Soccer Data Visualization Design Journal

Group 16 (Tahj Anderson, Elena Le, and Matthew Rundle)

CPSC 4030

Github Repository

GitHub Page (Zoom out for ideal layout)

Video Presentation

Overview and Motivation

Soccer is one of the most popular sports in the world, bringing in millions of fans yearly to stadiums, sports bars, and tv screens. Although soccer is seemingly less intricate than other sports, there are countless intricacies within the game. With eleven people on a squad, the soccer pitch transforms into a grid, with players being the nodes on the coordinate system and completed passes becoming the links. Shots taken become a position-based statistic intertwined with play scenarios and player skill levels. Viewing the beautiful game through an analytic lens leads us to a pivotal question. In the game of soccer, can we find patterns in soccer match data?

Narrowing down the scope of finding patterns, we want to focus on analyzing shots. We will attempt to answer questions like: “where have players been most likely to score from?” or “what are the different patterns when a player passes the ball? ” The goal of our visualization is for users to explore different shot scenarios and discover emerging patterns. Overall, we want to address where and when the best soccer shots are created. We want to see what teams in each league have high win percentages in the total games they played in the season.

The FIFA World Cup in Qatar is currently going on, which happens every four years. This inspired our group to look into soccer data. Particularly with soccer, there are many combinations for a team to reach a goal. With the richness of data, we knew our visualization would be challenging and yield cool visualizations.

We chose to work with this dataset because our group enjoys watching a soccer game.

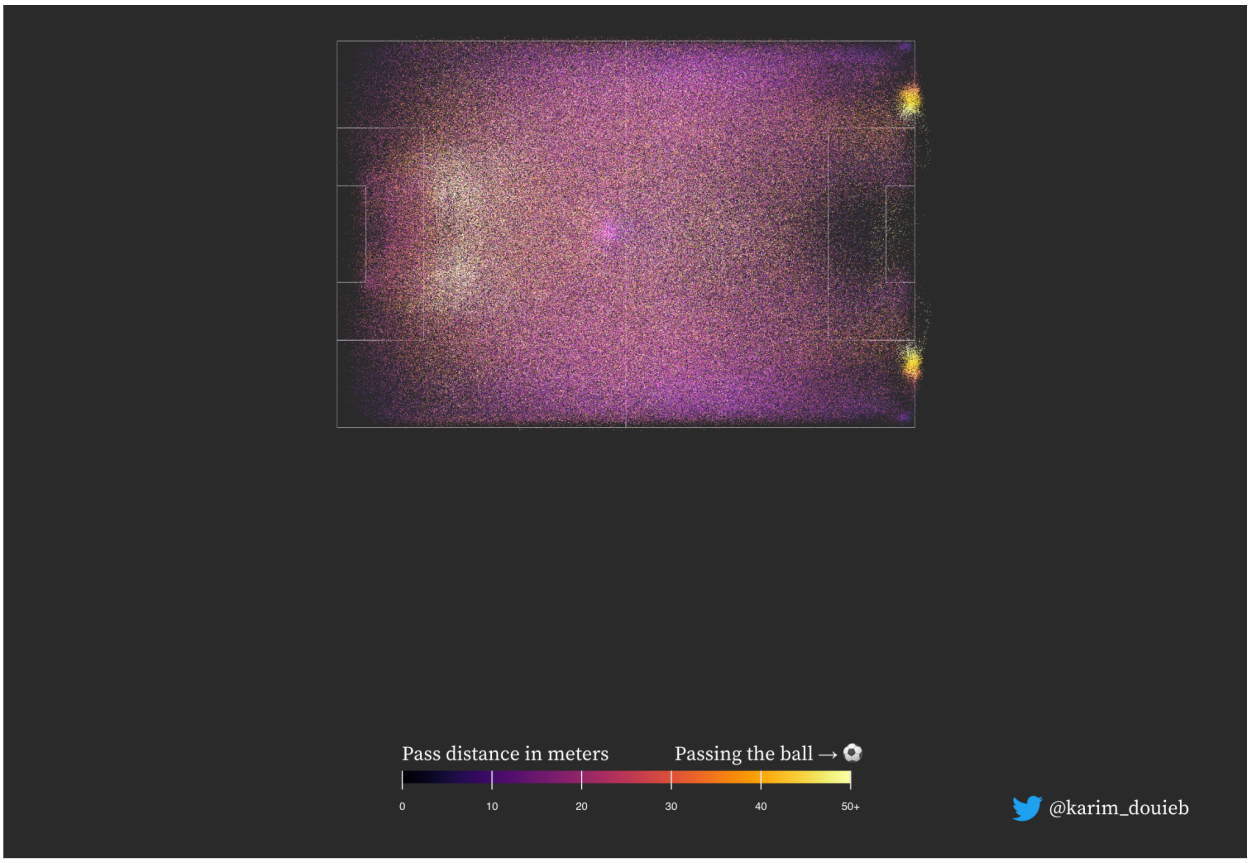
Matt is especially interested in the project because he is considering data analytics in sports as a possible future career. This, combined with the endless range of possibilities within the game of soccer, helped sway our group towards working with this dataset.

Elena also played soccer in high school. She played left-back. It is fascinating the techniques that players use to score. Soccer has many combinations to obtain a point or block the ball.

Tahj played soccer as a young child and again in middle school. He was a fanatic in middle and high school and would wake up early every Saturday morning to watch the EPL. He continues to watch and keep up with the popular European leagues and enjoys a game of FIFA now and then.

Related Work

In our research for visualizing soccer data, we came across a website that shared links to many different visualizations with different purposes. One visualization that inspired us was created by a guy named Karim Douïeb. He made a visual that shows 882,536 passes from 890 matches across various leagues and seasons. The visualization is stunning and inspired us to create our passing visuals since passing data is so abundant.



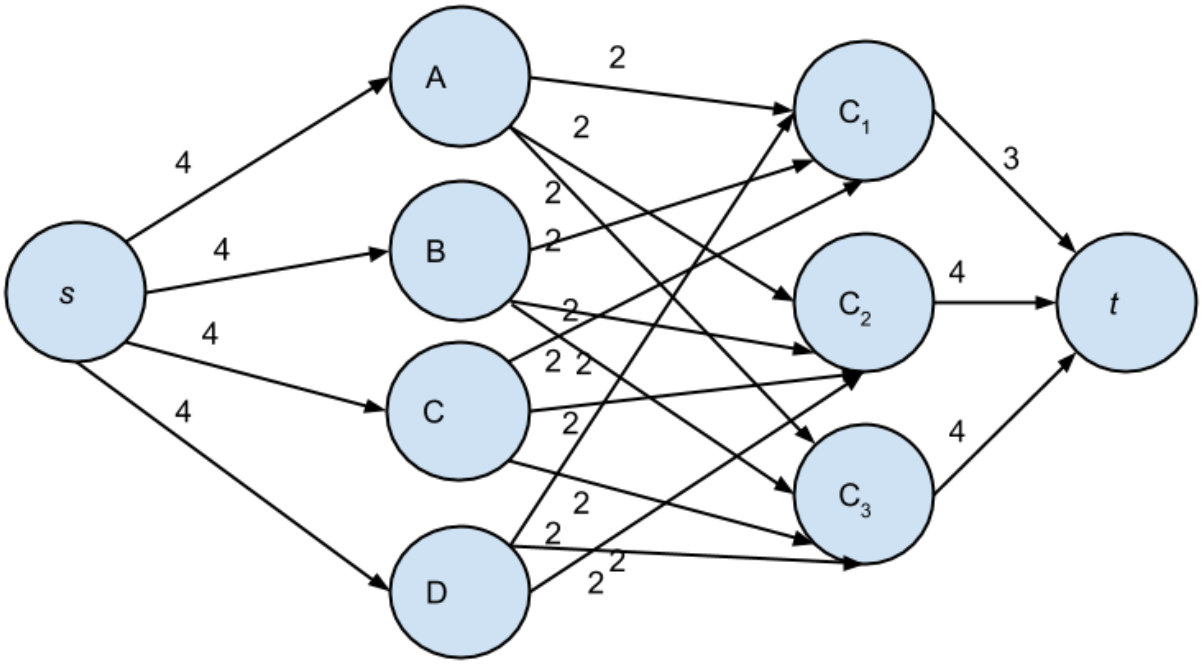
Source

Fig 1. Visualization of passes across many leagues and games.

The visual is animated and shows how passes permeate through the field. If you move your cursor over the field, a circle representing a user-controlled radius pulls all of the passes to that particular position and then disperses them to their intended location.

Karim's visualization inspired us to get excited about what we could do with soccer data. Our reaction was to create our passing network visualization. Since the game is built on passing, we knew that different teams had to have different passing behaviors. A team's success is built on

play in the defense, midfield, and offense. All three are connected yet separate parts of the field worthy of analysis.

