Is the FIFA World Cup Draw Truly Random?

Pratik Kulkarni (19400570)
Shrey Parekh (18706941)

Final Project Report for 301111 (MATH3004) Discovery Project

Supervisor: Dr. Russell Thomson

School of Computer, Data and Mathematical Sciences,
Western Sydney University

Spring, 2021

Table of Contents

1 Background

Is the FIFA World Cup draw truly random?

This section is written by Pratik Kulkarni (19400570)

This is the topic chosen for our Discovery Project. The topic was selected by our supervisor, Dr. Thomson while watching a soccer match and decided to investigate this topic. We intend to investigate whether the FIFA World Cup draw is truly random and if each group has an even chance of being drafted into the groups.

Researchers such as Cea et al. [2020], conducted research on a similar topic. In this paper, they found that one of those main factors that affects the countries chances of being drafted is their home-away status. While this does not directly affect the team ranking by points, a general belief is that teams always play better at home, this factor has not yet been considered in the draw of teams so far. This affects their game due to the home team feeling stronger with a crowd that is supportive of them. This in turn influences their world ranking. Moreover, another factor found in this paper's research was that until the 2018 World Cup, teams would avoid friendly games due to friendlies awarding very few points, teams that would play more friendlies had less chances of going further up the team ranking table as compared to teams who avoided friendly matches and climbed up the ranks.

In the 2018 World Cup, FIFA changed the draw system to a rank-based system, the groups would be split by the team's rankings. While this could be a better system, there are chances that the groups might become unbalanced, especially if a low ranked host country was placed in the top group with the best teams.

The knowledge gap in this topic would be that there is no fairness index created by any researchers. To fill this gap, we would have to make a fairness index for each team that did qualify for our chosen year of the World Cup based on their FIFA team ranking in the same groups.

2 Problem Statement

This section is written by Shrey Parekh (18706941)

The problem that this project is investigating is providing value to the fans of FIFA World Cup. This problem is valuable to address because it can put speculation of fans, statisticians, and data scientists to rest regarding the draw of the World Cup being truly random. This is a problem as it is always an interesting question during the World Cup which team will have matched each other. This raises a concern are the teams being treated fairly during the draw is a strong team matched with a weaker team. Some teams also gain an advantage due to hosting it in their Home Country.

This has always been a debatable question that we will address through the analysis in this project. This problem has not specifically been addressed in other projects. These projects have been predicting who will the World Cup for the respective year for instance winning the 2018 World Cup. The analysis being conducted is addressing the draw difficulty of the World Cup and predicting the winner of the 2022 FIFA World Cup.

3 Objective

The Supervisor's goals are centred on providing value and results to the fans. This can be seen through testing the randomness of the FIFA World Cup draw. The goal of this project is to determine the predictions of the simulated World Cup draw. Aim is to discover the fairness index for each team that qualified for the World Cup, based on the FIFA team rankings of the teams in the same grouping. This will allow us to determine that World Cup draw is based on chance or is done through an algorithm.

This section is written by Shrey Parekh (18706941)

We will establish whether the fairness index affects the performance through multiple linear regression or non-parametric tests. This will increase the scope of this project as we will be able to produce efficient statistical results to ascertain the influences of chance in the FIFA World Cup draw.

This can be further reinforced by simulating the World Cup draws to test if any countries have a fairness index that is higher or lower than expected by chance alone. This is to proclaim that certain countries will not have a favourable advantage while competing in the World Cup.

4 Hypothesis/question

Is the FIFA World Cup draw truly random? Can we test the randomization of the grouping?

This section is written by Shrey Parekh (18706941)

4.1 Why Ask this Question?

This is a concern fuelled by the excitement of the fans of the FIFA World Cup. This is important because it raises an interest about who will win this year's trophy. It also generates doubt in the mind of fans as the draw based on chance is fair or is it biased to some teams. They want to know the algorithm that is used to decide the matchups of the team.

The value of answering this question is that it will remove the doubts from people's minds. This will allow the fans to determine that the teams competing are fairly matched up. They will be able to then ascertain that the draws are random, and each team is given a fair chance to progress through the tournament. This will provide confidence for fans as the best performing team will win, not any team based on a biased draw.

4.2 Data Question

Can we accurately predict the winning team of the FIFA World Cup?

4.3 Data

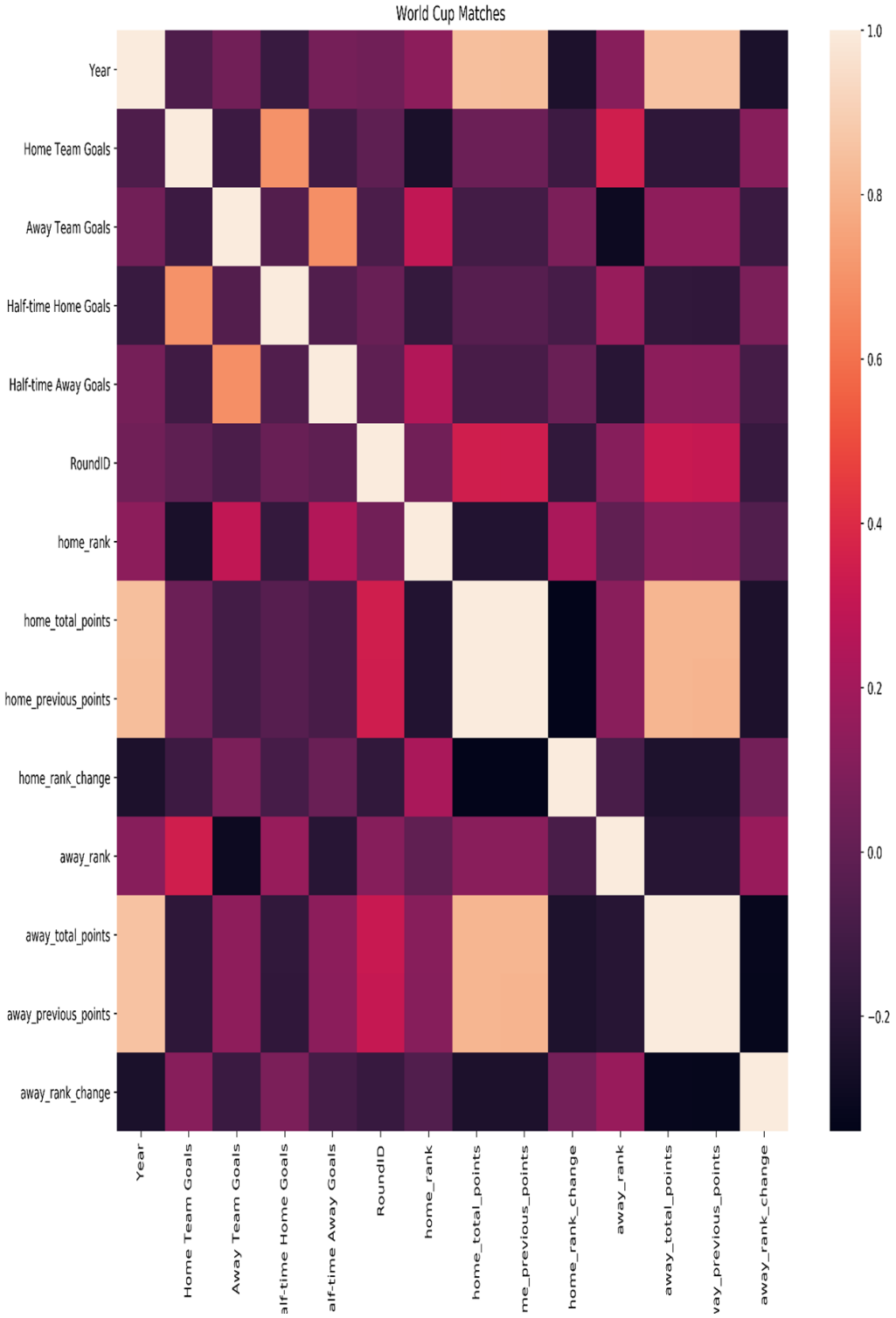
The datasets that have been used in this project are World Cup Matches dataset, team rankings dataset, final team positions dataset and 2018 World Cup Matches dataset. The World Cup Matches dataset is sourced from Kaggle and it has a usability index of 8.8. It demonstrates that dataset is reliable. This dataset contains

World Cup matches data from 1930 to 2014. In this project World Cups from 1992 onwards to 2018 will be used. This is due to FIFA has been official recording the rankings of teams from 1992. Prior to this there is discrepancy in the ranking data as there is no official source to justify that these are the ranks from 1930 to 1990 and the ranks are missing during this time. The 2018 matches dataset is sourced from GitHub. This dataset is merged to the existing World Cup matches data.

