MONASH University

UNIT CODE: FIT5147

DATA EXPLORATION AND VISUALIZATION

Project Title: FIFA 19 Data Exploration and Analysis

Name: Anand Rane

Student ID: 29934176

Tutor: Mohammad Haqqani

INDEX:

I. Introduction:

Football is not just a game, its an emotion for many. People follow their favorite Clubs no lesser than their Religion! Great Players are celebrated all over the world. But not everyone knows how much are these players' paid?, what contributes to their Market value?

The Fédération Internationale de Football Association (FIFA) is an organization which describes itself as an international governing body of association football. Using the 2019 Player Dataset by FIFA, we will try to answer some interesting questions:

1) What are the factors responsible for player's overall performance and market value.
2) Why are the young players being paid more than aged and experienced one's?
3) Does the preferred foot affect the positioning of the players, can the left footed players play on the right positions and vice versa? Is the fight for the position easier for the left footed players ?

I am myself a Football player and a enthusiast and hence chose this topic for my project, since a very long time I often wondered why is Messi (left footed) better than Ronaldo (right footed), what can be the reasons to that, are the left footed more talented ? And why are the young players like K. Mbappe are being highly paid, since they have just started their career, aren't they supposed to be payed less than the experienced one's? And also what are the factors which are assessed while counting the overall rating of a player, I had the curiosity in me to get the answers to these questions, and which is why I was motivated to take this up as a challenge to analyse FIFA datasets as my Exploration project.

II. DATA WRANGLING:

Here, instead of directly flicking the dataset from the Kaggle I scraped the data from the website called sofifa.com, which is a well-maintained website and has its dataset up to date.

I have the done the scraping using Python Jupyter Notebook. I have used Beautiful soup, pandas and regular expression libraries to extract the raw data from the website. Using the Request library I first got the source code, and using HTML parser in beautiful soup it was lot easier to extract the tags as compared to doing it by just using RE.

*(The entire scraping was done by me, and can be immediately provided on request. Due to page constraints cannot paste the code here.)

I could just extract the general information of the player like, Name, Age, Nationality, which for my project wasn't much of use.

Fig 1.1 Basic attributes

	ID	Name	Age	Photo	Nationality	Overall	Potential	Club	Value	Wage	Preferred F	Position	Height	Weight
0	158023	L. Messi	31	https://cdn	Argentina	94	94	FC Barcelor	€110.5M	€565K	Left	CF	5'7	159lbs
1	20801	Cristiano R	33	https://cdn	Portugal	94	94	Juventus	€77M	€405K	Right	LW	6'2	183lbs
2	190871	Neymar Jr	26	https://cdn	Brazil	92	92	Paris Saint-	€108M	€290K	Right	LW	5'9	150lbs

It had a limited of 7 rows which did not prove of any help for my exploration, so then I had to consider using more than one datasets, I needed individual player skills and attributes like Agility, Finishing, Interceptions, etc , then I used the player IDs to redirect the URL to the individual players page, which I had to do iteratively to get all the skills and values and then add them to a separate data frame.

Then finally I merged both my datasets using inner join at Player ID, for this merge method was used provided by the pandas library. The detailed information webpages has lot of unwanted attributes and values which weren't really required for our project and hence were excluded during the process of scraping, there were redundant information such as goalkeeping skills for a player who is not a keeper, all of which was taken care of during the scraping. The first basic dataset was easier since everything was provided on the single page, but the second dataset, took 5 and Half hours to scrape, since it had to browse through 18046 URL's and match the regular expressions.

