Laurel Ayuyao

Professor Barron

ITAO 40520

17 November 2020

Sports Analytics Project

# Introduction

In the world of sports analytics, volleyball has been one of the last sports to try to use data to drive strategy and to try to improve team and player performance. There is still very little data from volleyball matches that are publicly available. However, as evidenced by the increased adoption of analytics in other sports, including baseball and football, it would most likely be beneficial to use data analytics in volleyball in order to drive decision-making and to derive insights that can be used to better player performance. This project will analyze a volleyball team’s past performance and derive observations that can help shape strategy and training in the future.

For this project, I found a dataset that has all the play-by-play data for Harvard University's 2018 season which was recorded using a software called Data Volley. A user macfields uploaded this data to a [Github repository](https://github.com/macfields/mfields_finalproject). To import and parse my data, I used a package called [Datavolley](https://github.com/raymondben/datavolley) which was created by Github user raymondben.

The dataset has data from 21 matches throughout the 2018 season. Each play has data regarding which player completed the action, what the action was, an evaluation of the action, start and end coordinates of the ball, and information about points, among other things. There are over 80 variables in the dataset, but some are repetitive. One limitation of this dataset is that it is only for Harvard’s team, which will mean that the insights may not be generalizable.

The output of this project is various statistics and data visualizations that can be analyzed in order to see what types of factors led to better outcomes in the games. Most of the data visualizations will make use of the court location data, such as the ending x and y coordinates of the ball during a play.

# Related Work

I was unable to find many public projects making use of data from volleyball games. The Github user, [macfields](https://github.com/macfields/mfields_finalproject), who uploaded the dataset that I will be using for this project did some analysis using R Shiny, but most of the work focused on match-level analysis[[1]](#footnote-1), whereas my work will look at broader trends throughout the season. Additionally, when doing background research I found a single academic paper looking at using data analytics in volleyball. The paper written by Sarah Almujahed, Nicole Ongor, John Tigmo, and Navjot Sagoo created a Monte Carlo Simulation to analyze how random variations in transition probabilities during collegiate volleyball games, caused by extreme conditional scenarios can affect the team performance[[2]](#footnote-2), and they found that the best way to improve performance was to increase blocks. While this analysis was very interesting, the model would probably benefit by being more dynamic and easily explainable to a coach, which would possibly increase its usage.

# Methods

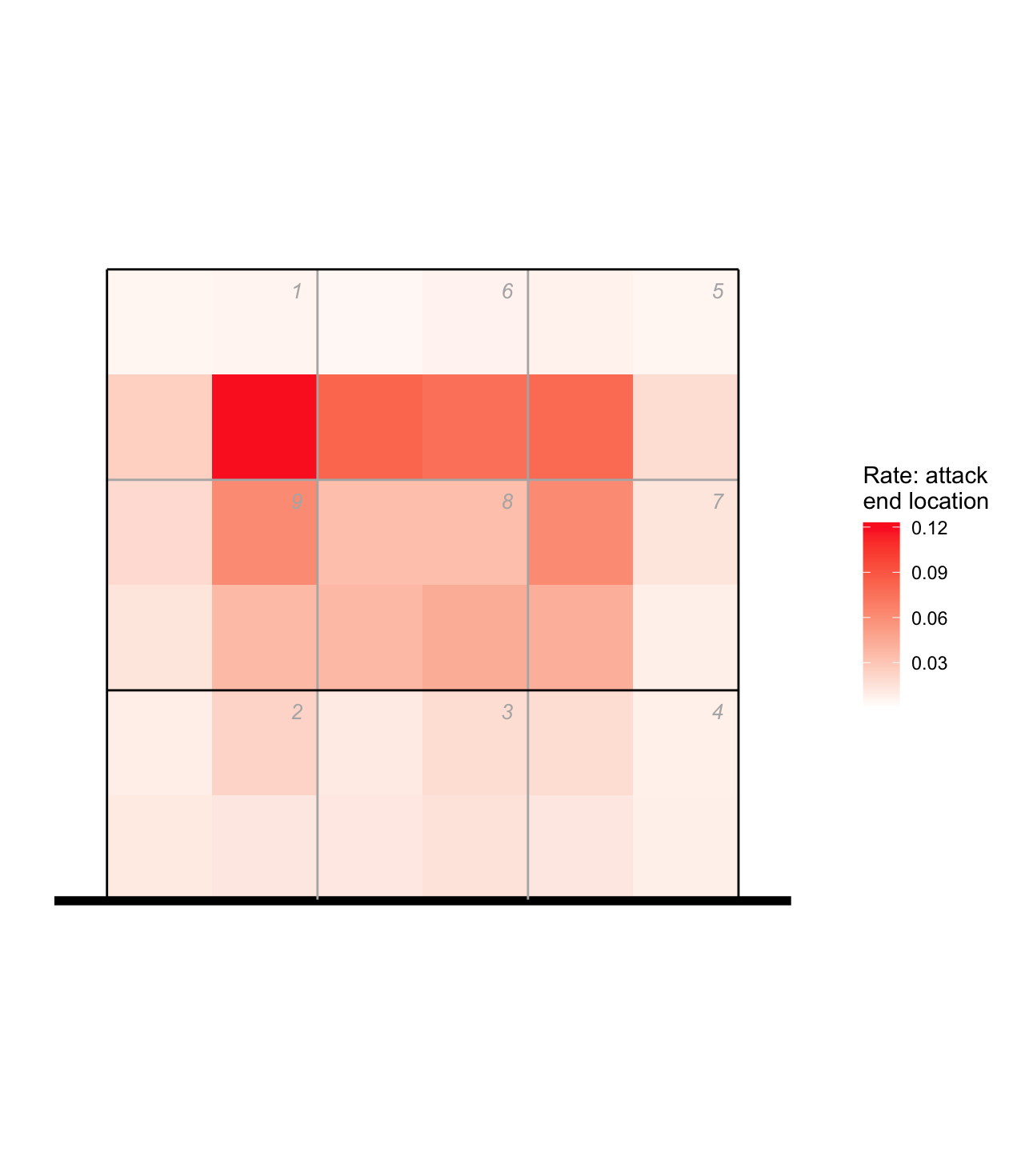
The first step of my project was to clean the data. Since the data was uploaded to Github as separate files for each match, I loaded the data into R and then combined all the data into a single dataset. I also created a dataframe that only had plays performed by the Harvard volleyball team players.

