VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
VIT Chennai
Vandalur - Kelambakkam Road, Chennai - 600 127

PROJECT REPORT

Programme: CSE

Course Code: CSE3020

Course Name: Data Visualization

Slot: D2

Faculty: Dr. Joshan Athanesious J

Title: History of Football

Arnav Singh 20BCE1731
Archit Agarwal 20BCE1773

ABSTRACT

The goal of this project is to develop an interactive platform to visualise and explore the data of international football matches played between 1872 and 2022. Users will be able to explore the evolution of international football through detailed data and statistics on match results, goals scored, team performance, and player performance on the platform. To achieve this goal, we will collect and analyse data from various sources, including official football associations, media outlets, and online databases. To present the data in an understandable and engaging format, we will use data visualisation techniques such as charts, graphs, and maps. Users will be able to filter the data by teams, players, tournaments, and time periods using the platform. Users will also be able to compare and track the performance of various teams and players over time. Furthermore, the platform will provide insights into historical trends and patterns in international football, such as team dominance, the impact of rule changes, and the evolution of playing styles.

Overall, this project will provide an educational and interactive platform for football fans, researchers, and enthusiasts to learn about the rich history of international football and gain insights into its evolution.

Keywords: Visualisation, international football matches, data analysis, match results, goals scored, team performance, player performance, filter data.

INTRODUCTION

Football is the world's most popular sport, with millions of fans and players worldwide. International football dates back to 1872, when Scotland and England played the first official international football match. International football has grown in popularity and significance since then, with countries competing in events such as the FIFA World Cup, the UEFA European Championship, and the Copa America.

Football data visualisation is a new field that uses data analysis and visualisation techniques to better understand and present the performance of football teams and players. With the increasing popularity of football, there is a greater demand for comprehensive and interactive platforms that allow users to explore data and gain insights into the sport. This project aims to create such a platform by visualising data from international football matches played between 1872 and 2022, providing users with a valuable resource for exploring international football's rich history. The visualization will include data on match results, goals scored, team statistics, and player statistics. The data will be presented in a user-friendly manner, allowing users to easily access and analyse the information.

LITERATURE SURVEY

1. State of the Art of Sports Data Visualization

Link:https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13447

Summary: This report summarizes recent developments and challenges in the field of sports data visualization, which is rapidly growing due to the importance and timeliness of sports data visualization research. Sports data visualization research covers a wide range of visualization tasks and goals, such as developing new visualization techniques, adapting existing ones to a novel domain, and conducting design studies and evaluations in close collaboration with domain experts, including practitioners, enthusiasts, and journalists. The report analyzes current research contributions through the lens of three categories of sports data: box score data, tracking data, and meta-data. The report concludes with a high-level discussion of sports visualization research that identifies critical research gaps and valuable opportunities for the visualization community. Sports visualization offers new approaches to exploring, understanding, and communicating sports data. The report highlights that visualization can be more accessible and more meaningful than traditional statistical analysis, and the number of visualizations of sports data has grown rapidly over the past decades. The report provides a classification of current sports visualization work and identifies the opportunities and challenges for the future of this fascinating subfield.

2. A survey of competitive sports data visualization and visual analysis

Link:https://link.springer.com/article/10.1007/s12650-020-00687-2

Summary: The paper discussed in this prompt provides a taxonomy of sports data visualization and summarizes the state-of-the-art research in this field. The paper first classifies sports data into two categories: spatiotemporal information and statistical information. Then, it proposes three main tasks for competitive sports data visualization: feature presentation, feature comparison, and feature prediction. Furthermore, the paper classifies competitive sports data visualization techniques based on data characteristics into five categories: high-dimensional data visualization, time-series visualization, graph (network) visualization, glyph visualization, and other visualization. The paper also introduces visual analysis research work of competitive sports, proposes the features and limitations of competitive sports data, summarizes multimedia visualization in competitive sports, and discusses visual analysis evaluation.

The paper emphasizes the importance of competitive sports data visualization and visual analysis in studying human behavioral patterns and activity habits. The rise of competitive sports data has given impetus to the development of research in competitive sports and has simultaneously provided a basis for the study of the law of human life and the habits of human beings. With the increasing requirements of data analysis, user interfaces based on visualization and visual analytics have become widely used. Thus, competitive sports data visualization and visual analysis are becoming a hot topic in the research field.

The paper concludes by identifying the challenges involved in visual analytics in the competitive sporting realm. It suggests that methods on how to choose appropriate visualization techniques for different data types, data attributes and derived features, how to better mix different data attributes to present data visualization and visual analysis results, and how to establish significant assessment frameworks are three urgent problems that need to be discussed in future research.

3. A visualization system involving NBA visual analysis and integrated learning model prediction

Link: https://www.sciencedirect.com/science/article/pii/S2096579622000833

Summary: The article describes the development of an interactive visualization system called NPIPVis that uses machine learning and data visualization techniques to help the general audience understand and analyse NBA game data and predict all-star players. The system includes dynamic hypergraphs of NBA team wins and losses, game plot narratives, and an integrated learning model called SRR-voting to predict all-star players. The article highlights the effectiveness and practicality of the SRR-voting model and suggests that NPIPVis can be extended to other sports events or related fields.
Overall, this study highlights the importance of data visualization and machine learning in the field of sports analysis, particularly in the NBA. By combining these techniques, the NPIPVis system can provide a more intuitive and user-friendly way for the general audience to understand and analyse NBA team and game data. Moreover, the SRR-voting model proposed in this study can predict all-star players with high accuracy, improving the efficiency of team selection and overall performance in the league.

In the future, this approach can be extended to other sports events and related fields, providing a valuable tool for coaches, players, and fans to analyse and predict performance. Additionally, further research can explore the integration of other machine learning models and data visualization techniques to enhance the interpretability and effectiveness of the system.

4. IPL Visualization and Prediction Using HBase

Link: https://www.sciencedirect.com/science/article/pii/S1877050917327023

Summary: The paper proposes a data visualization and prediction tool for IPL cricket matches using HBase, a distributed, open-source, and non-relational database. The tool can be used for player profiling and predicting the outcome of matches through various machine learning approaches. The paper suggests that the tool can be beneficial for team management in selecting the right team during player auctions. The paper addresses the challenges involved in predicting the accurate results of a game due to the varied parameters involved. The proposed approach uses key features such as luck and player strength for predicting the winner of a match. The paper concludes that KNN has been the most accurate machine learning algorithm used in the experiments.

5. Building an interactive visualization tool for athletes' performance data

Link: https://www.diva-portal.org/smash/get/diva2:1467010/FULLTEXT01.pdf

Summary: This paper discusses the design and implementation of an interactive visualization tool for sports data analysis using a problem-driven approach and continuous feedback from domain experts. The study found that visualizations enable intuitive insight generation and improve data analysis, making it more sophisticated and accessible. However, tailoring the tool to specific sports may affect transferability, and the multidimensionality of the dataset presented challenges for future work. Overall, the concept shows promise for future research, development, and implementation in collaboration with domain experts.

6. Why Is Data Visualization Important? What Is Important in Data Visualization?

Link: https://hdsr.mitpress.mit.edu/pub/zok97i7p/release/4

Summary: Data visualization is an essential aspect of data analysis that involves drawing graphic displays to represent data in a visual form. It aims to transform raw data into a more meaningful and understandable format to make it easier for users to analyze and interpret. The displays can be simple, such as a scatterplot, or more complex, such as a histogram, and may include statistical summaries of the data. Visualization can reveal data features that may be missed by statistics and models, such as unusual distributions, gaps, outliers, and local patterns. It also helps identify trends and clusters, detect outliers and unusual groups, evaluate modeling output, and present results. It is crucial for exploratory data analysis and data mining to check data quality and help analysts become familiar with the structure and features of the data before them.

Interpreting graphics requires experience to identify potentially interesting features and statistical knowledge to guard against overinterpretation. While static graphics are widely used in data visualization, dynamic and interactive graphics are in an exciting stage of development and have much to offer. They require their own article, as they allow users to interact with the data and explore different aspects of it. As such, data visualization is an important tool for data analysts, scientists, and decision-makers, enabling them to make informed decisions based on the insights they derive from the data.

7. What is data visualization and why is it important

Link:https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

Summary: Data visualization is the practice of presenting information in a visual context, such as a graph or map, to make it easier for humans to understand and extract insights. It is an essential step in the data science process and a key element of the broader data presentation architecture discipline. Visualization tools have become increasingly important in big data projects, allowing businesses to quickly and easily get an overview of their data. Data visualization provides a fast and effective way to communicate information using visual cues, and can help businesses identify patterns, predict sales volumes, and pinpoint areas that need improvement. Other benefits include the ability to absorb information quickly, make faster decisions, and eliminate the need for data scientists. By improving insights and understanding, data visualization can help organizations achieve success with greater speed and fewer mistakes.

8. Hybrid design for sports data visualization using AI and big data analytics

Link: https://link.springer.com/article/10.1007/s40747-021-00557-w

Summary: The study aimed to develop a video-based effective visualization framework (VEVF) for sports data analysis and visualization. The framework is based on artificial intelligence and big data analytics and is designed to extract both temporal and spatial features from sports videos. These features are then used to categorize the videos using machine learning, which helps to understand the collective tactical behavior of professional sports teams.

To evaluate the effectiveness of the VEVF model, the researchers compared its performance to that of other existing models. The experimental results demonstrated that the proposed VEVF model outperformed the other models in terms of accuracy, recall, F1-score, precision, error rate, performance, and efficiency ratios. Specifically, the proposed model achieved an accuracy ratio of 98.7%, recall ratio of 94.5%, F1-score ratio of 97.9%, precision ratio of 96.7%, error rate of 29.1%, performance ratio of 95.2%, and an efficiency ratio of 96.1%.

