Introduction

Data set

The 'International football results 1872 to 2019' dataset obtained from Kaggle details the matches played between countries, internationally, both - on a professional level (like FIFA World Cups) and friendly matches in the period 1872-2019 (Jürisoo, 2019). It consists of the number of goals by each team (the home teams and the away teams) in each match, on neutral/non-neutral terms (whether played on home soil or not). The dataset is large - 40,839 observations with 309 unique values, if filtered out using the distinct values for host countries, and has 93 distinct tournaments. It is clean (no missing values) and hence requires relatively less pre-processing. The features are:

Column	Description
Date	Date of the match
home_team	The name of the away team
away_team	The name of the away team
home_score	Full-time home team score including extra time
away_score	Full-time away team score including extra time
Tournament	The name of the tournament
City	The name of the city/town 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.

(Jürisoo, 2019)

Tasks

Following are the tasks that we finalized as part of Assignment 2 tasks:

TASK 1: Identify geographically and by win-rates, the countries that dominate FIFA.

TASK 2: Analyze the attack and defense strategy of these top FIFA ranking teams with respect to their winning margins.

TASK 3: Analyze whether playing on home soil affects the win-rates of the home teams and the away teams.

TASK 4: Explore whether the strategies adopted have an effect on the outcome of the match ending in a draw.

Aims and Contribution

The primary objective of the assignment is to visually analyze any relevant patterns that contribute to the success of teams that consistently perform well in FIFA organized tournaments by making use of apt visual tools, including Tableau, D3.js and python. We also plan to explain whether playing on home soil is advantageous for a team, and if so, why. This could help those teams that do not perform well rethink their strategy to perform better.

The strategies of the winning teams (top 50) were inferred from the visuals. However, more analysis and features are required to assess what really contributes to the higher win rates when the teams play on home soil.

Design

Analysis

This dataset is highly imbalanced in terms of the representation of the different types of tournaments, with ~50 percent of them being local/friendly tournaments. These tournaments involve the participation of a substantial number of local teams, a majority of which don't even qualify the FIFA qualifiers. As a result, these teams are not represented in the professional matches, significantly reducing the scope for further analysis.

Since the dataset is too large and varied, pruning was required in order to boost the aesthetic appeal of the visuals, without compromising on the integrity of the tasks defined. We subsetted the dataset to look into the details of the top fifty countries/teams (based on FIFA ranking). Further preprocessing was required to filter out all the friendly tournaments and FIFA qualifiers that involved the participation of these fifty teams.

The storyline revolves around the visual assessment of the performances of these teams based on the strategies (the extent to which they attack or defend) they adopt, whether the venue of these tournaments have a say in adopting a strategy that they are otherwise keen on sticking to. To explore and implement these tasks, feature engineering was required to calculate the win rates, strike/attack rates and the defense rates from the existing features, for each of the teams. These are given below:

Win rate = (Number of matches won by a team / total number of matches played by that team)

Attack rate = (Number of goals scored by a team / total number of matches played by that team)

Defense rate = - (Number of goals scored by the opponent / total number of matches by the other team)

Visualization

The graphs and charts in this section visualize the preliminary analysis/claims made in the previous section.

The graph below is drawn using Tableau (Recnac, 2019). It shows the win-rates of the current top fifty teams with the pie charts showing the win rates for the respective teams, with orange representing the rate when the country played on home soil and blue representing the win rate of the country on foreign soil. It can be inferred that orange dominates blue in every pie chart, confirming that home soil is advantageous. Further, most of the pie charts are clustered in the continent of Europe, indicating their dominance.

