|  |
| --- |
| FIFA World Cup:  Winning Teams |
|  |
| JULY 24  Team 9 Consultants (T9C)  SCS 3250 – 08  Julian Cheng, Hador Hasan, Robert McGowan, Aqib Salman, Kim Wu |

Contents

[Objective 3](#_Toc520150970)

[Data Preparation 3](#_Toc520150971)

[Data Source and Procurement 3](#_Toc520150972)

[Data Quality Challenges 4](#_Toc520150973)

[Data Formatting Challenges 5](#_Toc520150974)

[Data Variables 5](#_Toc520150975)

[Analysis 6](#_Toc520150976)

[Overview 6](#_Toc520150977)

[Variables 7](#_Toc520150978)

[Logistic regression 9](#_Toc520150979)

[Observations 10](#_Toc520150980)

[Top Three Variables 11](#_Toc520150981)

[Conclusions 13](#_Toc520150982)

[Appendix 14](#_Toc520150983)

[Data Samples 14](#_Toc520150984)

[Sample Code 16](#_Toc520150985)

[References 18](#_Toc520150986)

# Objective

Team 9 Consultants (T9C) were contracted by Shady Bookkeeping Incorporated to perform data-cleaning and preliminary analysis on FIFA World Cup matches to explore opportunities for increased yield on their business investments. The challenge was to analyze their data to find what makes a winning team.

After considering how to best predict what determines a winning team, T9C chose to focus on a few main themes: team attributes, elimination round matches and matches up to and including the 2014 world cup. Rather than focus on the results of specific teams, T9C focused what attributes might be accurate predictors of winning for all teams.

Some of the attributes considered include but are not limited to the following:

* number of wins
* nationality of coach
* number of cards received
* team experience

T9C elected to solely consider elimination round matches as elimination matches have a decisive winner and are therefore played at the highest level of competition. Group Stage matches do not necessarily carry the same intensity as many factors can determine a team’s desire or necessity to win.

The team reviewed the data, ran a logistic regression analysis on the chosen variables and proposed likely predictor variables for further analysis.

# Data Preparation

## **Data Source and Procurement**

The primary data source for the analysis was a standard machine learning data set obtained from Kaggle (<https://www.kaggle.com/abecklas/fifa-world-cup>). This contains data from the first World Cup (1930) up to and including 2014. The data was provided in three comma separated files:

* WorldCups.csv
  + Each row corresponds to a year in which the world cup was played and contains columns for the year, the total attendance, who won, and other facts.
* WorldCupMatches.csv
  + Each row corresponds to a specific match with columns for a unique match ID and match information such as which teams played, who was considered the home team, how many goals each team scored, when the match took place, etc.
* WorldCupPlayers.csv
  + Each row corresponds to a player in a specific match with columns for the match ID as well as the name of the coach, a string of noteworthy events each player was a part of (getting a card, scoring a goal, being substituted on or off, etc. ) and some other player attributes.

Data dictionaries were available on Kaggle for all three sources above.

## **Data Quality Challenges**

There were three major issues with the data:

* We decided to only analyze elimination round matches.
  + In group stage matches, there is the possibility of a draw which makes the analysis more complicated. Only elimination round matches (including the third-place play-off) guarantee there is a winner and a loser in each match.
  + There are often also occasions where teams are not motivated to win either because they have already qualified for the next round or because they are already out of contention.
* The supplied data dictionary for WorldCupPlayers.csv was not complete and contained false information.
  + The data dictionary stated that a player with an ‘O’ event had scored an own goal. But in fact, ‘O’ meant that the player was substituted out, while ‘W’ meant that the player had scored an own goal. There were a few other errors and omissions of the same sort. The data was corrected by referencing Wikipedia.[[1]](#footnote-1)
* The WorldCupMatches.csv was missing some information.
  + Without this information the winner of the match could not be determined. In these cases, we turned to Wikipedia to fill in the data.1

## **Data Formatting Challenges**

* None of the provided datasets presented the data as we needed it.
  + We wanted each row to represent a team playing a match and the columns representing the attributes of that team (whether they win, whether they are the home team, how many matches they had already played in their history, and others).
  + To generate our basic data table, we restructured the WorldCupMatches.csv and for each elimination-round match we created two rows: one for the team that won the match and the other for the team that lost. Relevant information from the match was kept in the new data table and additional columns were added.
* There were duplicate rows in the Players set.
  + Some observations in the Players set were duplicates. It appears as if a player played twice in the same game. The duplicates were identified and removed.
* Blank rows in the data sets.
  + WorldCupMatches.csv contained 3720 blank rows. The blank rows were removed.

## **Data Variables**

After looking over the datasets to see what information was available and possibly worth analyzing, we determined that we would want a data table as described above (one row per team in the elimination round matches) with the following columns. Most of these columns were not provided in the provided datasets. We had to engineer them by manipulating the provided data.

|  |  |
| --- | --- |
| VARIABLE | MEANING |
| Team | The name of the nation team |
| Team Initials | The unique three letter code for the nation team |
| Win | Indicates whether the team won or lost the match |
| WinLoss% | Percentage of games the team won team or zero if no previous games played because this gave the best analysis |
| Games Played | Count of games team played prior to match |
| Wins | Count of games team won prior to match |
| Losses | Count of games team lost prior to match |
| Foreign Coach | Boolean variable indicating whether the coach is from the same country as the team |
| All Substitutions | Count of all substitutions during the match |
| IH | Count of substitutions at halftime or going into extra time |
| I | Count of substitutions at other times received by the team |
| Y | Count of yellow cards in the match received by the team |
| R | Count of red cards in the match received by the team |
| All Cards | Count of all cards in the match received by the team |
| Home | Boolean indicating whether the team was identified as the home team |
| Foreign Soil | Boolean indicating whether the team is playing on foreign soil |
| XP | Sum of the total number of games each player has played prior to the match. If an individual player’s experience is the number of appearances in matches that player played prior to the match (whether he started or was substituted on), the (team) XP is the sum of all players who make an appearance (whether starting the match or substituting in) |
| Goals For/Match | The average number of goals scored by the team prior to the match or one if the team had no previous matches because this gave the best analysis |
| Goals Against/Match. | The average number of goals scored against the team prior to the match or one if the team had no previous matches because this gave the best analysis |

# Analysis

## **Overview**

After performing an analysis (to be explained in the next section) on the aforementioned variables, T9C suspected that these five variables would be the most predictive:

* Home team
* Number of wins
* Foreign coach
* Experience
* Win/ match percentage

