Soccer by the Numbers: A Multi-Stage Analysis of the FIFA World Cup

By Liam Elson, Kyle James, Pritha Das, Asif Hossain

University of Manitoba

DATA 2010

Ashani Wickramasinghe

10 April 2023

Table of Contents

1. Introduction

For our data analysis, our team has opted to analyze the FIFA World Cup, a subject of interest for our team due to our shared passions for world geography and sports. The two datasets being analyzed are *wcmatches.csv* and *worldcups.csv*. The former contains data regarding each World Cup match played from the inaugural World Cup in 1930 to the penultimate World Cup in 2018. This data set contains critical match details, such as the winning and losing teams, the conditions in which the match was won, and the match score. While the latter dataset covers broader aspects of the tournament, such as the top four teams, the number of teams which competed, the number of games played, and the number of goals scored. During the analysis, our team discovered several typos in *wcmatches.csv*, which we rectified by submitting a corrected version through a pull request. This revision was later approved and merged by the repository owner.

The primary objective of our analysis was to utilize winning-related statistics and develop a linear regression model that could predict match outcomes based on those statistics. Additionally, we aimed to create a standardized scoring model to rank each team's performance in a World Cup. These rankings would enable us to conduct several hypothesis tests to investigate several questions our team had. Specifically, we were interested in determining whether the division of Germany into East and West during the Cold War had any impact on team success, whether the dissolution of Yugoslavia affected team performance, and whether a team's stage of play (group or knockout stage) had any effect on their overall performance. A final statistic we wished to examine was the score range each team covered.

2. Analysis

A. Data Pre-Processing

We began our analysis by extracting relevant information from our two data sets (*wcmatches.csv* and *worldcups.csv*), using the *tidyverse* library. To obtain country-level statistics for each tournament, we utilized the matches data set which includes goals scored by each team in every match, the match winner and loser, and the year in which the match was played. The data set contained columns for goals scored at home and away, which required separate calculations for both home and away. Using the summarize function, we computed three key statistics for each team playing as both home and away: total goals scored, total goals against, and total games played. Following this, we combined the resulting datasets using a full join based on the country names. However, as some countries had not played both home and away games, the joining process resulted in "NA" values. To address this, we filled these missing values with 0 since "NA" was equivalent to no goals scored in this context.

Once the missing values were replaced with 0, we proceeded to extract several key statistics from the data. Specifically, we calculated the total goals scored by summing the goals scored by a country at home and away matches. Similarly, we calculated the total goals against as well as the number of matches played. Using these statistics, we derived the goal differential by subtracting the goals against from the goals scored. Finally, we calculated the average goals scored and against per game by dividing the total goals scored and against by the total number of matches played.

We then aimed to incorporate the team's record into our analysis. To achieve this, we utilized the *winning_team* and *losing_team* columns in *wcmatches.csv* and calculated the number of wins and losses for each team, resulting in two separate datasets. Additionally, we calculated the total number of draws. This required separate calculations for home and away games as there isn't a winner or loser in a draw, which resulted in four total datasets. We then combined these data sets with a full join and removed any "NA" values.

After creating the combined dataset (which included the team's records and goals scored and against), we computed a win percentage for each team using the following equation:

$$\frac{total\ wins + total\ draws \times 0.5}{total\ matches\ played}$$

The above equation accounts for the number of wins and draws for each team, with draws counting as half a win. This is then divided by the number of games played to calculate the team's win percentage.

After consolidating all relevant statistics into a single data set, a copy of the dataset was created and summarized to obtain overall statistics for each country. The purpose of this step was to generate larger, more long-term statistics that could be used to build a scoring model. By summarizing the data in this way, we were able to gain a more comprehensive understanding of each country's performances.

