ANALYSES OF INTERNATIONAL FOOTBALL RESULTS FROM 1872 - 2021

Project Report

Clodagh Doonan
clodaghdo@gmail.com

GitHub URL: [Bantyno2/UCDPA_CLODAGHDOONAN: Final report for DA Cert (github.com)]

Abstract

This project sets out to analyse the International Football Results for the period of 1872 to 2021. The dataset captures the tournament types across this period along with the results of games played in each tournament. From analysis and data manipulation of the dataset we can gain insights into the different types of tournaments played at an international level, the number of games played, goals scored and the team which have been the most successful in this period.

Introduction

One of the most popular sports represented at an international level is football. This dataset is of particular interest to me as I have a passion for football, it allows the user to get insights into the history of the sport. Currently the UFEA Euro 2020 competition is underway so this dataset is topical at the moment.

Dataset

The chosen dataset is titled “International Football results from 1872 – 2021”. Its source is “Kaggle”

[International football results from 1872 to 2021 | Kaggle]

I choose to use Kaggle as it is very user-friendly for beginners to explore and extract data to allow me to enhance my knowledge of datasets.

The original dataset consisted of 9 columns containing Objects, Integers and Booleans. There are a total of 42,183 rows, this is a large number of rows however given the period of the dataset this is to be expected.

	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
1	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
3	1875-03-06	England	Scotland	2	2	Friendly	London	England	False
4	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	False

Existing list of columns and what they represent:

Column Name	Description
Date	Date of Match
Home_Team	Name of Home Team
Away_Team	Name of Away Team
Home_score	Number of goals the home team scored
Away_score	Number of goals the away team scored
Tournament	The type of international tournament
City	The city where the game took place
Country	The name of the country where the match took place
Neutral	Weather the match took place at a neutral venue

In order to enhance my dataset for better manipulation for visualisation, I will be adding columns to the dataframe.

The additional columns include:

Column Name	Description
Results	The outcome of the game
Home wins	Games won at home
Away wins	Games won away from home
Drawn games	Games where result was a draw

When I viewed the dataset at a high level initially, and given my interest in the topic, I had a good sense of what I wanted to get from this data. This dataset would allow me to query and manipulate the data to produce visualisations from the international tournaments held in the 1872 – 2021 timeframe.

Implementation Process

I used Jupyter Notebook to load the dataset and explore the different insights prior to migrating this across to PyCharm. I found Jupyter Notebook very user friendly and useful.

When migrating across to PyCharm, first I created my PyCharm project folder, in order to explore, analyse and visualise the dataset I needed to import certain libraires into my script. For this I imported the following:

- Panda,
- Matplotlib,
- Seaborn,
- numpy and
- datetime.

The dataset was available in CSV format which I saved into my C: drive I then had to import this file into PyCharm in a readable format. For this I used "pd.read_csv".

At first, I indexed the columns with "index_col =0" so the date would be the index however I removed this as having the 'date' as the index was not beneficial. Instead, I looked at slicing the date and turning it into separate columns "month" and "year". With no need for the date column, I removed this to tidy up the table.

I explored the dataset by using a range of functions in order to see what the dataset I was going to be working with looked like. High level exploration showed:

- Head and tail function – showing 1st and last 5 rows of data
- Shape function – shows that there are 9 columns and 42,183 rows of data to work with
- Type – shows that I have columns containing 'objects', 'integers' and a 'boolean'
- Len – similar to outputs seen in shape the length of the dataset is x
- Describe – showing a synopsis of the statistics (i.e mean, std, min etc.)
- I also checked the index, columns and values

Searching for 'null' values

Given the large number of rows contained in the dataset I wanted to ascertain if there are any null values I may be dealing with. For this I used the 'isnull' function and we can see that there are no 'null' or 'NaN' values contained in this dataset. If there was, I would be using the 'fillna' function to populate these missing values.

```
date          0
home_team     0
away_team     0
home_score    0
away_score    0
tournament    0
city          0
country       0
neutral       0
dtype: int64
```