Fig 1.2 Detailed attributes

Crossing	Finishing	HeadingAc	ShortPassir	Volleys	Dribbling	Curve	FKAccuracy	LongPassin	BallControl	Acceleratio	SprintSpee	Agility	ShotPower	Jumping	Stamina	Strength	LongShots	Aggression	Inter
86	95	70	92	86	97	93	94	89	96	91	86	93	85	68	72	66	94	48	
84	94	89	81	87	88	81	76	77	94	89	91	87	95	95	88	79	93	63	
83	87	62	84	84	96	88	87	80	95	94	90	96	80	61	81	49	82	56	

After successfully scraping when expecting everything to be perfect, I found quite a few missing values in the attributes namely : Wage, Value, Club Name. In the dataset of 18000 entries, I tried my best to replace the Null values with the correct ones by looking up on the google and filling few with average values which I did using Excel, which was quite a bothersome task to do. After all of the scraping and wrangling, the final csv was then read into the Rstudio.

I entirely used R Markdown for Data checking, Exploring and visualizing the data for the insights.

III. Data Checking:

To start the exploration, I found some problems in the data, if you closely observe the Wages and Values, it weren't the usual number we would expect, they had "€", "K" and "M", I decided to leave the currency be Euros and imputed the "€", then since K and M meant something, erasing didn't make sense, so I multiplied the K digits with 1000 and M digits with 1000000. You can see the difference between fig 1.2 and 2.1 at wage and values columns.

Fig 2.1

Nationality <chr>	Overall <int>	Potential <int>	Club <chr>	Value <dbl>	Wage <dbl>
Argentina	94	94	FC Barcelona	110500000	565000
Portugal	94	94	Juventus	77000000	405000
Brazil	92	92	Paris Saint-Germain	108000000	290000

Next, the columns weight and height had values along with the units, for height it was Feet and for weight it was "lbs". I simply substituted the " ' " with a "." (Dot) for height and for weight I got rid of lbs, and then finally wrapped up by converting both of them to numeric using the famous "as.numeric() " method.

Height	Weight
5'7	159lbs
6'2	183lbs
5'9	150lbs
6'4	181lbs
5'11	154lbs
5'8	168lbs

Fig 2.2(a) In CSV

Height <dbl>	Weight <dbl>
5.70	159
6.20	183
5.90	150
6.40	181
5.11	154
5.80	168

Fig 2.2(b) In Rstudio

Next, during further reviewing of the dataset, I found that the Player's position wasn't general, they were abbreviated terms which a football enthusiast/player/fan can understand. The terms were like: LM, CDM, RB, RF, etc., these contributed to 27 different positions. For my exploration, this could have turned out nightmare if wasn't taken care of, so I had to generalize these positions in 11 different labels which are :

FWD, l_FWD, r_FWD, c_FWD, l_MID, r_MID, c_MID, l_DEF, r_DEF, c_DEF and GK.

Looking at the above positions its so easy understand what's going on here. All the players playing at left forward will be assigned to "l_FWD", players at CM, CAM, etc will be assigned to c_MID and so on. Below is the snippet of code for handling this,

```
x <- as.factor(df$Position)
levels(x) <- list(GK  = c("GK"),
                  l_DEF = c("LWB","LB","LCB"),
                  r_DEF=c("RB","RWB","RCB"),
                  c_DEF=c("CB"),
                  c_MID=c("CDM","CM","CAM",""),
                  l_MID=c("LM","LAM","LCM","LDM"),
                  r_MID=c("RM","RCM","RDM","RAM"),
                  FWD = c( "ST"),
                  l_FWD=c("LF","LS","LW"),
                  r_FWD=c("RF","RS","RW"),
                  c_FWD=c("CF"))
df <- mutate(df, Position_general = x)
```

Fig. 2.3 From Rstudio

Then similarly I found the need to break the Values and Wages into such levels as above, breaking the huge 6-7 digit numbers to brackets like “0-100k ”, “200k-300k” for wages and “40-50M”, ”60-70M” for values, so on.

wage_brackets <fctr>	value_brackets <fctr>
500k+	100M+
400k-500k	70-80M
200k-300k	100M+
200k-300k	70-80M
300k-400k	100M+
300k-400k	90-100M

This is where I finally concluded my data wrangling and checking.

