Analysis on “Football Player Statistics (Premier League from 2021-2022)”

1. **Introduction**

Football Analytics has become more and more obvious these days and has originated since the 18th Century. The main concept of bringing this data is to check the output of Goals and determining the players who perform well, gaining that competitive edge over others that has continued since World War II. Football managers and Directors spend a lot of time and money before they bring a particular player in their squad and build a team who could lead them to the League Cups and Titles there are to win.

Before Scouting a player, they must be aware of the position, type of player, strengths and weaknesses, and the past record of achievements that the player has. To portray a player as Excellent player and add him to the roster, the attributes of a player are analyzed by the desired team managers before they are recruited by other teams and have a good chance in the transfer market to aim to be the Top Clubs in the English Championship.

The ability to make sense of the gathered performance player data to scout the player’s style that also matches the coach, maintains a good relationship with the team enhancing mutual trust and respect among them. The signing of players based on their ratios helps them to discover innovative, counter-intuitive, and winning strategies with the results of the previous matches thereby improving granularity of their overall stats and performance.

The Expected Goals ratio has been one of the most revolutionary metrics to calculate the output of Goals for a team and the probability of scoring them is based on several factors like distance, angle of the shot, weak foot or strong foot, type of attack, direction of shot taken etc. Again, Goals are not the only measure of a players worth since, the one who passes, and the way that pass had been made for the scoring player, who has created the chance for the Goal scorer, is of the highest recognition and prominence. Some players score less Goals but the way they give away the ball so that a Goal is scored is called an Assist and it has the same value as a Goal in Football.

Another important metric that is considered as a key factor in Football is Defense or the ability to win the ball back from the opposition and not allow the other team players to score a goal by becoming the shield other than the Goalkeeper. In a technical aspect, football is generally categorized into Attack, Mid-field, and Defense where the stats of a player in detail can be seen below as follows:

Chart

Description automatically generated with medium confidence

Thus, the **Attack, Passing and Defense** constitute the main factors of a player’s profile. The way of playing of a player matters during analysis which includes **Vision** (seeing the formations and passing the ball, creating chances for more goals), **Dribbling** (using various skills needed to bring the ball into the Goal zone by getting past the defenders), **Heading** (scoring goals with the head and passing the ball using the head), **Crossing** (the ability of passing the ball the from far side to the center above the defenders), **Tackles** (the number of successful times player has won the ball from the opposition), **Cards obtained** (the yellow and red cards that are obtained when a player fouls another), and **Physical condition** ( the stamina and speed of the player to continue playing for longer duration of the game). Having objectives can help speed up the learning processes and create virtuous development cycles, making data analytics a powerful tool in Football to predict, identify and cultivate a players’ potential.

**About the Data Set:**

This dataset contains the Statistics of football players who played in the Premier League from (2021-2022). It has 692 rows and 29 columns consisting of various attributes and contributions of the players in detail.

Y = The Best Premier League Player of the Year 2021-2022.

The 29 Column names listed below are the key variables (X's) to conclude the Y variable and are Abbreviated as follows:

Player: Player's name.  
Team: Played club during 2021-2020.  
Nation: Player's nation.  
Pos: Position that one plays in.  
Age: Player's age.  
MP: Matches played.  
Starts: Matches started in the playing 11.  
Min: Minutes played.  
90s: Minutes played divided by 90.  
Gls: Goals scored or allowed.  
Ast: Assists.  
G-PK: Non-Penalty Goals.  
PK: Penalty Kicks made.  
PKatt: Penalty Kicks attended.  
CrdY: Yellow Cards.  
CrdR: Red Cards.  
Gls 90: Goals scored per 90 mins.  
Ast 90: Assists per 90 mins.  
G+A 90: Goals and Assists per 90 mins.  
G-PK 90: Goals minus Penalty Kicks made per 90 mins.  
G+A-PK 90: Goals plus Assists minus Penalty Kicks made per 90 mins.  
xG: Expected Goals.  
npxG: Non-Penalty Expected Goals.  
xA: Expected Assists.  
npxG+xA: Non-Penalty Expected Goals plus Expected Assists.  
xG 90: Expected Goals per 90 mins.  
npxG 90: Non-Penalty Expected Goals made per 90 mins.  
xA 90: Expected Assists made per 90 mins.  
npxG+xA 90: Non-Penalty Expected Goals plus Expected Assists made per 90 mins. Link: <https://www.kaggle.com/datasets/omkargowda/football-players-stats-premier-league-20212022>.

