"Behind the Goals: A Data-Driven Approach to the FIFA World Cup"

Brian Beasley, Matteo Calviello, Caleb Mosteller, David Utsis

February 21, 2024

Abstract

We analyzed data on international football matches from 1872 to 2023 encompassing a total of roughly 74,518 matches. The dataset included various match types, ranging from friendly matches to FIFA World Cup matches. Using the dataset, we determined some of the best players like Cristiano Ronaldo, Lionel Messi, and Robert Lewandowski along with some of the best teams such as Brazil, Germany, and Argentina. We then dived into a deeper analysis of the FIFA World Cup, looking into attendance records and progress of teams over time, looking both at the greatest teams of all time, especially European and South American teams, and some newer and smaller teams, like Belgium, Morocco, and Japan.

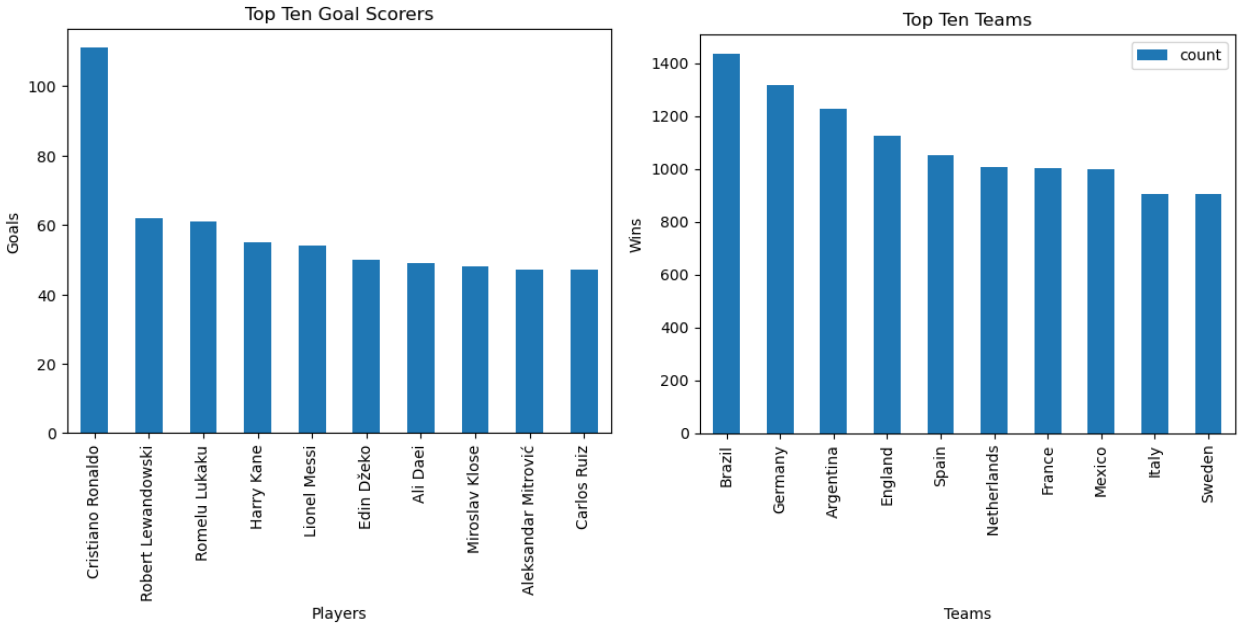
1 Introduction

Football has become the most watched sport in the world, creating a huge pool of fans and increasing the expectations for everyone's national team to perform well on an international stage. Recognized as the pinnacle of football, the FIFA World Cup showcases the culmination of skill, passion, and national pride. A data driven, statistical analysis of previous World Cups will help us draw conclusions about the top players, top teams, and diversity in the tournament. This information can allow us to make predictions about future World Cups and the future of international football. For this report, we will detail the top players, teams, diversity, and popularity of the tournament over the years, and take a look into which teams have historically improved or declined. Further, we will delve into one of the biggest rivalries of the World Cup, European teams vs. South American teams. Finally, we will also study some of the more recent World Cups to try and draw conclusions about recent team performance and what that means for future World Cups.