B. Constructing a Linear Regression Model

To create a model, we opted to use linear regression. Our first step was to identify the variables that possessed a high degree of correlation with winning. We selected wins as our gold standard. We then analyzed the data and generated a table displaying the correlation values for each variable, which can be found in *Table 1* below:

Variables	Correlation
Total Goals & Total Wins	0.9770
Total Goals Against & Total Wins	0.8585
Total Goals & Total Losses	0.8736
Total Goals Against & Total Losses	0.9623
Goals per Game & Total Wins	0.8428
Goals Against per Game & Total Wins	-0.4802
Goals per Game & Total Losses	0.5415
Goals Against per Game & Total Losses	-0.1608
Goals per Game & Win Percentage	0.8725
Goals Against Per Game & Win Percentage	-0.6971
Goal Differential & Win Percentage	0.4870
Goal Differential & Total Losses	0.0915
Goal Differential & Win Percentage	0.7553
Goal Differential & Total Draws	0.2428

Table 1. Correlation Between Two Variables

By examining the above table, we note that goals per game has a strong correlation with win percentage. On the other hand, we also found a strong negative correlation between goals against per game and win percentage. These findings suggest that teams that score more goals are likely to win more frequently, while teams that allowed more goals are likely to lose more often. These observations are consistent with the nature of the sport, where the winning team is the one that scores more goals than their opponent.

It is worth noting that we choose goals per game and win percentage rather than total goals and total wins, despite not having as strong of a correlation. This decision was made due to the stronger negative correlation between goals against per game and win percentage. By using normalized values, we are better able to establish meaningful relationships between the variables to generate a more accurate scoring model.

With an understanding of the variables that were highly correlated, we proceeded to build a linear model to test their predictive power. To assess the model's performance, we split the dataset into training and testing sets and fitted a linear model using the *lm()* function. We then used the model to predict outcomes for the test data and calculated the RMSE between the predicted and actual values. The resulting RMSE was 0.0919, indicating that the model performed very well.

C. Constructing a Scoring System

After discovering the most impactful covariates, we attempted to build a scoring system. Our initial approach was to simply subtract goals against per game from goals per game. However, upon closer examination of the data, we found that some World Cups had more scoring overall than others. For instance, both the 1954 and 1958 tournaments had an abnormally high numbers of goals scored for number of games played. Therefore, to create a more accurate and interesting score, we needed to normalize the goals for and against based on the year they were in.

To accomplish this, we began by grouping the dataset with each team's performance by year. Then, we calculated the mean and standard deviation of goals per game and goals against per game for each year. Next, we joined this data to the original dataset using a left join. Using the z-score method for normalization, we normalized each stat to the mean and standard deviation of the given year. We believed that the z-score method was the most appropriate approach, given the abundance of samples and the assumption that the data was approximately normal. We also generated a pair of graphs that showed that the data was mostly normal, with some outliers to the right. With normalized data, we calculated scores using the same formula as before:

$$score = normalized\ goals\ per\ game - normalized\ goals\ against\ per\ game$$

To answer the first question of our analysis, we sorted every country's scores in descending order. According to our scoring system, Germany during the 2014 World Cup was the best team to ever play at a World Cup. *Table 2* displays the 5 best teams to play at the World Cup according to our scoring system:

Ranking	Country	Year	Score
1	*Germany*	2014	3.5492
2	*Brazil*	2002	3.3798
3	*France*	1998	3.1156
4	*Argentina*	2006	3.0010
5	*Germany*	2010	2.9999

Table 2. Best World Cup Performances by Score

Similarly, we can sort each country's scores in ascending order to view the worst performances by a country at a World Cup. By examining table 3, we can conclude that Saudi Arabia in 2002 was the worst team to play at a world Cup.

