Date: 13th September 2019

Dataset: FIFA 2019

Data Storytelling and Visualization Report

Singh, Shekhar (11011694)

SRH Hochschule Heidelberg Student
Course: Big Data and Business Analytics

Abstract

FIFA 2019 is the latest rendition from EA Sports of its massively popular FIFA series, a football simulation video game. It stimulates the games and tournaments tracks (English Premier League, UEFA Campions League etc.) that we tend to follow in the real world. There are several modes to choose from for instance, players could select the 'career mode', where one can choose to build a team from scratch, and they play games with their team with the aim of winning tournaments offered in the game. Win or lose video gamers get various opportunities to trade players and ultimately build a highly effective and winning team. As a football and video games enthusiast I've performed preliminary data analysis on the comprehensive FIFA 2019 players dataset available on Kaggle in order to gain insight on trends and patterns present in the dataset.

This dataset comprises number of different attributes for around 18,207 players. Attributes include: basis statistics such as player name, club he is associated with, overall ratings, potential ratings, age, weight etc. as well as several player specific characteristics such as his market value, club wage, playing positions, skills such as movement, power and mentality. Findings from this exploratory analysis would cater to the need of other video gamers, who flock to buy this game or sport journalists in their process of reviewing football clubs and players.

Table of Contents

1. Introduction

FIFA video game series was first developed by Electronic Arts under the EA Sports label in the year 1993 and it was then named as ‘FIFA International Soccer’. It was made in order to celebrate the 1994 FIFA World Cup held in United States of America. The game has been a worldwide phenomenon for a very long time, every year an updated version of the game is released which is bought by more than 24 million gamers. These video gamers generally take lot of interest in gathering information regarding football clubs and players in the them. This information tends to help them in deciding the club and team they want to play with in the game. FIFA 2019 is the latest edition of this series; it has all updated details and statistics of currently active players in the game.

In this research paper I’ve used FIFA 2019 dataset to analyze and represent statistic information regarding football clubs and their players. This analysis would certainly help gamers in taking a more informed and calculative decisions while choosing clubs and players for a game, at the same time sport journalists would also benefit from it while writing their reviews. My calculations are based on the ratings provided by EA Sports; how specifically these ratings were calculated by them is somewhat opaque, but it definitely involves certain combination of performance statistics and subjective scout reports that must have been reviewed by a team of editors there.

2. Motivation

Ever since its release in 1993, FIFA football video games have soared to greater heights of popularity with every new edition. Pertaining to its popularity in the year 2013, the Spanish professional women’s footballer Vero Boquete started an online petition on Change.org, which requested Electronic Arts to introduce female players as well in the FIFA series. The petition received more than twenty thousand signatures in 24 hours. EA did take notice of this and finally FIFA 16, released on 25th September 2015, included female football national teams.

Latest edition of the game FIFA 19, released on 28th September 2018, introduces new competitions to the gam such as UEFA Champions League, Europa League and UEFA Super Cup. The game is programmed with lot of new functionalities and is regarded to have the world’s best graphical representation and seamless execution in the industry. The game is based on historical as well as new data and takes into consideration a player’s match performance, physical and financial statistics to provide ratings. As the players age, wage, market value, performance stats and physical attributes are constantly changing these need to be updated with every release. With every new edition there is always lot of curiosity among gamers as to which football player or club has better ratings or skill set and is going to perform better. Hence, we have used FIFA 19 dataset which has various elements and physical traits of several players to analyze this data and represent it in a more user-friendly manner. In this research paper we analyze players skill sets, wage, market value, overall and potential score using machine learning algorithms to find out interesting trends and correlation between them.

3. Problem Statement

Even though FIFA football video game series has been around for a fairly long time, users still do not have the application or means to judge who is currently the best soccer player, the most valued player, which is the best team or club in the world. Every individual has his or her own perspective when it comes to these questions, then there is a general public opinion which states that either Lionel Messi or Cristiano Ronaldo is the best player and that FC Barcelona and Real Madrid CF are the contender for best club in the world.

But these are presumptuous thoughts which are not backed by actual statistical information and often lead to misunderstandings.

The FIFA game does provide stats related to a player or club, but it does not offer its users the functionality to compare and evaluate all players and teams available in the game. Then are no clear evidence to suggest if a players age and nationality does have an impact on his market value and the amount of wage that he earns from a club. Similarly, there is no clarification as to how a player's overall and potential scores are calculated and if rest of his skills have an impact on these scores or not. Since data is readily available, these is a need to develop certain tool or application that could analyze this data and represent answers to these queries in an organized manner.

4. Literature Review

Although there have been quite a few analytical analyses on data related to FIFA football video game series but none of them really tackle the problem statement mentioned above. Neither have we come across any research paper based on this and it still seems to be an abandoned topic. Most of the analysis that have already been performed on this dataset took into consideration the entire data along with its inaccuracies and inconsistency hence the results obtained were not flawless. In this research paper we take into consideration a dataset of only top five thousand players after reviewing all dimensions of data quality such completeness, validity, accuracy and consistency. This gives a more concise approach to our research plan and better results.

5. Proposed Solution

Results that we desire to achieve from this research study is to find out answers to the following questions:

- Based on the 'Release Clause' data which football club can potentially earn the most by releasing its players?
- Players from which nationality are the most dominant in this sport?
- Analyze which club is the most valuable, has the best rated players, pays the highest wage on an average to its players?
- What is the age distribution amongst the players?
- Assess how much clubs spend based on players position, what are the dribbling speed, agility, shot power etc. of its high rated players?
- Is there any correlation between a player's age and overall score?
- Is there any correlation between a player's potential or overall score and his other skill sets, if yes, which skills tend to have more impact?

Our solution to the aforementioned problem is to use machine learning algorithms such as Linear Regression to look for correlations among various attributes of data. In addition, use R and Python programming language to visualize results and then finally come up with a dashboard in Tableau that graphically represents all requisite information.

6. Contribution

The findings of this research study will redound to the benefit the FIFA video game players as well as sports journalist in writing their reviews. Analysis provided in this paper does provide solution to all questions mentioned earlier. Any FIFA video game enthusiast would now easily be able to find out which club is the most valuable, pays the highest wage to its players, has best rated players, who is the best player, has the highest market value, etc. using this dashboard.

Even sports magazine journalists could use the report and dashboard for reviewing and writing articles on a particular player or club. The correlation techniques used does provide statistical evidence of there being relationship between player's potential or overall score and some of his other skill sets. There are several other interesting trends among data that could come in handy while publishing an article in the newspaper. Since research study was done only using top five thousand players, there is scope of using the entire dataset after preprocessing and cleaning keeping in mind data quality principles to achieve results on a larger scale.

7. Method

Below is the pictorial illustration of data pipeline that has been implemented in this research study.

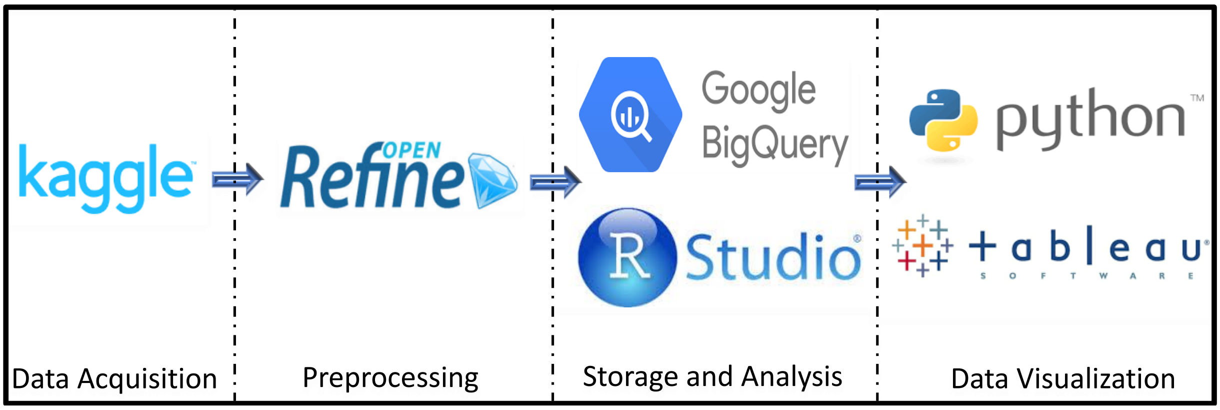

Figure 1: Data Pipeline

Following is detailed description of each step:

- **Data Acquisition:** The FIFA 19 dataset was downloaded from Kaggle which is an online community of machine learners and data scientists, owned by Google LLC.
- **Preprocessing:** This step included cleaning and preparation of data, which was executed in OpenRefine tool an open source desktop application.
- **Storage and Analysis:** After preprocessing data has been stored in Google BigQuery, which is a RESTful web services. It works in conjunction with Google storage and also enables interactive analysis of huge datasets. R programming language in R Studio tool has been used to execute certain machine learning algorithms to explore hidden trends among data.
- **Visualization:** Python programming language, the seaborn library, has been used to represent data for some of use cases while most of the data visualization is done using Tableau software. The final dashboard is also constructed in Tableau, it is an interactive graphical representation and outputs results based on user inputs.