2 Data Preparation

The initial dataset [1] from Kaggle, contained three separate datasets all relating to International Football Matches. This dataset included 74,518 results of international football matches starting from the very first official match in 1872 up to 2023. The matches were strictly men's full internationals and the data did not include Olympic Games or matches where at least one of the teams was the nation's B-team, Under-23, or a league select team. The first file contained the results of each match, highlighting what team won if they were the home or away team, and what the score was. The second file, 'shootouts.csv', contained data on penalty shootouts highlighting the specific results and teams involved in these decisive moments. Lastly, the 'goalscorers.csv' file included information on the goal scorers for each match. This was easily the biggest file, even though a majority of the games ended 0-0. the first step in the data preparation was to merge these three files. To merge these datasets, Python's pandas library was utilized. Both the 'results.csv' and 'shootouts.csv' contained columns for date, home_team and away_team. A left join was used to merge these two files in order to retain all records. Next, the resulting data frame was merged with the 'goalscorers.csv' file. this file also had the same common columns so another left join was used. Performing the merges was easy; however, there were some slight problems: the older matches from 1872, had significantly less recorded data than the newer and most recent matches, this meant that a lot of the data was null and it made harder to merge some of the attributes of the dataset.

Utilizing the merged dataset along with the Average and total attendance at FIFA football World Cup games from 1930 to 2018[2] from Statista we were able to create a merged dataset combining World Cup-specific statistics with their attendance data. Further adding in attendance information for the Qatar World Cup from the Qatar 2022 FIFA World Cup Attendance dataset[3] from Kaggle ensured that our merged dataset would contain information about all of the World Cups. In order to merge the datasets, the attendance dataset required dropping an unnamed row coupled with separating the year and location and renaming "USA" to "United States" to facilitate the merge.



(a) Top Scorers

(b) Top Teams

Figure 1: Interesting Stats

3 "Football" Statistics

While analyzing the international matches for the World Cup Games, we looked into some entertaining information like top players and top teams; the results that we were able to get from the dataset are shown in Figure 1.

We discovered that the top player, based on goals scored, is by far Cristiano Ronaldo followed by some other great names like Lewandoski, Kane, and Messi. These are major players that have influenced the world of soccer in the past decades and all played their last World Cup in Argentina in 2022. The best team, still based on goals scored, is Brazil. This is expected as football is a big part of Brazil's culture and they have had a lot of incredible players that have revolutionized the history and technique of football, like Pelé, Kaká, Neymar, Ronaldinho, Ronaldo, and Carlos Alberto. An interesting detail is that while Brazil has won over 1400 matches in the past 200 years, the number of ties is almost 10 times as much: there have been 14,000 ties in international matches. This is an extremely high number but is to be expected as many matches tend to end up with a very close score leading sometimes to penalty shootouts. Penalty shootouts occur on rare occasions, only during the final stages of major tournaments. 33 countries have never gone into penalty shootouts and the team that has won the most penalty shootouts is Argentina. Their last penalty shootout was the famous victory of the FIFA World Cup where they beat France in 2022.

Continuing to explore this immense dataset, we started to look into the diversity of team appearances in the World Cup. A graph exploring this topic is seen in Figure 2. This chart contains the appearance information for the top 10 teams in terms of appearances. Unsurprisingly some of the better teams like Brazil, Argentina, and Germany have the most appearances. Analyzing the data from 1930 to 2022, a total of 82 countries out of 195 total countries made an appearance in the World Cup. This number, encompassing a pure 42 % of countries is quite a small number proving the sheer difficulty of making the World Cup.

4 Results

In the past years there have been a lot of upsets and a lot of 'underdog' teams have made it past the group stages, eliminating some of the favorites. We decided to analyze team performance over the years to see if only the major teams that have a long history in football have been qualifying and making it to the end of the World Cup, or if newer and smaller teams have been exceeding expectations. We based our criteria for performance on two categories: Net Games (games won - games lost) and number of goals scored. For each of these criteria, we found the most improved team and the least improved team. The graphs for these teams can be seen in Figure 3.

