FIFA World Cup

Aniket Saha

2022-11-04

Introduction

Football is the most popular sport in the world. FIFA World Cup is the biggest sporting event in the world in terms of watching and people involved. According to FIFA, around 3.75 billion people around the world witnessed the previous edition in 2018, held in Russia. Now, that is almost half the planet. This year 2022, 22nd edition of this famed competition will be held in Qatar. We are doing this project to get some insights about the teams participating. Moreover, we'll try to get some idea about which players or teams will perform well in this upcoming mega event. Now, we have collected some data from official FIFA website and some other secondary data from Kaggle. We will perform our analysis based on these data sets.

Data Sets and Variables

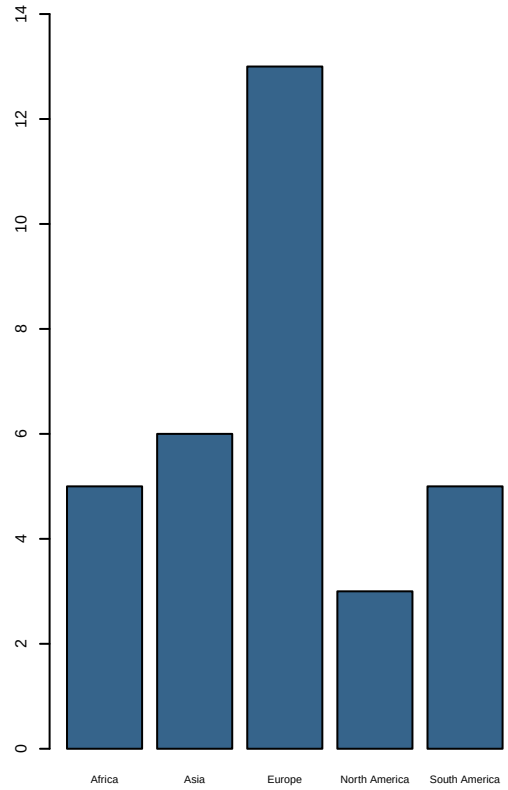
We have used the following data sets:

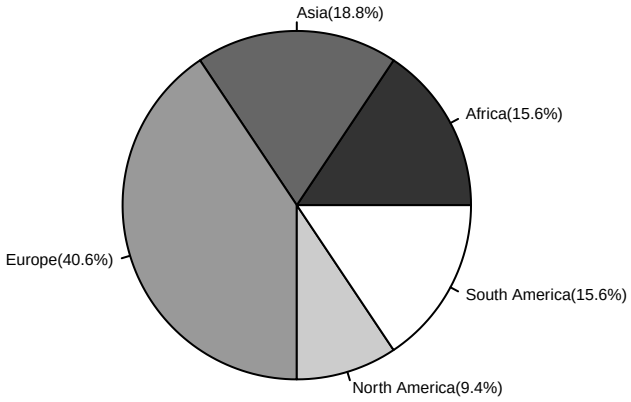
1. **Teams2022:** In this data set, we have:teams(country names), latest FIFA points of each team, ranks of each team based on points, region(continents). **Please note** that Australia participates through Asian qualification bracket, hence Australia is in group Asia.
   Source of Data: Self collected from FIFA website. Link: https://www.fifa.com/fifa-world-ranking/men
2. **WorldCup_All:** In this data set, we have the data of all the world cup matches since 1930. The variables are: year, host country, home team, away team, goals scored by home team, goals scored by away team, whether the game was played in a neutral venue and whether the home team won, drew or lost.
   Source of Data: Kaggle. Link: https://www.kaggle.com/datasets/ccsitai/worldcupdataset
3. **WorldCup_History:** In this data set, we have the years in which world cup were held over the years, number of teams, winner, runner up, whether host country became winner or runner up, total number of matches, total number of goals scored, average goals per game.
   Source of Data: Self collected from Wikipedia. Link: https://en.wikipedia.org/wiki/FIFA_World_Cup
4. **Players_FIFA23:** This is a subset of a very large data set containing all the details of all the football players around the world based on the EA Sports' game FIFA23 ratings which has very realistic data about the players.
   In this data set, we have the top 120 players in the game. The variables are: player name, country that they play for, overall rating, their position of playing (forward, midfielder, defender, goalkeeper).
   Source of Data: Kaggle: https://www.kaggle.com/datasets/bryanb/fifa-player-stats-database?select=FIFA23_official_data.csv

Analysis

Where They Are From

By the following plot we are trying to understand from which regions the teams come from. We are using the Teams2022 data set.
To begin with, we will see a couple of pictorial representation of this data.





