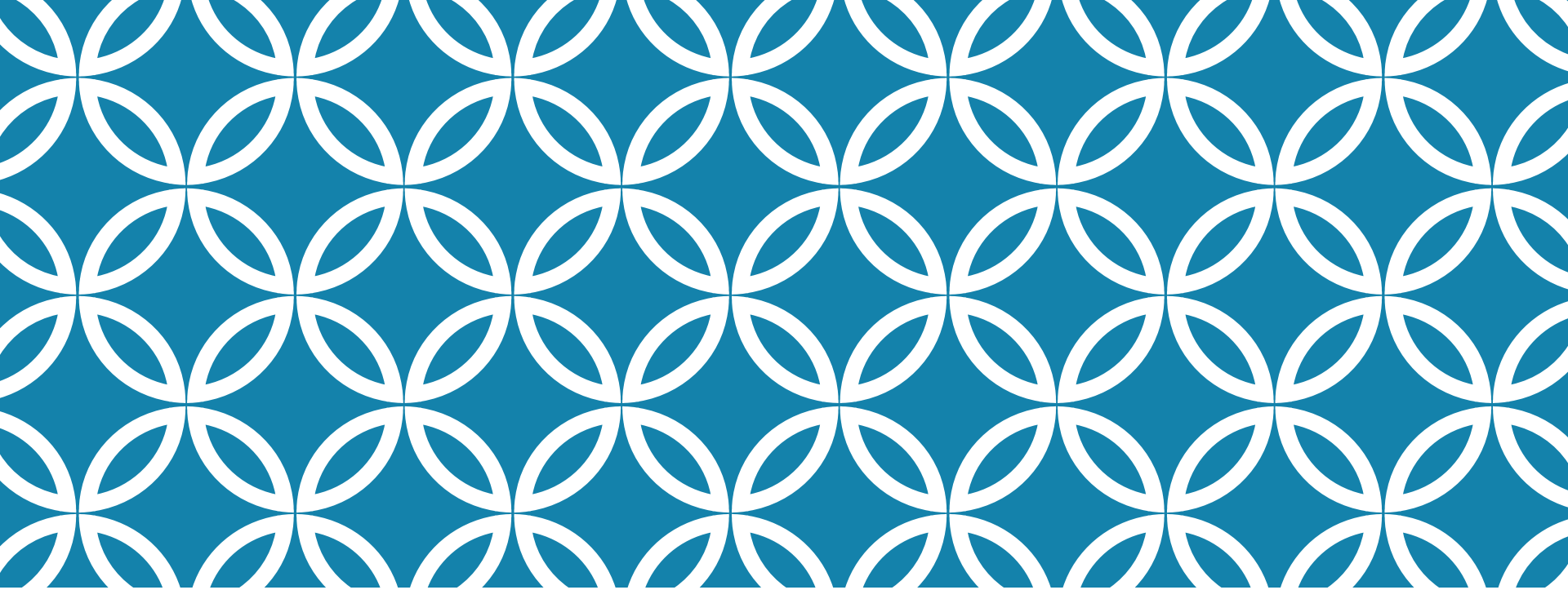


MONEYBALL FOR SOCCER

Instructor: Prof. Kylie Bemis

Yizhen Chen
Rohan Chouthai
Mayur Garudi

INTRODUCTION TO MONEYBALL

We came up with this idea from a film called “Moneyball”. In the film, the main character ‘Beane’ is facing a huge problem. He needs to build a team of undervalued talent by taking a sophisticated sabermetric approach towards scouting and analyzing players.

Suppose you are Beane, but this time you are lucky enough to have the whole world’s soccer players data in your hand. What will you do to change the game ??