To create the graphs we had to find every country that had ever played a home or away team in the World Cup and calculated the net number of games (games won minus games lost) for each team as well

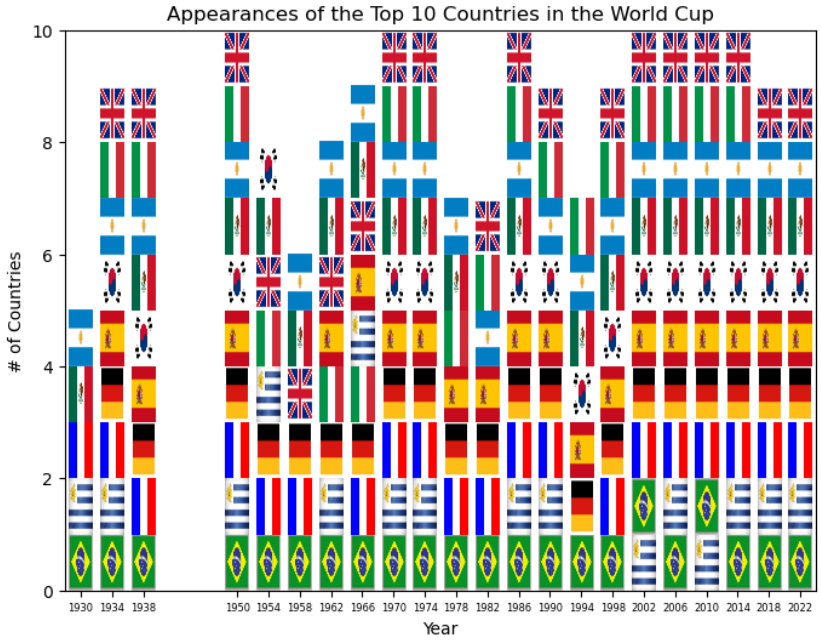
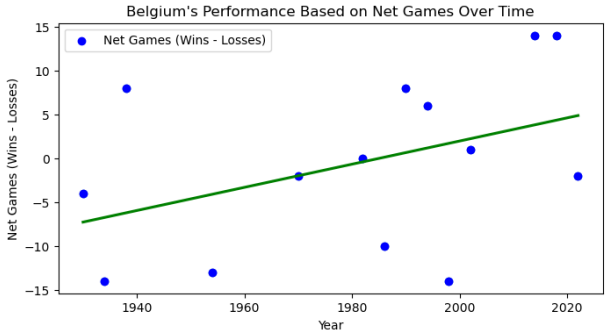
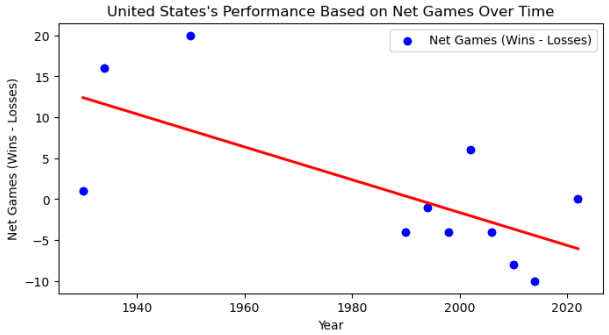


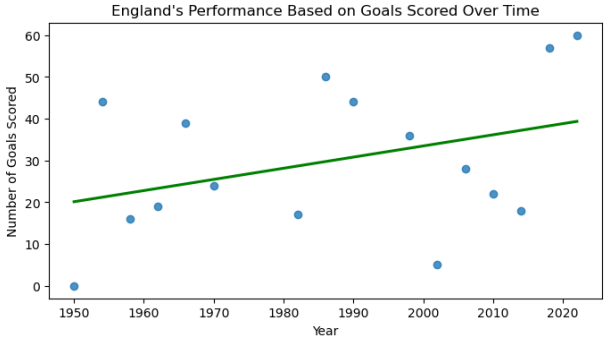
Figure 2: Team Appearances in the World Cup



