Python Project by Arjun Kumar- 045012

Submitted to: Professor Amarnath Mitra

FORE School of Management, New Delhi

BDA-04, Batch of 2023-2025

The **FIFA World Cup**, often simply called the **World Cup**, is an international association football competition between the senior men's national

teams of the members of the *Fédération Internationale de Football Association* (FIFA), the sport's global governing body. The tournament has been held every four years since the inaugural tournament in 1930, with the exception of 1942 and 1946 due to the Second World War. The reigning champions are Argentina, who won their third title at the 2022 tournament.

ANALYSIS OF FIFA WORLD CUP (MEN'S) DATA SET (1930-2022) FROM WIKIPEDIA

URL: https://en.wikipedia.org/wiki/FIFA_World_Cup

Libraries used:

Web Scrapping tools:

1. **Python**
2. **Beautiful Soup**
3. **Requests**

Cleaning tools:

Pandas

Matplotlib

Other tools:

Regression Analysis

Managerial Analysis

OBJECTIVES:

1.To show the total attendance in an edition and analyse our findings based on the observations.

2. To show the highest and lowest average attendance and analyse the findings and other insights and plot a graph to show relationship between matches and years using Matplotlib function.

3. To find the maximum number of times a country has hosted the world cup, showing the interest levels of football in these countries and the finances and support they receive from their governments.

4. To find the continents, from the set of countries hosting the World Cup and analyse why countries in Asia and Africa do not host the World Cup regularly.

5. Perform Regression Analysis between Matches and Years through analysis in detail.

6. Predict the number of matches in the 2026 edition using our regression model (subject to certain conditions)

7. Perform Regression Analysis between Total Attendance and years

8. Predict the total no of attendance in the year 2026 (subject to certain conditions)

9. Analyse top 10 player's performances in all editions through goals per game ratio.

10. Extract results of different countries in FIFA World Cup and analyse it by parameters like which team has been the most successful in the history of the cup.

ANALYSIS:

When we type the initial code, we get the following snapshot which is shown in the code section as well.

Here we have some points to note:

Points to note

1. Here we do not have data for 1942 and 1946 editions as the FIFA World Cup was not played due to the World War 2 crisis from 1939 to 1945.
2. Here the game column is the games which had the highest attendance
3. and with it is the corresponding venues, where these games with highest attendances were played.

	Year	Hosts	Venues/ Cities	Total attendance †	Matches	\
	Year	Hosts	Venues/ Cities	Total attendance †	Matches	
0	1930	Uruguay	3/1	590549.0	18	
1	1934	Italy	8/8	363000.0	17	
2	1938	France	10/9	375700.0	18	
3	1950	Brazil	6/6	1045246.0	22	
4	1954	Switzerland	6/6	768607.0	26	
5	1958	Sweden	12/12	819810.0	35	
6	1962	Chile	4/4	893172.0	32	
7	1966	England	8/7	1563135.0	32	
8	1970	Mexico	5/5	1603975.0	32	
9	1974	West Germany	9/9	1865753.0	38	
10	1978	Argentina	6/5	1545791.0	38	
11	1982	Spain	17/14	2109723.0	52	
12	1986	Mexico	12/11	2394031.0	52	
13	1990	Italy	12/12	2516215.0	52	
14	1994	United States	9/9	3587538.0	52	
15	1998	France	10/10	2785100.0	64	
16	2002	South Korea Japan	20/20	2705197.0	64	
17	2006	Germany	12/12	3359439.0	64	
18	2010	South Africa	10/9	3178856.0	64	
19	2014	Brazil	12/12	3429873.0	64	

	Average attendance	Highest attendances ‡	\
	Average attendance	Number	
0	32808.0	93000	
1	21353.0	55000	
2	20872.0	58455	
3	47511.0	173,850[94]	
4	29562.0	63000	
5	23423.0	50928	
6	27912.0	68679	
7	48848.0	98270	
8	50124.0	108192	
9	49099.0	83168	
10	40679.0	71712	
11	40572.0	95500	

