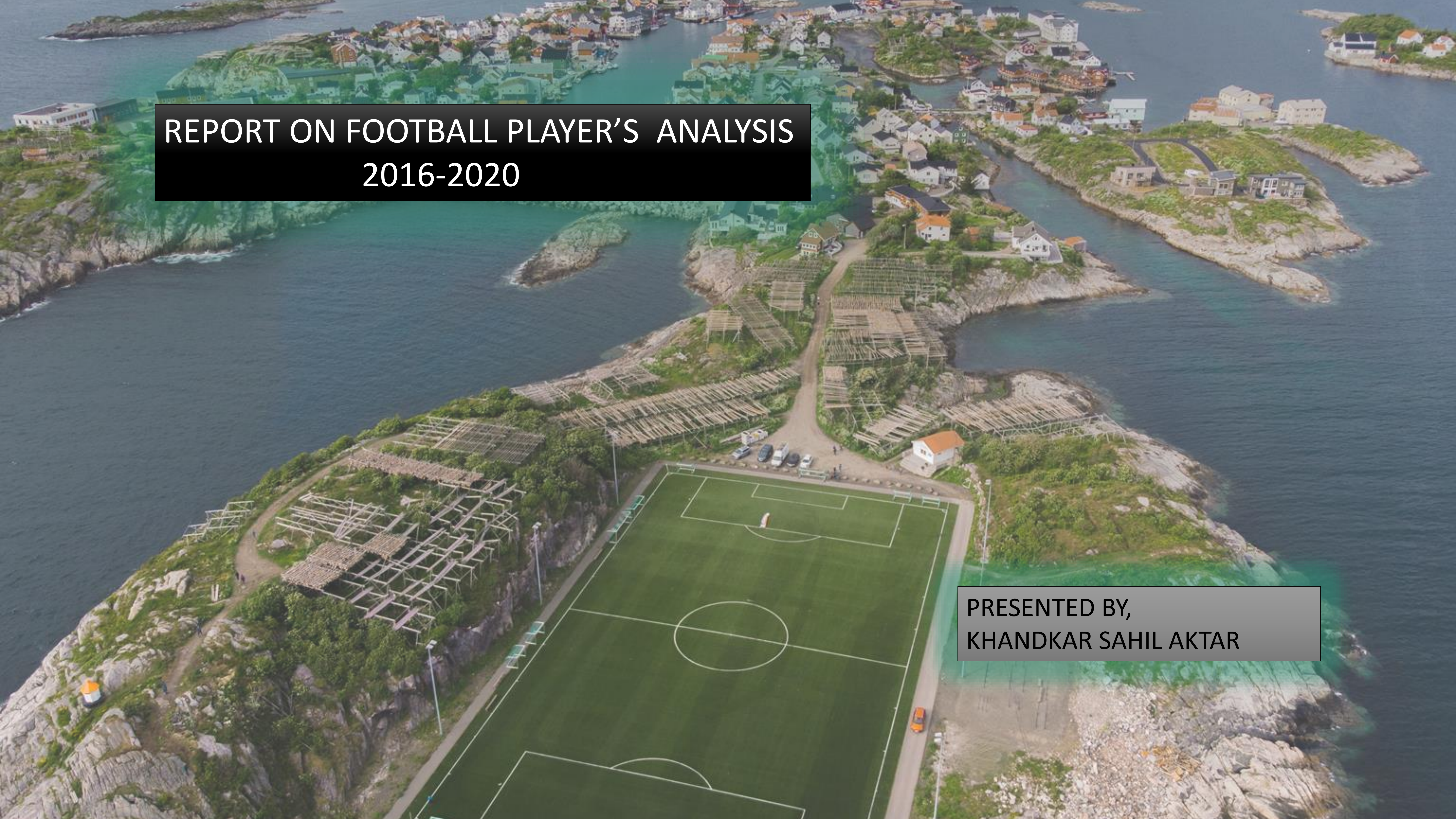


REPORT ON FOOTBALL PLAYER’S ANALYSIS 2016-2020

PRESENTED BY,
KHANDKAR SAHIL AKTAR

Football is a sport played between two teams of eleven players with a spherical ball. It is played by 250 million players in over 200 countries, making it the world's most popular sport.
The object of the game is to score goals by using any part of the body besides the arms and hands to get the football into the opposing goal.

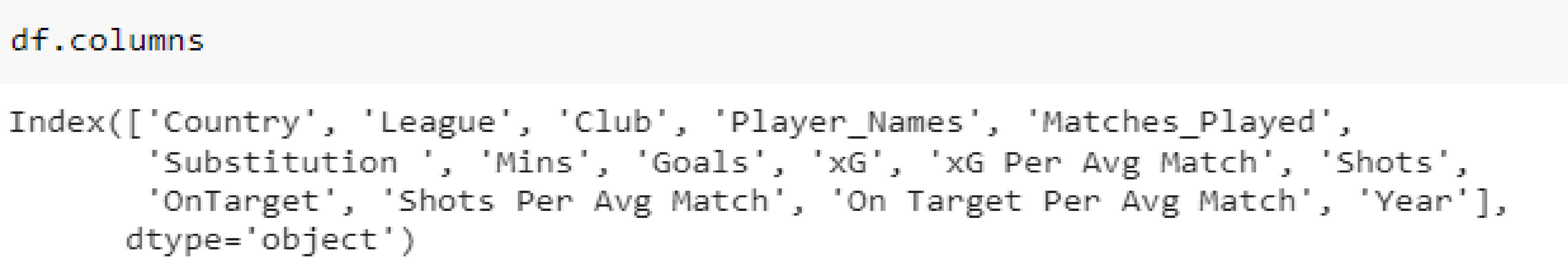
OBJECTIVE:

- The objective is to analyse the players performance.
- Top 5 Goal Scorers.
- Which Year was best.

ABOUT DATASET:

The dataset is collected from kaggle which is scraped from top website Infogol.
Infogol has league tables and statistics from some of the top competitions from all around the world, including the English Premier League, English Championship, Spanish La Liga, Italian Serie A, German Bundesliga, French Ligue 1, US MLS and Brazilian Série A.

The dataset includes columns such as:

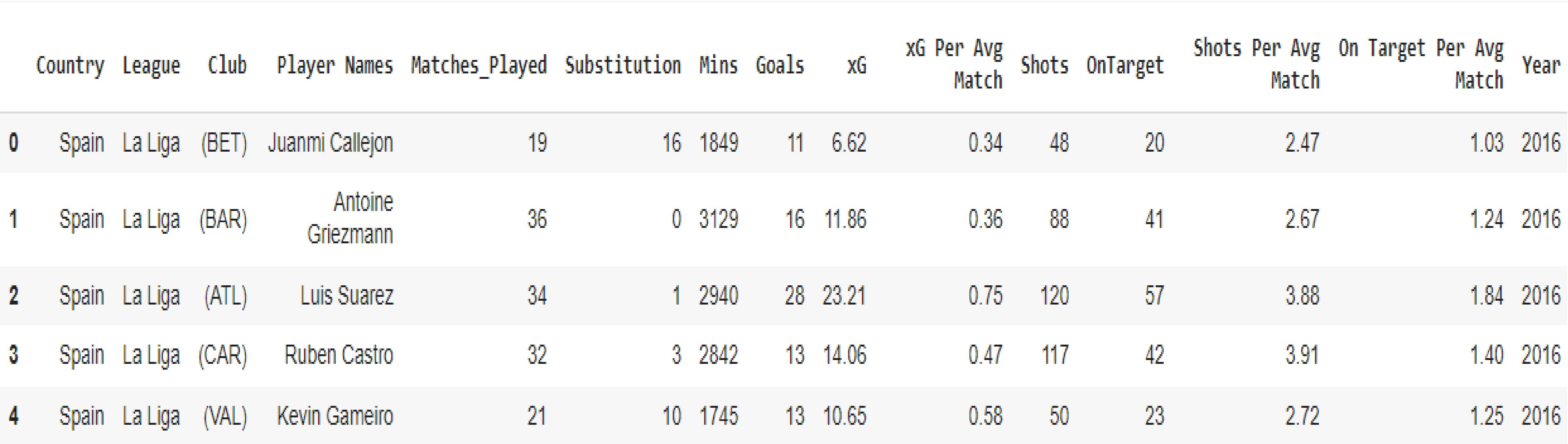


```
df.columns
```

```
Index(['Country', 'League', 'Club', 'Player_Names', 'Matches_Played',
       'Substitution ', 'Mins', 'Goals', 'xG', 'xG Per Avg Match', 'Shots',
       'OnTarget', 'Shots Per Avg Match', 'On Target Per Avg Match', 'Year'],
      dtype='object')
```

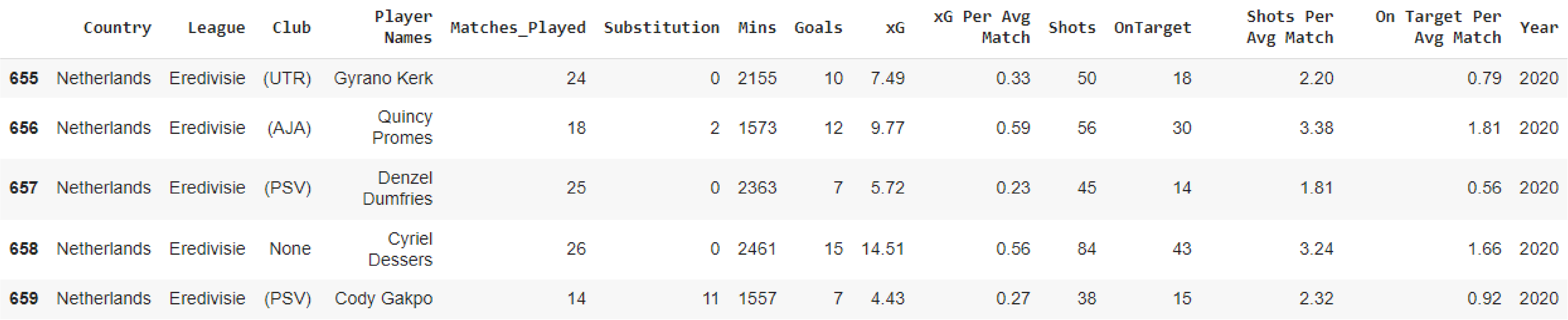
INSTANCE OF THE DATASET:

First five records:



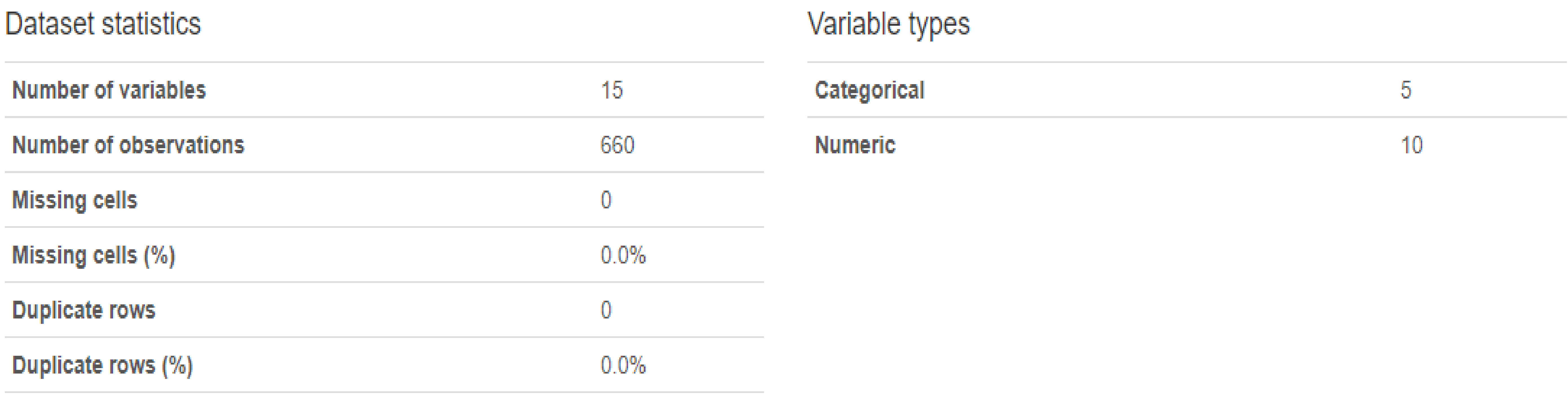
	Country	League	Club	Player Names	Matches_Played	Substitution	Mins	Goals	xG	xG Per Avg Match	Shots	OnTarget	Shots Per Avg Match	On Target Per Avg Match	Year
0	Spain	La Liga	(BET)	Juanmi Callejon	19	16	1849	11	6.62	0.34	48	20	2.47	1.03	2016
1	Spain	La Liga	(BAR)	Antoine Griezmann	36	0	3129	16	11.86	0.36	88	41	2.67	1.24	2016
2	Spain	La Liga	(ATL)	Luis Suarez	34	1	2940	28	23.21	0.75	120	57	3.88	1.84	2016
3	Spain	La Liga	(CAR)	Ruben Castro	32	3	2842	13	14.06	0.47	117	42	3.91	1.40	2016
4	Spain	La Liga	(VAL)	Kevin Gameiro	21	10	1745	13	10.65	0.58	50	23	2.72	1.25	2016

Last five records:



	Country	League	Club	Player Names	Matches_Played	Substitution	Mins	Goals	xG	xG Per Avg Match	Shots	OnTarget	Shots Per Avg Match	On Target Per Avg Match	Year
655	Netherlands	Eredivisie	(UTR)	Gyrano Kerk	24	0	2155	10	7.49	0.33	50	18	2.20	0.79	2020
656	Netherlands	Eredivisie	(AJA)	Quincy Promes	18	2	1573	12	9.77	0.59	56	30	3.38	1.81	2020
657	Netherlands	Eredivisie	(PSV)	Denzel Dumfries	25	0	2363	7	5.72	0.23	45	14	1.81	0.56	2020
658	Netherlands	Eredivisie	None	Cyriel Dessers	26	0	2461	15	14.51	0.56	84	43	3.24	1.66	2020
659	Netherlands	Eredivisie	(PSV)	Cody Gakpo	14	11	1557	7	4.43	0.27	38	15	2.32	0.92	2020

OVERVIEW OF THE DATASET:



Dataset statistics

Number of variables	15
Number of observations	660
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Variable types

Categorical	5
Numeric	10

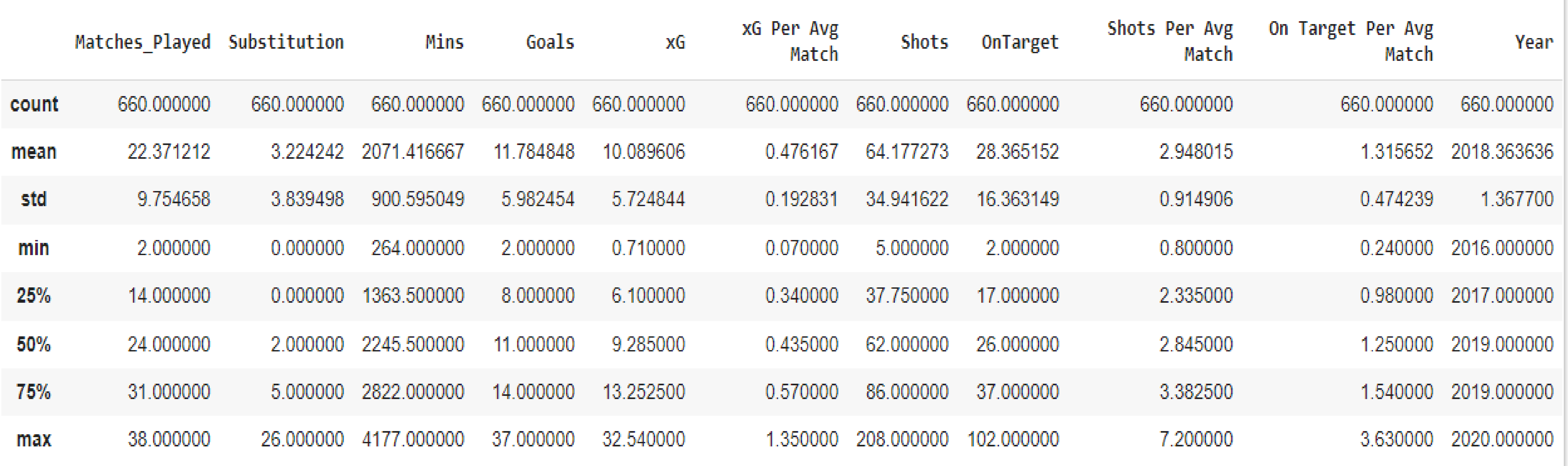
EXPLORATORY DATA ANALYSIS

EDA is one of the most important phases in data analysis since it helps us to obtain critical insights and statistical metrics. In general, EDA can be categorised in two ways.
The first distinction is that each method is either non-graphical or graphical.
Second, each method is univariate or multivariate in nature (usually just bivariate).