Overall, the study's findings suggest that the proposed VEVF model provides a solid foundation for developing fitness tools based on AI and big data analytics in professional team sports. The model can be used to evaluate fatigue, analyze performance potential, and reduce injury and illness risk by analyzing sportsperson monitoring data.

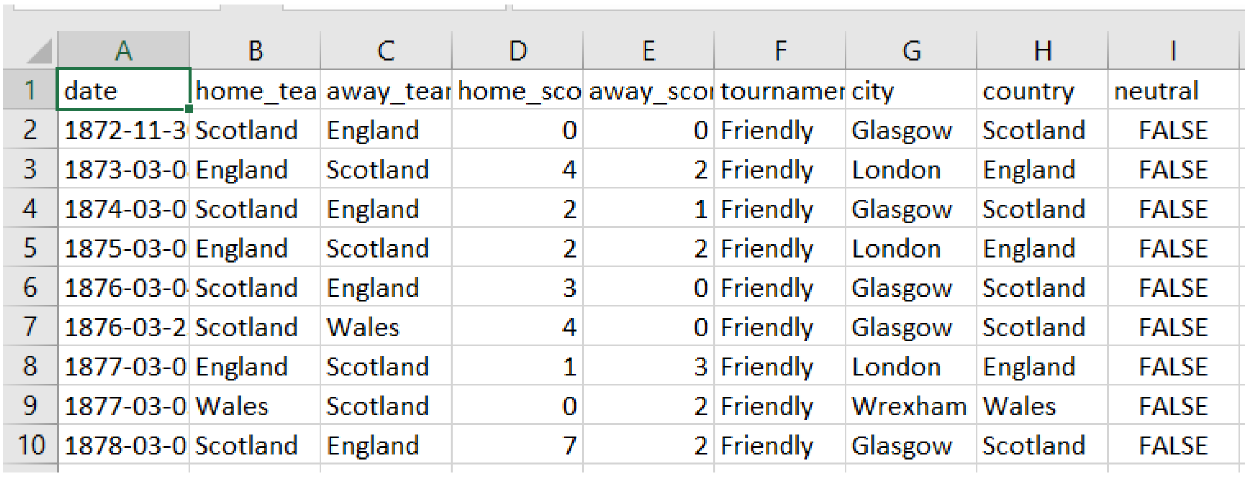
MATERIAL AND METHODS

Info about models

We have performed data analysis and visualisation using the Python and Pandas packages. Visualizations are also created using the Matplotlib and Seaborn packages.

- **Python** is a high-level programming language popular in data research and machine learning. It is well-known for its ease of use and simplicity.
- **Pandas** is a Python package that offers tools for data manipulation and analysis. It is based on the NumPy library and offers simple data structures for working with structured data.
- **Matplotlib** is a Python library that provides capabilities for data visualisation. It is commonly used to create charts, graphs, and other types of visualisations.
- **Seaborn** is a Python module that offers sophisticated data visualisation tools. It is built on Matplotlib and adds features for building more complicated visualisations.

DATASET



	A	B	C	D	E	F	G	H	I
1	date	home_tea	away_tear	home_sco	away_scoi	tournamer	city	country	neutral
2	1872-11-3	Scotland	England	0	0	Friendly	Glasgow	Scotland	FALSE
3	1873-03-0	England	Scotland	4	2	Friendly	London	England	FALSE
4	1874-03-0	Scotland	England	2	1	Friendly	Glasgow	Scotland	FALSE
5	1875-03-0	England	Scotland	2	2	Friendly	London	England	FALSE
6	1876-03-0	Scotland	England	3	0	Friendly	Glasgow	Scotland	FALSE
7	1876-03-2	Scotland	Wales	4	0	Friendly	Glasgow	Scotland	FALSE
8	1877-03-0	England	Scotland	1	3	Friendly	London	England	FALSE
9	1877-03-0	Wales	Scotland	0	2	Friendly	Wrexham	Wales	FALSE
10	1878-03-0	Scotland	England	7	2	Friendly	Glasgow	Scotland	FALSE

This dataset includes a comprehensive record of international soccer games between countries, starting from the very first game between Scotland and England in 1872 up until early 2018. It provides information such as the score, tournament, and host city and country for each game. This extensive dataset offers valuable insights into the evolution of international soccer and can be used for various purposes, including statistical analysis, research, and prediction modeling.

PROPOSED METHODOLOGY

Novelty:

The novelty of this project is its comprehensive analysis of the history of international football matches as well as its data visualisation to gain insights into the game. While there have been other studies on the subject, this project is notable for its use of programming and various libraries for data processing and visualisation. It also includes matches from 1872 to 2022, giving it a more comprehensive historical perspective on the game than many other studies.

Furthermore, the project focuses not only on identifying trends and patterns but also on how this information can be applied to the sport in practise. For example, the project reveals which teams have historically performed well in various aspects of the game, such as offence and defence, as well as which machine learning models are most successful in predicting match results. Teams can use this data to develop winning strategies and improve their performance.

Overall, this project is more beneficial than others due to its thorough analysis, practical applications, and use of cutting-edge technology. It offers useful insights that can assist teams in making data-driven decisions and gaining a better understanding of the sport.

Project contributions:

All of the team members have made significant contributions to the project. Arnav Singh’s visualizations have helped to communicate complex data in an easy-to-understand format, making the insights accessible to a wider audience. Archit Agarwal worked on pre-processing and handling the dataset which ensured that the data used in the project is accurate and relevant.

Together, our contributions have resulted in a comprehensive analysis of the history of international football matches. The project's visualizations have provided insights into the performance of various teams over time, the trends in the sport, and the strategies that can be used to win matches. This analysis has the potential to benefit football teams, players, and coaches by providing a deeper understanding of the sport and how to succeed in it.

RESULT AND DISCUSSION

Which teams play the most?

Surprisingly, Sweden has played the most games of any team. Major football nations like Brazil, Argentina, England, Germany, and France make up the majority of the top 10 nations. As participants in the first world cups (1930 and/or 1934), nations like Uruguay, Mexico, and Hungary are also considered veteran squads.

How many games per year?

There are few interesting things going on here:

- Number of games is rising, with high growth in the 80s/90s.
- It seems there is a peak around 2010, with a slight decrease since.
- We see a drop during world wars.
- Since the 80s, data is very spiky, likely due to the absence/presence of world cups or other events.
- World cup qualifications generates much more matches than the world cup itself

When do games occur?

Interestingly, the very first games mostly occur on Saturdays but a decent number also took place on Mondays. The first games mostly occur during Spring months and since then, some month have known some peaks of popularity for intenational games at different period (e.g. many games happened in December in the 1940s). In a more recent history, international games became less common in May but more in June.

Evolution of results

Let' know talk about sport and actual results! First let's check how the proportion of draws and home/away victories evolve through time. Main learnings are:

- A victory of the home-based team has always been the most likely event.
- A victory of the visitors is the second most likely outcome, although it tends to decrease in the second half of the 20th century.
- A draw has always been the least likely outcome, altough it has increased in share since the 1940's.

Best performing teams during soccer history

- **Best Defensive sides:**
    - Scotland also used to have a good defense.
    - England and Germany were solid during the 1930's and 1940's.
    - China and Tahiti were amongst the best defenses between the 1960's and 1980's.
    - Despite of being seen as an offensive team, Brazil was #1 and #3 best defense in the 1980's and 1990's.
    - Germany was the second best defense two decades in a row (2000's and 2010's)
- **Best Offensive sides:**
    - Scotland once, was one of the top scoring nations (OK, that was when max 10 teams were competing, but still) and slowly dropped from the top 6.
    - Sweden was consistently in the top 6 for 4 decades in a row (1910s to 1940s).
    - Fiji and Tahiti were at the top of the charts during some decades too, including some recent ones.
    - Zambia and China once were among the top scorers.
    - During the last 3 decades, Germany and Spain are the only major nations who made it twice to the top 6.

Are defense and attack correlated?

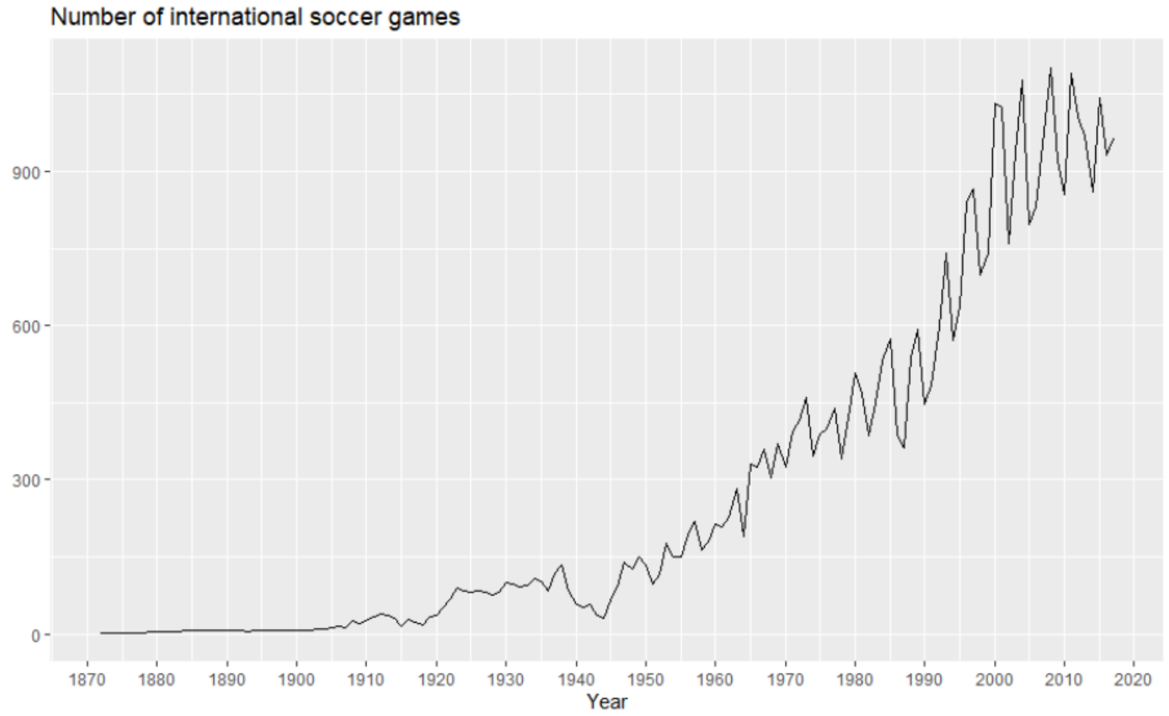
we can see that the teams scoring very few goals per game are also more likely to have a poorer defense. However, pat a given limit around 1.5 goals for per game, the quality of the defense remains rather constant. In general, teams above the line generally have a bad defense given their attack level and teams below the line have a better defense given their attack stats.

which team has the best win ratio?

