Project 1: Analysis of Top European Clubs

Shehzad Anwar

2022-06-10

Background and Problem Definition

The purpose for this project will be to analyse the playing statistics that separate these football clubs offensively and defensively while also observing which of them have been over-performing and/or under-performing based on their entire season's output of goals allowed and goals conceded. Furthermore, I will analyze which country is the dominant nation in European football at the moment based on the given rankings. UEFA picks the top 32 clubs to participate in its premier annual tournament, the UEFA Champions League, and so I will be using the top 32 ranked clubs to process the data in the final analysis.

The data being used in this project is stats of European football clubs that played in each of the top five ranked nations' first divisions during the 2021/22 season. Their ranks are based on how highly UEFA, the continental board for the sport, rate them in their standing in world football. The data was collected and made available to the public on FBref.com.

```
library(knitr)
library(ggplot2) # needed to plot
library(dplyr) # needed for data cleaning
library(tidyr) # needed to tidy the data up
library(plotly) # needed to plot
top5_stats = read.csv("Big_5_European_Leagues_Stats.csv") # reads the provided csv data file
top5_stats = na.omit(top5_stats) # gets rid of any NA values in the dataset
head(top5_stats)
```

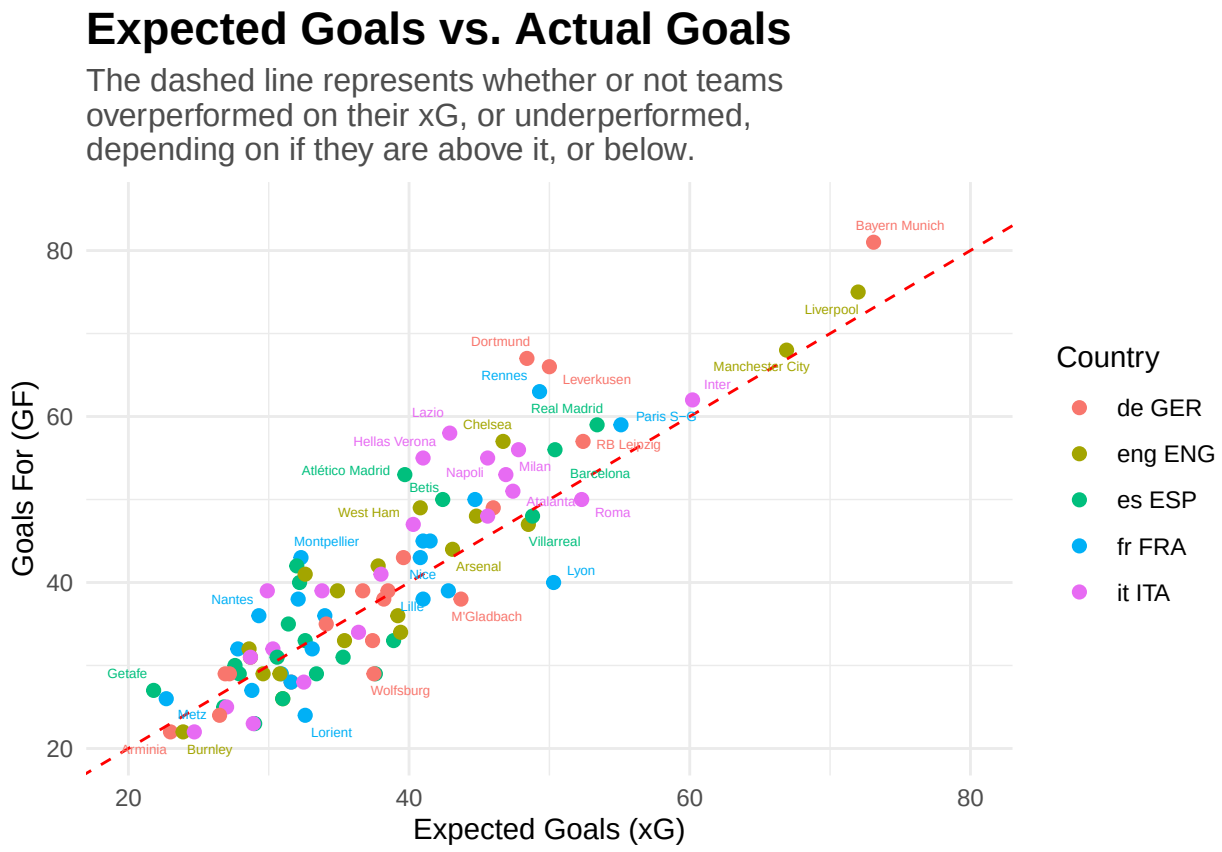
```
##   Rk           Squad Country LgRk MP  W D L GF GA GD Pts Pts.G   xG  xGA  xGD
## 1  1 Manchester City eng ENG    1 29 22 4 3 68 18 50  70  2.41 66.9 21.2 45.7
## 2  2       Liverpool eng ENG    2 29 21 6 2 75 20 55  69  2.38 72.0 27.6 44.4
## 3  3   Bayern Munich  de GER    1 27 20 3 4 81 28 53  63  2.33 73.1 27.4 45.7
## 4  4     Real Madrid  es ESP    1 29 20 6 3 59 25 34  66  2.28 53.4 32.8 20.5
## 5  5       Paris S-G  fr FRA    1 29 20 5 4 59 27 32  65  2.24 55.1 29.8 25.3
## 6  6           Milan  it ITA    1 30 20 6 4 56 29 27  66  2.20 47.8 29.4 18.4
##   xGD.90    Last.5 Attendance                    Top.Team.Scorer
## 1   1.58 W L W W D     52,681 Riyad Mahrez, Raheem Sterling - 10
## 2   1.53 W W W W W     53,459                 Mohamed Salah - 20
## 3   1.69 W W D D W     24,214            Robert Lewandowski - 31
## 4   0.71 W W W W L     38,892                 Karim Benzema - 22
## 5   0.87 L W L W L     40,513                 Kylian Mbappé - 15
## 6   0.61 D D W W W     35,359 Olivier Giroud, Rafael Leão... - 8
##         Goalkeeper
## 1          Ederson
## 2          Alisson
## 3     Manuel Neuer
## 4 Thibaut Courtois
```

```
## 5      Keylor Navas
## 6      Mike Maignan
```

