**Project Introduction**

Almost every February/March since my junior year of high school, I’ve taken on the challenge of trying to forecast the NCAA tournament. For quite a few years in a row, this ritualistic undertaking served as an excellent measuring stick of sorts. In high school, my big breakthroughs were learning about correlation coefficients, so there was plenty of progress to be made.

In 2019, however, I hit a wall. Armed with some elementary knowledge of statistical learning practices, I tried more or less every model I could find in [ISLR](https://www.statlearning.com/). Needless to say, I was quite disappointed when I found that the best model I could conjure up was a linear regression with just a few variables that re-packaged other rankings data.

The message was clear – experts like [KenPom](https://kenpom.com/), who rate college basketball teams for a living set a high bar – certainly nothing I could overcome with my shoddy understanding of machine learning and rudimentary programming skills.

Last winter, though, I began to embrace a new idea: perhaps forecasting the probabilities associated with each potential outcome isn’t the most important aspect of March Madness. Even perfect knowledge of the probabilities associated with each potential matchup could not, by itself, tell anyone how to fill out a “best” bracket. If the goal (as it is for most people) is to win a large bracket competition - the bracket that’s most likely to occur in real life does not necessarily have to be the most likely bracket to win a bracket pool competition.