This will enable us to predict the 2022 FIFA World Cup Match. The final team positions dataset is manually scraped by using beautiful soup and pandas from the fox sports World Cup history. These rankings demonstrate the first 4 positions of the World Cup holder ranging from champion, runner up, third place and fourth place. The team rankings dataset is sourced from GitHub fifa-world-ranking. This dataset contains is merged with the World Cup matches dataset. This is done because total points, rank of home team and away team is required to implement the fairness index algorithm.

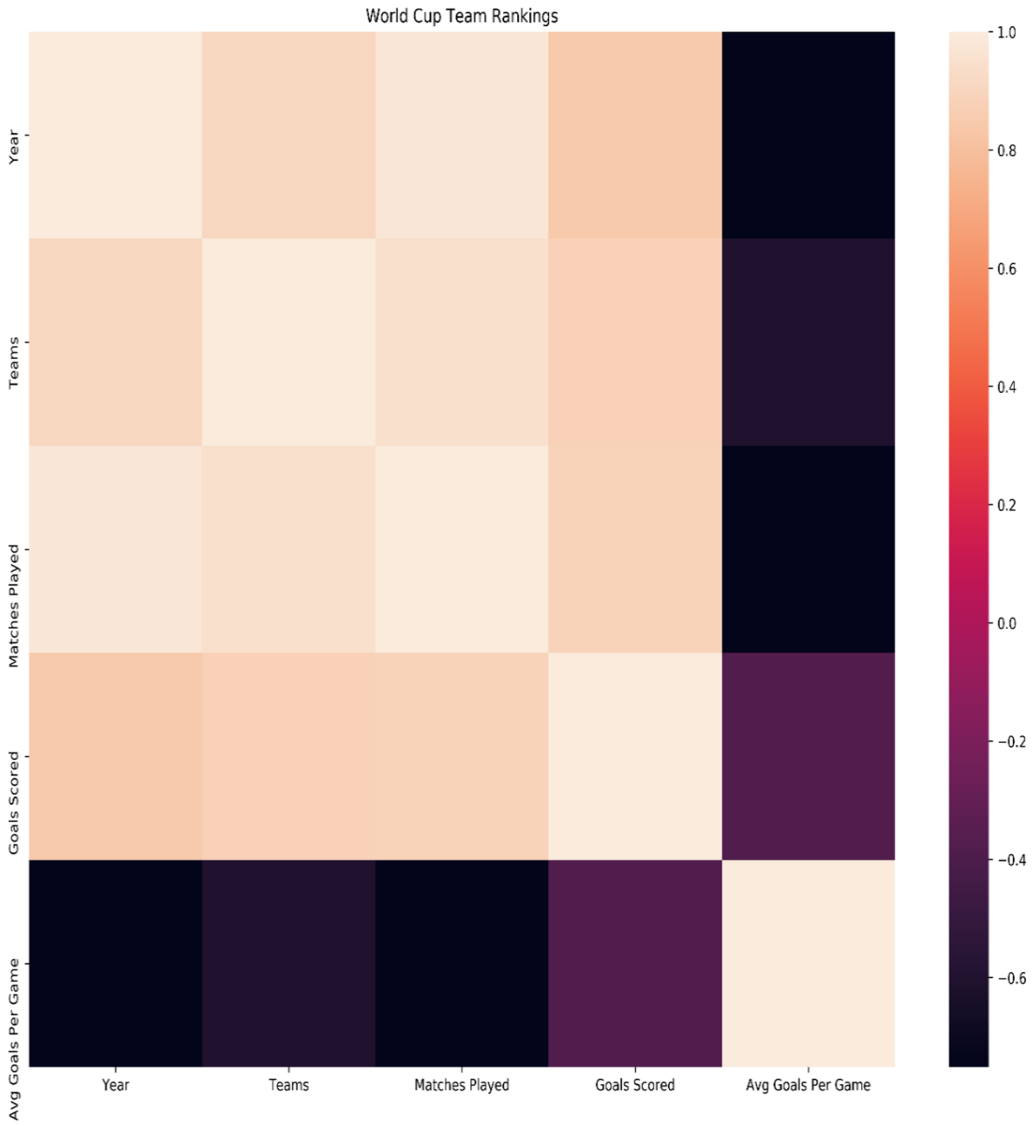
The 4 datasets features can be observed below through heatmaps. Overview of the features of the merged World Cup dataset can be seen below through the heatmap in Figure 1.

Figure 1


World Cup Matches
Year
Home Team Goals
Away Team Goals
Half-time Home Goals
Half-time Away Goals
RoundID
home_rank
home_total_points
home_previous_points
home_rank_change
away_rank
away_total_points
away_previous_points
away_rank_change
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

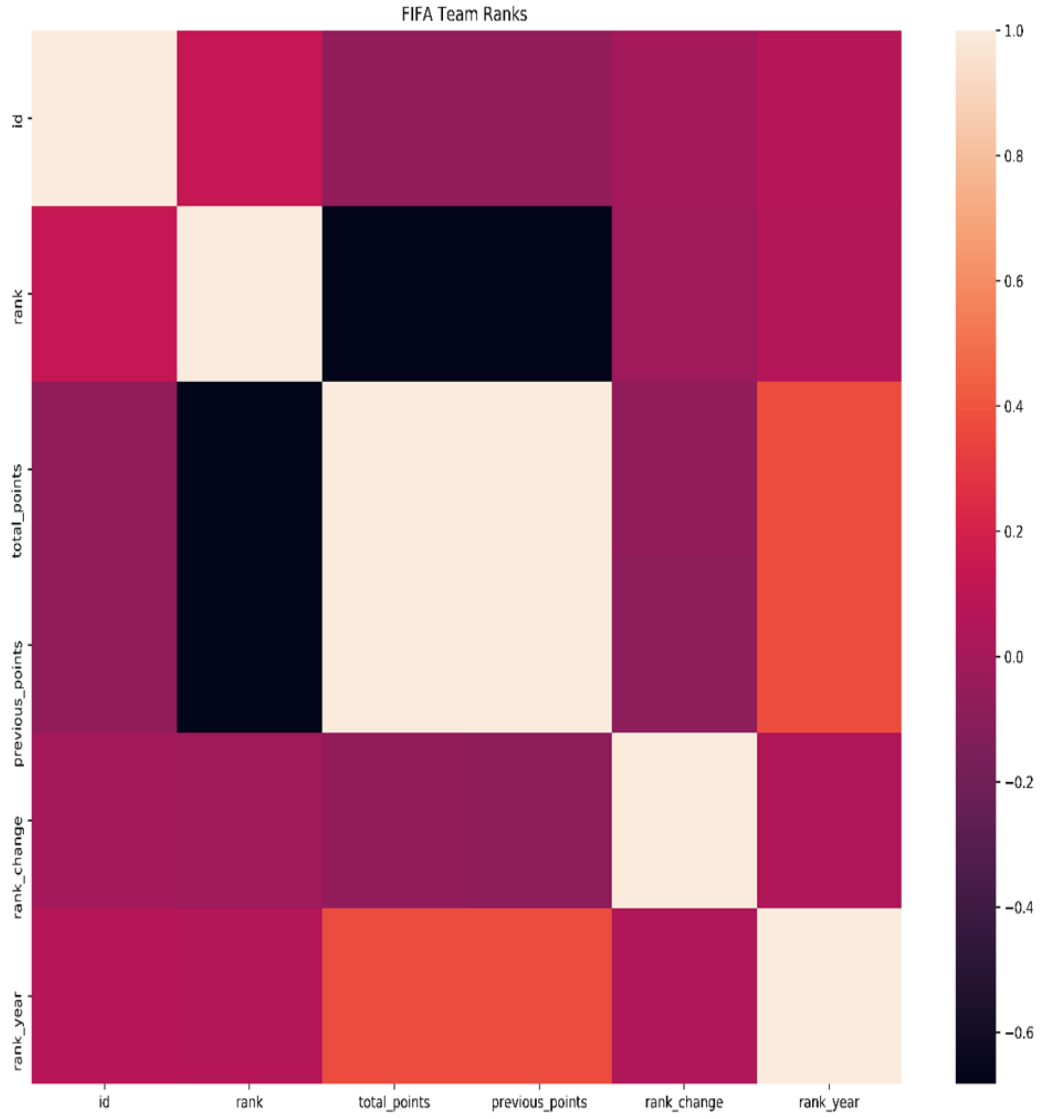
Another heatmap outlining the features of the final team rankings dataset is depicted below in Figure 2.

Figure 2



The team rankings features are depicted in the below heatmap in Figure 3.