Analysis of the data:

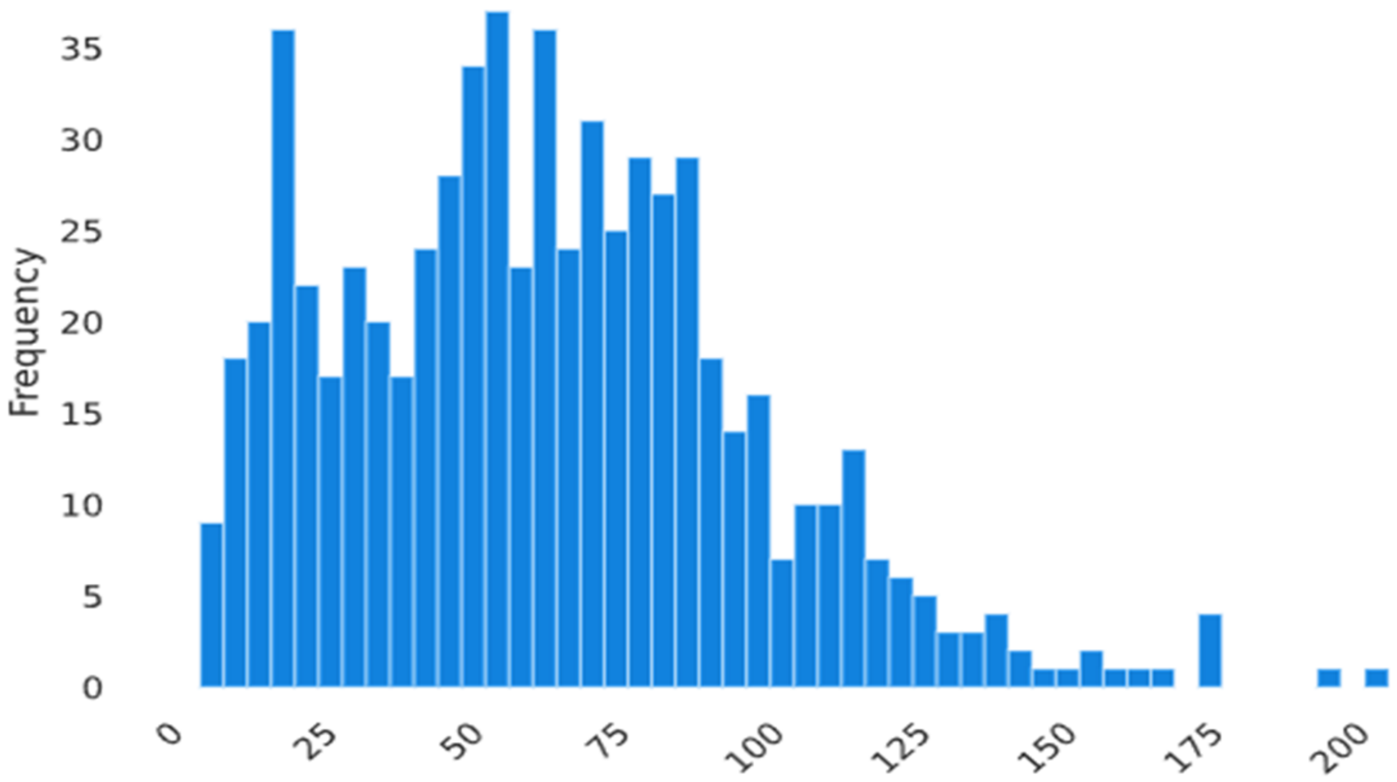
DESCRIPTIVE STATISTICS

Numerical Features

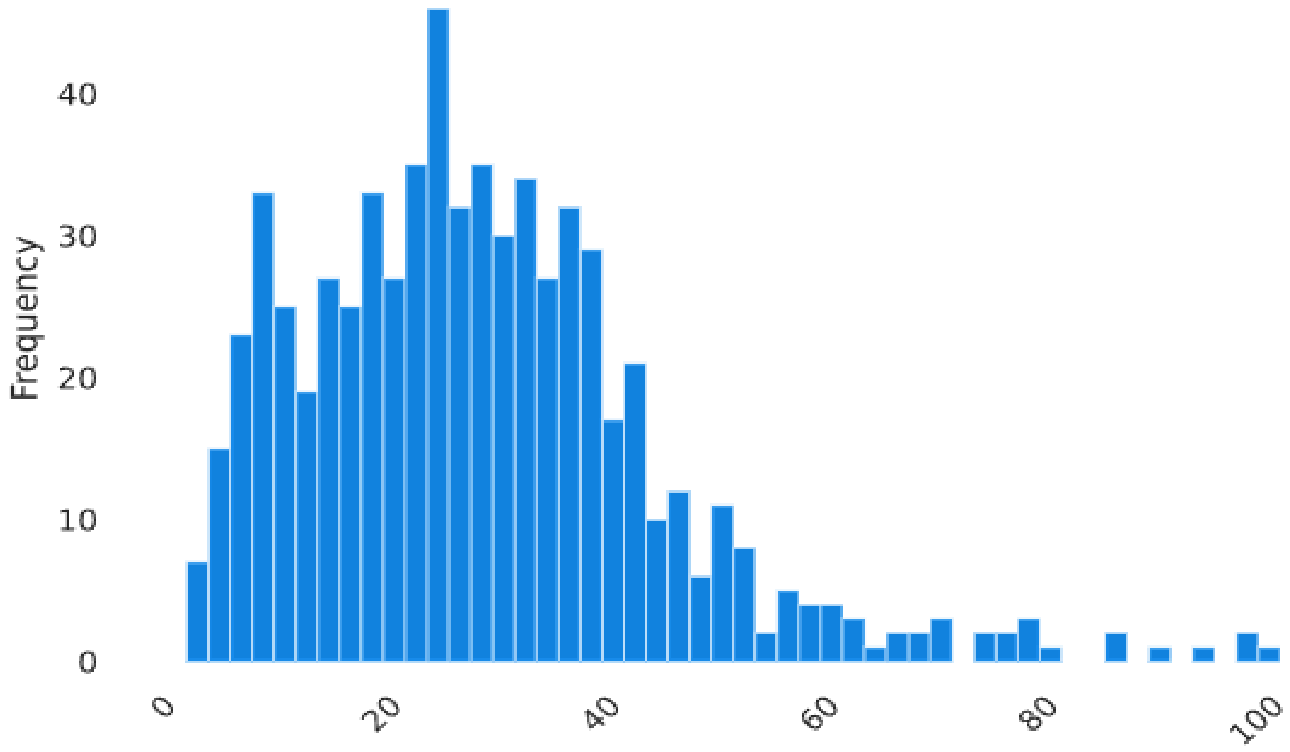


	Matches_Played	Substitution	Mins	Goals	xG	xG Per Avg Match	Shots	OnTarget	Shots Per Avg Match	On Target Per Avg Match	Year
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000
mean	22.371212	3.224242	2071.416667	11.784848	10.089606	0.476167	64.177273	28.365152	2.948015	1.315652	2018.363636
std	9.754658	3.839498	900.595049	5.982454	5.724844	0.192831	34.941622	16.363149	0.914906	0.474239	1.367700
min	2.000000	0.000000	264.000000	2.000000	0.710000	0.070000	5.000000	2.000000	0.800000	0.240000	2016.000000
25%	14.000000	0.000000	1363.500000	8.000000	6.100000	0.340000	37.750000	17.000000	2.335000	0.980000	2017.000000
50%	24.000000	2.000000	2245.500000	11.000000	9.285000	0.435000	62.000000	26.000000	2.845000	1.250000	2019.000000
75%	31.000000	5.000000	2822.000000	14.000000	13.252500	0.570000	86.000000	37.000000	3.382500	1.540000	2019.000000
max	38.000000	26.000000	4177.000000	37.000000	32.540000	1.350000	208.000000	102.000000	7.200000	3.630000	2020.000000

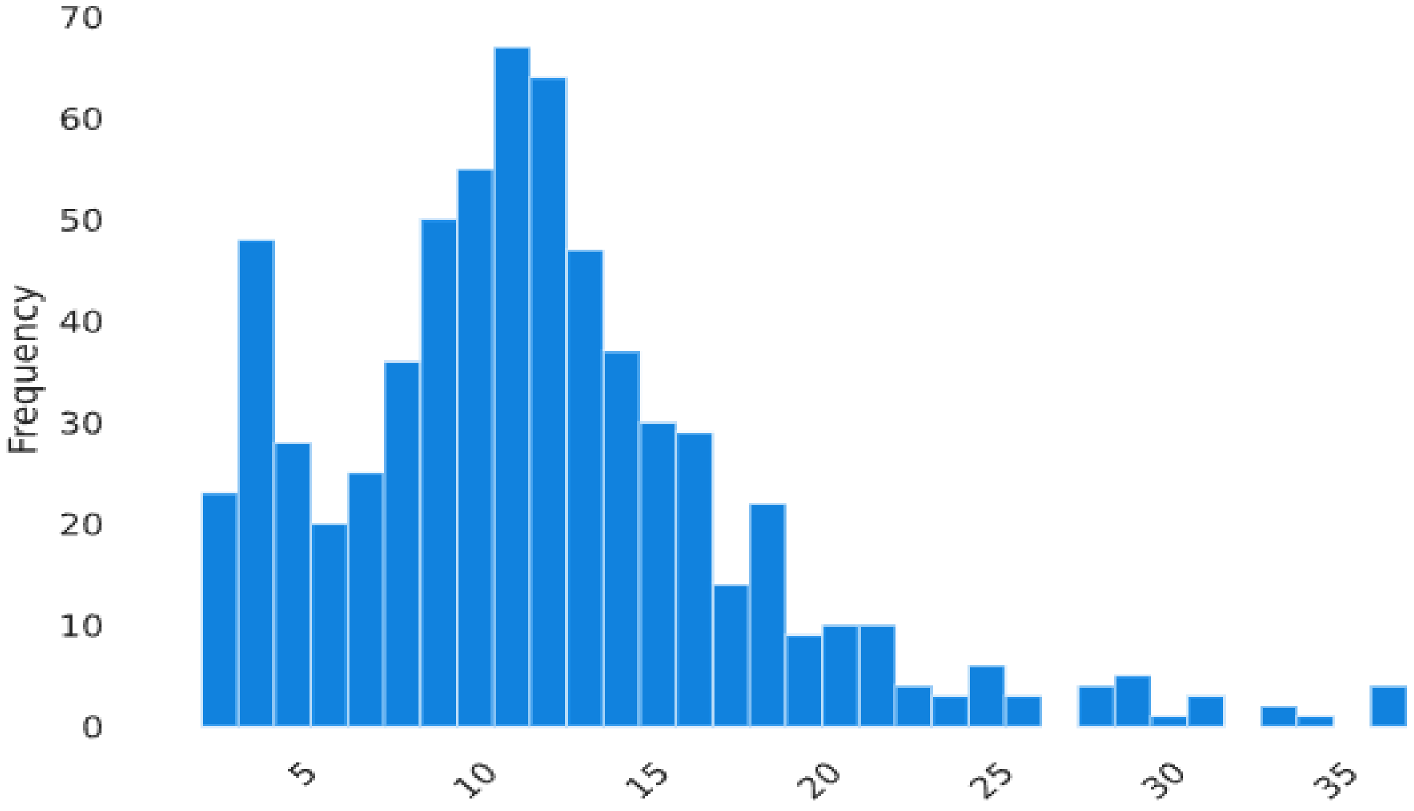
Frequency of the number of shots taken by player



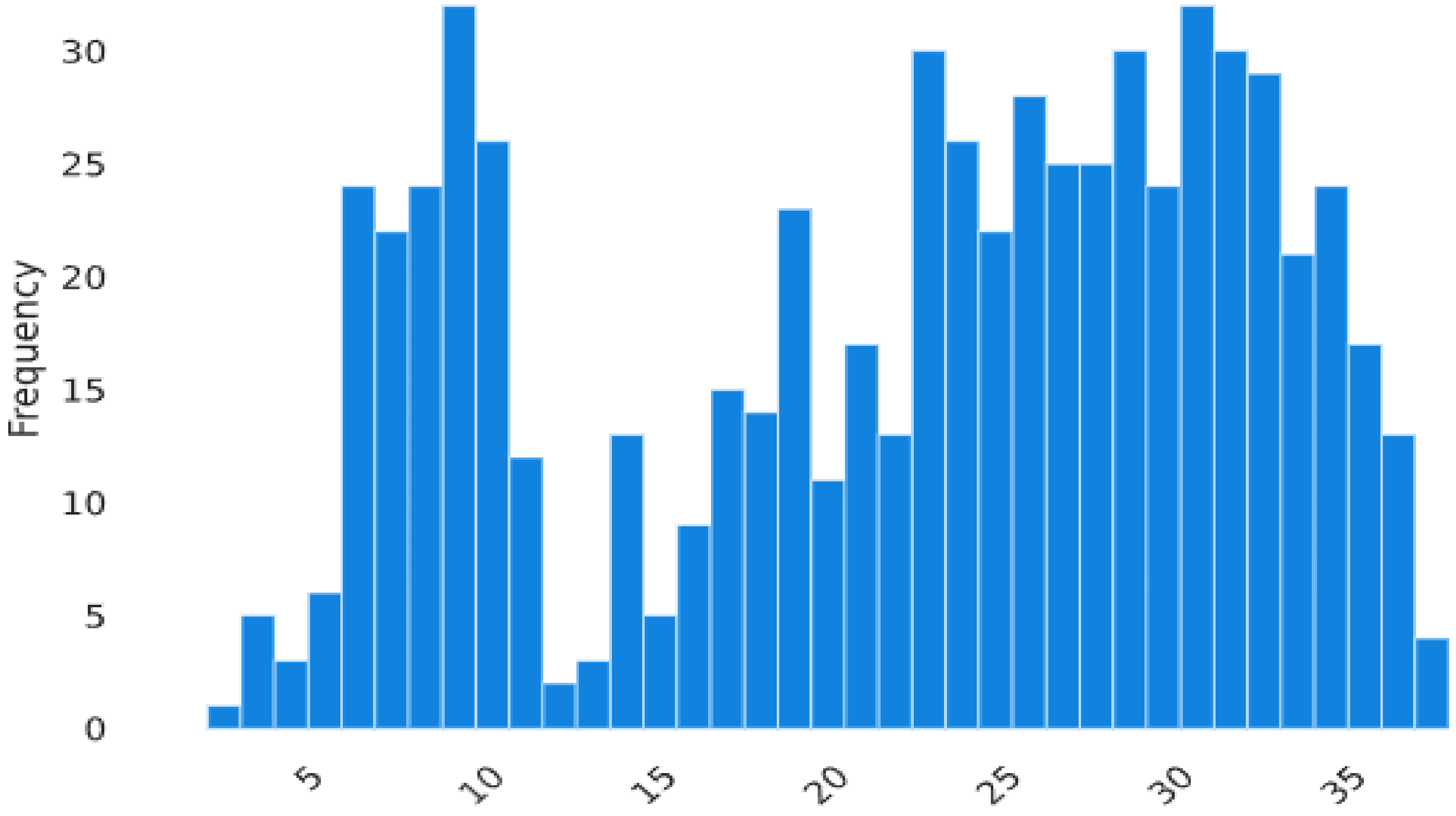
Frequency of the number of Shots On Target



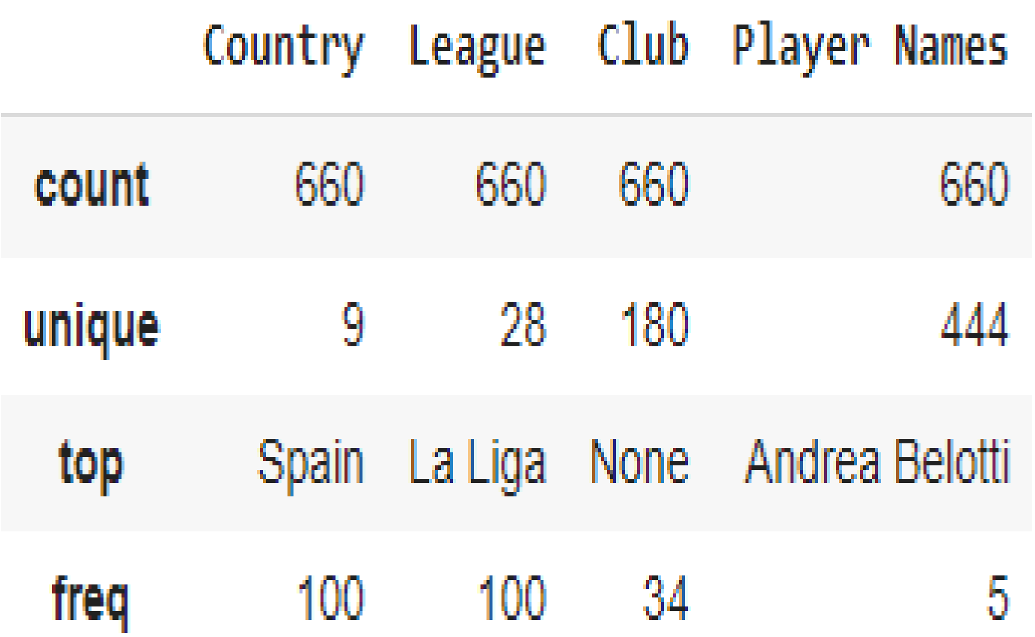
Frequency of the number of Goals scored by the players



