Domain: Sport

Datasets:

1. FIFA 18 dataset: <https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset>
2. FIFA 19 dataset: <https://www.kaggle.com/karangadiya/fifa19>
3. FIFA 20 dataset: [https://www.kaggle.com/stefanoleone992/fifa-20-complete-player- dataset](https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset)

Problem Statement:

Trying to understand the players population in the FIFA games. To understand the features/variable contributing towards the overall of the player and build a predictive model that could estimate of the overall of the player with accuracy.

Also, build a model to predict the value of the player with accuracy and least numbers of variables.

And build a model to recommend the most similar players to a particular player.

Three datasets are being used here, all three for different reasons.

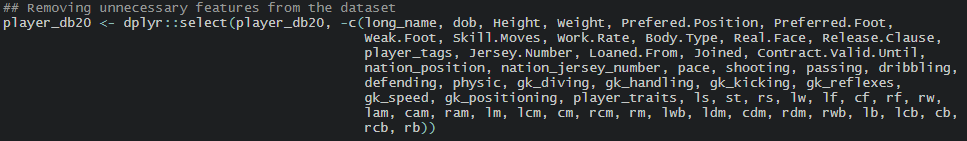
* FIFA 19 dataset is being used for exploration, visualization, build predictive model and the recommendation model.
* FIFA 18 data set is used only as training the data along with FIFA 19 data set.
* FIFA 20 dataset is only being used as the test dataset.

I have tried to make the project modular as such the data cleaning for each dataset is done in a different script and are later sourced to the main script. In the same way all the predictive models built are saved and then loaded to the main script. Scripts for the models are presented separately.

1. Initial Data Cleaning (Using Excel & R)
   1. Removing unnecessary fields (Using R)







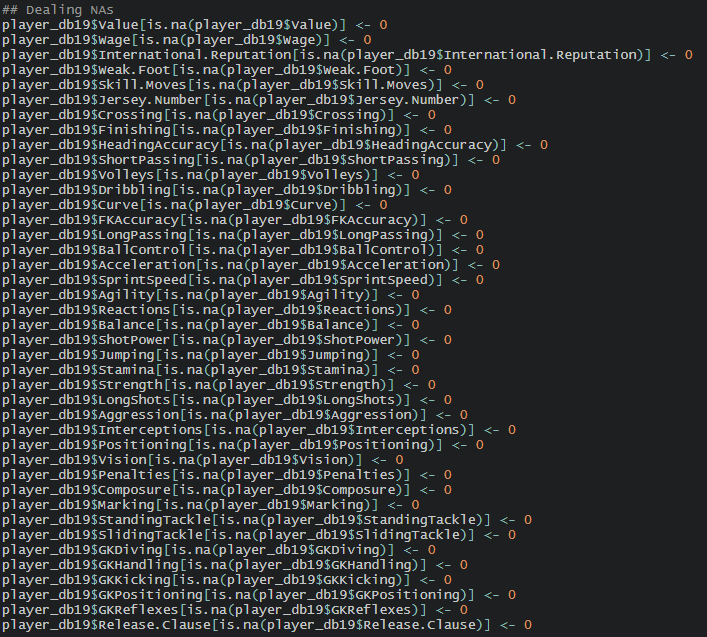
* 1. Simplifying the position into Goal Keeper (GK), Right Back (RB), Centre Back (CB), Left Back (LB), Defensive Midfielder (DM), Centre Midfielders (CM), Wingers & Forwards. (Using Excel)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CAM <- CM | LCB <- CB | RAM <- CM | GK <- GK | LS <- FORWARD | RS <- FORWARD |
| CB <- CB | LCM <- CM | RB <- RB | LAM <- CM | LW <- WINGER | ST <- FORWARD |
| CDM <- CD | LDM <- DM | RCB <- CB | LB <- LB | LWB <- LB |  |
| CF <- FORWARD | LF <- FORWARD | RCM <- CM | RW <- WINGER | RWB <- RB |  |
| CM <- CM | LM <- WINGER | RDM <- DM | RF <- FORWARD | RM <- WINGER |  |

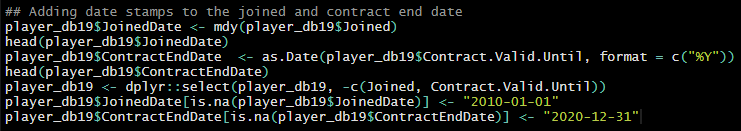
* 1. Dealing with NAs. (Using R)

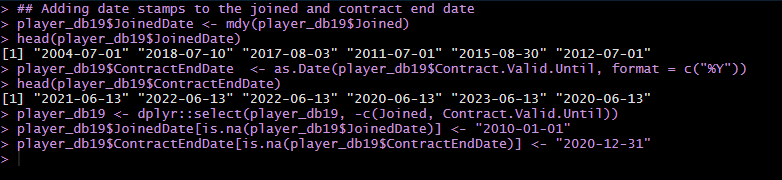




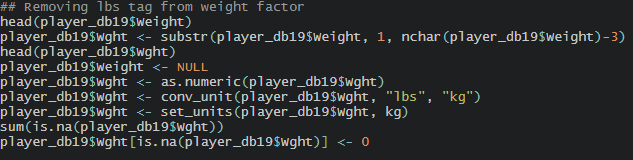


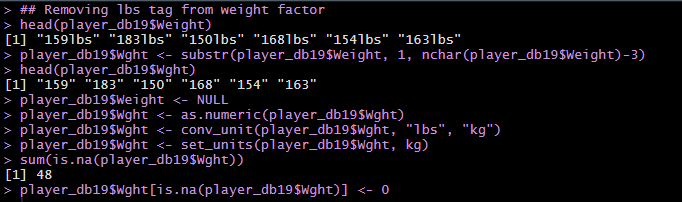
* 1. Changing joined date and contract valid date to time stamps. (Using R)



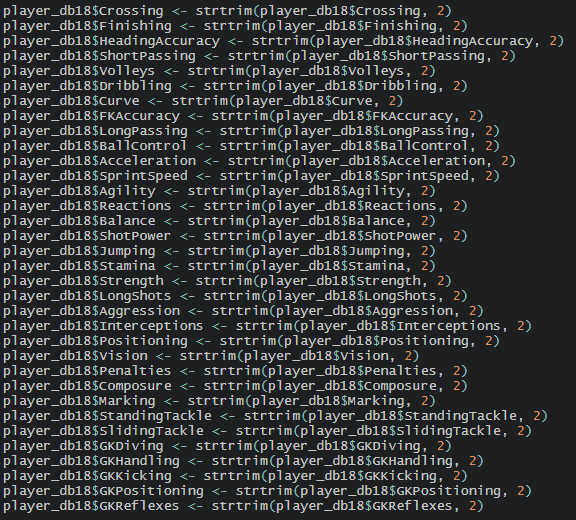


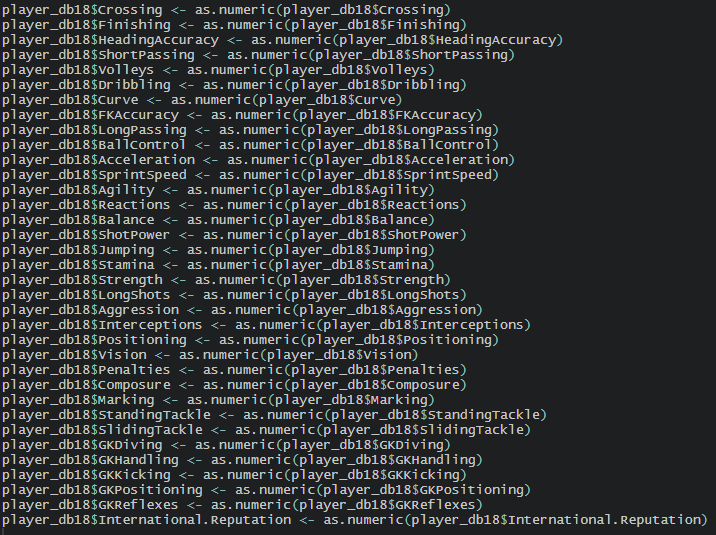
* 1. Removing units from weight feature. (Using R)

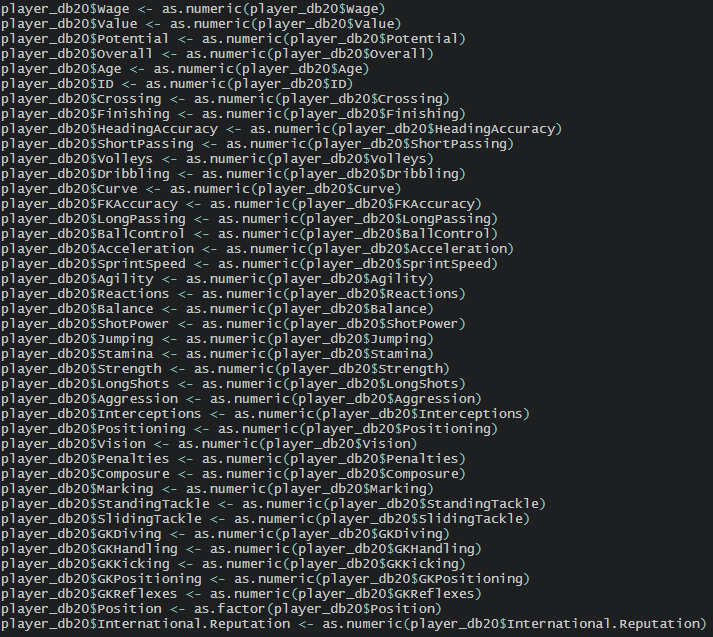




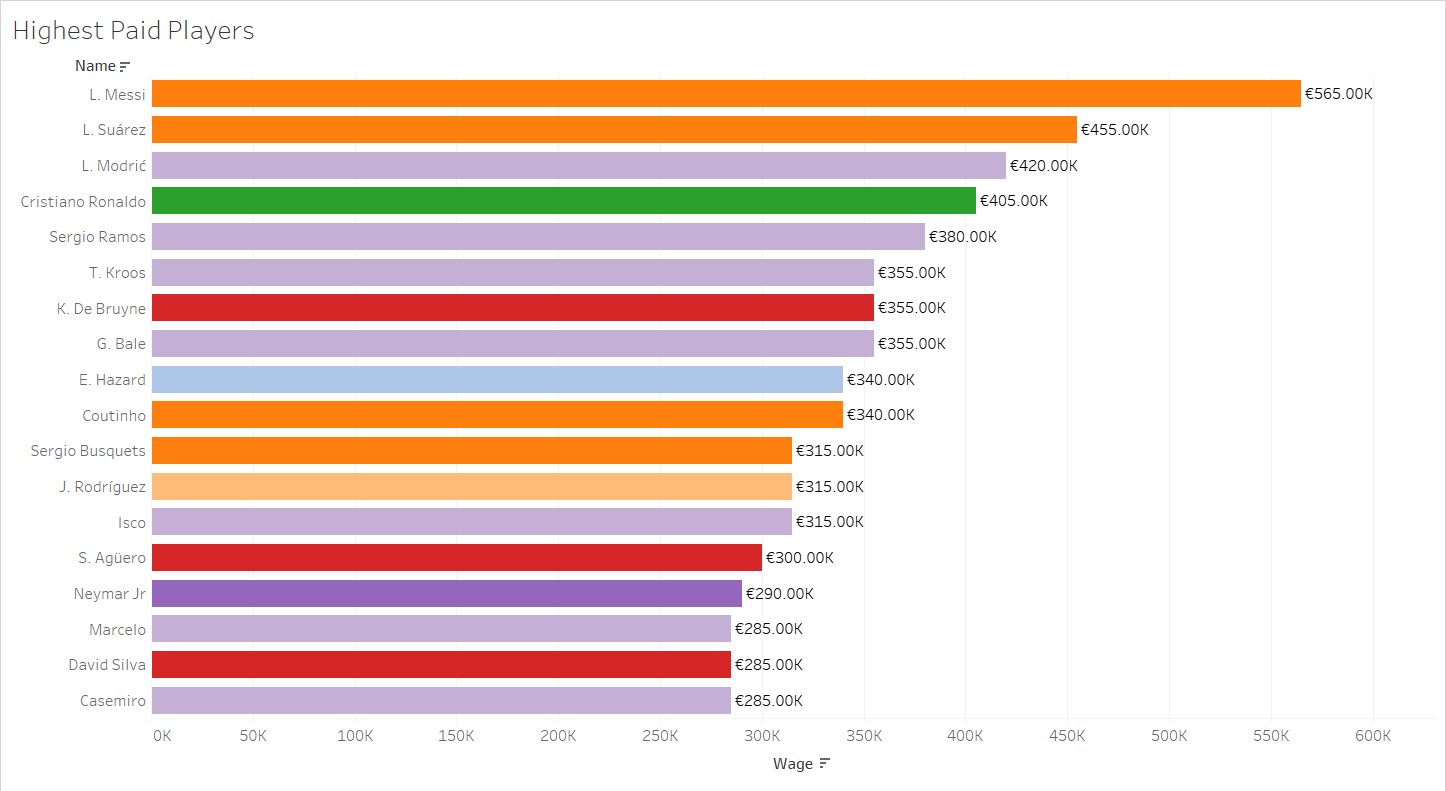
* 1. Changing Value, Wages and Release Clause from human form to real values. (Using Excel (LEFT and LEN function))
  2. Trimming data & changing the data type. (Using R)







1. Exploratory Analysis, answering various questions to learn the characteristics of the population using cleaned FIFA 19 dataset.
   1. Highest Paid Players



Top 18 highest paid players are shown in the graph & and the colors are defined to show the club of the player.

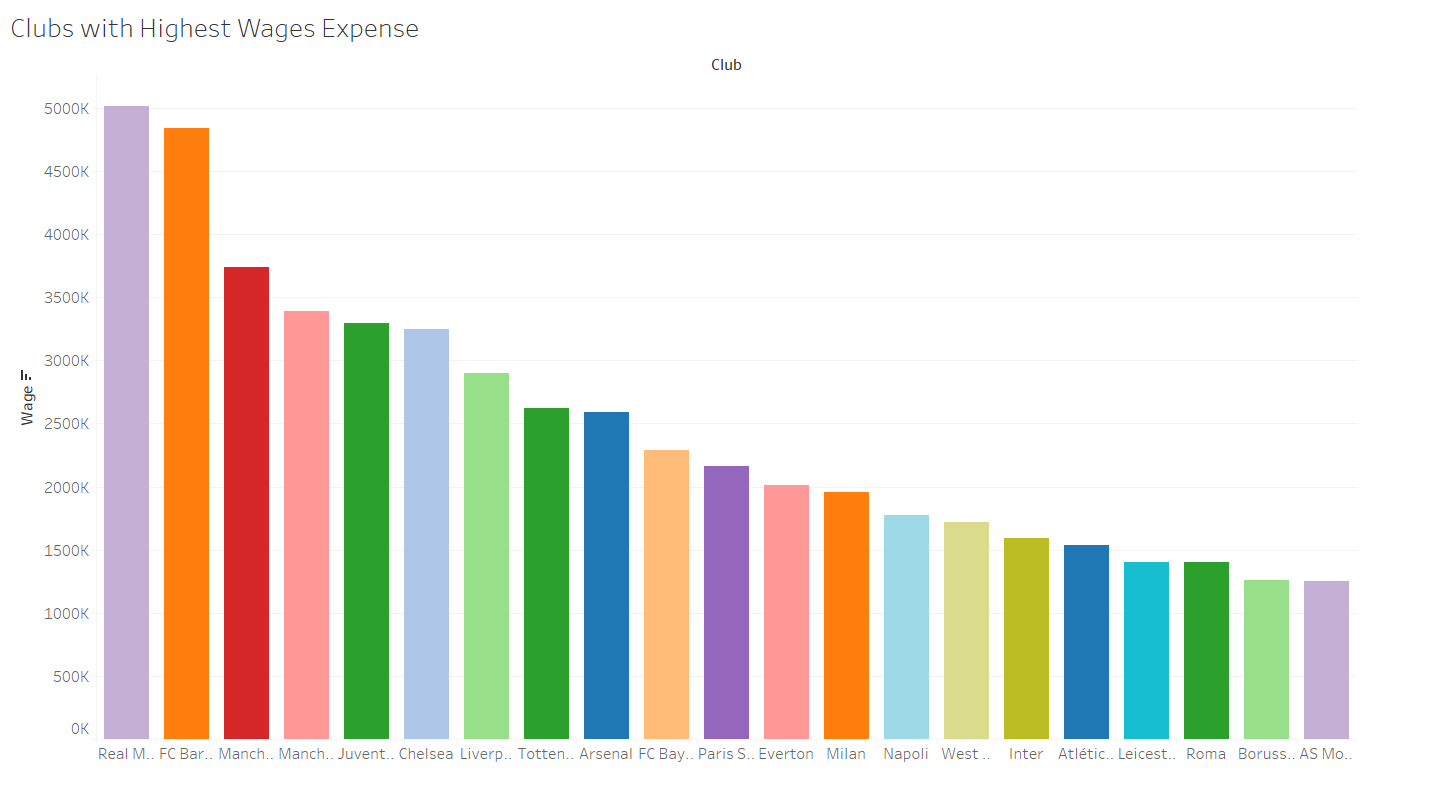
Lionel Messi is the highest paid player of the game, followed by Luis Suarez and Luka Modric.

