Project on Fifa 19 Analysis and Recommendation

Summer Training Project
National Institute Of Industrial Training,
Victoria Park, 37/2 Plot GN, Sector V, Kolkata-700091

Submitted By,
Subhabrata Kanjilal
[Institute Of Technology,Guru Ghasidas University]
[Bilaspur,Chhattisgarh-495001]

Acknowledgements

I am thankful to my institute for providing me with various resources and support during my training period. I am also very thankful to my training executives , Mr. Ranit Mishra and Mr. Sanjib Dutta for providing a beautiful oppurtunity to enroll in the summer training program in Artificial Intelligence with Machine Learning. I am Grateful to my mentor, Mr. Sayanton Chakraborty for providing his time and effort to guide me which helped me to complete my project . I am also invariably indebted to my parents whose sacrifice and constant support motivates me to pursue my career. I am also grateful to my colleagues for providing a supportive and interactive environment which really helped me to grasp the concepts of the topics clearly. I hope that my project might meet the expectations of my Institute .

Certificate

This to certify that **Subhabrata Kanjilal** underwent the summer training program on **Artificial Intelligence with Machine Learning** during the period **6th June,2019** to **25th June,2019** and successfully completed the project on the topic **Fifa 19 Analysis and Recommendation** and submitted it on **20th July,2019** under the guidance of **Mr. Sayantan Chakraborty.** This project is the original work of the student and is subjected to his copyright.

Seal of the Institution

Signature of the authority

Index

Introduction

FIFA is a football based game developed by game giants EA Sports. FIFA 19 contains a rich dataset with plenty of attributes covering all aspects of a real-life footballer in an attempt to immitate him as much as possible in the virtual world. This rich dataset provides a huge oppurtunity for us, data scientists or data analysts to analyze and come up with visualizations and patterns. In this paper, I will try to cover the following:

- The dataset distribution based on player nationality, player overall rating, age vs overall rating, player valuation and so on.
- The patterns in the dataset.
- Suggestions for playing Manager Mode.
- Building Models like Random Forest Regressor ,Linear Regressor , classification regrssor etc.
- Predicting the test results based on the models.

The Goal of this project is to build a recommendation system for managers via analysis of the downloaded FIFA data and building different maxhine learning models on the dataset.

Data and Preprocessing

The dataset consists data about players and their details regarding their Career information and in-field potentials ,all around the world. It consists names of 18207 players and their respective 89 characteristics/features such as Name, Age, Nationality, Overall, Potential, Club, Finishing , Header Accuracy, LB, CF etc.Such large amounts of features enable the managers to explore various techniques to predict the player characteristics.

First we import our data from dataset data.csv , which contains 18207 entries about details of players , each having 89 characteristics.There are 38 floating point datatypes, 6 integer type datatype and 45 object type datatype.

Exploratory Data Analysis

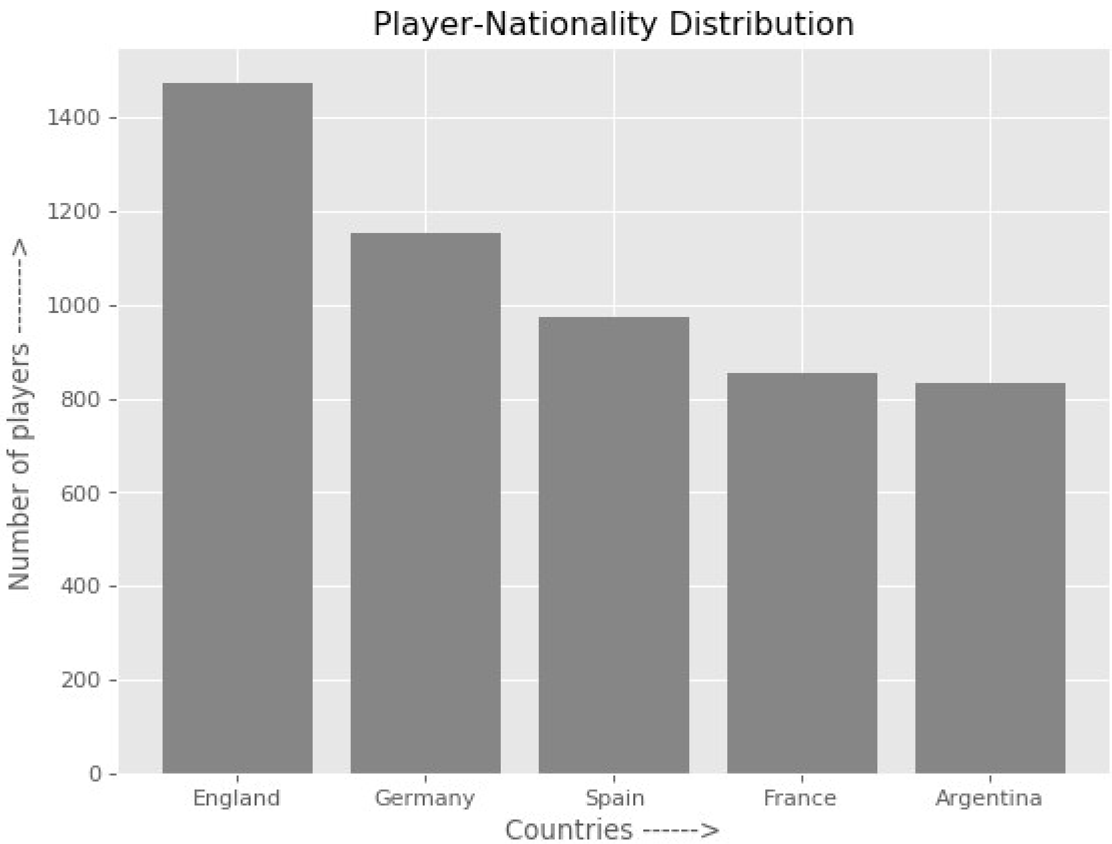
First we check the shape of the dataset :

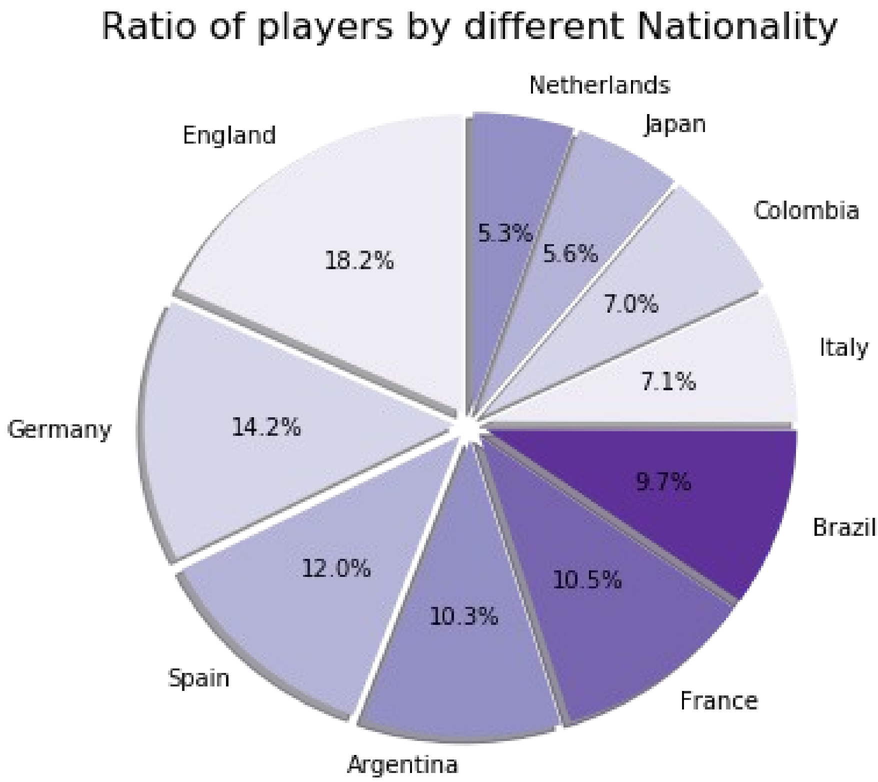
Dimension of the dataset is: (18207, 89)

We noticed that there are a number of fields which are of no use in our analysis. These fields include Unnamed:0, ID, Photo, Club Logo and so on. We also observed that there are number of fields whose values are missing in the dataset. These fields need to be dropped. There are also some fields like Value, Wage which needs modification so as to aid our analysis process.
We saw that all the columns have the same number of records. Hence, we started modifying some of the values in order to aid our visualizing process. One such column in the Value column. The Value column lists certain player valuations in thousands and some in millions. This poses two problems. First of all, we need to convert the values to a numeric type since we cannot create the required visualizations without doing so. Secondly, we need to modify the values in terms of a single unit i.e. either thousands or in millions. we modify the Value column values to a numeric type in terms of millions. This is done with the help of regular expressions and some simple mathematical calculations. We also convert the Wages field values to numeric types represented in terms of thousands.

