CASE STUDY 4:
SOCCER PLAYERS ANALYSIS

Team - Soccer Fanatics

Kartik Nautiyal - knautiyal@wpi.edu
Riddhi Thakkar - rathakkar@wpi.edu
Kratika Shetty - kshetty2@wpi.edu
Ashay Aglawe - alaglawe@wpi.edu

Introduction:

There are various soccer leagues being played around the world but Premier League is the most popular and viewed soccer league in the world. It is broadcasted in 212 territories to 643 million homes and a potential TV audience of 4.7 billion people. Sports companies have become a huge part of the soccer industry. Sports sponsorship is a huge business. Companies are ready to invest hundreds of millions of dollars in supporting clubs, leagues, events and athletes from their marketing budget to ensure that their brand is successful. From soccer shoes to coffee mugs, sports companies are selling a wide variety of products by keeping soccer players in their minds. This helps companies to drive sales and increase brand exposure. Some sports companies are specifically signing contracts with soccer players so as to have exclusive product endorsement from the player. Well established players certainly provide a decent option to these companies but can they cut down on the cost of signing a well established player while at the same time achieving their target? This is an attempt at an analytical approach to find the most optimum player who could potentially give the maximum returns on their investment and answer the question with a data centric approach.

Target Audience and Motivation:

The target audience for this project includes sports equipment selling companies like Nike who sign players every season for endorsement of their products. It is extremely crucial for the company to choose the right person for this purpose because endorsements drive a very high number of sales for the company.

By helping the company identify the right player for the job we can help the company save a lot of money. By the right person we mean a player who is not the obvious choice. A very popular player may attract a huge signing fee which will end up eating the profits of the company. Instead we intend to propose a player who is new to the league and is not very established already. Our idea is to bet on the potential of the player so as to tap on to the profits by signing a player with a low signing fee and drive maximum product awareness.

Being soccer fans ourselves and following the premier league, we really wanted to pick up this topic and see what we could do with it. We also wanted to know how data analysis could be applied in the sports industry.

Dataset:

We have used the 20/21 Premier League Season dataset. This was selected keeping in mind that the project should be relevant in the recent scenario. The 20/21 season is the latest completed season of the Premier League. Following is the table which shows all the features in the dataset:

FEATURES	**DESCRIPTION**
Name	Name of the player
Club	Club name of the player
Nationality	Nationality of the player
Position	Position that the players play in
Age	Age of the player
Matches	Number of matches played
Starts	Number of matches started
Mins	Number of minutes played
Goals	Number of goals scored
Assists	Number of assists provided
Passes Attempted	Number of passes attempted
Percent Passes Completed	Percentage of successful passes
Penalty Goals	Penalty goals scored by the player
Penalty Attempted	No. of penalties attempted by player
xG	Expected Goals
xA	Expected Assists
Yellow Cards	Number of Yellow cards issued to the player
Red Cards	Number of Red cards issued to the player

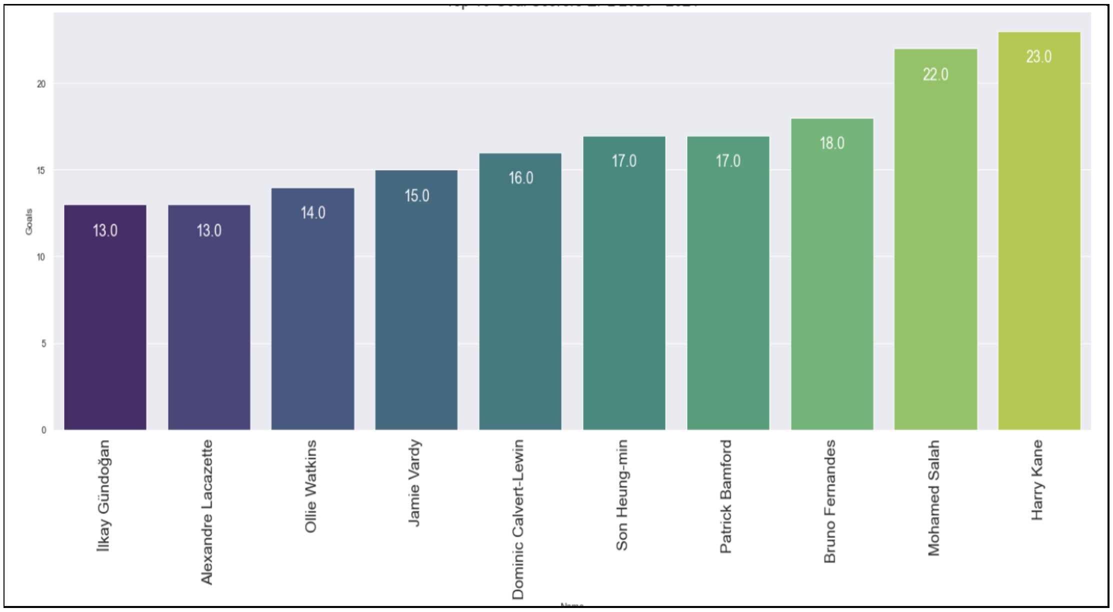
Data Preprocessing:

This dataset was a clean dataset and existed in the form of a CSV. We were able to download and import it directly into Python notebooks.

