DATA 5310: Data Visualization
Seattle University
Fall 2023

Exploring Trends in Men's Soccer: A Comprehensive Analysis of FIFA Matches from 1916-2023

McKenzie Maidl, Emma Oriol, Samikshya Pandey

Introduction

Football is one of the most watched sports in the world, with over 3.5 billion fans (World Atlas) tuning in to watch football every year. A popular sport around the world, it regularly features in the top two most watched sports in almost all countries, except in US, Canada and Philippines (Horberry). The global cultural, historical, and economic impacts of football are well known. Just the big five European Leagues generated 18.5 billion US dollars in revenue in 2021/2022 season and an estimated 450 million viewers had watched the Champion League final in 2023 (Dunbar).The Champions League is an annual club football competition contested by top-division European clubs. Football brings people together, and encourages international cooperation and understanding.

While there are several leagues and competitions that happen around the world, FIFA (Fédération internationale de football association) hosts one of the most prestigious and popular tournaments: Men's World Cup. The FIFA Men's World Cup has been held every 4 years since 1930 except in 1942 and 1946 because of the Second World War. Its popularity remains intact after all these years: the World Cup Final between Argentina and France, held in Qatar 2022, recorded 1.5 billion global viewers; the Super Bowl in comparison had close to 157 million viewership. (Shea). FIFA, along with the World Cup, also hosts other international matches as well.

With such a rich history, FIFA and its matches also hold enormous and interesting data about countries that have participated, goals scores throughout the years, geographic location of match, revenue and even viewership data. This paper aims to explore some of this data and investigates patterns and trends found in these soccer matches. This analysis investigates goals from FIFA matches spanning over 100 years from 1916 through 2023. The aim of this report is to understand how the sport has evolved over the years, identify standout teams and players, and understand what patterns and influences affect goal scoring.

Methodology

The dataset used in this analysis was sourced from Kaggle (Juriso). It includes 45,000 records from international football matches, ranging from FIFA World Cup to regular friendly matches, but excluding those from the Olympics or where at least one of the teams was a nation's B team.

The following fields were included in the original data:

- Date of the match
- The name of the home team
- The name of the away team
- The name of the team scoring the goal
- The name of the player scoring the goal
- The minute at which the goal took place
- Whether the goal was an own-goal
- Whether the goal was a penalty

Initial data cleaning was conducted to rename columns, update country names when required, and reformat the data types of some columns. The following additional fields were then calculated:

- Total points for each team per match
- The winner of teach match
- The pre-penalty points for each match
- Total goals scored in each match
- Total penalty scored in each match

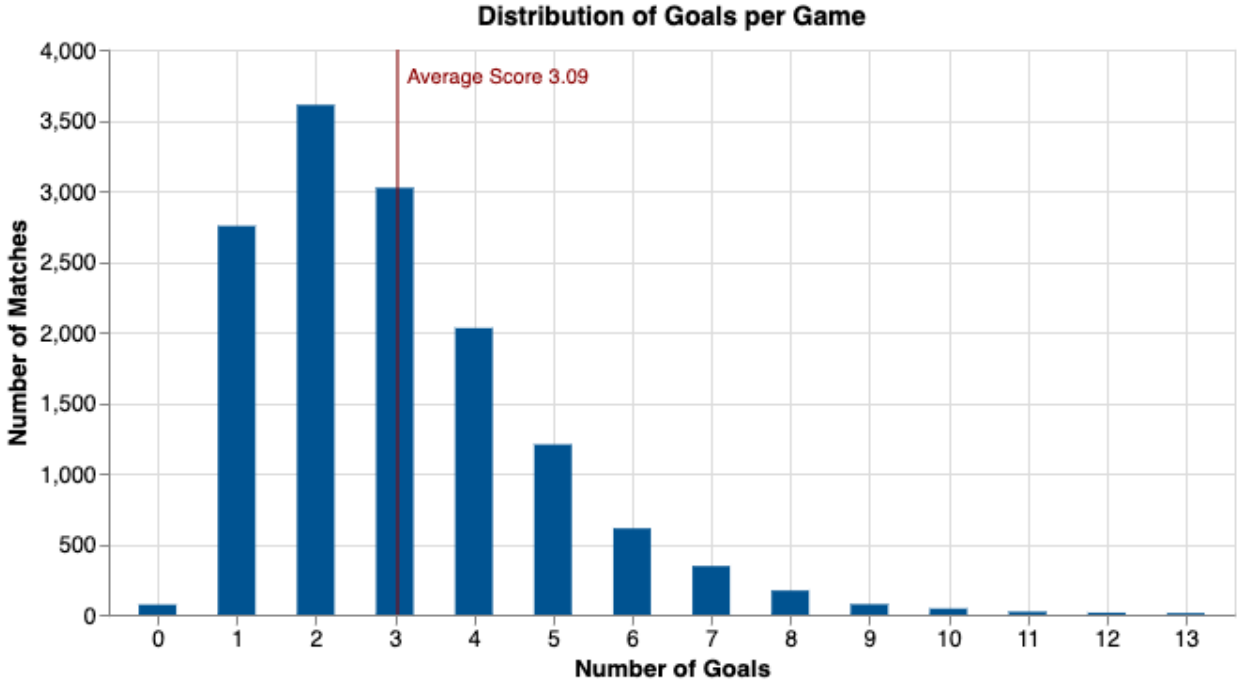
In addition to the initial data cleaning and processing, the data was further processed for each visualization to group rows by, for example, year or team, or to calculate additional fields as needed. Secondary data sources were also used in some cases, for example geographic country data for the map in Plot 5 below. Additional information about changes in football rules were added (Plots 3 and 4) to explore the general trends in goal scoring over the years.

Results

The plots created for this analysis cover three distinct categories. In the first section, each represents an aspect of the game. The second group focuses on team trends, and the third is about players.