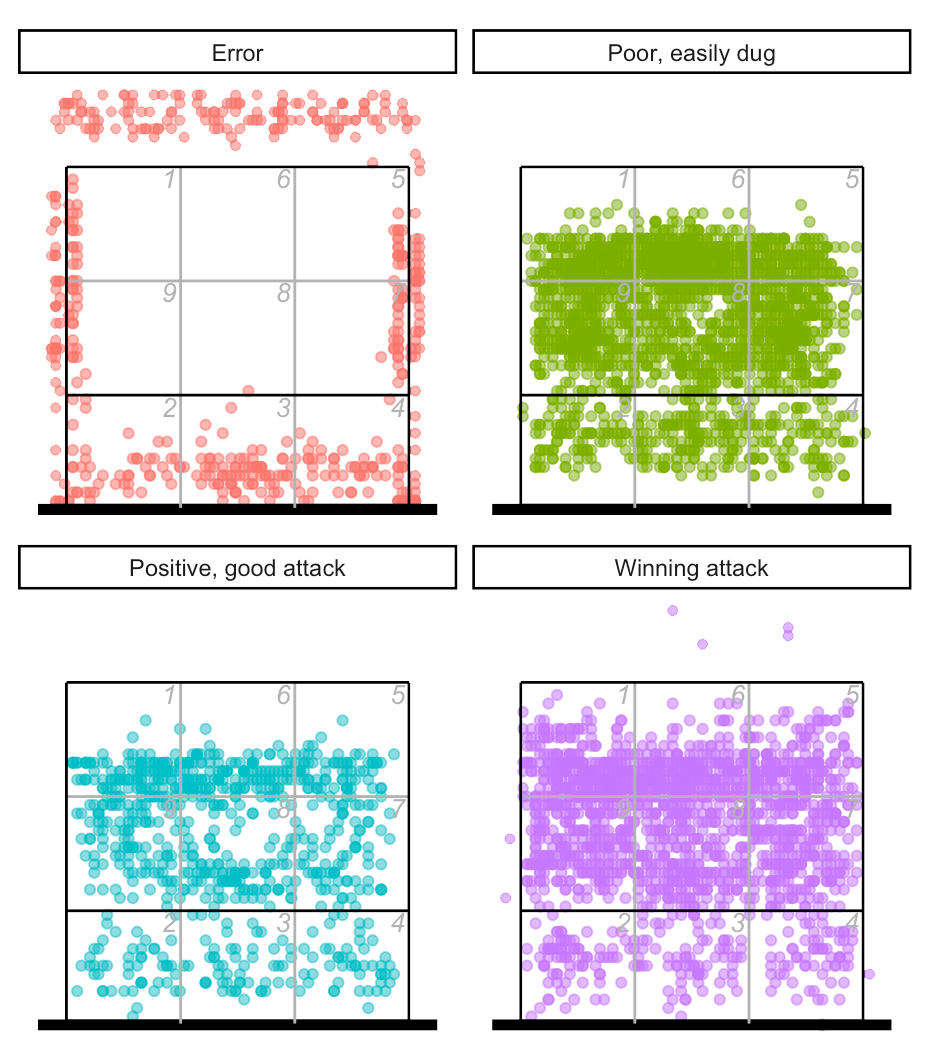
Some of the calculations that I included in my analysis were kill percentage and defensive errors percentage. Kill percentage is the ratio of the number of kills, or hits that land inside the court without a defensive player touching the ball or a hit where the defensive player commits an error, divided by the total number of hits. This percentage can be calculated per player or per end zone. Calculating the kill percentage for individual players will give an indicator of how effective that player’s hits are, with a higher percentage meaning that more of the player’s hits are evaluated as a kill. Calculating kill percentage per area can indicate which areas of the court a hitter should try to aim for and which areas have been more successful in the past. Likewise, defensive errors percentage can be calculated both for an individual and for a court zone. Defensive error percentage is the number of defensive errors, which are when a player’s pass cannot be passed by another player during serve-receive or while digging a hit, divided by the total number of defensive plays. The defensive error percentage for an individual indicates how good of a passer the player is, with a lower percentage indicating that they make less mistakes. For court zones, a high defensive error percentage could mean that the team needs to work on improving their passes in that area or that the position of the defensive players on the court is inefficient.

