Predicting Player’s Value: FIFA19

AKASH KUMAR

25th April, 2020

1. Introduction

1.1 Background:

FIFA 19 is a football simulation video game developed by EA Vancouver as part of Electronic Arts' FIFA series. Announced on 6 June 2018 for its E3 2018 press conference, it was released on 28 September 2018 for PlayStation 3, PlayStation 4, Xbox 360, Xbox One, Nintendo Switch, and Microsoft Windows. It is the 26th installment in the FIFA series. As with FIFA 18, Cristiano Ronaldo initially as the cover athlete of the regular edition: however, following his unanticipated transfer from Spanish club Real Madrid to Italian side Juventus, new cover art was released, featuring Neymar, Kevin De Bruyne and Paulo Dybala. The game features the UEFA club competitions for the first time, including the UEFA Champions League and UEFA Europa League. Martin Tyler and Alan Smith return as regular commentators, while the new commentary team of Derek Rae and Lee Dixon feature in the UEFA competitions mode. Composer Hans Zimmer and rapper Vince Staples recorded a new remix of the UEFA Champions League anthem specifically for the game. The character Alex Hunter, who first appeared in FIFA 17 returns for the third and final installment of "The Journey", entitled, "The Journey: Champions”. In June 2019, a free update added the FIFA Women's World Cup as a separate game mode.

1.2 Problem:

A player’s value is everything as it gives us a complete background picture of the player’s performance as well as his popularity among his fans. Any data that might contain the player’s value could be used to predict the player’s reputation in the market. This project aims to bring out that reputation through value prediction.

1.3 Interest:

Any club might want to know his player’s as well as other club’s players market value for competitive advantage and could also bring new emerging talent into the market. Others might also be interested are fans and other business people who are in sports industry.

2. Data Acquisition and Cleaning

2.1 Data Sources:

All the player’s data containing his name, photo, player’s position etc. can be found in the csv file of Kaggle’s FIFA19 dataset here.

2.2 Data Cleaning:

Data downloaded has 18207 rows and 89 columns and has many missing data needed to be filled, redundant attributes that needed to be truncated and some similar attributes to be aggregated. There was no duplicate data in the dataset, all names were unique. Firstly, unnecessary columns like unnamed: 0, ID were dropped and then features like photo, club logo, etc which contains links were dropped. Features which contains player's personal information like: - *'Real Face','Joined','Loaned From','Contract Valid Until','Jersey Number','Release Clause','Special'* were dropped. This led to the formation of Data Frame containing 18207 rows and 77 columns containing 40 numerical attributes and 37 nominal attributes in which some were also ordinal like '*Work Rate',* etc.

Secondly, missing data needed to be handled. I wrote a script to fill the missing values of numeric and nominal data separately. Also, in numerical data there was two separation some attributes were type(int) so they needed to be filled integer values only. I used the imputer method of python to fill these values. Numeric data containing decimal was filled by the mean values and integer data by median of the respective attributes. Nominal data was filled by the most frequent value of the attributes. Some nominal attribute like *'Club'* were given no club value to the missing data while in some attributes like *'Body Type'* there was some abrupt value that needed to be filled right.

Thirdly, there were some nominal attribute like: - *'Height','Weight','Value','Wage',* which contains numeric type data which needed to be converted into decimal values. I wrote functions to convert these string values into decimal values: Height in inches, Weight in lbs and Value, Wage in euros.

Fourthly, there were many similar attributes that could be aggregated to form a single column which determine the overall functionality of those attributes. I studied about the Football terminologies which provided me method to aggregate them. Features like *'Acceleration','Sprint Speed'* comes under the **Pace** attribute and similarly other attributes were also aggregated by taking means into **Shooting, Passing, Technique, Defence, Physical, Goalkeeper, Rating.** In this rating attribute was formed by aggregating player's potential to perform and his overall rating, shooting attribute was formed by aggregating player's shot power, long shots, penalties accuracy, etc. Passing by aggregating vision, accuracy, curves, etc. Technique by aggregating dribbling, balance, reaction, ball control, etc. Defence from tackle capabilities, physical from their strength, jumping, etc. And goalkeeper from aggregating goalkeeper's necessary skills.

Lastly, dropping all the columns that has been aggregated and also dropping In-Game-Stats ratings like: - *'RF','CAM',* etc. because they just determines a player's ratings according to its positioning in the game and it's also obvious that no club would make his attacker to play in defence position and we already has column *'Position'* containing best position in which players mostly play their game. After all the cleaning and aggregating the data we have reduced the Data Frame into 18207 rows and 23 columns containing no missing data.

3. Exploratory Data Analysis

3.1 Descriptive Analysis:

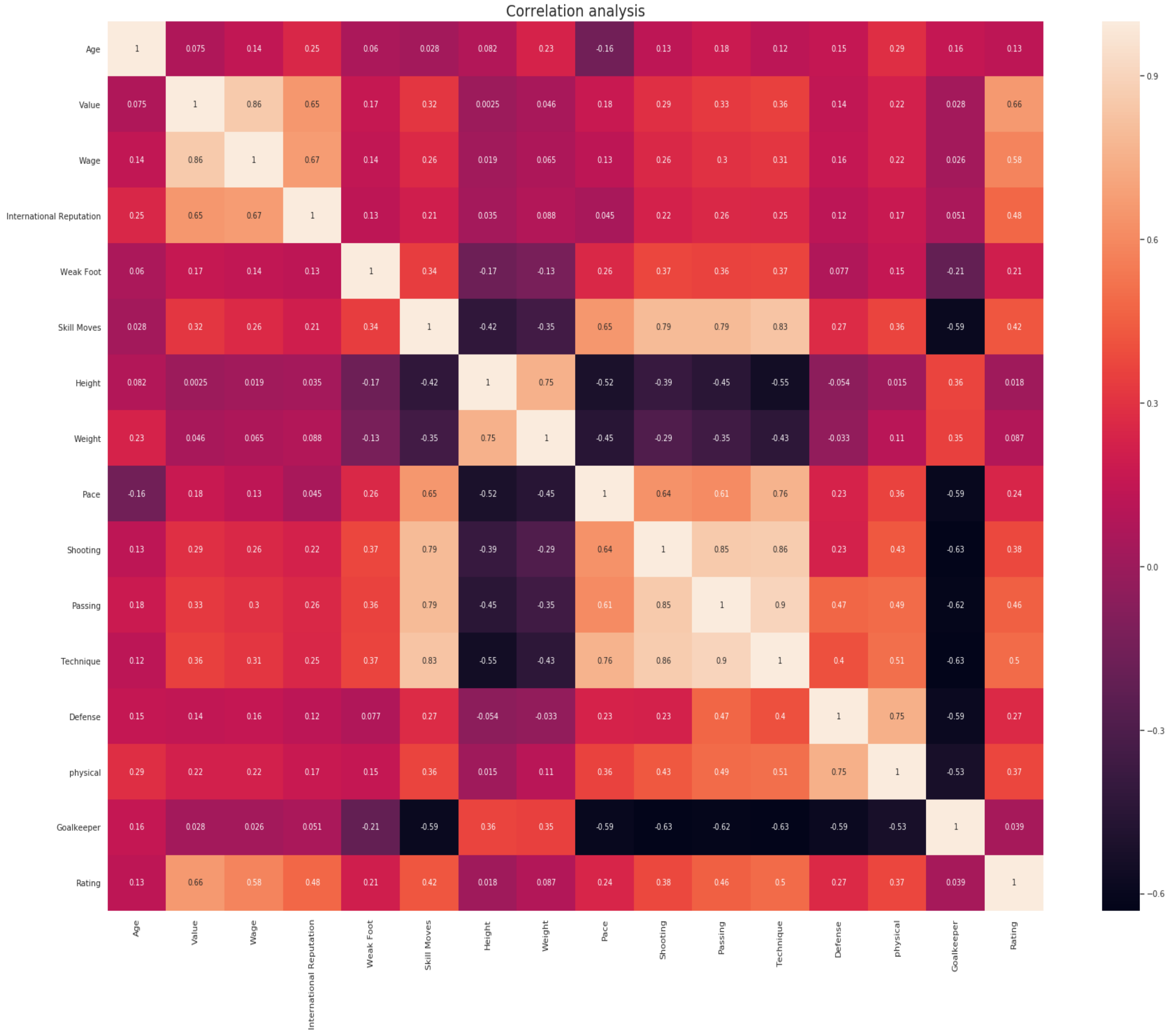
This include getting various insights about the data and players like: -

