# 2023 Rugby World Predictor

Peter Meihuizen

2023-06-23

## 1. Introduction

Machine learning has opened up the possibility to predict certain unknown outcomes by building models which can learn from histroical data and predict an outcome based on what has been the reuslt in the past. For this project I will attempt to predict the winner of the 2023 Rugby World Cup using a random forest machine learning model. This requires me to train the model on an existing data set in order to estimate the results for the particular fixtures which will be played at the 2023 World Cup. I will be extracting the necessary information for my training model from a historical data set which gives the result of every international rugby match played between the historically best 10 international men's rugby teams in the world, as well as from a online websites which tells me the ranking of each team before they played in a particular World Cup. The fixtures for the 2023 World Cup have already been released, with it consisting of group stages and knockout matches. Using the past match results from previous years, as well as the ranking of each team and their success at previous World Cups, as features, I will predict who will win every world cup game between the top 10 teams in the world, who will progress to each playoff martches, and ultimately who will win the 2023 Rugby World Cup Final.

## 2. Descriptive Statistics and Creating the Training Data Set

In order to conduct my analysis, I need to import the data set from excel format. This data set includes all matches which have been played since 1984, 3 years before the start of the first Rugby World Cup. Each observation represents a different match between two teams. The relevant variables which are given include the date of the match, the home team, the away team, the home teams score, the away teams score, whether the match was played at a nuetral venue and whether the match the match was palyed at a World Cup.

I decided to generate 4 new columns, in order to identify the winner of each game, the loser of each game, the points difference of each game and the winner of the World Cup for each World Cup final matches. The sign of the points difference value indicates the winner of the match, with positive values indicating the home team won the match and negative values indicating the away team as the winner.

I can now see the winners of each world cup up to date. As can be seen New Zealnd won the first World Cup Final in 1987 whilst South Africa won the most recent World Cup final (2019). Both of these teams have each won the Rugby World Cup 3 times. Otherwise Australia have won 2 World Cups (1991 and 1999) whilst England won 1 title in 2003.

In order to work out World Cup cycle specific values, I created a new variable specifying which World Cup cycle each match fell under in order to group the data set by each World Cup cycle. This enabled me to determine the win percentage for each world cup winning team for the 4 years leading up to their respective World Cup victories. In order to do this I looked at the number of wins over the total number of games played for each team in each over each 4 year period preceding the beginning of the world cup.

|   | competition | WC_winner |
|---|---|---|
| 1 | 1987 Rugby World Cup Final | New Zealand |
| 2 | 1991 Rugby World Cup Final | Australia |
| 3 | 1995 Rugby World Cup Final | South Africa |
| 4 | 1999 Rugby World Cup Final | Australia |
| 5 | 2003 Rugby World Cup Final | England |
| 6 | 2007 Rugby World Cup Final | South Africa |
| 7 | 2011 Rugby World Cup Final | New Zealand |
| 8 | 2015 Rugby World Cup Final | New Zealand |
| 9 | 2019 Rugby World Cup Final | South Africa |

Table 1: World Cup Winners

|   | WC_winner | WC_cycle | total_matches | total_wins | win_percentage |
|---|---|---|---|---|---|
| 1 | New Zealand | 2015 | 44 | 39 | 88.64 |
| 2 | England | 2003 | 35 | 30 | 85.71 |
| 3 | New Zealand | 2011 | 46 | 37 | 80.43 |
| 4 | New Zealand | 1987 | 18 | 13 | 72.22 |
| 5 | Australia | 1999 | 40 | 26 | 65.00 |
| 6 | South Africa | 2007 | 43 | 25 | 58.14 |
| 7 | South Africa | 2019 | 43 | 21 | 48.84 |
| 8 | Australia | 1991 | 23 | 11 | 47.83 |
| 9 | South Africa | 1995 | 22 | 9 | 40.91 |