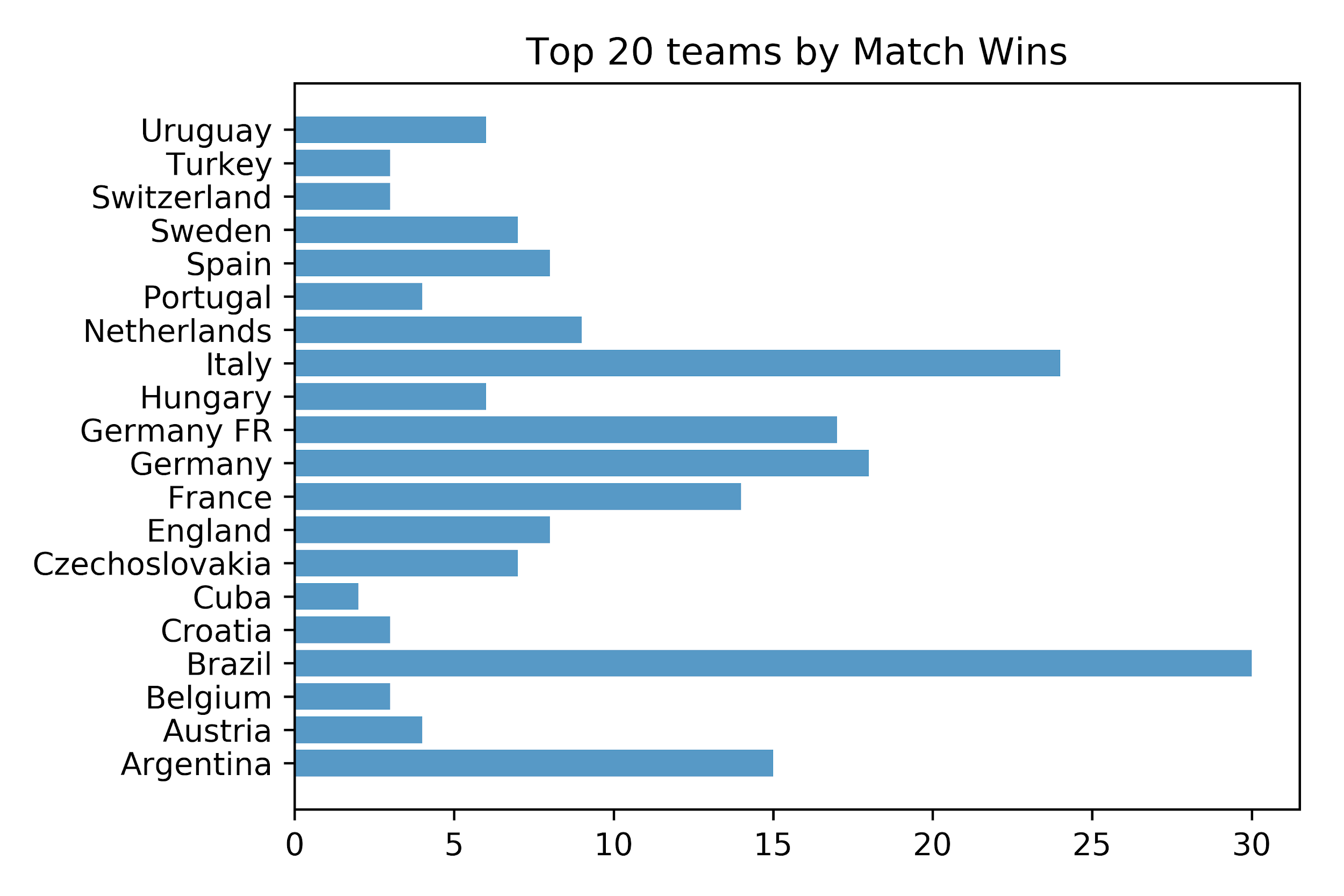
The following charts and plots show the number of wins and/ or win percentages of the above variable. This allows for a visual understanding of the significance of each variable as a predictor of winning.

Descriptions of all variables discussed below have been provided in the *Preparation* section under *Data Preparation*

