FIFA Women's World Cup 2019 - A Visual Journey through Soccer Analytics

CptS 575 - Data Science Project Report
Aryan Deshwal, WSU ID - 11624202
Reet Barik, WSU ID - 11630142

December 12, 2019

Abstract

With the increasing availability of data in the world of soccer, the field of soccer analytics has come alive in recent years. There is an emerging need to come up with visualization techniques that help quantify different aspects of the game, be it players' profile, the momentum of a game, opposition's style of play, etc. Another important requirement for these techniques is for them to be not demanding in terms of comprehensibility. This is because more often than not, the end consumer is a soccer coach or manager who has no formal training in the field of data science and analytics.

1 Introduction

The instance of statistics and computer science permeating an outdoor sport is not the first of its kind. Analytics has already been a mainstay in the field of basketball, hockey and baseball (made a part of popular culture by the movie 'Moneyball' starring Brad Pitt and Jonah Hill). Soccer, being a more fluid and dynamic sport compared to others, was considered to be especially challenging in terms of it being a suitable candidate for the use of statistical and other analytical techniques. This was compounded by the type of events possible in soccer being exponentially huge (throws, direct free kicks, indirect free kicks, offside, yellow cards, red card, penalties, etc.). But in this age of big data and the advent of other technological improvements in the field of player tracking, automated event data collection etc., soccer has become the latest of the big outdoor sports to be included in the list of those relying increasingly heavily on data driven decisions, be it game strategy formulation, opposition analysis or player scouting. Despite the recent surge of activity in the field of soccer analytics, most of the advances made are yet to make its way to mainstream media which comprises of post-match analysis and punditry on platforms like ESPN, Star Sports etc.

The broad aim of this project is to visualize the recently concluded 2019 FIFA Women's World Cup which was held in France and familiarise the audience with some of the more important visualization techniques currently in use in the field of soccer analytics. The above is done by making the United State's Women's National Team (referred to as USWNT from here on) the focus

of this project and visually investigating their superior and dominant performance throughout the tournament that made them the eventual winners.

2 Data description and Problem Setup

The main bottleneck and pain-point throughout the years has been the unavailability of quality data which hampered the development of sophisticated techniques. Finally, vendors like Opta, Wyscout and Statsbomb have superior technology at their disposal which makes the collection of quality data possible. Of the vendors mentioned, **Statsbomb** has made a part of their paid data repository publicly available to encourage independent analysts to get into this field and start playing around with actual real world data.

This project uses the FIFA 2019 WWC data available at: Statsbomb's Open Data repository. The documentation for the same can be found at the following link: Statsbomb Data Documentation.

The availability of event level data is what sets apart the chosen data set from the rest. All matches included have highly granular event data available with timestamps accurate to the milliseconds. Moreover, the x-y coordinates of such recorded events are also available along with other supplementary information like the name of the player who was responsible for that event, the name of the opponent player it was recorded against, whether the event was part of open play, etc. The availability of such a veritable gold mine of information makes the whole exercise much more rewarding at the end.

As mentioned earlier, the primary aim of this project is to visually investigate the USWNT's performance in the 2019 FIFA WWC and determine the various aspects of their game play that made them the clear winners of the tournament at the end.

3 Methods and Analysis

In this section, we present our analysis and main findings about USWNT's performance in the entire world cup. We present all the three main aspects of the game: passing, shooting and defense in both finer and big-picture details. Finally, we described few ways to summarize the entire game in one form of visualization.

3.1 Pass Sonar

We know soccer is a game in which passes play a very important part and their quality is almost always a factor in a teams overall performance. We employed a technique called 'Pass Sonar to judge the passing abilities of the teams. The underlying concept has been around in cricket analytics for a long time with the name 'Wagon wheel'. This was first introduced in 2011 as a technique in soccer analytics by Graham MacAree, the current interim editor-in-chief of SB

Nation. The final variant of Pass Sonar in its current form was due to the efforts of Eliot McKinleys work on Twitter and the American Soccer Analysis blog.

The way to interpret is it as follows: Lets say a player is standing at the center of this plot facing north which coincides with the opposition goal. The length of each sector is proportional to the number of passes by her in that direction. And the color gives the aggregate length of those passes. This technique, though deceptively simple to look at, gives a very comprehensive understanding of a team, or an individual player's passing style and makes visual comparison quite easy. For example, Figure 1 shows a cumulative (accumulated over all games) pass sonar of United States and the rest of the teams competing in the tournament. As evident in the figure, the US teams cumulative pass sonar is very rich and well-developed compared to the rest of the teams (ROW). The skew in the ROW profile suggests their players tendency to clear the ball up field unlike US whose passing was more organized.

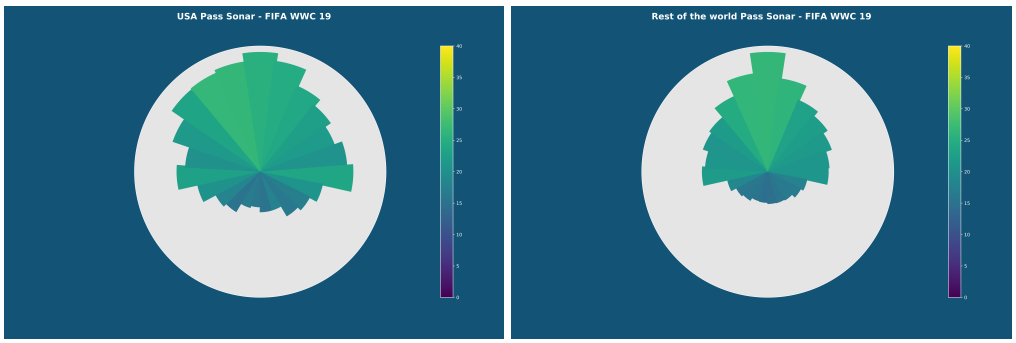


Figure 1: Comparison of passing performance of United States team with the rest of the teams using cumulative pass sonar.

The Pass Sonar of the USWNT is especially impressive when compared against that of Netherland's and Sweden's (first and second runners-up respectively). This is shown in Figure 2.

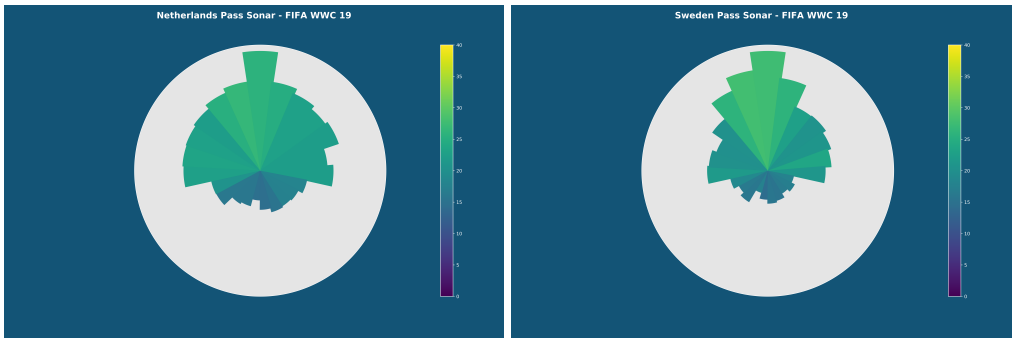


Figure 2: Cumulative pass sonars of Netherlands and Sweden in FIFA WWC 2019.