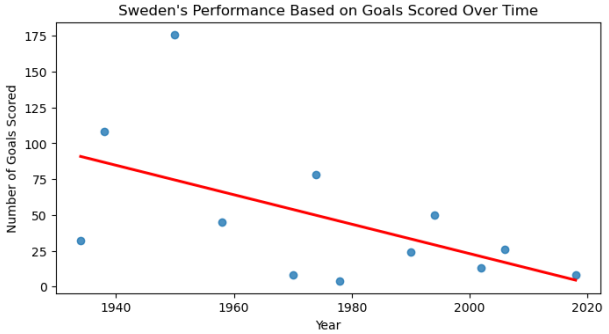
(a) Best Performance Games (Belgium)



(b) Worst Performance Games (USA)



(c) Best Performance Goals (England)



(d) Worst Performance Goals (Sweden)

Figure 3: Performance Comparison

as the number of goals scored. To ensure accurate analysis, we filtered out teams with insufficient data, considering only those with a minimum number of 10 matches played. We identified the most improved and least improved teams over the years and plotted the performance trends of these teams over time, utilizing regression analysis to visualize their performance trajectories.

The two graphs for performance on games show Belgium as the most improved team and the United States as the worst improved team. Belgium is a very small country and, while it has produced incredibly talented players like Romelu Lukaku and Kevin De Bruyne, it has always been very difficult for it to exceed in any international competitions due to the fact that there is a small population. Belgium has, however, had tremendous success as it is the only national team in the world to top the FIFA ranking without having won a World Cup or a continental trophy. On the other hand, the United States is ranked as the worst-improved team. This is mainly due to the fact that the United States has only had 11 appearances in the World Cup: between 1950 and 1990 the United States did not participate in any of the 10 editions. This has a huge toll on the performance graph as in the most recent years it has had a lot more successes.

The two graphs that analyze performance on goals scored highlight England as the most improved and Sweden as the least improved. England is a country with a lot of history in football and it has done extremely well in the past years too. Getting to the final matches in the World Cup doesn't always mean that you have scored a lot of goals, but England exceeded expectations by scoring over 100 goals in the last two editions of the tournament. The worst-performing team, based on this criteria, is Sweden. Sweden is a small country with a small team known for its extreme defensive style in play. Sweden has produced many talents, like Zlatan Ibrahimovic, but the trend line highlights that the team is having difficulty scoring goals and, furthermore, making it past the qualifying stages of the World Cup as they have qualified only 3 times in the last 7 editions.

