**Ella Heath**

**Data Wrangling**

**Michael Redmond**

**9 May 2025**

**March Madness Seed Analysis**

**1. Introduction**

March Madness is a yearly 3-week long single elimination tournament in NCAA Division 1 Men’s and Women’s basketball. The tournament consists of 68 teams that compete in 7 rounds to become national champions. The March Madness tournament has taken place every year (except the 2019-2020 season due to Covid-19) since 1939. There are two ways teams can participate in the March Madness tournament. First, if a team wins their division postseason conference tournament they receive an automatic bid into the tournament; there are 31 Division 1 conferences, so there are 31 automatic bids. The remaining and second way a team can play in the tournament if by receiving an at-large bid through an invitation; a selection committee on “Selection Sunday” decides 37 teams who were not automatic qualifiers based on stats and rankings to decide who will receive an invitation. Once the 68 teams are selected, four of these teams are eliminated in the opening round (Play-In or First Four) of the tournament which leaves the tournament with 64 teams. These teams are given a seed in one of four regions; each region has a 1-16 seed decided by the Selection Committee based on regular season and conference tournament performance (Wilco, 2025).

The tournament then starts with the top seeded team (No. 1) playing the lowest seeded team (No.16). This continues with the next highest team playing the next lowest team, leaving team No.8 playing team No.9. Single elimination leaves 6 rounds (Round of 64, Round of 32, Sweet 16, Elite Eight, Final Four, National Championship) for a team to be considered the national champion. March Madness if a highly anticipated sports event where millions of people create brackets in hopes of creating the perfect bracket (predicting the winner of each of the 68 games correctly). The chances of predicting a perfect bracket if you have some prior knowledge of basketball is about 120.2 billion. However, a perfect bracket has never been made, the longest perfect bracket was in 2019 when an individual correctly predicted 49 games until the 50th game after an upset (No.3 Purdue beats No.2 Tennessee). An upset in March Madness happens when a lower seeded team beats a higher seeded team (Benzie, 2025). Because of this, March Madness fans have become obsessed with trying to create a perfect bracket and correctly predicting upsets based on teams seedings going into the tournament (Wilco, 2025).

**2. Data**

This project uses two main sources of data: a comprehensive dataset of March Madness game results from 1985–2024 and ESPN/NCAA seed-based performance data.

*2.1 March Madness Game Results*

The primary data set used in this project is a combined dataset of March Madness games played from 1985-2024. The dataset tracks information about each game like the winning and losing team, their seed, the round the game was played, and the number of points scored. The data was sourced from a public GitHub repository:

<https://github.com/shoenot/march-madness-games-csv/blob/main/csv/combined.csv>

From this dataset the main cleaning and processing performed was:

* Dropping all rows with missing values
* Removing inconsistencies in formatting (e.g. team names with ‘#’)

From this dataset, a dataset containing only games that were considered an upset were derived called upsets\_df. This contains games where only a lower-seeded team beats a higher-seeded team. This allows for analysis focused on only upsets.

*2.2 Seed-Based Historical Performance Data*

The second major source of data was scraped using a python script from ESPN/NCAA’s seed performance summary page:

<https://www.ncaa.com/news/basketball-men/article/2025-04-16/records-every-seed-march-madness-1985-2025>

From this site, two tables were extracted:

* Seed Records: this table includes each seed’s overall win-loss records, best finish in the tournament, and the number of times they’ve reached that best finish.
* Seed Matchup Statistics: This table includes win-loss records and percentages for each seed vs. Seed pairing that takes place in the first round.

Cleaning for seed\_records\_df included:

* Turning “OVERALL RECORD” into separate “WINS” and “LOSSES” columns.
* Turning “BEST FINISH” into separate “BEST FINSIH TYPE” and “BEST FINSIH COUNT” columns.
* Text conversion (e.g. Champions to Champion) and number to integers (e.g. “one” to 1)
* Turning cleaned columns into integers and floats to be used for analysis

Cleaning for seed\_pct\_df included:

* Removing asterisk from one matchup where the game was canceled due to the COVID-19 pandemic.
* Turning the “SEED VS SEED” into separate “Seed\_1” and “Seed\_2” columns.
* Turning “W-L” into separate “Wins” and “Losses” columns.
* Turning cleaned columns into integers and floats to be used for analysis.

The final cleaned datasets used for analysis include:

* Project Proposal - Combined Games.csv (2,394 observations)
* combined\_upsets.csv (subset of games\_df)
* seed\_records.csv (16 records, one per seed)
* seed\_pct\_records.csv (matchups between 1–16 seeds)

These datasets formed a foundation for analysis, performance comparisons, and visualizations throughout the project.