7.1 Data cleaning and preparation

Firstly, the dataset is downloaded from Kaggle, it is a .csv file which includes latest edition FIFA 2019 players attributes like Age, Nationality, Overall, Potential, Club, Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Jersey Number, Joined, Loaned From, Contract Valid Until, Height, Weight, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, Crossing, Finishing, Heading, Accuracy, ShortPassing, Volleys, Dribbling, Curve, FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties, Composure, Marking, StandingTackle, SlidingTackle, GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes, and Release Clause.

The entire dataset comprised of 18,207 records and 89 attributes/columns. But some of had incomplete data or data was missing, whereas some had inaccurate and inconsistent record in them. The data wasn't standardized either as most of the amounts for wage and release clause columns were mentioned in Million Euros and some in Thousand Euros. Amounts for wage column were mentioned in Thousand Euros too, these had to be rectified and standardized as Thousand Euros for seamless calculations.

OpenRefine tool was used for preprocessing of data. We choose to proceed with only five thousand players hence all the incomplete or inaccurate records were first deleted. There were quite a few features that were not needed for our research study like LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB, players photo and flag therefore they these columns were deleted. The final dimensions of our dataset were 5000 records and 59 attributes.

7.2 Visualization techniques suitable for the chosen dataset

Visualization is certainly the most essential part of this research analysis, techniques chosen to symbolize data does provide actionable insights to users. Discovered patterns in the data helps users with brainstorming and taking rapid decisions. We have presented data mostly using bar charts, pie charts, line graphs and heat maps, information displayed in these forms allows discovering patterns which are not easily traceable otherwise. The dashboard that we have developed is a well-designed graphical visualization which helps users answer queries of any comprehension level in minimal time.

7.3 Justification of the chosen visualization technique

Majority of our focus was on comparing various attributes of different players or teams and for that bar charts or line graphs were the most suitable as they illustrate numerical values changing over a short or long period of time. Even user would find it easier to comprehend lines on a graph than column of numbers. Then we have used pie charts while comparing parts of a whole group. Heat maps were used to signify importance or impact of a particular record on the entire set. Since these visualizations are interactive it allows users to adjust the analysis boundaries on the fly, manipulate and tweak the data sets to extrapolate various results.

7.4 Marks and Channels

Marks used in this research study were points, lines, bars and area whereas channels used were mainly color, position and size. As the visualization techniques most used were bar charts, pie charts, heat map and world map. Differentiation of data is basically done by variance in size of the bars either in ascending or descending order so that visual representation is illustrative and self-explanatory. Colors are an important aspect of the entire visualization; therefore, lot of thought and emphasis has been put through it. Darker shades of color tend to highlight the higher value in charts and as the value goes down color gets lighter. The same regulations have been followed throughout the research paper.

7.5 Dashboard Design

Our dashboard is interactive and mainly has two components, one is meant to display all details regarding to players while the other one is meant to show data related to the clubs.
These would help users in comprehending complex and large amount of data more easily and efficiently.

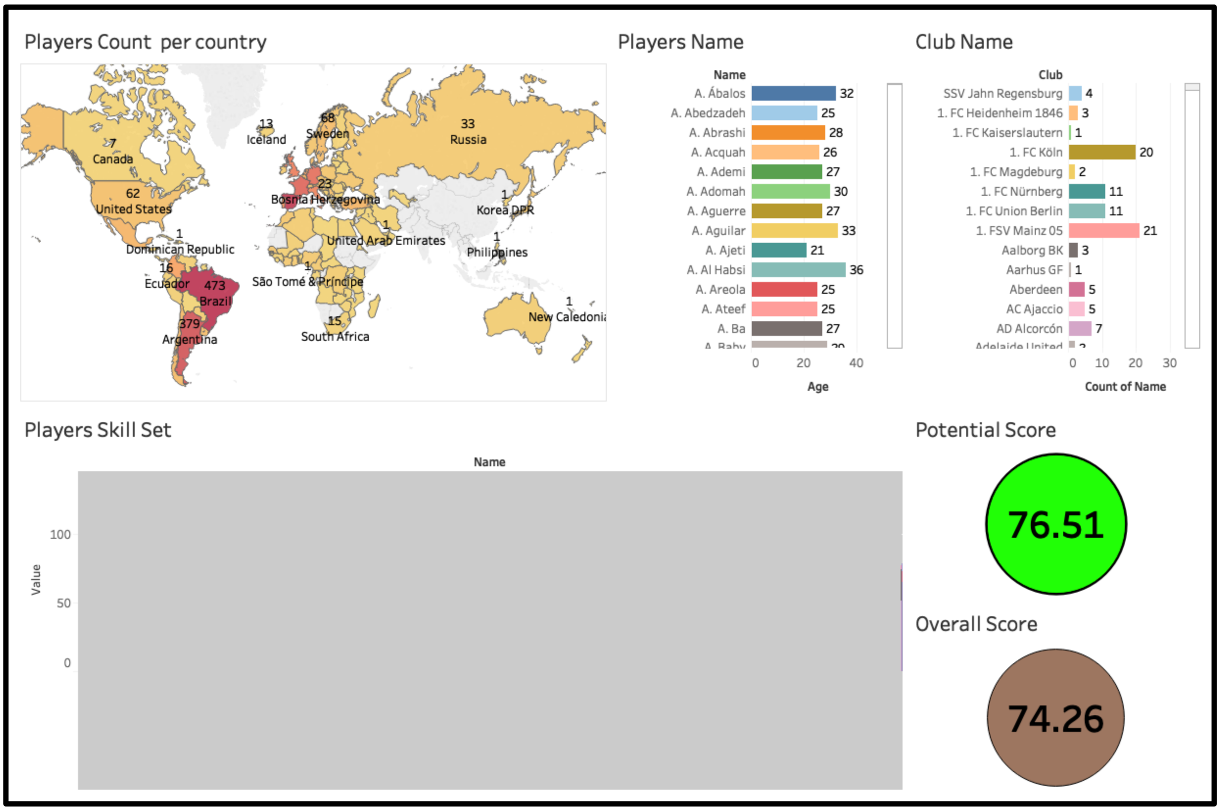

Figure 2: FIFA Dashboard 1

Above is the initial display, where we get to see a world map with players count as per their Nationality, all the players and clubs list. The potential score and overall score displayed here is an average of all five thousand players in the dataset.