- Highest paid players according to their wages
- Top Indian players who are in the big league
- Players with the most rated goalkeeping skills

- Players who are most skilled in using their week foot or un-preferred foot in the game
- Players who have the most skill moves ratings
- Clubs which are most popular according to their economic situations
- Most attacking and defensive clubs and their respective best player
- Correlation between the numeric attributes of the data

The correlation heatmap is here:

- It is clearly visible that value of the player is highly correlated to the wage of the player
- It is also mostly correlated to the player's rating, shooting capabilities, passing capabilities, International reputation, skill moves and technique
- This provides us an inference that most of the players who are mid-fielders or forwards have higher value



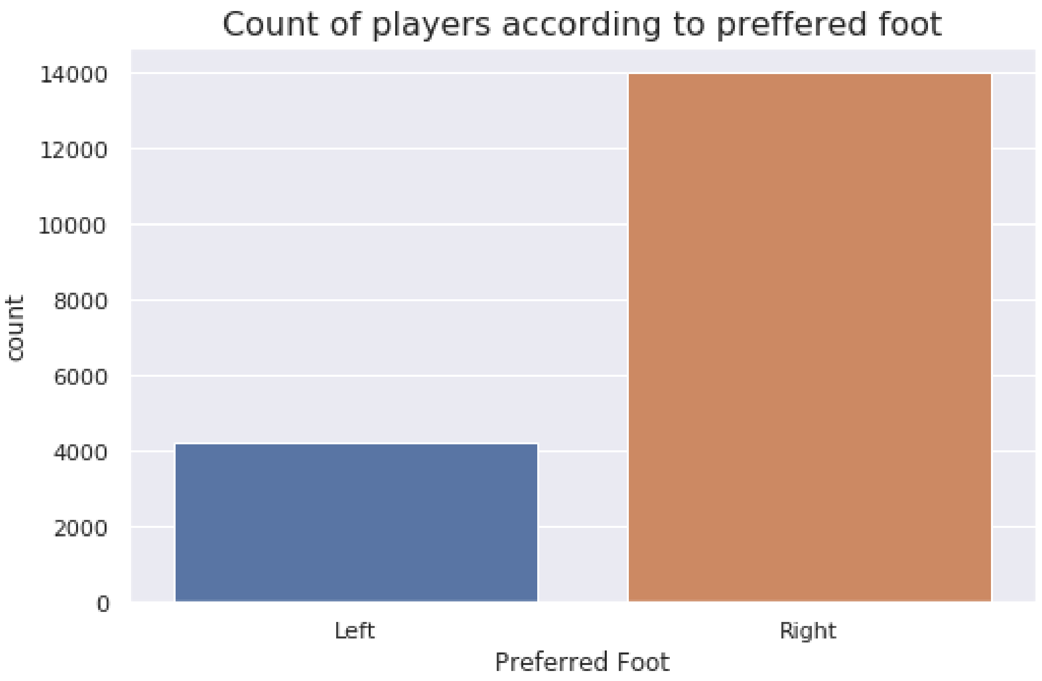
Correlation analysis

	Age	Value	Wage	International Reputation	Weak Foot	Skill Moves	Height	Weight	Pace	Shooting	Passing	Technique	Defense	physical	Goalkeeper	Rating
Age	1	0.075	0.14	0.25	0.06	0.028	0.082	0.23	-0.16	0.13	0.18	0.12	0.15	0.29	0.16	0.13
Value	0.075	1	0.86	0.65	0.17	0.32	0.0025	0.046	0.18	0.29	0.33	0.36	0.14	0.22	0.028	0.66
Wage	0.14	0.86	1	0.67	0.14	0.26	0.019	0.065	0.13	0.26	0.3	0.31	0.16	0.22	0.026	0.58
International Reputation	0.25	0.65	0.67	1	0.13	0.21	0.035	0.088	0.045	0.22	0.26	0.25	0.12	0.17	0.051	0.48
Weak Foot	0.06	0.17	0.14	0.13	1	0.34	-0.17	-0.13	0.26	0.37	0.36	0.37	0.077	0.15	-0.21	0.21
Skill Moves	0.028	0.32	0.26	0.21	0.34	1	-0.42	-0.35	0.65	0.79	0.79	0.83	0.27	0.36	-0.59	0.42
Height	0.082	0.0025	0.019	0.035	-0.17	-0.42	1	0.75	-0.52	-0.39	-0.45	-0.55	-0.054	0.015	0.36	0.018
Weight	0.23	0.046	0.065	0.088	-0.13	-0.35	0.75	1	-0.45	-0.29	-0.35	-0.43	-0.033	0.11	0.35	0.087
Pace	-0.16	0.18	0.13	0.045	0.26	0.65	-0.52	-0.45	1	0.64	0.61	0.76	0.23	0.36	-0.59	0.24
Shooting	0.13	0.29	0.26	0.22	0.37	0.79	-0.39	-0.29	0.64	1	0.85	0.86	0.23	0.43	-0.63	0.38
Passing	0.18	0.33	0.3	0.26	0.36	0.79	-0.45	-0.35	0.61	0.85	1	0.9	0.47	0.49	-0.62	0.46
Technique	0.12	0.36	0.31	0.25	0.37	0.83	-0.55	-0.43	0.76	0.86	0.9	1	0.4	0.51	-0.63	0.5
Defense	0.15	0.14	0.16	0.12	0.077	0.27	-0.054	-0.033	0.23	0.23	0.47	0.4	1	0.75	-0.59	0.27
physical	0.29	0.22	0.22	0.17	0.15	0.36	0.015	0.11	0.36	0.43	0.49	0.51	0.75	1	-0.53	0.37
Goalkeeper	0.16	0.028	0.026	0.051	-0.21	-0.59	0.36	0.35	-0.59	-0.63	-0.62	-0.63	-0.59	-0.53	1	0.039
Rating	0.13	0.66	0.58	0.48	0.21	0.42	0.018	0.087	0.24	0.38	0.46	0.5	0.27	0.37	0.039	1

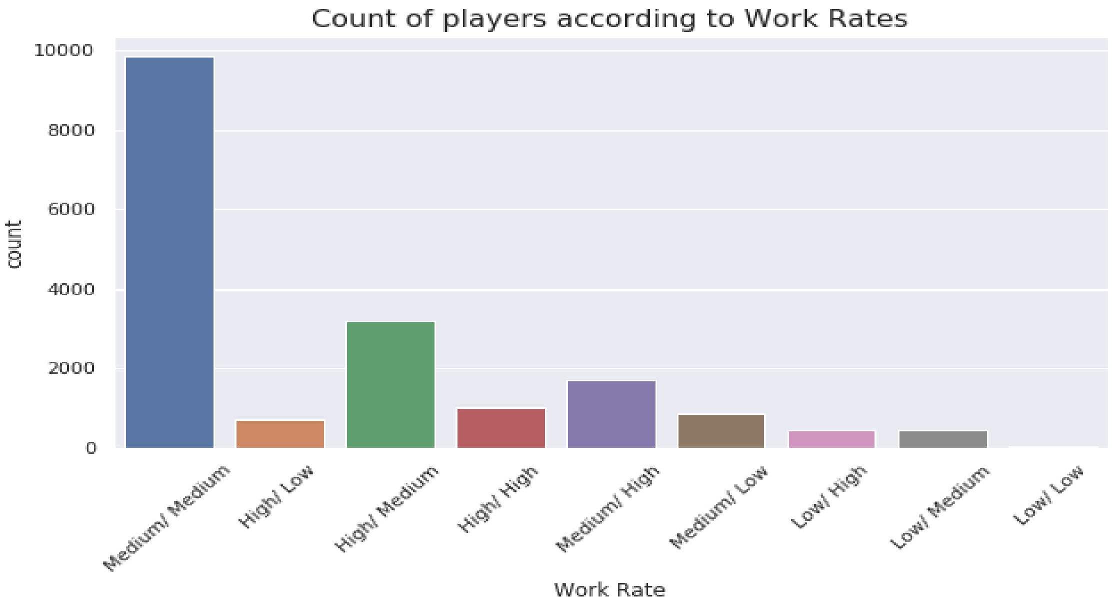
3.2 Visual Statistical Analysis:

Many methods were employed to extract patterns from the data

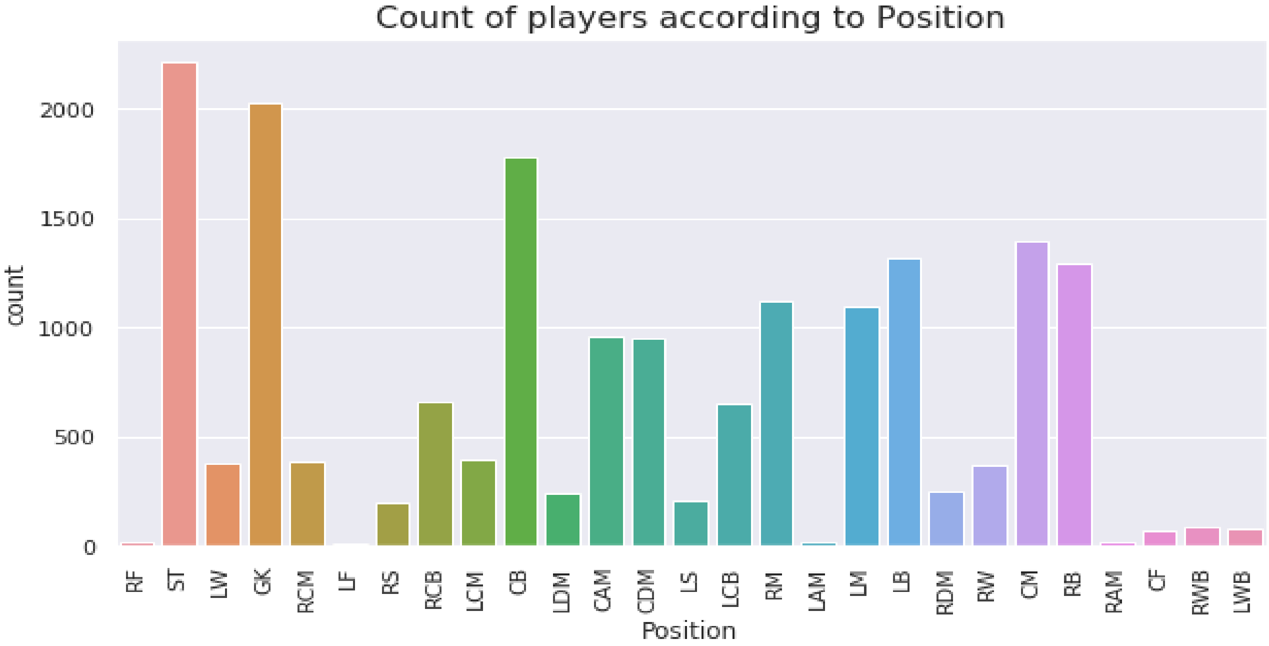
- **Count Plots**
  - This plot shows the variation of number of players according to the foot they use to play



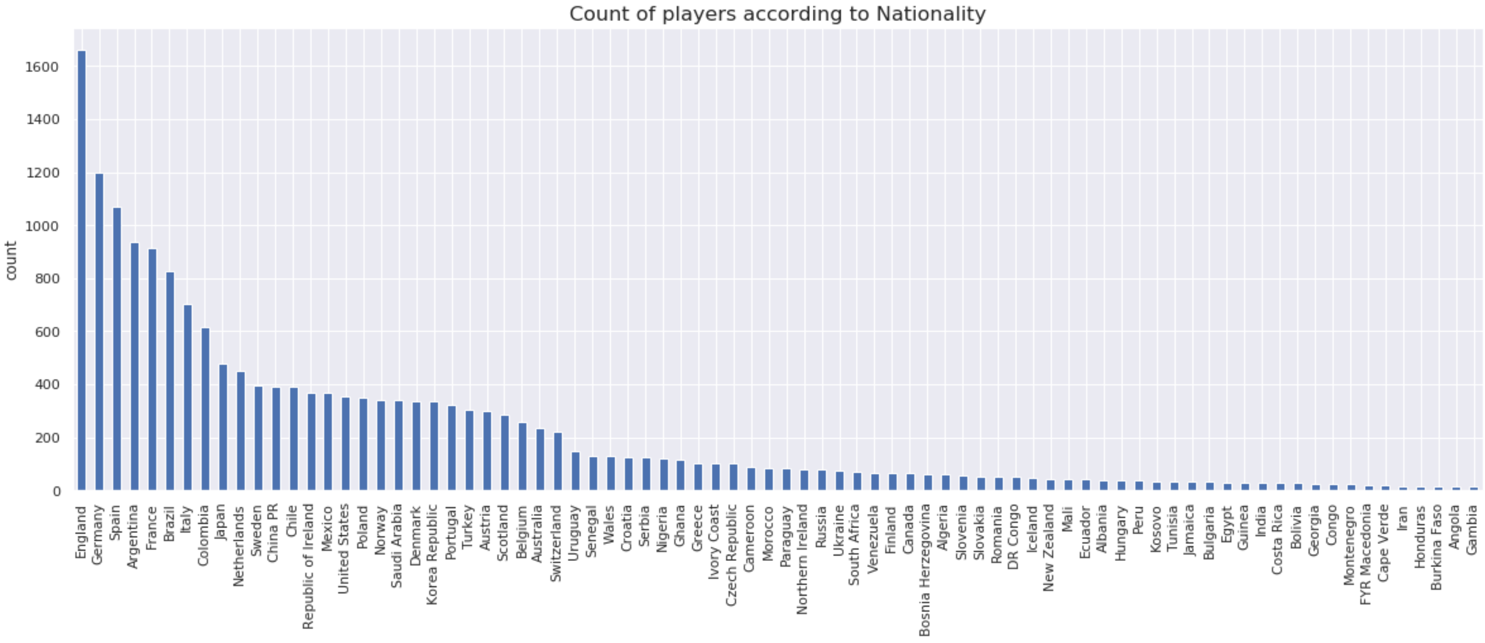
  - This plot shows the count of players according to their work rates in the field indicating the presence of large number of players are M/M work rating players



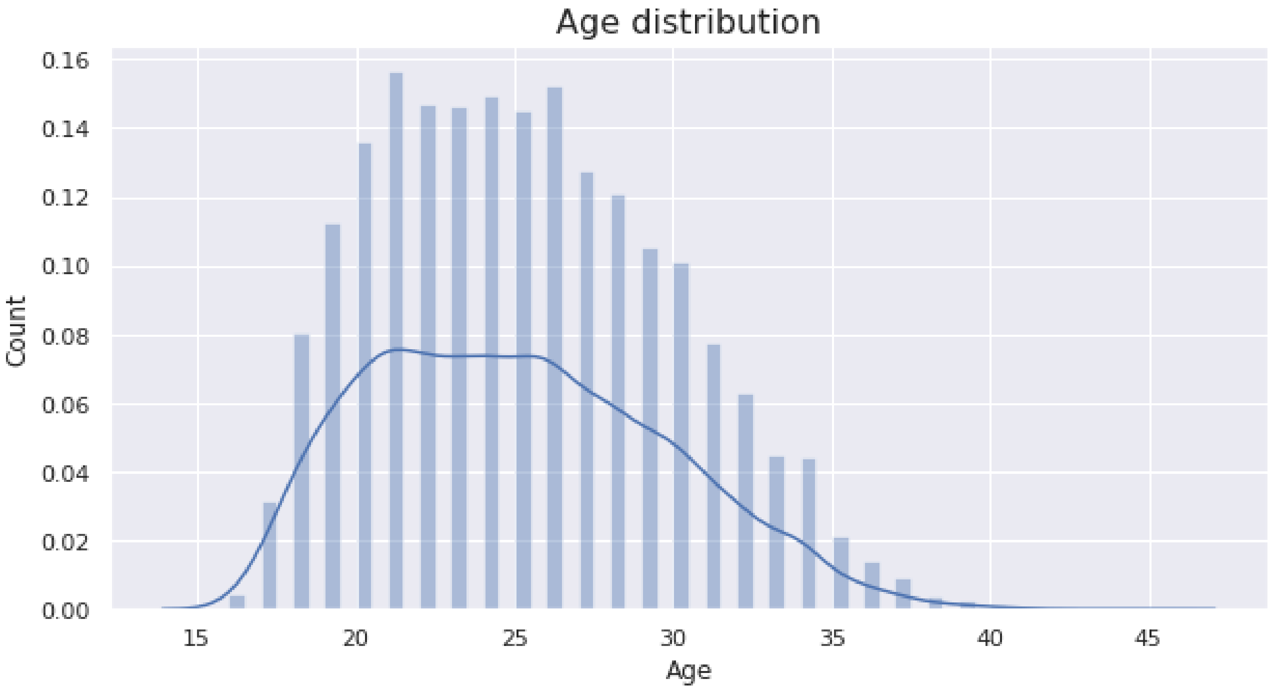
- From the plot it's clear that data contains many goalkeepers but strikers are largest in number having corner back position at third