	Venue
0	Estadio Centenario, Montevideo
1	Stadio Nazionale PNF, Rome
2	Olympique de Colombes, Paris
3	Maracanã Stadium, Rio de Janeiro
4	Wankdorf Stadium, Bern
5	Ullevi Stadium, Gothenburg
6	Estadio Nacional, Santiago
7	Wembley Stadium, London
8	Estadio Azteca, Mexico City
9	Olympiastadion, Munich
10	Estadio Monumental, Buenos Aires
11	Camp Nou, Barcelona
12	Estadio Azteca, Mexico City
13	San Siro, Milan
14	Rose Bowl, Pasadena, California
15	Stade de France, Saint-Denis
16	International Stadium, Yokohama, Japan
17	Olympiastadion, Berlin
18	Soccer City, Johannesburg
19	Maracanã Stadium, Rio de Janeiro
20	Luzhniki Stadium, Moscow
21	Lusail Stadium, Qatar

	Game(s)
0	Uruguay 6-1 Yugoslavia, semi-final
1	Italy 2-1 Czechoslovakia, final
2	France 1-3 Italy, quarter-final
3	Brazil 1-2 Uruguay, deciding match
4	West Germany 3-2 Hungary, final
5	Brazil 2-0 Soviet Union, group stage
6	Brazil 4-2 Chile, semi-final
7	England 4-2 West Germany, final
8	Mexico 1-0 Belgium, group stage
9	West Germany 1-0 Chile, group stage
10	Italy 1-0 Argentina, group stage
11	Argentina 0-1 Belgium, Opening match
12	Mexico 1-1 Paraguay, group stage Argentina 3-2...
13	West Germany 4-1 Yugoslavia, group stage
14	Brazil 0-0 (3-2p) Italy, final
15	Brazil 0-3 France, final
16	Brazil 2-0 Germany, final
17	Germany 1-1 (4-2p) Argentina, quarter-final
18	Spain 1-0 Netherlands, final
19	Germany 1-0 Argentina, final
20	France 4-2 Croatia, final
21	Argentina 3-3 (4-2p) France, final

1. After rearranging the data for total attendance in descending order we can infer the following:

```
Total attendance †  \
Total attendance †
        43936730.0
         3587538.0
         3429873.0
         3404252.0
         3359439.0
         3178856.0
         3031768.0
         2785100.0
         2705197.0
         2516215.0
         2394031.0
         2109723.0
         1865753.0
         1603975.0
         1563135.0
         1545791.0
         1045246.0
          893172.0
          819810.0
          768607.0
          590549.0
          375700.0
```

1. We can say that The US edition of the FIFA world cup in 1994 had the highest total attendance followed by Brazil in 2014 and Qatar in 2022.

#1.1. This is due to the combination of large stadiums in the United States and the sports loving culture of its people. #

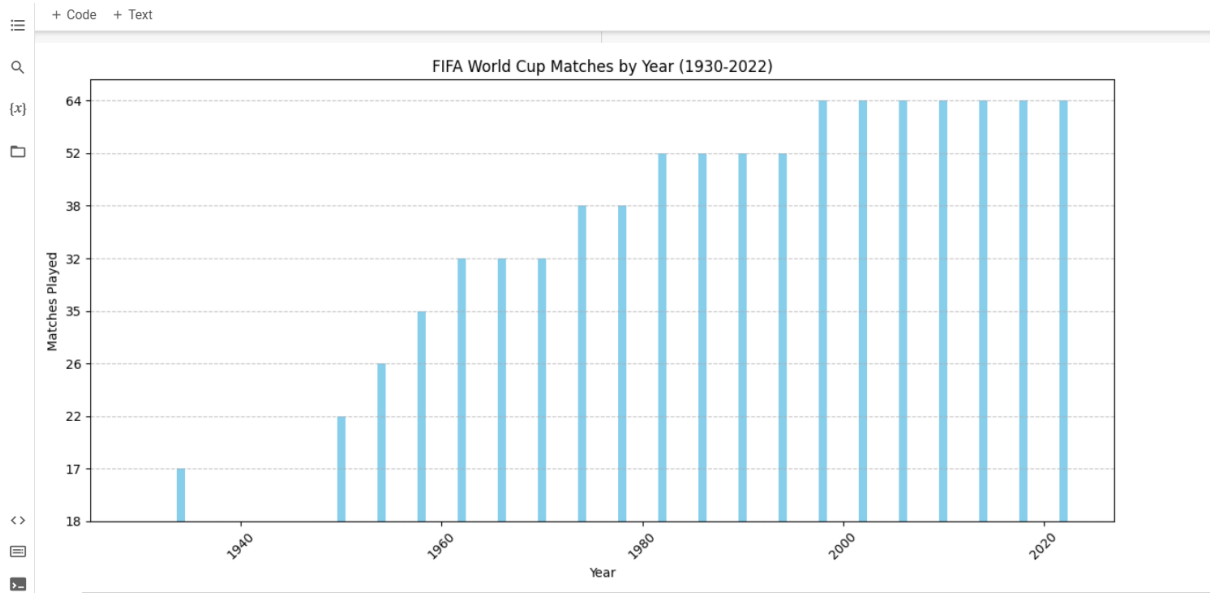
Next, we show the highest attendance in a single game, highest average and lowest average attendance and infer the findings:

