Data Visualization Milestone 3 Process Book

By: Georgios Fotiadis, Florian Genilloud, Gerald Sula

Table of Contents

Introduction

The goal of this project is to visualize various statistics related to football and betting. Specifically, we wanted to see if there is in fact a "safe" strategy when betting and in general if data can draw some definitive team dynamics from actual data. All of our team's members are huge football fans so it's only natural that we decided to do something relevant to it for this project. At the same time none of us has ever participated in betting because we were afraid that we're going to lose and that's why we decided to approach this issue from a scientific perspective and determine if our fear is reasonable or not.

Path followed to obtain the final result

To create our website, we had to follow a long process which included a lot of independent steps that we will analyze in detail in the rest of this section but in short the stages we completed that lead to our final result were: deciding which subject we want to cover, what aspect of this subject we'd like to explore, find data relevant to that, think about what we would like to visualize, see if these visualizations can be actually done with the data that we have and, finally, evaluate our final outcome to make sure it corresponds to what we wanted to achieve in the first place.

Deciding a subject was not easy. At first, we thought about doing something related to the COVID-19 pandemic, but a lot of work has already been done about this subject and we wanted our project to be original. Then, during a meeting we realized that all of us share the same passion for football, so we immediately decided to do something relevant to it. There were a lot of different directions we could follow like focusing on team performances,

individual player statistics or even trying to see how well video games like FIFA 2020 represent team and player dynamics.

Eventually, we decided to go with football and betting as, in our opinion, this also gives a social aspect to our project. Gambling addiction is a real modern problem, reportedly 350K people suffer from it in the UK alone and betting on football games represents a big percentage of those cases. With this project we want to showcase that the risk of a bet is disproportionately bigger than its possible gains or to at least provide some guidance on how to play in a safer way.

After pinpointing the exact topic we wanted to work on, we had to find data, a task that proved harder than expected. Although there's an abundance of data related to football, most of them were irrelevant to our objective and at the same time betting providers didn't offer any APIs. It took a lot of searching to find sources providing reliable, relatively clean and consistent data that spanned multiple seasons and different leagues.

Having collected the data and performed some initial cleaning we started exploring them to see what information we can extract and visualize from them, a subject we will analyze in detail in the following chapters. Then we needed to see if these visualizations are actually feasible and start implementing them. This was the most time-consuming step of the whole process and naturally some setbacks occurred (more details on the next chapter) but we resolved them and moved on to putting everything together and improving the overall design of our website. Finally, we looked at our outcome from a higher level, compared it with what was our initial idea and made the necessary adjustments to make sure that the two align.

Challenges we faced and design decisions

In this chapter we will be analyzing our design decisions as well as the challenges we encountered when developing our website and how we overcame them. As far as design decisions go, we tried to give our website a minimalistic look following the material design principles. We went for soft colors that are easy to the eyes and made sure we don't have color combinations that would be hard for a color-blind person to distinguish in any of our visualizations. We also decided to use dropdown menus to allow our users to customize the visualizations and to make their overall experience more interactive. Finally, we structured the whole website as one big page to make it easier for a user to explore it in its entirety.

As far as challenges go, we faced a few of them along the way. The first problem we encountered, as mentioned above, was the data collection. We searched extensively in Kaggle and the other sites that were suggested but most datasets were not at all related to betting. At the end, we concluded to 3 datasets: two from Kaggle and one from an independent site which contained a variety of statistics spanning many leagues and seasons. Then, during the implementation phase of our project, we realized that the data we had were not suitable for supporting one visualization we had in mind. We'll provide more details on this subject in the next section but in short, we decided to come up with a new visualization which shows similar information as the one we initially wanted to make.

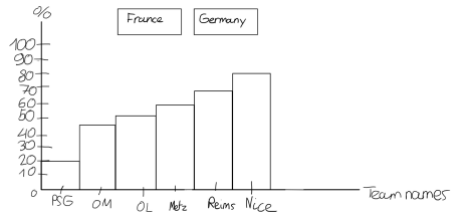
Finally, the last problem we encountered was to make all the visualizations follow the same design language to add a better “flow” to our website. For a detailed description of the per-visualization challenges and design decisions please refer to the dedicated subsections.