1. Dataset Distribution

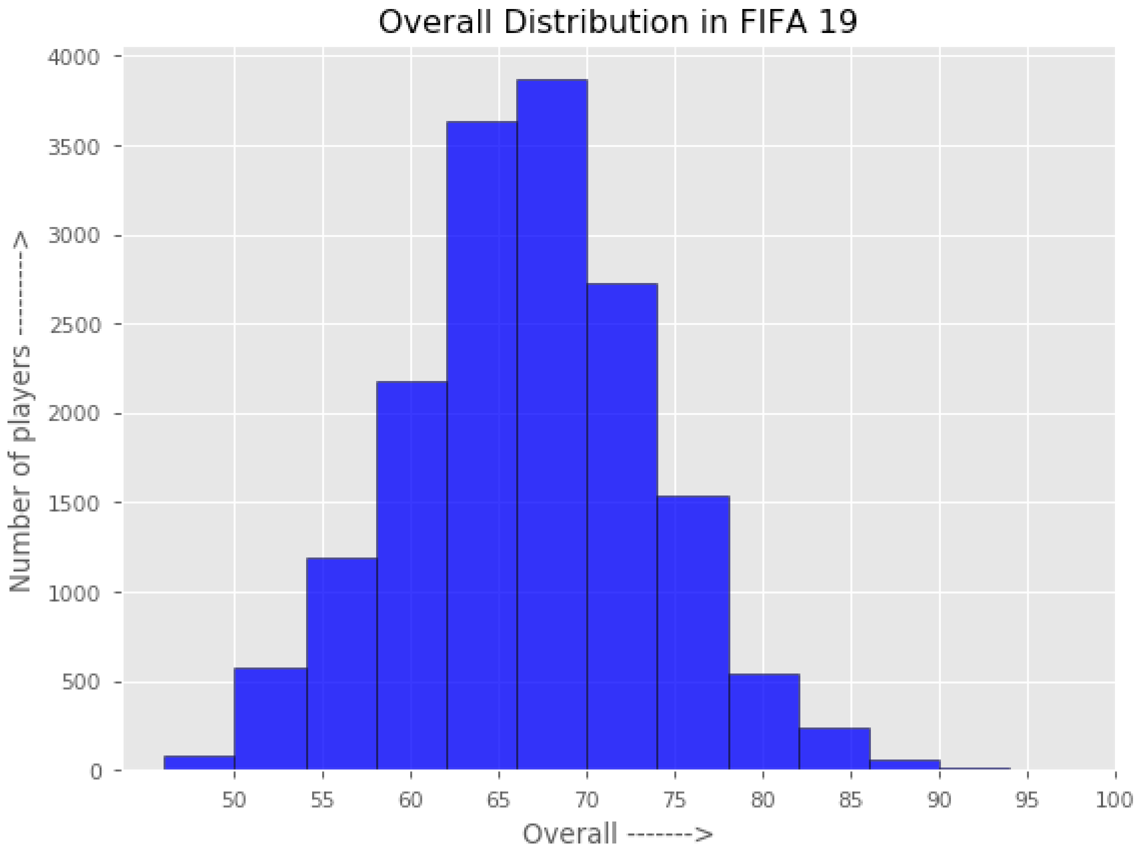
we will begin our analysis of the dataset by having a look at the distribution of the dataset. Football is a multi-national sport and is played in over 200 countries. So, naturally a large number of countries are represented in our FIFA 19 Player Dataset.First, we find out the Count of Players by top 10 nationality. Then, We find out the countries which have the most number of players in the game.





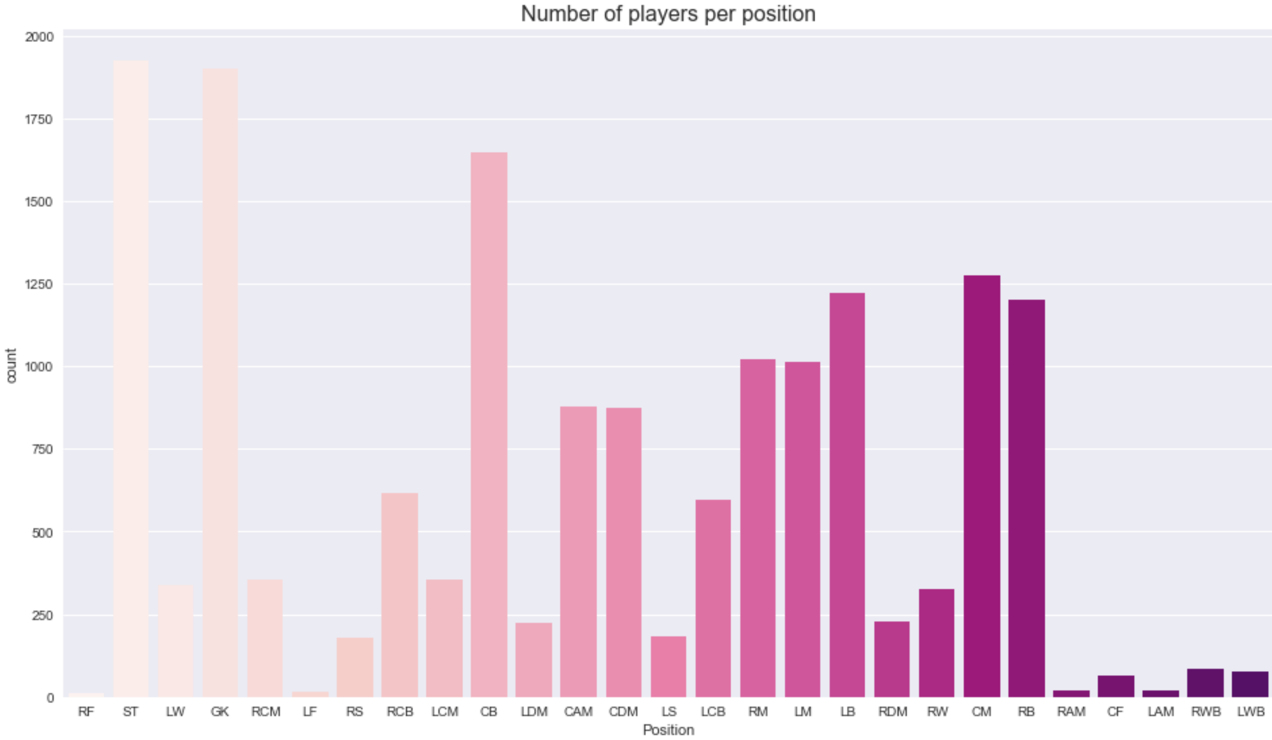
We notice from the bar graph that England is represented the most by over 1600 players in the dataset, followed by Germany, Spain, Argentina and France. In the pie Chart distribution , more than 50% of players come from popular countries like England, Germany, Spain, Argentina and France. This could be explained by the popularity and sizes of domestic leagues within these nations. There are some causes behind this and one of them is FIFA 19 contains upto three divisions of English football. Hence, a large number of English players find their place in the game. Similarly, two divisions of the German league and the Spanish league are also present in the game and hence these players are found in good numbers.

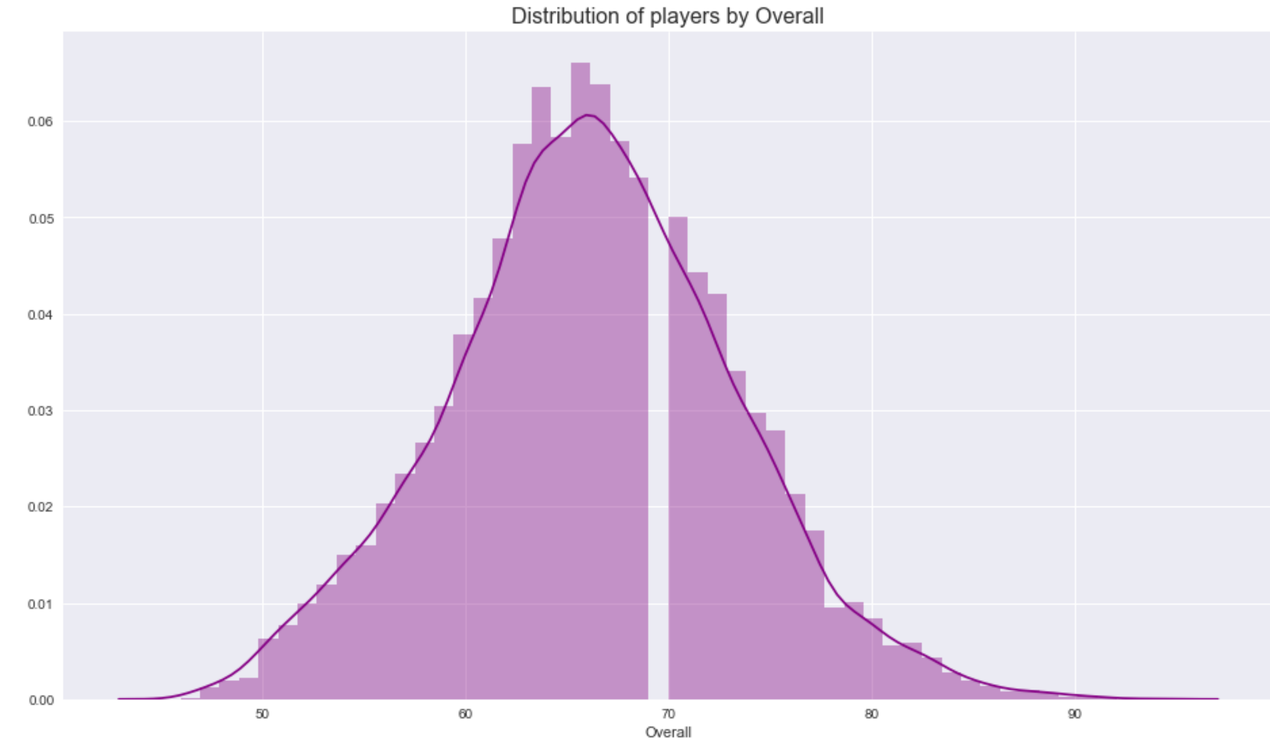
Let us now have a look at how player ratings are distributed in FIFA 19.



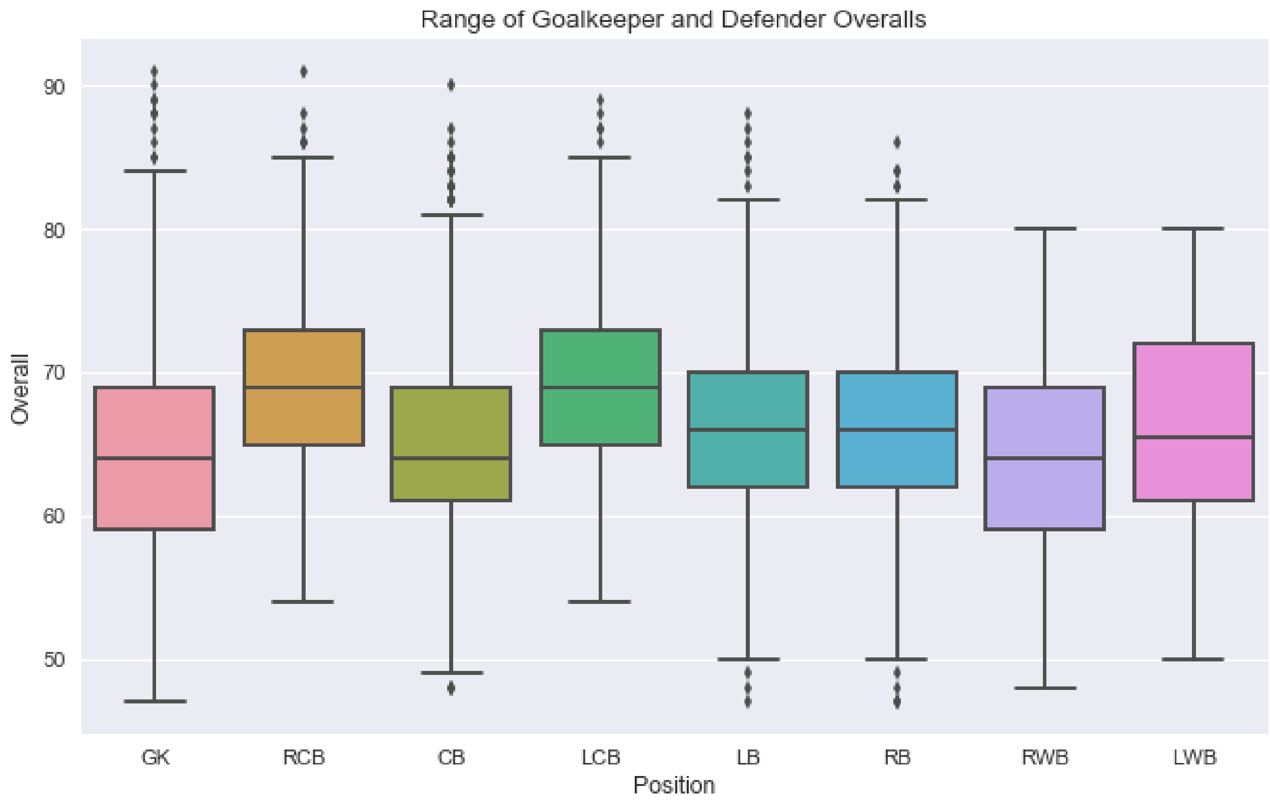
FIFA 19 has a fairly pyramidal distribution of overall with the highest number of players being concentrated in the range between 65 and 70. It is not surprising that only a few players have an overall rating of more than 90. These include the very best footballers of the planet including the likes of Cristiano Ronaldo and Lionel Messi.