	Matches	Average attendance	Highest attendances ‡	\
	Matches	Average attendance	Number	
14	52	68991.0	94194	
19	64	53592.0	74738	
21	64	53191.0	88966	
17	64	52491.0	72000	
8	32	50124.0	108192	
18	64	49670.0	84490	
9	38	49099.0	83168	
7	32	48848.0	98270	
13	52	48389.0	74765	
3	22	47511.0	173,850[94]	
20	64	47371.0	78011	
12	52	46039.0	114600	
23	964	45577.0	173,850[94]	
15	64	43517.0	80000	
16	64	42269.0	69029	
10	38	40679.0	71712	
11	52	40572.0	95500	
0	18	32808.0	93000	
4	26	29562.0	63000	
6	32	27912.0	68679	
5	35	23423.0	50928	

#2 We see that the US edition of the FIFA World Cup in 1994 had the highest average attendance of 68,991, followed by Brazil in 2014 and Qatar in 2022.
#2.1 This is due to the fact that USA has many large stadiums with highest capacities in the world, for example Rose Bowl. Also, they are a sports loving nation and believe that all round growth happens with the combination of sports and studies.
#2.2 Although Qatar had good average attendances, there was labor mistreatment in Qatar during building of the new stadiums, which portrayed a bad image of the World Cup 2022 edition and hence attendance was lower than expected.
#2.3 The final between Brazil and Italy in 1994 was the most viewed game according to our data with close to 94,000 fans in the stadium.
#2.3 The final between Argentina and France in 2022 was the highest viewed final. Although our data shows us 88,966 people attending it, there was online watching, making it 1.5 million viewers in total.
#2.4 This was due to the fact that it was the legend, Lionel Messi's (Argentina) last World Cup and fans did not want to miss it. He scored 13 goals in 26 matches in every edition he played.
#2.5 The least average attendance was in the 1938 World Cup in France with 20,872 fans.

#2.6 Because of anger over the decision to hold a second successive World Cup in Europe, Uruguay and Argentina refused to enter the competition.
#2.7 Spain meanwhile could not participate due to the ongoing Spanish Civil War. Due to these reasons the attendances in this edition was specifically low.

Given below is a bar graph showing the relationship between matches and the years in which the World Cup was played. Here we can see that the matches have increased year on year and have stagnated to 64 from the 2000's. So as the years progress, the matches have increased and stayed at 64 for a long period of time. It is reported that matches will increase due to increase in teams to 48 in 2026 so the matches will also increase. I have used regression analysis to predict the no. of matches in 2026 through my model.



Now, we have written a code which gives the list of the countries who hosted the FIFA World Cup the greatest number of times and here is the snapshot below:

```
Country(s) that hosted the maximum number of times:
(Hosts, Hosts)
Brazil    2
France    2
Italy     2
Mexico    2
dtype: int64
```

#3. From this we can analyse the countries which hosted the world cup the maximum number of times in history.

#3.1. Here, we can see that Brazil,France,Italy and Mexico have hosted the world cup twice and the others have hosted it only once.
#3.2. This tells us that these countries have a lot of wealth and stability as they can finance such a mega event.
#4.1 Hosting a World Cup of such stature requires huge government support and techincal expertise, which everyone cannot afford as they may not have enough finance and their priorities are different.
#4.2 This also shows how passionate these countries are about football and is also a major source of revenue to these countries through which their economy runs.

Now, we find the continents of the countries where they FIFA World Cup was hosted and find out why Asia and Africa have hosted the cup the least number of times, below is the snapshot:

	Hosts	Year	Continent
	Hosts	Year	
0	Uruguay	1930.0	South America
1	Italy	1934.0	Europe
2	France	1938.0	Europe
3	Brazil	1950.0	South America
4	Switzerland	1954.0	Europe
5	Sweden	1958.0	Europe
6	Chile	1962.0	South America
7	England	1966.0	Europe
8	Mexico	1970.0	North America
9	West Germany	1974.0	Europe
10	Argentina	1978.0	South America
11	Spain	1982.0	Europe
12	Mexico	1986.0	North America
13	Italy	1990.0	Europe
14	United States	1994.0	North America
15	France	1998.0	Europe
16	South Korea Japan	2002.0	NaN
17	Germany	2006.0	Europe
18	South Africa	2010.0	Africa
19	Brazil	2014.0	South America
20	Russia	2018.0	Europe
21	Qatar	2022.0	Asia

#5. Like in India, there is a huge craze for cricket and not much popularity for football, which is evident from the craze of IPL and Cricket World Cup.
#5.1 People take special leave from work to watch matches involving India.

