CS732/DS732: Data Visualization Assignment 4 Visual analytics Report

Daggubati Siri Chandana[1] and Dhvani katkoria[2]

[1]MS2019005
[2]MS2019007

December 4, 2019

1 Problem Statement

Perform data analysis on International football results data set collected from 30th Nov 1872 to 19th Nov 2019. We can use any analyzing or visualization techniques as per the requirement.

1.1 Data Explanation

This dataset includes 41,540 results of international football matches starting from the very first official match in 1972 up to 2019. The matches range from FIFA World Cup to FIFI Wild Cup to regular friendly matches. The matches are strictly men's full internationals, and the data does not include Olympic Games or matches where at least one of the teams was the nation's B-team, U-23, or a select league team. There are the following columns in the data table.

- date - date of the match
- home_team - the name of the home team
- away_team - the name of the away team
- home_score - full-time home team score including extra time, not including penalty-shootouts
- away_score - full-time away team score including extra time, not including penalty-shootouts
- tournament - the name of the tournament
- city - the name of the city/town/administrative unit where the match was played

- country - the name of the country where the match was played
- neutral - TRUE/FALSE column indicating whether the match was played at a neutral venue

The data is gathered from several sources, including but not limited to, Wikipedia, fifa.com, rsssf.com, and individual football associations' websites.

2 Analytical Tasks

2.1 Number of international soccer games

The motivation behind this task is to check how many games were played each year and how the total number of international games evolve with time. To perform this visualization we first separate date into day, month and year columns for the given data and then plot the Line graph to visualize the evolution of number of games. The graph is as shown below. By performing this plot we infer that number of games is rising, with high growth in the 80s/90s. It seems there is a peak around 2010, with a slight decrease since.

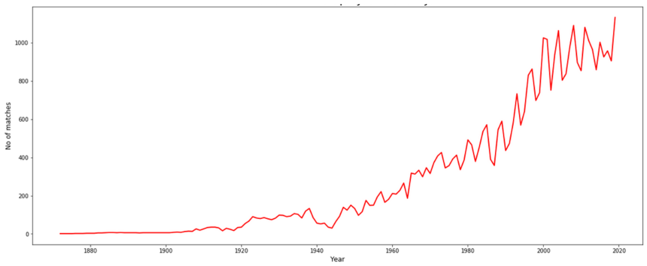


Figure 1: Number of international soccer games

2.2 Top countries with most wins

The motivation behind this task is to find out the countries with most wins from the games played with total number of wins to find out leading countries in the football game. To perform this visualization we add a new column named reults. The value in the column are names of the country that won that particular match. Winning/Losing is calculated based on the home-score and away-score. If home-score is greater than away-score result is updated with home-team name else with away-team name and in case of equal score it's updated with 'Tie'. Bar chart is plotted to visualize the top 20 countries with most wins that appear most times in the result column. By performing this plot we infer that Brazil, England, Germany are the top 3 leading teams in the game.

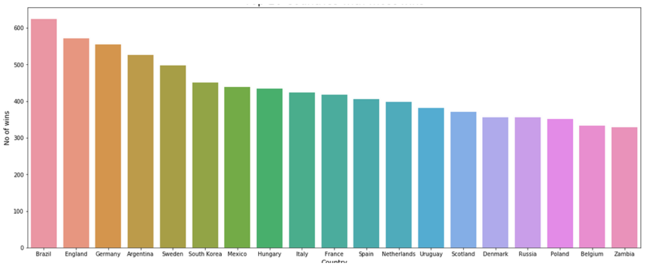


Figure 2: Top countries with most wins

2.3 Favourite ground for top country

The motivation behind this task is to analyze which city has been the most favourable for maximum number of wins for the leading country Brazil in the game. To perform this visualization we count the number of times a city comes in the game when Brazil has won the match based on the result column we previously created. Horizontal Bar chart is plotted to visualize the maximum wins of Brazil across different cities. By performing this plot we infer that Brazil has won most matches on the grounds of Rio de janerio, Sau Paulo and San tiago which indicates that its most wins have been on the home ground rather than the away ground.

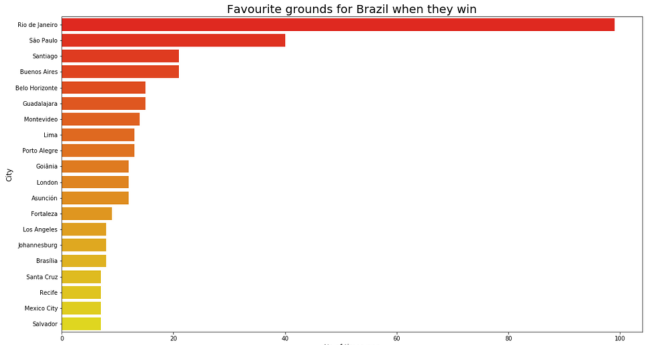


Figure 3: Favourite ground for top country

2.4 Most played leagues

The motivation behind this task is to analyze which are different tournaments that has been played most over time. To perform this visualization we count the number of times a tournament appears in the data. Pie chart with percentage of tournament type played is plotted to visualize the top 20 most played leagues. By performing this plot we infer that Friendly matches are the most common games that are being played followed by the FIFA world cup qualification and UEFA euro qualification tournaments.

