The Data

Data Description

The data was scrapped from the sofifa website using a python crawling script. The website contains the data from the EA Sports' game FIFA and gets updated regularly with the release of new versions of the game. data developed by Electronic Arts for the latest edition of their FIFA game franchise. Through several research projects done on soccer analytics, it has been established in the field of academia that the use of data from the FIFA franchise has several merits that traditional datasets based on historical data do not offer. Since 1995 the FIFA Soccer games provide an extensive and coherent scout of players worldwide.

For each attribute, we have an integer from 0 to 100 that measures how good a player is at that attribute. Examples of attributes are: dribbling, aggression, vision, marking and ball control. Observe that it seems to be unfeasible to accurately characterize players in these attributes automatically. Thus, all of those are gathered and curated by the company whose job is to bring the gameplay closer to reality as possible, hence preserving coherence and representativeness across the dataset.

The FIFA 18 dataset that has been used for this analysis provides statistics of about 16000 players on over 70 different attributes. These attributes are optimal indicators to determine the performance of a player at a particular playing position.

Data Collection

After extensively studying the structure of the website, the crawler was designed to scrape the website in two steps. The first step was to scrape the main page of the webpage which contains basic player information in a table. Apart from extracting the basic player info the crawler extracted the url of the player's detailed statistics page.

The table on the webpage had entries for 80 players on each page and the url of the page took offset for player id as a query parameter. The scrapper managed to algorithmically generate the urls of all the pages on website by manipulating the offset query in the url and iteratively visited all the pages to extract the details.

All Players 0 - 50

Toggle Columns: Basic ⌄ Attacking ⌄ Skill ⌄ Movement ⌄ Power ⌄ Mentality ⌄ Defending ⌄ Goalkeeping ⌄ Special ⌄

|						BASIC		SPECIAL	RESET
NAME		⌄ AGE	⌄ OVA	⌄ POT	⌄ TEAM& CONTRACT	⌄ VAL...	⌄ WAGE	⌄ TOT...	HITS / COMME...
🇦🇷 L. Messi CF ST RW		30	94	94	🔴 FC Barcelona 2004 ~ 2021	€118.5M	€565K	2161	0.2K / 31.4K
🇵🇹 Cristiano Ronaldo LW ST		32	94	94	⚪ Real Madrid CF 2009 ~ 2021	€95.5M	€565K	2228	0.1K / 38K
🇧🇷 Neymar LW		25	92	93	🔵 Paris Saint-Germain 2017 ~ 2022	€119.5M	€280K	2105	<100 / 20.6K
🇺🇾 L. Suárez ST		30	92	92	🔴 FC Barcelona 2014 ~ 2021	€97M	€510K	2321	<100 / 1.3K
🇩🇪 M. Neuer GK		31	92	92	🔴 FC Bayern Munich 2011 ~ 2021	€61M	€230K	1487	<100 / 2K
🇪🇸 De Gea GK		26	91	93	🔴 Manchester United 2011 ~ 2019	€74.5M	€295K	1465	<100 / 7.6K
🇧🇪 K. De Bruyne CAM CM		26	91	92	🔵 Manchester City 2015 ~ 2023	€104.5M	€395K	2234	0.1K / 7.3K
🇵🇱 R. Lewandowski ST		28	91	91	🔴 FC Bayern Munich 2014 ~ 2021	€92M	€355K	2151	<100 / 4K
🇧🇪 E. Hazard LW CF		26	91	91	🔵 Chelsea 2012 ~ 2020	€95.5M	€405K	2122	0.1K / 9.4K
🇩🇪 T. Kroos CM CDM		27	90	90	⚪ Real Madrid CF 2014 ~ 2022	€79M	€340K	2189	<100 / 3.8K
🇩🇪 M. Hummels CB		28	90	90	🔴 FC Bayern Munich 2016 ~ 2021	€62.5M	€215K	2063	<100 / 0.5K

The next step was to use the collected urls for individual players to scrape data from their details page. This page contained skill statistics, team affiliations and performance statistics for the player.
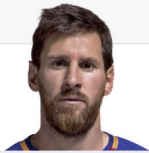
L. Messi(ID: 158023)