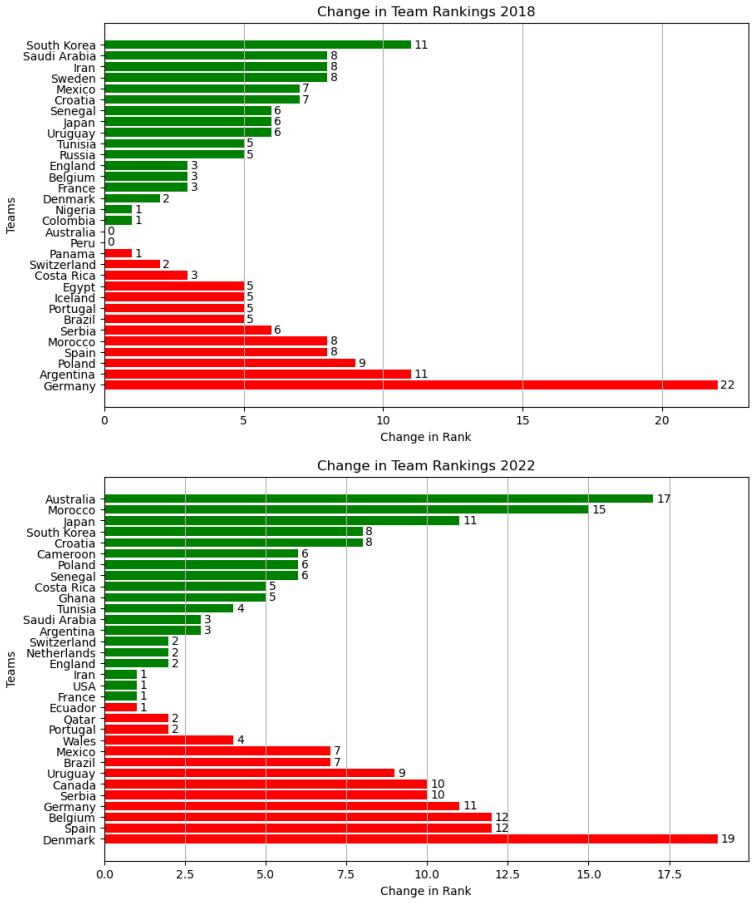


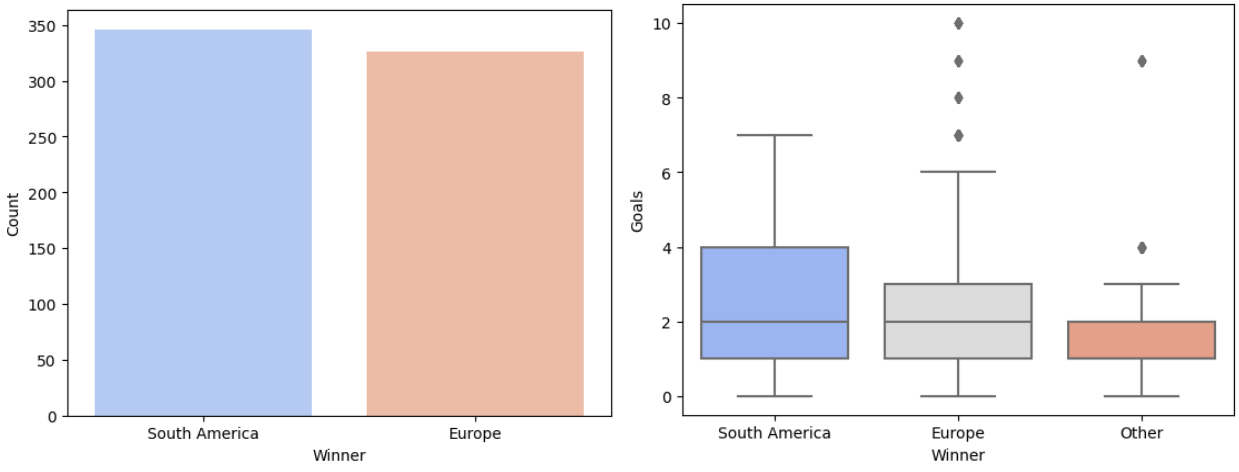
Figure 4: Team Performance in 2018 and 2022

While these graphs serve as valuable tools for visualization, their accuracy can be compromised when considering all past World Cups. This approach may overlook significant fluctuations in performance over time, as evidenced by the United States. We decided, therefore, to use another way to calculate performance over the most recent years. We decided to focus on the recent 2018 and 2022 World Cups to see which teams have been over-performing and which teams have been under-performing. FiveThirtyEight [4] provides data sets that have predictions for each team in the World Cup before the games begin. Predictions include

