Sports Analytics & Visualization

Go Soccer!

Hongju Lee (hongjlee@umich.edu)

Jungseo Lee (jungseo@umich.edu)

Yuanfan You (ivanyou@umich.edu)

Quishi Zhao (qiushiz@umich.edu)

Table of Contents:

Introduction:

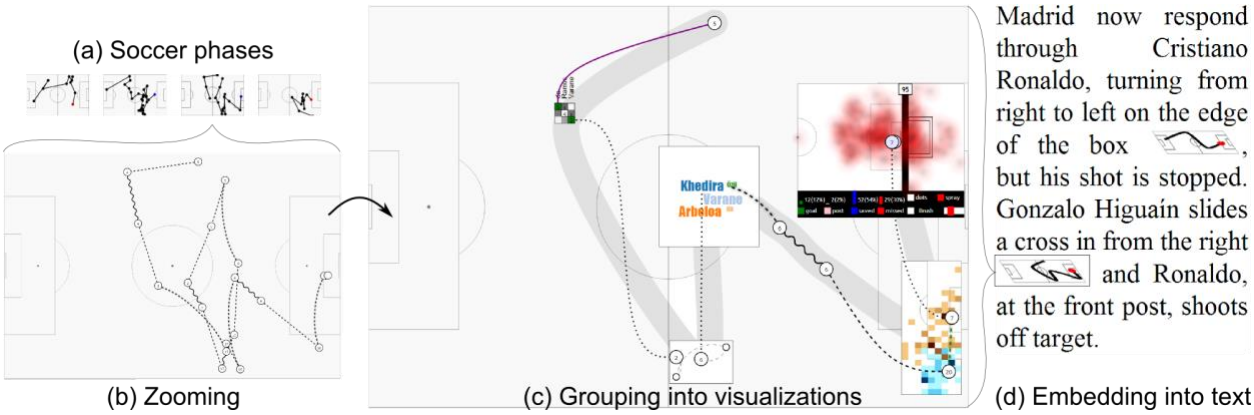
Sports analytics is not only a way of helping teams come up with winning strategies but also helping sports fans more engaged in the game. This applies data analysis techniques to the field of sports and analyzes various components of the sports industry including team and player performance. The introduction of camera-tracking technology and data visualization opened up a new era of sports analytics by helping sports fans and sports enthusiasts discover and compare the facts more easily. However, because it is still in an early stage, not many sports visualizations are commonly used. Especially, not like basketball, which is the 2nd favorite team sport in the United States. There are not many visualizations available of soccer, which is the number one favorite team sport in the world. The most commonly used visualization is a heatmap of the X-Y location of the players. However, when we think of the fact that it can't give comprehensive information about the player's performance and situations during the match, it was not hard for us to come up with a project idea to visualize soccer data.

Our project was focusing on visualizing sports data in the field of soccer to help soccer enthusiasts get a more comprehensive understanding of the team's and the players' performance during the match. We planned to analyze players' actions and performance during the game, evaluate how it has impacted the outcome as well as identify the players' overall strengths and weaknesses. By creating an interactive visualization that shows the overall team's movement, analyzes each player's skills that they have shown in a particular game such as shooting, passing, and so on, we hope to reduce the lack of detail and clarity existing in many of the visualizations and find ways to effectively analyze a player's performance. Furthermore, we may add dropdowns or annotations to see the detailed performance analysis of the player. The final comprehensive evaluation will be an analysis of each player's performance and its impact on the whole team.

Background Information and Why the Problem Matters:

Visualization based on key event:

Passing in football is the most frequent interaction between players and plays an important role in creating scoring opportunities. Therefore, much visual analytics work focuses on the analysis of passing events in football matches. SoccerStories is the first comprehensive and complete visual analysis system for football event data. As shown in Figure 1, SoccerStories takes a variety of visual forms to describe football games. Among them, the main view of the system adopts the visualization method of Focus+Context, which draws different local passing lines to the corresponding positions on the field, and draws the aggregated results of similar passing sequences on the field. Data analysts can understand common passing patterns in the game by aggregating the results, and explore the details of interesting passing lines through local specific passing lines. At the same time, SoccerStories also provides the display of statistical data and the automatic generation of text descriptions.