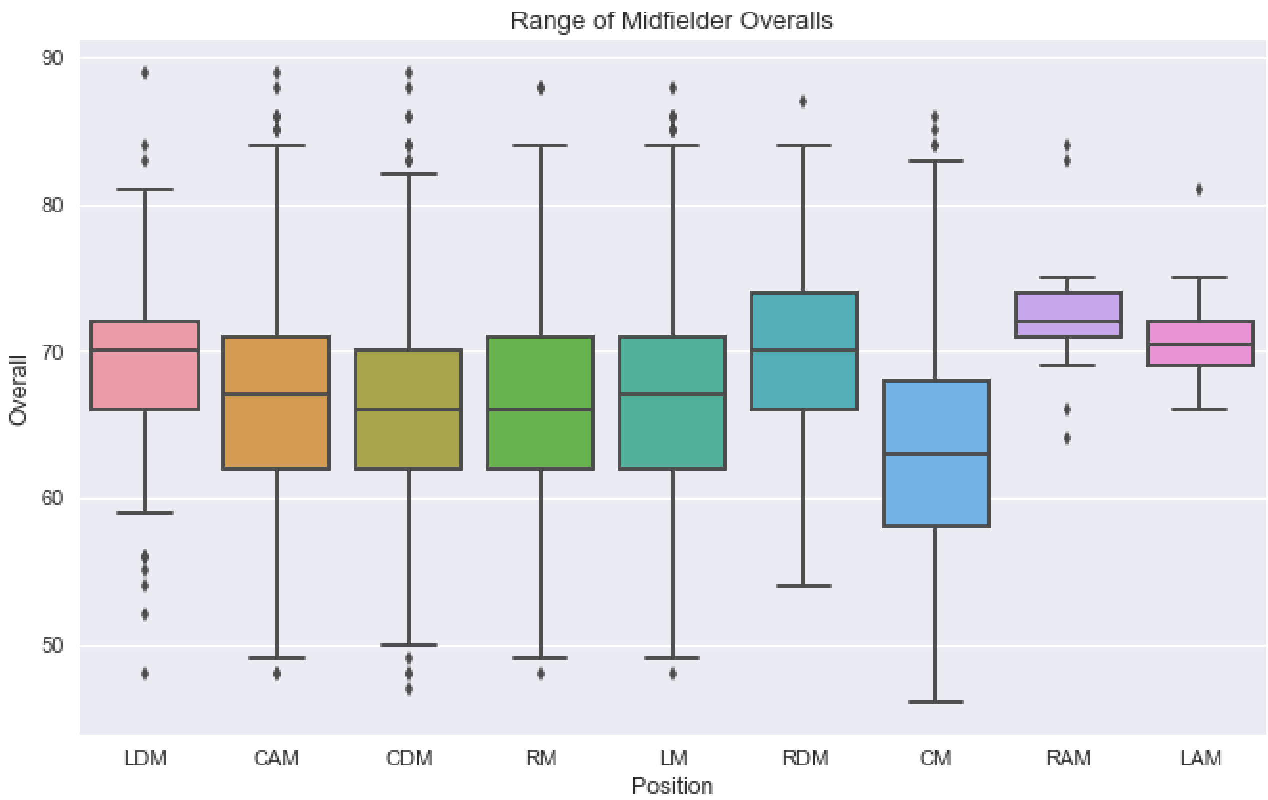
Now, Let us see the Count of players by position & Distribution of players by overall

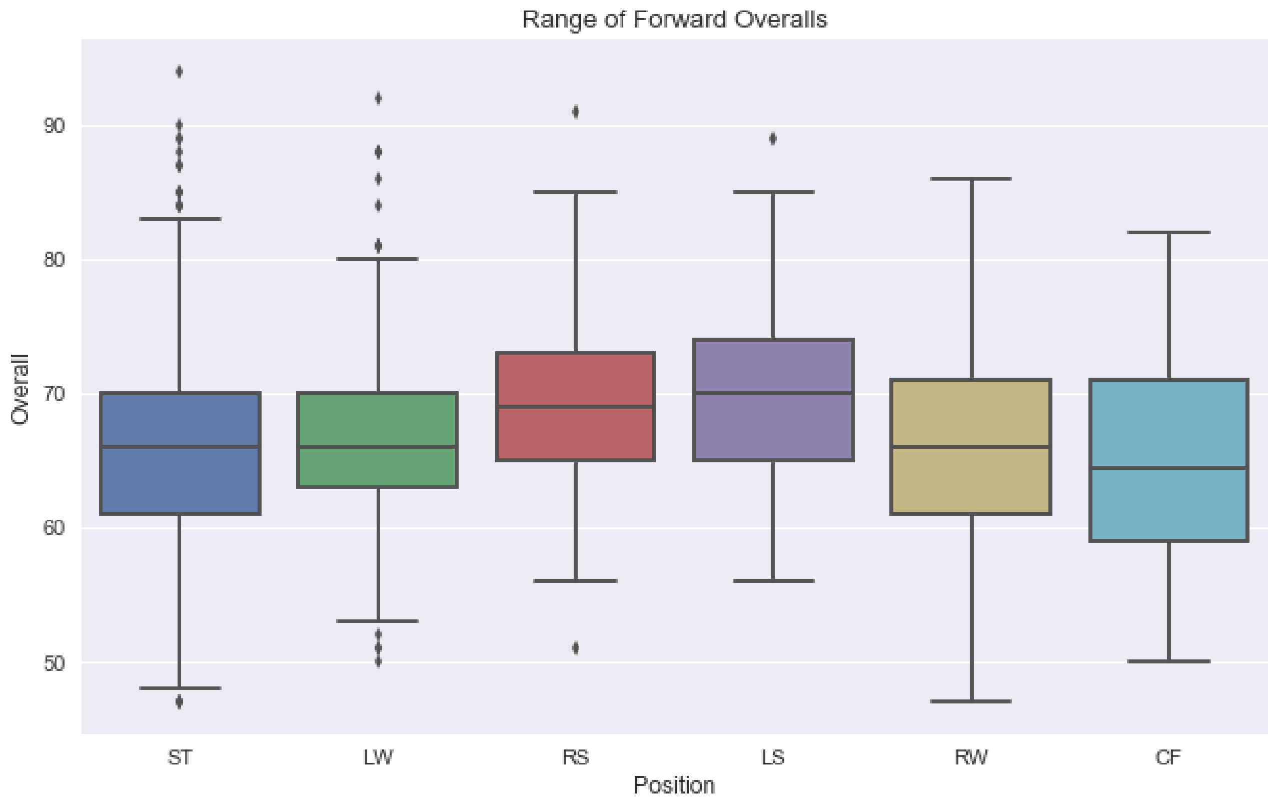




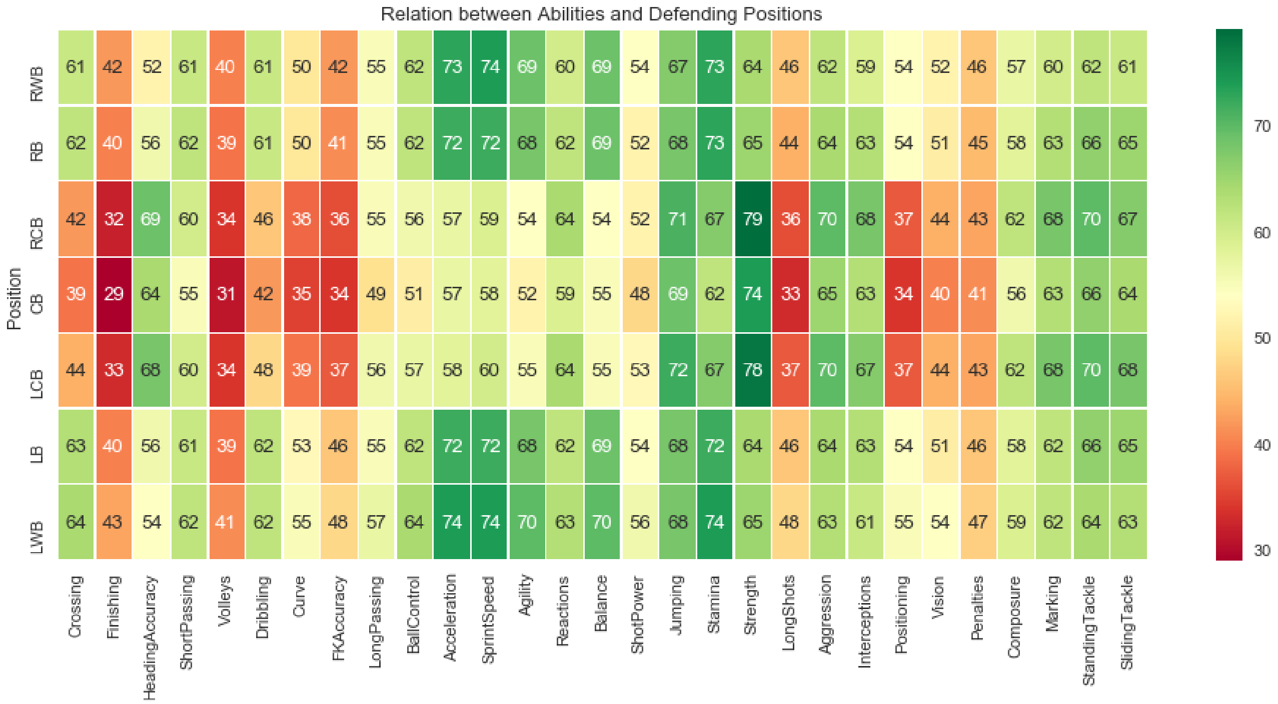
Let us now have a look at how the overall rating of footballers ranges in accordance to their playing position. We will group the positions into three main positions of a football i.e. Defender, Midfielder and Forward, and include goalkeepers inside the defenders grouping.

