# Predicting How Many Matches a Team Will Play in a World Cup?

Team 23 - Jigar Patel, Henry Yost, Davinia Muthalaly, Refugio Zepeda

2025-11-07

GitHub Repository: STAT107 Team 23 — World Cup Analysis

## Abstract

This project looks at the key factors that predict how many matches a national team plays in the FIFA World Cup. Using data from the last ten tournaments, we used pre-tournament performance metrics including ELO, goal ratio, win ratio, and ranking trends. After data cleaning and pre-processing, we conducted exploratory visualizations to better understand our data and also determine what models would be best. Additionally, we built linear models and a random forest model to evaluate predictive strength.

Our analysis shows that goal ratio and ELO rating are the strongest predictors of World Cup progression, showing a statistically significant relationship with match count. Linear models struggled with the nonlinear structure of the World Cup data, but the random forest model captured these complexities and predicted match counts within 1 to 1.5 games for the 2022 WC. Win ratio showed inconsistent significance in linear models, but remained relevant for non-linear modeling. These findings prove that predicting World Cup outcomes using only pre-tournament statistics remains a difficult challenge due to the nature of the game and tournament. Future work could incorperate player-levle metrics and group-stage performance to further improve predictive accuracy.

## Introduction

The purpose of this analysis is to examine the key factors that influence how many matches a national team plays in the World Cup. The primary question we seek to answer is: How many matches does a team play in a World Cup. We conduct statistical analysis using linear modeling as a baseline, to determine the impact of variables such as the rating/ELO, team rank, average # of goals scored in the season, etc… The analysis is valuable to sports betters and large sports-oriented industries aiming to understand the current status of teams, and how well they will perform in the upcoming world cup. We approached this project, by first selecting two datasets that include information all the way from the early 20th century. However, because soccer has changed so much in recent years, we decided it would be best to only use the last 10 world cups as data (1982 - present). First, we trained a regression model on our cleaned data to determine the variables that are most significant for predicting how many matches a team plays in the WC. Additionally, this was used as a baseline model and a sanity check to verify our data is accurate to what we would expect. The results will give us a clearer picture of understanding how teams perform in the WC, based on the variables in the datasets, helping stakeholders to make more informed decisions in world cup games.

## Data

The data used for this analysis was sourced both from open-source GitHub repositories, a popular platform for sharing data and code projects, that provides publicly available data. The first dataset is from JGavier

(Link), which includes data from 1901 to present day, with variables: rank, ELO, rating/rank changes, and total yearly match data.

ELO is a formulaic way to represent team strength ratings in a weighted manner for any tournament-based event. The formula used to calculate the ELO values comes from eloratings.net.

$$R_n = R_o + K \times (W - W_e)$$

Where, $R_n$ is the new rating, $R_o$ is the pre-match rating, $K$ is a constant weight for the tournaments played. $W$ is the result of the game (1 for win, 0.5 for draw, 0 for loss) and $W_e$ is the win expectancy from previous data and team strength.

The second dataset we used was sources from Jfjelstul (Link), a Ph.D developer interested in World Cup data. This dataset includes data all the way from 1930 - present, including variables: number of matches played for that team in the WC of that year, and their performance (categorical: `group stage`, `first round`, `semi-finals`, etc…) During the process of joining both datasets, we also calculated additional values we thought would have a significant impact in a teams performance in the world-cup: * `win_ratio`: the number of wins divided by the number of total games; this variable gives us an idea of their W/L ratio, which can be valuable in determining how far a team will get in the WC. * `goals_ratio`: the number of goals divided by the total number of games; this variable gives us an idea of the average number of goals per game.

## Data Cleaning

The data cleaning process involved several steps to ensure the final datasets quality and consistency. Originally, the dataset from JGavier had 17,201 observations, however, after filtering from 1981 - present (and only every 4 years), and updating missing entries to 0, we had a much cleaner and more accurate dataset for our question. The reason we extracted the data from 1981 (world-cup year) - present, every 4 years, is because the data was collected on Dec. 31st. Had we collected on the same year as the World Cup, then the data would've been after the World Cup. We extract the match information from the second dataset from Jfjelstul, and similarly, extract from 1982 - present. After extracting all the necessary variables, to start the data cleaning process, we first converted all the variables to numeric to ensure consistency, and also calculating the `win_ratio` and `goals_ratio` explored in the Data section above. During the data cleaning process, we made sure to capture the negative values by converting the sign to an ASCII negative sign. This data is extremely important, because it shows patterns of the teams that went down in ELO. Additionally, we converted all of the values (except for team name) to numeric, to make it viable when working with our models.