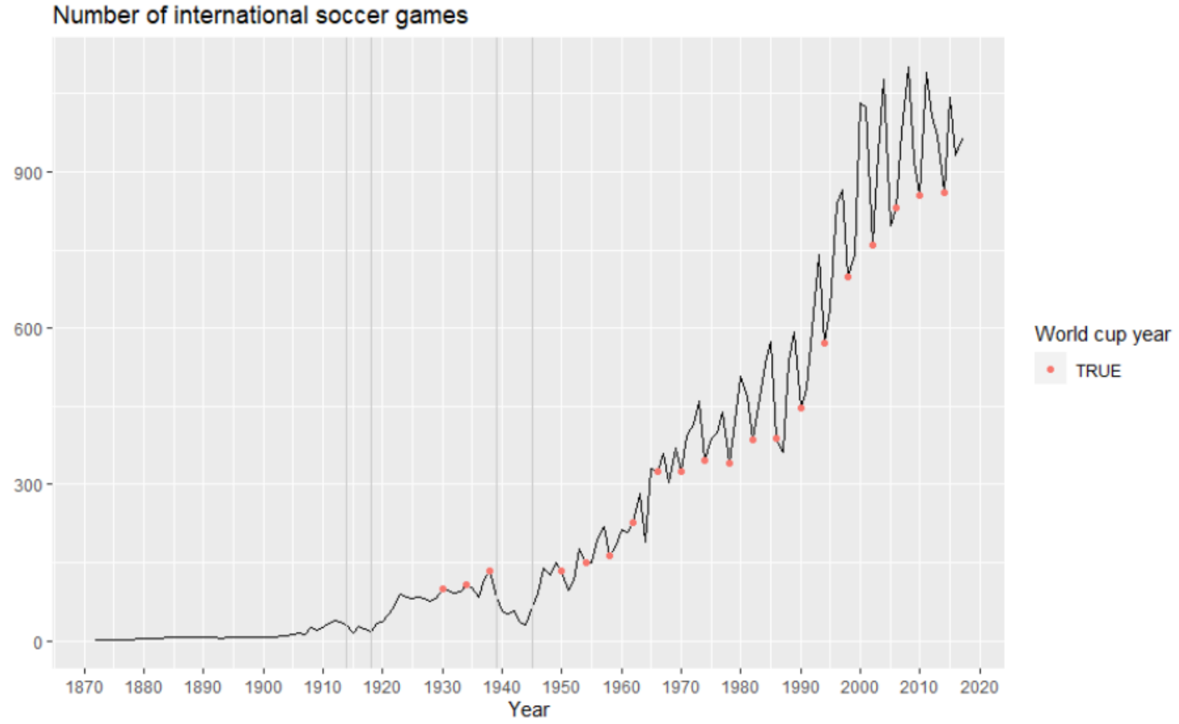
The top teams are not a surprise: Brazil, Germany and Spain. Some teams are more surprising such as Jersey or Northern Cyprus. Together with Brazil, Argentina and Iran are the only non-European countries in this top 10. Czech Republic and Croatia also make it to this top 10.

FIGURES AND COMPARISONS

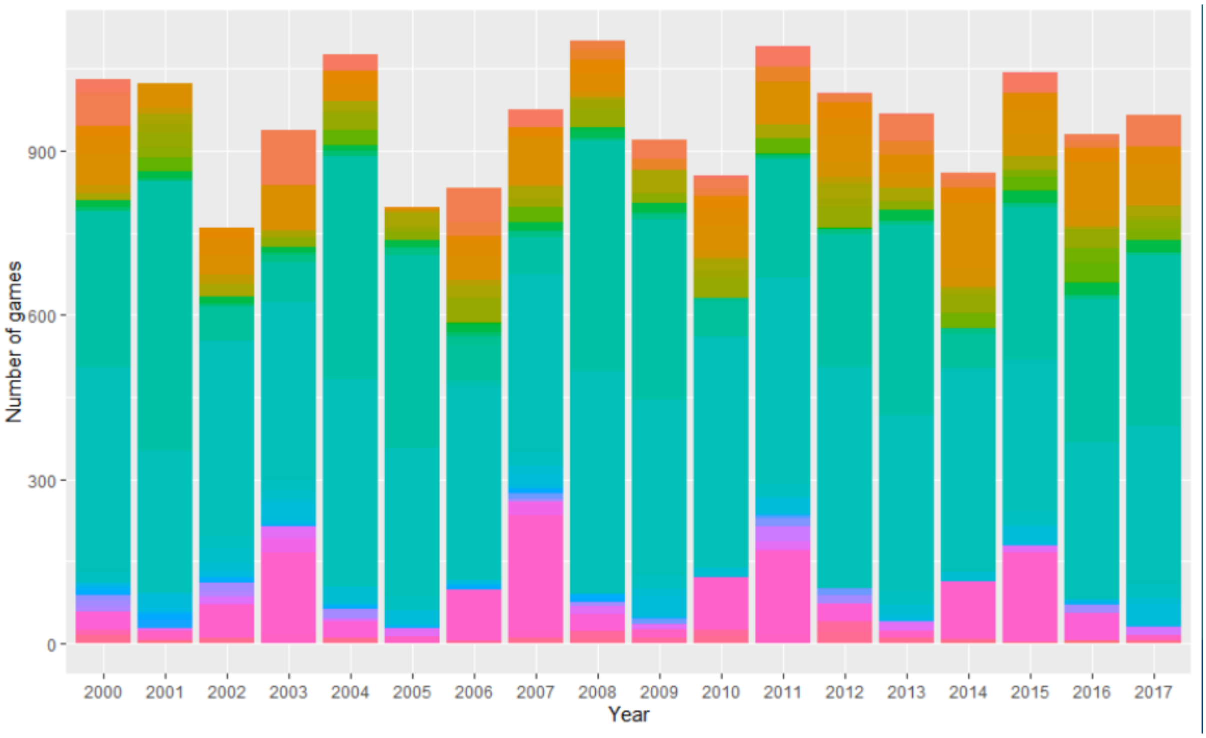
i. How games evolve with time?



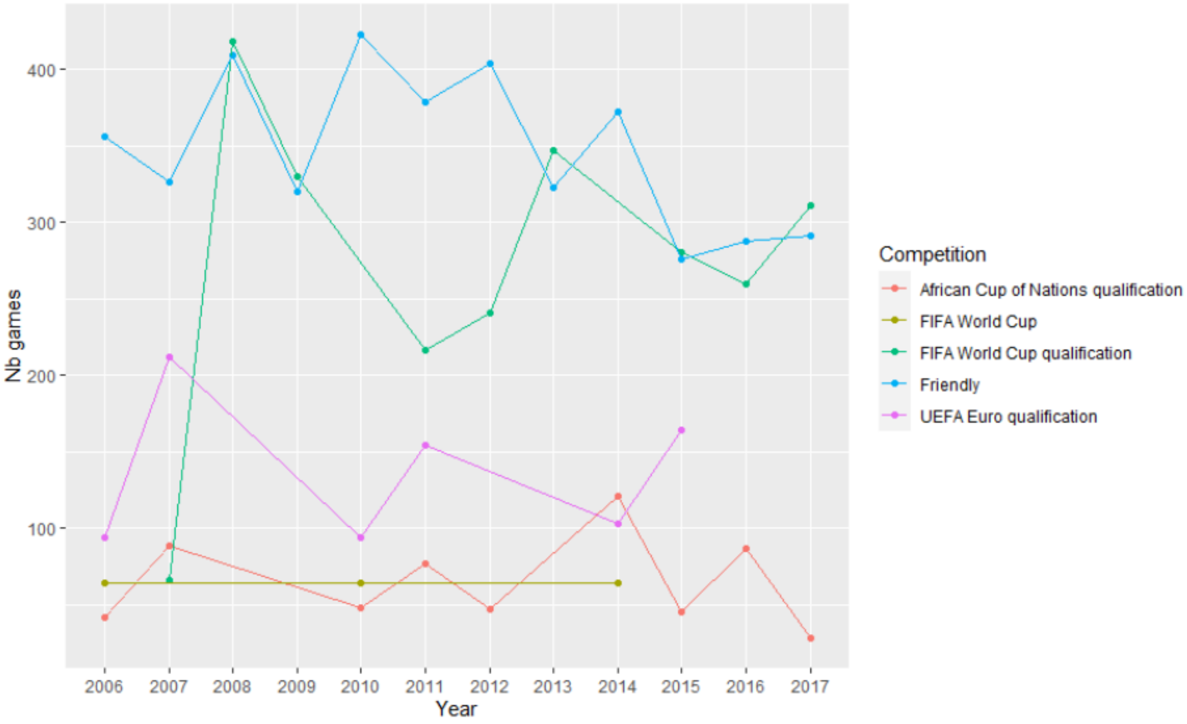
ii. How total number of world cup have evolved?