Now, we want to know about the number of representatives from each region. To do that we simply show this data in a tabular form.

Africa	Asia	Europe	North America	South America
5	6	13	3	5

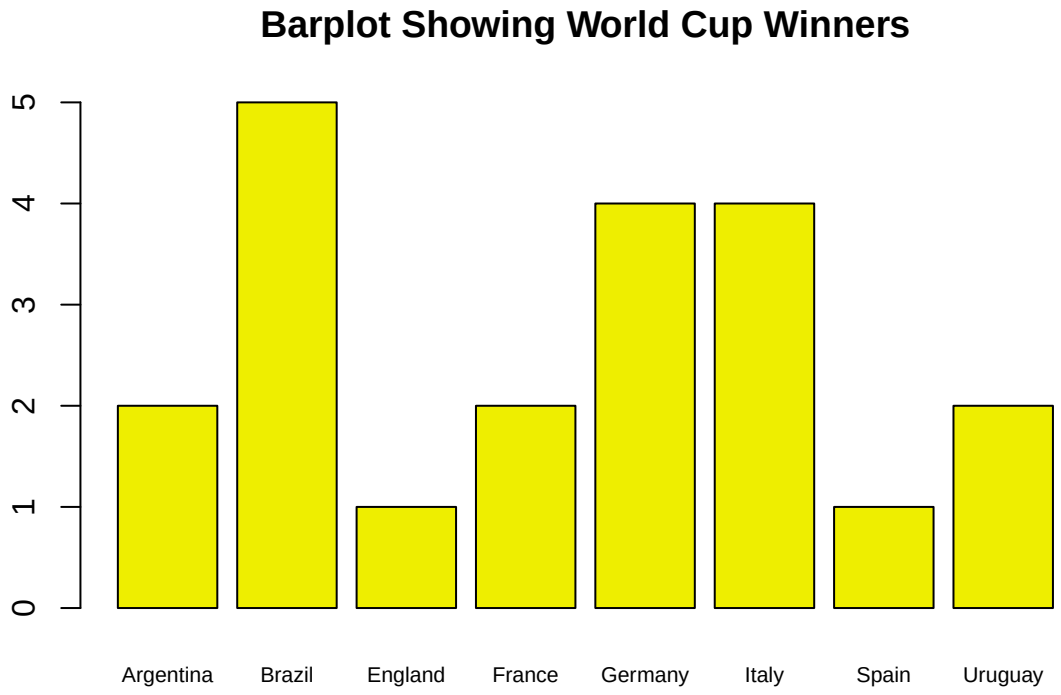
We can see that Europe has the most representative countries, 13, followed by Asia with 6, Africa and South America with 5 and then North America with 4.

Comment: Europe has more participants in the tournament as they have more nations compared to other continents like North and South America, and also the ranking of many European teams are much higher than that of Asian or African teams.

A Bit of History

Now, we want to see how the previous editions of the world cup went.
In this section we will know which team won, which team came runners up and the number of goals scored per game in each of the tournaments. We will use the WorldCup_History data set.