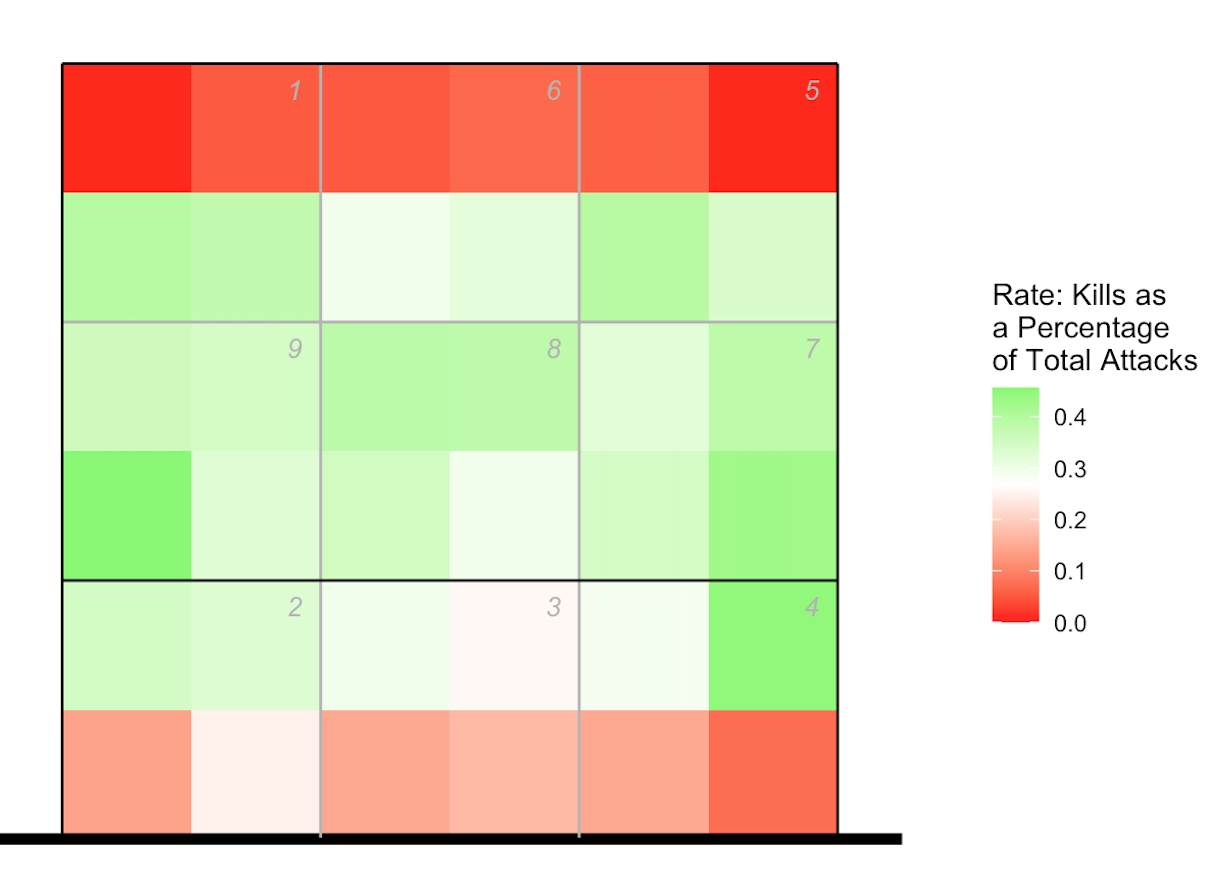
These statistics and others were then used for data visualizations, which can be easier to understand intuitively than these statistics alone.