Table 2: Best World Cup Winning Teams
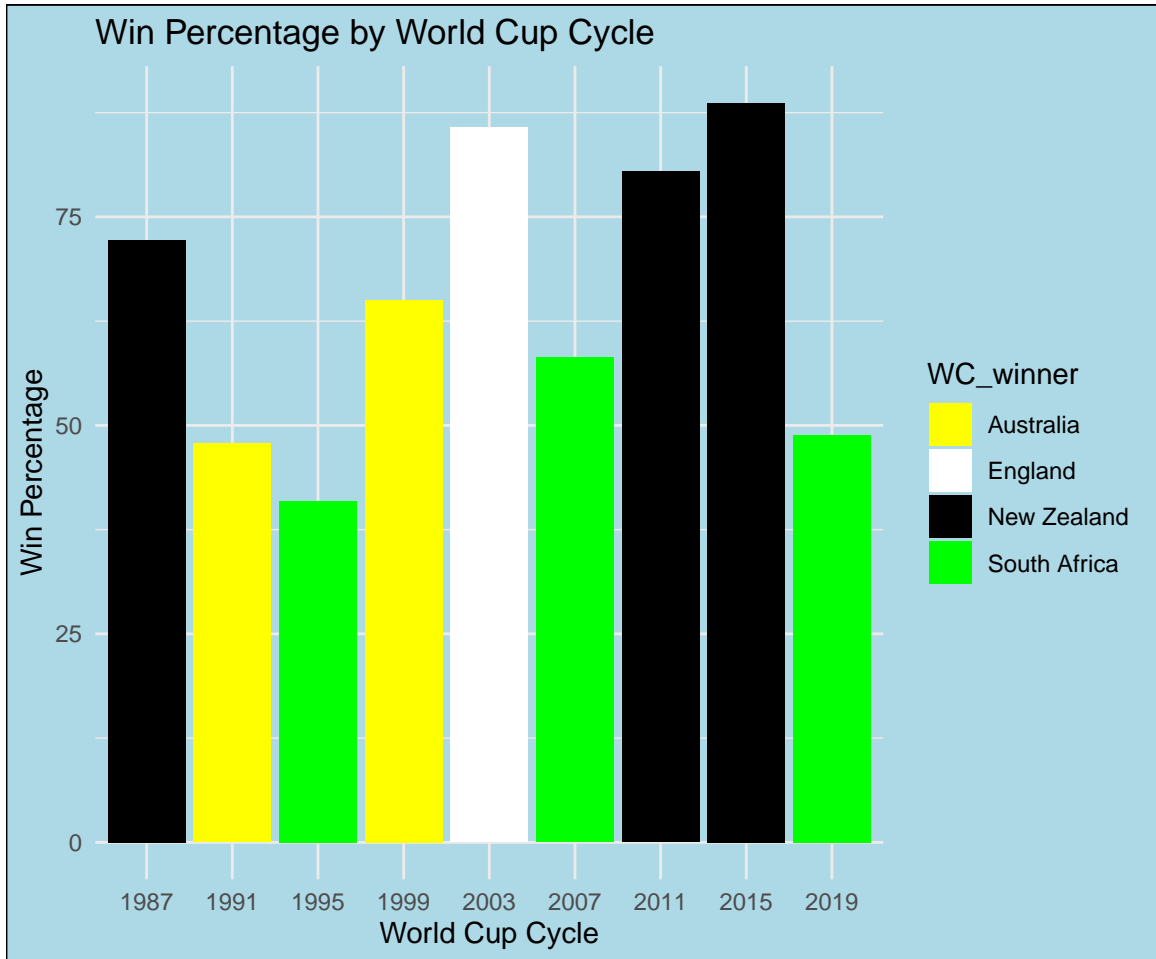
We can therefore see the respective win percentages of all World Cup winning teams against the over top 10 teams in the world in the 4 year cycle preceding each World Cup. Based on this information it can be seen that the New Zealand team which won the 2015 World Cup was had the best win percentage with an 89% win record. The second best was the England team that won the 2003 Rugby World Cup with an 86% win record, followed by the 2011 World Cup winning New Zealand team, who had an 80% win record. The worst performing team was the South African 1995 World Cup winning team who only managed to win 41% of their matched against top 10 opposition in the 4 years preceding their world cup campaign.