### Primary Cleaned Dataset

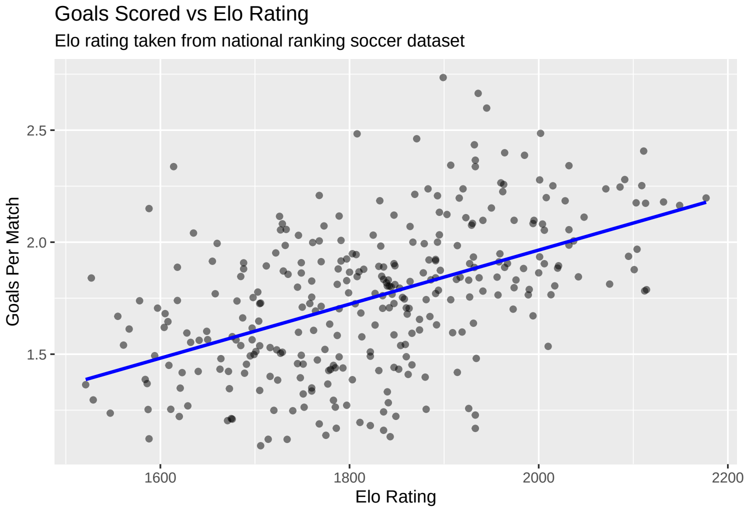
Our cleaned dataset includes the following variables:

- **world_cup_year** – Year of the World Cup
- **team_name** – National team name
- **elo_rating** – Team strength rating (ELO)
- **elo_rank** – Team rank based on ELO
- **elo_1yr_change_rating** – ELO rating change over past year
- **win_ratio** – Wins divided by total matches
- **goals_ratio** – Average goals scored per match
- **count_matches** – Number of matches played in the World Cup

## Visualization

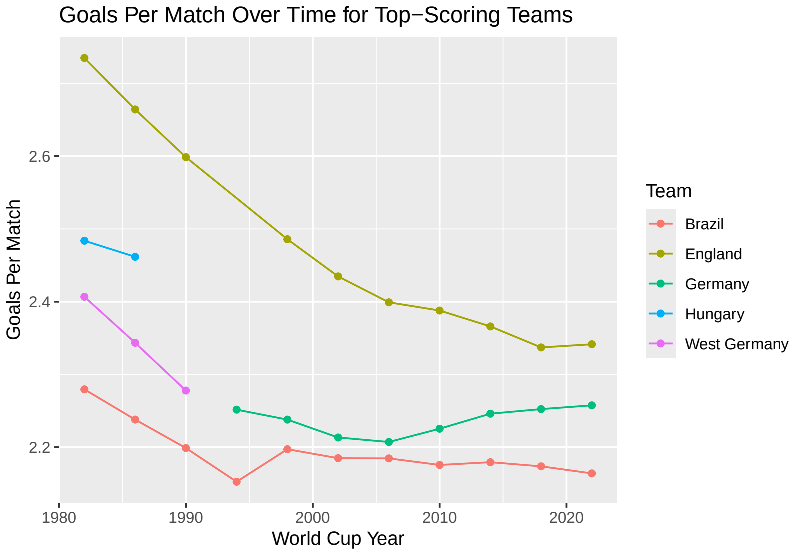
We used several visualization techniques in order to asses the correlation and nature of our variables and data to determine optimal tests and models to use for our project. These first set of graphs include visualizations, to explore the dataset, which will be further explored below:

### Goals vs. ELO Ratings



This scatterplot plots the Goals scored for per match in the World Cup versus the teams ELO ratings. It shows a clear, positive linear relationship between the two variables. While there is some noticeable variability, teams with higher ELO ratings generally tend to score more goals per match, as seen by the upward-sloping regression line. ELO summarizes long-term team performance and competitive strength before the tournament, and we can see that this rating is a relevant predictor for a teams offensive capabilities. Additionally, this scatterplot behaves as a ‘sanity’ check for our data, to verify that the expected relationship exists (linear and positive).