Figure 3

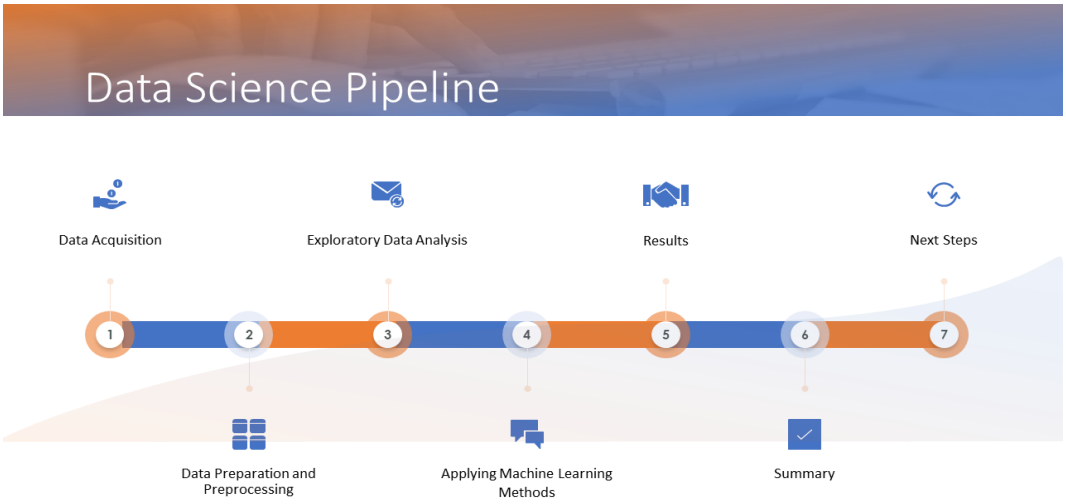


4.4 Data Science Process

4.4.1 Data Analysis

The Data pipeline that was used to wrangle the raw data is below in Figure 4:

Figure 4: Data Science Pipeline



5 Methodology

This section is written by Shrey Parekh (18706941) and Pratik Kulkarni (19400570)

The goal of our project is to create a fairness index for each team that qualified for the World Cup based on team rankings in the groups. Test the performance being affected by index through multiple linear regression tests and non-parametric tests. Simulating the World Cup draw to observe the draws are fair and not based on chance.

The measures that we need to take to conduct experiments to obtain precise results for this project are that we need to generate a fairness index from the points scored by the teams in their respective groups. This will be calculated using Python libraries such as NumPy and Pandas. Then we will look at the World Cup Result where the most successful team gets the highest amount. A predictive model will be generated based on the fairness index and World Cup Result. This will allow us to demonstrate that group rating and fairness index affects the winning team.

The calculation below is an example to use:

$$lm(result \sim points + fairness + features + residuals)$$

After this step our goal is to test the linear regression model for randomisation of the grouping by simulating the process in which the team members are grouped based on their ranks and points. This is achieved by writing a Python script to assess the randomisation of groupings.

The first step is to add the datasets together. We add the FIFA ranks dataset sourced from Kaggle with World Cup results. There are some changes that we must make to this dataset. We have to change the date-to-date time format, old names of countries to the latest names of the current countries. Then filter the rows based on date so we can easily access it through the loc function. This function helps navigate between rows easily.

The dataset output can be viewed in Table 1 below.

Table 1

rank	country_full	country_abrv	total_points	previous_points	rank_change	confederation	rank_date
104	Swaziland	SWZ	10	0	0	CAF	1992-12-31
42	Turkey	TUR	31	0	0	UEFA	1992-12-31
43	Northern Ireland	NIR	31	0	0	UEFA	1992-12-31
44	Finland	FIN	31	0	0	UEFA	1992-12-31
45	Australia	AUS	29	0	0	AFC	19922-12-31
74	United Arab Emirates	UAE	1326	1326	0	AFC	2020-12-10
75	China PR	CHN	1323	1323	0	AFC	2020-12-10
76	Curaçao	CUW	1313	1313	0	CONCACAF	2020-12-10

Then the extraction begins of the rank date into year and rank months. This is done by converting rank to date year format. This will help in implementing the fairness index algorithm. This can be seen below in Table 2:

Table 2

rank_year	rank_month
1994	2
1994	2
1994	2
1994	2
1994	2
...	...
2018	12
2018	12
2018	12

Then this data is joined with the FIFA ranks dataset as this is important for the calculation of the fairness index algorithm. This table is seen below in Table 3.

Table 3

id	rank	country_full	country_abrv	total_points	previous_points	rank_change	confederation	rank_date	rank_year
43943	109	Estonia	EST	11	11	0	UEFA	1994-02-15	1994
43916	165	St. Kitts and Nevis	SKN	0	0	1	CONCACAF	1994-02-15	1994
44002	58	Slovakia	SVK	30	2	92	UEFA	1994-02-15	1994
43817	57	Iran	IRN	30	31	2	AFC	1994-02-15	1994
43826	165	Macau	MAC	0	0	1	AFC	1994-02-15	1994
...	...	...	...	...	...	...	...	...	...
43877	137	Rwanda	RWA	1094	1094	0	CAF	2018-12-20	2018
43916	136	St. Kitts and Nevis	SKN	1105	1105	0	CONCACAF	2018-12-20	2018
43842	135	Yemen	YEM	1106	1106	0	AFC	2018-12-20	2018
43957	133	Lithuania	LTU	1111	1111	0	UEFA	2018-12-20	2018
43980	144	Solomon Islands	SOL	1073	1073	0	OFC	2018-12-20	2018

Next the World Cup matches dataset is loaded. This dataset needs to be cleaned as there are missing values and the names of old Countries need to be updated to the current names of the following Countries: for instance, team name IR to Iran. This needs to be done for the teams that have previous names changed over the years. This step is important for the data to produce efficient results and resolve discrepancies in the data. This will allow for accurate predictions. This can be seen below in Table 4.

Table 4

Year	Datetime	Stage	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	Win conditions	Half-time Home Goals	Half-time Away Goals	Home Team Initials	Away Team Initials
1930.0	13 Jul 1930 - 15:00	Group 1	France	4.0	1.0	Mexico		3.0	0.0	FRA	MEX
1930.0	13 Jul 1930 - 15:00	Group 4	USA	3.0	0.0	Belgium		2.0	0.0	USA	BEL
1930.0	14 Jul 1930 - 12:45	Group 2	Yugoslavia	2.0	1.0	Brazil		2.0	0.0	YUG	BRA
1930.0	14 Jul 1930 - 14:50	Group 3	Romania	3.0	1.0	Peru		1.0	0.0	ROU	PER

After cleaning the World Cup matches dataset, the data is then filtered from the date from 1994 to 2014 and this is stored into a new variable. Then we use the loc method so that it is easy to navigate into the dataset. This is saved in a new variable called FIFA clean. This can be seen below in Table 5.

Table 5

Year	Stage	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	...	Home Team Initials	Away Team Initials
1994.0	Group C	Spain	2.0	2.0	South Korea	...	ESP	KOR
1994.0	Group C	Germany	1.0	0.0	Bolivia	...	GER	BOL
1994.0	Group A	USA	1.0	1.0	Switzerland	...	USA	SUI
1994.0	Group E	Italy	0.0	1.0	Republic of Ireland	...	ITA	IRL
1994.0	Group A	Colombia	1.0	3.0	Romania	...	COL	ROU
...	...	...	...	...	...	...	...	...
2014.0	Quarter-finals	Netherlands	0.0	0.0	Costa Rica	...	NED	CRC
2014.0	Semi-finals	Brazil	1.0	7.0	Germany	...	BRA	GER
2014.0	Semi-finals	Netherlands	0.0	0.0	Argentina	...	NED	ARG
2014.0	Play-off for third place	Brazil	0.0	3.0	Netherlands	...	BRA	NED
2014.0	Final	Germany	1.0	0.0	Argentina	...	GER	ARG

This is the cleaned World Cup matches dataset. Similar steps are performed for the 2018 World Cup matches data is cleaned thoroughly for example unnecessary columns are dropped. This is saved in the new variable called FIFA 2018 clean. This can be seen below in Table 6 for the 2018 matches dataset.