## **Variables**

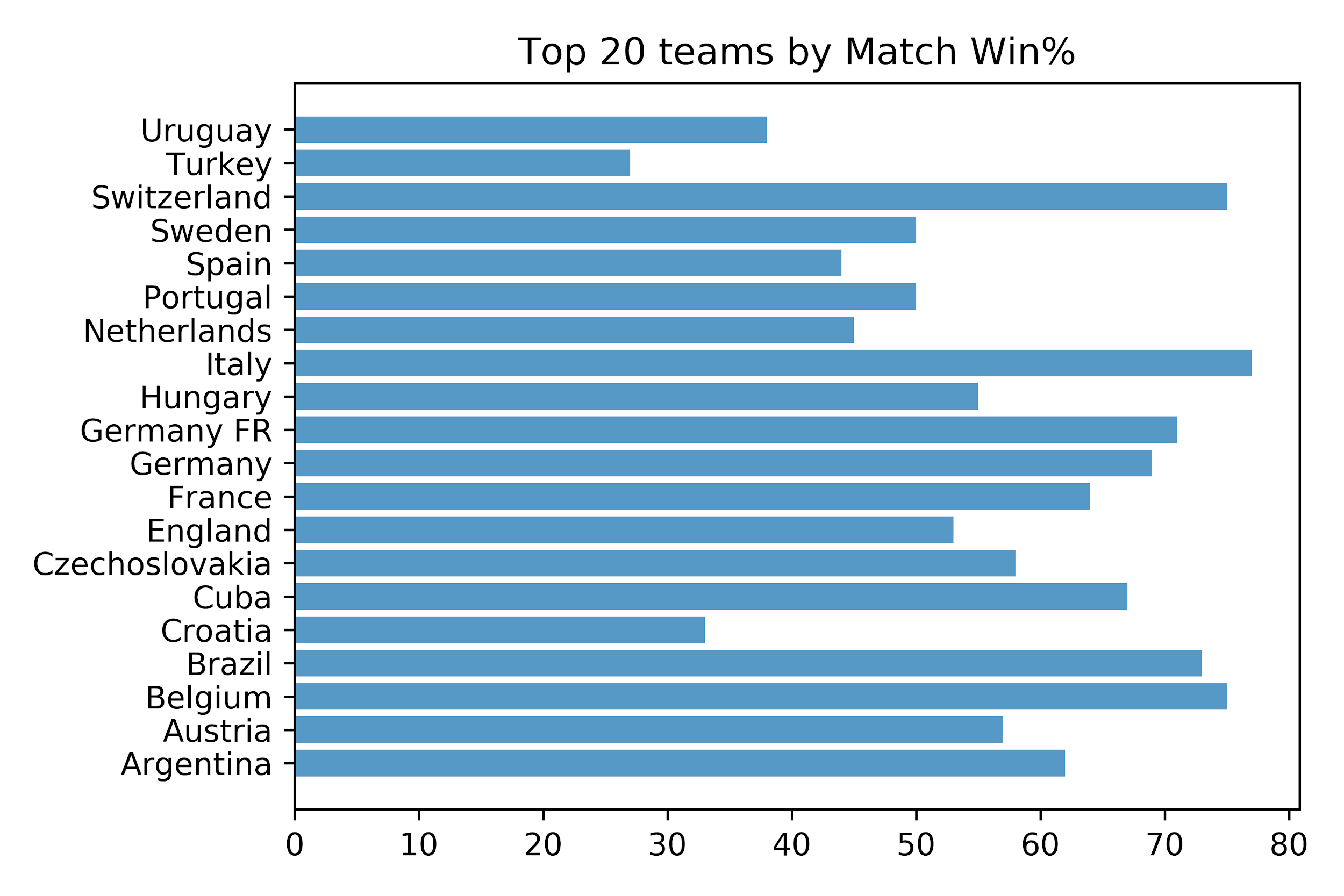
Number of Wins

The “Number of Wins” variable attempts to determine whether teams that have a higher number of elimination round wins tend to win more.



This plot shows the top 20 teams with the highest number of match wins. As we can see, some of the soccer giants, such as Brazil, Germany and France are included.

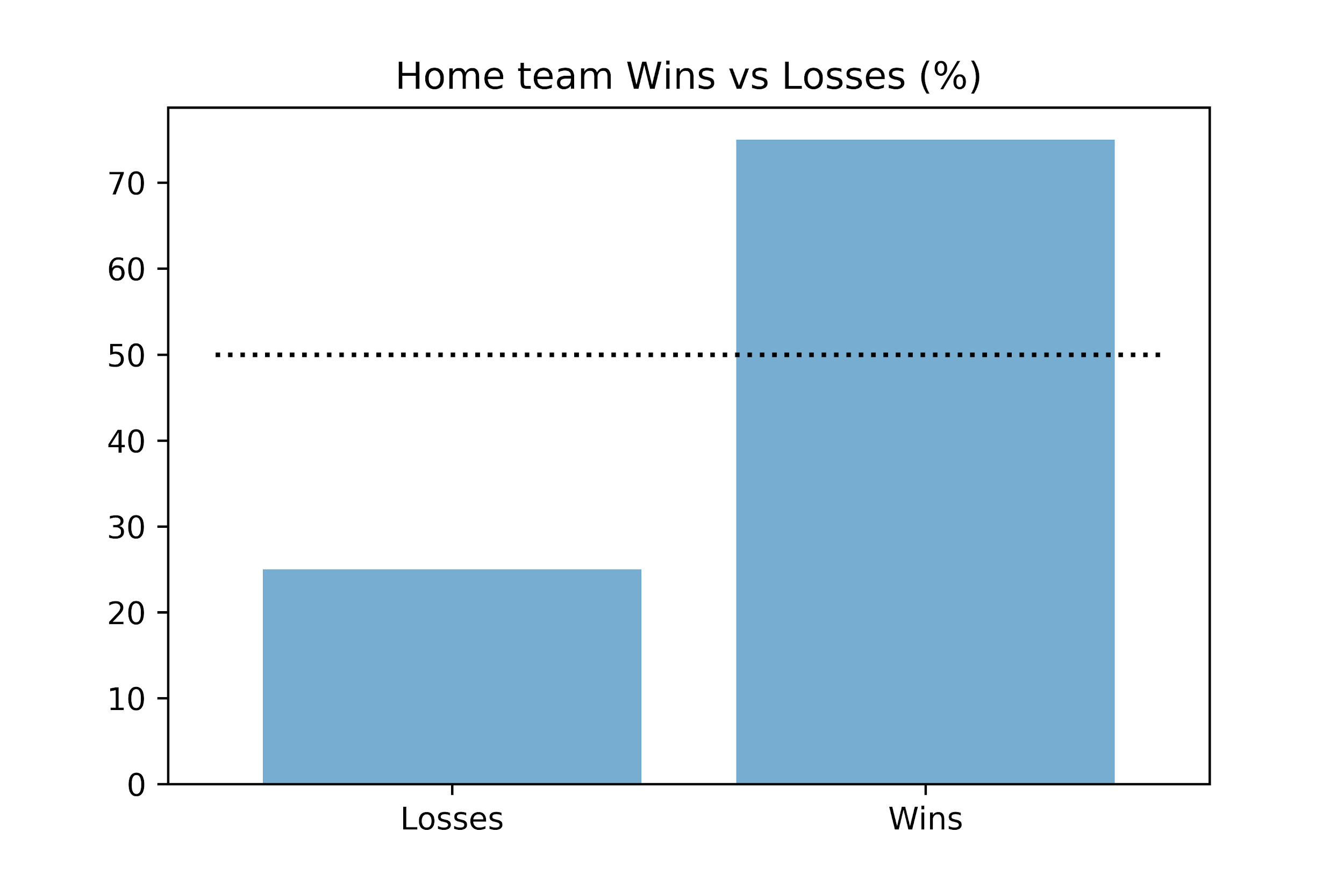
Win/ Match %

The “Win/ Match %” variable aims to determine whether teams with higher win to loss ratios have a higher chance of winning their matches.

This plot shows the top 20 teams with the highest wins to losses ratios. We can see that there are many teams included that often don’t perform well at the World Cup, have never won it or who often do not qualify. This shows that match win percentage may not necessarily be indicative of stronger teams.

