Football Player's Performance & Market Value

DATA MINING & MACHINE LEARNING PROJECT

Ohad Mavdali
Dotan Gotshtein

Introduction

A huge amount of money is involved with the transfers of top players in the big European football leagues. For various reasons, obtaining a good economic valuation of football players throughout the year is valuable, Furthermore, it is relevant to consider how the market value of a player relates to the performance of that player. these factors depend on various parameters of the player, that we gleaned from few sources on the web.

In this project, we will investigate how the market value and performance of the best 5 leagues players consists of, and try to predict their market value and scored goals.

Research questions:

1. What are the features that makes a player valuable?
2. Can we predict market value of a player?
3. Can we predict how many goals a player will score?

How did we create the database?

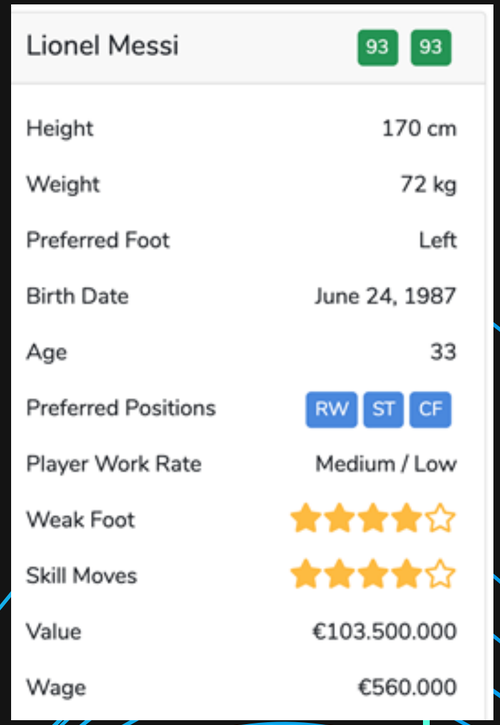
Web scraping players data using BeautifulSoup from FIFA index website.
total 52,500 data rows.

Scraping players data using Understat python library.
total 7 seasons data frames that merged to a big dataset of 52,500 data rows all over 7 seasons.

Describing the data: FIFA index.com

Collecting 7500 players data from seasons 2014 – 2020, including their Overall and Potential FIFA score, Market Value, Weekly Salary, Height, Weight, Age & Preferred Foot.

	Player	Overall Score	Potential Score	Market Value	Weekly Salary	Height	Weight	Age	Preferred Foot
0	Lionel Messi	94	94	110500000.0	560000.0	170	72	32	Left
1	Cristiano Ronaldo	94	94	77000000.0	400000.0	187	83	34	Right
2	Neymar Jr	92	92	108000000.0	290000.0	175	68	27	Right
3	Jan Oblak	91	93	75500000.0	115000.0	188	87	26	Right
4	Kevin De Bruyne	91	91	93000000.0	350000.0	181	70	28	Right
...	...	...	...	...	...	...	...	...	...
7465	Adam Matthews	68	68	750000.0	5500.0	178	71	27	Right
7466	Danny Wilson	68	71	925000.0	3200.0	188	79	27	Left
7467	Niklas Backman	68	68	700000.0	5600.0	186	80	30	Right
7468	Marcos Álvarez	68	69	1100000.0	2400.0	179	77	27	Right
7469	Lerin Duarte	68	68	925000.0	3700.0	169	71	29	Left



Describing the data: Understat.com

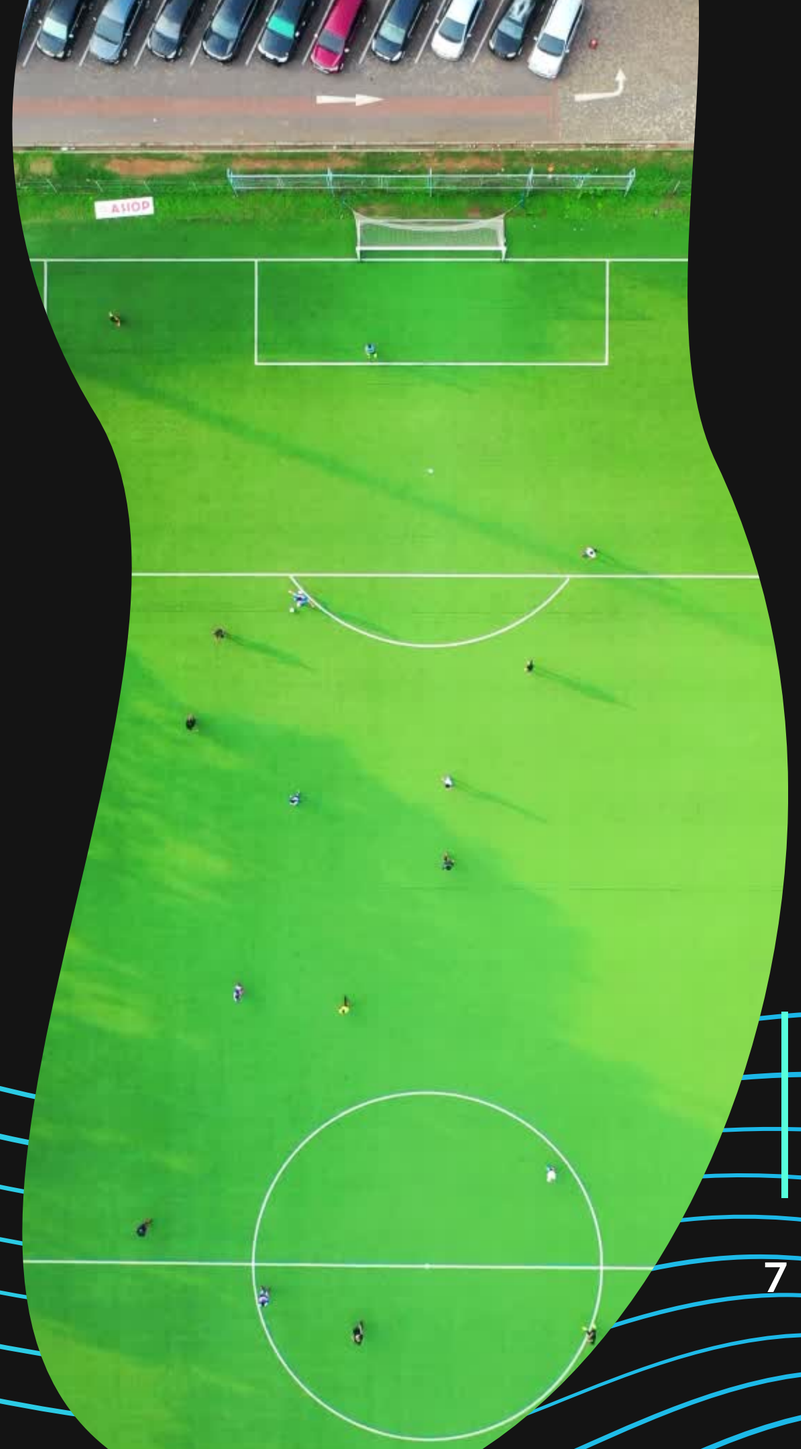
collecting 7500 players data from seasons 2014 – 2020, including their Names, goals, xG, assists, xA, shots, key passes, yellow cards, red cards, npg, npxG, xGchain, xGBuildup , minutes played, games.