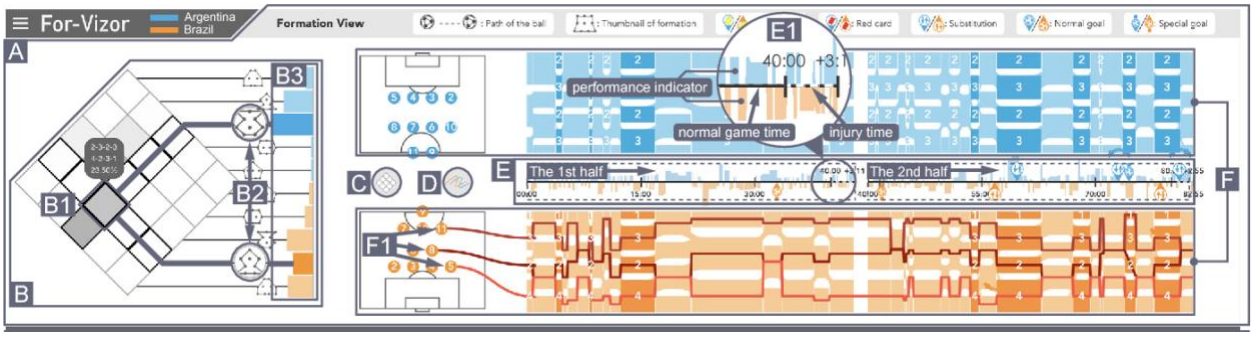


[Figure 1]

Analysis based on team formation:

The formation in a football game can reflect the tactics adopted by the team during the game. During the game, the team's formation changes with time and contains inherent spatial information. This spatiotemporal nature of football formations and other properties of football data such as multivariate features make football formation analysis a challenging problem. As shown in Figure 2, a novel visualization form of formation changes is designed in the visual analysis system ForVizor. The system processes the player's trajectory data through the formation

detection algorithm, and outputs the formation of each team corresponding to different moments in the game. The visual interface of the system includes formation view and display view. In the formation view, football analysts can understand the formations commonly used by the two teams in the game through the visualization of the matrix, and intuitively track the changes of formations and players in the whole game through the new visual design formation flow. Movements in the formation, and compare the formation change patterns of the two teams and observe the relationship between the formation and the situation of the game. Analysts can further obtain detailed team formation contextual information (such as the actual position of players on the pitch) in the presentation view, and perform detailed analysis with useful statistical indicators.



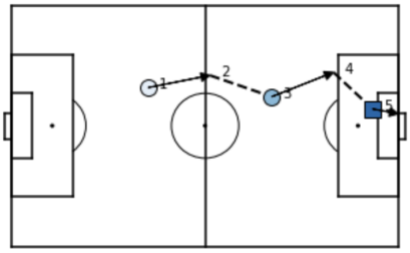
[Figure 2]

These visualizations all perform some kind of analysis of soccer games from a certain perspective such as passing events or change of tactic formations. However, sometimes the users need to understand the full image of the game in order to perform a better analysis on the performance of players or the whole team. For example, if we want to analyze the formation of a team at a certain moment in the match, in addition to the data about tactic formations, we also need to know about the relevant data of each player, such as passing success and the ability of tackling. So our group project will visualize and perform analysis of a soccer match from a variety of perspectives. Also, since the data of a soccer match is usually generated by the position trackers worn by players, we will visualize the data in the format of animation rather than traditional static images based on their time attribute.