Lionel Messi 🇦🇷 CF ST RW Age 30 (Jun 24, 1987) 5'7" 159lbs

Overall Rating 94	Potential 94	Value €118.5M	Wage €565K

Preferred Foot Left

International Reputation 5⭐

Weak Foot 4⭐

Skill Moves 4⭐

Work Rate Medium/Medium
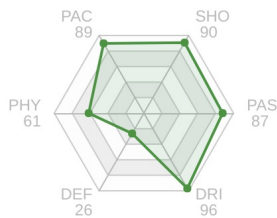
Body Type Messi

Real Face Yes

Release Clause €242.9M

#Dribbler #FK Specialist #Acrobat #Clinical Finisher

♥ Follow(758)   👍 Like(632)   👎 Dislike(197)

History Versio ⌄

Radar
PAC 89
SHO 90
PAS 87
DRI 96
DEF 26
PHY 61

🔵 FC Barcelona

85 ⭐⭐⭐⭐⭐

Position RS

Jersey Number 10

Joined Jul 1, 2004

Contract Valid Until 2021

🔵 Argentina

82 ⭐⭐⭐⭐⯪

Position RW

Jersey Number 10

Rankings

	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Mon												
Wed												
Fri												
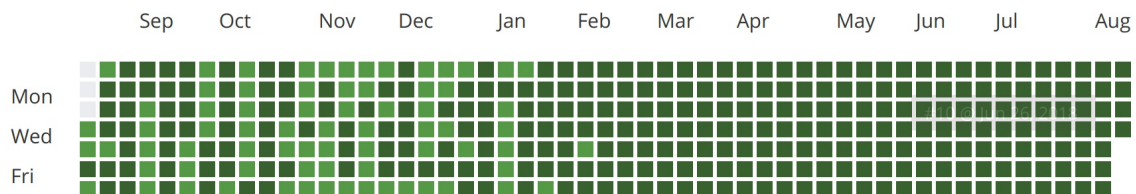
Attacking
77	Crossing
95	Finishing
71	Heading Accuracy
88	Short Passing
86	Volleys

Skill
97	Dribbling
90	Curve
92	FK Accuracy
87	Long Passing
96	Ball Control

Movement
92	Acceleration
87	Sprint Speed
90	Agility
95	Reactions
95	Balance

Power
85	Shot Power
67	Jumping
73	Stamina
59	Strength
88	Long Shots

Content

The scrapped dataset has the following properties:

- Every player featuring in FIFA 18
- 70+ attributes
- Player and Flag Images
- Playing Position Data
- Attributes based on actual data of the latest EA's FIFA 18 game
- Attributes include on all player style statistics like Dribbling, Aggression, GK Skills etc.

- Player personal data like Nationality, Photo, Club, Age, Wage, Salary etc.

```
data.columns
Index(['ID', 'Photo', 'Name', 'Age', 'Nationality', 'Flag', 'Overall',
       'Potential', 'Club', 'Club Logo', 'Value', 'Wage', 'Special',
       'Acceleration', 'Aggression', 'Agility', 'Balance', 'Ball control',
       'Composure', 'Crossing', 'Curve', 'Dribbling', 'Finishing',
       'Free kick accuracy', 'GK diving', 'GK handling', 'GK kicking',
       'GK positioning', 'GK reflexes', 'Heading accuracy', 'Interceptions',
       'Jumping', 'Long passing', 'Long shots', 'Marking', 'Penalties',
       'Positioning', 'Reactions', 'Short passing', 'Shot power',
       'Sliding tackle', 'Sprint speed', 'Stamina', 'Standing tackle',
       'Strength', 'Vision', 'Volleys', 'CAM', 'CB', 'CDM', 'CF', 'CM', 'LAM',
       'LB', 'LCB', 'LCM', 'LDM', 'LF', 'LM', 'LS', 'LW', 'LWB',
       'Preferred Positions', 'RAM', 'RB', 'RCB', 'RCM', 'RDM', 'RF', 'RM',
       'RS', 'RW', 'RWB', 'ST'],
      dtype='object')
```

Data Cleaning and Manipulation

As the data was scrapped from a website it had several inconsistencies and properties that made it unfit for appropriate exploratory analysis. In order to prepare the data, the following steps were performed:

- The Wage and Value of all the players were in form of strings with the symbol of the currency in front of the values - `€190K`. These had to converted to numeric values.
- The preferred positions of all the players was available as a space separated string (`ST LW`) which was converted to a list for easier retrieval.
- A new attribute `FieldPosition` was created by mapping the preferred positions of all the players with their respective roles in the team, like attack, midfielder, defence or goalkeeper.
- Using the nationality of the players, a new attribute was created that contained the continent the player belonged to.