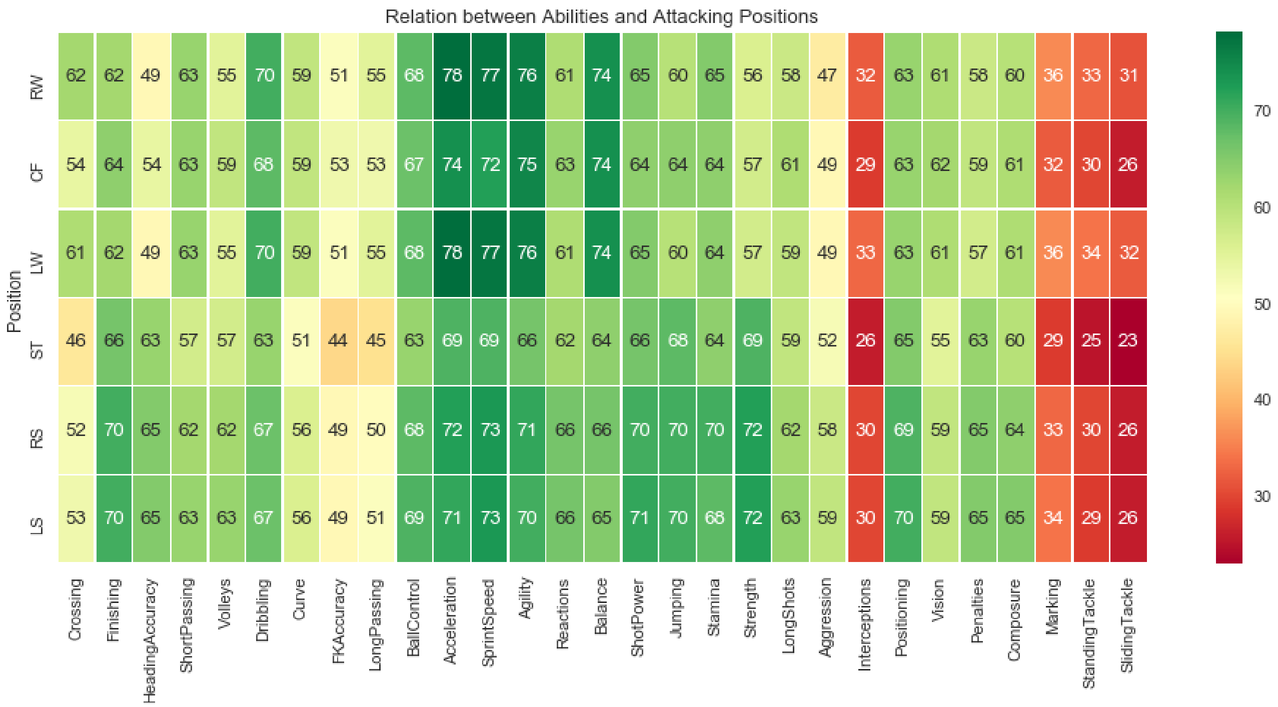

Range of Goalkeeper and Defender Overalls
90
80
70
60
50
Overall
GK
RCB
CB
LCB
LB
RB
RWB
LWB
Position

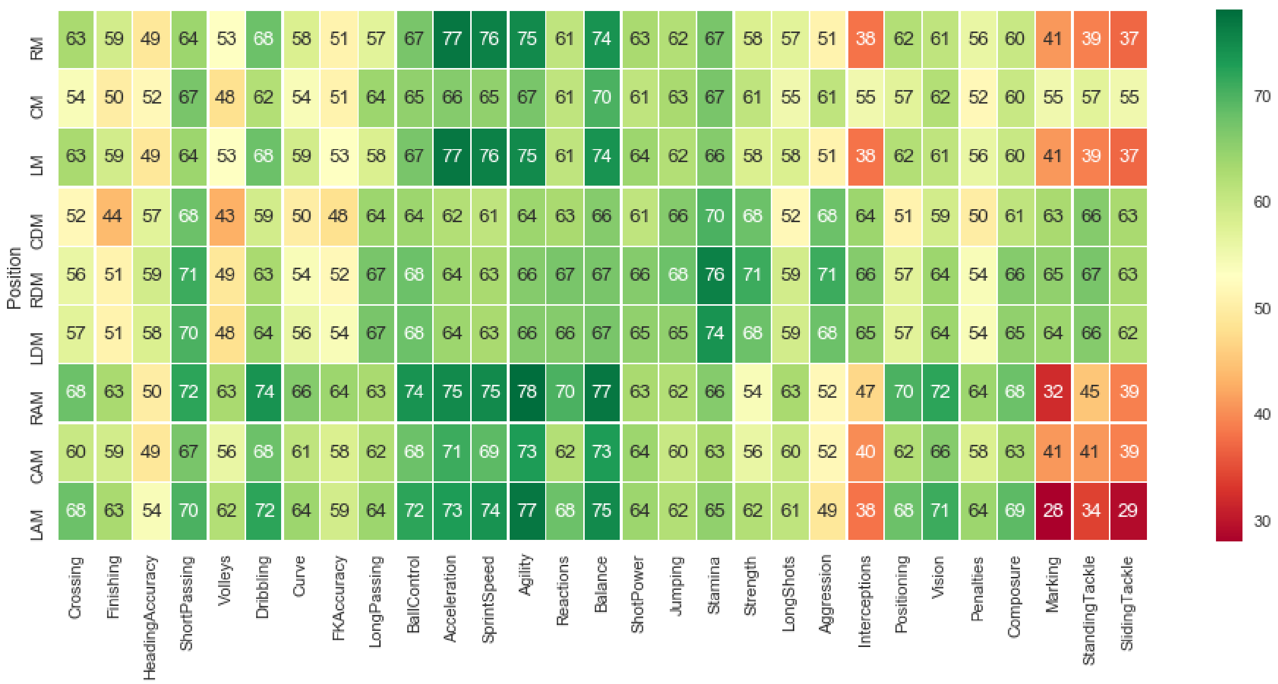

Range of Midfielder Overalls
90
80
70
60
50
Overall
LDM
CAM
CDM
RM
LM
RDM
CM
RAM
LAM
Position



Amongst the defenders, certain positions like CB and LB has a large amount of outliers which indicate a large range in overall values. Even goalkeepers have a large range of such values. Among the midfielders, positions like CDM have a good amount of outliers. LDM have a small interquartile range but have a large amount of outliers on the minimum side. ST also has a large amount of outliers on the maximum side. Certain positions like RAM and LAM have a small interquartile distance with low amount of outliers. This is primarily due to lack of records for such positions.We will now have a look at whether the position of a footballer has an adverse impact on his ability.


Relation between Abilities and Defending Positions
Position
RWB
RB
RCB
CB
LCB
LB
LWB
Crossing
Finishing
HeadingAccuracy
ShortPassing
Volleys
Dribbling
Curve
FKAccuracy
LongPassing
BallControl
Acceleration
SprintSpeed
Agility
Reactions
Balance
ShotPower
Jumping
Stamina
Strength
LongShots
Aggression
Interceptions
Positioning
Vision
Penalties
Composure
Marking
StandingTackle
SlidingTackle


Relation between Abilities and Attacking Positions
Position
RW
CF
LW
ST
RS
LS
Crossing
Finishing
HeadingAccuracy
ShortPassing
Volleys
Dribbling
Curve
FKAccuracy
LongPassing
BallControl
Acceleration
SprintSpeed
Agility
Reactions
Balance
ShotPower
Jumping
Stamina
Strength
LongShots
Aggression
Interceptions
Positioning
Vision
Penalties
Composure
Marking
StandingTackle
SlidingTackle


Position
RM
CM
LM
CDM
RDM
LDM
RAM
CAM
LAM
Crossing
Finishing
HeadingAccuracy
ShortPassing
Volleys
Dribbling
Curve
FKAccuracy
LongPassing
BallControl
Acceleration
SprintSpeed
Agility
Reactions
Balance
ShotPower
Jumping
Stamina
Strength
LongShots
Aggression
Interceptions
Positioning
Vision
Penalties
Composure
Marking
StandingTackle
SlidingTackle

A correlation heatmap is a great visualization to depict the dependency between different features in a dataframe. Amongst the defending positions, we notice that attributes like Sprint Speed, Acceleration and Stamina are the differentiators amongst players in RWB, RB, LWB and LB positions. This is expected since these players have to run across the flanks in order to help their side in both defence and attack. Some parameters like Balance and Stamina are maximum in players playing in central midfield positions. Attacking players like Forwards and Strikers have good Acceleration, Stamina, Sprint Speed, Strength and Shot Power among others. It is natural that these players are poor in defensive attributes like Marking, Standing Tackle and Sliding Tackle.

