EDA

BY ABHISHEK THOMAS

INTRODUCTION

- For my assignment I have chosen to talk about the greatest sport that has been played by mankind both offline and online. But for now I would like to talk about the online version of the sport 'FIFA 19' which is played by millions all over the world. I have been part of my college football team and I wanted you to know to some really interesting facts about the sport through a game based on it.
- I chose the game as it was the easiest way to obtain a clean and structed dataset of the all the players across the whole world. FIFA 19's game model is based on the real sport which has described the players attributes in the most EXACT way possible and I could not stop myself from analysing the hidden secrets of the sport using DATA SCIENCE!
- Without further intro lets EDA!

DATA

- The data was scrapped from sofifa website using a python crawling script. The website has data from the EA Sports' game FIFA and gets updated regularly with latest release. 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.
- Each attribute has a integer from 0 to 100 to measure how good a player is at that attribute. Examples of attributes are: dribbling, aggression, vision, marking and ball control. To make the game as realistic as possible the attributes describe each player in the most detailed and accurate manner .
- The FIFA 19 dataset that has been used for this analysis provides statistics of about 19000 players on over 70 different attributes. These attributes are optimal indicators to determine the performance of a player at a particular playing position.

DATA MANUPILATION

To perform analysis, I had to refine and tailor the data according to my needs, to do this

- I dropped the unnecessary columns that would not be of any use to me.
- All wage and value of the player are in ₤ , wage is thousands and the value of the player is in millions
- Final adjustments like clearing the missing values and duplicates using openrefine.

There are a total of **17,725 players** in the dataset after cleaning the data in jupyter notebook and openrefine.

In [1]:

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```python
fifa=pd.read_csv('fifa.csv')
```

In [3]:

```python
pd.set_option('display.max_columns', None)
```

In [4]:

```python
fifa.head()
```

t[4]:

	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club	Club Logo	Value	Wage	Preferred Foot	International Reputation
