University of Amsterdam

Final Project

A short history of the English premier league through 4 visualizations

Professor

Dr. Thijs Coenen

Student

Bobby den Bezemer

Status: Final draft / June 26, 2015

Introduction of the assignment

The Premier League is the most-watched football league in the world, broadcasted in 212 territories to 643 million homes and a potential TV audience of 4.7 billion people (Ebner, 2013). Its inception took place in 1992. Since then, 47 teams have competed in the league. The data visualization written for this course will deal with data from this particular football league. In short, its goal is to give a brief history of the Premier League through 4 data visualizations.

The remainder of this document is organized as follows. Firstly, I will discuss the problem domain of the visualization as well as the steps taken to acquire the data in order to treat this problem. Secondly, I will discuss the steps taken to translate the problem definition into computer language. Thirdly, I will describe the visual encoding and user interactions that I designed for this problem. Fourthly, I will describe the implementation of the visualization. Finally, I will wrap up with a section in which I reflect upon the final product and the choices that I have made during the process of coming to this product.

Target

Problem domain

The problem domain of the current data visualization pertains to the English Premier League. Although many football websites display certain statistics about this football competition, to the best of my knowledge very few website combine this information in an integrative and interactive data visualization. It has selected this information carefully such that the 4 goals of data visualization could be satisfied and that several hypotheses could be tested.

Firstly it serves the goal of recording information. For instance, the first part of the visualization combines information on the ranking of Premier League teams and Premier League top scorers on one page. Likewise, it serves the goal of analysis. By allowing the user to interact with a drop down menu, the user can explore team rankings and top scorer statistics. As such the user can create the argument that Manchester United has been the best team since the inception of the Premier League.

The second part of the visualization serves two different goals: revealing patterns and story telling. The interactive map that portrays the distribution of Premier League teams throughout the UK over time tells the story of industrialization and urbanization. As put forward by the book soccernomics, a disproportional part of England's Premier League teams are to be found in the old industrial areas of Merseyside and the greater Manchester area. The interactive map actually shows that this pattern holds with most of the teams to be found in this area while very few teams are to be found in the south of England. The linegraph on the right of the map (displayed in Figure 2) reveals the pattern of the ever increasing amount of foreign players in the Premier League. While in 1992 only 30% of the players were foreign, in 2014 this seems to have increased to a percentage well above 60. Furthermore this visualization, in collaboration with the map on its left, allows the user to compare different teams on their development of foreign players. It for instance shows the user that the clubs that have seen a influx of new wealth (e.g. Chelsea and Manchester City) are way above the average in terms of their amount of foreign players. This is easily explained by their expenditure drift in terms of buying foreign players in recent years.

Data collection scripts

In order to tell this story about the English Premier League, one needs data. Much of the data that was needed for the current visualization could be found on a range of different websites. Information on Premier League top scorers was found on wikipedia. The information on the location of the football stadiums of the English premier league was found on doogle.co.uk. The information on premier league rankings was found on statto.com and the information on the number of foreign players per premier league team was found on transfermarkt.co.uk. In order to extract the data from these websites, I used a common technique of web crawling. I made use of the python library pattern which was of great use. Not only did this library provide an easy way to extract the information, it also provided to be more powerful than some other libraries. For instance, I failed to extract the data from statto.com and transfermarkt by means of the R library rvest. Pattern on the other mimicked the behavior of a general website user more closely and by doing passed by certain security measures.

Translate

Tasks and operations supported by the visualization

After the problem domain had been formulated, one had to think about the visualization itself and what tasks it should support. The first part of the visualization that contains the Premier League Rankings and top scorers should support at the bare minimum an update function. This update function allows the user to update the data through selecting a new year. Furthermore this part of the visualization should have a mouse hoover function that allows the user to hoover the bar chart in order to gain more information about the particular team.

The second visualization contained a greater amount of user functionality. For instance, it should support various ways to interact with the timer. As such the visualization includes 3 different buttons: a pause button, a continue button and a restart button. Each button allows the user to interact with the visualization in its own way. Furthermore, in case a user wants to display a certain Premier League year, this visualization also contains a drop-down menu. Lastly this visualization has some hybrid functionality that allows the user to select teams on the map by clicking on them. In turn the line graph on the right side of the map will be updated and now includes the amount of foreign player per year of that given team. Another click on the same team will update the line graph and remove the line and the particular legend. In order to remain tidy, at most 2 footballs teams could be compared on the particular line graph.