Design Processes

Visualization of Passing Event & Tactic Formation

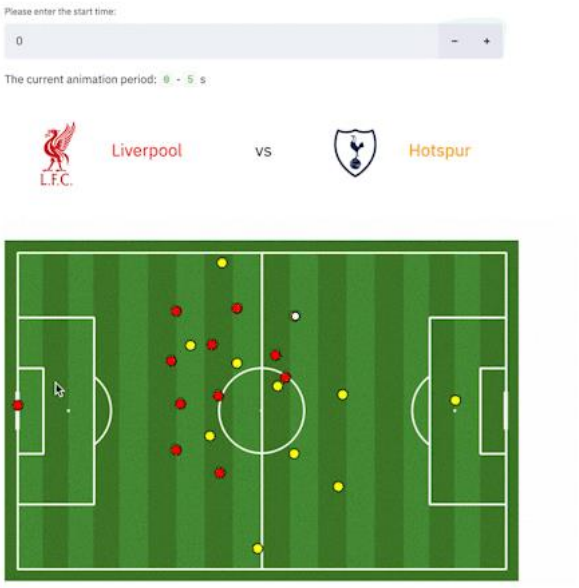
The first task of our group project is to visualize the tactic formation and passing events in a soccer match. At the beginning, we only wanted to visualize passing events between players and we thought it would be a good idea to visualize passing routes based on a series of key events, which is shown in Fig. 3. However, as we mentioned earlier, for professional users like trainers and coaches, they need to combine more information to evaluate an event. For example, it is hard to tell whether a passing is good or not without information about the current tactic formation and players' positions.



	time	actiontype	player	team
1	81m19s	pass	Axel Witsel	Belgium
2	81m21s	dribble	Kevin De Bruyne	Belgium
3	81m24s	pass	Kevin De Bruyne	Belgium
4	81m25s	dribble	Eden Hazard	Belgium
5	81m27s	shot	Eden Hazard	Belgium

[Figure 3]

When we were thinking about how we can do better, we noticed that there's another database on statsbomb which is based on the players' and soccer's position data at every 0.1 second during the match. Since this data is collected in real time by position sensors worn by the players or installed on soccer, we chose to visualize the data in the format of animation – one of the best and intuitive ways to visualize data having an attribute of continuous time. This visualization solves two problems at once, the passing between players can be reflected by the 不trajectory animation of the soccer players and the soccer itself, also, the users can combine other information like the tactic formation to perform a more reasonable evaluation of an event.



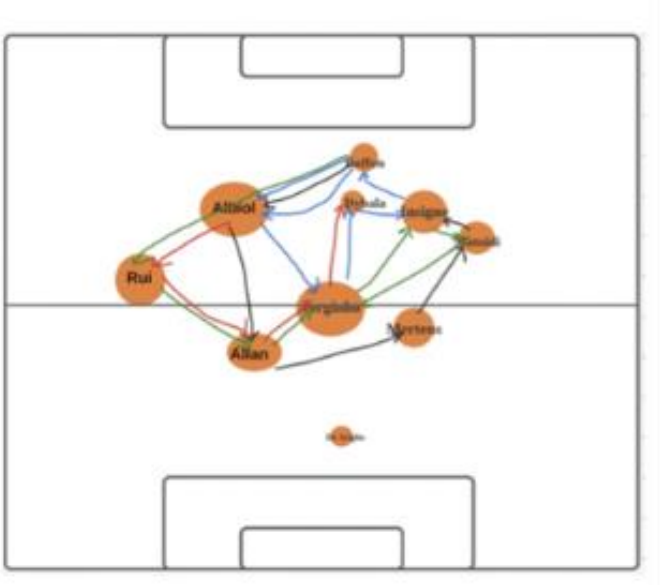
[Figure 4]

Our final visualization is shown in Fig. 4, where circles are colored in three different colors: the red and yellow circles represent the current position of players which belong to different teams, the white circles represent the position of soccer. Users can change the start time of the animation that they want to analyze in the input box above, the animation will be rerendered every time the input field is changed.

