Soccer Players Analysis

Kartik Nautiyal, Riddhi Thakkar, Kratika Shetty, Ashay Aglawe

Scope:

- There are soccer leagues being played across the world but Premier League is the most 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.

- To keep the study relevant in the recent context, we have used data from the premier league season of 2020-2021.

Target Audience and Motivation:

- The target audience for this project is a sports company (Nike), who is going to sign a brand ambassador for the upcoming season to showcase their new products (shoes, jerseys).

- This is an attempt to find a player who would give the most return on investment by reducing signing cost and drive maximum brand/product awareness.

- We are really interested in soccer and we wanted to understand how data analysis could be applied in the sports industry which we realised during the course of this project.

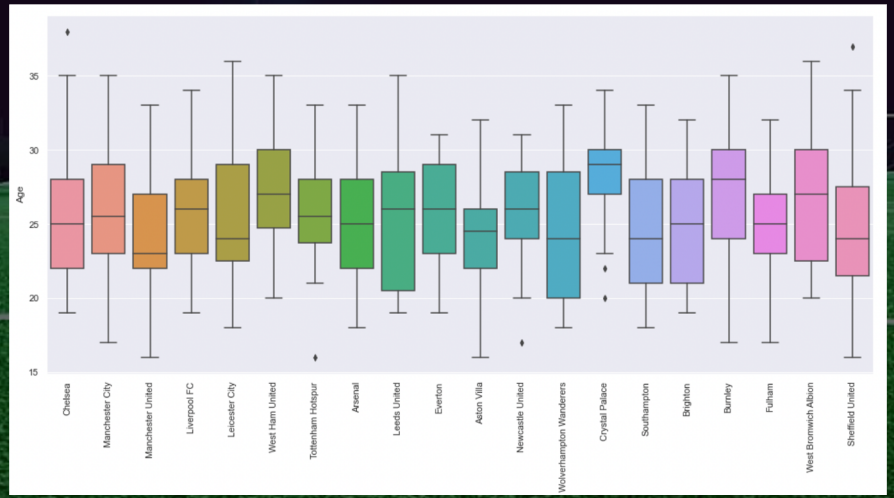

Exploratory
Data Analysis

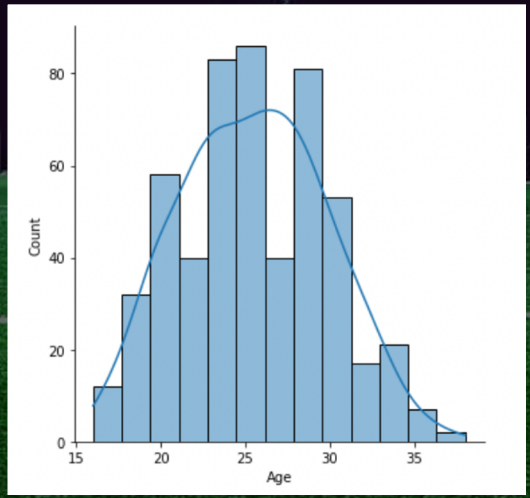
Dataset:

	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	Passes_Attempted	Perc_Passes_Completed	Penalty_Goals	Penalty_Attempted	xG	xA	Yellow_Cards	Red_Cards
2	Mason Mount	Chelsea	ENG	MF,FW	21	36	32	2890	6	5	1881	82.3	1	1	0.21	0.24	2	0
3	Edouard Mendy	Chelsea	SEN	GK	28	31	31	2745	0	0	1007	84.6	0	0	0	0	2	0
4	Timo Werner	Chelsea	GER	FW	24	35	29	2602	6	8	826	77.2	0	0	0.41	0.21	2	0
5	Ben Chilwell	Chelsea	ENG	DF	23	27	27	2286	3	5	1806	78.6	0	0	0.1	0.11	3	0
6	Reece James	Chelsea	ENG	DF	20	32	25	2373	1	2	1987	85	0	0	0.06	0.12	3	0
7	CÃ©sar Azpilicueta	Chelsea	ESP	DF	30	26	24	2188	1	2	2015	87.5	0	0	0.03	0.11	5	1
8	N'Golo KantÃ©	Chelsea	FRA	MF	29	30	24	2146	0	2	1504	86.6	0	0	0.04	0.05	7	0
9	Jorginho	Chelsea	ITA	MF	28	28	23	2010	7	1	1739	89.5	7	9	0.31	0.09	2	0
10	Thiago Silva	Chelsea	BRA	DF	35	23	23	1935	2	0	1871	93.5	0	0	0.05	0.02	5	1
11	Kurt Zouma	Chelsea	FRA	DF	25	24	22	2029	5	0	1720	91.9	0	0	0.08	0	3	0
12	Mateo KovaÄiÄ‡	Chelsea	CRO	MF	26	27	21	1815	0	1	1737	91	0	0	0.05	0.09	4	0
13	Antonio RÃ¼diger	Chelsea	GER	DF	27	19	19	1710	1	0	1476	90.7	0	0	0.06	0.02	0	0
14	Christian Pulisic	Chelsea	USA	FW,MF	21	27	18	1738	4	2	690	80	0	0	0.28	0.14	2	0
15	Kai Havertz	Chelsea	GER	MF,FW	21	27	18	1520	4	3	765	86.1	0	0	0.37	0.09	2	0
16	Andreas Christensen	Chelsea	DEN	DF	24	17	15	1371	0	0	1089	92.8	0	0	0.01	0.02	2	1
17	Hakim Ziyech	Chelsea	MAR	FW,MF	27	23	15	1172	2	3	734	74.7	0	0	0.15	0.28	3	0
18	Tammy Abraham	Chelsea	ENG	FW	22	22	12	1040	6	1	218	68.3	0	0	0.56	0.07	0	0
19	Marcos Alonso	Chelsea	ESP	DF	29	13	11	960	2	0	592	81.6	0	0	0.16	0.11	2	0
20	Callum Hudson-Odoi	Chelsea	ENG	FW,DF	19	23	10	1059	2	3	659	82.2	0	0	0.12	0.26	0	0
21	Olivier Giroud	Chelsea	FRA	FW	33	17	8	748	4	0	217	74.2	0	0	0.58	0.09	1	0
22	Kepa Arrizabalaga	Chelsea	ESP	GK	25	7	6	585	0	0	243	81.5	0	0	0	0	1	0
23	Billy Gilmour	Chelsea	SCO	MF	19	5	3	261	0	0	215	89.3	0	0	0.01	0.04	0	0
24	Willy Caballero	Chelsea	ARG	GK	38	1	1	90	0	0	26	92.3	0	0	0	0	0	0
25	Ruben Loftus-Cheek	Chelsea	ENG	FW	24	1	1	60	0	0	16	68.8	0	0	0	0	0	0
26	Emerson Palmieri	Chelsea	ITA	DF	25	2	0	90	0	0	63	81	0	0	0	0	0	0
27	Fikayo Tomori	Chelsea	ENG	DF	22	1	0	45	0	0	29	93.1	0	0	0	0	0	0