#5.2 In countries like Europe, North America and South America, people love football and prioritise it. This is shown in the table above, where countries like Brazil, Germany have hosted the world cup more than once.
#6. We can also see that Asia and Africa have hosted the World Cup the least number of times, Asia twice and Africa only once.
#6.1 This indicates that the countries in these continents lack the financial resources and government support, the players in this region are not developed to such an extent that they can impress their fans on a global stage.
#6.2 There is poverty and unemployment in Asia and Africa, greater than in Europe and America and they have to tend to their residents' requirements first like food, water, electricity, etc.
#6.3 Hence they do not have the resources and needs to host the World Cup.

Now, I have performed a regression analysis to find the relationship between the matches and the years and these were my findings.

+ Code + Text

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Matches   R-squared:                       0.879
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     153.1
Date:                Fri, 08 Sep 2023   Prob (F-statistic):           4.13e-11
Time:                        12:59:33   Log-Likelihood:                -78.164
No. Observations:                  23   AIC:                             160.3
Df Residuals:                      21   BIC:                             162.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1327.1395    111.016    -11.954      0.000   -1558.011   -1096.268
Year           0.6934      0.056     12.374      0.000       0.577       0.810
==============================================================================
Omnibus:                       23.100   Durbin-Watson:                   1.501
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               37.883
Skew:                           1.866   Prob(JB):                     5.94e-09
Kurtosis:                       8.060   Cond. No.                     1.39e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.39e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The OLS (Ordinary Least Squares) regression results provide valuable information about the relationship between the "Year" and "Matches" variables in our dataset. Let's break down the key elements of the regression summary and what they tell you about the data:

Dep. Variable (Dependent Variable - Matches): This indicates the variable you are trying to predict or explain. In this case, it's the "Matches" column.

Model: This section describes the regression model used. It's a simple linear regression model (OLS) that attempts to fit a straight line to the data.

Method: "Least Squares" indicates that the model was estimated using the least squares method, which minimizes the sum of squared residuals.

Date and Time: These indicate when the regression analysis was performed.

No. Observations: The number of data points used in the analysis. In this case, you have 23 observations.

Df Residuals (Degrees of Freedom - Residuals): The degrees of freedom for the residuals. It's the number of observations minus the number of estimated coefficients (21 in this case).

Df Model (Degrees of Freedom - Model): The degrees of freedom for the model, which is the number of estimated coefficients (1 for the constant term plus 1 for the "Year" variable).

Covariance Type: Indicates that the covariance matrix of the errors is assumed to be non-robust.

Now, let's focus on the key statistics related to the regression model:

R-squared (R^2): R-squared measures the goodness of fit of the regression model. In this case, R-squared is 0.879, which means that approximately 87.9% of the variation in the "Matches" variable can be explained by the linear relationship with the "Year" variable. A high R-squared suggests a strong linear relationship.

Adj. R-squared (Adjusted R^2): This is a version of R-squared adjusting for the number of predictions in the model. In this case it is 0.874, which is very close to the R-square value. It penalizes the inclusion of unnecessary variables in the model.

F-statistic: The F-statistic tests all significance of the regression model. A high F-statistic with a low p-value indicates that the overall model is statistically significant. In this case, the F-statistic is 153.1, and the corresponding p-value is very close to zero (4.13e-11), indicating that the model is statistically significant

Coefficients: The Coefficients section provides information about the intercept (const) and coefficient of the "Year" variable. In this case:

The constant (intercept) will be -1327.1395. The coefficient of "Year" is about 0.6934. These values represent the estimated values of the intercept and slope of the linear regression equation:

Mel = -1327.1395 + 0.6934 * years

p>|t| (p-value): The p-values associated with each coefficient test whether the coefficient is significantly different from zero. In this case, the p-values of both coefficients are very small (close to zero), indicating that they are statistically significant.

Confidence interval: The [0.025, 0.975] line gives the confidence interval for the parameters. They indicate the extent to which actual population projections fall within a certain confidence interval.