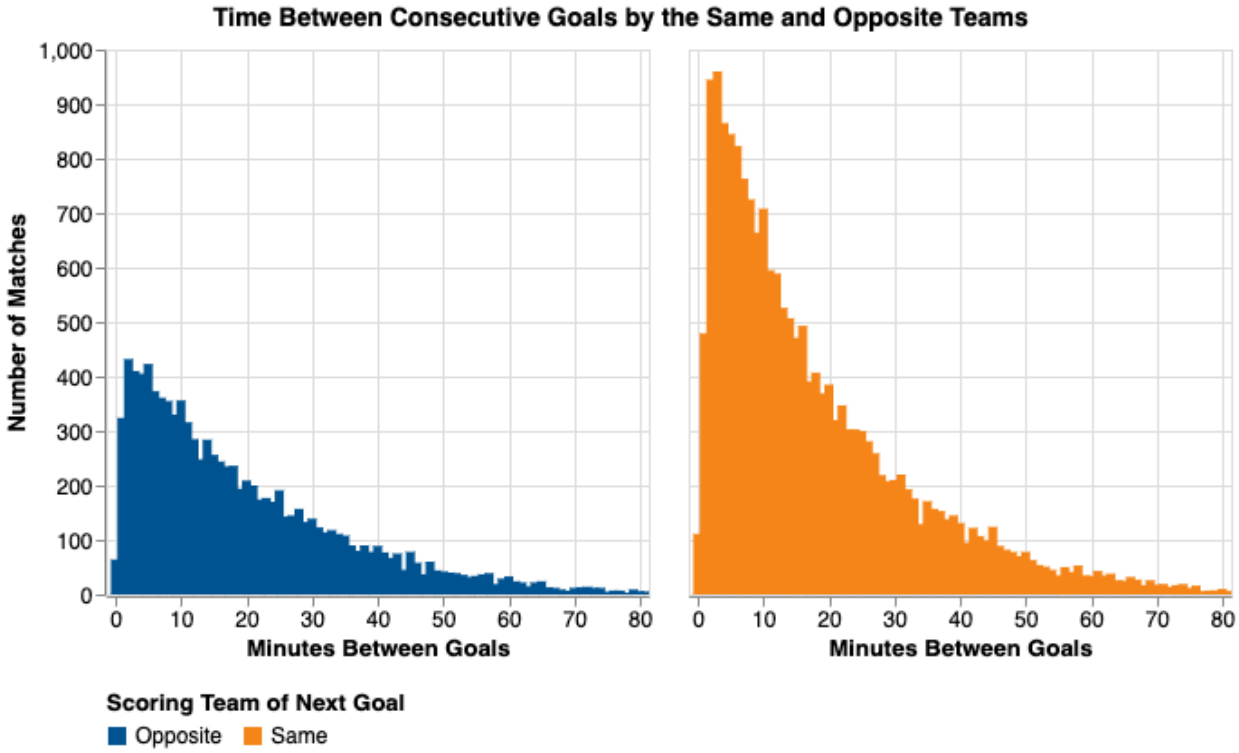
Game Plots

1: Average goal per game



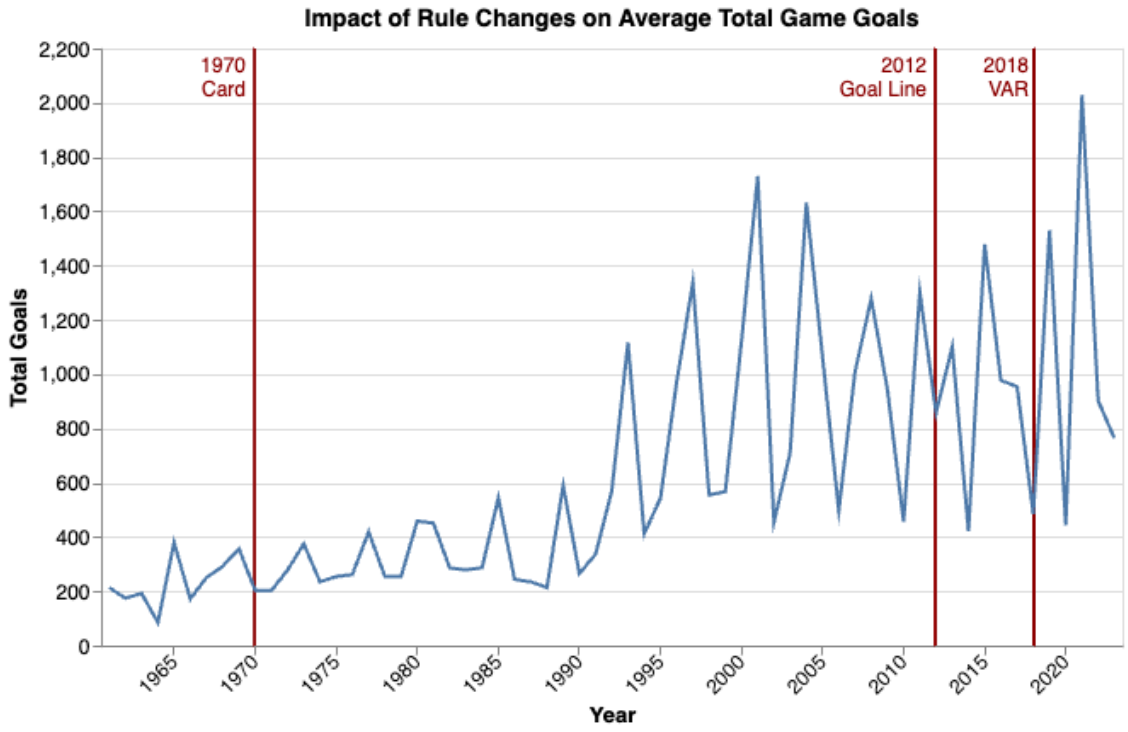
Plot 1 shows the distribution of total goals scored by both teams in a match. The red line shows the average number of goals scored.