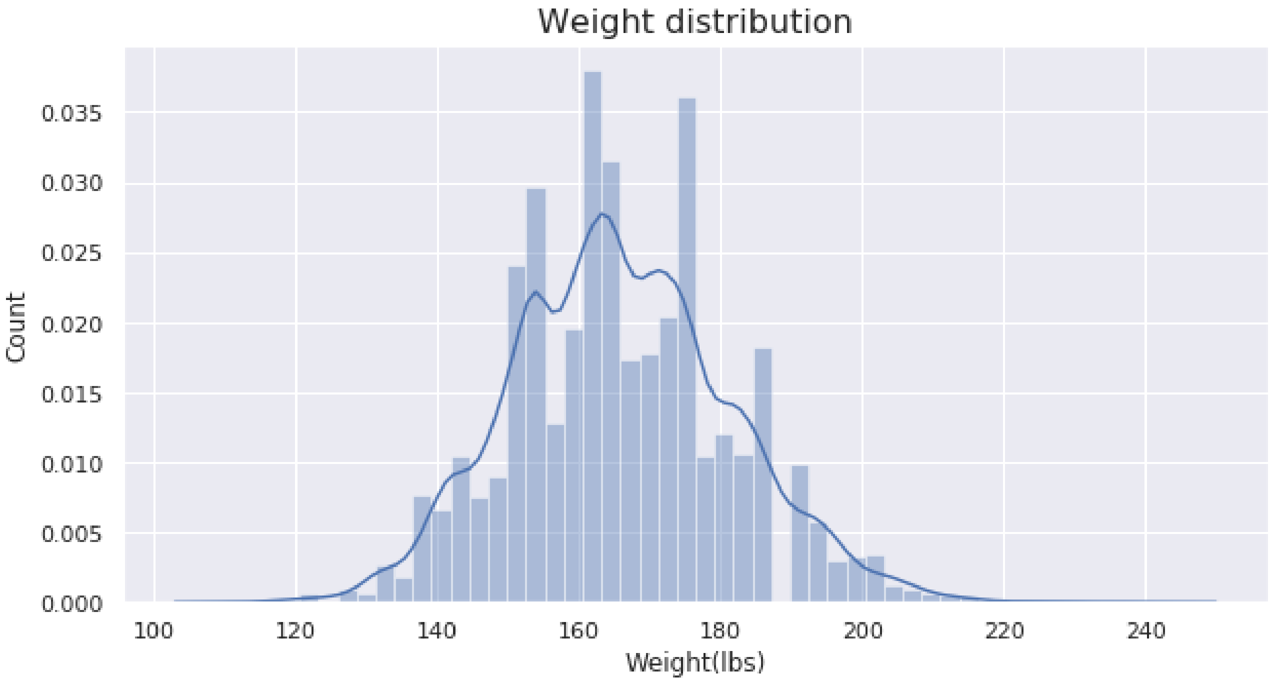
- Most of the players in football have a European or South American origin, England having the highest number of player origin.



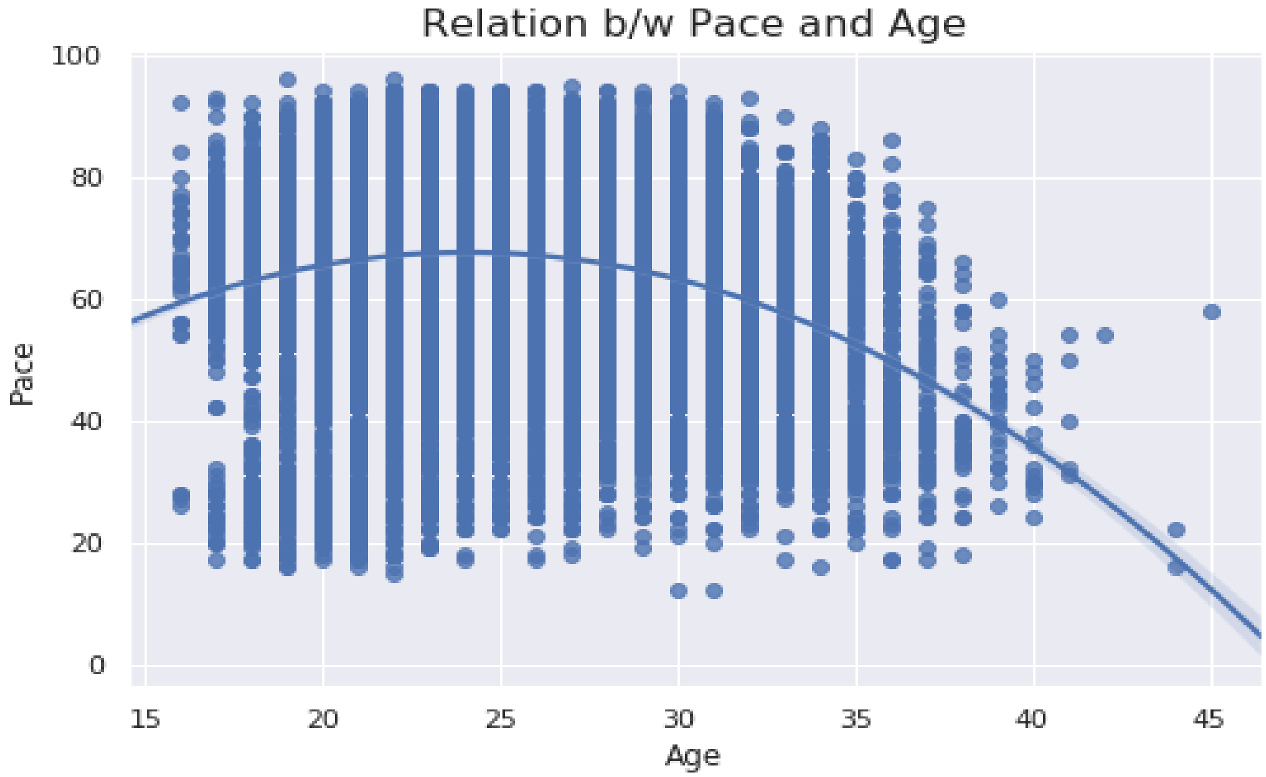
- **Distribution Plots**
  - Most of the players have age b/w 20 to 30 years