Finally, the "Notes" section provides additional information on standard errors, condition numbers, and other information relevant to regression analysis. Overall, the results show a strong linear relationship between the "year" and "match" variables, as given by a high R-squared value and a low p-value for the coefficients.

Prediction Analysis: Through our model, we get the following linear equation: Matches = -1327.1395 + 0.6934 * Year. If we want to predict the matches in the Year say 2026, the next edition (FIFA World Cup comes after every 4 years) then we plug in the value of 2026 in year in the equation, which gives us close to 80 matches. This is subjected to the number of teams participating. More the teams participating, more will be the matches played. Please note that this prediction assumes that the relationship between Year and Total attendance remains linear, and there are no significant changes or other factors influencing attendance that are not accounted for in this model.

Now I have also shown the relationship between the total attendance and years through regression analysis and below is the snapshot of my code:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Total attendance †   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.904
Method:                 Least Squares   F-statistic:                     197.7
Date:                Fri, 08 Sep 2023   Prob (F-statistic):           7.90e-12
Time:                        12:59:18   Log-Likelihood:                -310.44
No. Observations:                  22   AIC:                             624.9
Df Residuals:                      20   BIC:                             627.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -7.264e+07   5.31e+06    -13.682      0.000   -8.37e+07   -6.16e+07
Year        3.772e+04   2682.625     14.059      0.000    3.21e+04    4.33e+04
==============================================================================
Omnibus:                       10.669   Durbin-Watson:                   1.568
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.096
Skew:                           1.110   Prob(JB):                       0.0106
Kurtosis:                       5.234   Cond. No.                     1.45e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.45e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The provided regression results show a linear regression model for predicting Total attendance based on the Year given. Here's a breakdown of the key information from the regression results:

R-squared (R^2): R-squared measures the goodness of fit of the model. In this case, R-squared is 0.908, indicating that approximately 90.8% of the variation in Total attendance is explained by the Year variable in the model.

Coefficients:

Constant (intercept): -7.264e+07 Year: 3.772e+04 These coefficients represent the parameters of the linear regression equation: Total attendance = (-7.264e+07) + (3.772e+04) * Year P-values: P-value for the Year coefficient is very close to 0, indicating that Year is statistically significant in predicting Total attendance. To predict the Total attendance for the year 2026, you can simply plug in the value of 2026 into the regression equation:

Total attendance = (-7.264e+07) + (3.772e+04) * 2026

Now, let's calculate the prediction:

Total attendance = (-7.264e+07) + (3.772e+04) * 2026 Total attendance = -7,264,000,000 + 76,537,200

Total attendance = 69,273,200

So, based on this regression model, the predicted Total attendance for the year 2026 is approximately 69,273,200. (This can be true since the next World Cup will be played in Canada, USA and Mexico. USA has huge stadiums with large capacities like the Rose Bowl, Stanford Stadium, etc). Please note that this prediction assumes that the relationship between Year and Total attendance remains linear, and there are no significant changes or other factors influencing attendance that are not accounted for in this model.

Now, we have analysed the top 10 players' performances throughout the history of the FIFA World Cup, based on goals per game parameter.

	Rank	Player	Goals	Matches
0	1	Miroslav Klose	16	24
1	2	Ronaldo	15	19
2	3	Gerd Müller	14	13
3	4	Just Fontaine	13	6
4	4	Lionel Messi	13	26
5	6	Kylian Mbappé	12	14
6	6	Pelé	12	14
7	8	Sándor Kocsis	11	5
8	8	Jürgen Klinsmann	11	17
9	10	Helmut Rahn	10	10
10	10	Gabriel Batistuta	10	12

	Rank	Player	Goals	Matches	Goals per game
0	1	Miroslav Klose	16	24	0.666667
1	2	Ronaldo	15	19	0.789474
2	3	Gerd Müller	14	13	1.076923
3	4	Just Fontaine	13	6	2.166667
4	4	Lionel Messi	13	26	0.500000
5	6	Kylian Mbappé	12	14	0.857143
6	6	Pelé	12	14	0.857143
7	8	Sándor Kocsis	11	5	2.200000
8	8	Jürgen Klinsmann	11	17	0.647059
9	10	Helmut Rahn	10	10	1.000000
10	10	Gabriel Batistuta	10	12	0.833333

Here is my analysis based on the findings of my table:

#7. We see that Miroslav Klose scored the most goals during all FIFA world Cup editions in which he played from the data collected from 1930-2022. His goals per game is one of the lowest.
#7.1 This tells us that he was not so effective in every match he played. He was inconsistent, the reasons maybe tough opponents, physical fitness, less recovery time and so on.

