Data analytics in football sport to identify gaps for the improvement of quality opportunities throughout world-wide teams

Syed Ali Fathima S J, Sumathi V P, Sumanth S

Abstract: Football is a widely known sport. Billions watch and play the game around the world. Data Analytics has assumed a huge role in the world of Football. It has transformed how people approach games, team formation, player selection etc. Data analytics has enabled teams from around the world to understand their game better and perform better. Data analytics is also used to predict the outcomes of games enabling people to make educated guesses while betting. There is no doubt that Football is worldwide sport. However, there are so many teams worldwide who haven't improved when compared to some of the others. Few teams don't even manage to make into the main tournaments like FIFA. Some countries lack funding and some teams don't have the exposure to standard equipment, coaching opportunities etc. It is very important for a Football enthusiast to know that the game keeps evolving towards a point where there are more quality teams around the world. It is very important for data analytics to move into this direction of finding answers to the question "What can be done to provide quality opportunities to the teams worldwide?". The present paper discusses exactly that and looks to provide an answer to that very question.

Keywords: Data analytics, Football, Pandas, Players, Python, Sports, Statistics.

I. INTRODUCTION

Football game was always the most famous and popular sport in various European and South American countries that has been played and viewed. The popularity of the sport, however, recently started booming in the Indian subcontinent since the early 21st century, probably owing to the popularity of televisions and global broadcasting of the games. Data analytics has helped in many ways in the game. The first ever international match was played in 1872 between two neighboring countries – England vs Scotland. The match ended in a 0 – 0 draw. 146 years later the game of Football has grown from its roots in Europe to reach all around the world with 211 national men's teams currently playing every year. Data Analytics is playing a big role in the game of football. Current systems are being used to predict outcomes in a game which is made useful by people for betting on their teams. People are able to make an educated guess on who would win a match thus enabling their chances of making money. Data Analytics is also used to extract hidden information in the game. This enables team managers, coaches and the players understand their own game better, their mistakes, their opponent's strategies, weaknesses, etc. Teams are able to perform much better with all this information about their games. There's a 46% global interest in the game of Football. 20% of the world population take participation in the game. This means that only 20% of the world's population play football once every week. Out of this 20%, only 4% play football actively or professionally [1]. This shows us that there's a lot of room for growth. But that the challenge is that Football is already the most popular sport in the world and what can we do to improve the growth? This paper discusses how a geographical analysis is done to understand where international football matches are played. It is done also by looking at where the FIFA World Cup matches played. This has helped in realization of the first step that needs to be taken for the growth in popularity. It is also seen how playing friendly matches has changed over the years to understand what teams have felt about it. This will be very important information for upcoming teams.

II. BACKGROUND STUDY AND MOTIVATION

Football game was always the most famous and popular sport in various European and South American countries that has been played and viewed. With the popularity of global and television broadcasting of the games, football game recently started to popularize in the Indian subcontinent since the early 21st century. India is steadily becoming a global figure in the Football (American: Soccer) world, with more and more official football events happening and also major international stars participating in the new Indian Super League [15].To better understand the constraints that promote the sport or success in the game, game analysis has become a very important role in sports games. In football, performance can be defined as the interaction of different factors such as technical, tactical or even mental. Since the 1990s, and through the creation of international scientific societies (e.g., International Society of Performance Analysis of Sport), the edition of specialized scientific journals (e.g., International Journal of Performance Analysis in Sport; Journal of Quantitative Analysis in Sports) and the introduction of world conferences on notational analysis (currently named, "World Congress of Performance Analysis in Sport"), game analysis has become a prominent idea in the scientific literature of the game. Game analysis is changing the game as is.

Revised Manuscript Received on December 30, 2018.

Syed Ali Fathima S J, Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India.