Exploratory Data Analysis:

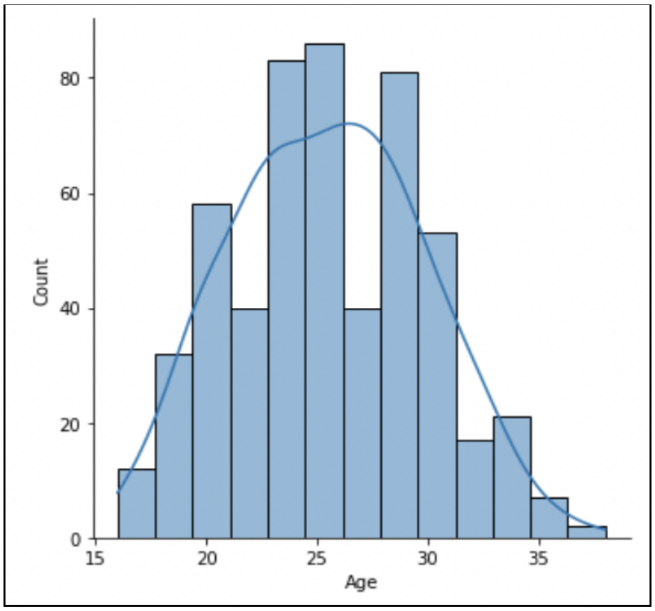
Going through a few variables, data was explored to find trends and get a summary of the dataset. Following are a few facts which were found out:

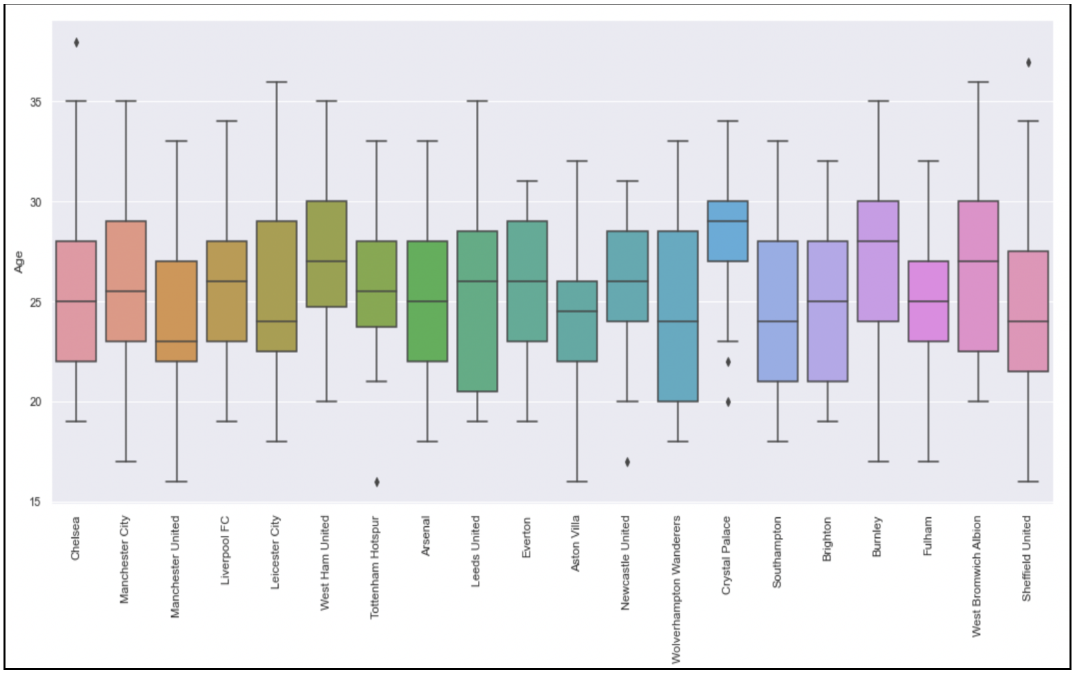
1. Top 10 Goal scorers in the 2020-2021 season:



From this plot we can see that Harry Kane scored the most number of goals followed by Mohamed Salah & Bruno Fernandes.