IV. Data Exploration:

After all this steps in wrangling and data checking and modifying, I am afraid I still found some inconsistencies in the dataset, which can be a hurdle to our thorough data visualization.

With data modelling I have learnt an excellent concept of statistical measures called Correlation for identifying the trends, patterns, redundancies or inconsistencies in the datasets.

Coming to the point if you look closely in the detailed dataset, the player skills which was extracted for the exploration turns out inconsistent, the Goal keepers have skills like Agility, stamina, strength, aggression, finishing, crossing and so on, which are not at all required for the last man, and are given low score to it, similarly a striker, midfielder and defenders are given attributes of GKDiving, GKReflexes, GKHandling and so on, this cause inconsistency, correlating these attributes, points out exactly what correlates and what doesn’t, the positive correlation is a positive R and a negative correlation gives out a negative R, which is also called as correlation coefficient.

The statistical correlation, will now help us find out factors which are responsible for the players overall rating and market value, which will also fetch us the answer to our first question :

1. What are the factors/attributes responsible for player's overall rating and market value.

Here, in the given dataset every player has values for all the attributes, even if they aren't related to the individual's preferred position, like explained above.

R has a method called "cor()" to find out correlation between the variables provided in the dataframe as Input to the function. R also provides a beautiful readymade package for correlation called "corrplot" for plotting these correlations on the graph. There are basically three methods for finding correlation, we would be using the default "pearson" method.

So, as we know GkDiving, GKReflexes, GKHandling, etc are the skills of a keeper and should be correlated to each other, hopefully positively and that is what we can observe in the following plot:

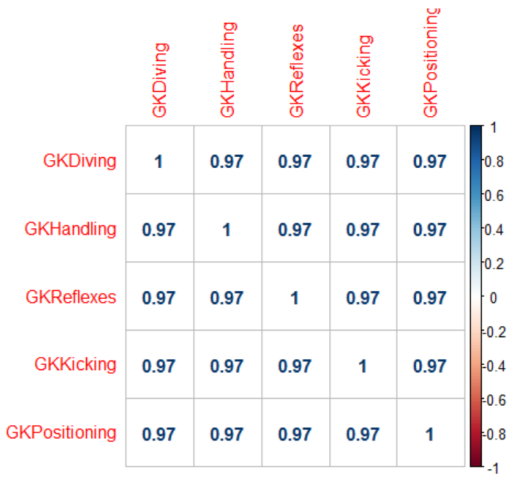

Fig.3.1

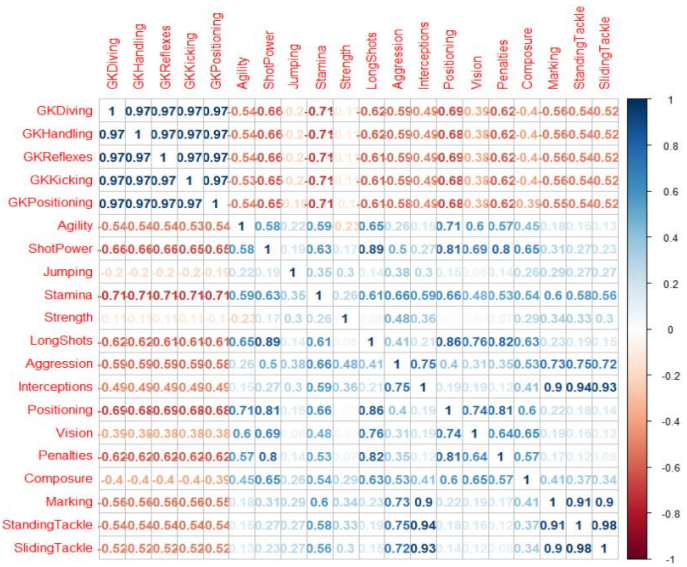

Fig. 3.2 From Rstudio

In fig.3.1, displays the correlation of the goalkeeper attributes, which to our no surprise correlates perfectly with 0.97 R value. To check the subset of these skills with other non-goal keeping skills which were being assigned to him, we plotted another correlation matrix and checked the correlation of these skills, few astonishing results came out, the Goal keeping skills showed no correlation at all, infact it showed a negative correlation with the other skills as shown in fig.3.2. With this, I found out that these Five attributes are the soul factors affecting the overall performance of a Goal Keeper, this means if you want to judge the performance and value of the player, these five skills needs to be assessed instead of assessing the other unwanted attributes for the keeper.