THE DATASET – FIFA 18

We used the dataset “Fifa 18 More Complete Player Dataset” from Kaggle

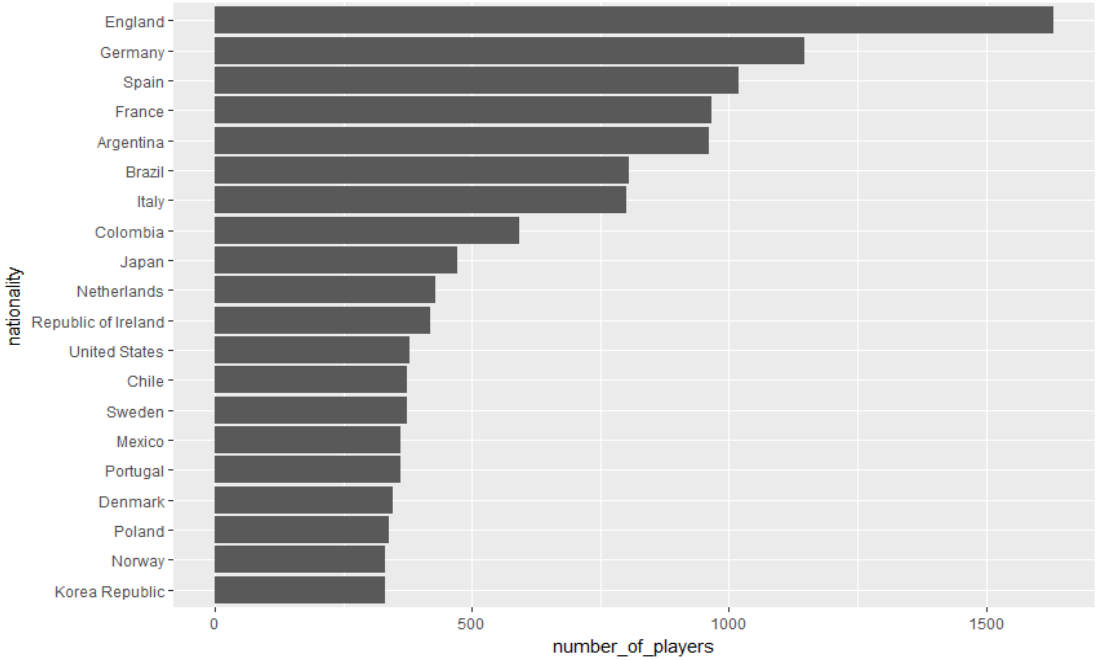
- It is a data consisting of around 18000 FIFA players worldwide with 180+ attributes
- These attributes can be generalized into 4 parts
  - Personal information like age, name, nationality
  - Physical attributes like attack, defense, speed, balance
  - Player’s performance on different positions on the field e.g. st, lcm, rb
  - Logical factors recording whether a player has a special ability or not

DATA CLEANING & FEATURE SELECTIONS

- First of all, we checked if data is in tidy format
- We checked for any NAs and class of data
- There were some logical attributes related to FIFA game which did not make sense for our analysis After identifying NA values and the undesired part from the dataset , we cleaned & subsetted the data
- We filtered the dataset with the player's age,league and nationality and formed new dataframes for further research.
- We extracted goalkeepers' data from the whole player dataset for goalkeepers modelling