Data processing

Before the data could be visualized, much data wrangling had to be executed. All data wrangling was done in python, and mostly through the pandas library. Most data wrangling was done according to the guidelines as formulated by Hadley Wickham (2014) author of many R packages on data wrangling. For instance, each variable had to constitute a column, each observation had to a row and each observational unit forms a table. As such most data had to be converted to long format to abide by these standards. This was done through the pandas melt and pivot functions. Also each observational unit was put in its own Dataframe. As such that data frames that I had to work with in javascript constituted of at most 5 columns.

Design

A first sketch of the visualization is depicted in figure 1. This sketch guided the development of the visualization. Note that this is a guide of the minimum viable product and has greatly been expanded in the final product.

As you can see a two row 1 column layout has been chosen for the minimum viable product. Each visualization was accompanied by a header and a short explanation of what it displayed. The final product is shown in figure 2.

As you can see, in terms of design a style dominated by steelblue has been chosen. The bars of the bar graph, the table, the clubs displayed on the map and the average line in the line graph all contain this colour. Thus this aspect is repeated throughout the visualization. By doing so, this abides by Robin William's principle of repetition. Likewise, in terms of positioning, those elements that are close to each other are related. For instance the bar graph is positioned next to table that contains the English Premier League top scorers. The premier league map and line graph are positioned next to each other as well as through interacting with the map, one can change the number of lines displayed in the line chart.

