Project Report on FIFA Data Analysis 2018

Introduction

FIFA is a non-profit organization that describes itself as an international governing body of association football, futsal, and beach soccer. It was founded in 1904 to oversee international competition among the national associations of Belgium, Denmark, France, Germany, the Netherlands, Spain, Sweden, and Switzerland. Headquartered in Zürich, Switzerland, its membership now comprises 211 national associations. It is the highest governing body of association football.

FIFA's organizational Statutes now include a variety of aims, including expanding association football internationally, ensuring that it is accessible to all, and campaigning for integrity and fair play. FIFA is in charge of organizing and promoting association football's biggest international competitions, including the World Cup.

FIFA, a football (soccer) electronic game series developed by EA Sports, is a division of the American gaming company Electronic Arts and licensed from the Fédération Internationale de Football Association (FIFA). EA Sports began the FIFA series in 1993 intending to promote the game. The player ratings, which are the overall points given to each footballer in the game, are one of the components of the FIFA series that gets fans so thrilled every year. This helps managers pick who to select on any particular game by determining who is the best player on each installment.

Data and Executive Summary

The data.csv (https://raw.githubusercontent.com/4m4n5/fifa18-all-player-statistics/master/2019/data.csv) file includes one of the latest editions FIFA 2019 players attributes. It includes 18207 rows and 88 columns with several player's performance describing characteristics. The source of our data was Kaggle which redirects to a github repository.

In this project, we aim to explore different factors including age, nationality, value, positions, stamina, agility, reaction time, clubs, skills, and many more for a player and analyze if different factors correlate with the overall performance on the pitch and if that performance matched the value and trust put on the player by different managers. We hypothesize that the higher the value put on the player, the higher their performance. Also, agility, reaction time, finishing, and skills have direct proportionality with performance while age has an inverse relation. It is presented as a short exploratory analysis of the FIFA 19 dataset using Python.

With the data of players' preferred foot, we wanted to see the popularity of a specific foot and if having a certain preferred foot impacted the overall performance of a player. For this analysis after some data

cleaning (which is shown in the main report), we observe that a clear majority of players were right-footed. Our analysis of preferred foot to overall ratings also showed that foot preference does not impact the overall performance of players.

Similarly, we wanted to see the age distribution of FIFA players. Our observations and analytics showed that most of our players' age distribution was between 20 to 30 years. The oldest player was 45 years old while the youngest players were capped from 16. Likewise, the most valued players were Neymar, Messi, and De Bruyne. England had the highest number of players with 1662 players while Germany was second with 1198 players. This could be due to the popularity of the English Premier League and Bundesliga all around the world.

Acceleration and stamina increased with age till 26-27 years old after which it declined as expected. However, the overall rating kept on rising with age maybe because of the experience gained from playing over the years and the statistics like goals and assists building up. It was also found that a reaction is one of the factors that strongly affect the performance of a player and has a strong positive correlation with overall performance. Similarly, penalties have a strong positive correlation with finishing. Finally, for a popular club, the median wage was over £100,000 and their median rating was in the high seventies (almost eighty). So, if a player can build up his statistics accordingly and join a popular club, we can say that he will earn more than 100, 000 per week which is very high compared to other professions.

Main report

Data Cleaning:

Categorical data:

For the categorical data, let's see the preferred data column. There were only 18159 non-null values while the total number of rows was 18207 meaning there were 48 non-values in the preferred foot column of our dataset. As done with categorical data, we replaced the missing/nan values with the mode of the entire column after which we performed the visualization using a bar graph.

Numerical:

We had many quantitative data but the columns that needed extensive cleaning and restructuring were Height, Weight, and Value of the players. In the dataset, the dtype of these numerical data was in an object which can be understood by the readers, but the computer cannot analyze it. So, we had to convert it into numerical data type i.e. int or float.
This was not the only case. The height and weight had some missing values and since it is supposed to be numerical data, we cannot replace it with the mode value. With this thinking, we came up with a function to convert the object data type into float and replace the missing values with the mean.