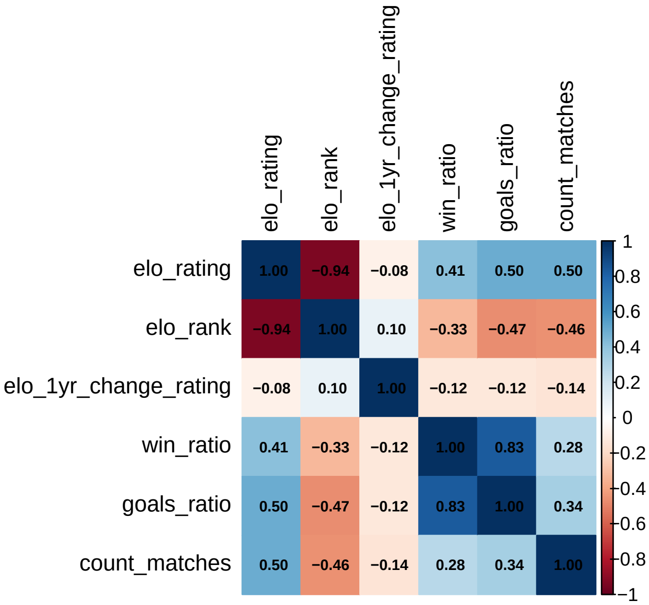
### Goals Per Match over years for Top−Scoring Teams



This line graph tracks how the highest-scoring teams in our dataset from various World cup years from 1982 - present (hence the West Germany). While the specific levels vary team by team, the majority of the teams show a gradual decline and/or stabilization in goals per match over time. Ultimately, this reflects more broad trends in global soccer, including more balanced tournament competition amongs the teams and potentially more emphasis on defensive plays. We can also see that historically strong teams, such as Brazil and Germany maintain a relatively consistent scores. These patterns suggest that past scoring behavior can be useful for predicting future scoring, which is parallel to our main question.

### Correlation Heatmap

Here, we can visualize and assess how the key numeric variables in our dataset (ELO rating, recent performance, match count, and goal scoring) all relate with one another.



| | elo_rating | elo_rank | elo_1yr_change_rating | win_ratio | goals_ratio | count_matches |
|---|---|---|---|---|---|---|
| elo_rating | 1.00 | −0.94 | −0.08 | 0.41 | 0.50 | 0.50 |
| elo_rank | −0.94 | 1.00 | 0.10 | −0.33 | −0.47 | −0.46 |
| elo_1yr_change_rating | −0.08 | 0.10 | 1.00 | −0.12 | −0.12 | −0.14 |
| win_ratio | 0.41 | −0.33 | −0.12 | 1.00 | 0.83 | 0.28 |
| goals_ratio | 0.50 | −0.47 | −0.12 | 0.83 | 1.00 | 0.34 |
| count_matches | 0.50 | −0.46 | −0.14 | 0.28 | 0.34 | 1.00 |

As seen in the heatmap, goals per match shows strong positive correlation with win ratio, and moderate positive correlation with ELO rating. This indicates that the teams who win more and have higher pre-tournament strength tend to score more goals. Match count is also positively correlated to goals and ELO, showing that stronger teams have the ability to not only advance further, but also have a stronger offensive. This is an interesting discovery, and something we can further explore with a linear model to determine which variables are most significant with match count beyond a correlation value. Additionally, ELO rank is strongly negatively correlated with ELO rating (i.e. 1989, 1244...) as higher ranks (i.e. Rank 1, 2, 3...) indicate stronger teams. Overall, these relationships should provide strong support for ELO-based and performance-based variables in our predictive models in answering our main question.

## Analysis

The main goal of our analysis is to understand which variables are most predictive of how many matches a team plays in the World Cup, so we can later build a stronger model for predicting tournaments. We are first using linear models as exploratory models, and not as final predictors. We are not fitting linear models to `count_matches` directly, because the relationship between match progression and performance statistics is non-linear and threshold-driven. Instead, we want to use linear models in an exploratory way, to find which predictors consistently carry statistical significance across related performance outcomes. We can use these findings to guide our decision making for the final model, and what predictors would be best.

### Linear Model 1: Predicting ELO ratings with performance based variables

In the first linear model, we will be predicting the ELO ratings with performance based variables: `win_ratio`, and `goals_ratio`. From the EDA, we know that the number of matches played in the World Cups is moderately positively correlated with ELO rating.

Table 1: Linear Model Coefficients: Elo_rating ~ Wins + Goals

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 1465.50392 | 38.49636 | 38.0686350 | 0.0000000 |
| win_ratio | -24.23886 | 143.35014 | -0.1690885 | 0.8658362 |
| goals_ratio | 212.04672 | 36.58418 | 5.7961314 | 0.0000000 |

This gives us an $R^2$ of 0.2496, or 24.96%, meaning 24.96% of the variance of the ELO ratings from the teams can be explained by performance based variables: `win_ratio`, and `goals_ratio`. The linear model struggling on this data makes sense, as soccer data is inheriently messy and non-linear, so this is something we expected.

However, interestingly, we can see that the `goals_ratio` variable is extremely statistically significant, with a $p < 0.001$. These are all observations that we need to take into account for the final model.