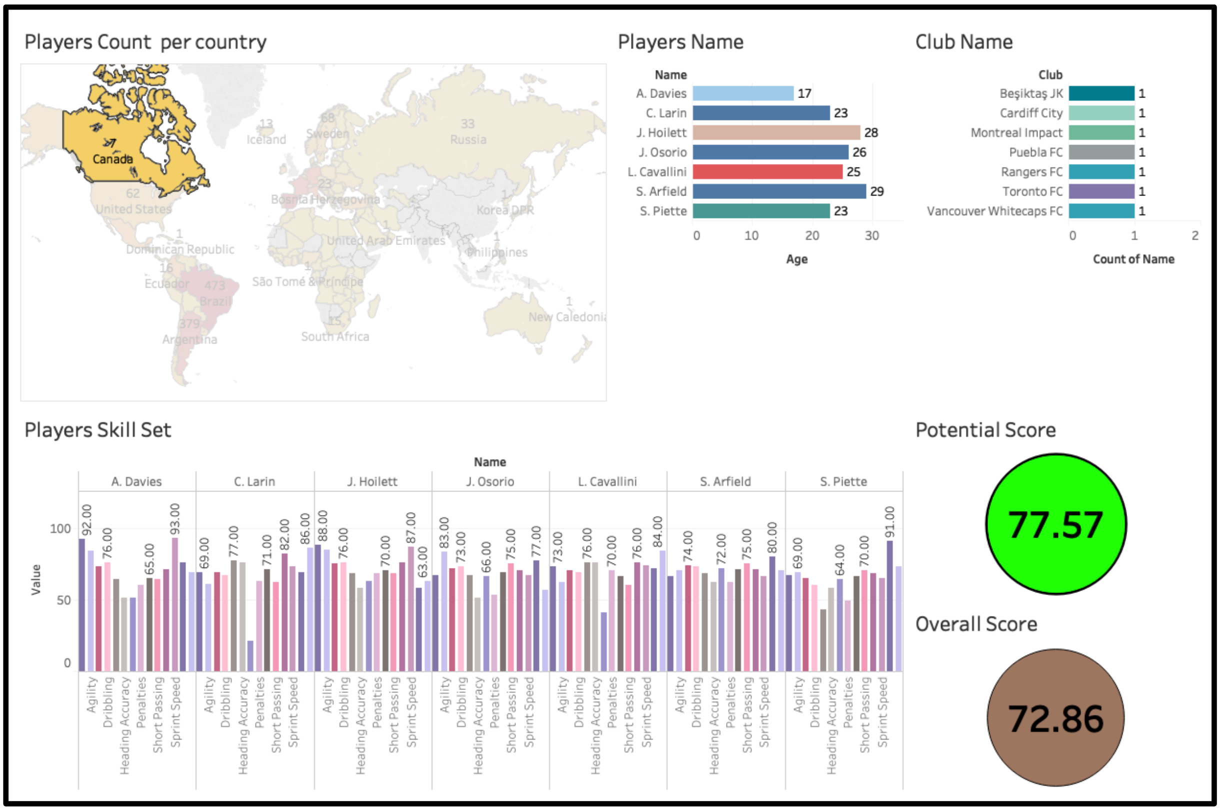

Figure 3: FIFA Dashboard 2

Once a particular country is selected, Canada in this case, then player name and club name list would display name of only those players who belong from this country and which all football clubs they are playing for. The players skill set section would displays skill scores for all those players only. The potential score and overall score displayed here is now the average of these seven players.

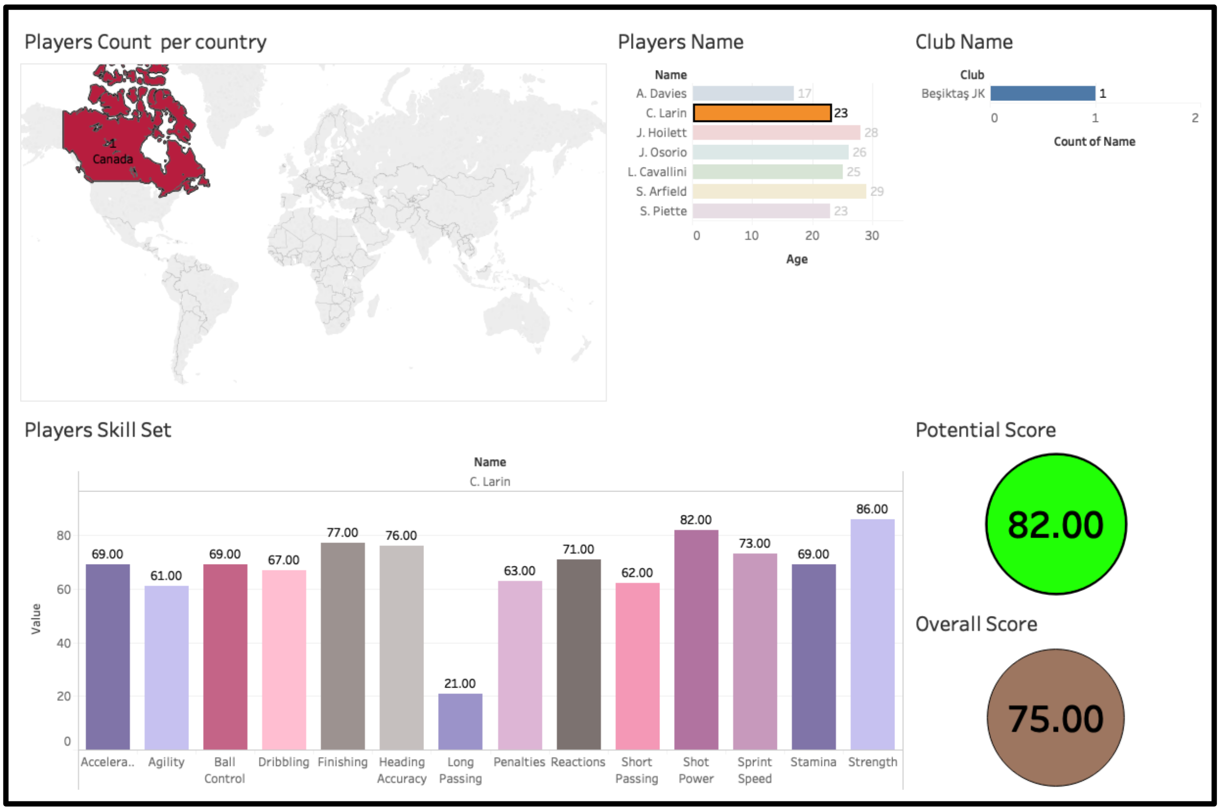

Figure 4: FIFA Dashboard 3

If we select any one out of all seven players in the list, then only his detailed skill set is displayed along with his overall score and potential score that he can achieve. Club name would display name of the club that he is playing for currently.

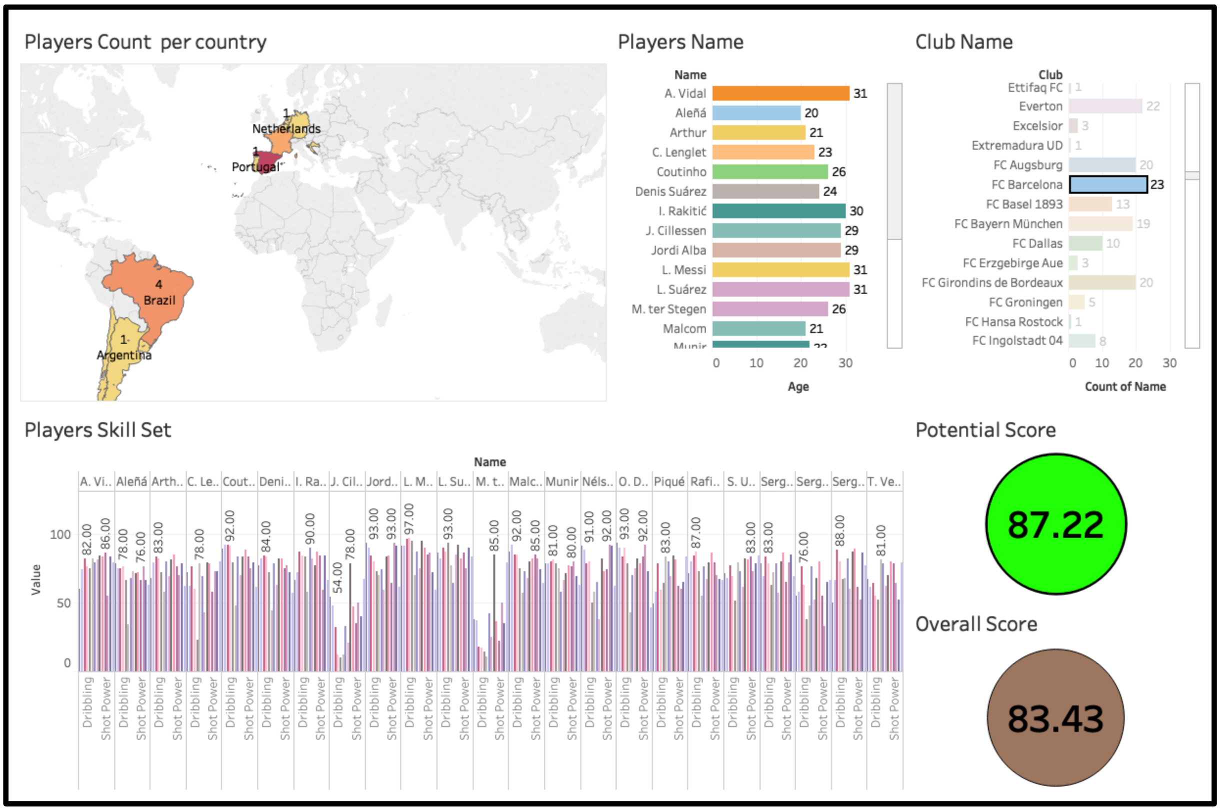

Figure 5: FIFA Dashboard 4

But it is not always necessary to go with the country first we can go the other way around as well, for instance, if we select a player from the list then all his details would be displayed directly. Or if we select a club first, then all players names along with which countries from would be displayed. The potential score and overall score that would be displayed then would be an average for all players in that club.

