Sophie Guo
UNC SAIL Application
March Madness 2019 Analysis
1/7/2023

For my analysis, I used Andrew Sundberg's College Basketball Dataset via Kaggle and focused on data from March Madness 2019. My goal was to find a more accurate way of predicting a team's rank when compared to the NCAA seed. I chose 2019 because I wanted to use a year that was unaffected by the COVID-19 pandemic. After limiting my dataset to the teams which participated in March Madness 2019, I sorted them in order of ascending SEED. Subsequently, I replaced the POSTSEASON string column with a RANK integer column so I could perform mathematical operations on this column. To compare the NCAA seed with RANK, I converted SEED into Standard SEED, using the accepted values for RANK: 1, 2, 4, 8, 16, 32, 64, and 68. Thus, Standard SEED and RANK would have the same values and could be directly compared.

After cleaning my dataset, I used linear regression and determined the correlation coefficient between multiple variables. I wrote functions called *standard_units*, *correlation*, *slope*, *intercept*, and *prediction_at* to make the process more efficient. First, I compared Standard SEED with the Adjusted Offensive Efficiency (ADJOE) average. ADJOE means the points scored per 100 possessions, or trips down the floor with the basketball, against an average Division I opponent. Since many teams have the same Standard SEED value—and thus RANK—I decided to use the average of ADJOE for each Standard SEED. Teams with the same Standard SEED or RANK are considered to be on the same level. Next, I used bootstrapping to predict the range for the ADJOE average, given a Standard SEED. I wrote a function called *bootstrap_prediction* to speed up the process. Then, I repeated these steps with RANK and ADJOE average.

Through the correlation coefficient, I found that there was a strong association between Standard SEED and ADJOE average (-0.9616538709786773) and RANK and ADJOE average (-0.9176629845352036). Additionally, from the bootstrapping prediction, I found that the 95% confidence interval showing where the ADJOE average was located for a 4th Standard SEED team [118.36, 121.62] experienced much overlap with the 95% confidence interval showing where the ADJOE average was located for a 4th RANKED team [116.31, 121.60]. I also found that the 95% confidence interval showing where the ADJOE average was located for an 8th Standard SEED team [117.71, 120.46] experienced much overlap with the 95% confidence interval showing where the ADJOE average was located for an 8th RANKED team [116.00, 120.42]. In both scenarios, the NCAA Seed (converted to Standard SEED), which was heavily correlated with ADJOE average, was effective at predicting a team's RANK.

Next, I used linear regression to determine the correlation coefficient between Standard SEED and Adjusted Defensive Efficiency (ADJDE) average. ADJDE refers to the points allowed per 100 possessions against an average Division I opponent. I repeated this process with RANK and ADJDE average.

Through the correlation coefficient, I found that there was a strong association between Standard SEED and ADJDE average (0.9261184488219529) and RANK and ADJDE average (0.9087902125187143). Additionally, from our bootstrapping prediction, I found that the 95% confidence interval showing where the ADJDE average was located for a 4th Standard SEED team [88.50, 90.43] experienced much overlap with the 95% confidence interval showing where the ADJDE average was located for a 4th RANKED team [87.39, 92.04]. I also found that the 95% confidence interval showing where the ADJDE average was located for an 8th Standard SEED team [89.64, 91.30] experienced much overlap with the 95% confidence interval showing

where the ADJDE average was located for an 8th RANKED team [88.42, 92.56]. In both scenarios, the NCAA Seed (converted to Standard SEED), which was heavily correlated with ADJDE average, was effective at predicting a team's RANK.

Lastly, I made an algorithm to predict which RANK a team will place into. I made two new tables (*ADJOE_sorted* and *ADJDE_sorted*) by sorting the March Madness 2019 dataset in order of descending ADJOE and in order of ascending ADJDE. For reference, it is more favorable for a team to have a higher ADJOE and a lower ADJDE. Therefore, I replaced each team's RANK with a New Rank, based on their ADJOE and ADJDE relative to the other teams. Then, I averaged these two New Rank values for each team, which produced a Raw Predicted Rank. I proceeded to sort the teams in order of ascending Raw Predicted Rank and converted the Raw Predicted Rank into Predicted Rank, which limited the accepted values for Predicted Rank to be the same as the initial RANK: 1, 2, 4, 8, 16, 32, 64, and 68. My *prediction_table* consisted of the columns TEAM, RANK, and Predicted Rank. My *NCAA_prediction_table* consisted of the columns TEAM, RANK, and Standard SEED.

Next, I wrote a function called *distance* which measures the distance between two points. I found that the sum of the distances between Predicted Rank and RANK for every team in prediction_table (144.79640879524604) was less than the sum of the distances between Standard SEED and RANK in *NCAA_prediction_table* (169.82343772283025). In order words, the *NCAA_prediction_table* showed greater differences between Standard SEED and RANK than the *prediction_table*. The Standard SEED calculated using my algorithm in the *prediction_table* seemed to be more accurate at predicting a team's RANK than the Standard SEED in the *NCAA_prediction_table*.

Lastly, I wrote a function called *check_matchingvalues* that returns the number of identical values when sifting row by row through two columns. The matching values between Predicted Rank and RANK exceeded the matching values between Standard SEED and RANK by 4 values. Thus, the algorithm I used to calculate Predicted Rank based on the average of a team's ADJOE and ADJDE rankings was slightly more accurate at predicting a team's RANK than the official NCAA seed.