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 explaination 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.

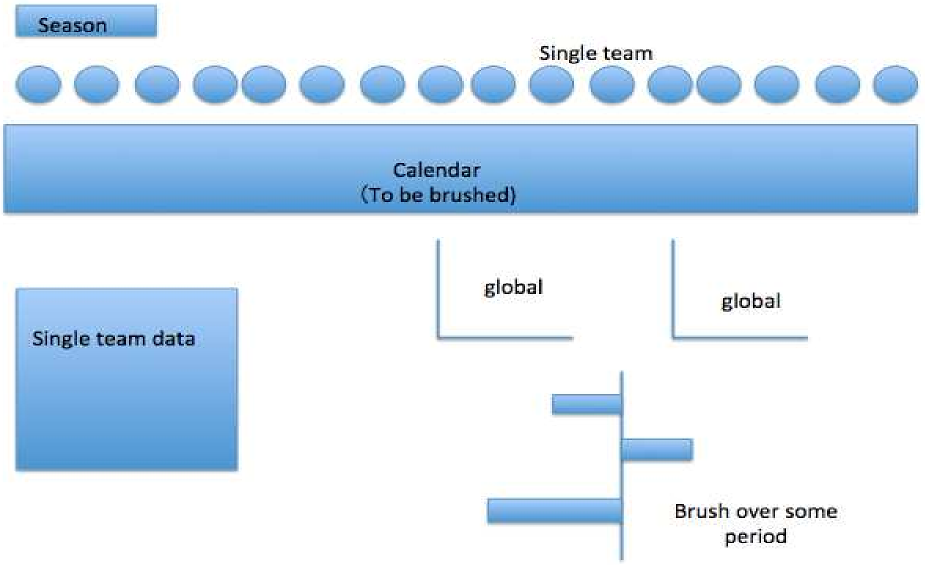
iii. Delete team when the user click on the chart and only display team info / hover point on the scatter graph. With the current team is lackluster, but I do not think it is a good idea to completely delete the chart, because the user might want to choose one team to see their team performance. However, 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

Below is the design of our interface(map's not included).



i. The first chart is called teamchart, as shown below in fig. 1. It's purpose is to allow people to select a specific team in a specific season. As shown, the season button is clickable to render selection of different seasons. Acutally team badges have been used to replace those circles. Tooltips have been added as well to display the name of the team when mouse hovers over a specific team badge. When a specific team is clicked, the property chart, which is used to display specific properties of a single team, would respond.

ii. The property chart includes a rough introduction of the performance that a team has. Our initial intention was to use a ploygon to display those properties. However, we negated that idea in the end due to the small size of the chart might make the changes in the chart very hard to notice. So we use what's shown below instead. The attack and defense abilities are measured and ranked by the goals that a team made/conceded during a season. More bars in the rows of attack/defense indicates better performance a team has achieved. At the same time, we compared and summarized the goals that a team made/conceded in the first and second half of the matches. The result would be discussed in the part of Evaluation.

iii. The next part of the project is called month chart (Fig. 2). The chart is displayed after click on a specific team. For any team during a season, it's supposed to have 38 matches in total. And the intention of this chart is to color the matches according to the match results (win, draw and lose) and show it to readers, so that readers would have a general idea of what performance a team had for a match or matches over a period. The chart would be brushed, and the results of brushed matches would be summarized in another chart.

iv. The global chart intends to explore factors that might affect final rankings of the teams. The factors listed include attack, defense, capability to create chances (Chanced make) and capability to take the chances created (Chances take). The ability to attack is measured by the goals that a team made during a season, and the ability to defend is measured by the goals that a team conceded during a season. The ability to create chances is measure by the total number of shots of a team during a season, while the ability to take the chances is measured by the total number of shots on target divided by the total number of shots. The reason to measure the four aspects of a team is because the final ranking of a team is affected by multiple factors. If a team didn't rank high in the end, it is probably because it could not shot enough goals to ensure victory, or it might be due to its bad defense. If it could not shoot enough goals, then is it because the team failed to create enough chances or because the forward players coul3d not turn chances into goals? The questions that we asked are key to the final rankings of the teams and that's where we need visualization to solve the puzzles.

v. The

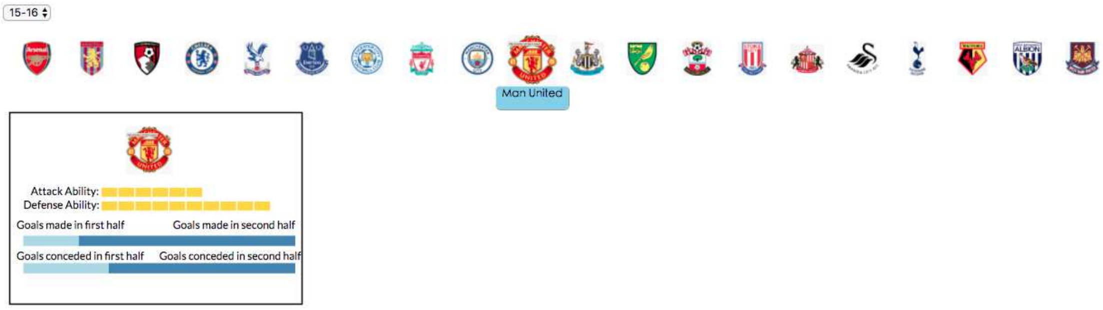

15-16
Man United
Attack Ability:
Defense Ability:
Goals made in first half
Goals made in second half
Goals conceded in first half
Goals conceded in second half