0	158023	L. Messi	31	https://cdn.sofifa.org /players /4/19/158023.png	Argentina	https://cdn.sofifa.org /flags/52.png	94	94	FC Barcelona	https://cdn.sofifa.org /teams/2/light /241.png	€110.5M	€565K	Left	5.0
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org /players /4/19/20801.png	Portugal	https://cdn.sofifa.org /flags/38.png	94	94	Juventus	https://cdn.sofifa.org /teams/2/light /45.png	€77M	€405K	Right	5.0
2	190871	Neymar Jr	26	https://cdn.sofifa.org /players /4/19/190871.png	Brazil	https://cdn.sofifa.org /flags/54.png	92	93	Paris Saint-Germain	https://cdn.sofifa.org /teams/2/light /73.png	€118.5M	€290K	Right	5.0
3	193080	De Gea	27	https://cdn.sofifa.org /players /4/19/193080.png	Spain	https://cdn.sofifa.org /flags/45.png	91	93	Manchester United	https://cdn.sofifa.org /teams/2/light /11.png	€72M	€260K	Right	4.0
		K. De		https://cdn.sofifa.org		https://cdn.sofifa.org			Manchester	https://cdn.sofifa.org				

In [9]:
```
def wage_split(x):
    try:
        return int(x.split("K")[0][1:])
    except:
        return 0
fifa_complete['Wage'] = fifa_complete['Wage'].apply(lambda x : wage_split(x))
def value_split(x):
    try:
        if 'M' in x:
            return float(x.split("M")[0][1:])
        elif 'K' in x:
            return float(x.split("K")[0][1:])/1000
    except:
        return 0
fifa_complete['Value'] = fifa_complete['Value'].apply(lambda x : value_split(x))
```

In [17]:
```
fifa_complete.loc[:,['Value','Wage']].head(5)
```

Out[17]:

	Value	Wage
0	95.5	565
1	105.0	565
2	123.0	280
3	97.0	510
4	61.0	230

```
In [36]: fifa_complete = fifa_complete.drop(['RWB','ST'],axis = 1)
         fifa_complete.head(5)
```

Out[36]:

	Name	Age	Nationality	Overall	Potential	Club	Value	Wage	Acceleration	Aggression	...	Reactions	Short passing
0	Cristiano Ronaldo	32	Portugal	94	94	Real Madrid CF	95.5	565	89	63	...	96	83
1	L. Messi	30	Argentina	93	93	FC Barcelona	105	565	92	48	...	95	88
2	Neymar	25	Brazil	92	94	Paris Saint-Germain	123	280	94	56	...	88	81
3	L. Suárez	30	Uruguay	92	92	FC Barcelona	97	510	88	78	...	93	83
4	M. Neuer	31	Germany	92	92	FC Bayern Munich	61	230	58	29	...	85	55

5 rows × 42 columns

QUICK EDA

PLAYER COUNT BY COUNTRY

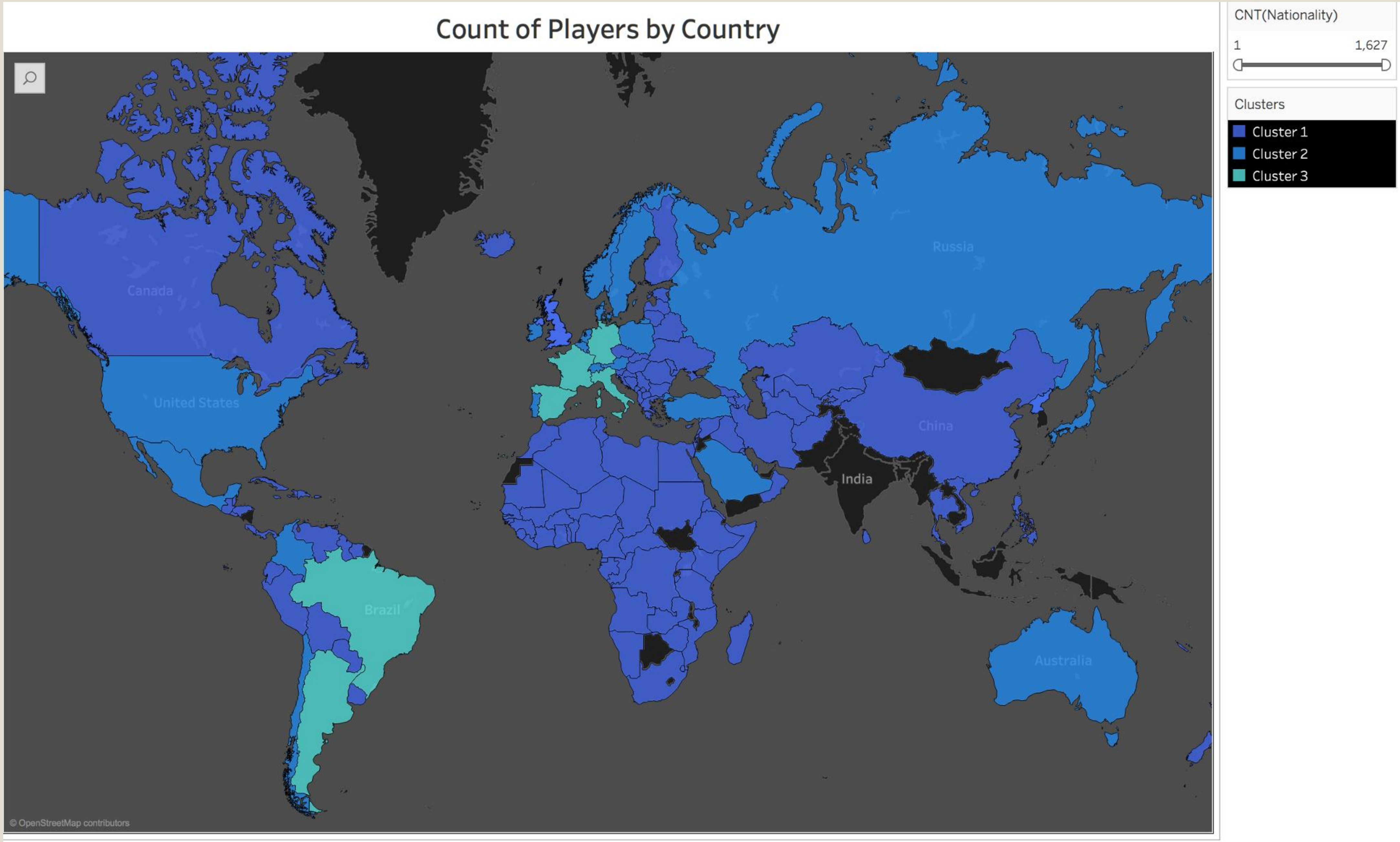
- In Tableau, I divided the players data into 3 clusters, with cluster 3 containing countries with the most players, and cluster 1 containing countries with the lowest count of players.