Exploratory Data Analysis

After preparing the dataset for analysis, explorations were made on a macro level like continent and player nationality level analysis as well as on player level. Through the process attempts were made to derive interesting correlations and trends by the use of visualizations.

# Geographic distribution of players

---Insert text here----
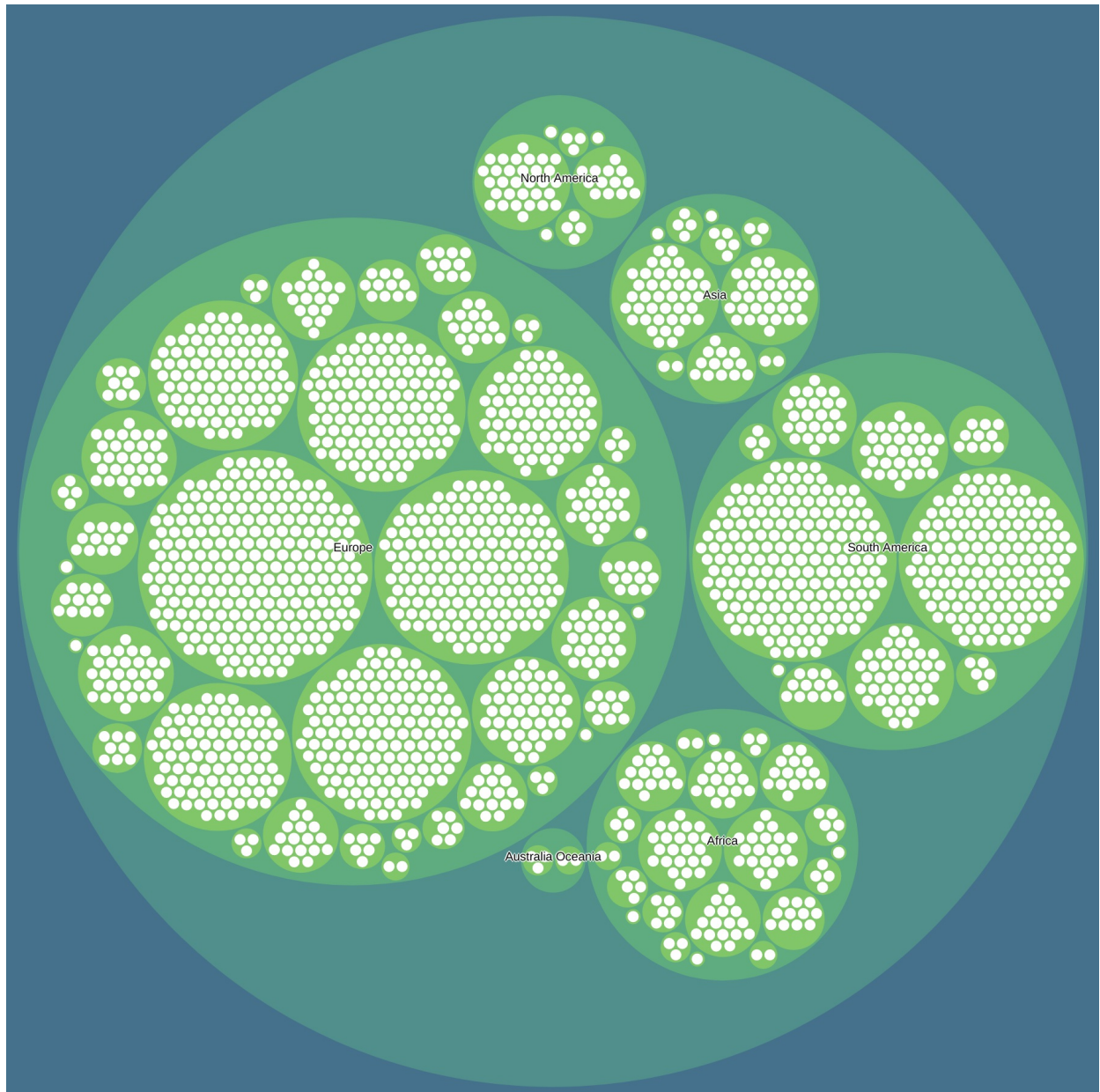
---Insert text here---