Next, I wanted to include the World rankings of each team which have been collected since the conclusion of the 2003 Rugby World Cup. In order to do this I scraped the data from Wikimedia page, showing the world ranking for every team for every week since after the 2003 Rugby World Cup. I turned the list which was scraped into a data frame.
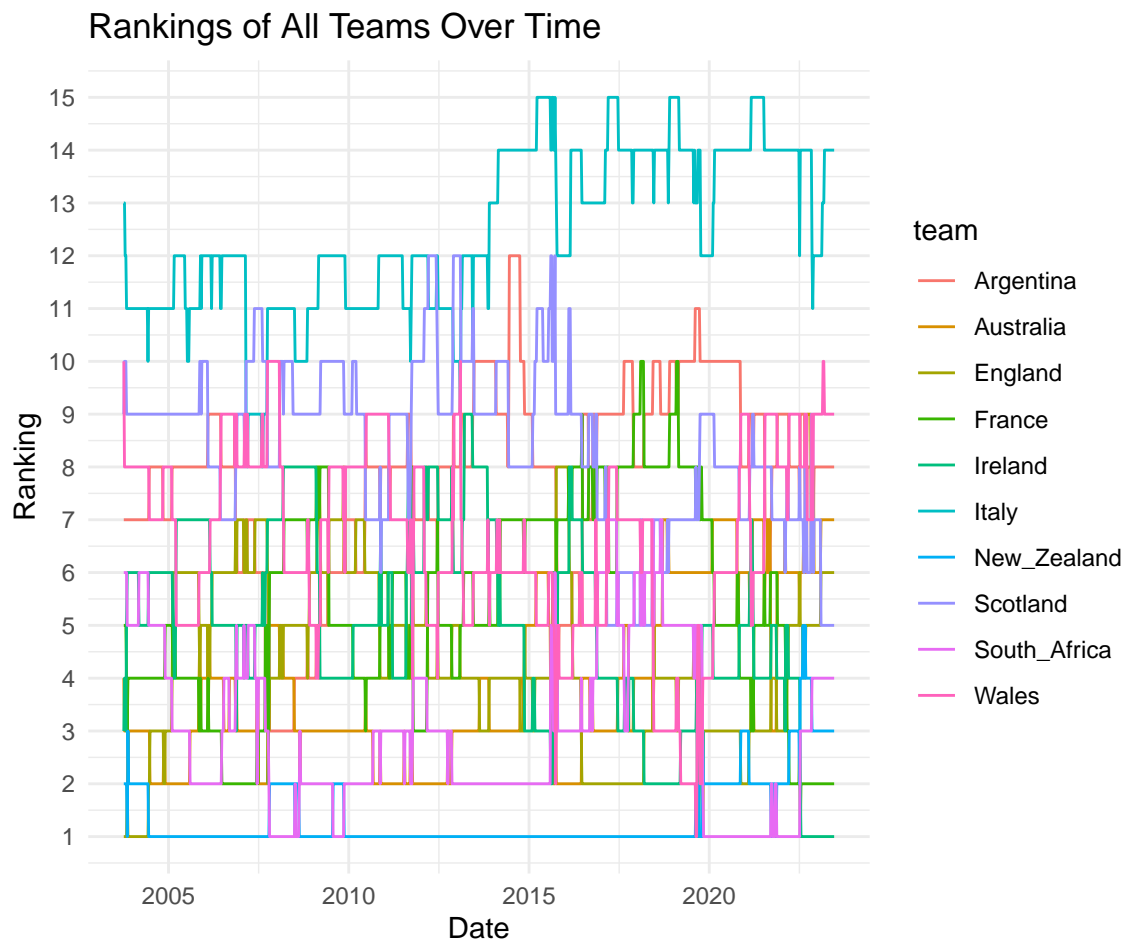