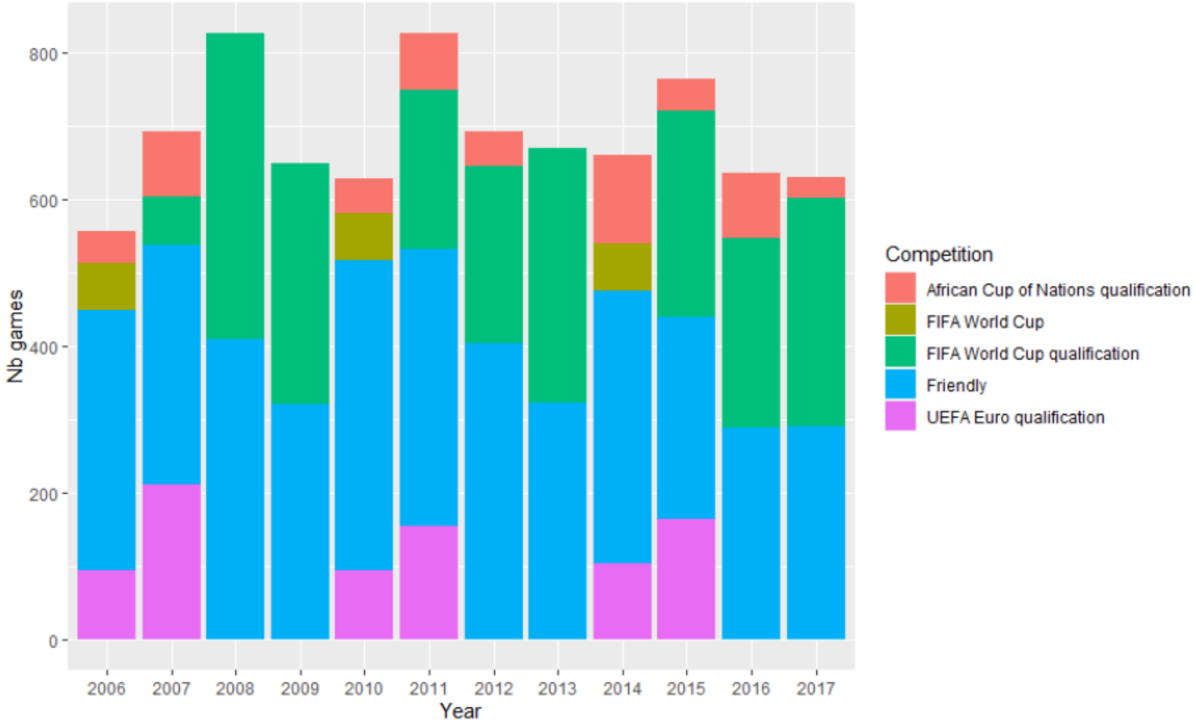
iii. A stack bar chart. We can see some events are more frequent on non-world cup year.



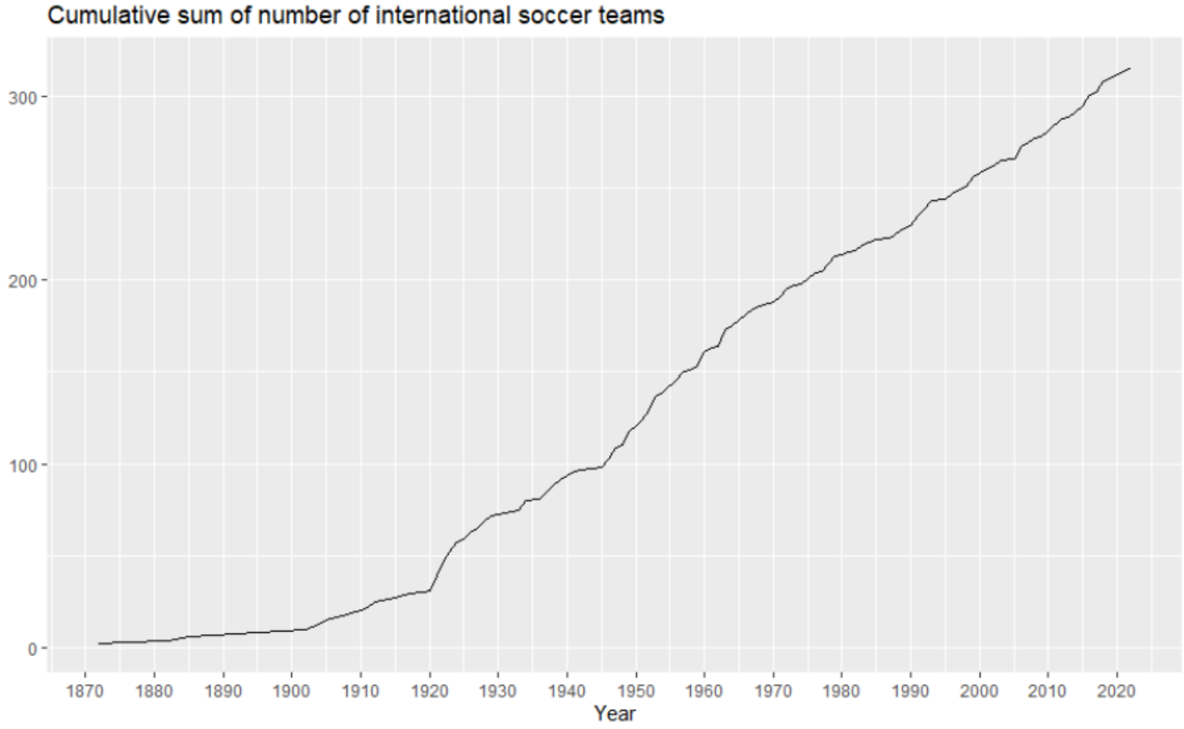
iv. World cup qualification generate more matches than world cup itself.



