**Written Report:** March Madness Analysis ([Link to App](https://mjmasl01.shinyapps.io/data334_final_project_March_Madness/)) 5/10/24

In my shiny application, I delved into data from the 2024 March Madness tournament, examining teams’ offensive and defensive statistics alongside their win probabilities against Division 1 teams. Additionally, I analyzed another dataset spanning from 1985 to 2019, detailing tournament winners and losers for each playoff round, including score results over the years. My analysis revealed that in the 2024 March Madness tournament, teams with higher two-point percentages and lower defensive two-point percentages tended to achieve more success. I also observed that in the first round of playoffs from 1985 to 2019, most games had score differences below 20 points, with only one game exceeding a 50-point difference. The subsequent rounds generally featured score differences below 30 points, except for the Final Four, where one game in 2016 saw a 44-point difference, with Villanova defeating Oklahoma 95 to 51. Furthermore, only one winning team in these years scored over 100 points in the Final game.

The data being used in my shiny application is on this past March Madness tournament, 2024 playoff, along with tournament results from each playoff round starting in the year 1985, when the league officially switched over to 64 team format, all the way to 2019, where it contains each results of every playoff round, including the winning/losing team, winner/loser score, along with a calculated difference in the two scores, to look at score variation among between the competing teams.

For the March Madness ’24 dataset, I am interested in exploring a comparison table where you could stack multiple teams on top of each other and compare defensive or offensive stats to see which teams are expected to put up more points or prevent points from being scored on the defensive end. More specifically, this dataset will contain the team’s rank in the tournament, their playoff seed, team name, the conference they play in, and number of games played (an indication of how far the team made it). Adjusted offensive and defensive efficiency is the number of points allowed/scored per 100 possessions against average Division One offense/defense. Barthag is a power Rating measured by a team’s chance to beat an average Division I team. EFG% and EFGD% represent the effective field goal percentage scored/allowed—offensive and defensive rebound rate, measured as a percentage. FTR and FTRD represent the free throw rate on offense and defense and are measured by free throws attempted per field goal. There is an offensive and defensive measurement for two- and three-point percentages, one for how many scored and the other for points allowed. Adjusted tempo is an estimate of tempo (determined by possessions per 40 minutes) against a team playing at an average Division One tempo. Finally, wins above the bubble (WAB) mark the cut-off between making the tournament or not. The primary purpose of using this data is to explore what factors impact teams, specifically to recognize what makes a team more successful when playing in the tournament.

For the data covering tournament results from 1985 to 2019, I am interested in finding any trends in the winning/losing scores as well as differences in the scores between compete, as well finding any score records in the tournament, particularly in the fifth and sixth round of playoffs(final two rounds). This dataset will contain the year, playoff round number, playoff seed number for both teams one and two, scores for both teams, and names of both teams differentiated between one and two, then mutate the data to make a column for teams with winning and losing scores, names for each game, and lastly a column representing the difference between the score of the winning and losing team. I will use my shiny application, a user-friendly and intuitive platform, to produce plots of each  playoff round. Users can easily adjust the round, year, and region filters for desired outputs, empowering them to conduct in-depth analyses and explore the data. This will allow the user to easily identify any games that seems to be an outlier, which is evidence of a high-score record, if the point is found on higher y-values.

Using the application, I will look more in-depth at the data, viewing the final four teams from the 2024 March Madness data: Connecticut, Alabama, Purdue, and North Carolina State. More specifically, I will look at how their offensive and defensive stats compare to one another and see if any inferences can be made based on the tournament's result.

A screenshot of a sports results

Description automatically generated

Figure i - Offensive Statistics

Starting with the base statistics, Connecticut and Purdue ranked one seed, Alabama was fourth, and NC State ranked 11th. NC State is considered the underdogs as they are seated significantly higher than the rest, which would most likely result of an upset. In addition to the playoff seed, the team's Barthag rating would compare similarly where Purdue and Connecticut are 0.97ish while Alabama is 0.92 and NC State 0.84, which assesses the team's probability of beating a division one team. In the 2024 tournament, Connecticut beat Alabama in the final four, and this can be supported by the fact that Connecticut has a higher adjusted offense efficiency, two-point percentage, and offensive rebound, implying that Connecticut has the overall stronger offense compared to Alabama. Like that game, Purdue and NC State had a similar result, leaving the two one-seeds to face in the championship. While Purdue may have the higher offensive rebound rate and better percentage from outside of the arc, Connecticut still has the more robust offensive efficiency and the better two-point percentage, which is the more common basket to be made in a game. In the end, Connecticut beat Purdue and stayed in the lead the whole time, which shows that their offensive statistics could prove to be a more promising combination. Furthermore, looking at each team's wins above the bubble, one can conclude that Connecticut, the champions, started the tournament with the best chances of making the cut. At the same time, NC State originally had no chance of making it, but it became the underdog instead. However, to fully grasp how a team might have been more successful than others, we will also have to look at the defensive half of the game.