**Data Visualization:**

Let us start with structure and description of our variables as shown in the diagram below:

Text

Description automatically generated

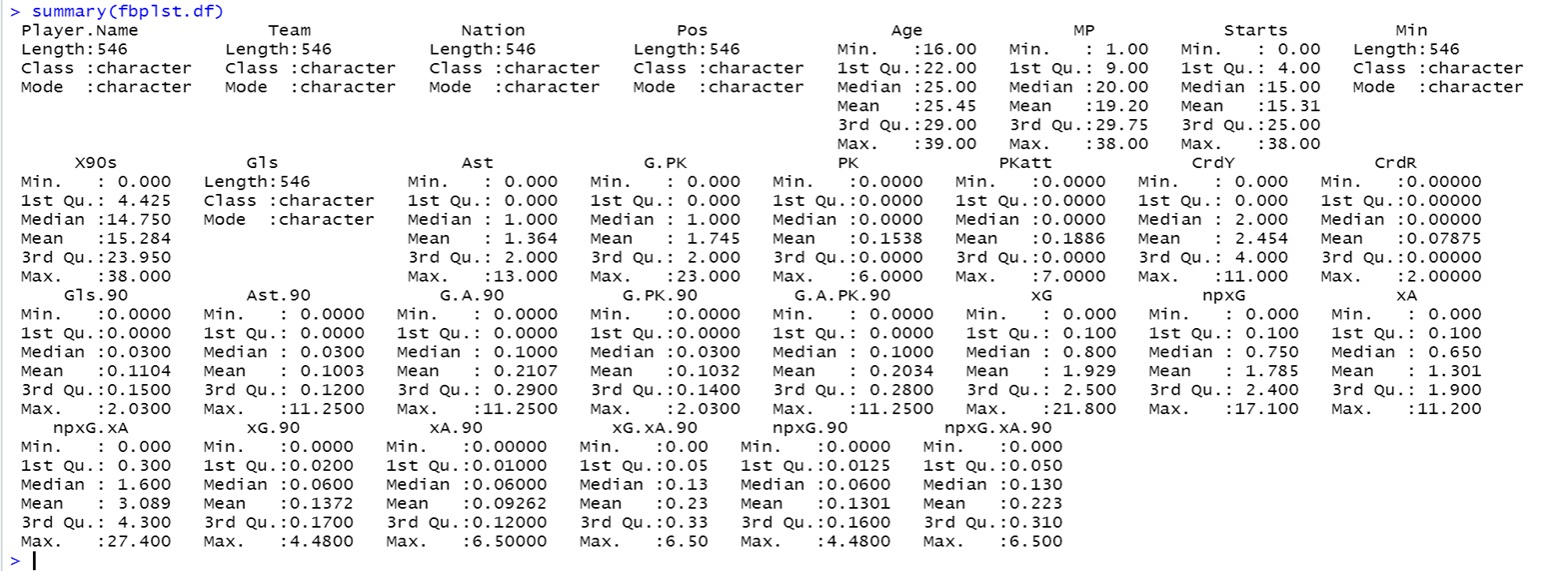
The Data frame has 695 Observations 30 Variables(columns) where only 6 of them are character based while the rest of the columns only has integer or numeric values in it. The Player name is the column where our Y is, and the numerical values will describe the player as discussed above in Introduction.

Now, removal of NULL values and empty rows of the extra players is done from the footballplst.df to minimize the difference in the summary statistics in the new data frame i.e., fbplst.df where we get only 546 observations when compared to the 695 in the previous one. The next step is the Generation of the summary statistics of the data frame and drawing various graphs and plots based on the derived data.

Table

Description automatically generated

As seen in the figure, there is a elimination of 149 variables using the command na.omit(footballplst.df). Now, summary statistics is generated and then we see the mean, variance, min-max values, and the quarters of each column in brief.



