Christopher Bussen
3/1/2024

Midterm Project Report

Dataset Selection and Preparation

Due to my deep interest in sports, I picked a dataset from Kaggle about college basketball statistics. From this dataset, I used two different csv files, one for the 2023 NCAA men's basketball season and another that combines data from the 2013-2019 and 2021-2023 seasons (leaves out 2020 due to COVID-19). These datasets contain all Division I teams, each with the following 23 variables: team, conference, games played, games won, adjusted offensive efficiency, adjusted defensive efficiency, power rating, effective field goal percentage, effective field goal percentage allowed, turnover rate, steal rate, offensive rebound rate, offensive rebound rate allowed, free throw rate, free throw rate allowed, two-point shooting percentage, two-point shooting percentage allowed, three-point shooting percentage, three-point shooting percentage allowed, adjusted tempo, wins above bubble, postseason finish, and tournament seed (the combined dataset also contains year).

The original author scraped this data from https://barttorvik.com/trank.php# and then added the postseason, seed, and year columns. When I selected this data, I knew I would want to create a heatmap which meant I would need each team's location. I then created my own web scraper using Python in Google Colab to add a location column by finding each team's state from https://www.ncsasports.org/mens-basketball/division-1-colleges. Despite this being my first time writing a web scraper, it worked quite well as it filled in the location for over 80% of the teams (the other 20% was due to inconsistencies in team naming between the two sources).

Data Exploration Findings

Despite having a good base understanding of the dataset and an idea of potential trends from the beginning, I began by conducting some initial exploration into the data by looking through the values as well as creating visualizations to get a better feel for the data. I wanted to make sure I fully understood each variable so I could form hypotheses about potential relationships, so I researched BARTHAG (power rating) and WAB (wins above bubble). From this research I learned power rating is the projected win percentage against an average team on a neutral court and WAB is how many more wins a team has compared to what a bubble team (a team on the verge of making or not making the tournament) would be expected to have with the same schedule. I also identified what I believed to be the most important variables in the datasets - wins, conference, adjusted offensive and defensive efficiency, power rating, three-point percentage, wins above the bubble, seed, and postseason. After this, I created several scatterplots to see if any variables might be strongly correlated with wins. These visualizations confirmed what I expected. Of the in-game related statistics, adjusted offensive/defensive efficiency and effective field goal percentage seem to have the strongest correlation with wins whereas the rest of the variables do not appear to have a particularly strong correlation to wins. Because more wins typically leads to a better seed in March Madness and often a better finishing position, this initial exploration made me curious about the potential relationships between efficiency on both sides of the basketball and how teams perform in the tournament. During data exploration, I also quickly realized that power rating and WAB would typically have an important relationship with wins, seed, and postseason finish as they provide a high-level overview of how good teams are compared to others. Overall, it appeared that higher WAB and power rating usually indicated more wins, higher seeding, and a better March Madness finishing position. However, conferences tend to complicate things and throw off these trends a bit. This is because some conferences are filled with strong teams and get several tournament bids while other conferences have much weaker teams and only get one bid (each conference gets one automatic bid for the

winner). For example, the B10 can get 8 bids and have a team with 17 wins get a bid because they had a very difficult schedule, and therefore has a high power rating and WAB despite a lower win total. On the other hand, in the Ivy League usually only the conference winner gets into the tournament, normally with a win total in the mid-twenties, but will have a low power rating and WAB due to playing weaker teams throughout the season. These trends and relationships were all important to keep in mind when creating my visualizations and analyzing them.

Visualization Insights and Analysis

In this section, I will go figure by figure and provide a brief analysis of the visualizations and the insights they provide. Overall, I decided to mostly use a sequential orange color scheme because most of the data I'm emphasizing through the use of color is ordinal, interval, or ratio data that does not have a critical midpoint to emphasize deviation from. However, when I use color to emphasize WAB, I used a diverging color scheme with orange and blue because it is important to be able to tell whether WAB is positive or negative when interpreting its value (Dr. Chen recommended getting rid of the color for this figure since it fits the trend of the bars but I decided to leave it in because I wanted to emphasize that the pattern is the same since that was one of my hypotheses). Lastly, when trying to make clear separations between conferences or clearly show teams within each conference, I used qualitative color schemes. I will talk more about the specific color scheme decisions in the next section which will also go over accessibility, but I tried to revolve my color schemes around orange because I felt that between the rims, hardwood floors, and the ball itself, basketball is closely connected with orange and would allow audiences to easily connect with the data.

Figure 1 shows a heatmap of the average wins for each state for using all teams in 2023. When hovering over a state, it will also display the number of teams in the state, average wins, minimum number of wins, maximum number of wins, and average tournament seed. It is especially important to consider the number of teams when looking at what state might have the strongest teams. For example, Vermont has the highest average wins at 23, but this is because it only has one team (and we can see it was a 15 seed). We also see that larger states typically have a very wide range of win values, such as California which ranges from 3 to 32 wins between 26 teams. Overall, the heatmap indicates that Washington, Utah, Arizona, Kansas, and Ohio are the most successful states in the 2023 season. I chose this type of visualization because many people often like to debate which states have the best college basketball, and this allowed me to combine many aspects of that argument into one map. For ease of reference, I also included a table of descending wins by state in the following sheet.

