Football Analysis

Archit Jain

15 March 2019

Problem Understanding

As we see these days the use of analysis is not just limited to traditional domains(Health, Finance, etc) but, it has also engaged into sports, and in return it has made sports more enjoyable. Its not only done at the national level for Olympics but it is also enjoyed (at regional level) by having leagues and different form of medium. And with the rise in Sports we need better analysis for managing the game, statistics and capital to take it to next level.

In this section we are going to look at the dataset of FIFA'19 and we are going to run some analysis on players and teams and see if we can get any fruitful results. Majorly we will be focusing on applying the Mining techniques to get some insights about the players, teams and countries and see how they are related. And as I wrap up the process I'll try to throw some anlysis on the Champions league final, as it is approaching this June'19 I would try to see whos in the better position (statistically) to win the finals.

Data Understanding

Lets try to unerstand the data for further analysis. First of all, we'll fetch the data from the link given here. If you are further interested in FIFA'18 or FIFA'17 or so data for, you can even scrape the data from here, which is the original source of data.

For Importing out data

```
library(readr)
fifa <- read_csv("F:/ML/seed/Fifa19/data.csv")
```

The size of our data

```
dim(fifa)

## [1] 18207    89
```

For looking at the information of our data and statistical distribution related to each label and their type

```
summary(fifa)
```

This would give the 5 rows of dataframe

```
View(fifa[1:5,])
```

After importing the data we can see the name of players, their respecting nationality and the club they play for, also they have have some attributes associated to them according to the position they play for, and the wages they get.

Data Preparation

As we saw earlier that the data was not clean (as in, it had numeric and characters value mixed), plus it had some unnecessary columns (like “flags”, “face type”, etc) which doesnt play any role in our dataset instead makes it more complex. So, In this section we are going to clean our data and make sure its ready to be tested on at the other section

First of all we will remove the “euro” sign from the Value and Wage column and replace the ‘M’ and ‘K’ (million and thousand) with respective numbers of zero by muliplying, and convert to numeric then, below I have just shown for **Wage** column, for ‘Value’ it can be interpreted in the same way

```
fifa$Wage <- gsub('[???]','',fifa$Wage)
fifa$lastW <- sapply(strsplit(as.character(fifa$Wage), ""), tail, 1)
extracting <- function(z){
  regexp <- "[[:digit:]]+"
  str_extract(z, regexp)
}
temp1 <- sapply(fifa$Wage, extracting)
fifa$Wage <- as.numeric(temp1)
fifa$Wage <- ifelse(fifa$lastW == "M", fifa$Wage * 1000000, fifa$Wage * 1000)
```

Same will be done for height and weight as well, we will convert them to numeric values and remove the wild characters we have in between, it will be simlpler than the previous case.

Modeling

Lets, come to an interesting part now, where we get to see what we can do with clean data. It can help us in analysis of the game without even knowing it too deeply.

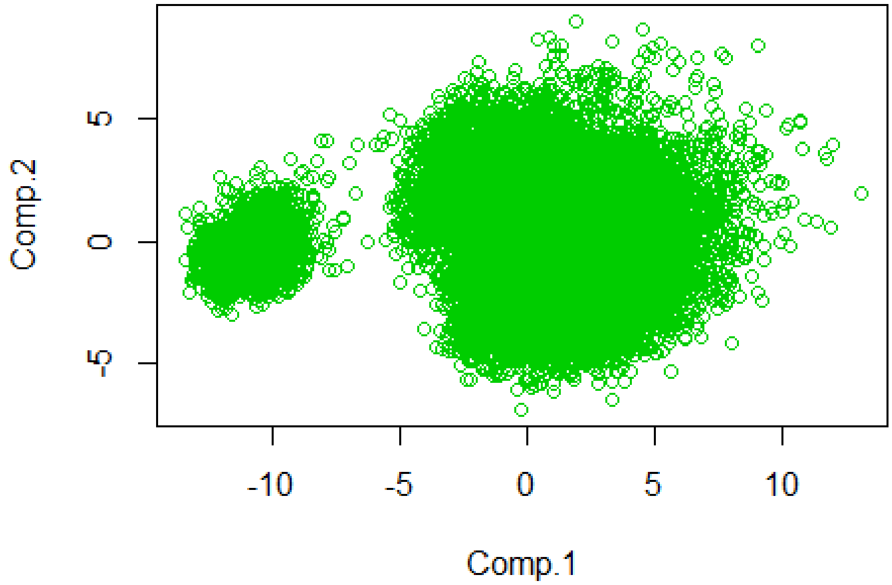
First of all, I’ll try to run the PCA and thereby reduce the number of columns into less features. The idea PCA works on is, it tries to reduce the number of dimension by maximizing the variance of components, which are then used to represent our data, and we select the principal components in a way that they cover 75% of our variance