Distribution of players on age

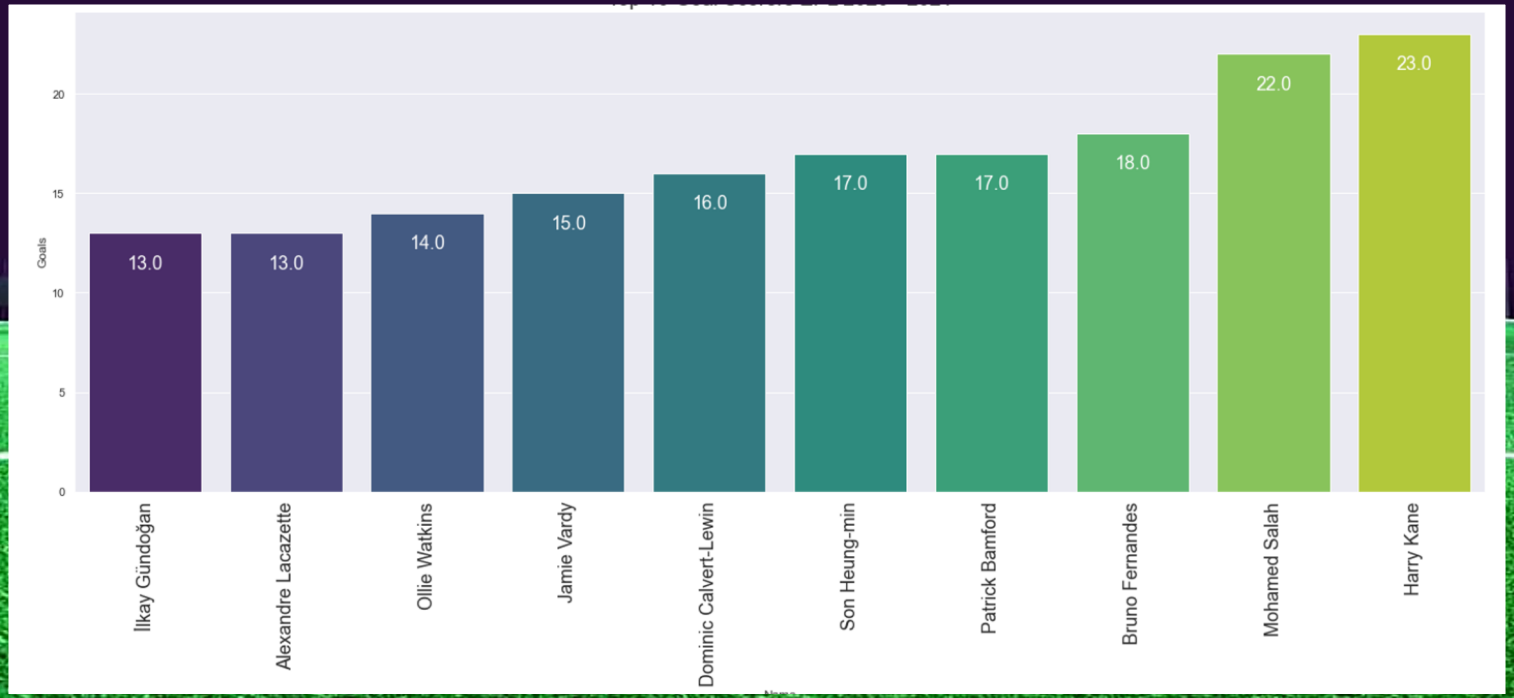
- Most players are in their mid-twenties
- Crystal Palace has one of the eldest squads percentile
- Manchester United has one of the youngest squad with average player age of 23