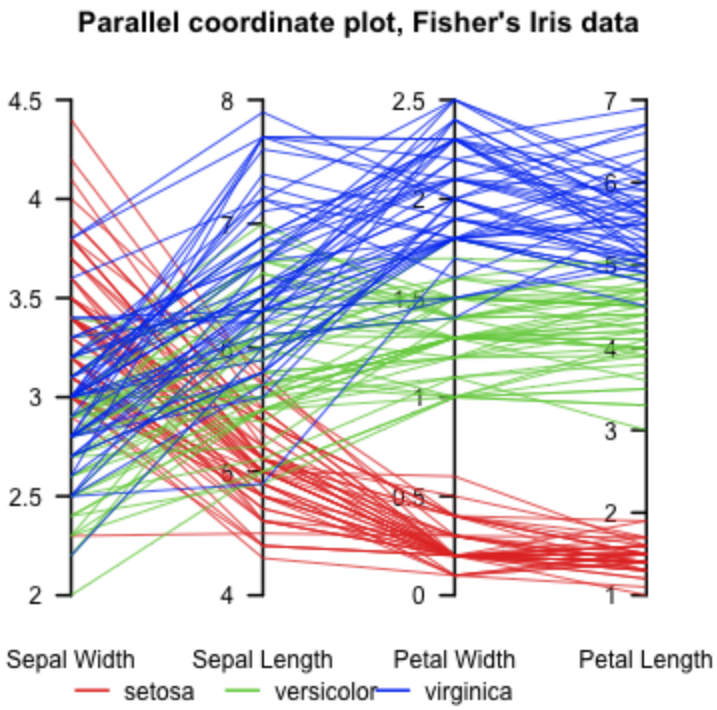


Source

Fig 2. Algorithmic graph

The structure of a graph visualization sparked our interest in creating a passing network visually. Each node above can be visualized as a player and each connection as a passing direction. In our minds' this was how we could imagine a soccer team and their passes.

However, we quickly recognized that soccer positions change more times than data can keep up. So we switched our approach from using nodes to map player positions since this was unrealistic and irrelevant to passing data. We gained inspiration from the lecture when Professor Iuirich mentioned parallel coordinate graphs.



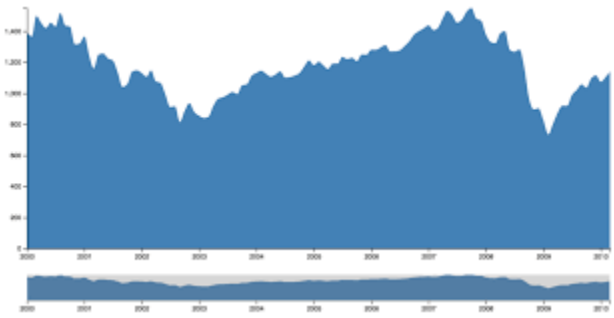
Source

Fig 3. Inspiration for passing networks.

Based on the four parallel lines in Fig 3, we decided to mimic this behavior to visualize the goalie, defense, midfield, and attacking lines. We hypothesized that the passes would connect the lines, and out of this will emerge patterns based on the space or overlap of passes. Also being able to isolate each parallel line helps us view specific positions and their interactions within and to other positions.

Fig 4. Playing With D3 Class Assignment

The bubble visualization was inspired by incorporating the class activities using forces with simple objects. The visualization shows a general overview of all the teams in the leagues. We were curious to see what teams had a higher win percentage than others.



Source

Fig 5. Inspiration for brushing

This example was seen to help follow through with and create the brushing effect in our passing network visual. Enabling the user to filter different passes based on the vertical position adds another layer of complexity that was not there when they could solely filter based on player position.

The lecture in class on heat maps and choropleth really strengthened the idea that we could visualize where players are more likely to score via color. This seemed like it could create a very pleasing visual that is easily interpretable. Since shot data is highly dependent on spatial position, a choropleth can pull needed context into understanding why some shots are more successful than others.

Questions

Original Questions:

1. What sort of differences do we see between soccer leagues (in Europe)?
2. What passing patterns are used most frequently?
3. Where are soccer players most likely to score from?

These original questions were somewhat general, so we needed to find a way to narrow in on more answerable questions. When looking for differences between soccer leagues (1), one key aspect they all have in common is a winning distribution. We then wanted to analyze how these winning distributions show within and across leagues. The next question (2) was intended to portray an essential aspect of the gameplay. This was another vague question, but when aggregating all of the passing data together in our visualization, it wasn't easy to discern patterns because of the sheer amount of passes. This led us to consider if we could compare teams because that would allow us to hone in on smaller amounts of data where patterns could emerge. Our last question (3) was to display another quintessential part of soccer. There is an important distinction between where players are more likely to score from and more likely to shoot from. Because players score at a higher percentage at a certain spot on the field, that does not mean most goals are scored because fewer shots could be taken there. This led us to want to reveal that quality, so we wanted to compare the percentage of goals at different parts of the field (goals/shots) with the total number of shots.

Final Questions:

1. How are team win percentages distributed within and across the top 5 European Leagues?
2. How do different teams' passing networks compare to one another?
3. Across all leagues, where are players more likely to shoot and score form?

These questions overall help answer the patterns within soccer. The questions transformed into a readable and interactive visualization that will help display the data in a way that allows the user to improve their understanding of soccer strategy. Our original questions were broader questions related to soccer, but then we focused on the significant parts of the dataset that will help visualize them better.

Data

We found a Kaggle dataset containing multiple csv files structured like a relational database. You can check out the source here.

Based on the research paper that inspired our work, we used their dataset to create visualizations the Kaggle data could not. You can check out the source here.

Both datasets contain similar data, allowing us to augment our dataset and patch any holes one dataset has. This gives us a more significant degree of freedom to create our visuals. The unfortunate trade-off is that the complexity of creating our datasets has increased.

To create the dataset, there needed to be some data cleaning done. We used a Jupyter notebook for data processing. Our data engineering tasks were completed inside the notebook using Pandas and Numpy, open-source libraries for Python. This provides high-performance data manipulation. We can reshape our dataset and allow data filtration.

The win percentages still needed to be included in the dataset. The calculation is expressed by the number of wins of each team divided by the total number of games each team played.

To visualize the passing networks for each team, we had to break our data up into ways that made our visuals easier to create and faster to load data. The data came in five separate json files, each representing data from the top 5 European leagues. Each json file had various columns and data under them based on events in games. For the passing networks, we mainly focused on passes and their starting and ending x and y positions. A few issues came with the data set.

1. The dataset did not have a team attached to an action.
    a. This caused us problems in being able to visualize individual team passing networks.
2. The dataset contained nested data, which made data unreachable by D3.
    a. JS is not dictionary-friendly and hinders us from quickly visualizing data, specifically the action start and end coordinates.
3. Each event within the dataset did not have the player or position attached.
    a. Since this was not included, we could not draw lines from the position they needed to be in.

With these issues, the dataset needed to be worked so D3 could easily reach data. In the Soccer Match Event Dataset, there were tables that had specific information and IDs to cross reference values from table to table. To add players to each action, we performed a data join on each event's associated player ID with the player's data set to attach each player and their position to each action. Next, we broke out the coordinates by looping through each row, breaking out

coordinates, and assigning them to their own respective columns. Finally, we added each team to an associated action by performing a merge on team ID using the events dataset and the team's dataset (hosts every team and their unique identifier). Although this section is quite lengthy, this decision made our design possible. Without engineering a more reliable data set, our visualization would have been impossible, and we would have had to undertake a different endeavor.

Our dataset needed cleaning to rid our data set of NaN values. D3 did not handle NaN values well and mistook them for points, causing our visualization to become faulty. We turned every NaN value into something D3 could filter to combat this.

After we created a dataset that had what we needed, the overall size of it was too big. So we decided to break every team into its own csv file. This created 95 different csv files with each team's events for the 2017-2018 season.

Within these csv files, actions went from event to event and we needed to use .filter() to only look at the desired events. Because one event was connected in some way to the next, we needed to access the actual index each event was at (because using .filter() re-indexes when using (d, i)). This meant we needed to add an index column to each of the 95 csv files.

Our visualization works because D3 does not have to parse through large amounts of data. Overall our data engineering and cleaning allowed our visualizations to come to life.