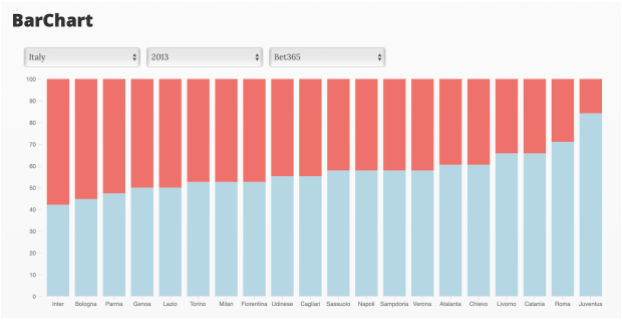
Original sketches and their implementation

Introduction

In this section we will be presenting and explaining our visualizations. We split this section in 6 smaller subsections, one for each visualization, and in each one of them we will be showing the original plan/idea of each visualization alongside its final implementation, the reasons we diverged from our original idea (in the cases that we did) and a description of what message we’re trying to pass to the user.

Barchart with performance predictability per team





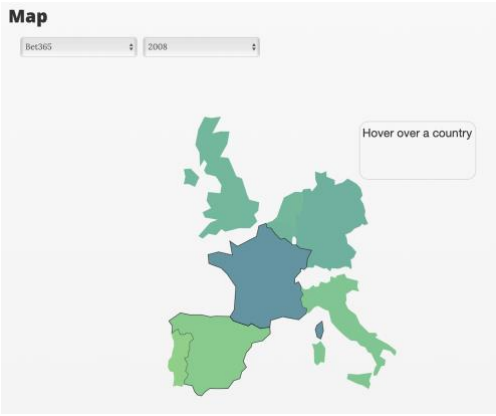
In the pictures above you can see the original sketch we had in mind and our implementation of it. In this first visualization graph, the aim is to show how predictable the performance of each team in the major European leagues is. The idea is to use the data from the different betting companies we have in our dataset and look at how they had decided on the coefficients of the results of each match (Win - Draw - Lose): the lower this coefficient, the more confident the bookies are on that result. We then compare with the actual full-time result of each match and see if the bookies were right or not.

We want to do this for every one of the major leagues in Europe, and make the visualization interactive, so that the user can pick the country/league using a set of dropdown menus. We also sort the teams based on their predictability to make more evident which teams are the most “safe to bet” on. Furthermore, there are some interesting insights we can see from this barchart. It looks like the most predictable teams seem to be either the best teams of the championship, or the worst ones, meaning that it’s easier to say that these teams will constantly win, or lose respectively.

The only change we made was to fill the empty space of each column with red to make it easier for the user to understand how often the bookies are wrong for each team. Although our default color choices were green for correct predictions and red for false ones, we

decided against it as it would make the visualization hard to understand for color blind people.

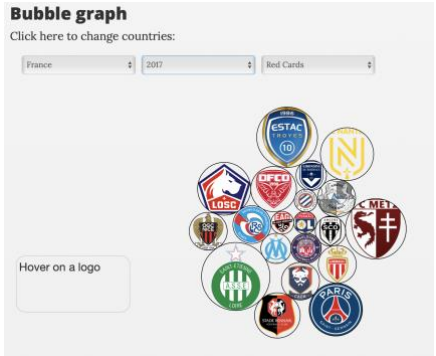
Map with biggest prediction accuracy per country



In the picture above you can see the original sketch we had in mind and our implementation of it. The idea here is that we have a map of Europe and each country's color shifts from dark to light depending on the accuracy of the bookie's predictions for this country's league. This provides the user with an easy guide to select on which country's championship they'd like to bet on and which ones they should avoid. To make this visualization interactive we allow the users to select which provider's predictions they'd like to evaluate and for which year. Additionally, when the user hovers over a country with his mouse, the opacity of the other countries reduces and the bookies' prediction accuracy for the selected country appears in an information box, next to the map.

As far as changes go, we decided to have an interactive map instead of a static one, to make the user's experience more fun and in terms of design, after testing both options we concluded that the best way to show the actual accuracy value was in an info box outside the map in order to avoid rendering the entire visualization bloated and confusing. Finally as far as challenges go, the biggest one was to combine data from different datasets in order to add more countries in the map and for the visuals, finding a color scale that offers enough granularity to highlight the small differences between the accuracy percentages of different countries.