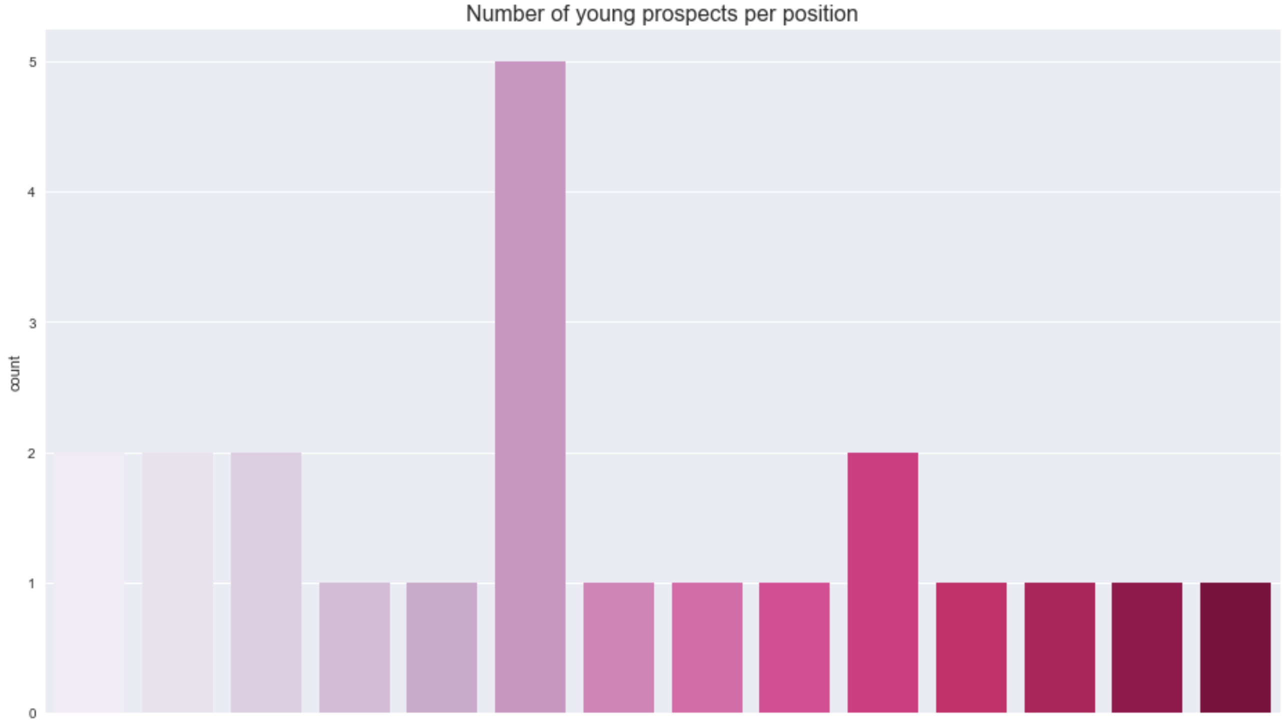
2. Youth prospects with high potential growth

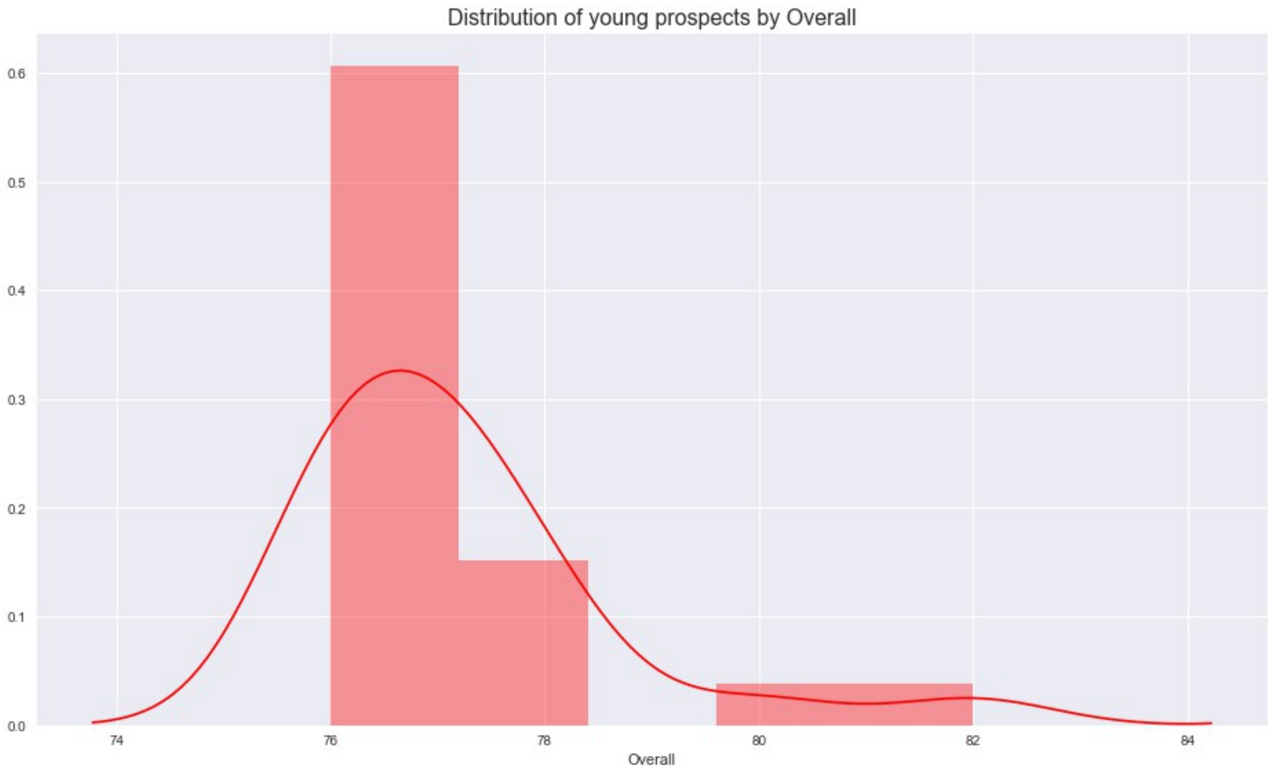
Young players are the most crucial asset for any club for the future. Let us start observing the trends by first start to create the a new dataset on young players.

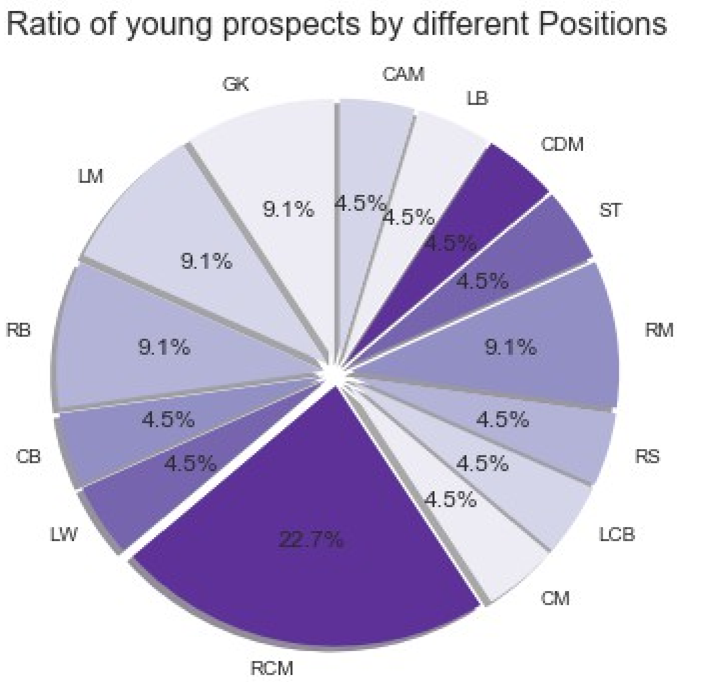
S>No.	Name	Club	Age	Overall	Potential	Position	Value
229	G. Donnarumma	Milan	19	82	93	GK	29.0
415	H. Aouar	Olympique Lyonnais	20	80	90	LM	23.5
734	A. Lafont	Fiorentina	19	78	90	GK	14.0
735	T. Alexander-Arnold	Liverpool	19	78	88	RB	14.0
744	D. Calabria	Milan	21	78	88	RB	14.5
755	J. Gomez	Liverpool	21	78	88	CB	14.5
1143	Vinícius Júnior	Real Madrid	17	77	92	LW	17.5
1172	N. Barella	Cagliari	21	77	89	RCM	15.5
1156	A. Diawara	Napoli	20	77	87	CM	14.0
1149	R. Bentancur	Juventus	21	77	89	RCM	15.5
1110	D. Upamecano	RB Leipzig	19	77	88	LCB	13.0
1070	P. Cutrone	Milan	20	77	89	RS	16.0
1029	C. Ünder	Roma	20	77	87	LM	14.5
1004	J. Sancho	Borussia Dortmund	18	77	89	RM	14.5

S>No.	Name	Club	Age	Overall	Potential	Position	Value
1231	Wendel	Sporting CP	20	76	86	RCM	12.5
1312	Gedson Fernandes	SL Benfica	19	76	86	RCM	12.0
1343	I. Sarr	Stade Rennais FC	20	76	87	RM	13.5
1422	B. Embolo	FC Schalke 04	21	76	86	ST	12.5
1450	Marc Roca	RCD Espanyol	21	76	87	CDM	11.5
1479	K. Tierney	Celtic	21	76	86	LB	11.0
1522	A. Ćorić	Roma	21	76	88	CAM	14.0
1564	S. Berge	KRC Genk	20	76	87	RCM	11.5

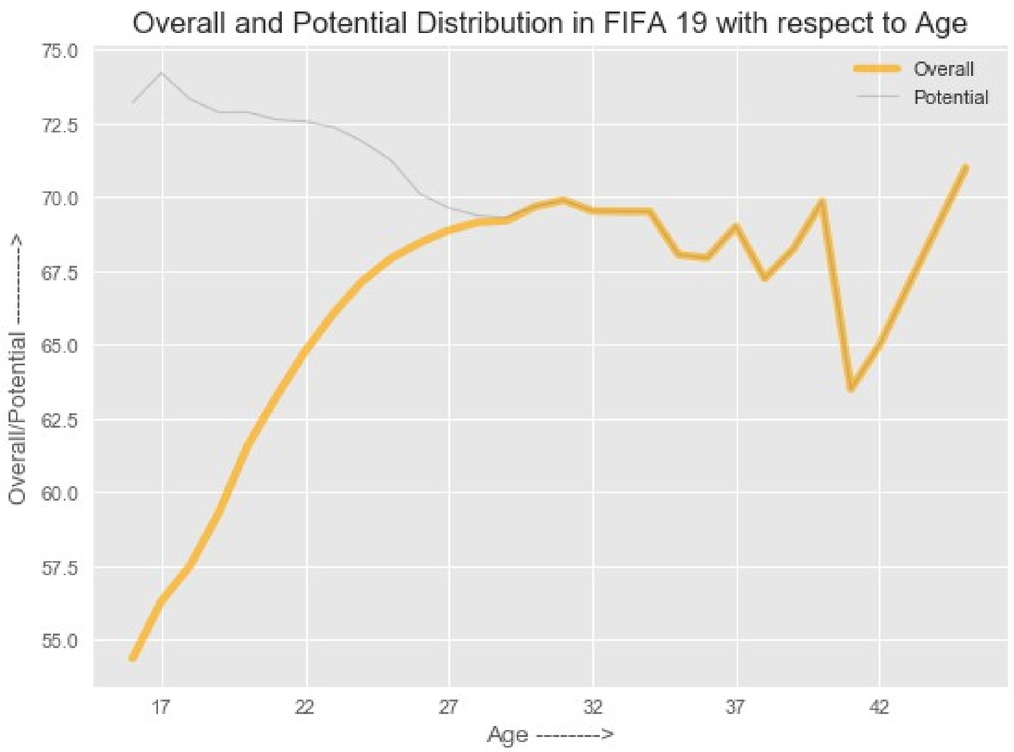
We saw that there are a quite number of youth players with some potential. Now let us see the distribution of youth players according to their positions