Sumathi V P, Assistant Professor, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India, ,

Sumanth S, UG Scholar, Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore, Tamilnadu, India

Players are improving; Managers are able to make good decisions, be it strategic or tactical. The game of football itself is growing so much so that it has now become one of the most if not the most popular sport in the world. Billions watch, millions play. To ensure successful execution of all tactics, a coach has to take into account the status of the team, the status of the opposition, as well as external factors like playing at home or even the weather. This is a place where Data Analytics plays a big role. Teams can be assessed based on different factors to define a winning strategy for the game.Football has played a huge role in the lives of people. It is and always has been one of the most popular sports played across the games. Teams that have been playing from the very beginning or those that have had very good exposure to international standards have improved over the years and have now established themselves as dangerous competitors. A lot of teams have either failed to catch on or have failed to receive any standard exposure to the world of football. A lot of the important football matches get played in countries that have done well themselves in the tournaments. A lot of matches have been conducted around the world but still a lot of teams fail to receive standard exposure. The motivation for this analysis is to find answers to these questions, i.e., how has the progress of the teams been over the years, why have teams not progressed, when did they lose momentum or did they receive quality exposure. It is also seen how playing Friendly matches has affected teams over the years by studying its trend each year. Such kind of exploratory analysis has been carried out in crowd estimation analysis at a social event using call data records [17].

III. OBJECTIVE

The major objective of this paper is to perform data analytics using Python Pandas in two fold on football sport dataset.

A. Geographical Analysis

The matches are analyzed based on their venue and its diversity in venues. This is done so to facilitate further analysis on the performance of teams to understand how conducting international matches in different countries changes the quality of football played. This helps to realize where to conduct matches.

B. Friendlies Trends

The number of teams choosing to play friendly matches has changed over the years. The objective of this analysis is to understand how this change has happened.

IV. FOOTBALL SPORT DATASET DESCRIPTION

Two football datasets have been used for the analysis. The main dataset contains the results of international football matches starting from the very first official match played in 1872 - England vs Scotland up to 2018 [16]. The matches range from World Cup to Baltic Cup to regular friendly matches. The other dataset includes the GPS coordinates of each city for geographical analysis. The two datasets have been combined to produce the football dataset and the sample dataset is given in Fig. 1. The dataset has 38903 records with 12 attributes. The dataset has been pre-processed to extract/add new attributes such as year, winners, latitude and longitude. Attributes like year and winners were extracted for convenience. The latitude and longitude coordinates were mapped to each city using another dataset that contains these data. The attributes in the final dataset are match_date, home_team, away_team, home_score, away_score, tournament, city, country, year, winner, latitude, longitude, Refer Table 1. The dataset is analyzed in different aspects to gain an insight into the game of Football.

	date	home_team	away_team	home_score	away_score	tournament	city	country	year	winners	latitude	longitude
0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	1872	Draw	55.833333	-4.25
1	1873-03-08	England	Scotland	4	2	Friendly	London	England	1873	England	42.983333	-81.25
2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	1874	Scotland	55.833333	-4.25
3	1875-03-06	England	Scotland	2	2	Friendly	London	England	1875	Draw	42.983333	-81.25
4	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	1876	Scotland	55.833333	-4.25
5	1876-03-25	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	1876	Scotland	55.833333	-4.25
6	1877-03-03	England	Scotland	1	3	Friendly	London	England	1877	Scotland	42.983333	-81.25
7	1877-03-05	Wales	Scotland	0	2	Friendly	Wrexham	Wales	1877	Scotland	53.050000	-3.00
8	1878-03-02	Scotland	England	7	2	Friendly	Glasgow	Scotland	1878	Scotland	55.833333	-4.25
9	1878-03-23	Scotland	Wales	9	0	Friendly	Glasgow	Scotland	1878	Scotland	55.833333	-4.25

Fig. 1. Sample of the Football dataset

Attribute	Description
match_date	date of match
home_team	team playing at home
away_team	team playing away
home_score	home team's score
away_score	away team's score
tournament	tournament type (Friendly, FIFA World Cup, Baltic cup etc)
city	venue of the match
country	country in which the match is being played
year	year of play(extracted from match_date)
winner	winner of the match(extracted from home_score and away_score)
latitude	latitude coordinate of the city in which match is being played
longitude	longitude coordinate of the city in which match is being played