TOP 15 AVG OVERALL RATING BY COUNTRY

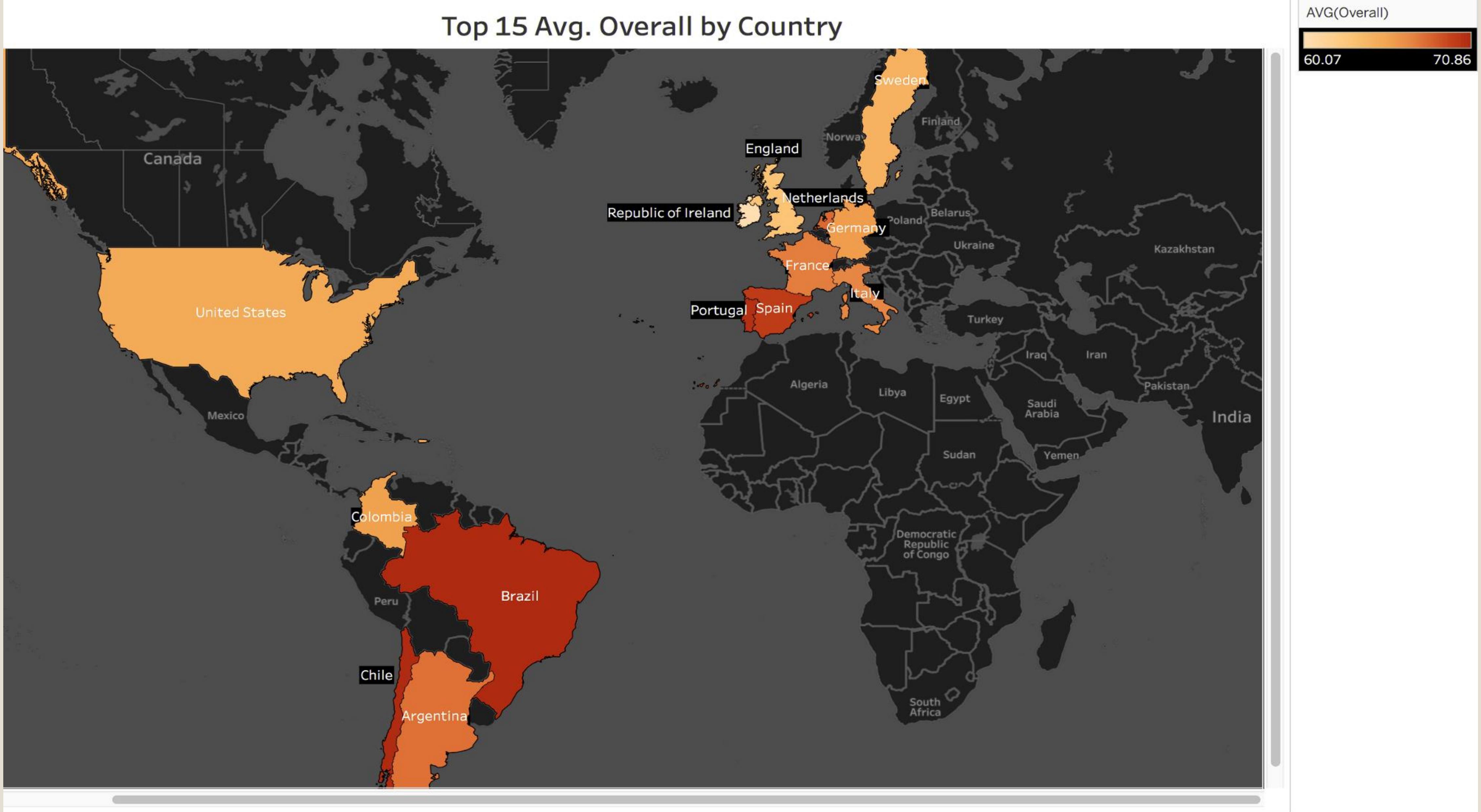
- In Tableau, I took the players with the top 15 overall rating and categorized them the country they were from

From the visualizations we can see that **Western Europe and South America** are the football powerhouse regions of the world.