Nationality wise analysis

This section explores aggregate statistics of players hailing from all major football playing nations. Country wise mean, maximum and minimum values were identified for all the numeric attributes and plotted on an interactive world map.

(Feature - Stat):

Overall - mean ▼
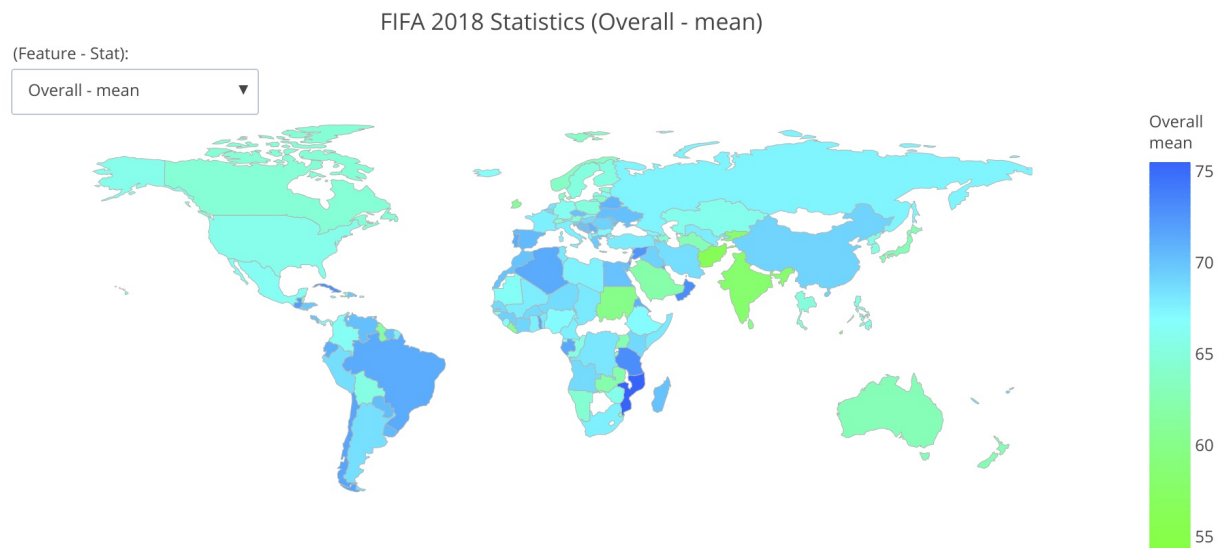
The above plot demonstrates that South American, African and European players are generally rated higher than their counterparts from Asia, Australia or North America. While the nationwide average rating for USA stands at 65.81 and 58.06 for India it is about 71.24 for Brazil and 69.11 for China.