This is how the components look like, here I have shown just 3 rows for illustration purpose

```
##         Comp.1   Comp.2   Comp.3    Comp.4    Comp.5
## [1,] 13.09823 1.969048 21.39557 -8.363483 -13.57854
## [2,] 11.79636 3.341420 17.96809 -3.435331 -10.97631
## [3,] 11.94903 0.560576 17.77507 -7.114749 -11.24405
```

lets look at the plot drawan by PCA and see what its trying to say

```
plot(df1, col=3)
```



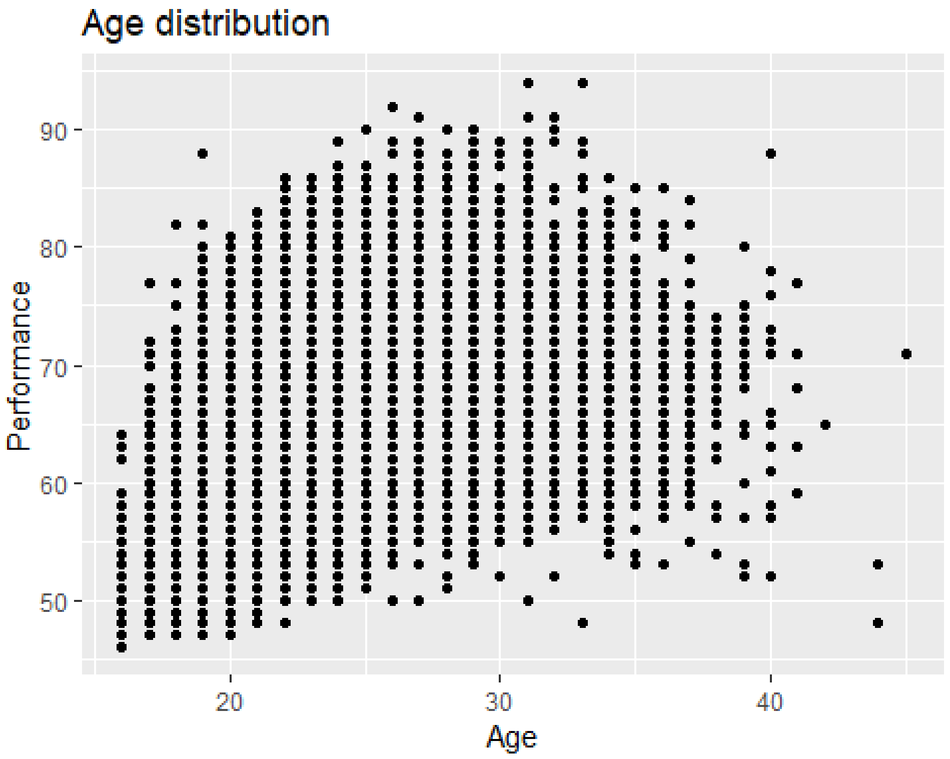
Here we can see primarily two clusters, one cluster represents goalkeepers and another cluster (which is in return is the pack of few cluster) represents another members like “defender”, “striker”, “midfielder”. And hence, we can see how the players are seperated without even having them to classify

Evaluation

This will be the most exciting part, where I have tried to guess the UEFA champions for 2019 and young player (which every club would be targetting for) and some stats analysis related to country and their performance

As we know that performance of player increases untill a certain age(which can be see below) and then start to decline, and this strategy can be used by managers to hire young players which have high potential and sign them for 5 years

```
plot
```