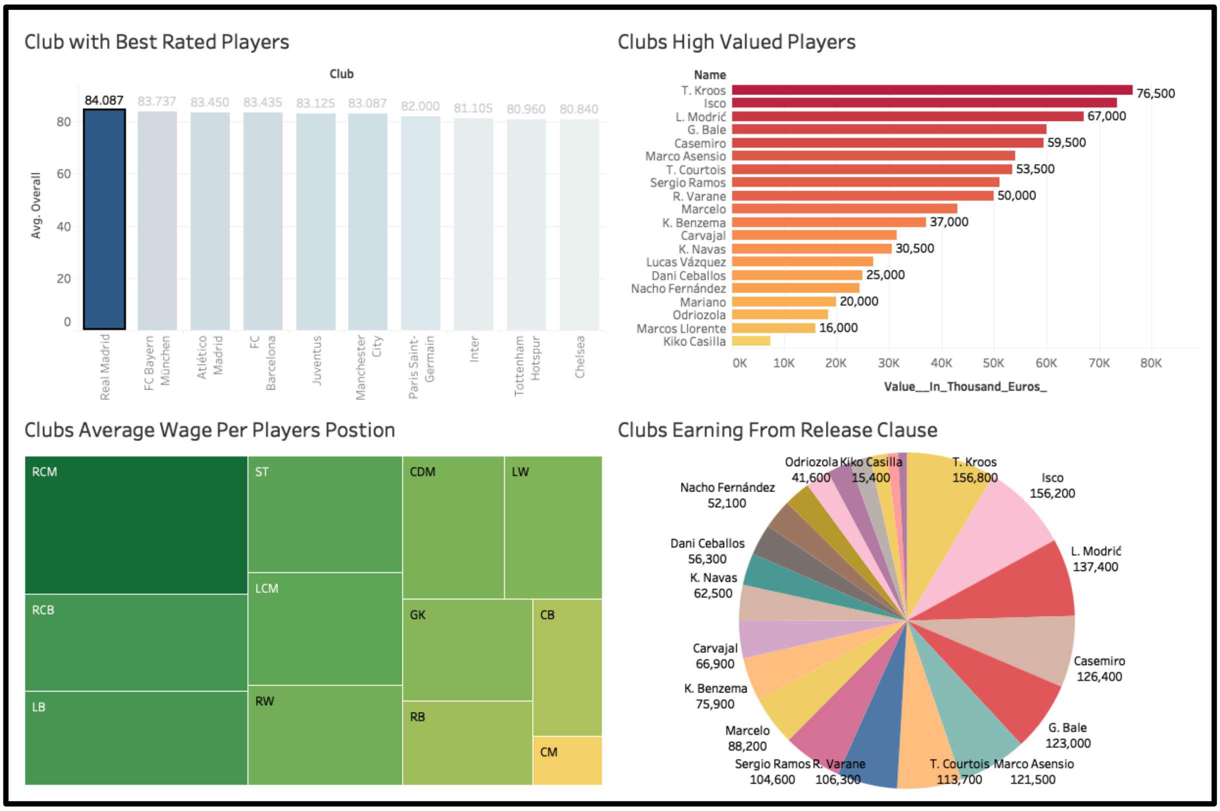

Figure 6: FIFA Dashboard 5

This part of the dashboard represents data related to the football clubs. Firstly, we start with the top 10 clubs based on average overall score of its players. We can choose any of the clubs, once selected its most valuable players are shown on the right in descending order via a bar chart. On the bottom left, heat map shows at which players position the club is paying the highest wages on an average. Pie chart on the bottom right depicts all the players release clause in thousand euros.

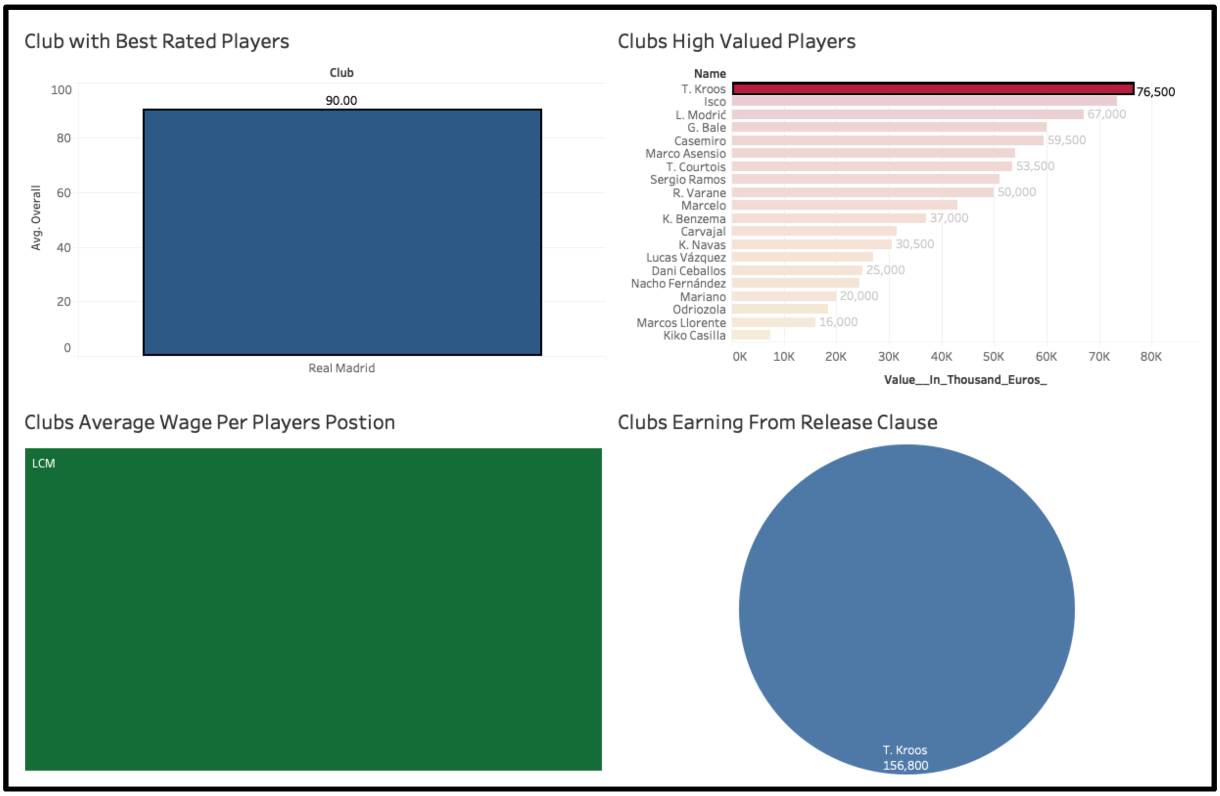

Figure 7: FIFA Dashboard 6

If a particular player is selected from the list of players his individual details would be displayed on screen as can be seen above. Here T. Kroos of Real Madrid is selected, and from this dashboard we get to know that his overall score is 90, he plays at LCM position and his Value and Release Clause amounts.

