# Final Project for Team 07-7: Analyzing NCAA Women's Basketball Records

Amanda Cai, James Gao, Conor West

## Introduction

While the NCAA Division I Men's Basketball Tournament has drawn millions of spectators yearly since 1939, the emergence of a national tournament for Division I Women's Basketball is a much more recent addition. Started in 1982, the NCAA Division I Women's Basketball Tournament has the same basic structure as the Men's Tournament: 64 teams compete in a single-elimination bracket for a shot at the national title. There are a few key differences, most notably the lack of "play-in" games and there being only 32 "at-large" playoff berths awarded compared to 36 in the Men's Tournament.

However, the most exciting and statistically interesting feature of both the tournaments is the single-elimination style. Unlike, for example, the NBA Playoffs, which consist of multiple seven game series, every single match-up is a best-of-one game. This feature of the Tournament creates an environment where "upsets" are a common and often expected outcome of many games and where the seeding of the tournament can have a significant impact on the performance of the teams. Furthermore, this makes the actual selection of the seeds a critical process, with the NCAA's "Selection Sunday" becoming an important cultural event for many sports fans across the globe.

This leads us to our central research question: "What makes a good seed?" or, perhaps more accurately, "What factors should the selection committee of the NCAA Division I Women's Basketball Tournament consider most strongly when determining the seeding?" We will divide this question into five main areas of analysis. First, we will analyze conference records and conference placings as a predictor of future success, before then examining the role of regular season records and comparing the two. Third, we will draw distinctions between at-large and autobid teams, then draw conclusions about the strength of conferences in women's basketball by grouping their results together. Finally, we will analyze seeding, and see which seeds statistically overperform, underperform, and which schools have been historically overseeded or underseeded.

We hypothesize that the strongest predictors for future tournament success will be conference and regular season records. These are generally the best indicators of the relative strength of each team, especially because they allow us to remove potential human bias that a selection committee may have when assigning seeding by hand. Additionally, we hypothesize that between the two, regular season records will most likely be the best indicator of tournament success. In-conference play can often obscure the true strength of the teams within each conference, especially within conferences that may only have a few strong teams dominating the rest of the competition (the Big East comes to mind, with a large discrepancy between the top-5 and bottom-5 teams).

## Data Description

We obtained our data from FiveThirtyEight, who used it for their story "The Rise and Fall of Women's NCAA Tournament Dynasties." The statistics themselves come directly from the NCAA. Our data contains information on the individual seasons for every team participating in the NCAA Division I Women's Basketball Tournament for every year since 1982, although for the purposes of our analysis, we are only looking at years after and including 1994 (when the tournament expanded to 64 teams) since they best simulate the conditions of present and future tournaments, which our research question hopes to analyze. Some relevant variables in the data are the year, school, seed, and conference, information on conference performance (wins, losses, percentages, placement), regular season performance (wins, losses, percents), method of qualifying to the NCAA tournament, whether the first NCAA tournament game was at home or not, tournament wins, tournament losses, the ultimate tournament placement, and full win, loss, and win percentages for the season.