Exploratory Data Analysis

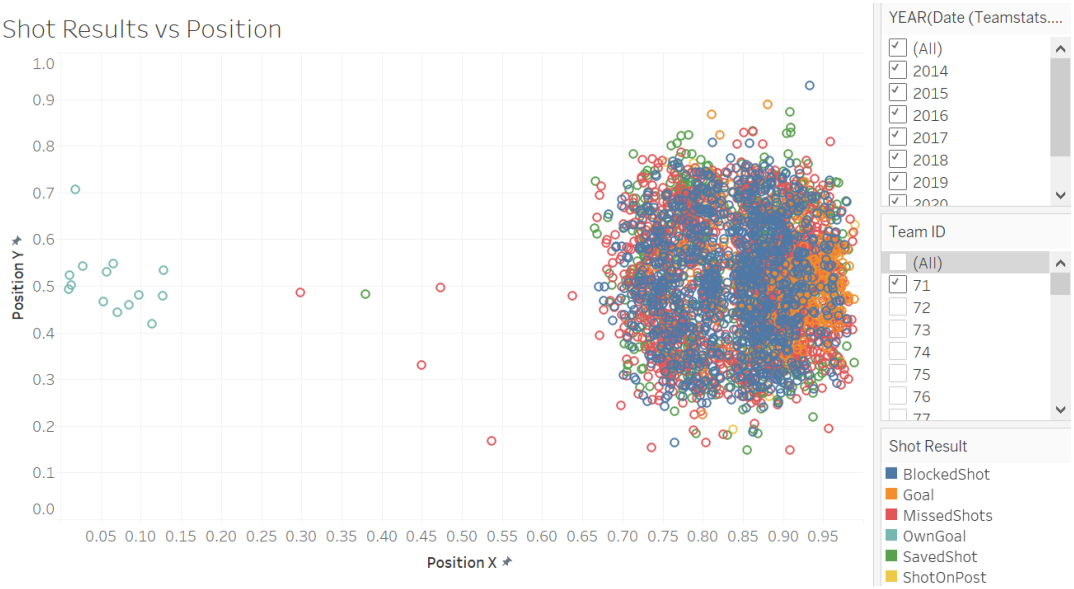


Fig 6. Tableau Sketch

When we were initially exploring the data, we used tableau on the shots data to get a look at how all of the different shots were distributed. When doing this, it was obvious that we were going to be working with an incredible amount of data and needed some sort of way to aggregate it. This facilitated the idea of using a choropleth to ensure that all of the shots did not merge into one blob of uninterpretable data.

Another aspect dealing with the shots data (when visualized in tableau) is that there was an apparent difference in where shots were taken versus where shots were scored. This helped reinforce the idea that we wanted to create a visualization to highlight that difference.

For passing networks and win percentages, before we could explore the data, we had to produce a visualization for a project milestone. Plus, with the added cleaning and engineering operations needed, we did not take the time to explore the data in another tool and instead chose to create our visualization to explore the data.

Design Evolution

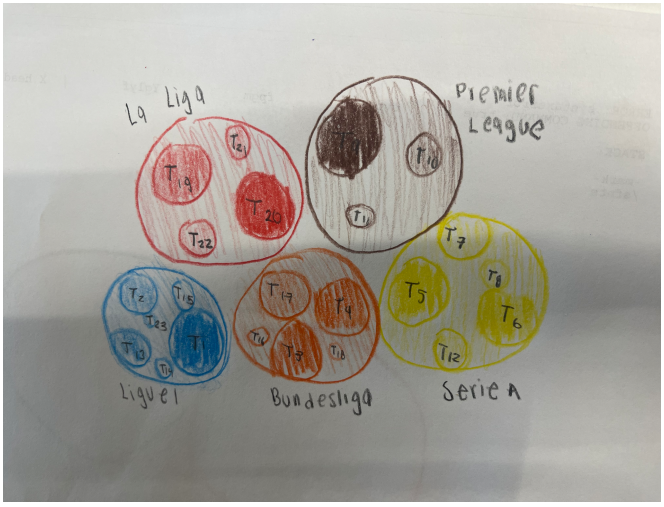


Fig 7.1. Bubble Distribution Draft

From going through a written design and actually implementing it, there were a few changes that were made. With the bubble distribution visualization, the first design was to have the bubbles made of teams under one big bubble based on the league. Originally we were going to have a gradient based on the win percentage with the win percentage being calculated by the total win of each team divided by the total games in all the leagues.

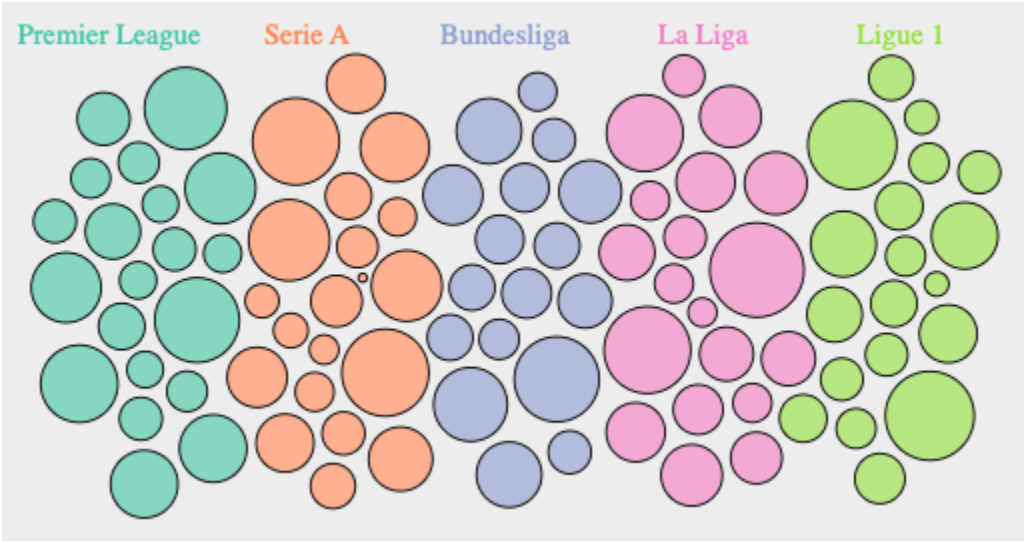


Fig 7.2. Bubble Distribution 2nd Draft

The bubbles are now the win percentage based on size and the calculation is based on the number of wins of each team divided by the total number of games each team played.

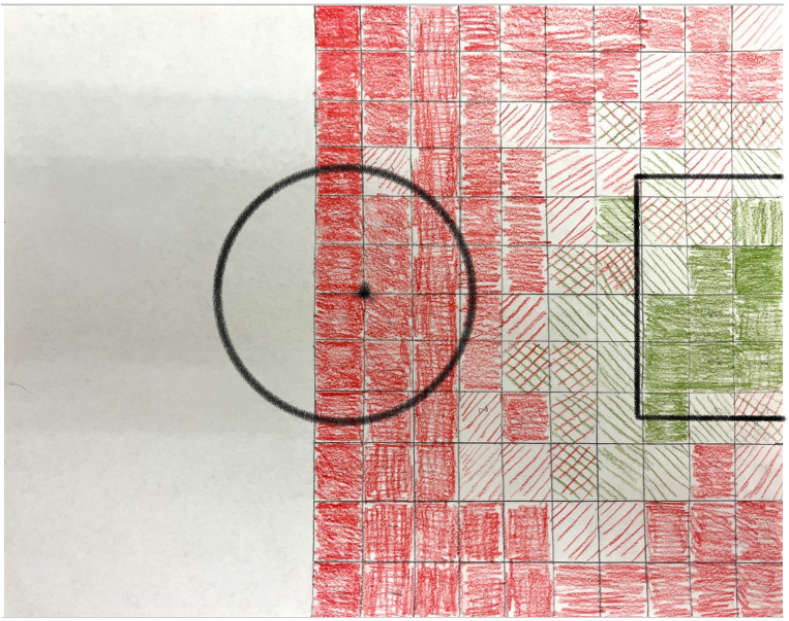


Fig 8.1 Choropleth Draft

Fig 8.1 was the initial design for the choropleth of percentage of shots made in different regions of the field. Regions that players score less from were colored a darker shade of red and regions that players score more from were colored a darker shade of green. The diverging color scheme really helps to highlight the difference between various regions of the field.

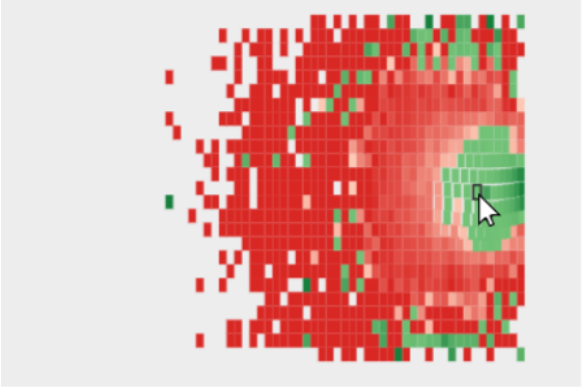


Fig 8.2. Choropleth FishEye Implementation

One of the ideas that was tested was implementing a fisheye to distort the choropleth, making rectangles around the cursor appear larger. Unfortunately it made it difficult to distinguish boxes around the cursor and did not really help side-by-side comparisons that were desired. It also ruined the alignment. Instead, the penalty box button described later seemed a better option for finer comparisons.