2. Age Distribution of Players:

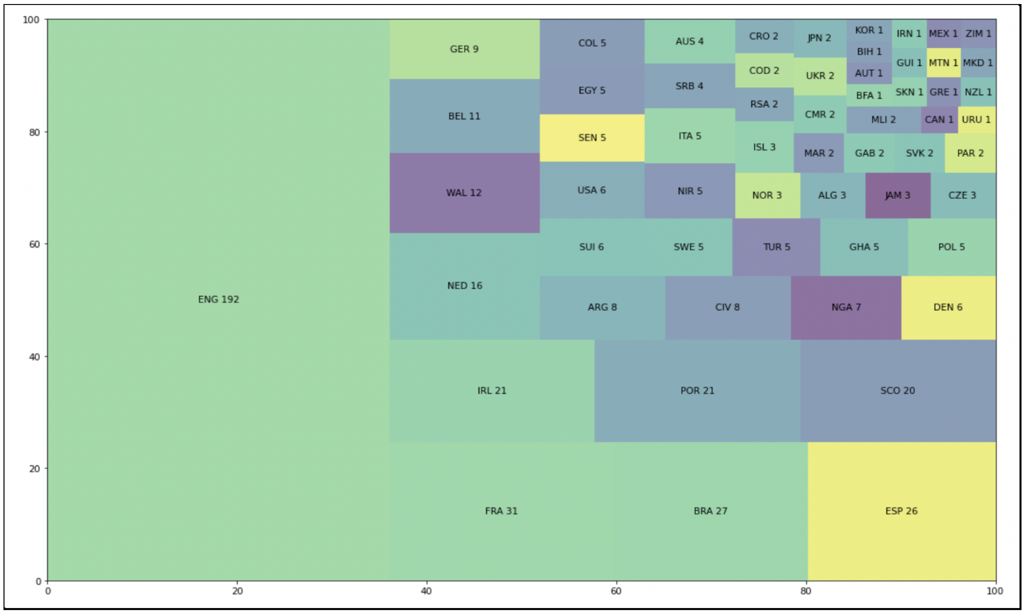




From these plots we infer that:

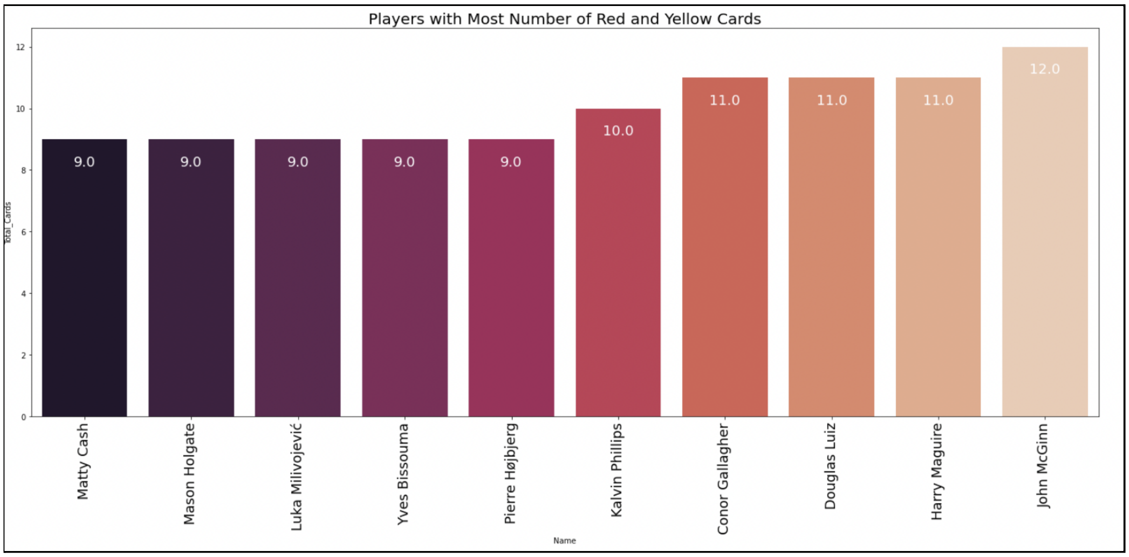
a. Most players are in their mid-twenties
b. Crystal Palace has one of the eldest squads percentile
c. Manchester United has one of the youngest squad with average player age of 23