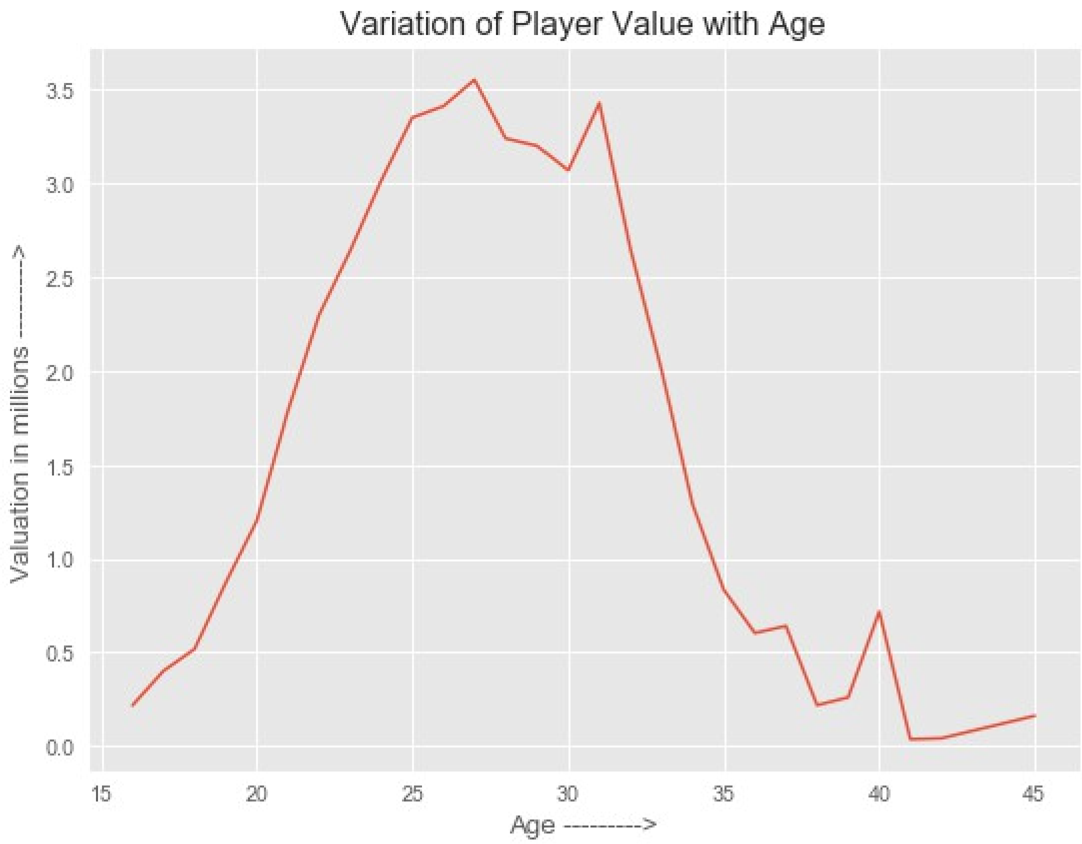




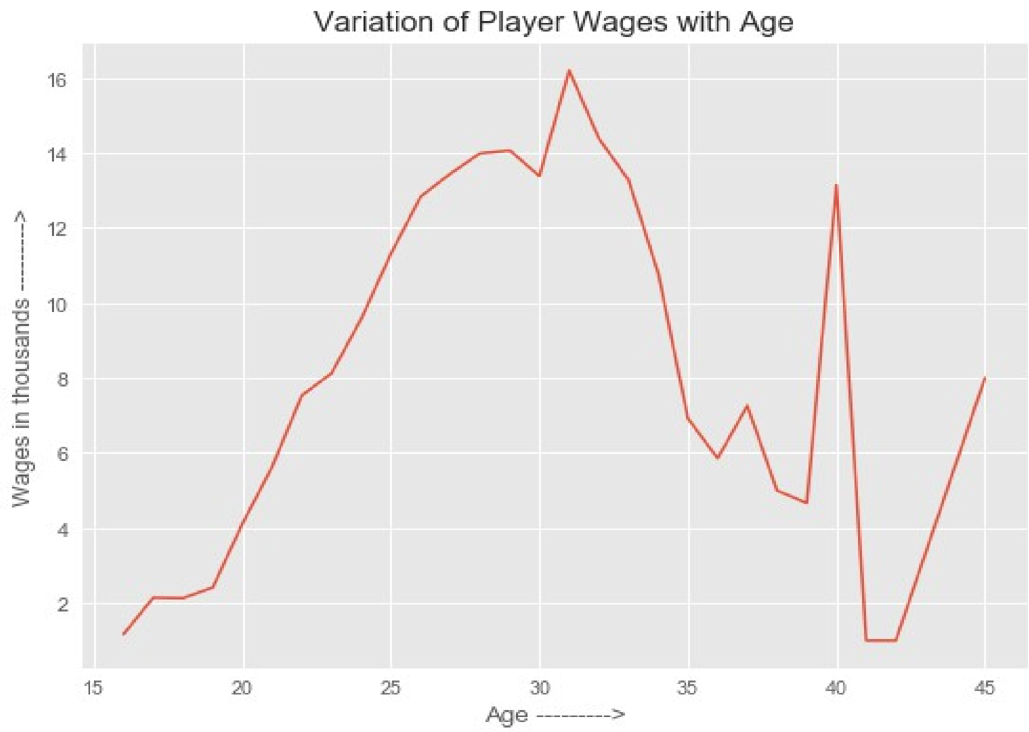


We noticed a few patterns in the data present in the FIFA 19 dataset and this may prove to be very handy for players. One such pattern is the relationship between Overall Rating and Age.
We notice that the overall of a player increases as he ages. This continues till about the age of 30 or 31 when the player is at his prime. Post 30, a player starts to decline in performance. We also see that the actual overall and the potential of a player in FIFA 19 come to equal terms at about the age of 29. The potential of a player is a reflection of his overall rating if he reaches his prime.
Now, let us have a look at how age affects the valuation of a footballer.

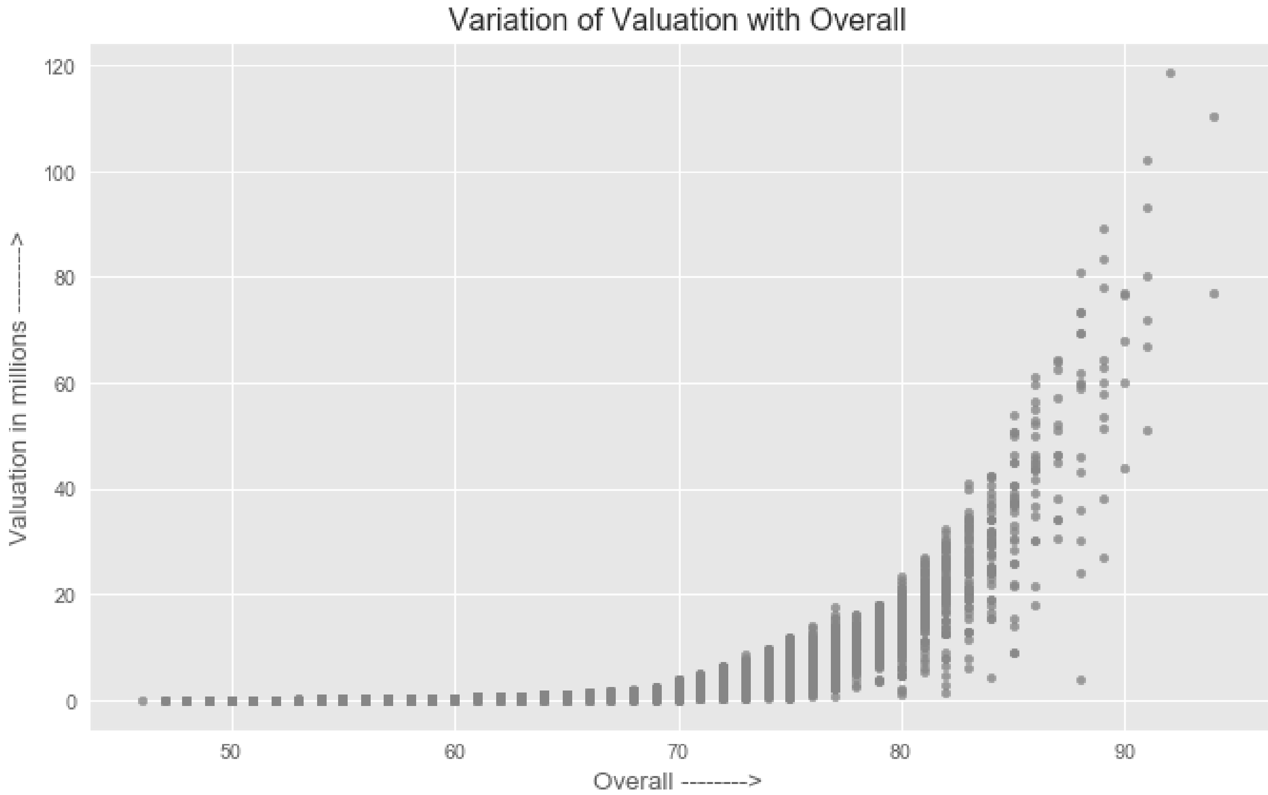




Unsurprisingly, the value of a player also increases with age and reaches a maximum at around the age of 31. After 31, the value of a player starts dwindling and hence it is important for future FIFA 19 managers to note this. The relationship between player wages and age also follows a similar pattern as depicted below:



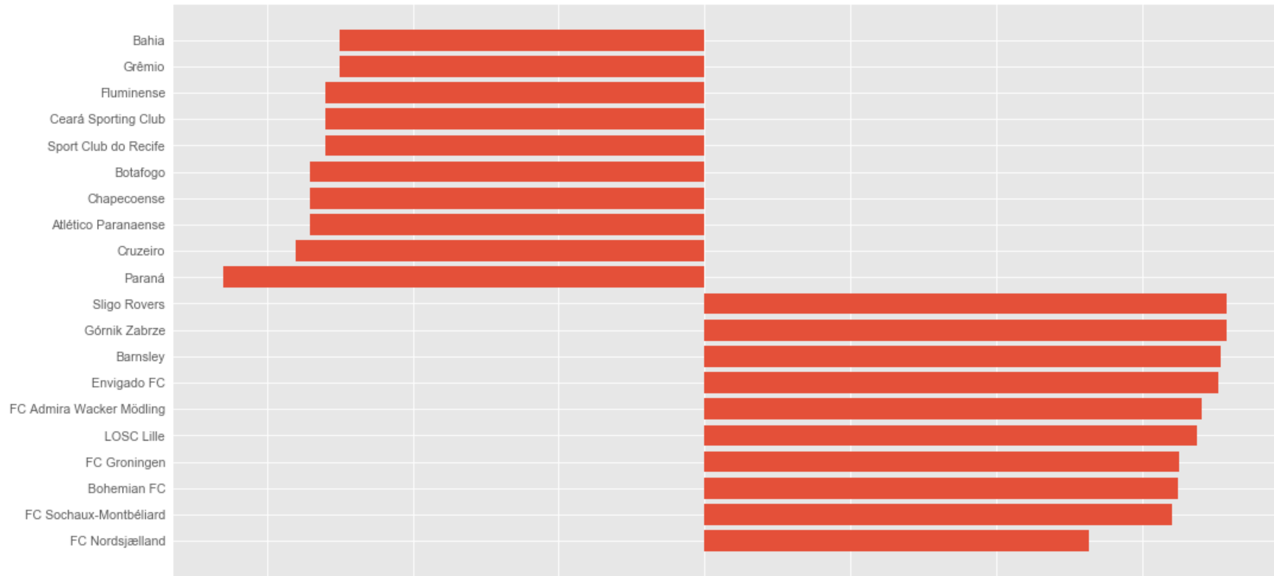
Next up, we will have a look at how overall ratings affect the valuation of a footballer with the help of a scatter plot.



It is easily noticeable that the valuation of a footballer increases with his overall rating.

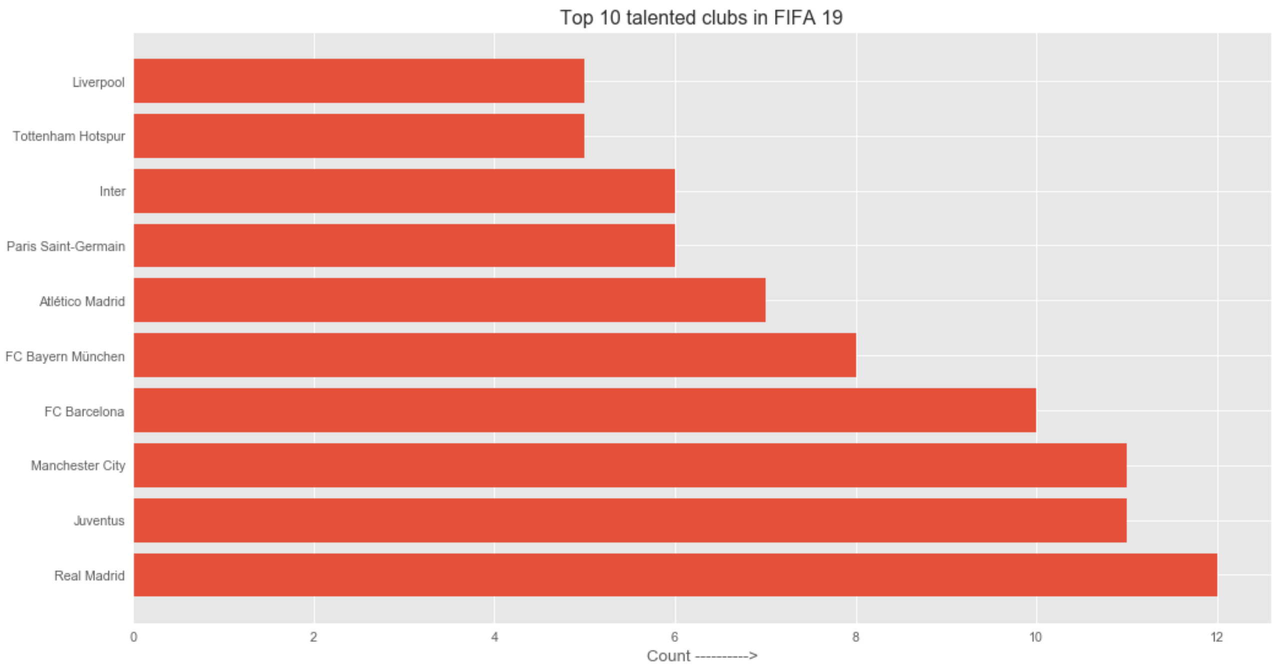
Playing Manager Mode

The Manager Mode of FIFA19 is one of the most popular features of the game among gamers. It puts a gamer on the managerial seat of his/her favourite football club. Different managers have different preferences when it comes to selecting the football club he/she will manage. Some players prefer to start off with a club having a young squad, thus having more scope of developing footballers. Here, we have a look at favourable clubs for these managers and the clubs he/she will stay away from.

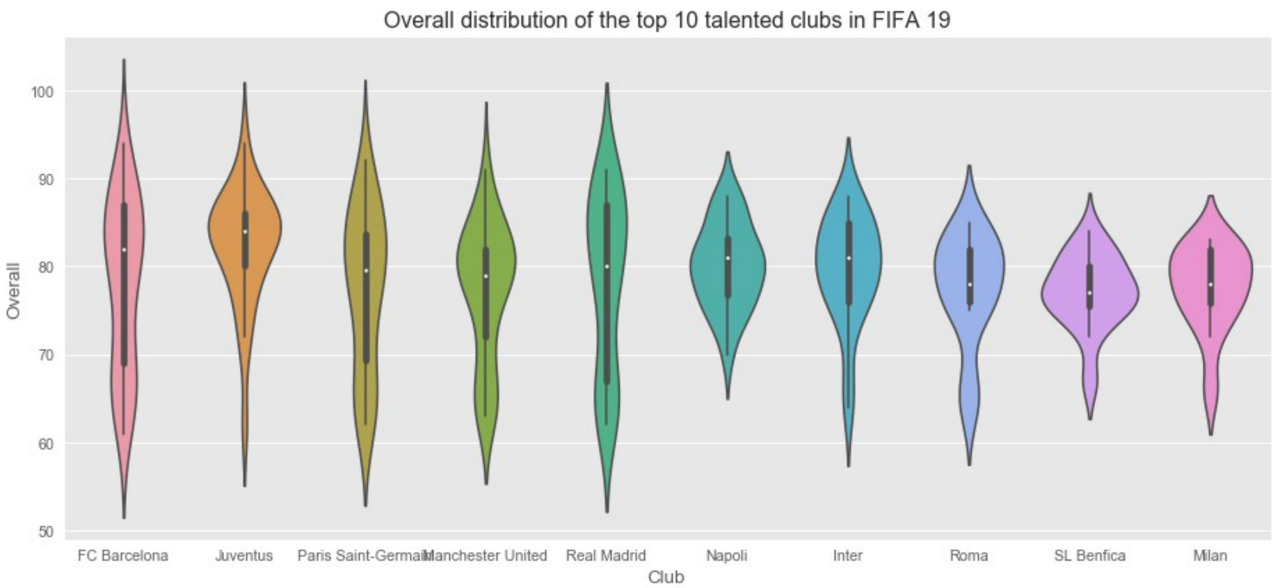


FC Nordsjaelland of Denmark has the youngest squad in FIFA 19, followed by FC Sochaux of France and Bohemian FC of Ireland. In general, South American clubs have the oldest squads in the game with all of the 10 such clubs being from South America.

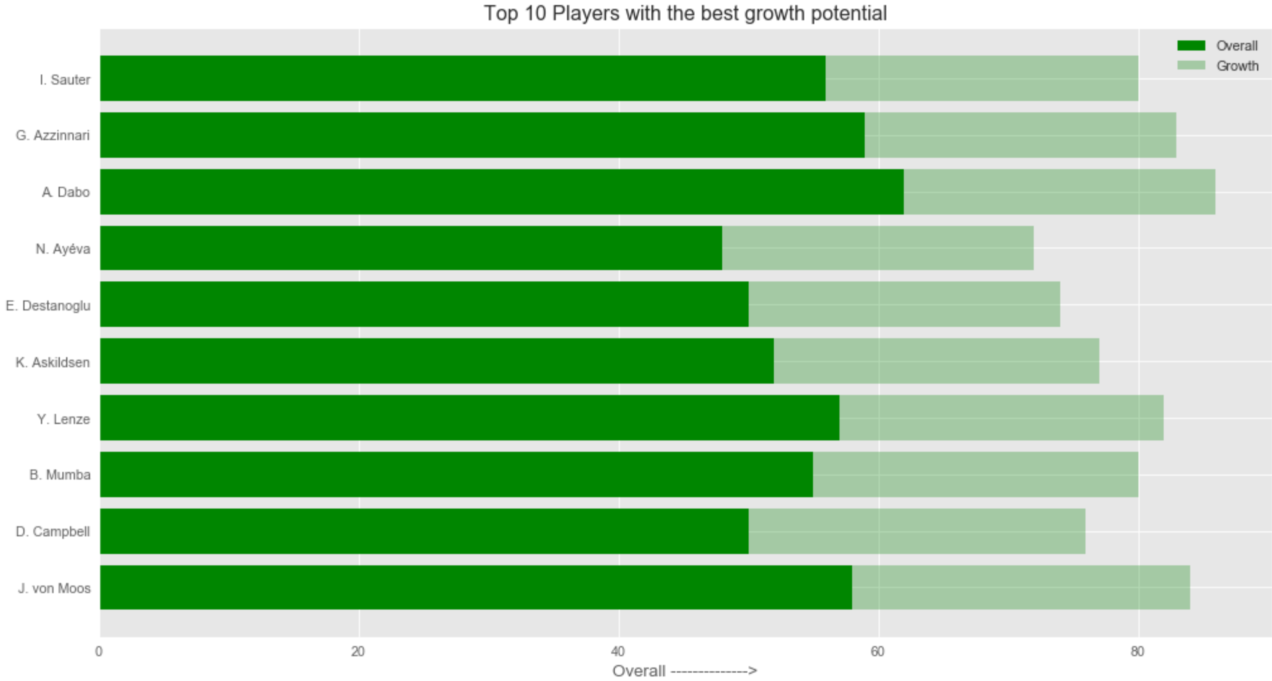
Some managers prefer to have the most talented players in their disposal. Hence, such managers will prefer managing clubs with the most talents. But how do we define a talent? Is it the number of superstars in a squad? Or does it mean having a higher average player OVR rating? In this section, we will try to answer both the above questions. If we define talent as a footballer having a rating of above 85, we get the following visualization.