#8 He was a contributor in victories in some matches but not all the time.
#8.1 This tells us that his team had greater cohesion and was not dependent on a single player to win matches. He was a part of the winning campaign for Germany in 2014 where he scored only 2 goals.
#9. Though, Sandor Kocsis and Just Fontaine played the least amount of matches, they scored the most goals in the matches they played.
#9.1 This tells us they were highly efficient and were a key contributor to their teams successes in the particular editions in which they played.
#9.2 Their teams depended a lot on them for victories and these players were cool and collected under pressure and delivered whenever the team required them to do so.
#9.3 Sandor Kocsis was a part of the Hungarian National Football team in 1954 and took his team to the finals single handedly where they lost to Germany.
#10. Lionel Messi has the lowest goals per game ratio i.e. 0.50, and that is why Argentina were not able to win a cup since 1986 (see table below) as they were solely dependent on him to score.
#10.1 In 2022, the team was united and reduced pressure from Messi's shoulders and won the World Cup.

Now, we look at the history of winners and runners up of the FIFA World Cup and analyse our findings, here is the snapshot:

	Team	Titles	\
0	Brazil	5 (1958, 1962, 1970, 1994, 2002)	
1	Germany1	4 (1954, 1974*, 1990, 2014)	
2	Italy	4 (1934*, 1938, 1982, 2006)	
3	Argentina	3 (1978*, 1986, 2022)	
4	France	2 (1998*, 2018)	
5	Uruguay	2 (1930*, 1950)	
6	England	1 (1966*)	
7	Spain	1 (2010)	
8	Netherlands	NaN	
9	Hungary	NaN	
10	Czechoslovakia	NaN	
11	Sweden	NaN	

	Runners-up	Third place	\
0	2 (1950*, 1998)	2 (1938, 1978)	
1	4 (1966, 1982, 1986, 2002)	4 (1934, 1970, 2006*, 2010)	
2	2 (1970, 1994)	1 (1990*)	
3	3 (1930, 1990, 2014)	NaN	
4	2 (2006, 2022)	2 (1958, 1986)	
5	NaN	NaN	
6	NaN	NaN	
7	NaN	NaN	
8	3 (1974, 1978, 2010)	1 (2014)	
9	2 (1938, 1954)	NaN	
10	2 (1934, 1962)	NaN	
11	1 (1958*)	2 (1950, 1994)	

	Fourth place	Top 4 total
0	2 (1974, 2014*)	11
1	1 (1958)	13
2	1 (1978)	8
3	NaN	6
4	1 (1982)	7
5	3 (1954, 1970, 2010)	5
6	2 (1990, 2018)	3
7	1 (1950)	2
8	1 (1998)	5
9	NaN	2
10	NaN	2
11	1 (1938)	4

#11. We see that Brazil has been the most successful team with 5 trophies, they have also been runners up in 1950 and 1998.
#11.1. We can see that Brazil is losing its foothold on the cup as the new generation players haven't been able to win the cup after 2002, due to emergence of new teams like Belgium, Netherlands, Japan, Morocco, etc.

#11.2. We can see a resurgence in the France National Team which had won the 2018 edition and were the runners up in 2022 edition. They have been performing well lately and are a force to reckon with.
#12. England having the best league in the world, the English premier league (EPL), with best players having great experience and exposure have not won a world cup since 1966. They have choked in pressure situations giving advantage to teams below their ranking.
#12.1. Italy and Brazil are the only teams to have defended their title, i.e.win back-to-back editions. Italy from 1934 and 1938 and Brazil 1958 and 1962.
#13. With teams like Belgium, Netherlands, Japan, Saudi Arabia, etc, it will be difficult for the defending champions to defend their title. These teams are coming up in a good way and can shock the leaders on their day.

Managerial Insights:

From our Regression Analysis, we found out that the total attendance will be around 69000 in 2026 edition of the FIFA World Cup. This is the perfect opportunity for the organisers and the event managers to raise the average prices of the tickets for each game. By raising the prices, for each game, the organisers and the managers who look after the matches, will generate more revenue and also can get more money to themselves through additional revenue.

Also, through the additional revenue, FIFA can get new sponsors and provide luxury seats to some lucky fans and conduct fan engagement activities for the growth of the sport and create excitement and curiousness for the event.