All of the visuals point to the USWNT's clear superiority when it comes to the pass distribution and passing quality as compared to the rest of the teams.

3.2 Passing Networks

Although pass sonar is an effective high-level abstraction of the passing quality, it does not show us the chemistry between team-mates. Passing networks is an effective visualization to address this issue. Every player in the match is represented by a circle in a passing network. The size denotes the number of touches by that player and the thickness of edges represent the number of passes played between two players. They highlight connections between players in terms of pass frequencies and ball touches. It helps to show how a team is structured when playing with the ball and so can become very useful towards understanding the general team tactic. For example, we illustrate this approach by considering a group stage match between United States and Thailand in Figure 3. As clearly evident, US has a stronger passing network indicating better team chemistry as compared to Thailand.

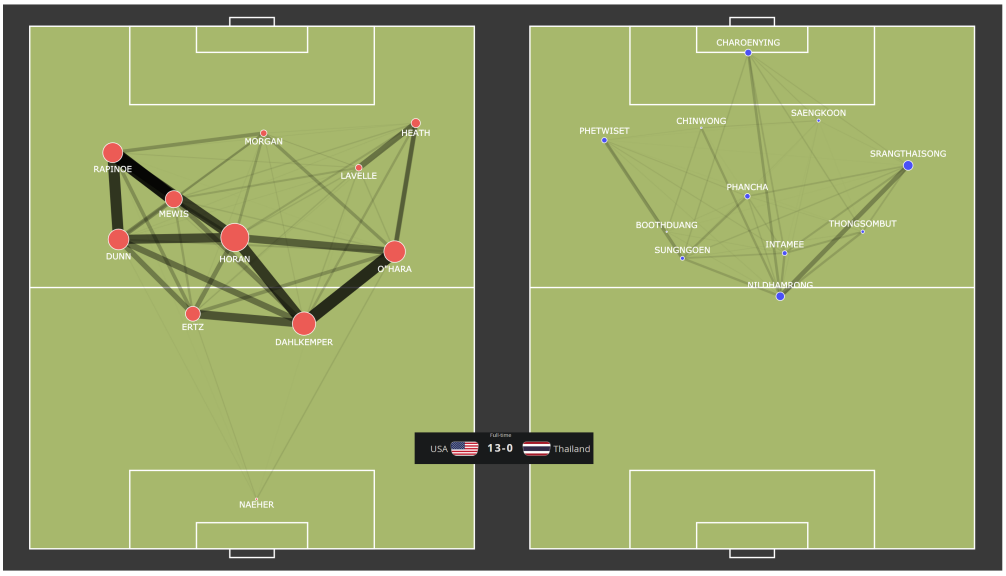


Figure 3: Passing networks of United States and Thailand for their group match.

The USWNT played a total of 7 matches in the tournament starting from the group stage to the final. The Passing Networks of the other matches are shown in Figure 4 below:

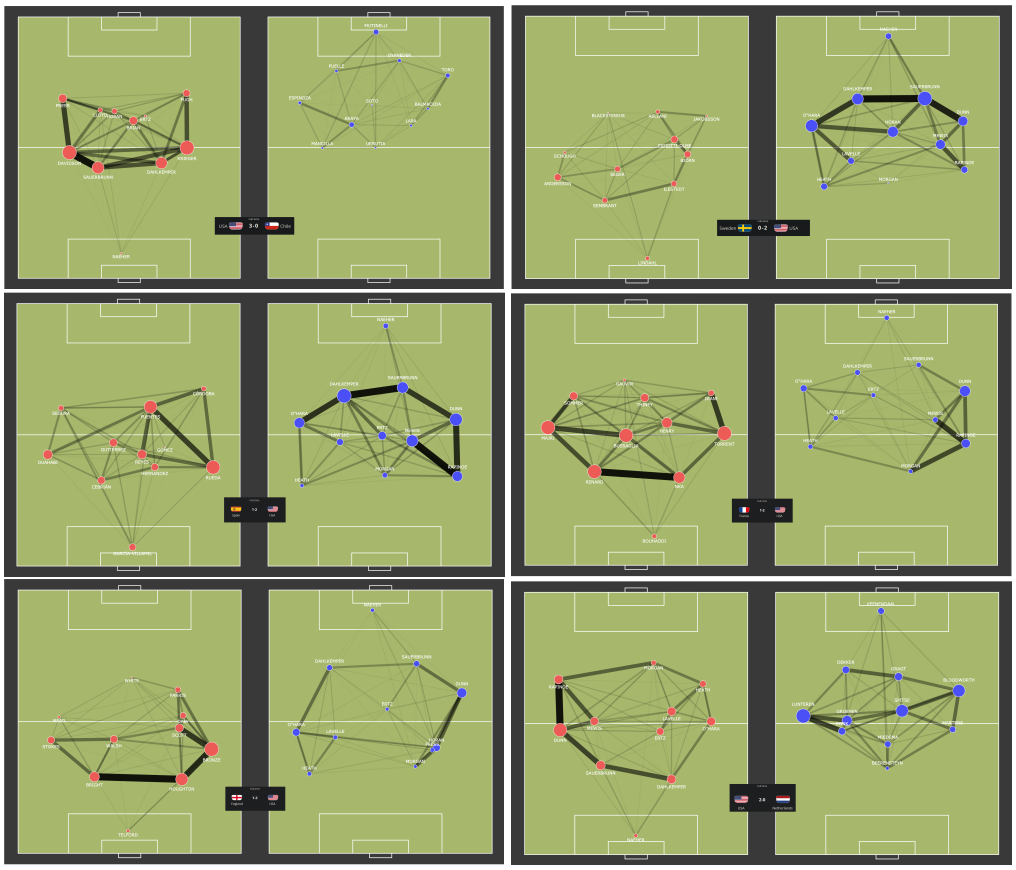


Figure 4: Passing networks of US vs the rest in FIFA WWC 2019

3.3 Shot Plot

We have been analyzing the passing aspect of the game till now and have observed US's superior passing abilities. Now, we analyze the scoring aspect (perhaps the most important part) of the game using a technique called 'Shot plot'. As the name suggests, it is a visual representation of the shots taken by the two teams on a background that represents a generic half soccer field. For better comparison, both teams are shown shooting at the same goal. Each dot is colored differently based on the team the player who took the shot belongs to. Black dots are the ones that resulted in a goal. The size of each dot denotes the Expected Goal Value (xG) of a shot. Expected goal value is a metric that measures the quality of a shot by computing the probability of it resulting in a goal based on multiple attributes like closeness to the goal, angle of the short, whether the player shot with her preferred foot etc.