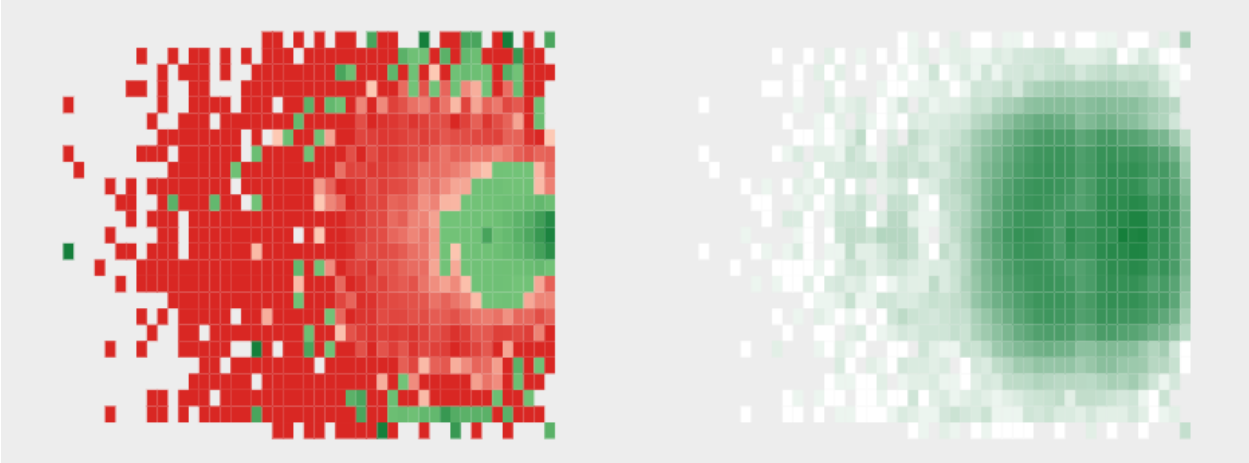


Fig 8.3. Final Choropleth Visualization

The result (Fig 8.3) from this design turned out the most similar out of all of our approaches. The main difference is that we added a second choropleth which is based off of total shots made. This way we can highlight not only where players score at different percentages from, but also where players score versus shoot from. Having the side by side comparison greatly strengthens the amount of information a user can take from our visuals.

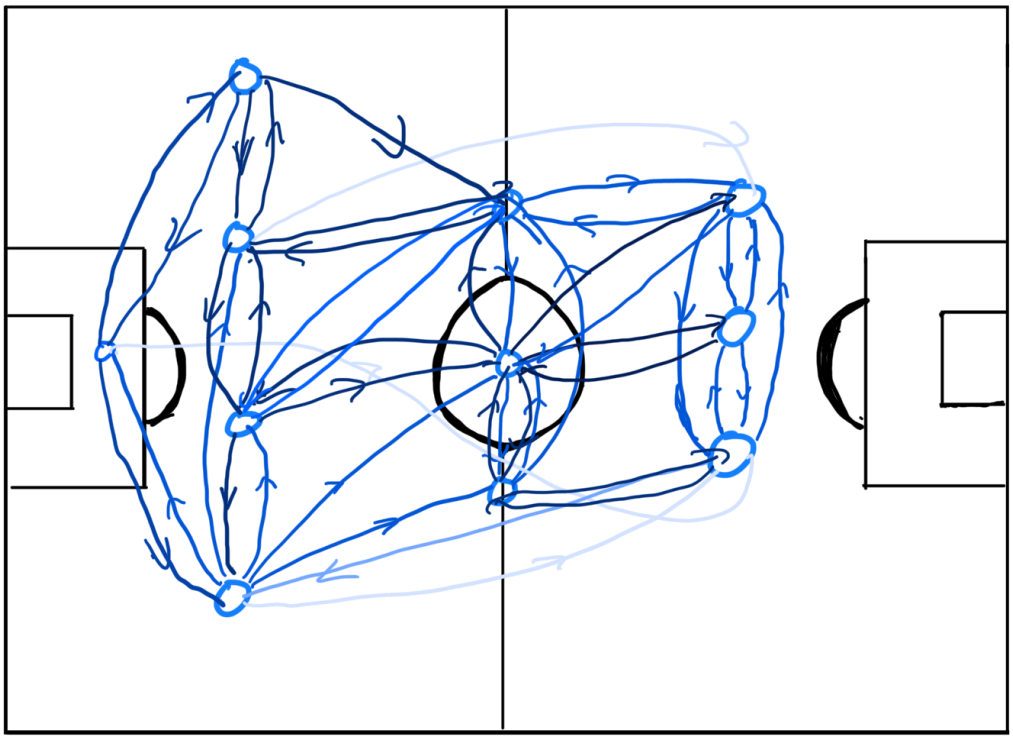


Fig 9.1. Passing Networks Draft

The above figure was the first design for the passing networks. As stated earlier, the idea was to have nodes and passing going from node to node. However, this design step was not feasible and required data that had formations in every game. Instead we focused on cutting out the nodes and separating each position line to its own parallel line. There we would draw passes based on their starting and ending coordinate and color code them based on the players position.

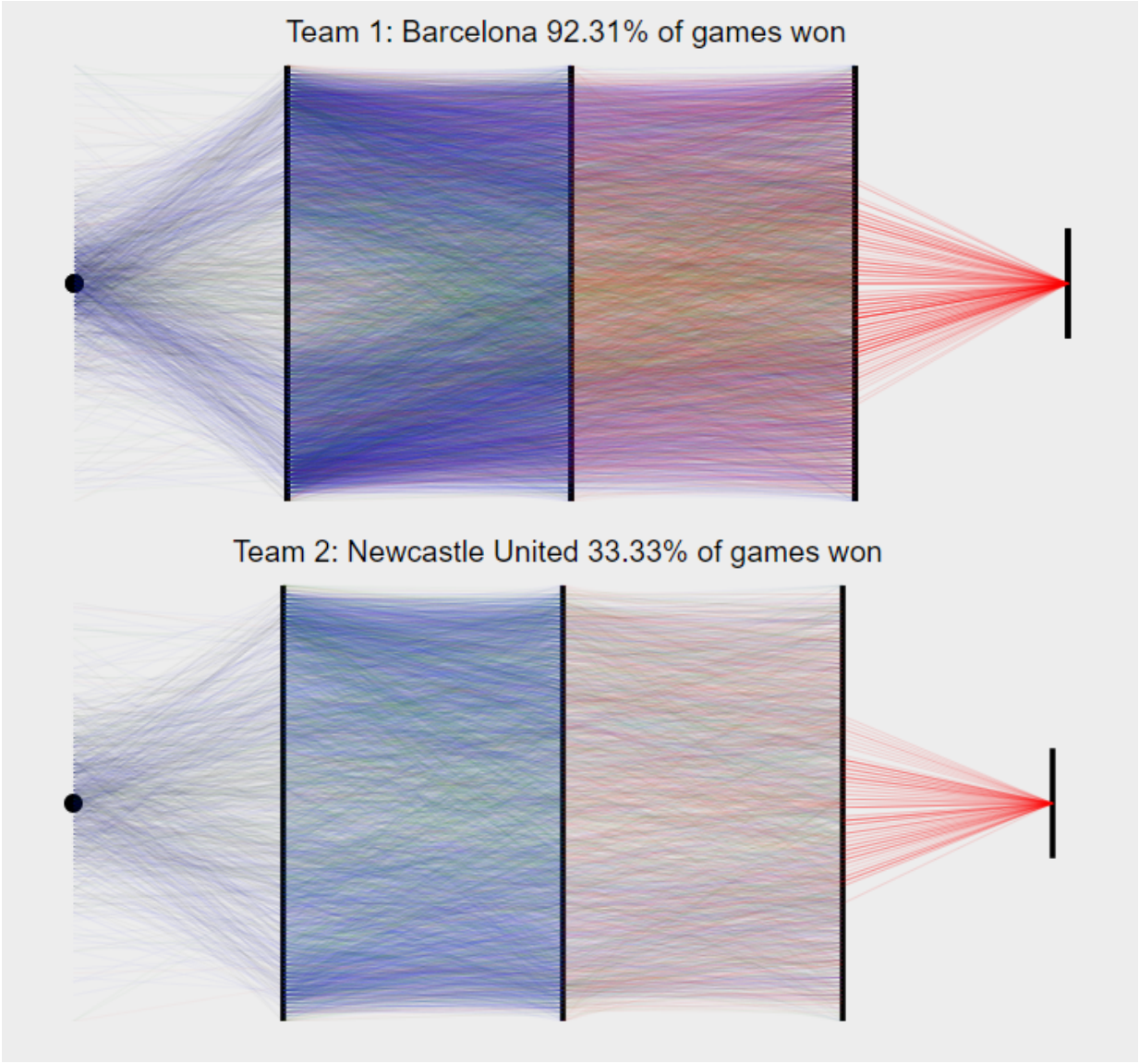


Fig 9.2. Pass Network Second Draft

Based on our first actual code design, we went without traditional soccer boundaries and created visuals consistent to the parallel coordinate plot. This visualization got our point across and showed patterns in passing. However, one weakness was that it did not reflect a soccer field, thus it was confusing to viewers. We did not anchor passes going to the goalie causing a weird square

cut off. Also, the visual was not appealing and looked too abstract. So to ground our visual we added traditional soccer borders to increase readability and make our visual more realistic.