Ranking	Country	Year	Score
1	*Saudi Arabia*	2002	-5.3728
2	*North Korea*	2010	-5.0999
3	*Bolivia*	1950	-4.5686
4	*Greece*	1994	-4.5059
5	*South Korea*	1954	-4.4479

Table 3. Worst World Cup Performances by Score

By taking the mean score of each country, we were able to determine the best team on average. A table of the best 5 countries on average follows:

Ranking	Country	Score
1	Germany	1.5813
2	Brazil	1.5373
3	West Germany	1.3334
4	Soviet Union	1.1317
5	Netherlands	0.9630

Table 4. Best Countries at the World Cup by Average Score

The scoring system was used to identify the best and worst teams in the World Cup. However, we also wanted to determine if these teams had a good finishing position in the tournament. To validate this, we extracted information on the top four finishers of each tournament from the *worldcups.csv* dataset using loops. Subsequently, we joined this new dataset with the scores data set using a left join, which produced the following values:

Ranking	Country	Year	Score	Placement
1	*Germany*	2014	3.5492	1
2	*Brazil*	2002	3.3798	1
3	*France*	1998	3.1156	1
4	*Argentina*	2006	3.0010	NA
5	*Germany*	2010	2.9999	3
6	*Colombia*	2014	2.9378	NA
7	*West Germany*	1990	2.9049	1
8	*Belgium*	2018	2.8899	3
9	*Brazil*	2006	2.8830	NA
10	*Uruguay*	1930	2.8585	1

*Where NA values signify placing outside of the top4 *

Table 5. Placements of the Best 10 Teams at the World Cup by Score

Table 5 reveals that the top 3 teams in the scoring system won the tournament, with 5 of the top 10 teams also winning. Furthermore, 7 out of those 10 teams were among the top 4 finishers in the tournament. However, it is important to note that some teams that performed well did not win the tournament due to a single bad game. Expanding the analysis to the top 25 teams, it is observed that 60% of the teams finished in the top 4. These results suggest that the scoring system is capable of identifying some of the best teams in the history of the World Cup.

D. The Effect of Partitioning and Dissolving a Country on World Cup Performance

Due to our team's interest in global geography and politics, we have decided to analyze the effect of partitioning and dissolving a country on its mean performance at the FIFA World Cup. To achieve this, we have conducted an ANOVA and two sample t-test on two case studies -

Germany and Yugoslavia. We have analyzed the mean performance of German teams before and after unification, as well as during the partitioned period. Similarly, we have analyzed the mean performance of Yugoslavia before its dissolution, as well as the mean performance of the countries that emerged from its dissolution. Through this analysis, we hope to provide insights into the impact of significant political changes on a country's ability to compete on the global sporting stage.

Our first hypothesis test was an ANOVA test conducted on the mean performance of German teams. The populations selected for comparison were the mean scores of East Germany, West Germany, and a unified Germany. We set a level of significance of 0.05 and determined that the null hypothesis would be rejected if the calculated p-value was less than 0.05. Our hypotheses follow:

$$H_0\colon\ \mu_{Unified\ Germany} = \mu_{East\ Germany} = \mu_{West\ Germany}$$

$$vs$$

$$H_A\colon Atleast\ one\ \mu\ Differs$$

Using R, we calculated a p-value of 0.55, meaning we fail to reject the null hypothesis. Therefore, we conclude that the partitioning of Germany did not have a statistically significant effect on the mean performance of its national teams at the World Cup.

We also conducted a second hypothesis test to analyze the mean performance of German teams, in the form of a two-sample t-test. In this test, we combined the performances of East and West Germany into a single split Germany population, to be contrasted against a unified German Population. Like the previous test, we set a level of significance of 0.05 and determined that the null hypothesis would be rejected if the calculated p-value was less than 0.05. Our hypotheses follow:

$$H_0\colon\ \mu_{Unified\ Germany} = \mu_{Split\ Germany}$$

$$vs$$

$$H_A\colon\ \mu_{Unified\ Germany} \neq \mu_{Split\ Germany}$$

A p-value of 0.52 was calculated for this two sample t-test, meaning we fail to reject the null hypothesis. Once again, we conclude that the partitioning of Germany did not have a statistically significant effect on the mean performance of its national teams at the World Cup.