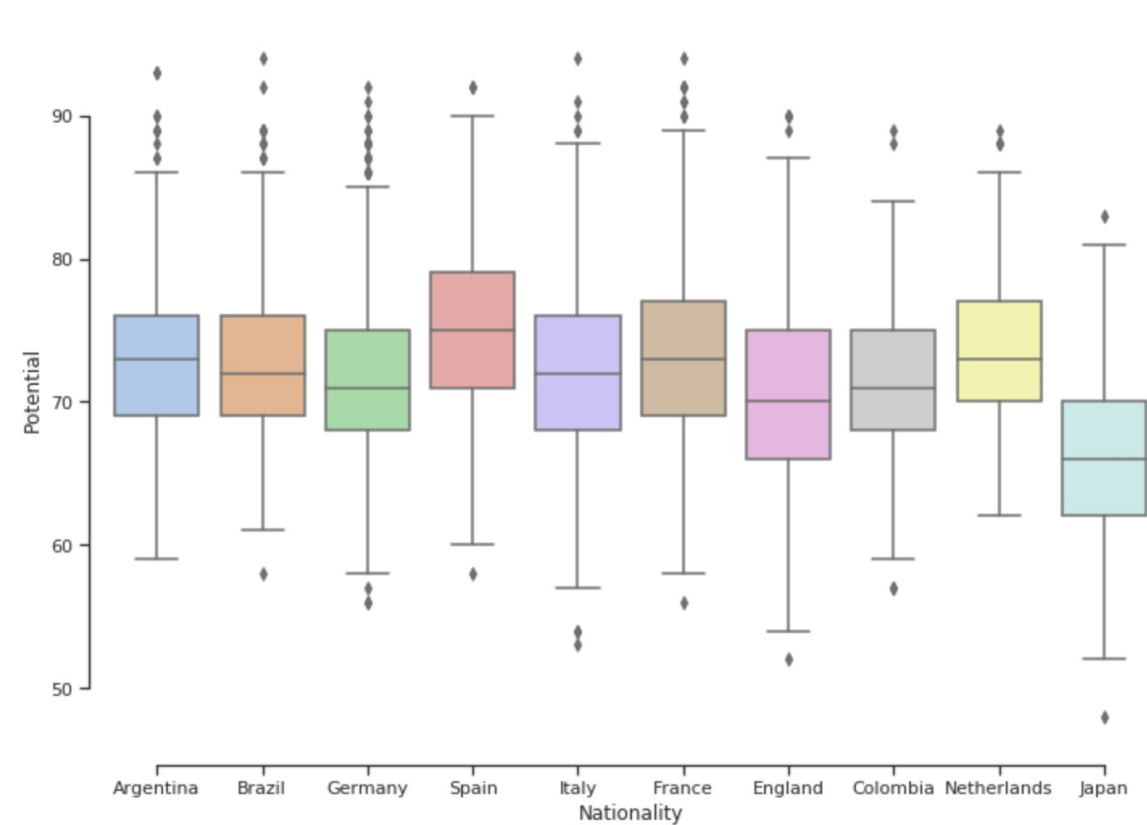## Countries with the best aggregate player ratings
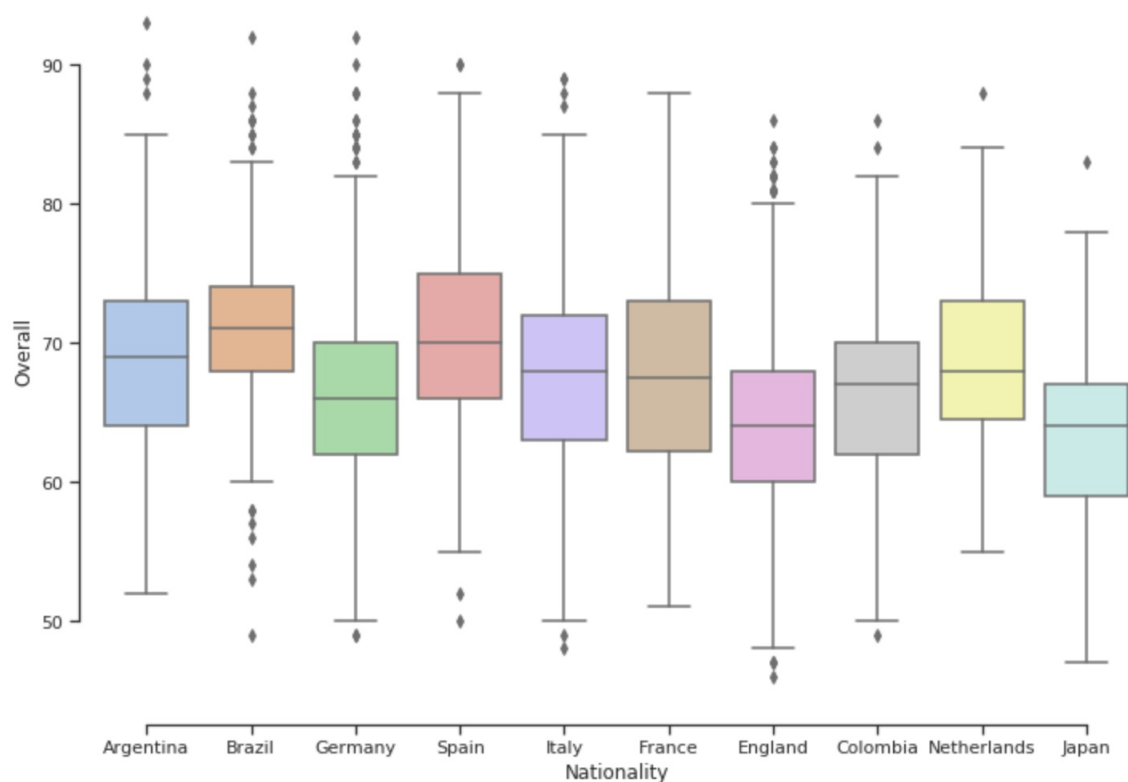
---Insert text here----