**Data Dictionary**

*Table 1 Data Dictionary for March Madness Games, 1985-2024, games\_df*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| year | Numeric | The year the game was played |
| round\_of | Numeric | Round the game was played |
| winning\_team\_name | Text | Name of the team that won the game |
| winning\_team\_seed | Numeric | Seed of the team that won the game |
| winning\_team\_score | Numeric | Score of the team that won the game |
| losing\_team\_name | Text | Name of the team that lost the game |
| losing\_team\_seed | Numeric | Seed of team that lost the game |
| losing\_team\_score | Numeric | Score of the team that lost the game |

*Table 2 Data Dictionary for upset March Madness Games, 1985-2024, upsets\_df (derived from games\_df)*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| year | Numeric | The year the game was played |
| round\_of | Numeric | Round the game was played |
| winning\_team\_name | Text | Name of the team that won the game |
| winning\_team\_seed | Numeric | Seed of the team that won the game |
| winning\_team\_score | Numeric | Score of the team that won the game |
| losing\_team\_name | Text | Name of the team that lost the game |
| losing\_team\_seed | Numeric | Seed of team that lost the game |
| losing\_team\_score | Numeric | Score of the team that lost the game |

*Table 3 Data Dictionary for Records of every seed in March Madness, 1985-2024, seed\_pct\_df*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| SEED VS. SEED | Text | Seed match-ups(1-16) |
| W-L | Numeric | The seed vs. seed options win-loss record |
| PCT. | Numeric | Seed vs. Seed win-loss percentage |
| Seed\_1 | Numeric | 1st seed in the matchup |
| Seed\_2 | Numeric | 2nd seed in the matchup |
| Wins | Numeric | Wins for Seed\_1 against Seed\_2 |
| Losses | Numeric | Losses for Seed\_1 against Seed\_2 |

*Table 4 Data Dictionary for Records of every seed in March Madness, 1985-2024, seed\_records\_df*

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| SEED | Text | Seed (1-16) |
| OVERALL RECORD | Text | That seeds overall record |
| BEST FINISH | Text | The best finish that seed has seen and how many times |
| WINS | Numeric | That seeds overall wins |
| LOSSES | Numeric | That seeds overall losses |
| BEST FINISH TYPE | Text | The seeds best type of finish |
| BEST FINISH COUNT | Numeric | The count of the seeds best finish |

**3. Analysis**

*3.1 First Round Upsets*

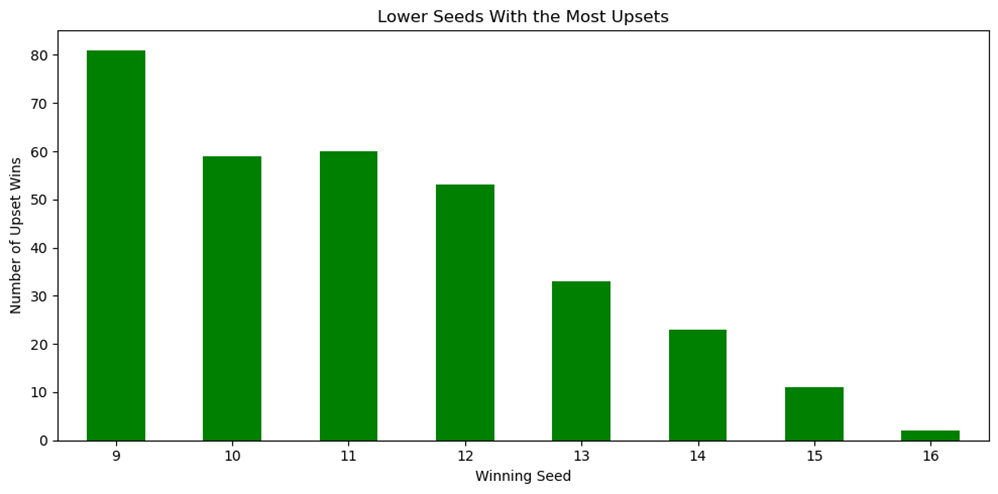
I wanted to analyze upset data from the round of 64 (first round) to identify which higher ranked seed is most likely to lose because of an upset. Seed 8 is most often upset, losing 81 times and seeds 6 and 7 losing around 60 times each. The average losing seed is 5.9, showing that middle seeds are more vulnerable to lose as result of an upset.