Visualization of Connectivity Network

The second visualization we were trying to come up with is a passing network. Soccer is a team sport, and it is important to analyze how well the players are connected. To visualize this, we decided to use a form of “network” to show interconnectivity.

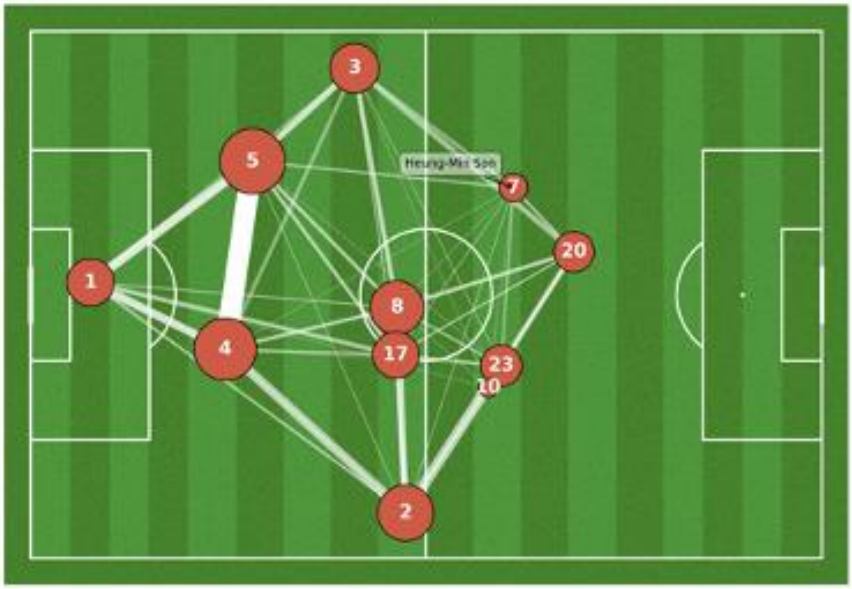
The initial sketch of the visualization is like below in [Figure 5].



[Figure 5]

The key specifications are the location of the players, connectivity that shows interactions between the players, and the activity of a single player. The position of the circle which is showing the location of each player is based on the average location of the player. We were planning to show interactivity between the players by connecting two dots (players) with lines. A line is supposed to show one time of interaction. Based on this principle, the two dots which have the largest number of lines will be the players that are having the greatest number of interactions. We were planning to encode the level of activity of each player with the size of the circle. As you can see on the initial sketch, there are different-sized dots. Each player has both action-in and action-out like passing-in and passing-out. Regardless of the type of action, we will sum up all the actions by each player and will reflect this on the size of the circle. We are expecting to see a level of activity by each player as a result. The last specification is an annotation. To specify players on the final visualization, annotation in the circle is needed. The example of the annotation in [Figure 5] is the name of the players.

Luckily, the match dataset contains most of the data that we need except the player's name. The final visualization of the passing network is like below in [Figure 6].



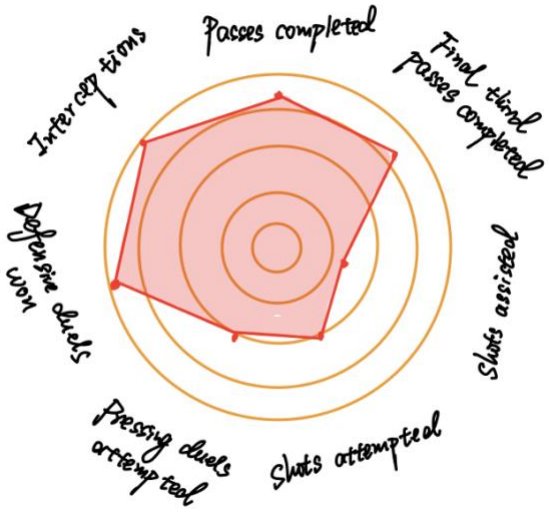
[Figure 6]