Home Team

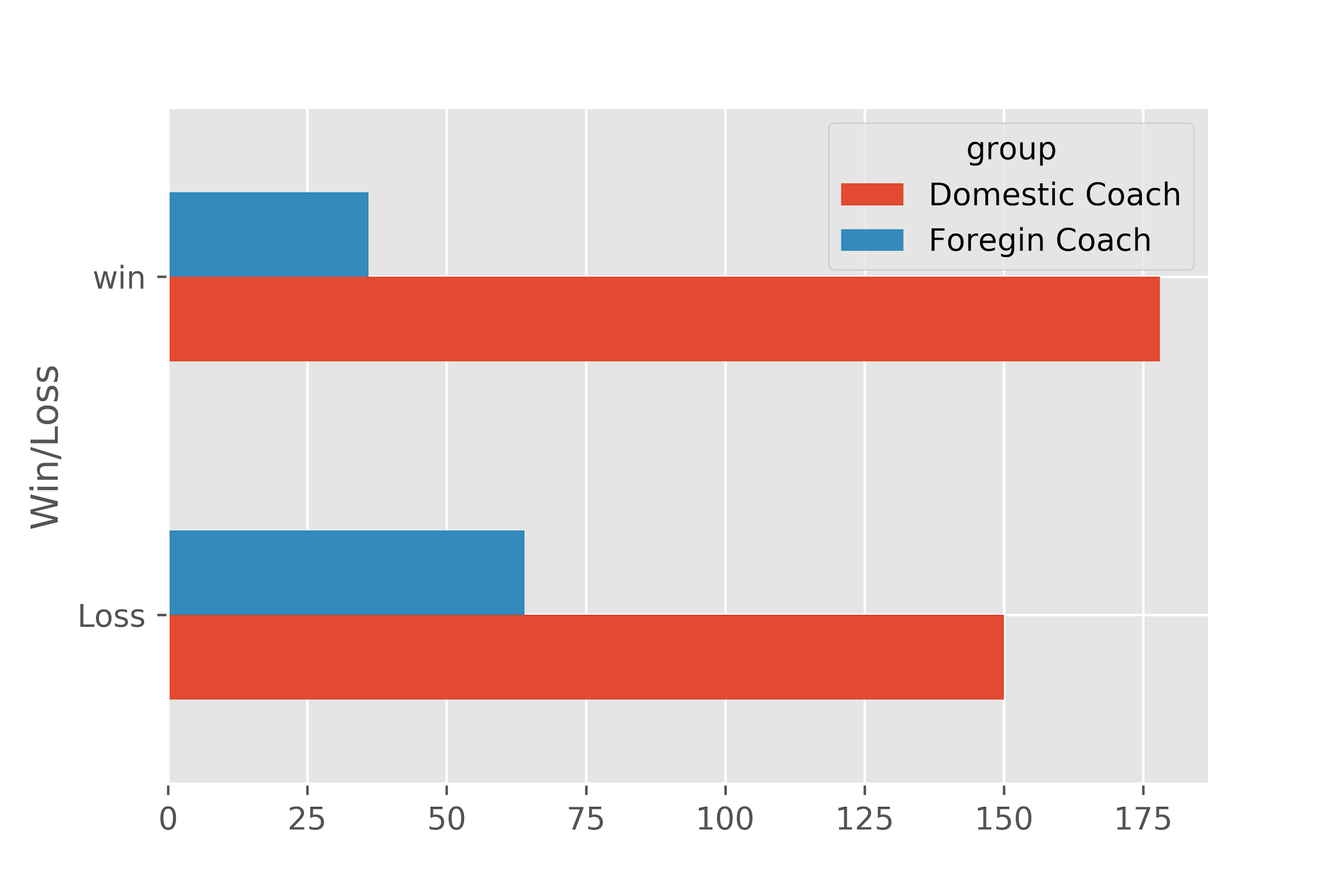
The “Home Team” variable attempts to show if there is an advantage associated with being labeled as the home team.

From this plot we can see that there is a significant advantage to being the home team. The home team tends to win over 70% of the time.

|  |  |  |
| --- | --- | --- |
|  | Wins | Losses |
| Home | 161 | 53 |
| Away | 53 | 161 |

Foreign Coach

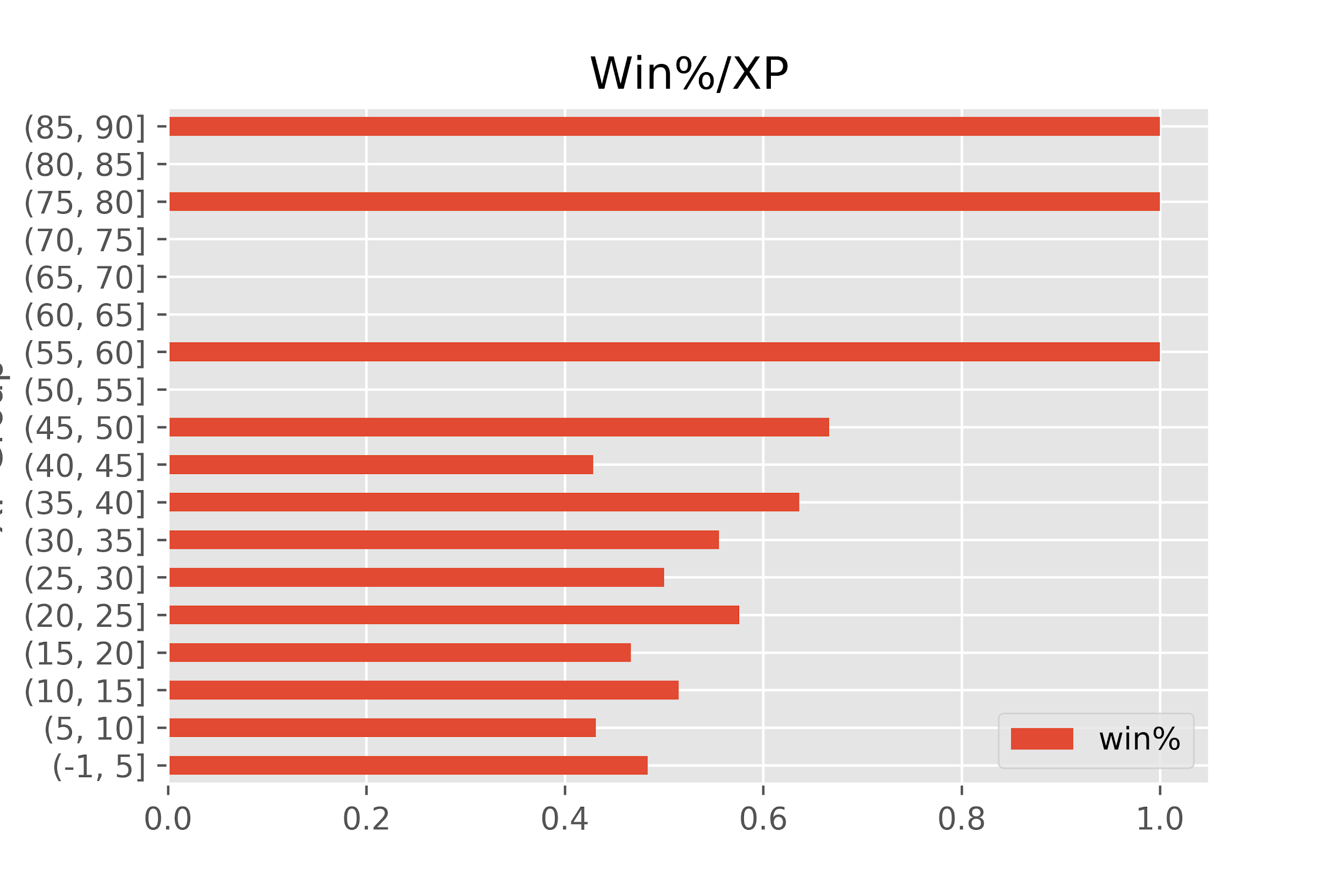
The “Foreign Coach” variable analyses the nationality of each coach and determines whether there is an advantage to a team having a coach of the same nationality.