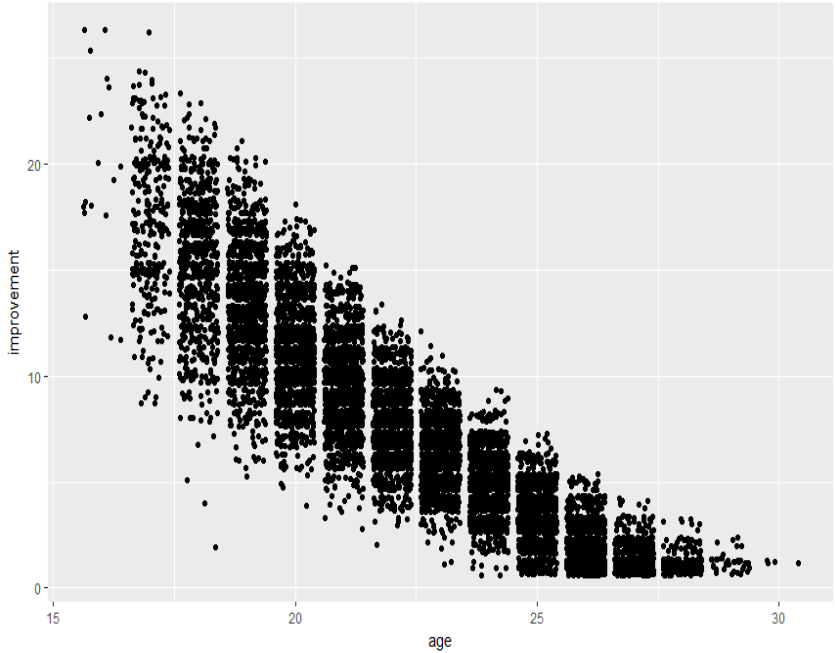
PRIMARY FINDINGS WITH DATA EXPLORATION

- Which country has the most number of soccer players in the world?
- First,we want to know the distribution of soccer players in the world by their nationality
- The result shows that most of the players belong to European countries followed by South American region



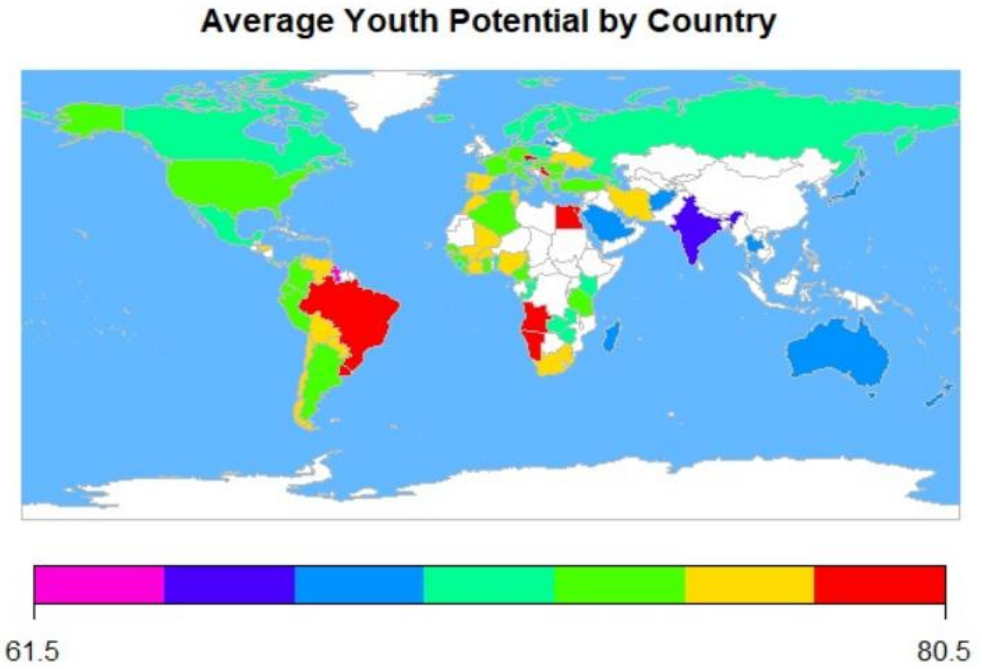
RELATIONSHIP BETWEEN AGE AND POTENTIAL

- As players grow older, we were curious to observe if there was any pattern between age & improvement window for him
- We can conclude that the potential is decreasing with the increment of age & we have a very small window of improvement for players above age 25
- It indicates that , clubs shall focus on the emerging talents. This will also enhance team budget strategy