Frequency of the number of Matches played by all the players

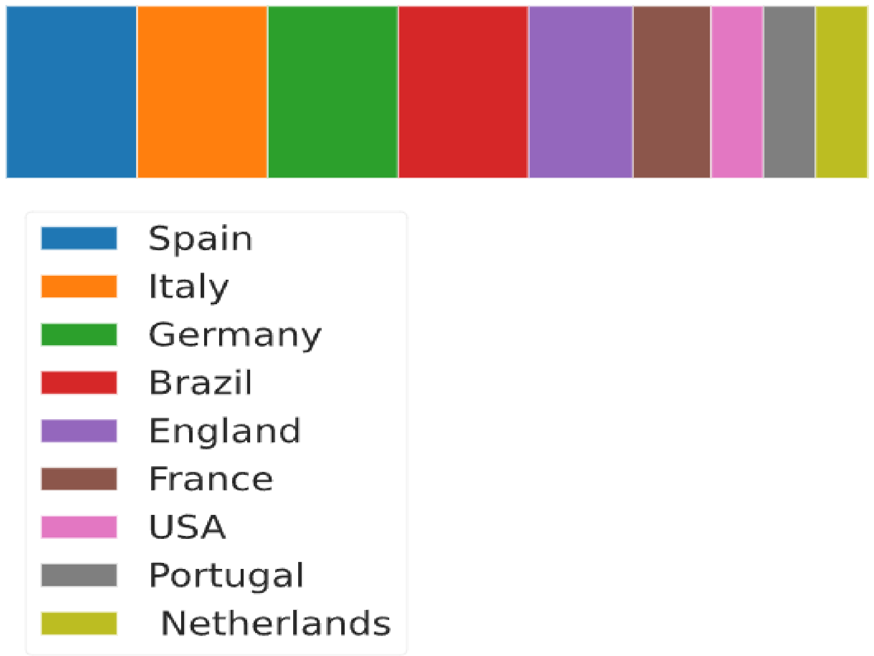


Categorical Features

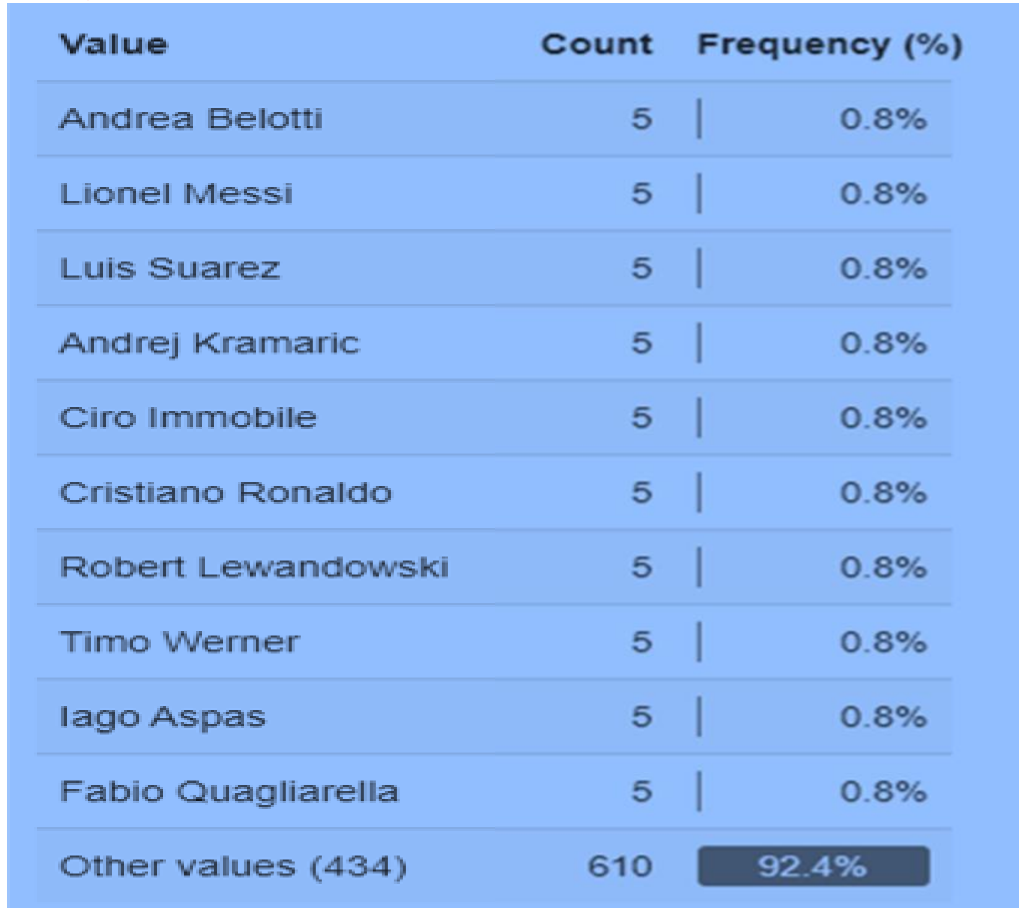


	Country	League	Club	Player Names
count	660	660	660	660
unique	9	28	180	444
top	Spain	La Liga	None	Andrea Belotti
freq	100	100	34	5

Country represented by the players

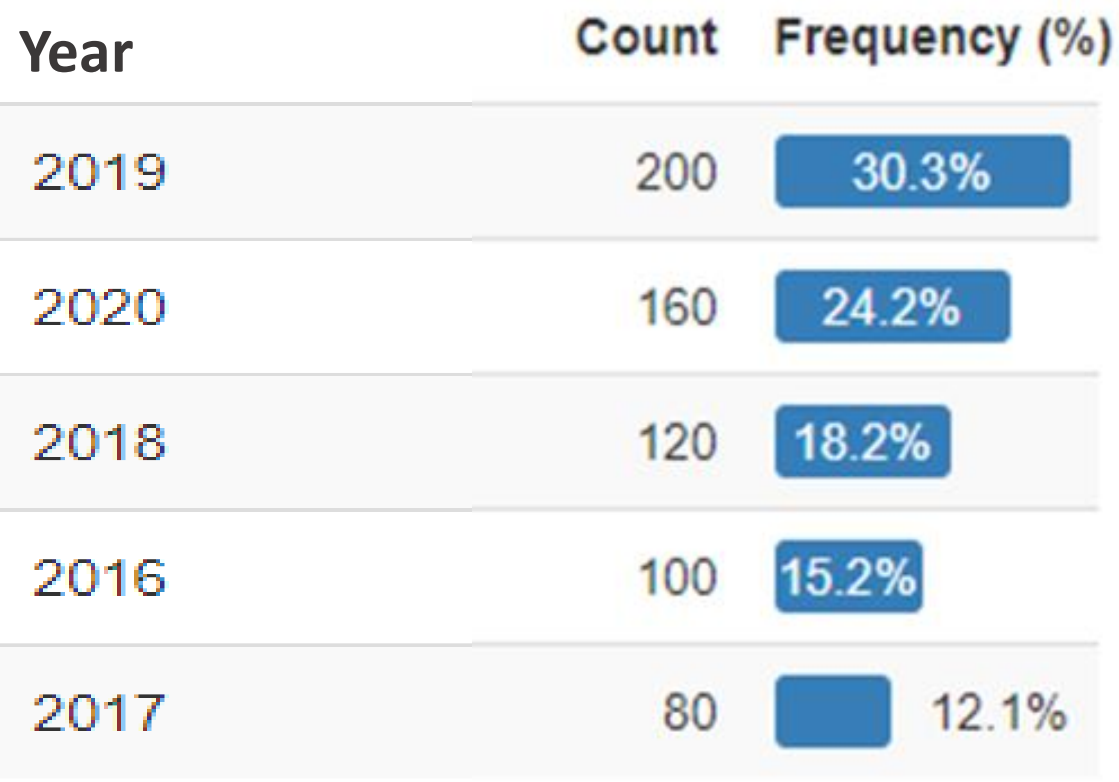


Player names



Value	Count	Frequency (%)
Andrea Belotti	5	0.8%
Lionel Messi	5	0.8%
Luis Suarez	5	0.8%
Andrej Kramaric	5	0.8%
Ciro Immobile	5	0.8%
Cristiano Ronaldo	5	0.8%
Robert Lewandowski	5	0.8%
Timo Werner	5	0.8%
Iago Aspas	5	0.8%
Fabio Quagliarella	5	0.8%
Other values (434)	610	92.4%

Years Count



Year	Count	Frequency (%)
2019	200	30.3%
2020	160	24.2%
2018	120	18.2%
2016	100	15.2%
2017	80	12.1%

OBSERVATIONS:

Matches Played: Lowest number of matches played by any player is 2 and maximum matches played by any player is 38 from 2016-2020.

Goals: Minimum goal scored by any player is 2 and maximum goal scored by any player is 37 from 2016-2020.

Shots: Minimum shots taken by any player from 2016-2020 is 5 and maximum shot taken is 208.

On Target: Lowest shot on target is 2 while maximum of 102 shots are on target from 2016-2020.

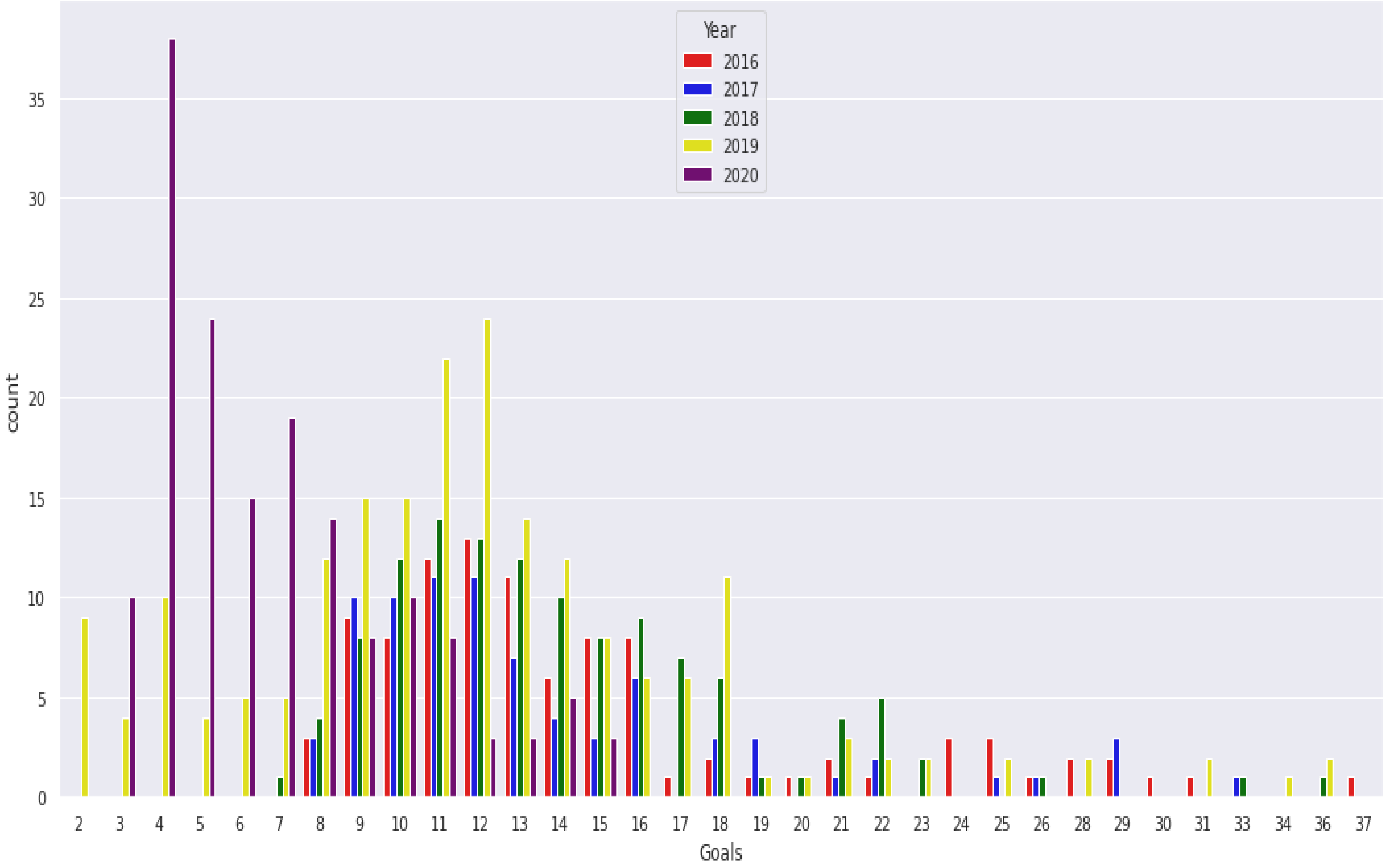
Country: Players from 9 countries were playing from which Spain, Italy, Germany and Brazil players were highest in apperance.

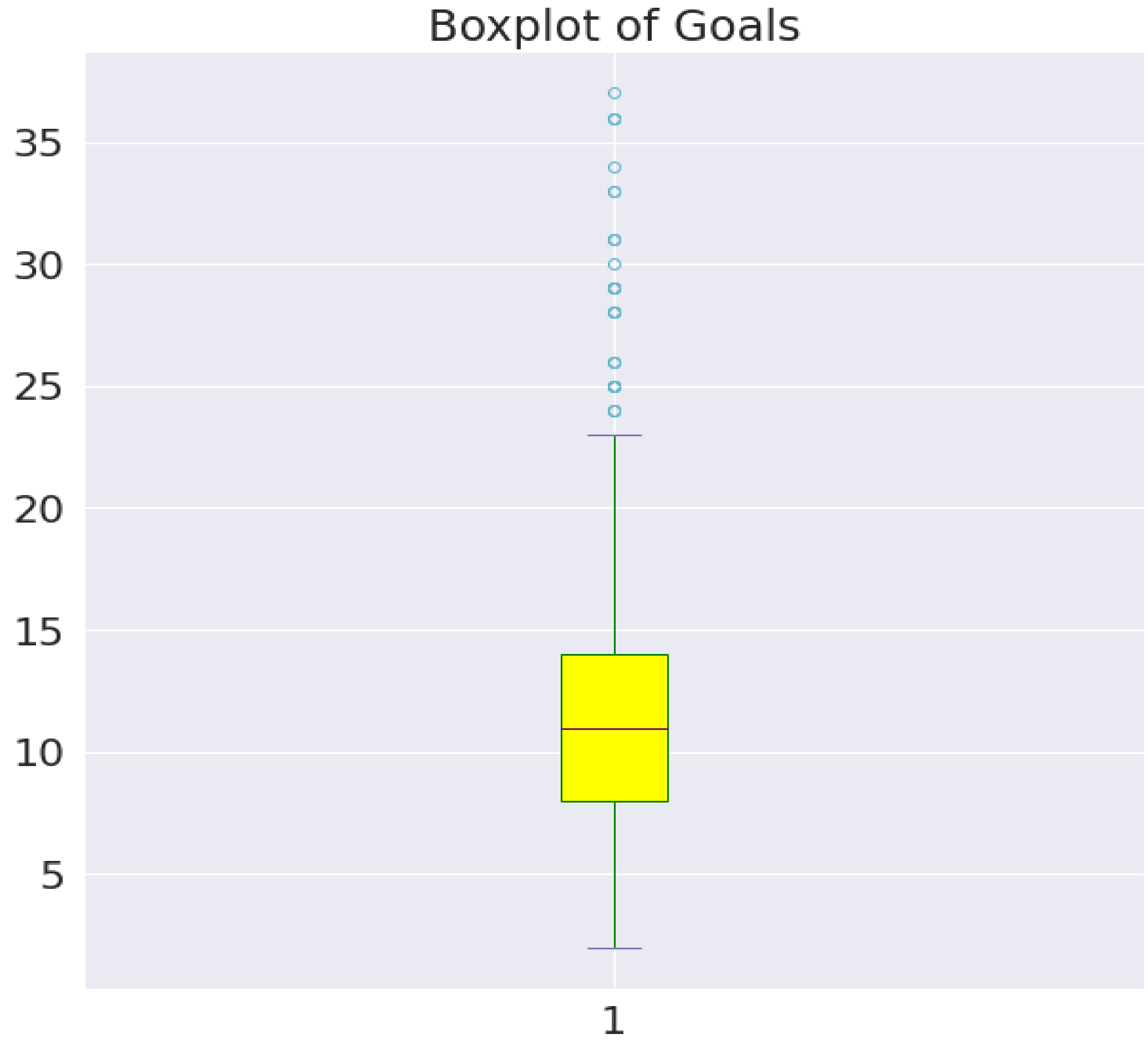
Player names: The count of the Player names shows the appearance of the player in each year. Eg. Andrea Belloti , Lionel Messi, Cristiano Ronaldo played in all five years.

xG: Expected goals reflects the average probability of scoring a goal with an individual attempt on goal.

Univariate Analysis

GOALS:

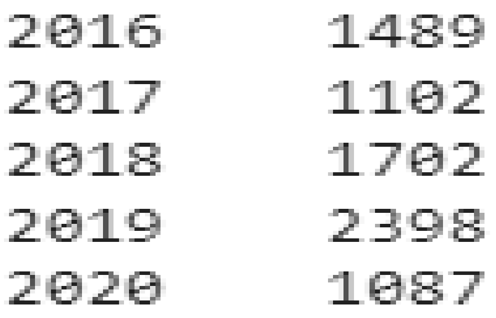




Total of 7778 Goals is scored by all the players.