Next, we see Cristiano Ronaldo, now it might seem weird as the Ronaldo’s overall is same as Messi’s only different being age, they both also play roughly the same position. There might be case to be made that the Barcelona the club for which Messi plays is generally considered to be richer that Juventus.

But I think age is the bigger factor here, I will be analyzing it further.

Also there seems to be a domination of orange (Barcelona) and pink (Real Madrid), so I would like to analyze the total wage expenditure of these clubs.

* 1. Clubs with Highest Wages Expenditure

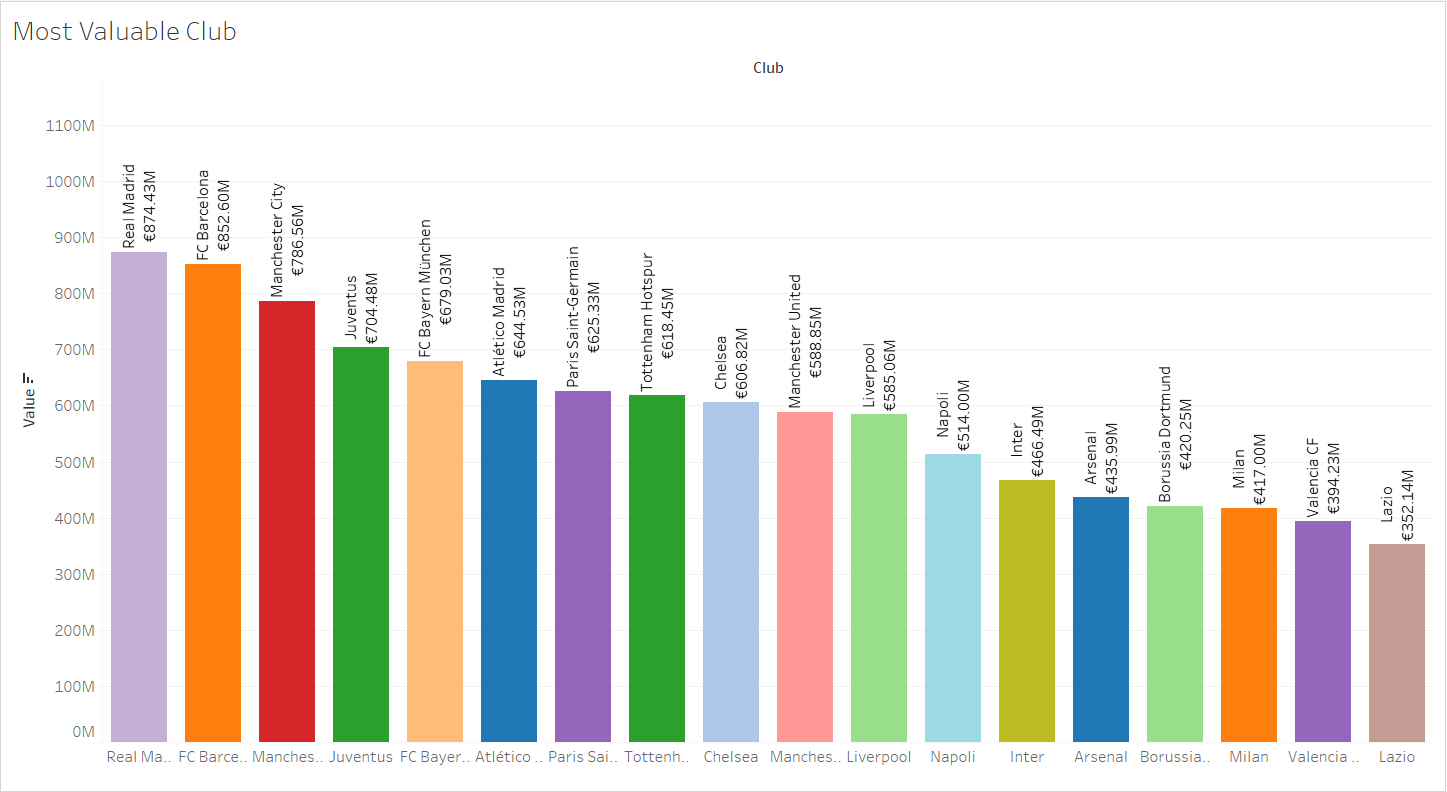


As we saw in the Highest Paid Players graph, it was visible that most of the players there were from Barcelona and Real Madrid. From our current visualization it is again clear that these clubs spend highest for wages.

If we ignore these two clubs from the graph, we can see that there is an incremental increase in wage expenditure. And there seems to be a lot higher increase from third to second place.

Maybe it could have something to do with the value of the player these clubs have. As it can naturally be assumed that the higher valued players would demand higher wages.

* 1. Most Valuable Club



Value of the player here is determined by the summing the values of all their player.

Here it is abundantly clear why the wages of Real Madrid and Barcelona were high; they have the most valuable squad in the game.

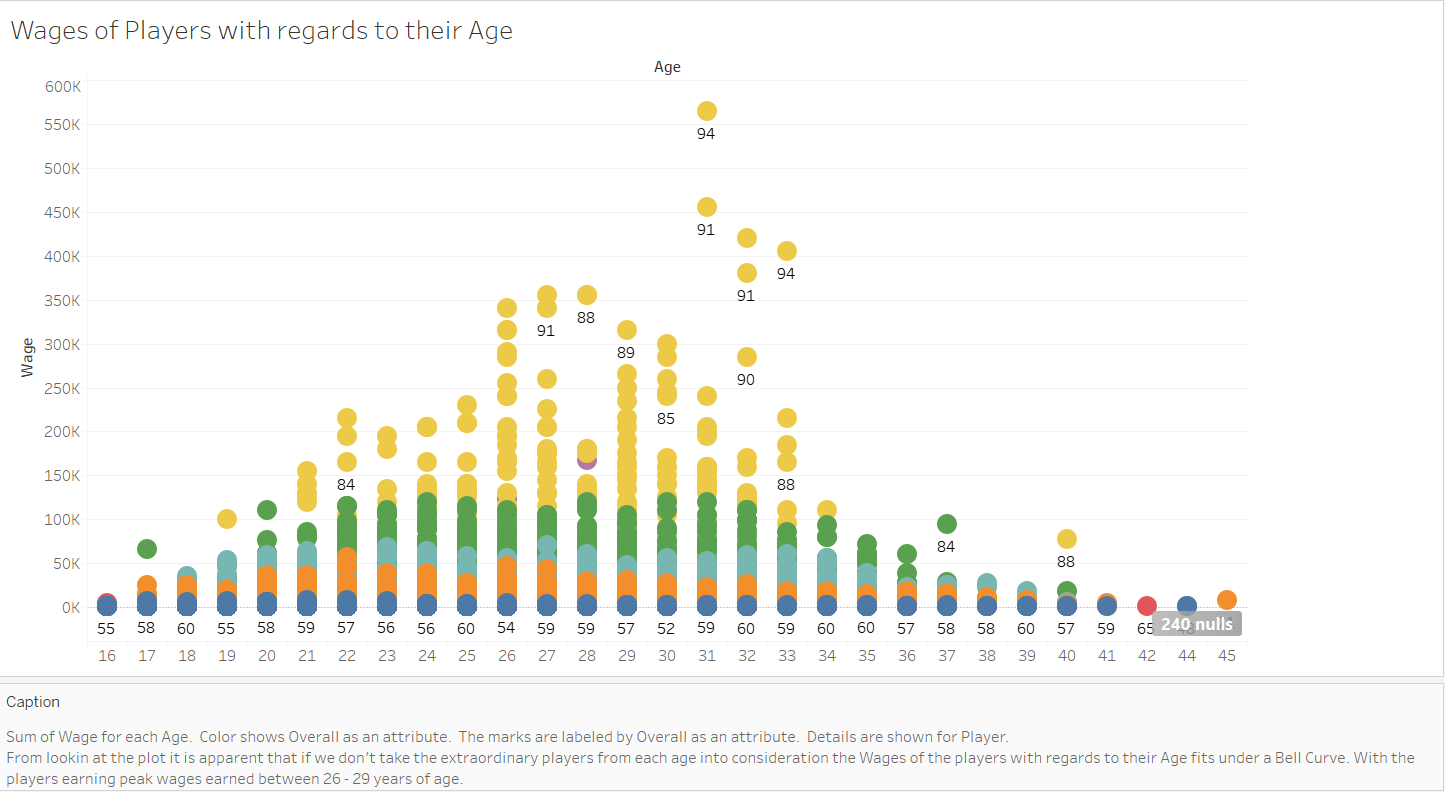
Also, almost all the clubs here keep the same place as the graph for highest wage expense. Expect that of Manchester United which had the fourth highest wage expenditure but has the tenth most valuable squad. Maybe they lacking in the negotiations or this might be because of their poor performance run in the last half decade or so.

On the other hand, we have FC Bayern who have the fifth most valuable club but are tenth highest in wages. These could be because of the near domination in their national league.

If we think along the same line, it could be argued that the reason of high value squad and high wage expense for Real Madrid and Barcelona could be result of their year on year competition for the national league title and their rivalry.

This could be resulting in their near domination of top two spot of their league.

* 1. Wages of Players with regards to their Age



Wages of all the players distributed by age.

We can see an apparent bell curve in the distribution, there are also outliers apparent.

It can be hypothesized that the player earns most when they are bet 25 – 30 years old.

There are of course exceptions to this theory but those can easily be considered outliers.

It is often said that the footballers are at their peak between 27 – 29. That would mean they are growing before that. That could be the reason that for the growth of the wages as the player are reaching 25-27 and once, they reach their peak the wages stagnate and soon after the wages

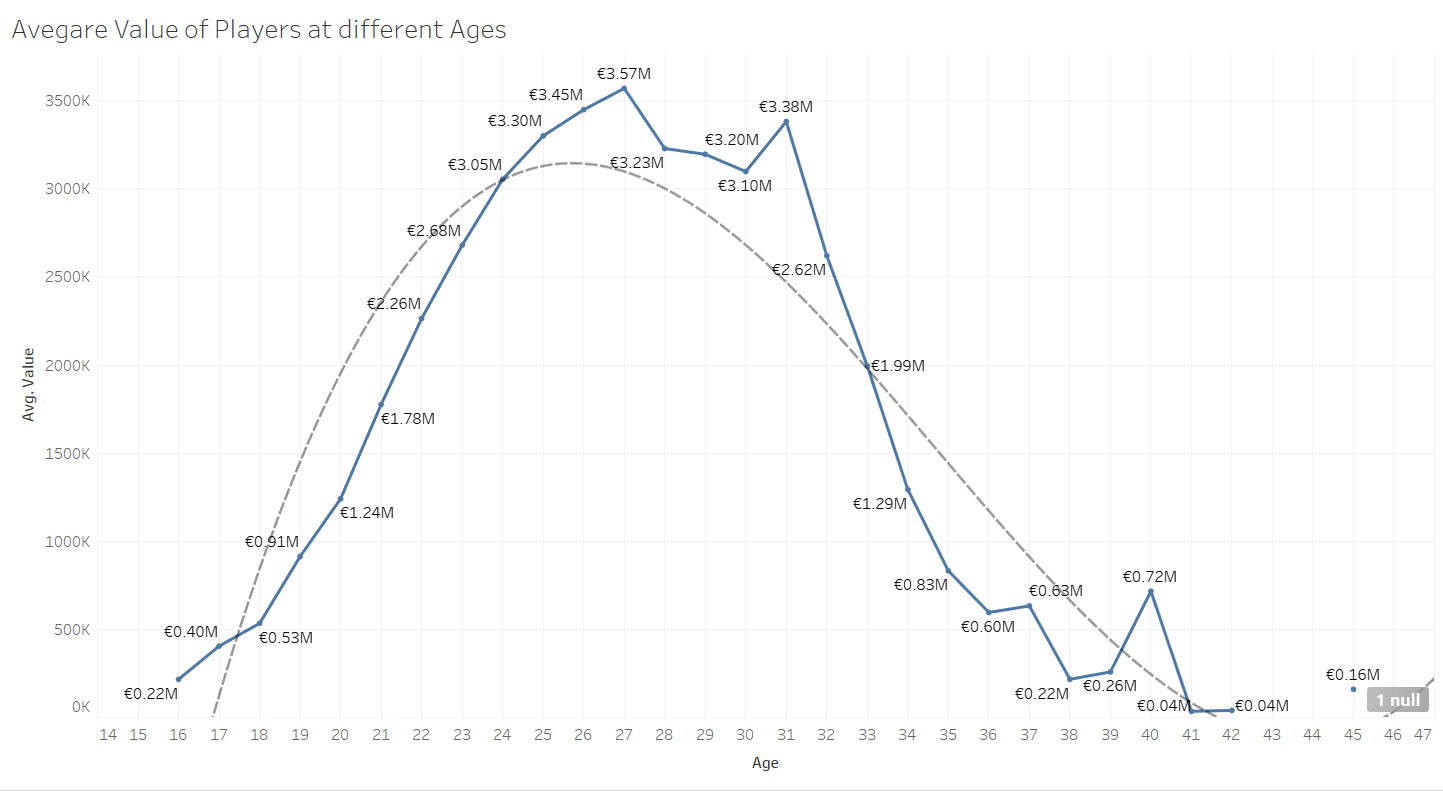
This shouldn’t be a surprise to anyone as football is a very physical game and the of course as the players grow they learn, gain more experience but once, they have passed their physical they would be very experience but their body would not be able to the same things they would have done I their prime.

This can of course be negated, the best example for this would be Cristiano Ronaldo.

Even at the age of 33 he is still earning more then most players in their prime.

This because he has adapted his playing style as he has grown older, relying less on his pace and more on his understanding of the game and efficient shorter runs.

