Charles Schatmeyer
DS5220 Project Abstract
MLB Batting Stance, Performance Analysis

In this project, I am looking to analyze Major League Baseball statistics from Baseball Savant and Google Cloud's Statcast, and specifically the newly released dataset on batting stances, or where a batter stands during their at-bat. Despite Baseball's reputation as potentially the most analytically-studied sport, there is no consensus on an optimal batting stance, and it is currently considered to come down to player preference. Through this project, I hope to find connections between batting stance factors, such as distance from the pitcher, angle, stance width, etc., and hitting outcomes, including hits, runs, times hit-by-pitch, and more. There is a great depth of conclusions that could be found within this dataset, from simple questions such as "Does standing closer to the plate increase the odds of getting hit by a pitch?" to harder questions like "Do power hitters have marginally different stances compared to contact hitters?" or "Can contact percentage be predicted from batting stance alone?".

For the exploratory data analysis of this project, I'll look to spend a solid amount of time visualizing the data to find strong correlations, both to eliminate overlapping predictors and to quickly find basic, interesting conclusions. As the project evolves, I'll look towards the logistic regression performed in class to find relevant, strong predictors from batting stances for a number of different targets from batting outcomes. Where the previous steps are successful, I'll expand by building decision trees, random forests, or XGBoost models as appropriate, to see if batting stance, as a whole or just a factor of it, can accurately predict batting performance in any specific way. Within this project, there is also potential for using clustering algorithms to find groupings of batting stance ideas, which may then be able to be labeled based on batting characteristics.

The batting stance dataset on MLB hitters contains a ton of useful predictors and targets that have the potential to lead to interesting conclusions, both high and low stakes. I am confident that through this project, at the very least, I'll gain useful insights into a key component of baseball, that which has up to this point remained mostly nebulous throughout the history of the sport.