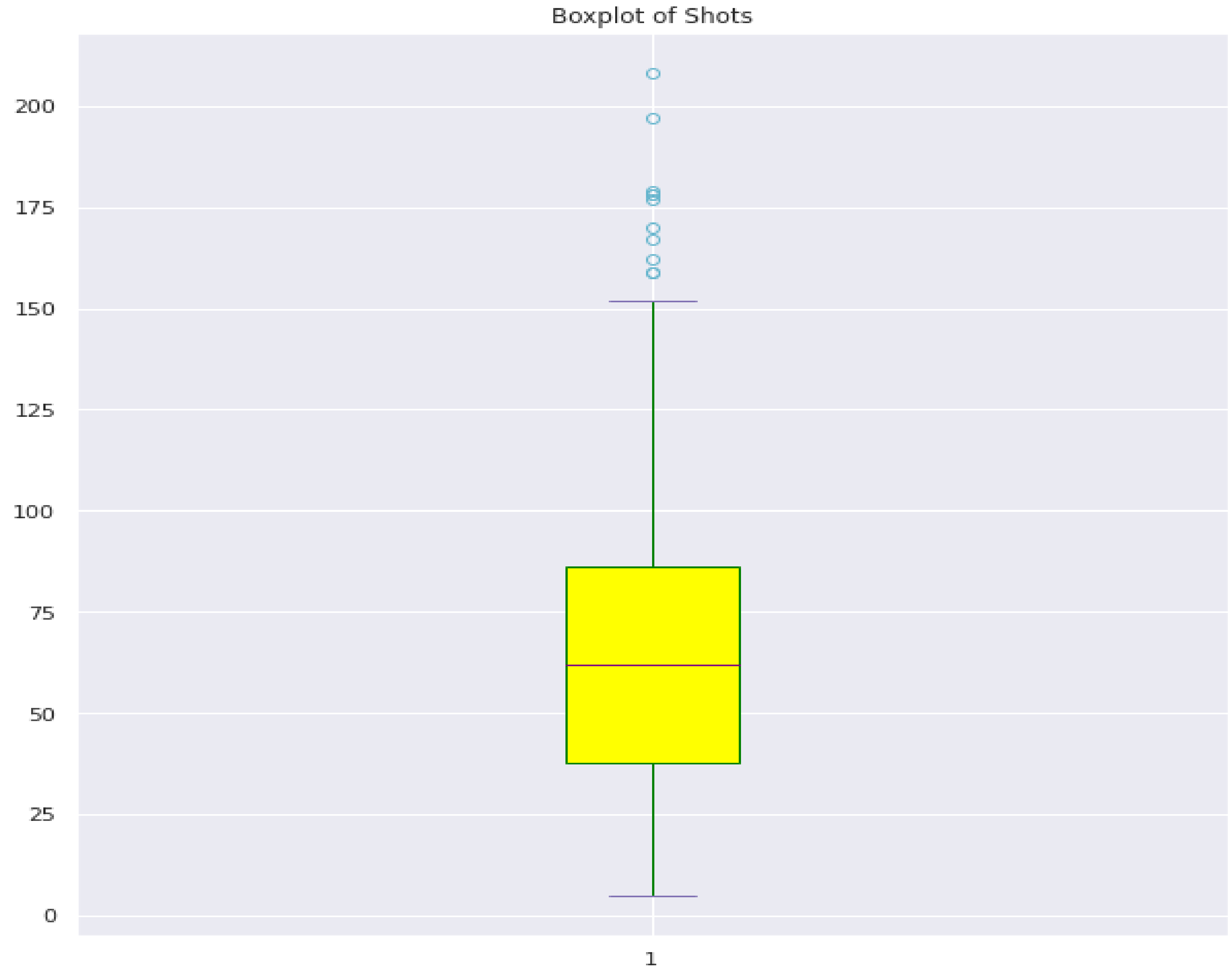
The plot shows the count of number of goals scored.

Highest goal scored is 37 in the year 2016.

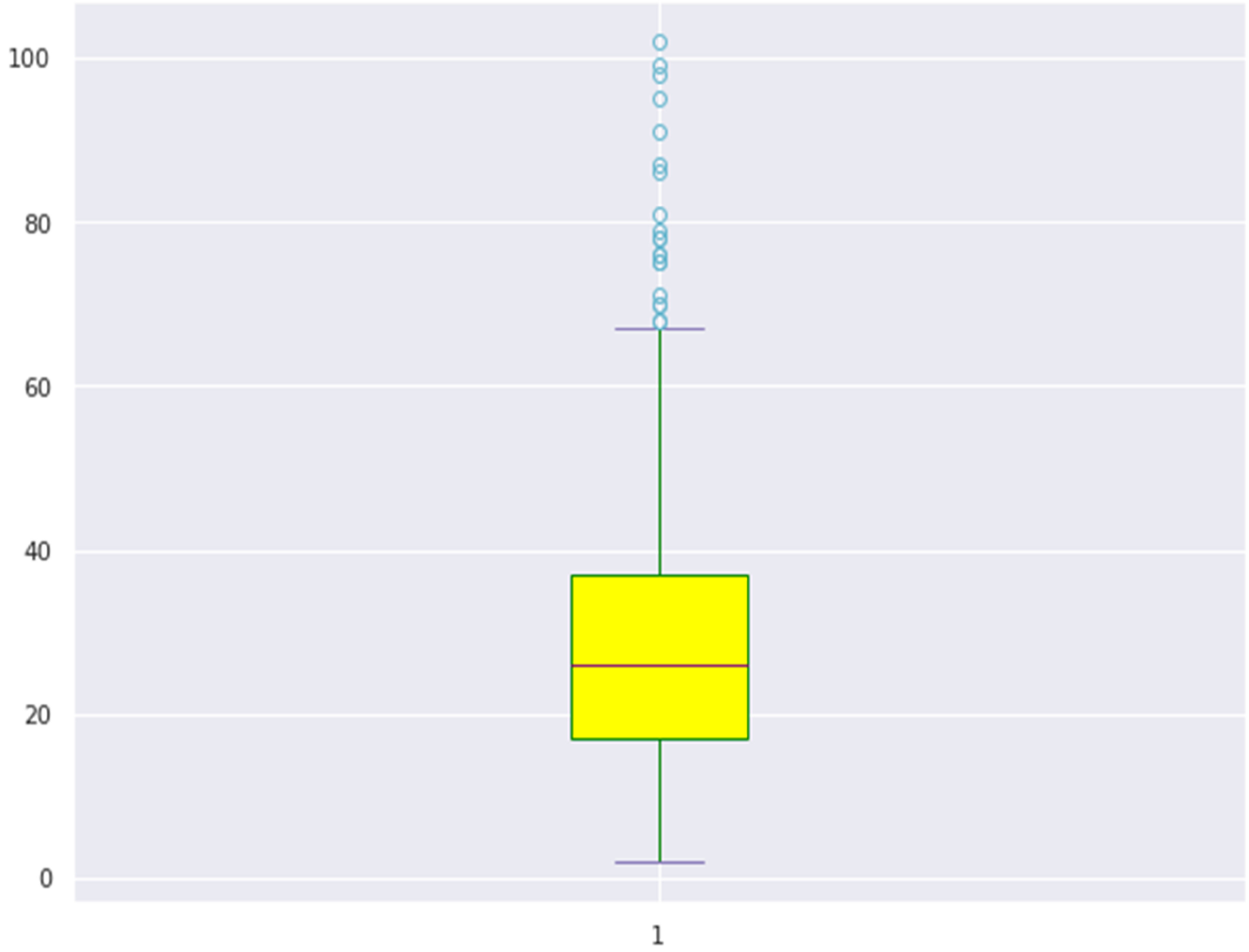


Year	Goals
2016	1489
2017	1102
2018	1702
2019	2398
2020	1087

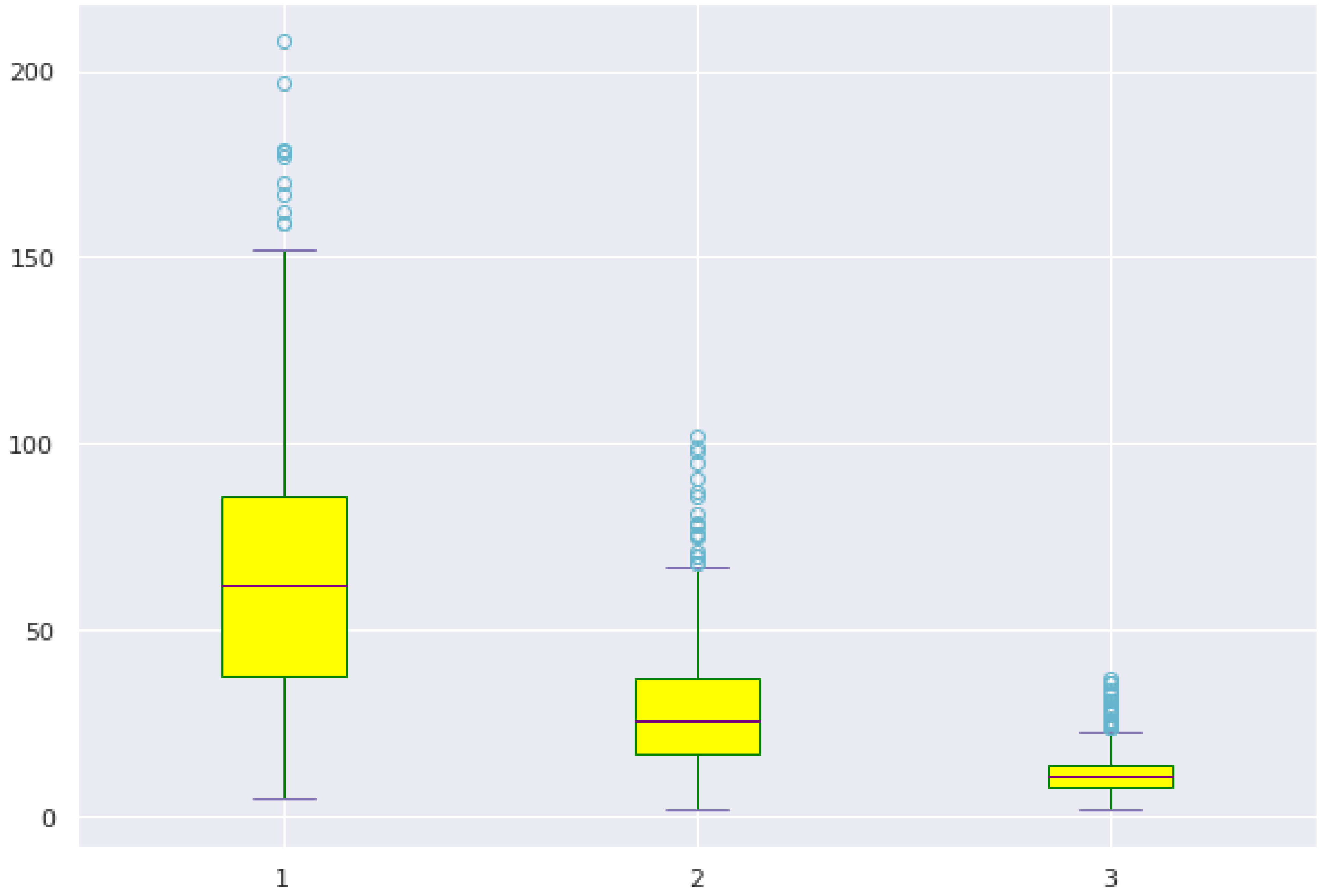
Shots:



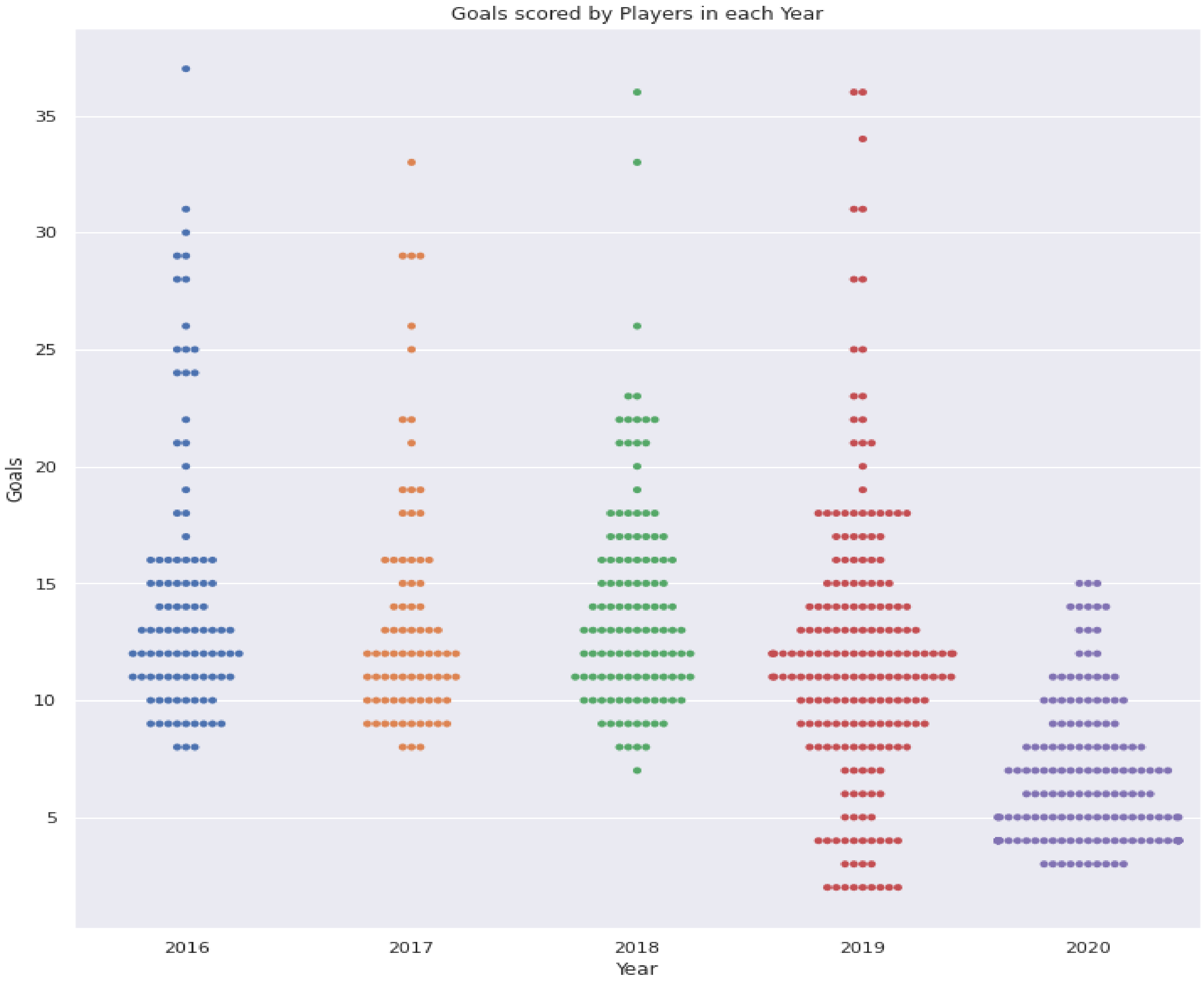
Shots On Target:



Boxplot of Shots, Shots On target, Goals:

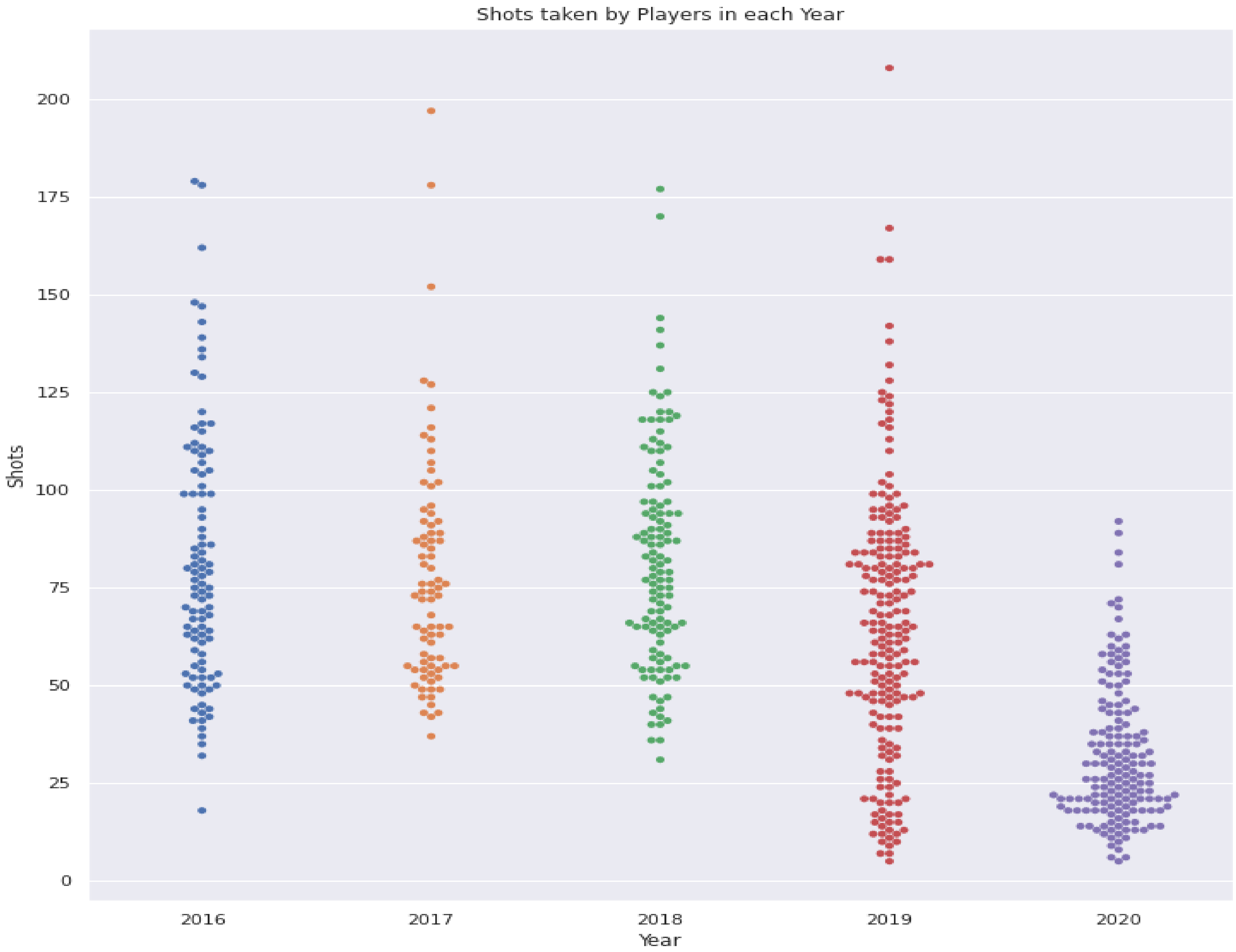


Goals scored by the Players in the 5 years using swarm plot:



All the Players combined scored more number of goals in the year 2019 than any other years.