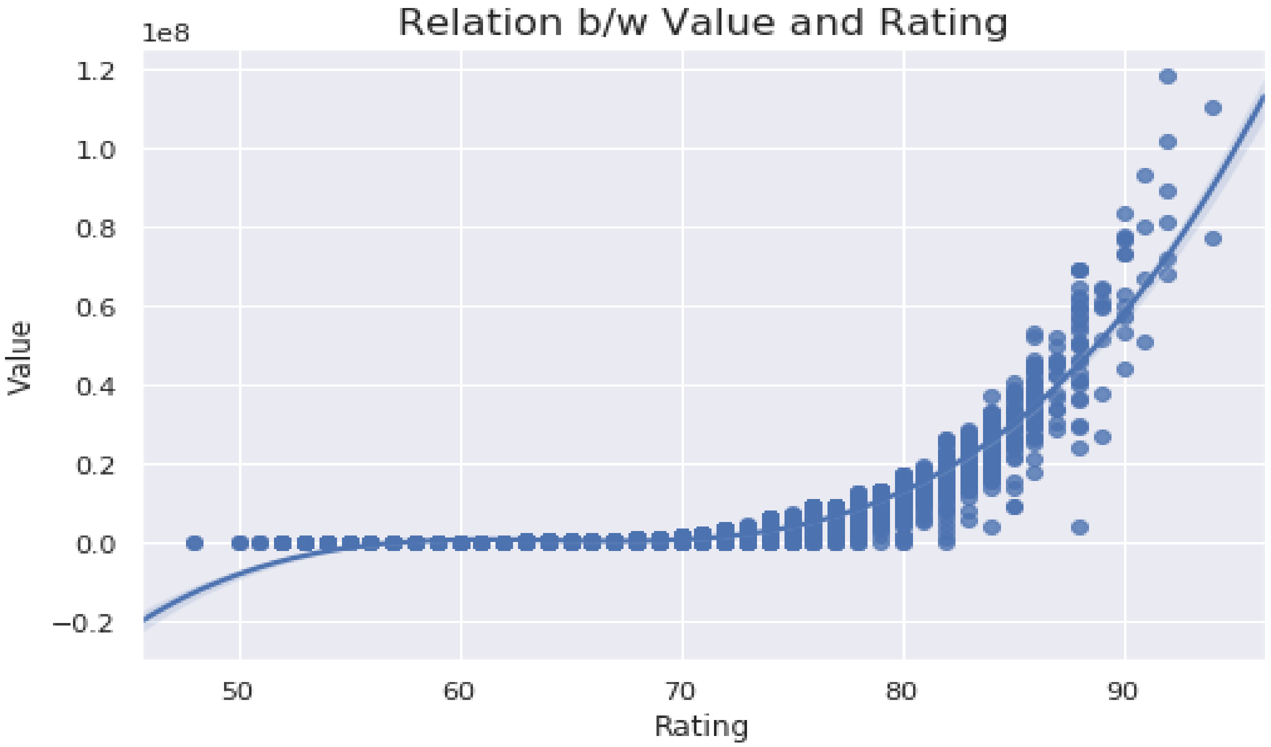
  - Most of the players have weight b/w 160 to 170 lbs



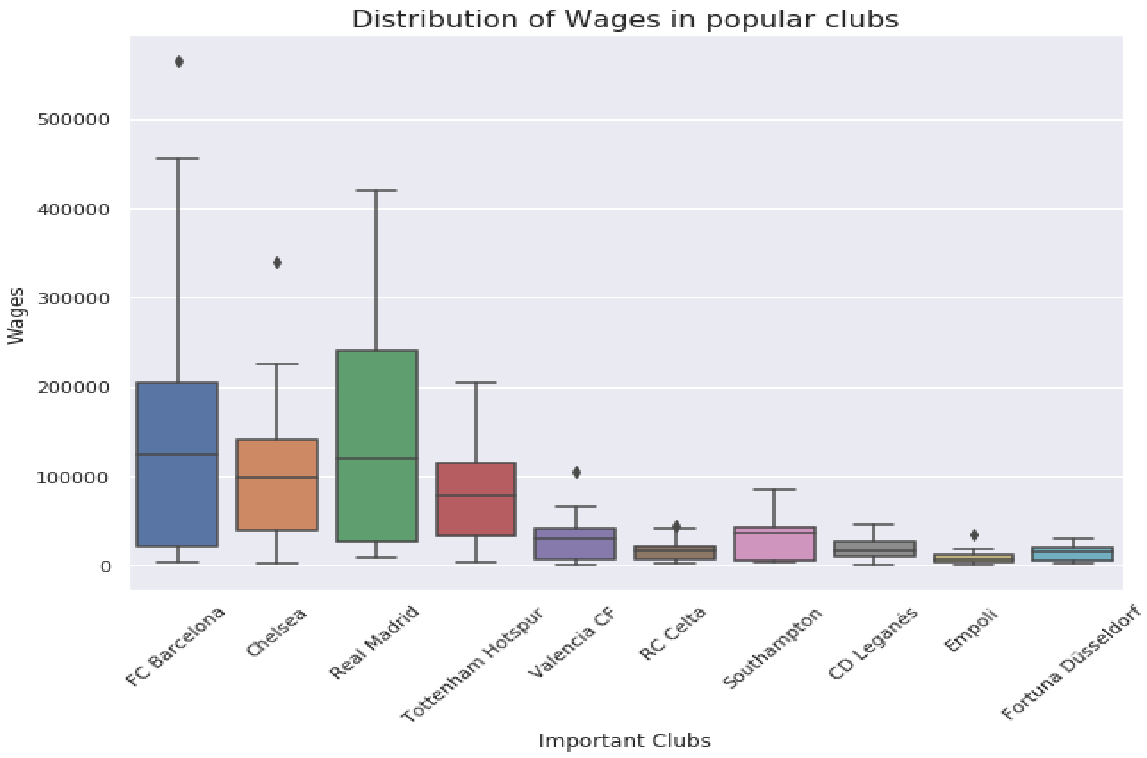
- **Regression Plots**
  - As the age of the player decline his pacing capabilities like sprint speed and acceleration reduces



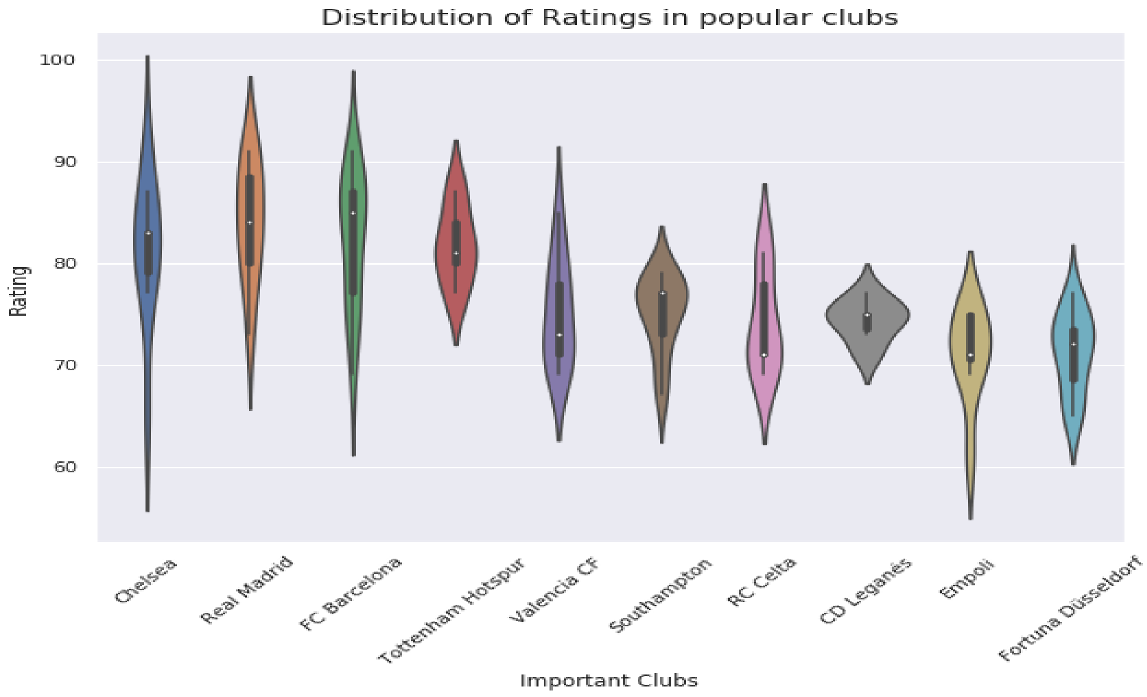
  - There is a high correlation between the rating of a player and its value. More skilled player have higher value in the market