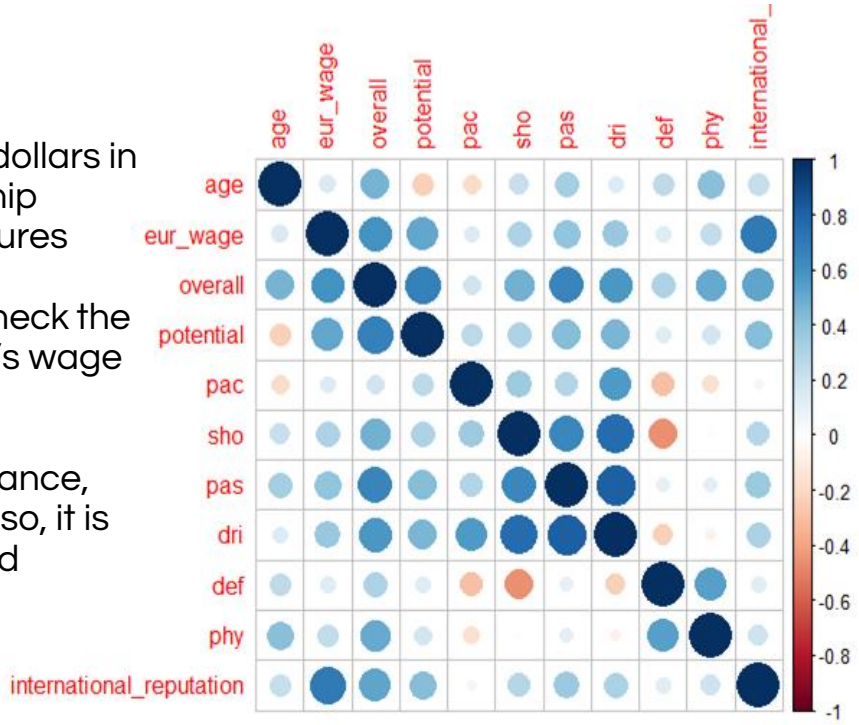
YOUTH POTENTIAL ACROSS GLOBE

- We can see that South American countries, West African countries and European countries have the players with highest potential
- We can observe the average young talent in the world of football is concentrated in Brazil, Poland, Angola & Turkey with >75%
- As European Soccer is getting enormously popular in North America , we can see the average potential of emerging talent is above 70 % in United States of America
- Venezuela has lowest young talent which is required to improve in future