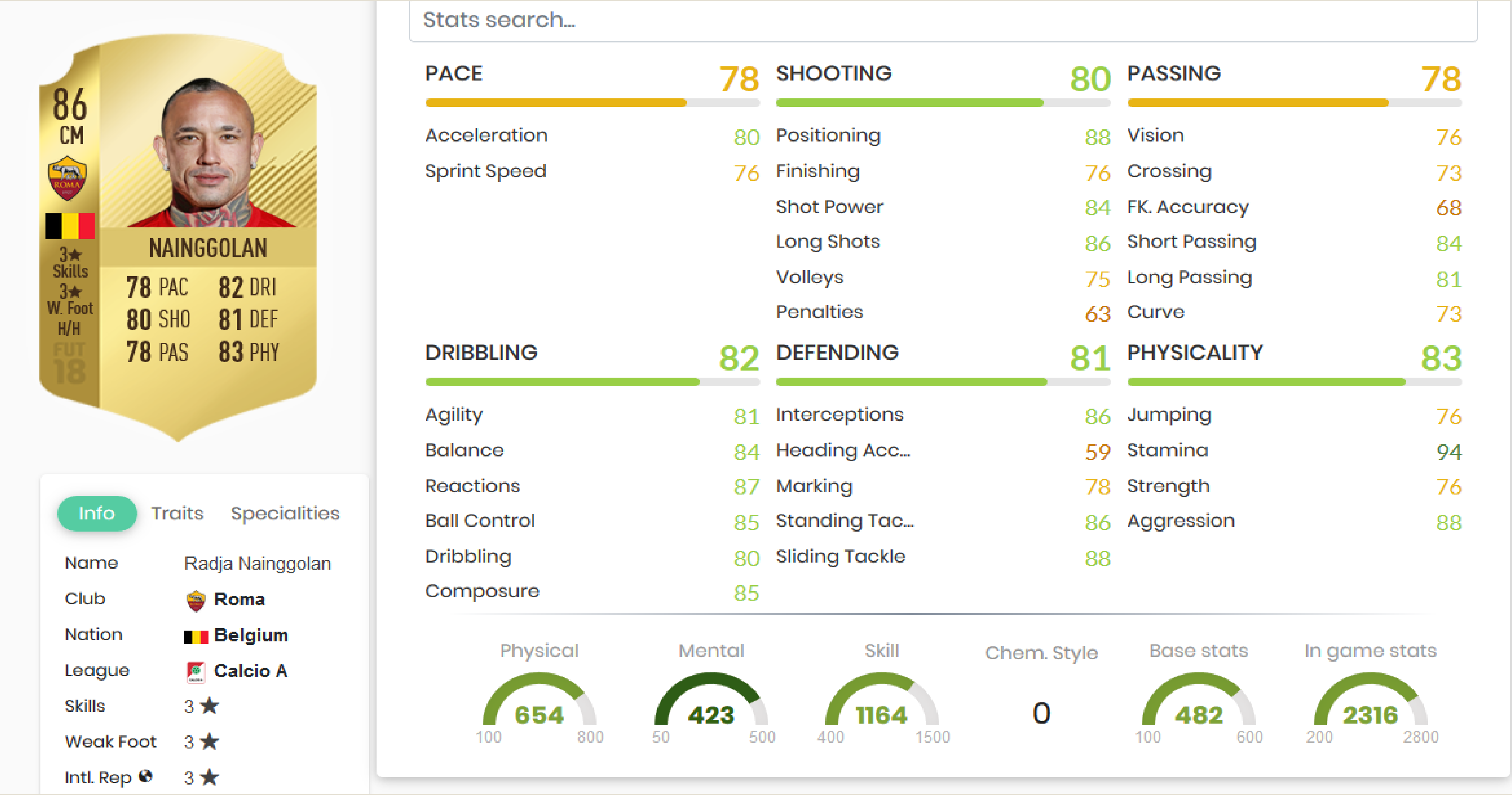
Count of Players by Country



Top 15 Avg. Overall by Country



PLAYER STATISCS



THE BEST DEFENDER

Every player has 6 core attributes which are pace, shooting, dribbling , defence, physical , passing. These are again subdivided into sub attributes which contribute to the arrtibute

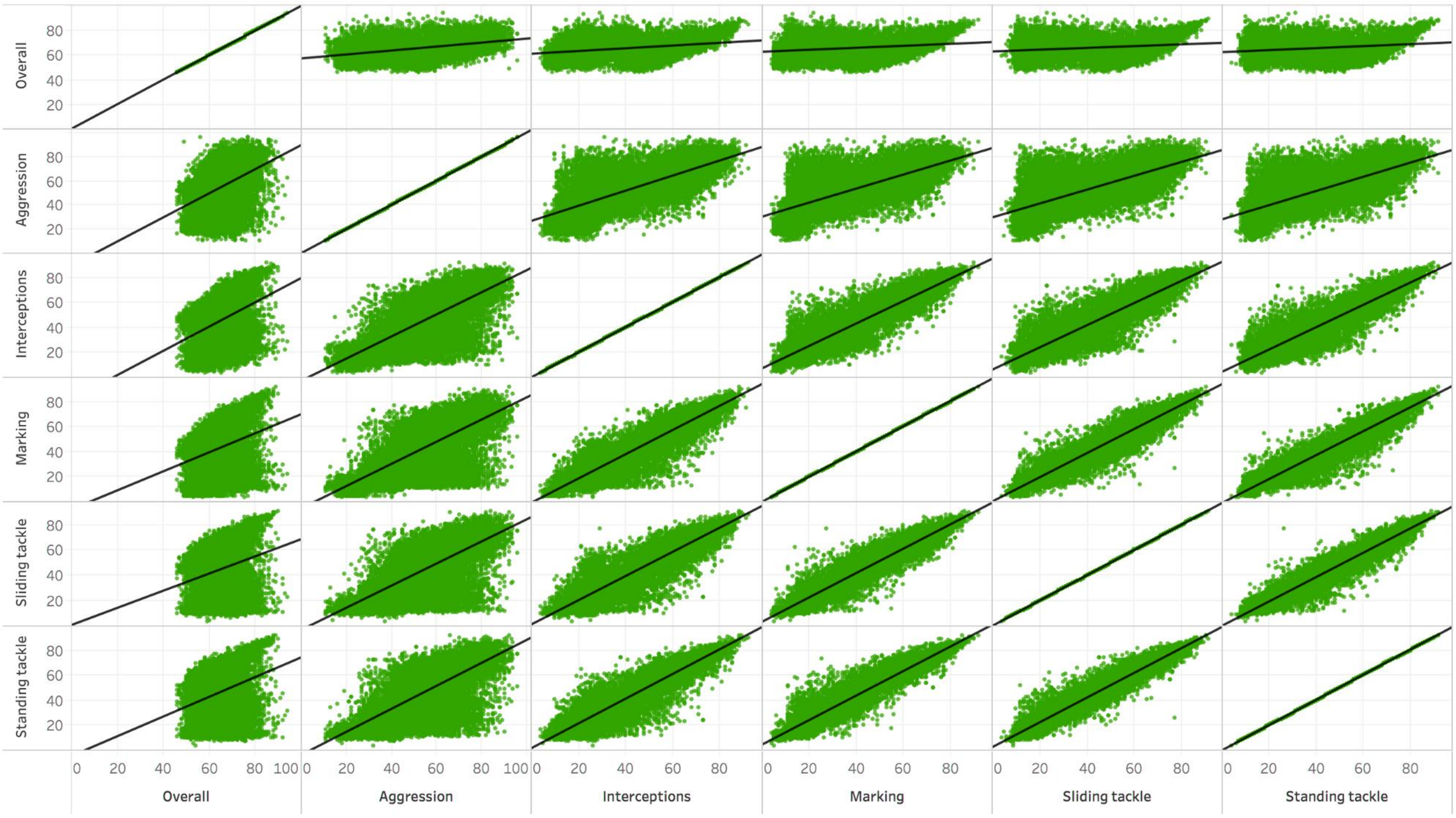
I used **Pearson's Correlation Coefficient Among Attributes**

- This was perhaps the most intriguing insight to me as I was exploring this dataset.The first question that popped into my mind even before I downloaded the dataset was which sub attributes, if any, influence the rating of other attributes? To examine this, I will attack it by position. To find the best defender we would look at the core attribute DEFENCE and how it correlates to all the other attributes .