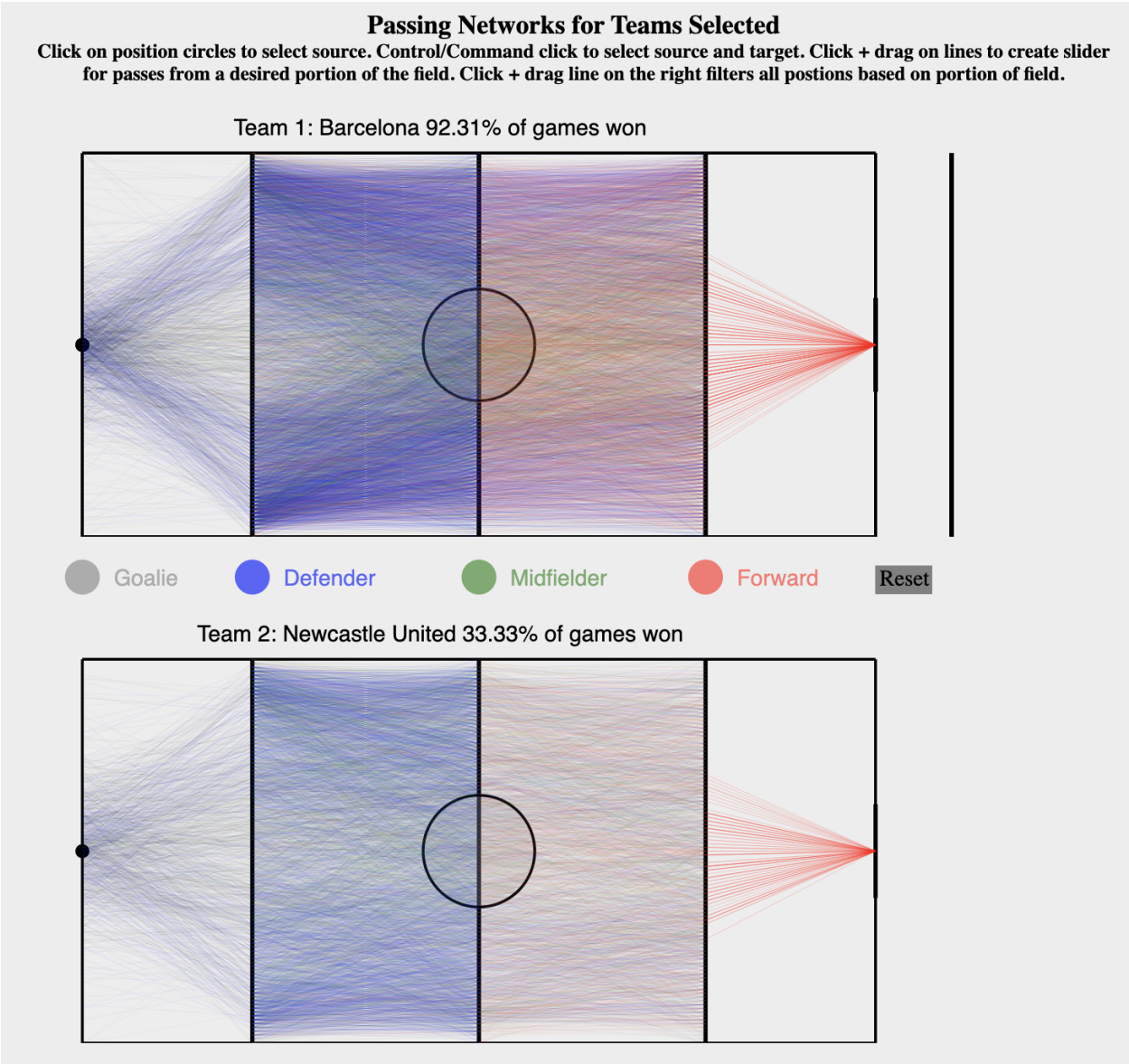


Fig 9.3. Final passing networks visualization

The final visualization contains the traditional football borders and markers, excluding a penalty box. We felt that including a penalty did not change the behavior of the passing and found it unnecessary to include. Adding the soccer field borders increases the readability of the passing networks and makes the visual more realistic. Also passing is constrained in borders, removing the abstract nature of the previous figure. Overall, we were more satisfied with this look than others. Clickable buttons and a sliding bar increase the implementation which will be discussed in the next section.

Implementation

The goal of the bubble distribution visualization is to analyze the number of games won by each team based on their league. This is to help show which leagues may be more competitive based on the winning percentages. Translating the data into a visual context is beneficial for users to create a general understanding of the best teams in their respective leagues.

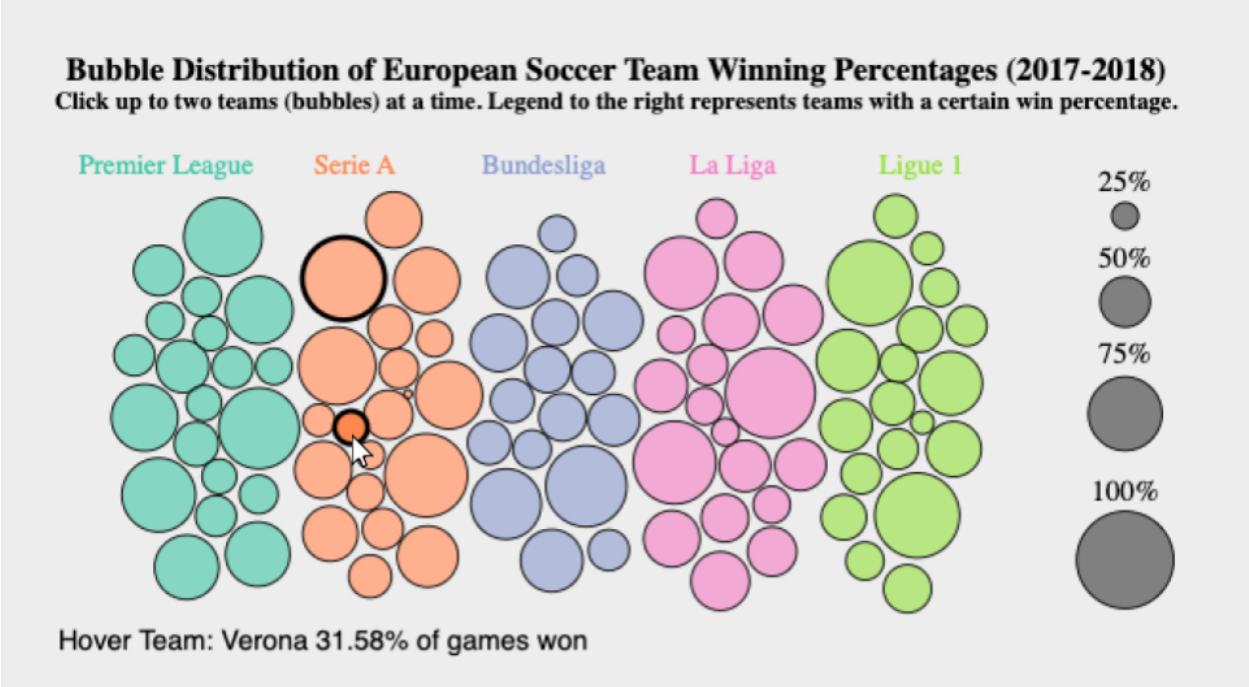


Fig 10. Final Bubble Distribution

Once a user hovers over the data, they can see the teams with their win percentages in the league. The size of the team bubble is based on the amount of games won in the league. There is also a legend on the right of the visualization to show the general comparative size of the bubbles. This visualization ties in with the passing network visualization. A user is allowed to click on two teams to compare the passing network data.