TABLE I
FOOTBALL DATASET – ATTRIBUTES AND DESCRIPTION

V. GEOGRAPHICAL ANALYSIS

The matches played over the years were analyzed based on where the matches were held to understand how diverse the organizers have been in extending the range of the venues. The analysis was done to understand the popularity of the game spread across the world. The matches were plotted on a world map based on their venues. Fig. 2. depicts all the different types of International matches played including Friendlies, World Cups, AFC Asian Cup etc. Fig. 3. on the other hand, depicts only the World Cup matches played.

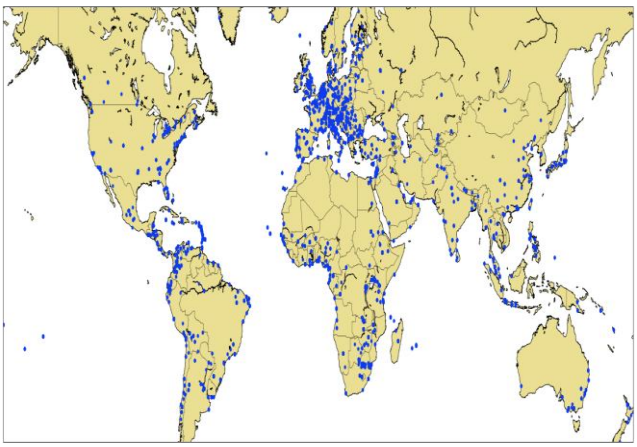

Fig. 2: World map plotted with venue of International Matches played

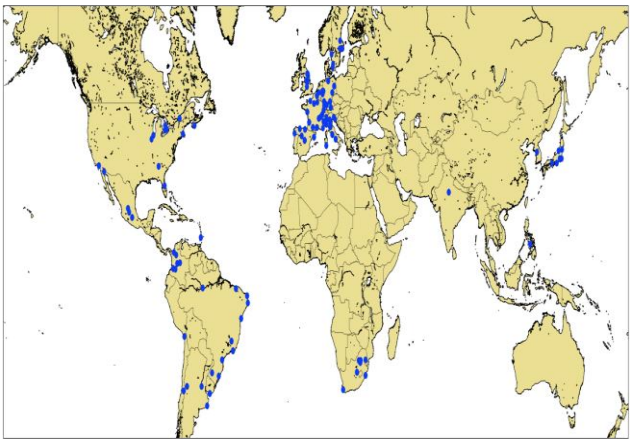

Fig. 3: World map plotted with FIFA World Cup matches conducted in the respective countries.

The analysis has helped to realize that the football matches have in fact been played around the world in a lot of different countries. But this has been as a result of which team is playing. When it comes to the world cup, the organizers have confined the venues to places wherein there is more popularity or in other words where the revenue generated will be most. It is seen that most world cup matches have been concentrated mostly in Europe and a little in South America. From this analysis it is realized that maybe if the tournaments like the World Cup is conducted in more places, the Global popularity will rise and there's a good chance of the quality of Football from these countries will improve exponentially.

VI. STATISTICAL ANALYSIS

A. Top 10 – All international matches won

In this analysis, teams are ranked in order of most matches won over the years refer Fig. 4. All types of matches have been considered in this analysis i.e., Tournament matches as well as the Friendly Matches. This analysis doesn't necessarily tell us which teams are the best given that friendlies are also considered.