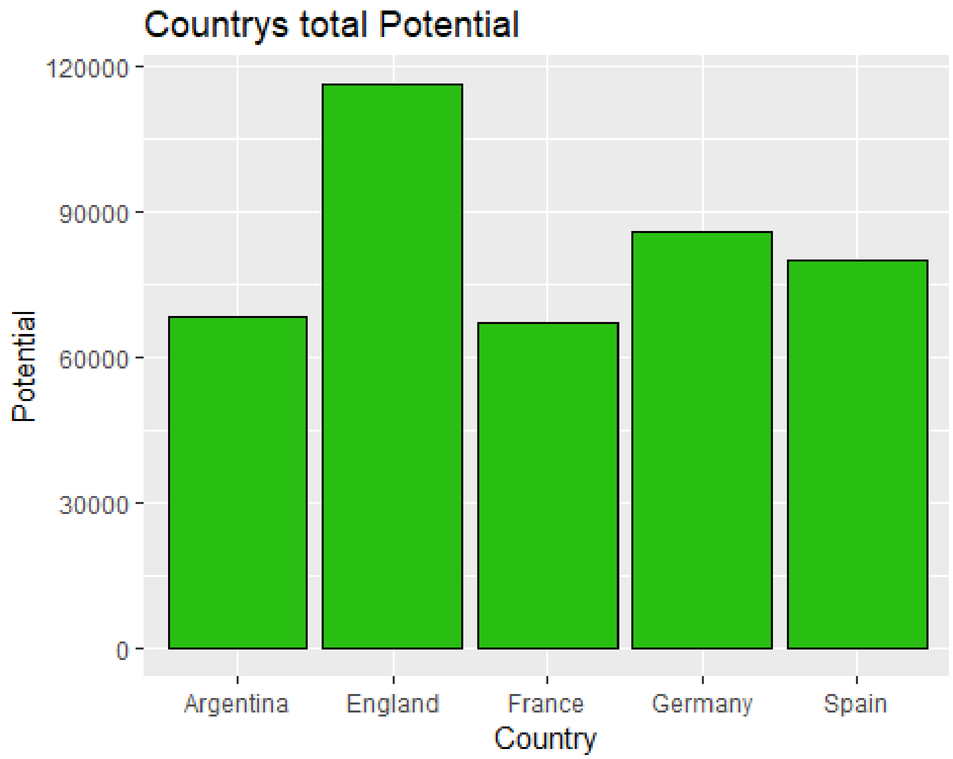
Now,lets try to see who would/should be the top 5 young players for FIFA'19, whos Overall is more than 80 and are less than 20

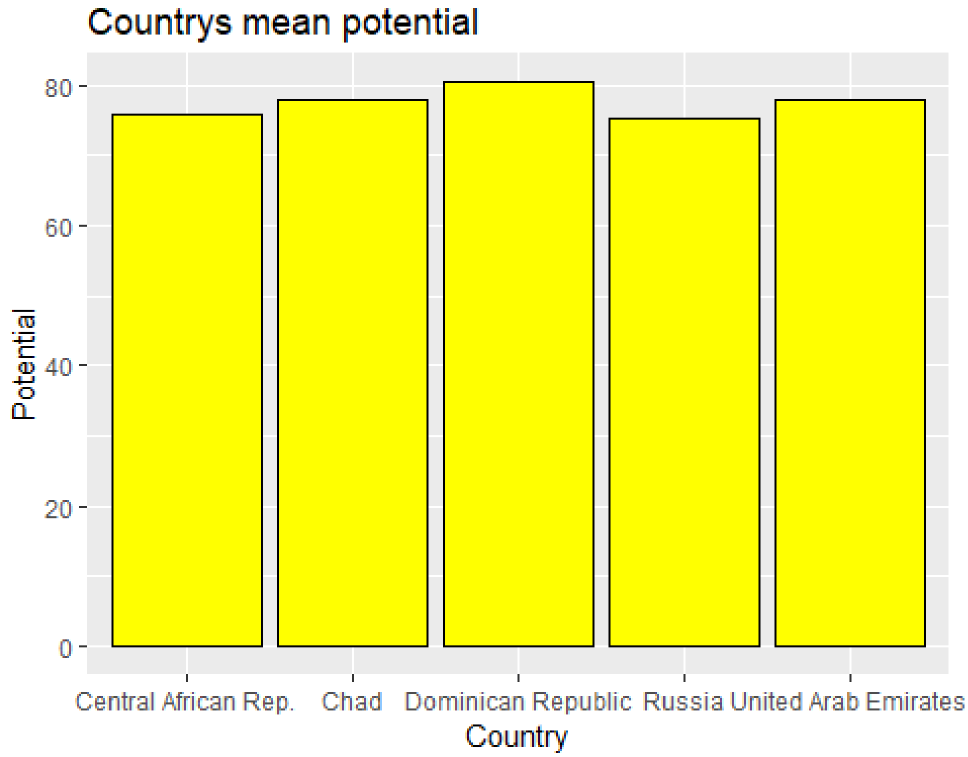
```
## # A tibble: 5 x 6
##   Name            Age Nationality Overall Potential Club
##   <chr>         <dbl> <chr>         <dbl>     <dbl> <chr>
## 1 K. Mbappé        19 France           88        95 Paris Saint-Germain
## 2 M. de Ligt       18 Netherlands      82        91 Ajax
## 3 G. Donnarumma    19 Italy            82        93 Milan
## 4 M. Rashford      20 England          81        89 Manchester United
## 5 L. Bailey        20 Jamaica          81        89 Bayer 04 Leverkusen
```