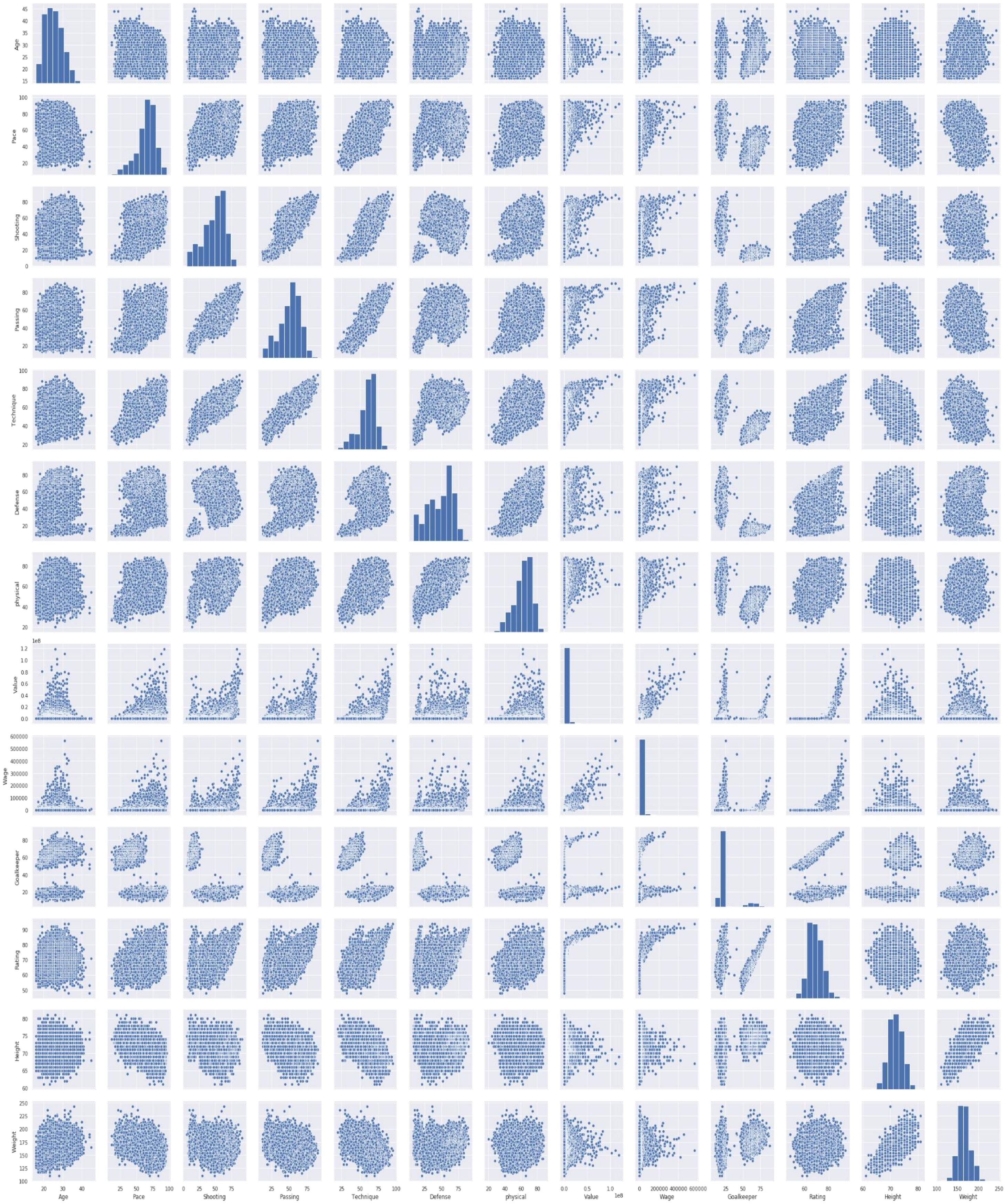
- **Club Plots**
  - Comparing among some popular club Barcelona and Real Madrid clubs are most economically profound clubs



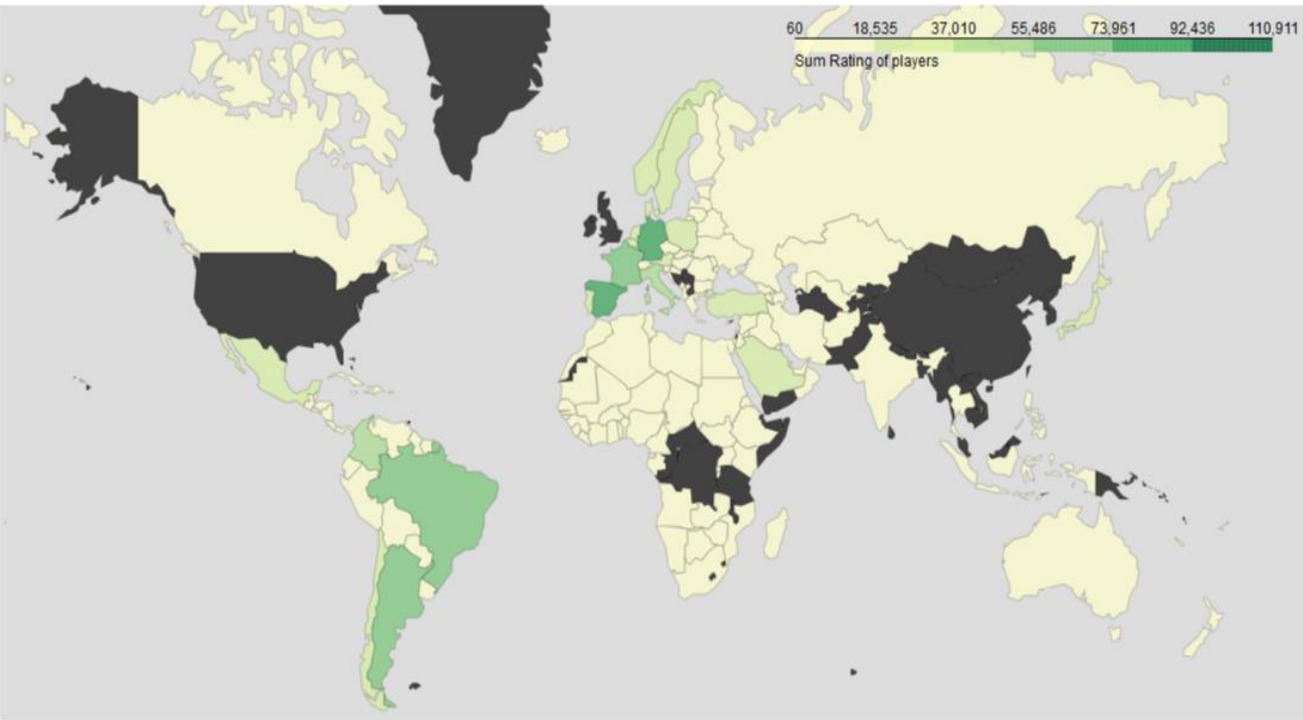
  - Ratings of Chelsea, Barcelona and Real Madrid are much higher than other