I've shown a snippet of the dataset to give an idea on how the rest of the data points are formatted. The libraries I've called are ggplot2, to allow me to plot the dataset, dplyr, to help in cleaning and arranging the data in a proper manner, tidyr to aid in that, and plotly to allow me to have more graphing options. 'top5_stats' will hold the data from the "Big_5_European_Leagues_Stats.csv" file and all NA values have been cleared off it.

Data Analysis and Visualization

How have clubs over-performed and under-performed in goals this season?



Expected Goals Scored vs. Actual Goals Scored

In order to get the plot, I extracted attack related columns from the original dataset and initialized them in top5_stats_attack, which I then used to implement into a ggplot scattergraph. I then set up the graph with a geom_point, geom_abline and a ggrepel label to differentiate between the points, as well as added color to the points to see what country each of the clubs are from to make it stand out more.

From the plot, we can observe that some of the clubs that outperformed their xG the most are Dortmund, Rennes, Levursken and Lazio. This tells us two things. It could mean that these clubs have high level forwards who finish most of their chances and have been doing brilliantly to do as they are. Let us look at the aforementioned clubs' top goalscorers, respectively:

```
top5_stats_attack[c(8, 17, 18, 29), ] # will select only the mentioned scorers' clubs
```

```
##         Squad Country LgRk GF   xG     Top.Team.Scorer
## 8    Dortmund  de GER    2 67 48.4 Erling Haaland - 16
## 17     Rennes  fr FRA    3 63 49.3 Martin Terrier - 16
## 18 Leverkusen  de GER    3 66 50.0  Patrik Schick - 20
## 29      Lazio  it ITA    7 58 42.9  Ciro Immobile - 21
```

A quick calculation will allow us to see what percentage of the total goals scored by the clubs will allow us to see how paramount their each team's forwards are and if they are overreliant on them or not. It is generally accepted that, for clubs in the modern day, that the entire squad generates goals and not just one player. As such, with 1 goalkeeper, 5 defenders, and usually 2 defensive midfielders, the rest of the 4 outfield players should bring in the goals. I'll take the mean of the total goals scored throughout Europe in order to assess the average amount of goals every club has scored and use it to calculate the percentage each team's 4 attacking players needed output of goals each.