But Managers also eye for the players which are young as 16 and 17 and have a good potential

```
## # A tibble: 5 x 5
##   Name              Age Nationality    Potential Club
##   <chr>           <dbl> <chr>              <dbl> <chr>
## 1 Vinícius Júnior    17 Brazil                92 Real Madrid
## 2 A. Davies          17 Canada                87 Vancouver Whitecaps FC
## 3 Paulinho           17 Brazil                86 Bayer 04 Leverkusen
## 4 Kangin Lee         17 Korea Republic        88 Valencia CF
## 5 C. Hudson-Odoi     17 England               87 Chelsea
```

Now, lets look at two analysis, which talks about the Potential of country in different terms

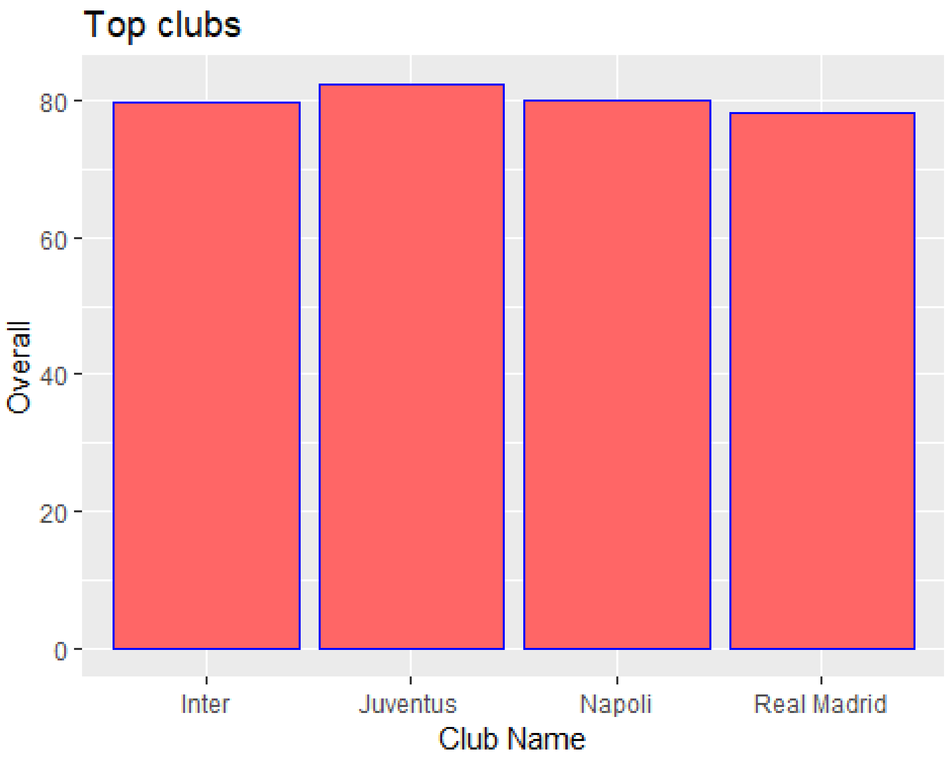




As we can see the results are really surprising, in terms of total potential (it was expected to have above countries), but, when we talk about the average performance of a country (i.e. total performace/no of players), the results are unexpected.

Now lets come to the **verdict** of our UEFA predictions, the predictions here are based on statistical performance of each player in a team

The results really looks promising by seing Juventus on top, with Ronaldo and Dybala on the same side, and the world class defence they have, they can really break the winning streak of Spanish teams this time.



Conclusions

The results look promising by seing Juventus on top, but still its mere an analysis (it didnt have Barcelona, which is again one of the strongest team), the reason being, it doesnt consider lots of factor like team work and compatibility and blunders. We can get close to our predictions by having a better data (eg: the matches played in past, the win/lose percentage). Also, there are lot of other factors like Manager, squad, formation, strategy. If we have data considering all this factors and somehow we could come up with a better way provided the complexity, our results would be more promising.

References

- https://www.datacamp.com/community/tutorials/15-easy-solutions-data-frame-problems-r
- https://www.kaggle.com/aishwarya1992/fifa-data-analysis-player-value-prediction/comments#L101
- https://data-flair.training/blogs/r-data-frame/