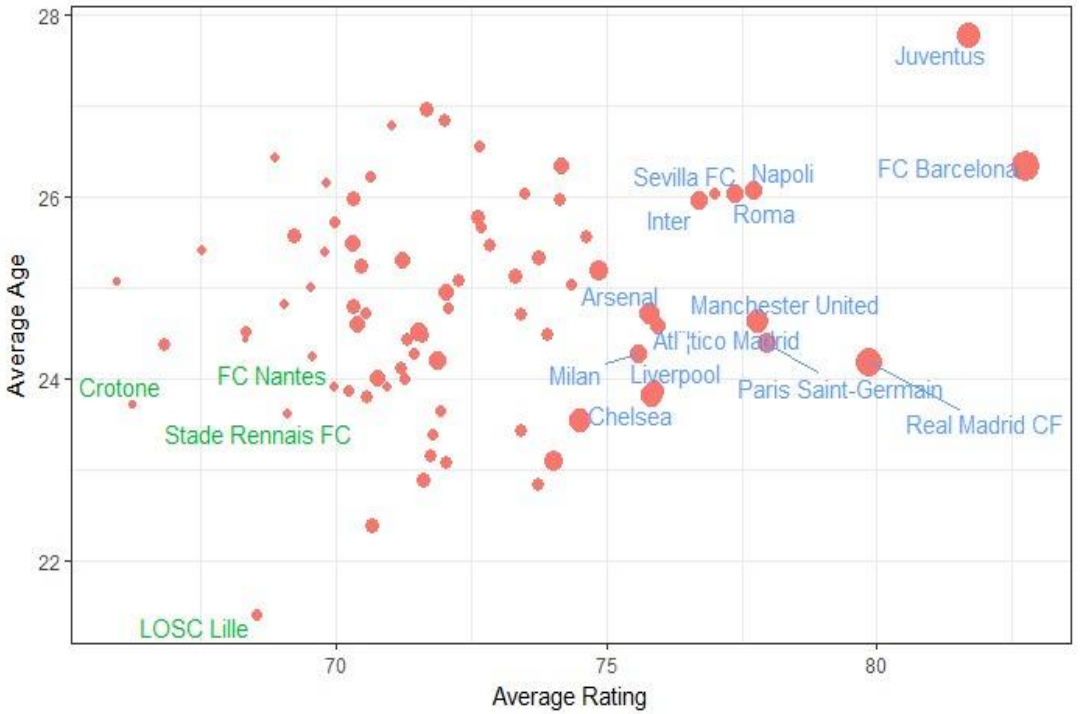
LET'S TALK MONEY

- For clubs which are investing billions of dollars in players, we wanted to explore relationship between wages and other possible features
- So we tried plotting correlation plot to check the impact of various parameters on player's wage
- We can observe that a player's wage is significantly affected by overall performance, potential and international reputation.Also, it is slightly affected by shooting,passing and dribbling



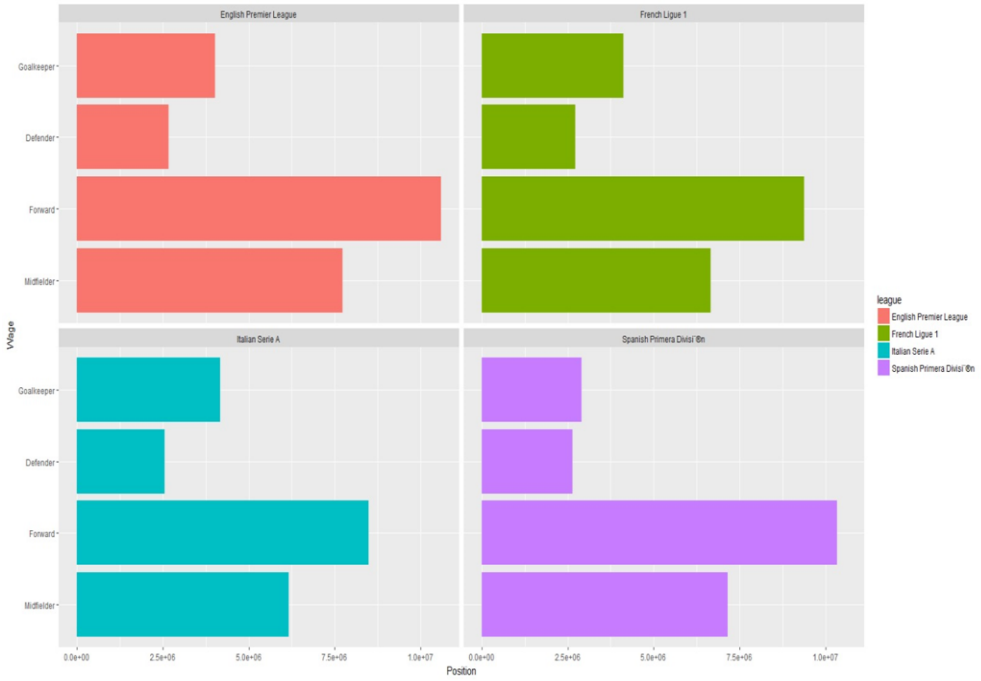
TOP CLUBS

- Every club has their financial budgeting and provides wages to players . we wanted to explore relationship between players ratings , their age & money involved in terms of wages
- Juventus and Barcelona do pay a lot of money for a relatively older squad
- Crotone & LOSC Lille seems to lack in both players rating & overall budget
- For others, the average rating of the players in the clubs is less for more aged players



