# FIFA vs Real Life

*A Data Analysis Of Football*

***COMP 30780 Data Science in Practice***

Michael Mallon(15138708), Mark Kennedy (15312186)

School of Computer Science

University College Dublin

*May 11, 2018*

**Abstract**.

Our project FIFA vs Real Life is an analysis of real life football and the world's most popular virtual football game EA Sports FIFA. In our analysis we test the accuracy of FIFA player statistics of each individual and compare them against the last 5 years of real life data. The inspiration behind our project is derived from our love for the game and a quote from an EA Sports Producer, Michael Müller-Möhring.

***'You can't rely on the stats 100%, because even if they are comprehensive, they don't always tell the whole story…***
***However, some players just don't work well with this formula, meaning they end up getting rated much lower than their real-world performance would indicate.'*** [1]

This quote sparked an interest in us both to investigate these claims, as we are avid players of FIFA and regular viewers of football.

**Declaration**. We (Mark Kennedy, 15312186 and Michael Mallon, 15438708) declare that this assignment is our own work and that we have correctly acknowledged the work of others. This assignment is in accordance with University and School guidance[2] on good academic conduct in this regard.

[1] See https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings

[2] See https://www.cs.ucd.ie/sites/default/files/cs-plagiarism-policy_august2017.pdf

# 1. Introduction

Soccer is said to be on of the most unpredictable sports in the world, with teams like Leicester City winning the English Premier League beginning the season at 5000/1 odds. It is this unpredictability that has lead to it being the most popular sport in the world. This volatility has installed itself in the core of football making it hard for bookmakers and statictions alike making accurate judgments. EA Sports FIFA a football video game applies rating to every professional player using said stats. The aim of this project is to test the accuracy of these ratings and to improve them.
Throughout the project we will be analysing and exploring if FIFA's ratings are authentic focusing on how real-life ratings reflect the video game and visa versa.

The aim of the project is to create a reproducible dataset containing real life football data and FIFA ratings that could be used to increase the accuracy of EA's rating but also have other uses. The dataset could be used when using the latest information on every league to find the most effective players to pick in virtual games like Fantasy Football. It could be used to identify underrated players that could be being sold less than their actual worth by professional clubs or used to predict who will win games for online gambling. It could be utalized with FIFA's career mode to predict players with high potential that will be cheap.

The project will be split into three main research questions, which look at comparing the game to real-life in different scenarios. The first will be seeing if we use an in game rating is possible to predict a player's real life form and are there any factors the indicate why a player may be underrated or overrated in FIFA. The second will look to see if there are any trends in FIFA between where a player is playing their football and how they perform in real life and a rated in the game. The final will set out to test the accuracy of FIFA's more lesser known attributes.

## 1.1. Crash Course in FIFA 18

Two fundamental items we will talk about frequently in the workings of this project are FIFA 18 game and Fantasy Football.

**FIFA 18** is a computer came which simulates the sport of football. It holds an official license given by FIFA (Fédération Internationale de Football Association) to over 700 teams from over 30 domestic leagues and 160 countries. There are over 18,000 players in the game who are each given an overall rating between 0 and 100. Ratings are generated using a collaboration of a network of 9000[3] game reviewers watching games rating each player's abilities. This data is then submitted to a group of 300 data editors who review the submissions and generate a player rating in 300 different fields. A player's final rating is a sum of different sub-ratings called attributes. Attributes are specific ratings each player has they include strength, speed, jumping ability and many more. A player is given an overall rating which is the sum of each attribute called special. This ultimately decides their rating when normalised with other players in the same position.

[3] See https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings

## 1.2. Crash Course in Fantasy Football

**Fantasy Football** is a game based on real life football. In the game you create a virtual team using a spending budget where each player has a value. Points are awarded to players when they contribute to a match in real life.

| | Points |
|---|---|
| Start | 2 |
| Goal | 5 |
| Assist | 3 |
| Clean Sheet | 4 |
| Clean Sheet Partial | 2 |
| Yellow Card | -1 |
| Red Card | -3 |
| Start as Substitute | 0 |
| Came On | 1 |
| Taken Off | 0 |
| Saved Penalty | 5 |
| Missed Penalty | -2 |
| Goals Conceded | -1 |
| Goals Conceded +1 | 0 |
| Own Goal | -3 |

*Table 1. Shows the Points System for Fantasy Football*

This breakdown of points is then accumulated for each player and their total points are utilized in the game. Each game week a team is given total points for every player in their starting team and the aim of the game is to get the highest points. Leagues can be created to play against friends and leaderboards exist in each country.

# 2. Motivations & Objectives

## 2.1 Background & Motivations

When we first began this module we both had the same mindset on the type of project we would choose. We wanted a project based around a topic which was of actual of interest to us, otherwise we felt we would lose motivation quickly. We decided that our project would be different from other groups and explore new areas rather than previously analysed data. After talking though ideas Football was the most prevalent. We both share a keen interest in the sport and are fond of playing games of FIFA to determine bragging rights. Recently there has been a surge in the amount of data readily available on football matches. Applications such as Livescore and Sofascore provide live updates and statistics at the tap of a button.

Our initial idea was to focus on these match statistics with the possibility of being able to predict the winner even the score of a game. However after research we found that it is an area in which a vast array of research has already been done. We would struggle to create a unique project original that was not a carbon copy of another. We later decided we would do our project on FIFA 18 alone as we thought any conclusions we could make would be interesting and useful to us. Our main concern with this project idea was that we would only be analysing the game which of course has been researched before. We struggled to see how we could turn this idea into a fitting project that would really give us our desired project. We came to the conclusion idea of comparing FIFA to real life football would be the most interesting and fruitful project we could make. Incorporating real life statistics into generated ones gave us the freedom to direct the project the way we desired. Having such a unique project would give us the individuality we wanted from the off. Utilizing a Fantasy Football dataset was our means to an end providing us with real life statistics but also another area that we could expand to, creating the best and cheapest XI for the game. This was an idea in which had never been examined before a combined dataset would have great promise.

## 2.2. Research Questions

### 2.2.1 RQ1: Underrated & Overrated Players

Are there underrated or overrated players within the game?

a) Can a player's in-game rating be used to predict real-life form? or is there previous average form a better indicator for this?
b) What factors lead to a player becoming underrated or overrated in the game?

### 2.2.2 RQ2: Homegrown Players vs Foreign Players

Do homegrown players generally perform the best in their respective leagues or are they inferior to their foreign counterparts players?

a) Can we spot any trends in the nationalities of players playing in leagues?
b) Can we spot any trends in leagues?
c) Are there any recurring general trends with Homegrown and Foreign players? Are the results reflected in FIFA's ratings?

### 2.2.3 RQ3: Accuracy of In-Game Attributes

Are player's attributes a fair indication of their abilities? Do a player's in-game attributes correlate with real-life statistics which back them up?

# 3 Data Wrangling

## 3.1. Data Acquisition

Our project contains two key datasets:

**FIFA 18 Dataset**[4]

Our searches for a FIFA Database began after identifying a website sofifa which was an ideal source that contained all the information we desired. After contacting a website administrator we ran into troubles as we were informed the data was split into multiple snapshots rather than one. We later found a publicly available dataset on Kaggle scraped from the sofifa website, this was ideal as it saved us the hardship of scraping and ethical considerations. The dataset contains over 18,000 from over 160 different countries. This database was previously scraped from the FIFA Official Player Database and published. Contained in the database was 185 columns, eighteen columns containing personal information (ie: Full Name, Age, Nationality and Position), all other columns contained the individuals FIFA generated ratings (ie: Overall, Potential, Speed and Shot power). Many of these generated ratings where irrelevant to our project and removed in the research question preparation process.

**Fantasy Football Dataset**[5]

The second dataset was acquired from Statbunker.com. This website provided real life football statistics for the past five years in each league we chose. This data was gathered using a simple pd.read_html function in python allowing us to read each of the tables from the website. Statbunker contained a ranking system awarding each player with points an accumulation of goals, assists etc and a column contain each individual statistic. The website used a universal point system across all leagues which was a huge advantage as other sites did not. The site was originally intended to be used for fantasy football a virtual game however it contained exactly the information we required to analyse players and was easily accessible. After originally attempting to use the information from official fantasy premier league sites for every league we learned that they all had a strict policy against scraping their data.This left Statbunker as an easy alternative method. The second dataset was split into six separate datasets for each of past five years for every league and the last being this years incomplete season.

[4] See https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset

[5] See https://www.statbunker.com

## 3.2. Data Cleaning & Preparation

The data cleaning and preparation contained three main stages. The most time exhausting stage in this project resulted in the joining of our datasets. A consideration we made at the beginning of our cleaning was to only consider players who have played in the Top Five European Leagues (England, France, Germany, Italy and Spain) for the past five years as this data would be the most accurate at worthwhile in our analysis.

There was little cleaning needed in the Kaggle FIFA dataset at it had been previously cleaned by the person who had published it. Two changes had to be implemented to the data and was only introduced an attempted unsuccessful first join of the datasets. The first being the standardisation of strings in name columns to remove any possibility of players statistics not merging. To accomplish this we created a lambda function that would standardise all the relevant columns. This change was implemented plainly for reproducibility purposes rather than our individual project as it would allow for mistakes in player names in a dataset. An example of where this could occur would be with the player Mesut Özil who contains an umlaut in his name may contain it in one database but not the other. The second change was to add two additional columns First Initial and Surname necessary for joining.

There were many small items that needed addressing in our Fantasy Football datasets as they had never been cleaned before due to it being taken from a website rather than Kaggle. This cleaning included changing datatypes of columns to their correct format, replacing hyphens with 0's when a player had no values in a section and adding new columns. The two columns we created were Apps (Appearances) which was the total appearances a player had accumulating their games started and games coming off the bench and Form which was the total points a player had in a season divided by their total appearances. Once again the two columns First Initial and Surname where created necessary for our joining method.

As stated above joining was the most challenging stage in our project, this was due to inconsistencies after joining.The problems were later identified to be duplicate names, the presence of diaeresis in a players name and if a player had transferred club mid-season.
Rectifying these problems was problematic but a solution joining players on their First Initial and Surname leaving only players with the same name. This was resolved by removing players that birth dates did not match giving us a final dataset of to 2400 players per season which eventually accumulated to 3400 players per research question when players who had played in the last five years in Europe's Top 5 leagues were included. The final dataset included over 150 columns of both personal information, real life football statistics and FIFA generated ratings.

# 4 Data Analysis & Results

## 4.1. RQ1: Underrated & Overrated Players

### 4.1.1. Datasets

In order to determine underrated and overrated players we needed a dataframe which contains both a player's overall and their real-life form (past & present).

**Filters Applied:**

This question only looks at players who have played at least 2 significant seasons (9+ games) within the top 5 European leagues, previous to this year, in order to assure that results are as accurate as possible.

We've also discarded players who have not made an appearance this season within those leagues, otherwise we would be unable to determine if they are underrated or overrated, as there would be no current form to compare with.

**Prerequisite:**

**Form:** Total Fantasy Points/ Total Appearances made

**Previous Average Form -** Average form, from the years the player has spent in the top five leagues within the last 5 years.

**Current Average Form -** The player's current form in the 2017/2018 season (current season)

**Predicted Form -** The form that is generated using linear regression, it is the form that a player's in-game overall rating suggests that they have.

**Distance -** The difference between a player's ' Previous Average Form' and their 'Predicted Form'

**Verdict -** Determines whether or not a player is underrated or overrated, by checking which is closer to their 'Current Form', is it their 'Previous Average Form' or is it their 'Predicted Form'

### 4.1.2 Approach

Our approach for determining underrated and overrated players was be broken up into 2 main sections.

1. Identifying overperforming and underperforming players
2. Determining whether or not overperforming players are in fact underrated, or if underperforming players are actually overrated.

For this question we were dealing with the relationship between a player's in-game rating and their real life form.

In this case their real life form is determined by their total Fantasy Football 'Points' divided by the total amount of 'Appearances' they made.

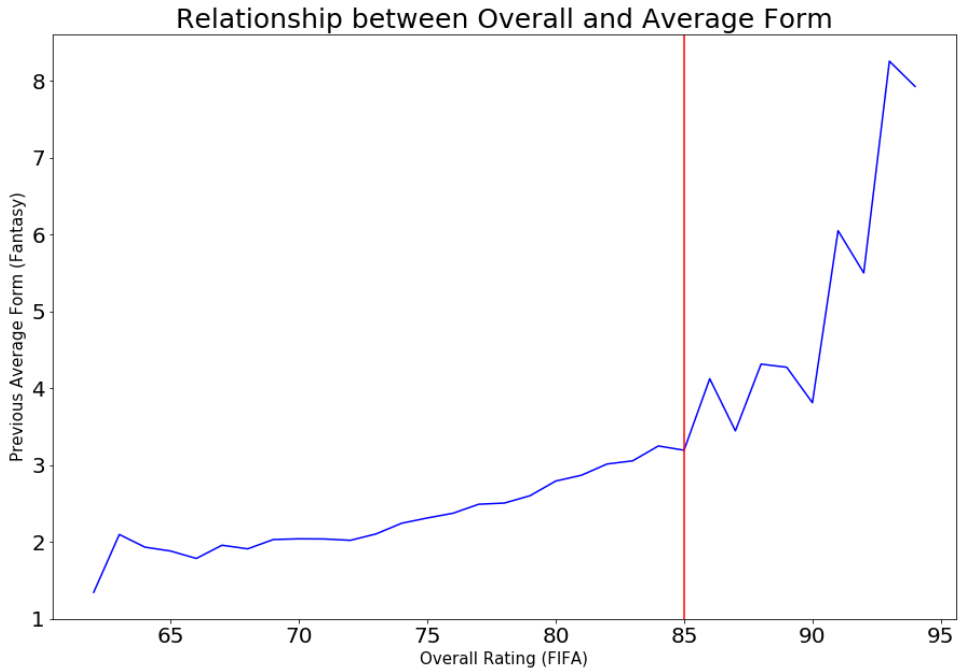



*Figure 1. A Graph showing Overall Rating vs Previous Average Form*

There is a moderately positive relationship between the two values for all players within or joined dataset (correlation of **0.62**). Notice how the rate of change in the line accelerates after 85. The graph is exponential. This is because the best players in the world are rated 85 or above in the game.
The original idea was to use linear regression on the data as it is and use the vertical distance of a point to the prediction line, to determine if a player was underperforming or overperforming in previous years.
However, the exponential nature of the graph means that the linear regression method would but inaccurate as all of the top players would be adjudged to be over performing for their rating.

Another factor which may affect the outcome is that in Fantasy Football players are generally scored different depending on what position they play, eg a forward is rewarded points for scoring and assisting goals, meanwhile goalkeepers are rewarded points for keeping clean sheets and saving penalties. As result players who play further up the field tend to score higher in the Fantasy Football table as goals are more common in real-life than clean sheets.
Taking these two points into consideration we decided it was best to divide the players up by what position they played, and also if they were rated 'less than' (<) 85 or 'greater than or equal to' (>=) 85.
The first step was to identify underperforming and overperforming players. ie, players who perform better than the majority of their counterparts with the same in-game overall rating, would be considered to be overperforming.
Plotting the data we have as a scatter plot and using linear regression to get a prediction line, we could use the vertical distance of a player from that line to determine if they are underperforming or overperforming.
For average players (overall < 85) we decided that players who are more than twice the average distance from the line will be determined as underperforming or overperforming for their rating. Meanwhile, world-class players (overall >= 85) we decided that players who are over the average distance will be determined as underperforming or overperforming for their rating.
From our lists of underperforming and overperforming players, we had to then identify the players who were overrated and underrated respectively. To do this we took their 'Previous Average Form' and 'Predicted Form' (the form which their in-game rating suggests they have), and seen which one was closer to predicting their 'Current Form' (Form for the current 17/18 season).
Underperforming players whose previous average form was closest to predicting their current real-life form were then determined to be overrated.
Overperforming players whose previous average form was closest to predicting their current real-life form were then determined to be underrated.
Finally, we analysed the underrated and overrated players to see if we could identify reasons as to why they are so.

**(Note: As of 11/05/2018 (Project Deadline) the 2017/18 season was still incomplete, results in notebooks may differ slightly from those shown in the report if run at the end of the season.)**

### 4.1.2. Results

#### 4.1.2.1. Identifying Underperforming and Overperforming Players

Below we can see the scatter plots which we used to identify underperforming (represented in green) and over performing players (represented in red) . We have two graphs for each position, one showing average players (overall < 85) and the other showing world class players (overall >= 85).