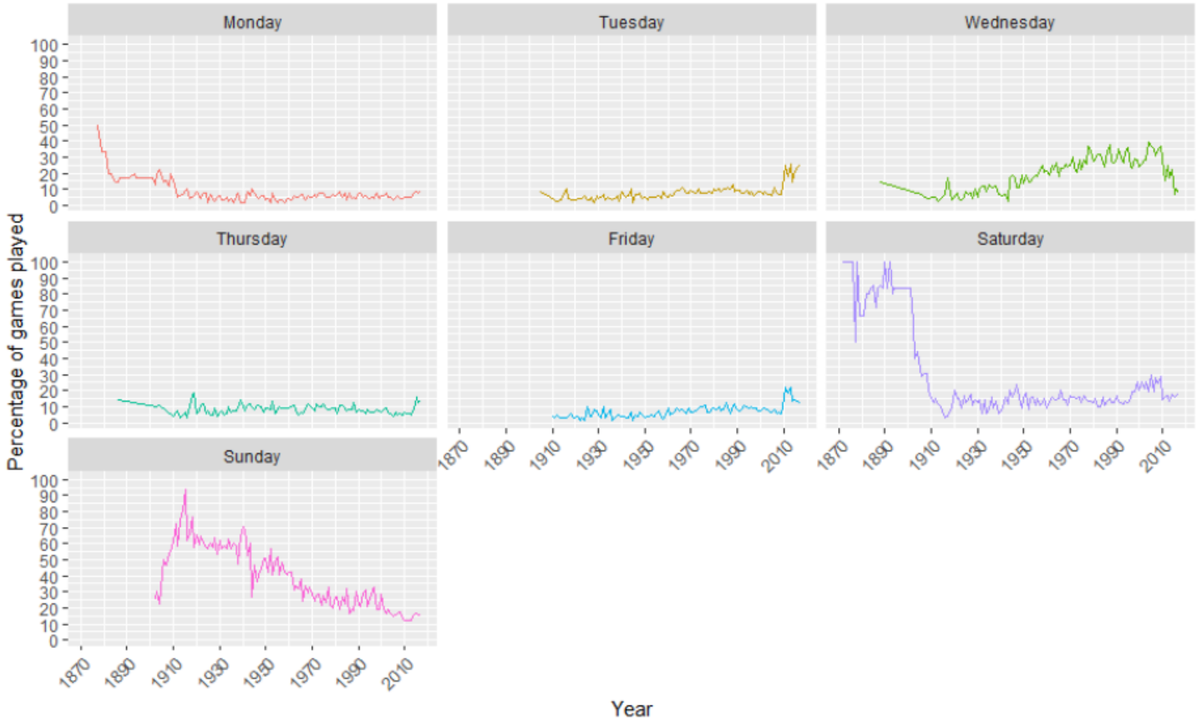
v. Non world cup events bar plot



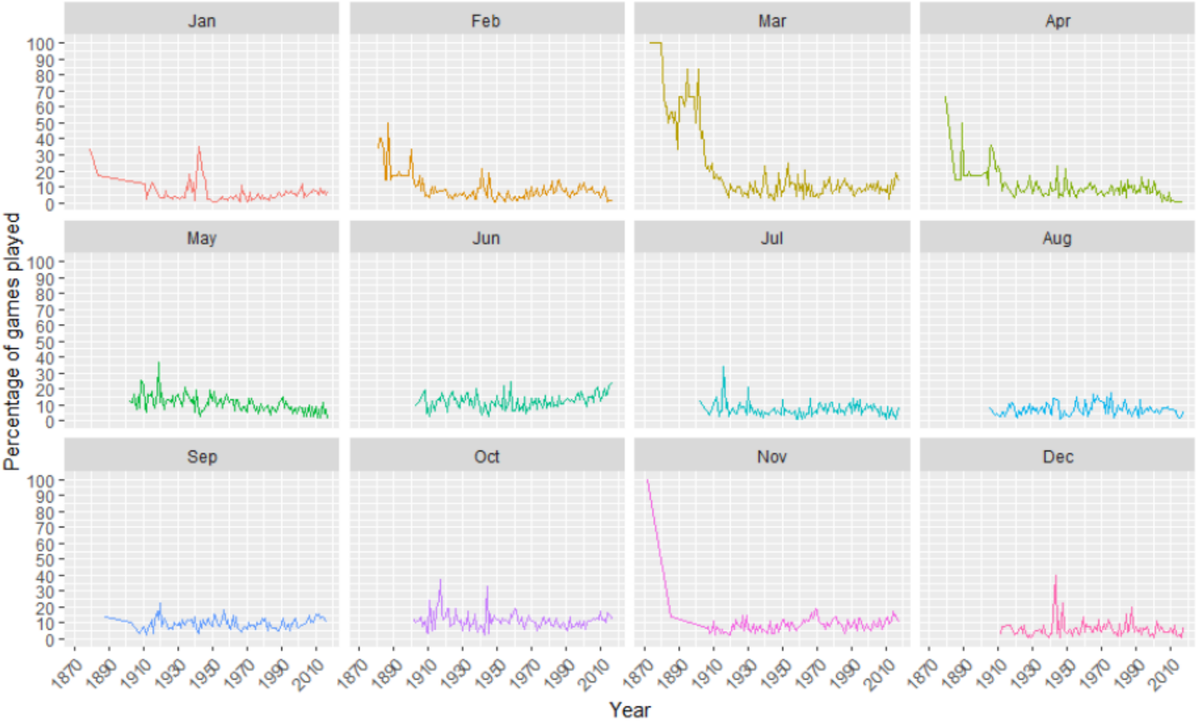
vi. Cumulative sum of no. of international soccer team



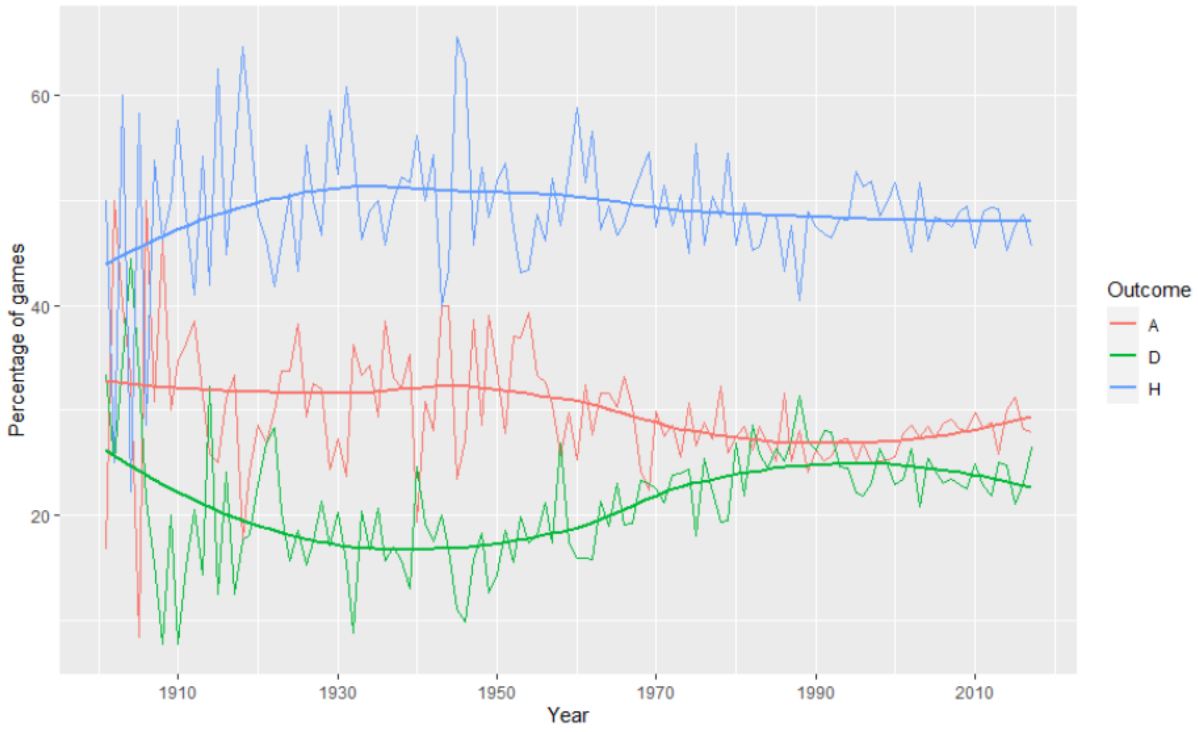
vii. Comparison between the days on the basis of number of soccer game played.



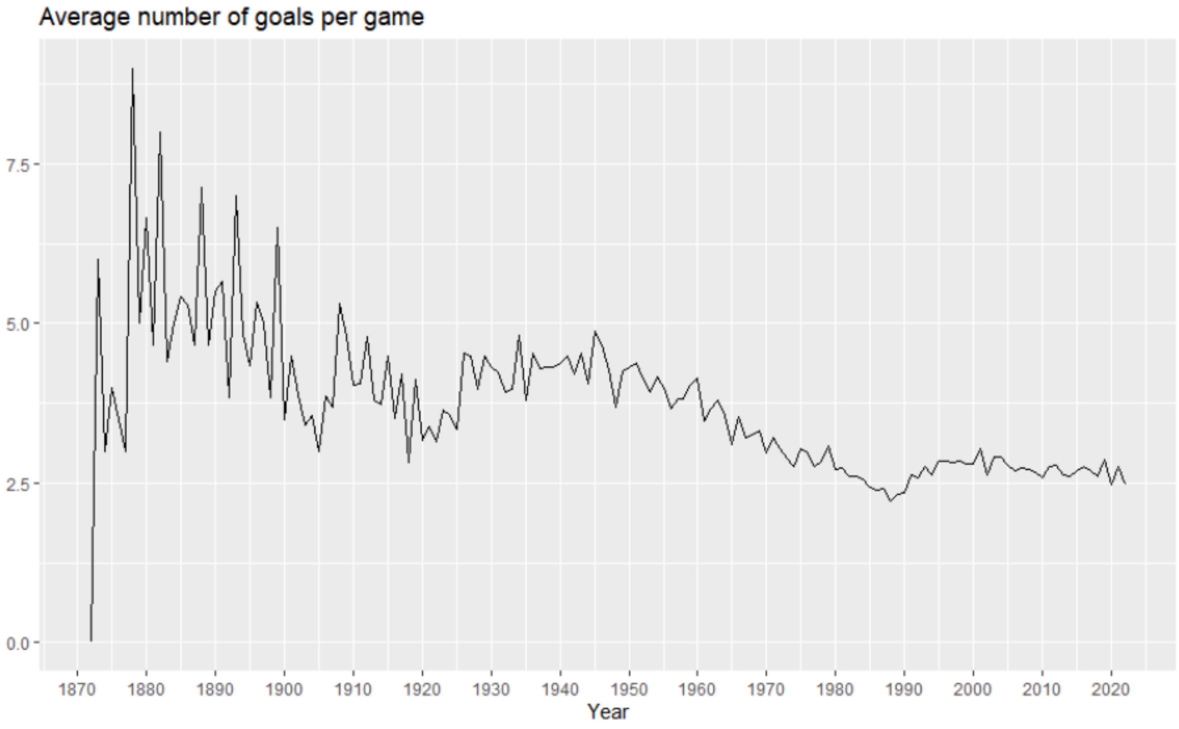
viii. Comparison between the months on the basis of number of soccer game played.



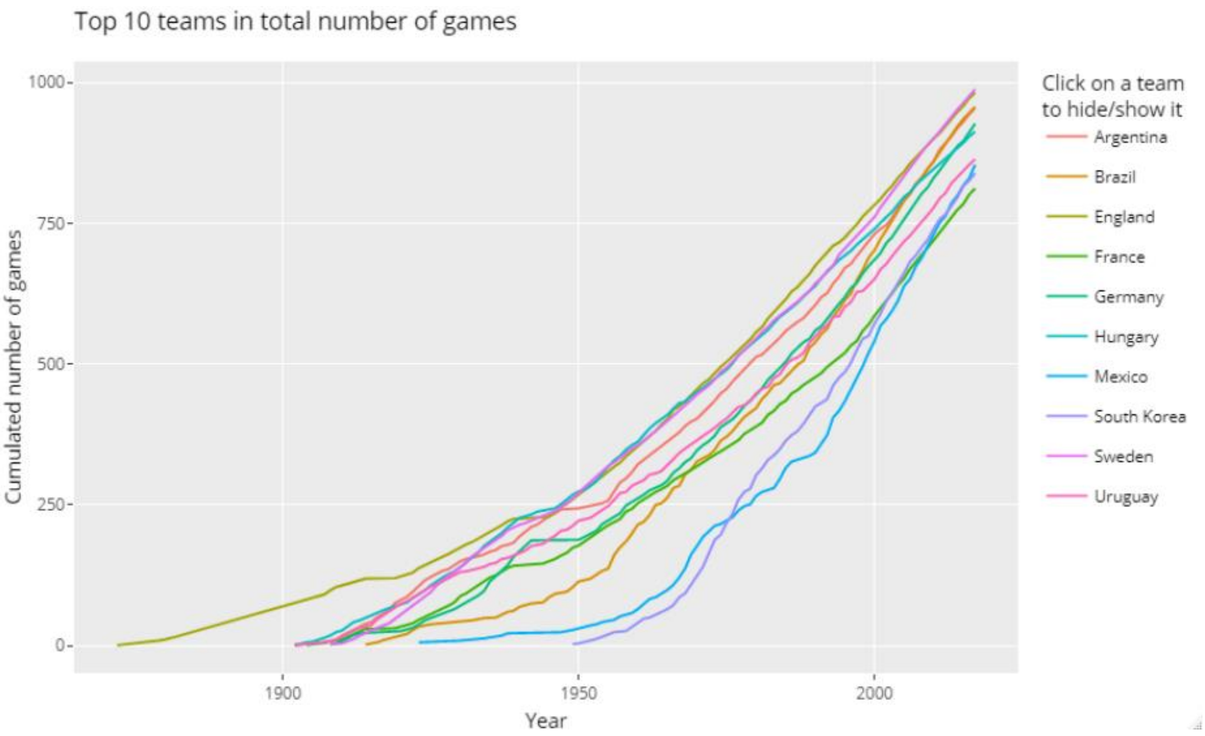
ix. The graph shows the percentage of games, i.e A, D, H