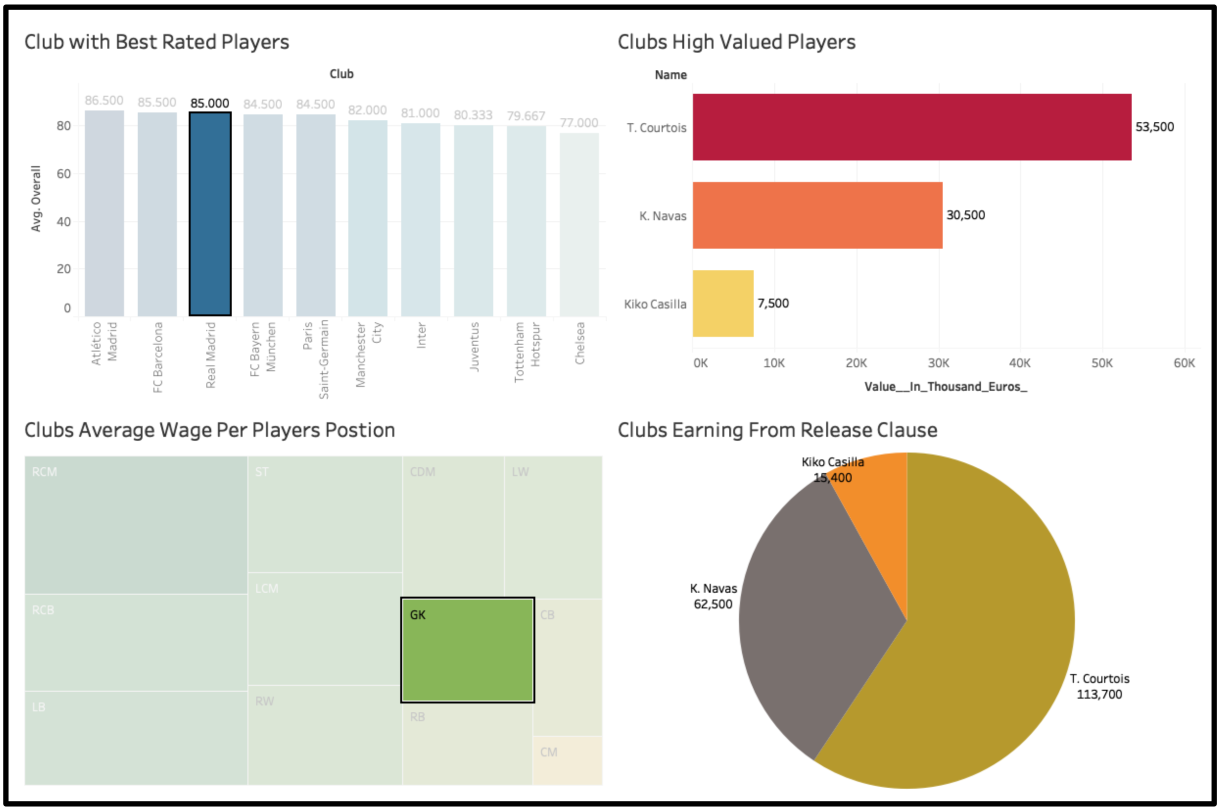

Figure 8: FIFA Dashboard 7

We can also choose to get information based on players position by selecting and of the positions. Here the position of Goalkeeper is selected, and it can be interpreted that there are three goalkeepers for Real Madrid, the most valued of them is T. Courtois and the club would earn the most by releasing him if they choose to release any of their goalkeepers.

The screenshots basically sum up all functionalities of our dashboard, but of course numerous trends can be explored using these visual representations.

8. Results

Results that we desired to achieve from the research study are as follows:

- Based on the 'Release Clause' data which football club can potentially earn the most by releasing its players?

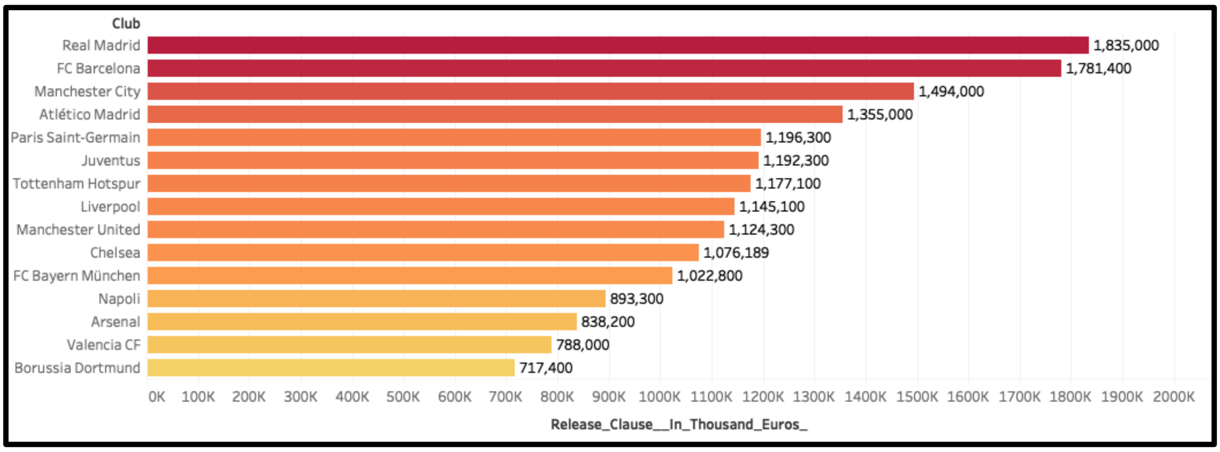

Figure 9: Club vs Release Clause

Bar chart above shows a list of top 15 clubs which would earn the most if thought of releasing all the players in it. Club names are mentioned on the x-axis against release cause amount on the y-axis, the release amount displayed here is in thousand euros. Real Madrid, FC Barcelona and Manchester City take up the first three positions respectively. These clubs generally have some of the best rated and high value players in them hence the graph that we have received is somewhat expected.

- Players from which nationality are the most dominant in this sport?

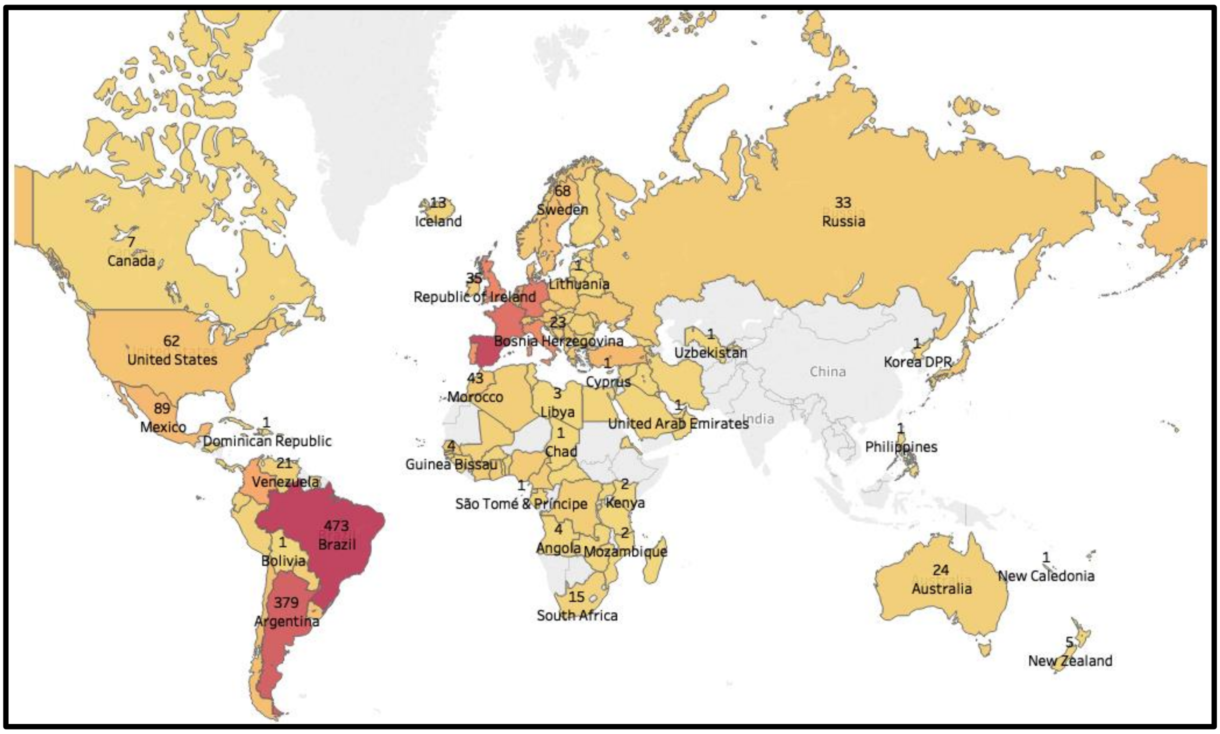

Figure 10: World Map representing players count as per Nationality

Out of five thousand players that are present in the dataset majority of them are either from South American regions or from European regions. Footballers from these parts of world seem to dominant in this sport. Countries like Brazil and Argentina contribute the most from South America whereas Spain, France, Germany, England and Portugal are of importance in the European region. Also going by the fact that football is generally the most popular sports in these areas, the results we got are obvious.

- Analyze which club is the most valuable, has the best rated players, pays the highest wage on an average to its players?