Shots taken by the Players each year using swarm plot:



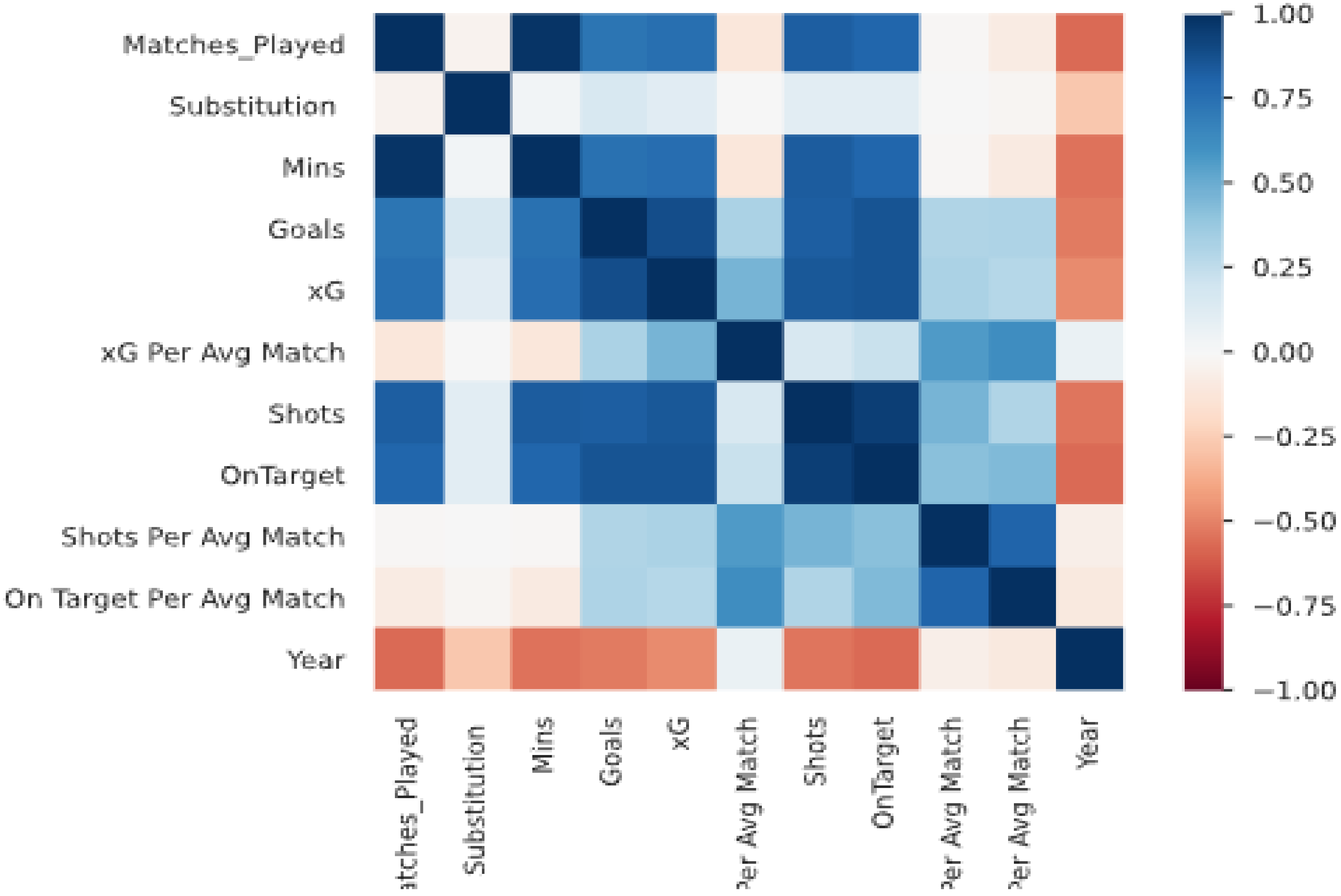
More than 200 shots were taken in the year 2019 by all the players which is the highest in all five years.

Correlations:

Spearman's ρ

The Spearman's rank correlation coefficient (ρ) is a measure of monotonic correlation between two variables, and is therefore better in catching nonlinear monotonic correlations than Pearson's r. It's value lies between -1 and +1, -1 indicating total negative monotonic correlation, 0 indicating no monotonic correlation and 1 indicating total positive monotonic correlation.

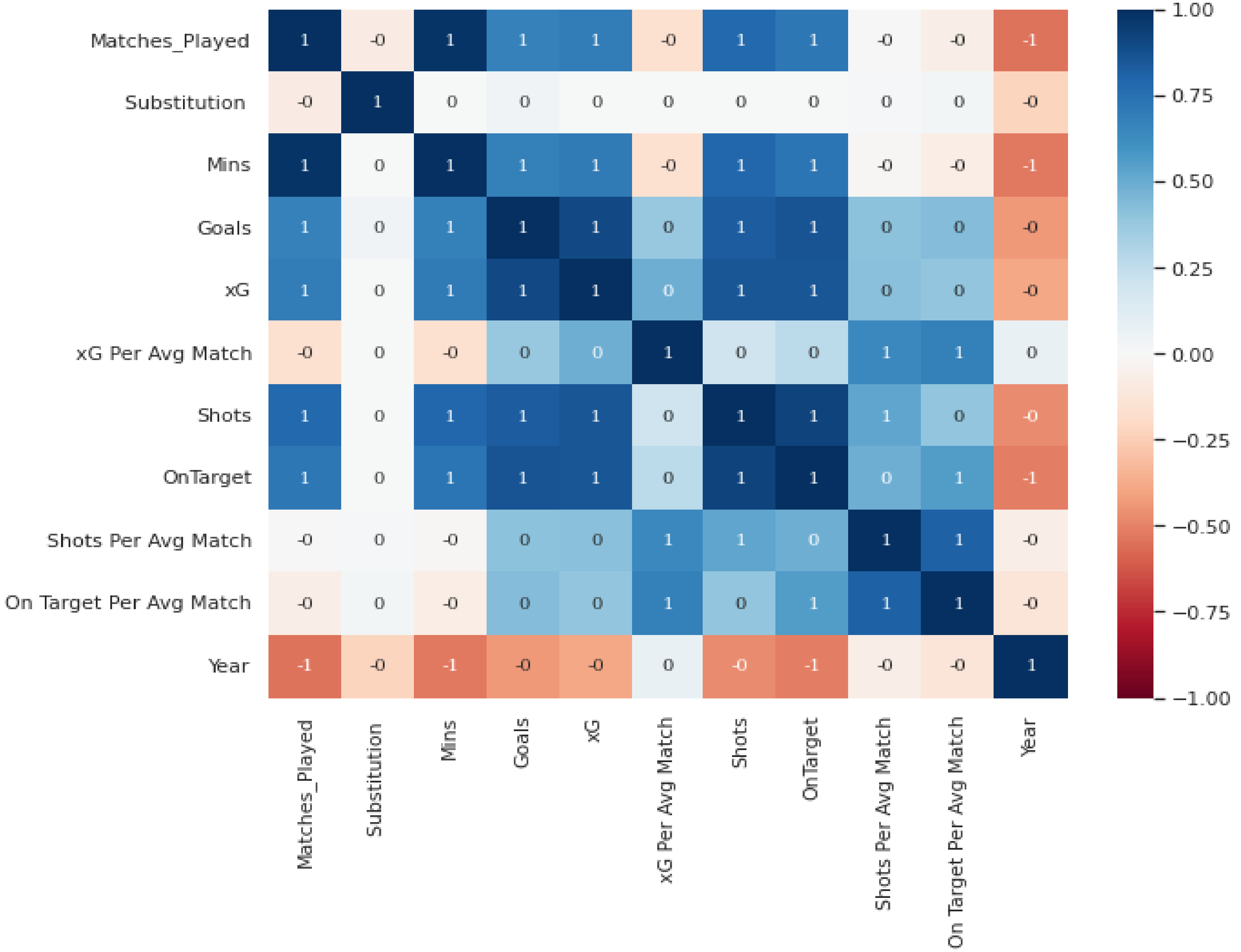
To calculate ρ for two variables X and Y, one divides the covariance of the rank variables of X and Y by the product of their standard deviations.



Pearson's r

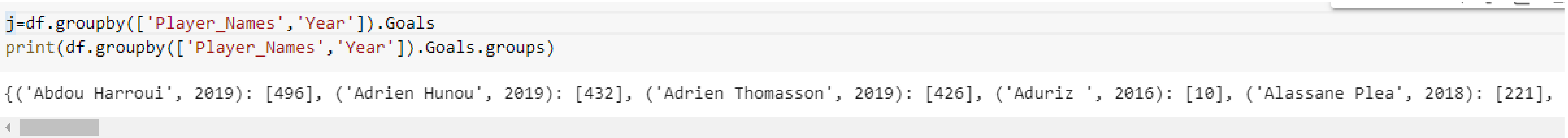
The Pearson's correlation coefficient (*r*) is a measure of linear correlation between two variables. It's value lies between -1 and +1, -1 indicating total negative linear correlation, 0 indicating no linear correlation and 1 indicating total positive linear correlation. Furthermore, *r* is invariant under separate changes in location and scale of the two variables, implying that for a linear function the angle to the x-axis does not affect *r*.

To calculate *r* for two variables *X* and *Y*, one divides the covariance of *X* and *Y* by the product of their standard deviations.



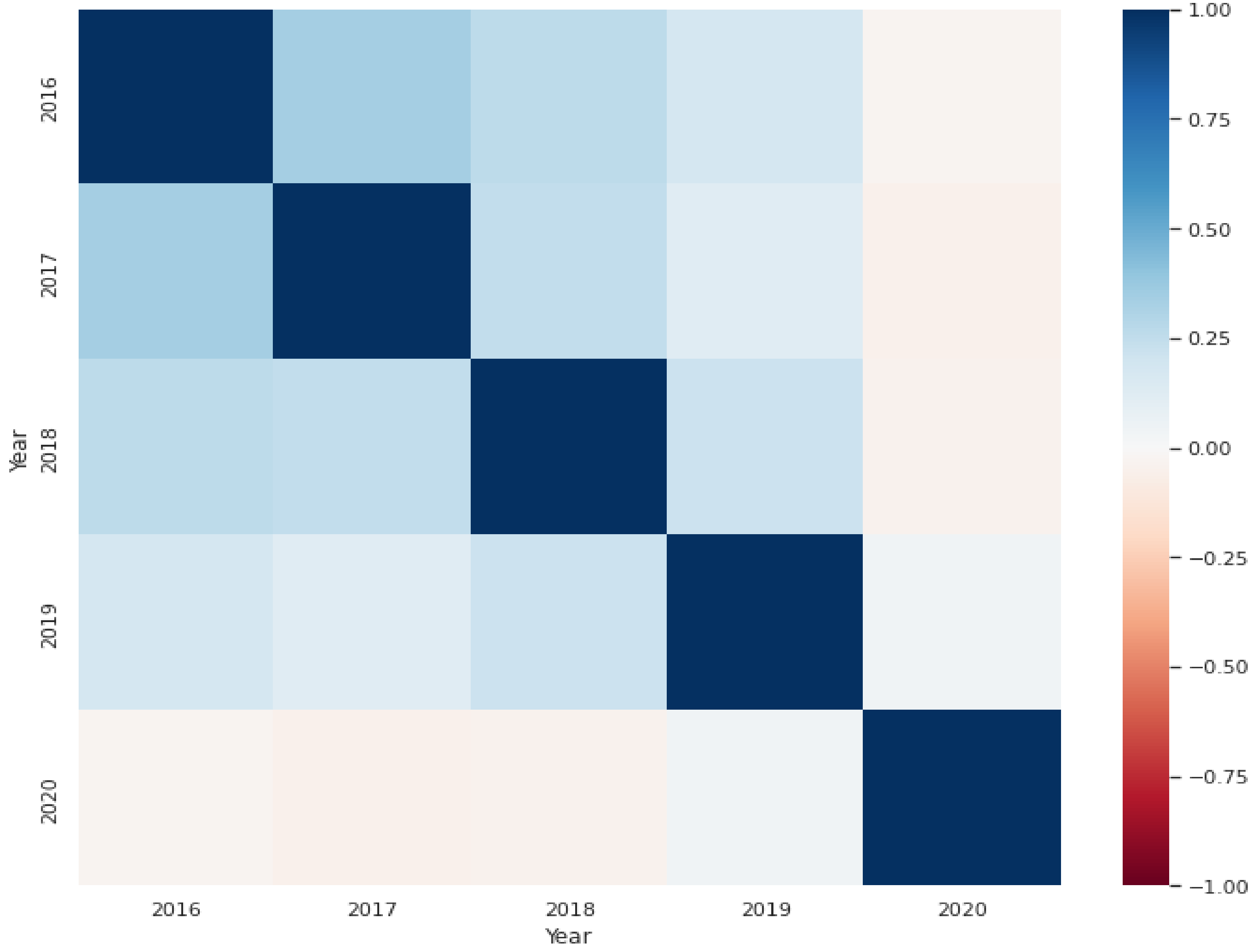
Correlation using groupby:

Coding Sample 1:

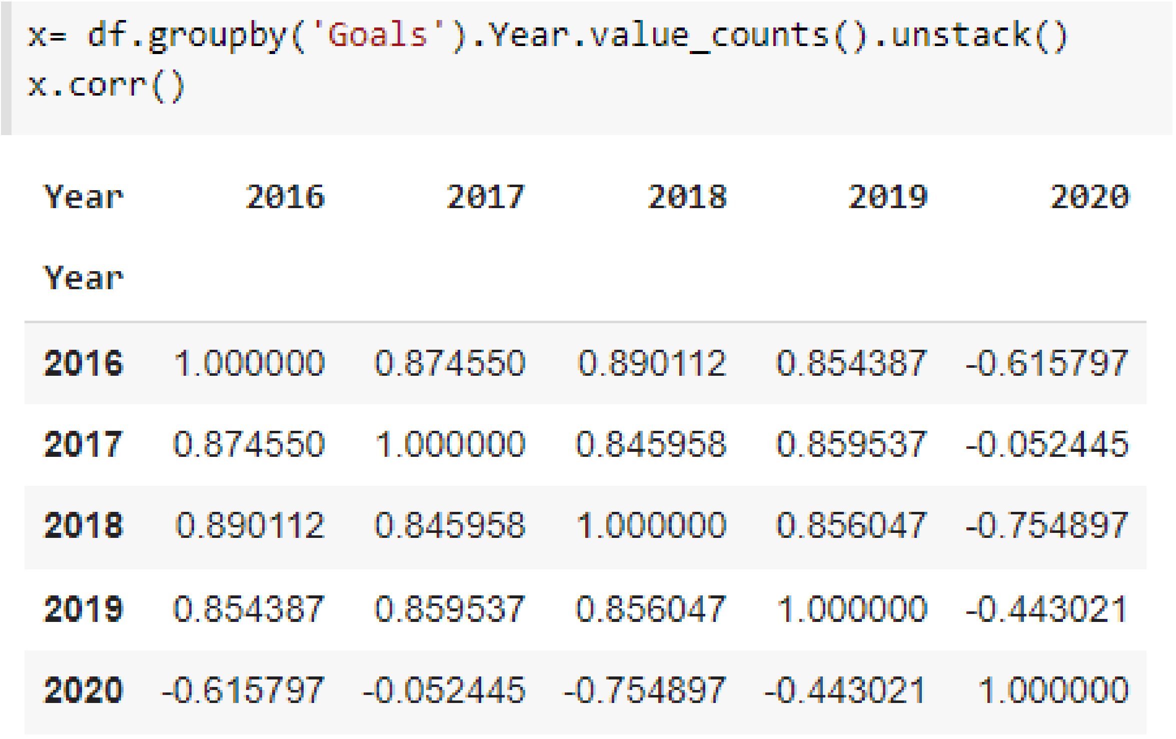


```
j=df.groupby(['Player_Names','Year']).Goals
print(df.groupby(['Player_Names','Year']).Goals.groups)
```

```
{('Abdou Harroui', 2019): [496], ('Adrien Hunou', 2019): [432], ('Adrien Thomasson', 2019): [426], ('Aduriz ', 2016): [10], ('Alassane Plea', 2018): [221],
```

