2023 Rugby World Predictor

Peter Meihuizen

2023-06-23

Introduction:

For this project I will attempt to predict the winner of the 2023 Rugby World Cup using historical data of every international rugby match played by between the the historically best 10 international men's rugby teams in the world. The fixtures for the 2023 World Cup has already been released, with it consisting of group stages and knockout matches. Using the past match results from previous years, I will predict who will win every world cup game between the top 10 teams in the world, including the Rugby World Cup Final.

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.

I am going to do is generate 4 new columns, in order to identify the winner of each game, the points difference of each game and the winner of the World Cup. First I create a function to generate the points difference between the home_team and away_team.

Next is a variable for the winner and loser of the game.

Then a variable which indicates the winner of the World Cup

I can now see the winners of each world cup up to date

```
## # A tibble: 9 x 2
##   competition                WC_winner
##   <chr>                      <chr>
## 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
```

Ok now I want to see what the win-loss percentage was for every world cup winning team in the 4 years leading up to their respective world cup victory. In order to do this I will look at the number of wins over the combined number of wins and losses over each 4 year period preceding the beginning of the world cup. First I'm going to create a 2 columns which indicate the year of each match as well as the world cup cycle of each non_world cup falls under.

There is a bit of an issue in that some games were played after certain world cup in the same year and therefore get assigned to that year's world cup even though the games were played afterwards. However this shouldn't be an issue because we can group the data by both the rugby world cup cycle and the

Ok now I want to calculate the win percentage for each world cup winning team in their respective world cup cycle. Lets first indicate the world cup winner for each world cup cycle vaue so that we can group it by world cup winner and world cup cycle.

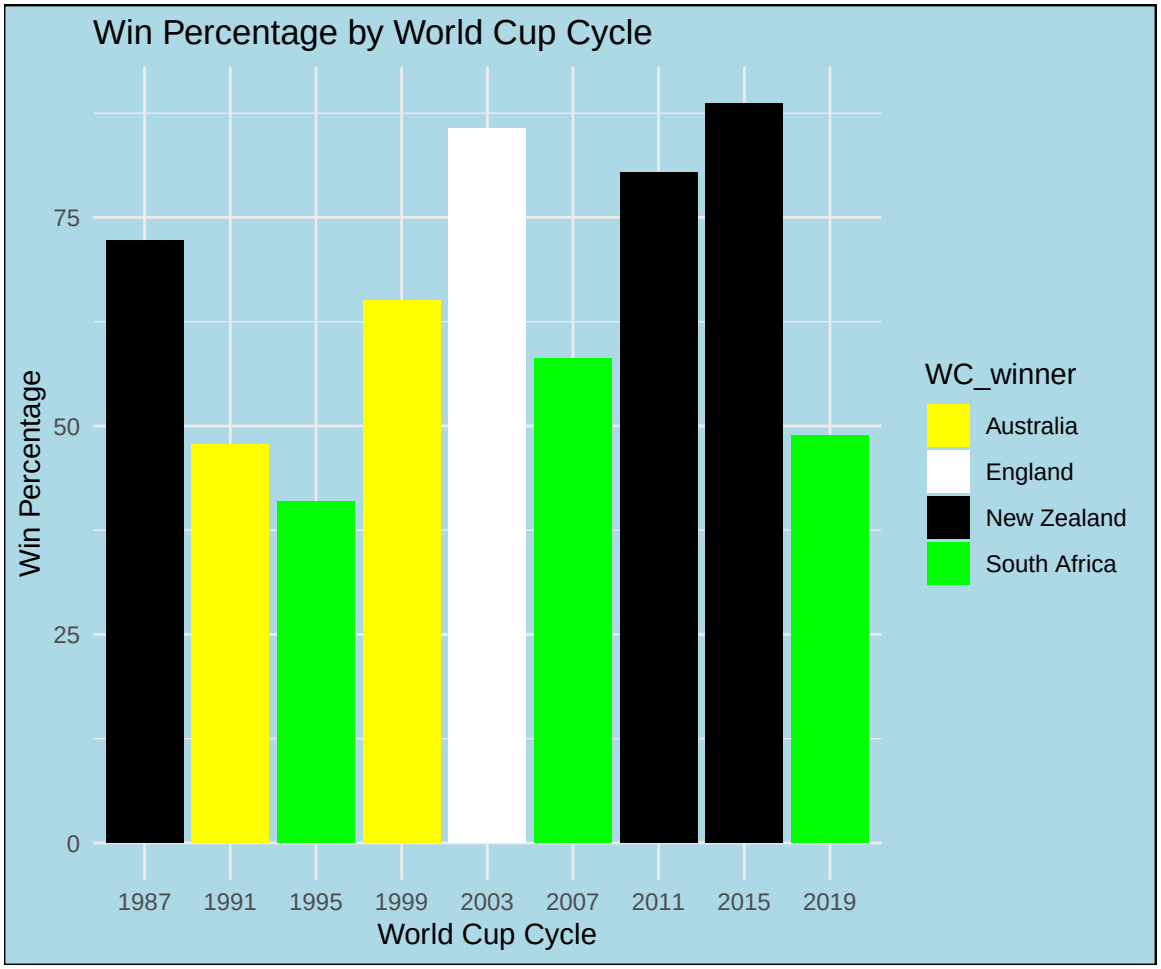
Ok so now to work out the win percentages of each WC winning tema against the top 10 teams in the world.