Player	Team	Apps	Min	G	NPG	A	xG	NPxG	xA	xGChain	xGBuildup
Mohamed Salah	Liverpool	23	1926	17	11	3	12.65 -4.35	8.08 -2.92	3.86 +0.86	15.36	6.07
Son Heung-Min	Tottenham	22	1909	13	13	6	8.07 -4.93	8.07 -4.93	4.39 -1.61	12.28	3.64
Harry Kane	Tottenham	20	1755	13	10	11	11.69 -1.31	9.41 -0.59	5.51 -5.49	14.02	3.25
Bruno Fernandes	Manchester United	23	1939	13	7	9	10.90 -2.10	5.57 -1.43	8.65 -0.35	17.35	7.03
Dominic Calvert-Lewin	Everton	20	1774	13	13	0	12.63 -0.37	12.63 -0.37	0.71 +0.71	12.05	2.63
Jamie Vardy	Leicester	20	1600	12	6	5	14.25 +2.25	8.92 +2.92	2.15 -2.85	10.71	0.84
Patrick Bamford	Leeds	22	1891	12	11	5	12.36 +0.36	11.60 +0.60	2.49 -2.51	16.00	2.72
Callum Wilson	Newcastle United	21	1818	10	6	5	11.65 +1.65	8.60 +2.60	2.44 -2.56	11.30	0.65
Ollie Watkins	Aston Villa	21	1890	10	9	2	12.25 +2.25	10.73 +1.73	2.94 +0.94	15.16	4.09
Ilkay Gündogan	Manchester City	17	1313	9	8	1	6.54 -2.46	5.01 -2.99	2.78 +1.78	11.71	5.84

Data Optimizing

We merged all data sets we've reached into a one major data set and we had to handle the data. For examples to drop missing values, get rid of duplicates, outliers, dropping unimportant columns and convert categorical columns to numeric.

Ending up with 11,207 rows, 28 columns.



Our Final Database

Player	Overall Score	Potential Score	Market Value	Weekly Salary	Height	Weight	Age	Preferred Foot	id	games	time	goals	xG	assists	xA	shots	key_passes	yellow_cards	red_cards	npg	npxG	xGChain	xGBuildup	year
Lionel Messi	94	97	78000000.0	450000.0	169	67	27	1	2097	38	3374	43	35.891764	18	17.611956	187	95	4	0	38	31.432098	54.753610	19.756724	2014
Cristiano Ronaldo	92	92	55500000.0	325000.0	185	80	29	0	2371	35	3103	48	39.308761	16	13.600642	225	76	5	1	38	30.389686	43.608020	7.520237	2014
Robin van Persie	89	89	30000000.0	240000.0	187	71	31	1	4387	27	2126	10	11.603472	2	2.986052	76	27	5	0	9	10.081134	12.917514	1.688729	2014
Sergio Agüero	88	90	42000000.0	190000.0	172	74	26	0	619	33	2551	26	25.270160	8	5.568922	148	33	4	0	21	20.703184	27.805155	6.878173	2014
Bastian Schweinsteiger	88	88	27500000.0	180000.0	183	79	30	0	907	20	1361	5	2.174514	4	2.219083	29	31	4	0	5	2.174514	7.016763	4.264381	2014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Matthieu Udol	68	75	1100000.0	5900.0	178	68	24	1	5745	10	797	0	0.063979	1	1.524224	3	10	1	1	0	0.063979	2.182267	0.692643	2020
Felix Passlack	68	78	1400000.0	17000.0	170	74	22	0	2663	6	343	1	0.510328	0	0.240143	5	3	0	0	1	0.510328	1.801331	1.050860	2020
Andrea Favilli	68	78	1500000.0	6500.0	191	78	23	1	1997	7	267	2	1.872083	0	0.112031	13	2	1	0	2	1.872083	2.520636	0.638041	2020
Hidde ter Avest	68	76	1200000.0	5300.0	182	78	23	0	7000	6	348	0	0.000000	0	0.000000	0	0	0	0	0	0.000000	1.647105	1.647105	2020
Oliver Burke	68	73	1200000.0	15500.0	188	82	23	0	5256	12	752	0	1.377128	1	2.462660	16	13	2	0	0	1.377128	4.340086	0.500298	2020

xG - Expected goals (xG) measures the quality of a shot based on several variables such as assist type, shot angle and distance from goal, whether it was a headed shot and whether it was defined as a big chance.

xA - Expected assists (xA) measures the likelihood that a given pass will become a goal assist. It considers several factors including the type of pass, pass end point and length of the pass

npg - Non-penalty goals

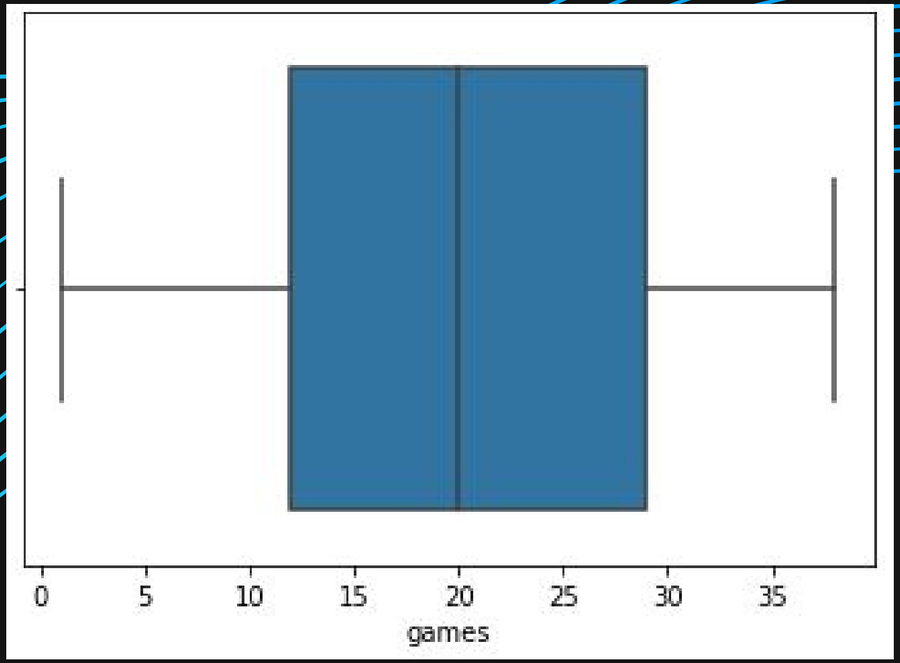
npxG - Non-penalty Expected goals

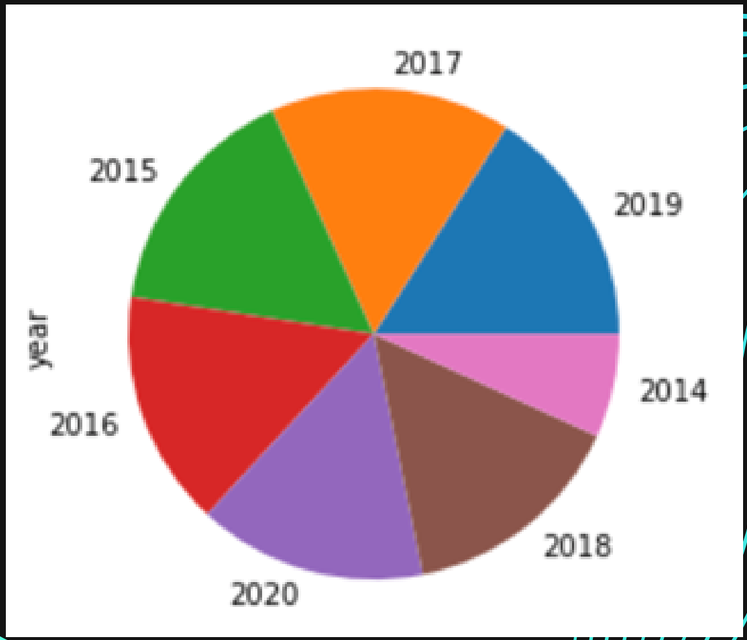
xGChain - is the total xG of every possession the player is involved in.