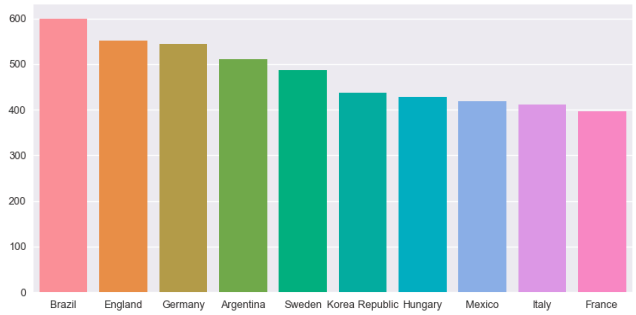

Fig. 4. Top 10 teams in terms of international matches won

B. Top 10 – Non-friendly matches won

The teams are now ranked on the basis of matches won excluding the friendly matches. This tells us the teams that have stepped up their game to perform in the tournaments. Right away it is seen that Hungary and France that were ranked 8th and 10th respectively are no longer in the top 10 refer Fig. 5. Instead Scotland and Uruguay who have made their way into the top 10 at position 5 and 6 respectively. This shows that Scotland and Uruguay have stepped when it was really of utmost importance and have performed well in the tournament matches.

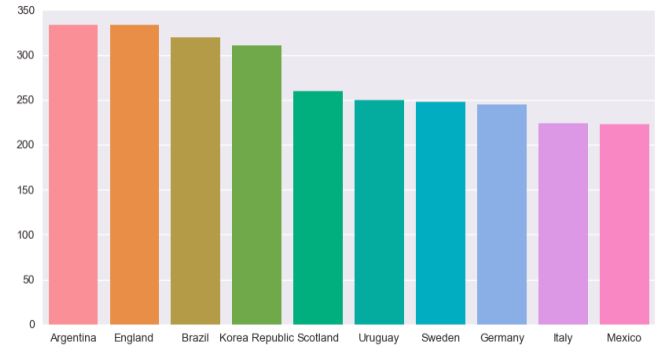

Fig. 5. Top 10 in terms of non-friendly matches won

C. Top 10 – Friendlies won

This analysis was done to rank the teams in terms of the friendly matches won. This helps us to understand which teams have performed the best in friendly matches and at the same time telling us which are the teams that have taken playing these Friendlies seriously refer Fig.6.

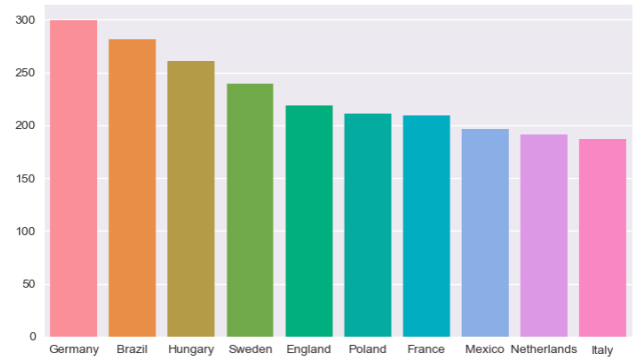

Fig. 6. Top 10 teams in terms of friendly matches won

D. Progress of the Top 10 over the decades

This analysis was done to understand how the top teams have progressed over the years to get to where they are now refer Fig.7. England and Scotland, being the very first teams to play an international match, started out in 1872 and have had a steady rise to the top. Around 1948 Scotland started lose its momentum and slowed down compared to England. Argentina and Uruguay started playing in the mid-1910s and Argentina has seen a very good rise to the top and Uruguay has been relatively slow. Brazil of course, has also made it to the top with a steady rise having started in the mid-1910s. Of the top 10, Korea Republic started the latest, mid-1950s, and has made it to the top 4 very quickly. It has seen the fastest growth of the rest of the teams.