Ok so now it has given a table of the win percentages of each WC winning team against all the teams in our data set. However this data needs to be cleaned due to certain matches being played in the same year but being in the same year as a world cup but being played after the world cup, thereby falling under the next world cup cycle.

```
## # A tibble: 9 x 5
##   WC_winner    WC_cycle total_matches total_wins win_percentage
##   <chr>        <chr>            <int>      <int>          <dbl>
## 1 New Zealand  1987                18         13           72.2
## 2 Australia    1991                23         11           47.8
## 3 South Africa 1995                22          9           40.9
## 4 Australia    1999                40         26           65
## 5 England      2003                35         30           85.7
## 6 South Africa 2007                43         25           58.1
## 7 New Zealand  2011                46         37           80.4
## 8 New Zealand  2015                44         39           88.6
## 9 South Africa 2019                43         21           48.8
```

```
## # A tibble: 9 x 5
##   WC_winner    WC_cycle total_matches total_wins win_percentage
##   <chr>        <chr>            <int>      <int>          <dbl>
## 1 New Zealand  2015                44         39           88.6
## 2 England      2003                35         30           85.7
## 3 New Zealand  2011                46         37           80.4
## 4 New Zealand  1987                18         13           72.2
## 5 Australia    1999                40         26           65
## 6 South Africa 2007                43         25           58.1
## 7 South Africa 2019                43         21           48.8
## 8 Australia    1991                23         11           47.8
## 9 South Africa 1995                22          9           40.9
```

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 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.