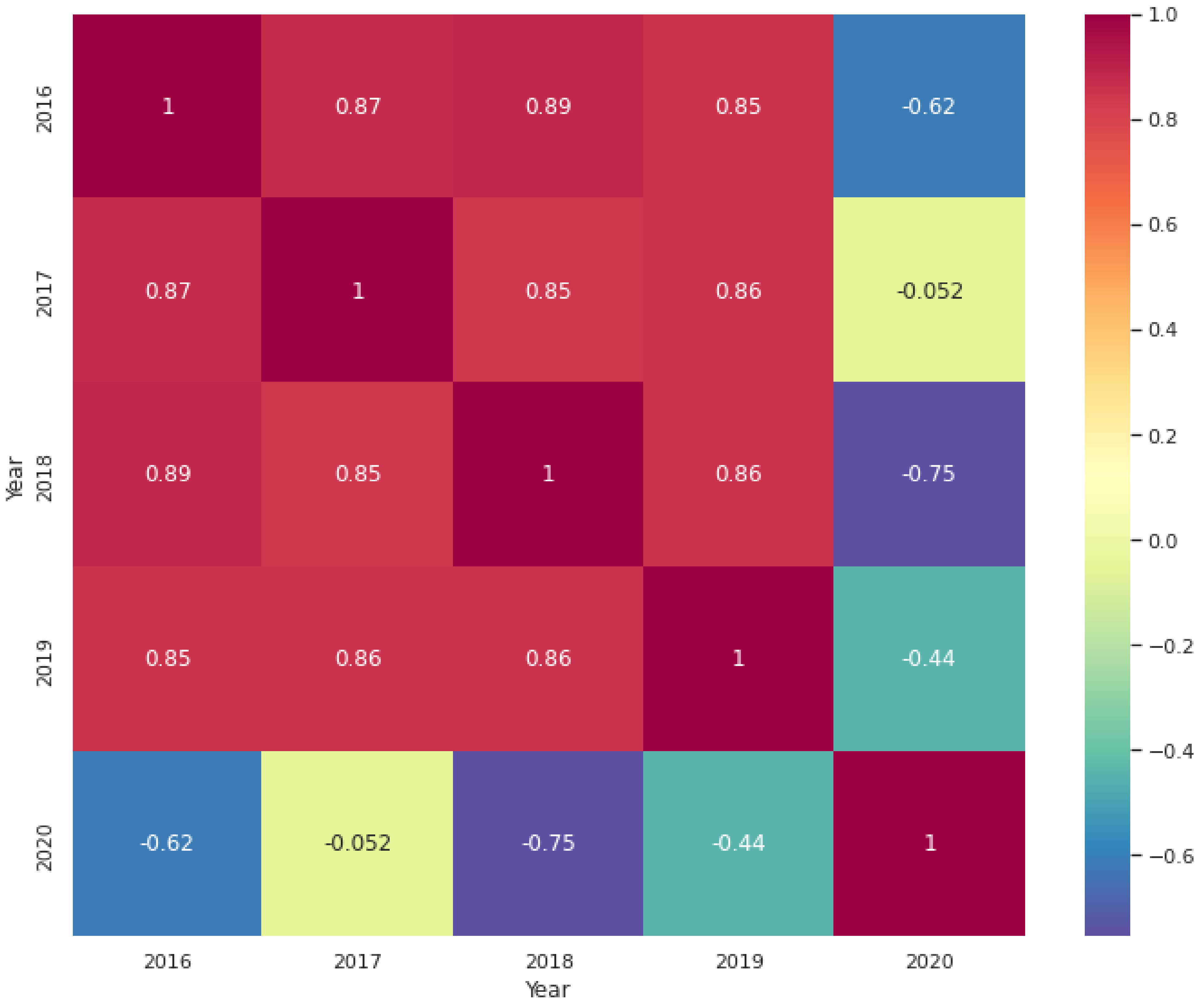


Coding sample 2:



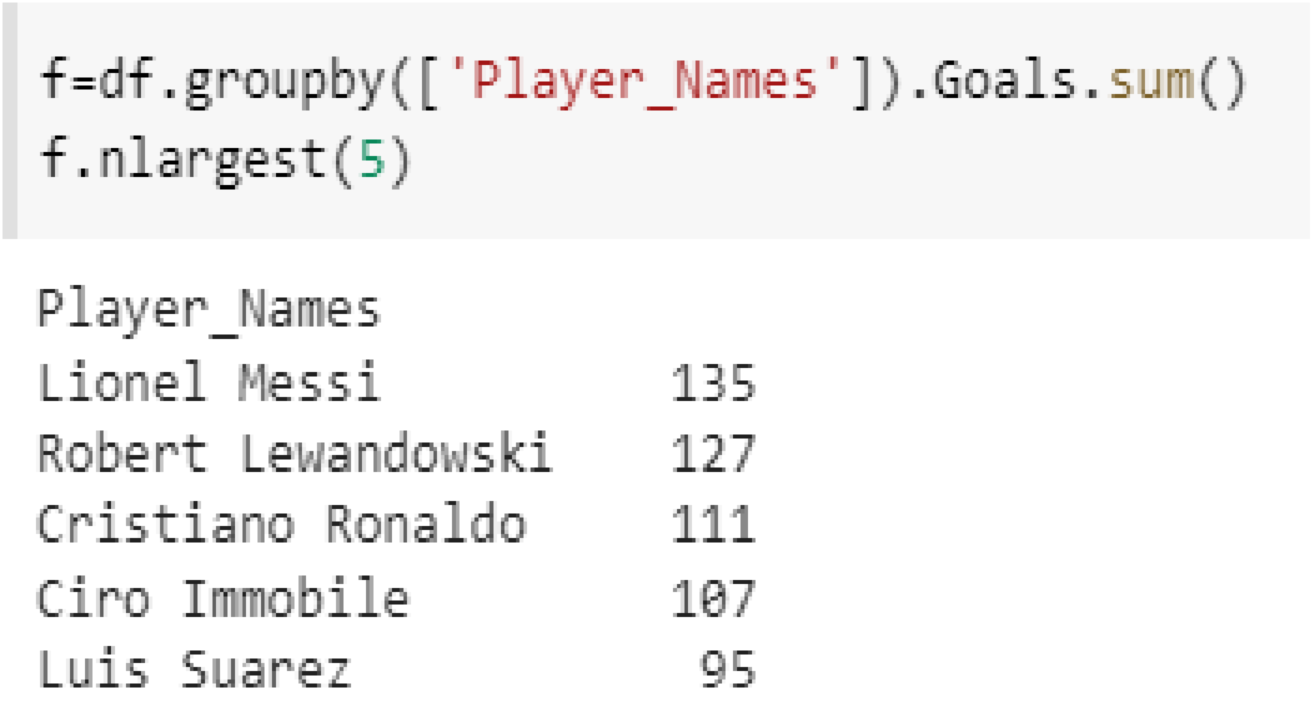
```
x= df.groupby('Goals').Year.value_counts().unstack()
x.corr()
```

Year Year	2016	2017	2018	2019	2020
2016	1.000000	0.874550	0.890112	0.854387	-0.615797
2017	0.874550	1.000000	0.845958	0.859537	-0.052445
2018	0.890112	0.845958	1.000000	0.856047	-0.754897
2019	0.854387	0.859537	0.856047	1.000000	-0.443021
2020	-0.615797	-0.052445	-0.754897	-0.443021	1.000000



ANALYSIS OF THE PLAYERS BY GOALS, XG, SUBSTITUTION, MATCHES PLAYED, XG_PER_AVG_MATCH:

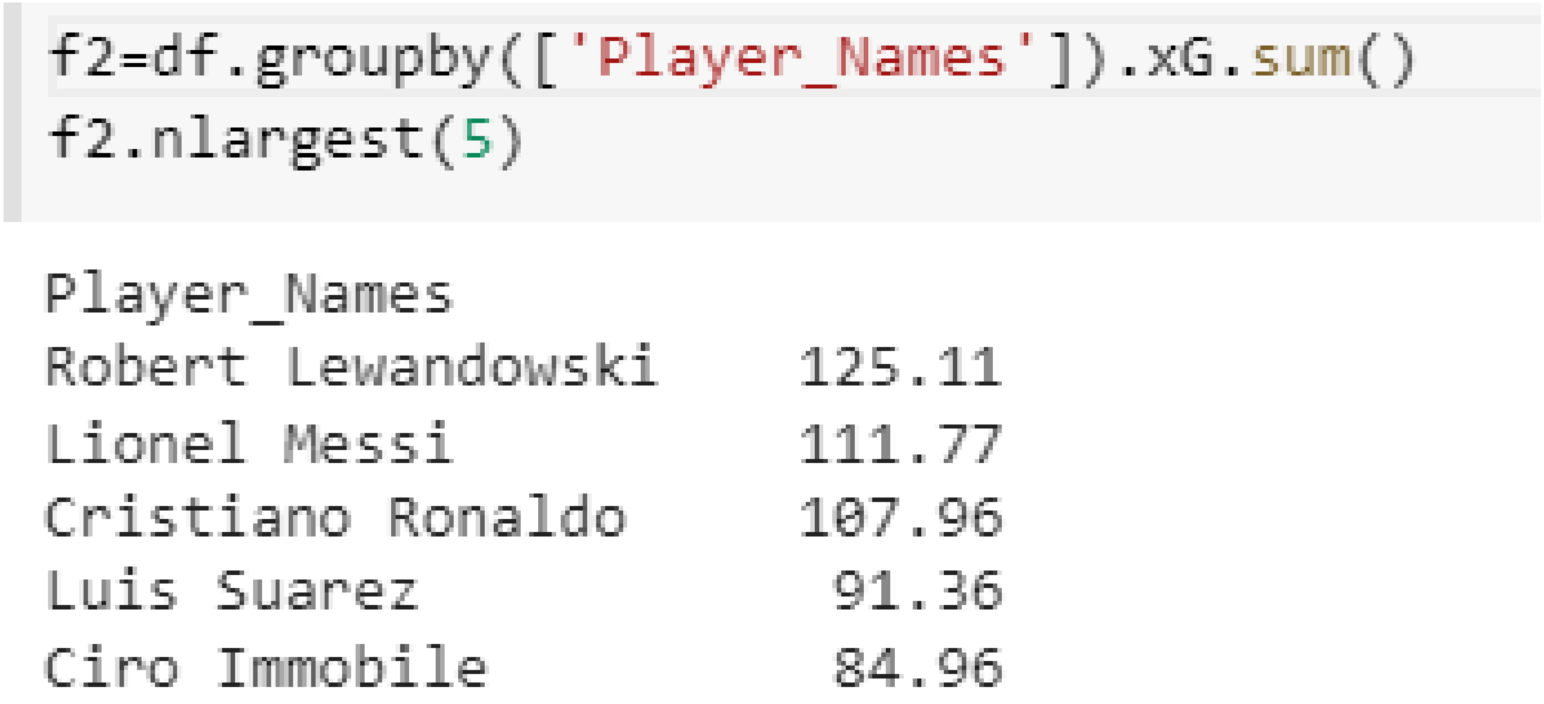
Top 5 Players by Goals:



```python
f=df.groupby(['Player_Names']).Goals.sum()
f.nlargest(5)
```

```
Player_Names
Lionel Messi          135
Robert Lewandowski    127
Cristiano Ronaldo     111
Ciro Immobile         107
Luis Suarez            95
```

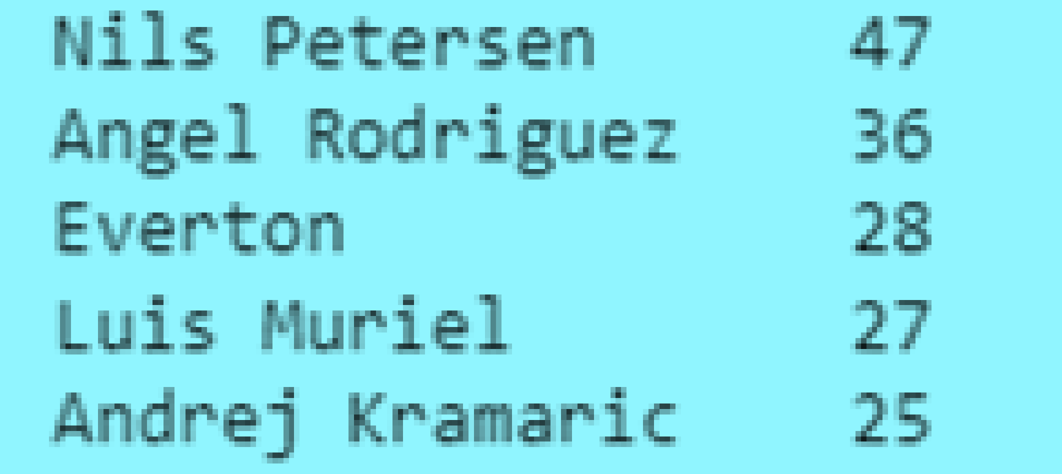
Top 5 Players by xG:



```python
f2=df.groupby(['Player_Names']).xG.sum()
f2.nlargest(5)
```

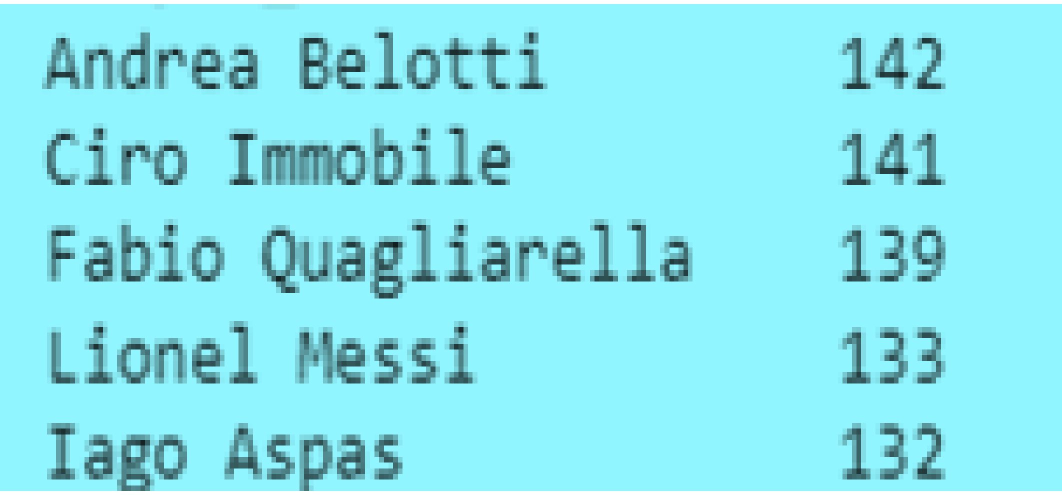
```
Player_Names
Robert Lewandowski    125.11
Lionel Messi          111.77
Cristiano Ronaldo     107.96
Luis Suarez            91.36
Ciro Immobile          84.96
```

Top 5 Most Substituted Players:



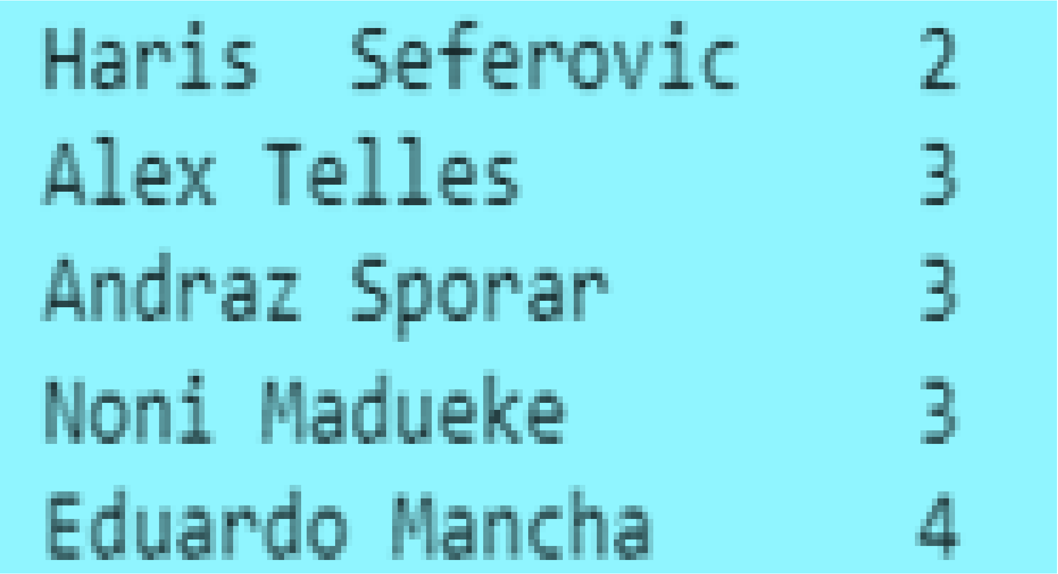
Player Names	Count
Nils Petersen	47
Angel Rodriguez	36
Everton	28
Luis Muriel	27
Andrej Kramaric	25

TOP 5 PLAYERS MOST MATCHES PLAYED:



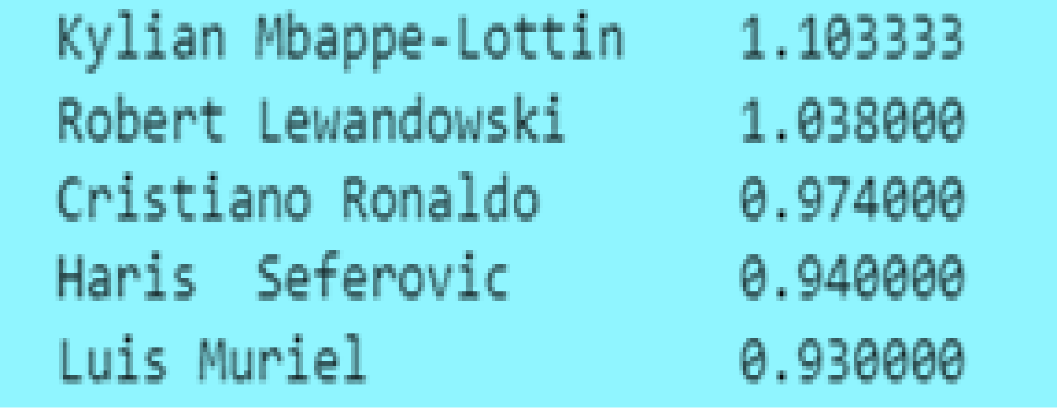
Player Names	Count
Andrea Belotti	142
Ciro Immobile	141
Fabio Quagliarella	139
Lionel Messi	133
Iago Aspas	132

PLAYERS WITH LEAST MATCHES PLAYED:



Player Names	Count
Haris Seferovic	2
Alex Telles	3
Andraz Sporar	3
Noni Madueke	3
Eduardo Mancha	4