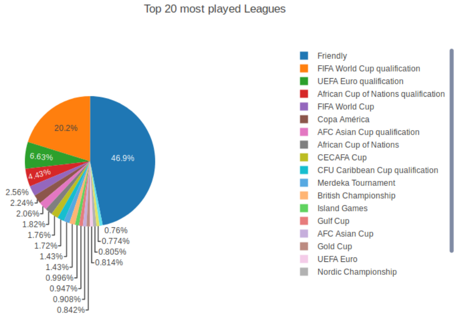


Figure 4: Most played leagues

2.5 World wide soccer adoption

The motivation behind this task is to analyze when did soccer start to be widely played, i.e. when do most nations start playing international games. To perform this visualization we groupby the data set acoording to the year column and count the number of unique teams that appear in the data set on the base of a particular year. Line chart is plotted to visualize the number of teams that has adopted football game over time. By performing this plot we infer that the number of rise in the teams is similar to the rise in number of games in the data as plotted earlier. The number of teams steadily increased 1902 and this increase accelerated up to 1920. From there, the pace of addition of new teams increase much faster and stalls a bit around the late 40's. The stalls around 1920's and 1940's are the periods when the world wars had taken place. Then we see a steady and rapid growth up to the mid 1990's.

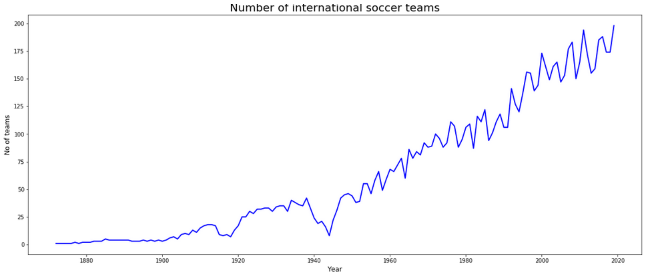


Figure 5: World wide soccer adoption

2.6 Evolution of goals scored

The motivation behind this task is to analyze the heart of soccer: goals. How did this evolve with time? To perform this visualization we add a new column 'Total-score' which has the values of total number of goals in the match by summing up the home-score and away-score for that same match. Then based on the month and year column we represent the total goals that appear in that month and year across time. To visualize this heat map is plotted to view the evolution of number of goals over month and years. By performing this plot we infer that although it started low then number of goals per games quickly skyrocketed. The first games mostly occur during Spring months and since then, some month have known some peaks of popularity for intenational games at different period (e.g. many games happened in December in the 1940s). Before 1900, the average number of goals per game per year were around 8. This average then stabilized around 4 until 1950 and then decreased down to 2.5 in a more modern era.

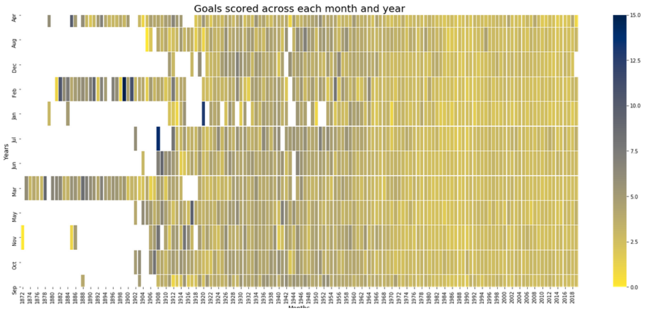


Figure 6: Evolution of goals scored

2.7 Total Home goals scored by countries

The motivation behind this task is to analyze the countries with highest total number of goals and home goals scored over time. To perform this visualization we groupby the data according to the country and arrange according to the highest number to total-score column calculated earlier. To visualize this stacked bar chart is plotted to view the top 30 countries with most total and home goals scored by each country. By performing this plot we infer that United states has the maximum total-score with most home ground wins followed by the France. Also, the country Finland has the more away wins than the home wins.

2.8 Number of matches played by countries

The motivation behind this task is to analyze the countries that has played most number of matches along with home matches and away matches.