Manipulating of the dataset

Analysis of Home v Away: The 1st obvious insight I want to produce is comparing the total home, away, draw columns to see if more games are won at home or away. I have created additional columns to enhance the dataset for this. This includes the following columns added:

- Home wins
- Away wins
- Drawn games

By creating these columns the outputs are returned as a set of Booleans (True/False) values. To interpret this and produce a visual insight I have counted the 'home_wins', 'away_wins' and 'draw' columns and used the print function to show the total in a string (str)

We can see from analysing this that more games are won by home teams than away teams. There is also a significant number of games that ended in a draw.

```
         date home_team away_team  ...  home_wins  away_wins   draw
0  1872-11-30  Scotland   England  ...      False      False   True
1  1873-03-08   England  Scotland  ...       True      False  False
2  1874-03-07  Scotland   England  ...       True      False  False
3  1875-03-06   England  Scotland  ...      False      False   True
4  1876-03-04  Scotland   England  ...       True      False  False
```

Home advantage

To explore if home advantage is a real concept in football, I created a python dictionary (dictionary over list being easy and efficient) with *key:value* pairs, the 'keys' = status and result and 'values' = "home" "away" "draw" and a list of integers corresponding to the result key.

```
data = {"Status":["home_wins", "away_wins", "draw"], "result":[20511,11944,9728]}
```

I added these to a dataframe and plotted the result on a simple bar plot for visual effect. (*See figure 1 in the Result section below)*

Another column I added to the dataset was 'total goals' which shows the total goals scored in each game.

I further enhanced the dataset by using the 1st column which contained date in format yyyy/mm/dd and splitting this out into 2 separate columns “Month” and “year”. Once this was done a short ‘loop’ was created o delete the “date” column as it was no longer required.

Once this was complete a quick look at the dataset following the above showed that it has increased from 9 columns to 14 columns.

```
   home_team away_team  home_score  away_score  ...   draw total_goals Month  year
0  Scotland   England            0           0  ...   True           0    11  1872
1   England  Scotland            4           2  ...  False           6     3  1873
2  Scotland   England            2           1  ...  False           3     3  1874
3   England  Scotland            2           2  ...   True           4     3  1875
4  Scotland   England            3           0  ...  False           3     3  1876

[5 rows x 14 columns]
```

Games/matches per year

To get the games per year I used the ‘value_count()’ function and reduced this to top 50. From this we can see in what year the most games were played. Further exploration of the data below will show a correlation between this and a view of the participating countries (*See figure 3 in the Result section below*)

Indexing/Slicing

By slicing the dataset to extract “country” , “home team” and “away team” and dropping the duplicates we can see that there is:

- 266 countries where tournaments were held
- 308 home teams and,
- 305 away teams

Plotting games/matches per year

See F*igure 2 in the Result section below* – by plotting the matches per year to see the trend or increase over the years. The top 5 years for games were 2019, 2008, 2011, 2004 and 2000.

Using Loc function

I used the ‘loc’ function to easily retrieve data from the dataset. Using this function, I wanted to see how many games England played at home and away. We can see that they played 507 home games and 515 away games.

Goals scored per month

As I had previously added a new column called ‘month’ I wanted to see if there was a correlation between the matches played and goals scored in certain months.

I used the ‘groupby’ function to extract the total goals scored column by month and got a ‘sum’ of these and for the goal per month extracted ‘count’ of the months from the dataset. By plotting these two visuals *(*See F*igure 4a and 4b in the Result section below)* we can see very clearly there is a correlation between the two.

Groupby – Hometeam wins

Further exploring the groupby function to show the home team wins and saving as a variable “top_team_scores”.

Next, I have looked at the participating countries over the period Figure 3 to show the correlation between participating countries over the years and the matches played over the years 1872 – 2021.