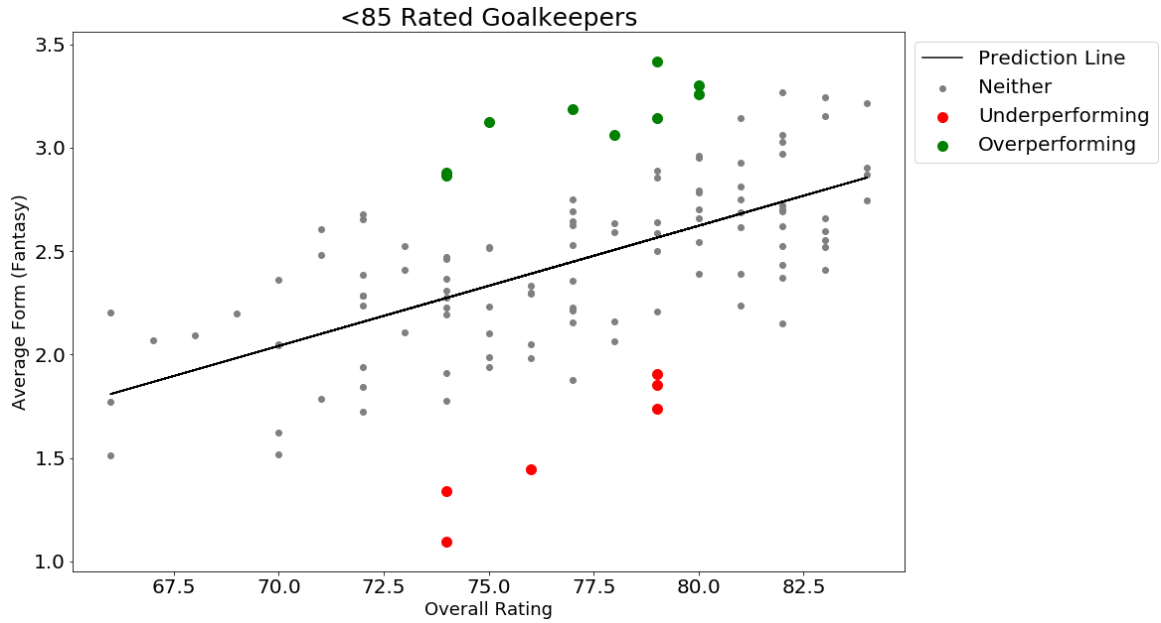



*Figure 2. A Graph showing Average Form vs Their Overall Rating for Goalkeepers rated < 85*

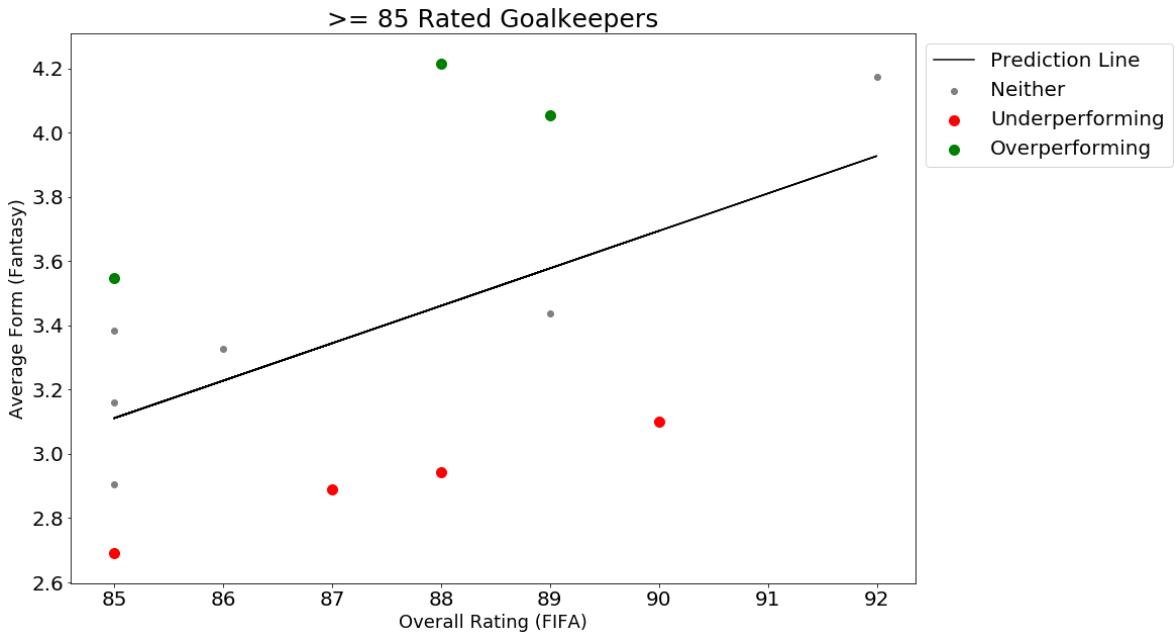



*Figure 3. A Graph showing Average Form vs Their Overall Rating for Goalkeepers rated >= 85*

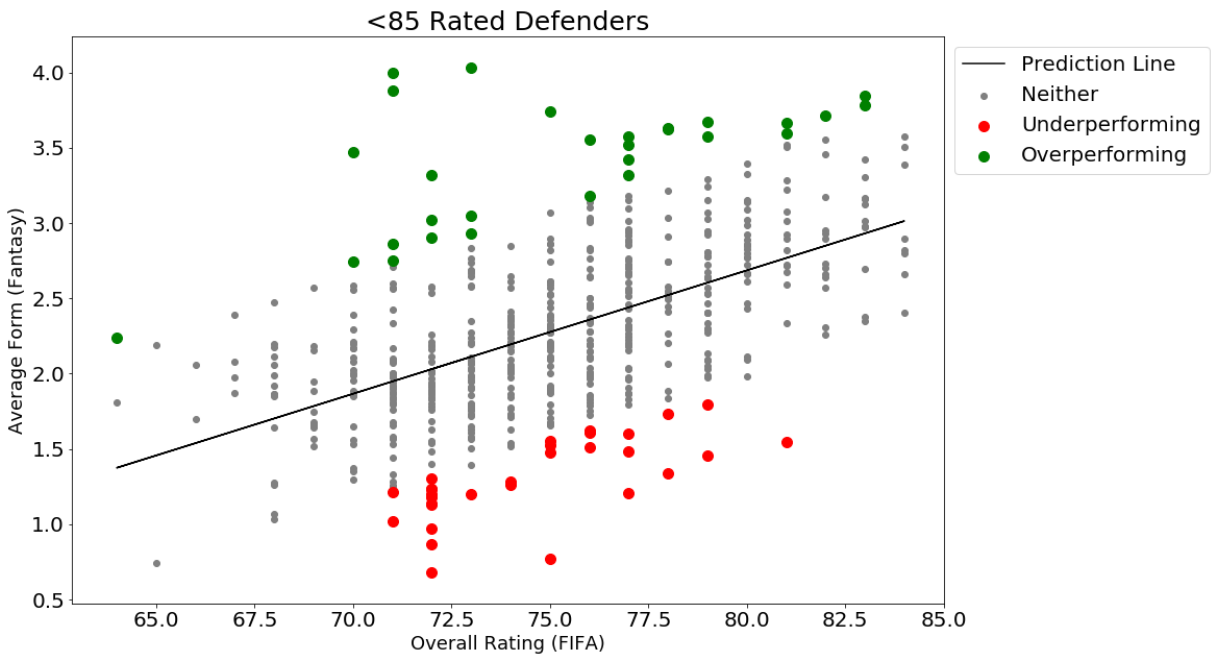



*Figure 4. A Graph showing Average Form vs Their Overall Rating for Defenders Rated <85*

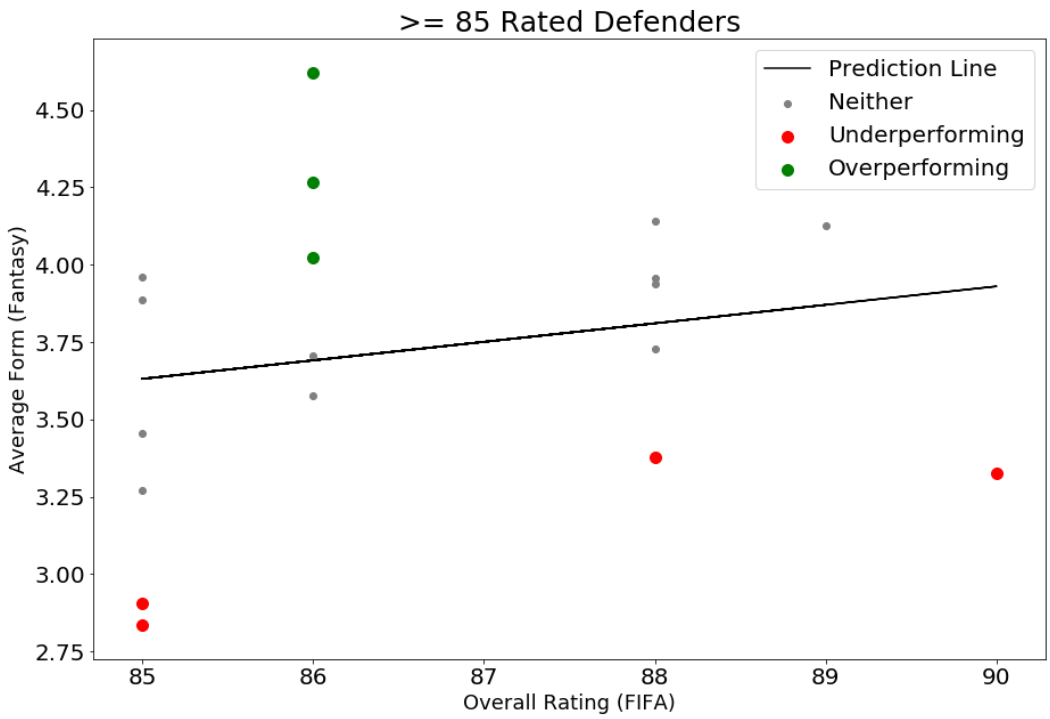



*Figure 5. A Graph showing Average Form vs Their Overall Rating for Defenders rated >= 85*

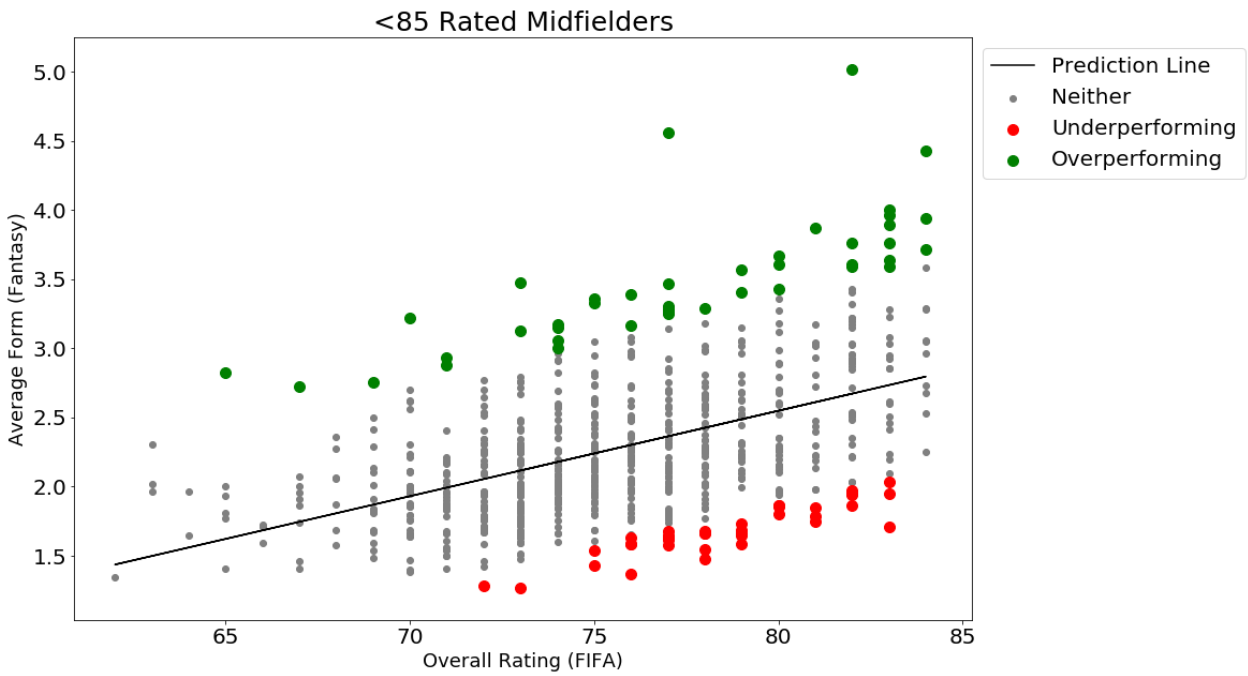



*Figure 6. A Graph showing Average Form vs Their Overall Rating for Midfielder rated < 85*

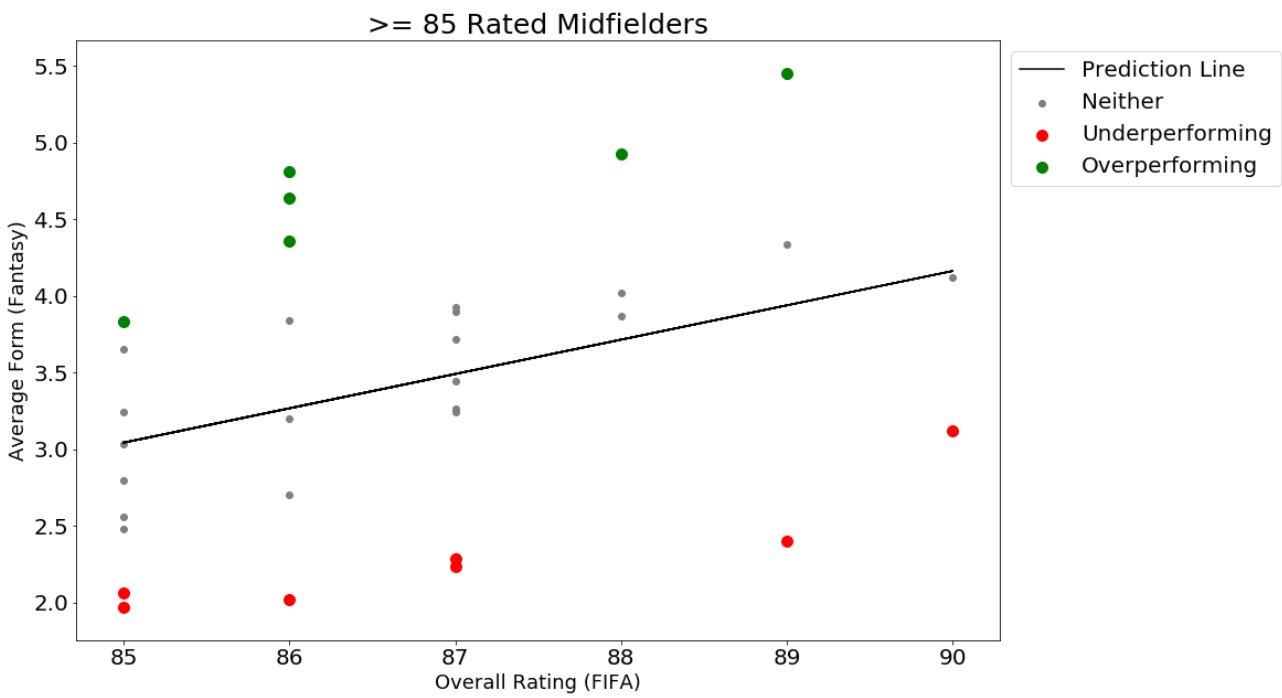



*Figure 7. A Graph showing Average Form vs Their Overall Rating for Midfielders rated >= 85*

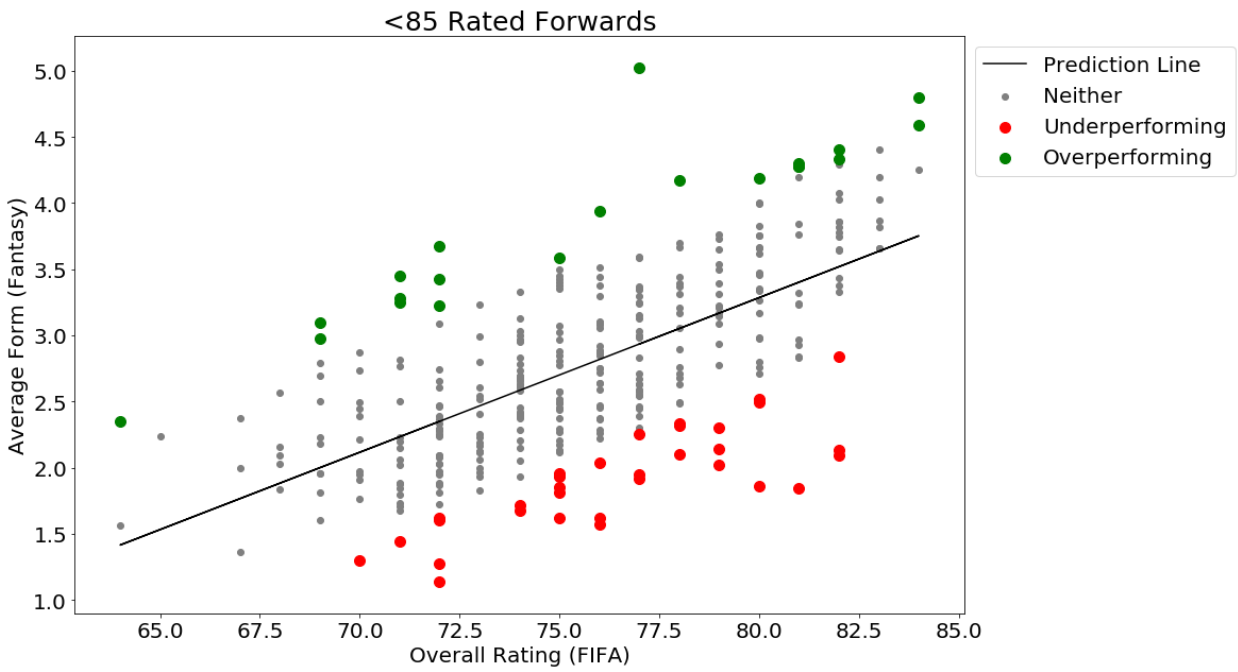



*Figure 8. A Graph showing Average Form vs Their Overall Rating for Forwards rated < 85*

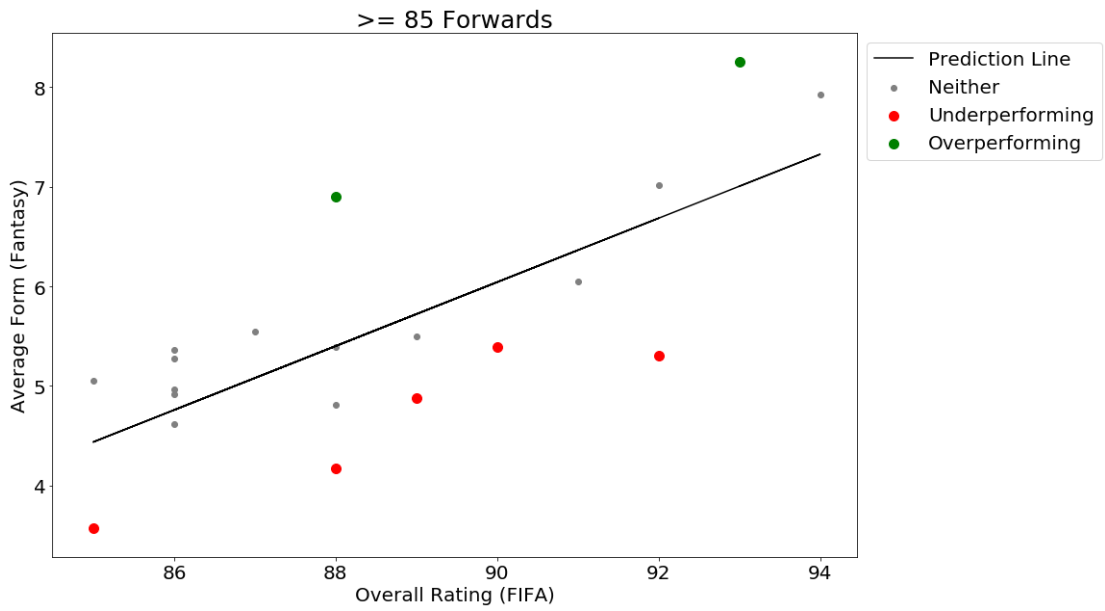



*Figure 9. A Graph showing Average Form vs Their Overall Rating for Forwards rated >= 85*

#### 4.1.2.2. Determining whether or not overperforming players are underrated.

Overperforming players whose previous average form was closest to predicting their current real-life form were then determined to be underrated. Otherwise we say that they were just 'over performing' in previous years, but not underrated.