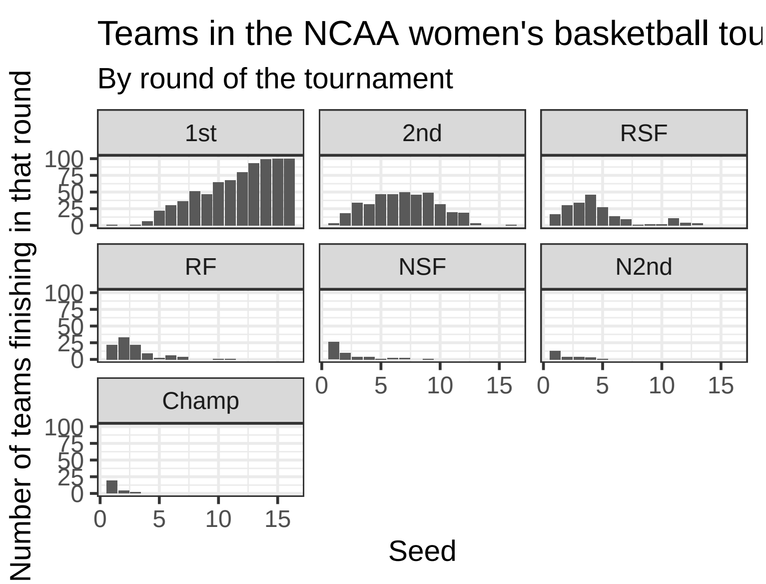
# Methodology

The biggest challenge in our data is the lack of flexibility in our response variable, tournament success (which will be represented by both tournament_finish and tournament_wins throughout our analysis, as both represent the same thing; we can glean one from the other.) As such, much of our analysis will be based on comparisons between rounds of the tournament. Essentially, how do the characteristics of teams in the Elite Eight compare to teams that make the Final Four? How about teams in the Final Four compared to those that make the championship? Because of the discrete nature of tournament_wins, linear regression is not a possibility. As you will see throughout, we will use other statistical tools and methods instead to make determinations about the importance of given factors in the NCAA women's tournament, including confidence intervals, hypothesis tests, and logistic regression. Each method of statistical analysis is justified based off the research hypothesis for the given area of analysis.

### Visualizations

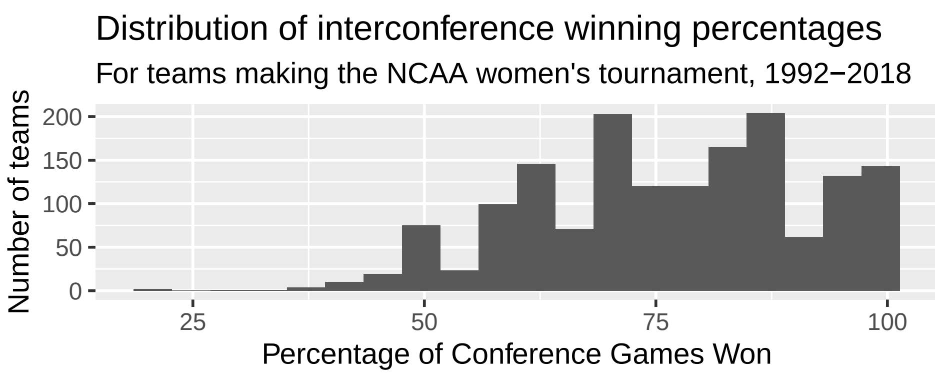
# Visualization 1: Tournament finish by seed

Our first visualization showcases how tournament seed correlates with success. We see immediately that a majority of the teams advancing to late outrounds in tournament are 1 and 2 seeds. Almost all of the lowest seeds lose in the first round, and most "middle" seeds lose in the first or the second round.



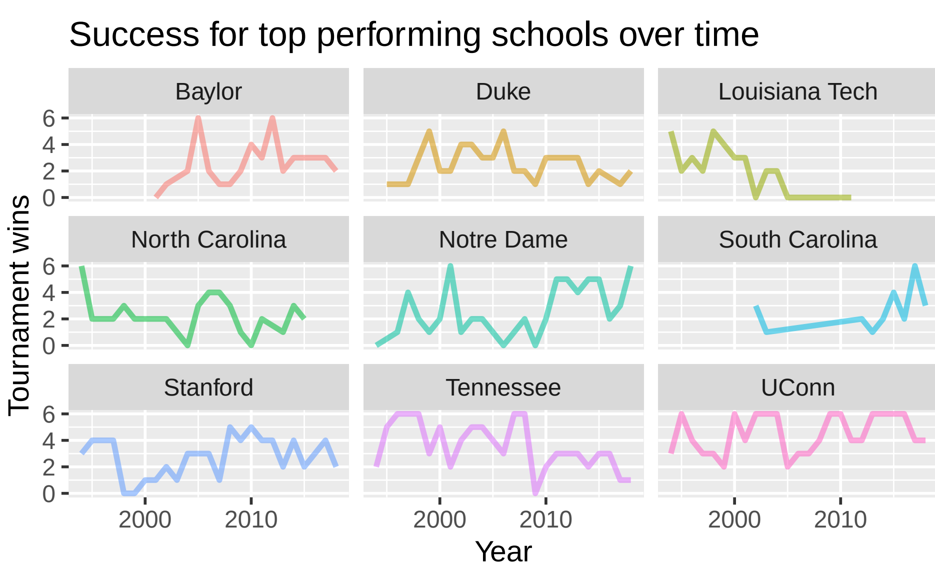
## Visualization 2: Distribution of conference

Our second visualization shows the distribution of conference records among all teams making the tournament across all years. We see that it is skewed left - which makes sense, as only the winningest teams even make the NCAA tournament.




