# Use R Trends in FIFA 2019 Analysis

How Visualization with R Helps Us to Analyze in FIFA 2019 Data

Professor: Dazi Chong

Social Media Analytics

**Anahita Bahraniasl**

**Chan Chou Yu**

**Meet Joshi**

**Mehra Sayah**

1. **Introduction (Mehra)**

Since FIFA 2019 is one of the most important cup games with huge fans so we tried to analyze different players attribute which includes interesting info for those fans.one of the most important cup games with huge fans so we analysis different players attribute which includes interesting info for those fans.

1. **Data (Mehra)**

The data that we have collected and chosen for our analysis is from here: <https://github.com/amanthedorkknight/fifa18-all-player-statistics/tree/master/2019>, which includes every player attributes such as club, age, overall, position and nationality who registered in the FIFA 19 database. The reason we believe this data is good for our analysis is that it allows us to do a variety of interesting data analysis for the soccer community. First of all, we clean the data so we could apply analysis tools which are beneficial to determining different comparisons.

1. **Methodology and results**

**2.1. Analyzing players age (Mehra)**

One of the interesting subjects for football consultants, coaches and fans is research into the topic of relative age effects. They need to analyze the age profiles of the teams involved in sports teams and find out how these may have impacted the competition.

We analyze players age in FIFA 19 dataset, the R code that we used to find this data is provided below:

First, we calculated the average age of players in the fifa2019 dataset which is approximately twenty-five years old.

**R code:**

> mean(data$Age)

[1] 25.12221

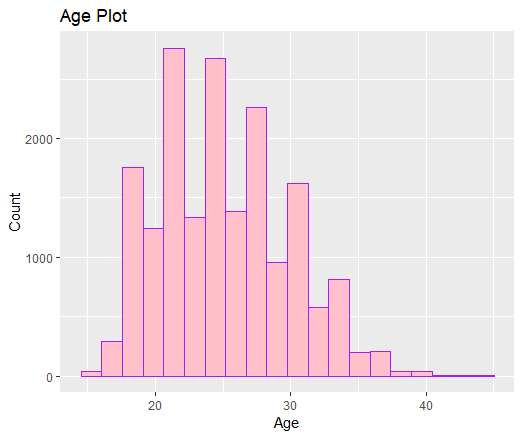
Second, plotting age for analysis

**R code:**

agePlot <- ggplot(data = data, aes(data$Age)) + geom\_histogram(color='purple',fill = 'pink',bins = 20) +

+ labs(x='Age',y='Count',title='Age Plot')

> agePlot



Third, Comparing of Total Rating by Age

**R code:**

#Creating Age bins

> for(x in 1:nrow(data)){

+ sample <- data[x,]

+

+ if(sample$Age >=16 & sample$Age <=19){

+ data[x,'Age\_Bin'] <- '16-19 years'

+ } else if(sample$Age >=20 & sample$Age <=24){

+ data[x,'Age\_Bin'] <- '20-24 years'

+ } else if(sample$Age >= 25 & sample$Age <= 29){

+ data[x,'Age\_Bin'] <- '25-29 years'

+ } else if (sample$Age >= 30 & sample$Age <= 35){

+ data[x,'Age\_Bin'] <- '30-35 years'

+ } else {

+ data[x,'Age\_Bin'] <- '35+ years'

+ }

+ }

>

#Comparing of Total Rating by Age

> data %>%

+ select(Overall,Age\_Bin) %>%

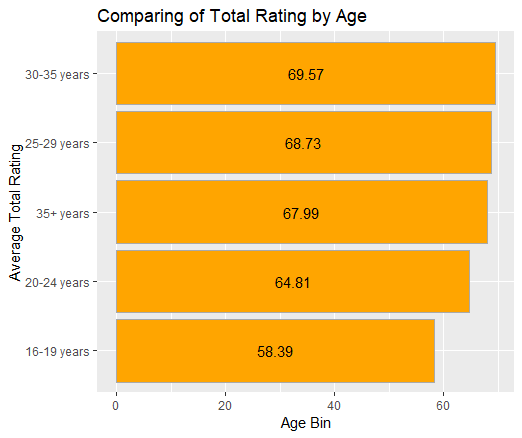
+ group\_by(Age\_Bin) %>%

+ summarise(n=mean(Overall, na.rm = TRUE)) %>%

+ ggplot(aes(x=reorder(Age\_Bin,n),y=n)) + geom\_bar(stat = 'identity',position = 'identity',fill='orange',color='darkgray') + coord\_flip() +