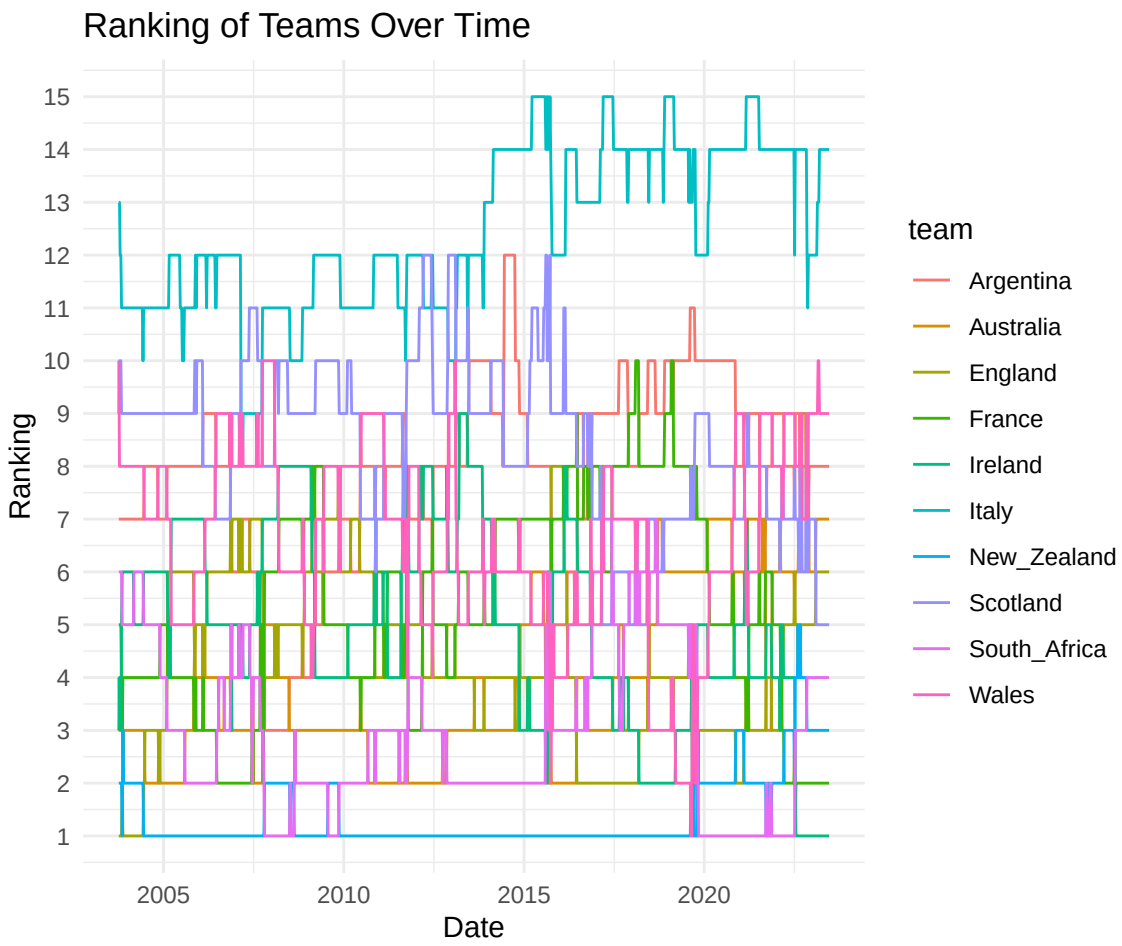


Next I want to include the World rankings of each team which have been collected since the conclusion of the 2003 Rugby World Cup. I need to web scrape this information.

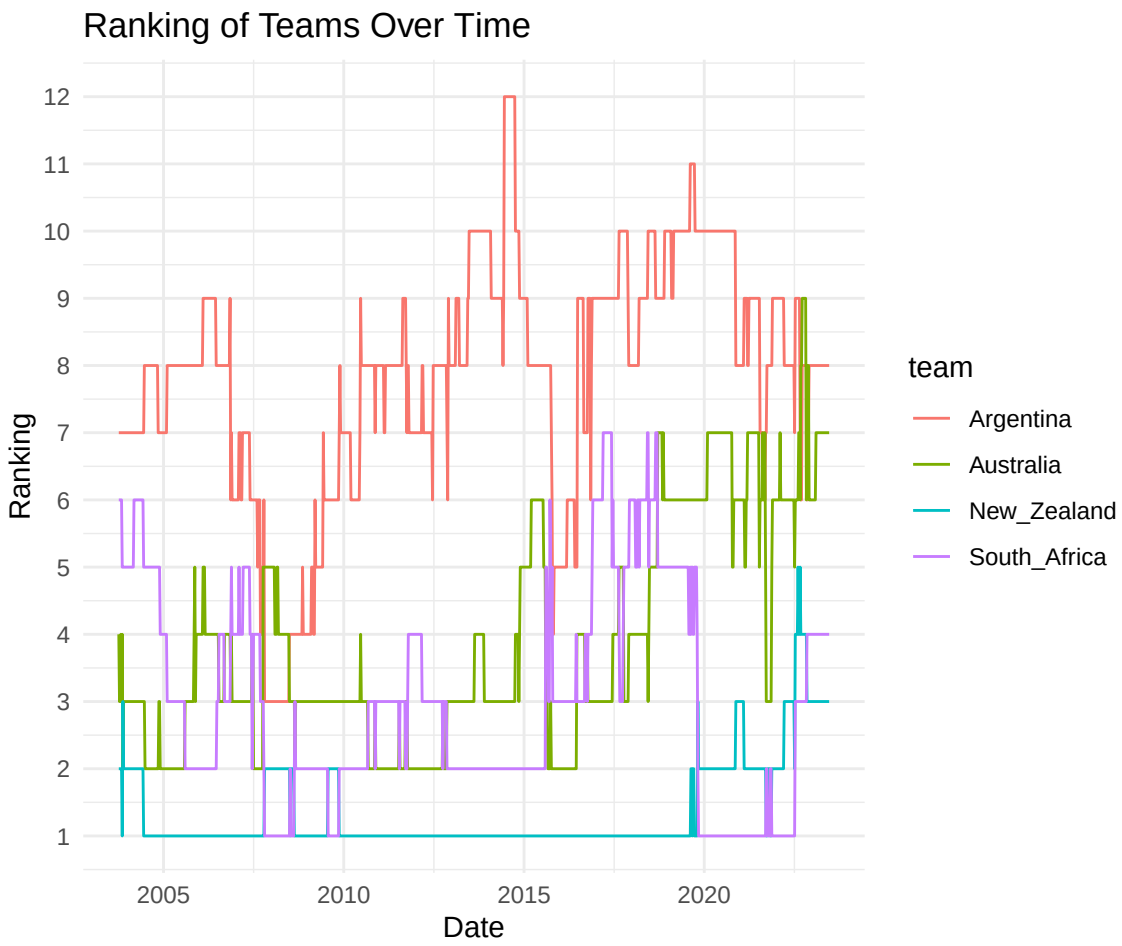
Now I turn that list which I scraped into a data frame and clean the data frame to only include the teams from the original data set.