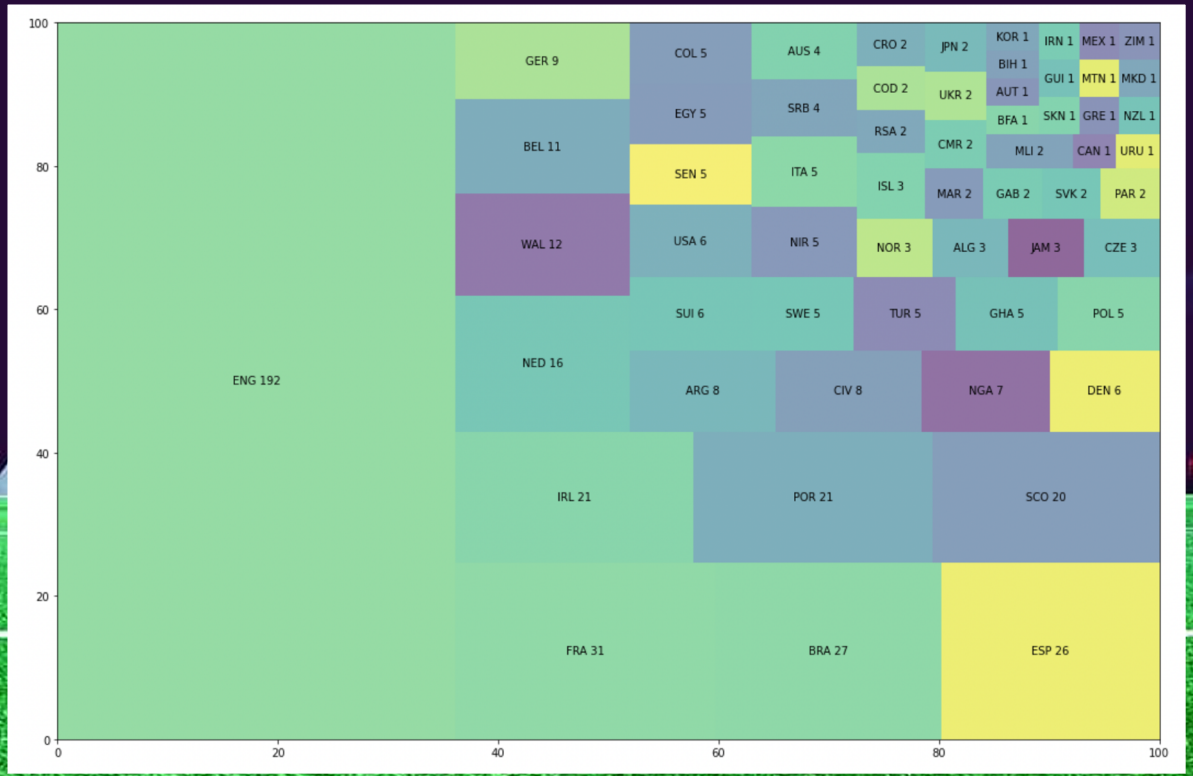


Goal Scorers:

- Top 10 Goal Scorers in EPL 2020-21

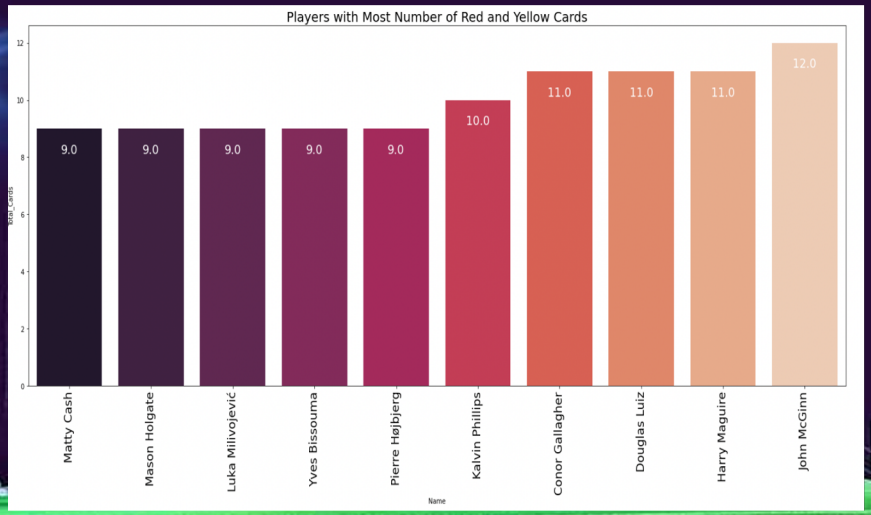


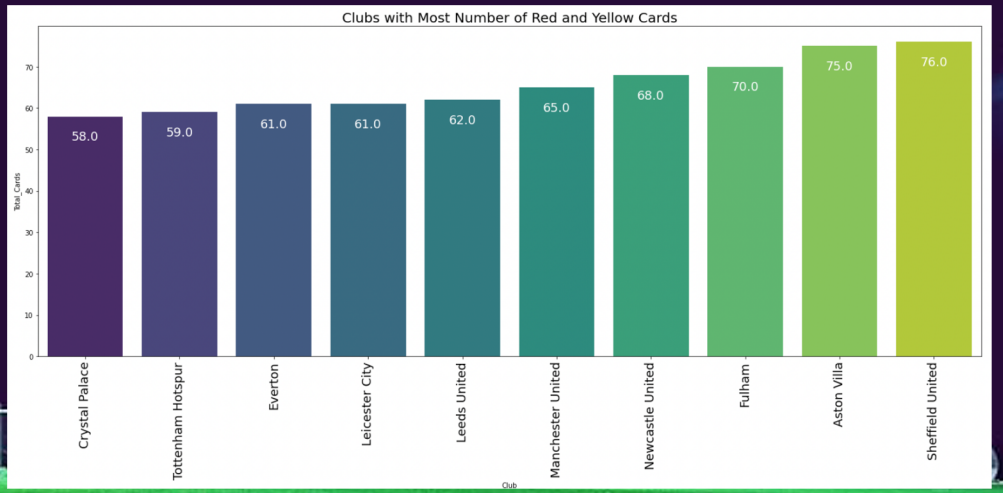
Nationality of players:



- Most of the players are from England followed by France, Brazil and Spain

Players with Most Number of Red and Yellow Cards



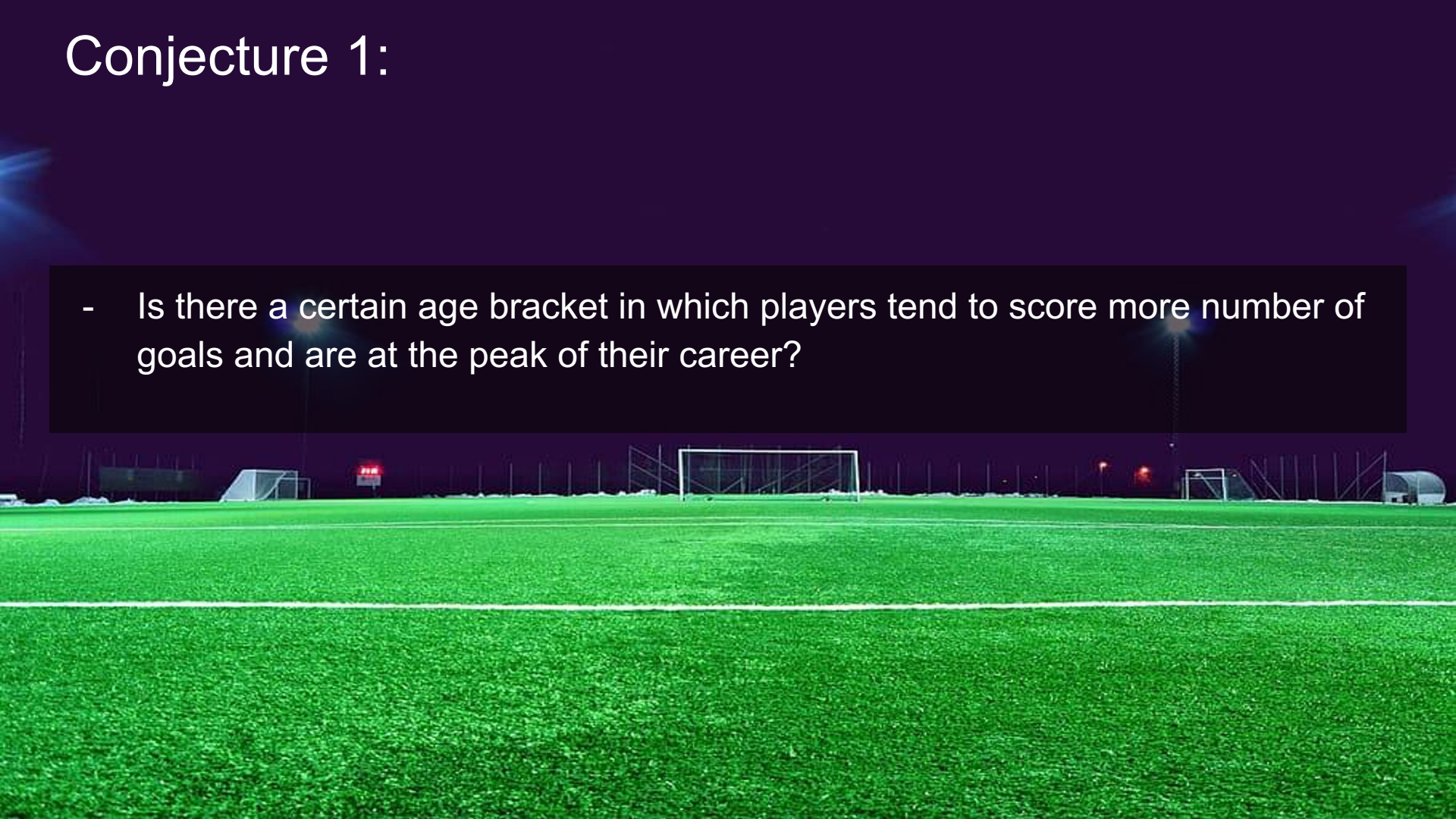


- John McGinnn had the most number of Cards issued against him
- It is not surprising that the players with the most number of cards against them are all defenders.
- Sheffield United and Aston Villa received the highest number of cards. The general trend seems to be that the clubs lower down in the league tables end up gathering more cards.

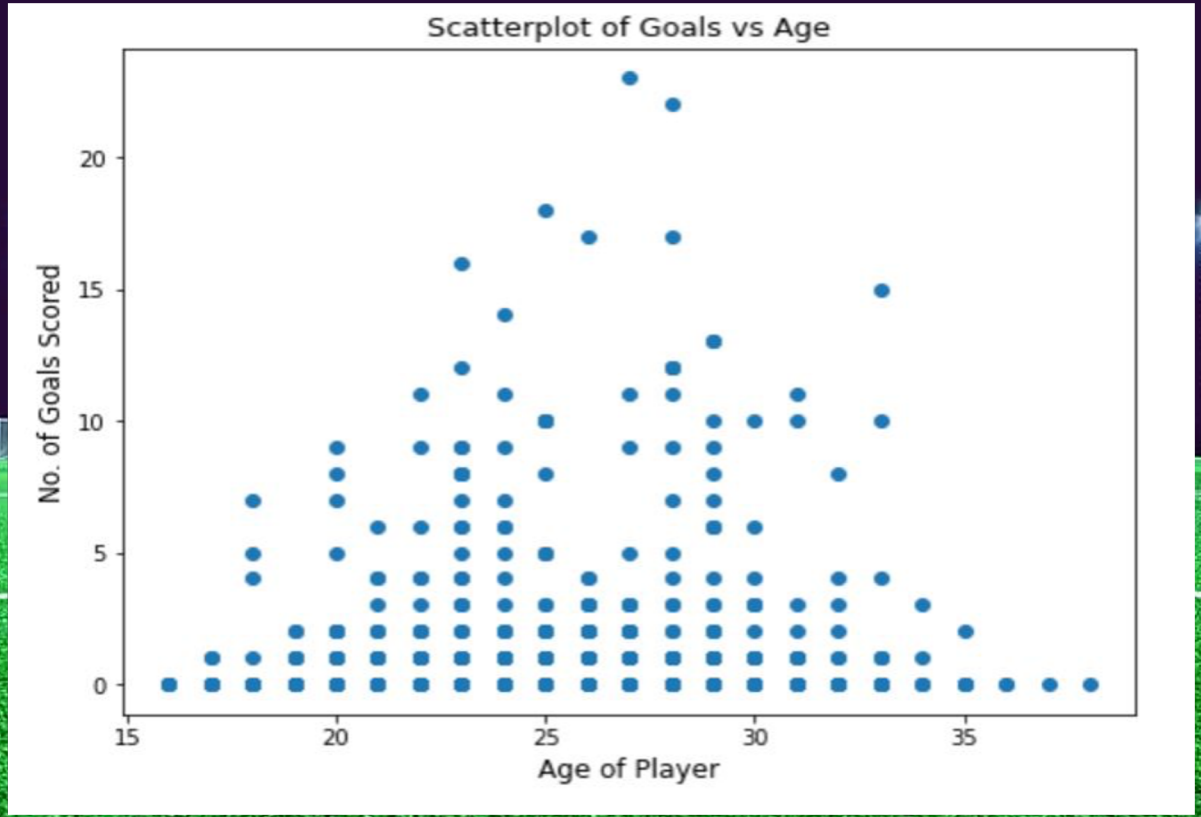

Conjectures

Conjecture 1:

- Is there a certain age bracket in which players tend to score more number of goals and are at the peak of their career?