2: Time between two consecutive goals



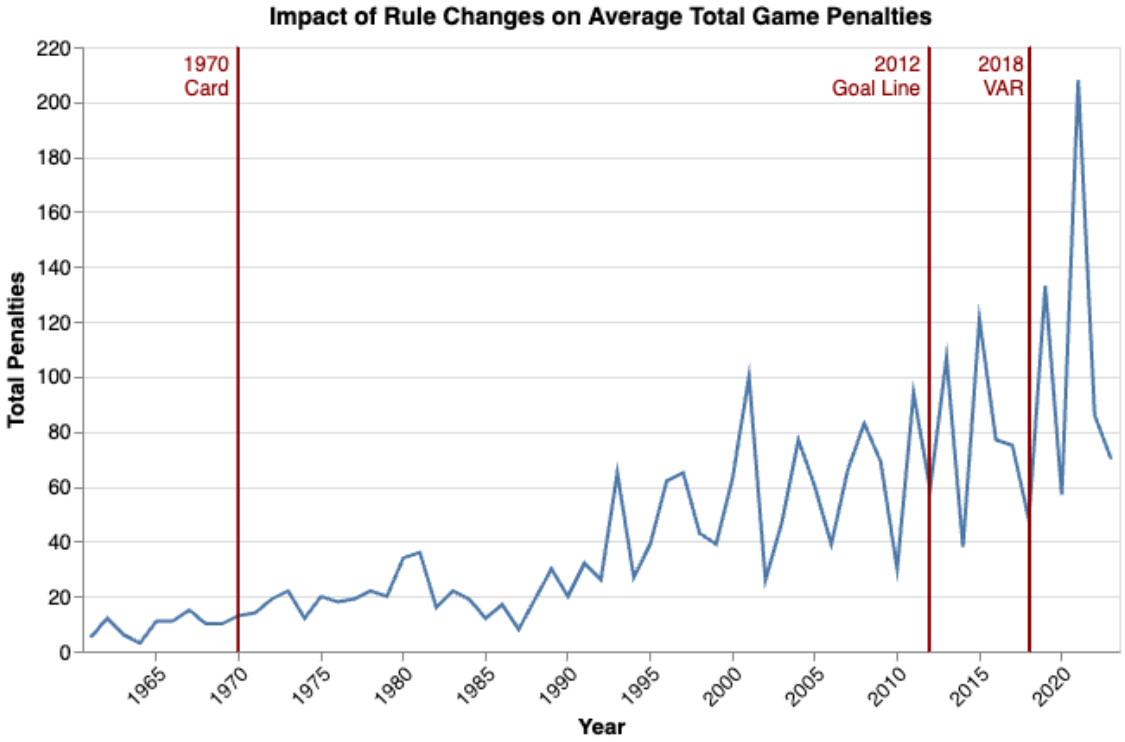
Plot 2 compares time between two goals,distinguishing if the next goal scored was from the same team or opposing team. For example, if team A scored the first goal and team B scored the next goal within 10 minutes, then the data would be plotted in blue ("opposite") graph at 10 minutes.

3: Years with significant change in football rules and its impact on average total goals scored



Plot 3 shows changes in average goals over the years (1960-2023). The red line marks years where significant rules regarding football matches were changed.

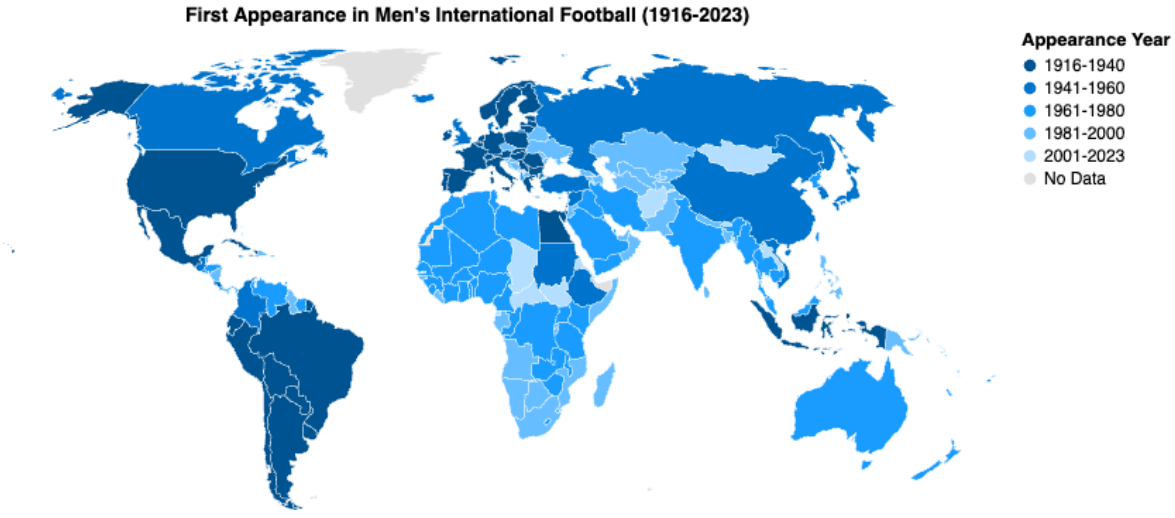
4: Years with significant change in football rules and its impact on average total penalties scored



Plot 4 also shows total penalties scored in a game over the years (1950-2023). Red line marks years where significant rules were introduced in the game.