+ labs(x='Average Total Rating',y='Age Bin',title='Comparing of Total Rating by Age') +

+ geom\_text(aes(label=round(n,2)),position=position\_stack(vjust=0.5))



**2.2. Analyzing the Positions of players (Anahita)**

In soccer players position themselves on the pitch based on specific formations. These positions generally define if a player has a defensive or attacking role. They also define whether a player tends to position to a side of the pitch or at the center. The table below provides the soccer player positions and their abbreviations (abbr.).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Abbr.** | **Position** | **Abbr.** | **Position** | **Abbr.** | **Position** |
| GK | Goal Keeper | CM | Centre Midfielder | CF | Centre Forward |
| CB | Centre Back | LM | Left Midfielder | LF | Left Forward |
| LB | left Back | RM | Right Midfielder | RF | Right forward |
| RB | Right Back | LCM | Left Centre Midfielder | LW | Left Winger |
| LCB | Left Centre Back | RCM | Right Centre Midfielder | RW | Right Winger |
| RCB | Right Centre Back | CAM | Centre Attacking Midfielder | ST | Striker |
| LWB | Left Wing Back | LAM | Left Attacking Midfielder | RS | Right Striker |
| RWB | Right Wing Back | RAM | Right Attachking Mid fielder | LS | Left Striker |
|  |  | CDM | Centre Defensive Midfielder |  |  |
|  |  | LDM | Left Defensive Midfielder |  |  |
|  |  | RDM | Right Defensive Midfielder |  |  |

To analyze player positions in FIFA 19 dataset, first, we need to find out which positions are in the dataset and what their frequencies are. The R code that we used to find this data is provided below. In this code first, we used the table function to extract the position categories and frequency. Next, we put the position data into a new variable named position count. In the end, we plot the results using the plot function. The resulted chart is provided below.

**R code:**

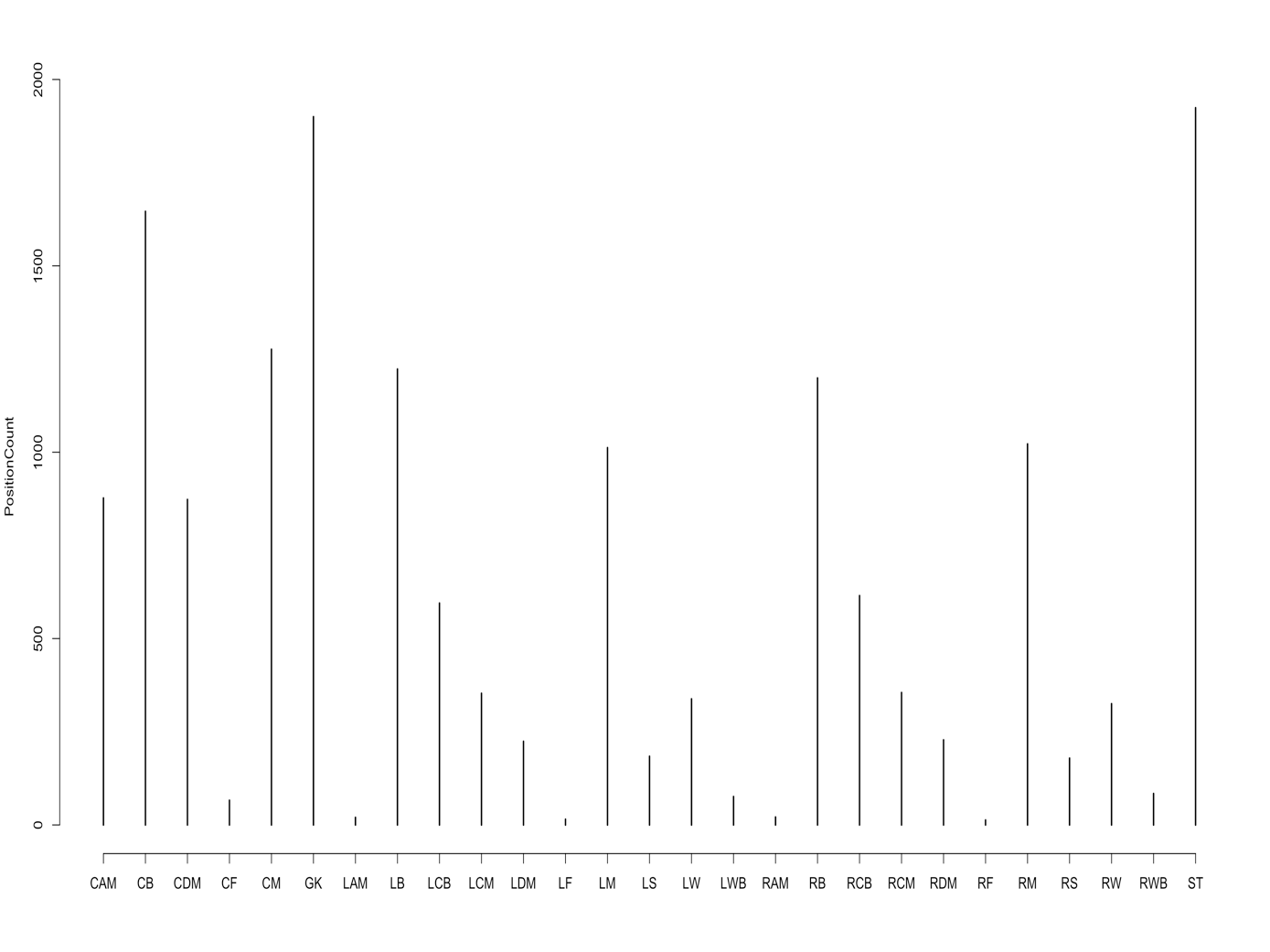
> table(fifa19clean$Position)