This plot shows that domestic coaches win more games than they lose while foreign coaches lose more games than they win.

|  |  |  |
| --- | --- | --- |
|  | Wins | Losses |
| Foreign | 36 | 64 |
| Domestic | 178 | 150 |

XP

The “XP” or “experience” variable attempts to show whether teams with higher levels of experience have a higher likelihood of winning.



From this bar plot we can see a slight trend that as team’s experience increases, the probability of winning also increases. The bins on the y-axis represent the group of experience associated with teams in those bins.

## **Logistic regression**

Because the outcome of the elimination round match is either a win or a loss, a logistic regression would be the most suitable predictive model. The goal of the model would be to determine which variables would have be best at predicting whether a given team would win a given match.

The challenges with applying a logistic regression include: current skill set in the team would not support a multivariable logistic regression or analysis of significance of the model, and the fact that the outcome of one team in a given match is perfectly and inversely correlated with the outcome of the other team in that match.

The following steps were taken to address each of the challenges above.:

* Multivariable logistic regression:
  + We analyzed the impact of the variables on the outcome of the match one at a time. Should management wish to analyze the impact of the combination of variables on the outcome of a match, T9C recommends engaging the team in a year’s time.
* Analysis of Significance and correlation of team outcomes:
  + To analyze the success rate of the model, the team chose to build the model for all matches up to and including the 2010 world cup, apply that model to the outcome of the sixteen games in the elimination round of the 2014 world cup and measure the accuracy. To eliminate the correlation of team outcomes, one team per match was chosen at random and a logistic regression model was performed. This was repeated 1000 times, essentially creating 1000 training data sets and 1000 models. Each one of the models was then tested for accuracy against the 2014 World Cup. Finally, the accuracy results were averaged for each set of regressions against each variable.
  + To determine a “good” average rate of accuracy, we created a simulation to evaluate the accuracy rate of a model consisting of sixteen random coin flips in predicting the outcome of 16 games, also simulated by a series of random coin flips. The simulation was run 1000 times. Based on the distribution of the results, the team chose a cutoff point of one standard deviation above the mean to determine the confidence level for analyzing the model. The chosen cut off accuracy rate was 62.5% with a confidence level of 84%.

Because we did not settle on a single model for each variable, the results of the analysis are a measure of the suitability of the logistic regression model for a particular variable.

## **Observations**

The below chart shows the result of the regression analysis on all variables over 1,000 iterations. The table outlines the advantageous characteristic, and the outcome shows the measure of how accurate each variable was as a predictor of a team to win a match during the 2014 World Cup.

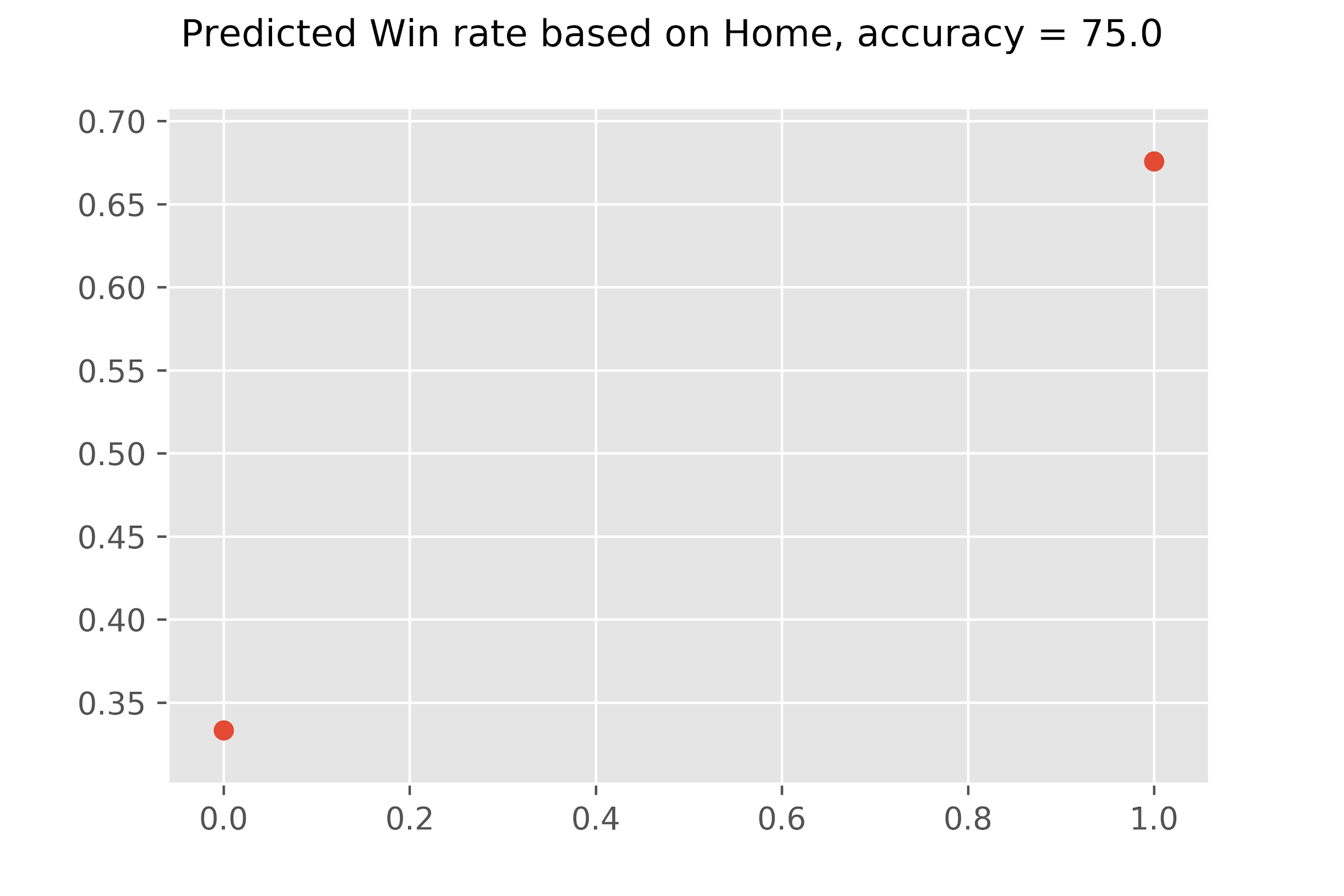
Based on the explanation above, only the first three variables (Home Team, Wins and Games Played) are appropriate predictors for the model, as they are above the cutoff accuracy of 62.5%.

