What statistical factors most attribute to success in March Madness for Division I basketball teams?

Division I basketball statistics over the last six years

We used a dataset from Kaggle that sourced data from Bar Torvick's website. The Kaggle user cleaned the and added POSTSEASON, SEED, and YEAR.

What Variables Predict the POSTSEASON Outcome of a Team?

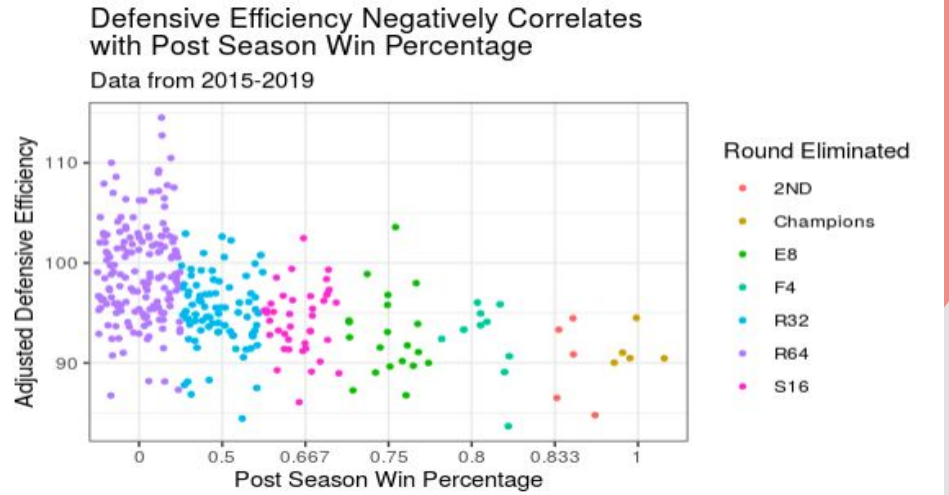
We formulated a linear model and then used backward selection to determine the variables that accurately predict the postseason win percent of a given team.

Logit and KNN Models to Determine Postseason Outcomes

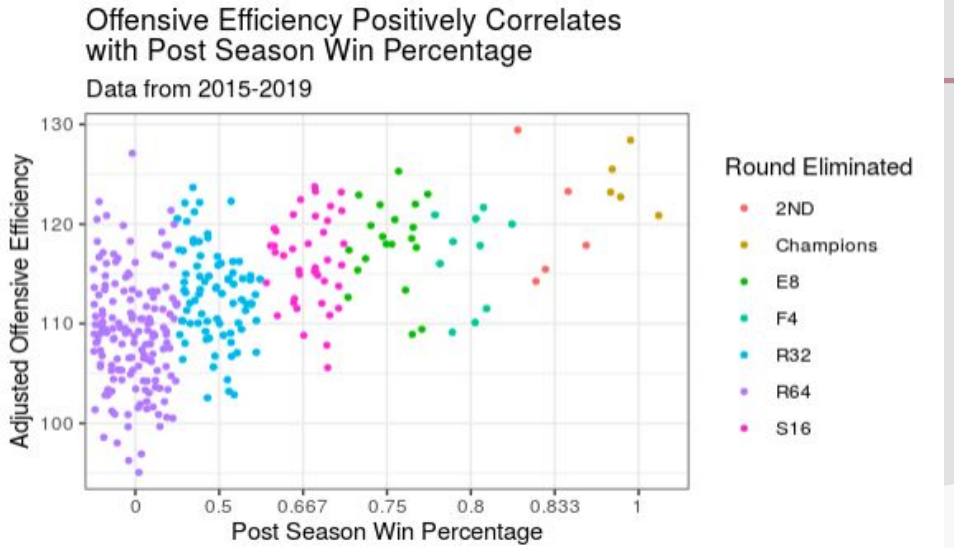
We developed a logit and a KNN model to test the prediction accuracy of the variables determined in the linear model.

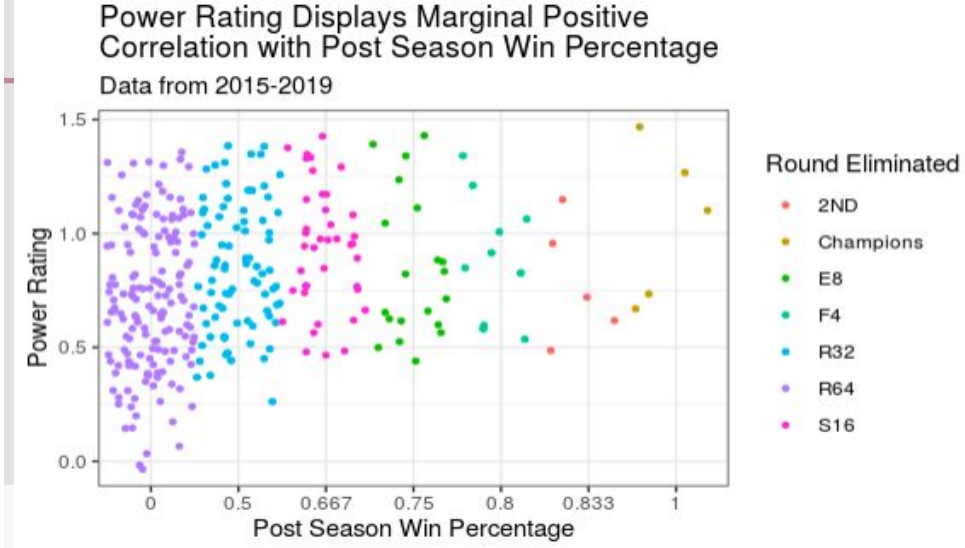
Would Duke Have Made It to the Final Four in 2020?

The motivation behind an analysis of Division I basketball statistics comes down to the one thing on everyone's mind: would 2020 have been the sixth? By using the methods outlined above, we will have an answer to the question that has been on everyone's mind.



These visualizations depict the effect of each of the three statistically significant variables on post season win percentage...





Conclusions

Significant Predictors

After backward selection, the linear model only included ADJOE, ADJDE, and BARTHAG to predict postseason win percentage.

Duke Does Not Make Final Four 2020

This is an accurate prediction due an 100% prediction accuracy of Duke Final Four appearances over the past five years. Our model predicted that the Final Four teams in 2020 would have been **Kansas, Gonzaga, Baylor, and Dayton.**

Relatively Low R-Squared Value

The linear model exhibited an r-squared value of 0.4687, which was lower than we had hoped. This shows maybe they call it March Madness for reason.

Prediction Accuracy

The logit model had a 94% prediction accuracy and an F1 score 0f 0.968. The KNN model had an F1 score of 0.958. This is most likely due to the limited number of true Final Four teams.