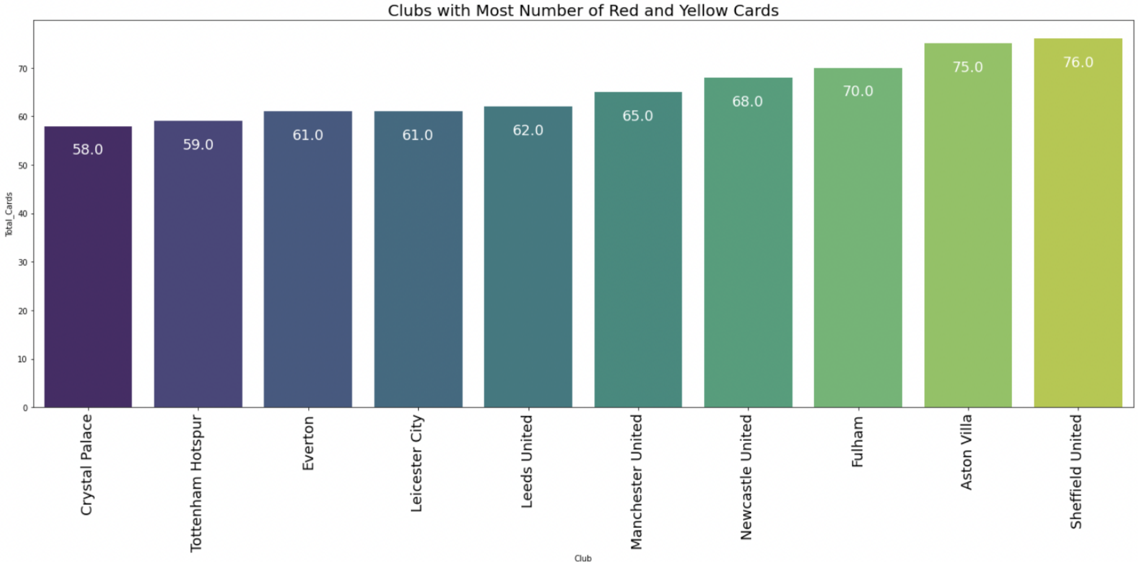
3. Nationalities of the players in the league:



Given that this is the English Premier League, we were expecting maximum players to be from England which came out to be true. This was followed by France and Brazil.

4. Players and Clubs with Most Number of Red and Yellow Cards





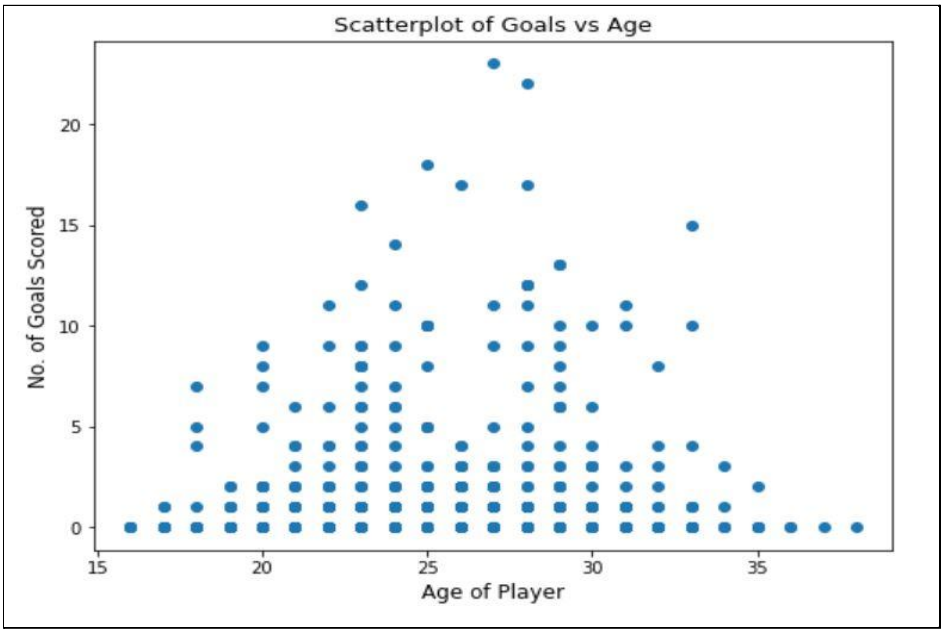
Above shown are players with the most number of cards. It is generally expected that defenders earn the most number of cards because they end up tackling most in a game. Looking at the players with the most number of cards, we can see that all of them are defenders.

Further, the second plot shows the clubs which have accumulated the most number of cards. Sheffield United leads the chart followed closely by Aston Villa. These clubs are usually at the bottom of the league table. A surprise inclusion is Manchester United who need to probably work on their discipline in the games.