* 1. Average Value of Players at different Ages



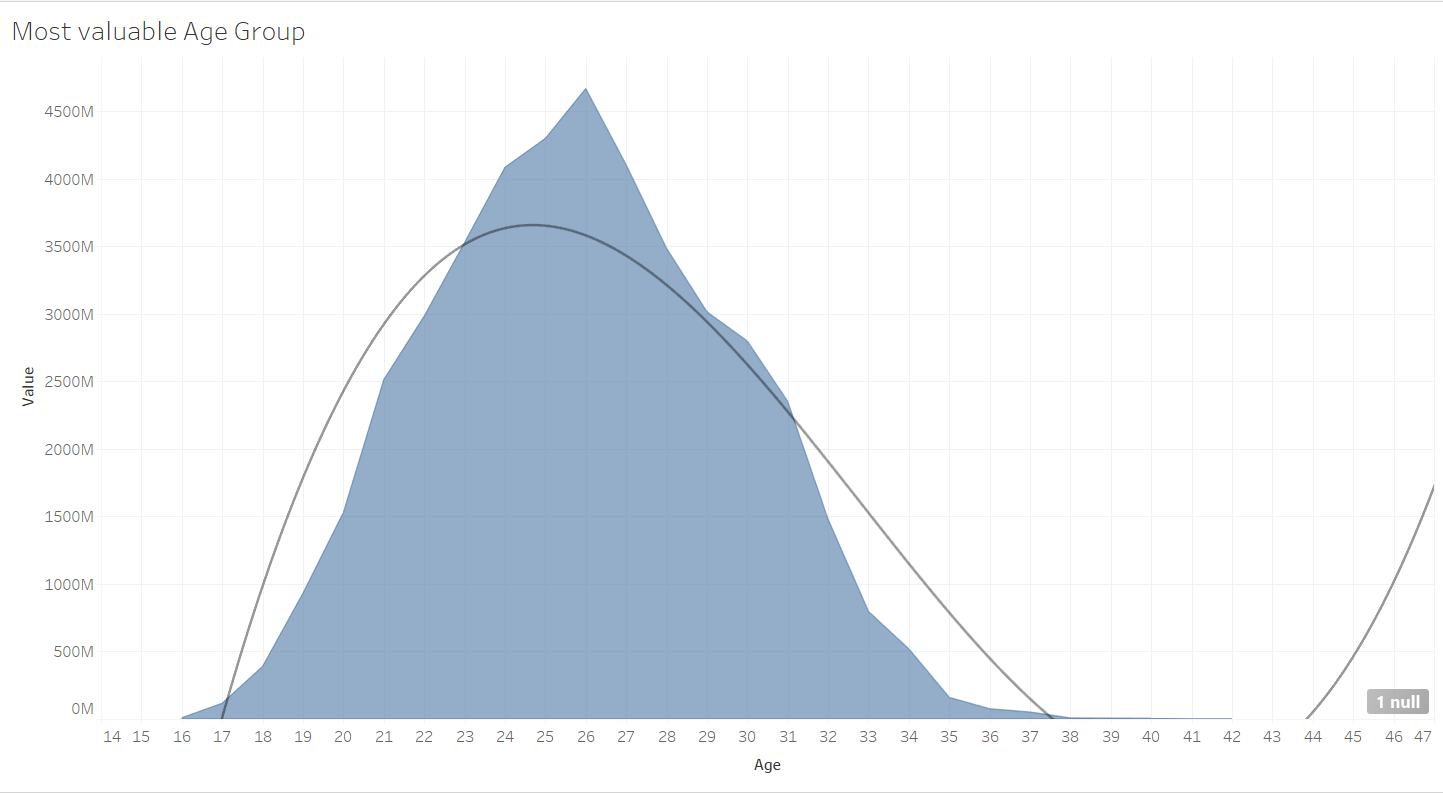
As we discussed earlier players reach their peak around 27-29 years of age. Which is also apparent from this visualization.

We can also see that the players value increase as they grow and reach their peak and then gradually fall off.

It is also apparent that the players in the game are reaching their peak at 27.

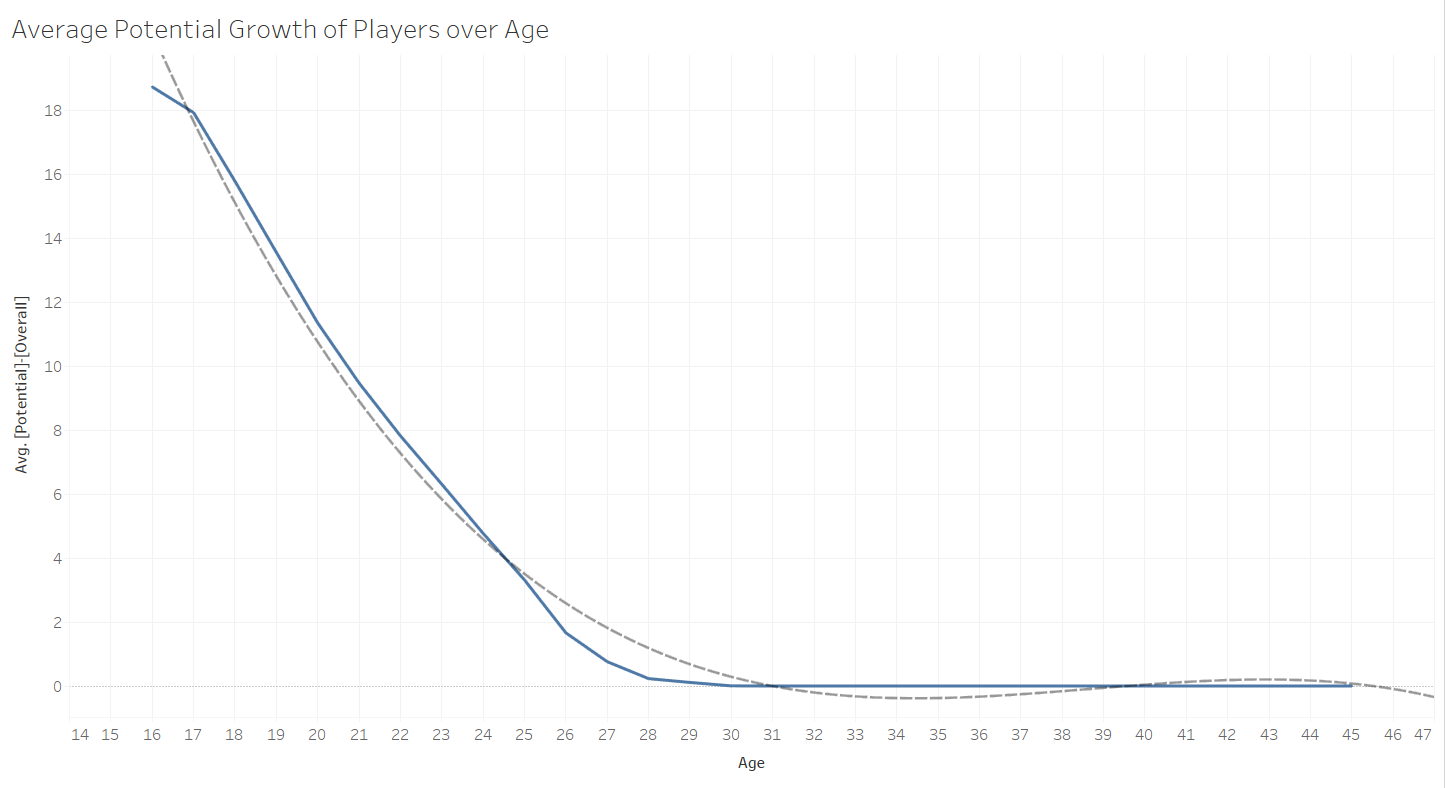
All of this also confirmed by the trend line fitted, which shows a steep increase in value of the player, stabilization for a short while and then a gradual decline.

* 1. Most valuable Age Group



For this visualization value of all the players in at a particular age is summed.

* 1. Average Potential Growth of Players over Age

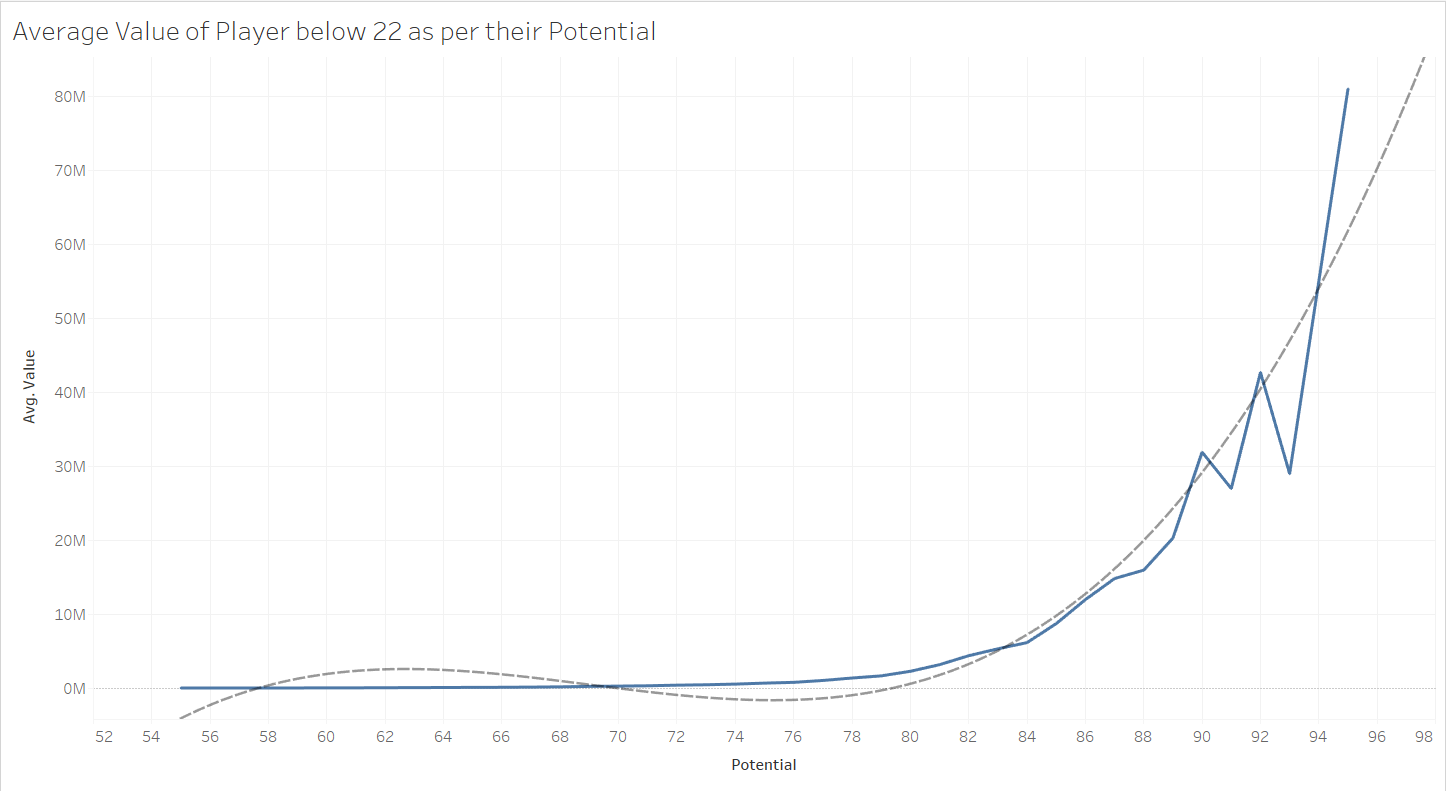


No surprises here the as it is abundantly clear that the younger the player the more growth potential he has.

But that is not why I wanted to visualize this. I wanted to visualize this because it confirms our hypothesis that the players reach their peak at 27 -29.

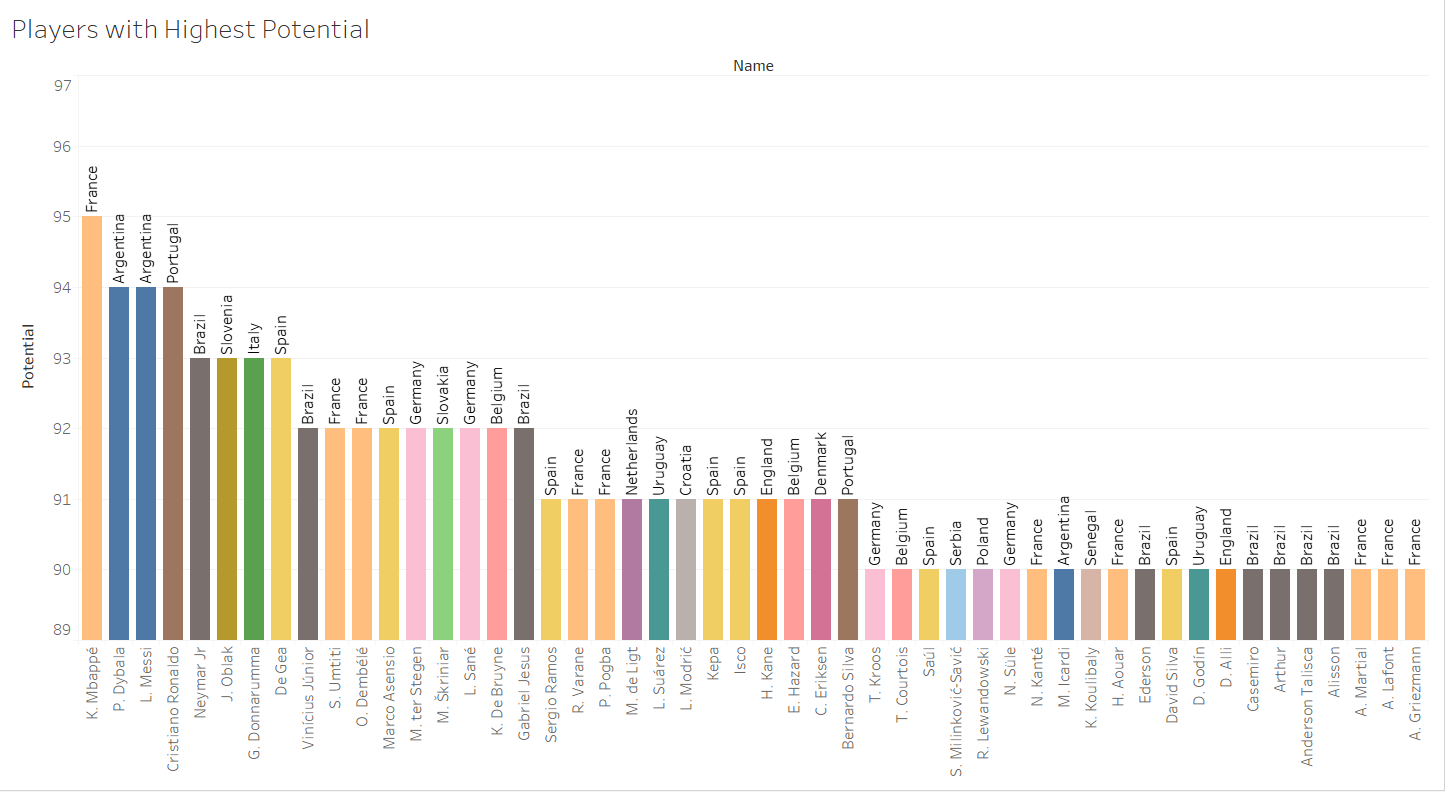
As that is when they stop growing. After that point the player only gathers more and more experience.