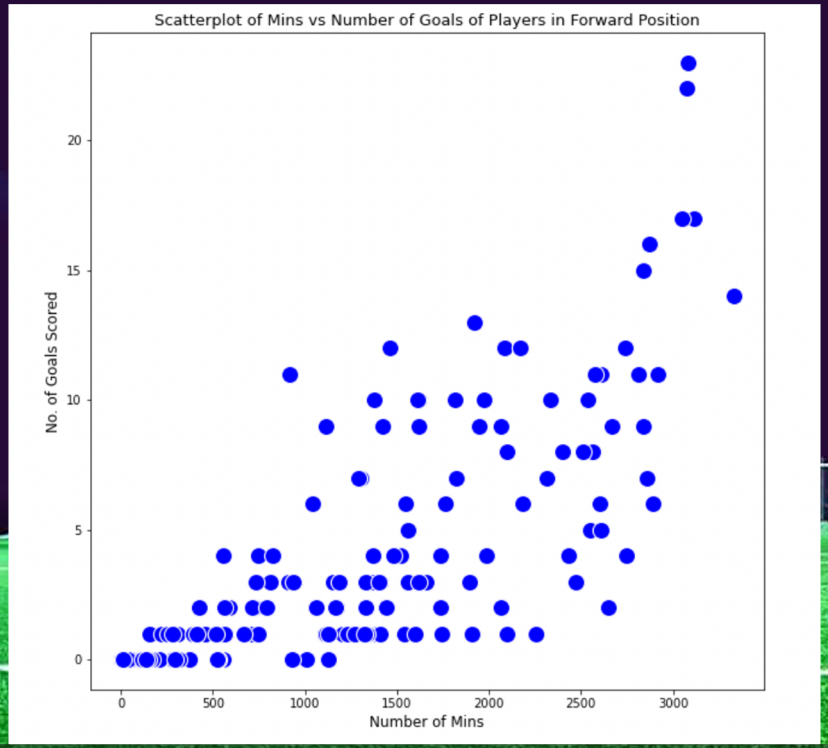
Conjecture 1

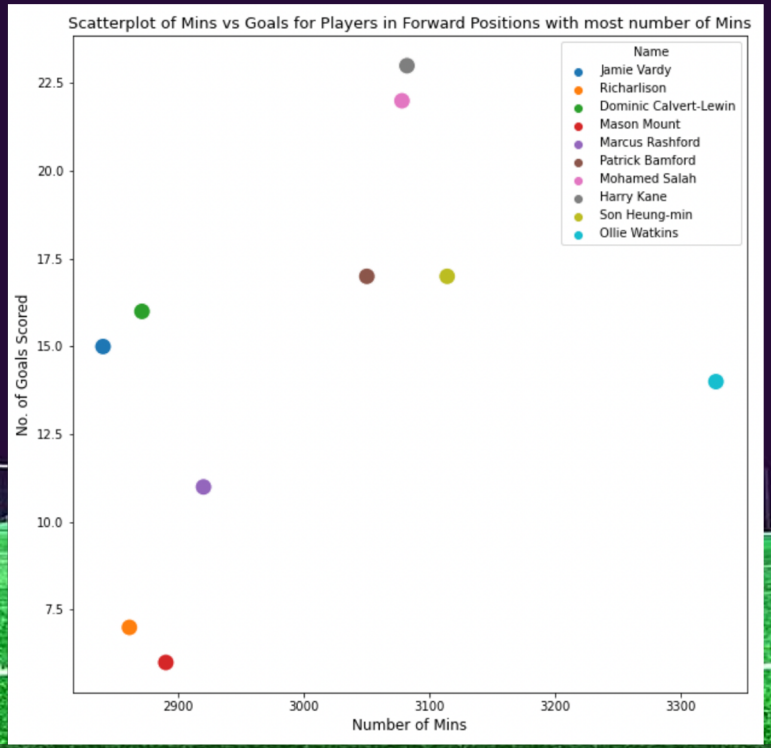


Conjecture 2:

- More minutes should mean more goals for players in forward positions?

Conjecture 2





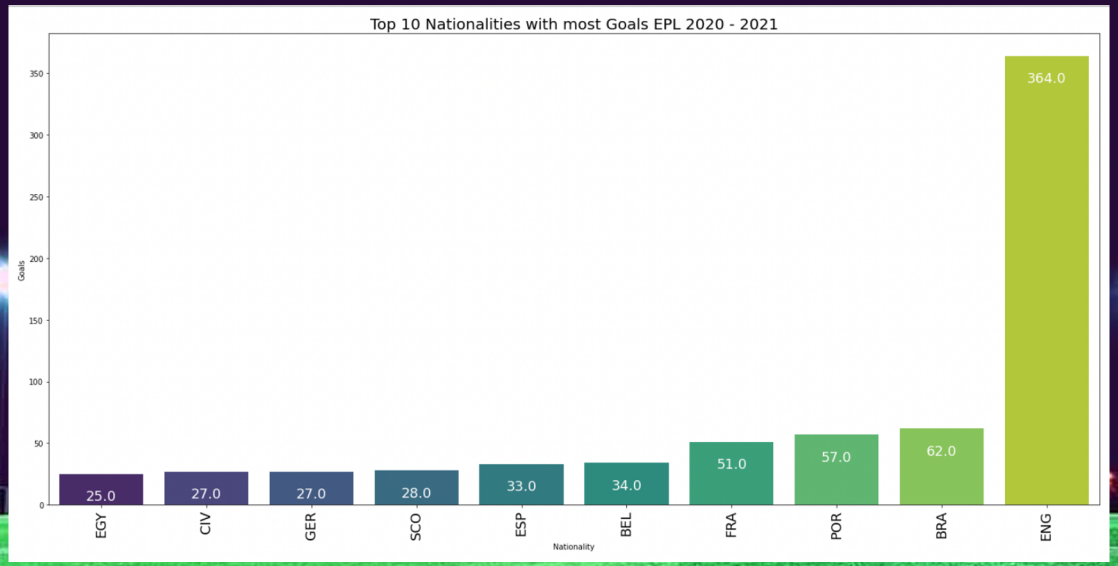
- Looking at the first figure, an increasing trend can be seen i.e. as the number of minutes played by a player in the season increases, the number of goals scored also tends to increase.
- Although interestingly the second plot shows that the top 2 goal scorers (Harry Kane & Mohamed Salah) did not play the maximum number of minutes but still ended up as the top goal scorers of the season.