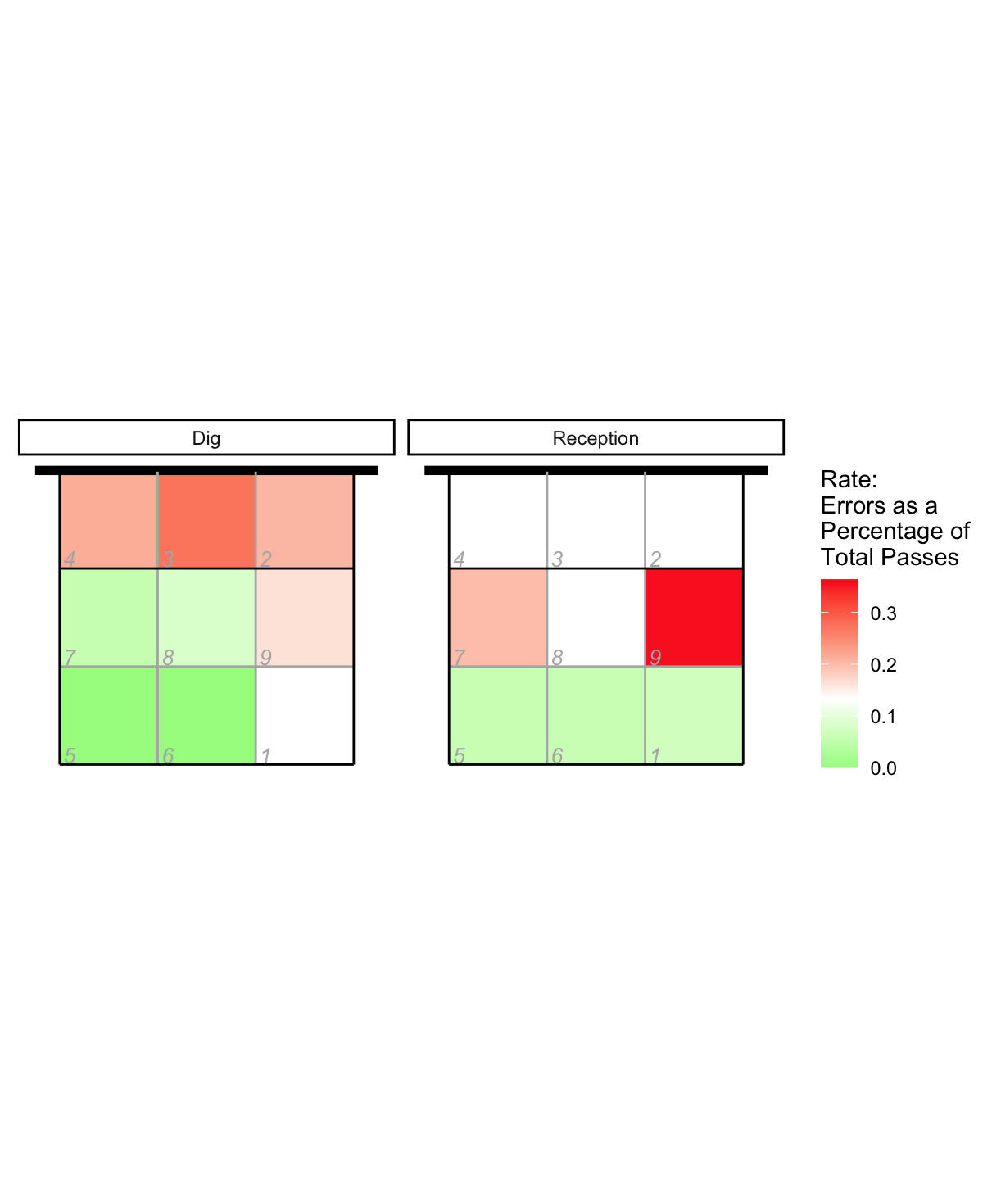
# 

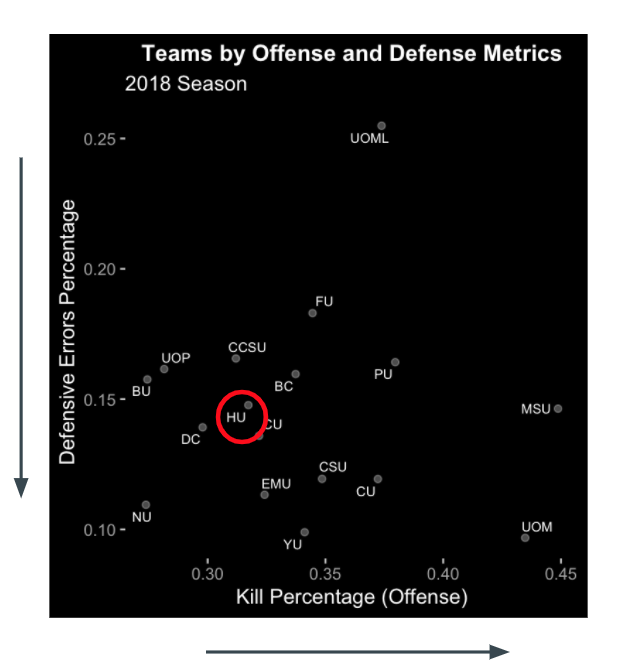
# Discussion

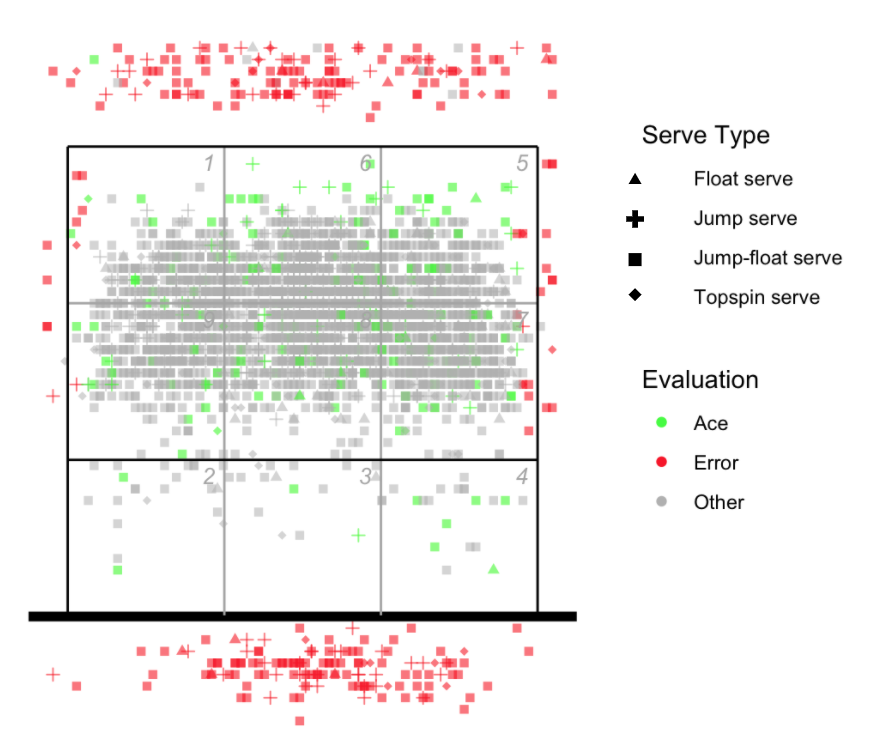
For my project I wanted to look at hitting, passing, and serving. First to explore where the ball was hit most often, I created a heat map of the court. The thick horizontal line represents the net, and darker red areas indicate . Looking at the graph, it looks like the most common area to hit toward is the back left and right corners, which aligns with traditional volleyball strategy. Surprisingly though, the ball is hit to the left back corner more often than the right, possibly suggesting that the team’s outside hitter often hits down the line or that the opposite hits often.