The graph represents the ranking of each team over the last 20 years with the lowet number (1) indicating the highest ranked team. I decided to decrease the number of teams in each graph so that they could be slightly more interpretable. I therefore decided to look at all the Southern Hemisphere teams (South Africa, New Zealand, Australia and Argentina) and Northern Hemisphere team (England, Ireland, France, Italy, Scotland and Wales) seperately.
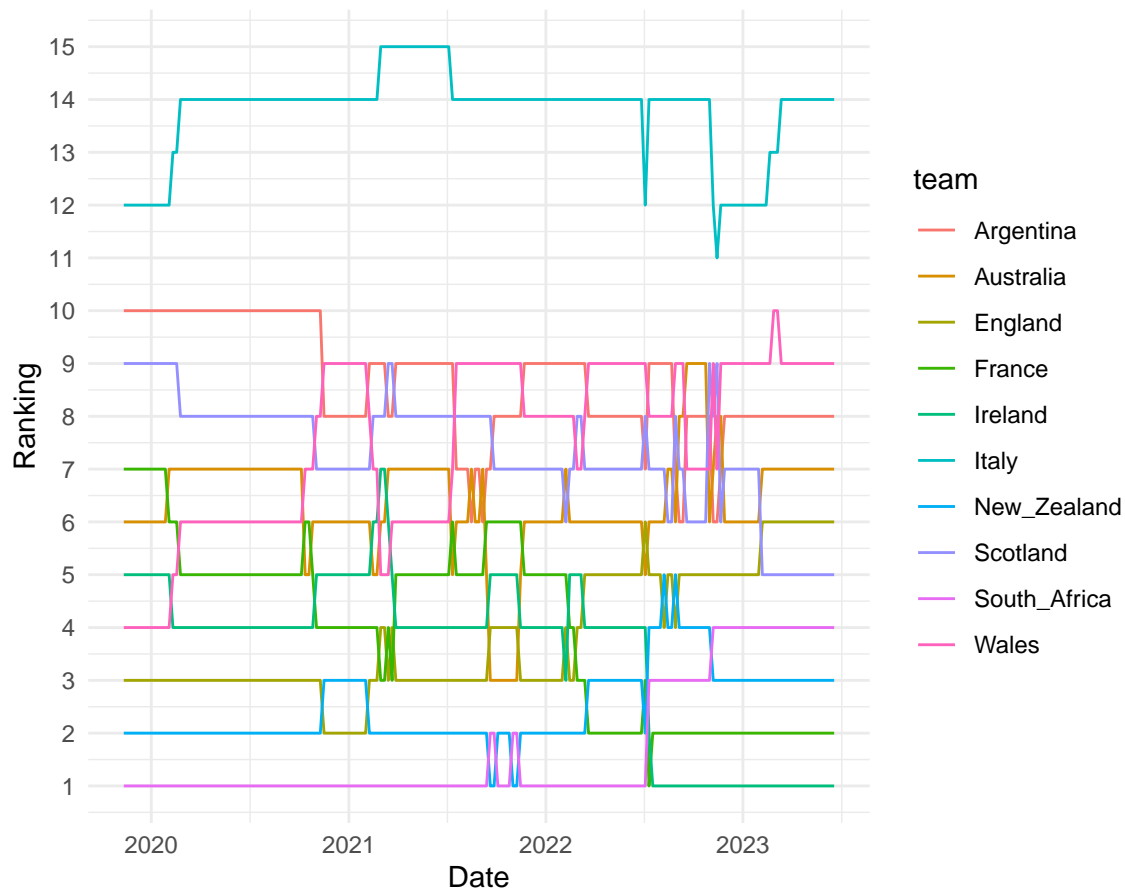
Rankings of Southern Hemisphere Teams Over Time