statistics about the teams, like goals for and goals against, as well as each team's probability of making each round of the tournament. Using the probabilities from the predictions, we ranked the teams from worst to best based on where they were predicted to place. In order to compare these predictions to results, we also needed to rank the teams based on their finish in the tournament. By analyzing all the matches in the original dataset for both the 2018 and 2022 World Cup we were able to find where they placed in the tournament. However, there are many teams that are knocked out in the group stage and the round of 16, so we analyzed all the games each team played and found their total goals scored and total goals against. We used this to calculate each team's goal differential for the tournament as a whole and that is what we used to rank teams who got knocked out in the same stage. If there were any more discrepancies, we used the original prediction data to rank the teams. For example, if two teams got knocked out in the group stage and both had an equal goal differential, then the team that was predicted to do worse before the tournament began would end up ranking higher than the team that was predicted to do better. This allowed us to get rankings for each team in the 2018 and 2022 World Cup before each tournament and after each tournament. We were then able to create Figure: 4 for the 2018 and 2022 World Cups. These graphs show the change in team rankings after the World Cup conclusion based on each team's predicted outcome. This graph is meant to highlight teams that are vastly under-performing or over-performing. For example, Japan, Saudi Arabia, Croatia, and South Korea are teams that have over-performed the past two World Cups. These teams are trending upwards and we would expect to see them in future World Cup tournaments and possibly become dominant teams. On the other side, teams like Germany, Spain, Serbia, and Portugal are trending downward as they have under-performed in the previous two World Cups. Germany won the 2014 World Cup and Portugal and Spain have been powerful contenders in previous years. However, this may indicate that those teams are going through a rebuilding process and may not even make the World Cup in the upcoming tournaments. These graphs and this data provide an interesting insight into the current state of the World Cup tournaments and perhaps allow us to make predictions about future World Cup tournaments and what they may look like.

5 Europe vs South America

Throughout history, Europe and South America have engaged in fierce competition, often vying for supremacy in major tournaments. Notably, the rivalry between the Netherlands and Argentina has captured attention, evidenced by their intense encounter in the Quarter Final stage of the most recent World Cup: the match was characterized by its physicality, resulting in a record-breaking 18 yellow cards. Beyond individual rivalries and matches, our examination extends to evaluating the overall performance of these continents by combining all the performances of each country within them. The graphs from this analysis can be seen in Figure 5. The analysis of the continents' performances in head-to-head match-ups during the World Cup highlights a near deadlock: South America stands at 484 wins to Europe's 427, a mere 60-game difference across 900+ matches. This disparity underscores the remarkable balance and prowess of both teams. Delving deeper into the analysis, we examined goal statistics, as depicted in Figure 5b. It is evident from the box plot that both continents boast an average of around 2 goals per game, yet South America has a higher tendency to score more goals compared to Europe.



(a) Win Record

(b) Goal Statistics

Figure 5: Europe vs South America

We decided to also show the goals statistics for the rest of the World (under 'Other' in the graph) to highlight the discrepancy and how dominant South America and Europe are over the rest. While South America maintains a winning record against Europe, their superior goal-scoring rate per match may also stem from distinct playing styles: Europe, exemplified by teams like the Netherlands, Belgium, and Germany, often adopts a defensive approach, capitalizing on counterattacks; in contrast, South American teams are renowned for their offensive prowess, using tactics like the 'tiki taka': rapid, one-touch passes to unsettle opposing defenses. Overall, it can be observed that the two continents have historically dominated and continue to dominate the world of football, consistently producing thrilling matches that captivate and inspire spectators.

6 Conclusion

The world of international football is very dynamic which comes with massive popularity throughout the world. Many teams such as Japan, Saudi Arabia, Croatia, etc, are on the uprising while other teams that have performed well in the past such as Germany, Spain, Serbia, and Portugal, are now facing challenges. Through the performance graphs, we can see a shifting dynamic of success from once teams considered powerhouses to new contenders. This shift of power dynamic leads to a very diverse selection of teams within the World Cup. It also reveals that the world of international football is very complex and it constantly evolving. As football continues to evolve, more exploration should be done to gain a better understanding of what makes football so popular.

References

[1] M. JÜRISOO, *International football results from 1872 to 2023*, Kaggle, 2023. DOI: XX.XXXX/XXX.XXXX.XXXXXXXX.

[2] S. R. Department, *Average and total attendance at fifa football world cup games from 1930 to 2018*, Statista, 2022. DOI: XX.XXXX/XXX.XXXX.XXXXXXXX.

[3] M. PARASHAR, *Qatar 2022 fifa world cup attendance*, Kaggle, 2023. DOI: XX.XXXX/XXX.XXXX.XXXXXXXX.

[4] N. S. JAY BOICE, *World cup predictions*, FiveThirtyEight, 2024. DOI: XX.XXXX/XXX.XXXX.XXXXXXXX.