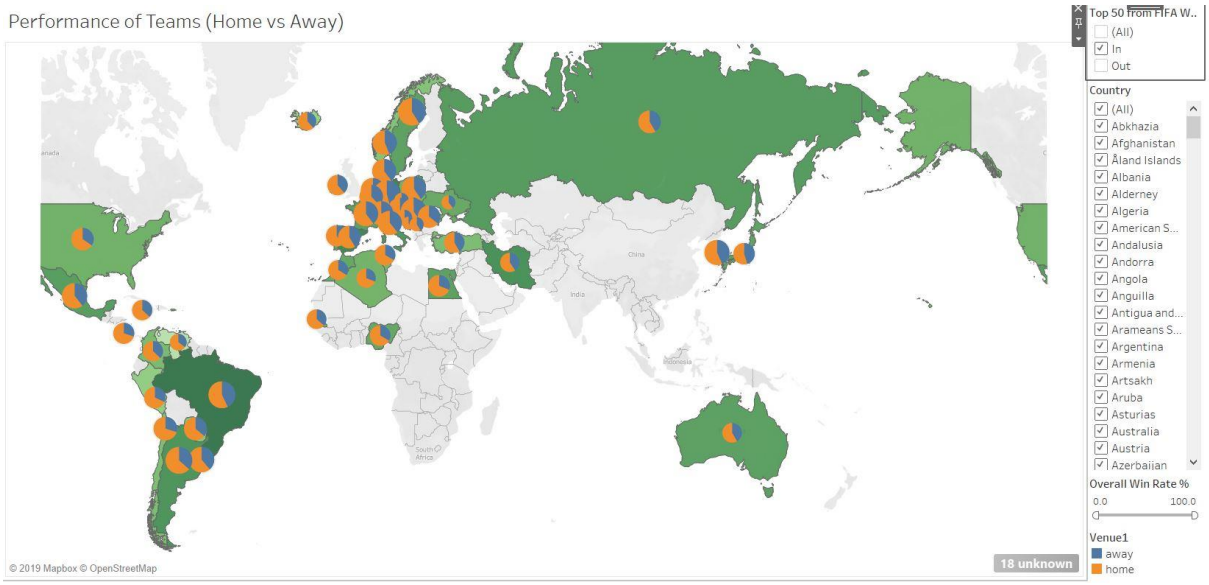

Figure 1: Win rate of top 50 teams

Aggregated over the course of the entire timeline, Fig.1 shows that the European countries, especially the western ones, have generally dominated the sport. The by-year dominance is explored better through the use of the drag-able scatterplot.

Implementation

Feature engineering was implemented using mostly python and Microsoft Excel. The visuals (donut graphs and line charts) used to analyze the dataset (prior to preprocessing) were created using Tableau and python. Tableau's Map charts and area graphs were important and useful in visually representing the win-rates of the countries and to spot the regions dominating the sport, at a glance and to confirm the conclusion drawn from the drag-able D3 scatterplot. Apart from these strengths of the Tableau, it was also more convenient to connect multiple CSV files using the tool.

In order to dynamically visualize the win rates, continents, and strategy patterns across the entire time period, the following drag-able D3 scatterplot was implemented.

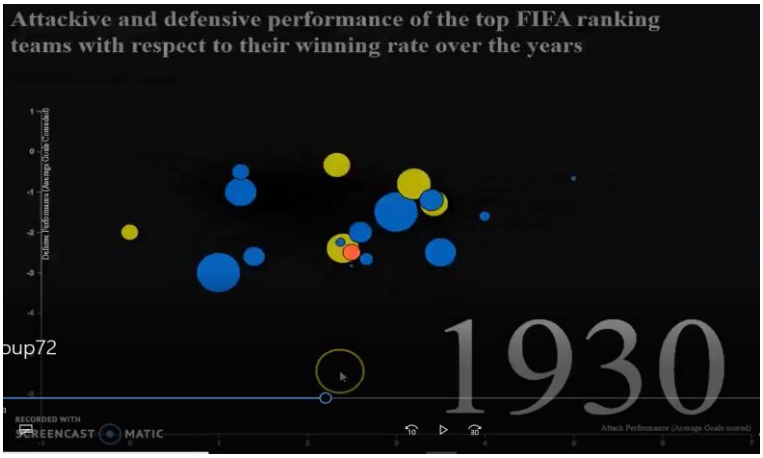

Figure 2: Attack/defense performance of the top 50 teams in 1930

During the early years (Fig. 2), the strategies adopted by the teams appear to be too random to draw any clear insight. Note that the size of the nodes indicates the win rate of the respective country/team, and the colors represent the continent. For instance, the blue nodes correspond to European countries and the orange nodes correspond to South American countries. It is evident that these two continents dominate the FIFA top 50 ranking list in the early stages.