Figure 5 shows the shot plot visualization of the final game between United States and Netherlands. It can be clearly observed from the figure that US attempted a lot more shots with higher chance of conversion.

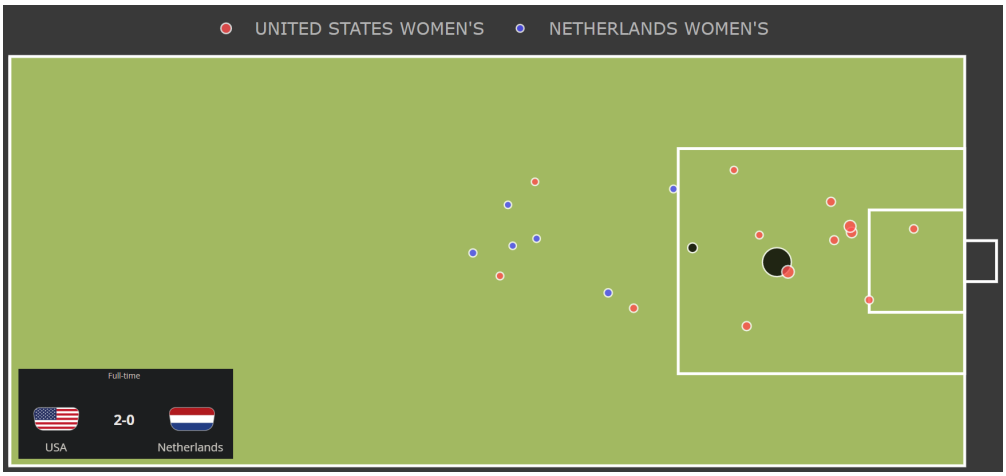
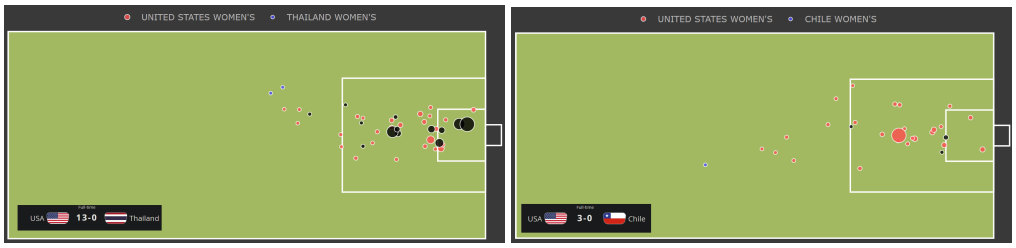


Figure 5: Shot plot visualization of the final game between United States and Netherlands.

The Shot plots from the rest of USWNT matches tell us a similar story. This goes on to say that the USWNT was not only good in possession as can be seen from both the pass sonars and the passing networks, but they followed it up with quality end product in terms of shooting at the opposition goal.



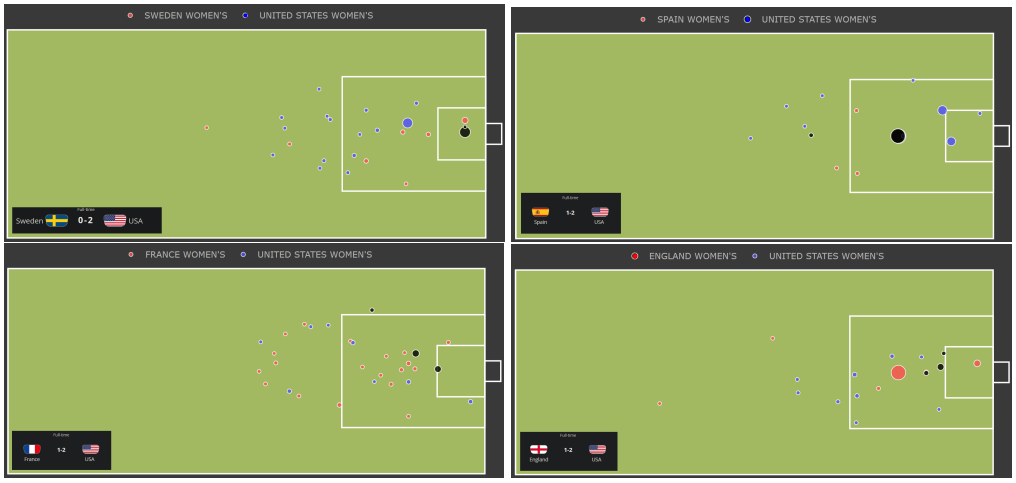


Figure 6: Shot Plot of US vs the rest in FIFA WWC 2019

3.4 xG Chart

This technique uses the metric 'Expected Goal' that was mentioned in the explanation of the Shot plot. There is much debate as to the introduction of xG into the soccer analytics literature. But we do find mention of it in the works of Howard Hamilton [2], CEO of Soccermetrics Research, LLC and another one by Sarah Rudd [3] who works in the analytics department of the footballing giant Arsenal FC. The primary disadvantage of looking only at shot plots is that they take into account shooting events exclusively, which, even in case of a free flowing exciting match, accounts for an insignificant proportion of the whole ninety minutes. One way to visualize the whole match and not just isolated shooting events, is by plotting the expected chance of a team scoring at any point in the game. Every action taken by a team throughout a match either leads to an increase in their chance of scoring or a decrease.

To this end, a model is designed that takes each event as an input and spits out the expected goal value at the end of that event as the output. A very primitive xG model, for the purpose of understanding, based only on shots would be: $xG = 0.10 * shots$. This treats every shots as equal and assumes one in every ten shots ends in a goal. A more sophisticated approach is to bucket the data according to shot type (e.g. in versus out of the box, headed versus non-headed) and assign a different probability for each class. Indeed, this is how some xG models work, and they perform better than the naive all-shots-equal model. Others perform regression on various shot properties, and one can end up with quite a complex set of equations. However, with large datasets and many variables describing each shot, this is an ideal application for machine learning/artificial intelligence.

In this case, gradient boosting is used to learn a set of weakly learned decision trees that give the end xG value as output after traversing the said trees. Some of the properties looked at are shot distance and opening angle, pattern of play, body part, previous actions, etc.

A visualization of that is by a cumulative xG plot, as is done for the final match in Figure 7. As we can see, United States always looked likelier to score as compared to the Netherlands.