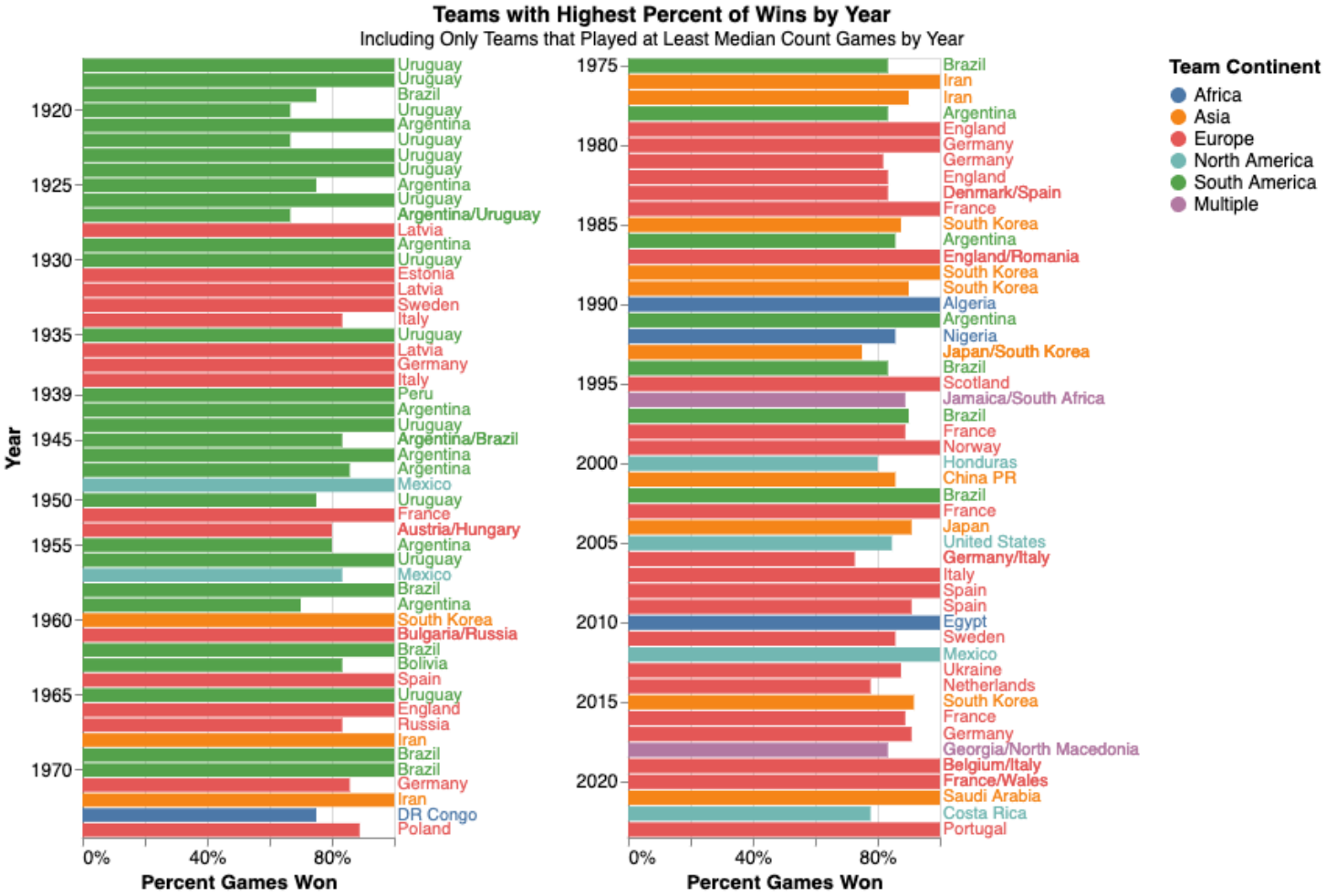
Team Plots

5: Map of First Appearance



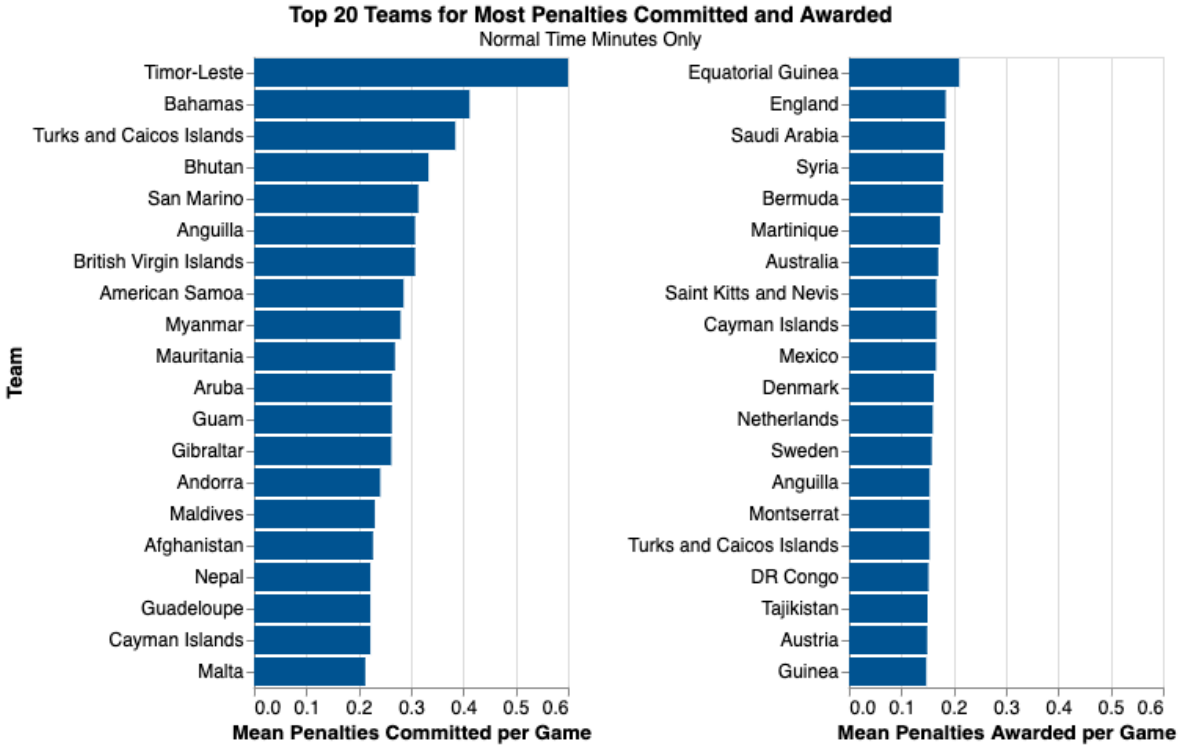
Plot 5 shows the range of years in which a team first appeared in Men's International Football, using 1916 as the earliest year. In general, teams from North America, South America, and Europe have been playing the longest.

6: Teams with Highest Percent of Wins by Year



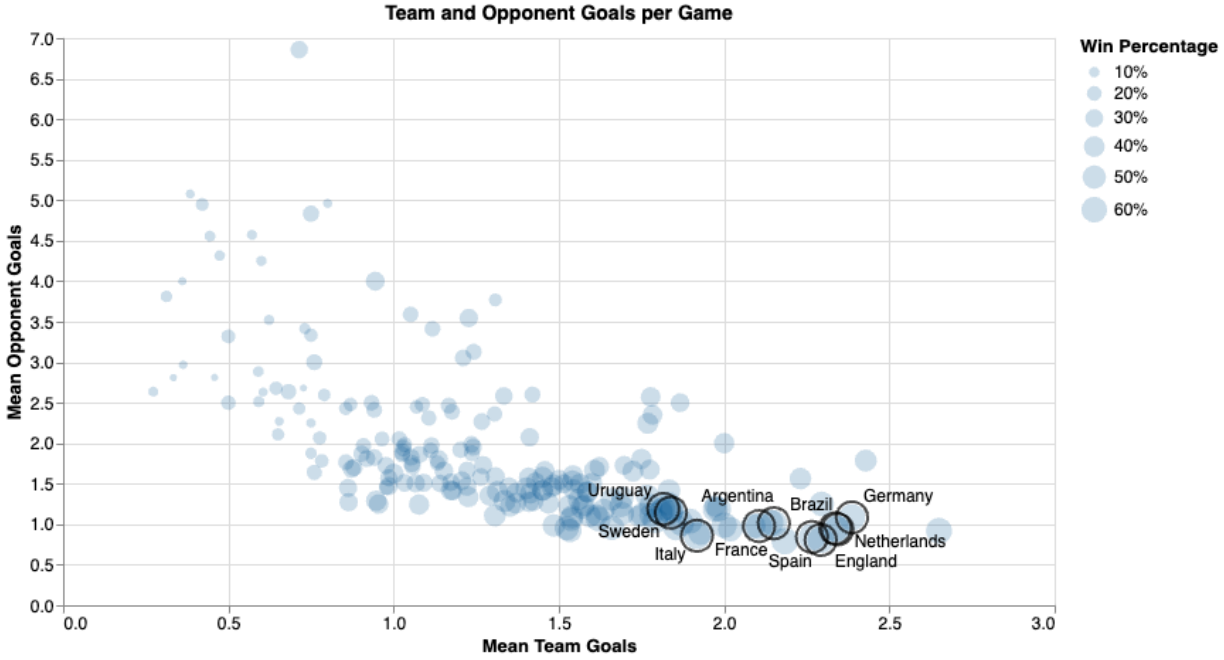
Plot 6 shows the team or teams that won the highest percent of their matches in each year. Only teams that played at least the median count of games per team for a year are included. This is to reduce bias in the data toward teams that played - and won - very few games.