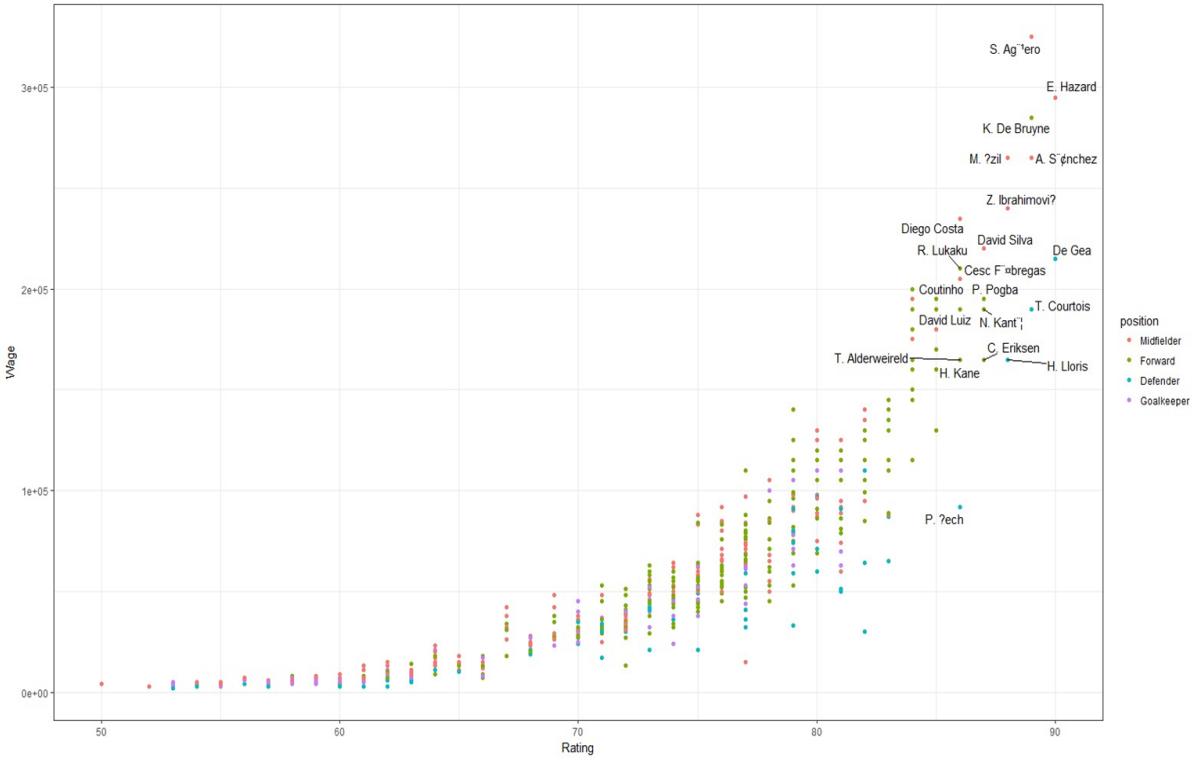
TOP LEAGUES- MOST VALUABLE POSITION

- When it comes to Soccer leagues , the players position plays a vital role towards match outcome
- As positions are not identified in dataset , we used K Means clustering (unsupervised learning module) to predict the preferred positioning of player
- We explored the wages of players in top 4 leagues with respect to their particular positioning on field
- The most valuable player is **Forward** in all leagues
- Defenders & goalkeepers are paid very less in comparison with midfielder & forward