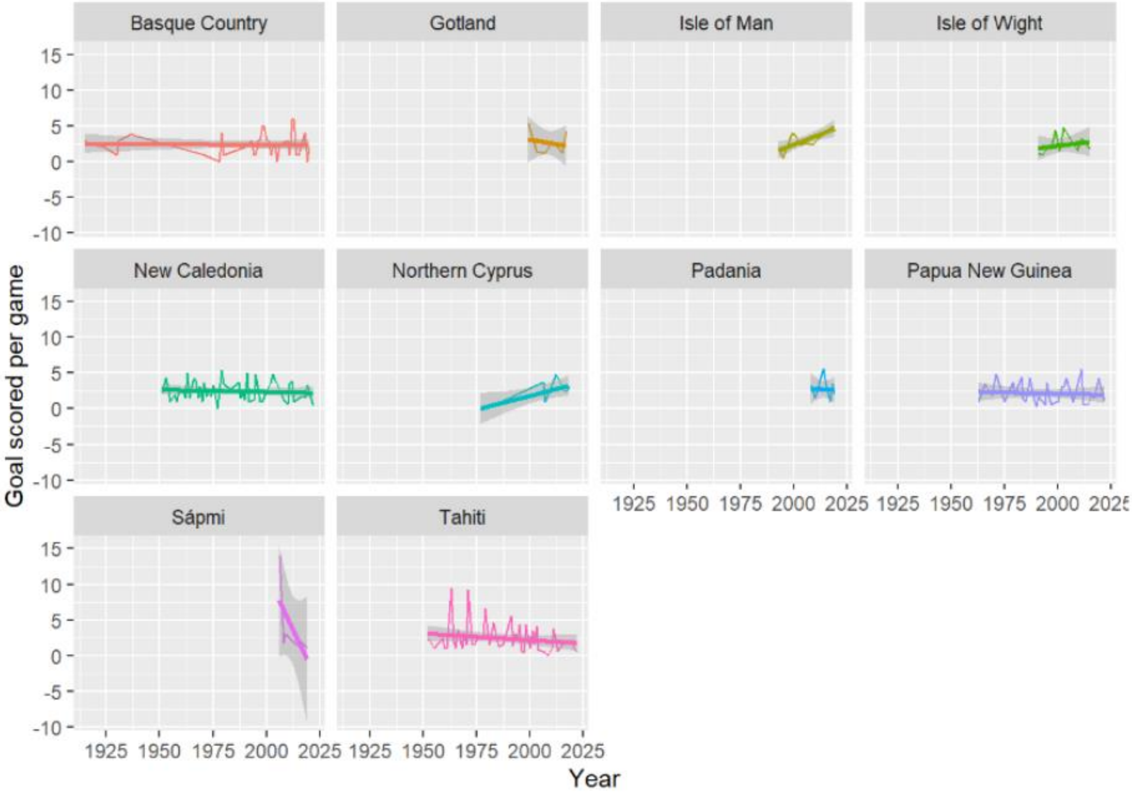
x. How the number of avg. goal per game have decreased over time.



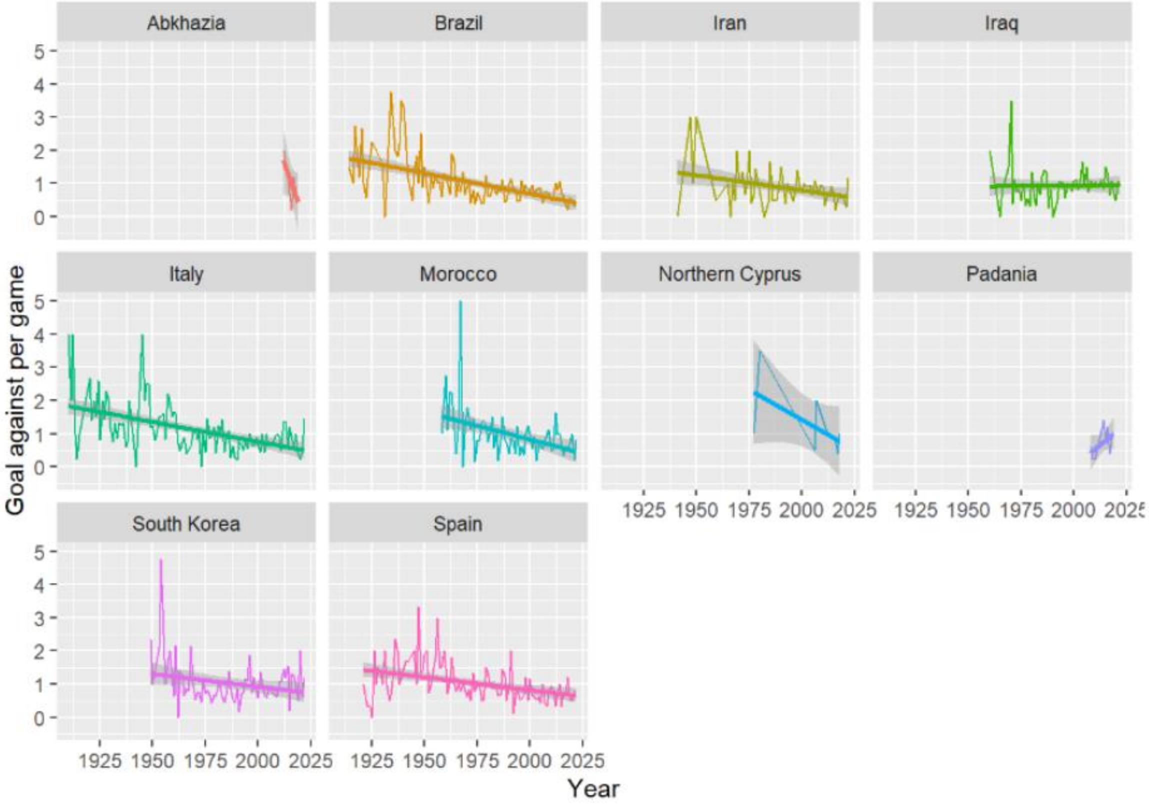
xi. This is an interactive graph, showing top 10 teams by wins



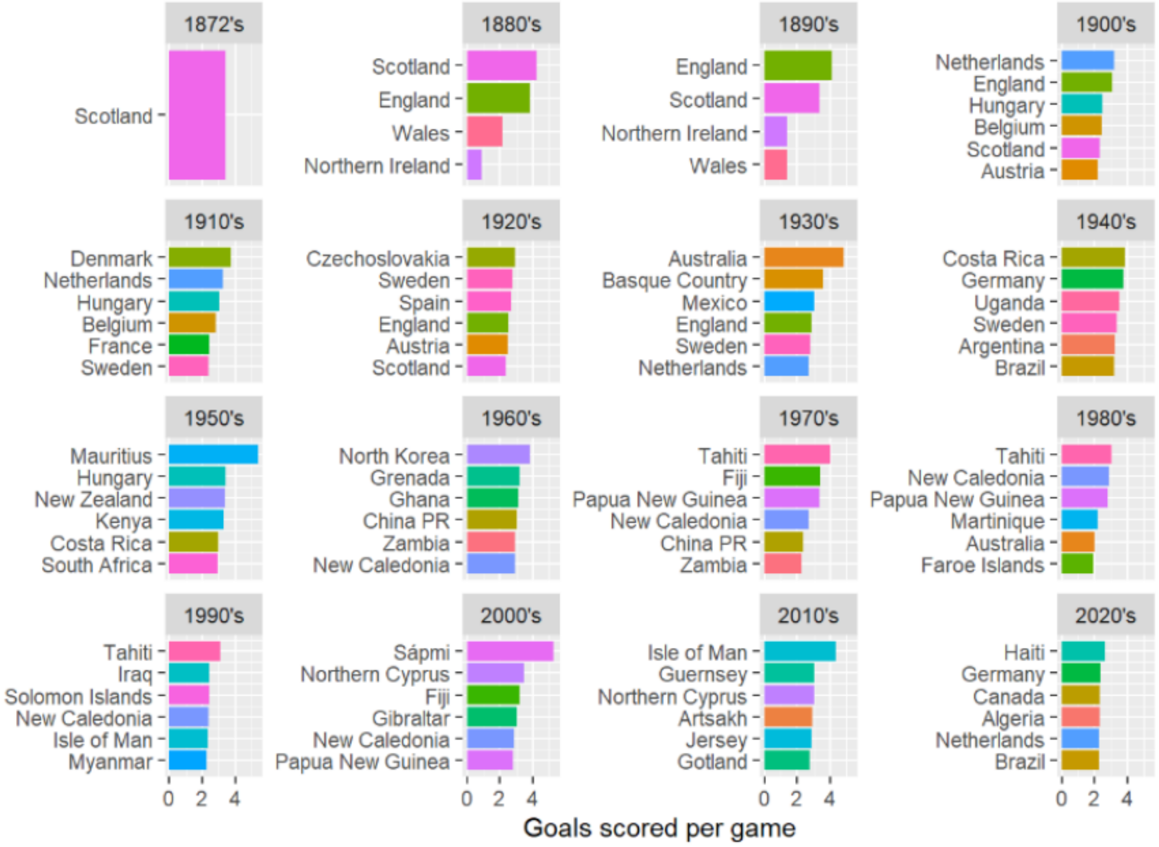
xii. Goals scored per team, prediction has also been shown