* 1. Average Value of Player below the age of 22 as per their Potential

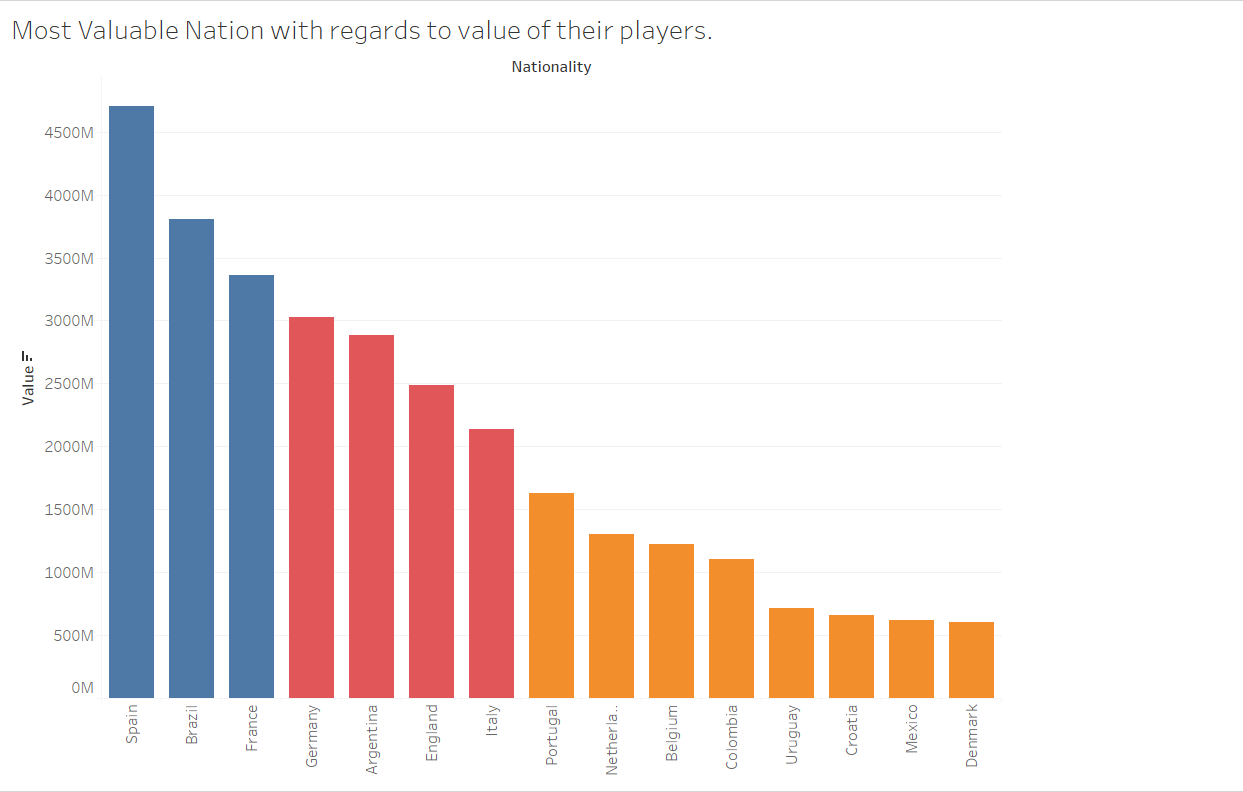


What I am try to visualize here is that if the potential of the player below the age of 22 is ascertained to be above 80 there would be exponential increase in his value.

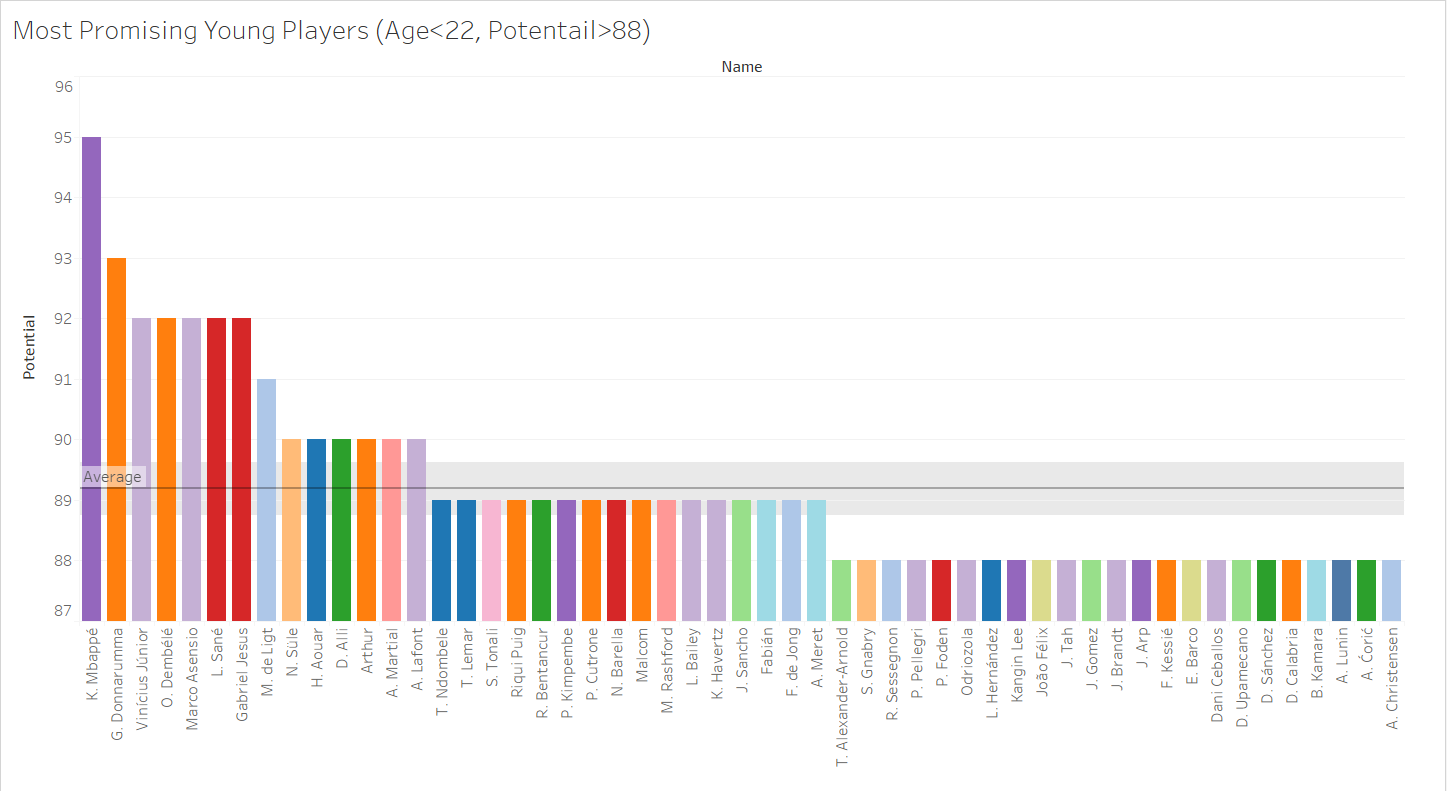
* 1. Players with Highest Potential



* 1. Most Valuable Nation with regards to value of their players



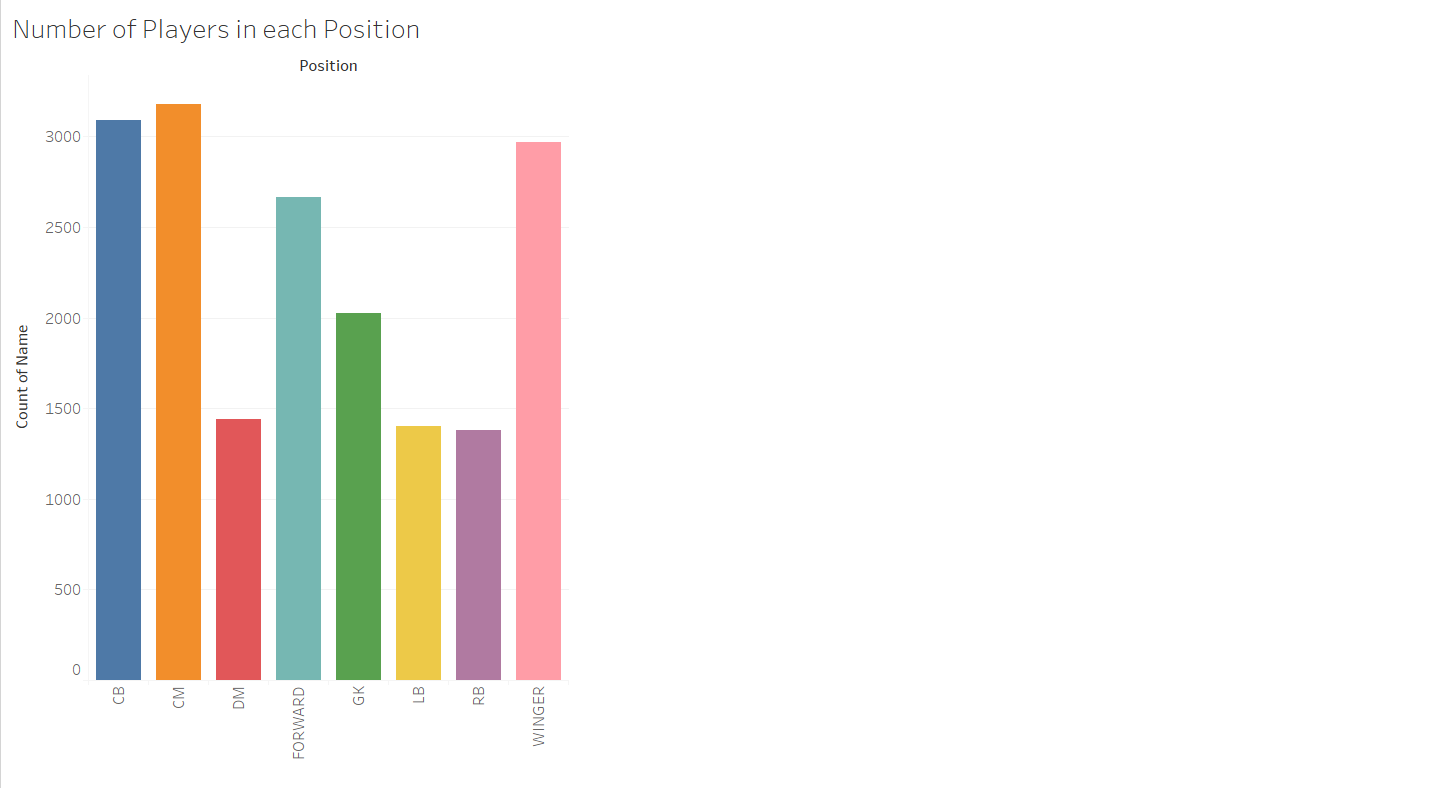
* 1. Most Promising Young Players (Age<22, Potential>88)



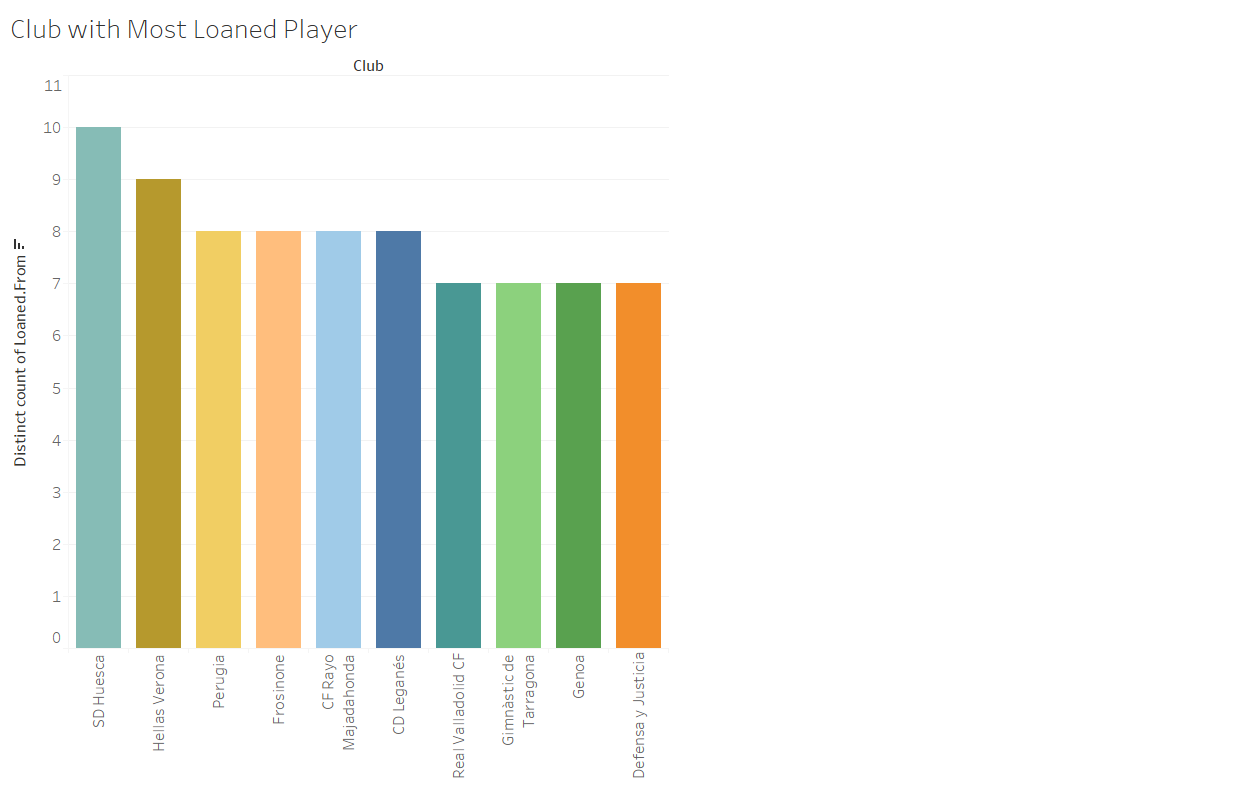
When you are playing career mode in the game who is young and has a high potential.

Here we have all the players below the age of 22 and potential higher than 88.

