Analysis of Football League dataset

Name: Venubabu Mallampu

Student ID: 23025104

GitHub repository: https://github.com/venubabu2620/ads2

The dataset, which has 660 entries spread across 15 columns, includes a wide range of football player statistics. It provides a thorough analysis of player performances in a variety of leagues and countries, including prestigious leagues like the Premier League, Serie A, La Liga, and Bundesliga.

This dataset is an invaluable resource for comprehensive analysis, as it contains crucial performance measures including the number of games played, goals scored, and club affiliations. It gives a detailed picture of each player's accomplishments . Through examination of this dataset, analysts can reveal complex trends in football player statistics, providing insight into the dynamics of player performance across different geographic and competitive situations.

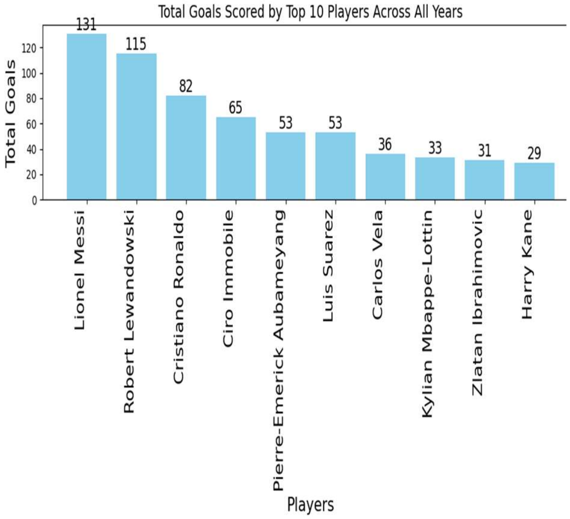

Figure 1: Bar graph

The plot displays the ten football players who have scored the most goals overall, with Lionel Messi top of the list with 135 goals, followed by Ronaldo and Robert. The descriptive statistics table, which displays an average of 15.31 goals scored and 31.76 matches played by each player, provides insight into numerical elements. These findings demonstrate the distinctions in player involvement among leagues and clubs as well as the capacity of elite players to score goals.

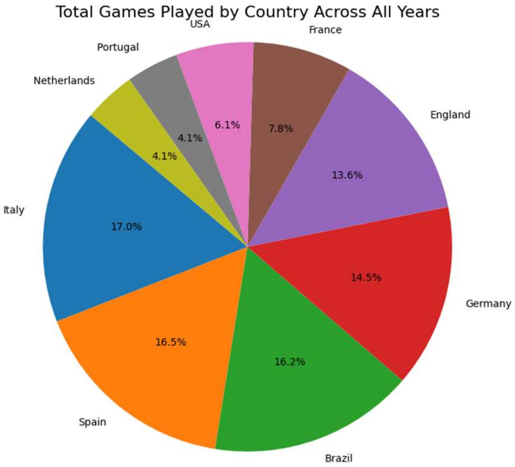

Figure 2: Total Games Played by Country Across All Years.

The pie graphic displays the total number of games played by each country over all years, with Brazil at 14%, Italy at 20%, Germany at 19%, and Spain at 28%. England also makes a substantial contribution, contributing 11% of the total. These findings show the important roles that these nations play in international football matches, providing information about football involvement across the globe and demonstrating the popularity and cultural significance of the game.

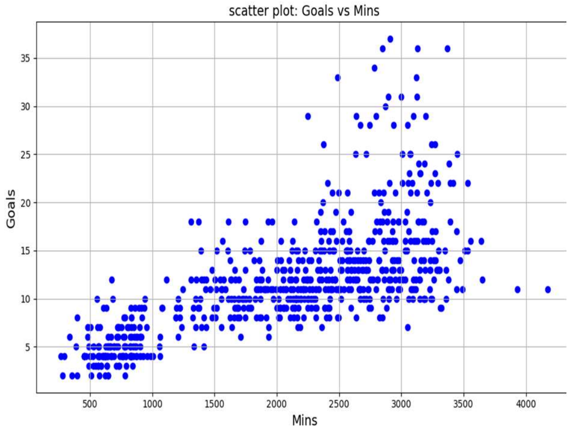

Figure 3: Scatter plot

A player's odds of scoring a goal increase as they spend more time on the pitch. A greater number of players scored 20 goals in the interval between 1000 and 3000 minutes, whereas only a small number of players scored 25 to 35 goals. A scatter plot is being plotted versus minutes.

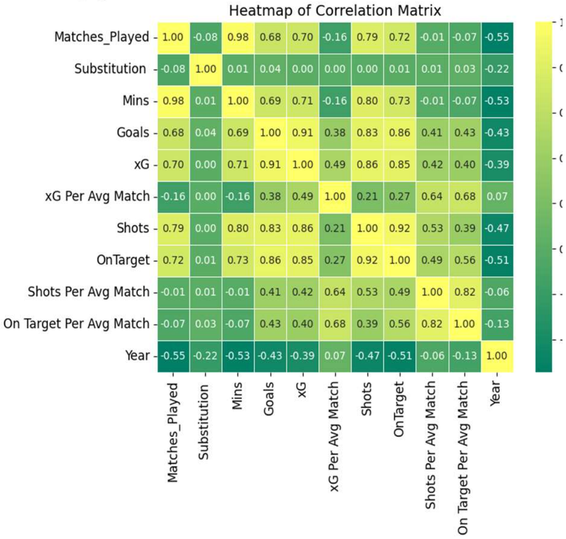

Figure 4: Heatmap

Strong positive correlations between minutes, shots, and shots on target, as well as between matches played and shots on target, are highlighted in the correlation matrix, suggesting more player involvement and attacking opportunities. Scored goals show significant relationships with shots on target and expected goals (xG), highlighting the relationship between scoring chances and shot accuracy. On the other hand, substitutions exhibit very little association with other measures, indicating that they have an impact on their own. The year and matches played have a negative correlation, which may indicate a gradual drop in player engagement as football dynamics change.