Conjecture 3

- Players from England should have higher average number of goals compared to players from other countries

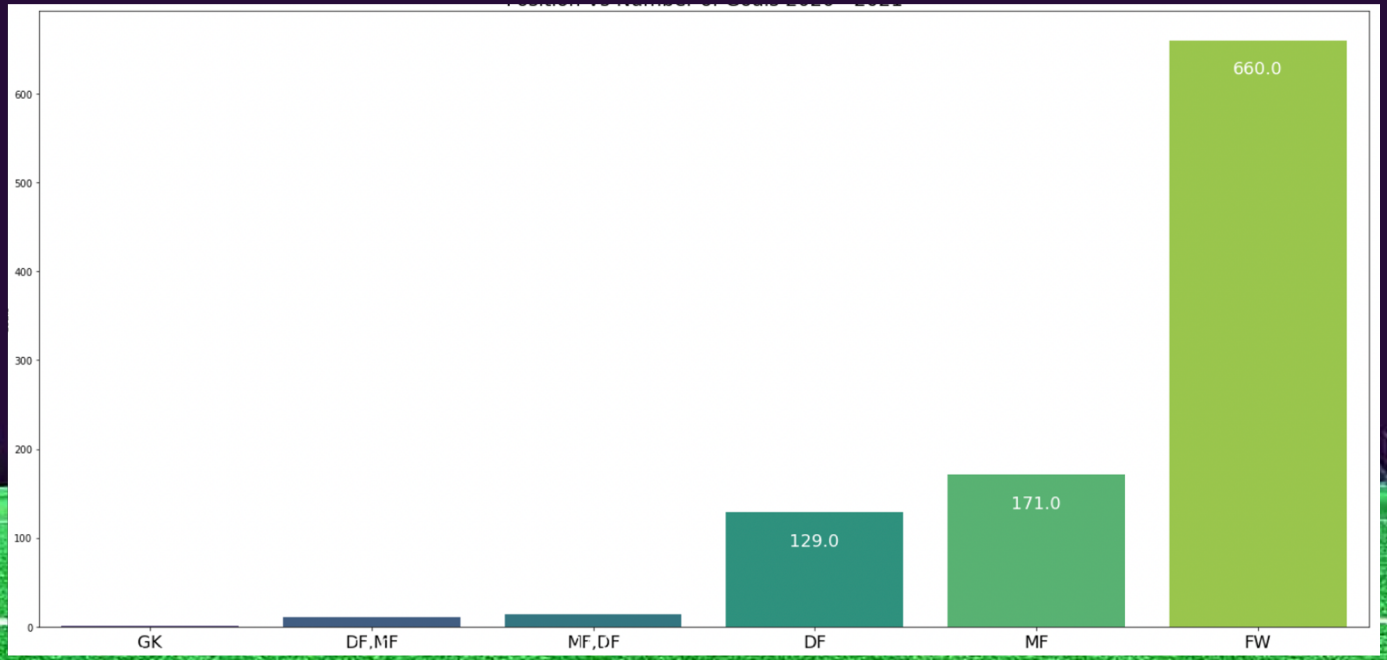
	Nationality	Number of Players	Average Goals
0	BEL	5	3.600000
1	BRA	13	3.769231
2	PAR	1	4.000000
3	NGA	4	4.000000
4	NED	3	4.000000
5	ESP	6	4.000000
6	FRA	9	4.000000
7	MEX	1	4.000000
8	POR	7	4.142857
9	ENG	62	4.145161
10	WAL	4	4.250000
11	COL	1	6.000000
12	CIV	4	6.000000
13	BFA	1	7.000000
14	SEN	2	7.000000
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

Conjecture 3



- 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
- These players attract viewers from their home countries. Deals with local sponsors and endorsements from their home nations can be targeted.

Conjecture 4: Top Goal Scorers should be Forwards



- 171 Goals were scored by the midfielders
- Forwards have scored 660 goals
- Position of the players would be significant in estimating the number of goals

Popularity score

Clubs:

Manchester United

rsenal

Liver

Chelsea

Clubs:

Fulham

Brighton

Popularity score

- Initially, we based the popularity score only on the performance of the player.

- Only, Performance is not a good to understand popularity of a player.

Other variables: Clubs

- Clearly some clubs are more popular than the other clubs (even for people who may not follow soccer). Hence, a player from popular club may be more popular than another player from some club which is not popular even if their performances are vice versa.

- We used the all time performance of clubs to score them on a scale of 5 with a score of 5 being allotted to the best performing clubs and 1 to the worst performing ones

Other variables: Stadium Capacities

- Some clubs have a very high stadium capacity and this allows a lot more people to be present at the venue.