According to the summary statistics, the minimum of all the variables is “0” while the maximum ranges from 2 – 38 depending on the factors and quality of the player. There is also a difference between the normal Goals variable and Goals per 90 minutes and the same applies to all other variables as well since a player may or may not play an entire 90-minute match and sometimes he may be substituted only for a particular time. It counts as a major variable as the player who has scored more goals in 90-minute matches is still less superior to the one who has scored the same number of goals playing half the time. Another factor that decides whether a player is good or not depends on the non-Penalty goals and the ability to score from anywhere.

It means that the Player with more Penalty Goals in his sheet (Penalty – short ranged) is not worth than the one who has same number of non-Penalty goals (Out of the box and others).

The data set consists of the Premier League players and their stats from 2021 to 2022 thus, Time series plot is out of scope. Major plots like Bar charts, Scatter plots, Histograms, Box plots, and Parallel Coordinate plots are drawn with the variables below:

Chart

Description automatically generated

As seen in the bar chart above, the Top players to score the most Goals in the season 2021-22 are ‘Son Heung min’ and ‘Mohamed Salah’ with 22 Goals each. ‘Cristiano Ronaldo’ has scored 18 Goals and is the second highest to make this feat while ‘Harry Kane’ is the third highest scorer with 16 Goals. The Golden boot Award for the most Goals in the Premier League will be shared by the Top 2 players and the rest are eligible for a consolation prize.

Chart, bar chart

Description automatically generated

Similarly, in the barchart above, the players with the greatest number of Assists are plotted where ‘Mohamed Salah’ takes the crown again. The second place is occupied by ‘Trent Alexander Arnold’ who is the Defensive player and a teammate of the Assist crown Salah.

Now, we plot a basic Scatter plot showing Goals of a player(with Index as point ) vs Matches Played to get a brief idea if the Player has played all the matches to score more.

Chart, scatter chart

Description automatically generated

The first plot shows us that the Goals scored are at a maximum of 25 when the players have played all 36 or 38 matches, they have during a single season. Most of them have Goals scored from 0 to 10 and get a lot dispersed since a few players are only capable of scoring 15 or more Goals in a year. The number of players who scored 0 to 5 are so many that the points are not clear.

Thus, using ggplot2 may give us the edge required to plot the dots more clearly and considering the Assists to show the Indices of the Player who gave more Assists as well. As seen in the diagram below, the player with most Assists has light colored dots and tend to be on the right side of the plot. The darkest blue plot indicates players with no assists and the deduction can be made that Players who have scored more Goals have played a minimum of 30 Matches in a single season.

Chart, scatter chart

Description automatically generated

Now, moving on to the Penalty Kicks and Goals, the next plot shows the Position of the player based on the Index vs the Minutes that he has played per game to score it. The Final dot shows a Mid Fielder scoring 6 penalty goals while the green color shows the Forwards who has scored 5 penalties each of them spending approximately 35 minutes for each goal. The blurry bar at 0 shows the players who have not scored or attempted the penalty out of the 540 remaining players.

Chart, scatter chart

Description automatically generated

Let us now make a box plot and then explore the various options that we can tell from it.

Chart, box and whisker chart

Description automatically generated

This box plot tells us that the Forwards has scored the most Goals followed by Midfielders and lastly Defenders. The Players who also play on 2 positions are also up to the mark competing the Forward Strikers but still a little less when checked with Goals.

A violin plot has been created considering the same features and points which tell us in a detailed when we compare.

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

In the above figure, the side-by-side box plots show the quality of the player at that Index and the median range and value based on the factors.

First one tells us that Expected Goals are more and lie in the outlier region than the Non-Penalty Goals in turn indicating that there could be more chances of scoring a Goal by the Player and that they missed.

Second one gives us the Avg age of the player who has the most goals and it ranges from 22 to 32.

Third one tells us about the yellow cards that are given to a Player which has increased as the Goals have with due no. of Matches and Assists.

The last plot tells us that Penalty Goals are not that important and are only smaller in value since they lie in the outliers and most Goals were not scored from Penalties.

**Summary and Conclusion:**

Thus, after all these graphs and plots, the result is clear on who the Best Player is, and a lot of factors must be considered before Analyzing the Player with most Potential. Mohamed Salah was the best player with most Goals, Assists, and Non-Penalty Goals scoring from a wide range of angles and supporting his teammates.

There are still a lot of options to explore on what else can be considered to see the Best in the Player.