|  |  |  |
| --- | --- | --- |
| Variable | Average Model Prediction Accuracy | Advantage |
| Home | 75 | Home Team |
| Wins | 69.960396 | Higher Wins |
| Games Played | 67.662466 | More Games Played |
| Losses | 60.70477 | Fewer Losses |
| Foreign Coach | 60.422142 | Domestic Coach |
| Win/Loss% | 59.692169 | Higher Ratio |
| Goals For/Match | 56.094509 | More Goals Scored |
| IH | 54.677768 | Fewer Substitutions |
| XP | 53.652565 | Higher Experience |
| All Cards | 53.288929 | Fewer Cards |
| Y | 52.681368 | Fewer Cards |
| R | 50.606661 | Fewer Cards |
| Foreign Soil | 50.377138 | Home Soil |
| Goals Against/Match | 50.070207 | Fewer Goals |
| All Substitutions | 50.049505 | Fewer Substitutions |
| I | 50.008101 | Fewer Substitutions |

## **Top Three Variables**

The following plots illustrate the analysis performed for the top three variables:

Home Team

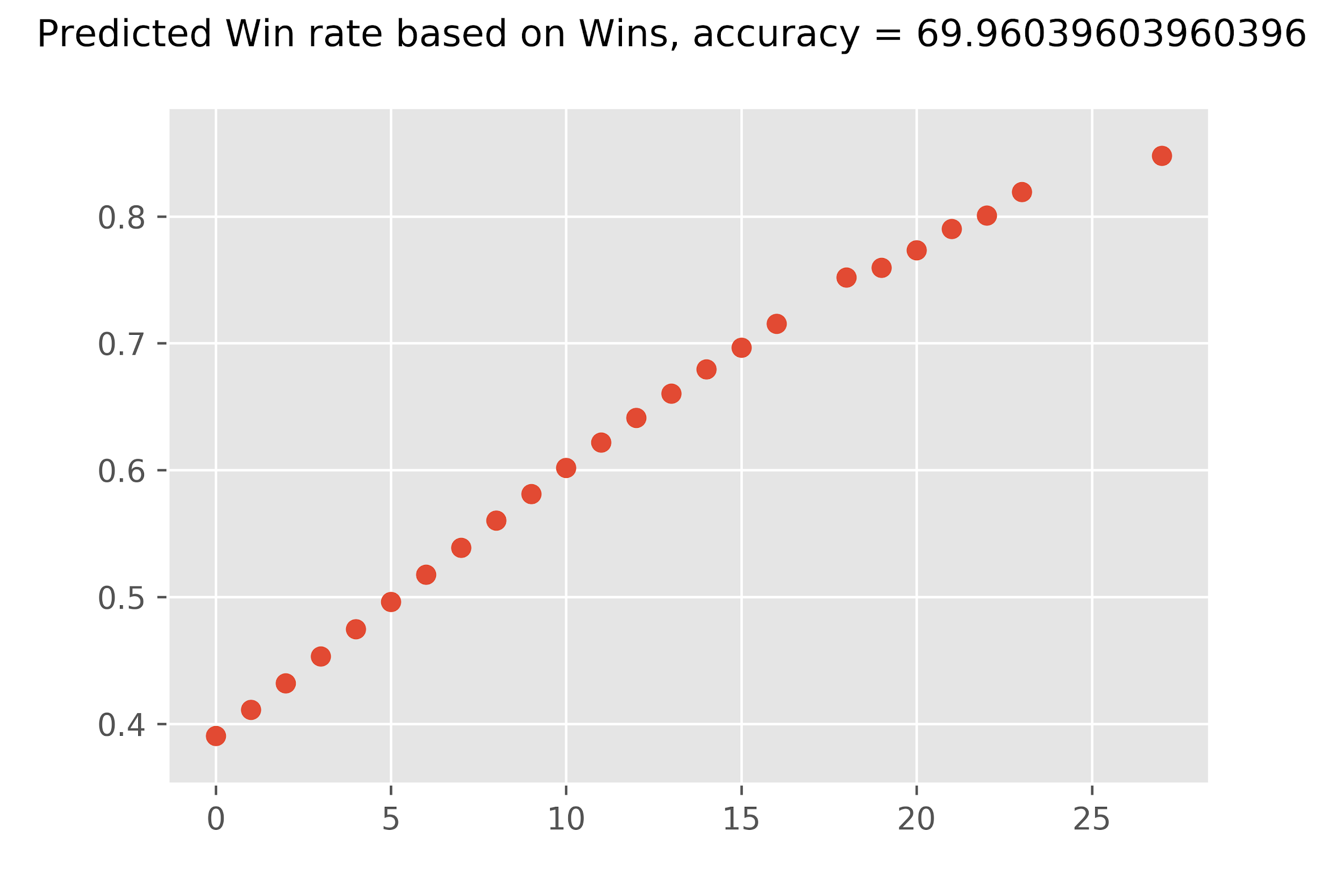


The “Home” variable resulted in being the top predictor for winning, at 75% accuracy. This means the being labeled as the “Home Team” offers a significant advantage.

There are two concerns regarding this outcome however. Firstly, preliminary research indicates that this label is randomly assigned before the start of any tournament. Secondly, every iteration of the model resulted in this variable being accurate at 75%.