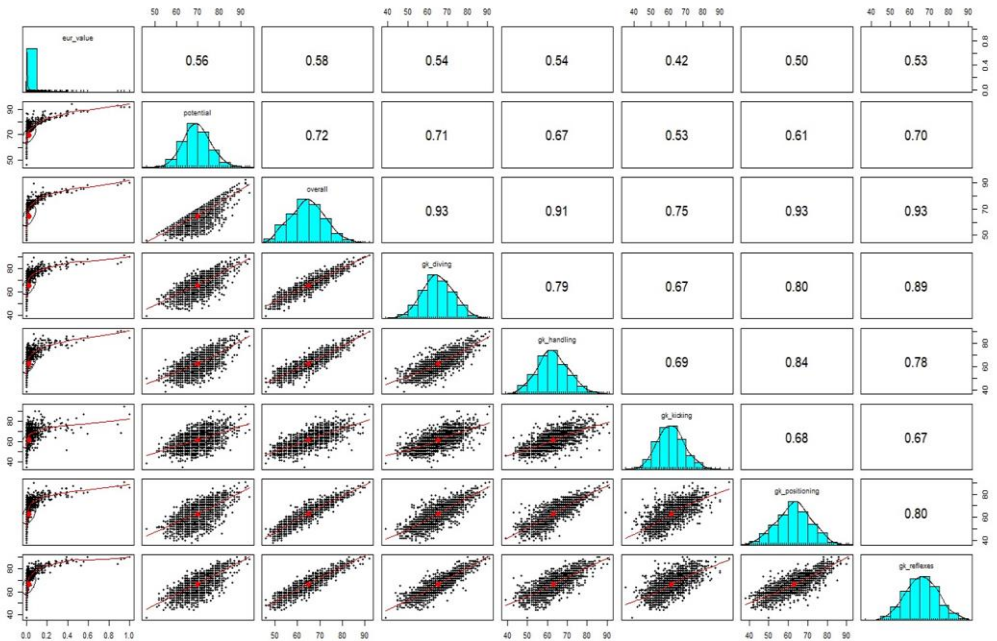
MOST VALUABLE PLAYERS- EPL

- We considered the players of English premier league for analysis of top paid players
- We can see that all the top rated players are strikers (Lukaku, Ibrahamovic, Aguero, Costa) right up there among the highest rating
- Liverpool and Tottenham Hotspur who are also big clubs pay their stars a bit less as can be seen from the graph (Coutinho and Kane).



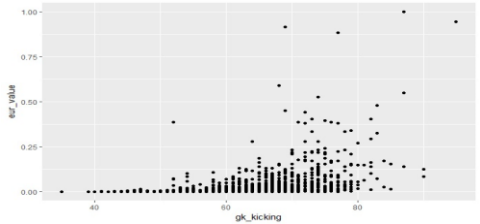
GOALKEEPER VALUE PREDICTION

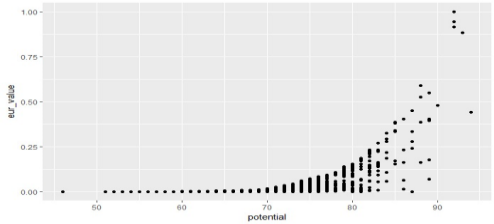
- For any club ,Goalkeeper is back bone
- The attributes for Goalkeeper are different than other positions
- We started plotting a correlation plot between goalkeeper's value against it's featured attributes.
- We observed a strong correlation with all the parameters (>0.5 in almost cases) . Hence individual pattern checking was performed

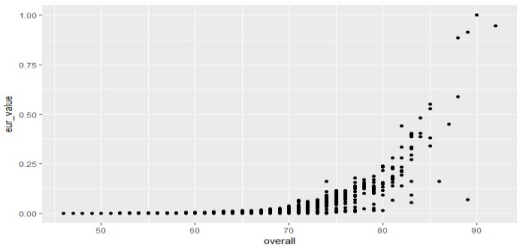


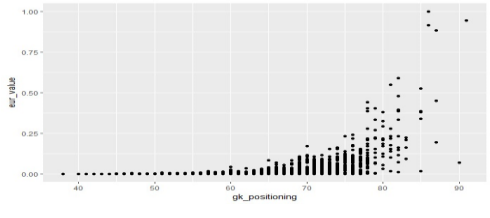
PREDICTOR PATTERN AGAINST ‘GK VALUE’

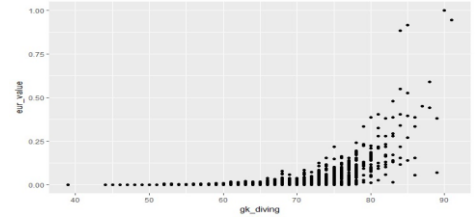
So we explored the pattern for each variable to shortlist predictors
We can see that the value of goalkeeper shows strong positive exponential pattern with ‘potential’, ‘diving’, ‘handling’ & ‘positioning’ but no significant pattern for ‘kicking’