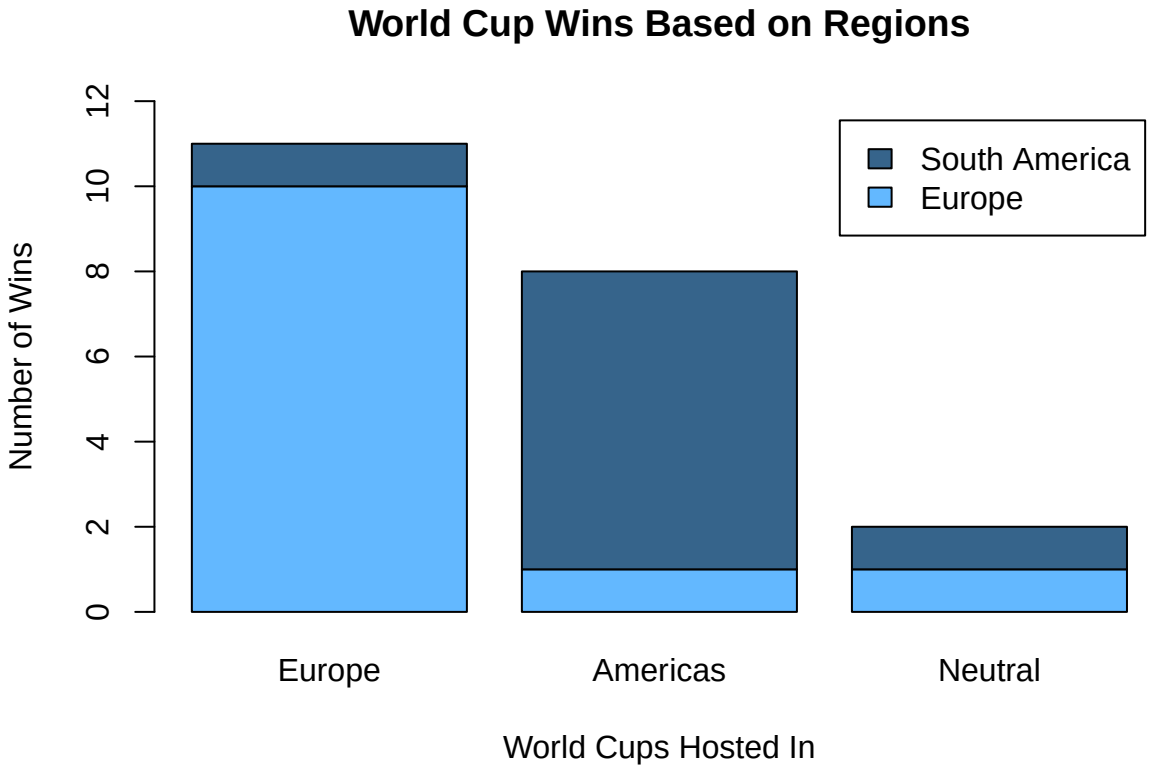
- **First of all, we will plot a barplot to see the teams, who have won the world cup before, and how many times they have won it.**



Comment: We can see that Brazil won the world cup 5 times, the most any country has achieved. Moreover, all the teams who have won the world cup are either from Europe or South America. Also European teams have won the tournament 12 times and South American nations have won it 9 times.

• **Now we will try to find out whether playing in a certain continent or region favours any team.**

We observe that among the 21 world cups so far, 11 were held in Europe, 8 were held in American continents and 2 were held in Neutral countries (One in each of Africa and Asia). This year the tournament will be held in Qatar which is in Asia and falls under the Neutral category in our classification. We will plot a subdivided barplot to depict the results of the world cup classified based on regions.