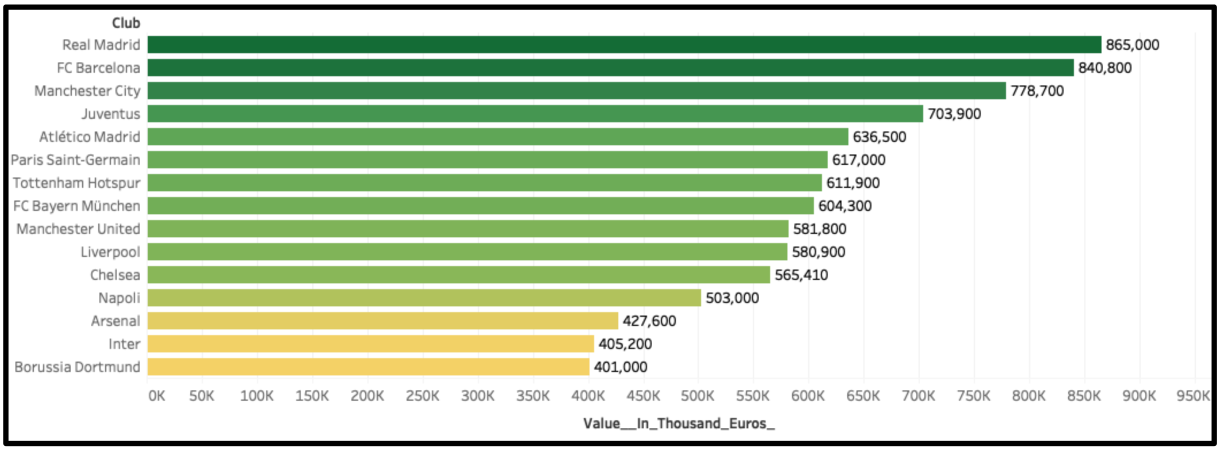

Figure 11: Club vs Value

The above Bar chart enlists top 15 most valuable clubs in the world of football. Total value of the club is aggregated as sum of value of all players assigned to a club in thousand euros and is plotted against the y-axis, club names are plotted on the x-axis. As seen in the above analysis Real Madrid, FC Barcelona and Manchester City retain the first three positions respectively here as well. Juventus and Atlético Madrid take up the fourth and fifth position respectively.

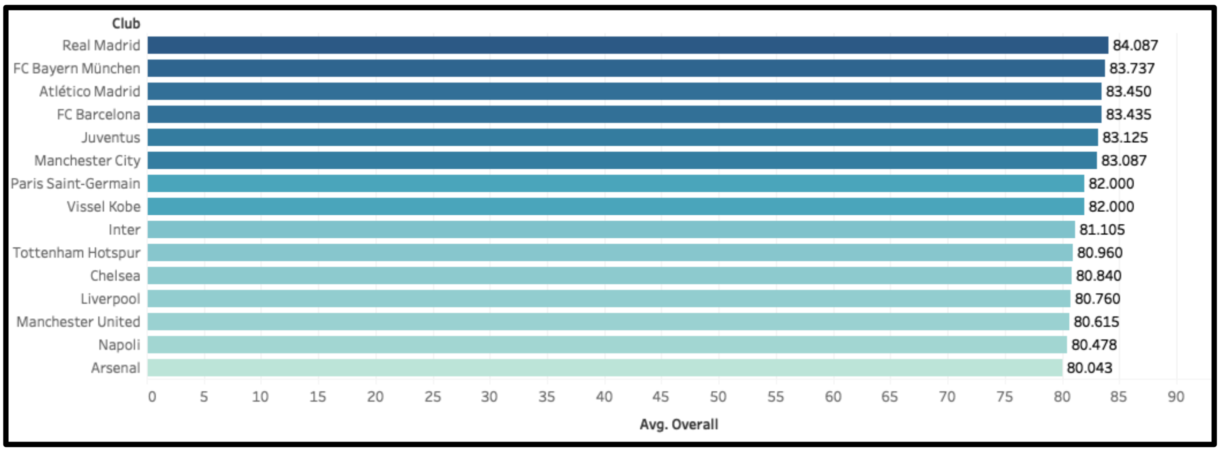

Figure 12: Club vs Average Overall Score

This bar chart represents top 15 clubs with the best rated players, numbers that can be seen here is an average of overall scores of all players in that particular club, remember that we have only considered five thousand players to get these results. As per the outcome above, Real Madrid with an average overall score of 84.087 tops the list. FC Bayern München with score of 83.737 and Atlético Madrid with 83.450 average score come in at second and third positions respectively. Even though Lionel Messi and Cristiano Ronaldo have the highest overall ratings, their respective teams FC Barcelona and Juventus do not feature in the top three as we are considering an average overall score of all players in their club and it seems like some players in these clubs don't have a good overall score.

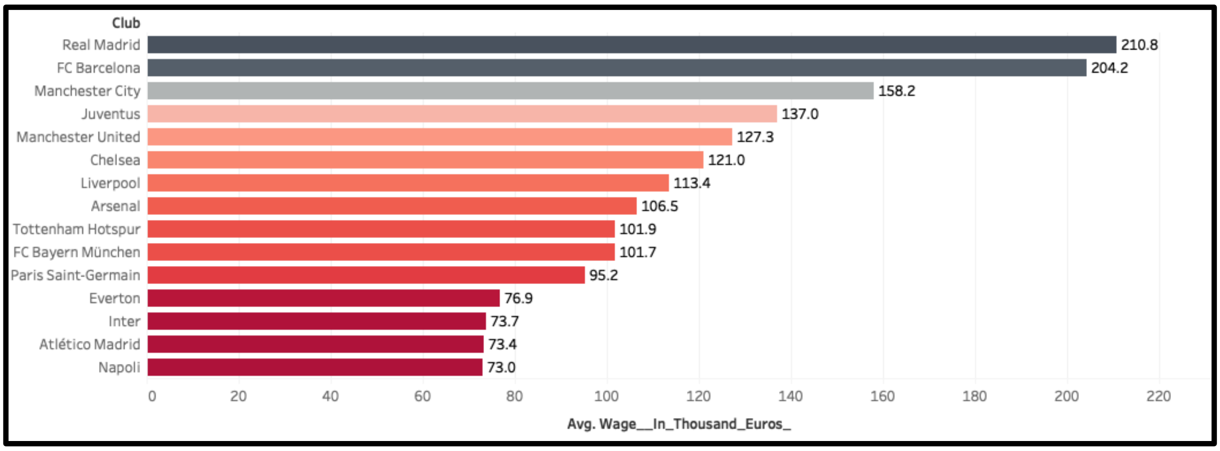

Figure 13: Club vs Average Wage

As is the common believe that since Real Madrid and FC Barcelona have the most valued players, they would be paying high wages to them and we do get the expected results from this analysis. As seen in the list above of top 15 clubs that pay the highest wage on an average, Real Madrid with an average wage of 210.8 thousand euros has acquired the first position, FC Barcelona with 204.2 thousand euros as average wage takes the second position while Manchester City with an average wage of 158.2 thousand euros is at the third in the list.

- What is the age distribution amongst the players?