On the flip side, I wanted to analyze which lower ranked seeds are most likely to upset a higher ranked seed. Seed No. 9 is the most common upset winner with 81 wins, followed by seeds No. 11 and times, both upset around 60 times. The average winning seed is No. 11, showing that the middle tanked lower seeds are the most likely to pull off an upset.

Overall, most first-round upsets involve matchups between seeds 6-11, with very few upsets involving top or bottom seeds. Figure 1 displays higher seeds that are most often upset, and Figure 2 displays lower ranked seeds with the most upsets.



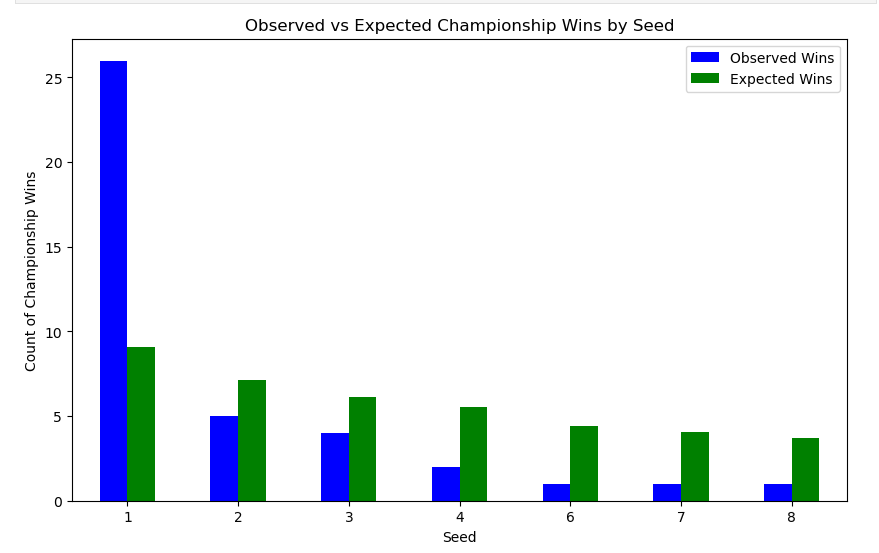
*Figure 1*



*Figure 2*

*3.2 Championship Success*

Next, I wanted to test whether certain seeds win the championship more often than expected based on how often they appear in the championship. To analyze this, I ran a chi-square goodness-of-fit test comparing observed championship wins to expected wins. The test revealed a p-value of almost 0 or less than an alpha level of 0.05. This led to the rejection of the null hypothesis. This result confirms that some seeds won the tournament championship more than expected. As shown in Figure 3 top seeds outperform their expected value and other seeds, winning the championship more than their appearance rate would predict. Figure 3 also shows that lower seeds underperform others, losing more championships than their appearance rate would suggest.

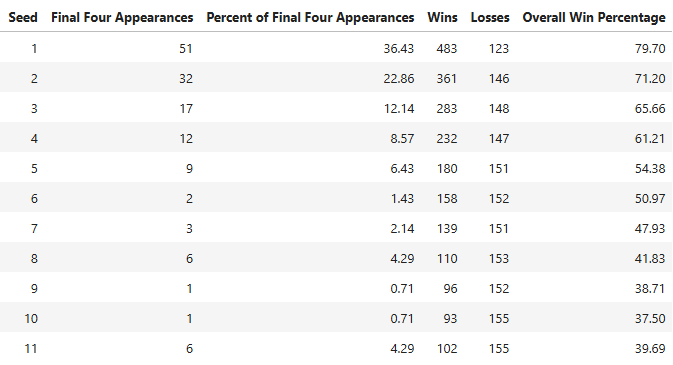


*Figure 3*

*3.3 Final Four Appearances and Win Rate*

Next, I wanted to examine whether a seed’s overall win percentage is related to its frequency of the most anticipated round of the tournament, the Final Four. I calculated each seed’s percentage of total Final Four appearances it accounts for and that seed’s overall win percentage. To see how strong or weak the relationship between the two variables is, I ran a Pearson Correlation test. This test showed a strong positive correlation and confirmed a statistically significant relationship (r = 0.895, p-value = 0.0002).

Overall, seeds that win more games overall are much more likely to reach the Final Four. Figure 4 shows this trend, with higher ranked seeds dominating both Final Four appearances and overall win percentage compared to lower ranked seeds, with seed 12-16 never having reached the Final Four round of the tournament.