Next, I wanted to see which areas of the court resulted in better attack outcomes. I created 4 separate scatter plots based on the evaluation of each attack. Not surprisingly, most attacks that were considered errors were hit out of bounds or into the net. For attacks that were considered poor or positive, hits were heavily concentrated along the back three zones, but there was no huge discernable differences between the two plots, suggesting that additional factors, like the speed of the ball or attack type, may play a larger role in determining if the attack is either poor or positive. Then looking at the attacks that were winning attacks, compared to the previous two graphs, there are a lot more balls hit closer to the end lines at both the back and sides. This suggests that hitting the ball close to the line is more likely to result in a winning attack, which is consistent with conventional volleyball knowledge. 

The next thing that I looked at for hitting was kill percentage by area, which was calculated using the formula mentioned in the Methods section. I created a heat map, where the red represents areas with a low kill percentage and green represents areas with a high kill percentage. This graph was slightly surprising because it shows that the back line has a very low kill percentage, meaning that hits to that area rarely result in kills. This is most likely because since it is near the endline, hits can easily go out, which would result in an error. This is also the case for attacks ending near the net, which can be the result of a block or hitting the ball into the net. Looking at the graph, the areas with the highest kill percentage were the middle sides. Comparing this to the previous scatterplot, it looks like it might be beneficial for players to work on aiming toward the middle side areas since there were not as many hits to that area before.

I then wanted to take a look at which areas the team is better at defending. I created a visualization of the court where red represents areas where passers more frequently commit errors and green represents areas where passers commit errors less frequently. The graphs are separated by whether the player is passing an attack or a serve. Looking at the graphs, the areas with higher error rates are slightly different for digging and serve reception. However, for both, the back line near areas 1, 5, and 6 have lower error percentages, meaning that the defensive players in the back row are doing a good job of passing. For digging, the area with the highest error percentage is the front of the court, specifically in the middle, likely because middle hitters tend to be less skilled at passing. For reception, the areas with the highest percentage are the middle sides, again where middles would likely be standing during serve receive. 

Then to see where all of the teams stand in regards to offense and defense, I created a scatter plot with their kill percentage and defensive error percentage on the axes. The Harvard volleyball team fell right in the middle for both, which aligns with their season record of winning 40 out of their 83 sets. However, based on Harvard’s position on the chart, I would still recommend that the team focus more on hitting and passing because there are a lot of other teams doing much better in terms of kill percentage.

The last skill that I wanted to examine was serving. I created a scatter plot where the green points represent an ace, the red dots represent a serving error, and the gray dots represent that the receiving team was able to pass the ball. The shape of the point also corresponds to the type of serve. Looking at the graph, errors are most common along the back line as they were served out and for being served into the net. There are also some errors along the sidelines. The serves classified as aces and other are pretty evenly distributed along the back ⅔ of the court, suggesting that there are likely additional factors including defensive player skill and serve speed that may more heavily affect the serve outcome. However, looking closer to the backline and in the front ⅓ of the court, it appears that there is a slightly higher amount of aces, suggesting that those specific areas might be good areas for servers to aim for.

# Conclusion and Future Work

Overall, this type of graphical analysis might be beneficial for teams to refine their strategy. While some of the data visualizations show that there is likely correlation between court area and play outcome, there are likely additional factors impacting the situation including player skill, team communication, and ball speed. This analysis can help show which areas of the court are more effective to aim for during serving and attacks and can expose areas which need improvement.

For future work for volleyball analytics, it would be beneficial to collect additional variables such as player position on the court, whether the setter is back row or front row, and ball speed and rotation. The importance and magnitude of these variables’ effects on the game could be compared to help coaches analyze which areas to focus on. Using these additional factors, teams would be able to get a more comprehensive understanding of how to improve their performance.

1. Maclaine Fields, “Harvard Volleyball 2018 Analytics”. 13 Dec 2018. [↑](#footnote-ref-1)
2. S. Almujahed, N. Ongor, J. Tigmo and N. Sagoo, "Sports analytics: Designing a volleyball game analysis decision-support tool using big data," 2013 IEEE Systems and Information Engineering Design Symposium, Charlottesville, VA, 2013, pp. 19-24, doi: 10.1109/SIEDS.2013.6549487. [↑](#footnote-ref-2)