Conjectures:

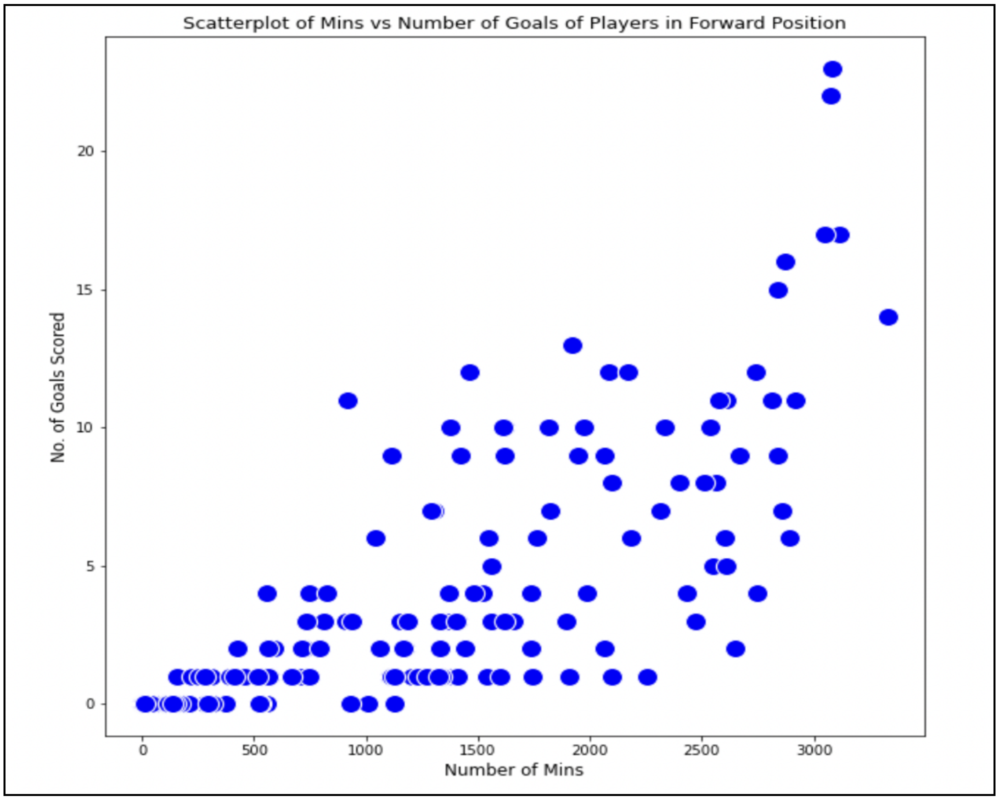
After exploring the dataset and looking at all the variables there were a few conjectures which were proposed. These were then approved or disapproved based on the visualizations created to understand them better. Some of the conjectures are as listed below:

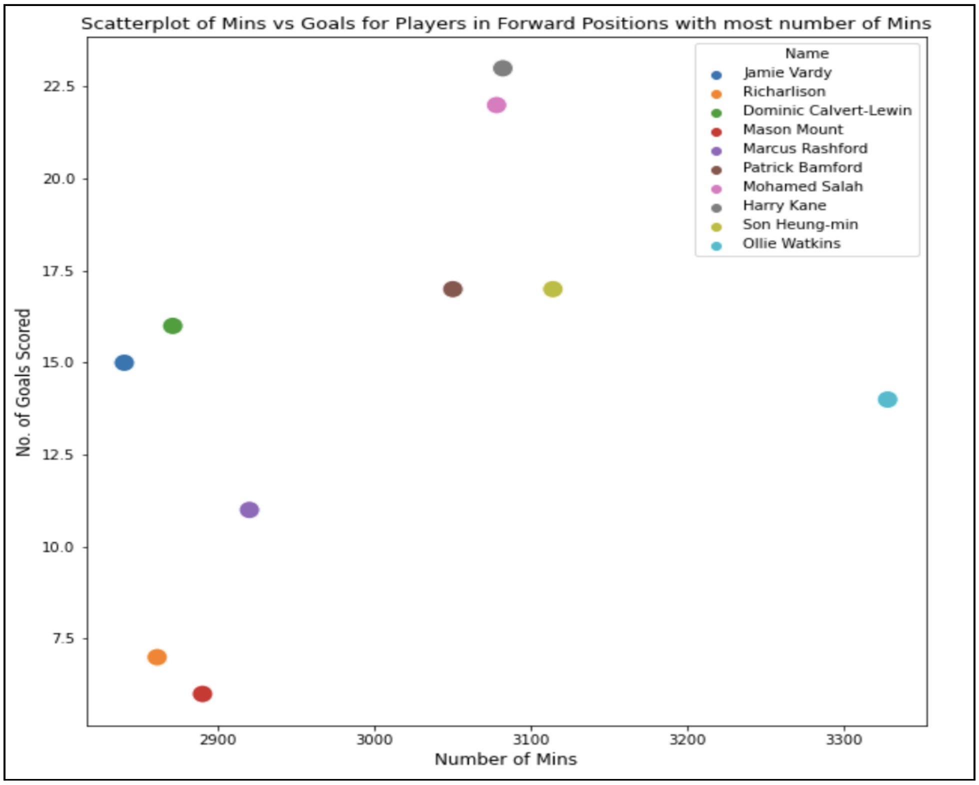
1. Players reach an age of peak performance after which their performance starts to reduce.