Tournaments

To see the tournaments within the dataset I have extracted the 'tournaments' columns and used value_count, from this I want to plot using seaborn with a barplot. I have restricted to the top 20 tournaments. (See F*igure 5 in the Result section below*)

```
Friendly                                17273
FIFA World Cup qualification             7378
UEFA Euro qualification                  2582
African Cup of Nations qualification     1719
FIFA World Cup                            900
```

Other analysis

To wrap up the exploration and analysis of this dataset I have used the following functions:

- For loop incorporating the range, If, elif with an append functions to extract and show the countries with the most wins
- Value_count on country to see who hosted the most games
- Sorting functions – sort by multiple variables in this case using country and city to sort the dataset to extract the capital from the country
- Subsetting :
    - I used subsetting to get the international games form 1960 onwards
    - To review the years when most games were played, I had previously plotted this in fig2. Top 5 years include 2019 = 1,155 games, 2008 = 1,092 games, 2011 = 1,083 games, 2004 = 1,066 games and 2000 = 1,026 games
- Adding new column to show the cumulative sum of the total goals scored

At the start the dataset contained 9 columns and at the end of the analysis and manipulation we end up with a dataset containing 15 columns.

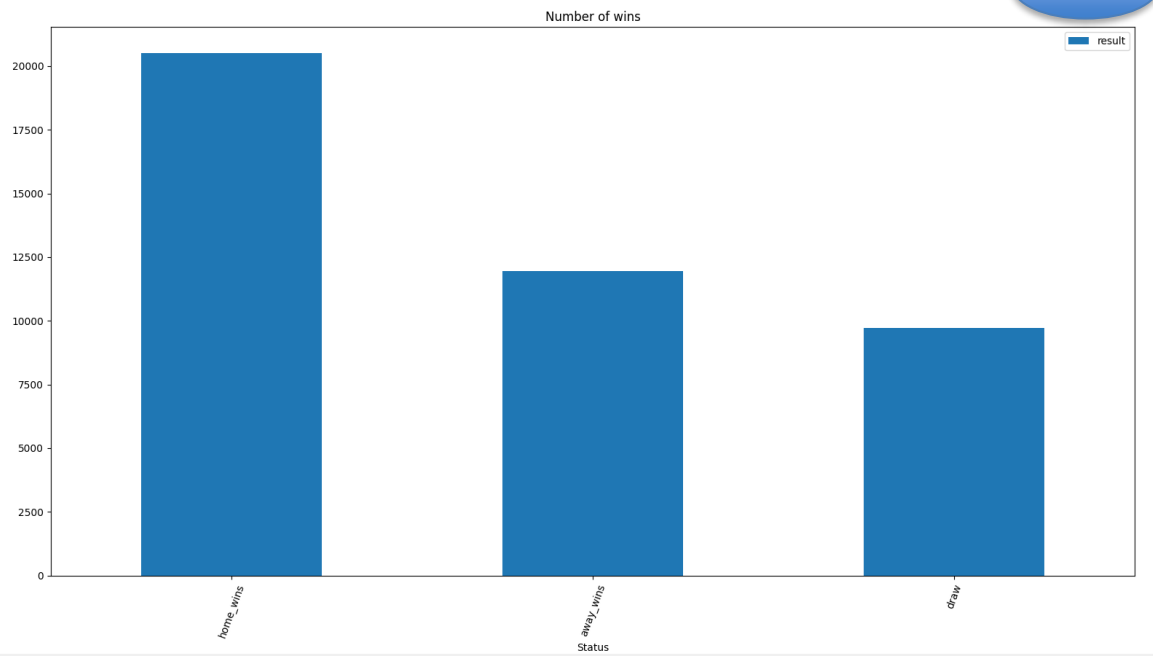
```
  home_team away_team  home_score  ...  Month  year goal_cum_sum
0  Scotland   England           0  ...     11  1872            0
1   England  Scotland           4  ...      3  1873            6
2  Scotland   England           2  ...      3  1874            9
3   England  Scotland           2  ...      3  1875           13
4  Scotland   England           3  ...      3  1876           16

[5 rows x 15 columns]
```