This is the sample of the Value column:

```
df['Value'].sample(5)
```

```
ID
174665     €250K
202769     €575K
236506     €1.4M
207862    €15.5M
169388     €575K
Name: Value, dtype: object
```

As we can see from the image that the value of a player has unnecessary signs like £, M, and K. If the value was in the same format as every value in either thousand or million, we could have just removed the £ and K/M and compare them. But as they are different, we had to come up with a function that removes all these and handles the K/M case correctly. Here's the sample of the Value column after restructuring:

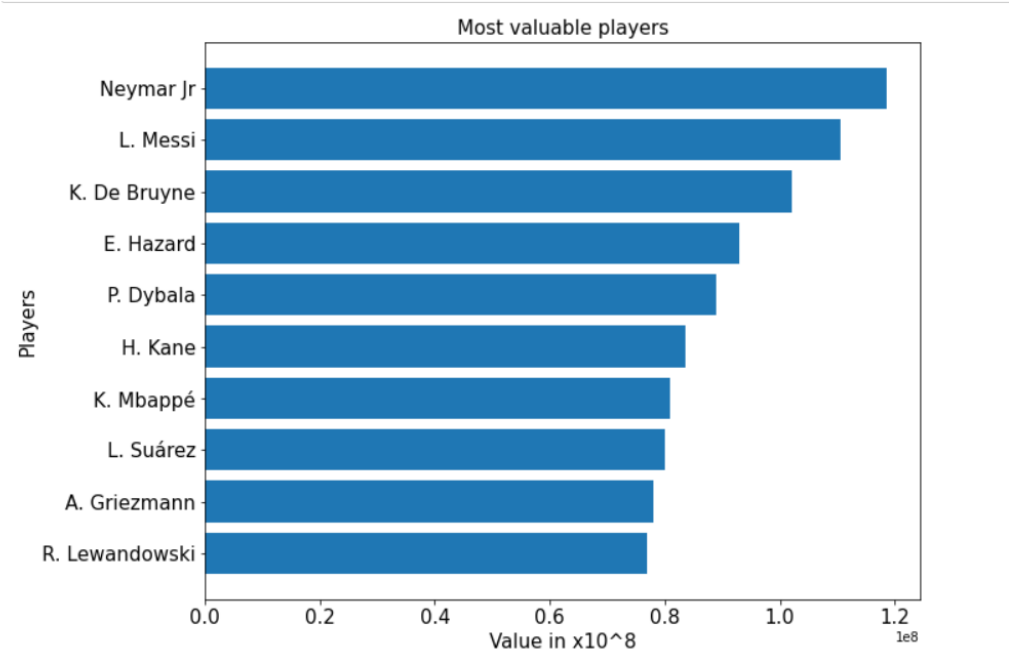
```
df['Value'].sample(5)
```

```
ID
213242    6500000
181971     775000
241096    1800000
242603     240000
169432     725000
Name: Value, dtype: int32
```

Data Analysis and Visualization

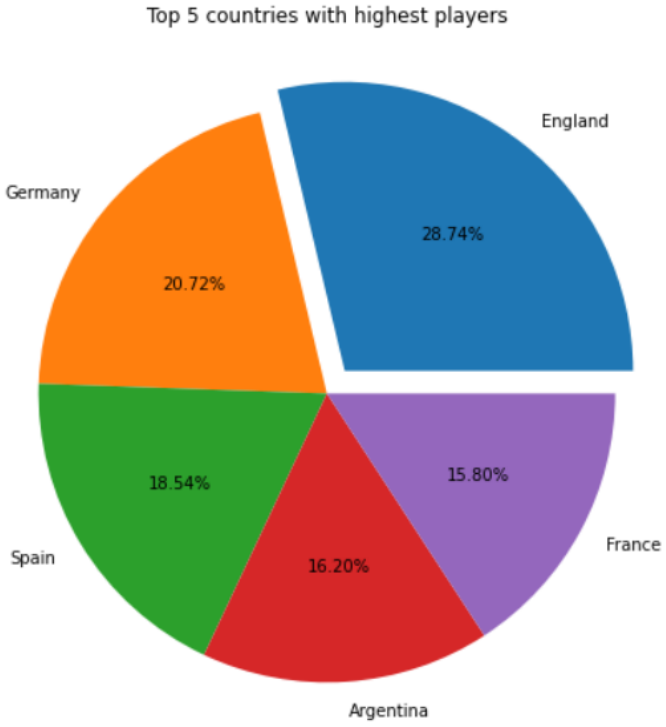
Valuation of players

The visualization shows how the top players in FIFA are valued. The bar graph shows that Neymar is at the top with 120 million followed by Messi. Here a bar graph was chosen to represent the data because of visualizing the amount.