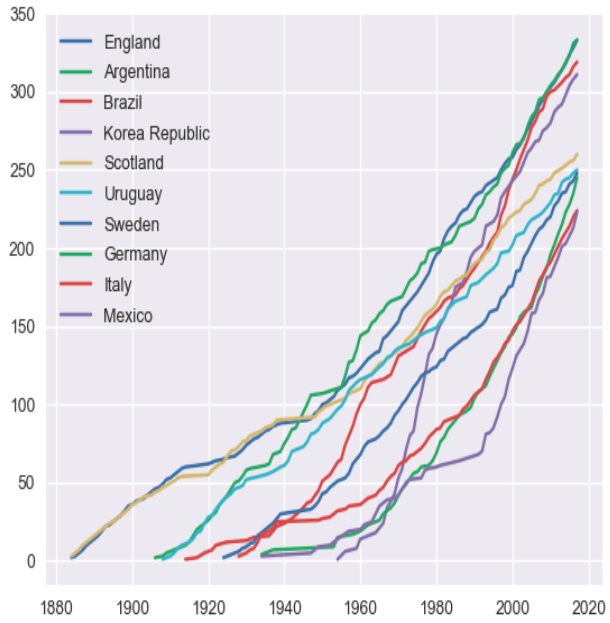

Fig. 7: Progress of the top 10 teams over the years

E. Top 5, Average 3 teams & India over the decades

This analysis shows that most of the average teams have started quite late as compared to the top 5 refer Fig. 8. With the exception of Czechoslovakia, which started much before Korea Republic. Korea Republic shot to the top 4 with extremely good growth rate while Czechoslovakia Football team couldn't progress much at all. On digging deeper it is realized that Czechoslovakia faced a lot of setbacks. Czechoslovakia was a sovereign state that existed from 1918 to 1992. The Czechoslovakia football team was formed in 1920. They continued playing till 1939. From 1939 to 1945, following its forced division and partial incorporation into Nazi Germany, the state did not de facto exist but its government-in-exile continued to operate. This is the reason why Czechoslovakia football team's progress plateaued from 1939 to 1945. After 1945 the team progressed well until Czechoslovakia's dissolution into Czech Republic and Slovakia in 1992. After that the Czech Republic football team was formed and as we can see that team has progressed very well owing to the team's years of experience. India started playing in the 1950s but showed slow progress as compared to a lot of other teams.

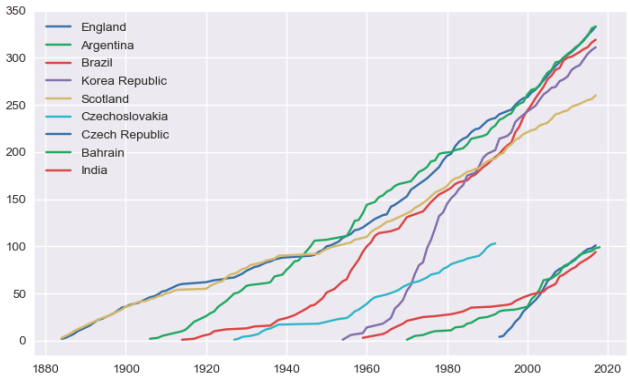

Fig. 8: Progress of the top 5 vs average 4

F. Trend of playing Friendlies over the years

We wanted to understand how the tendency of playing friendly matches has changed for teams over the decades. With this analysis it can be seen that the idea of playing friendly football matches started to catch on in 1920s. From then there has been a zigzag rise in the number of Friendlies played per year. In mid-2000s the world has seen the most number of Friendly matches played in a year, that's slightly 400 matches. Few of the reasons for playing friendlies could be – exposure, tactical or strategical analysis of the opponent teams or to understand the opposition's weaknesses. With time teams have realized the importance of playing Friendly matches and that is evident from the analysis as seen in Fig. 9.