CAM CB CDM CF CM GK LAM LB LCB LCM LDM LF LM LS LW LWB RAM RB RCB RCM RDM RF RM RS RW RWB ST

877 1646 873 66 1276 1900 20 1223 595 353 224 15 1012 184 338 76 21 1199 615 355 228 13 1022 179 325 84 1924

> PositionCount=table(fifa19clean$Position)

>plot(PositionCount)



This chart shows the frequency of player positions in the dataset. As the chart shows, the most frequent position is Striker (ST) followed by Goal Keeper (GK) and Centre Back (CB). The least frequent positions are Right Forward (RF), Left Forward (LF), and Left Attacking Midfielder(LAM) in order.

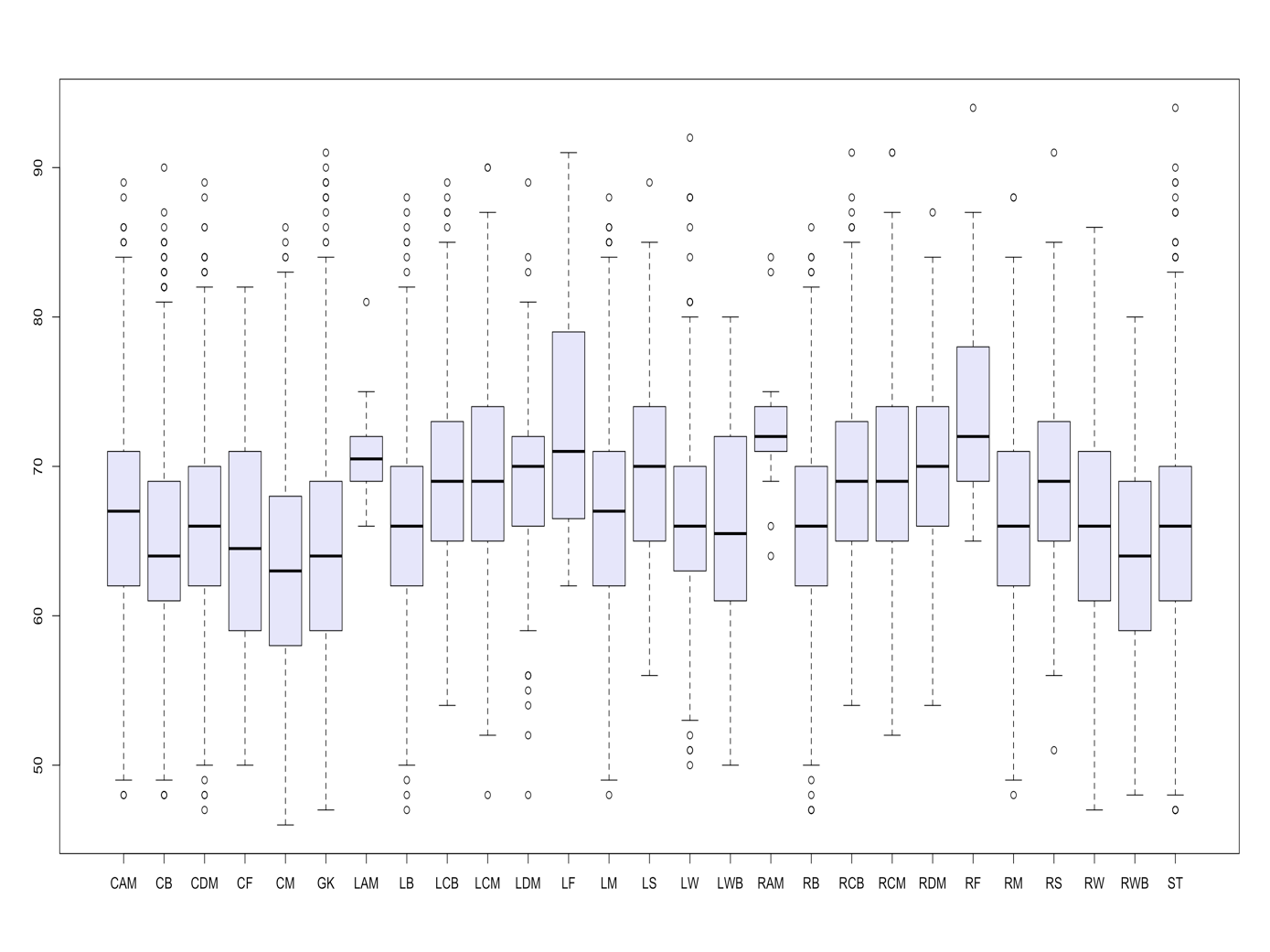
**Distribution of Overall rating of Players by positions**