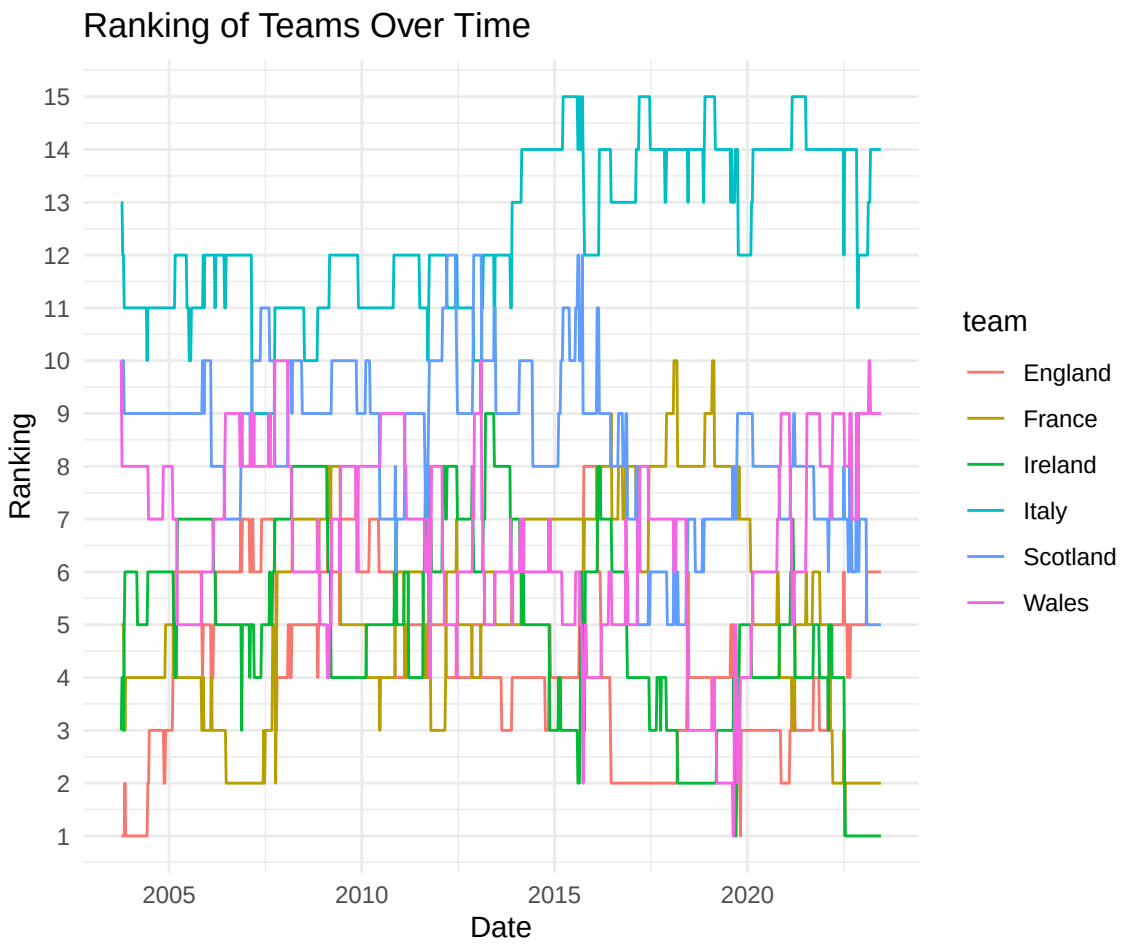
Next I need to change the date variables from characters to dates in order to merge the ranking data frame with the main data frame (df).