### Linear Model 2: Predicting win ratio with elo rating and performance based variables

In the second linear model, we will be predicting the win ratio with ELO rating performance based variables: `win_ratio`, and `goals_ratio`. From the EDA, we know that the number of matches played in the World Cups has a low positive correlation with the win ratio. This makes sense logically, especially because they're different observations with different targets.

Table 2: Linear Model Coefficients: Win Ratio ~ ELO + Goals + Matches

| term | estimate | std.error | statistic | p.value |
|---|---:|---:|---:|---:|
| (Intercept) | 0.0970750 | 0.0372319 | 2.6073090 | 0.0095642 |
| elo_rating | -0.0000041 | 0.0000244 | -0.1692447 | 0.8657138 |
| goals_ratio | 0.2131508 | 0.0093102 | 22.8943328 | 0.0000000 |
| count_matches | 0.0000815 | 0.0022294 | 0.0365388 | 0.9708761 |

This gives us an $R^2$ of 0.04654 meaning 4.65% of the variance in the win ratio from the teams can be explained by ELO and performance based variables: `win_ratio`, and `goals_ratio`. Similar to the first model, what we're seeing is that the linear model is too rigid for data that is constantly changing and does not move in a linear manner.

Similarly, interestingly, we can see that `goals_ratio` is a statistically significant predictor, with a $p < 0.001$. Again, this suggests that the goals ratio has a strong signal and will be valuable in our main model.

### Linear Model Results

The linear models unperformed, which is expected because of how World Cup progression works. Match outcomes are heavily non-linear. This suggests, that for our primary model, the best option would be a random forest, as it is appropriate for the data types and can handle non-linear relationships. Additionally, after the analysis, we think the best predictors for `count_matches` for a random forest are:

- `goals_ratio` consistently showed up as statistically significant across both models.
- `win_ration` is unstable and not significant linearly, but still relevant conceptually for non-linear models.
- `elo_rating` is moderately correlated with match progression, but the relationship is non-linear.
- `elo_rank` effectively the inverse of ELO rating, included as an aleternate representation of team strength.

**Random Forest Model**

Building on the insights from the linear models, we can shift to a random forest model. To verify the accuracy of our model, we're going to predict the number of matches each team played in the 2022 world cup, because we already know the statistics.