Table 6

Year	Stage	Home Team Name	Away Team Name	Home Team Goals	Away Team Goals
2018	Group-1	Russia	Saudi Arabia	5	0
2018	Group-1	Egypt	Uruguay	0	1
2018	Group-1	Morocco	Iran	0	1
2018	Group-1	Portugal	Spain	3	3
2018	Group-1	France	Australia	2	1
2018	Group-1	Argentina	Iceland	1	1
2018	Group-1	Peru	Denmark	0	1
2018	Group-1	Croatia	Nigeria	2	0
2018	Group-1	Costa Rica	Serbia	0	1
2018	Group-1	Germany	Mexico	0	1
2018	Group-1	Brazil	Switzerland	1	1
2018	Group-1	Sweden	South Korea	1	0
2018	Group-1	Belgium	Panama	3	0
2018	Group-1	Tunisia	England	1	2
2018	Group-1	Colombia	Japan	1	2
2018	Group-1	Poland	Senegal	1	2
2018	Group-2	Russia	Egypt	3	1
2018	Group-2	Portugal	Morocco	1	0

Then we can check for null values in the dataset. If there are none then we can progress to the next step. After this the column date is dropped and converted to date year similar steps performed with the matches dataset. Then names are changed in the 2018 dataset according to the World Cup matches dataset. This is important because by following this step there be no conflicts while merging the two datasets. Another step is to reorder all the columns according to the World Cup matches dataset. Now we append this dataset to the World Cup matches dataset. This will be added after the year 2014. This table can be seen below in Table 7.

Table 7

Year	Stage	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	...	Home Team Initials	Away Team Initials
1994.0	Group C	Spain	2.0	2.0	South Korea	...	ESP	KOR
1994.0	Group C	Germany	1.0	0.0	Bolivia	...	GER	BOL
1994.0	Group A	USA	1.0	1.0	Switzerland	...	USA	SUI
1994.0	Group E	Italy	0.0	1.0	Republic of Ireland	...	ITA	IRL
1994.0	Group A	Colombia	1.0	3.0	Romania	...	COL	ROU
...	...	...	...	...	...	...	...	...
2018.0	QFinals	Russia	2.0	2.0	Croatia	...	NaN	NaN
2018.0	SFinals	France	1.0	0.0	Belgium	...	NaN	NaN
2018.0	SFinals	Croatia	2.0	1.0	England	...	NaN	NaN
2018.0	Third-Place	Belgium	2.0	0.0	England	...	NaN	NaN
2018.0	Final	Croatia	4.0	2.0	France	...	NaN	NaN

The next step is to merge the ranks dataset and matches dataset. This is important as the team ranks is important in calculating the fairness index. In this process duplicate entries are dropped and only the last entry for each year is kept. Then drop all unnecessary columns in this scenario extra columns are dropped from the dataset. This table can be seen below in Table 8.

Table 8

Year	Stage	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	...	rank	total_points	previous_points	rank_change
1994.0	Group C	Spain	2.0	2.0	South Korea	...	2	62	61	3
1994.0	Group C	Germany	1.0	0.0	Bolivia	...	5	61	61	1
1994.0	Group A	USA	1.0	1.0	Switzerland	...	23	48	48	1
1994.0	Group E	Italy	0.0	1.0	Republic of Ireland	...	4	61	61	1
1994.0	Group A	Colombia	1.0	3.0	Romania	...	17	53	53	2

The rank columns added above all then have "home_" appended as a prefix so that they represent the home teams. The next step is to match the columns to get the data for away team for instance such as away rank, away total points, away rank change. This process is applied to capture data for the away team. This can be seen below in Table 9.

Table 9

home_rank	home_total_points	home_previous_points	home_rank_change	away_rank	away_total_points	away_previous_points	away_rank_change
2	62	61	3	35	41	41	0
5	61	61	1	44	37	37	1
23	48	48	1	7	58	58	0
...	...	...	...	...	...	...	...
48	1424	1424	0	4	1634	1634	0
2	1726	1726	0	1	1727	1727	0
4	1634	1634	0	5	1631	1631	0

After this a data frame function is created to extract data by year. This function is applied to get data on each year from 1994 to 2018. An example of this can be seen below in Table 10.

Table 10

Year	Stage	Home Team Name	Home Team Goals	Away Team Goals	Away Team Name	...	home_rank	...	away_rank	...
2018.0	Group-1	Russia	5.0	0.0	Saudi Arabia	...	48	...	69	...
2018.0	Group-1	Egypt	0.0	1.0	Uruguay	...	56	...	7	...
2018.0	Group-1	Morocco	0.0	1.0	Iran	...	40	...	29	...
2018.0	Group-1	Portugal	3.0	3.0	Spain	...	6	...	9	...
2018.0	Group-1	France	2.0	1.0	Australia	...	2	...	41	...
...	...	...	...	...	...	...	...	...	...	...
2018.0	QFinals	Russia	2.0	2.0	Croatia	...	48	...	4	...
2018.0	SFinals	France	1.0	0.0	Belgium	...	2	...	1	...
2018.0	SFinals	Croatia	2.0	1.0	England	...	4	...	5	...
2018.0	Third-Place	Belgium	2.0	0.0	England	...	1	...	5	...
2018.0	Final	Croatia	4.0	2.0	France	...	4	...	2	...

6 Results

6.1 Visualisations

This section is written by Shrey Parekh (18706941) and Pratik Kulkarni (19400570)

The visualisations provide a deeper insight in the World Cup matches and World Cup results from 1994 to 2018. It will provide the trends between the previous World Cups and the patterns of the winning team. The approach for all the visualisations is referenced from Kaggle Notebook shivan118. The first table is the results of the Countries in the World Cup. This table can be seen below in Table 11.

Table 11

Year	Host	Champion	Runner-Up	Third	Fourth	Teams	Matches Played	Goals Scored	Avg Goals Per Game
1994	United States	Brazil	Italy	Sweden	Bulgaria	24	52	141	2.7
1998	France	France	Brazil	Croatia	Netherlands	32	64	171	2.7
2002	South Korea, Japan	Brazil	Germany	Turkey	Korea Republic	32	64	161	2.5
2006	Germany	Italy	France	Germany	Portugal	32	64	147	2.3
2010	South Africa	Spain	Netherlands	Germany	Uruguay	32	64	145	2.3
2014	Brazil	Germany	Argentina	Netherlands	Brazil	32	64	171	2.7
2018	Russia	France	Croatia	Belgium	England	32	64	169	2.6

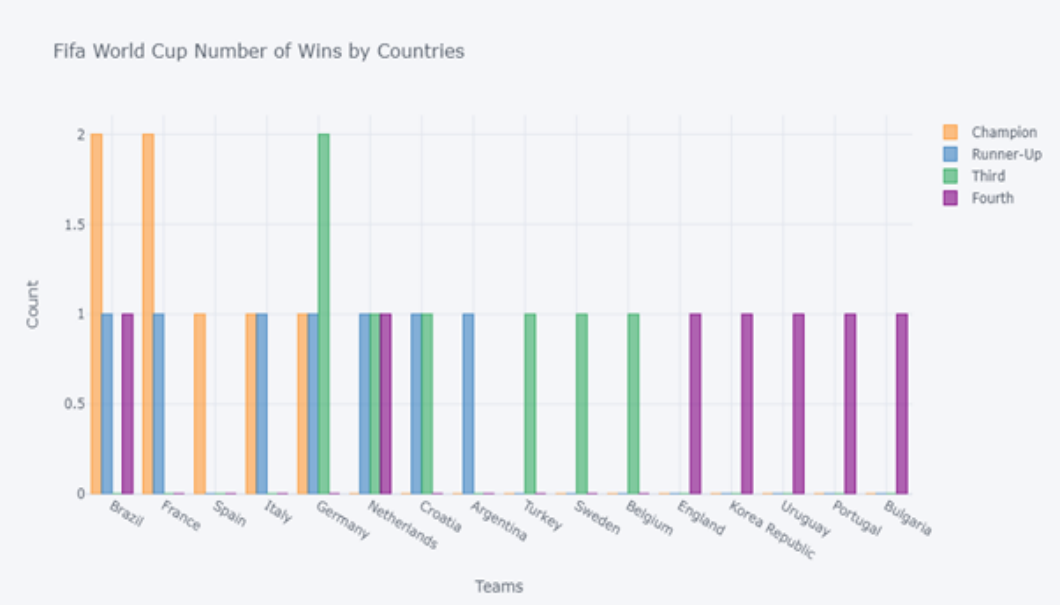
The second table is the total of Cups won by the Countries in FIFA. This table can be seen below in Table 12.

Table 12

	Champion	Runner-Up	Third	Fourth
Brazil	2	1	0	1
France	2	1	0	0
Germany	1	1	2	0
Spain	1	0	0	0
Italy	1	1	0	0
Netherlands	0	1	1	1
Argentina	0	1	0	0
Croatia	0	1	1	0
Belgium	0	0	1	0
Sweden	0	0	1	0
Turkey	0	0	1	0
Portugal	0	0	0	1
Uruguay	0	0	0	1
Bulgaria	0	0	0	1
England	0	0	0	1
Korea Republic	0	0	0	1

Now we will visualise the above table through a bar plot, seen below in Figure 5.