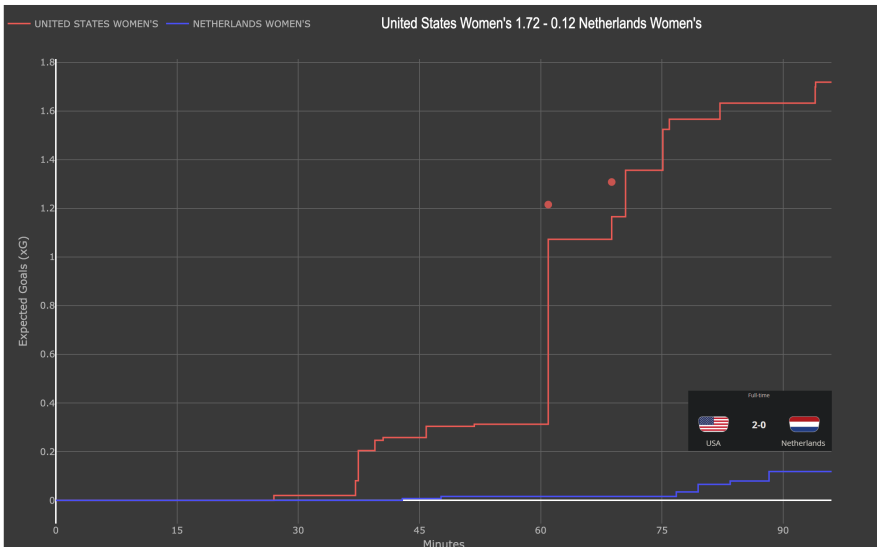
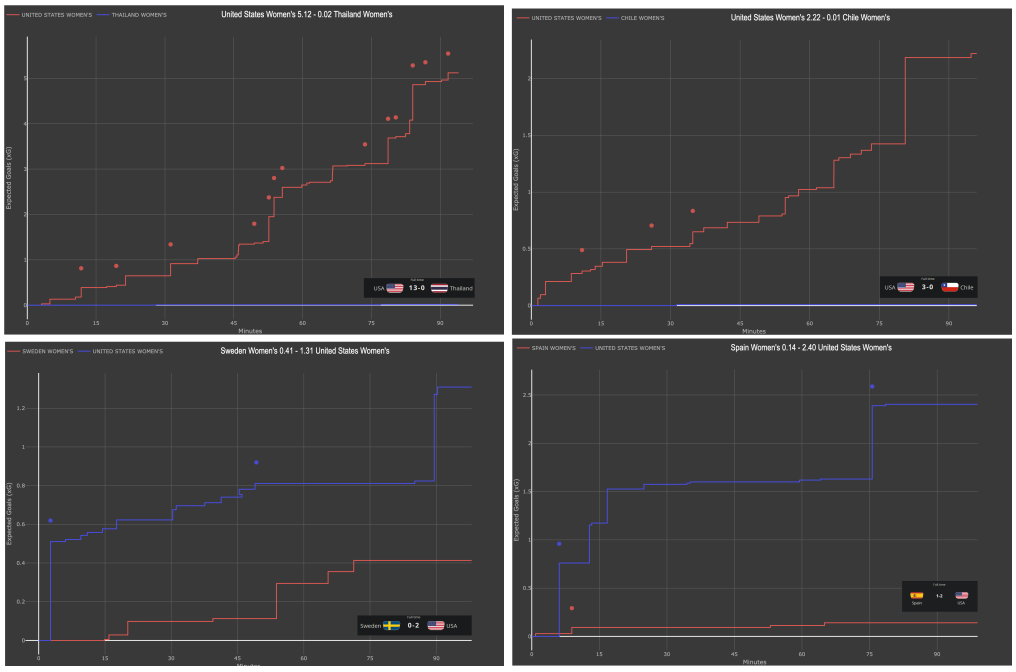


Figure 7: xG plot visualizing United States and Netherlands final match.

Just like in the other techniques, we can take a look at all of USWNT's matches in the tournament as shown in Figure 8. We can see that for the majority of time in all matches, US always look like the ones to score next. This shows that not only did they out-pass and out-shoot most of their opponent's they were in control the majority of the time in all the matches they played.



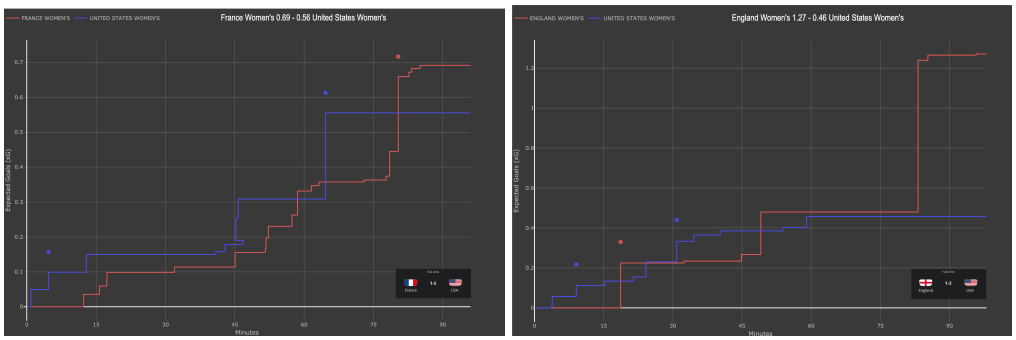
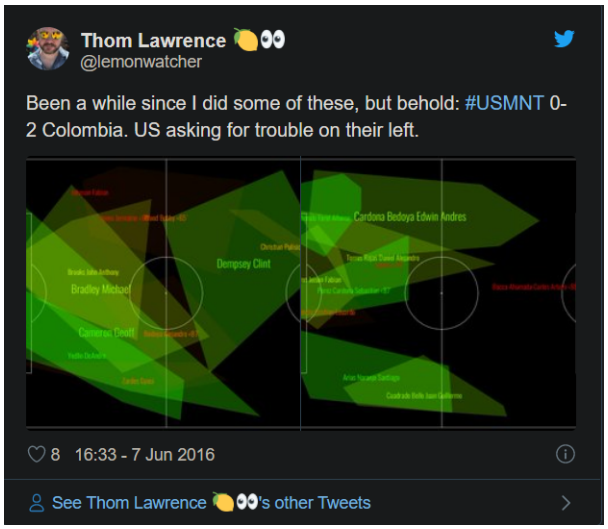


Figure 8: xG plot of US vs the rest in FIFA WWC 2019

3.5 Defense Patch

As the legendary manager Sir Alex Ferguson once said, 'Attack wins you games, defence wins you titles'. So far we have looked at the attacking aspect of the game. Now, the focus shifts to defense. In this section, we analyze the defense abilities of teams by visualizing their defensive shapes using a technique called Defense Patch. This was popularised in the soccer analytics community on Twitter by Thom Lawrence:



The technique uses convex hulls, which is defined for a set of points, essentially display the smallest area needed to cover a set of points. We construct the convex hull of defensive events like clearances and tackles of one player in a match to visualize the area of influence of that defender.