7: Penalties Committed and Awarded by Team



Plot 7 shows the average count of penalties committed and awarded per game by team. The top 20 teams in each category are shown. Only normal-time penalties are included to avoid bias in the data.

8: Average Goals Scored by Teams and their Opponents

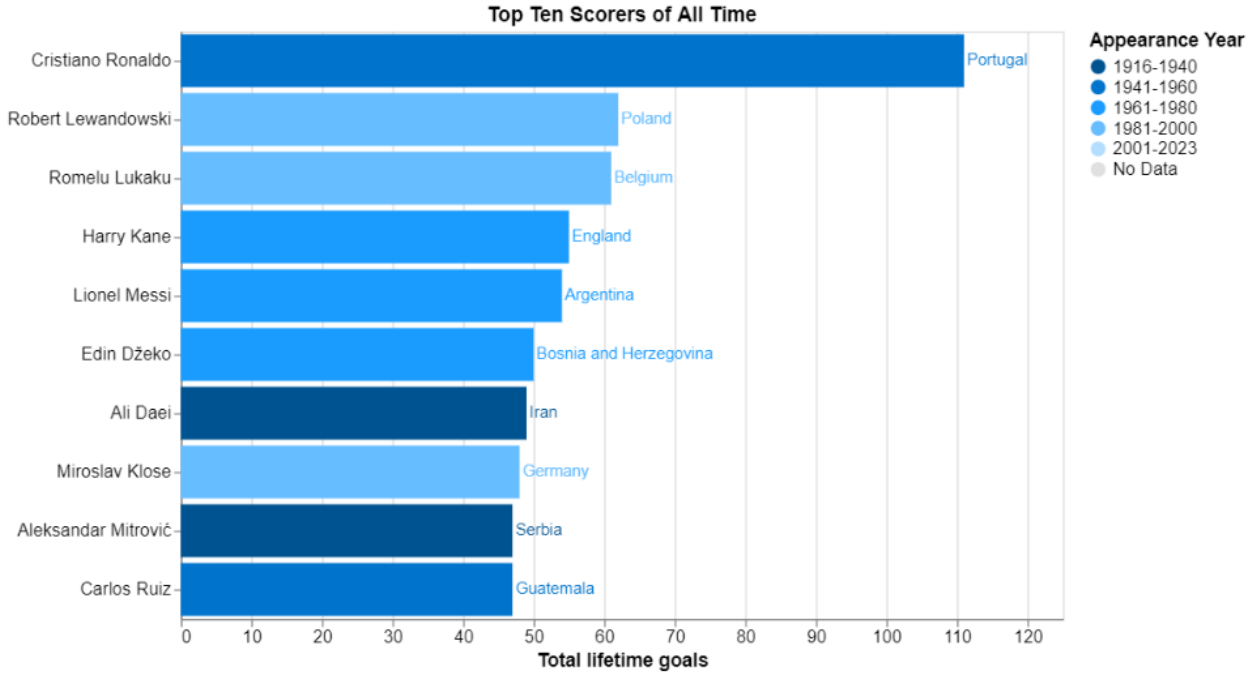


Plot 8 shows the average number of goals scored by teams and their opponents. In addition, a size encoding shows what the percent of matches each team has won overall. What are often considered the

top ten teams of all time are highlighted to show that better teams both score more goals and have fewer goals scored against them.