* 1. Number of Players in each Position



* 1. Club with Most Loaned Player

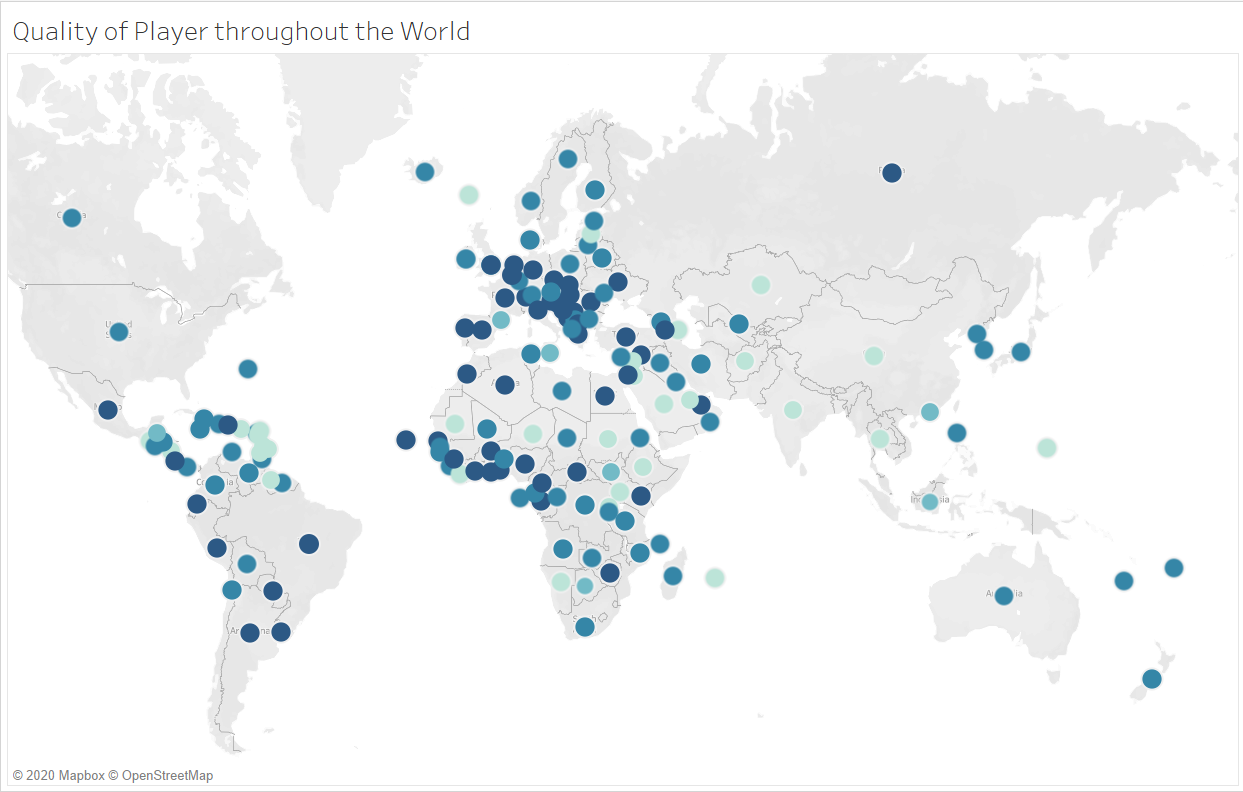


Not all clubs are equal in wealth and as such some prefer to buy the players they need while the others have to loan the players.

This graph shows the clubs with most players got

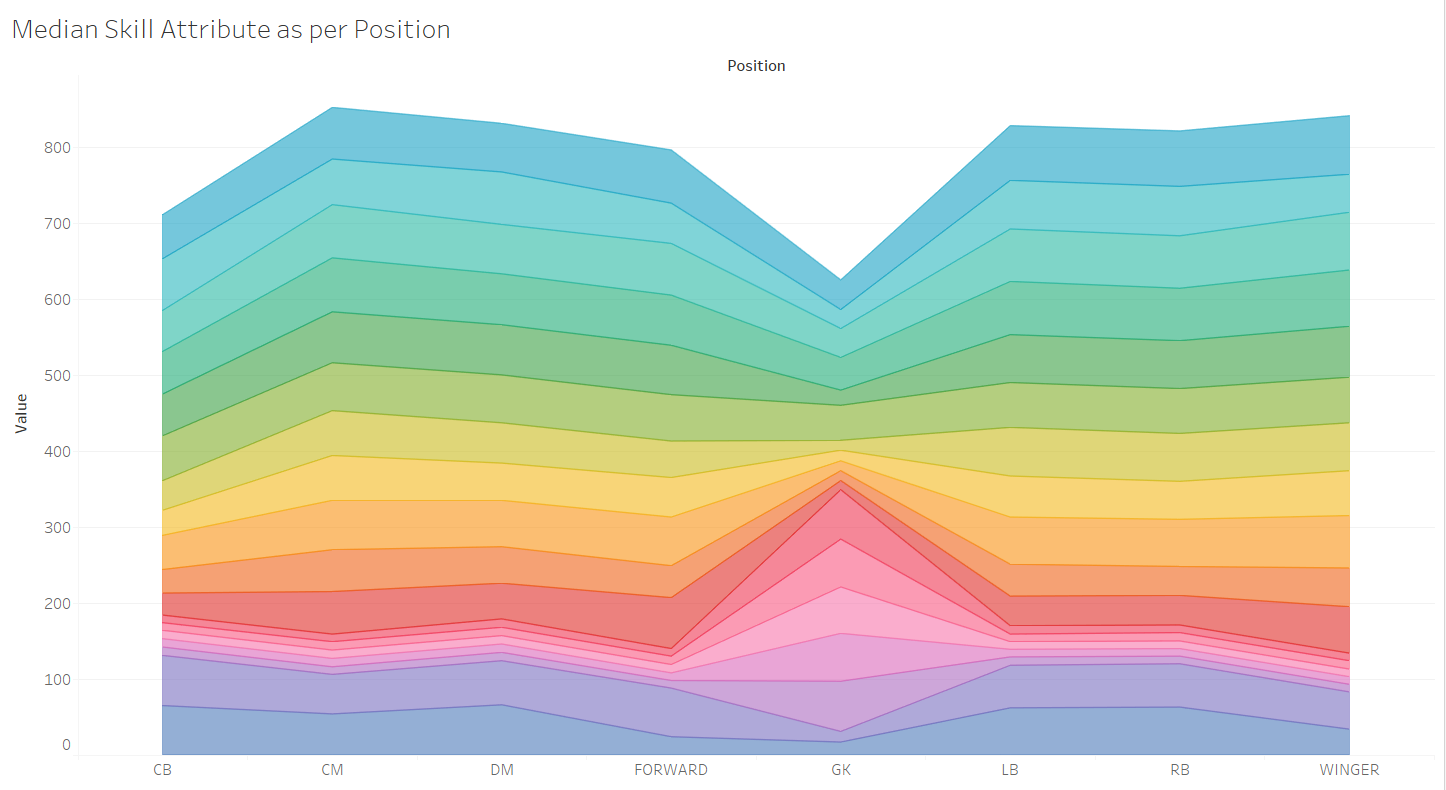
on loan.

* 1. Quality of Player throughout the World



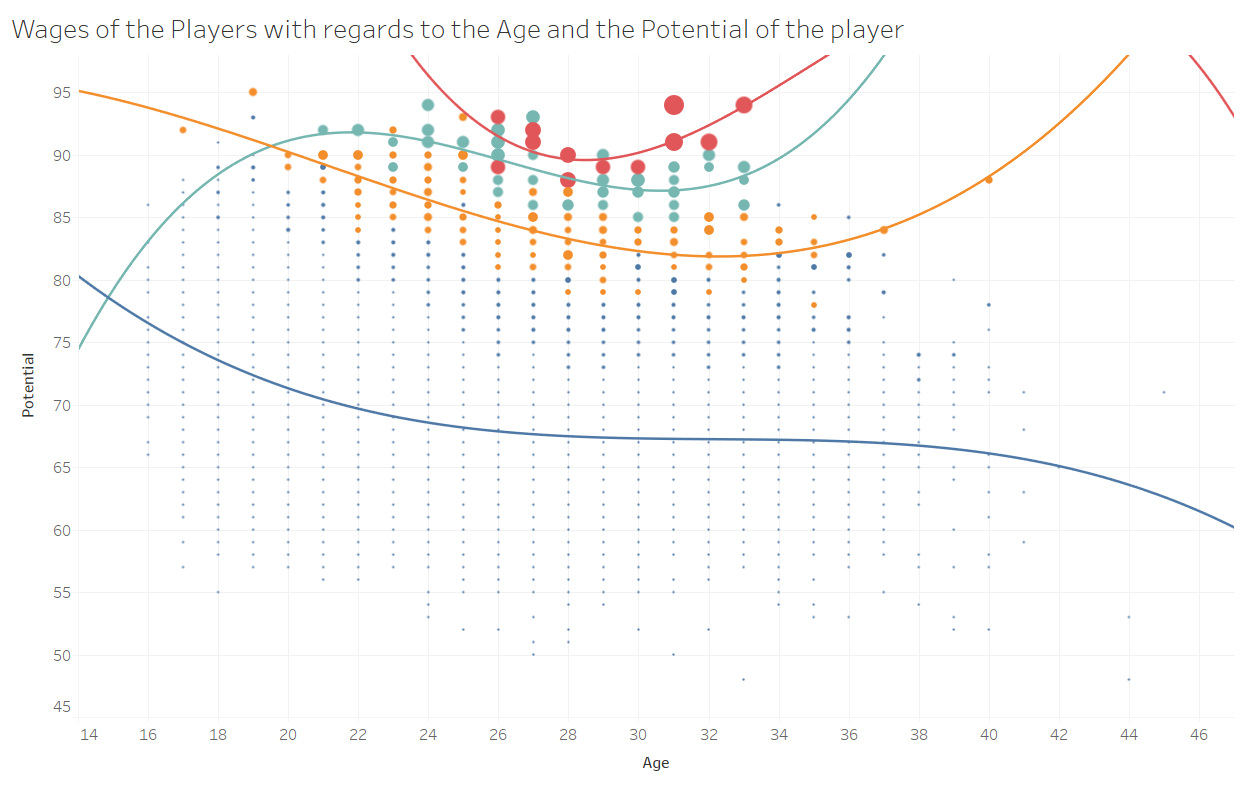
Here, median overall of the player has been taken for every nation. To determine the quality of players each nation have. Darker the point the better the quality of players.

* 1. Median Skill Attribute as per Position



Medians skills required for each position.

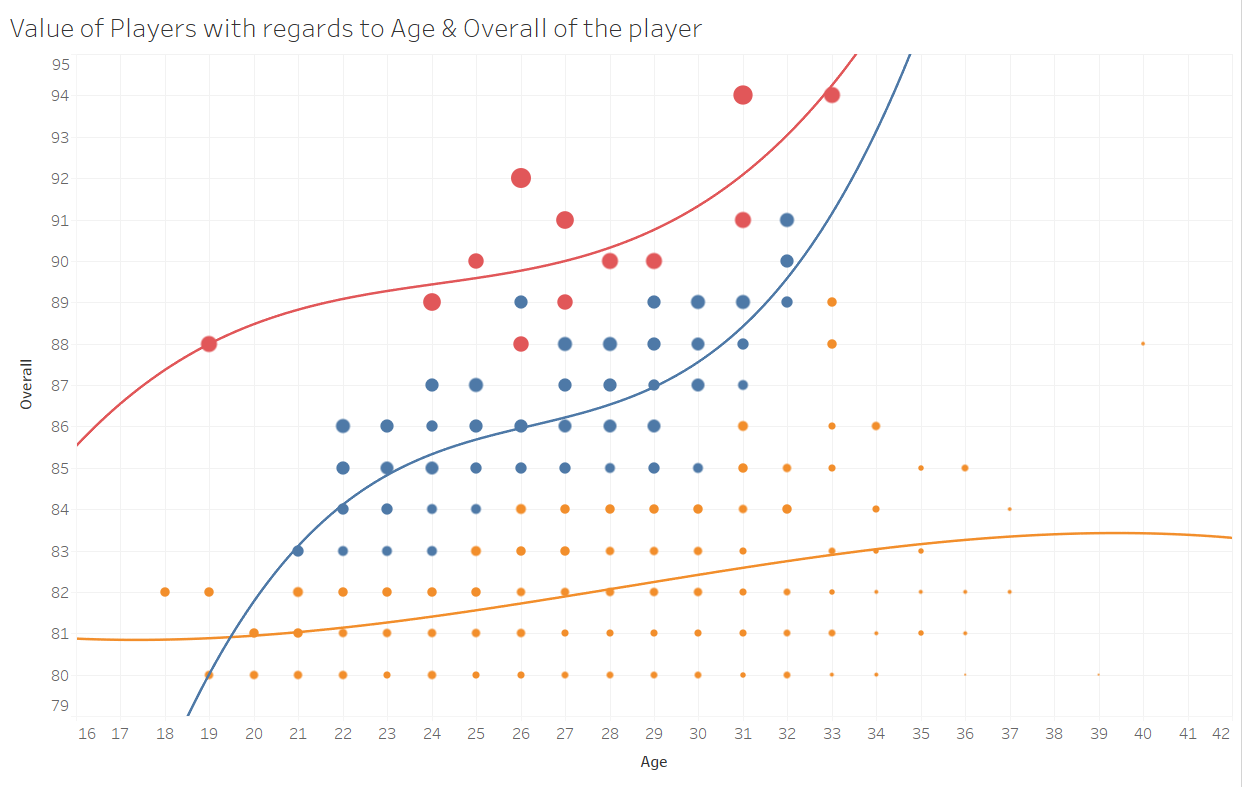
* 1. Wages of the Players with regards to the Age and the Potential of the player



Visualized is the wages of the players with regards to the age and potential of the player.

We can see four distinct clusters, or career path players could have.

* 1. Value of Players with regards to Age & Overall of the player



Similar to above this is the value of the players with regards to Overall and Age of the player.

There are three different clusters.

I like to think of them as;

Average players

Good players

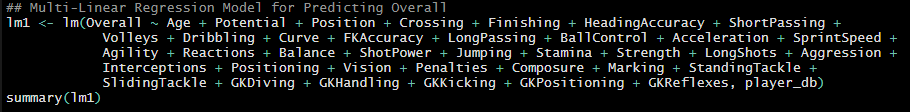
Extraordinary players

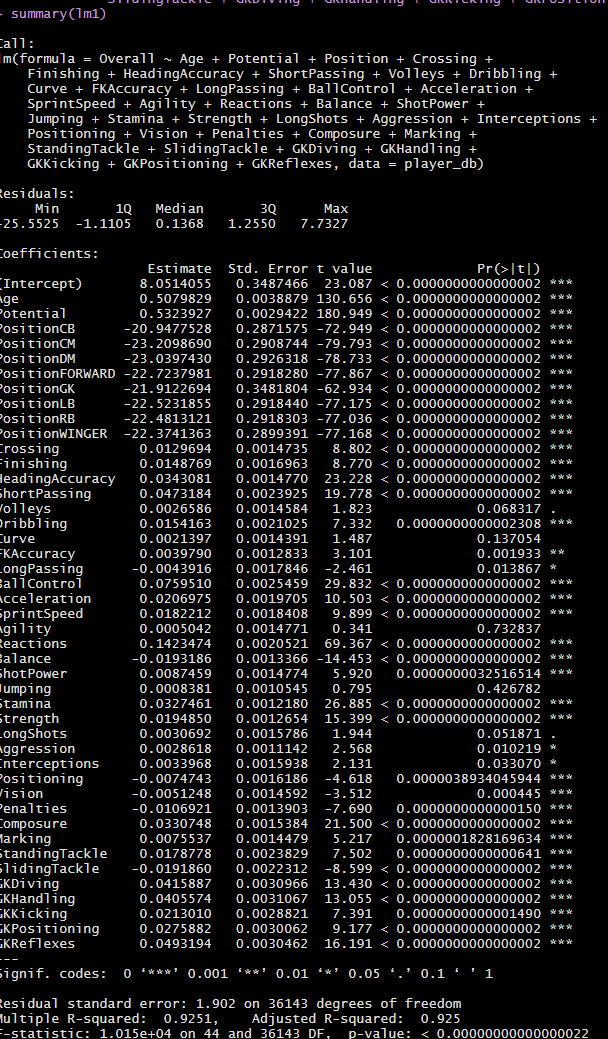
1. Finally, we get to building various ML Model to predict the Overall & Value of the players.