   To explore this conjecture, performance was defined based on the number of goals scored by a player. A scatterplot was plotted for the number of goals scored vs the age of players as shown above. By observation it is clear that players who were aged between 24-30 scored the most number of goals. Therefore we can infer that players are probably at the peak of their careers during this period and hence tend to score more goals. Hence this conjecture was approved.

2. More minutes played by a player during the season mean higher goals scored by the players in the forward position.


Scatterplot of Mins vs Number of Goals of Players in Forward Position
No. of Goals Scored
0
5
10
15
20
Number of Mins
0
500
1000
1500
2000
2500
3000

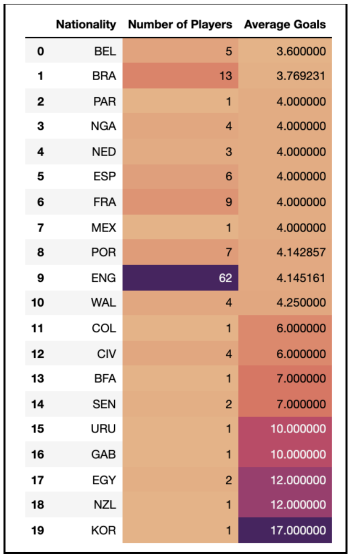

Scatterplot of Mins vs Goals for Players in Forward Positions with most number of Mins
Name
Jamie Vardy
Richarlison
Dominic Calvert-Lewin
Mason Mount
Marcus Rashford
Patrick Bamford
Mohamed Salah
Harry Kane
Son Heung-min
Ollie Watkins
No. of Goals Scored
7.5
10.0
12.5
15.0
17.5
20.0
22.5
Number of Mins
2900
3000
3100
3200
3300

Looking at the first figure, an increasing trend can be seen i.e. as the number of minutes played by a player in the season increases, the number of goals scored also tends to increase.

Although interestingly the second plot shows that the top 2 goal scorers (Harry Kane & Mohamed Salah) did not play the maximum number of minutes but still ended up as the top goal scorers of the season.

3. English players will dominate in terms of performance

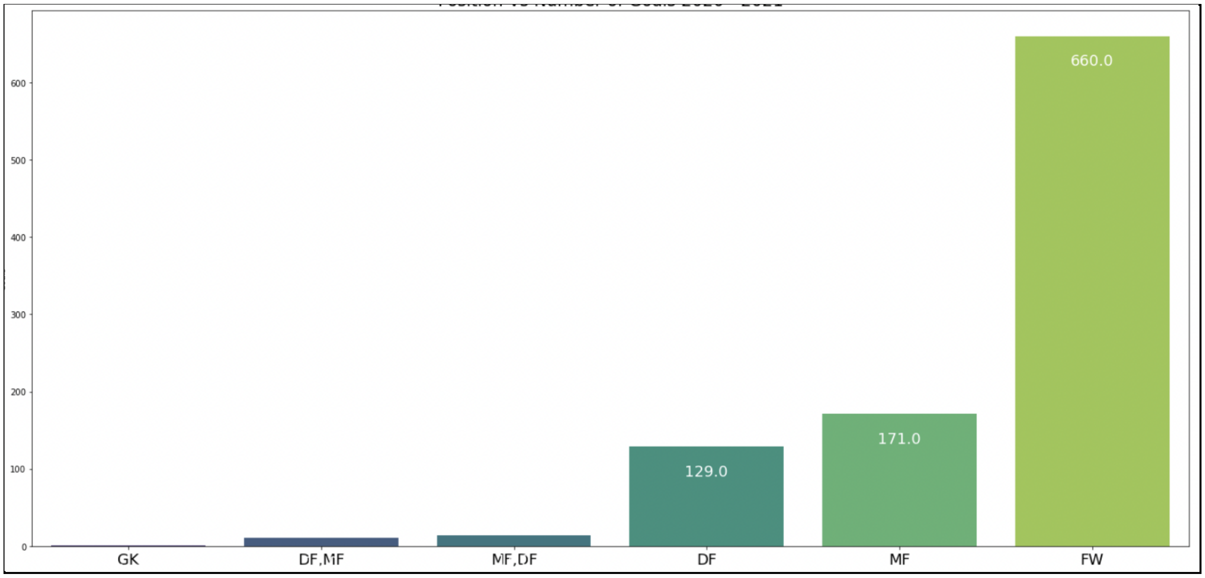
Given that this is the "English" Premier league, we expect english players to dominate because of home advantage. Players from other countries have scored significant number of goals compared to English Players