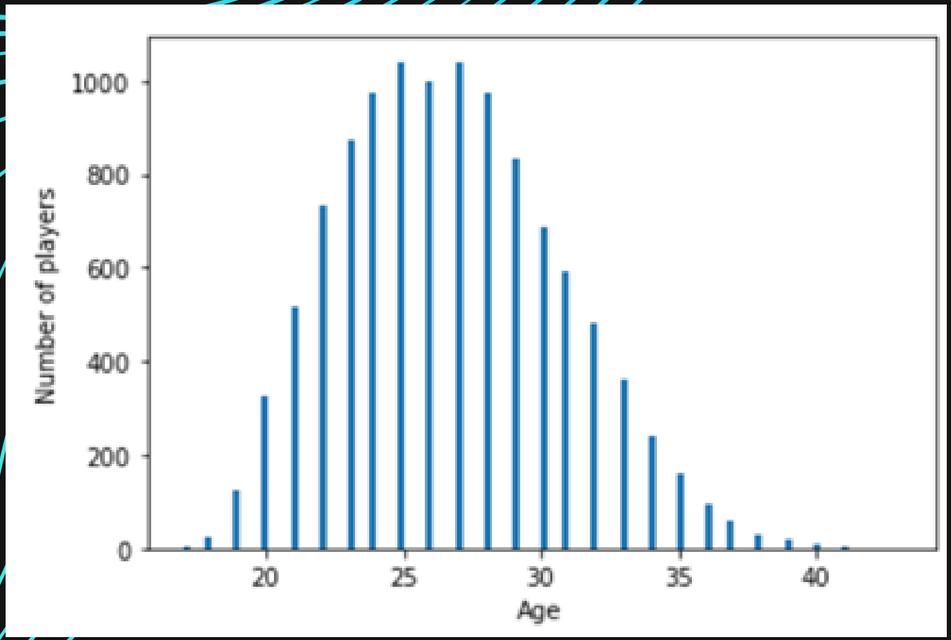
xGbuildup - is the total xG of every possession the player is involved in, excluding shots and key passes.

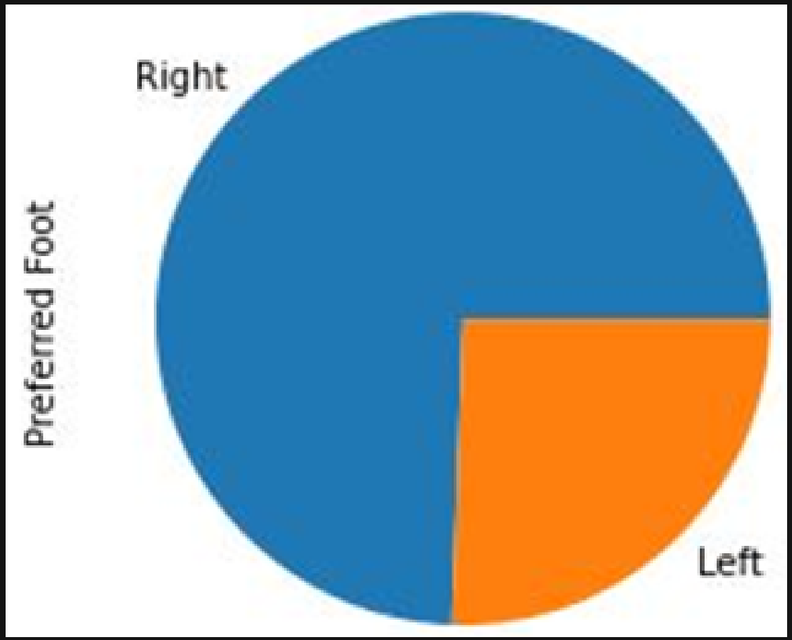
Exploratory Data Analysis:
Basic implementation

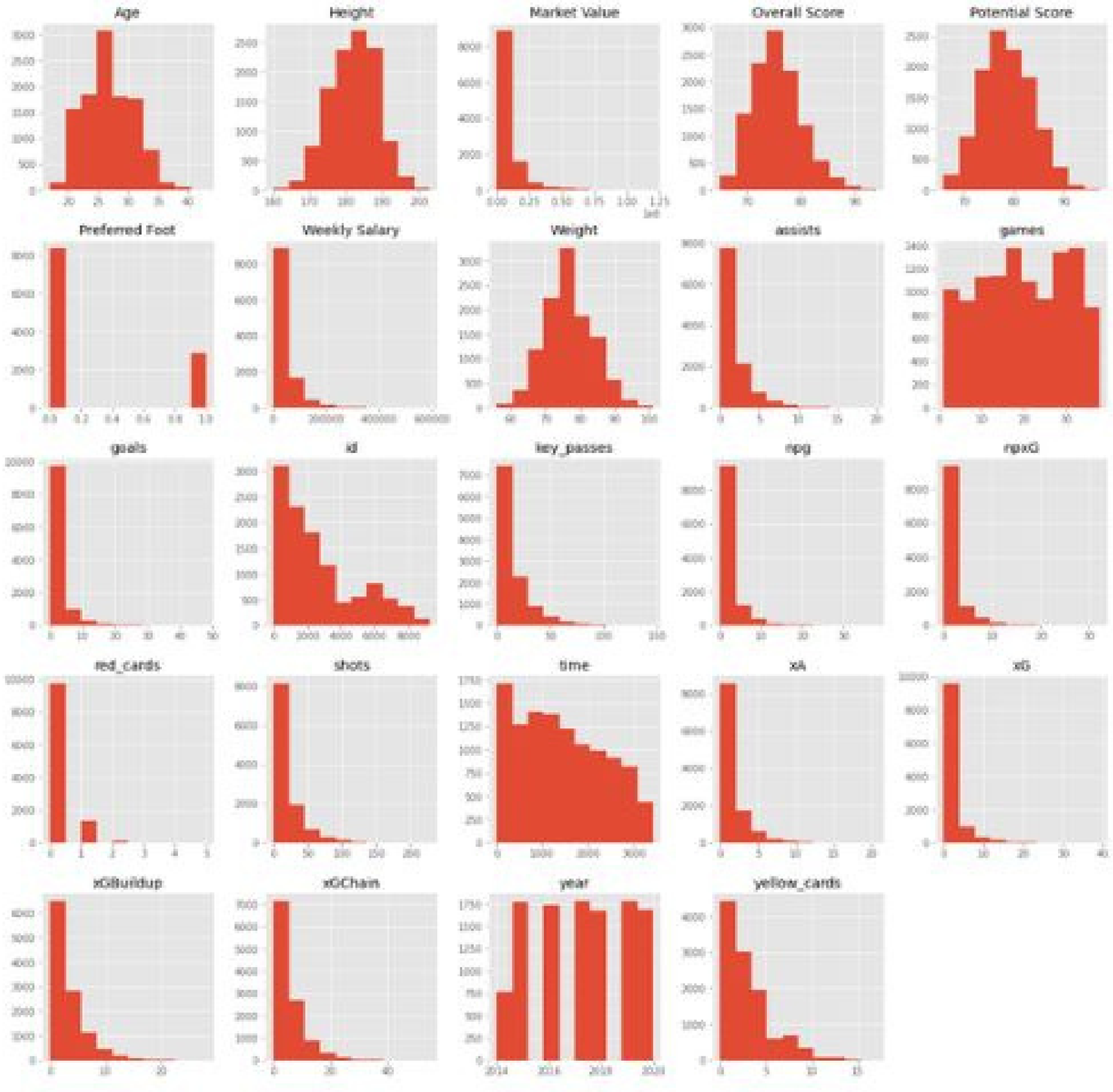
Here we can see how the database is divided by visualizations:



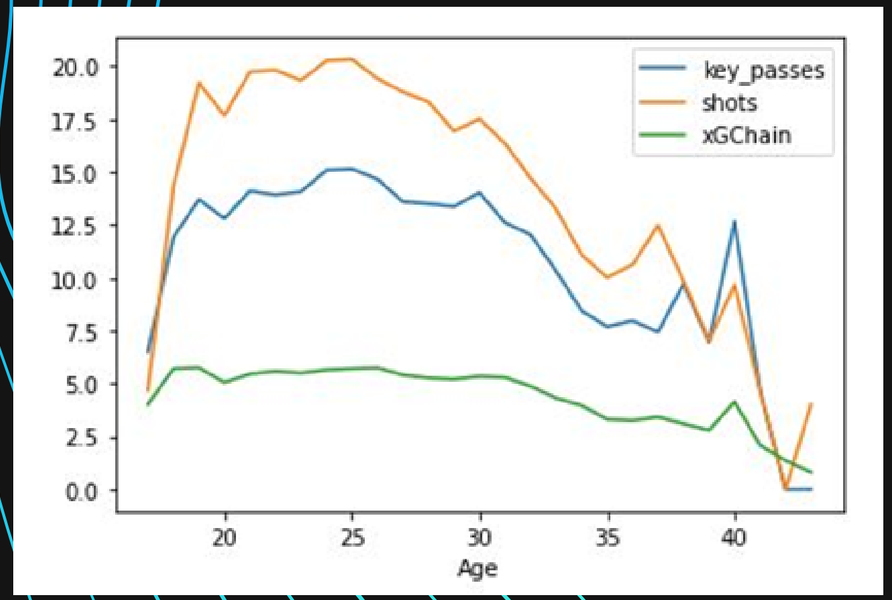


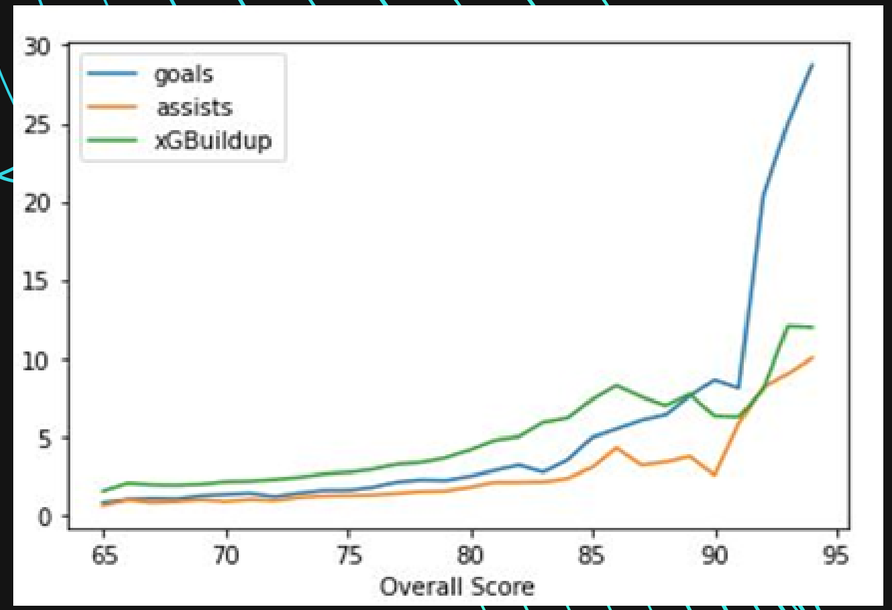


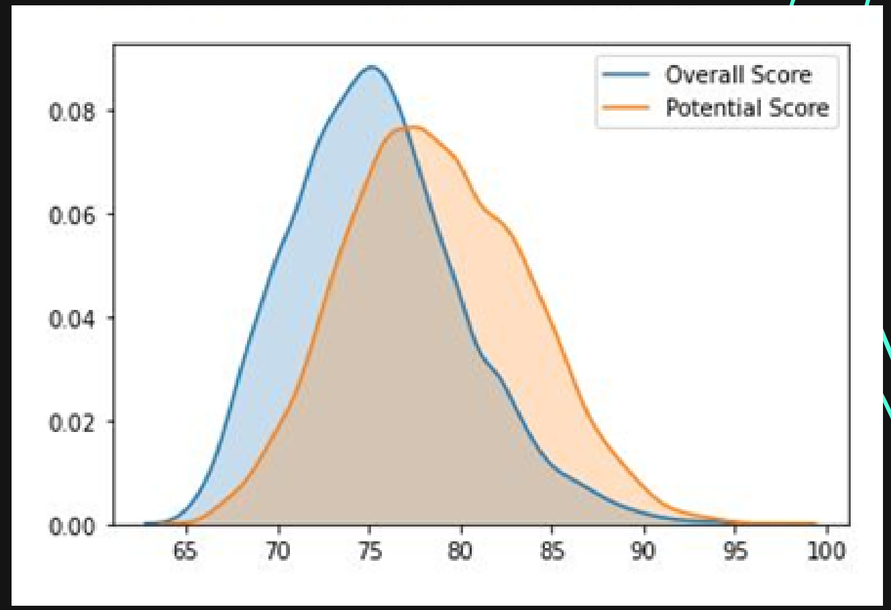



Age
Height
Market Value
Overall Score
Potential Score
Preferred Foot
Weekly Salary
Weight
assists
games
goals
id
key_passes
npg
npxG
red_cards
shots
time
xA
xG
xGBuildup
xGChain
year
yellow_cards