I have built multiple models of each type to ascertain the best model for the job. We will start with the model to predict the overall of the players and then move on to models to predict the value of the players. I have used Multi-Linear Regression, Regression Tree & Neural Networks. Let’s get started,

* 1. Building models to predict the Overall of the Player
     + Multi-Linear Regression Model

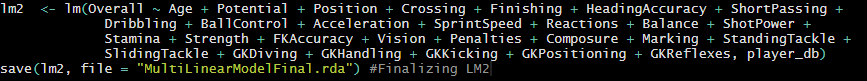
Creating the first model to identify the necessary variable for the model;

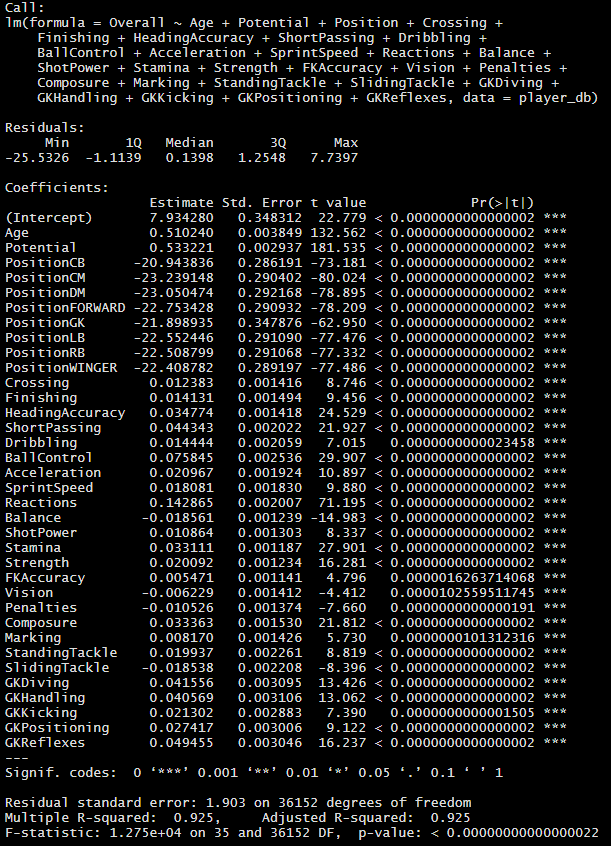




We have got good R-squared results and residual error is low as well.

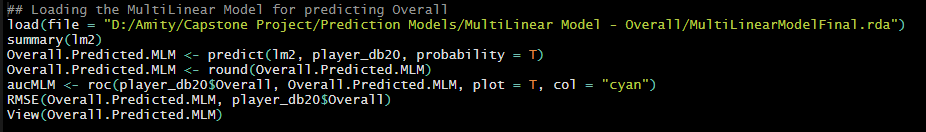
Next, to optimise the model the unnecessary variables have been dropped.





R-squared has not decreased much and neither has residual error increased a lot.

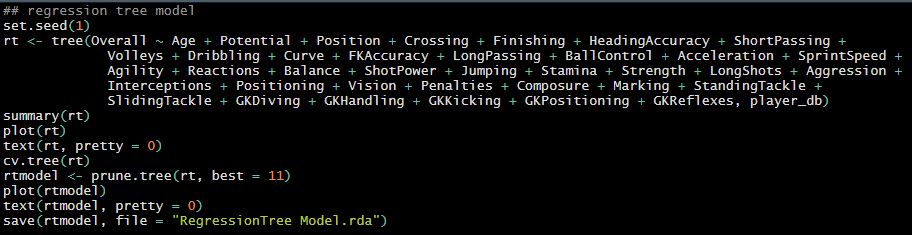
Testing the model;



Root-Mean-Error for this model is 1.83379, which is acceptable but we must look at other models built to determine the best from the lot.

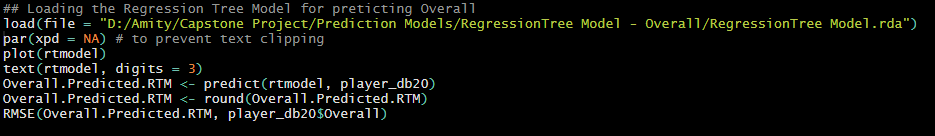
* + - Regression Tree Model

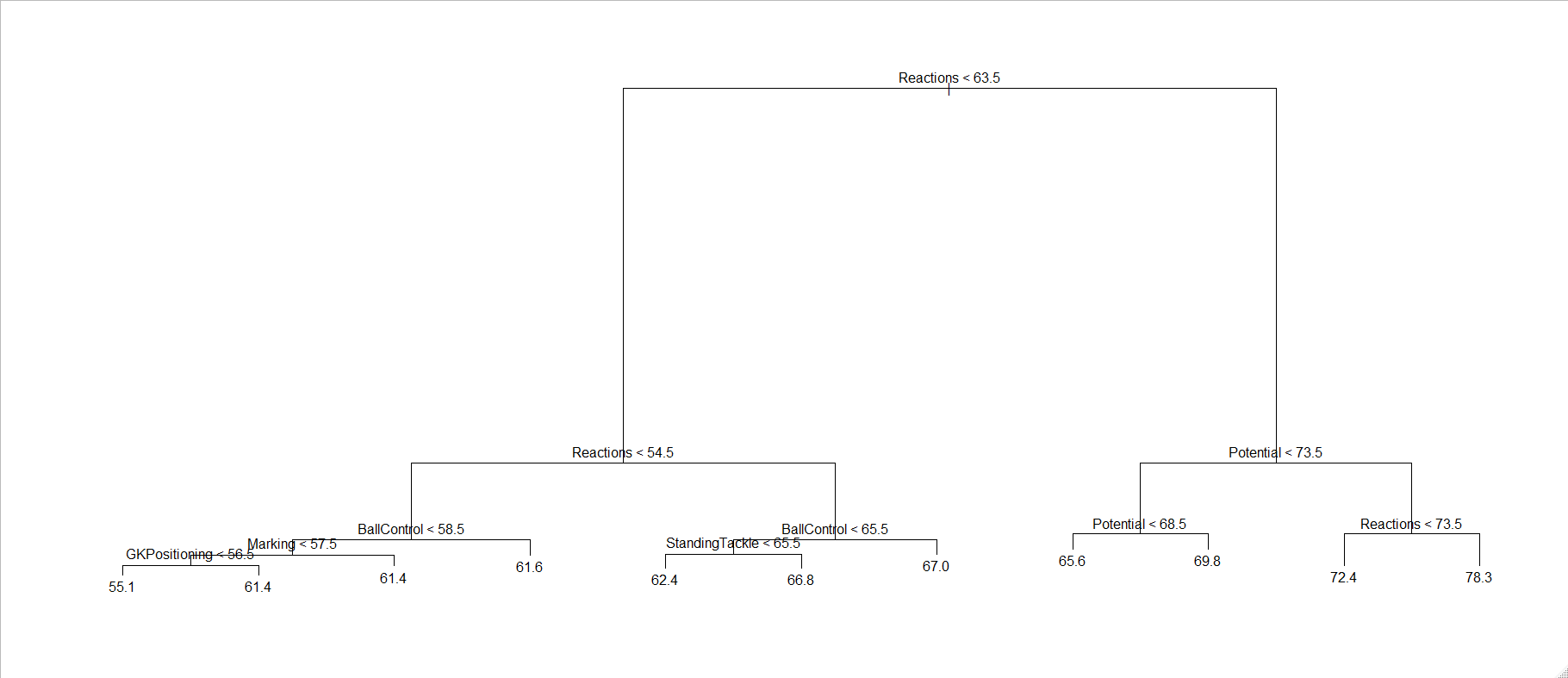
I have created only one Regression Tree Model using the same variables we used in the last model.



I have pruned the tree for optimization.

Let’s, test the model and have a look at how we have done.

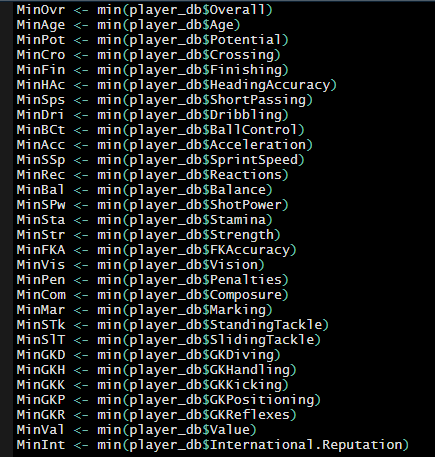


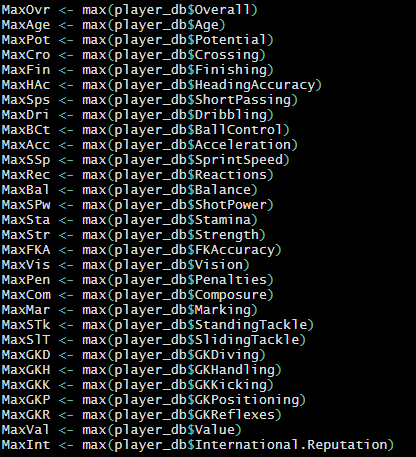


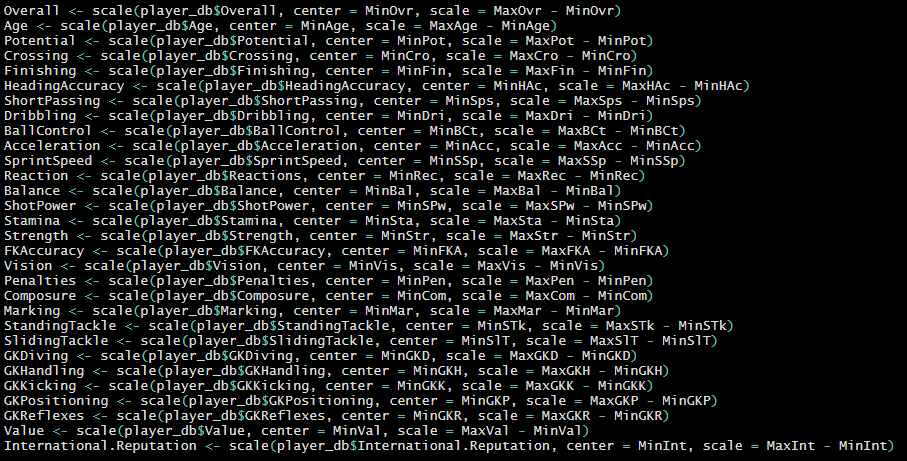
The Root-Mean-Squared-Error of this model is 3.146225 which is a lot higher than the Multi Linear Model (lm2) and thus lm2 seems like a better model till now.

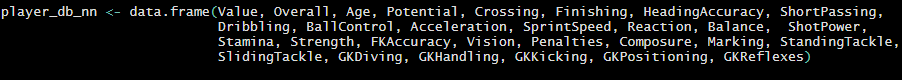
* + - Artificial Neural Network Model

Before, we start with the last model built for predicting overall of the players, we must standardize the data as it helps greatly with the reducing the model building time also, I have seen better accuracy using standardized data.



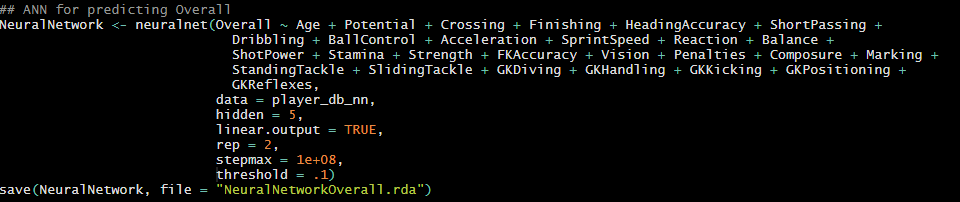


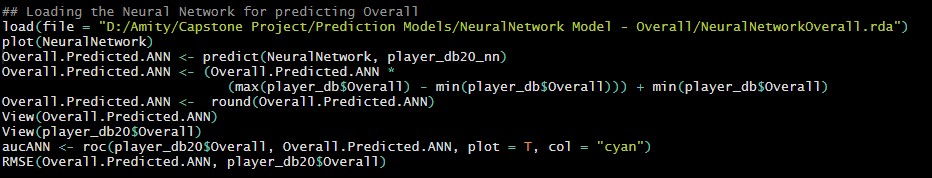


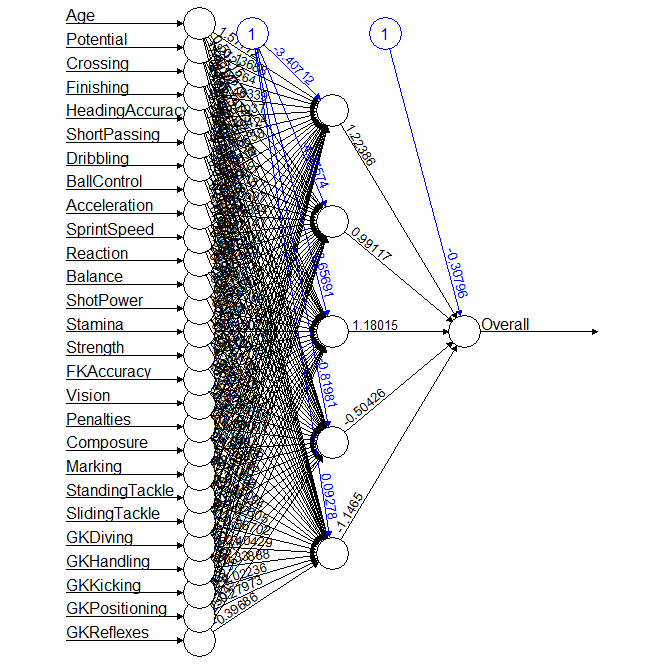


I am sure there are better ways to standardize the data. But this is the one I remembered at the time of doing this and since I had already done it, I didn’t bother changing it when I remembered other better methods.

Now to building the model







We also need to standardize the test data for the model to be able to predict, that is done in script file “FIFA\_20\_DataCleaning.R”.

Further after predicting the prediction results need to be rescaled to the scale of the original data to of any use to us.

RMSE is 2.133613 which more than it was for lm2 but is still acceptable.

* 1. Building models to predict the Value of the Player

While building the model for predicting the Value of the Player there were many difficulties to be faced.

First and foremost was that, I, for the life of couldn’t come with a model which would give me any form of satisfying results.

There were of course many reasons for this, not least of which is that valuation of a player is very delicate and many factors are considered such as Age, Overall, Potential, International Reputation these are the once used by me to build the model but, there are many other factors to be considered which can are not quantifiable or that we do not have access to such as squad status, league the player is playing, image rights, adaptability, etc.

So, keeping all the above things in mind I am only presenting the models that were even slightly satisfactory.

I made multiple Regression Tree and Neural Network but to no avail. At the end I decided to segment the data based on the overall of the players. I created 4 Segments;

First <- Players with Overall more than or equal 80

Second <- Players with Overall less than 80 but more than or equal to 70

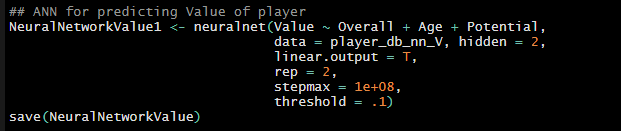
Third <- Players with Overall less than 70 but more than or equal to 60

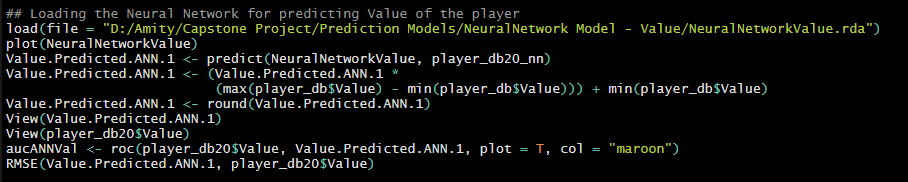
Forth/Last <- Everyone else

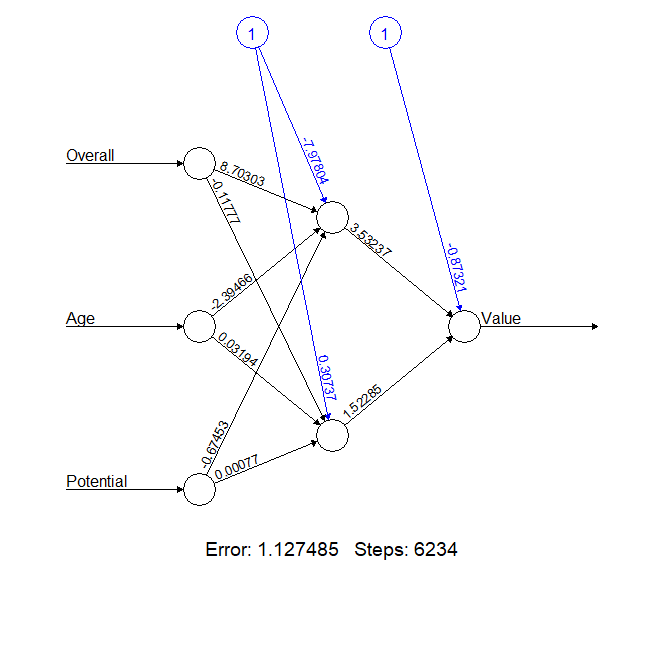
Doing same to the test data as well.