It contains X-Y location data of the players when they made any action, action type, and the time data. As planned, the circle is located based on the average X-Y location of each player. The size of the circle is encoded by the proportion of pass frequency of each player to show how active the player is. This is calculated by total_pass_count/total_pass_count.max(). Initially, we planned to draw a line per interaction between the players. However, we slightly changed this idea by using line width. This was because drawing multiple lines is not visible when there are many lines between the two players. At the same time, it is hard to make a comparison intuitively based on the number of lines. Line width is a clearer way of visualizing this. Like the size of the circle, the line width is encoded by the proportion of pass frequency between the players. Another specification that changed while doing a visualization process is annotation. We planned to annotate each player's name in the circle, however, it was not a good idea in many cases. The size of the circles varies, and the names of the players are too long to be fitted. We considered using the abbreviated name, but it might not be clear enough for some of the users. So we annotated the circle with a back number of each player and added an interactive tooltip that shows each player's name when the user hovers the mouse over the dots.

Visualization of Player's Performance:

Our group's third visualization is to show each player's performance. We concluded that a radar chart is a good way to compare player's various performance statistics. Our original design sketch was shown below in Figure 7. We planned to get some common attributes of soccer players and visualize each player's abilities on a radar chart. We expected the users to compare

players on each attribute more easily using a radar chart and it would enable the users to catch what are the strengths and weaknesses of the player at a glance.



[Figure 7]

While we were trying to build this visualization, we encountered some unexpected data collection issue. Since we couldn't access the raw data via StatsBomb's API, we slightly turned our direction and searched for other data resources. We were able to collect 11 common attributes to determine players' performance on FBref.com, which is a website contributed by StatsBomb. As this website is devoted to track statistics of football teams and players, we were able to create a soccer scouting report as our dataset that matched our teams with our previous resources. These attributes contain the tactical abilities and awareness that the best players possess in all areas of the pitch. We were coming up with these attributes based on the fact that players that lack these attributes will have a hard time getting into professional games at any position (*What Are the Main Attributes Needed to Become a Professional Footballer?*, 2015).

In the process of making radars, the two modules, "mplsoccer" and "radar_chart", helped us tremendously in creating the visualizations. Our inspirations of radars were mainly from StatsBomb. One of the most important decisions with radars is setting the radar's boundaries. StatsBomb (Knutson, 2016) popularized the use of radars and indicated that the top 5% and bottom 5% of all statistical production by players in the position. To increase the aesthetic quality, we also used mplsoccer's FontManager to load some fonts from Google Fonts and specified color scheme on the chart.

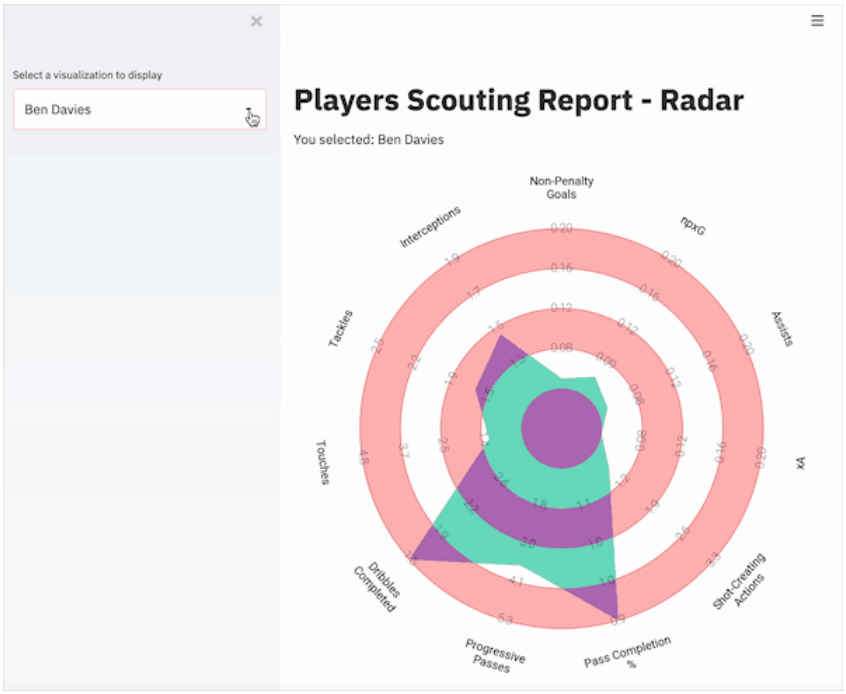
Our final visualization of the player's performance can be found below in Figure 8. This is one of the radar charts that the users can interact with in our final Streamlit code. The example