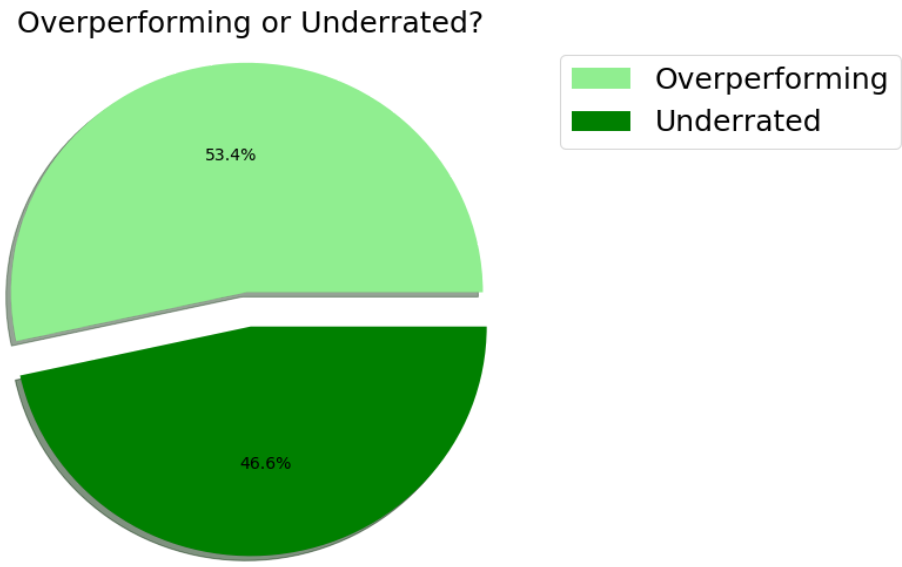



*Figure 10. Shows the percentage of underrated players among the overperforming players we identified in the first part.*

Of the overperforming player's 46.6% of them were actually underrated in this years game. The other 53.4% were just overperforming in the previous 5 years.
From this we created a new dataframe which contains all of the underrated players. We determined how underrated they were by finding the difference between their 'Previous Average Form' and their 'Predicted Form', the form their fifa rating suggest that they have. (Distance)

**Top 10: Most Underrated Players:**

| | Players | Position | club | overall | Predicted Form | Previous Average Form | Current Average Form | Distance |
|---|---|---|---|---|---|---|---|---|
| 0 | Marco Reus | Midfielder | Borussia Dortmund | 86 | 3.265931 | 4.806452 | 5.111111 | 1.540520 |
| 1 | Junior Stanislas | Midfielder | Bournemouth | 73 | 2.116653 | 3.476190 | 3.473684 | 1.359537 |
| 2 | Roberto Firmino | Midfielder | Liverpool | 83 | 2.734483 | 3.890909 | 4.571429 | 1.156426 |
| 3 | Sadio Mane | Midfielder | Liverpool | 84 | 2.796266 | 3.936170 | 4.370370 | 1.139904 |
| 4 | Lucas Hernandez | Defender | Atletico Madrid | 77 | 2.440202 | 3.576923 | 4.400000 | 1.136721 |
| 5 | Nabil Fekir | Midfielder | Olympique Lyonnais | 82 | 2.672700 | 3.764706 | 5.481481 | 1.092006 |
| 6 | James Rodriguez | Midfielder | FC Bayern Munich | 86 | 3.265931 | 4.357143 | 4.428571 | 1.091212 |
| 7 | Dele Alli | Midfielder | Tottenham Hotspur | 84 | 2.796266 | 4.428571 | 4.156250 | 1.632306 |
| 8 | Mauro Icardi | Forward | Inter | 84 | 3.752074 | 4.793548 | 6.483871 | 1.041475 |
| 9 | Anthony Martial | Midfielder | Manchester United | 82 | 2.672700 | 3.593220 | 3.620690 | 0.920520 |

*Table 2. Lists the top 10 most underrated players*

#### 4.1.2.3. Determining whether or not underperforming players are overrated.

Underperforming players whose previous average form was closest to predicting their current real-life form were then determined to be overrated. Otherwise we say that they were just underperforming in previous years, but not overrated.

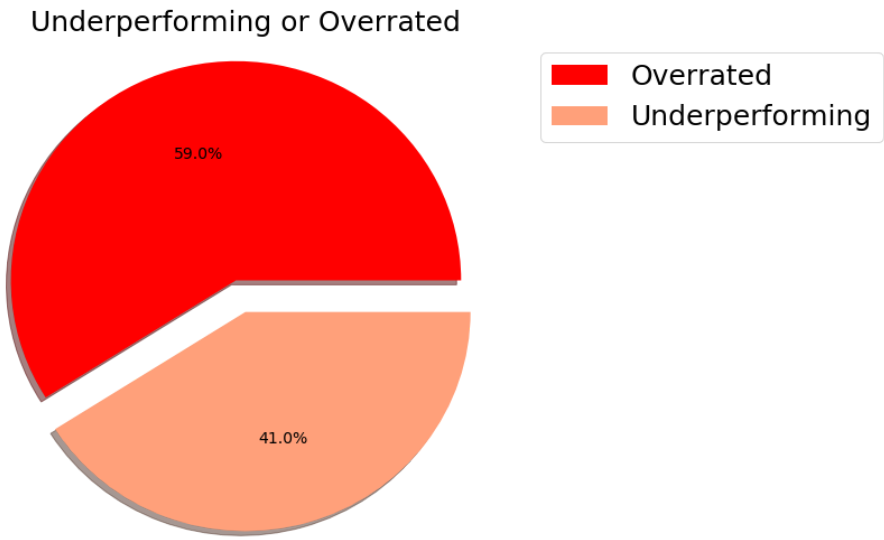



*Figure 11. Shows the percentage of overrated players among the underperforming players we identified in the first part.*

Of the underperforming player's 59% of them were actually overrated in this years game. The other 41% were just overperforming in the previous 5 years.
From this we created a new dataframe which contains all of the overrated players. We determined how Are you going to maths. overrated they were by finding the difference between their 'Previous Average Form' and their 'Predicted Form', the form their FIFA rating suggests that they have (Distance)

**Top 10: Most Overrated Players:**

| | Players | Position | club | overall | Predicted Form | Previous Average Form | Current Average Form | Distance |
|---|---|---|---|---|---|---|---|---|
| 0 | Manu Trigueros | Forward | Villarreal CF | 82 | 3.518352 | 2.094891 | 2.129032 | -1.423461 |
| 1 | Marco Verratti | Midfielder | Paris Saint-Germain | 87 | 3.489731 | 2.231343 | 1.909091 | -1.258388 |
| 2 | N'Golo Kante | Midfielder | Chelsea | 87 | 3.489731 | 2.284404 | 2.161290 | -1.205328 |
| 3 | Luka Modric | Midfielder | Real Madrid CF | 89 | 3.937332 | 2.400000 | 2.608696 | -1.537332 |
| 4 | Erik Moran | Forward | CD Leganes | 72 | 2.349742 | 1.275862 | 1.000000 | -1.073880 |
| 5 | Jorge Mere | Defender | 1. FC Koln | 78 | 2.522151 | 1.339286 | 1.428571 | -1.182865 |
| 6 | Muhamed Besic | Midfielder | Everton | 76 | 2.302002 | 1.371429 | 1.000000 | -0.930574 |
| 7 | Nordin Amrabat | Forward | CD Leganes | 79 | 3.167769 | 2.300000 | 2.000000 | -0.867769 |
| 8 | Tobias Strobl | Defender | Borussia Monchengladbach | 76 | 2.358253 | 1.513761 | 0.333333 | -0.844491 |
| 9 | Alexis Sanchez | Forward | Arsenal | 89 | 5.720774 | 4.879518 | 3.862069 | -0.841256 |

*Table 3. Lists the Top 10 most overrated players*

### *4.1.2.2:* Reasons why a player might be Underrated & Overrated

**Appearances:**

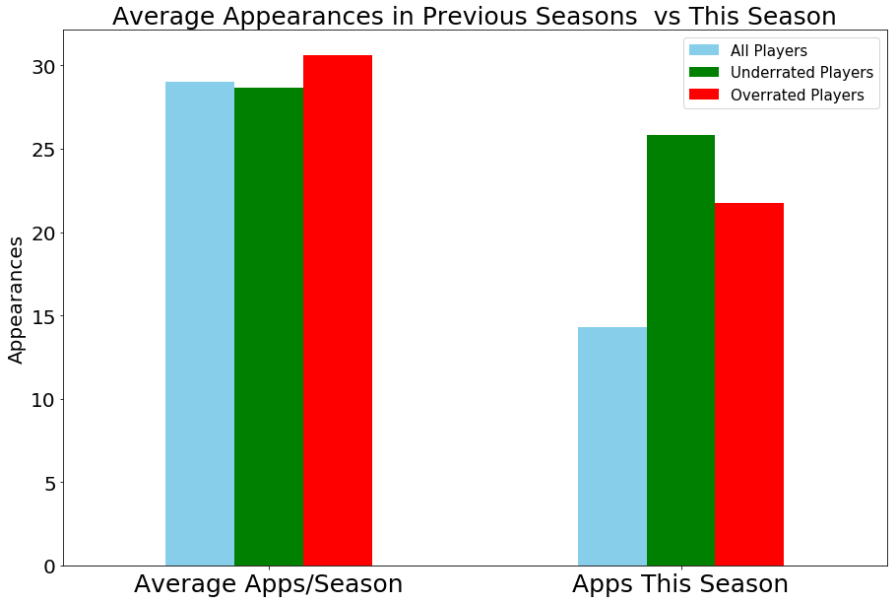



*Figure 12. Shows the average appearances made by players in previous seasons and this season*

We can see that in previous years underrated players made fewer appearances than the average. Whereas overrated players played more than average. However this season underrated players are playing the most games. Perhaps this is something that EA could not have foreseen, as players win their place on the starting team in training sessions behind closed doors.

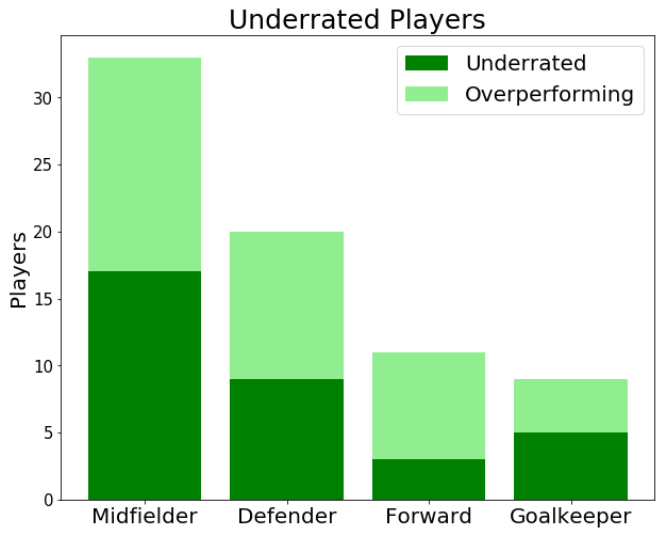
**Age:**

| | Average Age |
|---|---|
| All Players | 27.914 |
| Overrated Players | 25.8913 |
| Underrated Players | 26.5588 |

*Table 4. Shows the Average age of All, Overrated and Underrated Players*

Here we can see that Overrated players are the youngest, perhaps they are overrated because they have extra time to reach their full potential. Underrated players are older perhaps because EA believes they are beyond their best form.

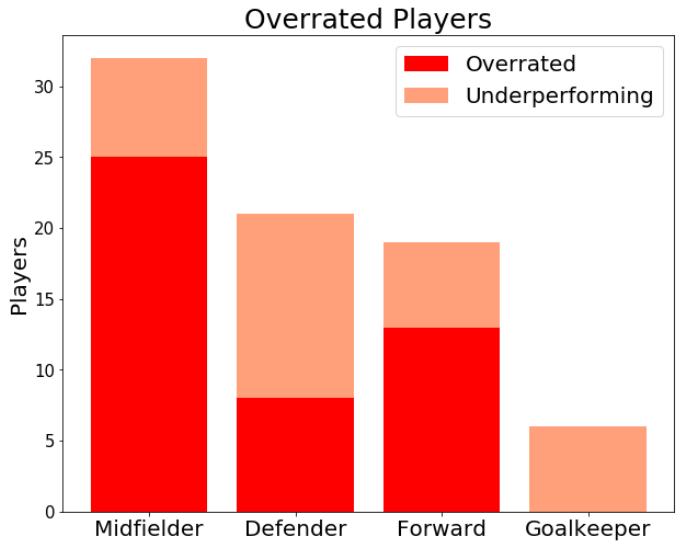
**Position:**




| Position | |
|---|---|
| Midfielder | 51.52 |
| Defender | 45.00 |
| Forward | 27.27 |
| Goalkeeper | 55.56 |

*Figure 13 & Table 5 . Show the percentage of underrated players among the overperforming players for each position.*

This stacked bar chart shows the percentage of underrated players within the group of overperforming players, for each position. The key finding to take from this graph is how low the percentage of underrated forwards there is in the overperforming forward's group (27.27%).

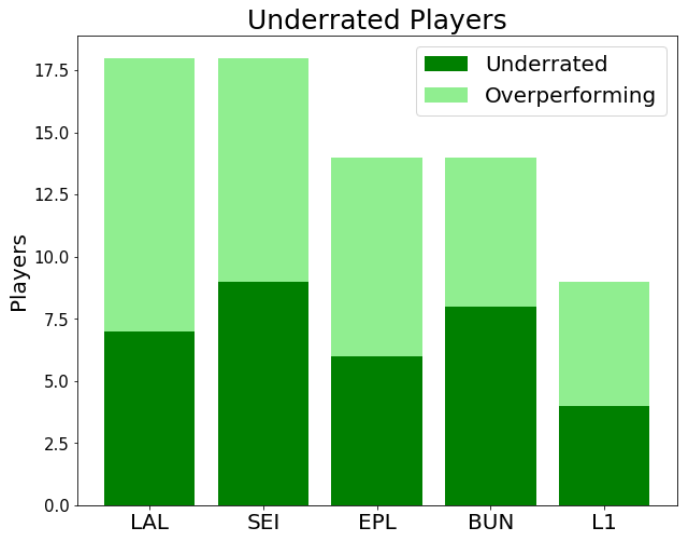



| Position | |
|---|---|
| Midfielder | 78.12 |
| Defender | 38.10 |
| Forward | 68.42 |
| Goalkeeper | 0.00 |

*Figure 14 & Table 6 . Show the percentage of overrated players among the underperforming players for each position.*

This stacked bar chart shows the percentage of overrated players within the group of underperforming players, for each position. The key finding to take from this graph is how high the percentage of overrated midfielders there is in the underperforming midfielders group (78.12%). And also the from the 6 underperforming we were unable to identify 1 overrated goalkeeper.

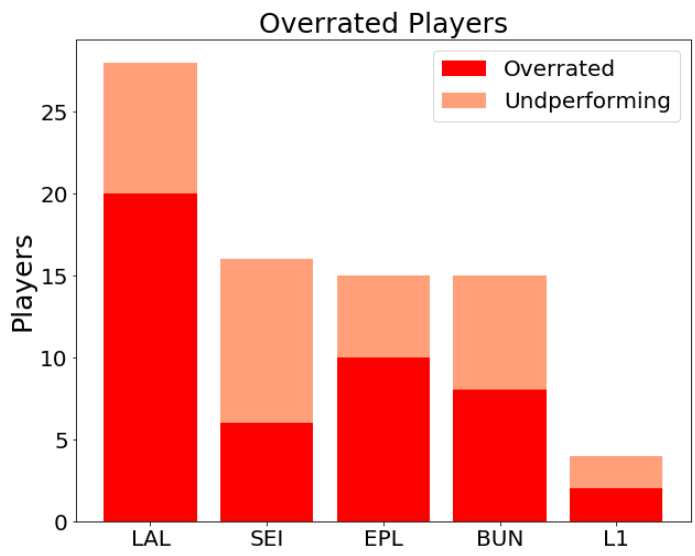
**League:**




| league | |
|---|---|
| English Premier League | 38.89 |
| French Ligue 1 | 50.00 |
| German Bundesliga | 42.86 |
| Italian Serie A | 57.14 |
| Spanish Primera Division | 44.44 |

*Figure 15 & Table 8 . Show the percentage of underrated players among the overperforming players for each league.*

This stacked bar chart shows the percentage of underrated players within the group of overperforming players, for each league. Note how the top three leagues (EPL, BUN, LAL) have the lowest percentage of underrated players. This may because EA has spent more time studying these leagues as they are definitely the most used leagues within the game.




| league | |
|---|---|
| Spanish Primera Division | 71.43 |
| Italian Serie A | 37.50 |
| English Premier League | 66.67 |
| German Bundesliga | 53.33 |
| French Ligue 1 | 50.00 |

*Figure 16 & Table 9 . Show the percentage of overrated players among the underperforming players for each league.*

This stacked bar chart shows the percentage of overrated players within the group of underperforming players, for each league.  Also note how the top three leagues (EPL, BUN, LAL) have the highest percentage of underrated players. One explanation for this may be because EA allows more overrated players in these leagues in order to make them more fun to play.

***Injury Prone Trait:***

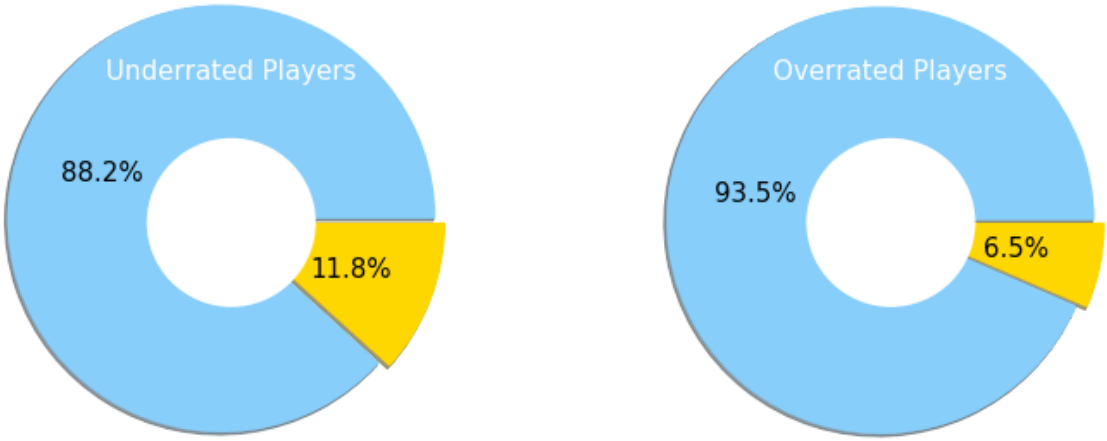



*Figure 17 . Shows the percentage of overrated and underrated players who have the injury-prone trait*

These graphs prove that if you are injury prone in real-life, it will increase your chances of being underrated in the game. A prime example of this is Marco Reus (who according to our method is the most underrated player in the game), he has missed 62 games in the last 2 seasons justifying his in-game injury prone trait. (Source: Transfermarkt).Note how low the percentage of injury prone players that are overrated. This would indicate that the injury-prone trait does influence your chances of being underrated and overrated.

### 4.1.3. Discussion

The FIFA franchise has been on the market for 25 years now, and EA has clearly put a lot of time into trying to perfect their method of generating player ratings. However their current system isn't perfect. Our results show that 59% of all underperforming players are actually underrated within the game, whereas only 46.6% of overperforming players are in fact underrated. But if we look at it from EA's point of view, they have to release the game to millions of football fans around the world annually. Which would be more fun to play and as result sell more copies for EA? A game where the majority of players are overrated or a game where most of the players are underrated. The answer is obvious, the game with the overrated players would be more fun to play. Our results hint that this is in fact the mindset of EA themselves as when we look at the most overrated league they just so happen to be the most popular league in the game.