This I thought would give the model a chance to understand the population with less fluctuation of the dependent variable. Which has worked out better than anything else I tried.

* Model 1 (NeuralNetworkValue)





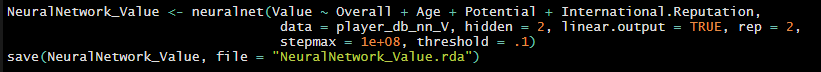


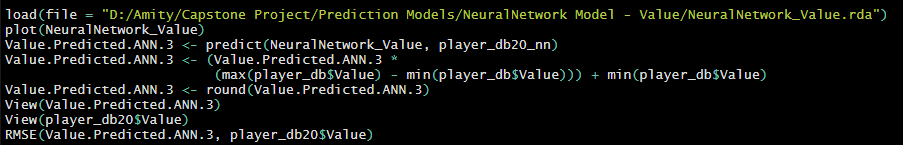
RMSE <- 1737018

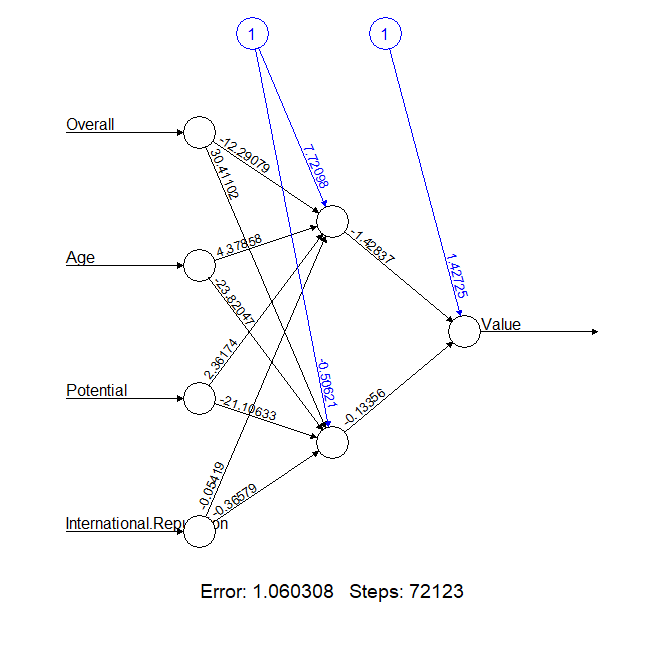
I have only used 3 variables here with scaled data, this was but a test run. A baseline of sorts.

RMSE is higher than I would prefer, but lest move forward to and have a look at how other models do.

* Model 2 (NeuralNetwork\_Value)

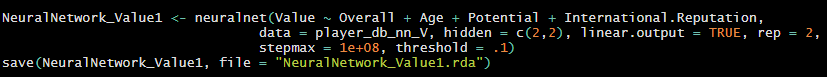


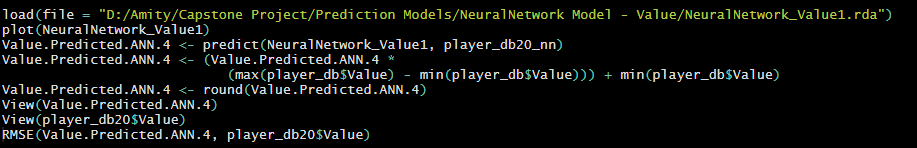


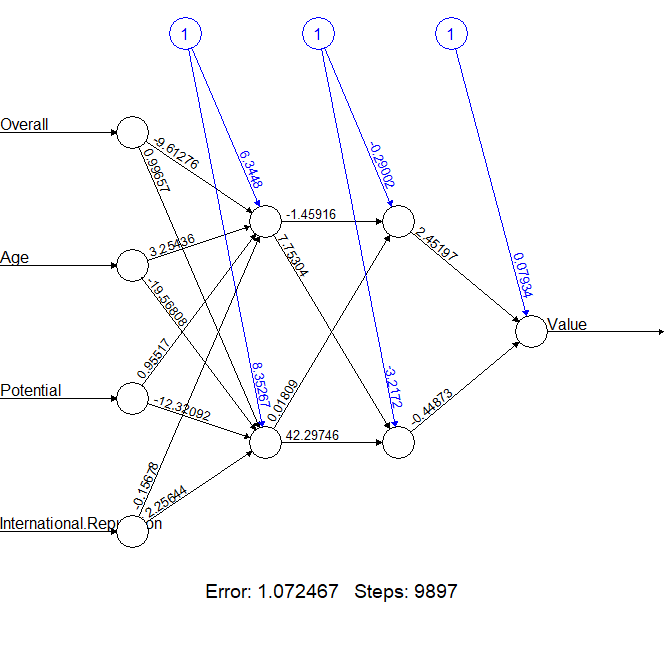


The complexity of the model has been increased, number of repetitions has also been increased but, the RMSE (1877101) has also increased for this model.

* Model 3 (NeuralNetworkValue1)

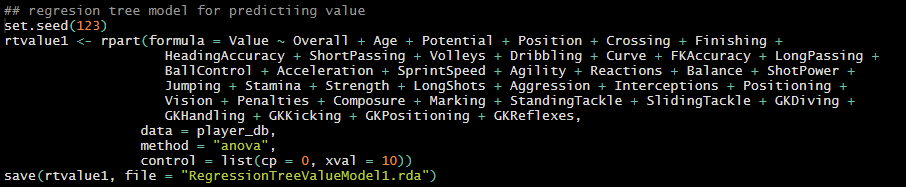


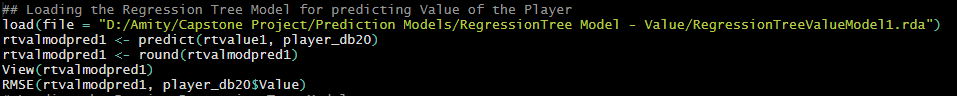




RMSE <- 2193584

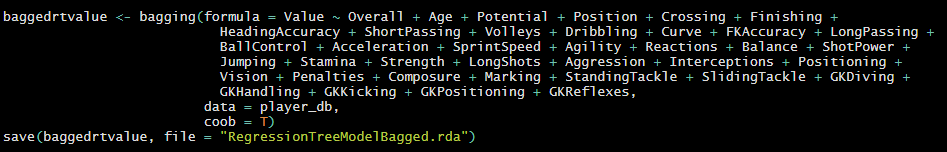
* Model 4 (rtvalue1)

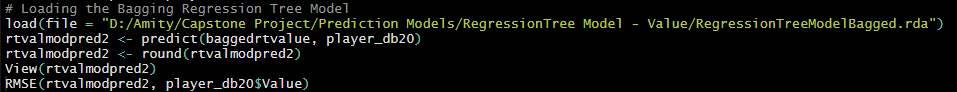




RMSE <- 1018627

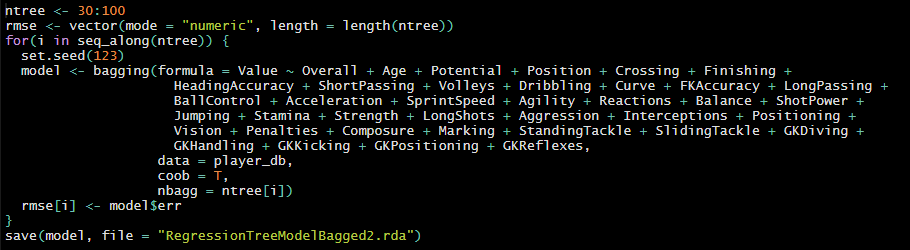
* Model 5 (baggedrtvalue)

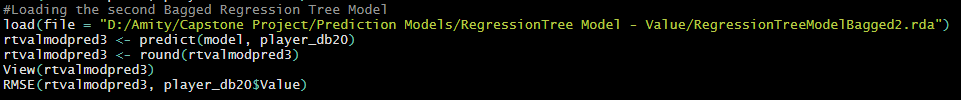




RMSE <- 1686963

* Model 6 (model) [Bagged Regression Tree Model]

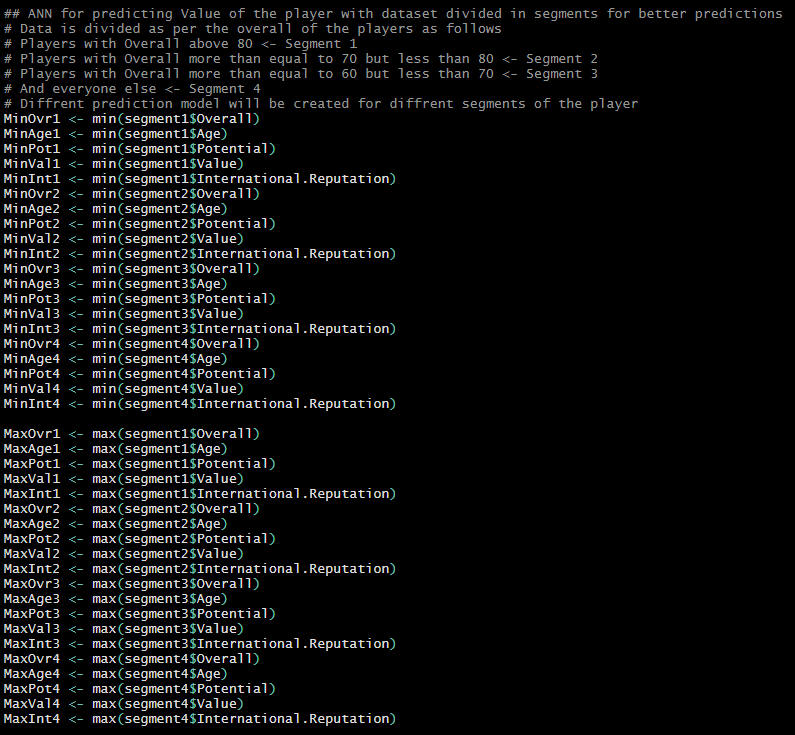


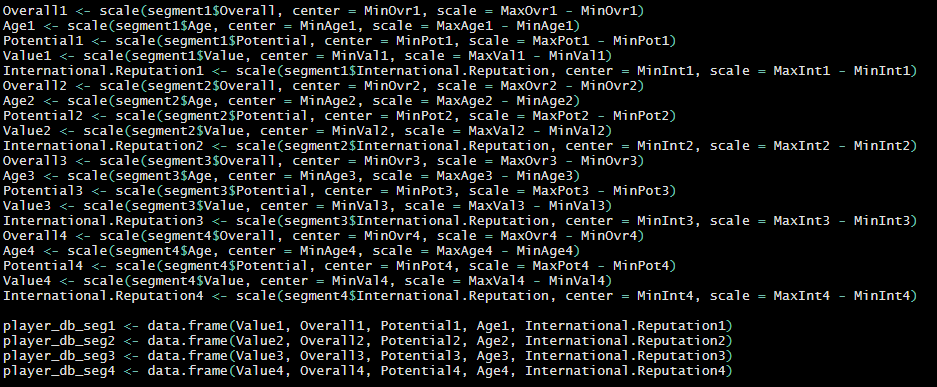


RMSE <- 1663853

* Model 7 [Neural Networks with Segmented Data]

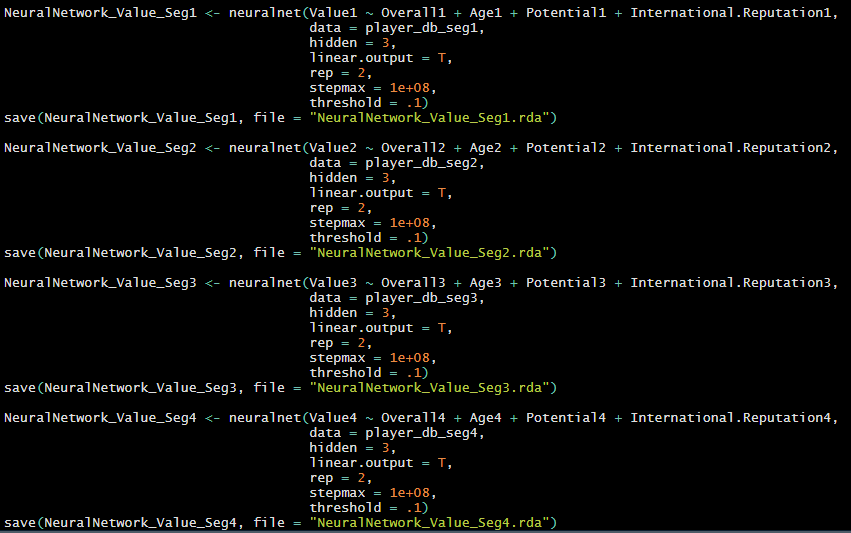
For this model again the data needs to be scaled, it is of course not ideal, but it does help in reducing model building time and the accuracy of the model does not suffer a lot. It is done in an act of balancing accuracy to compute time required.

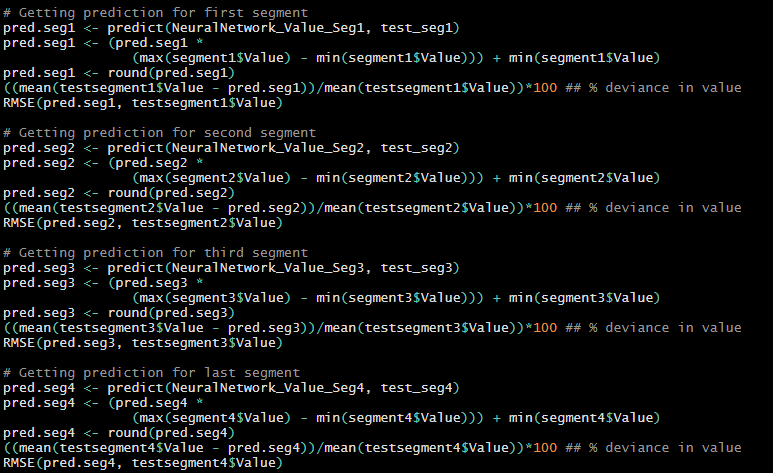




Once the data is scaled the we need to move on to building the models.

Here, since the data is segmented, we need to will need to create separate models for each segment.