Similarly, we conducted an ANOVA test on the mean performance of the national teams of Yugoslavia and countries which emerged from its dissolution. The populations selected for this test were the mean scores of Yugoslavia, Croatia, Serbia, and Slovenia. It should be noted that although more countries emerged from Yugoslavia, only these countries have participated in multiple World Cups. As with the previous tests, the level of significance was set to 0.05, the null hypothesis would be rejected if the calculated p-value was less than the level of significance, and our hypotheses were stated as the following:

$$H_0:\ \mu_{Yugoslavia} = \mu_{Croatia} = \mu_{Serbia} = \mu_{Slovenia}$$
$$vs$$
$$H_A: Atleast\ one\ \mu\ Differs$$

This test rendered a p-value of 0.13, meaning we fail to reject the null hypothesis again. We conclude that the dissolution of Yugoslavia did not have a statistically significant effect on the mean performance of the Yugoslavian and emerging countries' performance at the World Cup.

Our final hypothesis test is a similar two sample t-test with Yugoslavia's mean score as our first population and the mean scores of the emerging countries which were formed as our second population. We used the same level of significance and decision rule as our previous tests, and used the following hypotheses:

$$H_0:\ \mu_{Unified\ Yugoslavia} = \mu_{Dissolved\ Yugoslavia}$$
$$vs$$
$$H_A:\ \mu_{Unified\ Yugoslavia} \neq \mu_{Dissolved\ Yugoslavia}$$

This two-sample t-test returned a p-value of 0.1635, meaning we fail to reject the null hypothesis and conclude that the dissolution of Yugoslavia did not have a statistically significant affect on the mean performance of the Yugoslavian and emerging countries' performance at the World Cup.

E. Performance Differences in Group and Knockout Stages

Our group aimed to investigate the differences in team performance during group and knockout stage games. To accomplish this, we utilized two distinct approaches to examine the data. To start our analysis, we used *matches.csv* to identify which games qualify as group and knockout stage game. A data frame named *matches_rename* was created, which renamed the stage column value to “Group” or “Knockout.”

Approach 1:

The *matches_rename* dataset was utilized to extract year, *home_team*, *away_team*, *home_score*, *away_score*, outcome, and stage to calculate the total number of goals scored by each team in both Group and Knockout games. To compare the performance of teams between these two game modes, a subset of the group dataset was considered from the knockout dataset by creating two separate functions: one that calculated the total goals scored by each team (score) and another that compared if a team plays better in Group or Knockout (comparison). The group and knockout datasets were then merged using inner join to avoid any NA values. A bar plot was created to compare the total goals scored in Group and Knockout games.