Regarding reasons as to why players are underrated or overrated, we can see that overperforming midfielders are more than likely underrated and underperforming forwards are more than likely overrated. However there is a factor that influences this. At times in Fantasy Football a player is given the wrong position ie. listed as a midfielder but plays as a forward in real-life, therefore when they score a goal they are awarded 5 points for being a midfielder instead of the correct 4 points which forwards get for scoring. An example of this would be Anthony Martial for Manchester United, who is listed as being a midfielder but when he plays in real-life he plays further up the pitch, so he is more likely to score goals that the average midfielder.

There are other factors that influence a player being underrated or overrated within the game, such as age and injury prone trait. A possible future study that we would like to do is study how the popularity of players affects their in-game ratings. Our dataset did contain a column called 'fan_favourite_trait' which we hoped could give us an indication of popular players, however it was 'false' for every player within the game.

## RQ2: Homegrown vs Foreign Players

### 4.1.4. Dataset

The dataset used for this research question contained relevant personal information of a player (including name, position, age, defending and attacking statistics) and newly created columns listed below.

**Homegrown** - is a boolean column. If true it signifies a player who is playing in a league in their home country. ie: An English Player Playing in England.
**Abroad** - is a boolean column. If true contains all players that live in the location of the Top 5 Leagues (England, France, Italy, Spain, Germany) that play in one of the other Top 5 Leagues in these locations. Examples: A French player playing in Germany is said to be an Abroad player. A Home player is the same as a homegrown player, ie: A French player playing in France.
**Overall Apps** - contains the total number of appearances a player has made.
**Starts/Apps** - is the probability of a player starting a game. It is calculated by the total number of starts divided by the overall appearances of a player.
**Subbed Off/Apps** - is the probability of a player being subbed off in a game. It is calculated by the total number of times subbed off divided by the overall appearances of a player.
**Average Form** - is the total average form of a player. This is calculated by dividing a player's total points by their overall appearances producing average points per game.
**Goals/Apps** - is the probability of a player scoring a goal in a match. This is calculated by dividing a player's total goals by their overall appearances.
**Assists/Apps** - is the probability of a player assisting a goal in a match. This is calculated by dividing a player's total assists by their overall appearances.
**Clean Sheets/Apps** - is the probability of a player keeping a clean sheet in a match. This is calculated by dividing a player's total clean sheets by their overall appearances. This is only applicable to Goalkeepers and Defenders.

The final dataset included 3445 players with 23 columns of information who have all played in the Top 5 European leagues in the past five years.

### 4.1.5. Approach

***Do homegrown players generally perform the best in their respective leagues or are they inferior to their foreign counterparts players?***

The use of graphs was crucial in the answering of this question, giving real visualisations of what the data revealed was very helpful.
To identify if Homegrown players perform better in their home leagues or their Foreign counterparts the first step was gaining a better understanding how players typically perform from certain countries and leagues. Analysing these performances using Average Form in the case of real life football and player overall in FIFA gives a greater insight into how players normally play and are expected to play. After considering each analysis on its own comparing them side by side gave a good representation of how FIFA's ratings reflect on real life performances in both cases.When looking to count of Homegrown and Foreign players in leagues and countries pie charts showed the most effective. Bar charts were the best ways of representing this data when looking more in-depth analysis when breaking down statistics into each position.

## RQ2: Homegrown vs Foreign Players

### 4.1.4. Dataset

The dataset used for this research question contained relevant personal information of a player (including name, position, age, defending and attacking statistics) and newly created columns listed below.

**Homegrown** - is a boolean column. If true it signifies a player who is playing in a league in their home country. ie: An English Player Playing in England.
**Abroad** - is a boolean column. If true contains all players that live in the location of the Top 5 Leagues (England, France, Italy, Spain, Germany) that play in one of the other Top 5 Leagues in these locations. Examples: A French player playing in Germany is said to be an Abroad player. A Home player is the same as a homegrown player, ie: A French player playing in France.
**Overall Apps** - contains the total number of appearances a player has made.
**Starts/Apps** - is the probability of a player starting a game. It is calculated by the total number of starts divided by the overall appearances of a player.
**Subbed Off/Apps** - is the probability of a player being subbed off in a game. It is calculated by the total number of times subbed off divided by the overall appearances of a player.
**Average Form** - is the total average form of a player. This is calculated by dividing a player's total points by their overall appearances producing average points per game.
**Goals/Apps** - is the probability of a player scoring a goal in a match. This is calculated by dividing a player's total goals by their overall appearances.
**Assists/Apps** - is the probability of a player assisting a goal in a match. This is calculated by dividing a player's total assists by their overall appearances.
**Clean Sheets/Apps** - is the probability of a player keeping a clean sheet in a match. This is calculated by dividing a player's total clean sheets by their overall appearances. This is only applicable to Goalkeepers and Defenders.

The final dataset included 3445 players with 23 columns of information who have all played in the Top 5 European leagues in the past five years.

### 4.1.5. Approach

***Do homegrown players generally perform the best in their respective leagues or are they inferior to their foreign counterparts players?***

The use of graphs was crucial in the answering of this question, giving real visualisations of what the data revealed was very helpful.
To identify if Homegrown players perform better in their home leagues or their Foreign counterparts the first step was gaining a better understanding how players typically perform from certain countries and leagues. Analysing these performances using Average Form in the case of real life football and player overall in FIFA gives a greater insight into how players normally play and are expected to play. After considering each analysis on its own comparing them side by side gave a good representation of how FIFA's ratings reflect on real life performances in both cases.When looking to count of Homegrown and Foreign players in leagues and countries pie charts showed the most effective. Bar charts were the best ways of representing this data when looking more in-depth analysis when breaking down statistics into each position.

### 4.1.6. Results

#### 4.2.3.1 **Form vs Overall (Nationality)**

When comparing the average form of nationalities against the average rating many conclusions can be drawn. The first and most apparent conclusion that can be drawn is that French players are extremely underrated in general and are the best real life performers. When comparing both their form and rating side by side there is an evident difference. The graph tells us that French players on average are nearly two ratings underrated than you would expect.
It can be said that both Spanish and Italian players seem to be overrated. When observing Spanish footballers their form indicates when being compared to French they should be two ratings less and one rating less in the case of Italians.
German and English players are accurately rated when comparing form to FIFA overall.

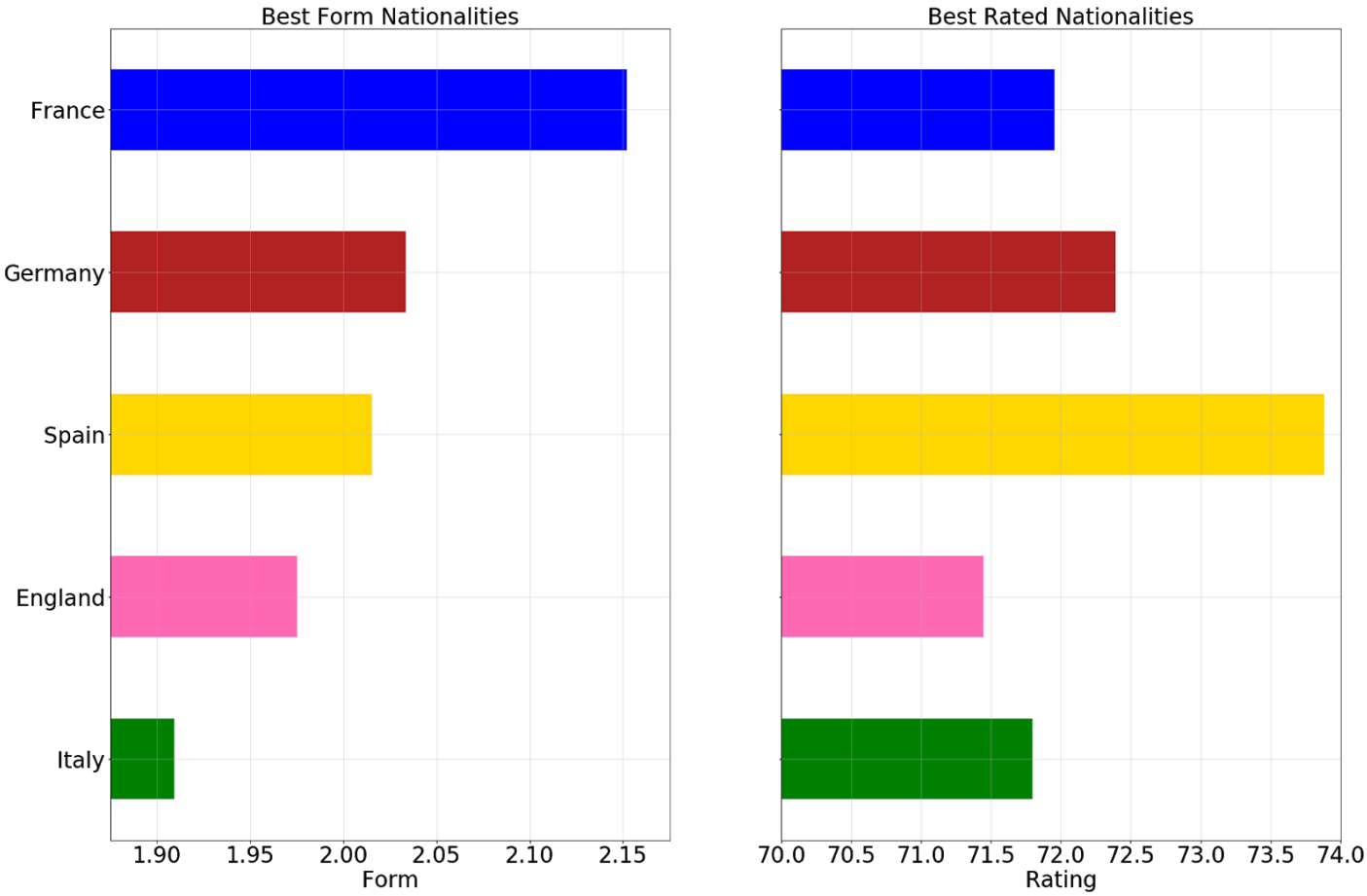



*Figure 18 . Best Form Nationalities vs Best Rated Nationalities.*

#### 4.2.3.2 **Form vs Overall (Leagues)**

The English Premier League is the best performing of all leagues and joint first in FIFA rating. When comparing the average form of leagues rather than nationalities against the average FIFA rating some very similar trends can be seen to *Figure 18*. Once again the English and German domestic leagues are very comparable like the nationalities. The pattern can be seen again were the French league seems to be underrated and the Spanish and Italian leagues overrated just as in *Figure 18*. This is a recurring trend in our data.

It is our belief that is inaccuracy is attributed to both the Spanish and Italian league being difficult to play in and the players being of a higher quality when compared to the French league. This is an assumption that frequent viewers of European football could conclude so it was affirming to see this in our data. The accuracy in both the German and English domestic leagues could be traced to the fact that both leagues have a larger viewership and number of scouts accounting creating more accurate feedback on players.

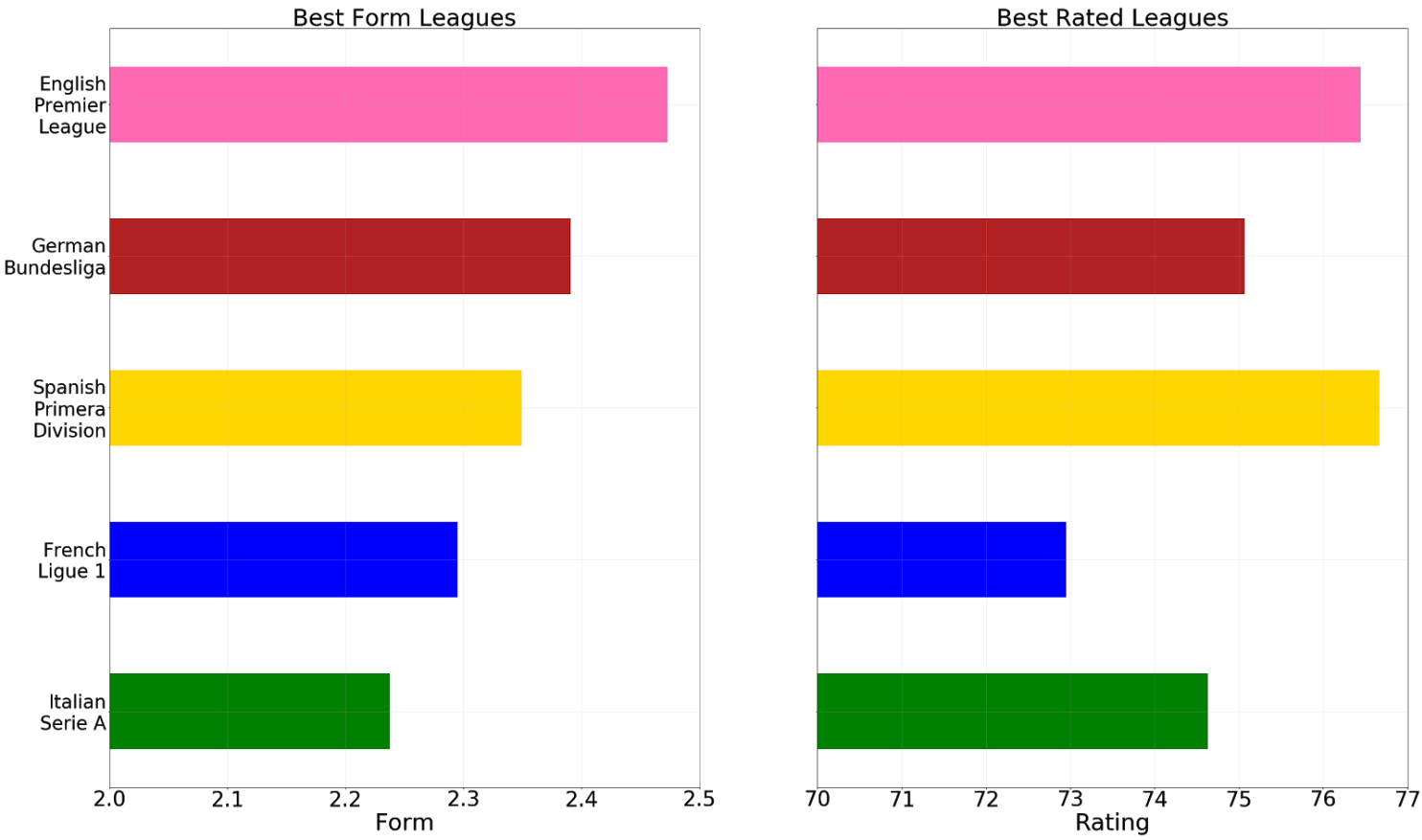



*Figure 19 . Best Form Leagues vs Best Rated Leagues.*

#### 4.2.3.3 **Homegrown/Foreign Player Count**

When looking at the visualization below there is a very apparent difference in the case of the English Premier League in comparison to every other league. Only 35% of their players are actually Homegrown. Could this be the reason why the English Premier League is rated higher in both form and overall. All other leagues have a balance close to 50/50 with Homegrown and Foreign players.

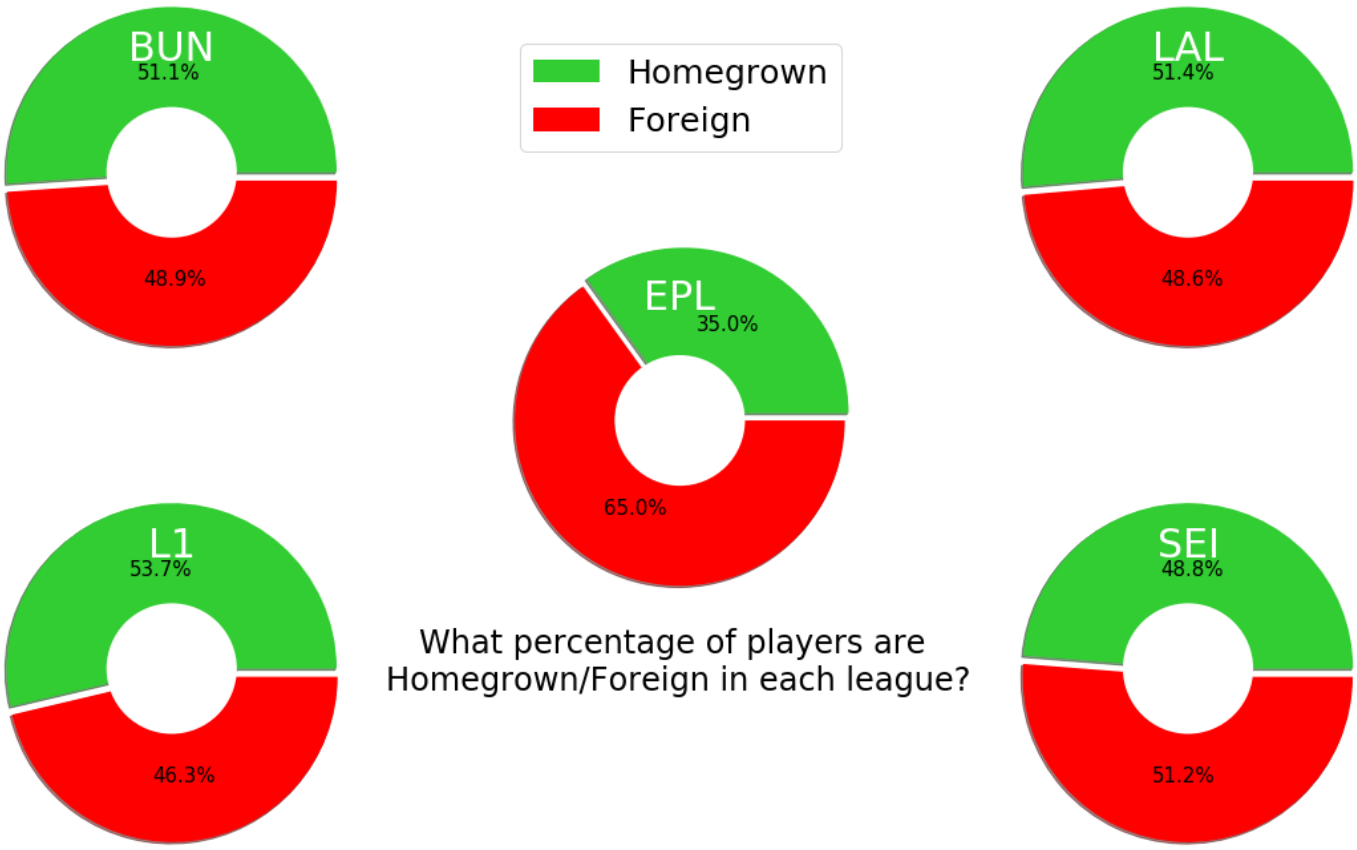



*Figure 19 . Player Count Homegrown/Foreign Leagues*

#### 4.2.3.4 **Home/Abroad Player Count**

Another shocking piece of information can be seen in this visualisation. Only 0.3% of the English players play Abroad even though it holds the best performing league. After looking into the data it could be seen that the 0.3% was actually only one player Reece Oxford, what makes him so different to every other English player?. It can be seen that he is extremely young as football player. He has only made 7 appearances in the Bundesliga and achieved a low average form of 1.43. After looking online for more information we came across an article[6] in the Independent detailing that he is only on loan to a club in the German league from West Ham an English club.It is common that teams loan youth players out to other countries to gain experience. This technically means no English players play in any of the Top 5 leagues outside England.