	Nationality	Number of Players	Average Goals
0	BEL	5	3.600000
1	BRA	13	3.769231
2	PAR	1	4.000000
3	NGA	4	4.000000
4	NED	3	4.000000
5	ESP	6	4.000000
6	FRA	9	4.000000
7	MEX	1	4.000000
8	POR	7	4.142857
9	ENG	62	4.145161
10	WAL	4	4.250000
11	COL	1	6.000000
12	CIV	4	6.000000
13	BFA	1	7.000000
14	SEN	2	7.000000
15	URU	1	10.000000
16	GAB	1	10.000000
17	EGY	2	12.000000
18	NZL	1	12.000000
19	KOR	1	17.000000

Since this dataset is from the English Premier League which is played in England, one would expect the English players to be the top performing players of the league because they would know the conditions better than the players from the other countries. The above table shows us that the conjecture is not true. Even though the number of English players is very high, the average number of goals scored by players of other countries is also significant.

4. Top goal scorers will be the forward players



Since the players who play forward are in front of goal and are the ones who usually find themselves in goal scoring positions, it is expected that the maximum goals will be scored by the players who play forward.

The above graph shows that the forwards are indeed the ones most involved with the goals. Hence this conjecture comes out to be true.

Popularity Score:

To find the most popular players, a popularity score was created. Various features have been used to come with a popularity score. Initially, a performance based popularity score was considered. But it was soon realised that performance cannot be the sole criteria to decide if a popular is popular. Some of the parameters were not included in the original dataset but were added (mentioned in the references) because they were crucial for the analysis. Following are a few parameters which were felt significant in contributing to the popularity of a player:

1. Number of Minutes
   Number of minutes a player plays during the whole season gives us a direct estimate of the time he is seen on television and in the stadium. More the number of minutes a player plays the more is the brand engagement.

2. Club Historic Performance
   Clubs which have performed really well historically have fans across the world. Everytime a player signs up for a club which performs well, there is more new surrounding the transfer and hence more attention to the player. The Premier League

specifically has a collection of ‘Top 6’ clubs. These 6 clubs have been the ones performing really well historically and these have fans across the world. This parameter can also reflect on the television engagement of the club and hence the players of that club.

3. Goals
Everytime a player performs well in terms of scoring more goals, there is news surrounding the player everywhere. Goals can also reflect the performance of the player and hence it was felt necessary in contributing to the popularity.

4. Yellow Cards and Red Cards
Yellow cards and red cards gathered during the course of the season reflect negatively on the player. A player who has more cards against him is often considered aggressive and a sports brand definitely does not want to engage with a player who is considered aggressive. Hence cards will have a negative impact on the popularity of the player.

5. Stadium Capacity
Each club has a home stadium where all the home games of the club are played at. A club which has a very high stadium capacity is able to house a very high number of fans. This increases the viewership of the matches and hence the brand engagement as well. This also to a certain extent explains the size of the club. Generally bigger clubs have a higher stadium capacity.

Each of these factors were considered for all the players and each factor was scaled using min max scaling. Following this a weightage in terms of their importance to contributing to the popularity score was given to each of the factors. The weightages used are as shown in the table:

Variable:	**Multiplier:**
Minutes	5
Club popularity (Hist Performance)	5
Number of goals scored	5
Yellow Cards	-3
Red Cards	-5
Stadium Capacities	10

These were then summed and the final scores were then again scaled using min max scaling and the final score was calculated on a scale of 100.

Model Development:

Since the idea is to be able to calculate popularity scores beforehand for players who have not yet played in the premier league, it calls for some method of estimating a few parameters which contribute to the popularity score. One of the major factors contributing to the popularity is the number of goals scored and the target variable for modelling will be goals scored by a player.

Looking at the conjectures, it was clear that some features were related to the target variable. Age, number of minutes played and the position of the player were found to be the most important predictors. Multiple models were trained by using these as the input variables and goals as the target variables. Since position of a player was a categorical variable, it could not be directly included. Considering that forwards end up scoring more goals than midfielders, defenders and goalkeepers, it was considered as an ordinal variable and each position was assigned a number based on this. Model training was done using a training set which was 80% of the data, and 20% of the data was kept as the validation set.

