

Visualization Project Proposal

Basic Info

Project Name: Data Visualization of English Premier League

Group Member 1:

Name: CHEN YANG email: chen.yang@utah.edu uID: u0738066

Group Member 2:

Name: Hao Sha email: u1078499@utah.edu uID: u1078499

Link to Github repository: <https://github.com/lucifer2012/dataviscourse-pr-premier-league>

Background and Motivation

The Premier League is an English professional league for men's association football clubs. At the top of the English football league system, it is the country's primary football competition. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League (EFL; known as "The Football League" before 2016–17). Welsh clubs that compete in the English football league system can also qualify. According to statistics, it's currently the most watched and welcomed soccer league in the world, which is broadcast in 212 territories to over 4.7 billion audience. In the past season, its average match attendance exceeds 36,000. With 47 clubs having competed in the league since 1992, there are only six teams are able to win the championship, including Manchester United, Chelsea, Arsenal, Manchester City, Blackburn Rovers and Leicester City^{reference}.

Having supported Manchester United for over 15 years, I am and still will be a fan for England Premier League. For each season, one of the biggest and most interesting question is which team is going to win the championship. Does the champion teams usually score the most goals or do they win the title because they have the best defence? For the teams downgraded to a lower rank, is it because they have the worst defence or because they could not find a way to score more goals? During each seasons, there are also many "big" games, like the derby between Manchester United and Arsenal. Are we able to predict the result more or less based on their performances against each other in the past years? These are all the keen questions that bother soccer fans. Thanks to visualization, now we are able to extract and display some key features of each team to explore why they can/can't win championships.

Project Objectives

There is an old saying that "good attack would help you win the audience, but good defense would help you win the champion". So our first two objectives are to testify if this is true.

- i. The first goal is to explore the relationship between the defense and final ranking of the season. So the idea is to plot those two rankings of the season, and then to decide if they are related.
- ii. The second goal is to decide if there exists a strong relationship between the goals that a team makes and its final ranking.
- iii. Another great application of this visualization is to make a simple prediction of a specific game based on the rival history of the corresponding teams. In the world of football, there used to be a strong pattern that people could follow when it comes to prediction of a game. For example, Aston Villa could only defeated Manchester United once since 1999. Therefore, we are going to display the game results of a team pair for six seasons to help with predictions for the new games.

Related Work

We got a lot of useful input from our TA and peer review. Here are a few of interesting advises that we decide to take:

i. Calender View

Previously we had little knowledge of how to organize all the matches of a single team. But thanks to help from our TA, we decide to incorporate the calender view into our project. Therefore, we would be able to study the performance of a team over a period of time more easily by brushing.

ii. Fifa games

We used to want to display different properties of a team, like attack or defense, through polygons. However, due to the scale of the webpage, we thought this might be difficult for people to figure out subtle changes. Therefore, we decide to show the results by segments. And now we think it's easy for people to spot differences between teams.

iii. Maps

We also want to highlight the stadiums of the teams. We add this as complimentary component to our project as we haven't got time to dig out the json.geo file. And this idea comes from our homework.

Questions

There are multiple questions that we want to answer. Firstly, we want to have a more direct way to monitor the performance of a team over a period. The reason that we want to answer this question is there are many factors influencing a team's performance. For example, Arsenal used to have a bad performance in every April. And some teams tend to play bad at the end of the year. It's yet mature to connect their performance with some factors, but our work is simply the beginning, which is to find a way to find these special periods first. Secondly, we want to know what's influencing the final ranking of the team, defense or attack? For each season, one of the biggest and most interesting question is which team is going to win the championship. Does the champion teams usually score the most goals or do they win the title because they have the best defence? For the teams downgraded to a lower rank, is it because they have the worst defence or because they could not find a way to score more goals? If we have enough data, we also want to test a few more things. For example, the saying is "good attack would help win the audience, but good defense would help you win the title". So once we have the audience data of each time, we also want to learn whether the number of audience is positively related to the attack ability of a team.

Data Source

Our data would be a relatively complete dataset of each team for 6 seasons. For each team, there would be pretty detailed information of its individual games, including half time results, full time results, corners and home/away team shots on target and so forth. The data of each season would be stored in a independent csv file, with the names of columns being the abbreviation of specific results statistics.

i. Full explanation of abbreviations in column names

<http://www.football-data.co.uk/notes.txt>

ii. 2015-2016 Season

<http://www.football-data.co.uk/mmz4281/1516/E0.csv>

iii. 2014-2015 Season

<http://www.football-data.co.uk/mmz4281/1415/E0.csv>

iv. 2013-2014 Season

<http://www.football-data.co.uk/mmz4281/1314/E0.csv>

v. 2012-2013 Season

<http://www.football-data.co.uk/mmz4281/1213/E0.csv>

vi. 2011-2012 Season

<http://www.football-data.co.uk/mmz4281/1112/E0.csv>

vii. 2010-2011 Season

<http://www.football-data.co.uk/mmz4281/1011/E0.csv>

Data Processing

Due to the original data was used for betting, it has some information that we do not need to include in our visualization, such as odds and rates. We expect to clean those unnecessary data away during our visualization. Besides, we probably need to extract some useful information of single teams to make a summary of each team during each season. The column names come in abbreviations in the original dataset, so we might need to change column names if necessary.