Ok so we have the rankings for all the top 10 teams since after the 2003 Rugby World Cup. Let's plot these ranking over time to see how each teams world ranking have changed.

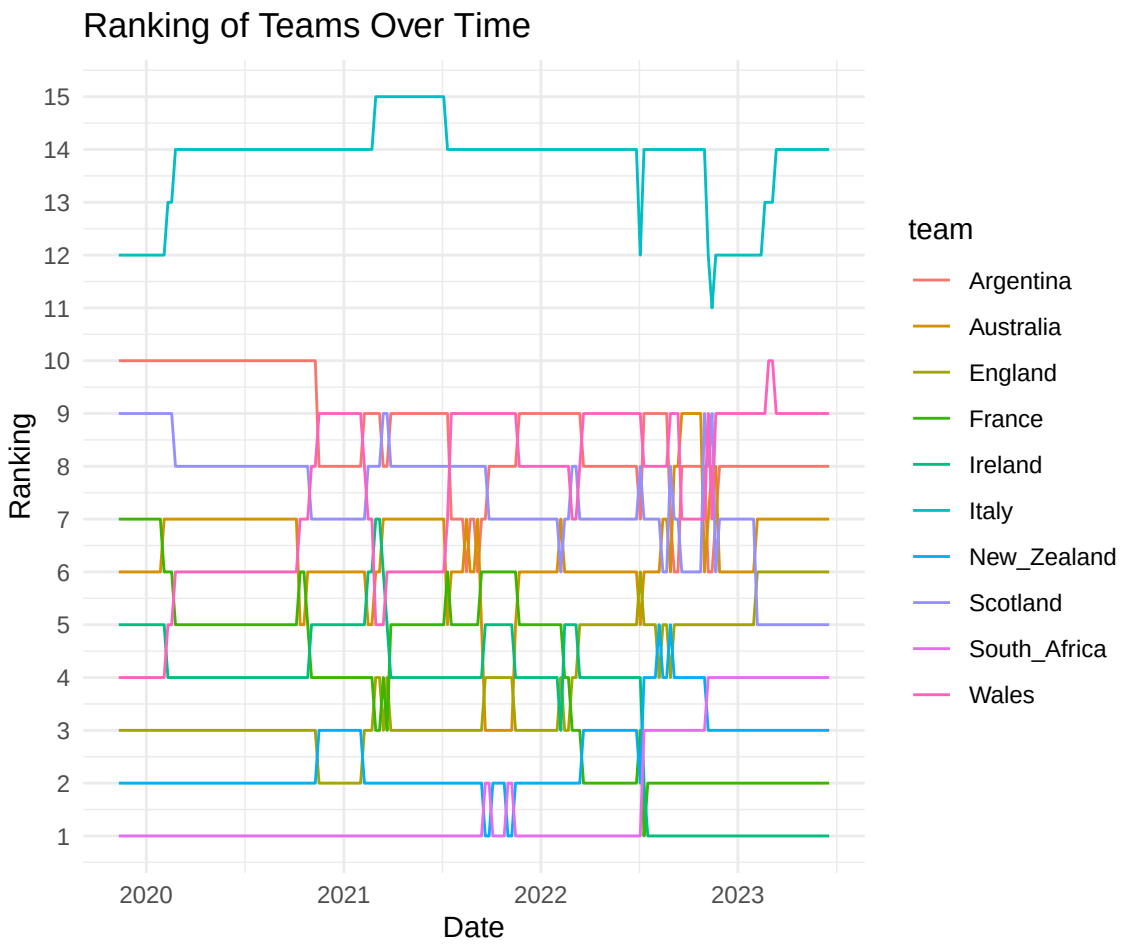


Let's decrease the number of teams in each graph so that it is slightly more interpretable. I'm first going to look at all the Southern Hemisphere teams (South Africa, New Zealand, Australia and Argentina). This graph will also show the ranking progression of the last 4 World Cup winning teams, two representing South Africa (2007 and 2019) and two representing New Zealand (2011 and 2015). In order to do this I am first going to split the rankings data set into a South And North data set.


Ranking of Teams Over Time
Ranking
12
11
10
9
8
7
6
5
4
3
2
1
2005
2010
2015
2020
Date
team
Argentina
Australia
New_Zealand
South_Africa



So as can ben seen the Southern hemisphere teams have dominated the rankings historically, with New Zealnad in particular holding the number 1 spot for a long time. Let's see how the ranking have been in the last World Cup cycle.



So as can be seen, in this world cup cycle South Africa have be the top team for most of the time, however Ireland has held the number 1 rank for the last year. The rankings stand as such as of the 12 of June 2023.

```
##            Team Ranking
## 1       Ireland       1
## 2        France       2
## 3   New Zealand       3
## 4  South Africa       4
## 5      Scotland       5
## 6       England       6
## 7     Australia       7
## 8     Argentina       8
## 9         Wales       9
## 10        Italy      14
```

So as can be seen the rankings show Ireland to be ranked number 1 on the most recent available rankings.

Next I want to look at the win percentage of each team against all other teams in the 4 year cycle before the world cup started. I also want to look at the win percentages between two teams for each respective match up between the teams. I will then import these results into a data set showing all matches which have been played at previous World Cups, what the result was and what was the total win percentage for the teams in that game, as well as their win percentage aginst each other.

Ok so there have been 132 matches played between these teams at World Cups. As an input to my model I need to calculate the win percentage of each team against these other teams in the 4 year cycle preceding this particular world cup.

```