Countries with the highest number of players

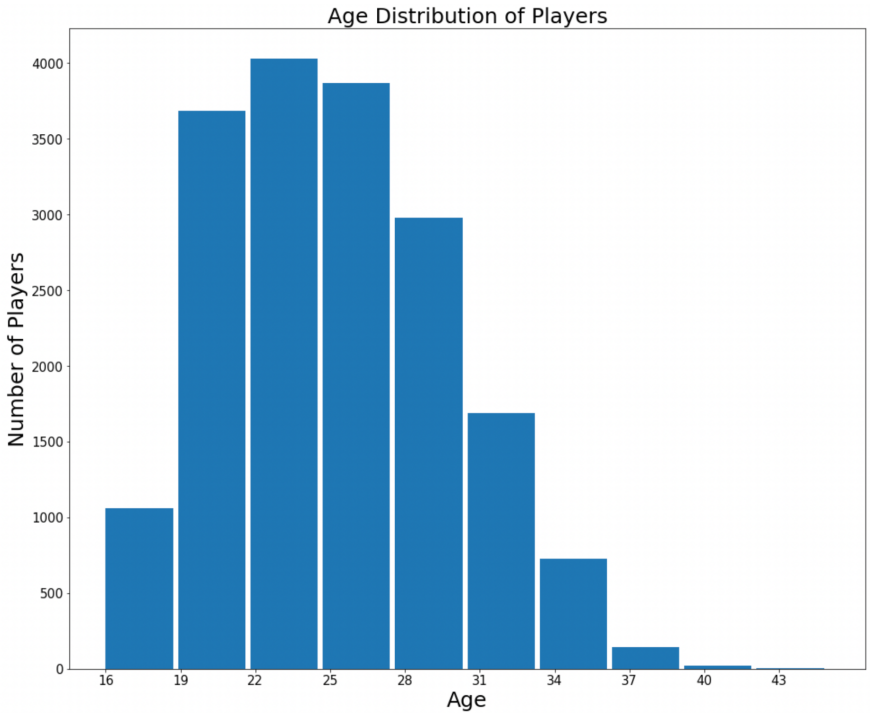
This figure is a pie chart showing the top 5 countries with the largest number of players in FIFA. This figure portrays that England has the highest percentage followed by Germany. To verify this figure, we have the tabular data representing the actual figures. This could be due to the popularity of the sport in particular nations.



	index	total
0	England	1662
1	Germany	1198
2	Spain	1072
3	Argentina	937
4	France	914

Age groups of players

Then we have a histogram to represent the various age groups of people in FIFA. As observed, the players start to join FIFA from the age of 16 years old and the data has a large group of players from 20 years old to 30 years old. Then, the histogram diminishes around the age of 45, this is the speculated retirement age for FIFA players.



Oldest and youngest players

Here is tabular data showing the top five oldest players and top five youngest players in FIFA. We can see that the oldest player is 45 years old and the youngest player is 16 years old. We can expect that almost all players retire after the age of 45 because their bodies just cannot handle the physical labor involved in the sport.

```
oldest5 = df.nlargest(5, "Age")
oldest5[['Name', 'Age','Nationality']]
```

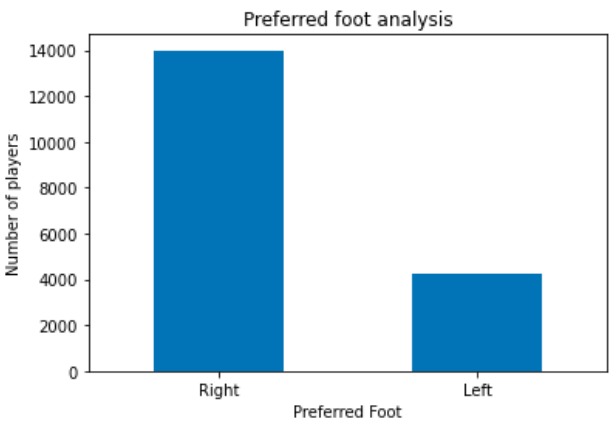
ID	Name	Age	Nationality
140029	O. Pérez	45	Mexico
51963	T. Warner	44	Trinidad & Tobago
53748	K. Pilkington	44	England
140183	S. Narazaki	42	Japan
156092	J. Villar	41	Paraguay