The goal of the choropleth visualization is to answer where players have the highest percentage of goals based on their location and where players shoot from the

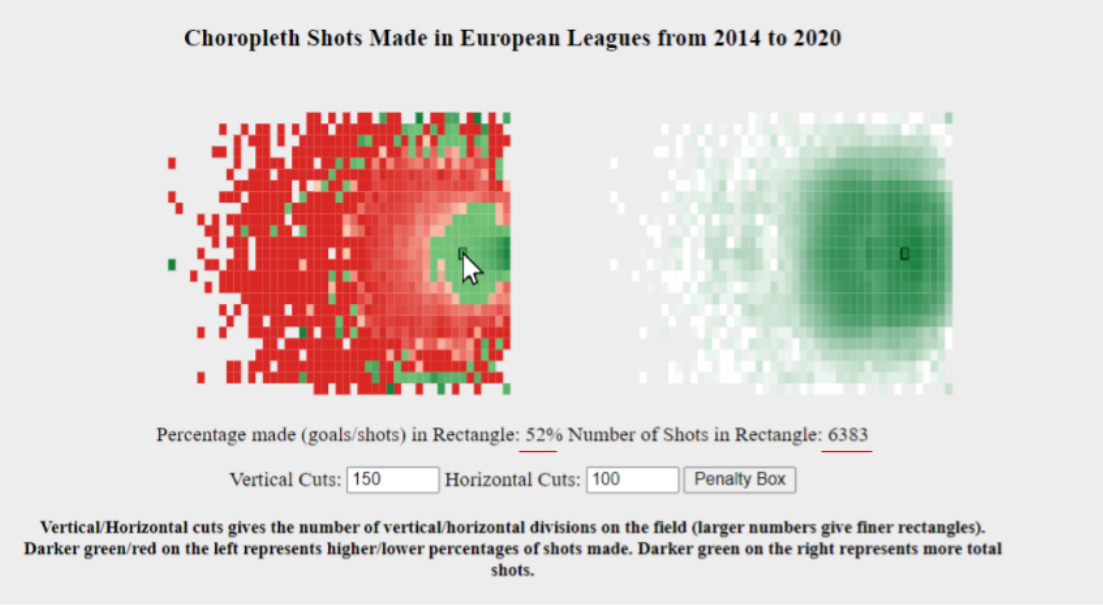


Fig 11.1. Choropleth Number of Shots

To interact with this visualization, the user is allowed to hover over different rectangles with their cursor. When this happens, as shown above, the text will show the percentage of goals/shots and total number of shots scored within that rectangle.

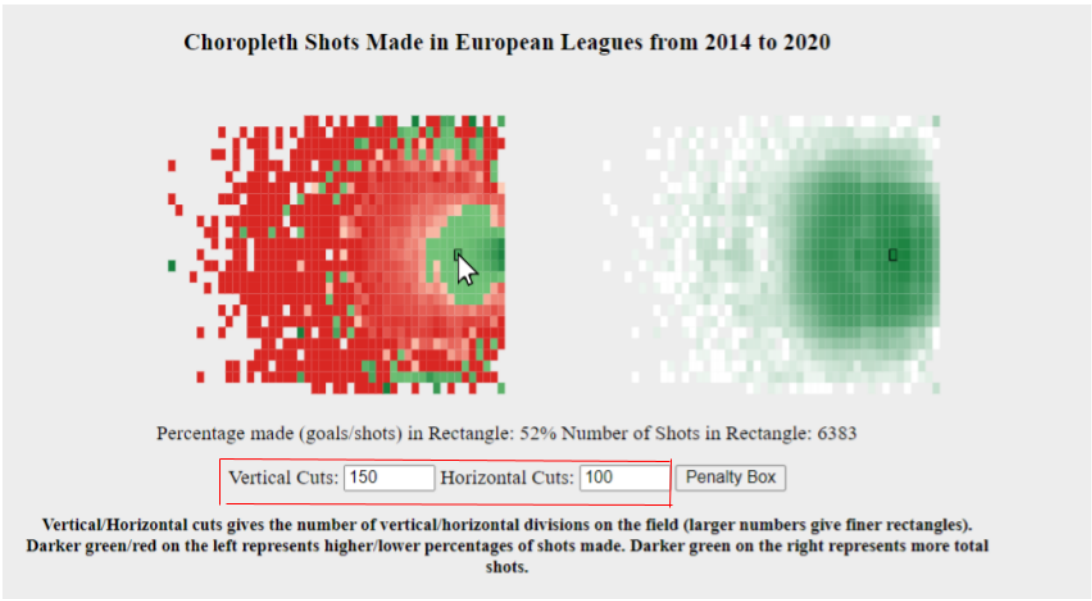


Fig 11.2 Choropleth Cuts on Field

The user can also change the number of vertical and horizontal cuts on the field. Increasing these numbers will make the rectangles smaller which can reveal more subtle changes on the field. You

can also make one of the two cuts incredibly small, like 1, and analyze the data by just horizontal or just vertical position on the field.

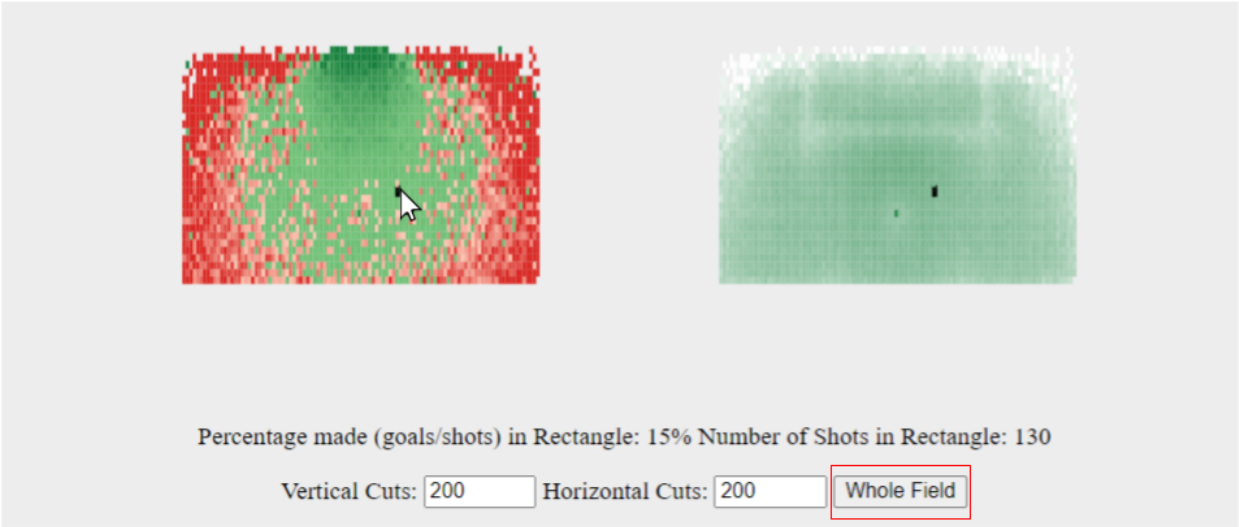


Fig 11.3. Choropleth Shot - Whole Field

The final interaction is that the user can click on the penalty box button and the visualization will zoom in and rotate to view the same data but only rectangles that are within the penalty box. This is where most of the interesting data (relating to shots) lies which is why it felt important to highlight it. The user can then click on the same button (which now says “Whole Field”) to revert back to the view of the entire field.

The goal of the passing network visualization is to analyze the pass patterns of teams. We theorize that different teams will have different patterns due to their unique play style. To explore this we created interactions for the visualization.