Figure 5



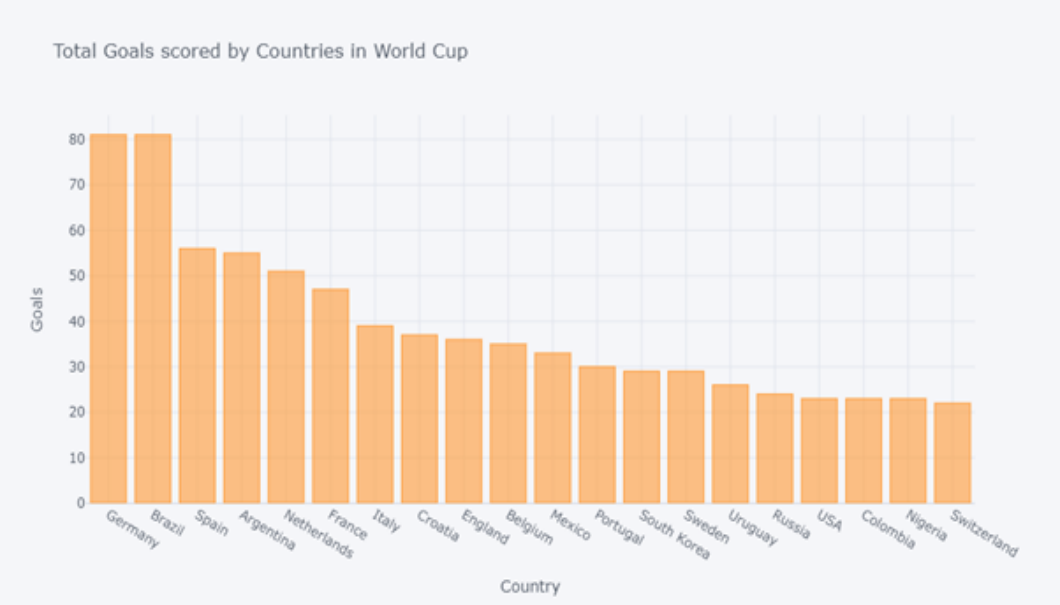
From the above graph the trends can be gathered such as Brazil and Germany have been crowned the highest times champion from 1994 to 2018. We can also assume that Brazil and Germany are strong FIFA candidates as they have also earned other positions in the World Cup such as runner up, places such as third or fourth. This makes them a high candidate for winning 2022 FIFA World Cup.

Next, we will visualise the number of goals scored by Countries as seen in Table 13 below.

Table 13

Countries	Goals
Germany	81
Brazil	81
Spain	56
Argentina	55
Netherlands	51
...	...
Korea DPR	1
Bolivia	1
Angola	1
Trinidad and Tobago	0
China PR	0

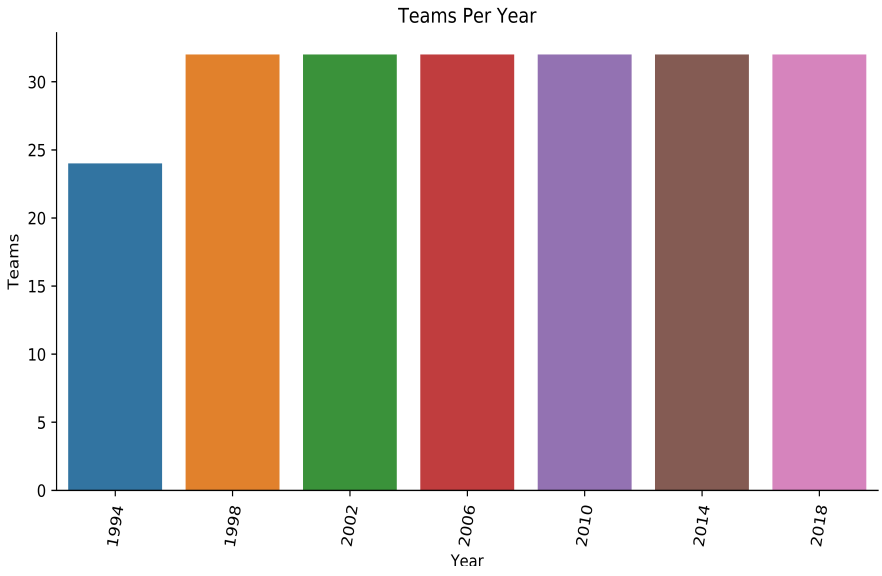
Figure 6



From the graph above in Figure 6, we can depict that the highest scoring teams to the lowest scoring teams from left to right. This can be seen from Germany, Brazil being top contenders for the 2022 World Cup. This is due to the number of goals scored they are tied to 81 each. Then Spain scoring 56, Argentina 55 and Netherlands scoring 51 goals. Similarly, the rest of the team scores goals.

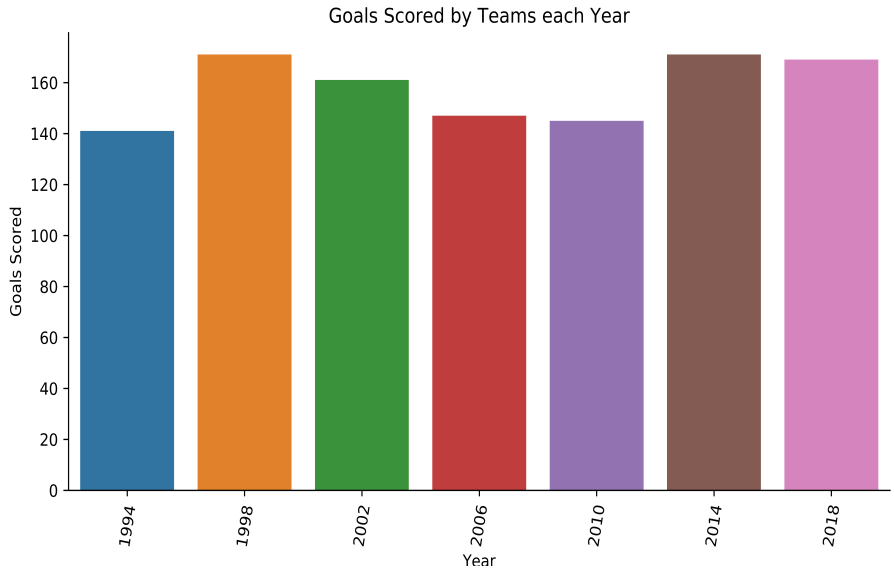
Next visualising the goals in each World Cup from 1994 to 2018. The first visualisation is the teams in each World Cup. This is seen below in Figure 7.

Figure 7



The above visualisation demonstrates there have been 24 teams competing in the FIFA World Cup. Then from 1998 to 2018 there have been 32 teams participating in the FIFA World Cup.

Figure 8



The above visualisation in Figure 8 highlights goals scored by teams from 1994 to 2018. In 1994 the goals scored were 140, then increased to 170 in 1998, 160 in 2002, 150 in 2006, 140 in 2010. Then 170 in 2014 and 169 in 2018. These insights give us a speculation that FIFA World Cup has been competitive in the 4 years between 2014 to 2018 World Cup. Also, we can assume that 2022 World Cup will be challenging.

The below plot in Figure 9 shows the matches played on average from 1994 to 2018. The insights we gain from the plot is that an average of 60 matches are played from 1998 to 2018 before only 50 matches were played.