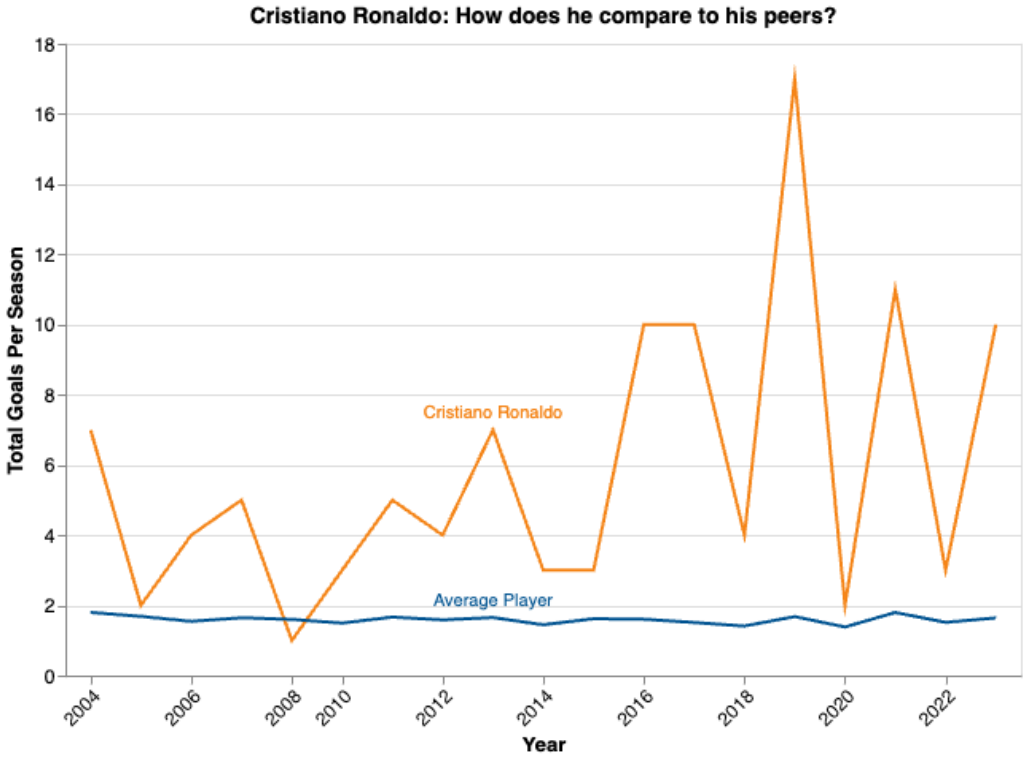
Player Plots

9: Top Ten Scorers of All Time



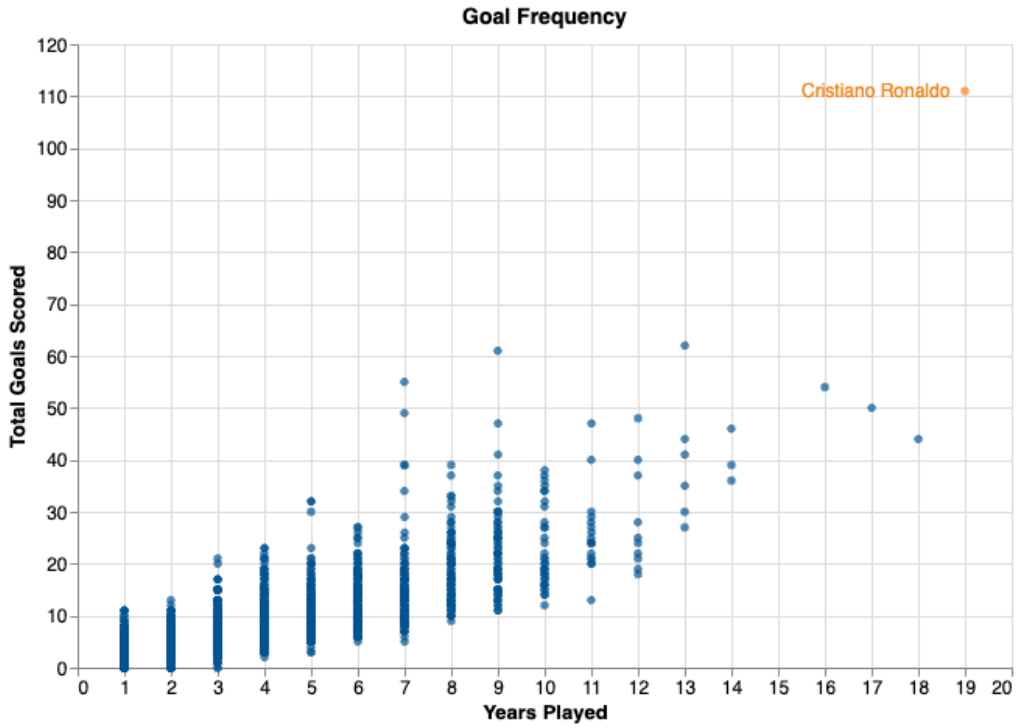
Plot 9 shows the top ten players with the most lifetime goals, and includes the team and appearance year of each player.

10: Cristiano Ronaldo: How does he compare to his peers?



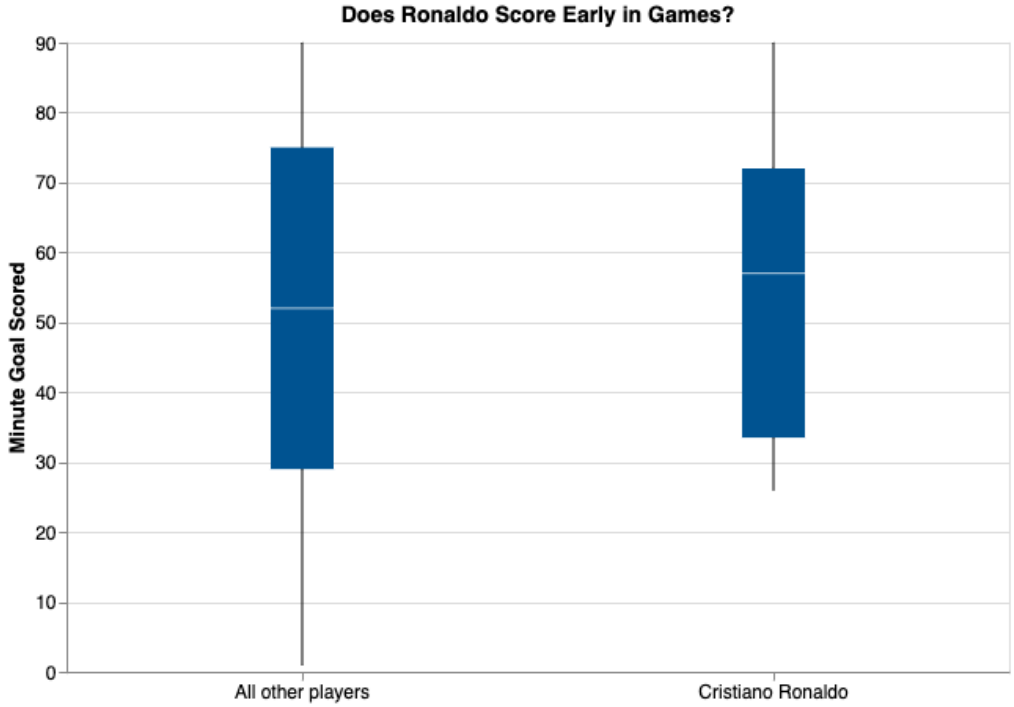
Plot 10 shows the total goals per season of Ronaldo compared to an average player during the years that Ronaldo played (2004 to 2023).

11: Goal Frequency



Plot 11 shows both the number of years played and total goals scored per player with Cristiano Ronaldo highlighted.

12: Does Ronaldo Score Early in the Game?



Plot 12 shows the distribution of goals scored by minute for Cristiano Ronaldo compared to all other players.

Design

The plots are designed to be cohesive. A custom theme was defined to ensure each plot had consistent sizes, fonts, and colors. The primary color used in each graph matches FIFA's branded blue. Additional colors and size changes were utilized when required. Plot-specific design encodings are explained below.