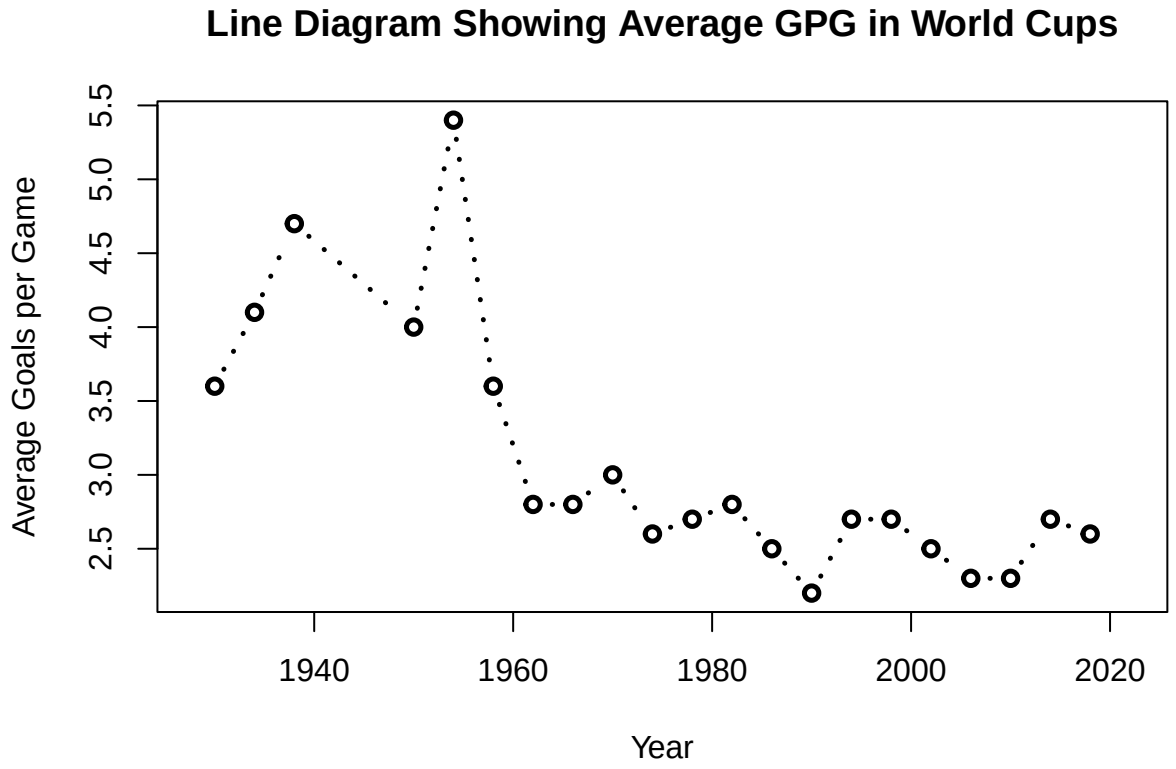


Comment: Here we can see that when the world cup is hosted in Europe, European countries dominate over others. Whereas while the world cup is hosted in American continents, South American teams dominate over others. In neutral countries, we get that the contest is even.

What to Expect This Year

- **Now we will try to predict the number of goals to be scored in this year's world cup.**

First we will plot a line diagram showing the GPG or goals per game of each edition of the world cup. Since the number of matches and teams are not equal in all the tournaments, we use this GPG variable for our analysis.

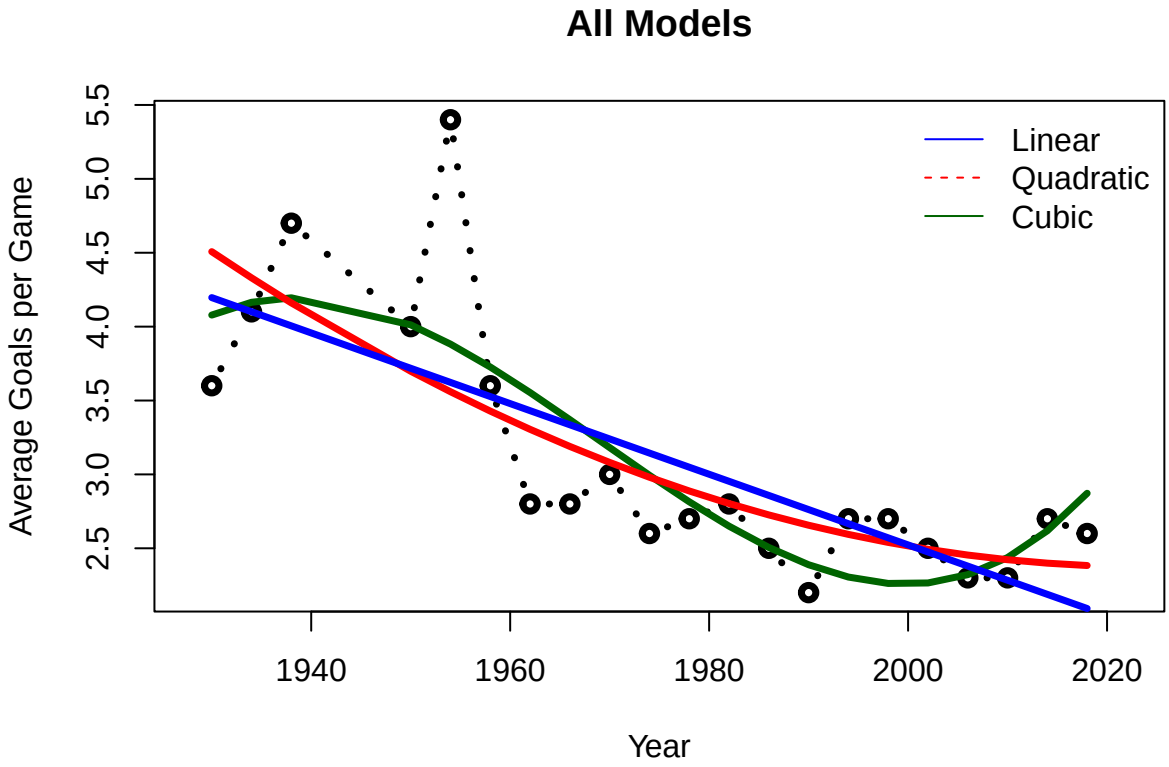


We observe that the average number of goals per game has decreased over the years. We will fit three types of models:

1. Linear
2. Quadratic
3. Cubic

Based on these models we will try to predict the average goal per game amount of 2022 world cup.
We have constructed our models and all the three plots are merged together along with their interpretation in the following page.

Plots are prepared to predict the average GPG of world cup 2022



We have made the following observations and conclusions from the graph:

1. **From the Linear Model:** We see that the blue line depicting the linear model actually decreases to 0 and beyond before reaching Year=2022. Hence no conclusion can be made based on the linear model as average goal per game can not be negative.

2. **From the Quadratic model:** From the quadratic model, we can at least make some inference. It also, however, decreases but doesn't go to 0 before Year=2022.
   According to this model, the average GPG of the 2022 world cup will be approximately 2.5. This is less than the last two world cups, but more than the 2006, 2010 world cups.

3. **From the Cubic Model:** The cubic model shows that the average goal per game is increasing since 2000 and the average GPG for 2022 world cup will be approximately 3.0, which is higher than last 12 world cups.

Comments: So the three models are quite varied in their predictions. However, based on the plots, we can say that the quadratic model and the cubic model fit the data better than the linear model. And hence we take the mean of these two estimates as an estimate of average GPG of 2022 world cup. Therefore the average GPG of this year's world cup will approximately be $(2.5 + 3)/2 = 2.75$

Home Sweet Home?

- **Now we will try to understand whether playing on home turf results in performing well in the world cup.**

We represent the following tabular data to see how much the home team finishes in top2 or top4.

Table 1: World Cup Winners

Year	Host	Winner	Second	Third	Fourth	Host_Top2	Host_Top4
1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	1	1
1934	Italy	Italy	Czechoslovakia	Germany	Austria	1	1
1938	France	Italy	Hungary	Brazil	Sweden	0	0
1950	Brazil	Uruguay	Brazil	Sweden	Spain	1	1
1954	Switzerland	Germany	Hungary	Austria	Uruguay	0	0
1958	Sweden	Brazil	Sweden	France	Germany	1	1
1962	Chile	Brazil	Czechoslovakia	Chile	Yugoslavia	0	1
1966	England	England	Germany	Portugal	Soviet Union	1	1
1970	Mexico	Brazil	Italy	Germany	Uruguay	0	0
1974	Germany	Germany	Netherlands	Poland	Brazil	1	1
1978	Argentina	Argentina	Netherlands	Brazil	Italy	1	1
1982	Spain	Italy	Germany	Poland	France	0	0
1986	Mexico	Argentina	Germany	France	Belgium	0	0
1990	Italy	Germany	Argentina	Italy	England	0	1
1994	USA	Brazil	Italy	Sweden	Bulgaria	0	0
1998	France	France	Brazil	Croatia	Netherlands	1	1
2002	Korea, Japan	Brazil	Germany	Turkey	South Korea	0	1
2006	Germany	Italy	France	Germany	Portugal	0	1
2010	South Africa	Spain	Netherlands	Germany	Uruguay	0	0
2014	Brazil	Germany	Argentina	Netherlands	Brazil	0	1
2018	Russia	France	Croatia	Belgium	England	0	0