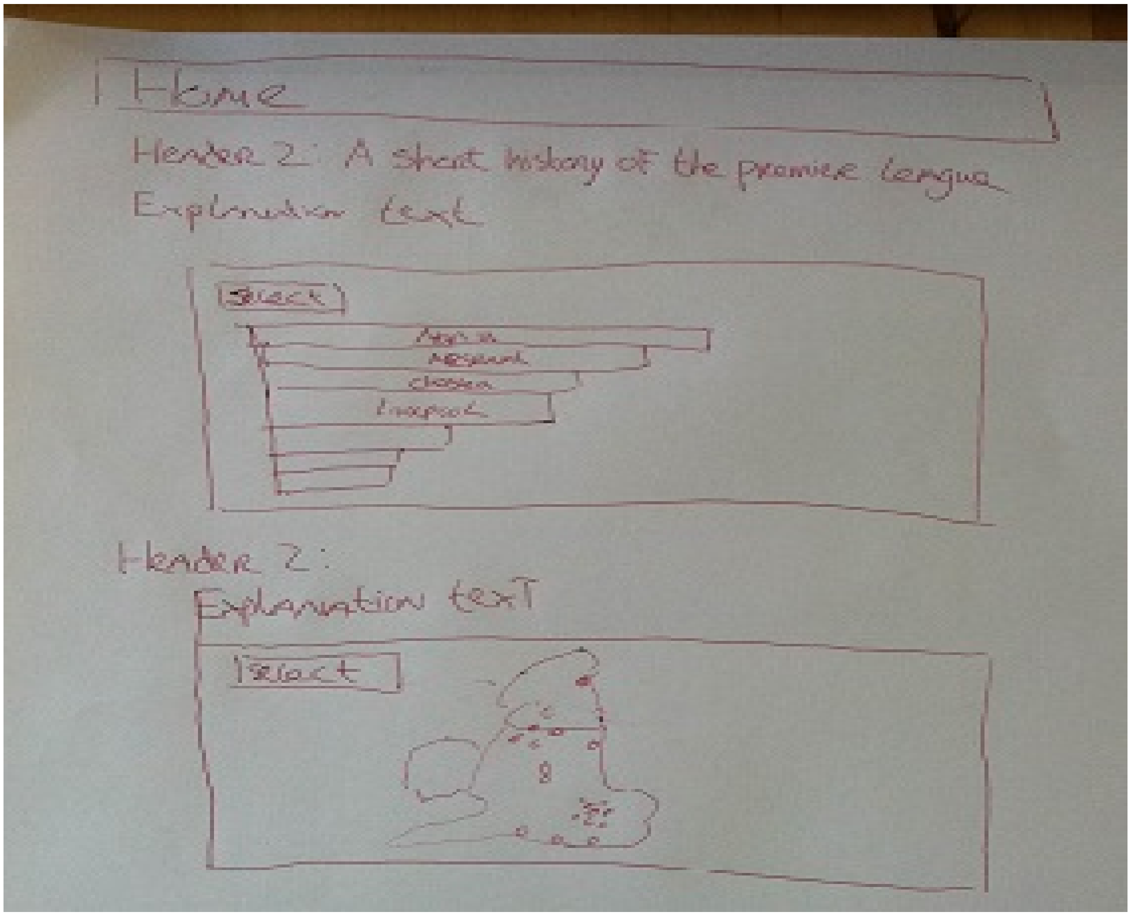
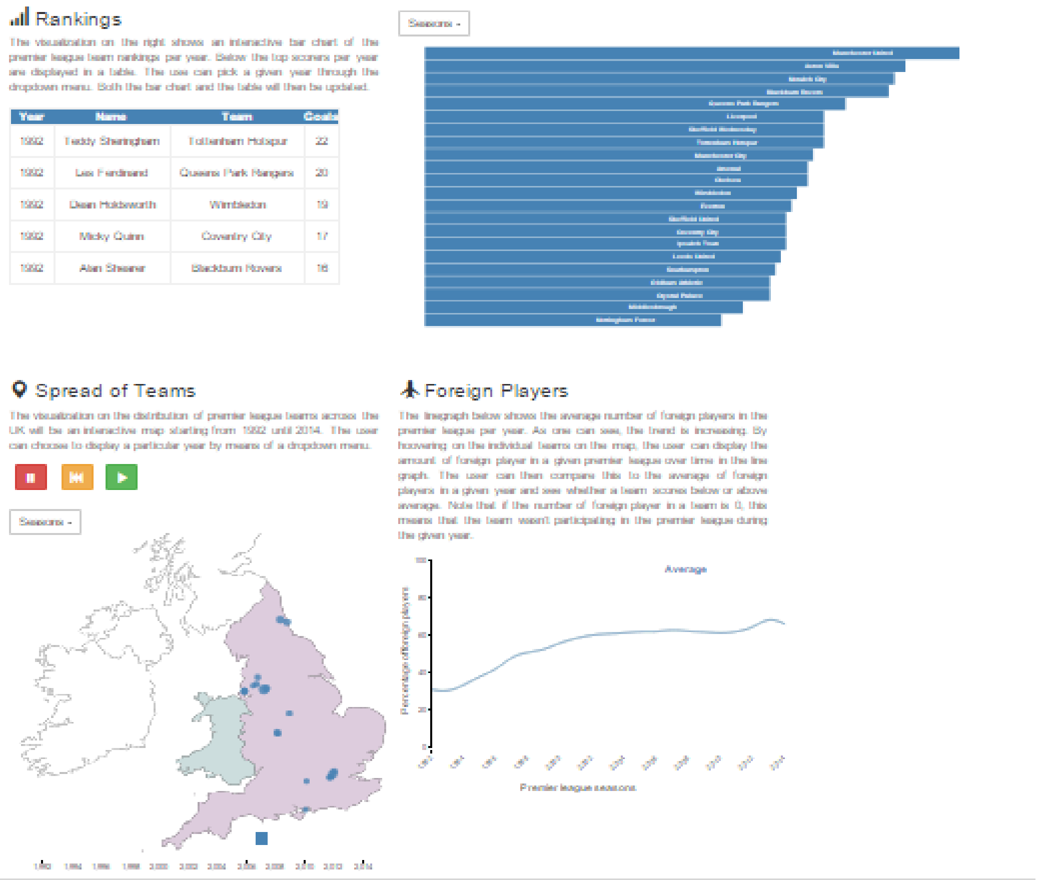


Figure 1. Figure 1: A sketch of the minimum viable product

The reason why premier league rankings are displayed in a bar graph is that humans find it easy to discern between position and size. This bar chart combines both elements. Firstly it is ordered such that teams that are higher on the page have acquired more points during the given year. Likewise, the size of the bar corresponds to the amount of points acquired such that the wider the bar, the more points the given team acquired. In case the user wants more information, she can hoover the bar chart. The rank of the team, the given team name and the amount of points acquired will then light up.