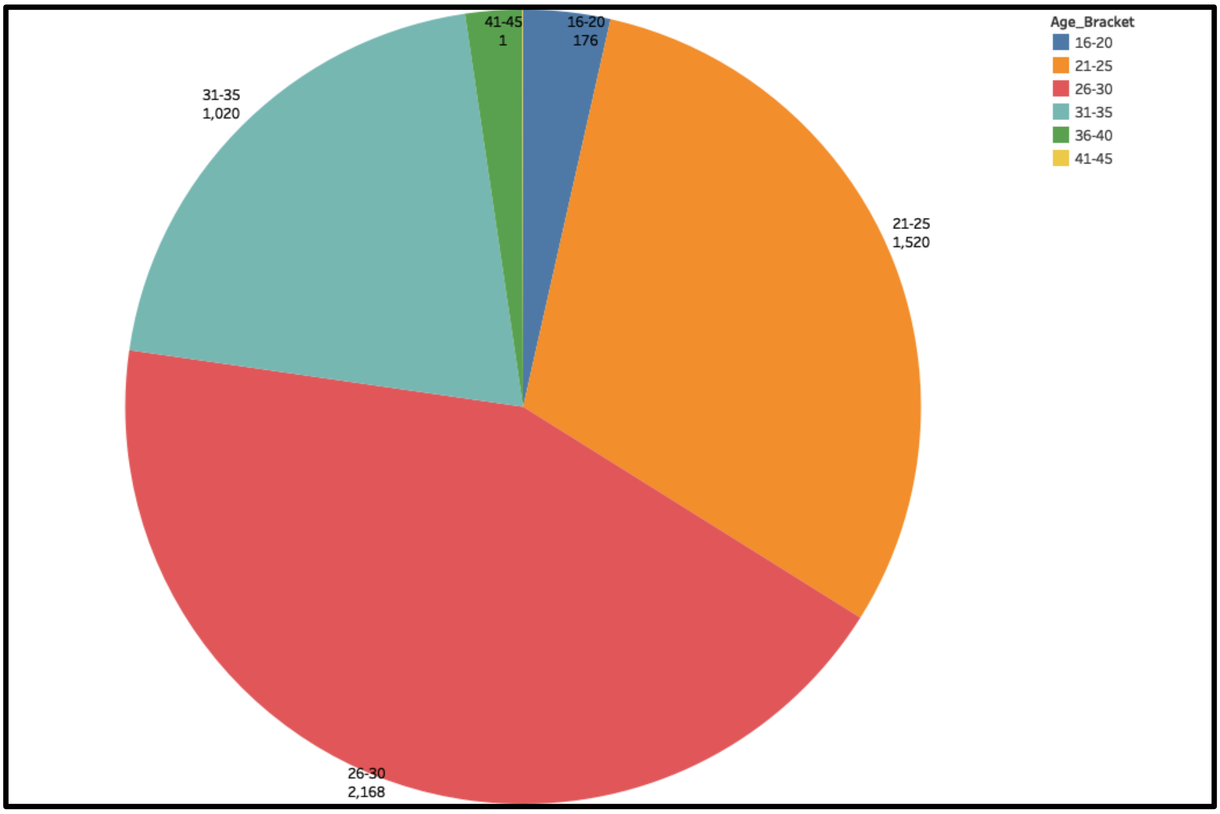

Figure 14: Players Count vs Age Bracket

As can be seen from the pie chart above, majority of the players fall in the age bracket of 26-30 it is also probably that during this phase they are at the peak of their careers. Many footballers being their careers at an early age hence we can see a good number of them between the 21-25 age bracket. This being a sport of strength and agility, not many continue playing it for a very long time even though we can see a decent count of players this is 1020 between the age of 31-35, there are hardly few players beyond that.

- Assess how much clubs spend based on players position, what are the dribbling speed, agility, shot power etc. of its high rated players?

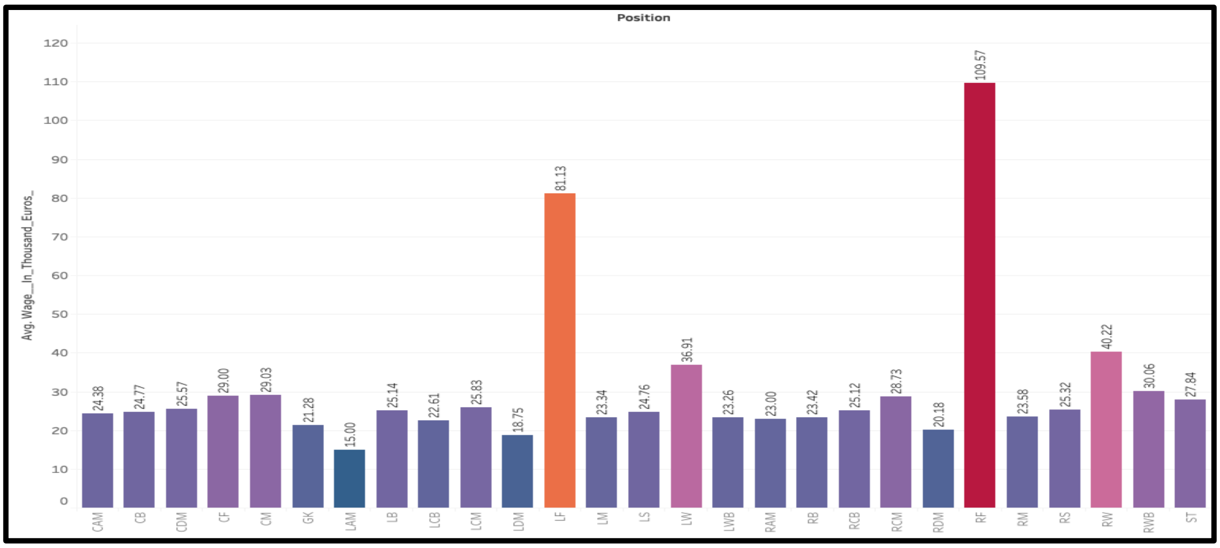

Figure 15: Average Wage Vs Players Position

If we take wage of all five thousand players into consideration and aggregate their average based on position at which the players play, we observe that right forwards (RF) end up getting the highest wage while left forwards (LF) receive the second highest wage on an average scale.

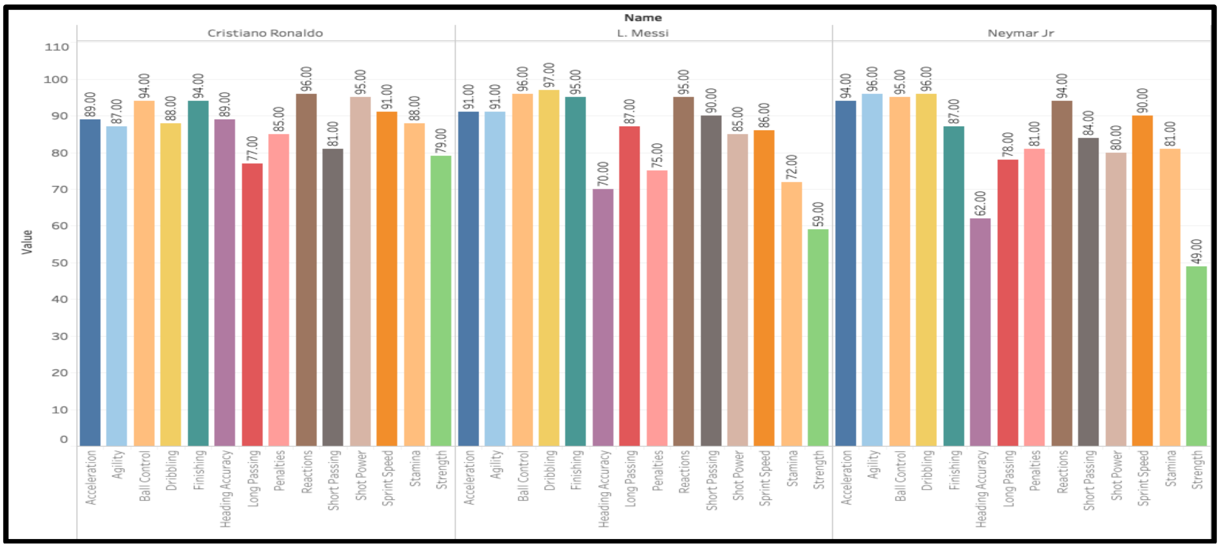

Figure 16: Count vs Skill Set

Histogram above shows various skill sets of a player which in return determine the overall score of a player. Above we have displayed details of only the top three players as per their overall score. Data can similarly be analyzed for all other players as well.

- Is there any correlation between a player's potential or overall score and his other skill sets, if yes, which skills tend to have more impact?