The average goals scored across Europe is:

```
## [1] 39.76531
```

A well balanced team's four attacking players should be scoring at least this amount of goals:

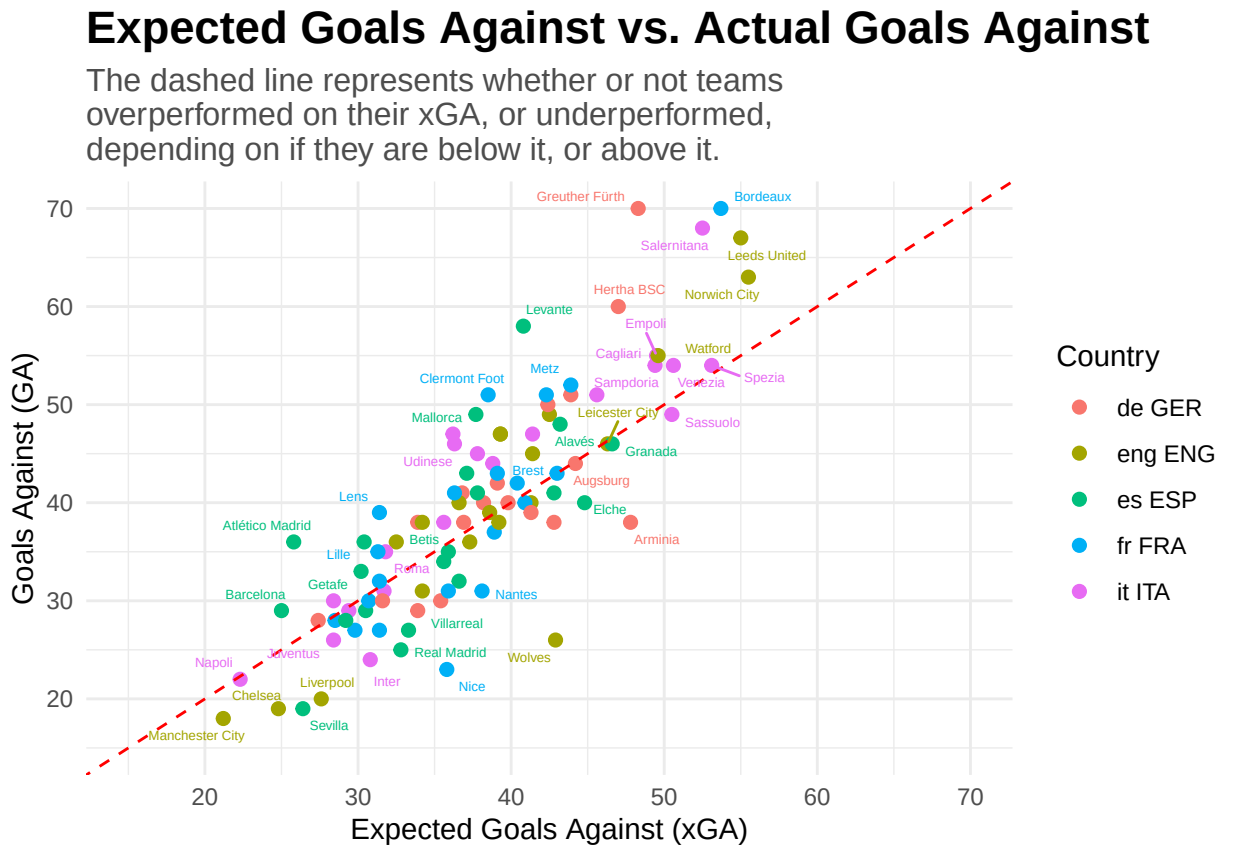
```
## [1] 9.941327
```

What this means is each of the 4 attacking players should score at least 9 goals in order to stay competitive in their leagues. Let us now see how reliant each of the 4 teams are on their top goalscorer:

```
##   Forwards Percent.Of.Goals
## 1  Haaland         23.88060
## 2  Terrier         25.39683
## 3   Schick         30.30303
## 4 Immobile         36.20690
```

From the data, we can tell that Haaland scores ~24%, Terrier scores ~25%, Schick scores ~30% and Immobile scores ~36% of their clubs' total goals scored. This proves that each of these teams are overreliant on one of their forwards for their goals, and needs to bring in another forward to keep from risking their goals scored next season from falling off a cliff. Any injuries, sales or unforeseen scenarios that may act on any of these forwards will cause each of these teams to have a drop in performance massively.