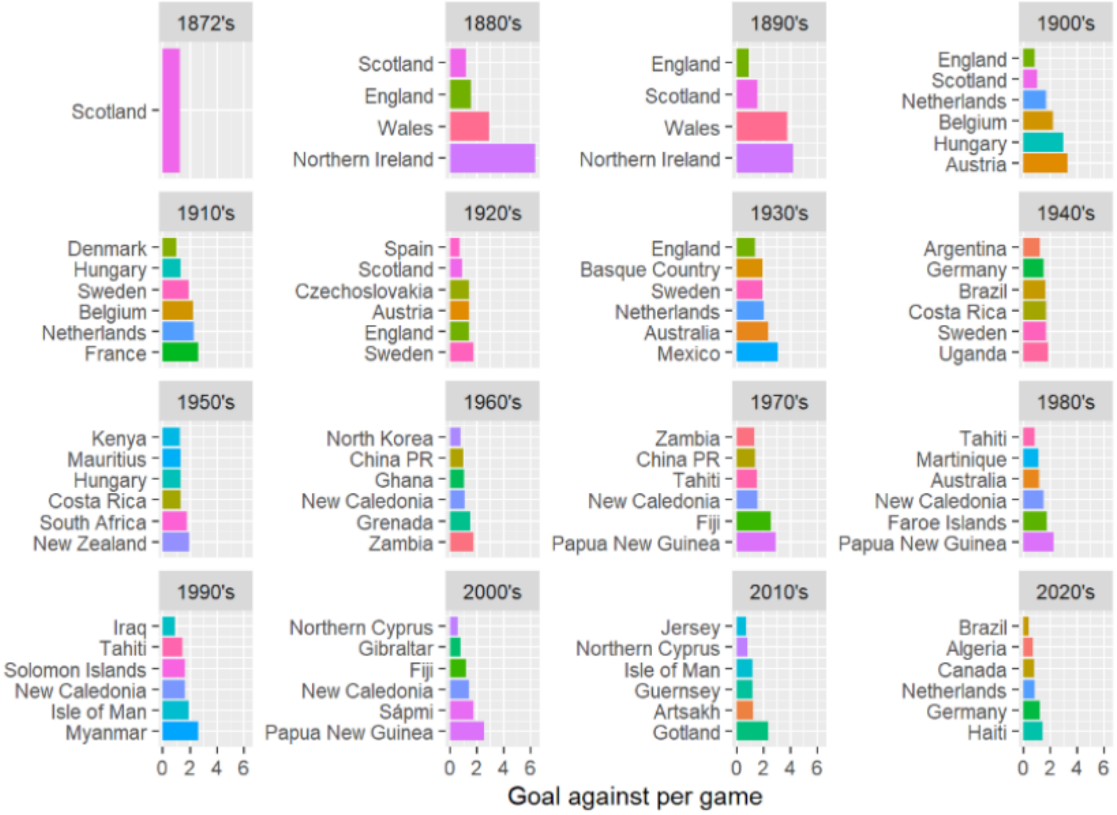
xiii. Number of graph against per game



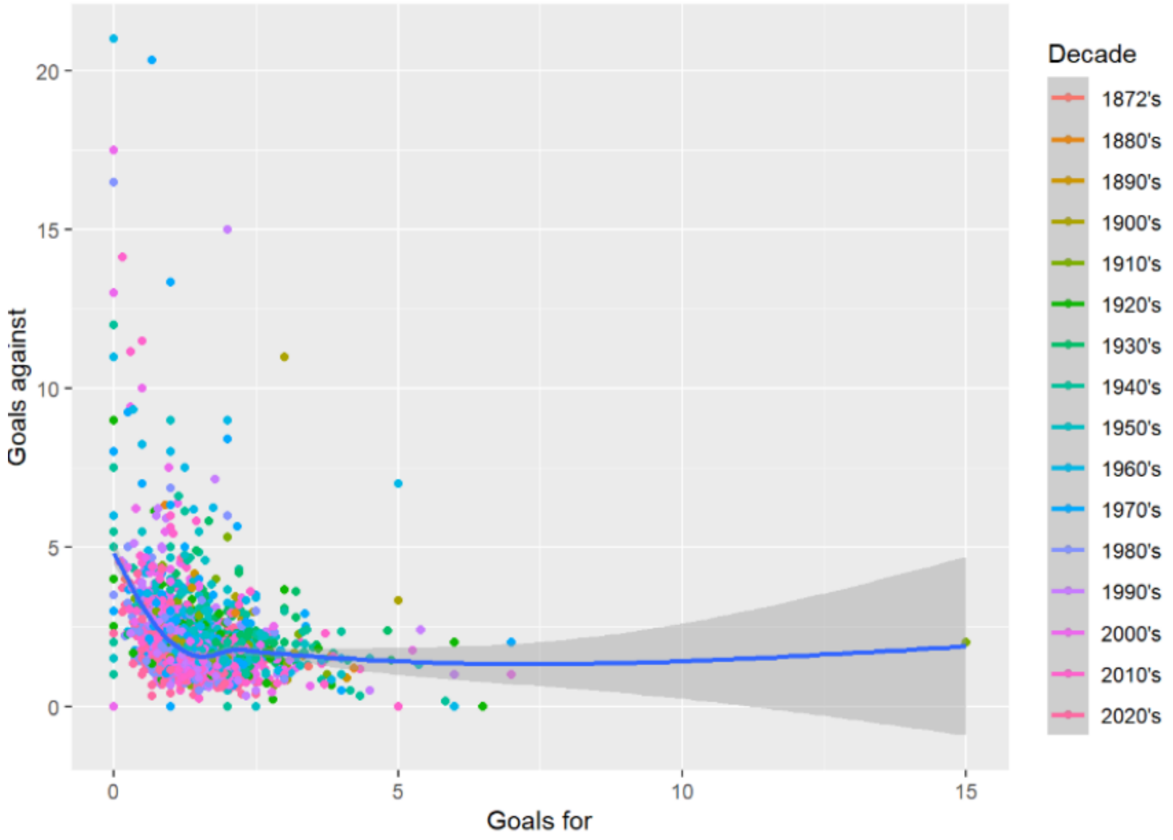
xiv. Number of goal scored per game, by decade



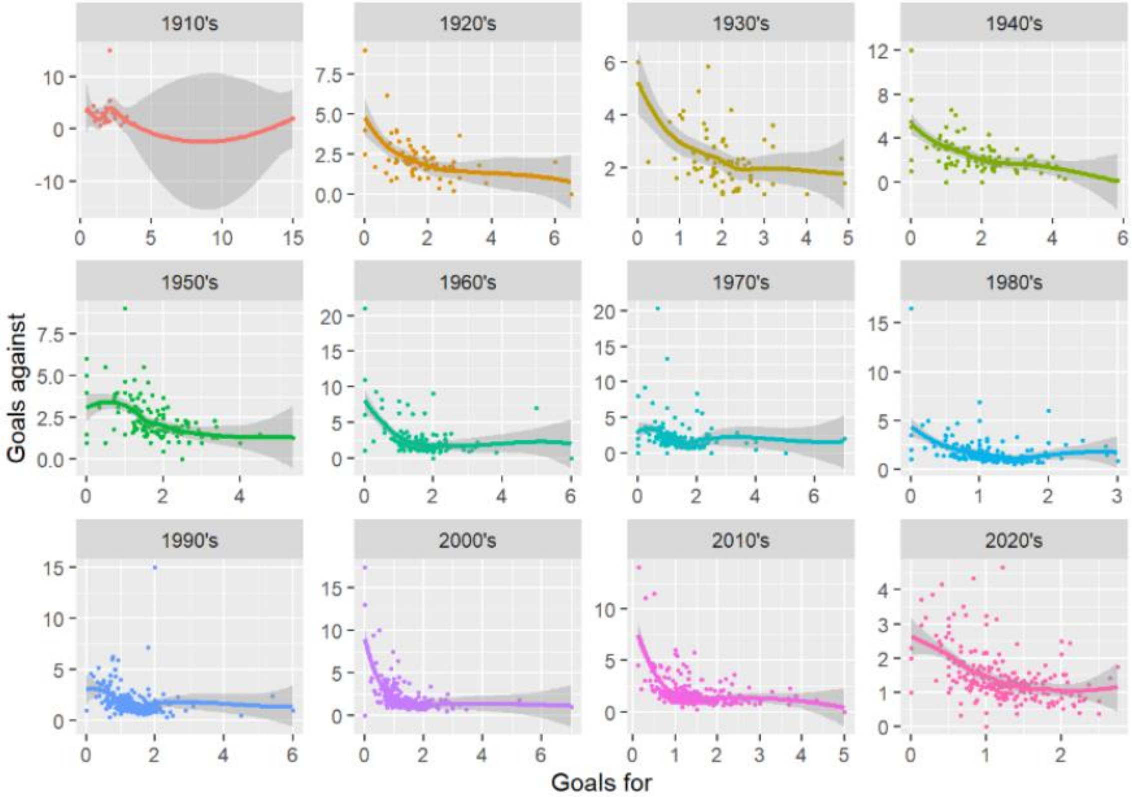
xv. Number of goal against per game, by decade