Exploratory Data Analysis

The data files that we have are all csv files. So in order to save time, we decide to look at the final rankings of teams online directly. In the commercial websites, we found most of the time, it's very hard to obtain match results over a period of time. So that basically gives us confidence in the usefulness of the calendar view. What's more, we also found that the teams with the worst defense used to rank bottom. So after we finish our implementation, we hope we could find somewhat linear relationship between the defense and final rankings.

Design Evolution

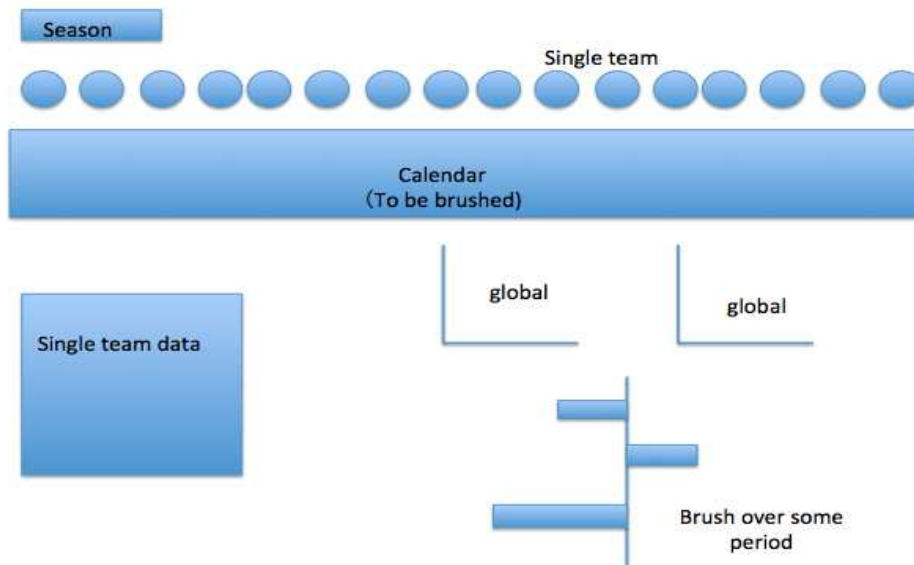
Our design changes profoundly thanks to our teaching assistant and peer review.

- i. Suggested apply to linear regression we were on the point of our scatter graph. This is a fair and helpful feedback. We are trying to see the link of the attack / defense / performance, so a linear regression is very useful in our graph, and we will definitely implement this feature. This was also mentioned in our TA's feedback on possible improvements.
- ii. Change into the pie chart, bar chart. This is a fair feedback. Although the pie chart, tag might be the best way of win / lose / draw rates, it is better to use a bar chart to show the goals / goals against numbers. Those numbers Meaningful specific numerical values, using pie chart will display those values to insufficient.
- iii. Delete team when the user click on the chart and only display team info / hover point on the scatter graph. With the current team is lackluster, but I do not think it is a good idea to completely delete the chart, because the user might want to choose one team to see their team performance. However, I think it might be the better team Visualize to chart the team better by putting more information into the team chart. The following design interests me a lot. <http://www.presentation-process.com/flat-design-powerpoint-org-chart.html>

- iv. Might add our Visualization back to the radar graph. This is a fair feedback. Personally I Liked the radar graph, but it lacks details and performs similar functionality as the bar chart, so we decided to drop it off. Instead, we decide to used segments to show the properties of a team.
- v. Add calendar view to our project as suggest by our teaching assistant.

Implementation

Our current work centers on the first half of the project. Below is the design of our interface(map's not included).



- i. Season button is working. The circles in the top row indicate single teams. Now single teams could be selectable and their team badges have been used to replaced the circles. The tooltips work as well. When the mouse hovers on the badge, the name of the club would disply in the floating box.
- ii. When a single team is clicked in the top row, the information of a team would display in the single team data box. There are a bunch of properties to be displayed, but now for convenience we have only tested the attack and defense. There would be more properties to be added.
- iii. One of our team member is working on the calendar chart as well. The intended effect is to not only brush the calendar chart, but once a specific month is clicked, the corresponding month would spread out to give people opportunities to monitor specific matches. This is a half done work.
- iv. The two scatter plots named global in the figure above is what we use to test the relationships between final ranking and defense/attack. It would be finished in a few days.
- v. The chart below the two scatter plots is what we use to display the performance of a team over a period. The result of brush effect on the calendar view chart would be summarized in this chart. And it would be done next week.

A little representation of our work is shown below(Calendar view chart is not included!)

England Premier League

contributors: CHEN YANG E-mail: chenyang@utah.edu UID: u0738066

contributors: HAO SHA E-mail: u1078499 u1078499@utah.edu

13-14 ▾



Evaluation

I think this part should be finished after the whole project is completed. So I'll leave this part blank until the whole work is done.