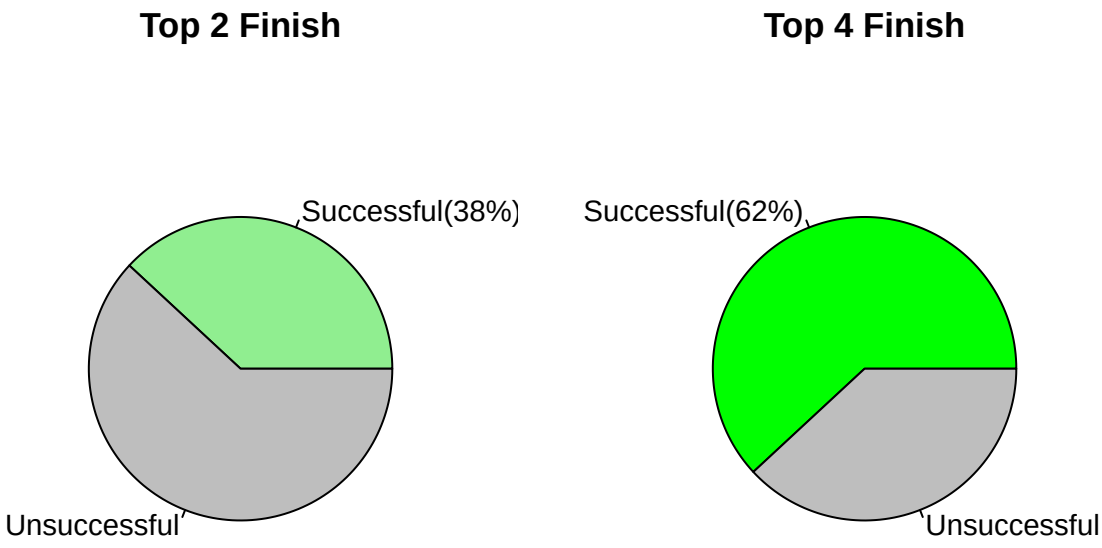
Here in the data, we should note that:

1. In 2002, South Korea and Japan were joint hosts of the world cup.
2. "Host_Top2" and "Host_Top4" show whether the host country finished in top 2 or top 4 respectively. 0 means they did not and 1 means they did.
3. Some countries like Yugoslavia, Soviet Union, Czechoslovakia do not exist anymore but they played in the mentioned tournaments many years back.
4. Germany participated as West Germany in some editions of the world cup, however, for easier analysis we have kept the name Germany in our data throughout. For more information on this please refer to https://en.wikipedia.org/wiki/Germany_at_the_FIFA_World_Cup

We observe that the total number of times the host country have finished in top 2 is 8
On the other hand, the total number of times the host country have finished in top 4 is 13
If we want to find the proportion of times a host country has finished in top 4, then we have:

$$p = 13/21 = 0.6190476$$

Which is quite high.
Now we will depict this data via a subdivided barplot as follows in the next page.



We can see that the host teams have finished in top 2 in more than 35 times of all world cups and they have finished in top 4 more than 60 of times.

- Now we will accompany this analysis with home team's match by match records obtained from the 'WorldCup_All' data set.

We have the data of all the matches ever played in the world cup. First of all, we will see how many matches have been won by the home teams, percentage of that and then we will see the data in the form of a pie chart.

To begin with, we will use the measure:

- Expected_Outcome=[(# wins)(1) + (# losses)(-1) + (# draws)(0)]/(Total Number of Matches Involving Hosts).

If this comes out positive then we can say that the host team is more likely to win a game, otherwise if it comes out negative then we can say the host team is more likely to lose a particular game.

Now, from the data, we get: Expected_Outcome = 0.440678

We can see that the Expected_Outcome gives a positive number.

Comment: Hence we can conclude that the host country is more likely to win any particular game. So, we can expect Qatar to perform well in this year's world cup.

Now we will take a look at the number of goals scored per game involving the host team. Number of Goals per Game Involving Host Team = 2.8305085

The average Goal per Game of all teams = 2.8311111

We see that there is a little difference in average GPG between these two categories.

However, we still have very little clue about which teams are more likely to win the world cup. Hence we move on with our analysis.

The Big Prize!

- **Now we will use the team rankings and the FIFA points system to try to predict which teams are likely to win the prestigious trophy.**