chart was implemented with Heung-Min Son's data. The chart indicates that Heung-Min is a good forward player with high non-penalty goals and dribbles completed values. He also shows good performance in shot-creating actions and high percent of pass completion.

There was a significant difference between the values of xA (expected assist) and Assist. The likelihood of him scoring assists is high, but his award is much lower than his expected value. This chart can be a valuable contribution to the entire visualization in that the users can compare the performance differences between players and catch strengths and weaknesses at a glance.

[Figure 8] Performance of a player: Heung-Min Son

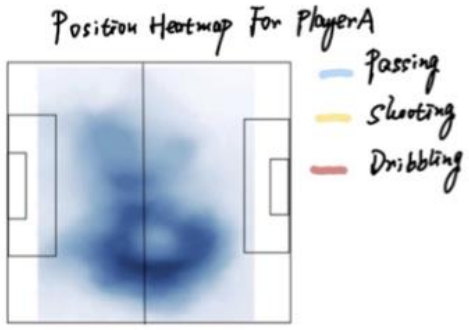
For the interactive component, we extracted the array of values for each player with Pandas and visualized our work on Streamlit framework. The final dynamic visualization of player's performance can be found in Figure 9.



[Figure 9]

Visualization of Player's Action Type:

In order to best show each player's location in a particular game by action type, we believed that it was necessary to plot the data points in a heatmap format or use a scatter plot. Our initial sketch for each player's action, shown in figure 10, shows the basic idea of a heatmap created on top of the soccer field background with interactivity on the action type where users can choose an action type that they want to view.



[Figure 10]

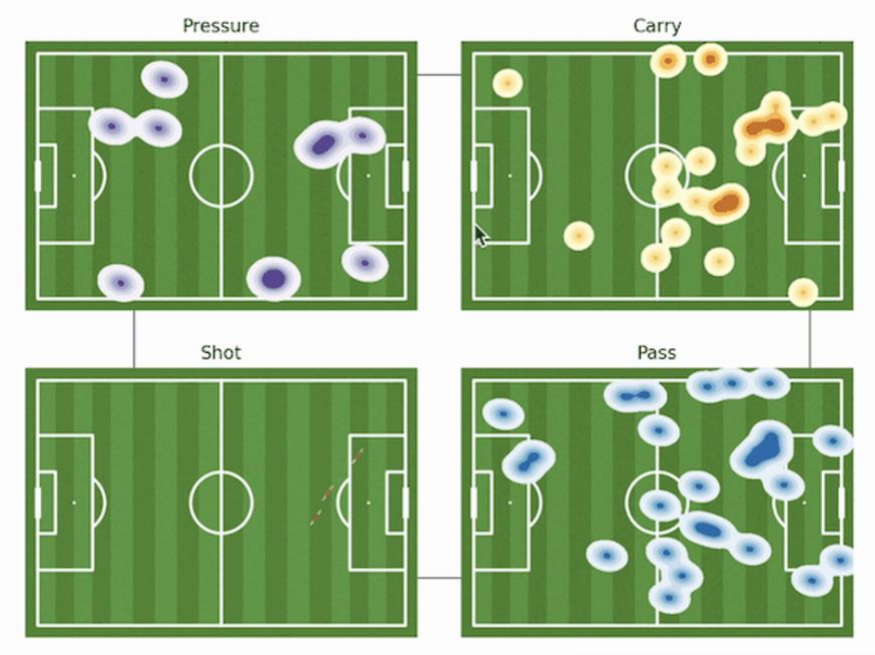
The obstacles came from finding a way to duplicate the initial sketch. One of the issues with utilizing the regular seaborn heatmap is that the color distribution occurs on the entire graph in square blocks. With heatmap square blocks covering the entire graph, there was a chance that the users won't be able to understand the purpose of the visualization and unable to analyze the graph.

