How many matches will teams play in the next 26' World Cup?

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

2025-11-07

Abstract

Write Abstract Here

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 the 26' 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, 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, 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 patters 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

```
## Dataset size:  316 rows and 8 columns
```

```
##    world_cup_year    team_name elo_rating elo_rank elo_1yr_change_rating
## 1            1982 West Germany       2111        1                    -12
## 2            1982       Brazil       2091        2                     46
## 3            1982    Argentina       2006        3                      7
## 4            1982 Soviet Union       2001        4                     49
## 5            1982       Poland       1959        5                     12
## 6            1982      Belgium       1917        6                    -36
## 7            1982   Yugoslavia       1916        7                     42
## 8            1982        Italy       1914        8                   -102
## 9            1982      England       1899       11                   -101
## 10           1982      Austria       1895       12                    -36
##    win_ratio goals_ratio count_matches
## 1  0.5660377    2.406709             7
## 2  0.6272727    2.279545             5
## 3  0.5381883    2.053286             5
## 4  0.5573770    1.934426             5
## 5  0.4248705    1.948187             7
## 6  0.3862069    1.843678             5
## 7  0.4805492    2.196796             3
## 8  0.5318066    1.984733             7
## 9  0.6031746    2.734921             5
## 10 0.4432314    2.032751             5
```

This chunk is to show the cleaned Dataset size (316 x 8) and also give a peek into the dataset, so that its easier to understand what sort of data is being used in this project.

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

Insert image of graph here

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

Insert image of graph here

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.

Insert image of heatmap here

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.