Below is my brief write-up as to how I beat the Marcel the Monkey Projection System in order to complete the Level 3 SABR analytics certification course. Word limit was 500 words. The code used to generate the model can be found on my GitHub here

My dream career is to work in MLB player evaluation processes, and since I don't have a sports/business/math academic background, I've been taking these courses to build my skills. I approached this challenge with that goal in mind, exploring multiple methods including Bayesian modeling, random forests, and gradient boosting. I also tried incorporating Statcast variables, adjusted projected OPS based on hitter tiers generated by PCA clustering, and position-specific age adjustment curves. While not all methods worked, I learned a lot through my failures.

The Marcel calculations are simple yet effective, so I decided to stick with the Marcel skeleton since developing something more complex proved difficult. MLB parks vary in hitter-friendliness, and even deeper, hitting from a certain side of the plate in a park makes a difference as well. Thus, to beat Marcel I first adjusted the previous three seasons of each player's OPS by Statcast park factors, considering specific left- and right-handed park effects. Using these adjusted OPS values in place of their actual OPS, I calculated a 5-4-3 weighted average. Next, I applied Marcel's reliability, regression, and age adjustment calculations to get the regressed OPS. To generate the preliminary projected OPS, I adjusted the regressed OPS based on the park factor for the team they began 2021 with. For example, a right-handed batter on the White Sox receives a 1% boost to regressed OPS based on Guaranteed Rate Field's park factor.

For switch hitters, I applied 72% of the left-handed park factor and 28% of the right-handed factor when calculating the park factor-adjusted OPS and preliminary projected OPS. This was based on research showing around 72% of pitches are thrown by right-handers in a given season. Although I considered using individual switch hitters' left/right splits, I opted for this method due to display issues on Fangraphs. Using play-by-play data from Retrosheet to calculate the exact proportion of plate appearances could have improved this step, but I felt this approach was ok for the purposes of this learning assignment.

To further improve projections, I calculated a weighted, regressed wRC+ for each player over the past three seasons and applied a final OPS adjustment. Since wRC+ and OPS share some of the same components, the idea behind using it was to "fight" regression to better project hitters at the extremes. For example, to get the final projected OPS for a player with a 140 weighted wRC+, I did 1.4 times league average .720 OPS, minus .720, divided by 5, resulting in a 0.0576 boost to preliminary projected OPS. Additionally, while there may be some overtuning to this specific range of years with this calculation, I felt it helped me go beyond just adjusting for park factors. Ultimately, I believe I was able to beat Marcel by recognizing how a player's home park can affect their numbers (especially when switching teams), and by accounting for players at the extremes. Overall, this project taught me a great deal about projection model construction, and I thoroughly enjoyed it.