---Insert text here---
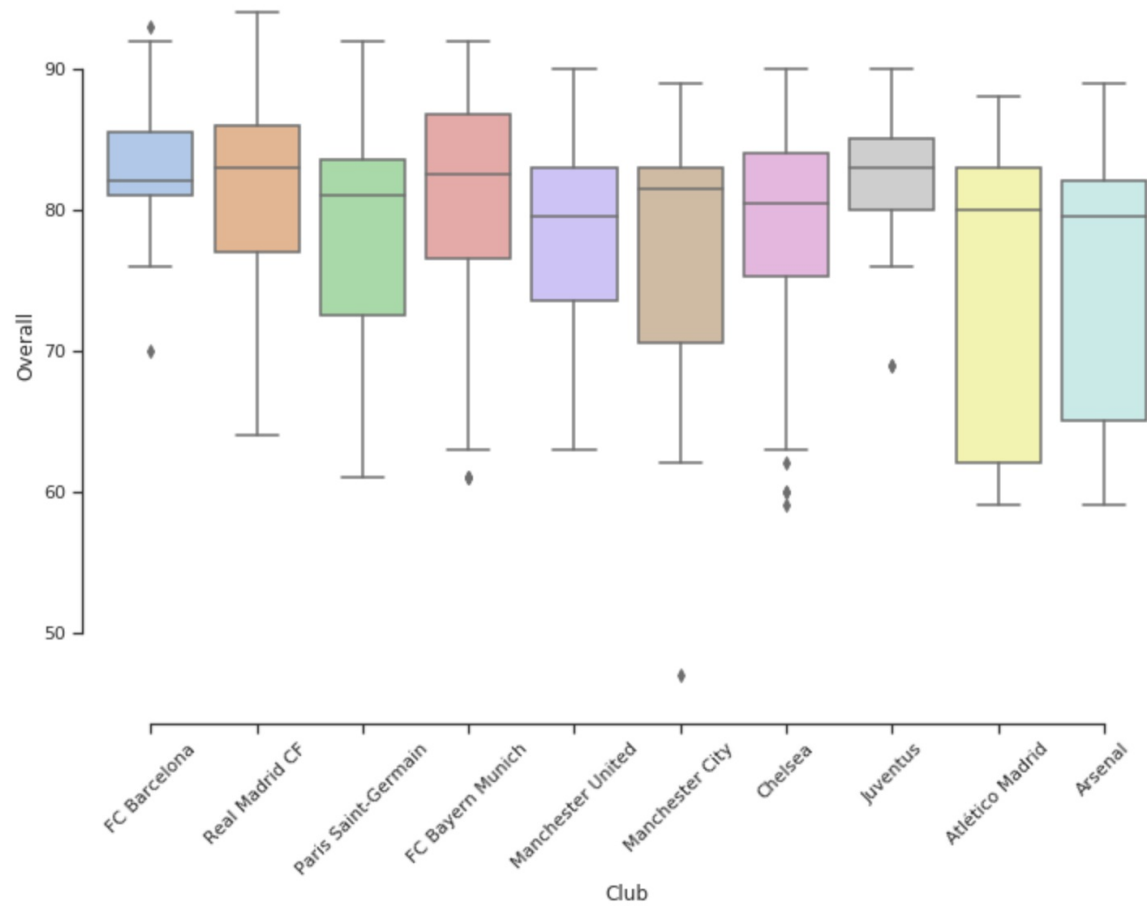
Countries with the best aggregate player potential