Similarly, after excluding the goalkeeper skills from the next correlation plot, I discovered, skills like Interception, Aggression, Marking, Standing Tackle, Sliding Tackle, again had outstanding correlation coefficients ranging from 0.73 to 0.98, which brings us to the next finding and conclusion that these skills are of a Defender with no confusion.

The next one is bit trickier, since a midfielder and the forwards have pretty much same skills at the position they play, now with the remaining dataset left, after plotting them we find attributes with lower correlations and few stronger. Attributes like Agility, Sprint Speed, Acceleration, Finishing, Dribbling, Positioning, Ball control and crossing had a good correlation whereas a really low or no correlation with the skills like Longshots, Volleys, Vision, FKAccuracy, Curve and short passing. This hints us to the conclusion that this position should be of a striker. In fig.3.3 we can see how well these attributes correlate and hence we can rightly claim that these attributes are factors affecting a strikers performance.

With some more experimenting, I figured out that few skills like Dribbling, Positioning, Ball control and crossing, correlate very well with the remaining attributes and found out that these are attributes of our last remaining position- Midfielders.

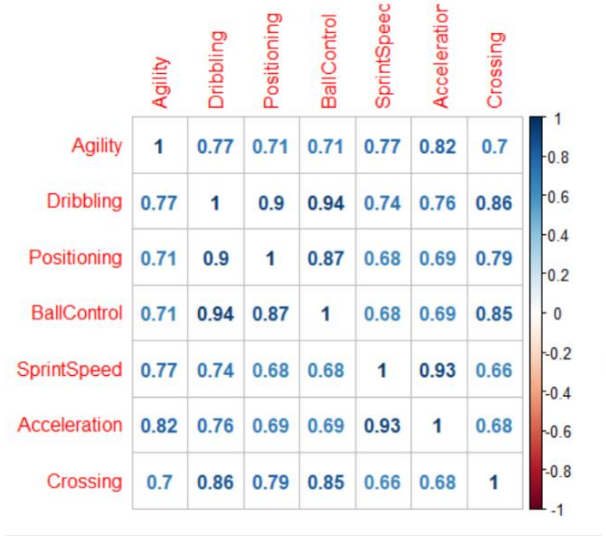

Fig.3.3

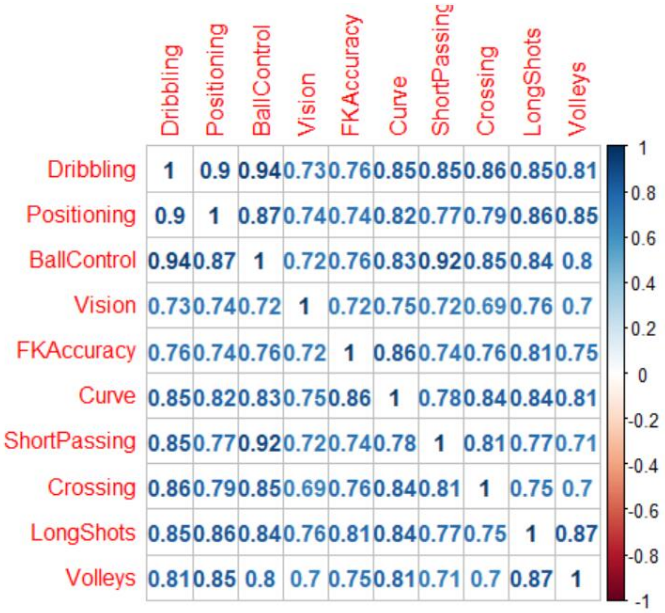

Fig.3.4

Hence, we successfully found out what are the factors which affect the performance of the player at their respective position.