# Rankings of Northern Hemisphere Teams Over Time
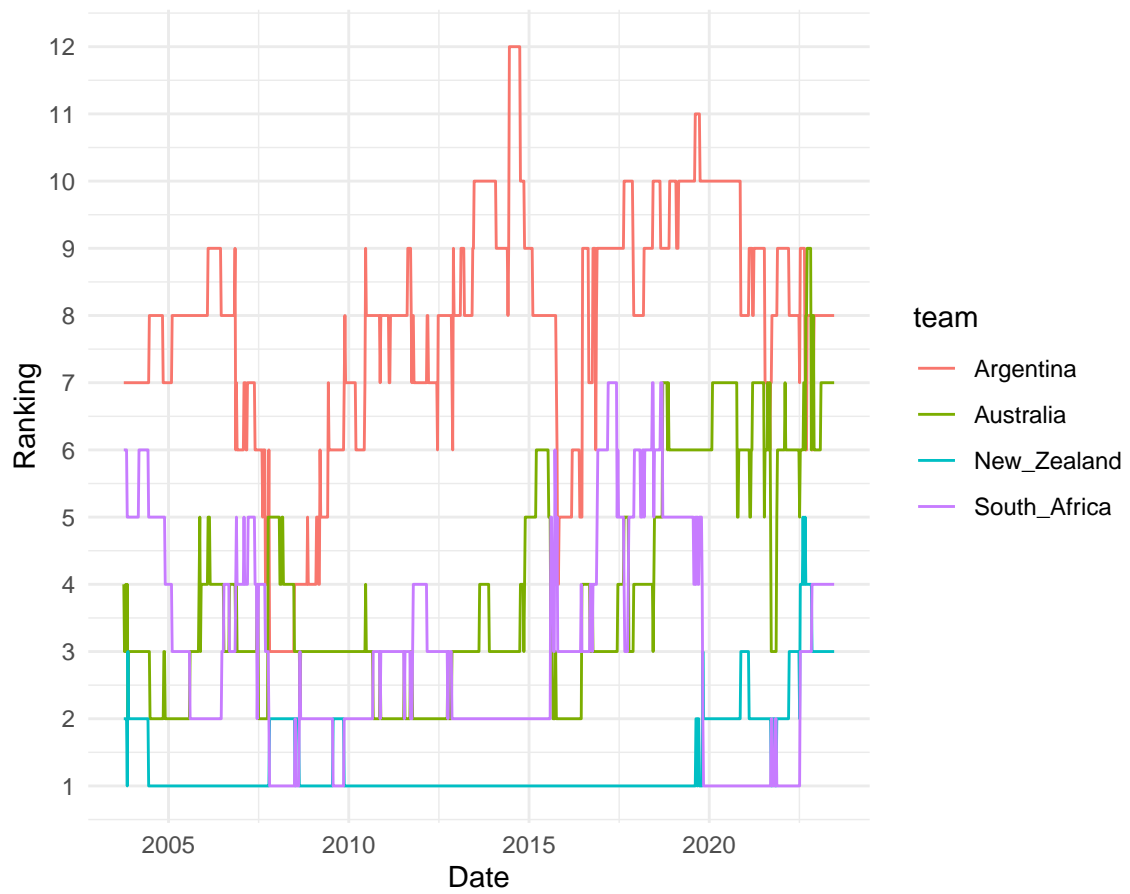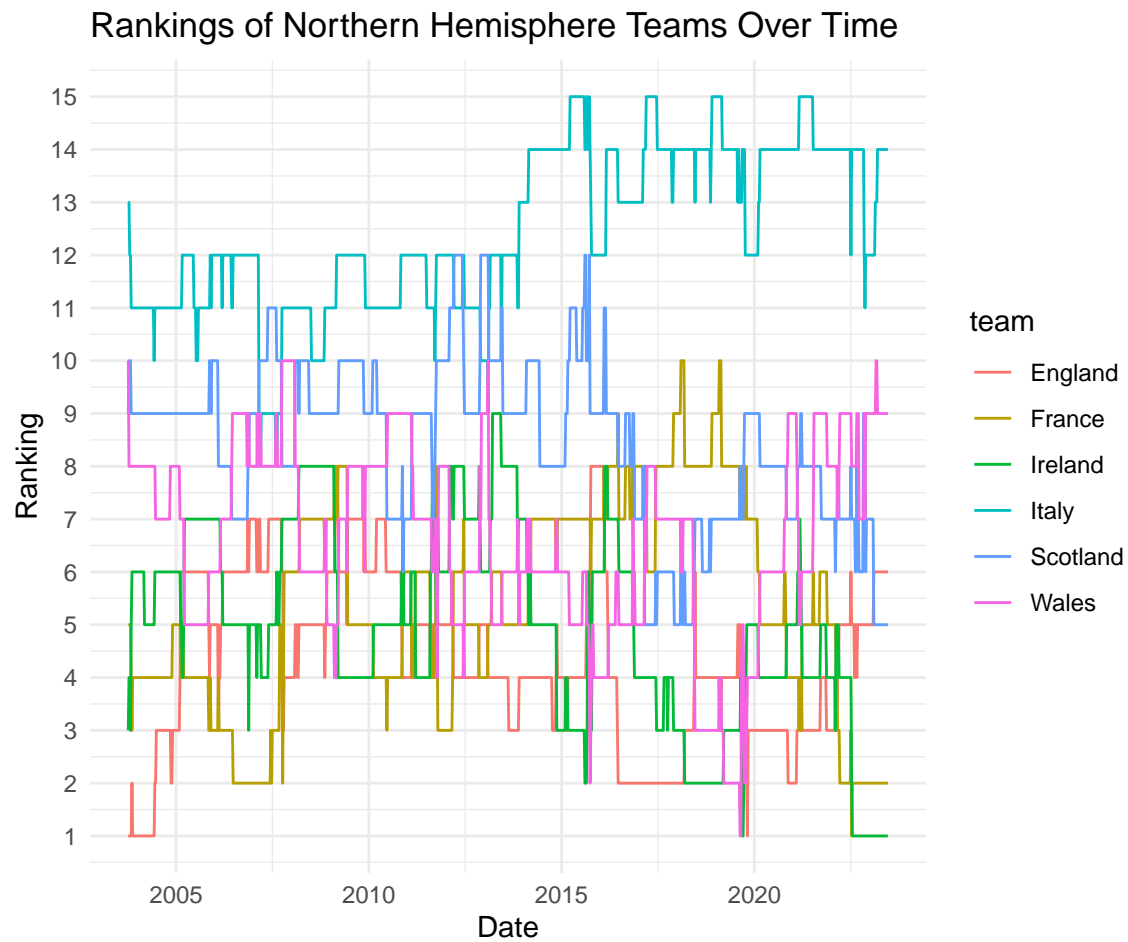