Two-Dimensional Graphs

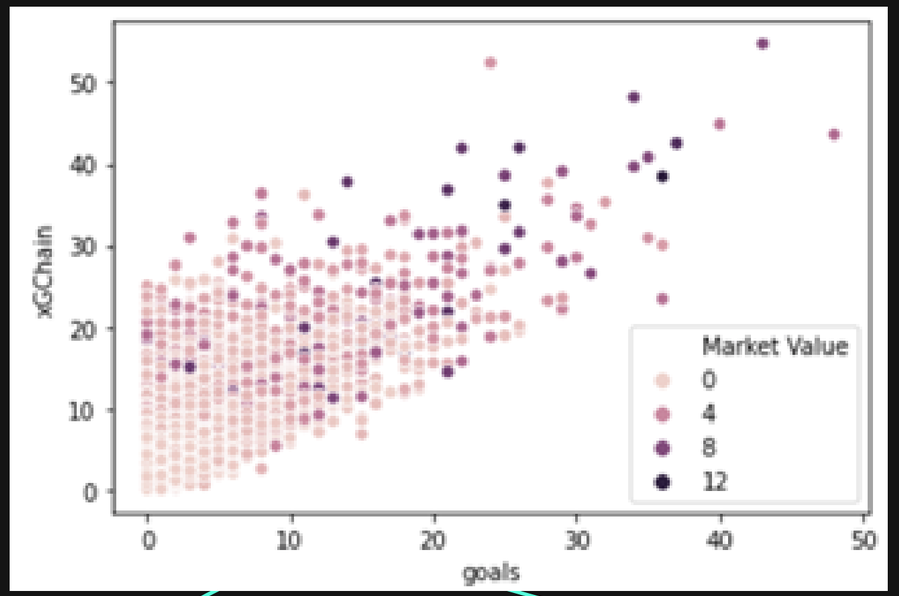


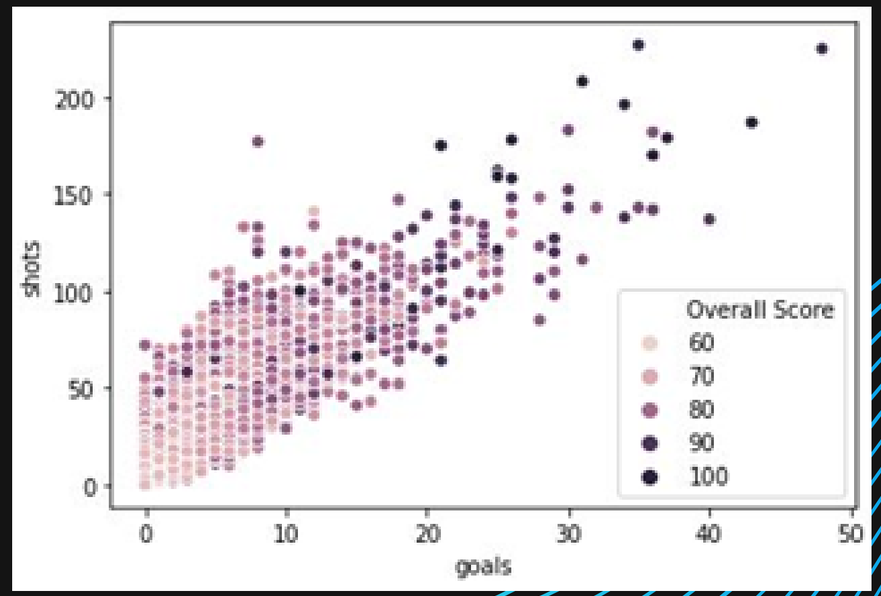


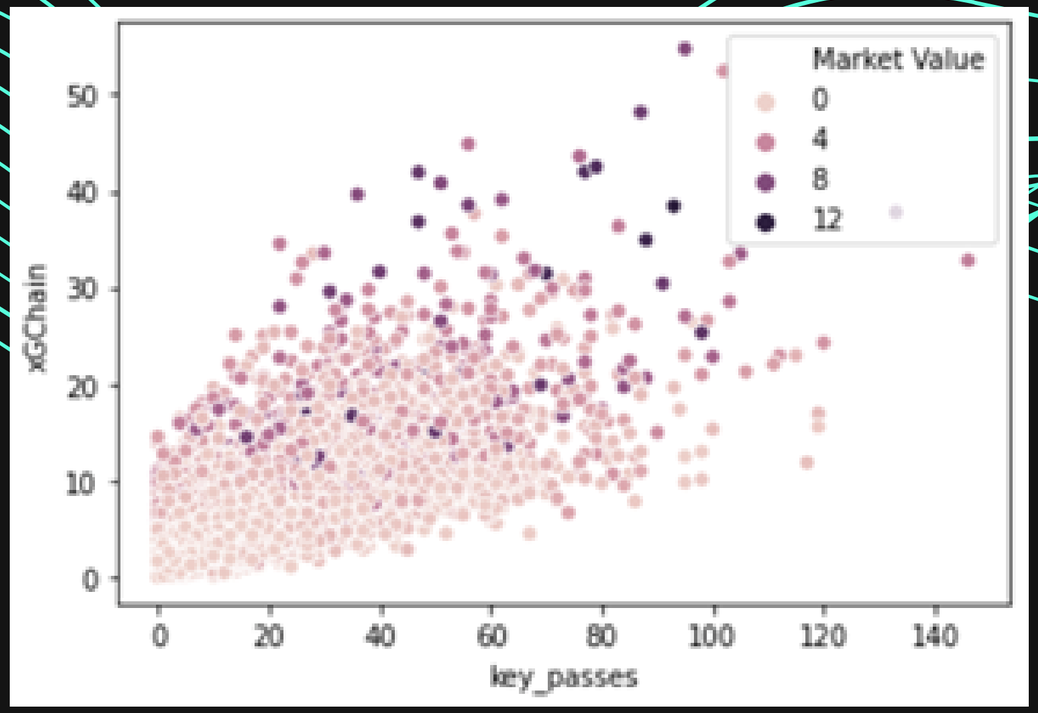


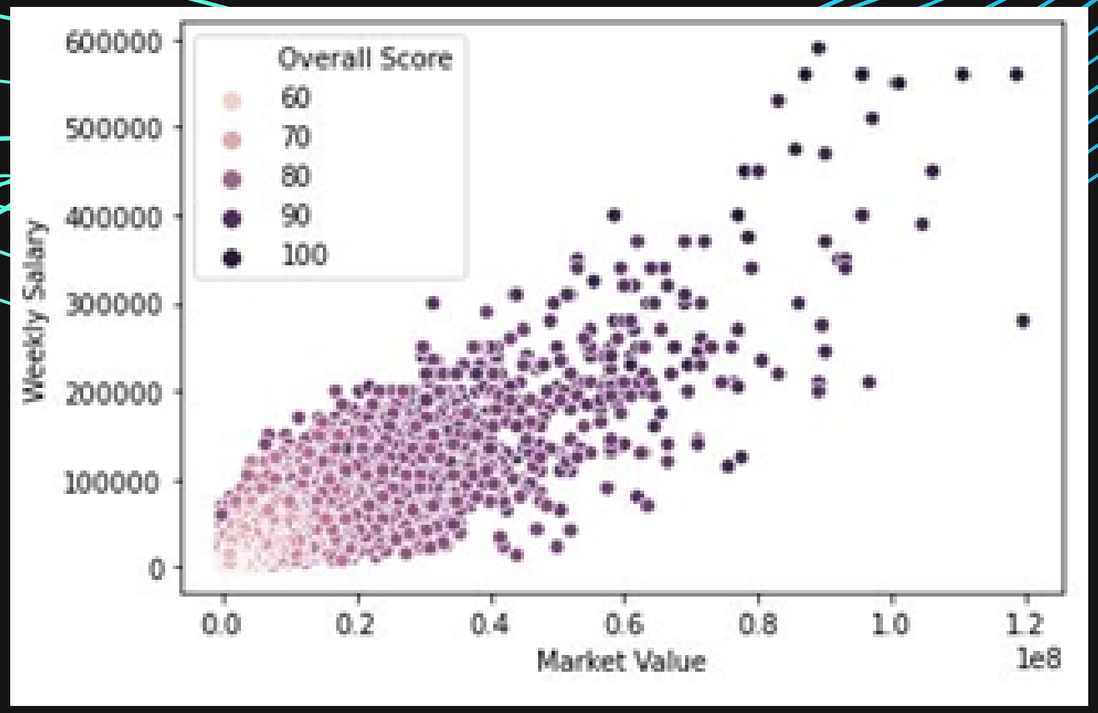
Scatter Plots

Describes the association between variables



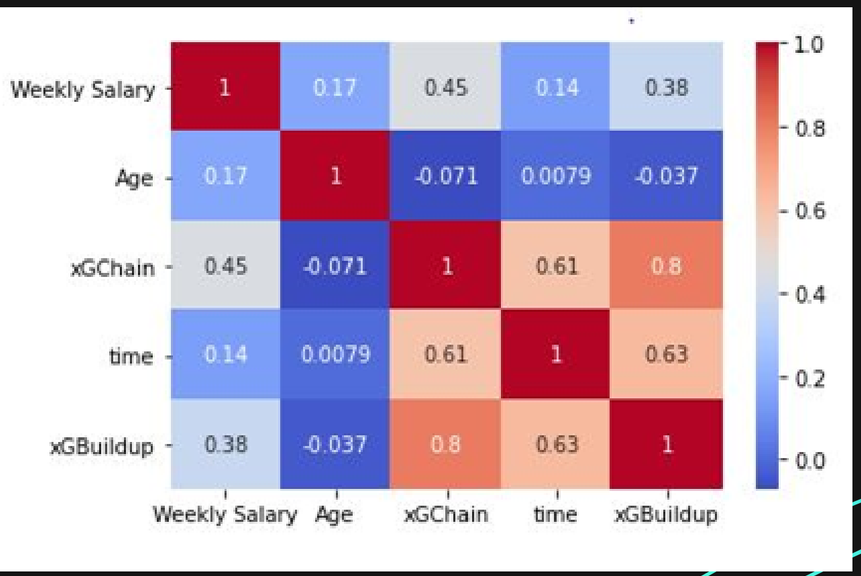


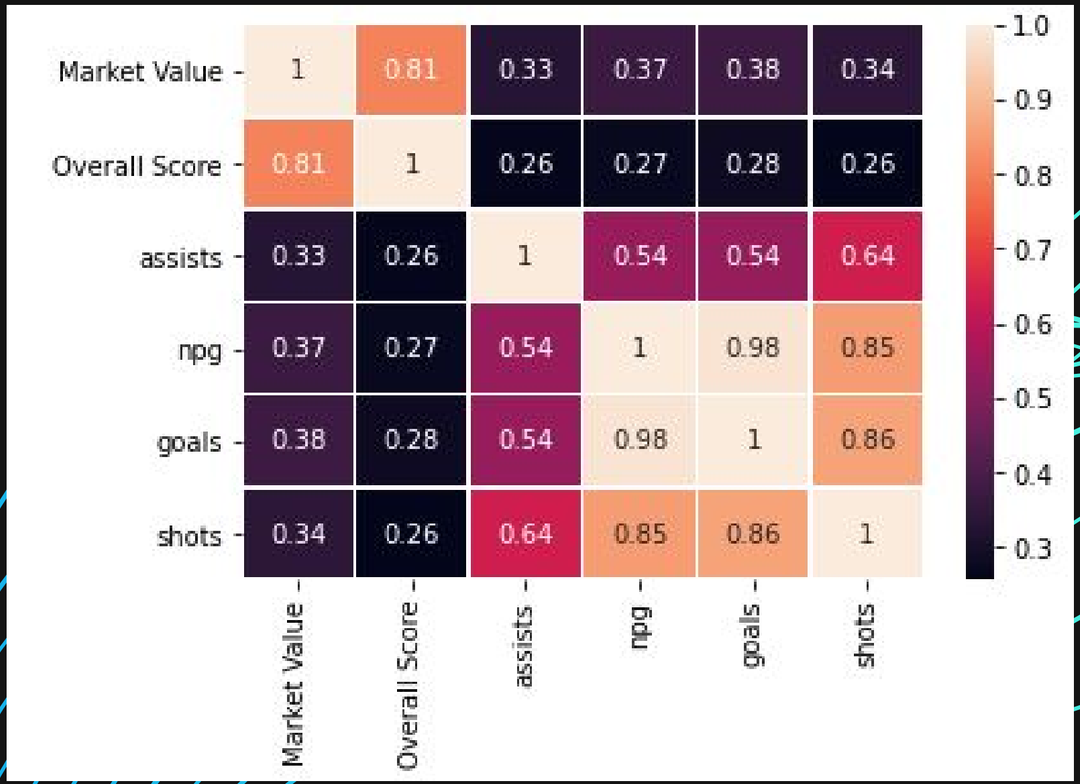


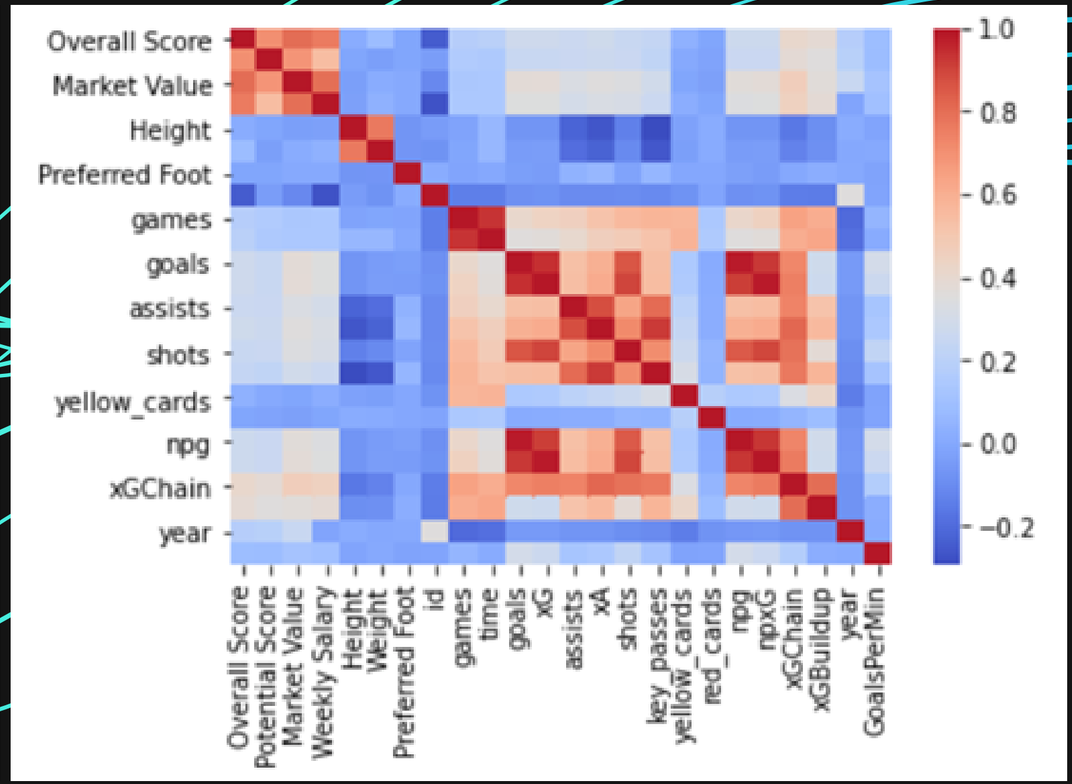


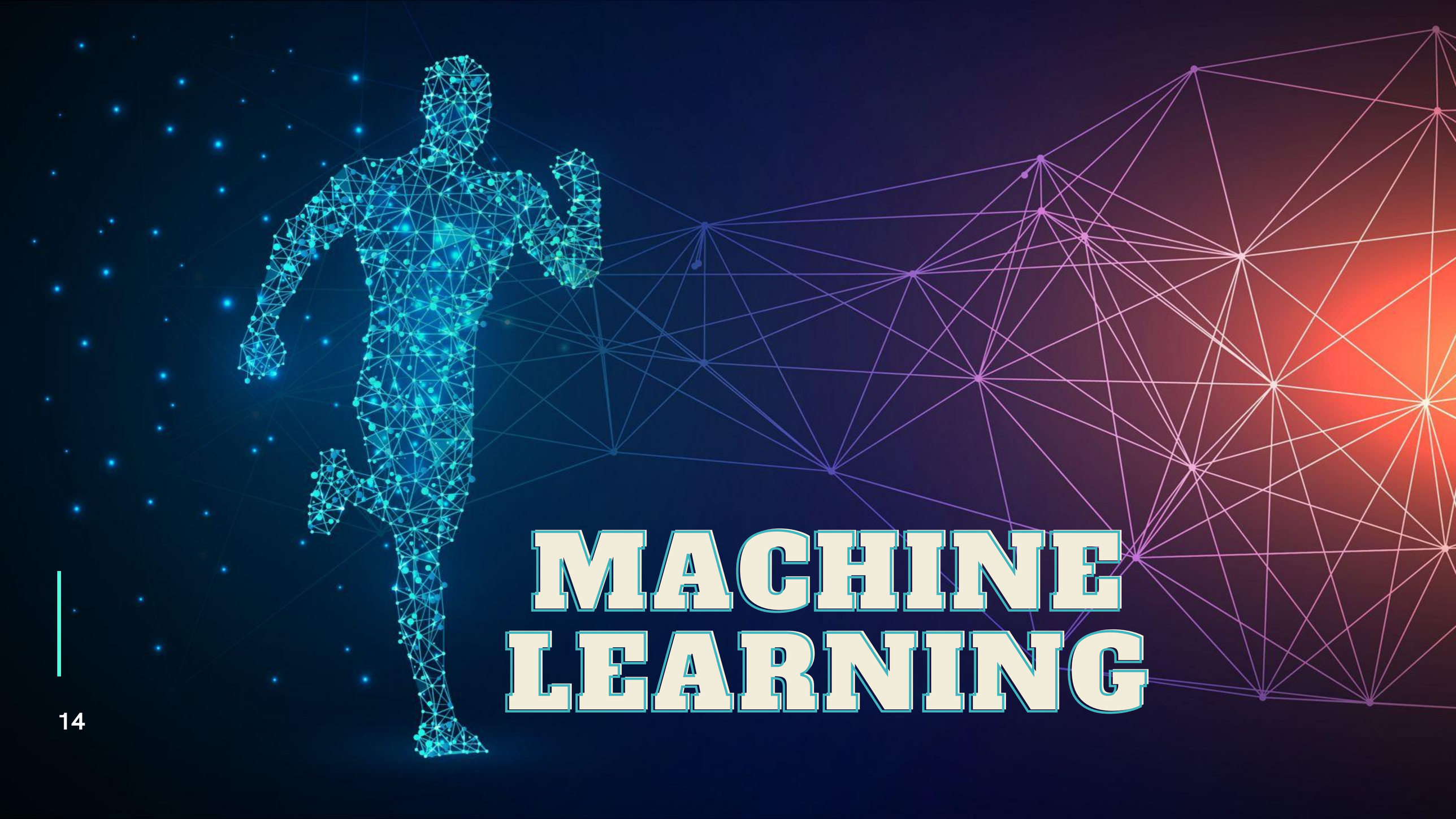
