FIFA'19 Analysis & Visualization

ABSTRACT

Federation Internationale de Football Association (FIFA) is an organization which describes itself as an international government body of association football. FIFA 19 is a product of EA Sports and its first video game was launched in 1994 and since then every year a new video game is added to the series. We researched and used Kaggle to get the dataset and it contains 18.2K rows and 89 columns. It consists of variables like Player name, age, nationality, club and other important statistics.

FIFA launches its new video games every year with new features embedded. Considering this fact, we came across the dataset which can be used to analyze and come up with the solutions of the following:

1. Determining the most valuable club
2. The football club which has a higher pay rate
3. The clubs which have maximum number of known faces
4. The significance of jersey number
5. Predicted value of the players in future
6. Average salary on age, popularity, potential, skills
7. Perform clustering on current age and experience
8. Potential average rating based on player's performance
9. Predicting future market value analysis of player
10. Skill analysis on rating
11. Future salary prediction

We successfully implemented the techniques learnt and visualized the data to get useful inferences. We performed visualization, clustering and regression using RStudio. The different plots and regression output generated can be used to draw conclusions regarding the players considering various factors contributing to improve the player and team performance.

DATASET DESCRIPTION AND TOOLS

- Dataset Description

  The dataset that we have used is the Fifa Dataset. The dataset was taken from Kaggle, and the dataset has a total of 89 variables and 18207 rows. Out of which the variables like Overall, Potential, Age, and a few others were used. The rest of the variables that were not required like Photo, Club Logo were cleaned. Apart from these variables, there were other attributes which were used, which helped in the visualization of the data clubsand for the clustering as well as for the regression and the prediction of the value of the players.

- Tools and Techniques

  The tool that we used was Rstudio. We performed various techniques on the dataset. The data consisted of many irrelevant attributes and values that needed to be eliminated.

1) First, data cleaning was performed on the dataset and all the unwanted variables were eliminated.
2) Next, for the visualizations we used the Clubs and their wages were used to determine the most valuable Club out of all the clubs present, to find the top wages, to find superstars in FIFA, determining the clubs with the youngest players.
3) Another graph was plotted to find the age distribution among the most valued clubs.
4) Linear Regression was performed and the value of the players for the future was predicted.
5) Next, we did clustering on the data using K-NN, to determine the range of the ages of the players in the clubs.

RESULTS OBTAINED

Graph 1: The Most Valuable Club
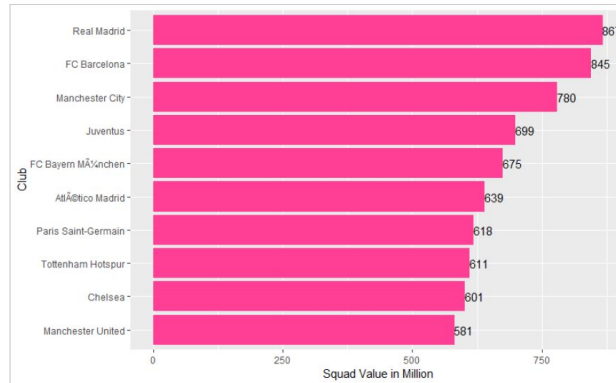
According to our analysis, we found out that Real Madrid is the top most club with value of 867 Million Dollars followed by FC Barcelona with a value of 845 Million Dollars, followed by Manchester City with a value of 780 Million Dollars and so on.

Graph 2: Top Wage Bills
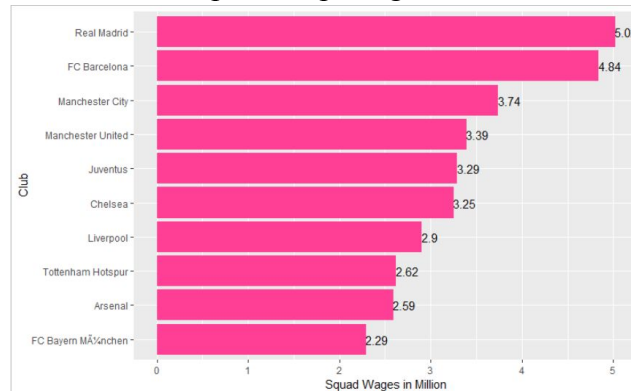
According to our analysis, we found out that Real Madrid is the top most club with a wage bill of 5.02 Million Dollars followed by FC Barcelona with a value of 4.84 Million Dollars, followed by Manchester City with a value of 3.74 Million Dollars and so on.