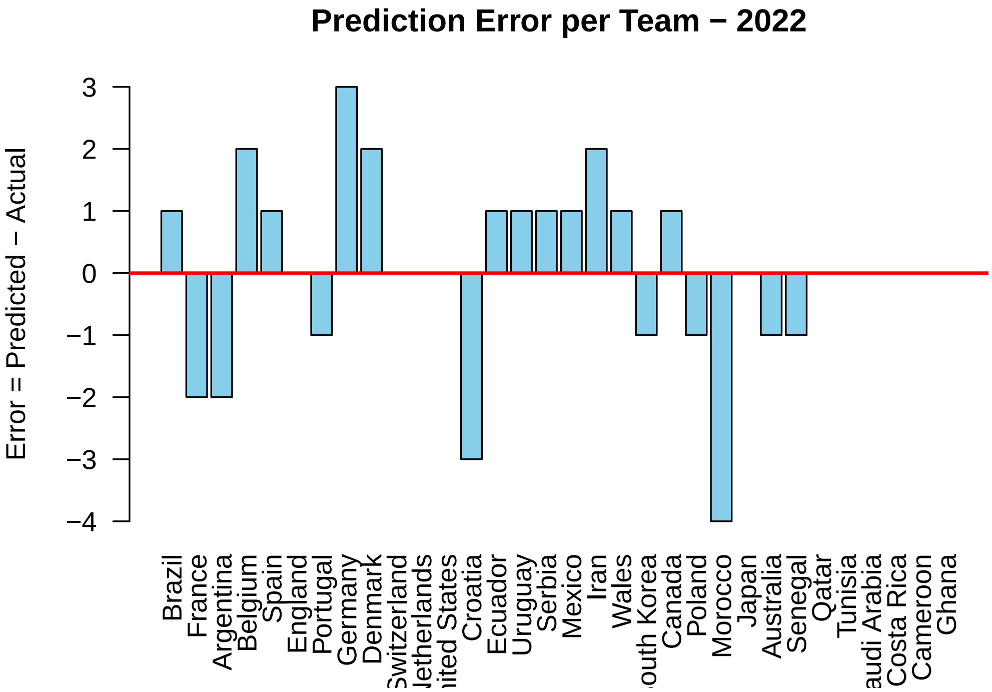


Table 3: Top 5 Predictions for 2022 WC

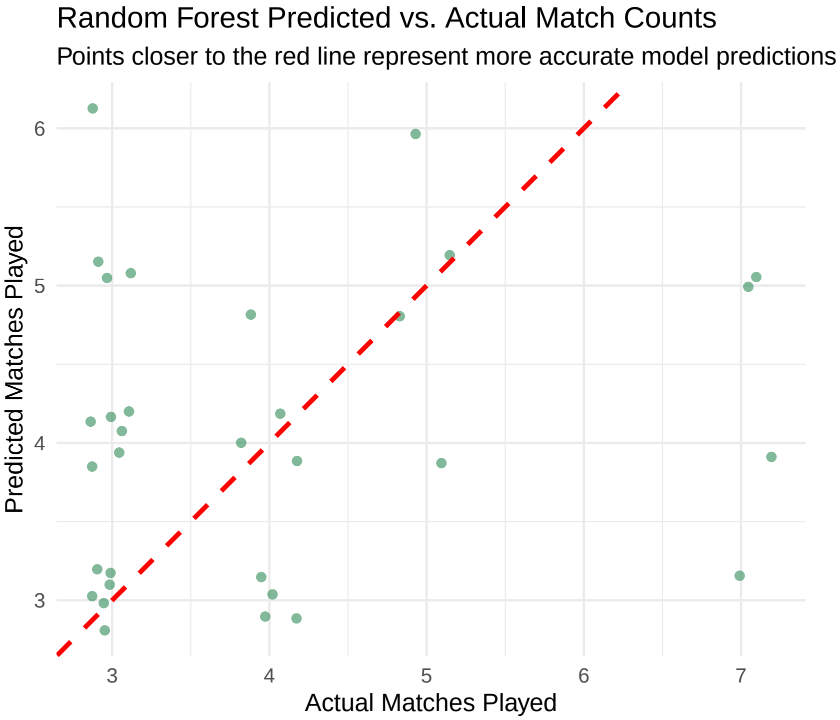
| team_name | count_matches | predicted_count_matches | error |
|---|---:|---:|---:|
| Brazil | 5 | 6 | 1 |
| France | 7 | 5 | -2 |
| Argentina | 7 | 5 | -2 |
| Belgium | 3 | 5 | 2 |
| Spain | 4 | 5 | 1 |

Table 4: Model Accuracy Metrics

| RMSE | MAE | MAPE |
|---:|---:|---:|
| 1.44698 | 1.03125 | 25.69196 |

The Random Forest model predicts World Cup match counts with reasonable accuracy: RMSE 1.43, meaning predictions are typically within one match to 1.5 matches; MAE 1, indicating average errors are about 1 match; and MAPE 24.65% which is relatively high magnitude considering the small range of match counts. Overall, this suggests that the random forest model is moderately strong in predictions with an average of +/- 1.5 games for the majority of the teams.

**Random Forest**



Additionally the scatter plot confirms the Random Forest model's accuracy, and additionally some of our EDA findings. The majority of the points cluster near/around the red dashed line, which represents perfect prediction from the model. This scatter plot validates our findings of an RMSE of 1.43 matches. Additionally, we're able to immediately identify outlines, such as Germany and Morocco that were significantly over/underestimated. This suggests that the model failed to learn from highly specific data that made the team win/lose more. This is an interesting finding, and similar to what our EDA found in the scatter plot and heat map, that because of the randomness of the tournament, the model would benefit from additional data: player metrics, injuries, etc…

## Conclusion

This analysis shows that pre-tournament performance metrics particularly, `goals_ratio` and `elo_ratig` are strong predictors of how far a team will progress in the World Cup. The linear models helped confirm which features mattered, but they were limited by the non-linear data. However, the Random Forest model was able to learn the complex relationships, predicting typically within one match of the actual results of the 2022 World Cup.

The model's biggest misses (Germany, and Morocco) show the limits of our data, and purely relying on team-level statistics for the past 10 world cups. Human-nature, such as injuries, player substitutions, rivalries, all introduce significant noise that exist in performance metrics. As a result, the errors aren't a failure on the Random Forest, but it rather reflects the relatively chaotic nature of the World Cup.

Future steps we'd be interested in pursuing are incorporating additional features such as player statistics, injuries, and additional group-stage metrics to improve the predictive accuracy. This would enable us from going to "decently correct" to a competitive model for forecasting the upcoming World Cup.

## Contributions

- Henry Yost (25% Final report)
- Jigar Patel (25% LM + Random Forest Model)
- Davinia Muthalaly (25% EDA)
- Davinia Muthalaly (25% Data Preprocessing)