Deviance of various model predictions to their actual values are as follows;

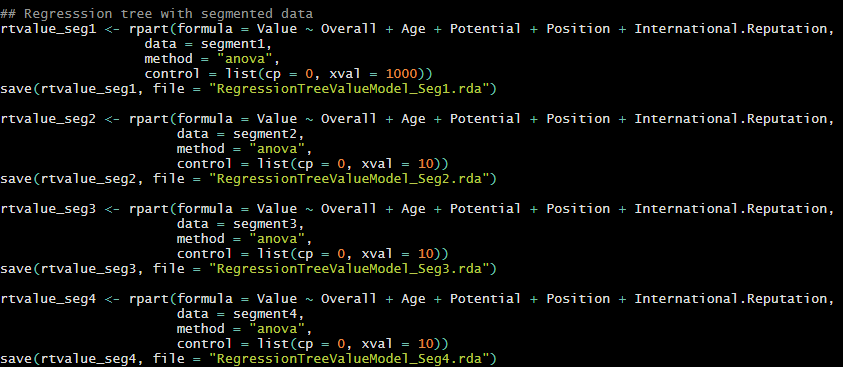
For Segment 1 = 3.554717% RMSE = 4045410

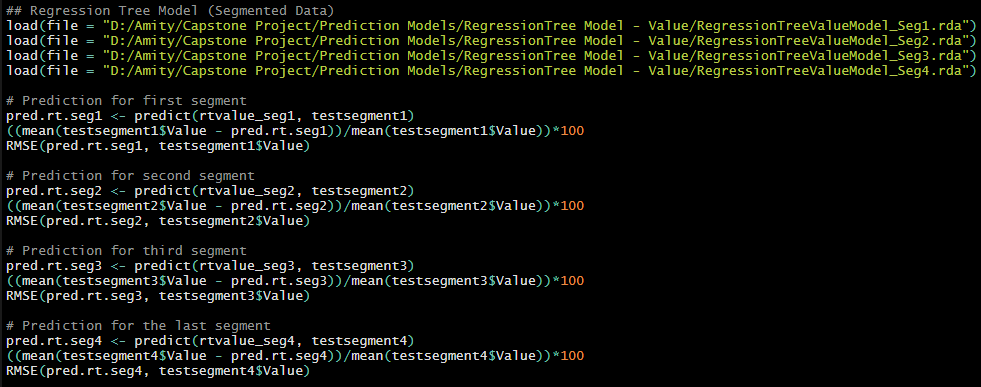
For Segment 2 = 2.486% RMSE = 1566107

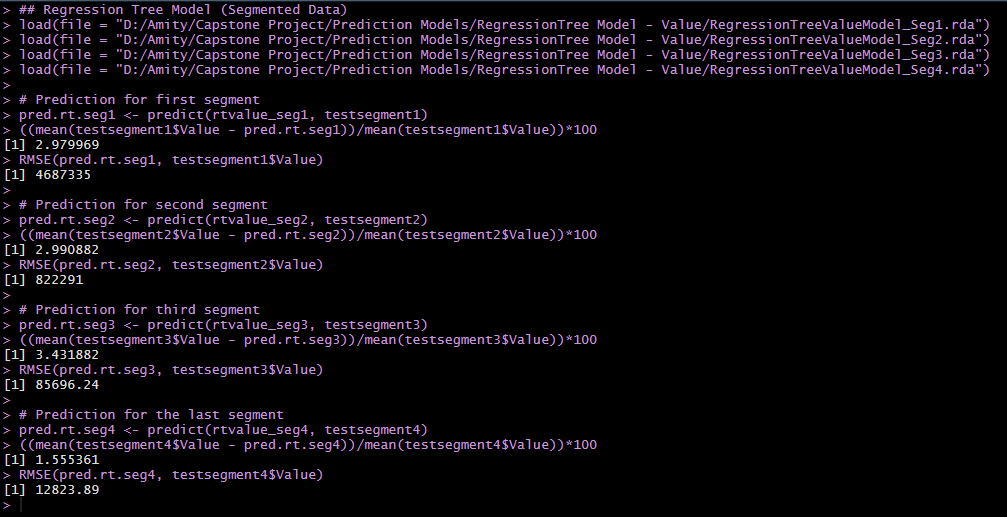
For Segment 3 = 4.587903% RMSE = 111627

For Segment 4 = 3.6521% RMSE = 28000

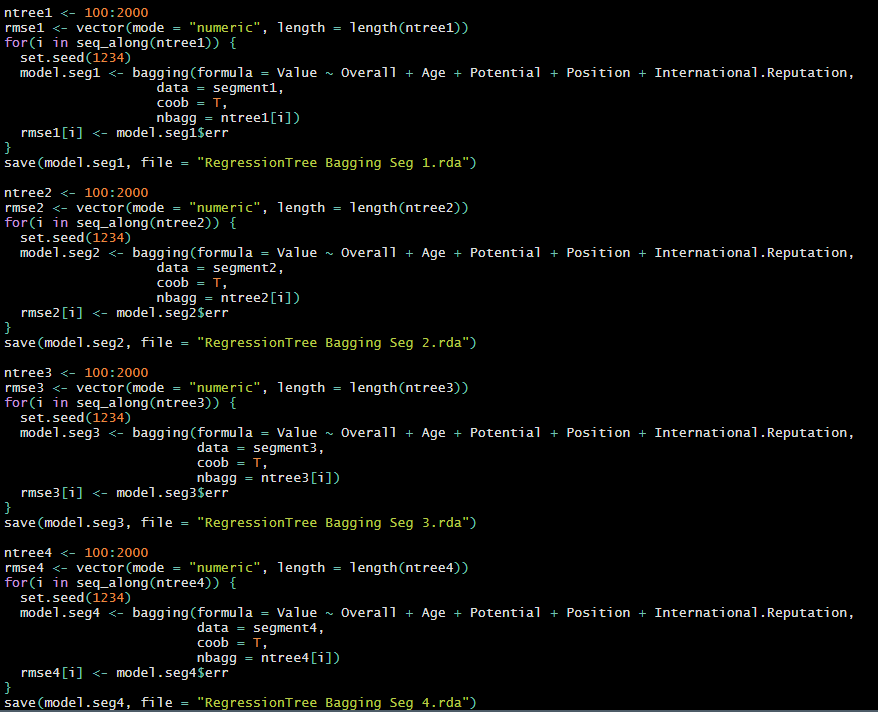
* Model 8 (rtvalue\_seg1,2,3&4) [Regression Models with Segmented Data]

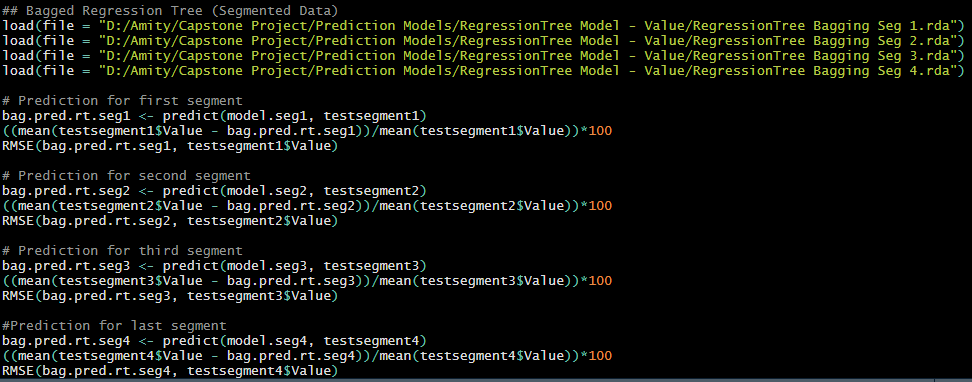


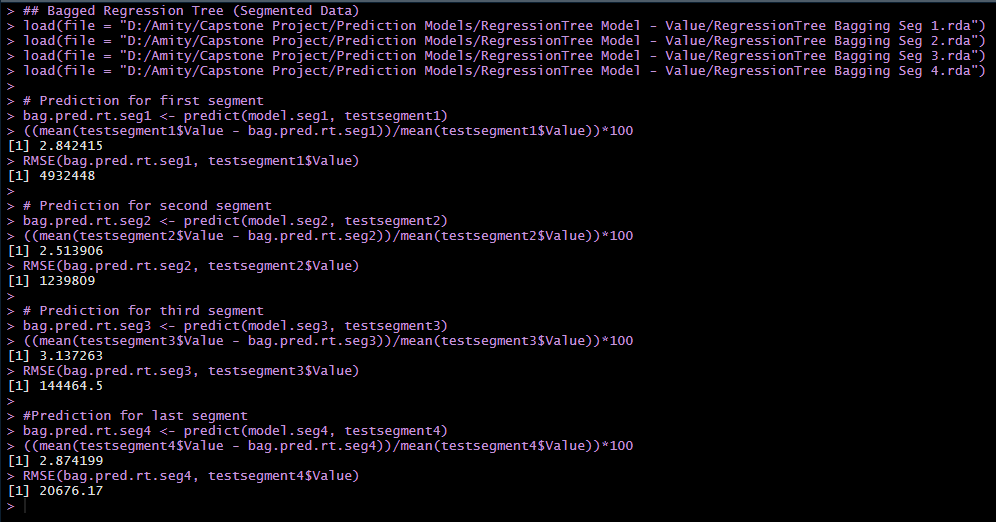




* Model 9 (model.seg1, 2, 3 & 4) [Bagged Regression Tree Model with Segmented Data]



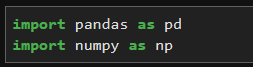




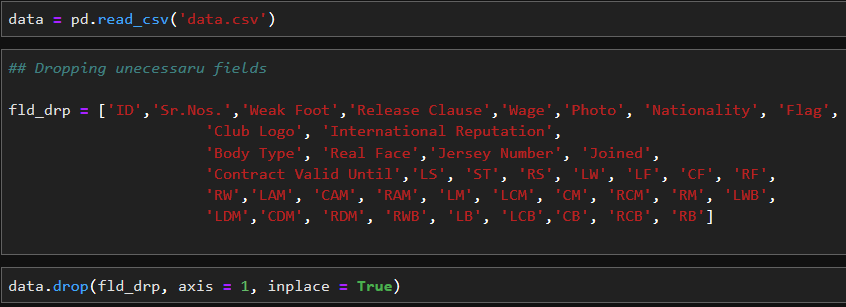
I ended up building 9 models to predict the value of the player to get a satisfying result and that came from Model 9 (model.seg1, 2, 3 & 4) [Bagged Regression Tree Model with Segmented Data].

It has been able to predict data from all 4 segments with least deviance, although that being said, these models take up about 3.5GB of space alone. I understand this might be a breaking point for many. In that case Model 8 (rtvalue\_seg1,2,3&4) [Regression Models with Segmented Data] can be used as it takes up significantly less space and can predict close enough to Model 9.

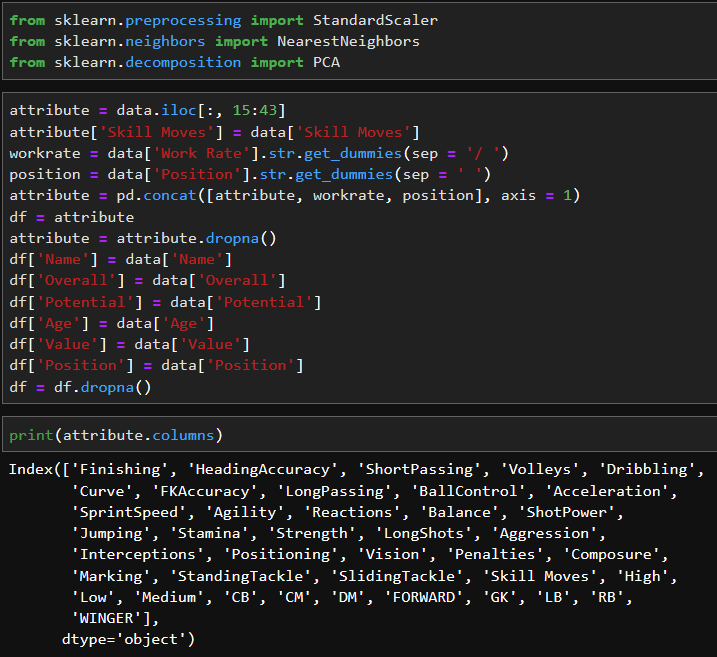
1. Creating a Recommendation Model. (Using Python)
   1. Importing “pandas” & “numpy”



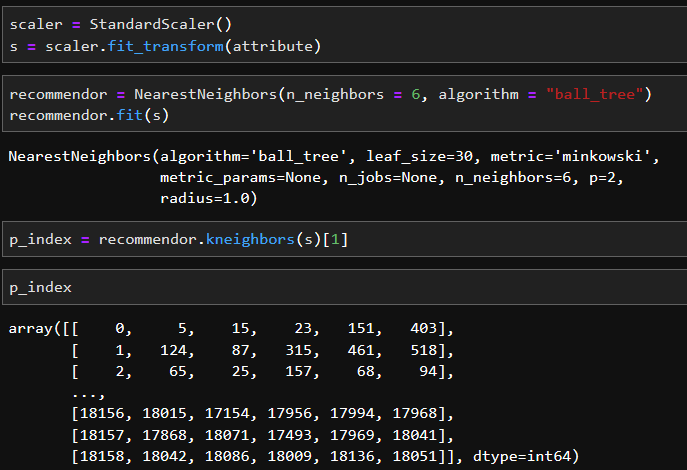
* 1. Importing data and dropping unnecessary fields.



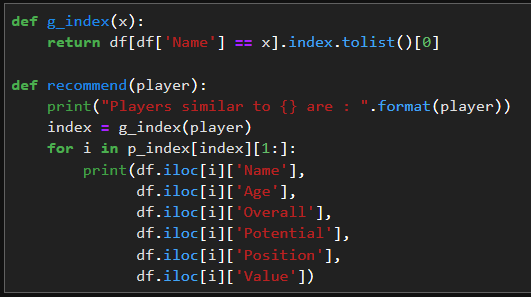
* 1. Importing necessary libraries for building the model & creating sub-data base necessary for the same.



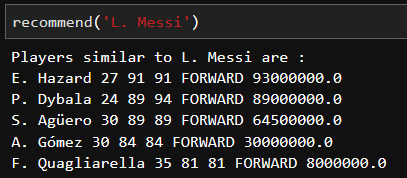
* 1. Creating the Model



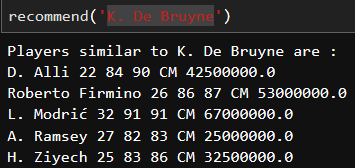
* 1. The Recommendation Function



* 1. Getting Recommendations
     + L. Messi



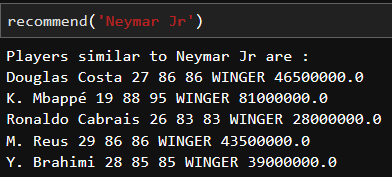
* K. De Bruyne



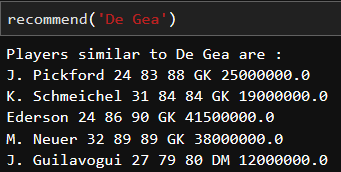
* Cristiano Ronaldo



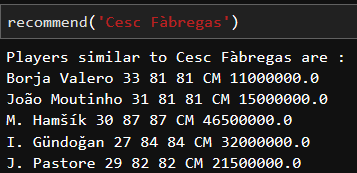
* Neymar Jr.



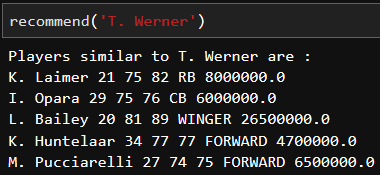
* De Gea



* Cesc Fabregas



* T. Werner



Conclusion:

\*All the file of this project can be found on (Github link) and the visualization can be found on (https://public.tableau.com/profile/yash.panchal3644#!/vizhome/FIFA19PlayerDatabase/Sheet1)