Figure 9

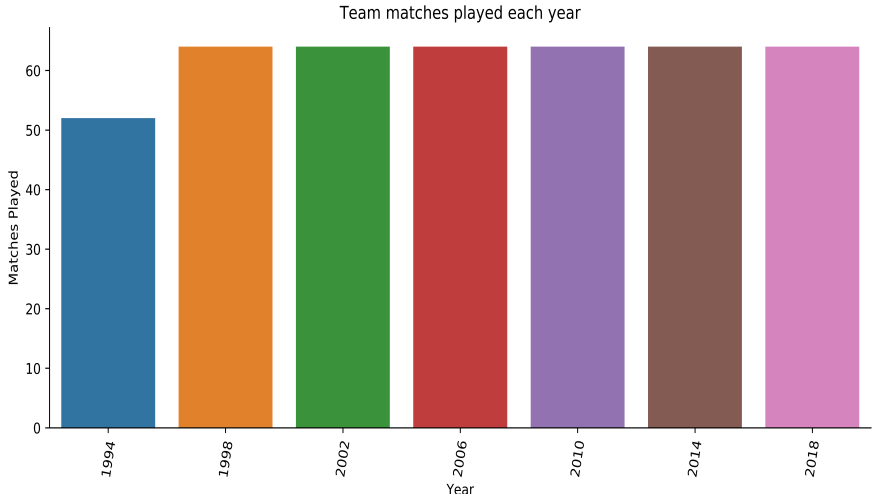
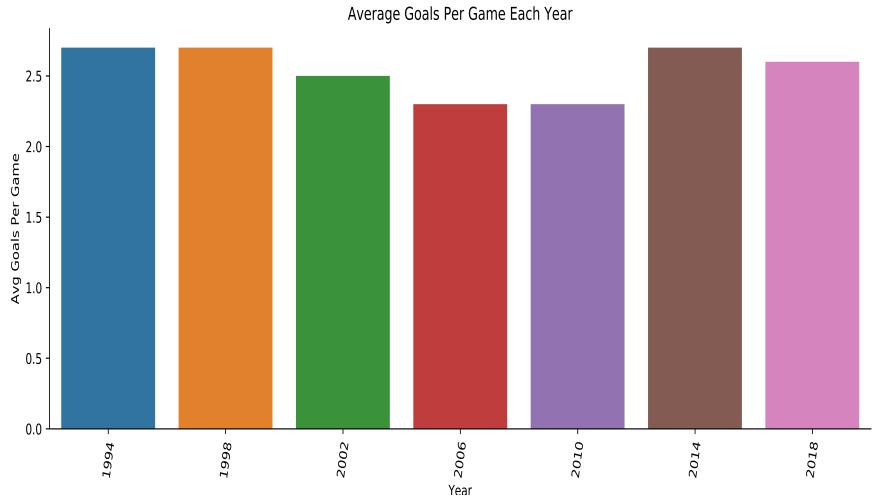


Figure 10



The trends depicted through the above visualisation in Figure 10 is average goals per game is from 2.0 to 2.5 from 1994 to 2018.

Now we will visualise goals of teams in each FIFA World Cup. This can be seen from the below Table 14.

Table 14

		Goals
Year		
	Argentina	8
	Belgium	4
1994	Bolivia	1
	Brazil	11
	Bulgaria	10
...	...	...
	Spain	7
	Sweden	6
2018	Switzerland	5
	Tunisia	5
	Uruguay	7

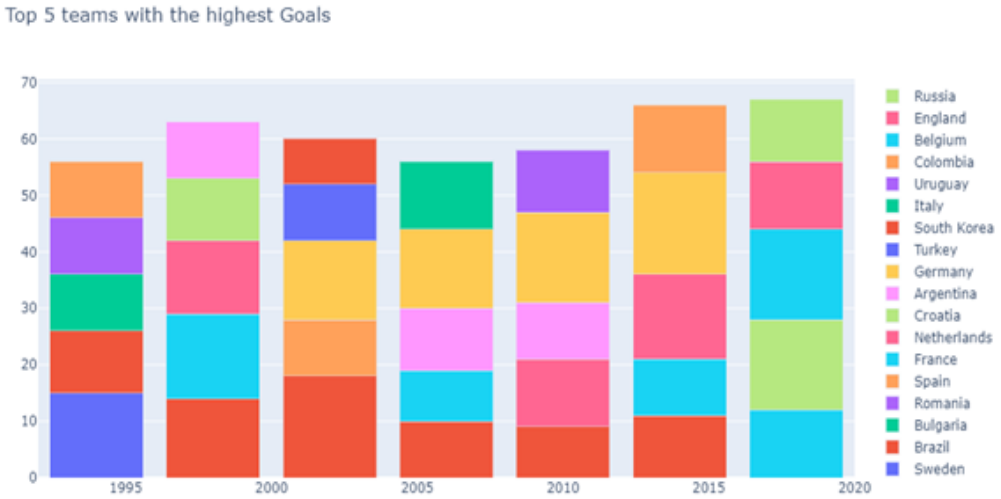
Next, we will view the merged table that consists of the Country, Goals and Year. This table can be seen below in Table 15.

Table 15

Year	Country	Goals
1994	Sweden	15
1994	Brazil	11
1994	Bulgaria	10
1994	Romania	10
1994	Spain	10
...	...	...
2018	Panama	2
2018	Peru	2
2018	Poland	2
2018	Saudi Arabia	2
2018	Serbia	2

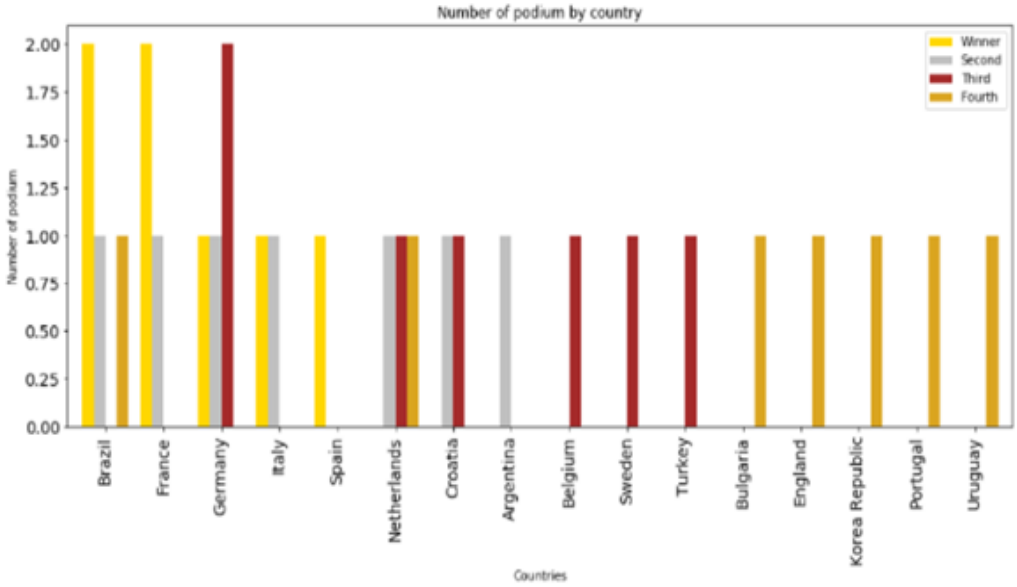
Now we will visualise the top 5 teams of the FIFA World Cup from 1994 to 2018. The trends that can be gathered from the below plot is the top 5 contenders based on the number of goals scored by each team. This allows us to see who the top teams in 2022 World Cup that will reach in the qualifying rounds based on the number of goals scored. It allows to assess the performance of teams. Their winning ratio to win the World Cup in 2022. This is seen below in Figure 11.

Figure 11



The next visualisation is about the Countries that have won the FIFA World Cup. This is seen below in Figure 12.

Figure 12

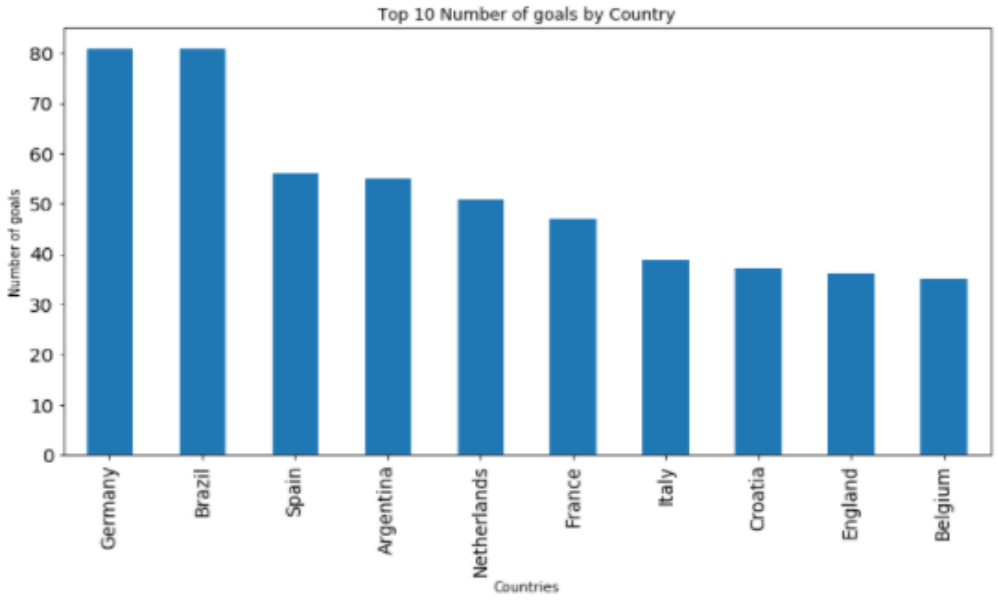


From the above plot it can be deduced that Brazil and France have won the most FIFA World Cups and have kept positions in the World Cup such as second place and third place. Then Germany, Italy and Spain have won titles and Netherlands has not won any titles but has kept position from second to fourth in the cups. This establishes concrete evidence from the observations of the strong teams to look out for to win the 2022 World Cup and the teams holding third and fourth place. They

can move to third, second place in the 2022 World Cup. This also demonstrates that some of the teams might go to qualifying rounds.