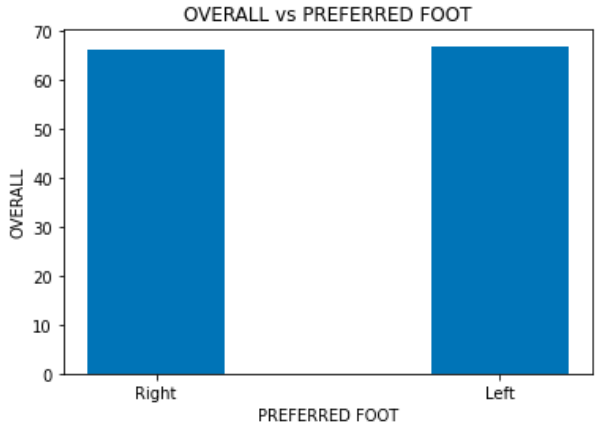
```
youngest5 = df.nsmallest(5, "Age")
youngest5[['Name', 'Age','Nationality']]
```

ID	Name	Age	Nationality
241266	W. Geubbels	16	France
244403	A. Taoui	16	France
245616	Pelayo Morilla	16	Spain
246465	Guerrero	16	Spain
246594	H. Massengo	16	France

Preferred Foot vs Overall Rating

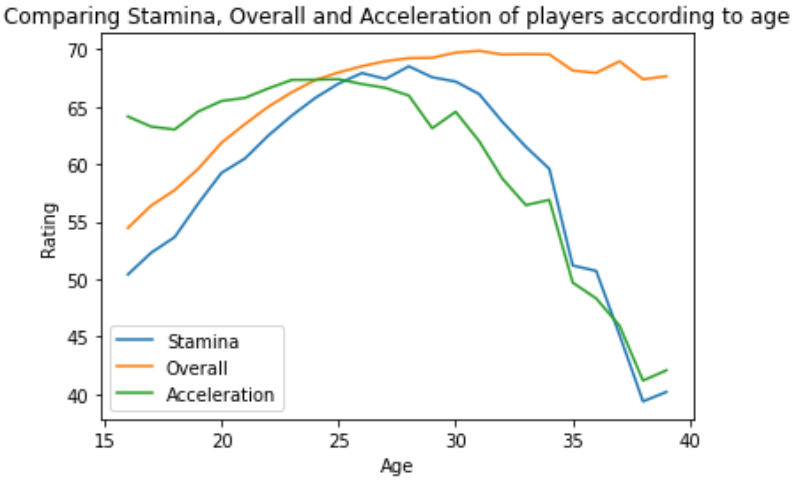
The bar graph between preferred foot and number of players and preferred foot and overall rating was created.





After data visualization, we came to the conclusion that the right-footed players were significantly higher in number than the left-footed ones. There were almost 14000 right-footed players while only around 4000 players were left-footed. Then to answer the question: having such a vast difference, did being right-footed or left-footed have any influence on the overall performance of a player, we used a bar graph to look at the average overall ratings of right-footed players and left-footed players. The results were as expected, there was no significant difference in the overall rating. Meaning, that preferred feet do not influence the overall performance of a player.

Age vs Stamina, Acceleration and Overall



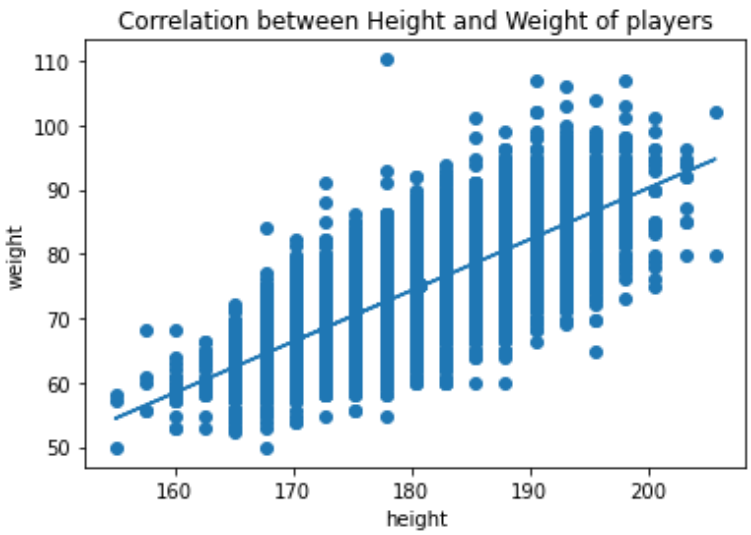
Then we have a line graph comparing the age with stamina, overall, and acceleration of different players. We can see the result that we normally expect like stamina and acceleration increasing with time, which peaks from the age of 20 to 30 which is adulthood, and gradually decreases as people age. However, the visualization shows that the overall performance of the players does not go into gradual decline which was a surprise to us. As we analyzed the data further, we realized that the overall does not only depend on stamina and acceleration as we know, there are various other skills like pass, and shoot that increase with experience causing this result. We are using a line graph to see the changes over time.