Fig. 1

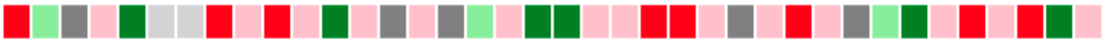


Fig. 2

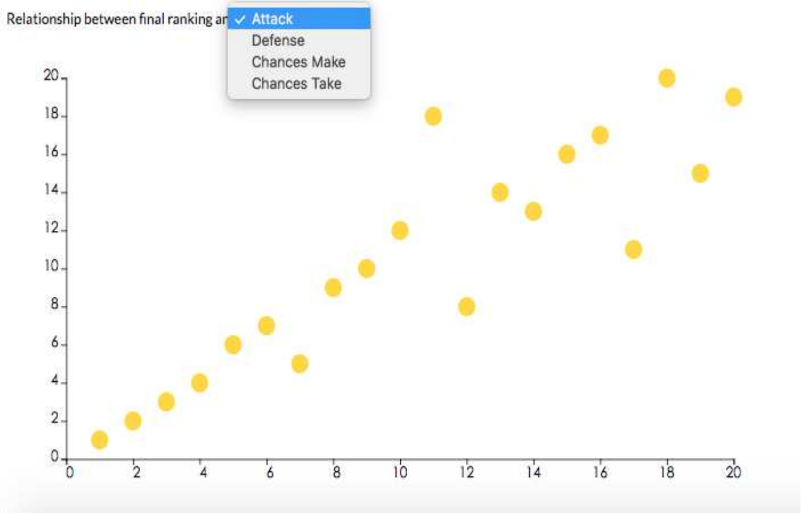


Fig. 3

Evaluation

As this project includes several charts, we think we'd better evaluate them one by one.

1. Team chart: As shown in fig. 1, I think team chart has met our demand. The purpose of this chart is to demonstrate all the teams participated in games for a specific season. According to performances of teams, the last three teams would downgrade to a lower league and the

top three teams of the lower league would get promoted after every season ends. So that basically results in the changes of teams every season. So with team chart, readers would have a direct understanding of what teams play in that particular season. If the readers are familiar with English Premier League, then they could identify teams by team badges. Otherwise, they could identify teams by putting the mouse over the badge icon, which would display the names of every single team. With the dropdown box, readers could select a particular season that they wish to investigate.

2. Property chart: It's believed that property chart is pretty straight forward. The chart would be produced after a click on the team badges in the team chart. The intention of this chart is to display some specific properties of the selected team, so that readers could have a general idea of the performance of the team. The focus of this chart is to show how well the team does in attack and defense when compared to other teams in the league. More gold bars in the attack/defense indicate a better job. Therefore, we believe properties like attack and defense could successfully deliver the idea of how a team does in those aspects. In this chart, we also include the goals that a team made/conceded in the first/second halves. By browsing the data in the first and second halves, we found that teams tend to score and lose less goals in the first half. We think this might be because in the first half, team players, especially those who play defense, are more energetic and more focused, so that it is more difficult for players responsible for attack to penetrate the defense of the other team. However, when it comes to the second half, we noticed that more goals could be socred or lost. We think this is relatively easy to explain: team players lost much energy so that they would be easily distracted from defense. Players responsible for attack would take advantage of that and scoring became relatively easy.

3. Global chart: This chart is intended to explore the relationship between final rankings and other facters, including attack, defense, capability to create chances and capability to take the chanced created. We've got some interesting findings. Let's take season 2014-15 for example with y-axis representing the factor variable, x-axis being the final rankings. For that particular season, attack seems to have the the strongest influence over the final rankings, where it seems to exist a linear relationship (Fig. 4). Defense is important too, but not as important as attack (Fig. 5). The final rankings are as well positively correlated with the other two factors, capability to create and take chances. In addition, we could figure out where some teams need to improve if they want to promote their rankings. For example, for those teams with fair ability to take chances but not to create chances, what they need to do next season is to introduce talented players to increase the chances they can make.