## Visualization 3: Success for the (historically) best women's basketball teams

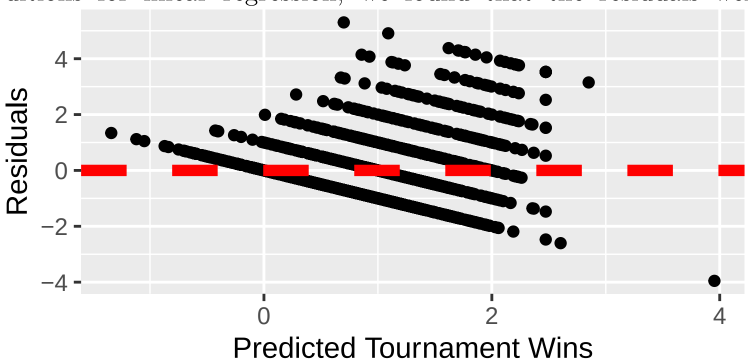
Our third visualization showcases the nine best-performing teams since 1992 (how we made that determination can be found under Analysis 4). We can see that some teams have been consistently impactful (UConn, Stanford, Notre Dame) while others have fallen off recently (Louisiana Tech, Duke, North Carolina) and others are still on the rise (South Carolina). The best programs of all time are UConn and Tennessee by a wide margin, with multiple back-to-back championships.




## Results

## Analysis #1: Conference and Non-Conference Winning Percentages

For our first area of analysis - conference and non-conference performance of tournament teams - we hypothesize that non-conference winning percentages will be a more meaningful predictor of tournament success than conference winning percentages.Unfortunately, the discrete nature of the number of tournament wins makes it difficult to do linear regression. When we evaluated the conditions for linear regression, we found that the residuals were not distributed equally around y = 0.




As such, we have limited tools for statistical analysis. We will attempt to see if there is a meaningful difference between non-conference performance and conference performance as a determinant in NCAA success by looking at previous data, which cannot help us fully answer our above hypothesis but can tell us if either statistic is significantly higher for the winningest teams. We can do this by conducting hypothesis tests for a difference in means of winning percentages. We will conduct three sets of paired t-tests, one for the Sweet Sixteen, one for the Final Four, and one for National Champions (across all years in our dataset). Each one of them has the same hypotheses:

$$H_0 : \mu_{conference\ percent} = \mu_{non-conf\ percent}$$

$$H_a : \mu_{conference\ percent} < \mu_{non-conf\ percent}$$

They are all conducted at the $\alpha = 0.05$ level.

#### Paired t-test for means: Conference vs. non-conference games as predictors

For the purposes of not exceeding the page limit, we will manually evaluate the p-values resulting from paired t-tests rather than printing tibbles.
For the Round of 16 data: $p = 0.485$, with a test statistic of -0.037
For the Round of 4 data: $p = 0.873$, with a test statistic of 1.15
For the Champions data: $p = 0.447$, with a test statistic of -0.133

```
##   mean_conf_pct mean_nonconf_pct
## 1      82.19549         82.22481

##   mean_conf_pct mean_nonconf_pct
## 1        89.499            87.82

##   mean_conf_pct mean_nonconf_pct
## 1        92.564           92.996

## # A tibble: 1 x 6
##   statistic  t_df p_value alternative lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1   -0.0372   398   0.485 less            -Inf     1.27
```

```
## # A tibble: 1 x 6
##   statistic  t_df p_value alternative lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1      1.15    99   0.873 less            -Inf     4.10
```

```
## # A tibble: 1 x 6
##   statistic  t_df p_value alternative lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1    -0.133    24   0.447 less            -Inf     5.11
```

Because our p-values of $0.485, 0.873, and 0.447$ are all larger than $\alpha = 0.05$, we fail to reject the null hypothesis at the Sweet Sixteen, Final Four, and National Champion level. For all teams making each of those rounds, we fail to find a statistically significant difference between their conference winning percentages and non-conference winning percentages.