Results

1) Total Home/Away/Draw games

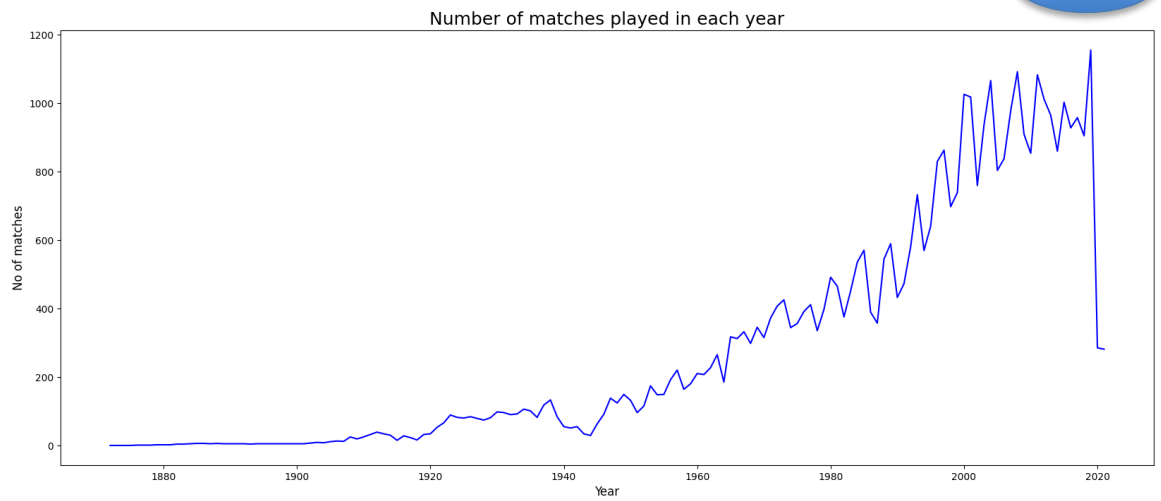
Fig.1



The above chart shows analysis of games won at home and away, in addition drawn games are shown. We can see that teams playing at home have an advantage with more home games being won (20,511). Teams playing away are less likely to win, away wins and drawn games are closely aligned with 11,944 away wins and 9,728 drawn games.

2) Trend of matches played across the years

Fig.2

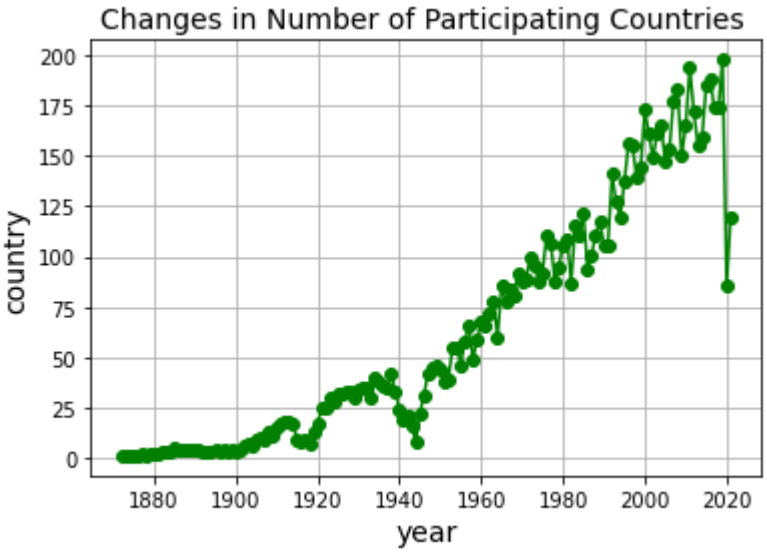


Since the start of the collation of this data in 1872 there has been a huge increase in the number of matches played. We will see in the next graph that this correlates with the increase in the number of participating countries. From 1920 onwards the trajectory of the graph heads in an upward direction peaking in 2019 when the most matches were played. It then heads in a downward trajectory however data only goes up to the start of 2021 and would not be a true reflection of the full year.

Given the Covid-19 pandemic and delay of the Euros 2020 this would also contribute to the downward trajectory.