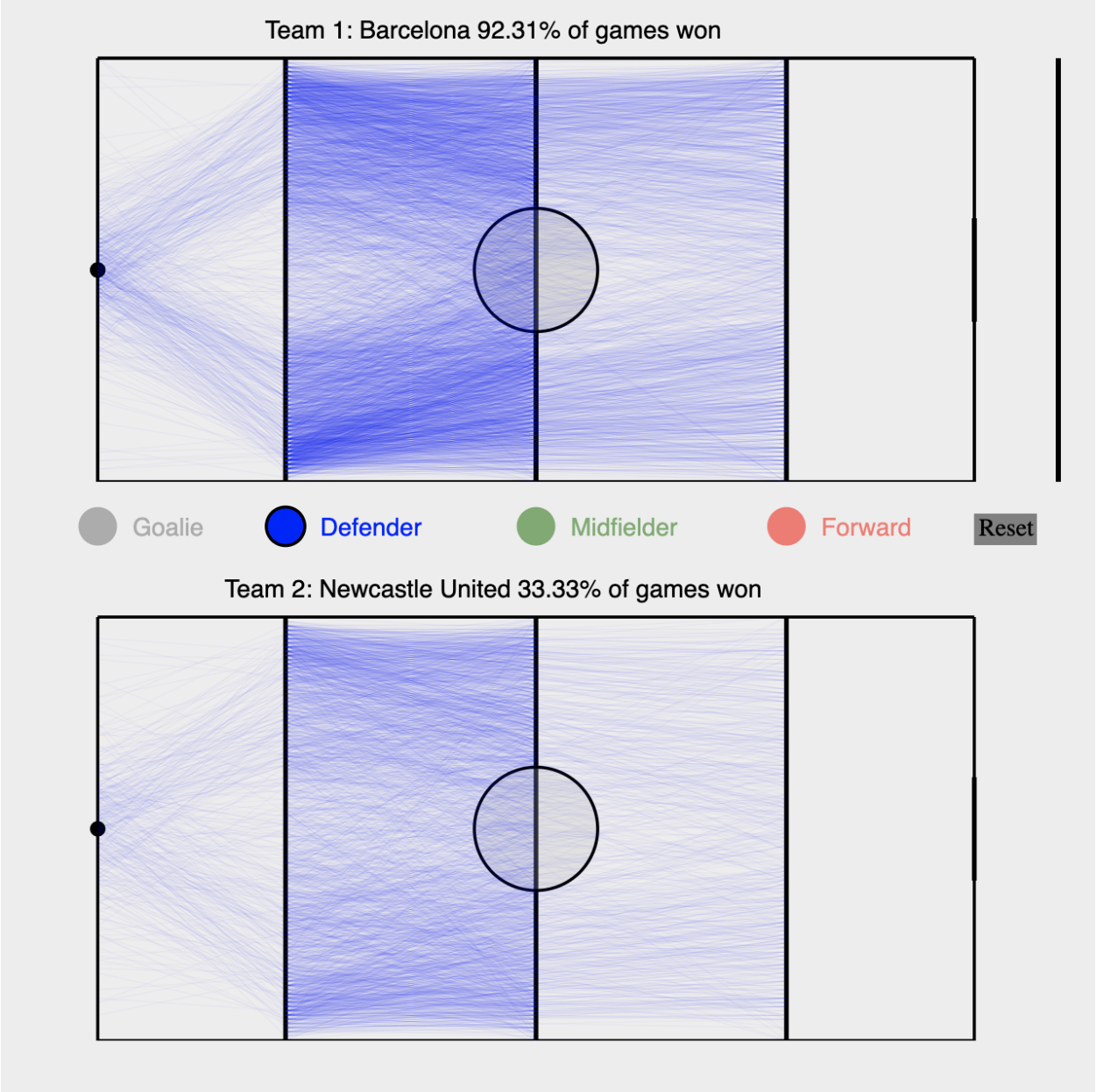


Fig 12.1. Passing Network Filter Passes

Users can visualize position specific passes by clicking on the circles labeled with positions. The above figure allows users to key in on defensive passes and explore passing habits for defenders. We can also compare the two teams' defensive passes with one click.

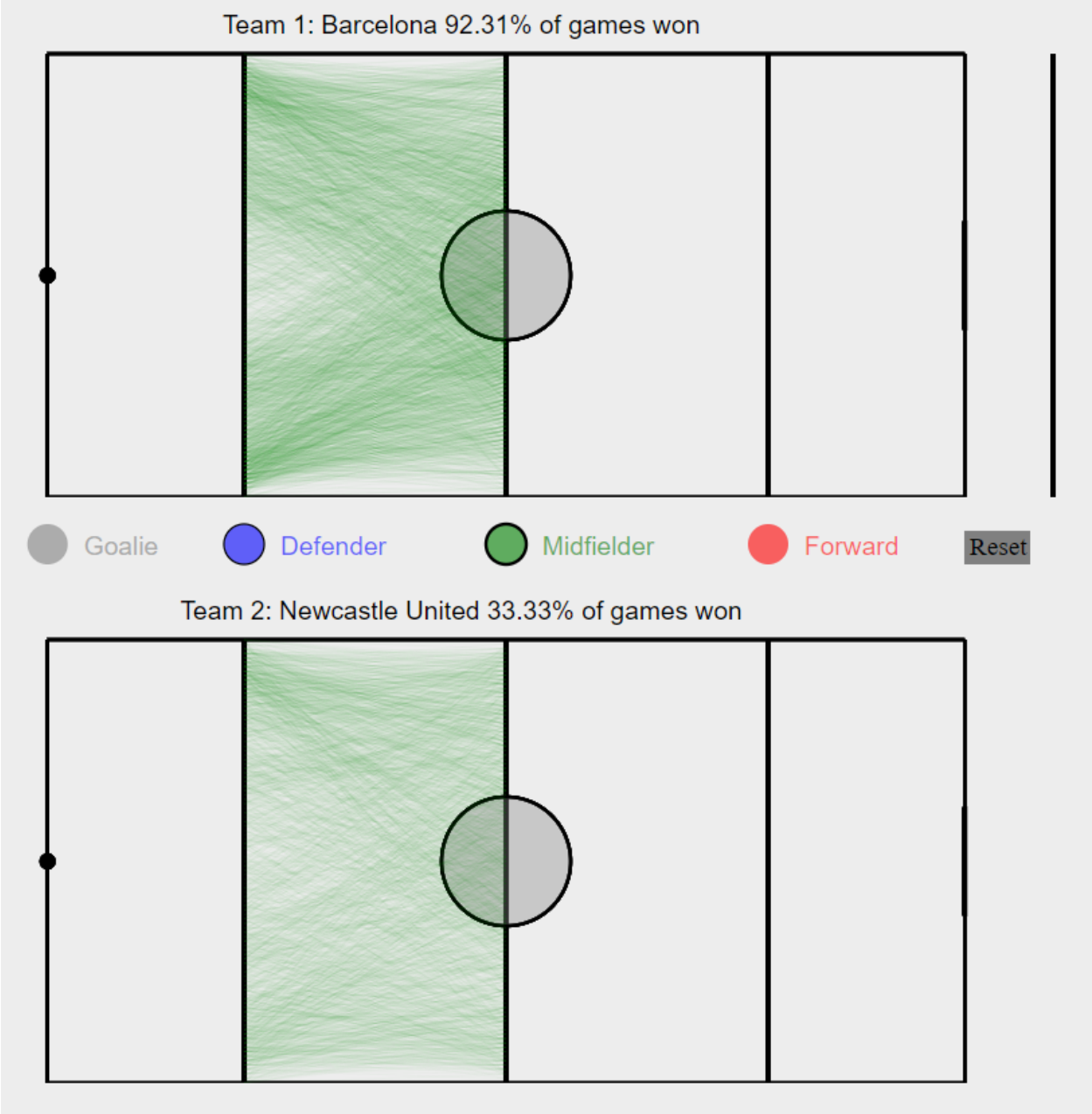


Fig 12.2 Passing Network Filter Target

Users can also control + click to select a source and target for a pass. The picture above shows what happens when the user first clicks the midfielder circle (holding control) then clicks the defender. The source and target then have borders for which positions are represented.