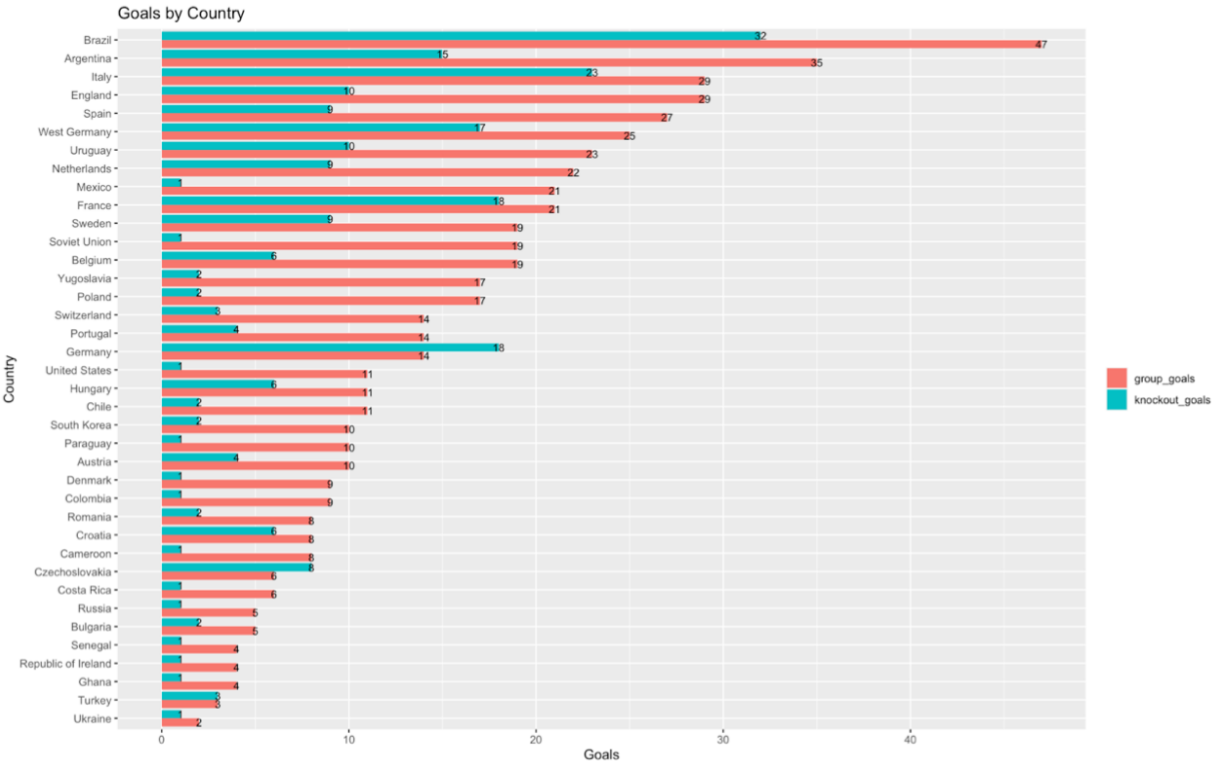

Figure 1. Stage Goals by Country

As observed, most teams scored more goals in Group games compared to Knockout games. A strong positive correlation of 0.7871 was found between the total goals scored in Group and Knockout games, indicating that these two variables are highly correlated. This suggests that an increase in goals scored during the Group games leads to a corresponding increase in goals scored during the Knockout games and vice versa.

Approach 2:

Scoring the overall performance of each team using our scoring model, we first created *Home_Group_Goals* by grouping all the teams and years and calculating all the goals scored by the home team, goals against the home team, and the total number of matches played by the home team in that World Cup. Similarly, *Away_Group_Goals* was made by grouping all the teams and years, and calculating the goals made by the away team, goals made against the away team, and the total number of matches played by the away team in that year. These two datasets were joined using an inner join to avoid any NA values and resulted in the creation of the *Group_Goals* dataset. Likewise, a *Knockout_Goals* dataset was also created.

Additional statistics for these two datasets were also mutated, which were required by our scoring model. These goals were normalized, and the scoring model was applied to both datasets. The scores obtained from both datasets were then used to create a bar plot.