Game Plots

Plot 1 *('Average goal per game')* is an exploratory analysis of total goals scored in a match. The dataset initially was in a wide format; each row represented a goal scored. The first step was therefore, to aggregate total goals scored by both teams and create a new column to identify the pair the team that played together. These variables helped to identify total goals scored per game. After dropping the duplicates, the position encodings were used to visualize the total goals scored by both team and count of the matches. Total goals were encoded in x-axis with count of total match in y axis as total goals and the count of these goals were the most important relationship that plot1 wanted to showcase. Total goals scored in a game is a discrete data (Total goals are whole numbers) so creating bar charts was the most viable graph choice. Later, the graph was layered with line and text data to contextualize the distribution of the goal. Both color and text were utilized to highlight the average of the total goals scored in a game.

Plot 2 ('*Time between two consecutive goals')* uses position, color and facet encodings. The graph wanted to answer the question: after a team scored how quickly is the next goal scored. Either the same team or the opposite team could score the next goal and distinguishing the team that scored the next goal could reveal some patterns in the style of game and goals scored. Therefore, time between goals was used in x-axis encoding and count of goals that were scored in these time differences was used in y-axis. New column that distinguished if the next goal was scored by the same or different team was calculated and

used as color encoding. To make comparison of which team scored the next goal easier, facet encoding for same or opposite team was used as well.

Plot 3 (‘*Years with significant change in football rules and its impact on average total goals scored’)* wanted to explore if total goals scored in a game changed after significant rules were introduced in the game. In 1970, a system of red and yellow cards were implemented to penalize the team that committed fouls. Yellow cards were given to players that committed fouls as a warning. If the referee gives a red card then the player has to leave the pitch, leaving the team with only 10 players on the ground. This could give advantage to opposite teams. In 2012, goal line technology was introduced which determined if a goal was scored automatically and did not rely on referee’s discretion or players. If the ball crossed the goal line, regardless of the goalkeeper catching it or defending it, the goal would be counted. In 2018: VAR line (Virtual Assistant Referee) was introduced; an electronic aid that referees would reference to make decisions regarding penalty awards, player’s foul etc. All these rules could change the game and the plot wanted to explore if we can trace the impact of these rules in the total goals scored over the year.
Since it tracks changes in variables over time, the x-axis uses year encoding with y axis tracking total goals scored in that year. Since the graph traces changes in total goals over time, line graph encoding was suitable. The x-axis ticks and label were customized to show every 5 years so while the graph looked less busy, readers could still be able to track changes over the years. Use of title, label axis was also used to communicate findings of the graph. To highlight the significant rules added, a red line was used with text encodings. This encoding instantly informs the reader the year where rules were added and the text label introduces the rule and captures attention as well.

Plot 4 (‘*Years with significant change in football rules and its impact on average total penalties scored’)* also uses similar encoding choices as it explores how the same rules impacted the total penalties scored over time. Some of the fouls and VAR rules could grant the opposite team a penalty opportunity. So to track if these rules changed the total penalty scored, the graph has years and total penalty scored aggregated in the year in the position axis. The layer of year the rules were introduced was encoded with color and text choices.

Team Plots

In plot 5 (‘*Map of First Appearance*’), each team was encoded with the geographic position of its country. In addition, a binned color encoding was used to show the year in which the team first appeared in the data. Color encodings are effective in quickly showing geographic patterns.

Plot 6 (‘*Teams with Highest Percent of Wins by Year*’) also uses both position and color encodings. Year is encoded on the y-axis, while the percentage of games won is encoded on the x-axis. The color encoding shows which continent a team is located in. This color encoding is effective because it links back to regional patterns in the map, and quickly shows the viewer which global regions dominated men’s international football through time. South America, for example, had a clear dominance through about 1930. Team name is included as a text attribute as to not generalize countries and provide more granular information.

Plot 7 (‘*Penalties Committed and Awarded by Team*’) utilizes position encodings to show the teams that committed or were awarded the most penalty kicks. Team name is encoded on the y-axis so the labels are easy to read, and mean penalty kicks per game are encoded on the x-axis. The ordering of the y-axis by the values on the x-axis reduces cognitive load and allows the user to quickly see which teams had the most penalty kicks. Penalty kicks committed versus awarded are faceted into two columns to both improve the data-to-ink ratio with a shared axis and facilitate comparisons of the teams in each ordered position.

In Plot 8 (‘*Average Goals Scored by Teams and their Opponents’*), position and size encodings are used to compare the mean goals scored by a team per match to the goals scored by their opponent. The size encoding shows the percent of games each team has won. The position encodings are effective here because quantitative variables are being compared. The size encoding is effective because it points out that teams that both score goals and defend other teams from scoring goals are more likely to win.

Player Plots

A bar chart is appropriate and effective for the Plot 9 (‘*Top Ten Scorers of All Time*’) because ‘Total lifetime goals’ is a discrete variable and player is a categorical variable. These were both the most prioritized variables so they were both given a position encoding. I added a label to encode each player’s team because it is clear, simple, and accessible. Lastly, I encoded ‘Appearance year’ using color. I binned color to make it easier to distinguish which range of years each player falls into. I chose to make earlier years a darker shade of blue to signal an older year.

Plot 10 (‘*Cristiano Ronaldo: How does he compare to his peers?*’) is a line plot showing the total goals per season of Ronaldo compared to an average player during that year. Both year and total goals per season were given position encodings to most effectively communicate them, and a line chart was appropriate because we are examining a trend in a variable over time. Lastly, I encoded the player type (Ronaldo vs average player) using color, however I also added a direct label to improve accessibility for people with color impairment. Color and text were simple and effective encoding choices because there were only two categories of player type.

The plot (‘*Goal Frequency*’) (Plot 11) encodes ‘Years Played’ and ‘Total Goals Scored’ both as position variables. I created a scatter plot because it is effective and fully expressive. I double encoded Cristiano Ronaldo’s data point with both color and text to draw the viewer’s attention and clearly label his point.

Finally, Plot 12 (‘*Does Ronaldo Score Early in the Game?*’) is a boxplot showing the distribution of when goals are scored within a 90 minute game. A box plot is appropriate because it is very effective at communicating important and succinct information about a quantitative variable.

Analysis

Game Plots

Plot 1 (‘Average goal per game’) shows the distribution of the total goals scored in a game. The distribution of the score follows the trend of normal distribution. The quantitative value of the total data is discrete, however, the average goal scored per game is 3.09. So, on average, the total goals scored in a game is approximately 3.