Another parameter that we used in order to analyze player positions is the overall rating of players. Although teams ideally want to have players with the highest ratings, these ratings differ based on the position of the players. For instance, it can be hard to find a player with the overall rating above 85 in certain positions while for other positions you can find multiple players with the rating above this number.

In order to find a better understanding of the distribution of ratings over player positions, we used R code to create a box plot. For this purpose, first, the split function was used to divide the overall rating values based on the position groups. We put the result of this function into a variable called posbyoverall. Then, we created a boxplot using the data in posbyoverall. The R code and box plot diagram is provided below.

**R code:**

> posbyoverall<-split(fifa19clean$Overall,fifa19clean$Position)

>boxplot(posbyoverall,col="lavender")This boxplot diagram helps us see a wide variety of information. We can see the distribution of the overall rating for each position. We are able to measure the variability of overall ratings by analyzing the interquartile range of overall ratings for these positions. We can also compare the median overall ratings of different positions. Furthermore, this diagram shows us the minimum and maximum ratings and outliers. For instance, based on this diagram we can see that Left Forward (LF) position has the highest maximum overall ratings while the Striker (ST) and Right Forward (RF) positions have the highest outliers. This information can help the team managers to know what player rating range do they need to look for when they are scouting for different positions.

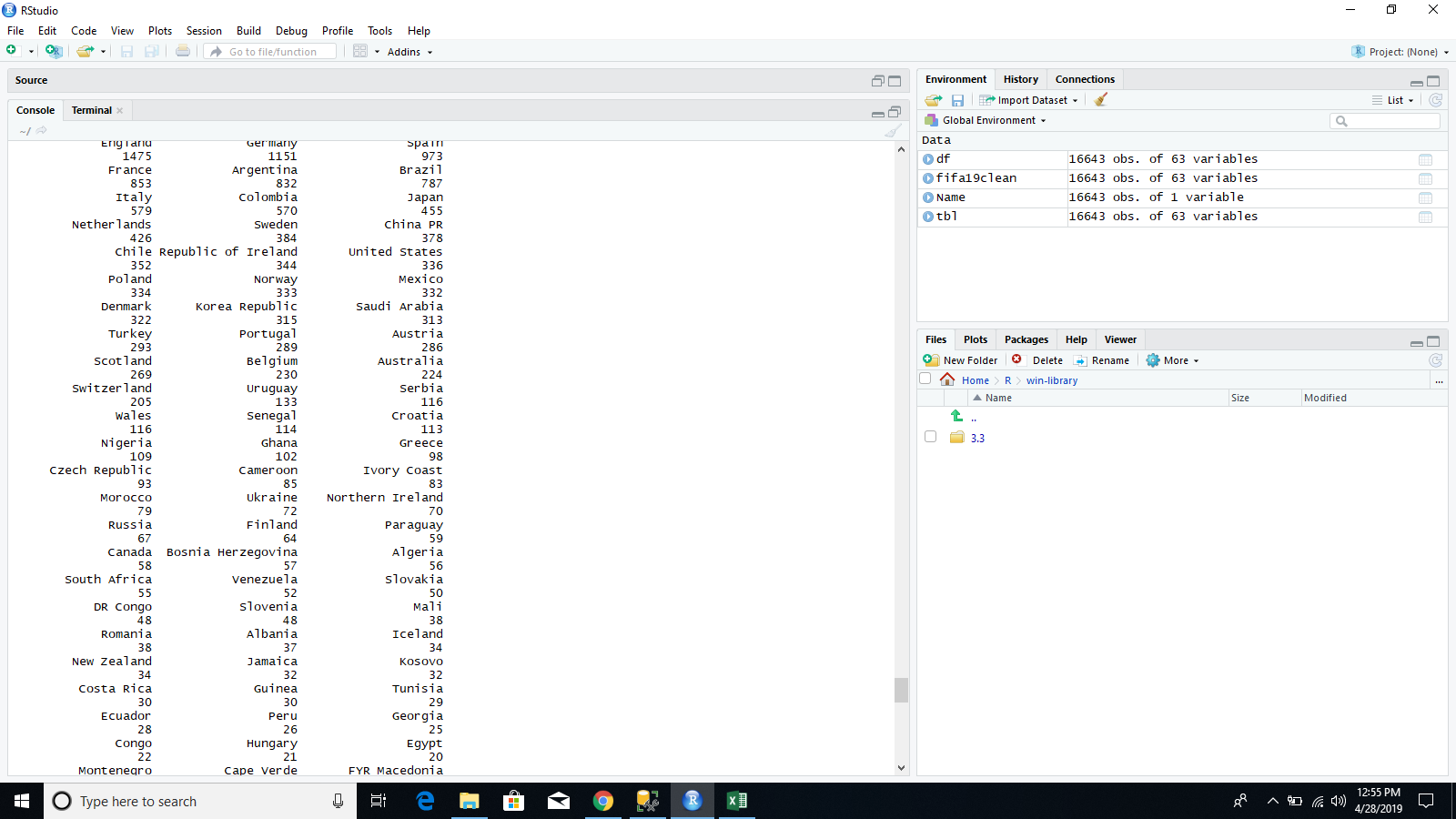
**2.3. Analyzing the player and their nationality (Meet)**

As we all know that soccer is a very popular game around the world and almost in every country, this game is played. In the international level tournament, the teams divide by their country. Here, we have used R code to analyze players nationality. The result below shows that England has most numbers of players playing soccer which is 1475 following by Germany and Spain which are 1151 and 973.

**R code:**

Name <- tbl["FIFA19CLEAN"]

print(summary(Name[-1:-5,])

****

**Using ggplot, analyze the number of players with their Potential.**

We have used R code to analyze the potential of players based on their capability of scoring goals. We have used ggplot to visualize it in a better way. ggplot2 is a data visualization package for the statistical language R. Its an implementation of Leland Wilkinson's Grammar of Graphics—a general scheme for data visualization which breaks up graphs into semantic components such as scales and layers.

**R code:**

Parsed with column specification:

cols(

.default = col\_integer(),

Name = col\_character(),

Nationality = col\_character(),

Club = col\_character(),

Preferred.Foot = col\_character(),

Work.Rate = col\_character(),

Body.Type = col\_character(),

Position = col\_character(),

Joined = col\_character(),

Loaned.From = col\_character(),

Height = col\_character(),

Weight = col\_character(),

LS = col\_character(),

ST = col\_character(),

RS = col\_character()

)

See spec(...) for full column specifications.