The second thing these clubs' GF to xG comparison tells us is that they need to bolster their squad with better creative midfielders. A consistent exceptional output such as these clubs have done throughout the season is unsustainable and will eventually catch up to them for the worst.



Expected Goals Against vs. Actual Goals Against

From the data, it can be observed that Bordeaux, Levante, Greuther Furth, Salernitana, Leeds United and Norwich City are all wildly underperforming on their xGA. The aforementioned clubs have all been relegated to the second division as well, except for Leeds United (who, as mentioned in the background, are a club that have gotten extremely lucky not to have been relegated with their stats):

```
top5_defense[c(79, 93, 95, 97, 98), ]
```

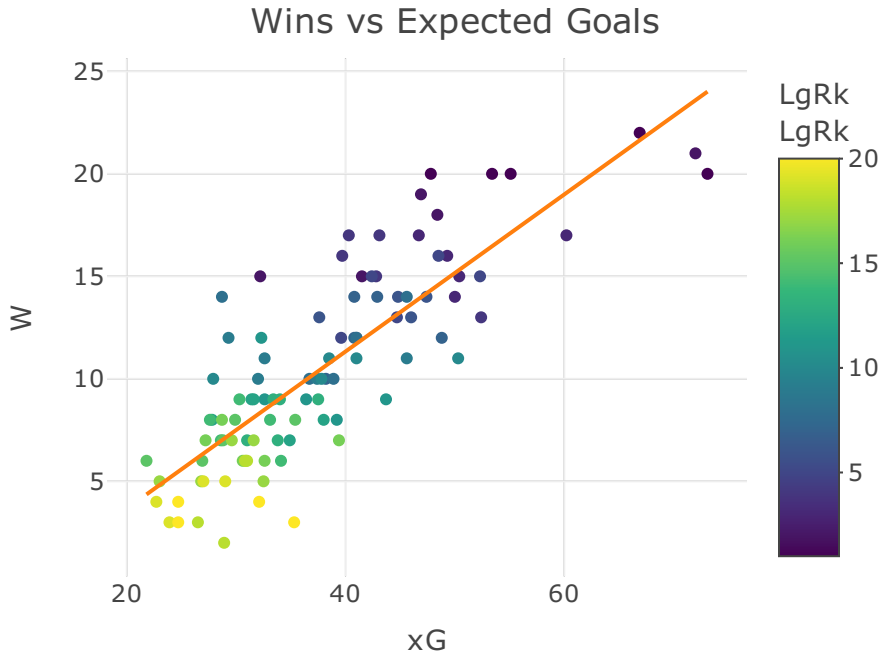
```
##             Squad Country LgRk GA  xGA
## 79   Leeds United eng ENG   16 67 55.0
## 93       Bordeaux  fr FRA   20 70 53.7
## 95        Levante  es ESP   20 58 40.8
## 97 Greuther Fürth  de GER   18 70 48.3
## 98    Salernitana  it ITA   20 68 52.5
```

Simply put, this means that each of these teams must improve their squads or style of play if they want to continue to fight in Europe's top flight. Whether they improve the defense or the goalkeeper remains a question that must be answered with data from other sources.

Do more goals scored throughout the season mean more wins?

For the following question, I will be creating another dataframe that only contains the columns I will need. The expected goals (xG) statistic will be used as it is an indicator as to how many goalscoring opportunites

a team, and the higher the league rank of each team, the higher their xG. I will aim to answer whether offense is more important than defense.



The summary of the linear regression model can be found below:

```
##
## Call:
## lm(formula = W ~ xG, data = top5_winsToGoals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5319 -1.8369 -0.1498  1.4046  6.9963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.99003    1.03496  -3.855 0.000209 ***
## xG           0.38306    0.02637  14.526  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.653 on 96 degrees of freedom
## Multiple R-squared:  0.6873, Adjusted R-squared:  0.684
## F-statistic:   211 on 1 and 96 DF,  p-value: < 2.2e-16
```

Assigning the linear regression to fitW, I created a plotly scatterplot and inserted the fitted value of fitW to see how the model would line up against actual data. It is accurate and seems to prove that focusing