We take a look at the group stage match between the US and Sweden as shown in Figure 9. Each polygon here corresponds to a defender. We see here that the US exhibits the shape of a proper strategic defensive unit which is seldom under pressure as opposed to Sweden who, from the looks of it, had to defend for their lives all over the pitch.

Defense Patch, though intuitive, is not a good indicator of defensive performance in cases of a vast mismatch in performance. For example, the first group stage match between USWNT and Thailand ended with US going on to rout Thailand 13-0. The Defense Patch of that performance is shown in Figure 10 along with the rest of the matches played by the US in the tournament. In that

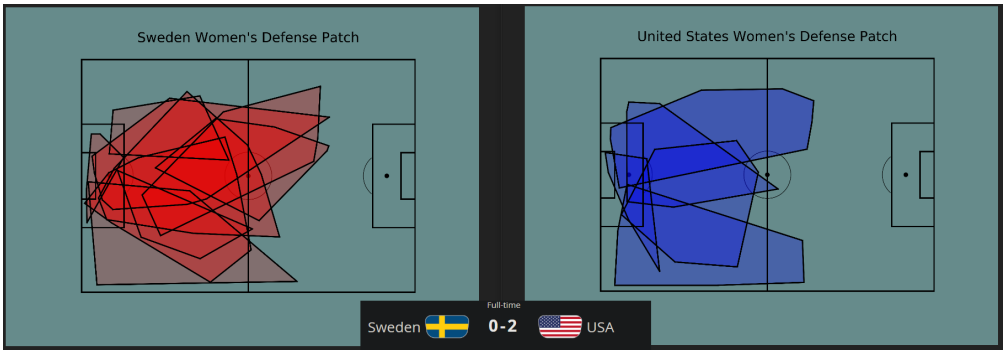


Figure 9: US Defense Patch vs Sweden in FIFA WWC 2019

match Thailand had to defend for their lives when faced with the barrage of attack from the US as evident from the Shot plot and the xG chart of that match. But the Defense Patch is inconclusive because the US hardly had to exert themselves defensively. This makes Defense Patch unreliable when comparing the performance of two competing teams with one team far far superior than the other. The reason for this has been explained in the conclusion section of this report.

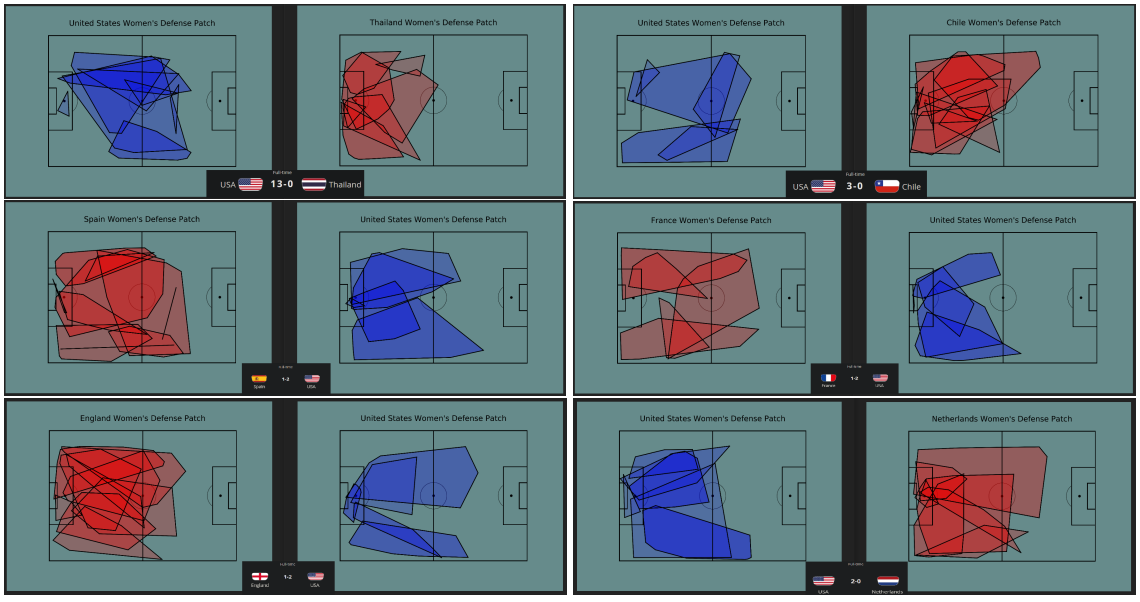


Figure 10: Defense Patch of US vs the rest in FIFA WWC 2019

3.6 VAEP

Now, we look at a framework which was recently introduced in KDD 2019. This paper [1] aims to assign a value to every action in a match depending on the effect it has on increasing its own team's chance of scoring a goal and decreasing the opponent's chance of the same. To this end, it takes as input event level data from three known data vendors (Opta, Wyscout and Statsbomb) and converts it into a language called SPADL (Soccer Player Action Description Language) as an attempt to

unify the existing event stream formats into a common vocabulary that enables subsequent data analysis. Each action is a tuple of nine attributes:

- StartTime: the actions start time,
- EndTime: the actions end time,
- StartLoc: the (x,y) location where the action started,
- EndLoc: the (x,y) location where the action ended,
- Player: the player who performed the action,
- Team: the players team,
- ActionType: the type of the action (e.g., pass, shot, dribble),
- BodyPart: the players body part used for the action,
- Result: the result of the action (e.g., success or fail).

Next, the VAEP (Valuing Actions by Estimating Probabilities) framework trains a probabilistic binary classifier that outputs the probability of a team scoring and conceding a goal as a result of every action. This is then used to assign a value to events in a match.

This provides a very fine level analysis of the data. Figure 11 shows a frequency histogram of the action values in the final game between US and Netherlands. It is clear from the figure that US executed more positive valued actions than Netherlands.