Metric bubbles

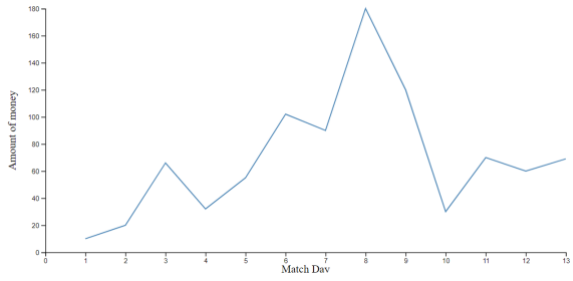


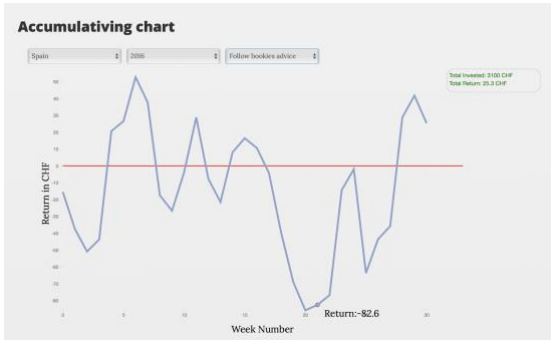
In the picture above you can see the original sketch we had in mind and our implementation of it. The idea here is that we have a variety of metrics (wins, red cards, corners etc.) and a bubble for each team that participated in a major league at some season. The bubble's size denotes how much of that metric a specific team achieved in that season (the more, the bigger the bubble) and this way we aim to show how teams perform on those metrics and reveal possible small correlations between them in a more innovative and interesting way.

Regarding the design decisions we took for this visualization they are mostly concentrated on the level of interactivity of it. First of all, as in the majority of our other visualizations, we include dropdown menus to enable the user to see the information he's most interested in and have an info box to present that information. Additionally, we included a "mini-game" inside this visualization: the users can drag around the bubbles, which bounce with each other, but they always try to move to the center of the screen, as if there's a gravitational force being applied on them. Again, we preferred having the actual information on an info box which is separate from the rest of the visualization to make the latter less cluttered.

Implementing this visualization was very challenging. Compared to the others, cleaning and formatting the data was fairly easy but the visual aspect was very detailed and hence needed a lot of attention. Finding images of the logos of all the teams that participated in all the leagues in the last six years that are not bounded by copyrights was a rather tedious and time-consuming process while modeling the physics of the visualization was also something that we had no previous experience with.

Diagram with how profit evolves based on betting strategy

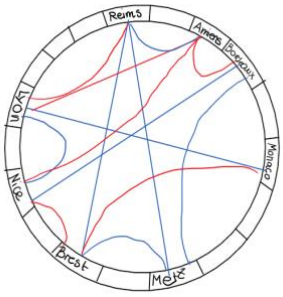


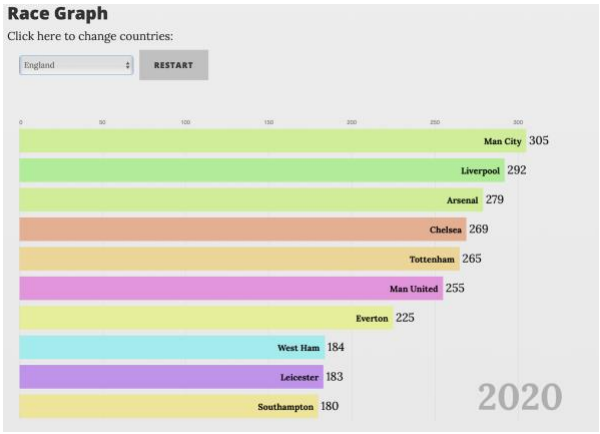


Once again, above you can see the original design of this visualization and next to it our implementation. This is in our opinion the flagship of our website and the visualization that will result in the biggest social impact. The idea here is to give the user the option to see for himself how different betting strategies perform. For example, the user can choose to either always follow the bookie advice while betting 10CHF every day or play in a riskier way (but with a bigger profit perspective) by always betting against the bookies. After extensive testing of our simulation algorithm we concluded that making a profit is a very rare event and even then, the profit never exceeds 10% of the original investment. In other words, the risk is way too high and the return too small and that's exactly the message we're trying to communicate to our users. We strongly believe that if people that have an addiction problem with betting see with their own eyes, in a very simple visualization that they can't make money this way, it's going to encourage them to change their behavior or ask for help.

We were very careful when making design decisions regarding this visualization. We wanted to keep it as simple as possible but at the same time offer all the information we judged necessary. Having a red line denoting that the user has basically lost all his investments adds a dramatic effect that in our opinion makes the image more impactful. As in the previous graphs we have dropdowns to help the user choose what information he wants to visualize but now instead of having an info box we show the current return on the investment directly on the graph. Nevertheless, we maintain an info box showing the total return on the investment and the result of these two design decisions is that the user has two different granularities of information on the same graph but placed in such a way that avoids bloating it. Compared to the original design, we extended it by showing more information and added animations when recalculating the curve to make the whole experience more visually appealing.