Heatmaps

Describe the correlations between features








MACHINE
LEARNING

Supervised Learning

Using “sklearn” library, the model which fits the most for predicting a player’s **market value** over the years is the Linear Regression Model.
We split the data set by years: 2014-2019 for the train set, and 2020 for the test set. It is about 80%- 20% ratio.

At first, we applied the algorithm on our raw data set to get a first perspective of the model.
The first R-Squared score that was achieved is 0.796.

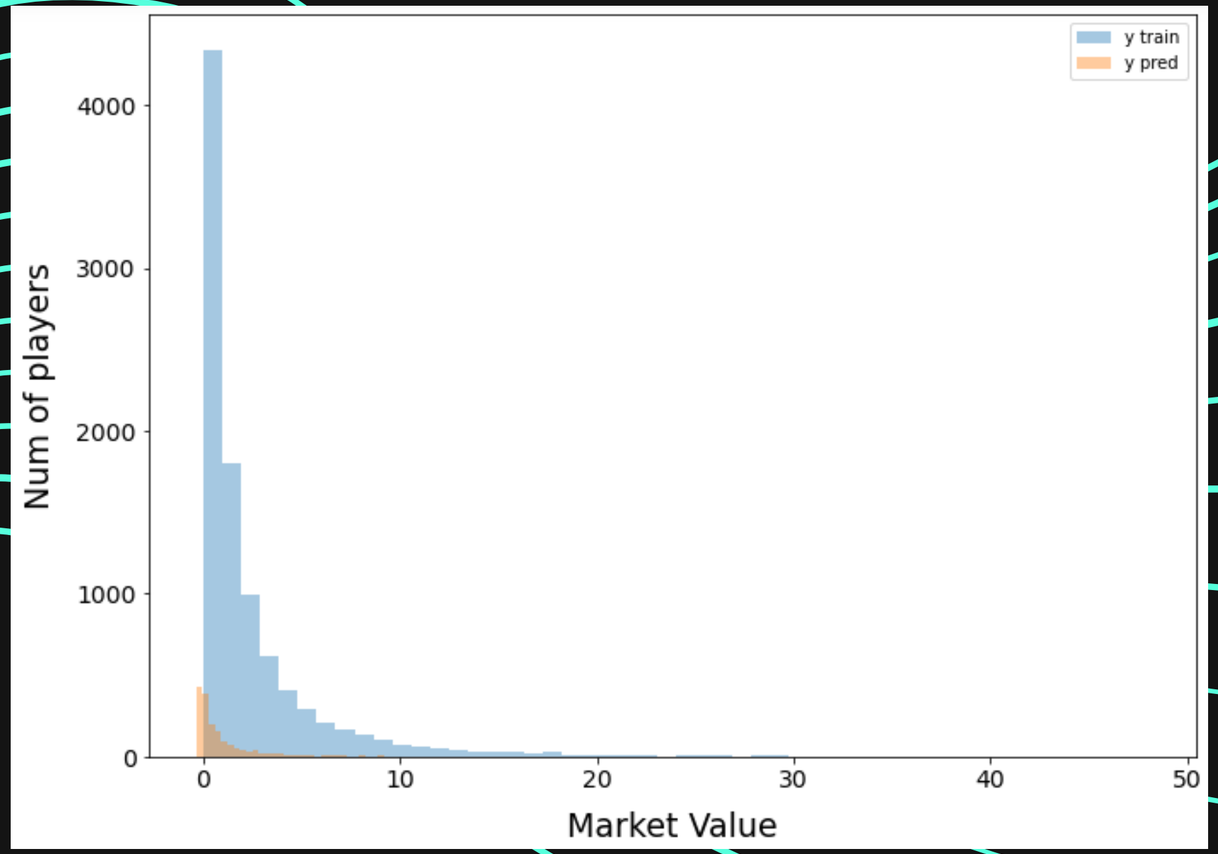
Next, we tried to improve our score by removal of features, creation of new features which contain some combinations of the original features and other manipulations.

The optimal score we got is 0.798 which means that the improvement wasn't significant.