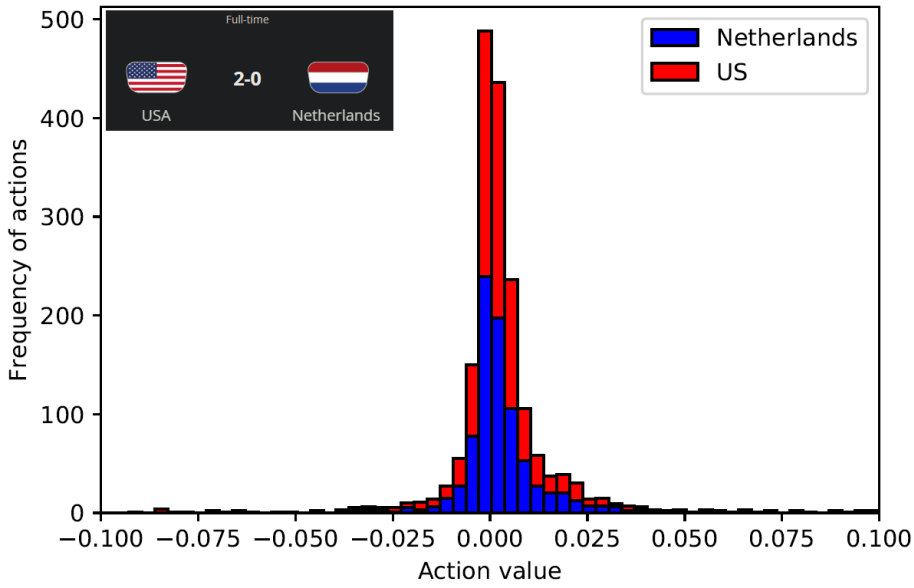


Figure 11: Aggregated Frequency histogram of VAEP values comparing United States and Netherlands in the final match.

A similar treatment is given to all the matches played by the US as shown in Figure 12:

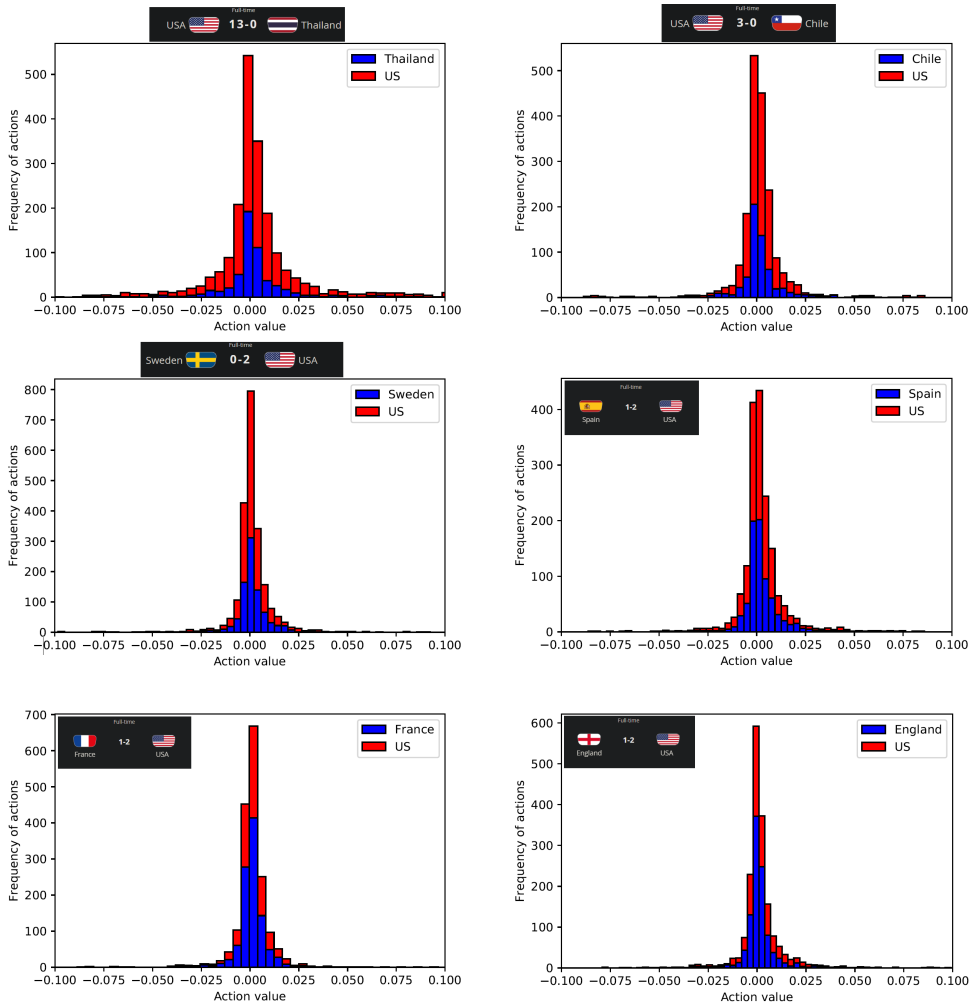


Figure 12: Aggregated Frequency histogram of VAEP values comparing United States and Netherlands in the final match.

The trend is clear to see: US executed a lot more positive valued actions, so much so that it outweighed the number of negatively valued actions executed. This is shown with the help of Figure 13, which computes the aggregated VAEP values for each team in the entire tournament. This aggregated analysis shows that US had the biggest proportion of high-value actions among all teams.