### Ordinal Logistic Regression Model

An ordinal logistic model can help us predict the probability of success or failure for a success condition that has multiple ordered levels, such as the rounds in the NCAA women's tournament. We develop an ordinal logistic model using conference winning percentages and non-conference winning percentages as predictors.

```
## # A tibble: 8 x 5
##   term        estimate std.error statistic coef.type
##   <chr>          <dbl>     <dbl>     <dbl> <chr>
## 1 conf_pct      0.0216   0.00334      6.47 coefficient
## 2 nonconf_pct   0.0766   0.00441     17.4  coefficient
## 3 1st|2nd       7.30     0.428       17.1  scale
## 4 2nd|RSF       8.62     0.442       19.5  scale
## 5 RSF|RF        9.61     0.455       21.1  scale
## 6 RF|NSF       10.5      0.469       22.3  scale
## 7 NSF|N2nd     11.3      0.484       23.2  scale
## 8 N2nd|Champ   12.0      0.508       23.6  scale
```

Our ordinal logistic regression model is as follows (keeping in mind that percentages are on a scale from 0 - 100):

$$log(\hat{P}(Wins\ >\ J) = Intercept - 0.022 Conference\ Win\% - 0.077 Non - conference\ Win\%$$

For the sake of concision and clarity, we will simply list out the intercepts rather than listing seven different model equations. - For $J = 0$ (Probability of advancing to the Round of 32): $Intercept = 7.3$
- For $J = 1$ (Probability of advancing to the Round of 16): $Intercept = 8.62$
- For $J = 2$ (Probability of advancing to the Round of 8): $Intercept = 9.61$
- For $J = 3$ (Probability of advancing to the Round of 4): $Intercept = 10.5$
- For $J = 4$ (Probability of advancing to the Championship): $Intercept = 11.3$
- For $J = 5$ (Probability of winning the Championship): $Intercept = 12.0$

However, it is difficult to evaluate the correlation between the log-odds of advancing in a tournament and conference/non-conference win percentages. Because our current statistical tools cannot allow us to compare the predictive power of the two, and because our statistical test on past data found no meaningful difference between them, our hypothesis is inconclusive. However, it appears that non-conference and conference records are similarly predictive of tournament success.

## Analysis 2: At-Large vs. Autobid Placements

For our second area of analysis, we examine the qualification method for teams across all years and see if one team does better than the other. We hypothesize that at-large teams do better than autobid teams, in part because they are hand-selected by an NCAA committee. In order to test this hypothesis, we will do a confidence interval for the difference of means across all teams and all years.