As can be seen the Southern Hemisphere teams have dominated the rankings historically, with New Zealand in particular holding the number 1 spot for the most time overall.

## Rankings of All Teams in 2023 RWC Cycle



The graph above shows the rankings progression for the last 4 years, representing the 2023 Rugby World Cup cycle. As can be seen, in this World Cup cycle South Africa have be the top team for the majority of the period, however Ireland has held the number 1 rank for the last year. The rankings stand as such as of the 12 of June 2023.

So as can be seen the rankings show Ireland to be ranked number 1 on the most recent available rankings. In second place is France followed by New Zealand and then South Africa. Italy are not even ranked in the top 10, however I belive this is due to them playing more top 10 teams in a year relative to other teams above them, therefore I feel they are harshly represented at number 14.

Next I wanted to look at the win percentage of each team against all other teams in the 4 year cycle before the world cup started. I did this so that I could import these win percentages into a data set showing all matches which have been played at previous World Cups between the top 10 teams. This could them be used for the training set of my model. When restricting the main data frame to only World Cup matches I saw that there have been 132 matches played between these teams at previous World Cups.

```
## \begin{table}[ht]
## \centering
## \begin{tabular}{rllrrr}
##   \hline
##  & WC\_winner & WC\_cycle & total\_matches & total\_wins & win\_percentage \\
##   \hline
## 1 & New Zealand & 2015 &  44 &  39 & 88.64 \\
```

```
##    2 & England & 2003 &  35 &  30 & 85.71 \\
##    3 & New Zealand & 2011 &  46 &  37 & 80.43 \\
##    4 & New Zealand & 1987 &  18 &  13 & 72.22 \\
##    5 & Australia & 1999 &  40 &  26 & 65.00 \\
##    6 & South Africa & 2007 &  43 &  25 & 58.14 \\
##    7 & South Africa & 2019 &  43 &  21 & 48.84 \\
##    8 & Australia & 1991 &  23 &  11 & 47.83 \\
##    9 & South Africa & 1995 &  22 &   9 & 40.91 \\
##     \hline
## \end{tabular}
## \caption{Top 10 Teams of All Time}
## \end{table}
```

The top 10 teams table shows the win percentages of teams from each world cup cycle to show who was the best performing teams from all combined world cup cycles, including teams that di not win the World Cup. As can be seen New Zealand take up 5 of the top 6 spots in the highlighting their dominance outside of world cups. The 2015 World Cup winning team takes the top spot with a win percentage of 88.6% against all the top teams in the world, thereby making a strong case for being classified as the greatest rugby team in history. Interesting to note is that two teams from the 2023 world cup cycle are included in the top 10, France in 7th place with a win percentage of 77.4% and Ireland with a win percentage of 76.7%. This shows the good form both of these teams have had leading up to the 2023 Rugby World Cup.

The next step was to include the rankings of each team from before each world cup began for the last 4 world cups, seeing as these are the only world cups where the rankings existed. I included 2 variables for the rank of the home team and the away team. I also included a column indicating is the game was played at a nuetral venue or not. The only games to not be neutral were those played by the team hosting the World Cup. Lastly I included a variable which indicated the number of World Cup each team had won before the particular World Cup which was played in that year. For example South Africa's value would be 0 for the 1995 World Cup and before, 1 for the 1999 to 2007 World Cups, 2 for 2011 to 2019 World Cups and 3 for the 2023 Rugby World Cups. Again a variable was put in for both the home and away teams. I decide to include this variable because I believe that previous World Cup wins has an effect on teams winning considering many teams have won multiple World Cups and there are so few teams to have won the competition.