Some X-Y relations of player attributes:

We are using scatterplot while analyzing x-y relations.

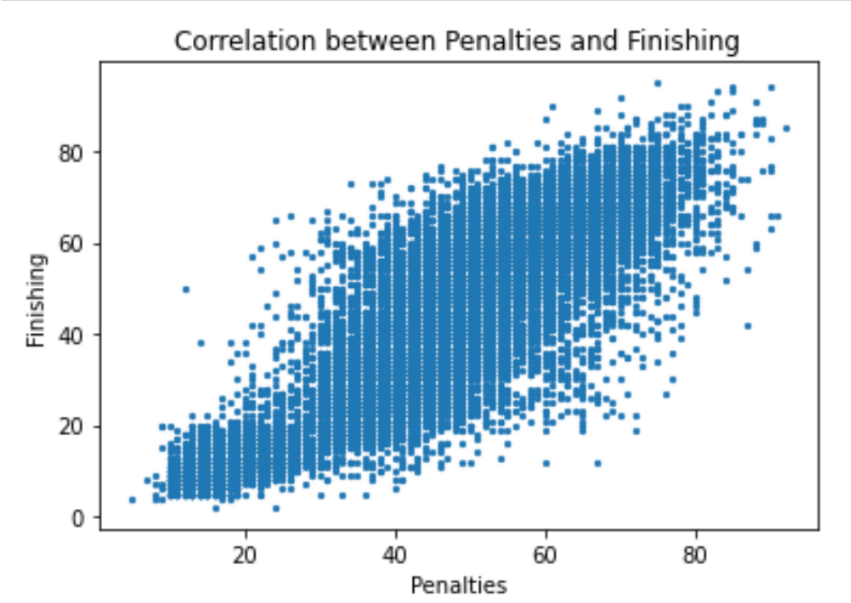
Height vs Weight:

We wanted to verify the correlation between height and weight using a real world dataset. We found out that it had a high positive correlation of 0.75.



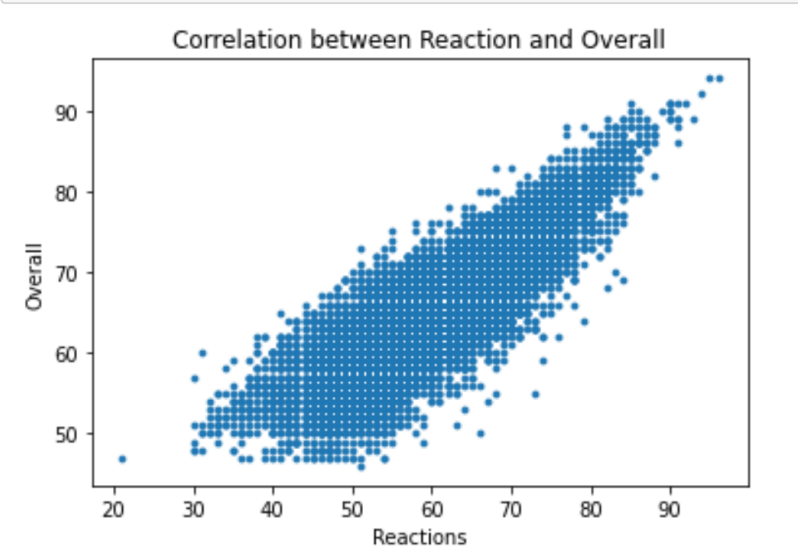
Penalties vs Finishing:

Finishing in soccer is the process of scoring a goal from inside or outside the penalty box. While analyzing data we found out that some players had high finishing even though they have average overall. We were concerned about this and started researching the reason behind this. We found out that penalties are one of the reasons. Penalties have a high chance of scoring a goal which also increases the finishing rate. We found a positive correlation of 0.83 between them. Hence, players with high penalties had higher finishing points regardless of their overall performance score.