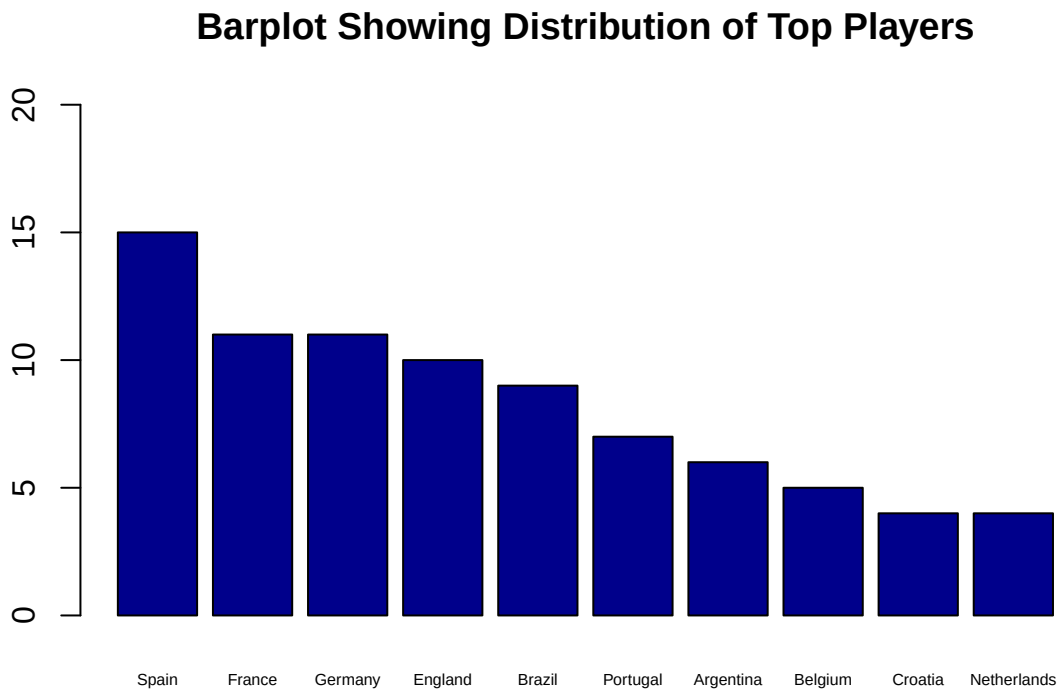
We will also use the FIFA23 data set of players to see which teams have the strongest players.

Our original data set had 120 players. However, all of them are not from countries which are participating in the world cup this year.

Now we filter out the players who could feature in the world cup.

We have 98 such players.

Now we will depict the top 10 countries with most players among these 98 players:



We see that Spain have the most players among the aforementioned 98 players, followed by France, Germany, England, Brazil.

Now we will see the ranks and FIFA points of the top 10 teams in this year's world cup. The data is shown in a tabular form in the following page.

Table 2: Countries Along With Their FIFA Points

Country	FIFA_Points
Brazil	1841.30
Belgium	1816.71
Argentina	1773.88
France	1759.78
England	1728.47
Spain	1715.22
Netherlands	1694.51
Portugal	1676.56
Denmark	1666.57
Germany	1650.21

• **Now we will try to come up with a team which is more likely to win this year's world cup than others.**

First of all, we find out the countries which are in top 10 in terms of having elite players and FIFA points. These are:

Spain, France, Germany, England, Brazil, Portugal, Argentina, Belgium, Netherlands.

So we can expect one of these teams to win the world cup this year.

Among these, teams who have won the world cup previously are:

Spain, France, Germany, England, Brazil, Argentina.

We need to distinguish among these teams and come up with our predicted winner.

Now we will use the following index to use in our prediction.

Team_Index = (wonp) + (# elite players)/10 + ((FIFA Points)-mean(FIFA Points))/sd(FIFA Points)

Here, **(wonp)** is a variable, which takes value 1 if the team has previously won the world cup, 0 otherwise. **(# elite players)** is the number of elite players the team has out of top 120 players according to the FIFA23 data set.

We have decided to normalize **FIFA_Points** because it has very high values and the distance between the values is not much. So in a linear model (as ours), adding or subtracting the FIFA points to construct our measure would not be a good idea.

How Team_Index Measure Worls: The measure takes into account whether the team has experience of winning such a big tournament. It is reflected by the 'wonp' variable. Since a pformer world cup champion country will have the mentality and experience to do it again, we have taken the coefficient of 'wonp' to be positive.

It takes into account the number of elite footballers each team has, which is reflected by 'Elite_Players' variable.

Most importantly it normalizes FIFA points, which is a measure of current strength and form of a country.