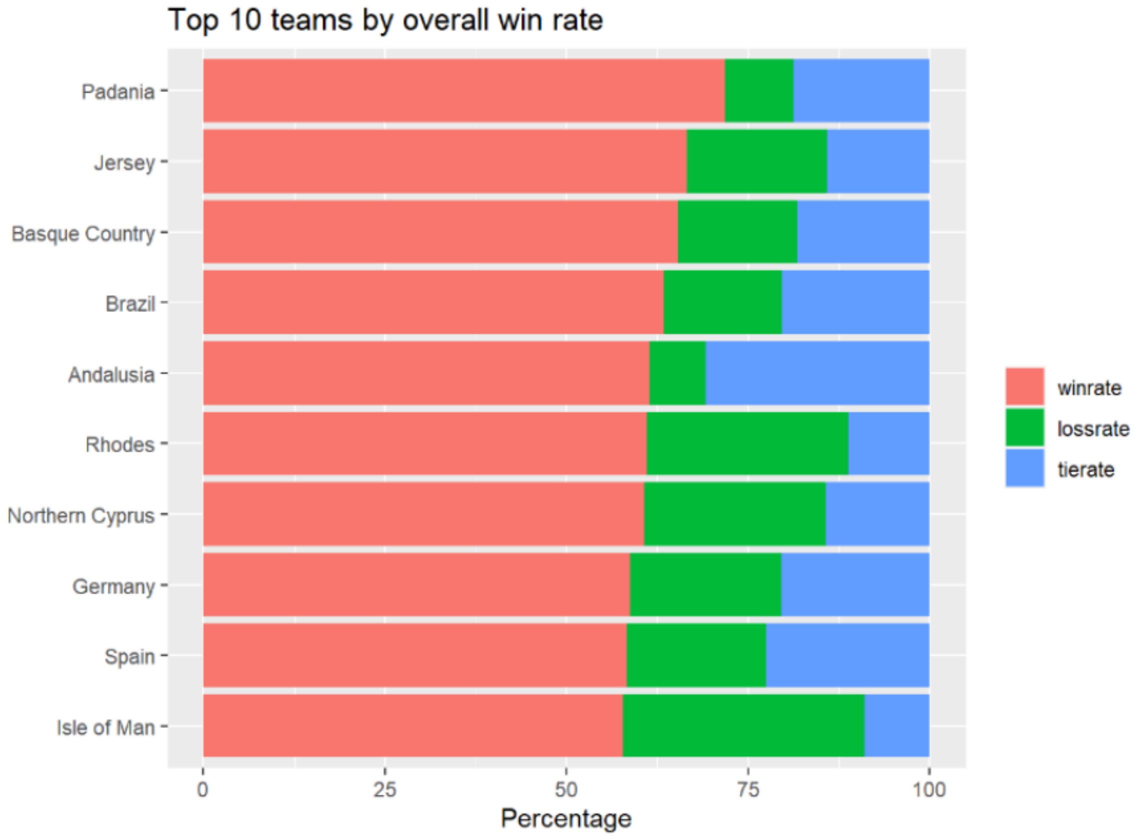
xvi. A beautiful visualisation of goals for vs goals against



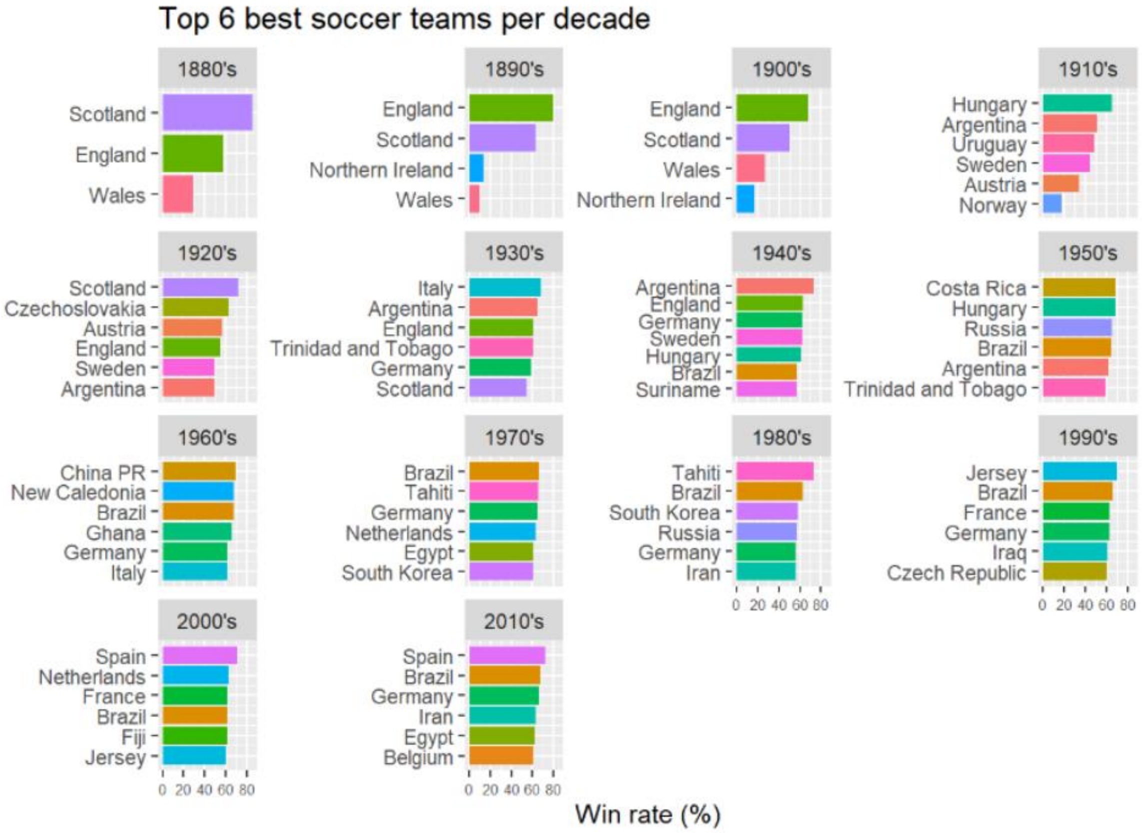
xvii. A delightful graph showing goals against vs goals for, by decade



xviii. Top 10 teams by overall win rate



xix. Top 6 soccer team, by decade



CONCLUSION

Data visualization is increasingly important in football due to the availability of data and the need to analyze it to make informed decisions. It helps coaches and analysts to track and analyze player and team performance, enabling them to make data-driven decisions that improve performance. Data visualization also engages football fans and provides them with new insights on their favorite teams and players. It can help teams and leagues to attract sponsors and improve their marketing efforts by demonstrating the value of their brand.

This project offers a comprehensive and interactive platform for football fans, researchers, and enthusiasts to explore the data of international football matches from 1872 to 2022. Through data visualization techniques, the platform presents historical data on match results, goals scored, team performance, and player performance in a user-friendly and engaging format. Users can filter and compare the data to gain insights into the evolution of international football, including trends, patterns, and playing styles over time. The project is a valuable resource for anyone interested in the sport and provides a unique opportunity to explore the rich history of international football.

GitHub Link: https://github.com/arnav-2606/Data-Visualisation

REFERENCES

Kitching, Gavin. (2015). The Origins of Football: History, Ideology and the Making of 'The People's Game'. History Workshop Journal. 79. 10.1093/hwj/dbu023.

Khimenes, Khrystyna & Briskin, Yuriy & Pityn, Maryan & Hluhov, I. & Drobot, K.. (2020). Monopoly and Rivalry in American Football in History and Nowadays. Ukraïns′kij žurnal medicini, bìologìï ta sportu. 5. 364-370. 10.26693/jmbs05.05.364.

Hognestad, Hans & Wergeland, Even. (2023). Nordic national football stadiums: past and present. Soccer & Society. 1-12. 10.1080/14660970.2023.2179196.

Lenartowicz, Michał & Karwacki, Adam. (2005). An overview of social conflicts in the history of Polish club football. European Journal for Sport and Society. 2. 97-107. 10.1080/16138171.2005.11687771.

Kitching, Gavin. (2015). The Origins of Football: History, Ideology and the Making of 'The People's Game'. History Workshop Journal. 79. 10.1093/hwj/dbu023.

Njororai Simiyu, Wycliffe. (2022). The Origins, Status, Contributions and Contradictions of Association Football in Uganda. 10.1007/978-3-030-94866-5_12.

Cin, Hülya & Özbek, Oğuz. (2022). The Bans Of Women's Football: The Real Struggles Are Besides The Pitch.

Allen, Stephen. (2023). Brenda Elsey and Joshua Nadel. Futbolera: A History of Women and Sports in Latin America .. The American Historical Review. 128. 555-556. 10.1093/ahr/rhad074.

Cookinham, Brittani & Swank, Chad. (2020). Concussion History and Career Status Influence Performance on Baseline Assessments in Elite Football Players. Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists. 35. 257-264. 10.1093/arclin/acz012.

Moradi, Reza & Amirseyfaddini, Mohammadreza & Amiri-Khorasani, Mohammadtaghi. (2022). Effect of Fatigue on Some Kinematic Characteristics During Gait, Balance, and Accuracy of Football Shots in High School Boys of Kahnooj Nomads With a History of Coronavirus. The Scientific Journal of Rehabilitation Medicine. 1352-1365. 10.32598/SJRM.10.6.3.