---Insert text here----

---Insert text here---

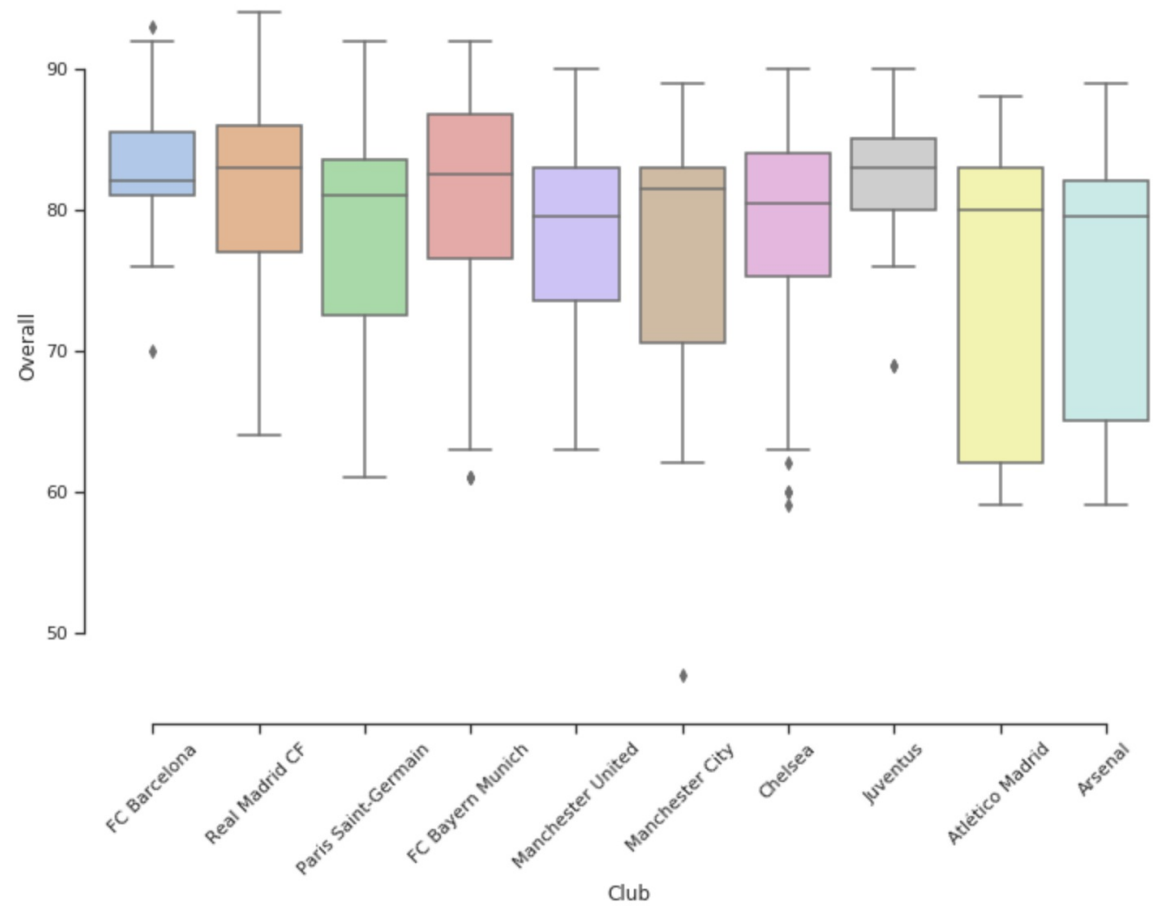# Clubs with the best aggregate player ratings

---Insert text here----

---Insert text here---

Clubs with the best aggregate player potentail

---Insert text here----

---Insert text here---