Plot 2 (‘*Time between two consecutive goals’)* explores the time between two consecutive goals and investigates if one can expect the next goal to come from the same or opposite team. According to the analysis, if a team scores a goal then there is a higher chance that the same team would score the next goal within the next 20 minutes, compared to the opposite team. Average time between minutes when two consecutive goals were scored by the same team is close to 18 min, and for the opposite team is approximately 19.5 minutes. There is however, some biases in the data that needs to be considered. For example, if the match is between a really strong team and comparatively a weak performing team, there is higher chances that the strong team would score more consecutive goals. Strong team would also have a strong defense so the counter attack by the opposite team is easily blocked. Further explorations on

number of attacks by same and opposite team as well as ranking of the team could make this analysis more robust.

Plot 3 (‘*Years with significant change in football rules and its impact on average total goals scored’)* traces the change in total goals scored per year and changes in rules in 1970, 2012, 2018 lead to significant changes in the patterns of total goals scored per year. From the graph, we can see that total goals scored in a year is trending upward in general. However, the total goal scored per year is also volatile and it is hard to determine if the year where the rules were introduced and year following the implementation of rules significantly impacted the game. Therefore, with current data and analysis, it is difficult to conclude if the changes in the rules impacted the total game scored because of the volatile nature of the total goals scored per year.

Plot 4 (‘*Years with significant change in football rules and its impact on average penalty scored’)* traces the change in total penalty scored per year and if changes in rules in 1970, 2012, 2018 lead to significant changes in the patterns of total penalty scored per year. Similar to plot 3, the nature of the total penalty scored is trending upwards, however, the total penalty every year is very volatile and it is difficult to study the impact of new laws and on total penalty scored per year with current data. Calculating penalty awarded as opposed to penalty scored might help to narrow down the impact of rule change on total penalty.

Team Plots

Plot 5 (‘*Map of First Appearance*’) aims to show geographic patterns, if they exist, in when men’s football teams began playing on the international stage. Regional patterns are easily seen in the map. Teams from South America and Europe in particular have been playing the longest, which should be no surprise considering these teams are often still the best in the world.

Plot 6 (‘*Teams with Highest Percent of Wins by Year*’) also aims to show spatial as well as temporal patterns in winning teams. South American teams win most frequently in the earlier years of the data, while Europe gains more dominance over time. It is also interesting to see the years when other teams have particularly successful years.

Plot 7 (‘‘*Penalties Committed and Awarded by Team*’’) aims to show which teams tend to commit penalties versus those that have penalties committed against them. It is notable that the teams that are smaller and win less often commit more penalties, while teams known to be good kick more penalties.

Plot 8 (‘*Average Goals Scored by Teams and their Opponents*’) aims to compare the mean goals scored by each team to those scored by their opponents. This graph accomplishes a few things. First, it shows that teams who score more goals than their opponents on average win more games, as is to be expected. Second, it shows that teams that win more games effectively defend opposing teams from scoring many goals. In general, there is also more spread in how many goals a team’s opponents score on average (about 0 to 5.5 if excluding the outlier at 7) than in the goals a team scores on average (between about 0 and 3).

Player Plots

Plot 9 (‘*Top Ten Scorers of All Time*’) aims to capture the most valuable and notable players of all time and to check if there are any standout players. It does this by showing the total lifetime goals scored for the top ten highest scoring players and the year of the players debut and player’s team in order to show if there are any trends in when/where the best players played. The results show that Ronaldo is a standout

player with about double the goals as the next highest scorer. There is no apparent pattern of appearance year or team among the top ten players.

Plot 10 (‘*Cristiano Ronaldo: How does he compare to his peers?*’) aims to assess Ronaldo as a player across the span of his career compared to an average player. It shows a line plot of the number of goals scored per year by Ronaldo as well as the average across all other players for each year that Ronaldo played. It is easy to visualize the trends in time of goals scored and make a comparison between Ronaldo and the average players. The results show that while Ronaldo’s scoring fluctuates every year, he is usually far above average for number of goals scored.

Plot 11 (‘*Goal Frequency*’) aims to answer the question of how frequently (i.e. goals per year) Ronaldo scores compared to other players. This scatter plot shows the number of lifetime goals scored on the y axis and the number of years played on the x axis. Ronaldo’s point is highlighted to show how he has more longevity and a high rate of scoring compared to all other players. This plot is particularly effective at highlighting Ronaldo’s value as a player, because he is such an extreme outlier in terms of longevity and goals scored.

Plot 12 (‘*Does Ronaldo Score Early in the Game?*') assesses the validity of Ronaldo's well-established reputation for scoring goals in the early stages of matches. It is a box plot showing a side by side comparison of the distribution of the time of a goal scored for Ronaldo compared to all other players. This plot allows for easy comparison of the median, IQR, and range for time of goals scored between Ronaldo and all other players. The results show that Ronaldo’s reputation is, in fact, unearned. He does not score goals earlier in games than the average player.

Conclusion

In conclusion, this analysis aims to explore and understand the evolution of men’s international soccer. Our primary goal was to gain insight into changing dynamics in football, identify standout teams and players, and understand patterns and influences in goal scoring.

To accomplish this, we analyzed a dataset containing over 45,000 records with information on each goal scored in international men’s soccer from 1916-2023. We performed data cleaning, processing, and transformation then examined patterns in teams, players, and goal scoring over time.

By accomplishing these objectives, we improved understanding of the history and evolution of football and gave insight to patterns and influences in goal scoring, which could inform improved game strategy. The global cultural, historical, and economic importance of football brings people together, and encourages international cooperation and understanding. Through this analysis, we hope to provide value to the football community that will enrich understanding, appreciation, and strategy within football and continue to bring people together globally.

Works Cited

Dunbar, Graham. “Champions League final set to reach 450 million broadcast viewers worldwide.” *AP News*, 10 June 2023, https://apnews.com/article/champions-league-final-uefa-television-audience-d4c7c3ca39cf63602fa61edd7eec88e6. Accessed 7 December 2023.

Horberry, Roger. “What Are The Most Watched Sports In The World? - GWI.” *GWI Blog*, 6 April 2023, https://blog.gwi.com/chart-of-the-day/worlds-most-popular-sports/. Accessed 7 December 2023.

Juriso, Mart. “International football results from 1872 to 2023.” *Kaggle*, 2023, https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017/data. Accessed 7 December 2023.

Shea, Bill. “Super Bowl vs. World Cup: What the ratings say about the U.S. and global TV kings.” *The Athletic*, 15 February 2023, https://theathletic.com/4198045/2023/02/15/super-bowl-world-cup-tv-ratings/. Accessed 7 December 2023.

World Atlas. “The Most Popular Sports In The World - WorldAtlas.” *World Atlas*, 13 September 2023, https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html. Accessed 7 December 2023.