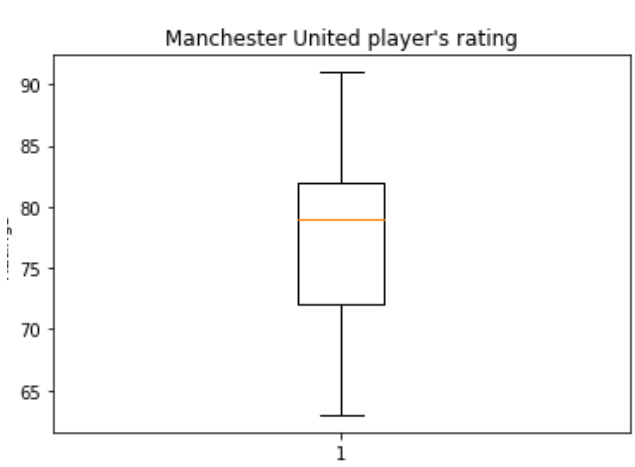
Reaction vs Overall:

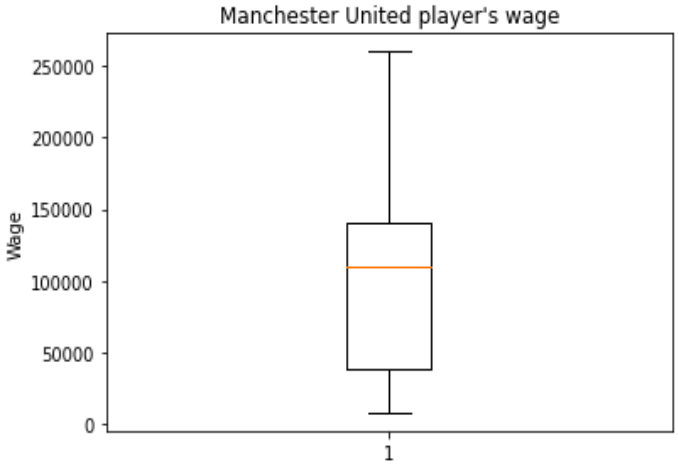
The overall performance of a player defines how a player plays. It also determines their value and wage. We wanted to know what increases a player's overall performance. We found out that a reaction is one of the factors that strongly affect the performance of a player. The correlation between reaction and overall performance is 0.85. This is because football is a spontaneous game and if a player takes too much time thinking about where to pass or what to do, he's not a good player.



Specific club player's rating and wage analysis

Filtering out the players from one of the top clubs in the world, Manchester United, and looking at the players' ratings and wages through box and whisker diagrams, we saw the following graphs





Here the rating is out of 100 and the wage is pounds per week. For a top club like Manchester United, we see that the median of ratings was in the high 70s, which is considered a very good rating. Likewise, the third

quartile was in the low 80s which again is very good ratings. Also, the first quartile is around 70 which is considered above average. The best players with a rating over 90 are also there in this club. As expected, top players earn more wages and this is seen in the second diagram above. The median salary was above 100, 000 and the first quartile was around 45,000 range. The highest-earning player earned more than 250,000 weekly and also the lowest-earning was around 10,000 weekly which is expected to be the beginner players/youth just entering the club at a base salary which even for a base salary is pretty good.

Conclusion

Using the FIFA dataset 2018, our data analytics process was able to draw various inferences and findings. The first step of data analysis was data preparation where we did a hawk-eye view of the data. Then, we did data cleaning which included cleaning the data format, changing the units, filling the missing data with appropriate statistical models, and so on. After that, we did the data analysis where we found various results. We used the matplotlib library extensively to visualize the findings using the most suitable graphs and visualizations. Our findings showed that most players were right-footed, but having a particularly preferred foot does not impact the overall performance. Similarly, most of the players were from England followed by Germany, maybe due to the popularity of the English Premier League and Bundesliga throughout the globe. The age distribution showed that the majority of players were from 19-29 years old. We found that acceleration and stamina increased with age till 26-27 years old after which it declined as expected. Overall rating kept on rising with age maybe because of the experience gained from playing over the years and the statistics like goals and assists building up. We also found out that a reaction is one of the factors that strongly affect the performance of a player.

The future of this data visualization can be taken up by analysis of more recent data (updated every month or so). Certain machine learning techniques can be implemented to predict the value of a player or which player to invest in based on the current statistic of an individual. By training models from the most recent data set we can get to more accurate conclusions on how a player could develop and be valued in the future.

Data analysis and visualization by: Sanskar Adhikari, Ankit Kafle, Anish Bhurtyal and Aayush Shrestha