$$H_0 : \mu_{atlarge} > \mu_{auto}$$

$$H_0 : \mu_{atlarge} = \mu_{auto}$$

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.239    0.578
```

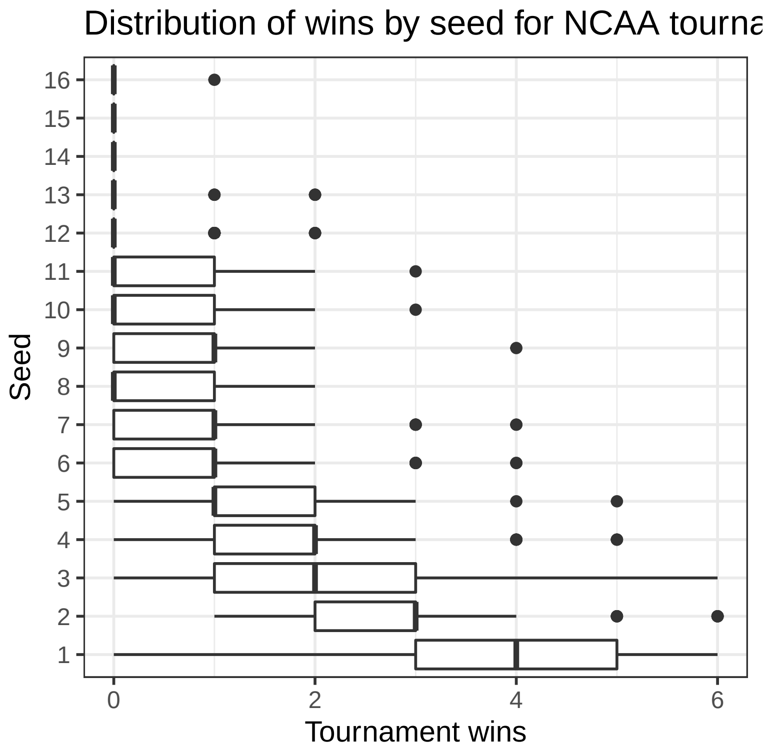
Because zero is not on the confidence interval, we can conclude that there is a statistically significant difference in means between autobid and at-large teams. Our hypothesis is supported - at-large teams do better than autobid teams, and we are 99% confident that this difference is between 0.24 and 0.58 mean wins.

## Analysis 3: Seed Overperformance/Underperformance

For this section, we hypothesized that the seeds would be a strong predictor of tournament success, with the seeds' expected win totals falling in predicted order. To assess seed accuracy, we first found general confidence intervals for the 16 different seeds' tournament win number at a 95% confidence level. For clarity and conciseness, the function evaluating confidence intervals for all sixteen seeds is not run; we will instead write them manually:
- 1 seeds: (3.53, 4.11)
- 2 seeds: (2.39, 2.88)
- 3 seeds: (1.89, 2.36)
- 4 seeds: (1.62, 2.02)
- 5 seeds: (0.98, 1.34)
- 6 seeds: (0.83, 1.21)
- 7 seeds: (0.69, 1.03)
- 8 seeds: (0.39, 0.60)
- 9 seeds: (0.45, 0.70)
- 10 seeds: (0.27, 0.51)
- 11 seeds: (0.31, 0.59)
- 12 seeds: (0.15, 0.35)
- 13 seeds: (0.02, 0.17)
- 14 seeds: Does not exist
- 15 seeds: Does not exist
- 16 seeds: (0.00, 0.03)

For the most part, as the seed number increased, the confidence interval decreased, with the exception of Seeds 9 and 11, which surprisingly did better than the seed above them. The non-existence of the confidence intervals for the 14 and 15 results from the fact that neither team with those two seeds has ever advanced in the NCAA tournament. Because the lower bound for the confidence interval of 1 seeds is higher than the upper bound for the confidence interval of 2 seeds, the difference in the mean number of tournament wins for 1 seeds and 2 seeds is statistically significant. The same goes for 2 and 3 seeds, 4 and 5 seeds and 7 and 8 seeds.

## Seed Distribution Visualization




A boxplot visualization of the data can help us reaffirm the conclusions we have made above. There is a large cutoff between the 1 and 2 seeds' performances, and 1, 2, and 3 seeds are the only seeds to ever win the championship. Other cutoffs seem to exist between the 7 and 8 seeds (with the exception of seed 9, whose comparative overperformance is visible graphically) and between the 11 and 12 seeds. Overall, our hypothesis is supported, although we were unable to account for some small nuances, like the overperformances of 9 and 11 seeds and the dominance of 1 seeds. On the whole, the NCAA does a pretty good job seeding, it seems!

## Analysis 4: Conference/School Performances

Lastly, we wanted to look specifically at individual conferences and schools see if their seeding was an accurate representation of their tournament success. We hypothesized that "Power Five" conferences and schools (Atlantic Coast, Pacific-10/Pac-12, Big 12, Big Ten, and Southeastern) would be overseeded compared to how they actually performed, and other schools and conferences would be slightly underseeded. We did this first by mutating the "mean_tourney_wins" and "mean_seed" variables for each school & conference, arranging by mean tournament wins and mean seed, and joining the two variables together. We then assigned a rank to each school/conference's mean tournament and seed, and calculated a "disparity" variable to see how much each seed rank differed from its respective tournament rank. The "disparity" column shows the difference between where a conference should be ranked based on its mean tourney wins and where it is ranked by the tournament committee, based on its mean seed.

### Conference Data

Since there were only 49 total conferences in our data, we only looked at the ten most successful conferences (by mean tournament wins) for our analysis.

```
## Joining, by = "conference"
```

```
## # A tibble: 10 x 6
##    conference     mean_tourney_wi~ conf_rank mean_seed conf_seeds_rank disparity
##    <chr>                     <dbl>     <int>     <dbl>           <int>     <int>
##  1 American Athl~             2.58         1      4.67               2         1
##  2 Atlantic                   2            2      4                  1        -1
##  3 Pac-12                     1.91         3      4.71               3         0
##  4 Southeastern               1.90         4      4.73               4         0
##  5 Big East                   1.82         5      5.62               7         2
##  6 Atlantic Coast             1.71         6      4.83               5        -1
##  7 Big 12                     1.42         7      5.33               6        -1
##  8 Pacific-10                 1.35         8      6.19               9         1
##  9 Southwest                  1.2          9      7.1               10         1
## 10 Big Ten                    1.19        10      5.98               8        -2
```

```
## # A tibble: 1 x 1
##   mean_disparity
##            <dbl>
## 1              1
```

Our results show that the seeding of schools within some conferences is not as accurate as others. Ideally, the “mean_seed” value should increase as the list goes down (as it is arranged according to tournament success). However, conferences such as Atlantic, Atlantic Coast, and Big Ten all seem to underperform as they have a lower seed average than conferences that fare better than them, while the Big East and Pacific-10 are slightly underrated. The mean of the absolute values of each disparity number is the “mean_disparity”, which represents the overall “inaccuracy” of the seeding. The mean disparity of the conferences was 3.55.

### School Data

Since there were 271 schools in our data set, we only looked at the 25 most successful schools (by mean tournament wins).