Graph 3: Superstars in FIFA
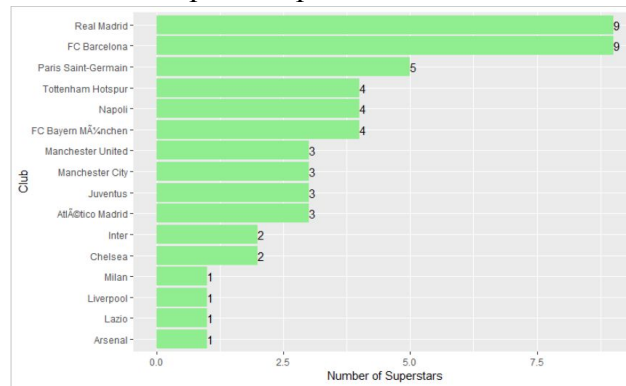
According to our analysis, we found out that Real Madrid & FC Barcelona have 9 superstars whereas all other clubs have 5 or less than 5 superstars.
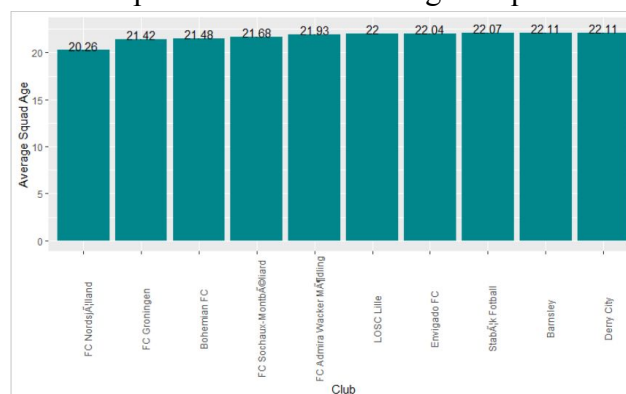
Graph 4: Age Distribution Among Top Valued Clubs
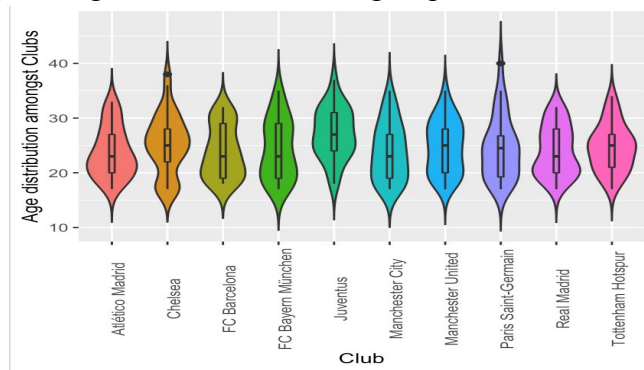
1.

In the above graph, we found out that age 20 to 30 is the ideal age for players and Club Vitória Guimarães has the perfect age graph while all other club graphs are not perfect.

Graph 5: Clubs with Youngest Squad

In the above graph we have found out the average age of all individual clubs.

Graph 6: Jersey Number related to Overall Rating
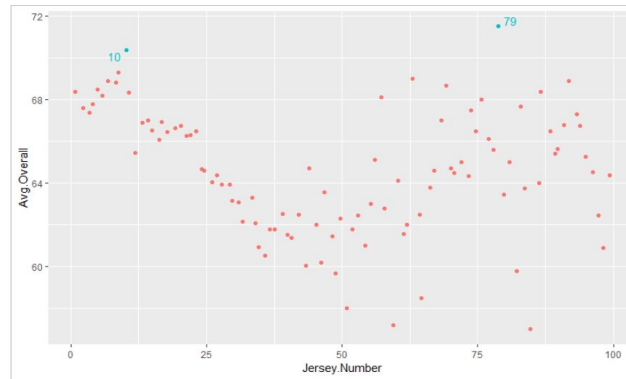
In the above graph as we can see Number 10 & 79 are the most uncommon numbers in group of Jersey Numbers, this is because 10 is the most valuable jersey number and it is not given to everyone in the club and 79 is the outlier that we've detected in the graph.

CONCLUSION

- Any limitations and future work:
  Some of the limitations and challenges:
1. Data cleaning: As the data consisted 18K records with 89 columns, cleaning was an essential task as it would affect the regression accuracy achieved.
2. As the dataset has a huge list of players, there were a few of them which were unknown.
3. As the jersey numbers of players were overlapping, it was challenging to plot the scatter plot of jersey numbers 10,79
4. Running a number of regression model with different variables to get the final model with highest accuracy.

- Business Objective Accomplished
  The main objective of all the computations performed was to come up with a regression model with a good fit and has some kind of significance. Also, we performed clustering to determine the age ranges along with the player performance to determine the potential age of the players. All of these analysis performed can be referred by the club managers to analyze and decide amongst the pool of players to be selected for their respective teams.