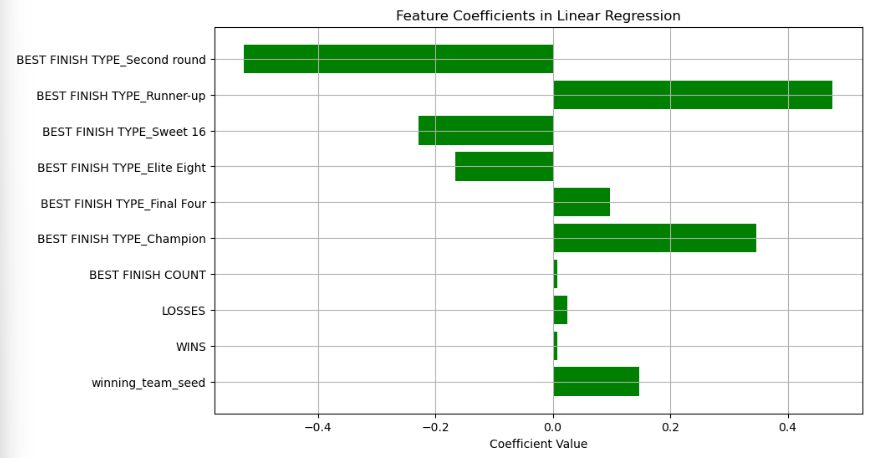
*Figure 4*

*3.4 Predicting Round Advancement*

To predict how many rounds a team will go in the March Madness tournament, I decided to utilize multiple linear regression using a team’s seed, wins, losses, best finish type, and best finish count. To be able to use the best finish type as a variable, I had to create 6 dummy variables. The multiple linear regression model had an R-Squared value of 0.25, meaning that 25% of variation can be explained by the model. The model also had a mean squared error of 1, meaning the model could be off by about 1 round.

The model suggests that Best Finish Type (Runner-up, Final Four, and Champion) and winning team seed positively impact how many rounds a team will go. For example, historically a No.1 seeded team will advance 3 rounds in the tournament (Sweet 16). I also utilized Ridge and Lasso regression to see if results were similar; linear regression yielded the highest R-squared and Ridge yielded the lowest MSE, however, results were very similar.

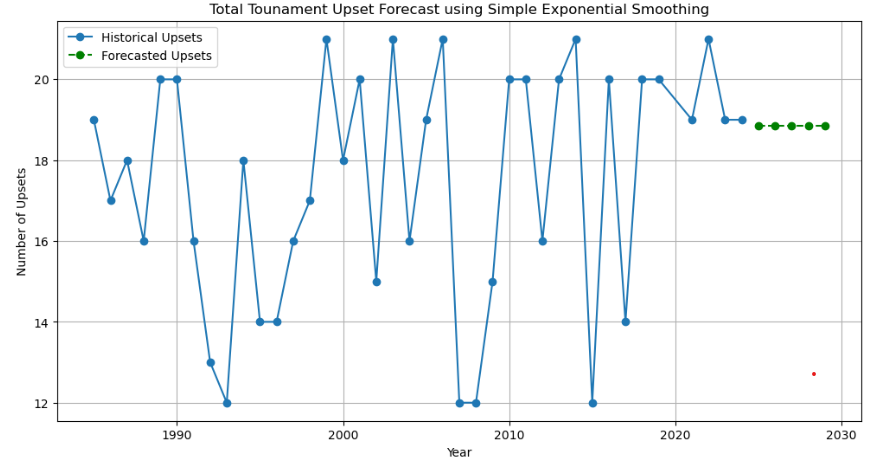
Figure 5 shows the feature coefficients of the multiple linear regression model, championship finishes, and seed had a strong positive effect, while a seed only making it to the second round negatively affects their tournament performance.



*Figure 5*

*3.5 Project Future Total Number of Upsets*

To forecast the total number of upsets from 2025-2029, I used simple exponential smoothing using historical upset data. The model forecasts approximately 19 upsets per tournament over the next 5 years. I analyzed multiple smoothing levels to find the model with the lowest mean absolute error. I found that a smoothing level of 0.2 yielded the smallest MAE of 2.262, meaning total upset predictions could increase or decrease by 2-3 upsets. Figure 6 visualizes historical data up total upsets and the forecast number of total upsets for the next 5 years. The results indicate that there will be a fairly consistent pattern of total upsets in upcoming tournaments.



*Figure 6*

**4. Conclusion**

In this project, I analyzed March Madness tournament outcomes and how a team’s seed affects their performance and the outcome of the tournament. In Summary, from the analysis questions presented in my proposal, I found the following results:

1. In the first round, is there a higher seed that is more likely to lose as a result of an upset/ is there a lower seed that’s more likely to win as a result of an upset?