> head(fifa19clean[1:5])

# A tibble: 6 × 5

ID Name Age Nationality Overall

<int> <chr> <int> <chr> <int>

1 158023 L. Messi 31 Argentina 94

2 20801 Cristiano Ronaldo 33 Portugal 94

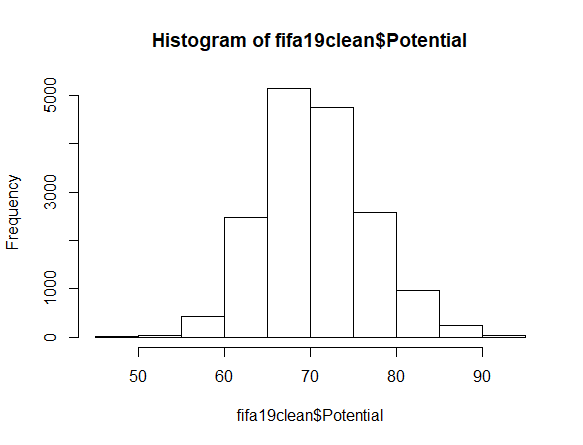
3 190871 Neymar Jr 26 Brazil 92

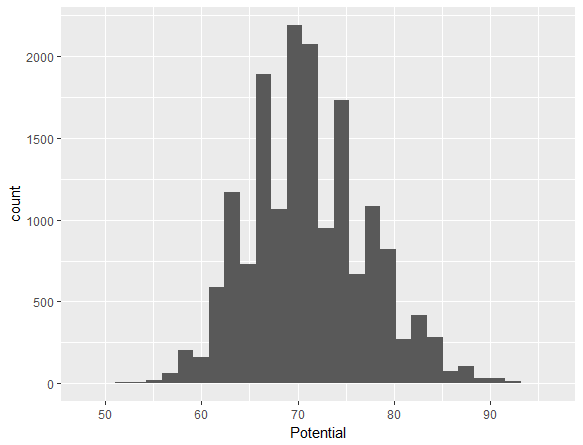
4 193080 De Gea 27 Spain 91

5 192985 K. De Bruyne 27 Belgium 91

6 183277 E. Hazard 27 Belgium 91

hist(fifa19clean$Potential)



****

In the Histogram above, we can see that the majority of players potential during the Fifa 2019 lies between 65 to 80. The potential is measured by player’s overall rating and their current rating.

**Analysis of players wears the same number of Jersey**

We have used R code to analyze which number of jersey soccer players wears and the result shows that jersey number between 1 to 40 are favorite for most of the players. We have also used Histogram for better visualization.

R code:

Parsed with column specification:

cols(

.default = col\_integer(),

Name = col\_character(),

Nationality = col\_character(),

Club = col\_character(),

Preferred.Foot = col\_character(),

Work.Rate = col\_character(),

Body.Type = col\_character(),

Position = col\_character(),

Joined = col\_character(),

Loaned.From = col\_character(),

Height = col\_character(),

Weight = col\_character(),

LS = col\_character(),

ST = col\_character(),

RS = col\_character()

)

See spec(...) for full column specifications.

A tibble: 6 × 17

ID Name Age Nationality Overall Potential Club Value Wage Special

<int> <chr> <int> <chr> <int> <int> <chr> <int> <int> <int>

1 158023 L. Messi 31 Argentina 94 94 FC Barcelona 110000000 565000 2202

2 20801 Cristiano Ronaldo 33 Portugal 94 94 Juventus 77000000 405000 2228

3 190871 Neymar Jr 26 Brazil 92 93 Paris Saint-Germain 118000000 290000 2143

4 193080 De Gea 27 Spain 91 93 Manchester United 72000000 260000 1471

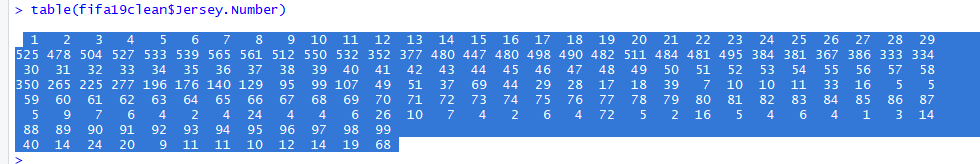
5 192985 K. De Bruyne27 Belgium 91 92 Manchester City 102000000 355000 2281

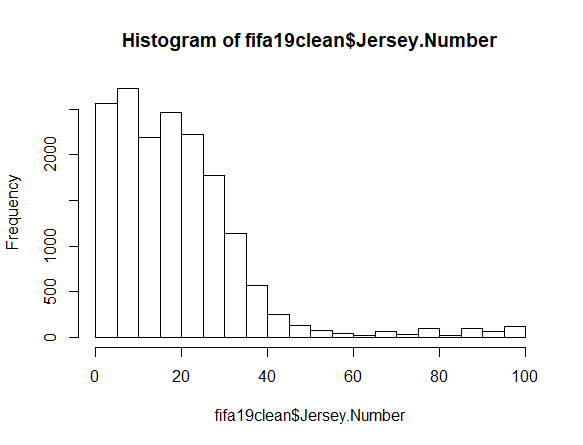
6 183277 E. Hazard 27 Belgium 91 91 Chelsea 93000000 340000 2142