Since now we have all the information that we need to determine the performance of the player at any position, we can make an attempt to find the Ultimate Team using the data we have. Doing some calculations with the skills and ratings in the dataset I found out top eleven (This was done in R). Below is the visualization of eleven players at different positions on the field .

As observed in the fig.1.1, you can find the URL's to the photos of every player, with the help of Magick library I uploaded a png image of Football field and then using methods in it, I plotted all the player on the field using the X and Y coordinates.

The code to everything here can be provided on request since it's is done by me in R.

2. Why are the young players being paid more than aged and experienced one's?

How is the growth of the player measured, and why are some inexperienced young players worth more than the experienced aged player, what are the factors here ?

Answering this question is like looking out for the outliers, which disobeys the usual trend. Lets start with what growth is, growth as I looked up, its basically measure of how well a player can do as compared to his current performance.

Its basically, the difference between potential rating and the overall rating of the player. So, Growth = Potential rating - Overall Rating. Knowing this I felt the need to find out the count of the players at every position who still had some rooms for improvements, with some r coding, I managed to get the exact count of players at every positions, which are as follows:

```
[1] "Strikers: 22"
[1] "Center Forward : 31"
[1] "Left Forward : 23"
[1] "Right Forward : 26"
[1] "Center Midfielder : 20"
[1] "Left Midfielder : 27"
[1] "Right Midfielder : 24"
[1] "Center Back : 25"
[1] "Left Back : 30"
[1] "Right Back : 20"
[1] "Goal Keeper : 30"
```

Fig. 3.5 Output from Rstudio.

In the fig.3.6, point graph, the player age are plotted against the growth, and one thing we clearly notice that, the growth of the player decreases with age and stops after a certain age, to be exact after 30. There are no exceptions or outliers here.

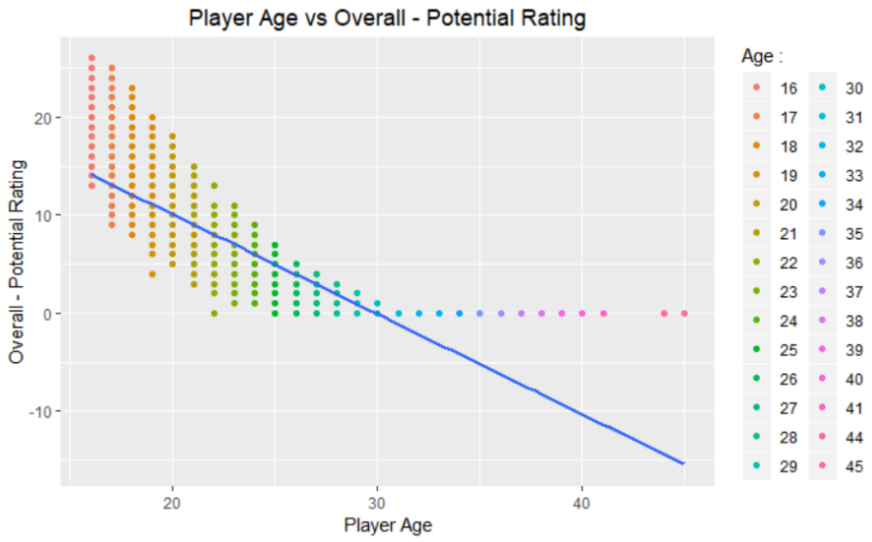

Fig. 3.6

What about the value and wage, how does it behave with age?