Figure 2 displays how many tournament teams each of the 32 conferences had in the 2023 season while also displaying the average power rating of all teams in each conference using color as well as the average tournament seed when hovering over each conference. From this visualization, we see that there are nine conferences that received multiple bids, including five with over five bids. The B10 and SEC were tied with the most bids at eight while also having some of the highest power rating values. B12 had the second most bids at 7 but had the highest average power rating at just over 0.87 (meaning that on average, B12 teams have an 87% chance of beating an average team on a neutral court). B12 also has the best average seed of conferences with more than two bids, at 4.29. Lastly, we notice that all multi-bid conferences have an average power rating of at least 0.65 and an average seed below 10. On the contrary, no single-bid conference has an average power rating above 0.6 with most well below 0.5 and all but one

getting seeds higher than 10. I chose a treemap because it was able to easily show the conference hierarchy between the 68 tournament teams.

Figure 3 shows how many tournament appearances in the last 10 years of March Madness each team has and groups them by conference using different colors to make it easier to digest and separate each conference. This visual shows us that only three teams have made the tournament all ten years - Michigan St., Kansas, and Gonzaga. We can also see that most of the more well-known conferences have had over 10 different teams with tournament appearances while several smaller conferences have as little as three teams with appearances, indicating a lack of competition. I chose a side-by-side bar chart because it allowed me to make a chart for each conference and then easily separate them while still having them in the same graphic.

Figure 4 tests the common saying “defense wins championships” by showing the distribution of defensive efficiency for each category of postseason finish. While it is not a direct causation, the figure indicates that teams that go further in March Madness typically have a better defensive efficiency (lower is better). Teams that do not make the tournament (null) have the largest variation in defensive efficiency and easily the highest median. As teams advance, we notice that the median defensive efficiency continues to decrease and the variation gets smaller (which is also partially due to there being less teams). The champions have the lowest median and eight of the ten values are smaller than the next smallest median, which comes from the second place team. I decided to do boxplots here because I wanted to be able to show the distribution of ADJDE across multiple categories.

Figure 5 looks at how team efficiency on both offense and defense relates to how teams have done in the postseason over the last ten years by using color to indicate how far a team goes in March Madness (excluding teams that did not make the tournament). Generally, teams with a lower offensive efficiency and higher defensive efficiency (again, higher is worse for defense) are eliminated in the round of 68 or 64 - this can be seen in the upper left part of the scatterplot. As we move into the middle of the plot, there is a large amount of variation and it is harder to make out trends. However, as we move to the bottom right of the plot where teams have higher offensive efficiency and lower defensive efficiency we see that many of them fall somewhere in the Elite 8 to champion range, which we would expect. It is also expected that there are some outliers due to upsets. I opted for a scatter plot because I knew that I was plotting two continuous variables against each other and wanted to show the relationship between while using color to display another dimension.

Figure 6 shows us the average seed for tournament teams in each conference throughout the last ten years while also using a diverging color scheme to display the average WAB values (once again, I decided to leave the color scheme for WAB in to emphasize this pattern that I was originally curious about). The graph shows that there are three conferences who are almost always 16 seeds (NEC, SWAC, MEAC) whereas the top seeded conference on average is the B12 as a 5 seed. Average WAB ranges from about -9 to 5, and as expected, it generally increases as the average seed of the conference increases. This makes sense because we would expect higher seeded teams to perform better than teams that are on the bubble. I decided to use a bar chart here because it is a simple way to show a numerical value for each category and allows for each bar to be a different color to provide the average WAB as well.

Figure 7 provides the total number of 2023 season wins that each conference had from teams that made the tournament. It also shows how many wins each individual team that made the tournament within each conference had. We can see that the SEC, B10, and B12 easily had the most wins among tournament teams, with each conference being over 160 total wins. On the

opposite end, the SWAC only had 14 wins from tournament teams (indicating that one of the more average teams won the automatic bid) and the OVC was also under 20 wins. It is also worth pointing out that of conferences with only one team to make the tournament, total wins ranges from 14 to 35. I used a stacked bar graph so that I could show each conference while also showing how each conference got to its total win count.

Lastly, Figure 8 simply shows the distribution of seed for each category of postseason finish. We see that the median seed shrinks round by round and the range gets notably smaller starting with the Final 4. The median seed of champions is 3 while the highest seed to win in the last ten years is a 7 seed. I chose to use a boxplot for this figure because it adequately shows the distribution of seed across the eight different categories.

Accessibility Reflection

Overall, I did not think that the process of making my visualizations accessible was particularly difficult. While I did consider using tools like Color Oracle and Color Brewer, I ended up primarily using the premade Tableau color blind color palette which I have heard works quite well for the most common types of CVD. I used this color palette for Figure 3 to differentiate each conference in the side-by-side bar chart as well as for Figure 7 to show each individual team within the stacked bar chart. Additionally, when I decided to use orange as my primary color for these visualizations, I did some research to make sure that it is generally a CVD friendly color. I came to the conclusion that it works well for most types of CVD, which helped me decide to stick with sequential orange color schemes. During this research, I also discovered that along with orange, the combination of orange and blue is CVD friendly, which is why I selected the diverging color scheme used in Figure 6 to show the average WAB values for each conference. Overall, Tableau having a premade CVD friendly palette helped me avoid many challenges with making my visualizations accessible. The only challenges I ran into were that it originally took me some time to find this color palette and in Figure 7, due to the order of the original data and the color palette, there were two instances of teams stacked on top of each other having the same color. This originally made it look like it was only one team but I addressed this issue by adding a label for each team name and adding borders to each component of the stacked bars.