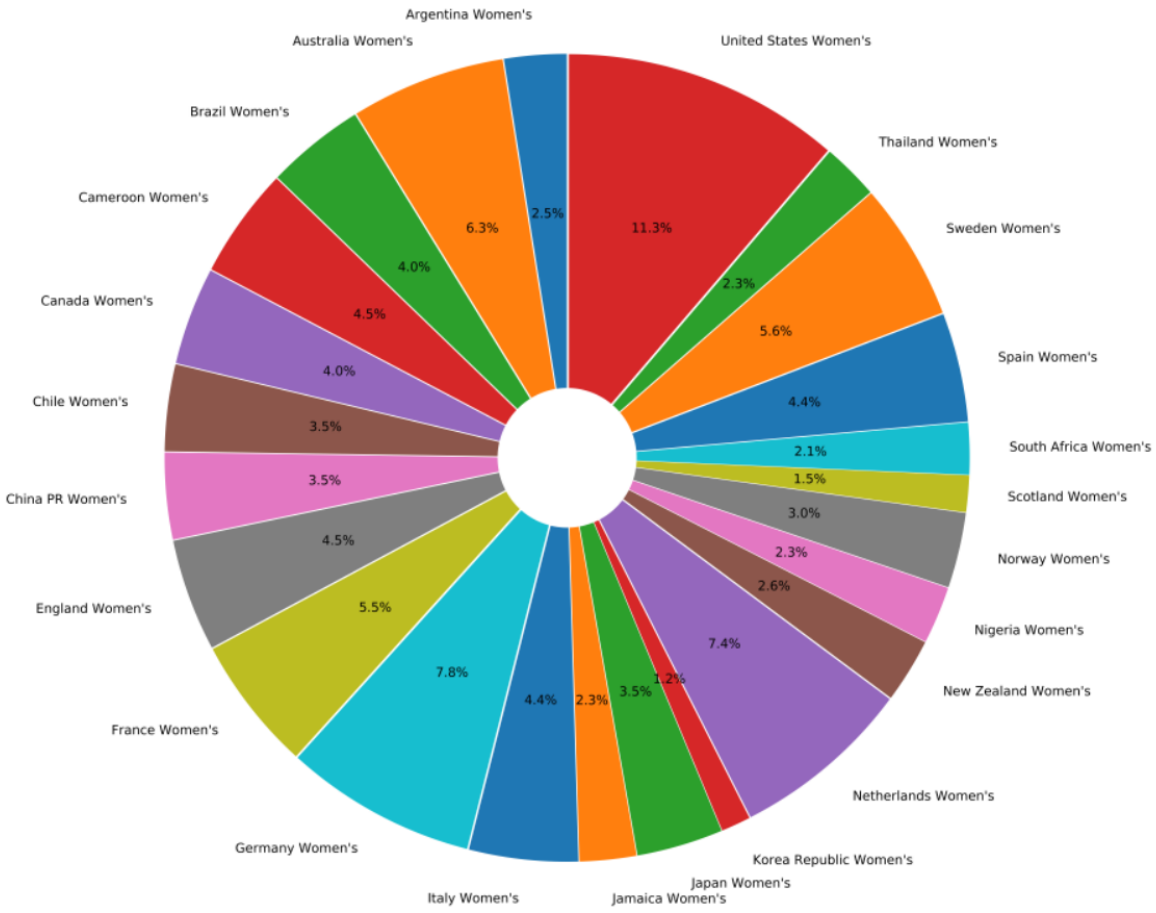
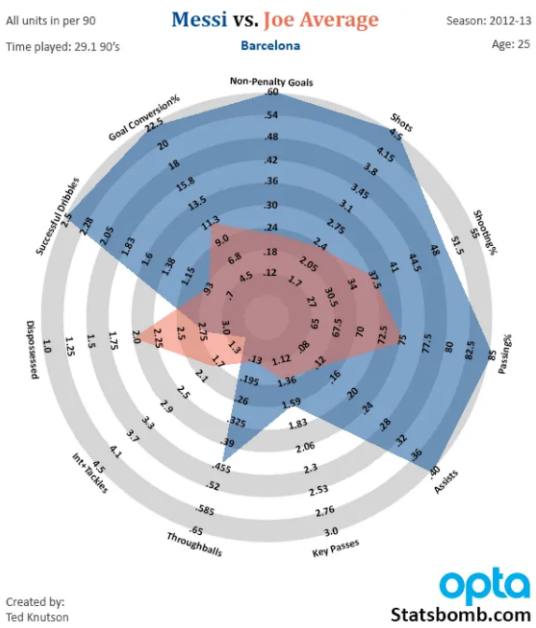


Figure 13: Proportion of total action values for different teams in FIFA WWC 2019

3.7 Radar Chart

A good way to visually summarize each and every aspect of a teams performance for a single match, against their opposition is a radar chart which are quite self explanatory by nature. This has become a staple in the soccer visualization community as explained in an article by Ted Knutson [4], the CEO of Statsbomb. A sample radar chart by Statsbomb is shown:



In our variant, the four different aspects of the game: attack, possession, aggression and defense, are grouped together to make the plot more intelligible from the team's overall performance's perspective. If we take a look at the final match again in the form of a radar chart in Figure 14, we see that US outperformed the Netherlands in all departments except the defense because Netherlands was the one that needed to defend more.

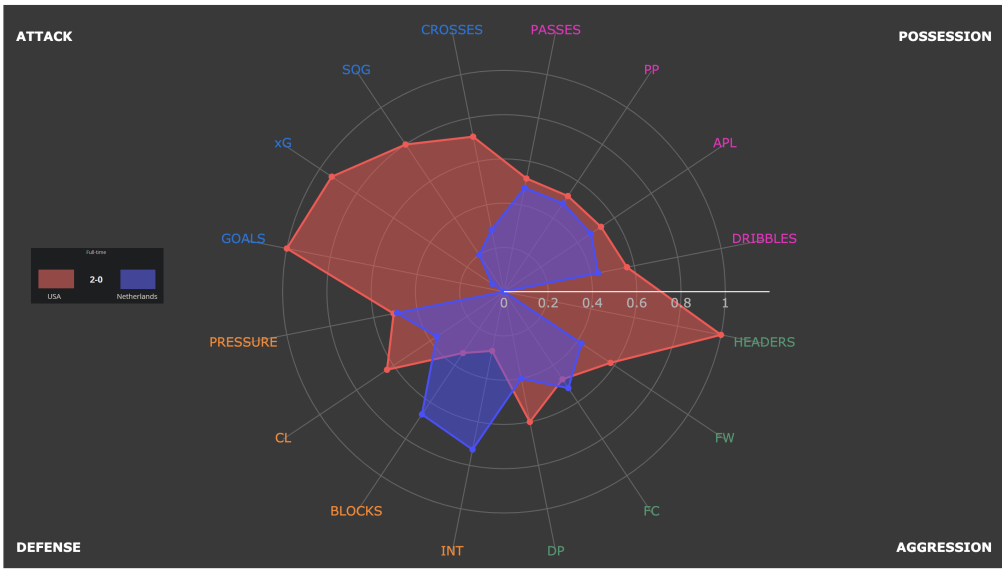


Figure 14: Radar Chart of US vs Netherlands in FIFA WWC 2019

As has been done for the other visualization techniques, Figure 15 takes a look at the US's performance against the other teams in the remaining six matches it played in the tournament.