Using a scatter plot also generated some concerns on overall user experience and effectiveness of the visualization. Because of the characteristic dataset and the large number of rows that it contains, it was hard to distinguish the circle plots from each other and see which point the player was more frequently present in comparison to others.

Through extensive research and a number of trial and errors, we were able to settle on a visualization function in the seaborn library called kdeplot. Kdeplot is a Kernel Distribution Estimation Plot that depicts the probability density function. By this way, we were able to visualize the distribution of a given dataset along the x and y-axis on top of the soccer field background without any distortion and color overlap. Also, with this method, users would be able to distinguish the player's location and its frequency more clearly compared to past trials.

One other major difference between the initial sketch and the final design is user interaction. In order to successfully deliver our designing intention and purpose to the users, we have simplified our visualizations to prevent any information overload. By pre-selecting the match and the team we want to focus on and adding one significant interaction at a time, we were able to avoid complications and successfully build our dashboard by well structured code that could be easily used for further implementations.

Furthermore, we originally planned for the users to select the action type of a pre-designated player. However, for our final visualization, we have decided to allow the users to pick the player that they want to analyze in contrary to the initial plan and arrange the main action types in one view. Rather than having users to move in between the action types graphs, we believed that it would be more efficient to have the plots laid out in one view.



[Figure 11]

As briefly mentioned before, seaborn kdeplot was the main visualization tool used to plot the location of each type. Kdeplot is a method for visualizing the distribution of observations in a dataset that requires x and y vector coordinates. With kdeplot, we were able to visualize the actions of the players without any color overlap. Also, in order to further prevent any color overlay with the green soccer field background, contradicting colors were chosen for each action type. The density of the colors represents the number of occurrences of average x and y data points throughout the match which shows how often the player showed the action type at a specific location.

The heatmap shows 4 main types of a player's actions in a particular match: pressure, carry, shot and pass. The different colors of each type are used for the users to easily distinguish the types by a glance. The statsbomb API dataset provides over 3400 events per match with various variables associated with each row. Each row is clocked by time(seconds) and contains valuable player data that can be used to analyze players' movements. The variables used to create the heatmap are start_x, start_y, end_x, end_y, type_name, player_name. For this particular visualization, we conducted additional calculations with the start x, y data and end x, y data to find the average coordinate data for each row. Because each player is continuously moving throughout the match, it was critical to find a specific point for the player's location at each time capture. This calculation was used to distinctively plot a player's location at each second so

that the action patterns users can analyze the frequency of different action types at a certain location.

This visualization not only helps to understand the chosen player's action types in a particular game but also contributes to understanding the overall play patterns of the player. There are 22 action types included in this dataset that are associated with each player at each time frame. By just looking at the 4 main actions types of each of the players, the users will be able to analyze what kind of action each player carried out the most in a particular game and where it happened the most. This will be a major contribution for the entire visualization when users are analyzing players' performance at the end of each game and finding places to improve.