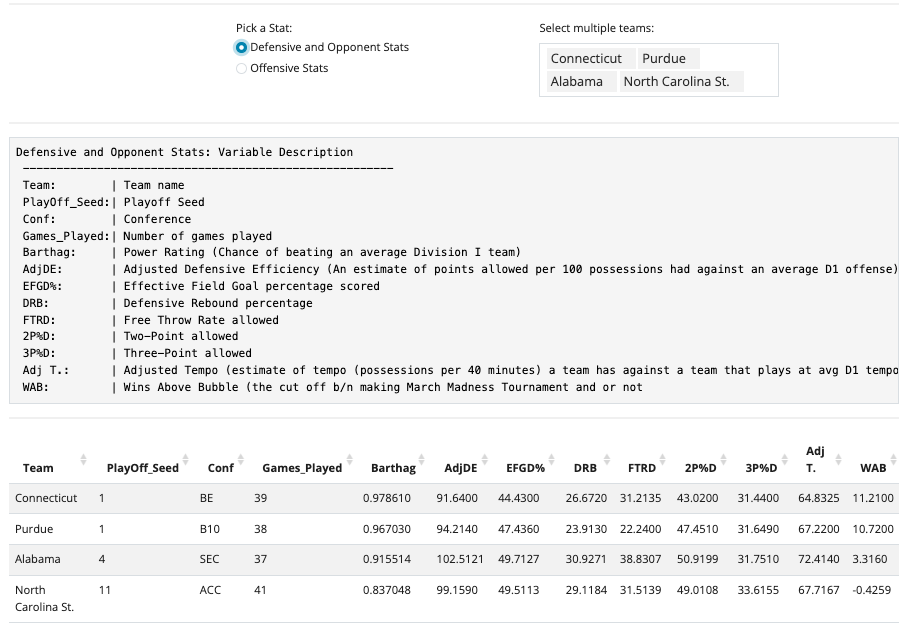


Figure ii - Defensive Statistics

      On the defensive side of the games, looking at the final four teams again, the Barthag rating, WAB's, and adjusted tempos will compare the same as seen in the offensive stats, however the rest will be different. Now, starting with the final four matchups of Connecticut versus Alabama, defensively, Alabama is stronger based on their adjusted defense efficiency, where they got 102.5, and Connecticut got 91.6, meaning Alabama allowed roughly 103 points every 100 possessions. Also, Alabama has a higher allowed free-throw rate, which implies that they tended to foul more, causing free points for the other team. One can infer that Alabama is weaker on defense overall than Connecticut. This can be confirmed by the fact that Alabama's two-points-allowed percentage is more significant than Connecticut's, and since most points scored are two-pointers, it would make sense that Connecticut ended up winning that game. Looking at the other semi-final, North Carolina State has better adjusted defensive efficiency than Purdue, where NC State tends to allow more points scored than Purdue per every 100 possessions. Similarly to the other game, the allowed free-throw rate for NC State is much higher than Purdue's, likely because NC State could have a high rate as well. In all, we can conclude that Purdue was the overall better team. Looking at the championship game, Connecticut gave up fewer points per 100 possessions, and the allowed-two-point percentage is less compared to Purdue statistics; therefore, we can argue that Connecticut had the better defensive game, which could have contributed to their success and winning the 2024 tournament. Next, we will look at historical tournament data, specifically the scores of the championship game.

A graph showing the results of a historical tournament

Description automatically generated with medium confidence

Figure iii - 1985 to 2019 Championship Round Score Differences

Looking at the plot, where it has the years on the x-axis while the difference in the winning score versus the losing one. I chose the differences to captivate both the team's scores and is a good measure of whether the game was a blowout or a close one. The trend of the scores has several close games where the difference is less than ten points, starting in 1985 and going up until 1990, where UNLV beat Duke 103-73. It was one of the largest score differences for the championship games, a 30-point difference. It was also the first big blowout in the finals in 1985 when the league switched to a 64-team tournament bracket. Following 1990, it dips down below ten again, then back up to the second-highest score difference in 1992, where Duke beat Michigan 71-51, a 20-point score difference—from 1995 to 2019, the score difference range between zero-to-20-point. Furthermore, during that duration, the scores may seem to fluctuate up and down a bit. However, the scores seem to increase gradually towards a twenty-point score difference. Also, following the 2010 game, the scores declined slightly until 2016, indicating that those championship games were gradually becoming more competitive. While some fluctuations occur, most championship games remain close and competitive, with score differences typically less than ten points.

In future work, I would focus on developing predictive models using the identified factors from the historical March Madness dataset to forecast tournament scores. Additionally, I aim to incorporate more advanced dataset to enhance the depth of analysis for the team’s defensive and offensive statistics for overtime, then using predictive models again to forecast the probability for future success. For these predictive models, I would incorporate time series analysis to accomplish this.