# ... with 7 more variables: Preferred.Foot <chr>, International.Reputation <int>, Weak.Foot <int>,

# Skill.Moves <int>, Work.Rate <chr>, Body.Type <chr>, Position <chr>

table(fifa19clean$Jersey.Number)

****

****

In the Histogram above, it is measured that the favorite jersey number for most of the players is between 1 to 40

**2.4. Analyzing Players physical fitness (Jim)**

After we collect data from FIFA, we can use supervised function such as classification or clustering to group them. One useful way is partitioning these data using clustering method, then compare the result to some classifier, then, we can try to find some interesting result.

There are lots of techniques to implement clustering analysis, including partitioning methods, hierarchical methods, density-based methods, grid methods. We choose to adopt partitioning approach because it’s easy and fundamental, which organizes the objects of a set into several exclusive groups or clusters.

In partitioning methods, we choose k-means as the method to cluster our data. K-means will treat all the columns as part of the data in one vertex, and choose some centric randomly, comparing the distance of all vertices with these centric and re-group them. The algorithm will do the process iteratively until all points belong to the cluster with shortest distance. Only useful physical fitness numeric data will be used in these clustering, since the other data isn’t involved in fitness examination and normalization them isn’t worthwhile.

|  |  |
| --- | --- |
| column | Data type |
| ShortPassing | int |
| Volleys | int |
| Dribbling | int |
| Curve | int |
| FKAccuracy | int |
| LongPassing | int |
| BallControl | int |
| Acceleration | int |
| SprintSpeed | int |
| Agility | int |
| Reactions | int |
| Balance | int |
| ShotPower | int |
| Jumping | int |
| Stamina | int |
| Strength | int |
| LongShots | int |
| Aggression | int |
| Interceptions | int |
| Positioning | int |
| Vision | int |
| Penalties | int |
| Composure | int |
| Marking | int |
| StandingTackle | int |
| SlidingTackle | int |
| GKDiving | int |
| GKHandling | int |
| GKKicking | int |
| GKPositioning | int |
| GKReflexes | int |

Since all the data has been clean and can be used directly, we used R-code and K-means function to cluster them. The R code is as below. In this case, we chose k=10 because lots of statistic about athlete performance are use 10 degree to categorize them.