Although many public visualizations used similar methods to show the location of each player at each game, they lack in detail in that they just focus on the analysis of the entire team or on the location of a player without specifying which action the player is performing. By providing the users with players' locations and frequency of actions by the action types, the users will be able to perform more intricate analysis on each player such as in which location does a player show a particular action type and how has the team tactic impacted player's actions.

The visualization objective of our project was to fill in the disparity between team analysis and individual analysis and provide the users with a continuous flow analysis environment on the soccer field. With this additional player analysis visualization, users can conduct a much more sophisticated analysis for a team and for each player.

Limitations

1) Compatibility in interactivity features on Streamlit

After we finished our basic design of our visualization with Matplotlib, we have implemented our visualizations on Streamlit to use more interactivity features. Dropdown menus for selecting team and player worked nicely on the first, third and the last visualization. However, we later found that the interactive feature with Matplotlib itself was not compatible with Streamlit. We used mplcursors to show the name of the players when the users hover their mouse over the dots(each player). The tooltip can not be shown on Streamlit because it tries to export the matplotlib image as a static image. Even though we have added a player information table at the side of the page on Streamlit, it would be better if we can enjoy some interactive tooltips and catch the information of the player more intuitively.

2) Run time issue of dynamic rendering

We used fairly complex code and multiple interactive features on our visualizations. Even though it provided us and possible users more enjoyable and informative visualizations, it caused some problems with computing power. Specifically, the first visualization, an animated tactic formation, we needed to transform animation into GIF to display it on Streamlit because of the limitations on Streamlit. This caused the code to need about 30 seconds to render a 5 second short animation.

3) Data availability

As we mentioned earlier, not all the datasets were available with the original StatsBomb API. We needed to search for other data sources while we were trying to visualize the performance of the players. As it needed to connect an extra data source from outside of the API, it caused some extra work. If we need to get information about a new player who was not available in the dataset, it will require us to go an extra step and build a connection to retrieve the relevant data for the player every time.

Future Direction

If we were to continue this project for further development, there are few improvements and changes that could be made. Some feature improvements would be,

1) Scaling up user interactivity features

Our visualization currently has some basic interactivities which can be improved to give the users the opportunity to conduct more personalized analysis. Additional user interaction features can be a supplement drop down feature for competition, season and year so that users can not only select a player from a team but also select any match from any season they would like to analyze.

Because our code is structurally refined for further development, additional features can be easily implemented to give the users more flexibility.

2) Pattern recognition with machine learning technique

Using machine learning prediction modeling, it would be possible to classify and predict a player's possible action tactics based on past patterns and visualize it. A few additional examples of this implementation would be,

- Prediction of higher passing or shooting occurrences based on tactic formation
- Contribution prediction of players based on their past performances and skills

3) Application to other sports

With an appropriate dataset, this visualization can be applied to different sports. In order to successfully apply this visualization to other sports, the most critical variables to look for are the x and y coordinates. Because this visualization is mostly based on the location of players throughout the entire game, it is important to check the dataset beforehand for the x and y coordinates of each player collected in a 1-3 second term.

References

Knutson, T. (2016, April 25). Understanding Football Radars For Mugs and Muggles. StatsBomb | Data Champions.

Charles Perin, Romain Vuillemot, Jean-Daniel Fekete. SoccerStories: A Kick-off for Visual Soccer Analysis. IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, 2013, 19 (12), pp.2506-2515. ff10.1109/TVCG.2013.192ff. ffhal-00846718f

Y. Wu et al., "ForVizor: Visualizing Spatio-Temporal Team Formations in Soccer," in IEEE Transactions on Visualization and Computer Graphics, vol. 25, no. 1, pp. 65-75, Jan. 2019, doi: 10.1109/TVCG.2018.2865041.

What Are the Main Attributes Needed to Become a Professional Footballer? (2015, March 30). The Soccer Store Blog. https://www.thesoccerstore.co.uk/blog/football-training/main-attributes-needed-become-professional-footballer/