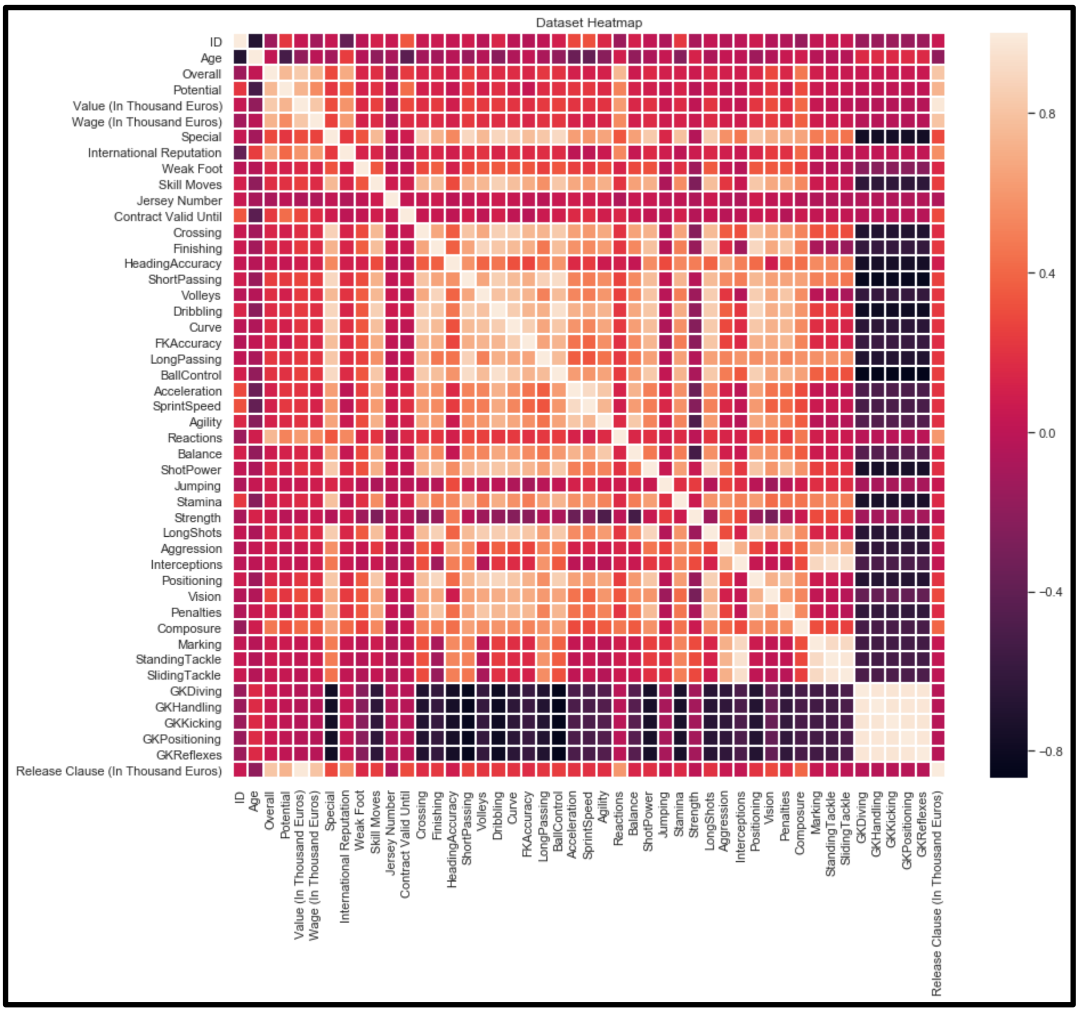

Figure 17: Dataset Heatmap

We have potted a heat map above for the entire dataset to find out hidden correlations between different attributes. As can be seen above potential and overall score does seem to be correlated with most of the

other skill sets. So, it can be stated that other physical attributes of a football player do have an impact on deriving the overall and potential score of that player.
Similarly, we can also see that age has a negative correlation with skills like sprint speed and acceleration, as expected players at younger age would be more agile. Release clause amount, Value and wage these all seem to have a negative correlation with age as well. Various other relationships can be sorted from this heat map.

- Is there any correlation between a player's age and overall score?

The scatter plot demonstrated below does reveal that the overall score of a player is somewhat dependent on the age of that person. It can be figured out that players between the age 25-35 seem to have a high overall score. This might be related to their physical capabilities as well as their performance in the international scenario. Thus, it can be concluded that the correlation between a player's age and his overall score is negative.

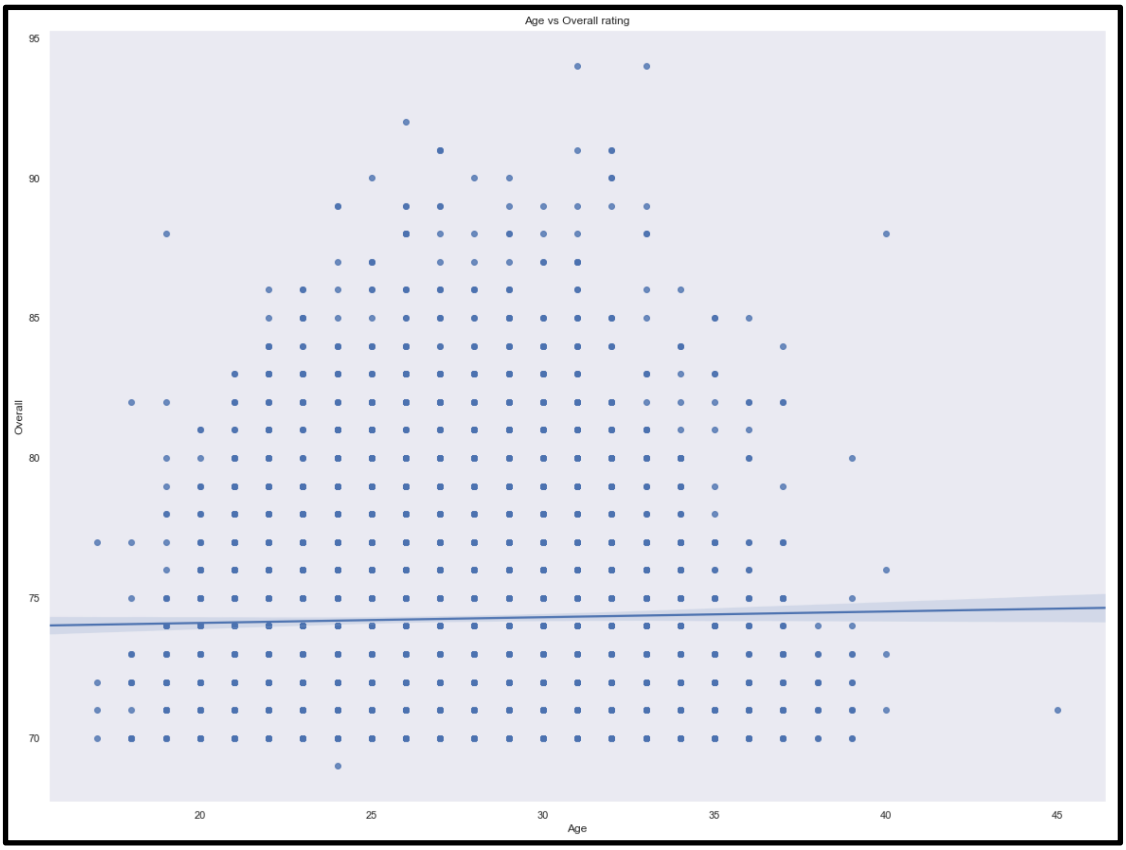

Figure 18: Age vs Overall Score

9. Evaluation

I believe the overall design of dashboard which provides a graphical representation of FIFA 19 players and clubs is at par with the requirements of our research study. It is suitable for all young and old gamers who wish to view some interesting stats about the players or clubs in order to make verified and enlightened choices before starting off a match in the video game. Even sports journalists can make use of this dashboard to write well-versed and informative articles. The positioning of different bar charts, heatmaps, pie chart or world map is done in a sequential manner to interpret the flow of data information which can easily be decoded by our users, but we do offer flexibility and it is not absolutely necessary to follow the determined flow of data. Correlations between various attributes are vividly represented and the colour scheme does make it easier for user to understand these relations hassle free and seamlessly.

Although we do like and appreciate our dashboard, but it could be improved and be even better. For instance, we have only used the best five thousand football players along with their fifty nine attributes for our analysis but if the entire dataset of 18,207 records and 89 characteristics is used then the results would definitely show more accurate results. Maybe few more analytical statistics could have been added in the dashboards for better understanding of data. In future we could definitely like to work on these points.

10. Discussion and Conclusions

Our findings suggest that Real Madrid and FC Barcelona are not only the most popular clubs in the world but in fact the best two clubs to play with or for. Research analysis do show that these clubs have the best rated players and are the ones who pay the most wage on an average to their players. When it comes to players, it is observed that most of the players come either from South American region or from the European region. It's also a known fact that football is a sport is most popular generally in these regions only. As stated in our research paper earlier, there is a public notion that either Lionel Messi or Cristiano Ronaldo is the best player and as analyzed from the data Lionel Messi does have the best ratings and he gets the first position whereas Cristiano Ronaldo is the second best in the world. They are also currently few of the most valued and highly paid players in the world. Even though these results were obtained from five thousand data records, we do not expect the output to change much even if the entire dataset is considered only the number might change a bit. Some of the correlations that we came across did surprise us, especially where there is negative or positive correlation between data. Machine learning algorithms such as linear regression was used to these correlations.

In this study, we examined correlations between different datapoints. We have made comparisons amongst various football clubs and players with the intention to explore hidden trends in them. After examining the outputs, we can now compute which team would earn the most by releasing all its players, what is the age distribution, which nationality or region is the most dominant in football, etc. The dashboard the we have developed does provide all these functionalities but there is always scope to improvise and enhance the dashboard. Future researches should consider exploring entire dataset instead of just the five thousand records that we have researched with. Regardless, our results are on point with the problem statement we have stated above.

References

1. https://en.wikipedia.org/wiki/FIFA_(video_game_series)
2. https://www.ea.com/games/fifa
3. https://www.kaggle.com
4. http://openrefine.org
5. https://cloud.google.com/bigquery/
6. https://www.rstudio.com
7. https://www.python.org
8. https://www.tableau.com
9. https://seaborn.pydata.org/tutorial.html
10. https://towardsdatascience.com/linear-regression-understanding-the-theory-7e53ac2831b5
11. https://imotions.com/blog/analyze-heat-maps/