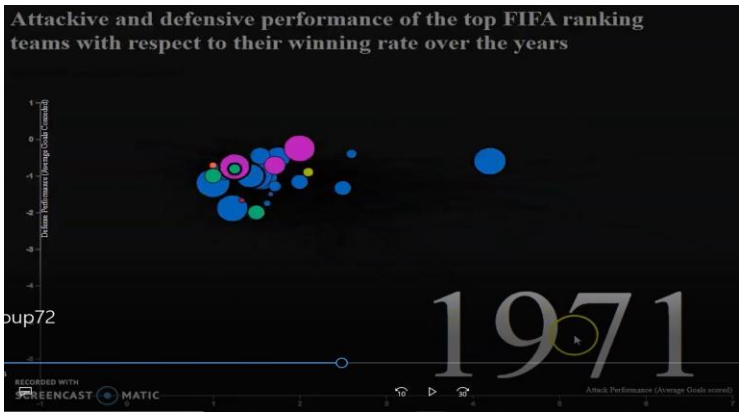

Figure 3: Attack/defense performance of the top 50 teams in 1971

By the early 1970s (Fig. 3), the teams have started to adopt a more consistent strategy - making their defense stronger (notice how the teams have reduced the number of goals conceded, which is why the nodes have clustered to the top of the y-axis, compared to the performance in 1930). This is true for almost all the countries. Other continents had also started making their mark in the ranking.

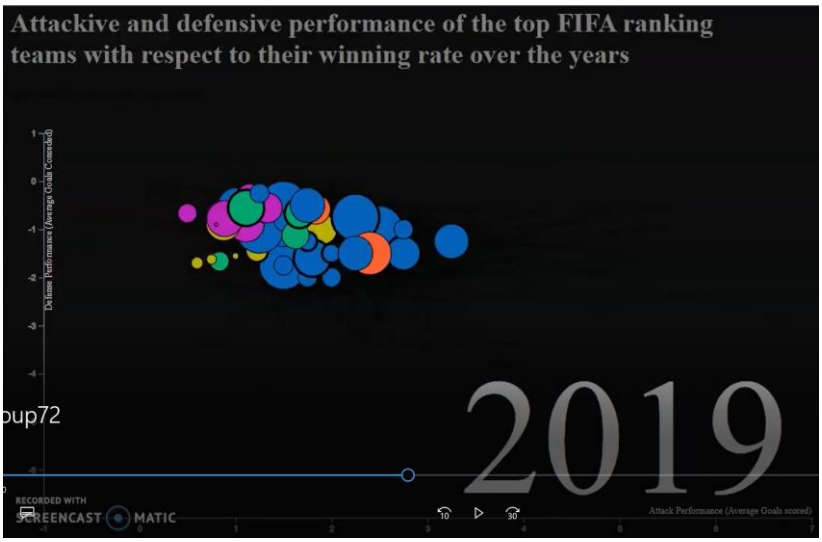

Figure 4: Attack/defense performance of the top 50 teams in 2019

This strategy has lasted in the long run (Fig. 4), with countries from other continents adopting the same to improve their performance. It can be intuitively understood that as the defensive skills of a team improves, it becomes more difficult for the opposing teams to score goals against them. Hence, initially, their strike rate/attack rate decreases as a result of the decrease in the number of goals conceded by the opposing, before converging to a more stable attacking performance (number of goals scored). This is why the nodes have started to cluster towards the left, horizontally closer to the y-axis. An extension to this theory is analysed further in the following graphs.