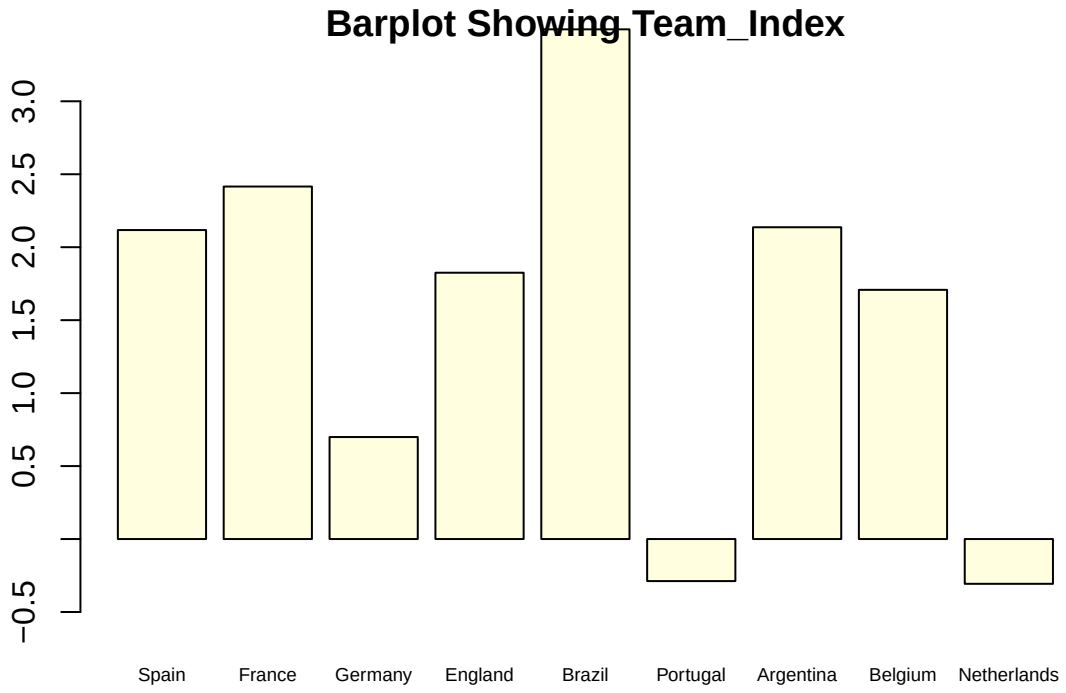
We have used a linear model to find Team_index since it is easy to understand and easy to compute.

The table showing Team_Index is given by:

Table 3: Table Showing Team Index Values

Teams	FIFA_Points	Elite_Players	wonp	Team_Index
Spain	1715.22	15	1	2.1175892
France	1759.78	11	1	2.4157683
Germany	1650.21	11	1	0.6989935
England	1728.47	10	1	1.8251941
Brazil	1841.30	9	1	3.4930475
Portugal	1676.56	7	0	-0.2881470
Argentina	1773.88	6	1	2.1366912
Belgium	1816.71	5	0	1.7077642
Netherlands	1694.51	4	0	-0.3069011

We will see this data using a bar plot.



Comment: We observe that Brazil has the highest Team_Index value, followed by France, Belgium and Spain.

Hence we predict Brazil to win the world cup this year, which would be their sixth world cup in history.

Conclusion

We have done some exploratory data analysis on the history of FIFA World Cup and tried to predict the winner of this year's world cup. Also we predict that the average goal per game in this year's world cup will be 2.75, which is lower than the previous two world cups.

At the end of our analysis, we have found that Brazil is most likely to win the world cup. However, Belgium, France, Germany and Spain would be tough competitors for the trophy.

Reference

We have used the following references to do this project.

1. https://www.google.co.in/
2. https://stackoverflow.com/
3. https://www.fifa.com/
4. https://www.kaggle.com/
5. https://www.wikipedia.org/

The End: Addendum

This is the end of our project.

However, for some additional information, we will present a table showing every world cup winner so far.

Table 4: World Cup Winners

Year	Winner
1930	Uruguay
1934	Italy
1938	Italy
1950	Uruguay
1954	Germany
1958	Brazil
1962	Brazil
1966	England
1970	Brazil
1974	Germany
1978	Argentina
1982	Italy
1986	Argentina
1990	Germany
1994	Brazil
1998	France
2002	Brazil
2006	Italy
2010	Spain
2014	Germany
2018	France

Thank You