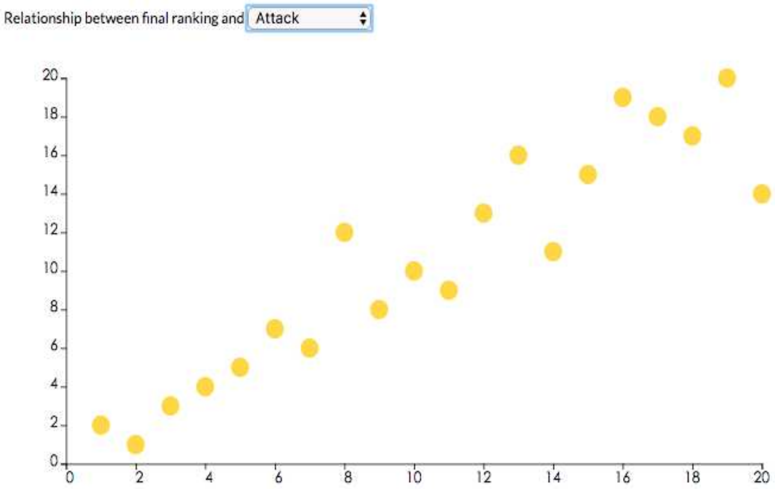
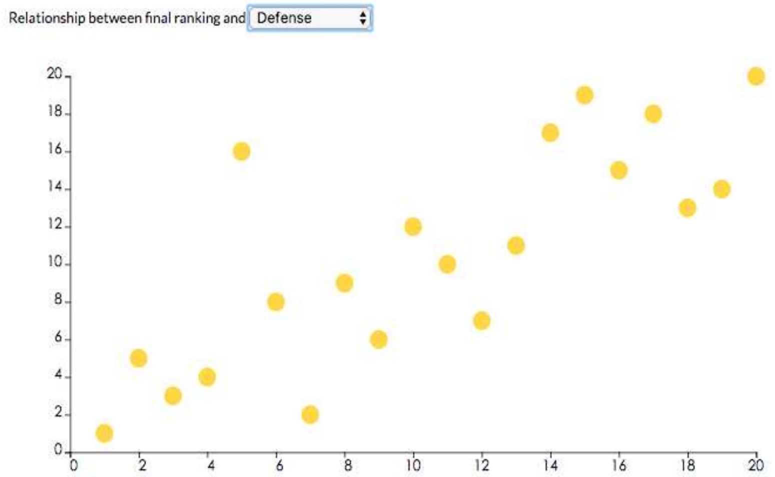
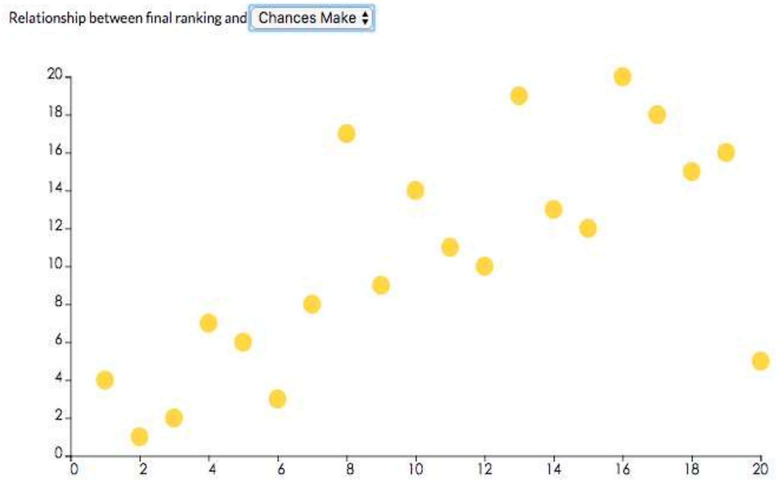


Fig. 4



Relationship between final ranking and Defense

Fig. 5



Relationship between final ranking and Chances Make

Fig. 6

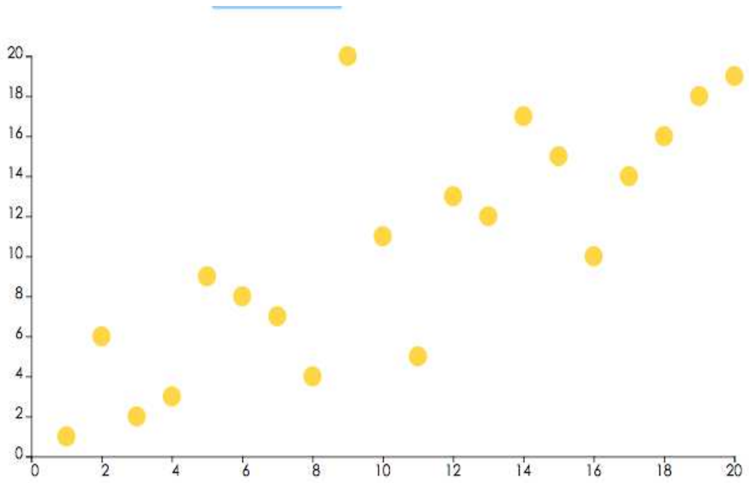


Fig. 7