It can be concluded that if a player's **STANDING TACKLE(s.t.)** attribute rating is high, then that player's other important defensive attributes are likely to be high as well

The **BEST DEFENDER** goes to **MATT HUMMELS(s.t.-93)** ,followed by **GIORGIO CHIELLINI (s.t. 92).**

Defensive Correlation Matrix



Standing Tackle, Sliding Tackle Corr	0.9725
Standing Tackle, Marking Corr	0.9611
Standing Tackle, Interceptions Corr	0.9350
Standing Tackle, Aggression Corr	0.7321
Standing Tackle, Overall Corr	0.2521

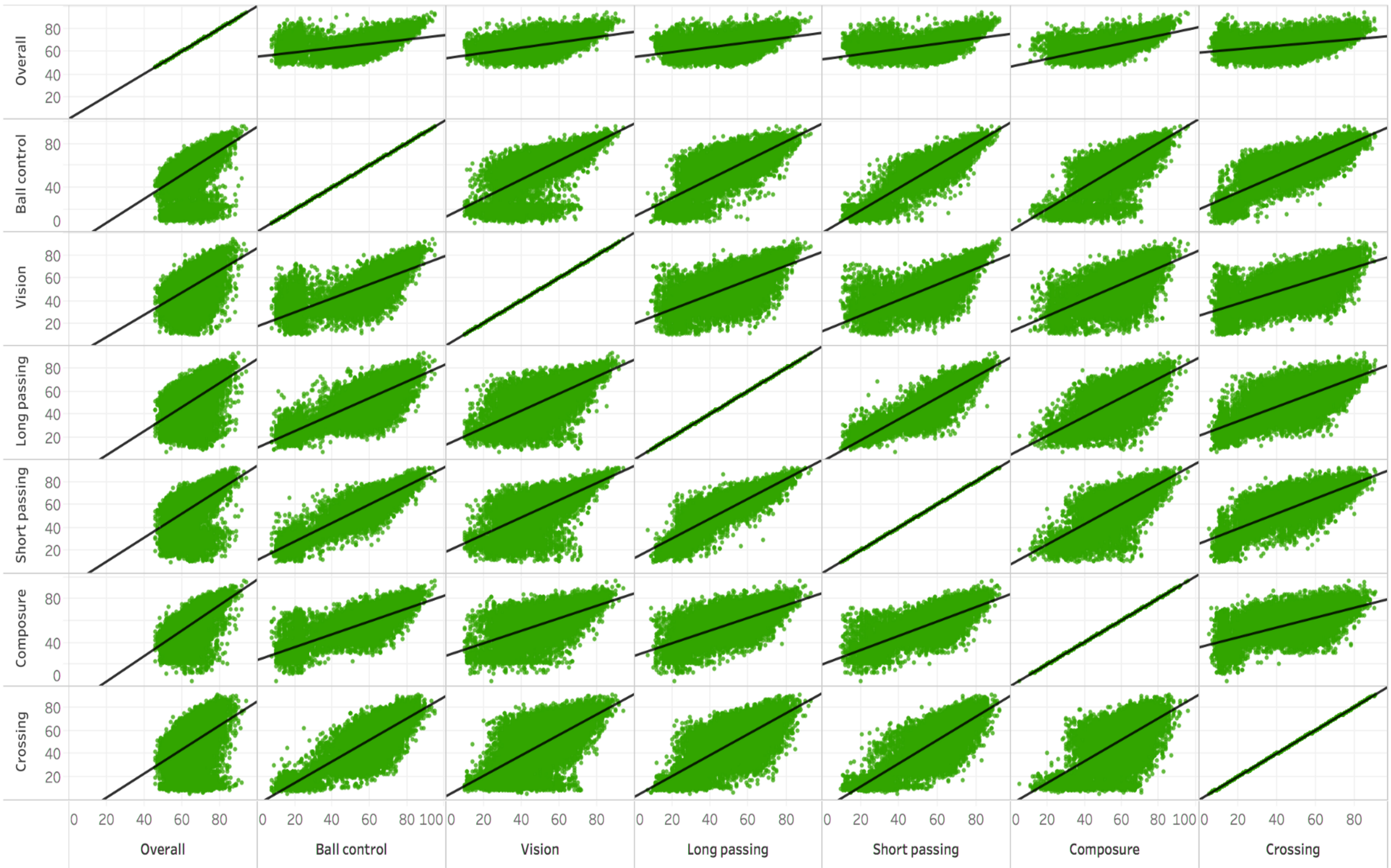
THE BEST MIDFIELDER

To find the best midfielder we would look at the core attribute PASSING ,DRIBBLING and how it correlates to all the other attributes .

