	Demonstrates the goals scored or allowed
Ast	Demonstrates the assists made
G-PK	Demonstrates the non-penalty goals
PK	Demonstrates the penalty kicks made
PKatt	Demonstrates the penalty kicks attended
Crdy	Demonstrates the yellow cards
CrdR	Demonstrates the red cards
Gls	Demonstrates the goals scored per 90 minutes

Ast	Demonstrates the assists per 90 mins
G+A	Demonstrates the goals and assists per 90 mins
G-PK	Demonstrates the goals minus penalty kicks made per 90 mins
G+A-PK	Demonstrates the goals plus assists minus penalty kicks made per 90 mins
xG	Demonstrates the expected goals
npG	Demonstrates the non-penalty expected goals
xA	Demonstrates the expected assists
npG+xA	Demonstrates the non-penalty expected goals plus expected assists
xG	Demonstrates the expected goals per 90 mins
npG	Demonstrates the non-penalty expected goals made per 90 mins
xA	Demonstrates the expected assists made per 90 mins
npG+xA	Demonstrates the non-penalty expected goals plus expected assists made per 90 mins

Table – 2

Q#	QUESTIONS
Q1	Is there enough evidence to support the claim that the average expected goal of champion team players greater than the rest of the team players?
Q2	Is there enough evidence to support the claim that the average expected assists of the Manchester City players are not equal to the Liverpool team players?
Q3	A sample of 691 players showed that 322 players minutes played in 2021-2022

season is greater than 1000. Is there enough evidence to claim that more than 50% of the european football league players' in 2021-2022 season played minutes are greater than 1000 at the 0.05 significance level?

Q4 The distribution of the defence players of the first and last 10 teams according to the positions differ from each other. Is there any significant difference?

Q5 Is there a significant relationship between goals (dependent variable) and expected goals (independent variable) at a 0.05 significance level?

Q6 Is there a significant relationship between matches played (dependent variable), non-penalty expected goals(independent variable 1), expected assists (independent variable 2) and age (independent variable 3) at a 0.05 significance level?

Q7 Whether the means of getting yellow cards differ according to the positions.

Figure-6

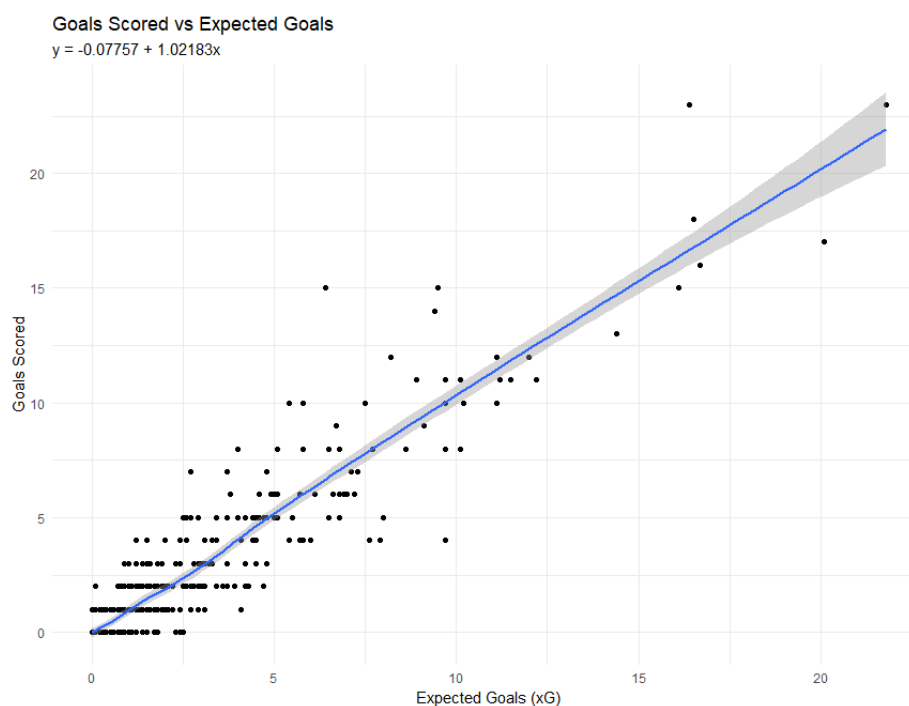


Figure-7