Edge Bundling / Accumulative champion race graph





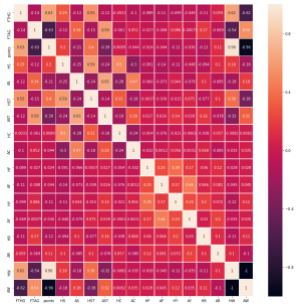
As it is obvious from the two images above, we ended up doing something that is visually completely different than the original idea. Originally, the idea was to use a circular edge bundling schema to show the different rivalries and typical results between two teams throughout the years. We would have a circle containing the teams of a championship on the edge, and different lines connecting teams that either have a tendency to win or lose against another team. This way we hoped to reveal patterns of certain rivalries or interesting other information that can relate to the historical rivalries of the teams. The problem was that when we started implementing it, we realized that bringing the data to the form necessary to visualize what we wanted was very hard and when we did it for just a single championship the overall result was underwhelming, so we concluded that the user would have a hard time understanding what we're showing him.

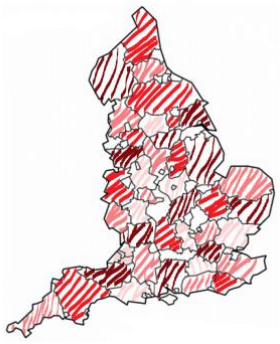
That's why we decided to implement a different visualization which maintains the same theme of presenting cumulative historic data and after performing extensive research on what we could do, we stumbled upon a video showing a race graph like the one above. We immediately concluded that it's a perfect match with what we aim to show and started implementing it. The idea is that we gradually build a graph with the cumulative sum of

points each team acquires through all the major European leagues across the last 7 years. This visualization is very interesting as it shows very clearly some events like a season where a team performed very well or when the exact opposite happened. It's true that it's not directly related to betting, but it looks very cool and we believe that every true football fan will enjoy it as much as we do.

As for challenges, the race graph was originally implemented with version 5 of D3 when the other visualizations used version 4, so when we merged all of them together a lot of problems were born out of this incompatibility. Therefore, we had to reimplement the race graph in version 4, an overhead that we could have avoided. Finally, it was very challenging to deal with the JavaScript's timer because it is totally different from other programming languages that we're used to.

Heatmap with metric correlation and distances map





These are the two visualizations that didn't make it to the final cut of our website, each one for different reasons. For the heatmap, the idea was that we would show how different metrics correlate with the outcome of a game and there were two main reasons why we didn't include it in the final version of our website. First, it's not as innovative and interesting as far as visualizations go: it's cluttered with too much information and we doubted that any user would spend the time necessary to take a look at it and understand it. Second and most importantly, we were afraid that such a visualization would encourage the creation of more sophisticated betting strategies thus going against our social mission. An argument could be made that this is the case for some of our previous visualizations too, like the barchart, but in fact those graphs highlight how unpredictable the outcome of a game can be, rather than offering practical betting tips.

Regarding the distances map, the original idea was to see how much each team travels in a season and how this affects its performance. The main reason it wasn't included in our website is the same one as why we included it as a bonus figure in our milestone 2 report in the first place: we couldn't find any data with this information. We spent days searching for datasets containing this kind of information, but our search wasn't fruitful. The only solution to make this visualization a reality would be to manually scrape data from an application like Google map which proved to be a very tedious process and consequently we abandoned this whole idea.

Conclusion

For our project we decided to visualize data concerning betting and football. Our team is consisted of avid football fans so making something relevant to this topic was only natural, but at the same time we wanted to give a strong social aspect to our project. We present a wide variety of visualizations, all showing a different aspect of our topic of choice. We put a lot of effort into implementing these graphs and even though some challenges arose (as it is natural in such a project) we believe that we overcame them successfully and managed to present a complete, interesting and entertaining website. We are very proud of it and hope that you'll find some value in it too. Thank you for giving us the chance to work on something like this.

Peer Assessment

George

- Wrote the process book
- Implemented the map visualization
- Cleaned and prepared the data for the map visualization and barchart
- Refined some details in the barchart

Gerald:

- Implemented the barchart, the accumulating chart with the investment strategies and enhanced the initial version of bubble chart
- Cleaned and prepared the data for the 3 visualizations mentioned above (using Python and Pandas)
- Helped creating the website, added the wording content and made sure that all visualizations were bug-free and worked together

Florian:

- Build the front-end of the website
- Implemented the core of the bubble chart
- Implemented the race graph