Reece Oxford
Age: 18.0
Nationality: England
League: German Bundesliga
Average Form: 1.43
FIFA Overall: 66.0
Overall Apperances: 7.0

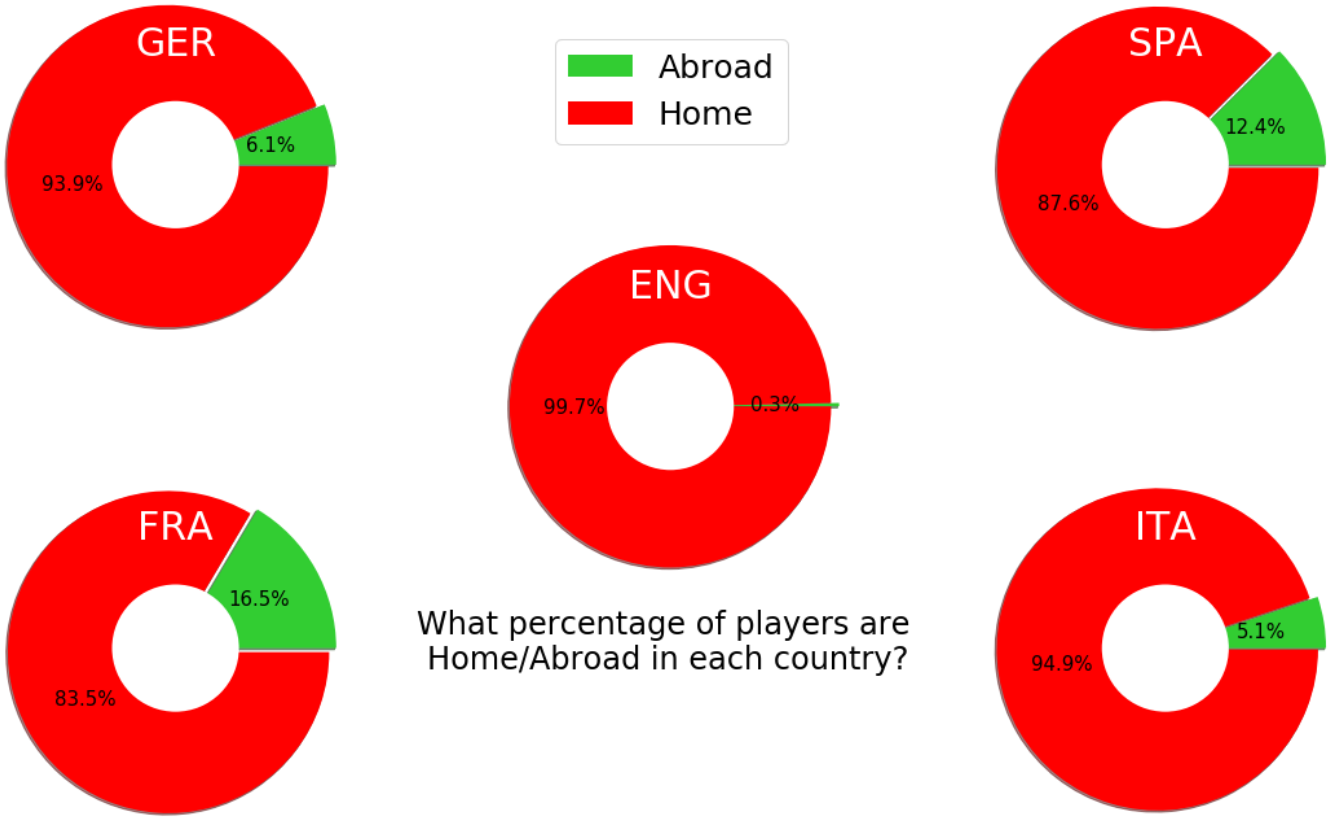



*Figure 20 . Player Count Home/Abroad Nationalities*

[6]https://www.independent.ie/sport/soccer/reece-oxford-returns-to-borussia-monchengladbach-on-loan-36553090.html

#### 4.2.3.5 **League Position Breakdown**

The next step in understanding if there is any trends between Homegrown and Foreign players was investigating each position individually using both form and overall. In the graphs below we can see from right to left Goalkeeper, Defender, Midfielder and Forward statistics in both average form and FIFA overall for all of the Top 5 leagues.

General trends we discovered were that across the board Foreign players on average have much higher ratings that Homegrown players in both real life form and FIFA ratings. Of the 32 comparable bar charts in the graphs below in only one case does a Homegrown player take over a Foreign player which is a surprising finding. There is a substantial difference in average form of Forwards when compared to any other position in every league by close to 0.5. This statistic is not reflected in FIFA's ratings. Our belief is that FIFA normalise ratings for each position hence the inaccuracies. Keeping this in mind we then look to individual leagues to see if patterns form. Although Foreign players are superior to Homegrown players FIFA overestimate Homegrown player overalls.

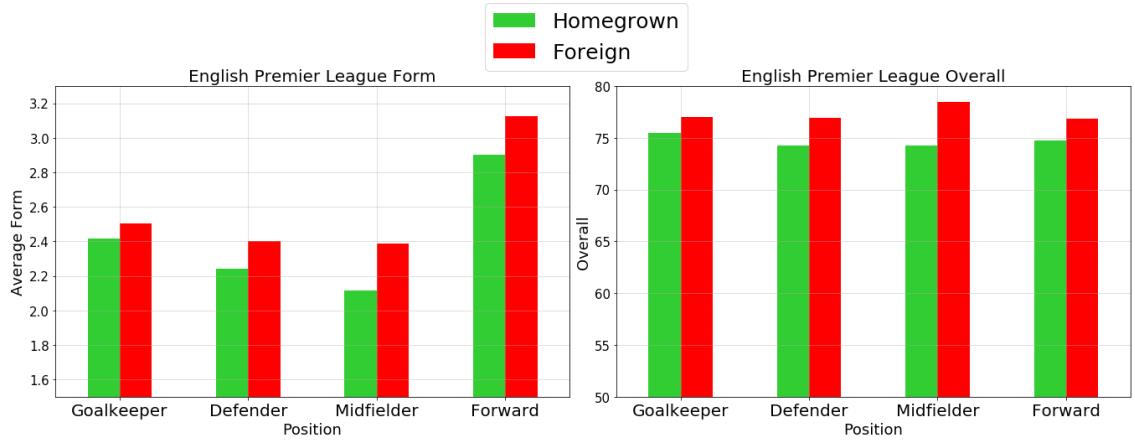



*Figure 21 .English Premier League Position Form & Overall Breakdown*

When analysing the English Premier League keeping in mind the normalisation of player ratings in their respective positions FIFA's reflection of real life is accurate. The difference between Homegrown and Foreign players is nearly exact in every case. Once again the general trend of Foreign players being superior to Homegrown can be seen.

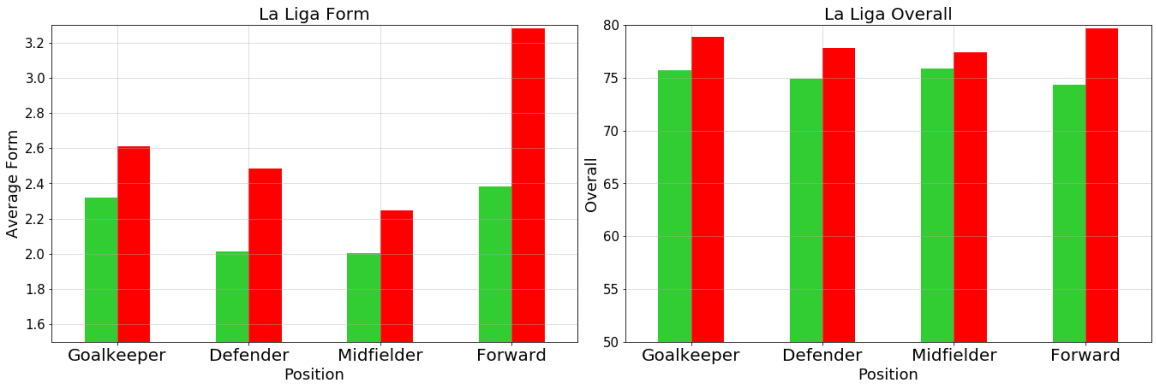



*Figure 22 . Spanish La Liga Position Form & Overall Breakdown*

Analysing the Spanish La Liga once again Foreign players surpass Homegrown players. Unlike the EPL Homegrown players are overrated when compared to Foreign players. The distances in the Form bar charts are not reflected in Overall bar charts. The most serious case of this can be seen in the Forward position.

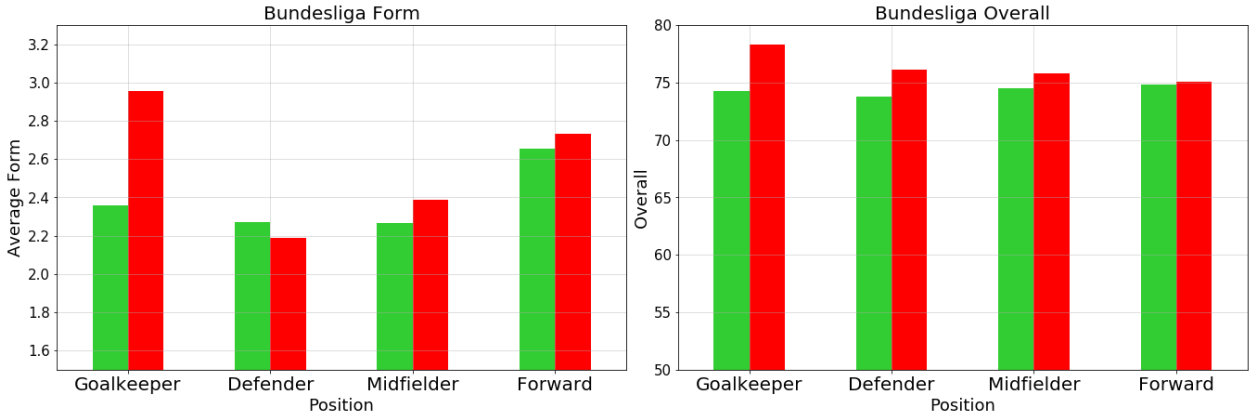



*Figure 23 .German Bundesliga Position Form & Overall Breakdown*

When analyzing the German Bundesliga there lies the one case where Homegrown players perform better than Foreign, this occurs in the position of Defenders. This is not reflected in FIFA's ratings which is hard to understand as we have seen the trend of Homegrown players being overrated generally. Homegrown goalkeepers also seem to be very overrated in this league. This is due to the player Manuel Neuer who is said to be the best performing goalkeeper in the world rated at 92 increasing the average.

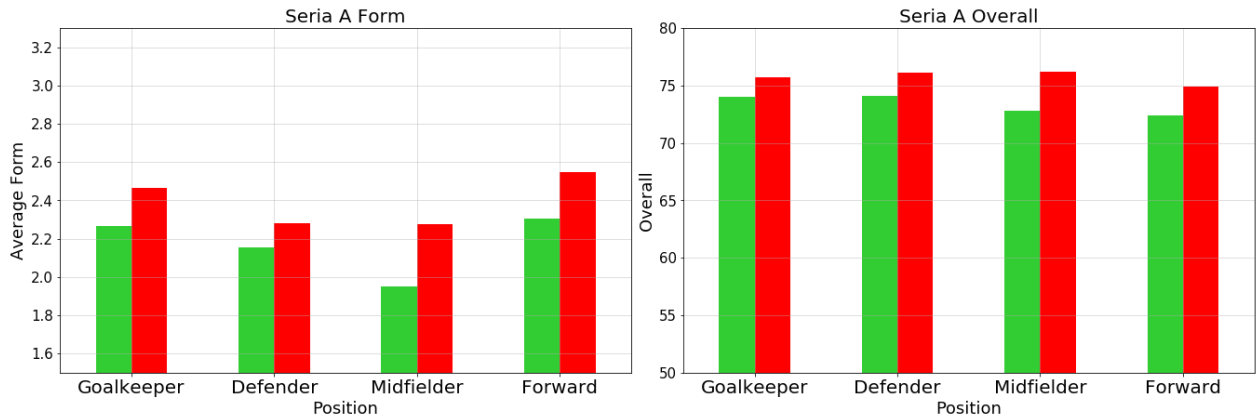
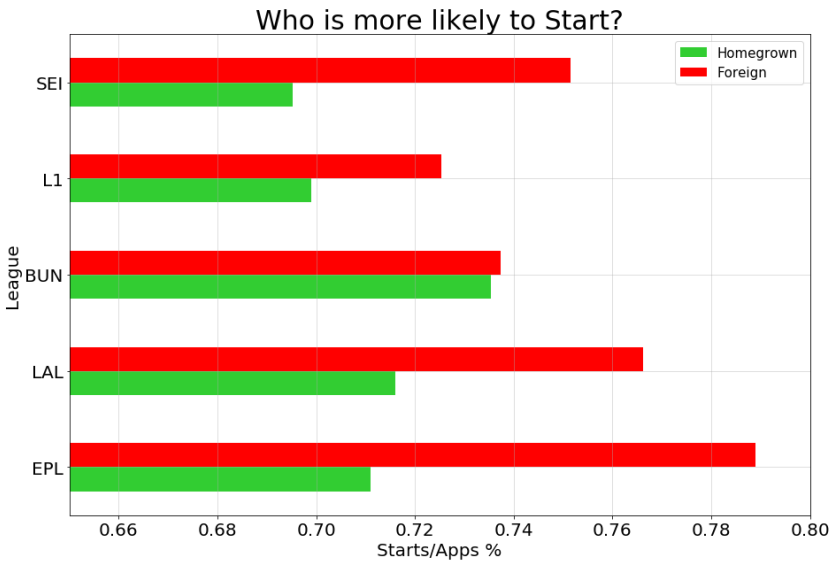



*Figure 24 . Italian Serie A Position Form & Overall Breakdown*

The same trends as the Spanish La Liga can be seen in the Italian Serie A where distances in the Form bar charts are not reflected in overall with Homegrown players. This pattern repeating in both Spanish and Italian football with Homegrown players being overrated appears to be the deciding factor in why they are overrated in the graphs in *Figure 18 and 19*.

#### 4.2.3.6 **Who's Better?**

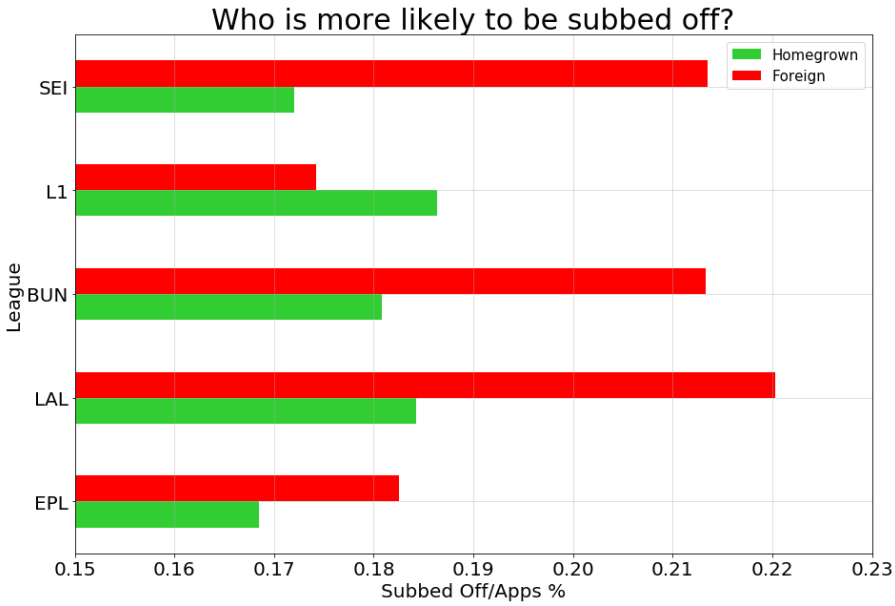
In this section of our research question, we evaluated if a Homegrown/Foreign is more inclined to be rated higher than the other in different areas in each league.




| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 0.711031 | 0.788975 |
| 1 | LAL | 0.715971 | 0.766175 |
| 2 | BUN | 0.735374 | 0.737333 |
| 3 | L1 | 0.699078 | 0.725429 |
| 4 | SEI | 0.695266 | 0.751528 |

*Figure 25 & Table 10 .*

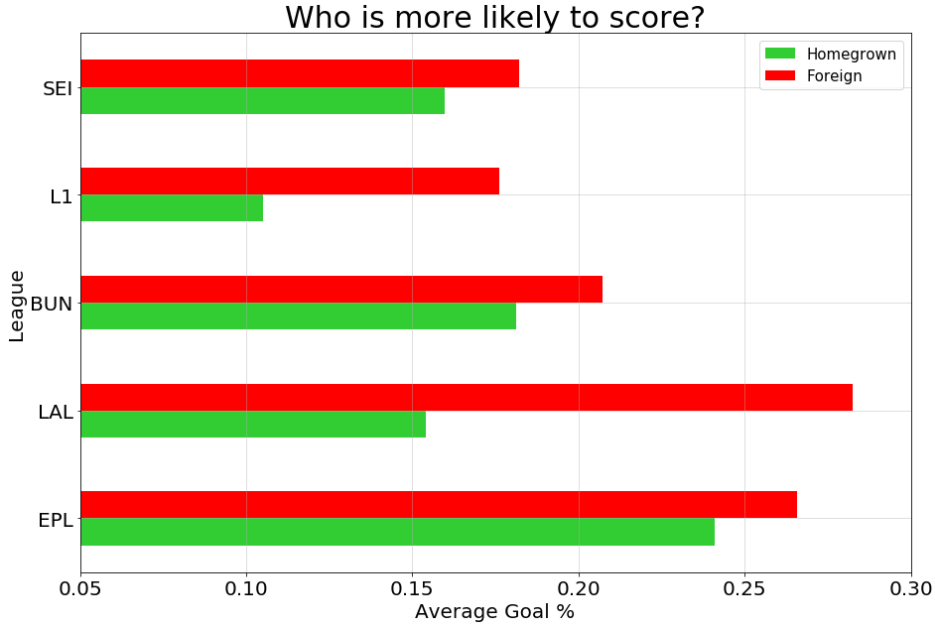
In this graph it can be said that Foreign players always start more than Homegrown players especially in the English Premier League where a Foreign player is 8% more likely to start. The German Bundesliga is very balanced in this respect. A reason for Foreign players starting more could be surmised by the fact that a certain amount of Homegrown players are unproven youth.




| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 0.1685 | 0.182508 |
| 1 | LAL | 0.184304 | 0.22032 |
| 2 | BUN | 0.180851 | 0.213323 |
| 3 | L1 | 0.186398 | 0.174236 |
| 4 | SEI | 0.172065 | 0.213468 |

*Figure 26 & Table 11 .*

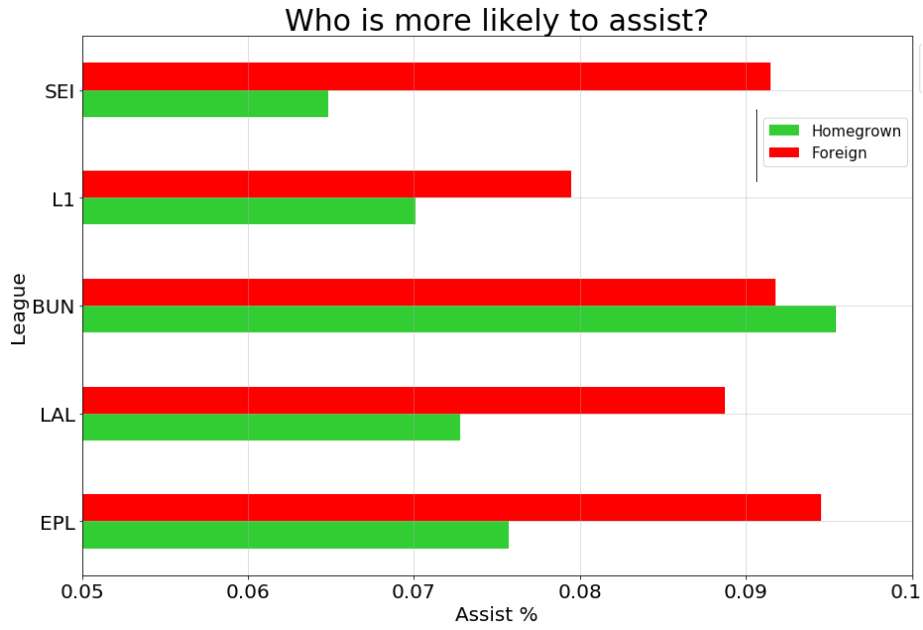
In this graph Foreign players are subbed off in all cases except in the French Ligue 1 where they are .02% more likely to be subbed off. This statistic could tell us that Foreign players do not have the same stamina as Homegrown players and have to be subbed off due to tiredness.




| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 0.240909 | 0.265736 |
| 1 | LAL | 0.154038 | 0.282512 |
| 2 | BUN | 0.181261 | 0.207093 |
| 3 | L1 | 0.105024 | 0.176152 |
| 4 | SEI | 0.159793 | 0.18205 |

*Figure 27 & Table 12 .*

For this graph we excluded all positions except forwards for more relevant results. Foreign forwards are more likely to score than a Homegrown forward in all of the leagues. In the Spanish La Liga a Foreign forward is 10% more likely to score than a Homegrown forward which is a substantial amount as teams can play up to 60 games a year meaning 6 more goals.