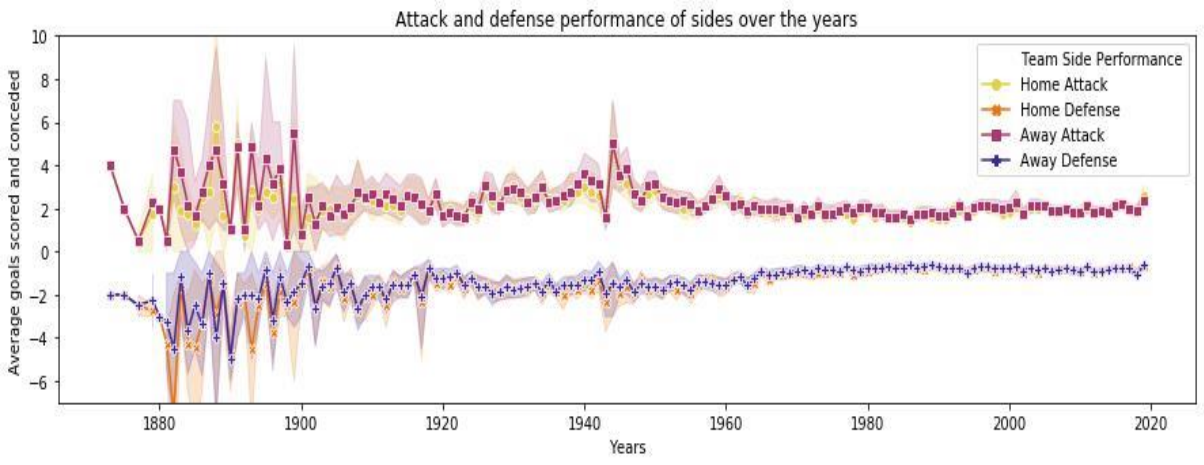

Figure 5: Attack/defense performance of home and away teams (line graph)

Fig. 5 is a more detailed study of whether the venue of the tournament plays a significant role in the teams' choice of strategy by sub-classifying each of the two strategies further into home teams and attack teams. There is no substantial change in the strategy or the attack/defense rates with respect to the venue, visually, and the patterns observed (a general improvement in the defensive skills) is evident here. This is further confirmed by the area graph (Fig. 6) below.

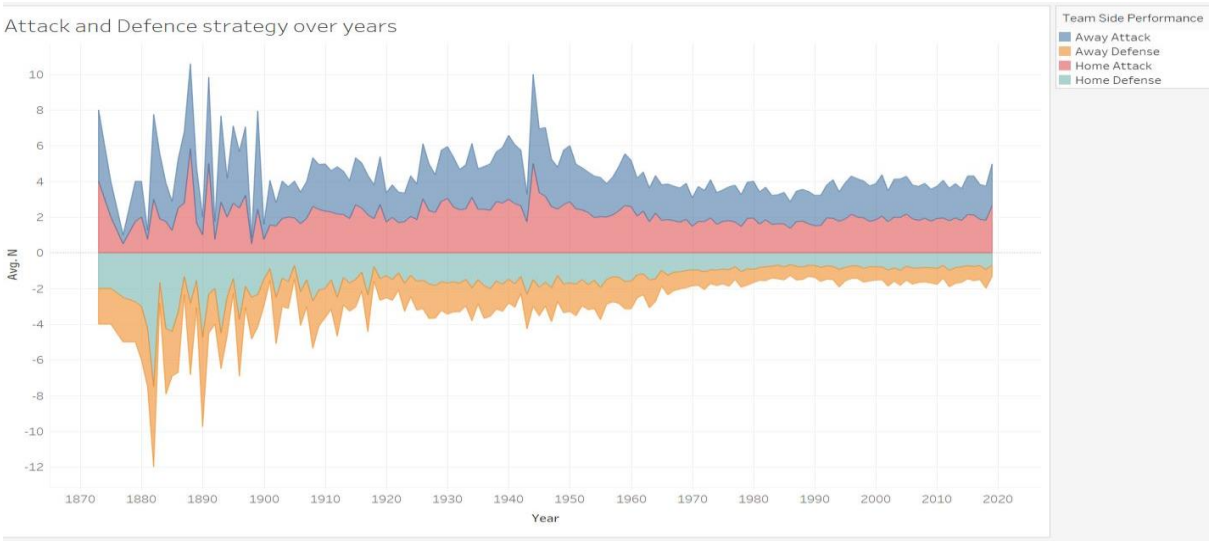

Figure 6: Attack/defense performance of home teams and away teams (area graph)