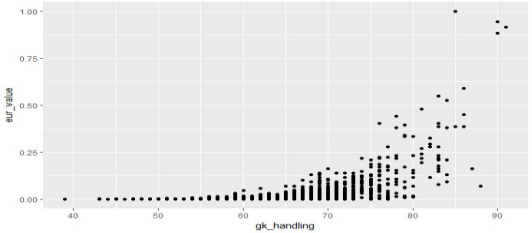






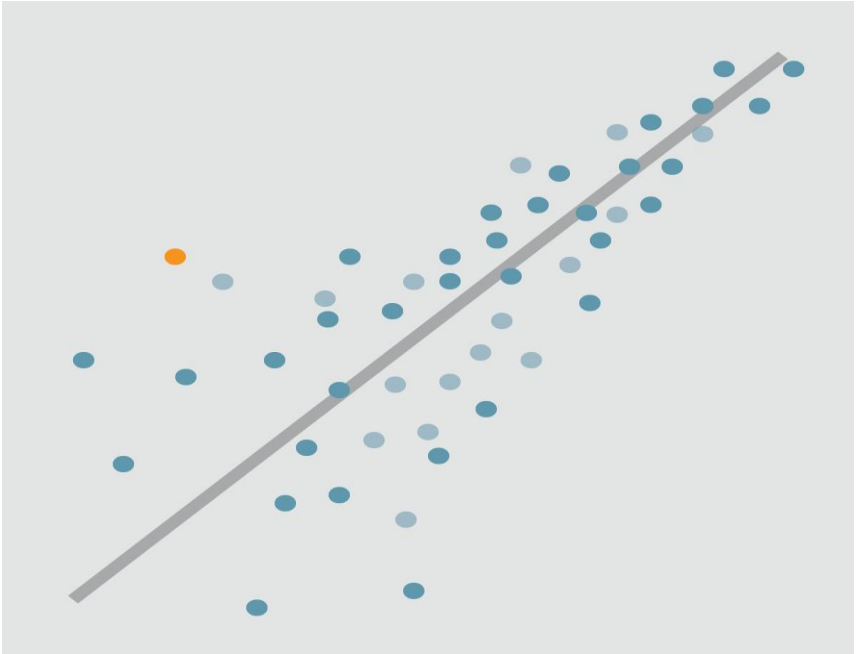






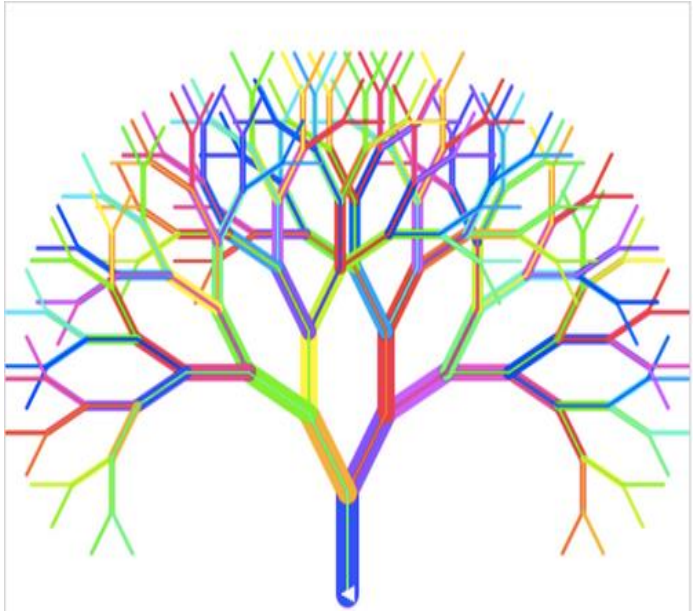
MODELLING- LINEAR REGRESSION

- By fitting linear regression model on train data to predict the value of goalkeeper on test data , the value for goalkeeper came out to be 45 million euros
- The rmse was 0.095 indicating a good quality of model



MODELLING- RANDOM FOREST & FUNCTION ACCURACY

- As we get the predicted value of goal keeper, we started exploring data for each player to check the accuracy of predictions
- Initially, we had 52 variable on players attributes. We then performed step regression to cut down predictors to 29 with the help of 'P value'
- Performed Random forest algorithm with model outcome of 96.31 % variance explained.Used accuracy function to get 91.49 % accuracy for ½ Million Euros tolerance



THANK YOU