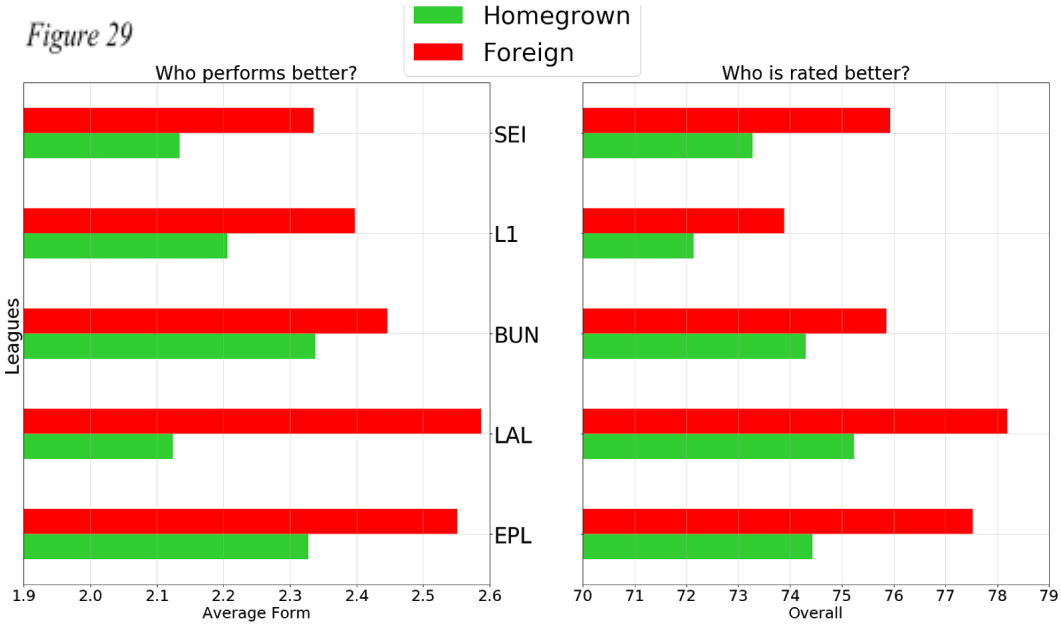

| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 0.0757235 | 0.0945436 |
| 1 | LAL | 0.072769 | 0.088729 |
| 2 | BUN | 0.0954345 | 0.0917738 |
| 3 | L1 | 0.0701031 | 0.0794721 |
| 4 | SEI | 0.0648131 | 0.091497 |

*Figure 28 & Table 13 .*

For this graph we excluded all positions except midfielders for more relevant results. Homegrown midfielders are only more likely to assist in the German Bundesliga in all other leagues they are between 1% and 4% more likely to assist.

#### 4.2.3.7 **Form vs Overall**

*Figure 29*




Form:*Table 14*

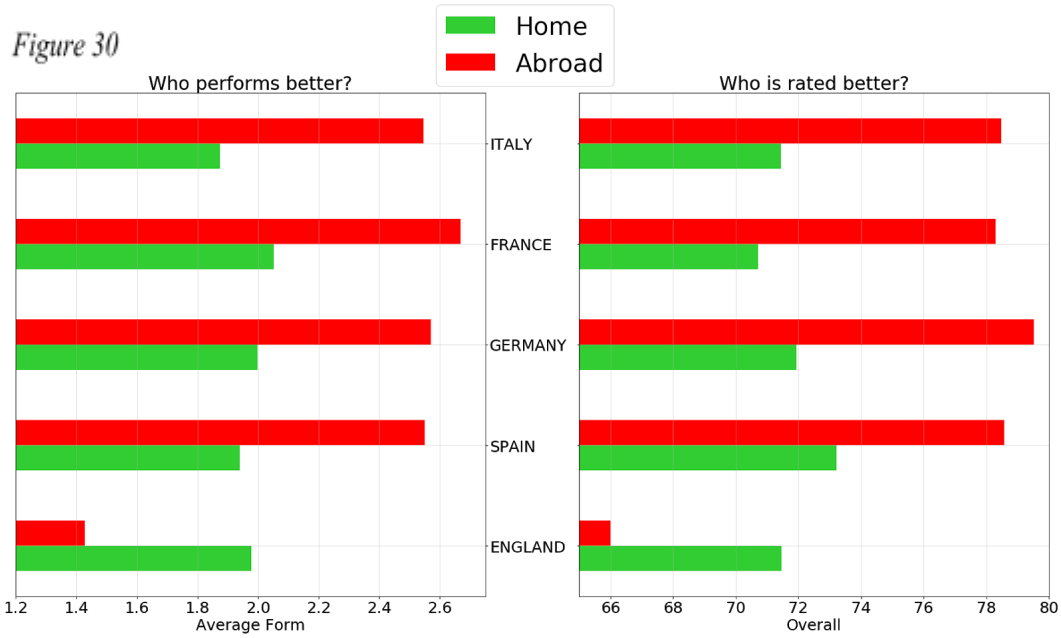
| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 2.32796 | 2.55078 |
| 1 | LAL | 2.12445 | 2.58726 |
| 2 | BUN | 2.33793 | 2.44611 |
| 3 | L1 | 2.20652 | 2.39744 |
| 4 | SEI | 2.13404 | 2.33633 |

Overall:*Table 15*

| | League | Homegrown | Foreign |
|---|---|---|---|
| 0 | EPL | 74.4333 | 77.5269 |
| 1 | LAL | 75.2395 | 78.1899 |
| 2 | BUN | 74.3069 | 75.8674 |
| 3 | L1 | 72.1373 | 73.892 |
| 4 | SEI | 73.2704 | 75.9369 |

When looking at Homegrown and Foreign players as a whole rather than positions the same general trends we saw in *Figure 21-24* can be seen. The same general trend that Foreign players perform better than Homegrown players. When looking at the length difference in the bar charts between Homegrown and Foreign players all expect the La Liga seem fairly represented where Homegrown players are rated about 3 ratings higher than indicated.

*Figure 30*




Form:*Table 16*

| | Country | Home | Abroad |
|---|---|---|---|
| 0 | ENG | 1.97711 | 1.43 |
| 1 | SPA | 1.93952 | 2.55114 |
| 2 | GER | 1.9988 | 2.56976 |
| 3 | FRA | 2.05047 | 2.66799 |
| 4 | ITA | 1.87494 | 2.54697 |

Overall:*Table 17*

| | Country | Home | Abroad |
|---|---|---|---|
| 0 | ENG | 71.4664 | 66 |
| 1 | SPA | 73.2199 | 78.575 |
| 2 | GER | 71.9288 | 79.5238 |
| 3 | FRA | 70.7026 | 78.2933 |
| 4 | ITA | 71.4371 | 78.4737 |

A trend that Abroad players perform better than Home players is evident in the case of this graph. Further research exists in our analysis questions that is not included in our final report for further information. This trend does not exist in England where there is only one Abroad player Reece Oxford who we looked at earlier. It could be said Abroad players are slightly overrated than you would expect.

#### 4.1.7. Discussion

The graphs in this research question tell us a great deal about real life football and also FIFA. There is a clear bias for Foreign players across the Top 5 leagues. It is our belief this exists across all leagues. It could be said that the best leagues will always have a higher percentage of Foreign players as they have a choice of players from all over the world if affordable. This is not reflected in parts of this question, we see that the English Premier League has a huge 65% of its players Foreign whereas in the remaining Top 5 there is a 50/50 balance in *Figure 19*. There is a distinct difference in performance of Foreign players in every league trumping their Homegrown counterparts in most cases. When seeing if these trends are reflected in EA's ratings we see some inconsistencies. Although we can see the general trend of Foreign players being rated higher there seems to be a pattern of occurrences of Homegrown players being rated higher than they should. This could be attributed to the fact EA try to balance the scales and rate Homegrown players higher across the board. Another reason for this could be because FIFA are overrating youth players from the countries of the Top 5 leagues, this is because they are expected to perform better as rising through the ranks as a Homegrown player in these countries is close to impossible.

When looking to Home and Abroad players the same general trend can be seen where Abroad players perform better expect in the case of England in *Figure 20*. These findings indicate that if there is a player playing in one of the other Top 5 leagues away from home there is a high probability they are a high performing player. When looking at England the case may be that their best players stay in England to play in the English Premier Leagues. All of these findings indicate that in a statistical sense the English Premier League is the most interesting and possibly the 'best league' due to is higher average overall and average form.

Lastly, when looking at leagues and nations being overrated in Spain and underrated in France there are many reasonable explanations for these results. The Spanish league as stated before seems to be overrated in general due to its difficulty hence why their players are given a higher rating than their form suggests in *Figure 18 and 19*. This occurrence happens again in the French league but in the opposite direction rating players lower than they perform. Another explanation for these ratings could be outliers in the data. When we speak of outliers we are talking about players who perform immensely better than their counterparts like Cristiano Ronaldo and Lionel Messi who play in the Spanish La Liga. These outliers could also be used in the opposite direction with players underperforming in average, this could explain why French players are so underrated in FIFA.

## 4.2. RQ3: Accuracy of In-Game Attributes

### 4.2.1. Datasets

For this research question we've focused on the player's in-game attributes columns. We also use real life statistics such as goals scored and clean sheets kept.

Newly created columns are listed below:
**Attacking Attributes:** Mean average of the following, Crossing, Finishing, Heading Accuracy, Short Passing & Volleys

**Defending Attributes:** Mean average of the following, Marking, Sliding Tackle, Standing Tackle

**Goalkeeping Attributes:** Mean average of the following, GK Diving, GK Reflexes, GK Handling, GK Kicking & GK Positioning

(Note: these are how the attributes are grouped on the sofifa website, where our FIFA 18 dataset originated from)

### 4.2.2. Approach

Initially, we looked at the individual attributes associated with the different areas of the game eg, Finishing Attribute vs Average Goals/Game or GK_Reflexes Attribute vs Clean Sheets. Our aim was to identify the most important attributes for players to have. However we were finding poor correlations as maybe our real-life data wasn't precise enough for this. When exploring the sofifa website (where our FIFA dataset was originally sourced) we noticed that the attributes were already arranged into groups eg. Attacking, Defending & Goalkeeping. We then decided compared these grouped attributes against player's real life abilities, in order to check if they actually reflect a players real-life ability.

### 4.2.3. Results

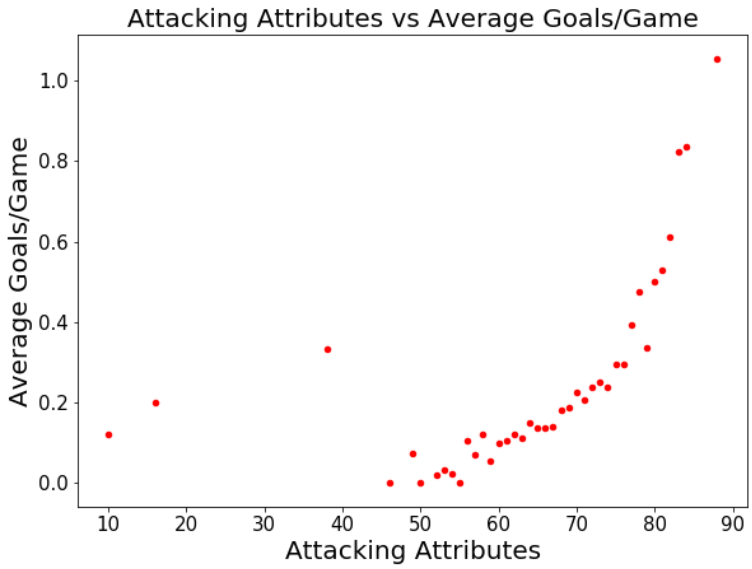



*Figure 31. A Scatterplot showing Attacking Attributes vs Average Goals/Game*

There is a moderately positive relationship between the two values (Correlation of **0.56).** According to this graph the higher a players 'attacking' attributes are within the game, the more likely they are to score goals.

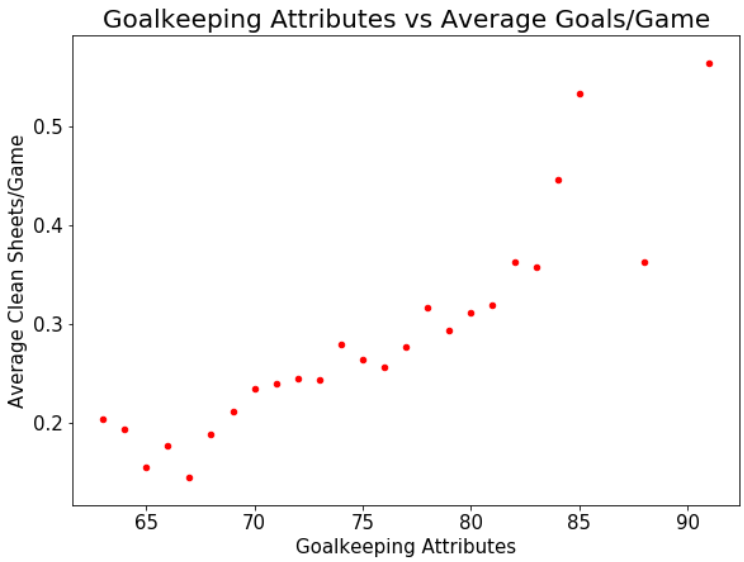



*Figure 32. A Scatterplot showing Goalkeeping Attributes vs Average Clean Sheets/Game*

There is a moderately positive relationship between the two values (Correlation of **0.60).** According to this graph the higher a players in-game 'goalkeeping' attributes are, the more likely they are to keep clean sheets.

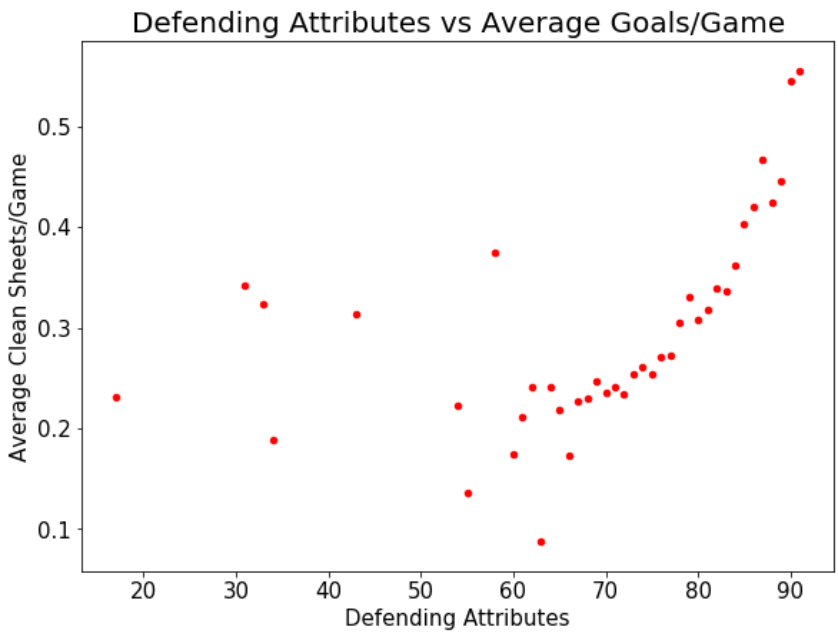



*Figure 33. A Scatterplot showing Defending Attributes vs Average Clean Sheets/Game*

There is a moderately positive relationship between the two values (Correlation of **0.37**). According to this graph the higher a players in-game 'defending' attributes are, the more likely they are to keep clean sheets.