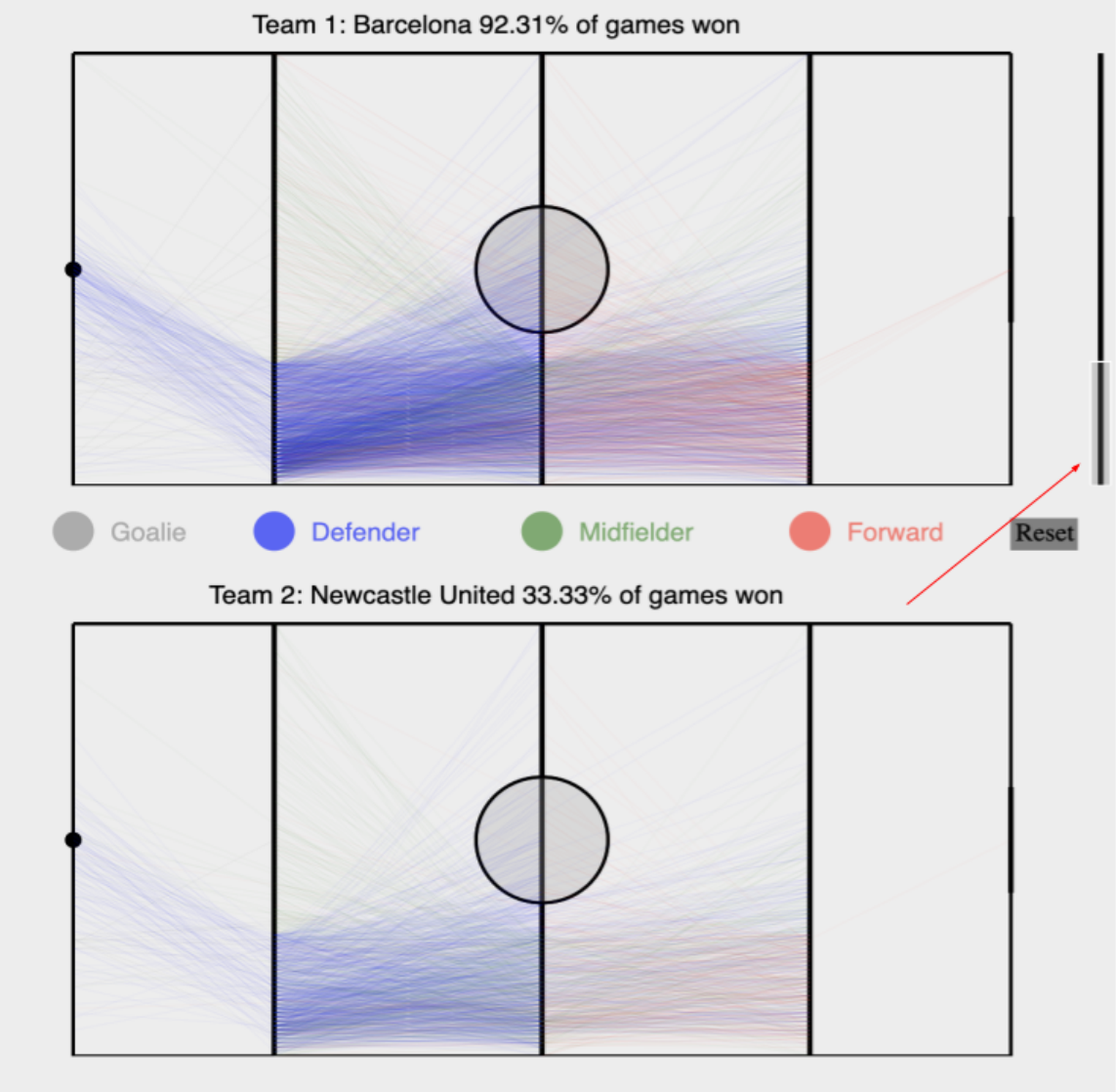


Fig 12.3. Passing Network Slider

Users can create and use a slider to section off the vertical portion of the field where certain passes are from. This allows the user to interact with the passing networks, exploring where the behavior of each team from different sections of the field.

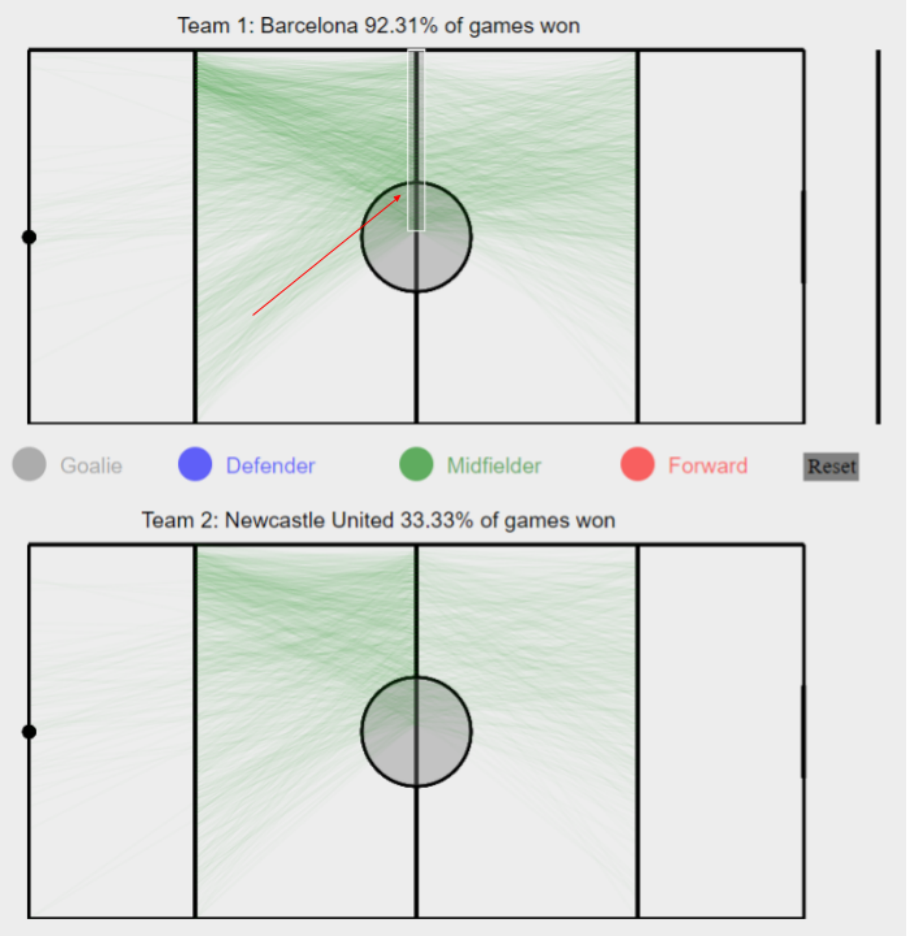


Fig 12.4. Passing Network Slider Based on Position

Moreover, the user can create sliders on any of the goalie, defender, midfielder, or forward lines. This will section off which portion of the field the passes stem from with respect to certain positions too and allows for greater data exploration.

Finally, the reset button resets both networks and should be used once the control + click has been done or if you want to get rid of old sliders.

Evaluation

In the bubble distribution visualization, it is clear that in every league, there are distinct 3-4 teams with high win percentages, and the others are roughly the same amount. An aspect of each league that is subtly hinted at in the win distribution is relegation. The lowest three teams of every season drop-down, while the top three teams from the bottom league are ushered in. Due to this rule, teams are incentivized to maximize their performance. In other sports, the worst team performance equals more star players from a draft. In soccer, there are no drafts, so teams have no incentive to "tank". We believe that the win percentages are fairly close due to the competitive nature of the teams wanting to stay in the top leagues. Bubble distribution allows users to compare teams that have or will not play against each other. We could improve this visualization by adding team logos to the center of each bubble. This will make the search process much easier on our users. We could also add a search feature so teams are easily found. More information on teams via user click would offer more context to win percentages. For example, if we could implement a "team profile card" that detailed information about the team's salary cap, stadium size, fan base, the average age of players, etc., then users can draw more conclusions as to why some teams have the edge over others.

In the passing networks, it is apparent right away that the teams with higher winning percentages have a more dense passing network. This is likely due to better team communication leading to more completed passes and the fact that they will have made the playoffs and thus played a couple of extra games. When exploring the data, you can see differences between teams and how some will tend to pass to different field sections more often than others. That can be attributed to a star player, or a play style that the team believes will work. Some times find success utilizing their wingers. Other teams find more success using centralized play. Other times it depends on the opponent. What our visualization shows is passing is not random. Teams utilize their unique passing habits to succeed in offensive and defensive situations. This visualization could be improved by creating moving passing networks showing progressions. That way, passing is not static, and users can explore how teams pass in the game. Next, we could create visuals that show individual games. Viewing games will show what teams have an in-game strategy for handling the opponent.

In the choropleth, it is evident that as players are closer to the goal, they are more likely to score. The total shots get smaller as players get very close to the goal but are densely populated around where players shoot penalties. There are more shots (and a higher percentage of shots made) from the left (top) corner than the right (bottom) corner, which is interesting. Overall, the shots data appears relatively symmetrical, but when toning in on the penalty box, there is a slight skew towards a higher percentage of shots scored being on the right side. The visualization answers the question about where players shoot and score by comparing shots made versus shots side-by-side. It does well to show what was wanted as it represents all the data on a 2D field that could not otherwise be seen by looking at raw data. The way the data was aggregated also

allowed for an understanding of what is going on at different spots on the field. The visualization works well overall, but the color scheme for the shots made could be improved. It was difficult to come up with a scale that could cover such a wide range of total shots made. Moreover, it would have been nice to have the choropleths line up with the teams selected from the bubble distribution. Unfortunately, the Kaggle data did not have a connection between who took the shot and what team they were on. We thought about trying to use the other dataset, which also contained shot data, but this one did not tell whether or not the shot went in. This is why the choropleth remained relatively unconnected from the rest of the visual system. It still, however, provides excellent insight into what is going on within a soccer field.