Next visualisation of Top 10 goals performed by Countries in the World Cup. This is seen below in Figure 13.

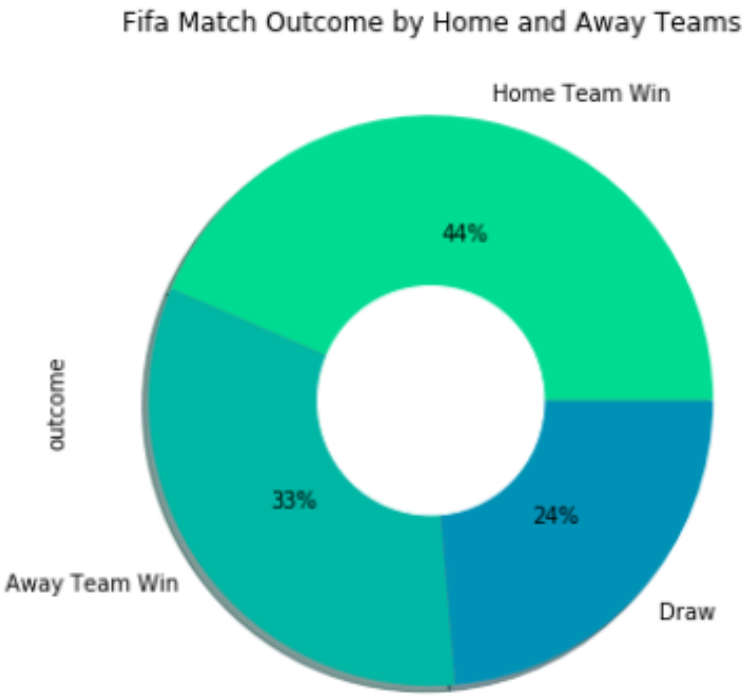
Figure 13



This is a simple visualisation. The highest scoring teams are from left to right the lowest scoring teams. This demonstrates that Brazil and Germany are top contenders to win and then teams Spain to France based on goals scored, then Italy to Belgium.

The below visualisation in Figure 14 indicates what teams are likely to win the FIFA World Cup when hosting the tournament or playing at an away ground. The plot demonstrates that Home Teams have a 44 percent chance of winning the Cup. Away Teams have a chance of 33 percent winning the cup. There is also 24 percent chance of the game being a draw. This demonstrates that the host of 2022 FIFA World Cup has a high percent of winning due to playing on home grounds but there could also be a chance of opponent team winning, or the game can also be a draw. These 3 factors are unpredictable and could be determining in the winner of FIFA World Cup 2022.

Figure 14


Fifa Match Outcome by Home and Away Teams
Home Team Win
44%
outcome
33%
24%
Away Team Win
Draw

6.2 Fairness Index Implementation

This is an important step we will use the equation to calculate the fairness index:

This section is written by Pratik Kulkarni (19400570)

$$lm(FinalPos \sim TeamRank + Fairness)$$

The fairness index is created for home and away teams. In the dataset the fairness index is created from home and away teams, grouped by stage. The formula to calculate this is:

```
# Load in data per game
games <<- read.csv(
paste("fifa", year, "Fairness.csv", sep = ""),
header = TRUE)

# Make Data frame that only includes group games
groupgames <- games[grepl("Group",games$Stage),]

# Function to calculate Fairness index per team
fairness <- function(teamname) {
        return( sum( groupgames$away_total_points
        [groupgames$Home.Team.Name==teamname],
        groupgames$home_total_points
        [groupgames$Away.Team.Name==teamname]))
}

# Create data frame of teams
teams <- data.frame(name = levels(factor(games$Home.Team.Name)))


# Apply function to each team in the World Cup
teams$fairness = sapply(teams$name,fairness)
```

Then we apply this function to each of the years to get the fairness index. This method is applied inside the R file. The R file is accessed through the python notebook. A function is created where we retrieve the year, country, number of games played, fairness index, team rank and the final position. It is then applied similarly to all datasets to generate the fairness index. A sample of the first 10 rows can be seen below in Table 16.

Table 16

Year	Country	NumGames	Fairness	TeamRank	FinalPos
2018.0	Argentina	4	4513	11	12.5
2018.0	Australia	3	4833	41	24.5
2018.0	Belgium	7	4450	1	3.0
2018.0	Brazil	5	4544	3	6.5
2018.0	Colombia	4	4437	12	12.5
2018.0	Costa Rica	3	4756	36	24.5
2018.0	Croatia	7	4461	4	2.0
2018.0	Denmark	4	4680	10	12.5
2018.0	Egypt	3	4368	56	24.5
2018.0	England	7	4546	5	4.0

All the datasets from 1994 to 2018 are merged by the concat pandas' method. This can be seen in the table below. This is then exported to a csv file as seen below in Table 17.

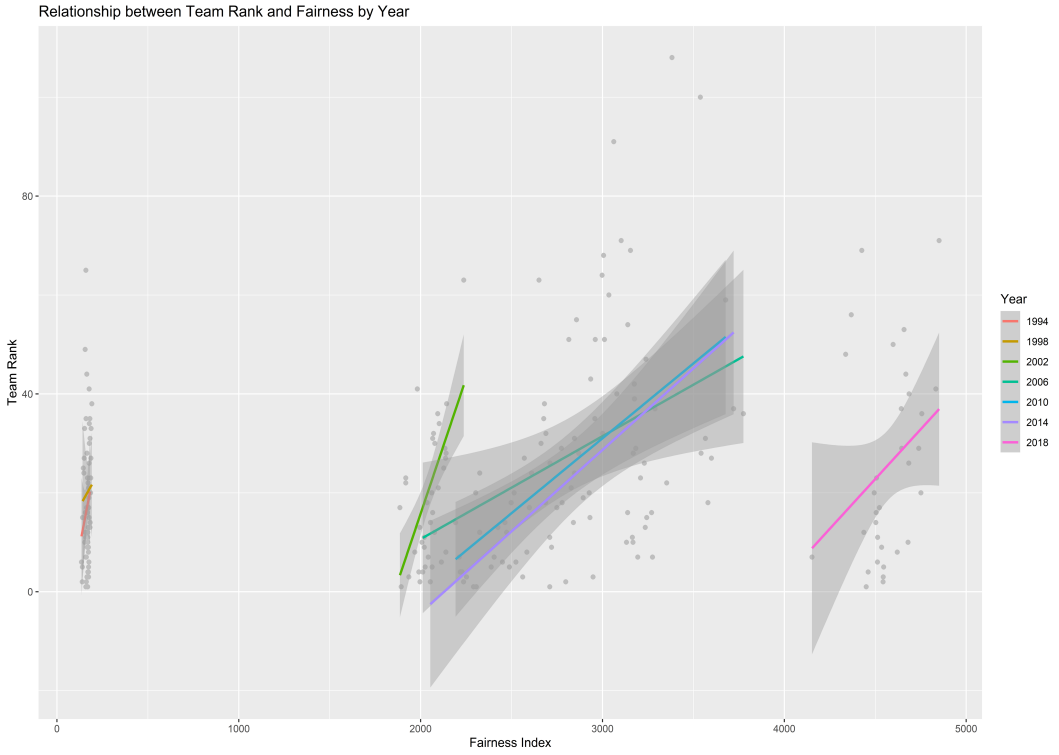
Table 17

Year	Country	NumGames	Fairness	TeamRank	FinalPos
1994.0	Argentina	4	153	10	12.5
1994.0	Belgium	4	148	24	12.5
1994.0	Bolivia	3	164	44	24.5
1994.0	Brazil	7	159	1	1.0
1994.0	Bulgaria	7	156	16	4.0
...	...	...	...	...	...
2018.0	Spain	4	4535	9	12.5
2018.0	Sweden	5	4503	14	6.5
2018.0	Switzerland	4	4621	8	12.5
2018.0	Tunisia	3	4684	26	24.5
2018.0	Uruguay	5	4152	7	6.5

6.3 Linear Modelling

Figure 15 below shows the relationship between Team Rank and Fairness by Year. It can be seen that lower ranked teams are associated with a higher fairness index.

Figure 15



Successive predictive models with FinalPos as response variable were created, to assess and compare the predictive accuracy at alpha symbol = 0.05.

TeamRank was entered as the predictor in Model 1 and it was significant, p-value < 2e-16, explaining 36.6% of variation in FinalPos outcome.