## # A tibble: 10 x 5
##    WC_cycle team        total_wins total_matches win_percentage
##       <dbl> <chr>            <int>         <int>          <dbl>
##  1     2015 New Zealand         39            44           88.6
##  2     1991 New Zealand         21            24           87.5
##  3     2007 New Zealand         32            37           86.5
##  4     2003 England             31            36           86.1
##  5     2019 New Zealand         35            41           85.4
##  6     2011 New Zealand         37            46           80.4
##  7     2023 France              24            31           77.4
##  8     2023 Ireland             23            30           76.7
##  9     1995 England             16            21           76.2
## 10     2019 England             31            41           75.6
```

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. 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.

Ok now I need to import the winning percentages into the WC_matches data set so that I can show the win percentages of each team as an input into each game they played.

So now we've edited the WC_matches data set which shows the win percentages for each game, I want 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.

Ok now I have each teams ranking before the start of each world cup, now to combine this with the WC_matches data set to show the ranking of each world cup winning team and their opposition for before each world cup.

Ok now the onlly thing remaining is to indicate is the game was played at a neutral venue or not, this will only be the case if a particular team hosted the world cup and played.

Now I know that previous world cup wins has an effect on teams winning considering many teams have one multiple and there are so few teams to have won. Therefore I am going to include a variable which indicates how many world cups each teams has won before that particular world cup.

Now the learning data set is ready. 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 with the value 15, as this is larger than all other ranking values.

Now I need to import the data set for the 2023 Rugby World Cup fixtures.

I can see that I have 15 games which need to be predicted in order to determine who will win the 2023 Rugby World Cup.

So firstly I need to put in the necessary information in order to predict who will win each game. First it's about working out the win percentages of each team against all the teams in the data set over the last 4 years. I can use the win_percentgae data frame to iimport the relevant data.

Now both 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 world cup.

```
## [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
```

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

Results: Ok so now editing the relevant excel sheet to include the quarter finalists.

```
##  [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
```

```
##  [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
```

```
##  [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
```

Therefore France will win the 2023 Rugby World Cup.