Midterm Progress Report

There have been a few adjustments to the previous timeline.

The first delay was with the dataset compilation. Originally, I was planning to scrape data using R, but there were some package/versioning issues that I ran into. I considered pivoting to a python package that used ESPN’s public API to scrape game scores for the 2022 and 2023 seasons, however the run-time for pulling the relevant data was excruciatingly long. I therefore found a new R package called “hoopR” to pull the relevant data. The data consisted of box scores for the 2022 and 2023 seasons. It also took some extra time to calculate the offensive and defensive ratings since they weren’t automatically provided through the APIs. Once I got those figured out, I did some data preprocessing and created an initial data to import into python with only the relevant columns. After the dataset was prepared, I created the function to sample from individual distributions of each team. Each team will have a sample taken from a normal distribution centered around their corresponding and standard deviation based. I chose a normal distribution since all game scores are approximately normally distributed. With the function prepared, I plan on simulating games this week before using the model to simulate last season’s March Madness matchups and build a bracket based on those predictions.

I am a bit delayed on the linear model front. I want to compare this Monte Carlo method against a baseline logistic regression but have not trained any of the models as I’d hoped. My goal is to train and test that model this week before predicting some games later this week. I’m planning to compare based on log loss since I’ll be using probabilities as the output. Below, I’m attaching an updated Gantt chart to represent my progress.