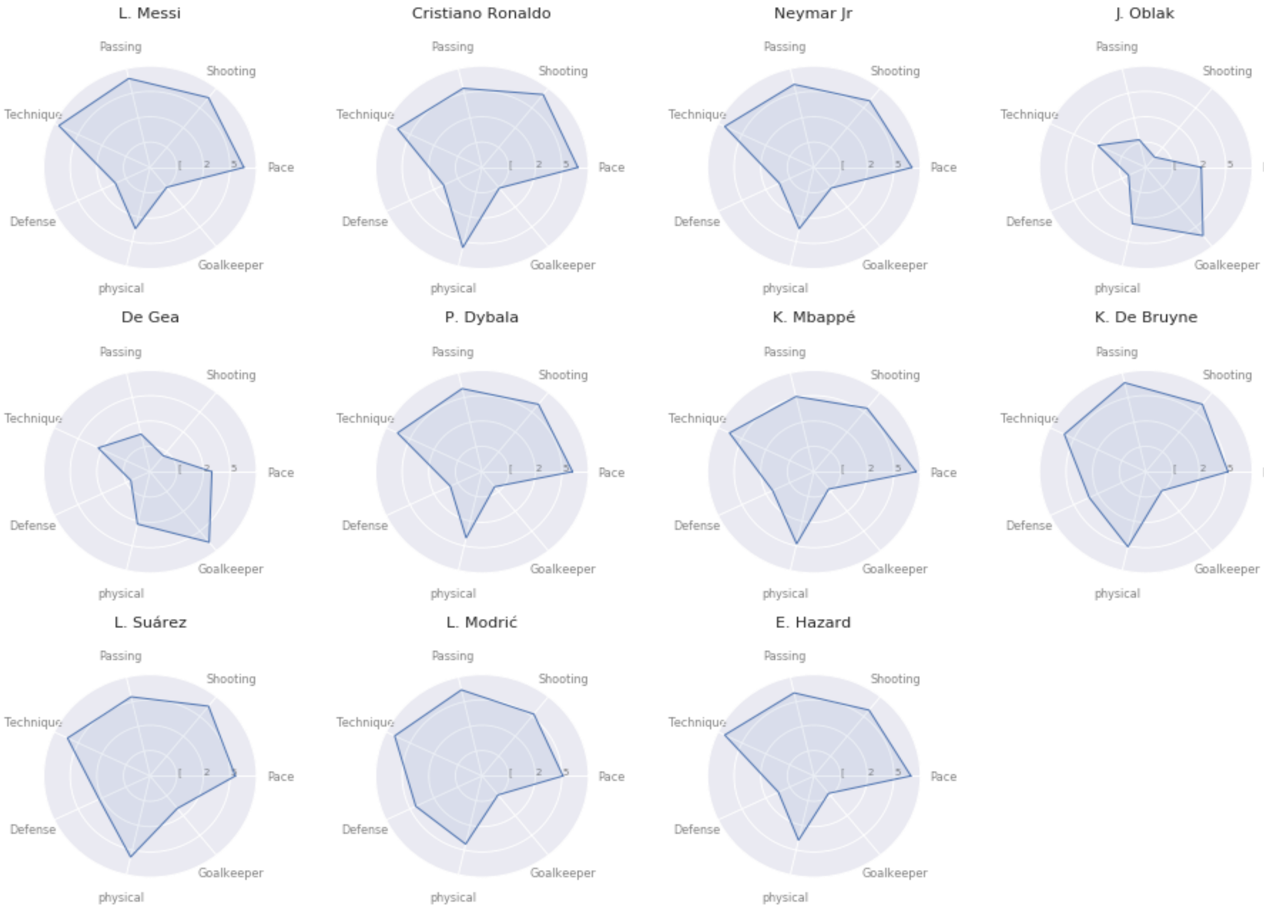
- **Pair Plot**
- Pair plot provides us all the correlated features in the data and its visualization on a single graph



- **Rating Map**
  - Folium map showing the top-rated players worldwide



- **Radar Plot**
  - Plot showing the skills of top 11 players according to rating



4. Data Modelling

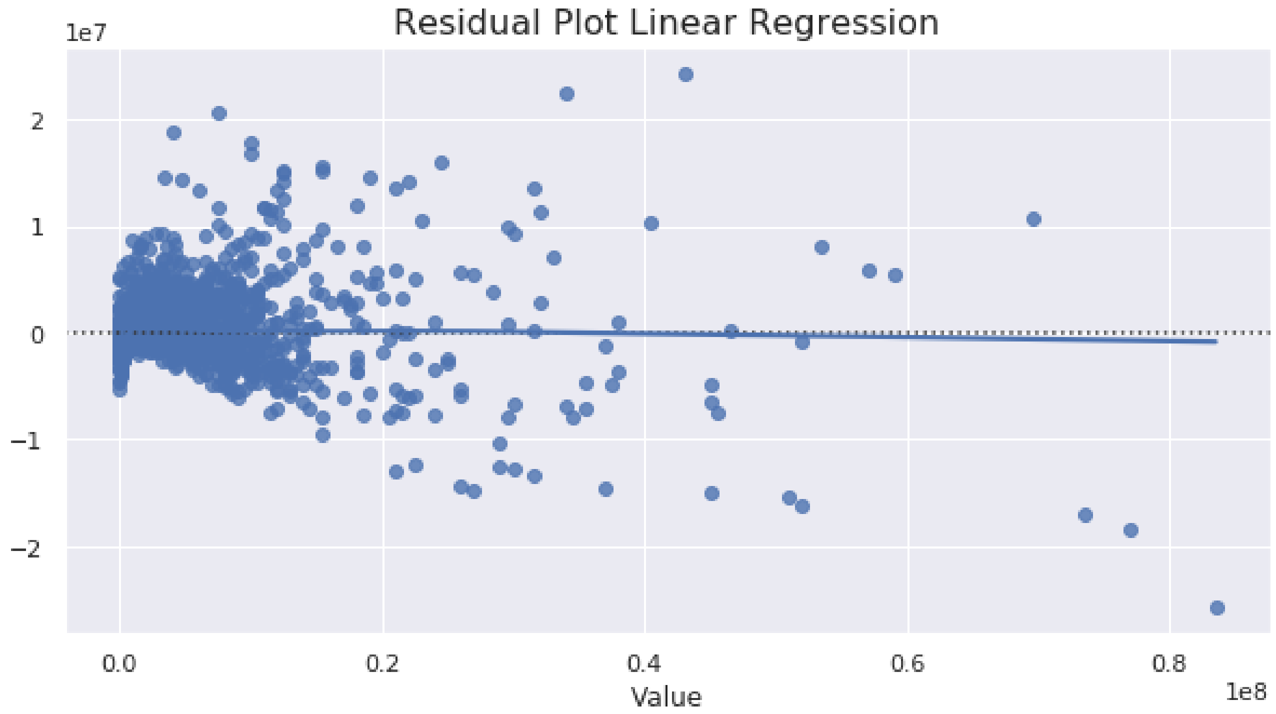
4.1 Data Preparation:

Firstly, name, club name, nationality, etc. are dropped for the preparation of data frame for the model only numerical data along with one feature having position data is taken because a player's position in the field determine whether he will get to score or not as we have already seen earlier that strikers have more value in the market than other players.

Secondly, the categorical variable needed to be converted into numerical for regression which led to the formation of final data frame containing 42 features including the target dataset. Then we split the data set into target and train variables after that further splitting was performed to create train and test sets and scaling of the data was performed before engaging the model.

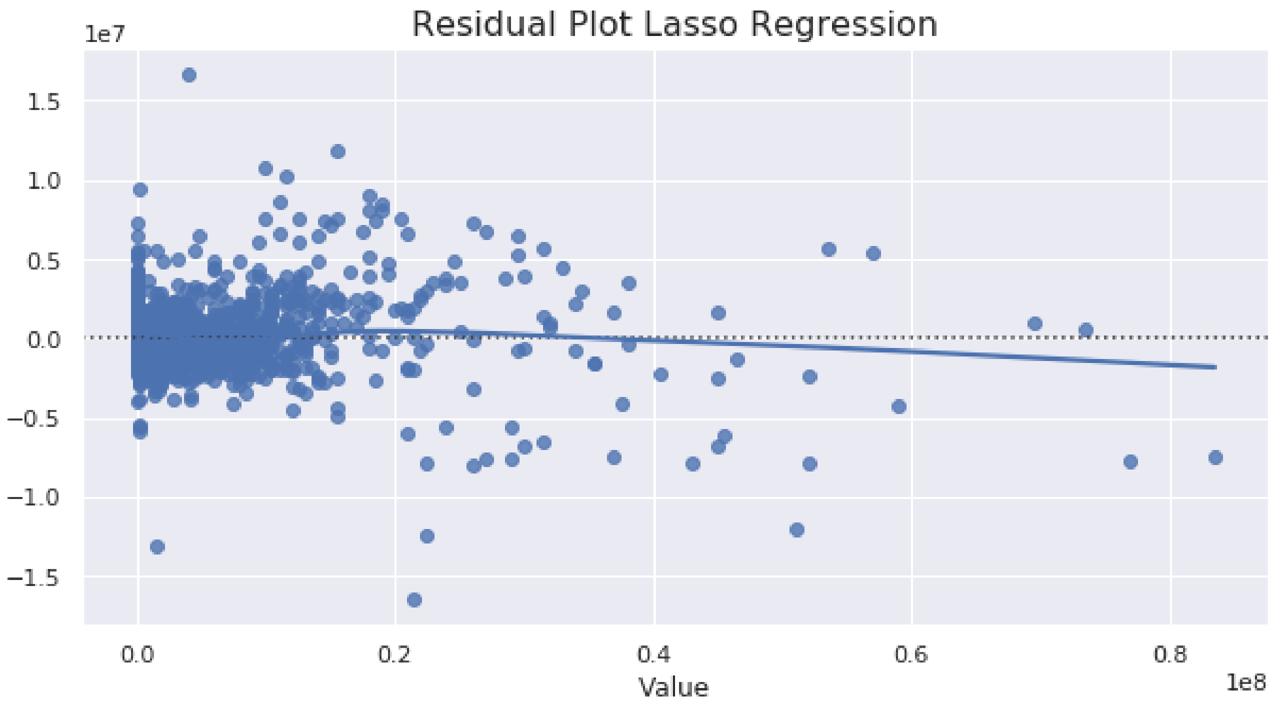
4.2 Linear Regression:

A simple linear regression model was employed on the training data which provided a R2 score of 0.77 on the test data. This could mean that a linear model is not the best model for the data. The residual plot is given below for linear Regression.