3) Changes in number of participating countries

Fig.3



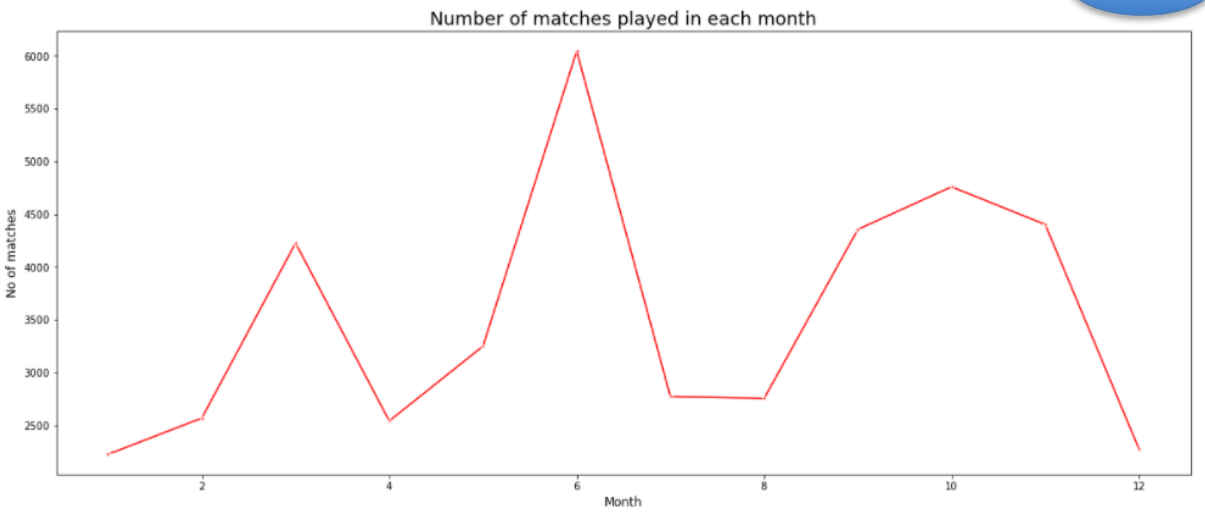
Similar to the graph above, we can see that the number of participating countries has increased over the years and can be linked to the increase in matches played.

4) Analysis of matches and goals per month

Here I wanted to see a breakdown of the months and if there is a correlation between the matches played with the number of goals scored. Once again, we can see there a commonality here in the top 2 months June and October however the 3rd month there is a slight difference. September has the 3rd month with most matches however is 4th month for goals scored.