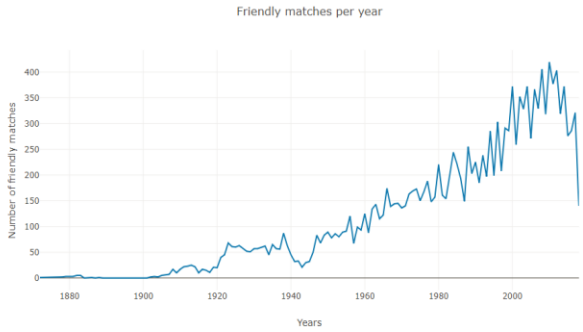

Fig. 9. Trend of Friendlies played each year

VII. Conclusion and Future Work

The game of football has been analysed in terms of geography, its top teams on the basis of different aspects, we have analysed their progress over the years and understood how these teams have performed in comparison to other teams that are as good as them. We have also made a comparison between teams in the top 10 and teams that are rather average. We have analysed the trends of Friendly matches played over the years helping us understand that teams around the world have realised that playing friendly matches with their oppositions is quite fruitful. We have seen that the most of the top 10 teams have more or less had big tournaments such as the FIFA World Cup in their home country. This has helped us realise the importance of conducting such tournaments in countries that are less productive in the game of Football. This can inspire and motivate more players to dream bigger and perform better.

With more data such as ball possession, momentum shifts during games, strategic team formation, player strengths etc., and further analysis can be done as follows.

- Analysing how playing Friendlies has affected each team's performance in tournaments.
- Machine Learning algorithm to predict the results of upcoming games/tournaments can be done

REFERENCES

1. M. Andrews, and P. Harrington, "Off Pitch: Football's Financial Integrity Weaknesses, and How to Strengthen Them", *Faculty Research Working Paper Series, Harvard Kennedy School*, 2016, Paper No. 311, [Online] Available: https://research.hks.harvard.edu/publications/workingpapers/Index.aspx
2. Davenport, T. H, "Analytics in sports: The new science of winning", International Institute for Analytics, 2014, no. 2, pp.1-28.
3. Gangal, A., Talnikar, A., Dalvi, A., Zope, V., and Kulkarni, A., "Analysis and Prediction of Football Statistics using Data Mining Techniques".,*International Journal of Computer Applications*, 2015, *132*(5), pp. 8-11.
4. Horne, J. D., and Manzenreiter, W., "Accounting for mega-events: forecast and actual impacts of the 2002 Football World Cup Finals on the host countries Japan/Korea." *International review for the sociology of sport*, 2004, *39*(2),pp. 187-203.
5. Rein, R., and Memmert, D., "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science", *SpringerPlus*, 2016, *5*(1) pp. 1410.
6. Sarmento, H., Marcelino, R., Anguera, M. T., CampaniÇo, J., Matos, N., and LeitÃo, J. C., "Match analysis in football: a systematic review.", Journal of sports sciences, 2014, 32(20), pp. 1831-1843.
7. Christopher Carling, Craig Wright, Lee John Nelson, Paul S Bradley. "Comment on 'Performance analysis in football: A critical review and implications for future research". Journal of Sports Sciences, 2014, 32:1, pp. 2-7.
8. http://nielsensports.com/global-interest-football/, Last Accessed 09/05/2018.
9. http://www.fifa.com/about-fifa/who-we-are/the-game/index.htm, Last Accessed 10/05/2018.
10. https://datafloq.com/read/how-big-data-is-changing-the-world-of-football/1796, Last Accessed 11/05/2018.
11. https://thesetpieces.com/features/football-and-data-the-future-of-analytics/, Last Accessed 11/05/2018.
12. https://www.footballhistory.org/, Last Accessed 12/05/2018.
13. https://www.theguardian.com/football/2014/mar/09/prmier-league-football-clubs-computer-analysts-managers-data-winning, Last Accessed 12/05/2018.
14. https://www.independent.co.uk/sport/football/premier-league/transfer-window-football-betting-analytics-moneyball-a7934181.html, Last Accessed 12/05/2018.
15. Manjula Sanjay, Shreyas Srinivasan and Rahul Kulkarni, Data mining technique for Best 11, International Journal of Conceptions on Information Technology and Computing Vol. 4, Issue. 2, 2016.
16. http://www.ai-man.ir/en/post/international-football-results-from-1872-to-2017-en/, Last Accessed 20/06/2018.
17. Sumathi, V P, Kousalya, K, Vanitha, V, Cynthia, J, "Crowd estimation at a social event using call data records", 2018, Int. J. Business Information Systems, Vol 28, Issue. 2, pp 446-461.