- More stadium capacity = more people looking at the product

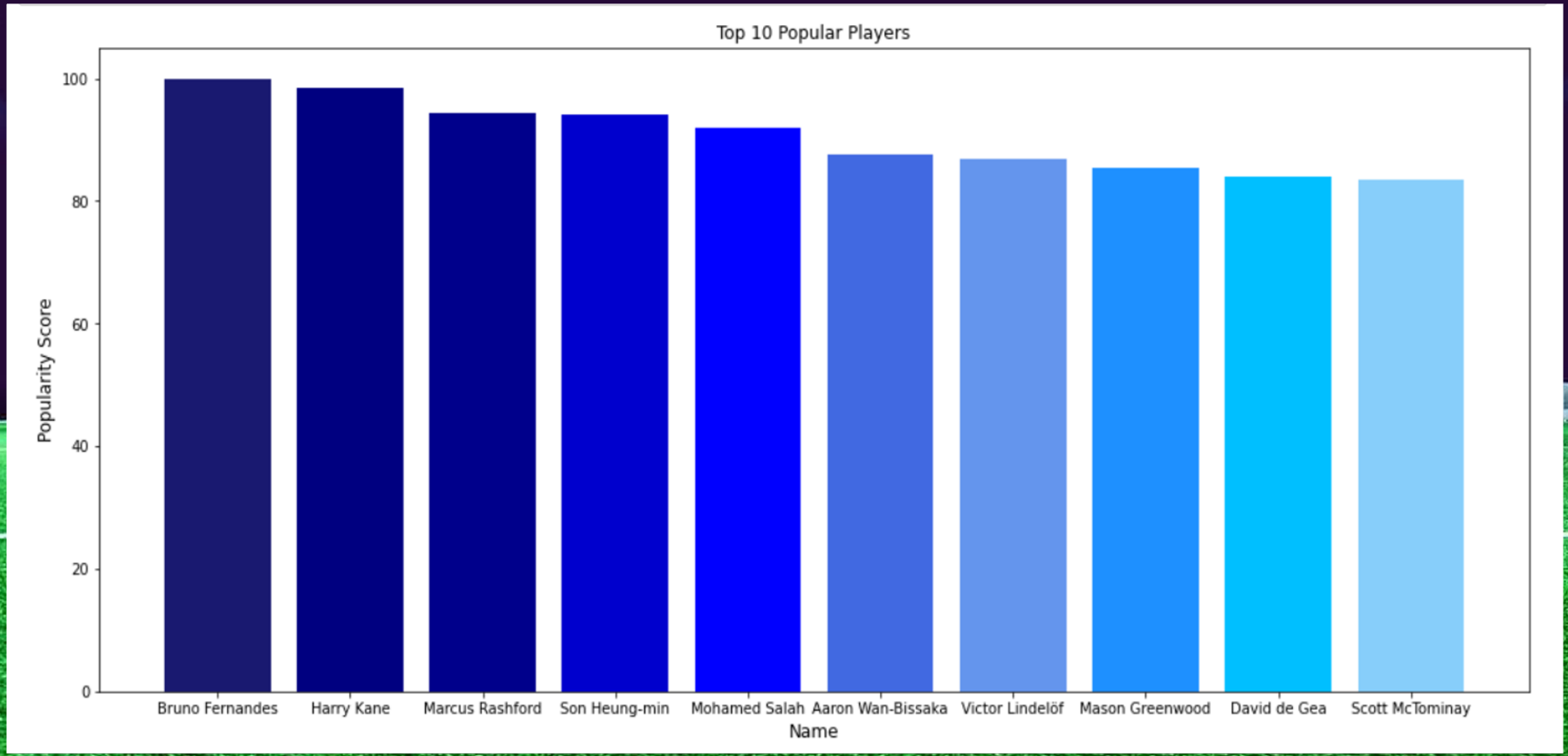
Total popularity score:

- Min-max scaling was done on all features and weights were given to features according to their importance as showing 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

- The total score was then again min-max scaled to a score of 100

Most popular players:




Modeling

Motivation:

- 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 other leagues.

- These players are not established, hence could be signed for less amount by the sports company.

- But, no idea about their performance and hence no idea about their popularity

Model:

- Used predictors of minutes to play, age, position and trained multiple models to predict the number of goals.

- Accuracies reported as follows:

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

Getting back

- Player A or player B?
  Using the model we predicted the number of goals for both of them and we found that player A will end the season with **18 goals** and B with **6 goals.**

	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

Player A

Cristiano Ronaldo

Player B

Martin ødegaard

Conclusion:

To sum it up, we:

- Looked at the data from the Premier League for the season 20/21.
- Explored the data to see trends.
- Created a popularity score by giving weights to different factors.
- Created multiple models to predict the number of goals a player might score on the basis of some features.

Limitations and Challenges:

- Similar aged players will end up with same number of goals. (Random Forest Regressor)

- Club name as a predictor could be used. We explored dummy variables and one hot encoding but could not improve the accuracy any further.

- Assists can be predicted and included in the popularity scoring.

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


Questions?


Thank you