## 

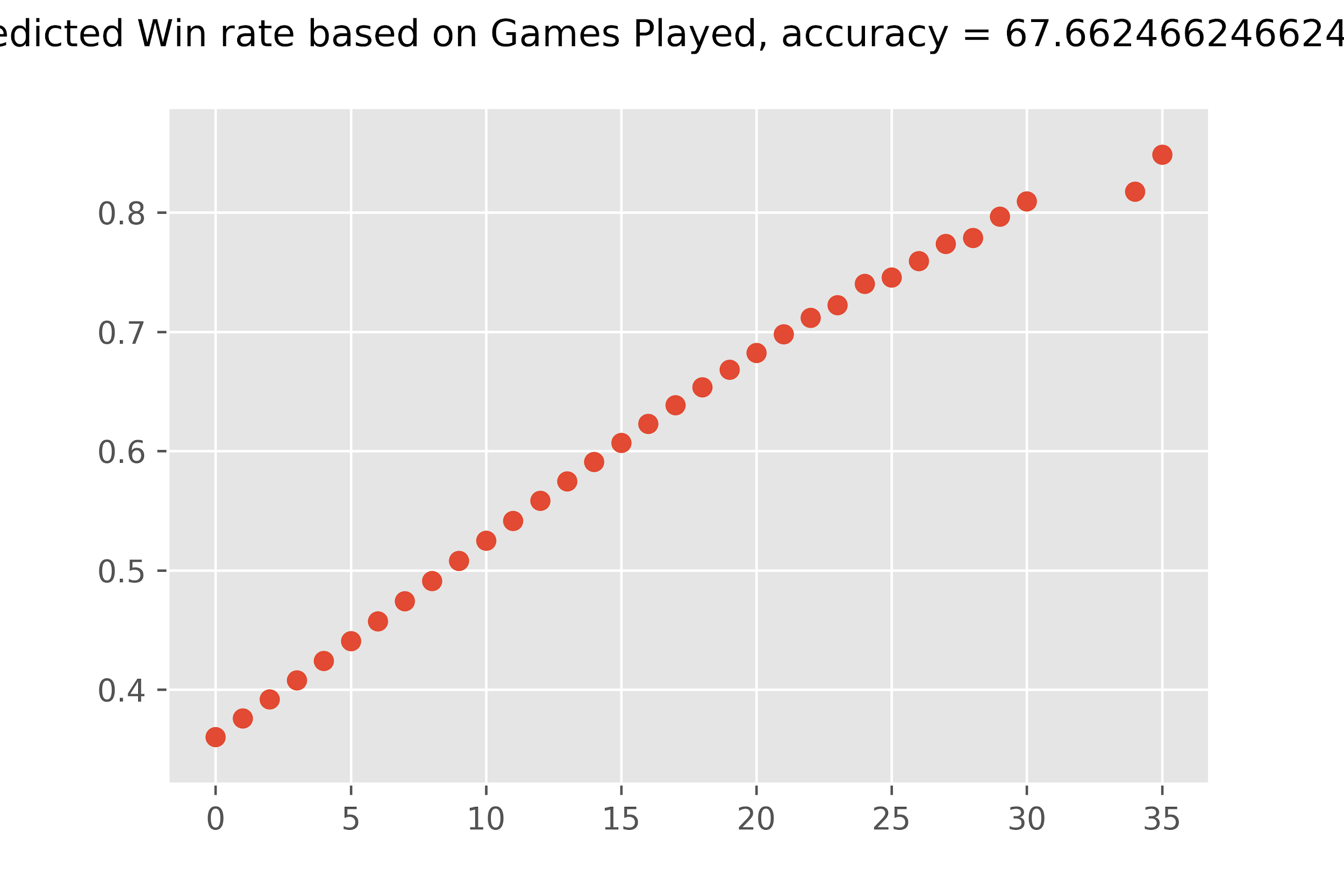
Previous Wins

Having more previous wins was determined to be a predictor of winning with 69.96% accuracy. This result remains an intuitive conclusion as general expectation would be that more previous wins results in higher likelihood of winning.

It was found that a team that has won 6 or more World Cup elimination round matches prior to the current match had a greater chance of winning than a team with 5 or fewer wins.

The plot shows a strong linear relationship between the variable and likelihood of winning.

Previous Matches



The number of games a team has played is a predictor of winning at 67.66%.

It was found that a team that has played 9 or more World Cup elimination round matches prior to the current match has a greater chance of winning than a team with 8 or fewer matches.

This plot also shows a strong linear relationship.

# Conclusions

Single variables do not provide enough information to confidently predict the outcome of a match.

By far the best single variables for predicting whether a team would win was whether the team was considered the home team. However, this could only be used to accurately predict the winner of a team 75% of the time.

The next best variable was the number of wins the nation had accumulated over all the world cups.

T9C recommends the following further steps for analysis.

* Including 2018 World Cup data
* Including player data from league games
* Analyzing the head-to-head win percentages.
  + For example, how likely will a team with 20 wins defeat a team with only ten wins.
* Doing multivariable logistic regression combining Home, Foreign Coach, Win/Loss%, and Wins
  + These four variables are the most predictive variables.
* Perform a Clustering or Classification analysis on the data to identify cohorts and then analyze how teams from one cohort perform against teams from another cohort
  + Here the analysis would answer questions like this. Suppose team A is an away team with over nine wins, a foreign coach, and a 50% win record. Suppose team B is a home team with fewer than nine wins, a domestic coach, and a 67% win record. How likely will team A defeat team B?

# Appendix

## Data Samples

The following tables are the heads of the dataframes used in the analysis. The first three were the base sets and last two were created.

Please note that due to space not all columns are shown, only the variables relevant to the analysis are displayed.

Cups table

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Country | Winner | Runners-Up | Third | Fourth | GoalsScored | QualifiedTeams | MatchesPlayed | Cups Attendance |
| 1930 | Uruguay | Uruguay | Argentina | USA | Yugoslavia | 70 | 13 | 18 | 590549 |
| 1934 | Italy | Italy | Czechoslovakia | Germany | Austria | 70 | 16 | 17 | 363000 |
| 1938 | France | Italy | Hungary | Brazil | Sweden | 84 | 15 | 18 | 3757000 |
| 1950 | Brazil | Uruguay | Brazil | Sweden | Spain | 88 | 13 | 22 | 1045246 |
| 1954 | Switzerland | Germany FR | Hungary | Austria | Uruguay | 140 | 16 | 26 | 768607 |

Matches table

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | Datetime | Stage | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | Win conditions | Attendance | RoundID | MatchID |
| 1930 | 13 Jul 1930 - 15:00 | Group 1 | France | 4 | 1 | Mexico |  | 4444 | 201 | 1096 |
| 1930 | 13 Jul 1930 - 15:00 | Group 4 | USA | 3 | 0 | Belgium |  | 18346 | 201 | 1090 |
| 1930 | 14 Jul 1930 - 12:45 | Group 2 | Yugoslavia | 2 | 1 | Brazil |  | 24059 | 201 | 1093 |
| 1930 | 14 Jul 1930 - 14:50 | Group 3 | Romania | 3 | 1 | Peru |  | 2549 | 201 | 1098 |
| 1930 | 15 Jul 1930 - 16:00 | Group 1 | Argentina | 1 | 0 | France |  | 23409 | 201 | 1085 |