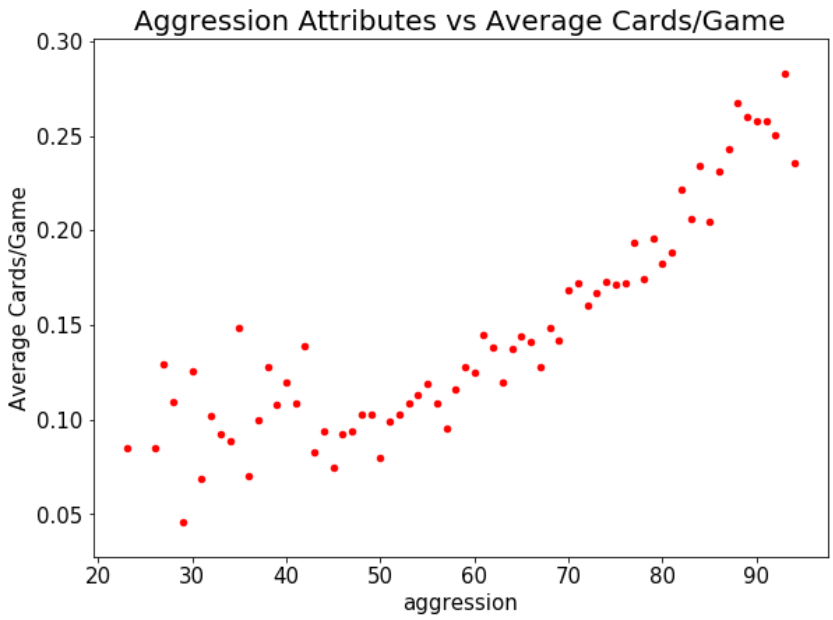
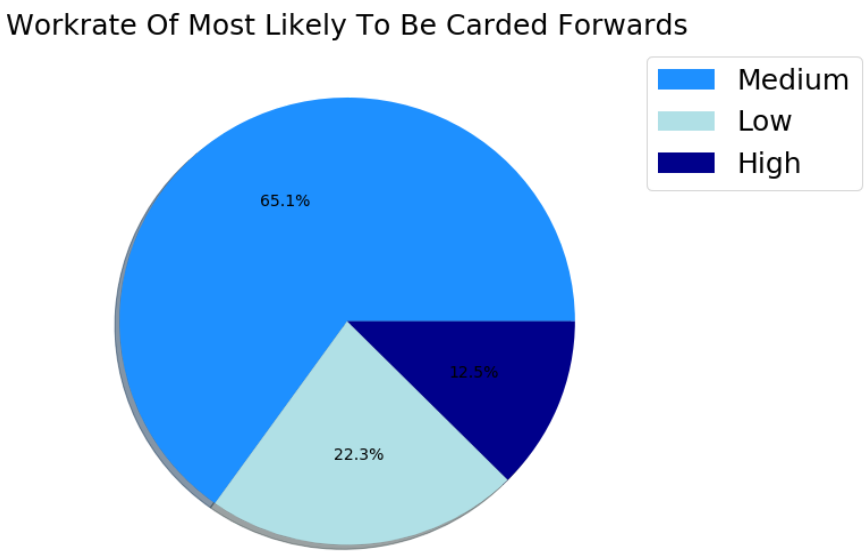



*Figure 34. A Scatterplot showing the Aggressive Attribute vs Average Cards/Game*

There is a moderately positive relationship between the two values (Correlation of **0.40).** According to this graph the higher a players in-game 'aggression' attribute is, the more likely they are to be carded in a game.



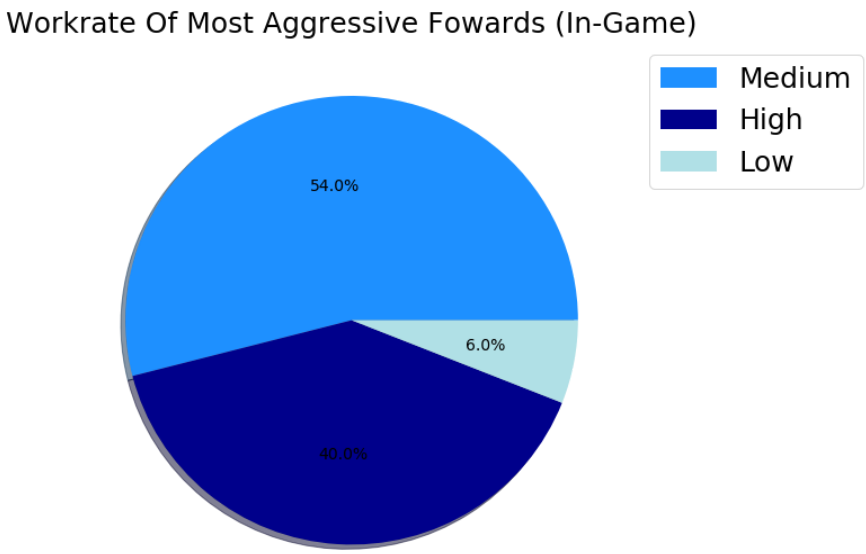




*Figure 35 & 36. A show the percentage breakdown of the defensive workrates for the most aggressive players in real-life and in-game (respectively).*

The purpose of these bar charts is to show the defensive work-rate of the most aggressive forwards both in the games and in real-life. Often forwards are criticized for having poor defensive workrates (ie, not tracking back to help out when the opposing team has possession). We used the aggressive attribute to find the most aggressive forwards within the game, and we used the Average Cards/Game to determine the most aggressive forwards in real-life.

From the pie charts we can see that the most aggressive players in real-life have a low/medium defensive workrate (87.3%). But this is not portrayed directly into the game as the most aggressive players in the game have a high/medium defensive workrate (94%).

#### 4.2.4. Discussion

These graphs show that EA have based their in-game attributes for players off of real-life statistics. Although the strength of the correlations would suggest that there is room for improvement. Perhaps if we had more precise information in our fantasy football dataset we could to examine the strength of the individual attributes in more depth. However from our research it is clear that some real-life trends are not portrayed in the game, eg. the defensive workrate of the most aggressive forwards.

The reason we picked Statbunker as the source for our fantasy football dataset was that it shared a universal scoring system for all of the leagues, this was vital for our project. However had we sourced our fantasy football data from the official sites for each league, we would have had access to more detailed information. And potentially got more from this research question.

# 5 Discussion

## 5.1. Ethical Considerations

There were very few ethical considerations to consider as football a very public topic and data is widely available for anyone to use.
Our biggest concern was that we were obtaining the datasets in an ethical fashion. The FIFA dataset was freely available on Kaggle for everybody to use. Fantasy Football dataset we retrieved from a website called Statbunker which had all of the information we wanted for each season, stored neatly within a table on their webpage. The Statbunker website has no terms and conditions on their website in relation to scrapping their data, so using pandas.read_html() we obtained the data from the website within minutes using each leagues unique page id.

When regarding ethical considerations during our research and report write up we were are not dealing with any sensitive topics hence we have no concerns about being ethically wrong. Football players play in front of thousands of fans every week (especially those who play within the top 5 leagues). The only personal information we are dealing with is their dates of birth and their nationality, which when you are a professional soccer player is made publically available. Hence there was no need for us to anonymise the names of the football players within our dataset.

## 5.2. Reproducibility

Our main reproducibility concern came with duplicate players (players with the same name) and mid-season transfers as referenced in Data Wrangling. The first method we used to correct this problem was to search the players manually using wikipedia to see if the player played for two different clubs within the same year and manually drop them if it was not true. This would not be the best practice for reproducibility as data scientists looking at our notebooks in the future may become confused to why we had dropped select players. We opted to remove any player that did not satisfy set conditions eliminating the chance of any incorrect data when joining. To resolve this we only kept rows that's FIFA 18 'full_name' column was an exact match to the corresponding 'Players' column in the fantasy football dataset if there was a possibility they were duplicated. This prevented the deletion of some players, the remaining players where removed. The resulting combined dataset contained over 3400 players so using this automatic no fail methodology was the best method for reproducibility purposes. Throughout our project we used GitHub as a version control. We applied for GitHubs student park using the University's details successfully. This allowed us to create a private shared repository worth 7 dollars per month normally. It allowed us to keep a record of file changes over time so that specific versions could be recalled later if required. It also served as a backup in case of deletion of our work. If this occurred we could access the repo and redownload the GitHub repository. Finally we have uploaded a version of our combined dataset to Kaggle, to allow other people contribute new work utilizing our dataset. We feel that our dataset is unique and is in itself the greatest achievement of this project. There were many possible paths we could have taken with our research, we hope that uploading it to Kaggle may help to exploit its potential and it's limitless applications.

## 5.3 Limitations

Prior to this project we had no experience with data science methodologies and algorithms such as linear regression aside from our 6 week course in Python. Hence we were very cautious about doing this wrong, as result our methods may come across as basic. Our fantasy dataset contained a fruitful amount of information, however it wasn't as detailed as other fantasy football websites, which included more precise information, such as total shots saved by a goalkeeper and performance based bonus points. These would have aided the findings of our research, especially for research question 3 as we could have compared individual attributes to more relevant aspects of the game and possibly see a higher correlation than what we had originally expected. We felt that time was the biggest limitation of this project as there was still so much more potential for us even after what we had done.It was difficult to know where to stop in our research and was easy to get lost in as there are so many unexplored areas we could have undertaken. Our joined dataset was very promising and it was hard choosing the route we wanted to use in our research. If more time was available it could have enabled us to get more familiar with data science methodologies such as machine learning, which would have enhanced and brought our project to the next level.

# 6. Conclusions & Future Work

This was our first experience with a full scale data science project.We were lucky enough to have some very interesting guest speakers over the course of the project.
Siobhan Grayson gave a very inspiring talk about her project 'Star Wars: A Social Network Analysis' which won her last year's 'Student Data Scientist Of The Year'. Her use of 'Gephi Networks' was something that intrigued us, we thought maybe we could apply it to our dataset to study the network of players. After a small trial using our dataset the results seemed confusing and time consuming for results that could be inconclusive hence we opted to take another path for the results it would give us. Perhaps this could be a possibility for future work others can develop with our dataset on Kaggle.
Derek Greene spoke of text analysis, we realised it was a methodology we could introduce to our project. Studying the popularity of players in real-life and how or if it can affect their in-game rating. We would obtain news articles for each league season in the countries of each league, then we could simply apply textblob for each players name (a built in python library that picks up on the sentiment of text) to determine just how popular they are (using both frequency and sentiment of text). However once again the main difficulty was time and we were unsure if these findings would actually affect a players rating the analysis would only be experimental. Unlike the 'Gephi Networks' we spoke of above we feel these findings would be very relevant to our current work if successful and could have the potential to enhance our project significantly in future work.
Brian Mcnamee talked about how our personal data can be easily obtained and used to determine what type of people we are, he touched on the recent scandal at Cambridge Analytica. The scandal involved using a person's Facebook profile to create a personality profile of that person without their consent, this information was then used to tactically place political adverts targeting those who seemed swayable. Wouldn't it be interesting if we used a similar method to generate new mentality attributes for players, based off of how they interact with fans and 'trolls' on social media. One of FIFA's biggest flaws is that its games can feel repetitive, almost as if your playing the same game over and over again. With new mentality attributes EA could introduce new features which give each game a unique feel. For example if a player scored two goals in quick succession, the opposition would crumble dependant on how strong their mentality attributes are. However this type of study could be controversial and bring with it many ethical considerations.
With the ever increasing amount of data being generated in football today there is a huge potential for the FIFA franchise to create much more accurate and realistic reflection of the sport.

# 7 Bibliography

**1/ 3** : Article title: EA explains how FIFA player ratings are calculated - VG247 Website title: VG247 URL: https://www.vg247.com/2016/09/27/how-ea-calculates-fifa-17-player-ratings/

**4** : Kaggle.com. (2018). FIFA 18 Complete Player Dataset | Kaggle. [online] Available at: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset [Accessed 11 May 2018].

**5** : Statbunker football. (2018). Football stats | Team stats | Player stats | Soccer stats. [online] Available at: https://statbunker.com [Accessed 11 May 2018].

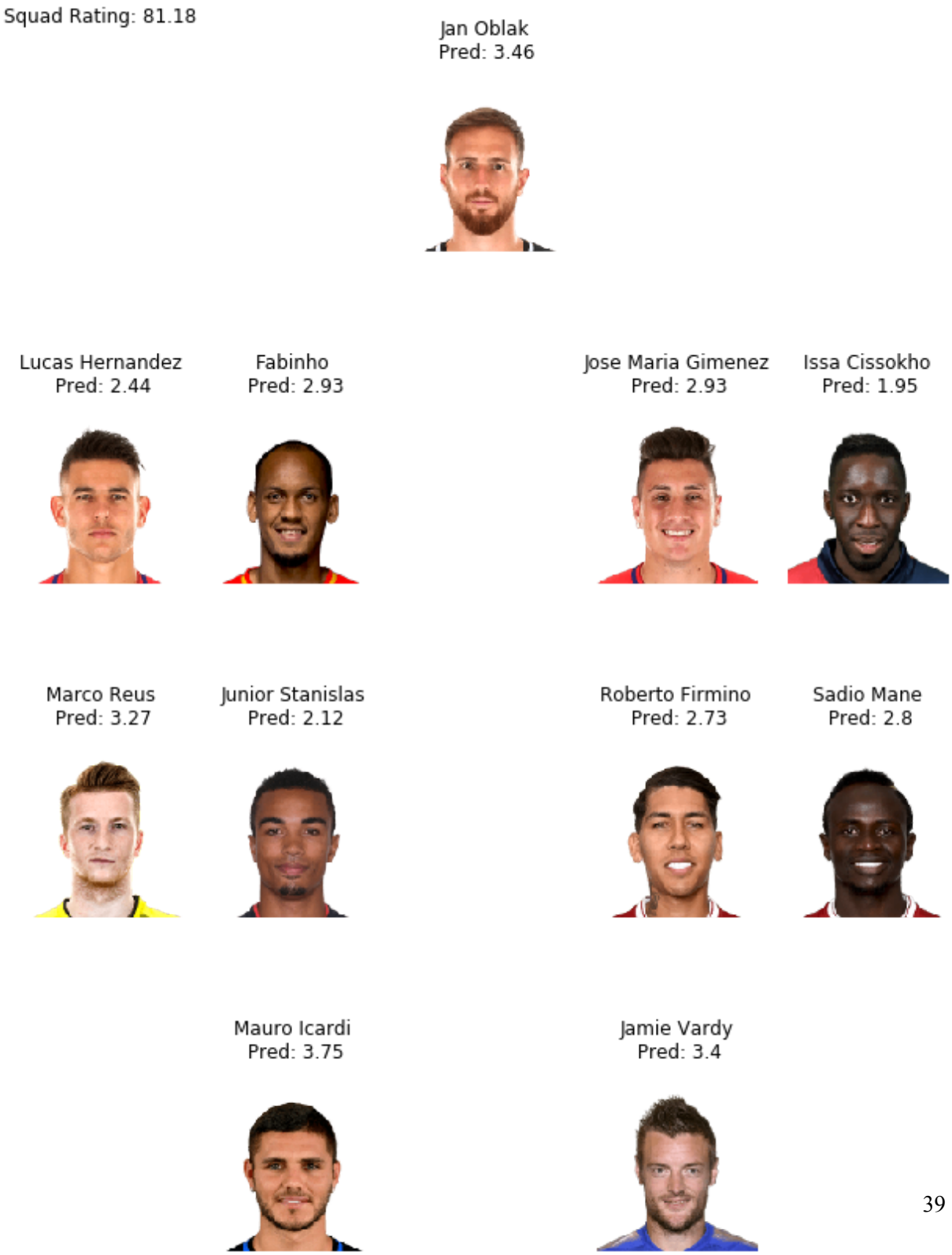
**6** : Independent.ie. (2018). Reece Oxford returns to Borussia Monchengladbach on loan - Independent.ie.[online]Available at: https://www.independent.ie/sport/soccer/reece-oxford-returns-to-borussia-monchengladbach-on-loan-36553090.html [Accessed 11 May 2018].

# 8 Appendix

## 8.1. Appendix RQ1:

We found it difficult to incorporate XI's into our project. It could possibly be an idea for later work. We thought it was an interesting idea.

Underrated XI:



Overrated XI:

Squad Rating: 73.27

Jorge Mere
Pred: 2.52

Ximo Navarro
Pred: 2.28

Raphael Varane
Pred: 3.63

Leon Balogun
Pred: 2.28

Marco Verratti
Pred: 3.49

N'Golo Kante
Pred: 3.49

Luka Modric
Pred: 3.94

Muhamed Besic
Pred: 2.3

Manu Trigueros
Pred: 3.52

Erik Moran
Pred: 2.35

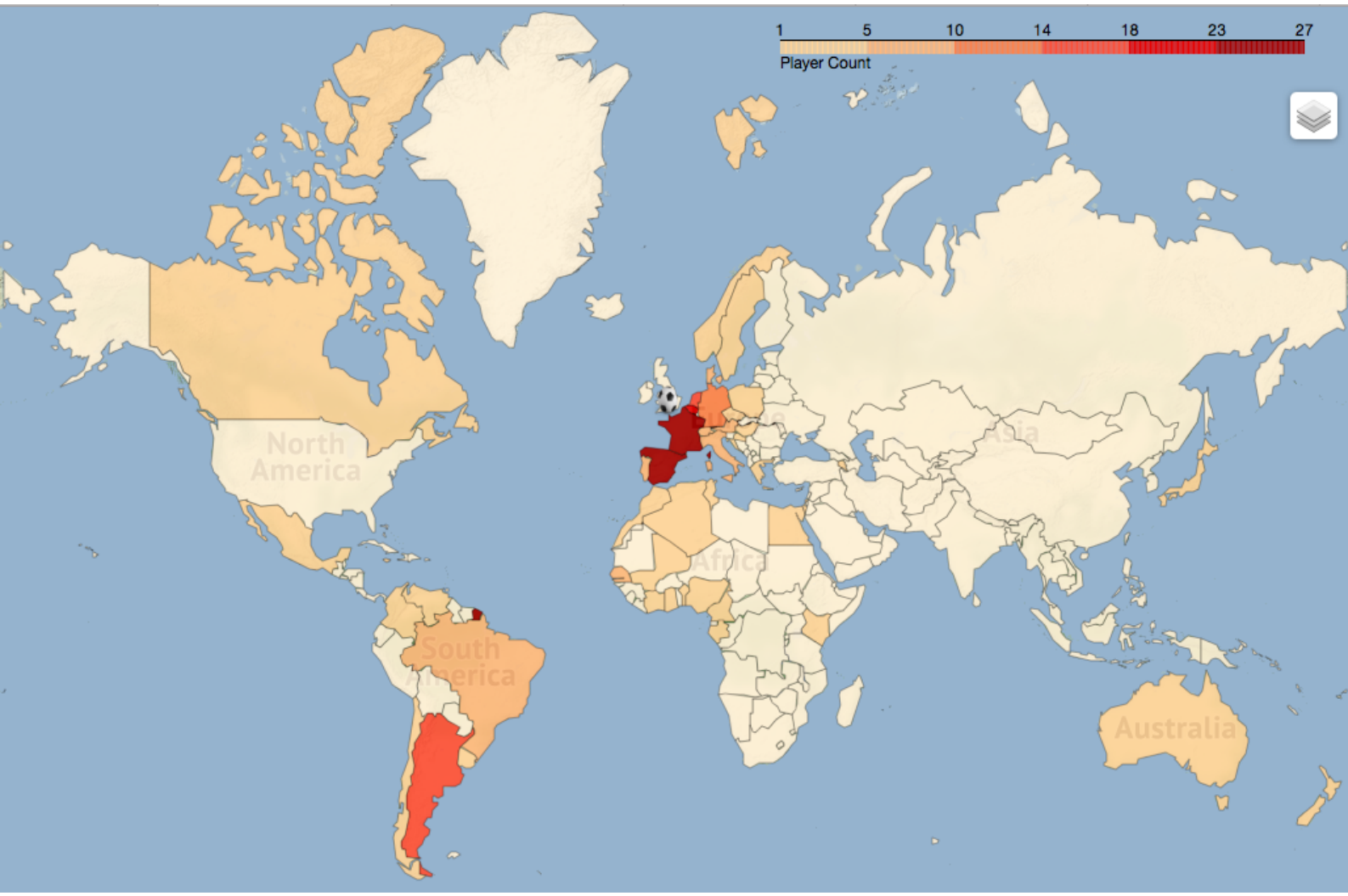
## 8.2. Appendix RQ2: Heat Maps

The following heatmaps contain all of the Top 5 leagues. For every league the Home country has been removed from the dataframe and we create a count for all other nations

**EPL Foreign Players Heatmap**

The majority of players playing in the English Premier League are European.This can been seen with France, Spain and Belgium the countries with the highest count. 8 of the top ten teams are European. The English Premier League contains little players from Asia and next to none from the US. Neighbouring countries like Ireland, Norway and the Netherlands also have a high count.