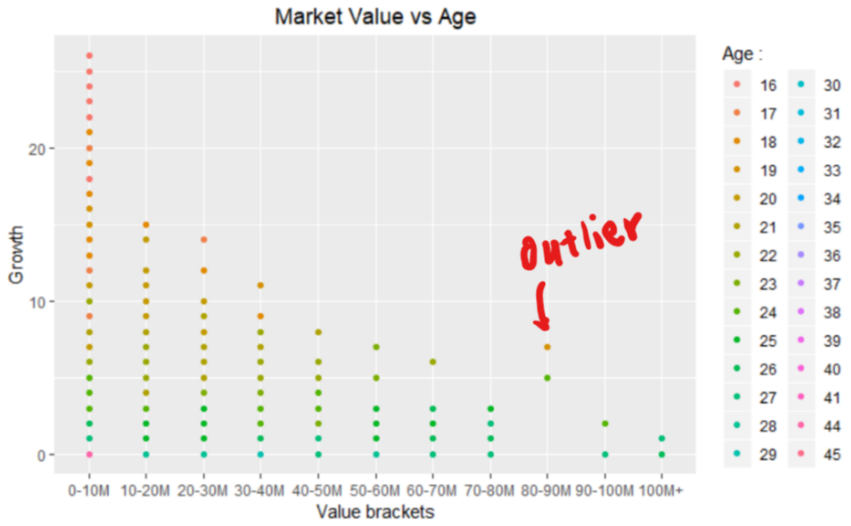

Fig.3.7

In the fig. 3.7, similar to the growth, the value of the player also decreases as the age increases.

From the above graph, we can see that our outlier doesn't have much growth left but still acts as a outlier and is one of the highly paid footballer. This is where the factor of age kicks in, if we check the Age legend, our outlier is under the age of 20, he is not even 20, and yet has higher market value.

From this visualization experiment, we conclude that, on the general basis the players which have higher growth initially have very less value and rating which gradually increases with age. But in our outlier's case it contradicts, top clubs are willing to spend higher budgets to recruit the talents like these, since starting the career with such excellent stats at a very young age is indeed the things which causes the skyrocketing of the player's market value.

3. **Does the preferred foot affect the positioning of the players, can the left footed players play on the right positions and vice versa? Is the fight for the position easier for the left footed players?**

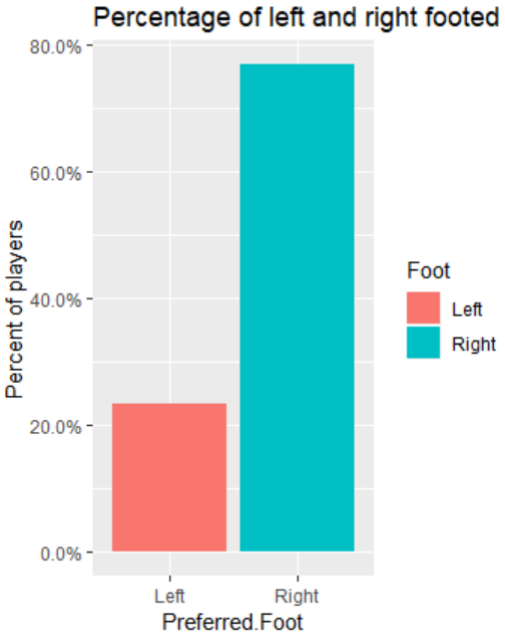

Fig.4.1

To start we need to understand the proportion of the left footed and the right footed players in the entire population.

In fig. 4.1, the bar graph tells us that the, left footed players are less than 25% of the total population.

Whereas the right footed have a proportion more than 75% which make the right footed players dominant.