Following are the models which were trained and the accuracies are as shown in the table below:

Model Name	Test Set Accuracy
SVR (RBF Kernel)	75.2%
Linear Regression	42.5%
Linear Regression (poly features deg = 4)	75.1%
Random Forest Regressor	**82.1%**

For a few of the models, scaling of features was also considered. Although there was not much improvement in the accuracy in the Random Forest Regressor, some improvement was noticed in the Support Vector Regressor. But since the best results were obtained on the random forest regressor, it was chosen for further analysis.

Model Relevance:

Given that there are 2 different players signed by 2 different clubs in the new upcoming season. Now, all the factors except the number of goals, yellow cards and red cards are known to calculate the popularity score of the player. Number of yellow cards and red cards is more or less the same for players over the years. Cards reflect on the aggressive nature of the player and an aggressive player will remain aggressive and the number of cards given to him from the previous season can be used. But the number of goals are unknown in the context of the premier league. Although the player's previous performance in some other league can be considered, to get an accurate estimate of the number of goals scored by a player in the premier league, the model can be used.

Hence, the model can be used to predict the number of goals a new player in the premier league will end up scoring. This can help the sports company identify a player with huge potential who is soon going to be a success beforehand. This gives the sports company an opportunity to sign this player for less amount as he is not yet very established.

An example of 2 players is taken to further explain the use case. Let us assume Season 21/22 is starting. Manchester United has signed a 36 year old forward player (A) and Arsenal have signed a 22 year old midfield player (B) from the other leagues. A good enough assumption is that these players will end up playing 3000 minutes in the season as they are one of the good signings for the clubs.

The model is used to predict the number of goals these players will score. The model predicts that player A ends the season with 18 goals and player B ends the season with 6 goals. Now the popularity score is calculated for both the players (shown in table below) to conclude which player should be preferred based on the popularity score.

	Player A	Player B
Minutes	3000	3000
Club popularity (Hist Performance)	5	5
Number of goals scored	18	6
Yellow Cards (Avg Last Season)	3	0
Red Cards (Avg Last Season)	0	0
Stadium Capacities	76212	60355
Total Score	**102.70**	81.22

A score of 102.7 means that this player will end up being the most popular player in the premier league. This is indeed true because player A in this case is Cristiano Ronaldo and player B in this case is Martin ødegaard.

Conclusion:

To sum it up, the premier league data of the season 20/21 was explored to find trends emerging in the data. A popularity score was created to assess the popularity of different players. Multiple different models were trained to predict the number of goals a player might score on the basis of a few features.

Limitations and Future Scope:

There are a few limitations in this approach which are as listed below:

1. The weights of the factors to calculate the popularity score are chosen by intuition and not backed by data. A data based approach for these weights would help us be more accurate.
2. A model can be developed to predict the number of assists. Further the number of assists can also be used to contribute to the popularity score.
3. More data can be included so as to get better results on the models.

Story of the Group

We initiated the case study by selecting Analysis of EPL 2020-21 as our topic for the case study. We came across the dataset in Kaggle and became very much intrigued by how the performance of players can be predicted given a set of predictors. As sports sponsorship plays a very big role in EPL, we realised that estimating the performance of players would be a game changer for companies and in the field of advertisement. Companies are willing to invest millions of dollars to ensure that their brands have the right players to endorse their products and to make sure that their brands are associated with success. After finalising the topic, we split the work among ourselves. Riddhi was able to explore the dataset and find trends, patterns and relationships between the inputs. She was able to come up with relevant visualisations which provided insights into the performance of each player. EDA performed helped the remaining team members to get an essence of data. Using this information, the entire team was able to collaborate and come up with the conjectures. Kratika went ahead and worked on getting evidence for our conjectures. She was able to prove some of the conjectures and was able to get necessary plots for these. These conjectures were essential in determining the predictors that we could use in data modelling. Kartik and Ashay got started on the modelling using the information from the conjectures. They tried models like SVR (RBF Kernel), Linear Regression, Linear Regression (poly features deg = 4) and Random Forest Regressor. It was found that Random Forest Regressor worked best on the dataset and gave the highest accuracy. Kartik was able to come up with the popularity score equation which considered various other factors along with the Goals predicted by the model for the estimating the popularity of the players.

References:

Dataset:
https://www.kaggle.com/rajatrc1705/english-premier-league202021
Additional Data:
https://www.footballcritic.com/premier-league/season-2020-2021/venues/2/41756
https://www.premierleague.com/stats/top/clubs/total_scoring_att?se=363
https://www.premierleague.com/stats/top/clubs/wins?co=1&se=-1&co=-1?se=-1