```
## Joining, by = "school"
```

```
## # A tibble: 25 x 6
##    school         mean_tourney_wins wins_rank mean_seed seeds_rank disparity
##    <chr>                      <dbl>     <int>     <dbl>      <int>     <int>
##  1 UConn                       4.52         1      1.2           1         0
##  2 Tennessee                   3.56         2      1.92          2         0
##  3 Stanford                    2.76         3      2.96          6         3
##  4 South Carolina              2.67         4      2.56          4         0
##  5 Baylor                      2.59         5      2.59          5         0
##  6 Notre Dame                  2.58         6      4.58         15         9
##  7 Duke                        2.48         7      2.35          3        -4
##  8 North Carolina              2.26         8      3.26          7        -1
##  9 Alabama                     2.17         9      3.83         11         2
## 10 Louisiana Tech              2.07        10      5.2          23        13
## # ... with 15 more rows
```

```
## # A tibble: 1 x 1
##   mean_disparity
##            <dbl>
## 1            5.8
```

We can see from the data here that Notre Dame, Louisiana Tech, Mississippi State, and Louisville are significantly underseeded, with the largest disparities of the top 25. By comparison, Colorado, Penn State, and Texas Tech were overseeded by the tournament committee. The mean disparity for schools was 5.52. Overall, we did not notice any strong trends with regards to over/underseeding of schools and/or conferences. Thus, we reject our initial hypothesis - while the NCAA could definitely improve in the way it treats specific schools, there does not seem to be any systemic conference bias one way or another.

# Discussion

We aimed to answer a few central research questions with our project, chiefly, "What factors should the selection committee of the NCAA Division I Women's Basketball Tournament consider most strongly when determining the seeding?" Overall, we conclude that the NCAA is doing a pretty good job; the factors they are considering right now are resulting in strong teams making the tournament and seeding that is mostly accurate. Strong conference and non-conference records correlate fairly well with tournament success (although, again, as a caveat, we were unable to support this information statistically.) The teams that the NCAA hand-picks to receive an at-large bid do meaningfully better than autobids, as confirmed through our second analysis, and higher seeds are generally more successful than lower seeds, with the highest seeds (1-4) winning most of their games in any given year. Lastly, there does not appear to be any systemic bias towards/against certain schools or conferences, although the tournament organizers ought to consider past overseeds and underseeds of specific schools when creating future brackets.

There are definitely limitations and room for improvement in our methods. For one, we were unable to come up any conclusive results for the conference placements, conference records, and regular season records due to the discrete nature of the data. Further, the data itself had its own issues–the main factor we wanted to look at with this project was tournament success, but there was only one variable that related to that (tournament finish). This makes it difficult to judge actual tournament success by team - for example, lower-seeded teams may consider being 'successful' in the tournament as simply making it past the first one or two rounds, while for higher-seeded teams anything short of making the national tournament could be considered an underwhelming finish. Furthermore, it is important to consider the impact that the actual seeding has on the tournament outcome. Higher-seeded teams traditionally have an easier path to advancing further in the tournament, which begs the question of whether it is the higher seeding that allows for tournament success, or rather the relative strength of the teams being correctly predicted by the seeding.

Especially compared to the men's tournament, the women's tournament has relatively lower amounts of upsets and usually results in one of the top-seeded teams being crowned the champion. One potentially confounding variable that may affect this analysis is the relative strength of each NCAA league. The men's league usually has a variety of "good" teams that are all potentially viable for title contention, as compared to the women's league which has comparatively fewer "good" teams, with several stronger teams at the top taking the majority of the wins. If we were to do this project again, we would have to consider several of these confounding variables and limitations. Most notably, we would likely have to devise some sort of metric for judging tournament success beyond just final placement, as well as devising a method for judging whether it was a team's seeding or a team's actual strength that predicted their tournament success. Regardless, from this analysis we can conclude that the Selection Committee is not making any obvious errors in their appointment of seeds. Additionally, it does appear that regular season and conference wins are the best predictor of relative team strength and future tournament success.