Model 2 had only Fairness as the predictor, and the model was significant, p-value = 0.0313, explaining 2.6% of variation in FinalPos outcome.

Model 3 had TeamRank and Fairness as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 36.6% of variation in FinalPos outcome. TeamRank was a significant predictor while Fairness was not significant.

Model 4 had TeamRank, Fairness and their interaction as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 40.7% of variation in FinalPos outcome. The interaction term was significant in this model.

Model 5 had TeamRank, Fairness and NumGames as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 87.2% of variation in FinalPos outcome. TeamRank and NumGames were significant predictors, while Fairness was

not a significant predictor.

The coefficients of the linear models are presented below in Table 18.

Table 18

	Predictors	Coefficient	Significance	R-Squared	MSE
Model 1				36.55%	50.14
	TeamRank	0.2776	***		
Model 2				2.64%	72.59
	Fairness	0.0012	*		
Model 3				36.55%	50.07
	TeamRank	0.2784	***		
	Fairness	0.0000			
Model 4				40.65%	56.71
	TeamRank	0.4842	***		
	Fairness	0.0018	**		
	TeamRank*Fairness	-0.0001	***		
Model 5				87.16%	10.09
	TeamRank	0.0572	***		
	Fairness	0.0000			
	NumGames	-5.4760	***		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The R-Outputs of all Models are in Appendix A.

7 Results and Conclusions

This section is written by Shrey Parekh (18706941) and Pratik Kulkarni (19400570)

The FIFA World Cup draw is not entirely random as it is dependent on a variety of factors such as the performance of the team in previous World Cups. This enables their rank to increase for the next World Cup, so they have a higher chance of being selected. This also affects the draw as it can be demonstrated through other factors such as hosting the cup, playing on home grounds or away grounds. This can be highlighted through the visualisations as it depicts that playing on home grounds has a higher win potential then away teams. This also implies that there is a slight chance of bias, and the draw is not entirely random as the Country that hosts the championship are qualified to participate. They are guaranteed a spot to participate in the cup.

The visualisations also indicate who the next winner of the cup might be due to their previous performance such as goals, number of the titles held by the team such as Brazil and Germany. Another aspect to consider from the visualisations is that to not underestimate average performance of teams such as France, Netherlands. Although they have not scored much, they have managed to grab the FIFA World Cup title. France has won the 2018 FIFA World Cup Championship. This statistic provides information that depending on the factors illustrated anyone can win the 2022 FIFA World Cup and draws are not entirely random.

It can be seen from the linear models that the most important predictors are Team Rank and Number of Games played while Fairness does not affect the overall outcome for Final Position in the World Cup. The lower ranked teams had shown higher fairness index over 1994-2018 period and the interaction between Team Rank and Fairness had a significant effect on the outcome. Furthermore, the number of games played was also a significant predictor. This indicates that the lower ranked teams are disadvantaged and get placed with the higher ranked teams in the group stage which helps the stronger teams progress into further rounds while the weaker teams get knocked out.

Through the tests conducted in R and Python on the randomisation of the groupings and the visualisations, we can deduce that the FIFA World Cup draw is not truly random. This is due to it depending on several factors such as hosting Country, performance, team rank obtained from the analysis of this project.

8 Future Goals

This section is written by Pratik Kulkarni (19400570)

The project can be enhanced further in the future if more data was sourced such as prize money, population size of country, past history of teams. These could be used as additional predictors to see the effect they would have towards predicting fairness or even final position of the teams in the FIFA World Cups. Additional data would help strengthen the models created, allow for new unique visualisations, and give a more thorough understanding of the data.

Additionally, in the future, a proper simulation can also be run to simulate the World Cup draw, to see if any particular countries have a fairness index that is higher or lower than expected by chance alone. This was originally planned to be covered in this project report but due to time constraints surrounding the project, this task was not feasible to be explored and completed thoroughly.

References

ccn8. cnc8/fifa-world-ranking: Fifa world ranking scraper. *GitHub*, Jan 2020. URL https://github.com/cnc8/fifa-world-ranking.

Sebastián Cea, Guillermo Durán, Mario Guajardo, Denis Sauré, Joaquín Siebert, and Gonzalo Zamorano. An analytics approach to the fifa ranking procedure and the world cup final draw. *Annals of Operations Research*, 286(1):119–146, 2020.

Launay Christian. World cup prediction. *Kaggle*, Oct 2018. URL https://www.kaggle.com/launay10christian/world-cup-prediction.

Disha. Fifa-worldcup data visualization. *Kaggle*, Jul 2018. URL https://www.kaggle.com/dicoderdisha/fifa-worldcup-data-visualization.

Shivan Kumar. Fifa world cup (data analysis). *Kaggle*, Sept 2020. URL https://www.kaggle.com/shivan118/fifa-world-cup-data-analysis.

Nathan Lauga. Data visualization of fifa world cup. *Kaggle*, May 2018. URL https://www.kaggle.com/nathanlauga/data-visualization-of-fifa-world-cup.

Mrthlinh. Fifa-world-cup-prediction/report.md at master · mrthlinh/fifa-world-cup-prediction. *GitHub*, Jul 2018. URL https://github.com/mrthlinh/FIFA-World-Cup-Prediction.

Rodrigo Nader. Using machine learning to simulate world cup matches. *Towards Data Science*, Jul 2018. URL https://towardsdatascience.com/using-machine-learning-to-simulate-world-cup-matches-959e24d0731.

University of Washington. How to write your discovery project proposal. *University of Washington School of Medicine III Requirement*, 2018. URL https://sites.uw.edu/somcurr2/iii-scholarship-requirement/scholarship-of-discovery/how-to-write-your-project-proposal/.

Abhinav Raghunathan. Simulating the fifa world cup 2022. *Towards Data Science*, Dec 2020. URL https://towardsdatascience.com/simulating-the-fifa-world-cup-2022-d363fad7da22.

9 Appendix A

Model 1 Summary:

Figure 16

```
summary(model1)

##
## Call:
## lm(formula = FinalPos ~ TeamRank, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3390  -5.0205  -0.7666   6.0984  14.1479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.79699    0.81400   12.04   <2e-16 ***
## TeamRank     0.27757    0.02773   10.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.932 on 174 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3618
## F-statistic: 100.2 on 1 and 174 DF,  p-value: < 2.2e-16
```

Model 2 Summary:

Figure 17

```
summary(model2)

##
## Call:
## lm(formula = FinalPos ~ Fairness, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.165  -7.604  -1.598   7.483  10.418
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.392e+01  1.174e+00  11.855   <2e-16 ***
## Fairness    1.160e-03  5.344e-04   2.171   0.0313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 174 degrees of freedom
## Multiple R-squared:  0.02637,    Adjusted R-squared:  0.02078
## F-statistic: 4.713 on 1 and 174 DF,  p-value: 0.03129
```

Model 3 Summary:

Figure 18

```
summary(model3)

##
## Call:
## lm(formula = FinalPos ~ TeamRank + Fairness, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4582  -5.0004  -0.7949   6.0332  14.1738
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.867e+00  1.040e+00   9.490   <2e-16 ***
## TeamRank     2.784e-01  2.896e-02   9.616   <2e-16 ***
## Fairness    -4.893e-05  4.505e-04  -0.109    0.914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.952 on 173 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3582
## F-statistic: 49.83 on 2 and 173 DF,  p-value: < 2.2e-16
```

Model 4 Summary:

Figure 19

```
summary(model4)

##
## Call:
## lm(formula = FinalPos ~ TeamRank * Fairness, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2249  -4.9781  -0.7494   5.5535  14.5968
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.833e+00  1.545e+00   3.774 0.000220 ***
## TeamRank           4.842e-01  6.600e-02   7.337 8.27e-12 ***
## Fairness           1.810e-03  6.942e-04   2.607 0.009950 **
## TeamRank:Fairness -8.407e-05  2.440e-05  -3.445 0.000717 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.743 on 172 degrees of freedom
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.3961
## F-statistic: 39.27 on 3 and 172 DF,  p-value: < 2.2e-16
```

Model 5 Summary:

Figure 20

```
summary(model5)

##
## Call:
## lm(formula = FinalPos ~ TeamRank + Fairness + NumGames, data = train1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.300 -3.392  1.570  2.598  4.933
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.698e+01  1.142e+00  32.381  < 2e-16 ***
## TeamRank     5.723e-02  1.558e-02   3.673  0.00032 ***
## Fairness     3.042e-05  2.033e-04   0.150  0.88122
## NumGames    -5.476e+00  2.103e-01 -26.039  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.136 on 172 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8694
## F-statistic: 389.2 on 3 and 172 DF,  p-value: < 2.2e-16
```