	Player	Actual	Predicted
9516	Lionel Messi	95,500,000.000	83,946,544.315
9517	Cristiano Ronaldo	58,500,000.000	66,017,063.521
9518	Virgil van Dijk	90,000,000.000	45,866,056.203
9519	Jan Oblak	77,500,000.000	33,480,773.262
9520	Kevin De Bruyne	90,000,000.000	62,293,564.883
...	...	...	...
11202	Matthieu Udol	1,100,000.000	2,724,415.018
11203	Felix Passlack	1,400,000.000	4,876,350.552
11204	Andrea Favilli	1,500,000.000	3,191,133.441
11205	Hidde ter Avest	1,200,000.000	2,716,524.134
11206	Oliver Burke	1,200,000.000	4,578,136.703

Mean Absolute Error: 0.5407897090876934
Mean Squared Error: 0.7637838166663377
Root Mean Squared Error: 0.8739472619479608

In advanced, we tried to predict how many league **goals** will a player score in 2020 season. For this purpose we applied again the algorithm. The optimal R-Squared score we've got is **0.813**

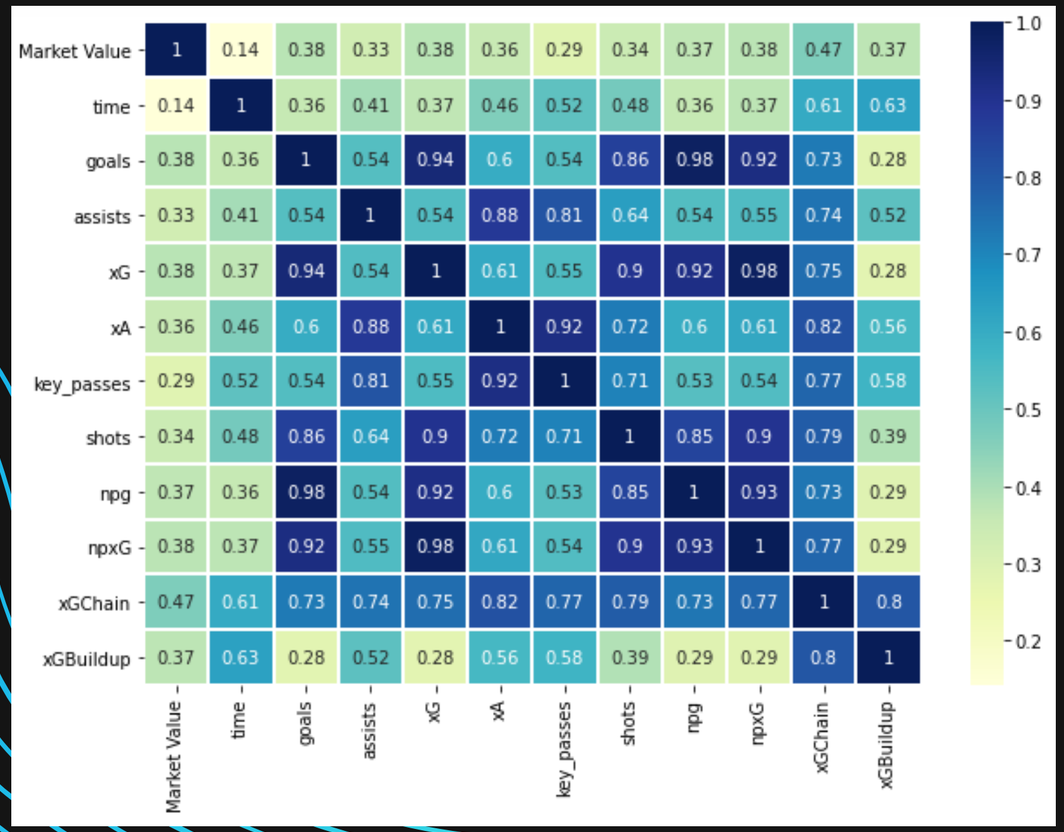


Player	Actual Goals	Predicted Goals
Robert Lewandowski	21	15.525
Lionel Messi	11	14.163
Cristiano Ronaldo	15	13.480
Jamie Vardy	11	13.267
André Silva	12	12.007
Romelu Lukaku	12	11.936
Harry Kane	12	11.876
Patrick Bamford	10	10.796
Mohamed Salah	13	10.658
Wout Weghorst	11	10.508
Memphis Depay	11	10.506
Wissam Ben Yedder	9	10.423
Bruno Fernandes	11	10.408
Dominic Calvert-Lewin	11	10.153
Callum Wilson	8	9.635
Ludovic Ajorque	9	9.485
Neal Maupay	7	9.232

Conclusions

1

Features that have the highest impact on a player's market value.



2

Although the score we got - 0.798 is not a bad result at all, we understand that there are probably other components that were not taken into account which affect the market value of a player, for example influence on social networks, merchandise and more...

3

With predicting goals we had a higher success, mainly because there is a close correlation between the features in our data and the number of goals a player will score.

WOW

OH WOW!

The 20th of April, 2020 was a kindly morning.
Drinking my coffee, I am getting a notification of a new mail.
Aston villa's scouting team heard about our new project and asked us to find an attacking player, to join their team for the 20/21 season.
Aston Villa FC is a famous English football club that competes in the Premier League, and we were more than happy to help them.
They asked for a player that plays as a striker, but can also play as a winger, at a specific age, to help them score more goals and to lead the team.

Their requirements were very clear:

1) Age: 24-28
2) Position: Attack
3) Goals: at least 10 last season
4) Transfer budget: 20M€
5) Salary budget: 70K€ per week.
6) Bonus: England Citizenship, Total transaction cost (low).

8 BEST MATCHING PLAYERS

We created for Aston Villa scouting team a list of 8 best players that match their requirements

Player	Overall Score	Potential Score	Market Value	Weekly Salary	Height	Weight	Age	Preferred Foot	id	games	time	goals	xG	assists	xA	shots	key_passes	yellow_cards	red_cards	position	team_title	npg	npxG	xGChain	xGBuildup	year
Gerard Moreno	82	82	23,500,000	49,000	180	77	28	Left	2120	16	1419	10	9	2	3	49	22	3	0	F M	Villarreal	6	6	12	4	2020
Wout Weghorst	81	81	19,500,000	60,000	197	84	28	Right	7052	16	1362	11	10	1	2	42	19	3	0	F S	Wolfsburg	9	8	11	3	2020
Luis Muriel	80	80	15,500,000	70,000	179	79	29	Right	1229	15	480	10	7	2	1	36	10	0	0	F S	Atalanta	9	6	9	2	2020
Andy Delort	79	79	13,500,000	41,000	181	83	28	Right	3530	18	1510	9	6	6	3	54	31	3	0	F	Montpellier	9	6	9	2	2020
André Silva	78	83	14,000,000	38,000	184	78	24	Right	6170	15	1285	12	12	2	3	53	13	0	0	F	Eintracht Frankfurt	7	8	11	2	2020
João Pedro	78	78	11,000,000	38,000	184	73	28	Right	1592	17	1514	10	7	2	2	37	16	1	0	F M	Cagliari	8	5	6	3	2020
Ludovic Ajorque	77	80	11,500,000	31,000	197	82	26	Left	6905	19	1599	9	9	4	2	38	20	4	0	F S	Strasbourg	6	6	9	1	2020
Patrick Bamford	72	75	4,000,000	43,000	185	71	27	Left	822	18	1556	10	11	2	1	65	9	1	0	F	Leeds	9	10	13	2	2020

OH WOW!

One player stood out above them all - Patrick Bamford.
Although his Overall score is the lowest, his stats are on the same level and his market value is significantly lower.
The Englishman is currently 27 years old and plays as a Striker for Leeds United which also competes in the Premier League in England.

Our machine learning model has predicted that Patrick Bamford would score between 10 to 11 goals in 2020, and indeed Patrick met the target.
By the way, not a bad statistic at all, compared to Lionel Messi who scored 11 goals at the same period.

Hat-trick Patrick!

- Aston Villa did start talks with Leeds United to sign Patrick Bamford, but eventually decided to drop the deal.
  In the last match-up between the teams that was played this season, Leeds United defeated Aston Villa 3-0, when Patrick Bamford scored a hat-trick, proved that the prediction was successful, and even took the ball home!

- This season (still running) Patrick Bamford has already scored 12 goals & 6 assists in 22 games and his current market value is 15M€.
  Good luck Patrick!

Let's work together

Dotan Gotshtein
Ohad Mavdali
HIT - Holon Institute of Technology

GitHub Link:
https://github.com/ohadmavdali
https://github.com/Dotangot

Contact Us:
www.linkedin.com/in/ohadmavdali
www.linkedin.com/in/dotangot