Pattern of the number of matches drawn across the years

As could be intuitively inferred from the line graph above, the number of goals scored or conceded would decrease and start to converge, as the year progresses. This could mean that in the later years, the two teams participating in a match would more likely score the same number of goals, thus ending the match in a draw. This in turn means that the number of matches that were drawn increases across the years, which is visually represented below. (task 4)

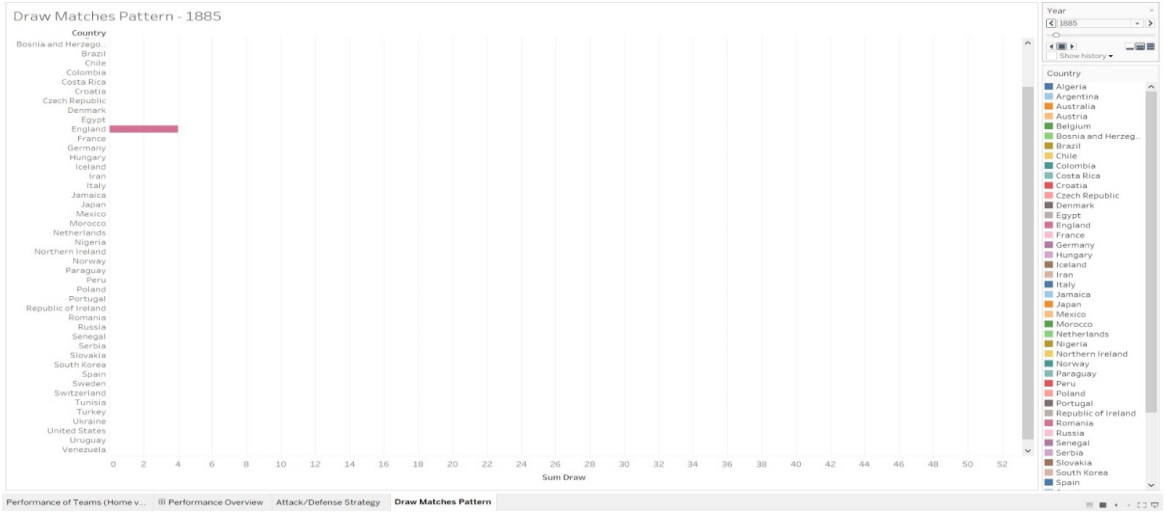

Figure 7: Number of matches drawn in 1885

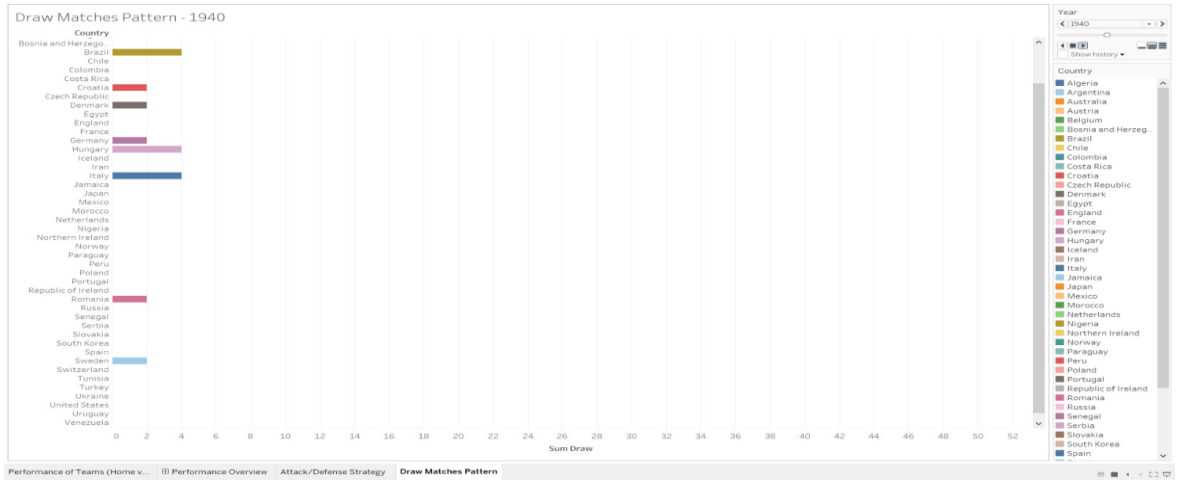

Figure 8: Number of matches drawn in 1940

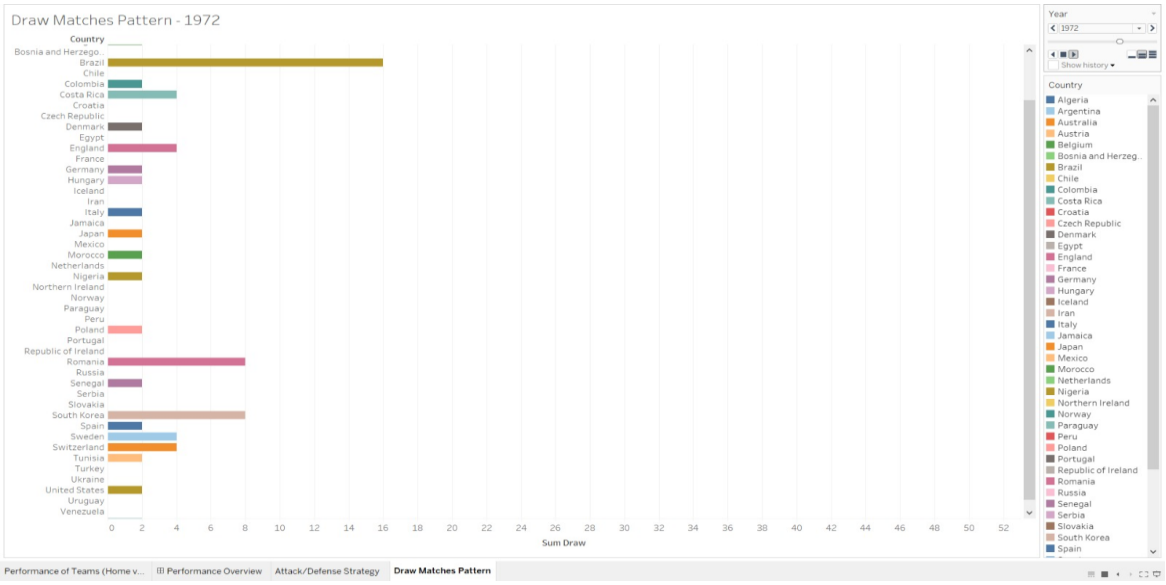

Figure 9: Number of matches drawn in 1972

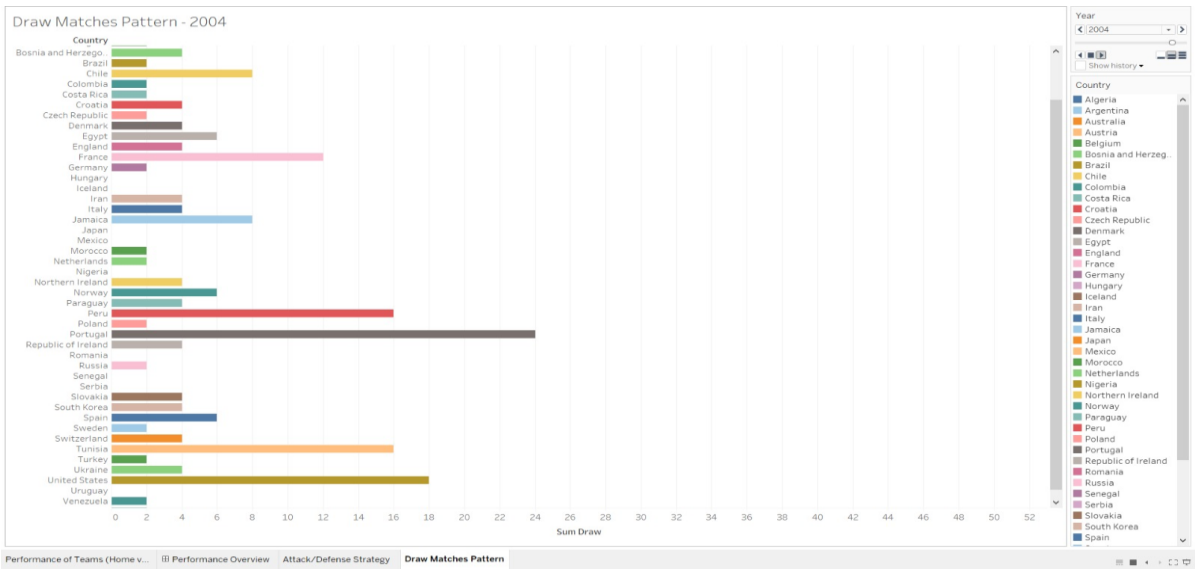

Figure 10: Number of matches drawn in 2004

In the figures 7,8,9 and 10, the vertical axis displays the top 50 countries based on the current FIFA ranking, and the length of the horizontal bar indicates the number of matches that were drawn during that particular year. Graphs corresponding to four different years have been shown to indicate the progress of time, and to explore the patterns that arise.