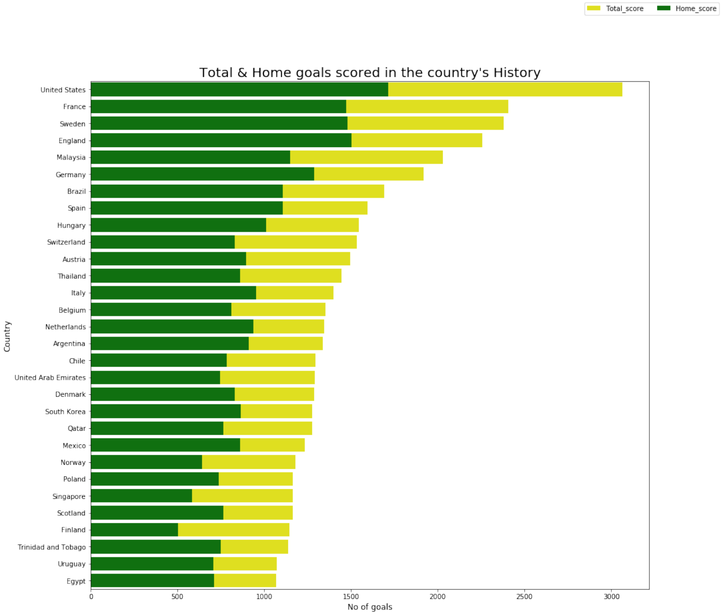


Figure 7: Total Home goals scored by countries

To perform this visualization we add 3 new columns 'Home-match', 'Away-match' and 'Total-matches'. The 'Home-match' has the count of home-teams while 'Away-match' has the count of the away-team after grouping by the country. Column 'Total-matches' has summed values of home-match and away-match. Based on this column the data is sorted. To visualize this scatter lines binded with cufflinks modules in python is used. By performing this plot we infer that Europe was the first continent to provide many soccer teams, followed by the American continent (North and South) which shows European teams have most number of matches played followed by American teams.

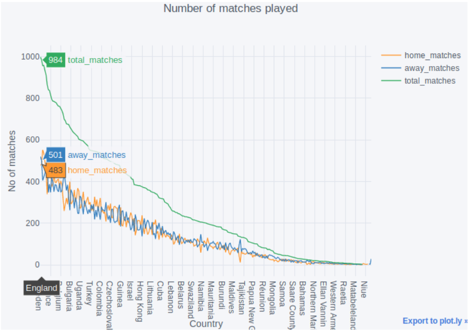


Figure 8: Number of matches played by countries

2.9 Win prediction based on history of victories

The motivation behind this task is to predict the results of the game played among different countries based on the historical victory ratios of each country team. To perform this visualization we calculate the home-wins by taking the ratio of the home-matches won to the total number of home matches played. Similarly, away-wins are calculated by taking the ratio of the away-matches won to the total number of away matches played. This is calculated based on each team in the data set. Further we add a new column total-victory ratio which is calculated by taking the ratio of the total count of wins in home and away matches to the total home and away matches played based on the above calculated home-wins and away-wins ratio. Then the top 16 teams are selected from the resulted data. We then randomly group these 16 team for a match between two teams based on the group of 2 created. Further, based on the total-victory ratio the winning team is predicted from each group. In the next set we have 8 teams for tournament. This process is iterated till final winning team is predicted. To visualize this layer wise node link diagram is used. By performing this plot we can predict the winning team based on their victories in the past. The plot below shows one such arrangement of win prediction based on one of the grouping possibilities for top 16 teams.

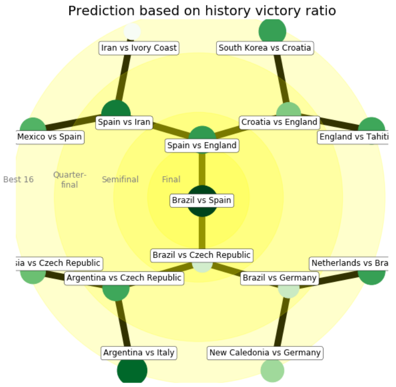
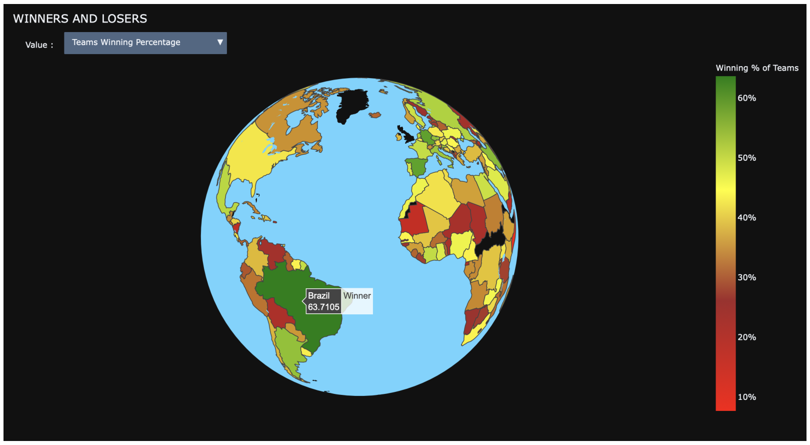


Figure 9: Win prediction based on history of victories

2.10 Predicting winning rate/losing rate

The motivation behind this task is to guess how probably a team wins/loses if it is the home team or away team or irrespective of teams. To perform this visualization Firstly, we exploit python and its main libraries to extract from the dataset. Calculate the result of teams based on home score and away score if it (wins, draws, losses). Now aggregate based on home teams/away team, then get win count and loss count and divide it by total count to get the percentage.