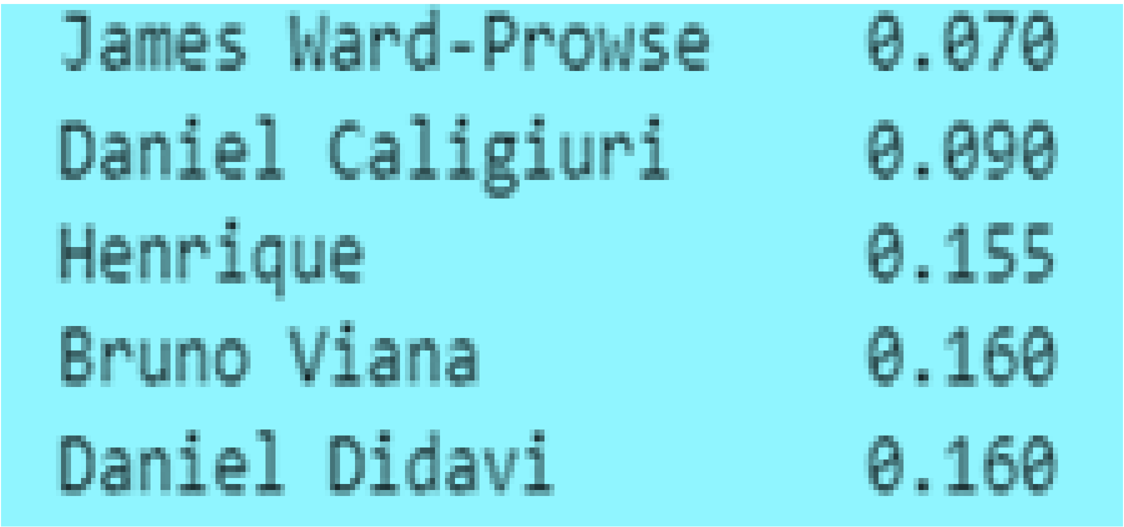
PLAYERS WITH HIGHEST GOAL SCORING EXPECTATION PER AVERAGE MATCH:



Player Names	xG_Per_Avg_Match
Kylian Mbappe-Lottin	1.103333
Robert Lewandowski	1.038000
Cristiano Ronaldo	0.974000
Haris Seferovic	0.940000
Luis Muriel	0.930000

PLAYERS WITH LOWEST GOAL SCORING EXPECTATION PER AVERAGE MATCH:

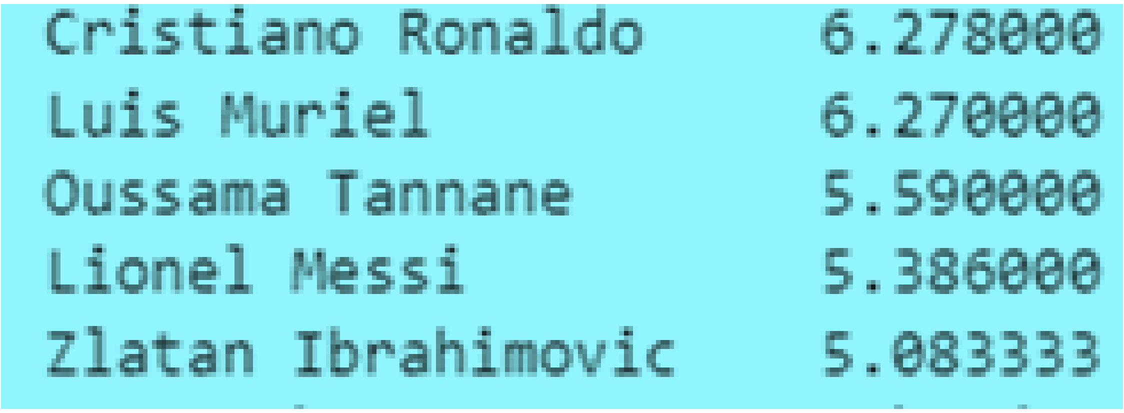
Player Names **xG_Per_Avg_Match**



```
James Ward-Prowse    0.070
Daniel Caligiuri     0.090
Henrique             0.155
Bruno Viana          0.160
Daniel Didavi        0.160
```

SHOTS PER AVERAGE MATCH (TOP 5):

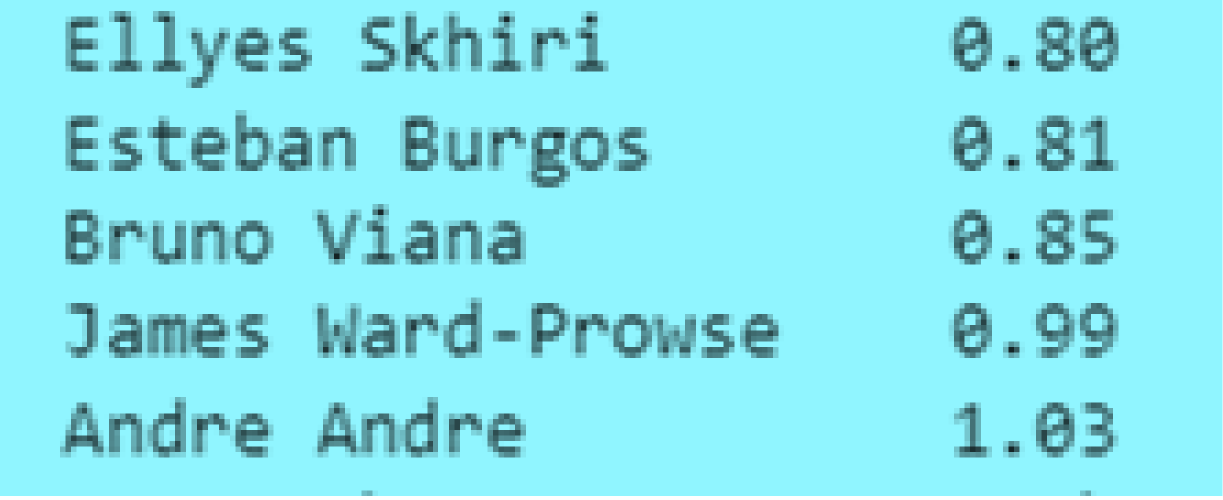
Player Names **Shots_Per_Avg_Match**



```
Cristiano Ronaldo     6.278000
Luis Muriel           6.270000
Oussama Tannane       5.590000
Lionel Messi          5.386000
Zlatan Ibrahimovic    5.083333
```

SHOTS PER AVERAGE MATCH (BOTTOM 5):

Player Names **Shots_Per_Avg_Match**

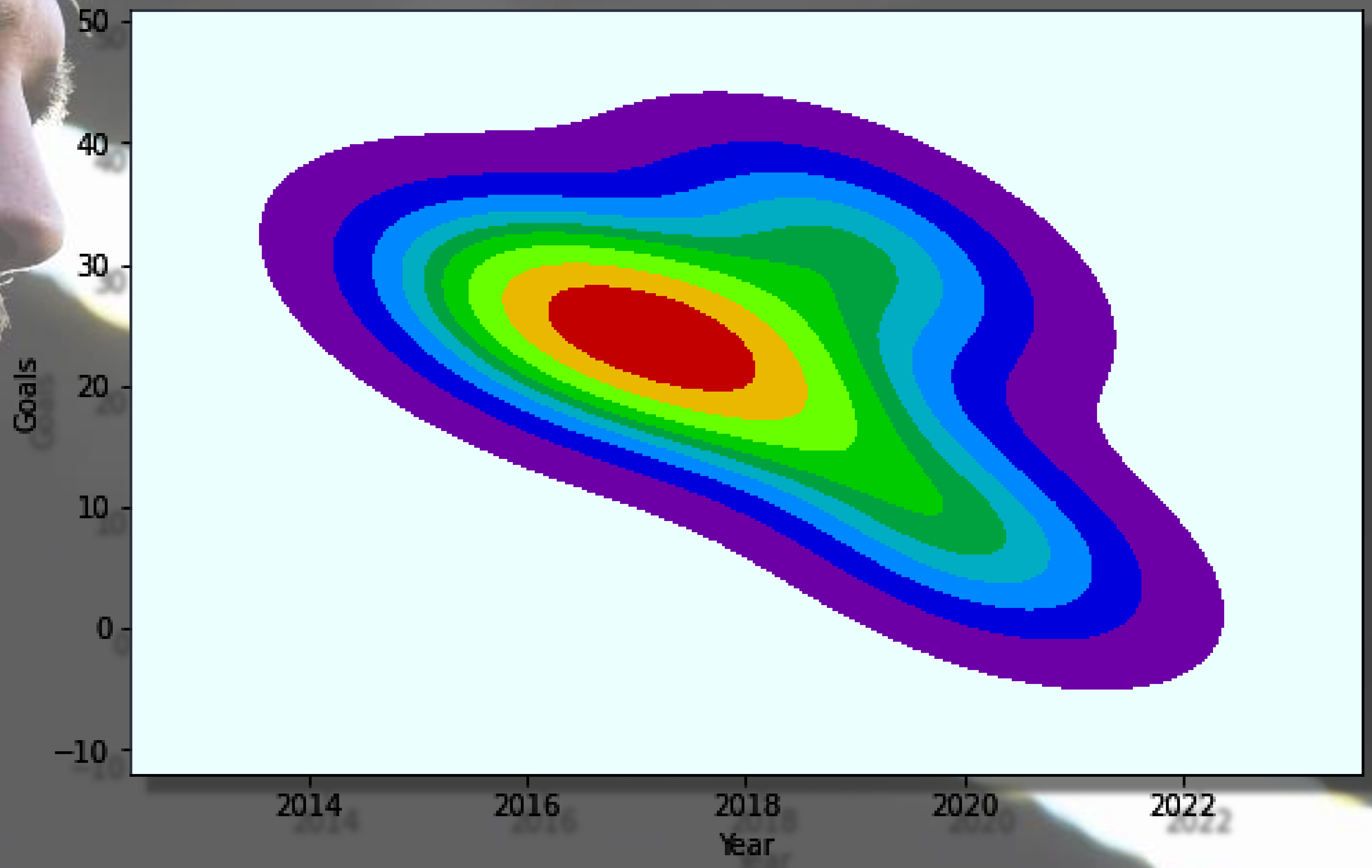


```
Ellyes Skhiri           0.80
Esteban Burgos          0.81
Bruno Viana             0.85
James Ward-Prowse       0.99
Andre Andre             1.03
```

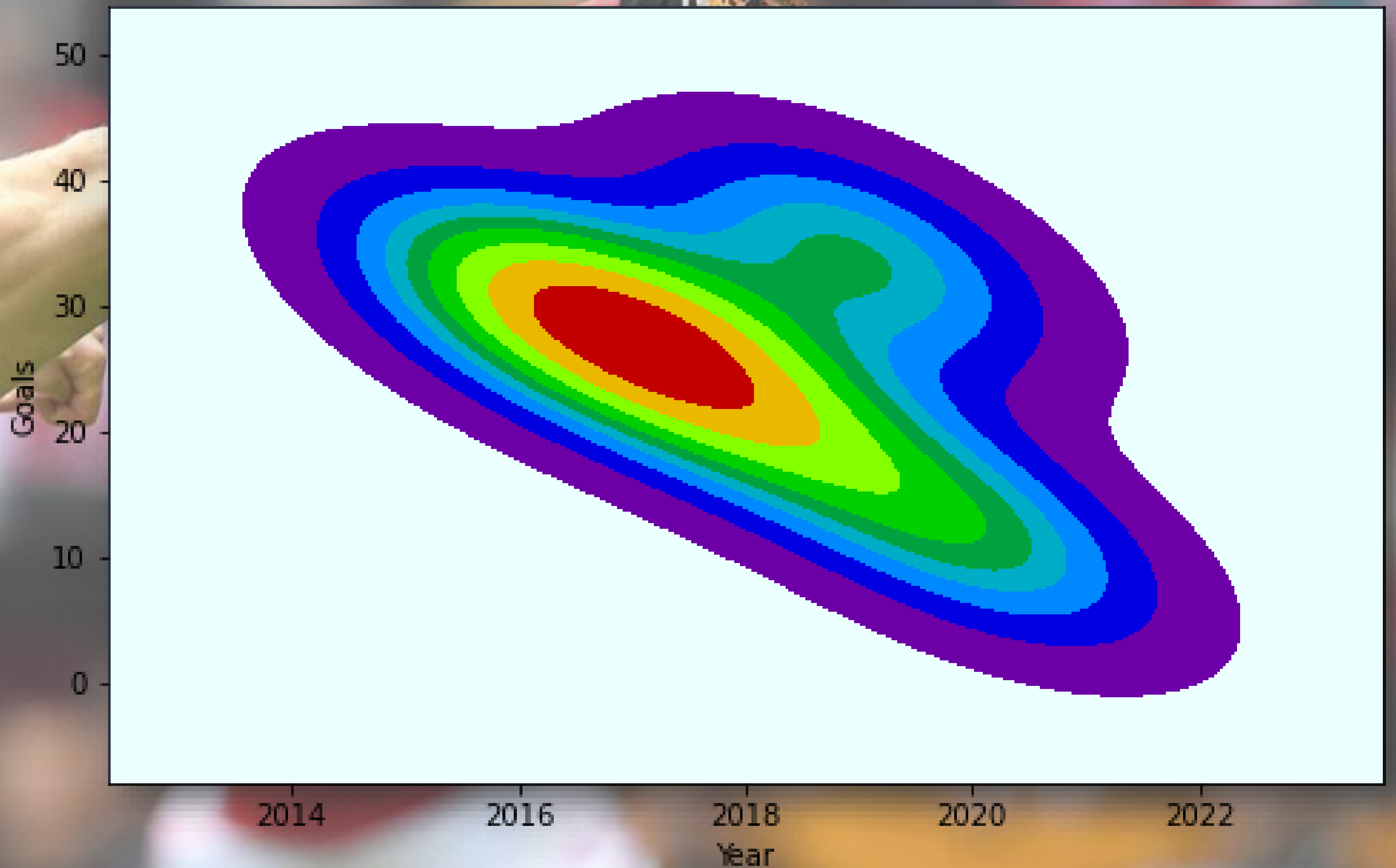
Analysis of top 5 players:

Analysis on the basis of Goals scored:

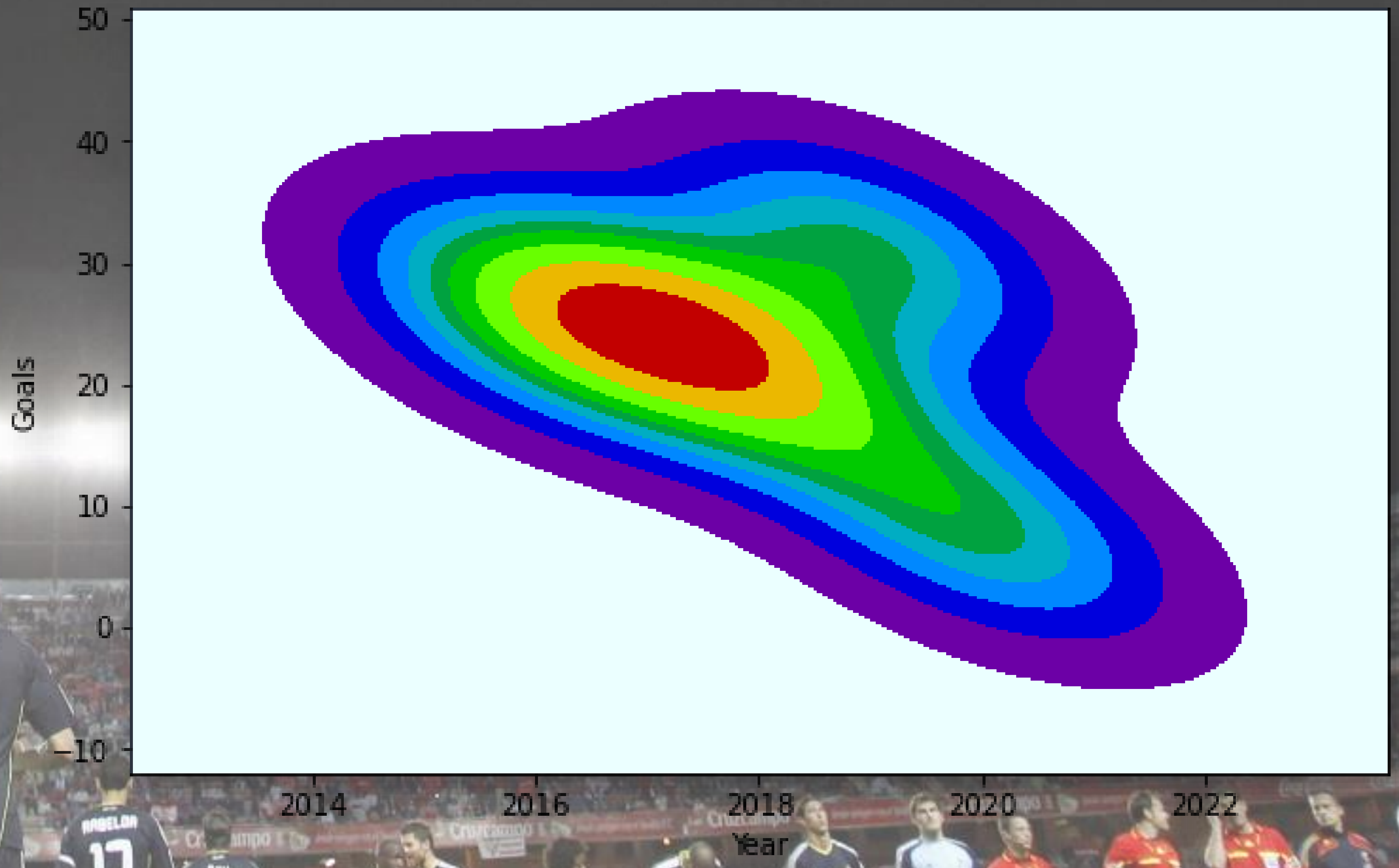
LIONEL MESSI



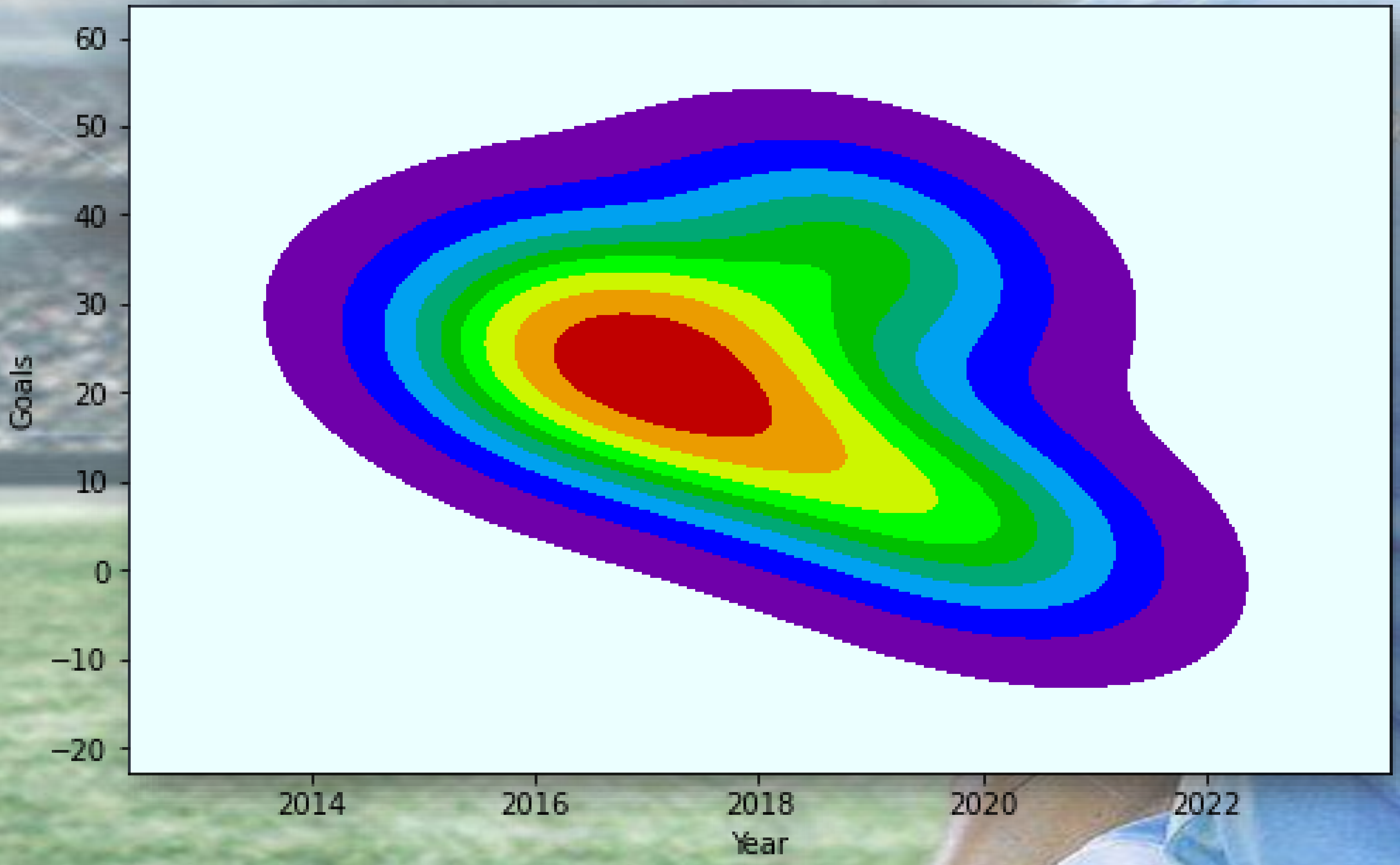
ROBERT LEAWANDOWSKI



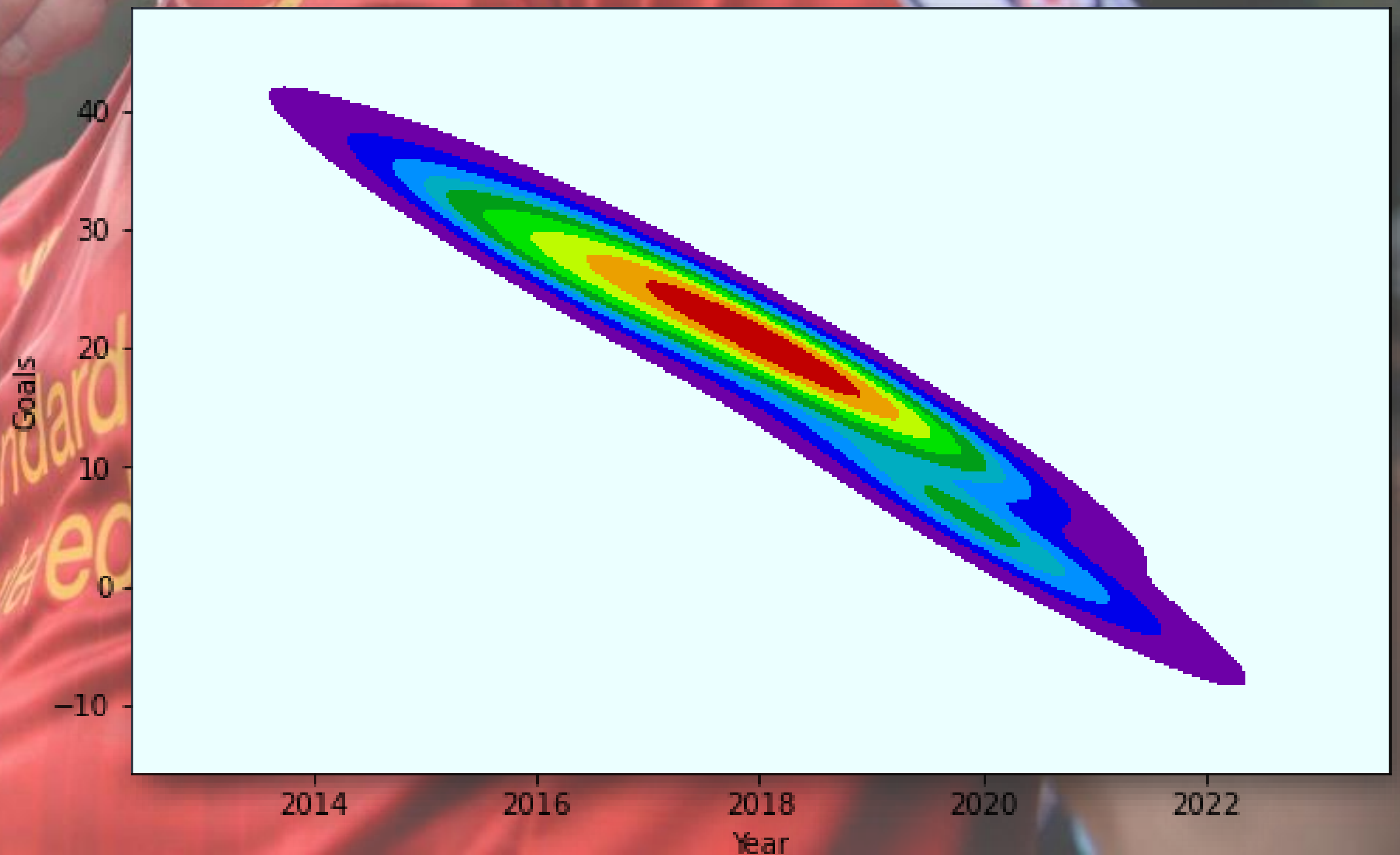
CRISTIANO RONALDO



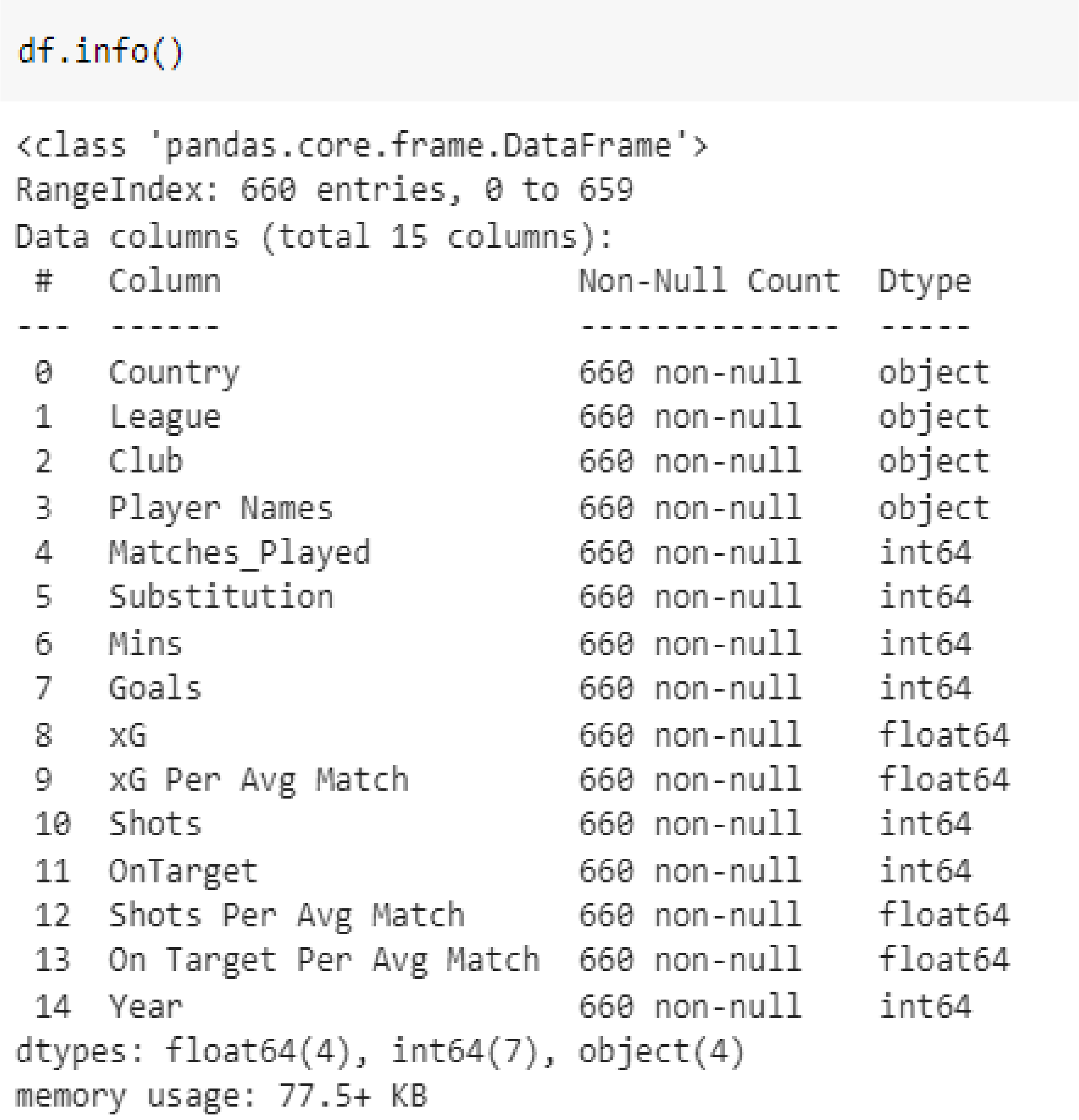
CIRO IMMOBILE



LUIS SUAREZ



DATA PREPARATION :



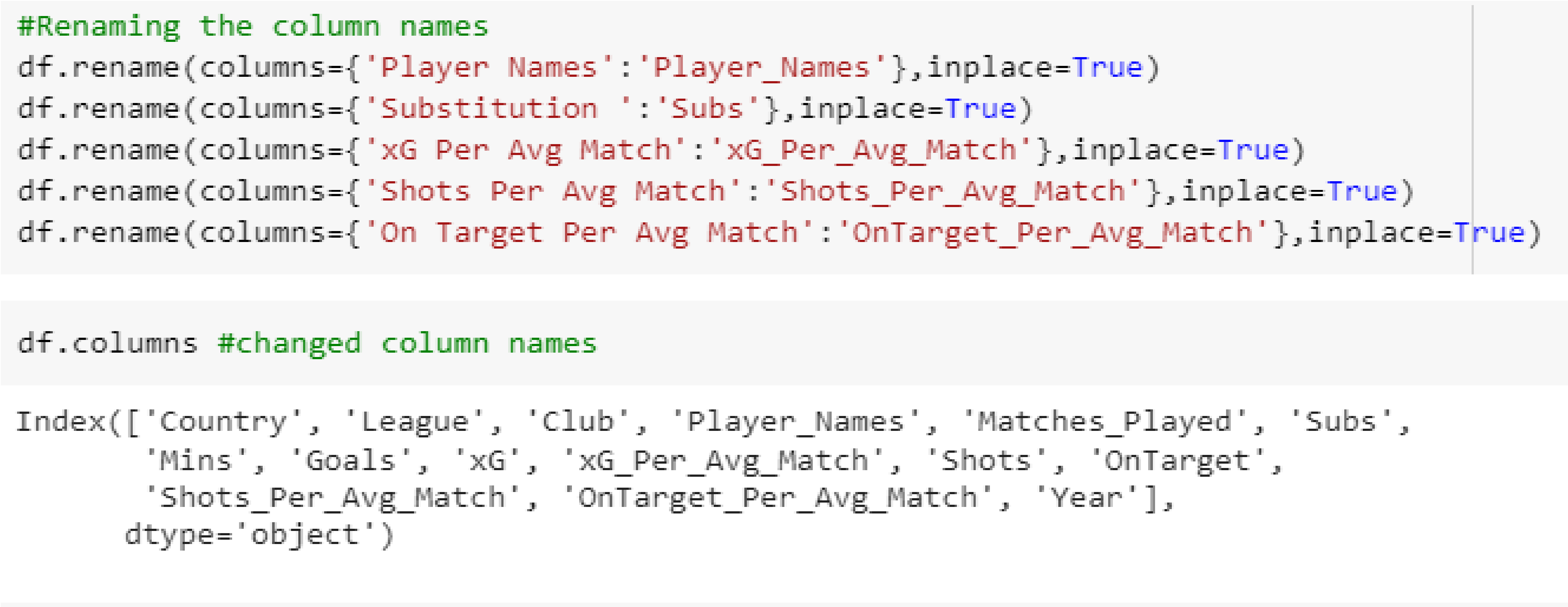
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Country                  660 non-null    object 
 1   League                   660 non-null    object 
 2   Club                     660 non-null    object 
 3   Player Names             660 non-null    object 
 4   Matches_Played           660 non-null    int64  
 5   Substitution             660 non-null    int64  
 6   Mins                     660 non-null    int64  
 7   Goals                    660 non-null    int64  
 8   xG                       660 non-null    float64
 9   xG Per Avg Match         660 non-null    float64
 10  Shots                    660 non-null    int64  
 11  OnTarget                 660 non-null    int64  
 12  Shots Per Avg Match      660 non-null    float64
 13  On Target Per Avg Match  660 non-null    float64
 14  Year                     660 non-null    int64  
dtypes: float64(4), int64(7), object(4)
memory usage: 77.5+ KB
```

The column names are not in the form of '_'
 E.g. Players Names

Python doesn't accept the column names having space while using few keywords like groupby.

Renaming the column names:



```python
#Renaming the column names
df.rename(columns={'Player Names':'Player_Names'},inplace=True)
df.rename(columns={'Substitution ':'Subs'},inplace=True)
df.rename(columns={'xG Per Avg Match':'xG_Per_Avg_Match'},inplace=True)
df.rename(columns={'Shots Per Avg Match':'Shots_Per_Avg_Match'},inplace=True)
df.rename(columns={'On Target Per Avg Match':'OnTarget_Per_Avg_Match'},inplace=True)
```

```python
df.columns #changed column names
```

```
Index(['Country', 'League', 'Club', 'Player_Names', 'Matches_Played', 'Subs',
       'Mins', 'Goals', 'xG', 'xG_Per_Avg_Match', 'Shots', 'OnTarget',
       'Shots_Per_Avg_Match', 'OnTarget_Per_Avg_Match', 'Year'],
      dtype='object')
```

CONCLUSION:

According to the data “Lionel Messi”, “Cristiano Ronaldo”, “Robert Lewandowski” are the best players in those five years.

Lionel Messi has scored the most Goals (135) from 2016 to 2020.

As per Analysis the Top 5 Players on the basis of goal scoring are :
1) Lionel Messi 2) Robert Lewandowski 3) Cristiano Ronaldo 4) Ciro Immobile 5) Luis Suarez.

Andrea Belloti has played the most number of matches (142) followed by Ciro Imobile (141).

Haris Sefarovic has played the least number matches (2).

Nils Peterson is the most substituted Player from 2016-2020.

Robert Lewandowski has a high goal scoring expectation (xG) than Lionel Messi and Cristiano Ronaldo.

Kylian Mbappe has the highest goal scoring expectation per average match 1.1003.

James ward-Prowse has the lowest goal scoring expectation per average match 0.07.

According to shots taken per average match Cristiano Ronaldo is at the top (6.2780) and with (6.270) Luis Muriel is on 2nd.

Ellyes Skhiri has the lowest shot per average match ratio (0.80)