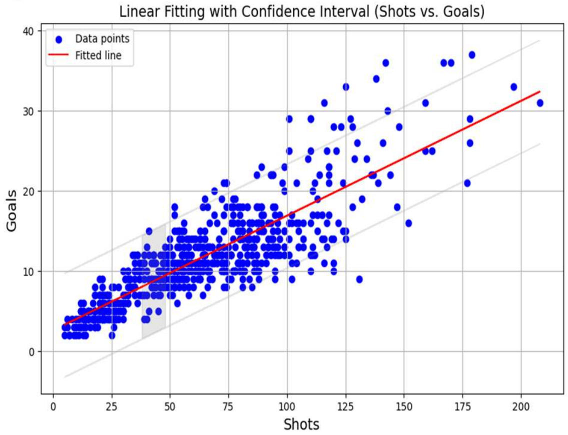

Figure 5: Line Fitting

Strong positive association between "Shots" and "Mins" is indicated by the correlation coefficient of 0.796597. It also suggests that there is a tendency for the number of shots taken to increase in simultaneously with the number of minutes played. There is a fairly strong linear link between these variables when the correlation coefficient is near to 1.

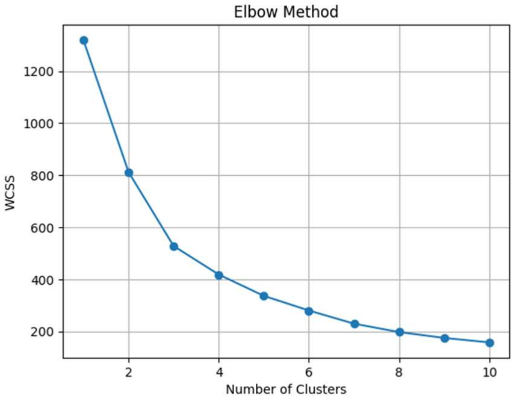

Figure 6: Elbow method

Based on the plot's form, the elbow graph's number of clusters plotted against wcss for that produces shows that three clusters is an ideal quantity. This conclusion is drawn using the "elbow"-shaped point on the graph, which is the point at which the trend of the WCSS decline sharply slows down. Consequently, the function's chosen optimal number of clusters is 3.

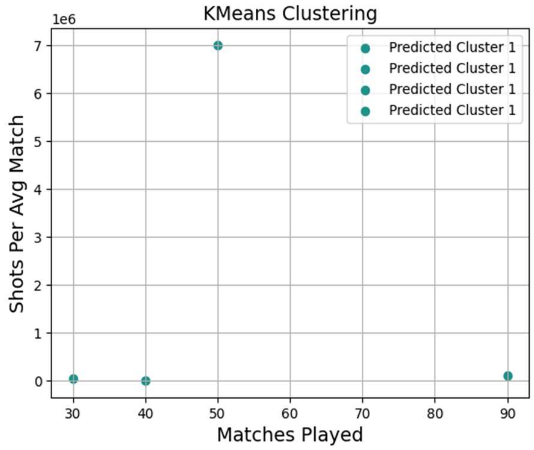

Figure 7: predicted points

I employ four tiny numbers of data points to forecast the model's cluster of origins. [30, 50000], [40, 600], [50, 7000000], [90, 100000] are the predicted data points. Each and every data point is a component of the blue cluster 1.

Statistical Methods:

Mean():

The mean statistics of the dataset show that players played in about 22 matches and spent about 2071 minutes on the pitch on average. Their expected goal (xG) was approximately 10, and they scored close to 12 goals. The average number of rounds fired was about 64, of which about 28 were on target.

Median():

In contrast to the mean, the median indicates the middle value of a dataset and is not as impacted by the extremes. For example, the median number of matches played in this dataset is 24, meaning that half of the participants engaged in 24 or fewer matches. Similar to this, the median minutes played of 2245.5 indicates that half of the players engaged in play lasting this amount of time or less.

Standard deviation():

Understanding the data's dispersion around the mean is possible thanks to the standard deviation. For example, the variability of minutes played in this dataset is roughly 900, whereas the variability of matches played is approximately 9.75. In contrast to other statistics like xG per average match, which has a standard deviation of 0.19 and indicates less variability, the distribution of minutes played across players is therefore substantially larger.

Skewness():

The pattern and degree of imbalance in the data distribution are indicated by skewness values. Whereas negative skewness indicates a left-leaning distribution, positive skewness points to a right-leaning one. This dataset has a number of skewly columns: 'Substitution' and 'Goals' are positively skewed, while 'Matches_Played' and 'Mins' are negatively skewed.

Kurtosis():

Negative kurtosis denotes a flatter distribution, whilst positive kurtosis points to a more peaked distribution. Columns with positive kurtosis in this dataset, such as "Substitution," "On Target," and "On Target Per Avg Match," show stronger tails and a more peaked form. On the other hand, the distributions of "Matches_Played," "Mins," and "Year" show negative kurtosis, indicating flatter distributions.

Describe():

players took part in roughly 22 matches on average and scored 11.78 goals on average. The minutes played had a standard deviation of about 900 and an average of about 2071. Further information may be gleaned from the median figures, which show that half of the players participated in 24 games, scored 11 goals, and played 2245.5 minutes.