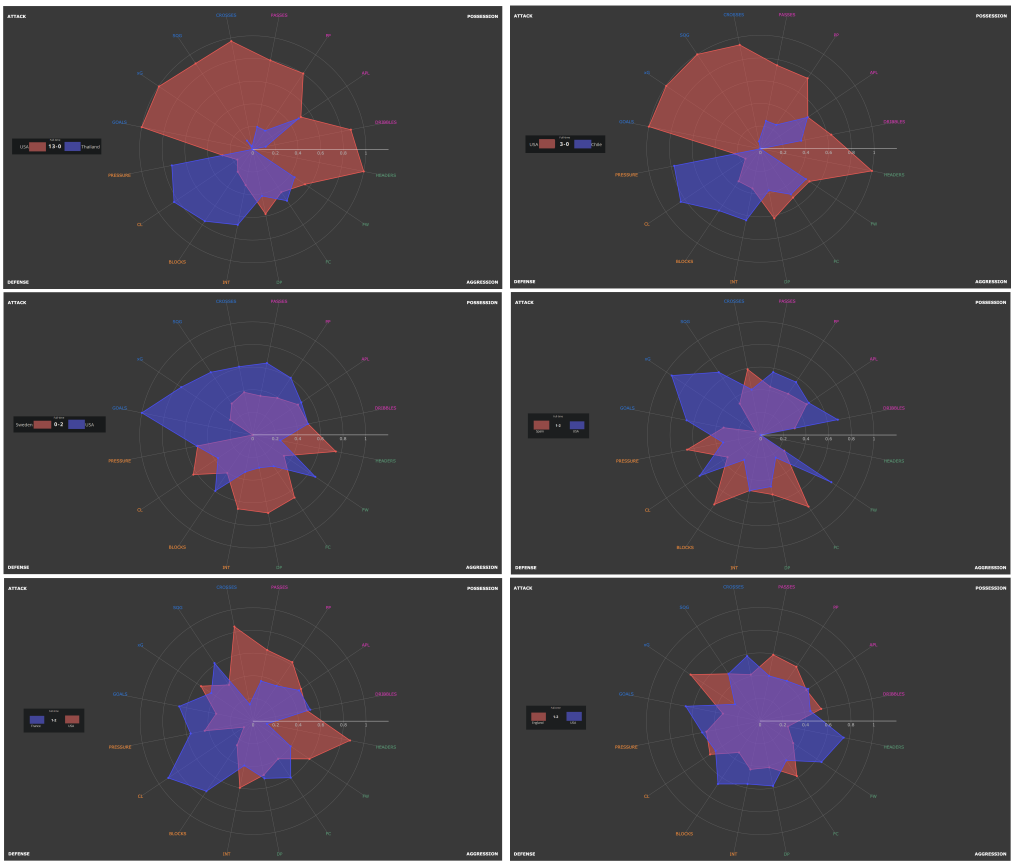


Figure 15: Radar Chart of US vs the rest in FIFA WWC 2019

4 Conclusion

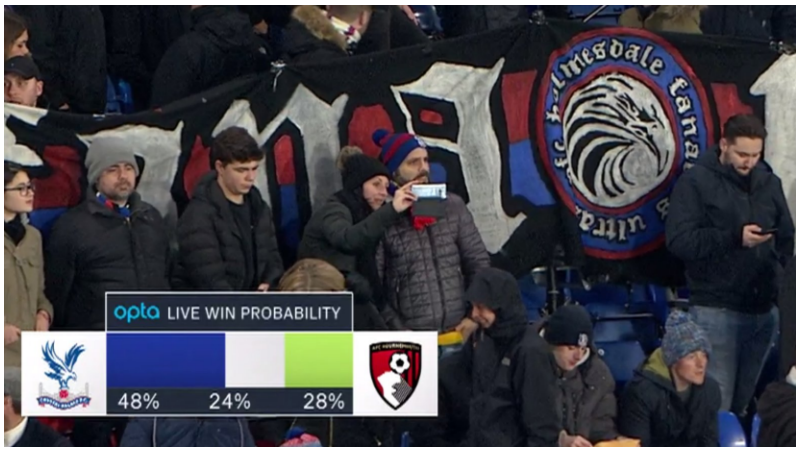
Throughout the report, we looked at how USWNT was tactically the superior team compared to the rest in every aspect of the game:

- Possession: Pass Sonar and Pass Network.
- Attack: Shot Plot and xG chart.
- Defense: Defense Patch.
- Overall: VAEP and Performance Radars.

Though these are intuitive and quite easy to understand, they are not sophisticated enough to differentiate between teams of comparable skill level. For example, a comparison between the Pass Sonars of the French Men's National team and Croatia Men's National Team from the FIFA WC 2018 final match would not be visually contrasting enough to give much information in terms of comparing the passing abilities of the two teams. Some of the existing techniques can certainly be improved to make them more sophisticated. Keeping this in mind, the following are some of the suggested modifications from our side:

- In case of Pass Sonar, an indication of the usefulness of the passes could be made a part of the template (Example: Passing Accuracy).
- An individual approach to Pass Sonar already exists wherein, all eleven player's with their individual sonars are shown on top of a soccer pitch. This idea could be combined with the one from Pass Network and their relative positioning on the pitch could be indicative of their mean position on the actual field throughout the match.
- This report takes into account xG while creating shot plot. This inherently focuses on just the end product while most of the times, the quality of the preceding pass or 'Assist' makes all the difference, so much so that the quality of assist becomes almost as indicative of that shot being converted into a goal as the shot itself. A model which computes the Expected Assist value of a pass might be an answer to this.
- Expanding the above point beyond just the penultimate pass before a shot will result in taking into account a team's chain of passes or possession that leads to a goal. Evaluating such chains or possessions by building an Expected Possession model could be the logical next step.
- Not much work has been done on the defensive side of things. Defense Patch, though intuitive, isn't sophisticated enough as was explained in the report. The very essence of defending can be summarized by a quote from Maldini, an Italian soccer legend and one of the greatest defenders the world has seen: "If I have to make a tackle, then I have already made a mistake". Defense Patch, by focusing solely on defensive actions like tackles, fails to capture the idea that defending is not always about putting your body on the line. This shows that there is a big scope of further research in this area.
- Radar charts with fixed templates for each position (central defender, winger, left back, striker etc) could be useful in visualizing the suitability of players in a specific position.
- A lot of other sports such as basketball and hockey have a rich body of analytics literature that has seen successful implementation in the real world. Given the similarities between those and soccer, there are bound to be concepts and techniques which are equally applicable in soccer. This could turn out to be a quite fruitful strategy in taking soccer analytics forward.

At the end, we can conclude by saying that soccer analytics is still in a very nascent stage. It was only recently that the term 'xG' was mentioned for the first time in the popular broadcasting channel BBC's 'Match of the Day' segment. A neat project by Opta which calculates the win probability of competing teams in real life during a match made its debut in November 2019 as shown:



Despite the slow start, with the fortunes that soccer clubs rake in as revenue, almost all prominent ones have started employing an in-house analytics department. The biggest example of the increasing influence of soccer analytics can be observed from the fact that three of the last four editions of MLS has been won by either Toronto FC (with an analytics department led by Devin Pleuler) or Seattle Sounders (with an anlytics department led by Ravi Ramineni). This can only indicate one thing: Soccer analytics is here to stay.

5 Appendix

The visualizations done as part of this project has been done using **Python** as the sole programming language and used the following plotting libraries: **Matplotlib, Plotly**.

Jupyter notebook was used as the coding platform because of the rapid prototyping required to come up with plots in a short amount of time. The code (in the form of the Juoyter Notebooks) for this project can be found at the following GitHub repository available at the following link: Project Code Repository which was forked from the open data repository of Statsbomb.

References

[1] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019

[2] Hamilton, Howard (8 January 2009). ”Moneyball and soccer”.

[3] Rudd, Sarah (24 September 2011). ”A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains”.

[4] https://statsbomb.com/2018/08/new-data-new-statsbomb-radars/