This finalizes the testing data set. In order to include each team's ranking as a variable which determines the outcome I needed to replace all the NA's in the years before there were rankings. I decided to replace them with the value 15, as this is larger than all other ranking values.

## 3. Creating the Testing Data Set

The next step was to create the testing data set with all the 2023 Rugby World Cup matches. After Importing the data set I can see that there are 15 games which need to be predicted in order to determine who will win the 2023 Rugby World Cup. So firstly I needed to put in the necessary information in order to predict who will win each game. This meant that the data set needed to include the following for each match: the year, home team, away team, home team win percentage, away team win percentage, home team rank, away team rank, neutral, home team titles, away team titles and lastly the points differences as the target variable.

For win percentages I used the win percentages of each team in the 2023 World Cup cycle which was worked out earlier. For rank I used the relevant ranking for each team as was given in the June 12 ranking table. For title I used the number of World Cup titles each team as won to date and the neutral column was FALSE for all observations, except for where France (the hosts of the tournament) were the home team.

# 4. Random Forest Model and Predictions

After creating the training and testing data sets have been set up meaning I can perform a random forest to determine who is going to win the 2023 Rugby World Cup. The random forest model learns from the training data set in order to predict what the the probable value of the target variable in the testing data set. In the case of this project, it sees how the other variables, or features, determines the points difference value in the previous World Cup matches data set. Based off of this information it predicts the points difference of each match in the 2023 matches data set, based off of the particular feature values of each observation. In order to do this

```
## [1] 16.65798
```

```
##    mtry min.node.size replace sample.fraction      rmse perc_gain
## 1     4            10    TRUE            0.50 16.20935 2.6931736
## 2     4             3    TRUE            0.50 16.31217 2.0758958
## 3     4             5    TRUE            0.50 16.39369 1.5865373
## 4     4            10   FALSE            0.80 16.47163 1.1186387
## 5     4             1    TRUE            0.50 16.54106 0.7018416
## 6     3             1    TRUE            0.50 16.54507 0.6777904
## 7     2             3   FALSE            0.63 16.55565 0.6142839
## 8     4            10   FALSE            0.50 16.60780 0.3012358
## 9     4            10   FALSE            0.63 16.60827 0.2984073
## 10    2            10   FALSE            0.63 16.61242 0.2734504
```

In order for this to work, I had to restrict my testing data set to only look at pool games initially. Once I ran the model on the pool games, the model predicted the points difference of each match, once again with a positive value indicating the home team won and a negative value indicating the away team won. Once the winners of each pool match was determnined, I was able to manually put the information into the quarter-final matches to correctly indicate who would play each other in which particular match. The modle was thereafter run again to predict the outcomes of the ppol matches and the quarter-finals. This process was repeated for both the semi-finals and final, with the final list showing the outcome for the entire World Cup.

```
## [1]  5.105836  7.895513  7.943690  6.713986 -3.484363 30.038368 30.018119
## [8] 16.895148
```

The results of the pools matches show the following outcomes (values were rounded to whole numbers):

This means that the quarter-final fixtures become as follows: QF1: QF2: QF3: QF4:

```
##  [1]  5.1058357  7.8955135  7.9436897  6.7139857 -3.4843635 30.0383683
##  [7] 30.0181190 16.8951484  2.4635230  0.9672175  7.9575563 10.1646984
```

After including the relevant teams for each quarter-final the results show as follows: QF1: QF2: QF3: QF4:

This means that ….. progress to the semi-finals with the following fixtures taking place: SF1: SF2:

```
##  [1]  5.1058357  7.8955135  7.9436897  6.7139857 -3.4843635 30.0383683
##  [7] 30.0181190 16.8951484  2.4635230  0.9672175  7.9575563 10.1646984
## [13] -9.9508897  5.6610040
```

The results of the semi-finals show as follows: SF1: SF2:

Therefore the final becomes the following match-up?

```
## [1]  5.7760595  7.8955135  7.9436897  6.7139857 -3.4843635 30.0383683
## [7] 31.1345302 16.8951484  2.4635230  0.9672175  7.9575563 10.8164730
## [13] -9.9508897  6.0855230  4.1272690
```

The result of the final shows as:

Therefore this model predicts that France will win the 2023 Rugby World Cup. I suspected that France being the hosting team had a significant effect on this outcome. In order to test this theory I checked which features were