The reason why the distribution of the premier league teams is displayed on a map is obvious. To the best of my knowledge, there is no better way to display geospatial data than on a map. The reason why a line chart has been chose to display the trend of percentage of foreign player per year is that line charts generally portray time series data



Rankings

The visualization on the right shows an interactive bar chart of the premier league team rankings per year. Below the top scorers per year are displayed in a table. The use can pick a given year through the dropdown menu. Both the bar chart and the table will then be updated.

Year	Name	Team	Goals
1992	Teddy Sheringham	Tottenham Hotspur	22
1992	Les Ferdinand	Queens Park Rangers	20
1992	Dean Holdsworth	Wimbledon	19
1992	Micky Quinn	Coventry City	17
1992	Alan Shearer	Blackburn Rovers	16

Spread of Teams

The visualization on the distribution of premier league teams across the UK will be an interactive map starting from 1992 until 2014. The user can choose to display a particular year by means of a dropdown menu.

Foreign Players

The linegraph below shows the average number of foreign players in the premier league per year. As one can see, the trend is increasing. By hoovering on the individual teams on the map, the user can display the amount of foreign player in a given premier league over time in the line graph. The user can then compare this to the average of foreign players in a given year and see whether a team scores below or above average. Note that if the number of foreign player in a team is 0, this means that the team wasn't participating in the premier league during the given year.

Figure 2. Figure 2: A display of the final product

in a good way. Also this line makes use of position to show this percentage which, according to Cleveland and McGill (1984) is the preferred is superior to length or angle. The line chart also makes use of Robin William's principle of contrast. When the user displays a new team line in addition to the average line, the team line will obtain a very different colour than the average. This is because if two items are not exactly the same, then make them really different.

Implement

In order to implement the visualization itself, several technologies where used. Firstly the body of the webpage was built in html. To make it responsive, the popular bootstrap framework was used. To further style the webpag, css was used. The entire visualization itself was built through the javascript d3 framework. Through d3, firstly some svg elements were created and put on the page. After loading the data itself, one main function was called. This main function contained the core functionality of the visualization. Within the main function several other helper functions were called. Also, all sorts of event handlers were added to html elements. Some problems arose at first in the data update functions. It took a week to implement a fading in of the new teams that were added to the spread of Premier League teams map. The problem was due to the ordering of the enter, update and exit selections in the d3 code. Other problems arose with regard to the interaction between the map and the line chart. I failed consistently at updating the line chart by clicking on individual premier league teams. The problem was overcome by attributing proper css classes to the lines and legend items.

Validate

After the establishment of the data visualization, one has to test this and reflect upon the final product. The testing of the visualization was done through panel interviews. More specifically, I have asked a group of friends to test the visualization and reflect upon its usability. During this process some bugs arose. For instance, by binge clicking on all the different buttons together, one could cause the visualization to crash. After this realizations, I went back to fix the code. Other things that arose during the testing phase was an inconsistent updating of the graph legend which had to be fixed as well. Apart from the panel interview, other testing took place through using and debugging took place through printing many things to the console.

All in all, I am satisfied with the final product and most of the design choices that I made. There is enough interaction for the user to keep themselves busy, and the visualizations yields some interesting insights as stated in the problem domain section of this paper. There are however several things that I could have improved upon. Firstly I could have changed the colour pattern of the map to better match the steelblue colour of the visualization. Secondly, in terms of the datasets that I acquired through crawling the web, I could have made some additional visualizations with the goal of finding associations between different variables. For instance, I would have make a scatterplot where I try to correlate the amount of points acquired by a team and the number of foreign players per team. Also I could have incorporated the market capitalization of premier league teams and correlate that with the amount of foreign players per team. Lastly, I could have tried to scrape data on wage expenses and transfer expenses and test some additional hypotheses such as which variable is mostly associated to the amount of points acquired per team.