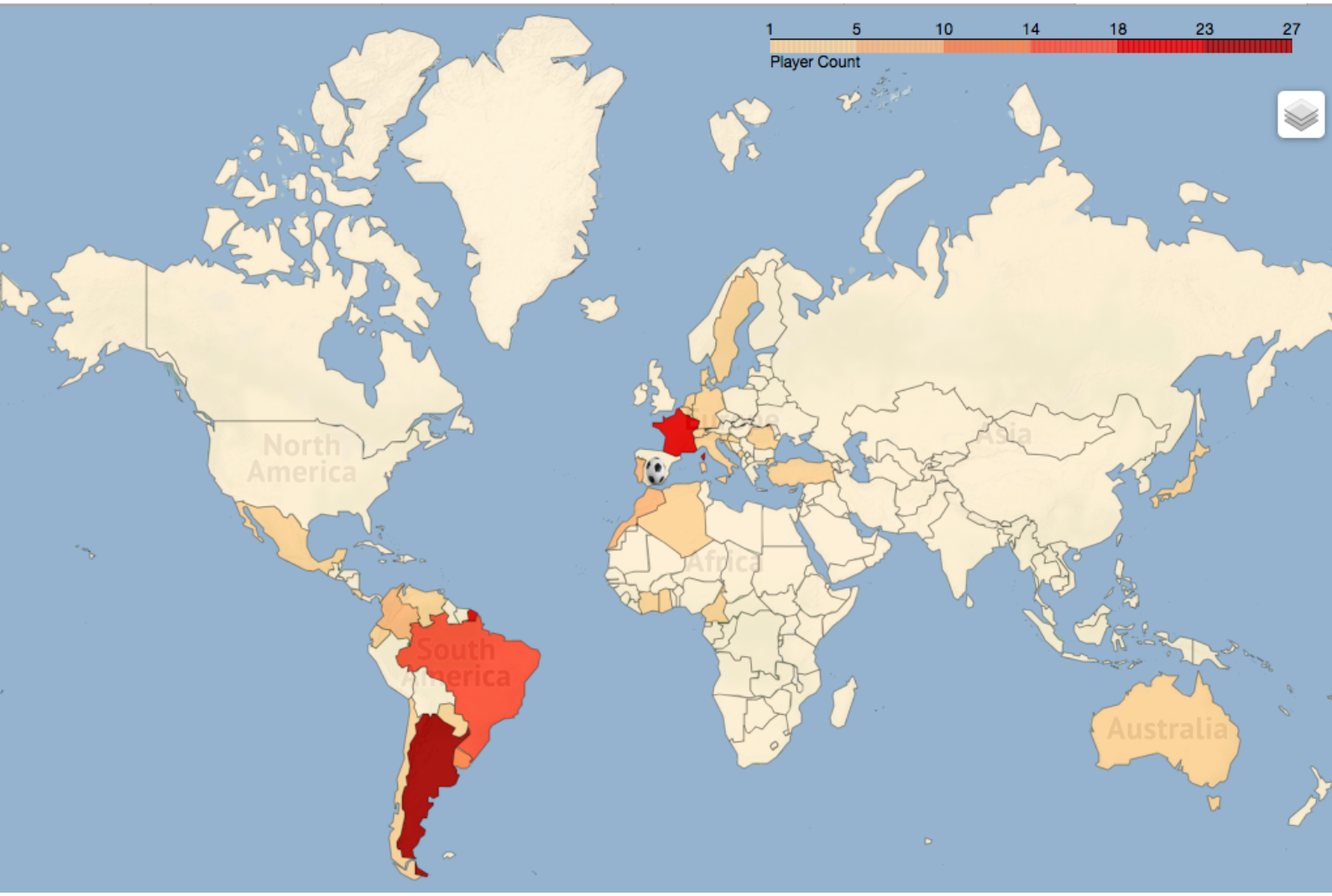
| | nationality | Players |
|---|---|---|
| 0 | France | 27 |
| 1 | Spain | 23 |
| 2 | Belgium | 19 |
| 3 | Argentina | 16 |
| 4 | Netherlands | 16 |
| 5 | Republic of Ireland | 14 |
| 6 | Germany | 10 |
| 7 | Senegal | 9 |
| 8 | Austria | 7 |
| 9 | Switzerland | 6 |



**La Liga Foreign Heatmap**

The majority of players playing in the Spanish La Liga are South American.This can been seen with Argentina, Brazil and Uruguay the countries with the highest count.Neighbouring countries like France, Portugal and Morocco also have a high count.

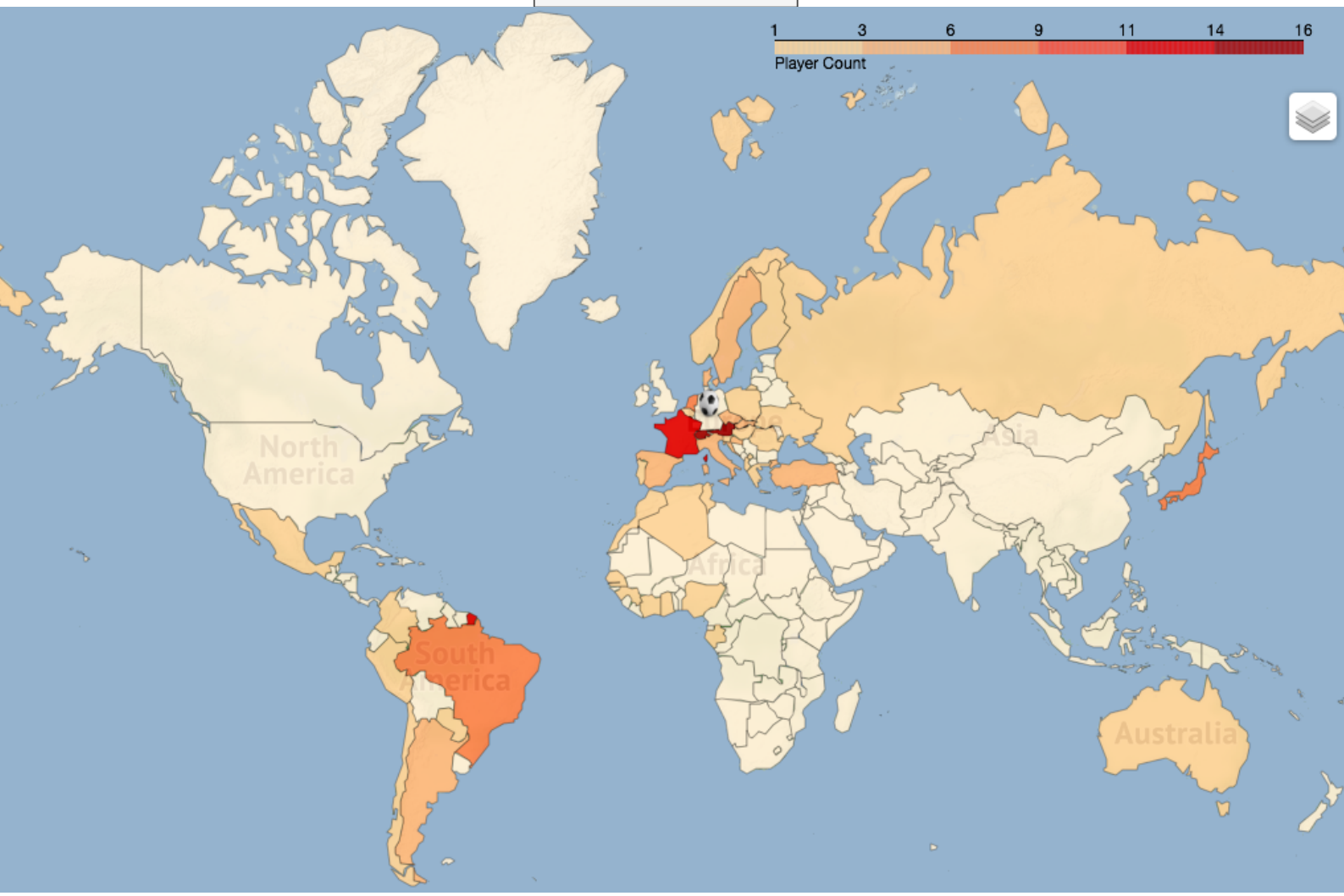
| | nationality | Players |
|---|---|---|
| 0 | Argentina | 27 |
| 1 | France | 20 |
| 2 | Brazil | 15 |
| 3 | Uruguay | 10 |
| 4 | Portugal | 9 |
| 5 | Colombia | 8 |
| 6 | Morocco | 6 |
| 7 | Serbia | 6 |
| 8 | Italy | 5 |
| 9 | Croatia | 5 |



**Bundesliga Foreign Heatmap**

The majority of players playing in the  German Bundesliga are neighbouring countries.
This can been seen with Austria, Switzerland,Serbia and France the countries with the highest count.
8 of the top ten teams are European.

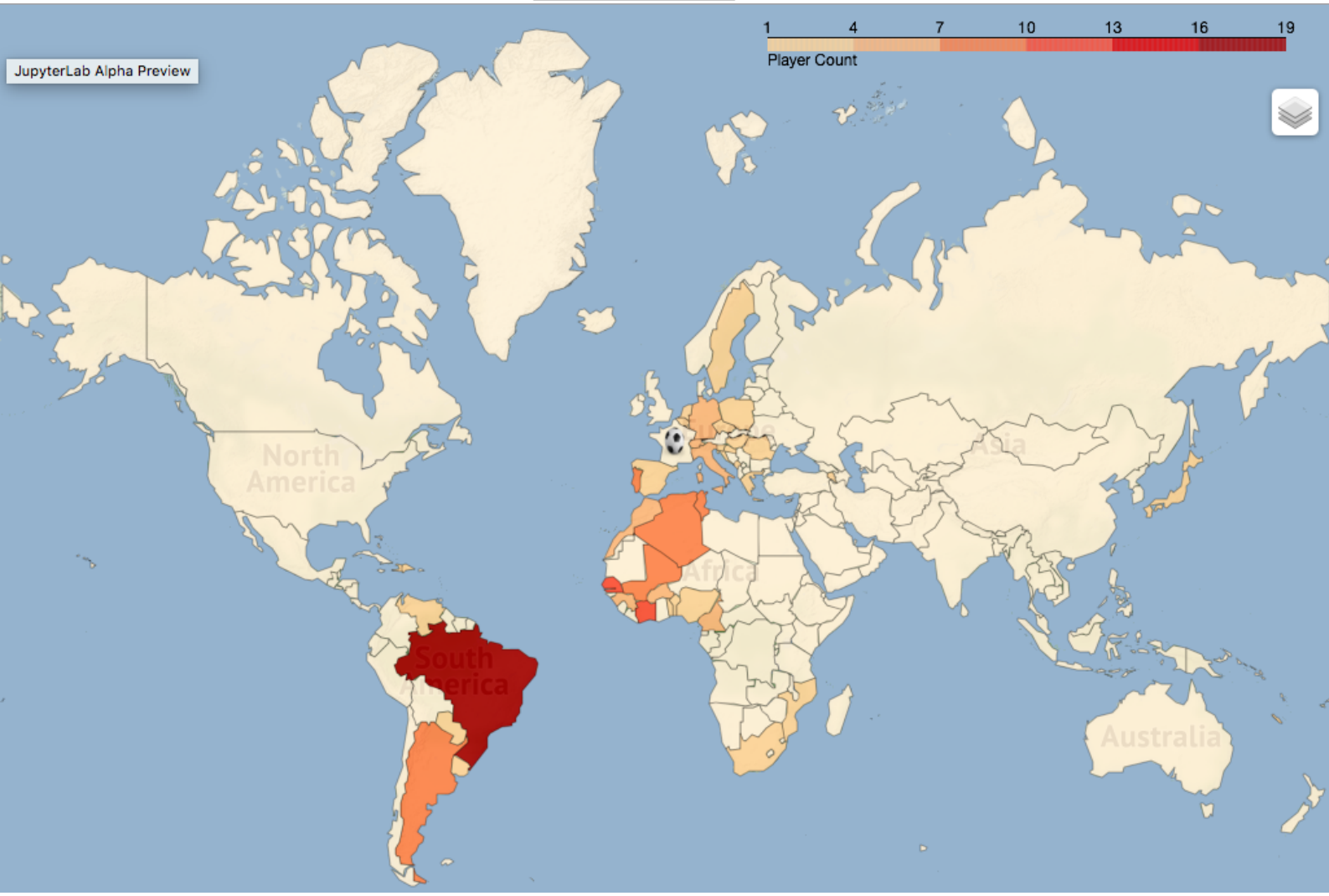
| | nationality | Players |
|---|---|---|
| 0 | Austria | 16 |
| 1 | Switzerland | 15 |
| 2 | France | 13 |
| 3 | Serbia | 9 |
| 4 | Brazil | 8 |
| 5 | Netherlands | 7 |
| 6 | Japan | 6 |
| 7 | United States | 6 |
| 8 | Denmark | 5 |
| 9 | Bosnia Herzegovina | 5 |



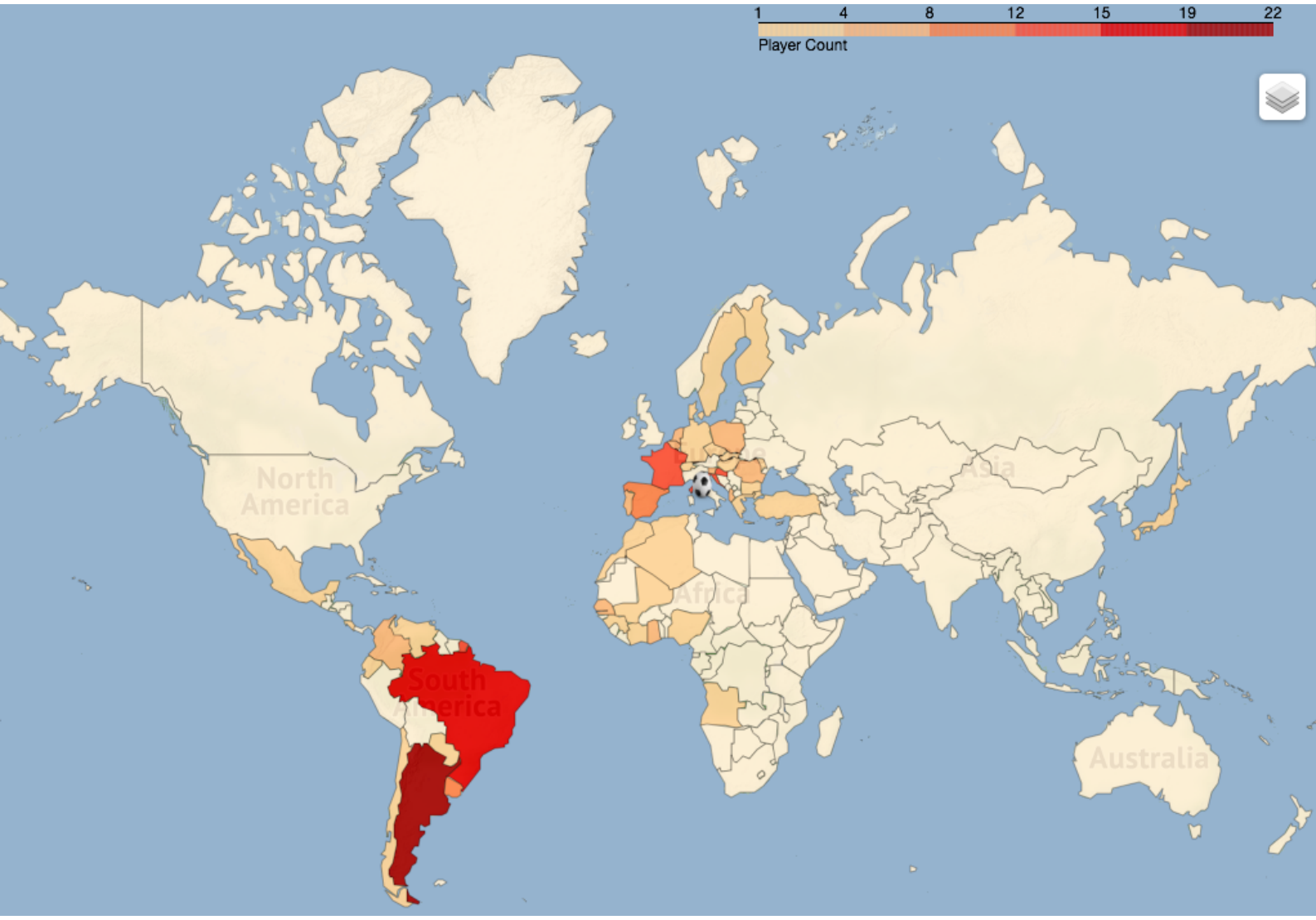

**L1 Foreign Players Heatmap**

The French Ligue 1 differs from other leagues.
The majority of players playing are African, this is the first time in the dataset Africans have a high count.Of the French Ligue 1's top 10 counts 7 are African.
They also have a high count of Brazilian players and little to no players from Asia.

| | nationality | Players |
|---|---|---|
| 0 | Brazil | 19 |
| 1 | Senegal | 13 |
| 2 | Ivory Coast | 13 |
| 3 | Argentina | 9 |
| 4 | Portugal | 8 |
| 5 | Mali | 8 |
| 6 | Algeria | 7 |
| 7 | Tunisia | 7 |
| 8 | Cameroon | 6 |
| 9 | Guinea | 5 |




**Serie A Foreign Players Heatmap**

The majority of players playing in the  Italian Serie A are South American.
This can been seen with Argentina and Brazil with the highest count.
Neighbouring countries like France, Croatia and Spain also have a high count.

| | nationality | Players |
|---|---|---|
| 0 | Argentina | 22 |
| 1 | Brazil | 17 |
| 2 | France | 15 |
| 3 | Croatia | 13 |
| 4 | Spain | 10 |
| 5 | Uruguay | 8 |
| 6 | Serbia | 8 |
| 7 | Slovenia | 7 |
| 8 | Poland | 7 |
| 9 | Ghana | 6 |