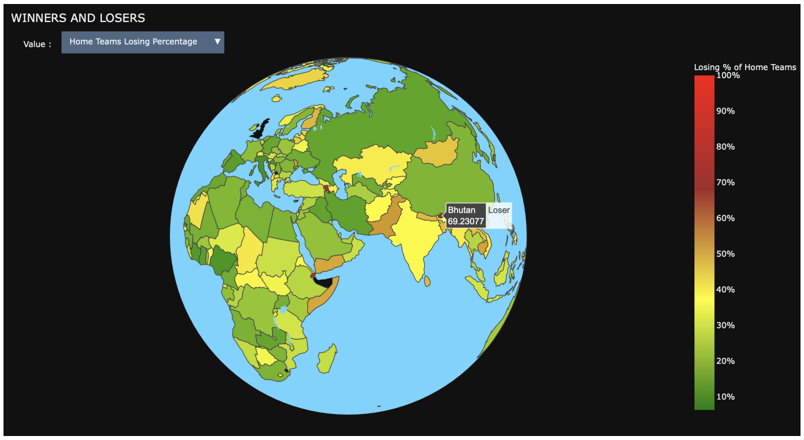
We plot the obtained percentage values of the home team, away team, or both using choropleth map in plotly with a color scale ranging green, yellow, brown, and red. The selection of winning / losing rates is given to the user based on the dropdown list. We also added the hovering effect of the mouse, which displays the country name and its probability rate of that particular polygonal region on the world map.



Value : Teams Winning Percentage

Figure 10: Choropleth map for winning rate of countries

The knowledge obtained by observing the below visualization is England, which is the country that has created the football game, has played more games than Brazil. But, the Brazilian victory rate is higher than the English one with 625 victories over 997 games. So we can declare Brazil the champion above any others. On the other hand, Liechtenstein, with 154 losses over 194 games played, is the worst National football team in history in terms of results.



Value : Home Teams Losing Percentage

Figure 11: Choropleth map for losing rate of home countries

2.11 Favorability of teams

The motivation behind this task is when a team wants to know about its rival team based on previous histories of matches of the rival team with other teams, we will plot the Favorability of the team's winning/losing the match with the other teams. To perform this visualization Firstly, we exploit python and its main libraries to extract from the dataset with 39672 matches played by 294 different teams. There are 6500 unique pairs in the dataset. Given that there are choose(294, 2) = 43071 possible pairs, this means that only 15.09% of possible matches took place. Here the order of pair is not concerned, and the aggregation performed is irrespective of order.

Calculate the result of teams based on home score and away score if it (wins, draws, losses). Then they are aggregated to obtain Favorability for all team pairs in the dataset. Favorability metric ranges from -1 (all games lost against given rival) to 1 (all games won), and in case of the draw of the match, it equals 0. The metric calculation is shown below

$$metric = \frac{won - lost}{total\_played} \tag{1}$$

We visualize the Favorability of country using choropleth map in a range of values 1, 0, and -1 on a color scale of green, white, and red, respectively. If there is no match with that country team, it is colored black.

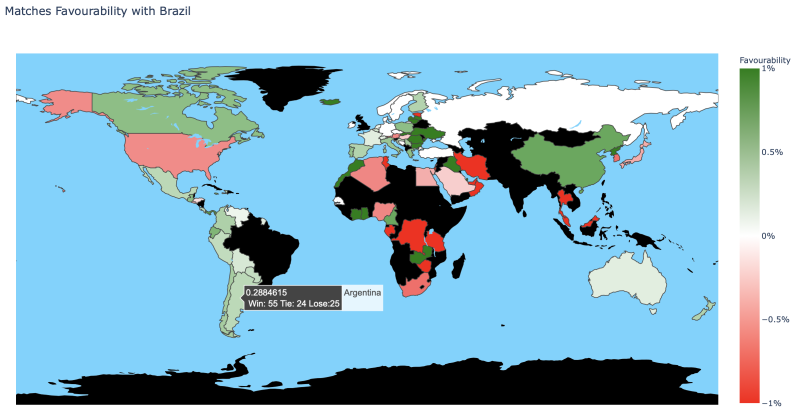


Figure 12: Choropleth map showing the Favorability of Brazil