**setwd("D:\\");**

**df<-read.csv("D:\\data0.csv")**

**//only use fitness data**

**df1<-df[55:87];**

**//run k-means**

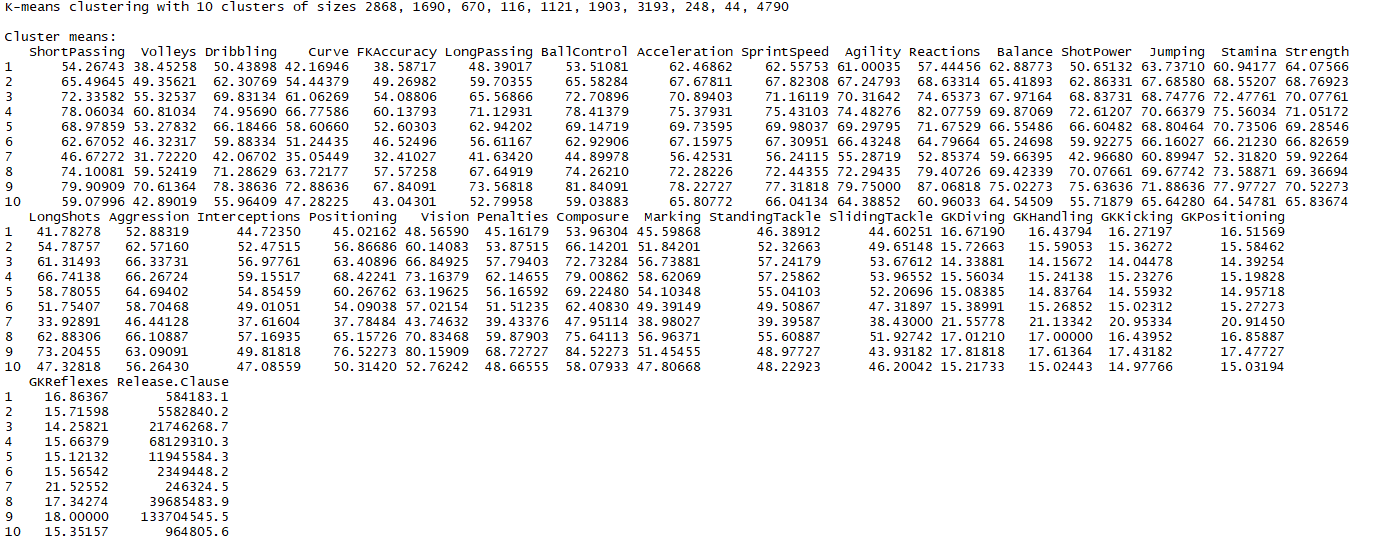
**kmeans.result <- kmeans(df1, 10)**

**print(kmeans.result)**

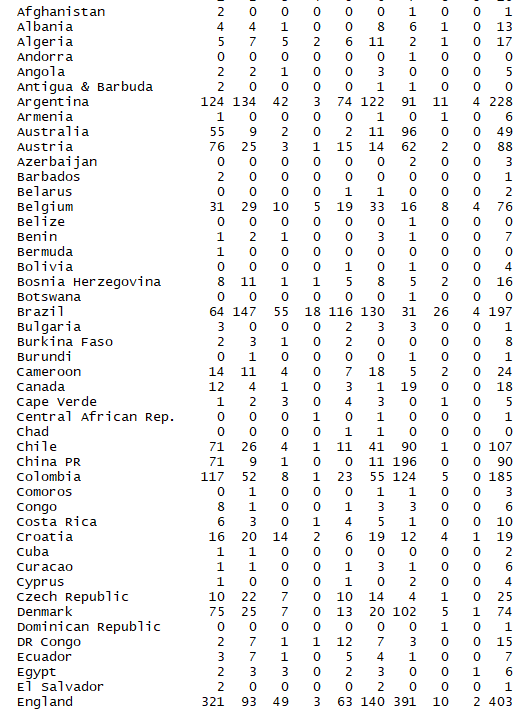
**//compare the clustering result with the classifier ‘nationality’**

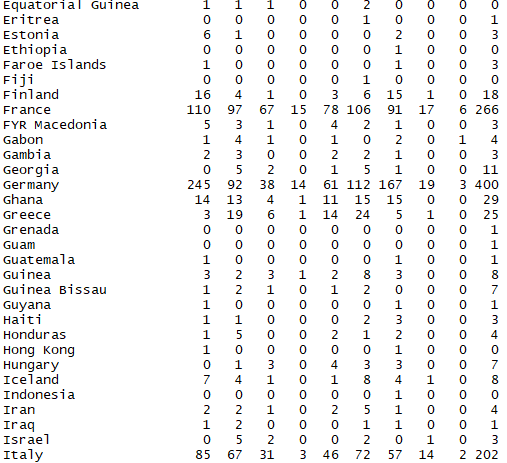
**table(df$Nationality, kmeans.result$cluster)**

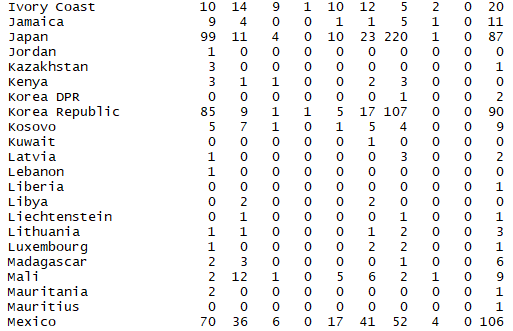
We can see the clustering result as below



And comparing with the classifier ‘nationality’.







We can analyze these clustering data and notice that there are few players in cluster 9, no matter in traditional soccer power, such as Argentina, Brazil, England, France, Germany, Italy, Portugal, Spain, or other countries. There are two possibilities, the first one is ‘these players are elite so they are minority’, we can use these data to compare with another classifier ‘salary’ in the future, if it’s the case, we can use the clustering as another classifier, and use cluster 9 as the standard to determine if a player is an ‘elite’. The other possibility is ‘these player are not useful, so only some few player in this category can play in FIFA.

1. **Discussion and conclusion**
2. **References**