With this we can expect to see right footed players dominating at almost every position on the field, however this turns out False when we look at the following line graph (Fig. 4.2)

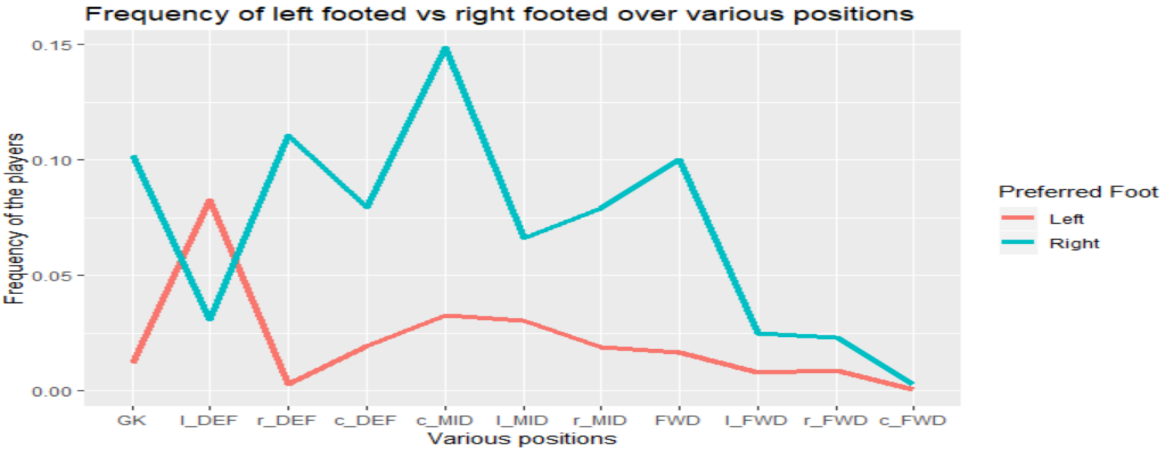

Fig. 4.2

From the above line graph we can see that the left defence position have left footed players dominating the right footed and almost all the Goal keepers are right footed.

Overall the proportion of both left and right footed players are quite same, with very few exceptions. This gives us an insight that, it doesn't matter much whether a player is right or left footed. The players can choose to play at any side they want, hence the demand for one position over the other will rough be same.

V. Conclusion

I would like to conclude the project by saying that, I successfully explored and visualized the data to satisfy the thirst of questions which I wanted the answers to. The facts which came out with exploration were quite astonishing, I was under the impression earlier that left footed players were preferred more over right footed ones since proportion of left footed were less in comparison and had higher demand, the fight for the position was higher for the right footed players. All of these beliefs were rejected by our exploration. The fight for any position on the field is pretty much the same. And every player has to strive hard for the position, nevertheless the player is a left footed or a right footed.

Another interesting fact which came out is that, every player has skills of an goal keeper or a defender though the player's instinct are that of a strikers. Such skills cannot be assessed to calculate the player rating, this can turn out biased and inefficient. With our Statistical correlation we were able to find out the correct skills respective to the player positions. Now the players can be assessed based on the skills they actually possess and which they train hard for.

New and young players like K.Mbappe, Marcus Rashford, etc were being highly valuated without wining world cups, and trophies nothing, I was confused with this anomaly, and answer to that was "YOUNG TALENT", that did not answer my question. So when I explored the question we realised that, in the trend of players starting with low overall rating and then gradually training hard to increase the overall rating and value, these young talents are already starting with such excellent stats, which is an exception from the regular pattern that we see.

VI. Reflections

I learnt a lot of concepts from this project and most importantly I enjoyed the entire process of exploration, since I chose the topic of my interest. This project portrayed how important exploration of the data can be, how deep any dataset can be explored, to be honest, in the this project with all the exploration and days I have spent, I could barely scratch the surface of this massive dataset. The dataset can be still worked on deeper levels, the exploration basically brings out the relationships between various factors and variables. I term it as the "THE MOST" important step of Data Analytics.

I found a issue with the massive dataset that I used, since I had to correlate the attributes with one another, the correlation had huge samples which led to normalizing of the data, I am not sure to what extent that could cause a problem but, I found this point worth making a note of.

Another Good thing is that, I found R programming very easy and efficient with all the packages and libraries it provides, before this I was very afraid of R programming since I had no hands on experience with it. I also got an opportunity to implement statistical measures like Correlation in this project. Exploration is very important to visualise any data.

VII. References

1) Wikipedia.com
2) TowardsDataScience.com