In the beginning of the timeline (Fig. 7), few matches were drawn, but as time progresses (Fig. 8,9 and 10), the number of such matches have increased for the top 50 FIFA teams, even if there is randomness in the patterns.

Attack and defense strategy over the years

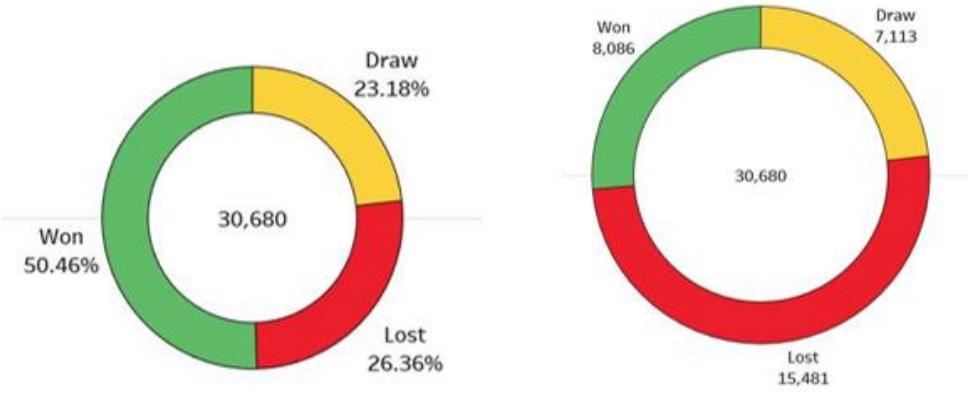

Figure 11: Home (left) and away (right) performance

Despite the seemingly lack of correlation between the tendency to change the strategy with respect to the tournament taking place on home soil or not, there exists a clear advantage for most (if not all) teams when they play in their home country, with ~50% matches won. An alternate narrative would be the 'away' teams winning close to only 25 percent matches. The reason(s) behind this observation cannot be inferred from this dataset and are practically beyond the scope of this assignment (Ron, 2016).

Evaluation

Results

The data was preprocessed and adequate feature engineering was implemented to keep in line with the objectives/tasks of the assignment. Despite the overlapping of the nodes in the scatterplot, a general understanding of the strategy was effectively communicated using the visuals (Pan and Wróblewska, 2016). According to the external test subject survey conducted, the results shown here are the mode of the respective attributes:

D3 scatterplot analysis

Appropriate color (black) used as background	Agree
Appropriate choice of colors and hues for the nodes	Agree
Node sizes appropriate and able to differentiate win rates	Neutral
Font size and appropriateness of the x-axis, y-axis fields	Neutral
Font size of node names (countries/teams)	Agree
Interpret the meaning of the strategy prior to explanation	Disagree
Understand after explaining the theory	Agree

Line graph (Average goals scored/conceded vs. years)

Differentiate between home teams and away teams at first glance	Agree
Differentiate between attack and defense rates	Agree
Sufficient colour contrast	Neutral
Interpret the meaning of the strategy prior to explanation	Disagree
Understand after explaining the theory	Agree