Seed No. 8 has been upset the most, followed by Seeds 6 and 7. The seeds most likely to win as a result of an upset are No. 9, 11, and 10. These findings show that middle seeds are the most volatile in early matchups.

1. Do some seeds win significantly more than other seeds?

There is a statistically significant difference in how often seeds win the championship compared to how often they appear. A hypothesis test returned a p-value of 1.6e-07, leading to the rejection of the null hypothesis. This indicates that certain seeds are more likely to win when they reach the championship game.

1. Is there a relationship between a seed’s overall tournament win percentage and the percentage of total Final Four appearances that seed accounts for?

A strong positive correlation (r = 0.895) was found between a seed’s overall win percentage and their Final Four appearance rate. This suggests that seeds with higher win rates tend to go deeper in the tournament more consistently. This is statistically significant, there is a less than 1% chance the results are due to randomness.

1. Predict how many rounds a seed will go based on their wins, losses, best finish, and best finish count

I used multiple linear regression, ridge regression, and lasso regression models to predict the number of rounds a team advances. The linear regression model had the best performance, with an R-squared value of 0.252 and a mean squared error of 1.14, indicating that about 25% of the variation in rounds advanced can be explained by the model. Ridge regression produced a similar result (R-Squared = 0.252, mean squared error = 1.14), while lasso regression had slightly weaker performance (R-Squared = 0.246, mean squared error = 1.15). Although the models capture some trends, a large portion of variability remains unexplained, suggesting that factors beyond team history and seed also play a significant role in tournament outcomes. The regression equation revealed that the teams seed No.1 are predicted to 3 rounds.

1. Based off historical data, how many total upsets will there be in the next 5 tournaments (2025-2029)

Answer: there will be a predicted number of 19 upsets per tournament for the next 5 years.

Using a regression model trained on past tournament data, I forecasted the number of first-round upsets for each of the next five years. The model predicts an average of 18.86 upsets per tournament from 2025 through 2029. The model’s mean absolute error is 2.62, indicating reasonably consistent performance. These results suggest that the number of upsets will remain relatively stable, with fluctuations likely falling within a range of 3 more or 3 less upsets.

**5.Limitations**

This project has a few limitations, including the fact that I wasn’t able to create a fully . integrated dataset. Some game records had to be dropped due to missing values (NAs), which may have slightly impacted the results. Additionally, this analysis didn’t include other factors that could impact a seed’s performance like injuries or changes to the team. Future work on this project could include improving data integration by merging more detailed performance metrics, such as point differentials or team statistics.

**Works Cited**

“The Longest an NCAA Bracket Has Ever Stayed Perfect.” *NCAA.Com*, NCAA.com, 24 Mar. 2025, [www.ncaa.com/news/basketball-men/bracketiq/2025-03-23/longest-ncaa-bracket-has-ever-stayed-perfect](https://www.ncaa.com/news/basketball-men/bracketiq/2025-03-23/longest-ncaa-bracket-has-ever-stayed-perfect).

“Records for Every Seed in March Madness from 1985 to 2024.” *NCAA.Com*, NCAA.com, 5 Feb. 2025, [www.ncaa.com/news/basketball-men/article/2025-02-05/records-every-seed-march-madness-1985-2024](https://www.ncaa.com/news/basketball-men/article/2025-02-05/records-every-seed-march-madness-1985-2024).

Shoenot. “March-Madness-Games-CSV/CSV/Combined.Csv at Main · Shoenot/March-Madness-Games-CSV.” *GitHub*, github.com/shoenot/march-madness-games-csv/blob/main/csv/combined.csv. Accessed 26 Mar. 2025.

“We’re Tracking Every Upset in the NCAA Men’s Tournament.” *NCAA.Com*, NCAA.com, 20 Mar. 2024, [www.ncaa.com/live-updates/basketball-men/d1/mens-march-madness-upset-tracker](https://www.ncaa.com/live-updates/basketball-men/d1/mens-march-madness-upset-tracker).

“What Is March Madness: The NCAA Tournament Explained.” *NCAA.Com*, NCAA.com, 22 Jan. 2025, [www.ncaa.com/news/basketball-men/bracketiq/2025-01-22/what-march-madness-ncaa-tournament-explained](https://www.ncaa.com/news/basketball-men/bracketiq/2025-01-22/what-march-madness-ncaa-tournament-explained).