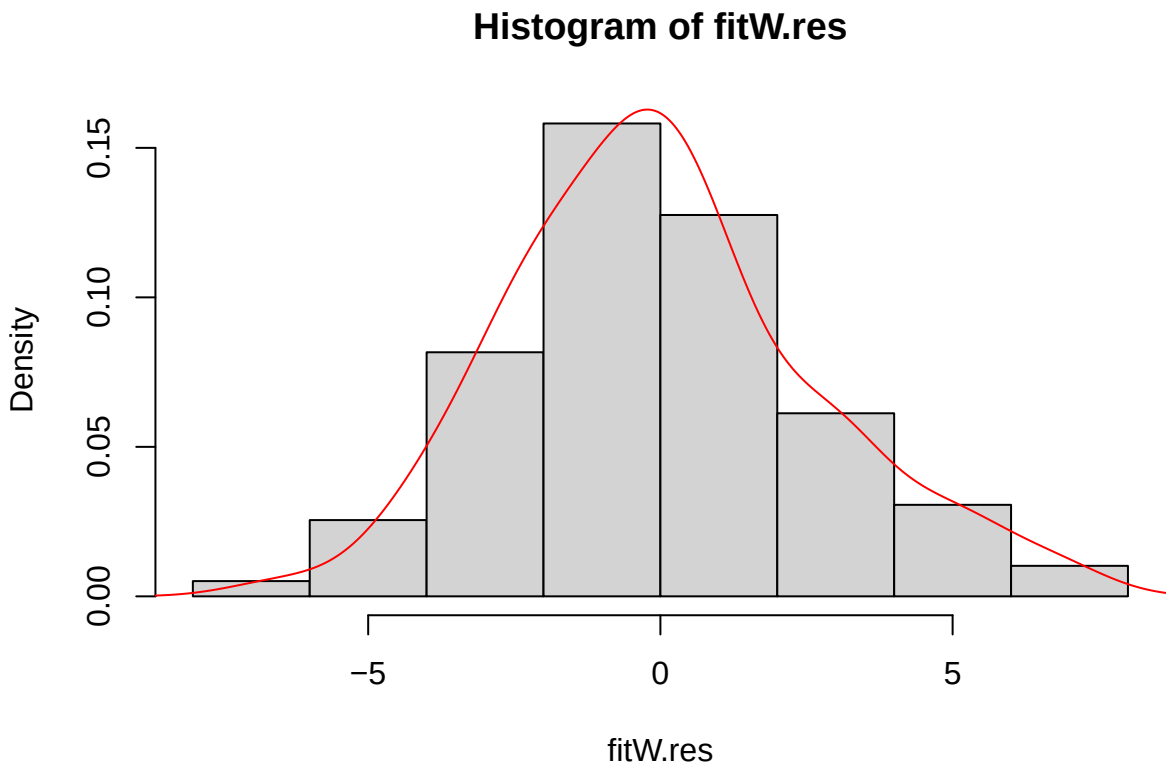
on defense is a better tactic than offense. The R^2, the correlation coefficient, is not too small and does show that there is a correlation between the two variables. The model generated here gives us a formula of *Number of Wins* = -3.99003 + 0.38306 * *xG*.

As can be observed from the plot and the linear regression model, it seems less important to be able to score more goals in order to get more wins, which leads us to the idea that it is more important to not concede. Liverpool and Bayern Munich, 2nd and 1st in their own leagues, respectively, are both behind Manchester City in number of wins. Liverpool are also direct competitors to Man City and proved that their defense is *slightly* better than Liverpool's. There is a difference of only 2 goals conceded between the two teams, while the difference in expected goals is much higher (5.5):

```
top5_defense[c(1, 2), ]
```

```
##             Squad Country LgRk GA  xGA
## 1 Manchester City eng ENG    1 18 21.2
## 2       Liverpool eng ENG    2 20 27.6
```

Such are the margins that they miss out on the top spot because of 2 goals they conceded, which ultimately led to a loss or draw. As such, it its important for Liverpool to bolster their defense in the coming seasons if they are to compete at City's level, and the same can be said for other teams that plan on doing the same.



The density graph is not really skewed in any direction and is mostly symmetric, which leads me to believe that the model that's been setup fits well.