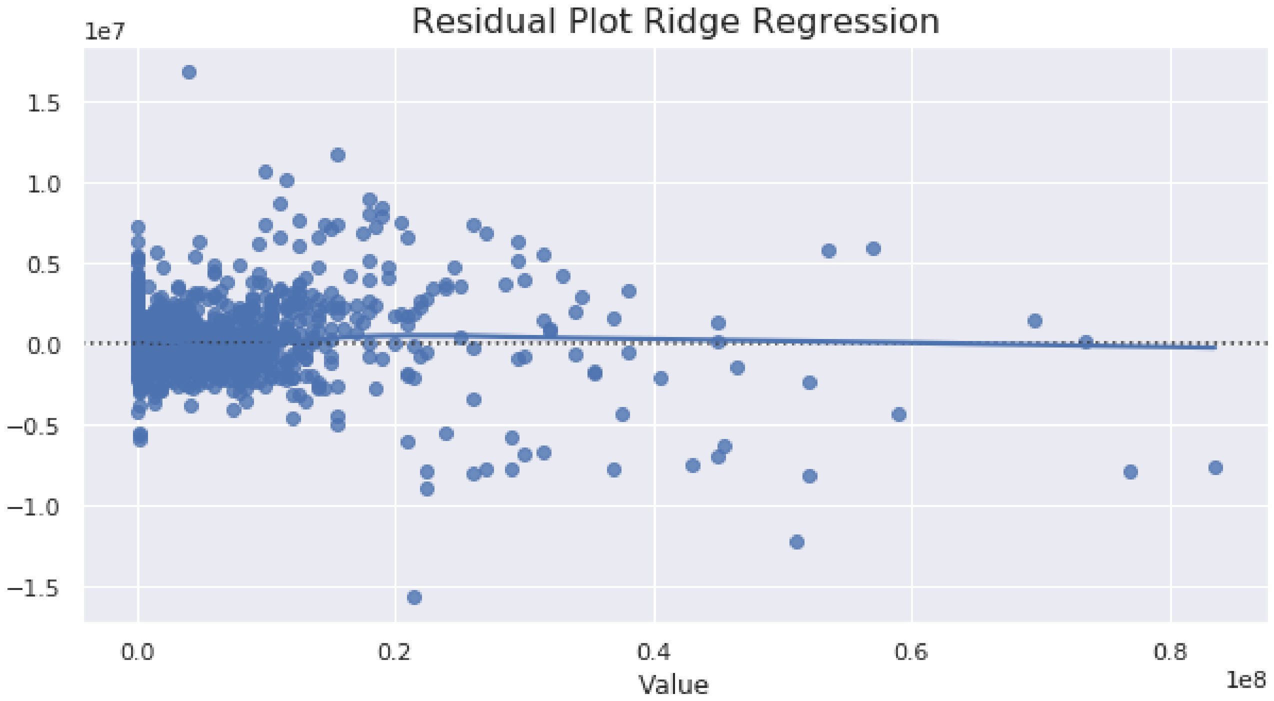
4.3 Lasso Regression:

The next regressor employed was lasso this time we took the polynomial feature into account and set the degree of polynomial to be quadratic which gave us a R2 score of 0.93 on the test data, which was really good value and shows the non linear behaviour of data. The residual plot is given below for Lasso Regression.



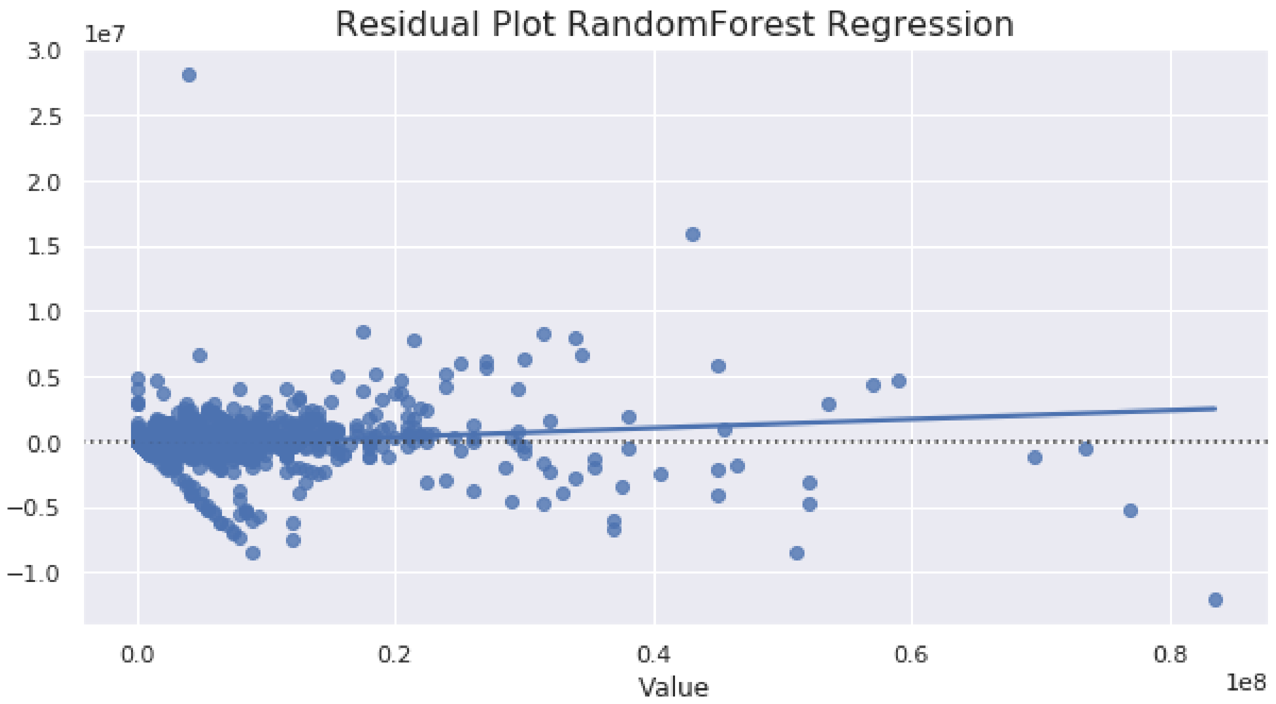
4.4 Ridge Regression:

Next was Ridge Regression with polynomial features having degree as quadratic providing a R2 score of 0.94 on the test data, which shows that L2 regularization worked better on the data. The residual plot is given below for Ridge Regression.



4.5 Random Forest Regressor:

This regressor model provided the best score as it is more robust to outliers providing a R2 score of 0.96 on the test data. The residual plot for Random Forest Regressor is given below.



5. Result

Regression Models	R2 Score
Linear Regression	0.77
Lasso Regression	0.93
Ridge Regression	0.94
Random Forest Regressor	0.96

6. Discussion and Conclusions

The models performed well on the test data but it could be much better by taking in account the outliers because there were many player's having very less value and some having very high value. These models could also be used to predict the ratings of the value according to their skills. The clubs could use these to predict their player's value in the market and to solve many business problems.