It can be concluded that if a player has a high **COMPOSURE RATING(c.r.),** they are more likely to be a **BETTER MIDFIELDER** and have a **OVERALL RATING.**

The **BEST MIDFIELDER** goes to **DAVID SILVA (c.r.-93)** ,followed by **ANDRES INIESTA (c.r. 92).**

Midfield Correlation Matrix

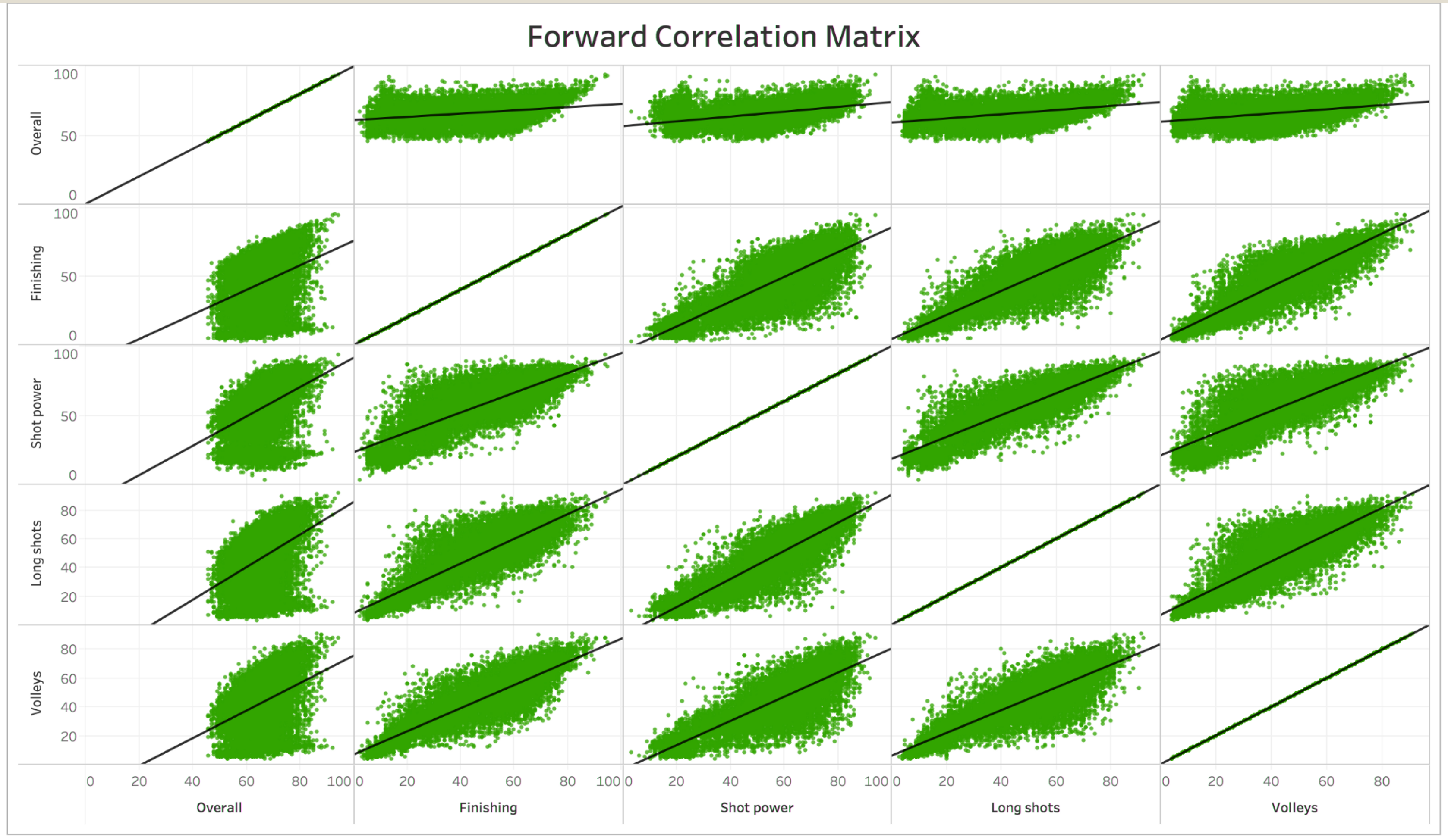


Composure	1,026,863
Overall	1,174,009
Overall, Composure Corr	0.6328

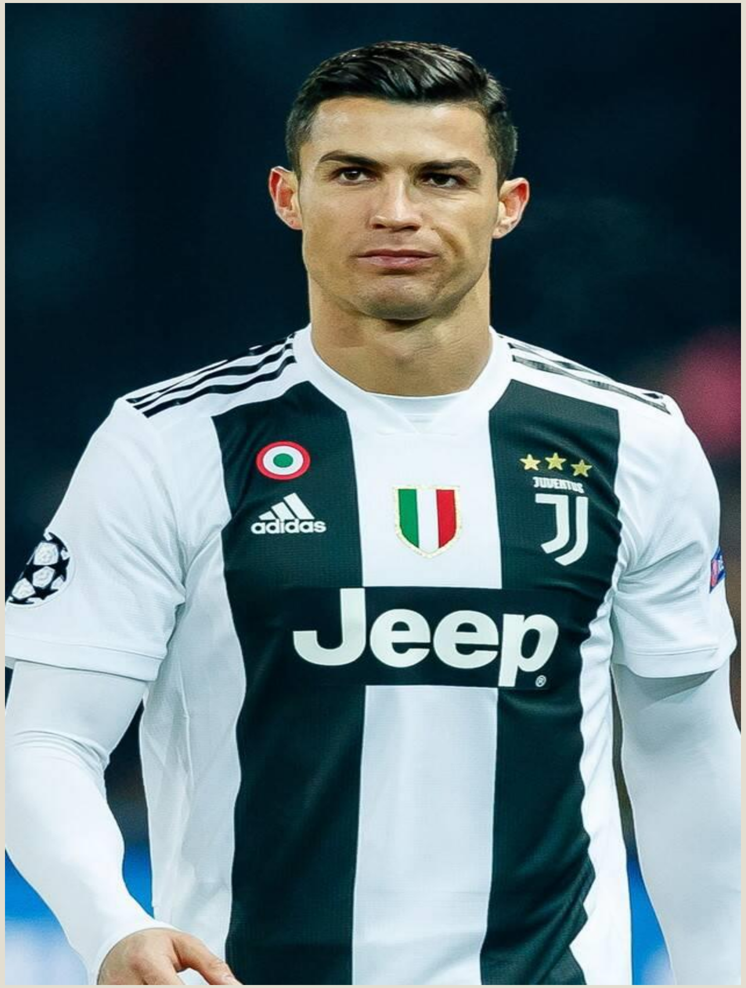
Ball Control, Short Passing Corr	0.9055
Ball Control, Crossing Corr	0.8392
Ball Control, Long Passing Corr	0.7836
Ball Control, Composure Corr	0.7637
Ball Control, Vision Corr	0.7227
Ball Control, Overall Corr	0.4544

THE BEST FORWARD

To find the best forward we would look at the core attribute SHOOTING and how it correlates to all the other attributes .



Finishing, Volleys Corr	0.8794
Finishing, Long Shots Corr	0.8666
Finishing, Shot Power Corr	0.7995
Finishing, Overall Corr	0.3222



It can be concluded that a player’s **FINISHING(f)** attribute is the most important attribute to look for, as it seems to have a positive relationship with all of the underlying Finishing attributes.

Finally ,The **BEST FORWARD** goes to **LIONEL MESSI (f-95)** ,followed by **CRISTIANO RONALDO (f- 94).**

THANK YOU

STAY TUNED FOR MORE INTRIGUING
EDA's