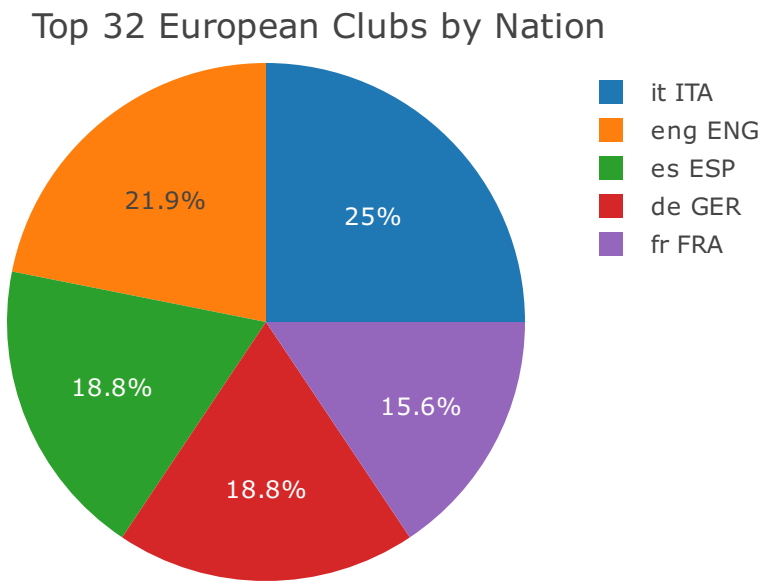
What country's clubs dominate the top 32 European rankings?

We'll take the top 32 clubs as they are the ones that UEFA recognizes them as they are the ones chosen to play in UEFA's annual tournaments and recognized as good enough to compete at the top level consistently.

```
countryRanks32 = head(top5_stats, 32)
head(countryRanks32)
```

```
##   Rk           Squad Country LgRk MP  W D L GF GA GD Pts Pts.G   xG  xGA  xGD
## 1  1 Manchester City eng ENG    1 29 22 4 3 68 18 50  70  2.41 66.9 21.2 45.7
## 2  2       Liverpool eng ENG    2 29 21 6 2 75 20 55  69  2.38 72.0 27.6 44.4
## 3  3   Bayern Munich  de GER    1 27 20 3 4 81 28 53  63  2.33 73.1 27.4 45.7
## 4  4     Real Madrid  es ESP    1 29 20 6 3 59 25 34  66  2.28 53.4 32.8 20.5
## 5  5       Paris S-G  fr FRA    1 29 20 5 4 59 27 32  65  2.24 55.1 29.8 25.3
## 6  6           Milan  it ITA    1 30 20 6 4 56 29 27  66  2.20 47.8 29.4 18.4
##   xGD.90    Last.5 Attendance                    Top.Team.Scorer
## 1   1.58 W L W W D     52,681 Riyad Mahrez, Raheem Sterling - 10
## 2   1.53 W W W W W     53,459                 Mohamed Salah - 20
## 3   1.69 W W D D W     24,214            Robert Lewandowski - 31
## 4   0.71 W W W W L     38,892                 Karim Benzema - 22
## 5   0.87 L W L W L     40,513                 Kylian Mbappé - 15
## 6   0.61 D D W W W     35,359 Olivier Giroud, Rafael Leão... - 8
##         Goalkeeper
## 1          Ederson
## 2          Alisson
## 3     Manuel Neuer
## 4 Thibaut Courtois
## 5     Keylor Navas
## 6     Mike Maignan
```

```
##   Country n prop lab.ypos
## 1  it ITA 8 25.0    12.50
## 2  fr FRA 5 15.6    32.80
## 3  es ESP 6 18.8    50.00
## 4 eng ENG 7 21.9    70.35
## 5  de GER 6 18.8    90.70
```



From the pie graph, it can be confirmed that the Italian league has the most number of clubs in the top 32. This supports the idea that Italy has been the best country in terms of overall club quality and can be confirmed further with the fact that Italy won the UEFA Euros 2020 (which took place in 2021 due to COVID). England comes at a close 2nd according to the pie chart, with 7 clubs in the top 32 as opposed to Italy's 8. This is further strengthened, again, as England were Italy's opponents in the Final of the Euros. It is a testament to both nations that both of them are at the pinnacle of the sport at the moment.

Conclusion

Through the use of scatterplots, linear regression models and a pie chart, we were able to demonstrate techniques that facilitate the observations of overperformance and underperformance in expected goals scored and expected goals conceded, as well as confirmed Italy as the current best nation overall according to UEFA's ranking system, based on how many Italian clubs were there in the top 32 teams.