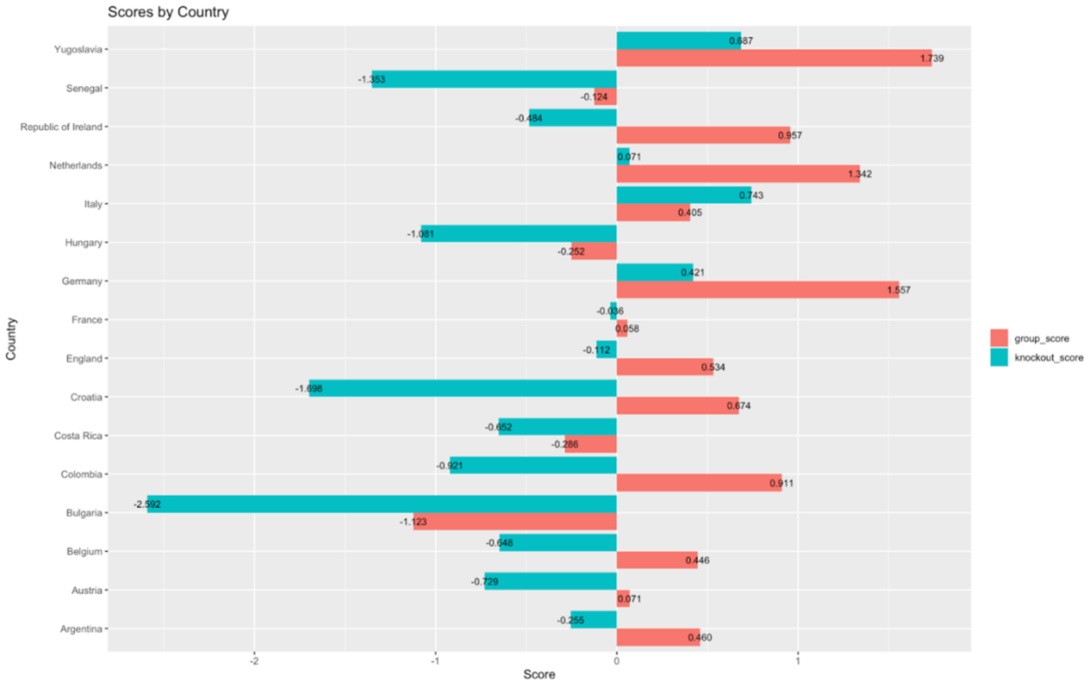

Figure 2. Stage Scores by Country

Our analysis revealed that, similar to approach 1, the scores of the team and their performance are much better in group games than in knockout games. This observation was further supported by the correlation taken between the normalized group scores and normalized knockout scores, which was found to be 0.6881.

To formally test the second approach, a matched pairs t-test was conducted. We first filtered the mean score of every team in the group and knockout stage, to be used as our two population. A 0.05 level of significance was selected and the following hypotheses were determined:

$$H_0\colon\ \mu_{Group\ Stage\ Score} = \mu_{Knockout\ Stage\ Score}$$

$$vs$$

$$H_A\colon\ \mu_{Group\ Stage\ Score} \neq \mu_{Knockout\ Stage\ Score}$$

Using R, a p-value of 1.997e-05 was calculated for this t-test. We hence reject the null hypothesis and conclude that playing in the knockout stage affects a team's performance.

F. Examination of Team’s Scoring Range

A final statistic we wished to explore was the range of scores for different teams. We started by taking the minimum and maximum scores for each country then finding the differences to find the range of scores. We were then interested in finding the teams with the widest scoring range. A table of this can be found below:

Ranking	Country	Max Score	Min Score	Score Range
1	Hungary	2.6152	-3.0563	5.6716
2	Saudi Arabia	-0.0963	-5.3728	5.2765
3	Brazil	3.3798	-1.8099	5.1898
4	South Korea	0.6918	-4.4479	5.1398
5	Mexico	1.4977	-3.616	5.1140

Table 6. Five Widest Score Ranges of Teams at the World Cup

3. Conclusion

Our analysis involved two datasets, wcmatches.csv and worldcups.csv, and aimed to develop a prediction model and scoring system to investigate several hypotheses. A linear regression model was created to predict the outcome of games, which yielded a strong model with a RMSE value of 0.0919.

Subsequently, a scoring system was developed to rank teams, and several hypothesis tests were conducted to explore the effect of partitioning and dissolution of countries on the success of their teams. ANOVA tests were conducted on German and Yugoslavian teams, and the results showed no statistical evidence to reject the null hypothesis. Two-sample t-tests were also performed on the same premise and yielded similar conclusions to the ANOVA tests.

Moreover, a matched pairs t-test was conducted to investigate if the tournament stage affects a team’s performance. The results of this test led us to reject the null hypothesis and conclude that the playing stage does influence a team’s performance. Lastly, we calculated each team’s minimum and maximum scores, to determine which team had the widest range of scores.