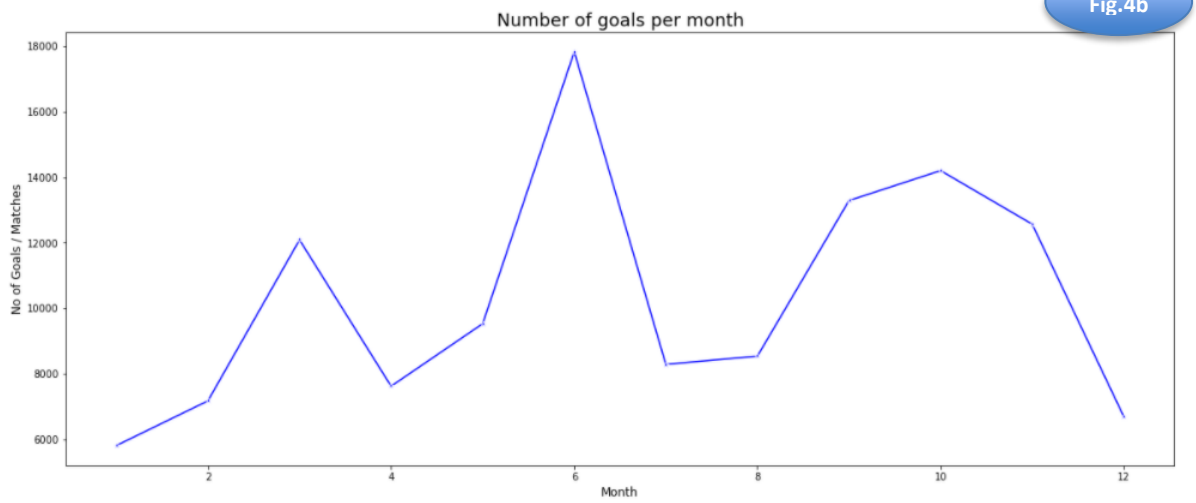
(a) Number of matches played each month

Fig.4a



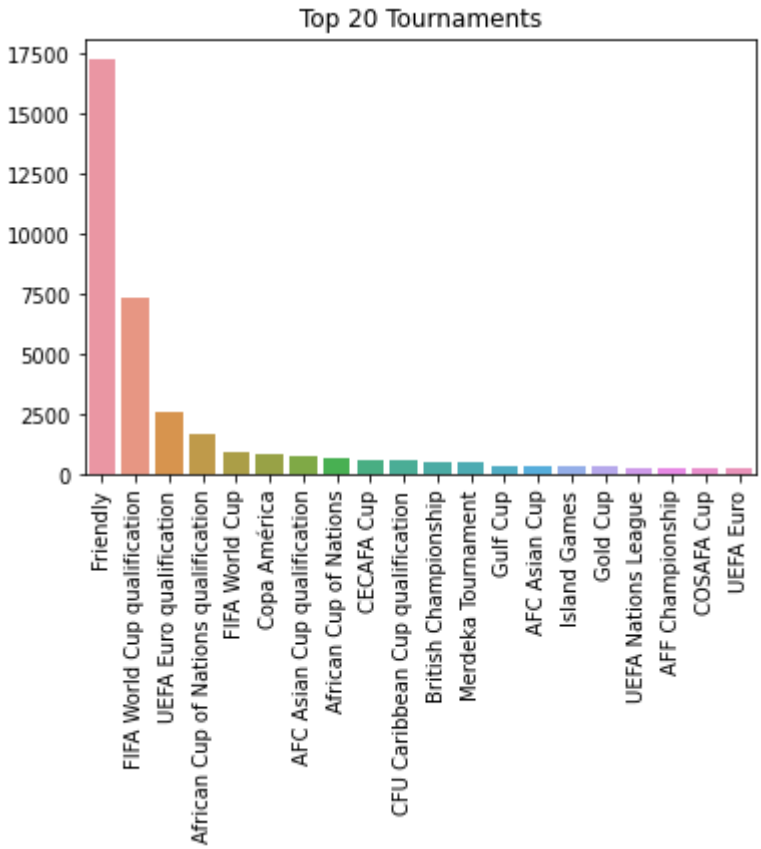
(b) Number of goals per month

Fig.4b



5) Top 20 Tournaments

Fig.5



Top 20 tournaments played across the period. The friendly tournaments are the most popular, although they may not be seen as a 'tournament' these games are vital towards teams preparing for bigger tournaments such as "Euros" and "world cup". FIFA world cup is ranked 5th on top tournaments based on games played with 1-4 bring friendlies and qualification tournaments.

Insights

By exploring, manipulating and analysing the dataset “international football results from 1872 – 2021” we can see the following insights:

1. Teams playing at home appear to be more advantageous to those away. There were 20,511 home games won with 11,944 away games won.
2. England have played 507 home games and 515 away games a total of 1022 games of which they have won 582 (57%)
3. Brazil is the country with the most wins at 631 followed by England(582), Germany(561), Argentina(529) and Sweden (508). Top 5 countries contain 2 countries from South America and 3 from Europe.
4. There is a direct correlation between matches played each year and the increase in the participants in tournaments.
5. 2019 is the year with most games played (1,155), the data for 2020 and 2021 shows a downward trajectory however given impacts of Covid-19 pandemic on the sporting world this is not surprising. An example of this would be the UEFA Euro 2020 which was postponed to June 2021.
6. The friendly tournaments are the most played games, followed by 3 qualification tournaments (FIFA, EUROs and African Nations). The 5th most played being the FIFA World Cup.
7. Tournament games are held throughout the year with June being the most popular month to play matches.
8. There is also a direct correlation between the goals scored per month and the games played per month. Again, June being the month where most goals were scored.
9. There are 266 countries which host the tournament
10. United States has hosted the most games (1,169)
11. A total of 123,583 goals have been scored across all tournaments in the period 1872 – 2021

References

[International football results from 1872 to 2021 | Kaggle]