Players table

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| RoundID | MatchID | Team Initials | Coach Name | Line-up | Shirt Number | Player Name | Position | Event | Year | PlayerID |
| 201 | 1096 | FRA | CAUDRON Raoul (FRA) | S | 0 | Alex THEPOT | GK |  | 1930 | FRAAlex THEPOT0 |
| 201 | 1096 | MEX | LUQUE Juan (MEX) | S | 0 | Oscar BONFIGLIO | GK |  | 1930 | MEXOscar BONFIGLIO0 |
| 201 | 1096 | FRA | CAUDRON Raoul (FRA) | S | 0 | Marcel LANGILLER |  | G40' | 1930 | FRAMarcel LANGILLER0 |
| 201 | 1096 | MEX | LUQUE Juan (MEX) | S | 0 | Juan CARRENO |  | G70' | 1930 | MEXJuan CARRENO0 |
| 201 | 1096 | FRA | CAUDRON Raoul (FRA) | S | 0 | Ernest LIBERATI |  |  | 1930 | FRAErnest LIBERATI0 |

Events table

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Coach Name | Player Name | Event | Year | Home Team Name | Home Team Goals | Away Team Goals | Away Team Name | Win conditions | Attendance | Country | Winner | GoalsScored | Cups Attendance |
| CAUDRON Raoul (FRA) | Alex THEPOT |  | 1930 | France | 4 | 1 | Mexico |  | 4444 | Uruguay | Uruguay | 70 | 590.549 |
| LUQUE Juan (MEX) | Oscar BONFIGLIO |  | 1930 | France | 4 | 1 | Mexico |  | 4444 | Uruguay | Uruguay | 70 | 590.549 |
| CAUDRON Raoul (FRA) | Ernest LIBERATI |  | 1930 | France | 4 | 1 | Mexico |  | 4444 | Uruguay | Uruguay | 70 | 590.549 |
| LUQUE Juan (MEX) | Rafael GARZA |  | 1930 | France | 4 | 1 | Mexico |  | 4444 | Uruguay | Uruguay | 70 | 590.549 |
| LUQUE Juan (MEX) | Hilario LOPEZ |  | 1930 | France | 4 | 1 | Mexico |  | 4444 | Uruguay | Uruguay | 70 | 590.549 |

## 

Teams table

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Team | Win | Home | Goals | Foreign Coach | Y | RSY | R | I | IH | All Cards | Country | Wins | Losses | Games Played | WinLoss% | All Substitutions | Foreign Soil | XP | Goals For/Match |
| Argentina | TRUE | TRUE | 6 | FALSE | 0 | 0 | 0 | 0 | 0 | 0 | Uruguay | 0 | 0 | 0 | 1 | 0 | FALSE | 0 | 1 |
| Uruguay | TRUE | TRUE | 6 | FALSE | 0 | 0 | 0 | 0 | 0 | 0 | Uruguay | 0 | 0 | 0 | 1 | 0 | TRUE | 0 | 1 |
| Uruguay | TRUE | TRUE | 4 | FALSE | 0 | 0 | 0 | 0 | 0 | 0 | Uruguay | 1 | 0 | 1 | 1 | 0 | TRUE | 10 | 6 |
| USA | FALSE | FALSE | 1 | FALSE | 0 | 0 | 0 | 0 | 0 | 0 | Uruguay | 0 | 0 | 0 | 1 | 0 | FALSE | 0 | 1 |
| Yugoslavia | FALSE | FALSE | 1 | FALSE | 0 | 0 | 0 | 0 | 0 | 0 | Uruguay | 0 | 0 | 0 | 1 | 0 | FALSE | 0 | 1 |

## 

## Sample Code

This block of code reads through the matches dataframe and creates a teams dataframe. This was the most important part of all our data preparation.

|  |
| --- |
| def beats(teama: dict, teamb: dict, i: int, index: int, home: bool) -> None:  """  teama beats teamb. home = teama == the home team.    Precondition: teama and teamb both contain Name and Initials  """  winners[index]['Team'] = teama['Name']  winners[index]['Team Initials'] = teama['Initials']  winners[index]['Opponent Name'] = teamb['Name']  winners[index]['Opponent Initials'] = teamb['Initials']  losers[index]['Team'] = teamb['Name']  losers[index]['Team Initials'] = teamb['Initials']  losers[index]['Opponent Name'] = teamb['Name']  losers[index]['Opponent Initials'] = teamb['Initials']  if home:  winners[index]['Goals For'] = matches.loc[i]['Home Team Goals']  winners[index]['Goals Against'] = matches.loc[i]['Away Team Goals']  losers[index]['Goals For'] = matches.loc[i]['Away Team Goals']  losers[index]['Goals Against'] = matches.loc[i]['Home Team Goals']  else:  winners[index]['Goals For'] = matches.loc[i]['Away Team Goals']  winners[index]['Goals Against'] = matches.loc[i]['Home Team Goals']  losers[index]['Goals For'] = matches.loc[i]['Home Team Goals']  losers[index]['Goals Against'] = matches.loc[i]['Away Team Goals']   stages = ['Final', 'First round', 'Match for third place', 'Play-off for third place', 'Preliminary round', 'Quarter-finals', 'Round of 16', 'Semi-finals', 'Third place'] #elim rounds only winners = {} losers = {} index = 0 problems = {}  for i in matches.index:  winners[index] = Series(matches.loc[i])  losers[index] = Series(matches.loc[i])   hometeam = {"Name": matches.loc[i]['Home Team Name'],   "Initials": matches.loc[i]['Home Team Initials']}  awayteam = {"Name": matches.loc[i]['Away Team Name'],   "Initials": matches.loc[i]['Away Team Initials']}   if matches.loc[i]['Stage'] in stages:  # home team beats away team  if matches.loc[i]['Home Team Goals'] > matches.loc[i]['Away Team Goals']:  beats(hometeam, awayteam, i, index, True)     # away team beats home team  elif matches.loc[i]['Home Team Goals'] < matches.loc[i]['Away Team Goals']:  beats(awayteam, hometeam, i, index, False)     # home team beats away team   elif hometeam['Name'] in matches.loc[i]['Win conditions']:  beats(hometeam, awayteam, i, index, True)     elif awayteam['Name'] in matches.loc[i]['Win conditions']:  beats(awayteam, hometeam, i, index, False)     # away team beats home team   else:  problems[index] = Series(matches.loc[i])    index +=1 for i in winners:  winners[i]['Win'] = True  losers[i]['Win'] = False  teams = pandas.concat([DataFrame.from\_dict(winners, orient='index'), DataFrame.from\_dict(losers, orient='index')], axis=0, ignore\_index=True) |

## 

## References

Source data:

<https://www.kaggle.com/abecklas/fifa-world-cup>

Supplementary Data:

<https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_own_goals>)

Logistic Regression Modelling Code:

Python for Data Analysis, 2nd Edition: Data Wrangling with Pandas, NumPy, and IPython, William McKinney, O'Reilly Media, October 2017

1. While Wikipedia is generally considered to be an unreliable secondary source of data, the team agreed that for the purposes of this analysis it was sufficiently reliable. [↑](#footnote-ref-1)