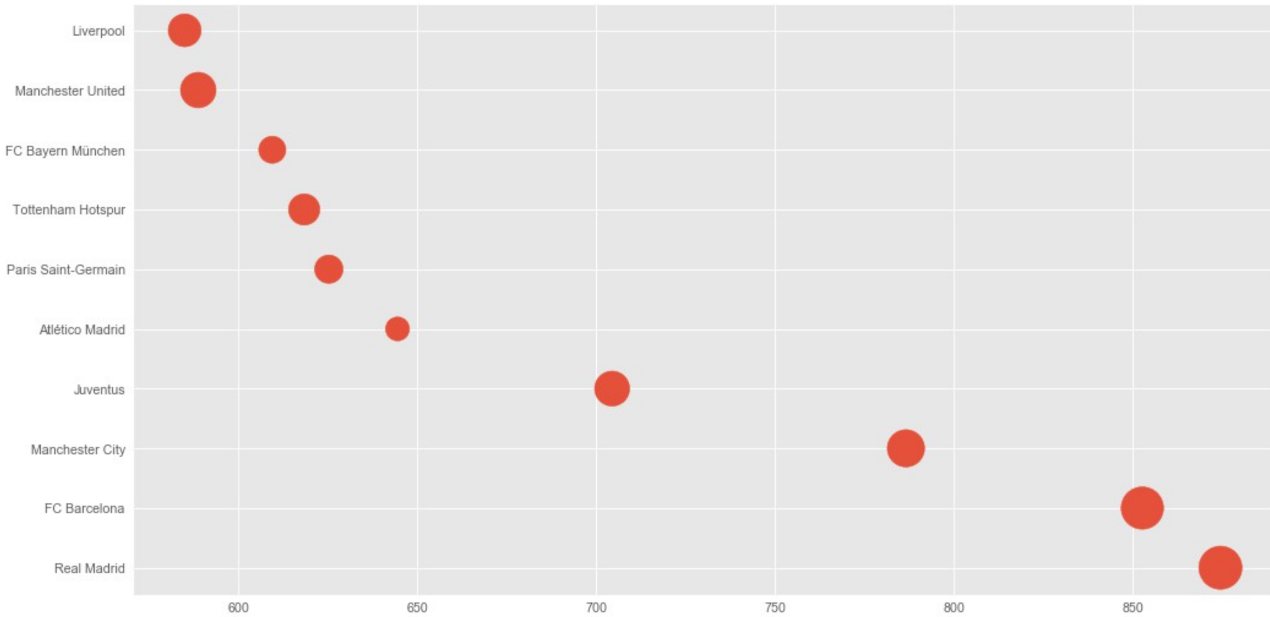
Spanish giant Real Madrid have the most footballers with an overall rating of above 85. They are followed by Juventus of Italy and Manchester City of England. The top 10 clubs include three each from Spain and England, followed by two from Italy (Juventus and Inter Milan). PSG and Bayern Munich make up the other two.



Now if we were to define the most talented club as the one having the maximum average player overall rating, then these are the top 10 clubs. We find that clubs like Napoli, Roma, Benfica and Milan have entered the mix here. It is interesting to note that four Italian clubs figure in this list. Serie A is known for having sturdy defenders and hence, managers to like to play attacking football may find managing one of these four clubs very interesting. Managers also need to be aware of young starlets in the transfer market. Here, we look at the youngest players to have the most to gain in terms of achieving their potential. Most of these players come at a low fee and demand low wages to make life easier for the managers.



We notice that A.Dabo has the maximum potential. Dabo currently has an OVR of just over 60 and has the potential to reach around 92-93. He is closely followed by Y.Lenze and J.von Moos. Managers will surely love to keep an eye on these players.
Some managers like to manage clubs with the maximum valuation. This allows them to sell some of the existing players of the club in order to make money for buying players which best fit their playing style. Here are the 10 most valuable clubs in the game.

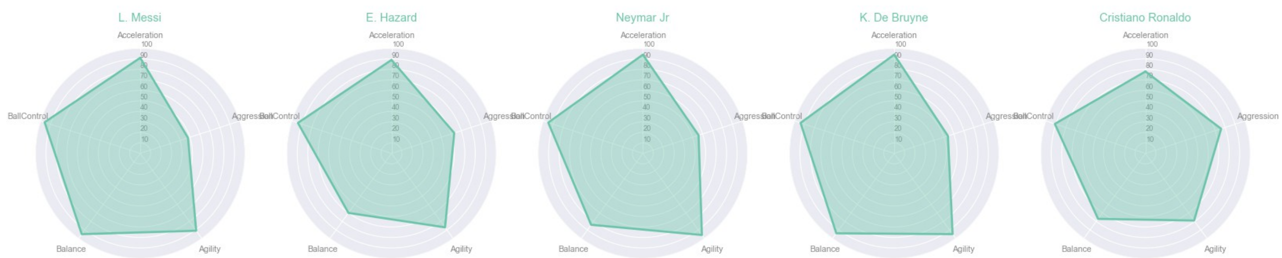


Quite unsurprisingly, Spanish giants Real Madrid and FC Barcelona are the most valuable football clubs in the game. These clubs also pay the most wages to their footballers. It is interesting to note that although Manchester United are 10th in this list, they pay more wages than other clubs in the list like Bayern Munich, PSG and Juventus among others. Atletico Madrid are one of the most well managed clubs. Although they are the fifth most valuable club, they pay the lowest wages among these clubs.

We also provided mechanisms for scouting players for best positions.For example, The top 10 players for RM position the field are

Name	Nationality	Club	Overall
L. Messi	Argentina	FC Barcelona	94
E. Hazard	Belgium	Chelsea	91
Neymar Jr	Brazil	Paris Saint-Germain	92
K. De Bruyne	Belgium	Manchester City	91
Cristiano Ronaldo	Portugal	Juventus	94
A. Griezmann	France	Atlético Madrid	89
L. Modrić	Croatia	Real Madrid	91
L. Insigne	Italy	Napoli	88
M. Salah	Egypt	Liverpool	88
K. Mbappé	France	Paris Saint-Germain	88

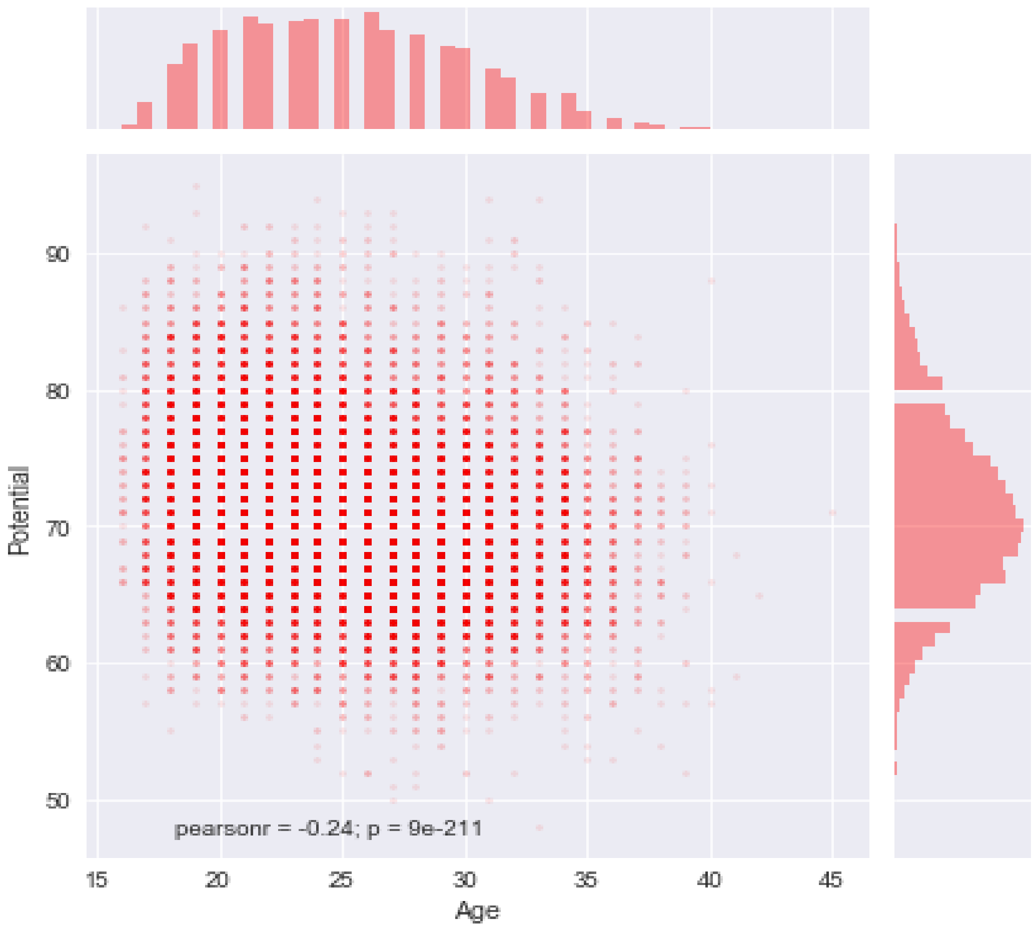
The comparison of top 5 are shown below



Models

We include a **Linear Regression** Model to Predict Player Valuation and got an r2 score of 92.926% and and RMSE of 1.85. From **Principal Component Analysis** we observed that **Overall** , **Age** and **Reactions** are the most important factor defining a player's characteristics.We also observed few trends in player's characteristics . For ex-

- Potential tends to fall as the player grows old



we used **Random Forest Regression** also to predict the Player Valuation and got the accuracy rate of **74.94%** and MAE 0f **0.45** and through **RandomizedSearch** , we got an accuracy of **75.01%**. Through **Nearest Neighbours** model we predicted the similar players for a specific positon.For ex-

- These are 3 players similar to **E. Hazard** :
  Name: **Neymar Jr** Position: **LW**
  Name: **P. Dybala**   Position: **LF**
  Name: **Ronaldo** Cabrais Position: **RW**

Future Scope

This Project focused mainly on representation as well as predicting the player's characteristics. Further more development is possible , Like Building a **Logistic Regression** model to predict player valuation more precisely. More new Features can be added like handling budget,player transfers , etc.

References

Dataset : www.kaggle.com