Title: Identifying Players Who Hit it Out of Their Ballpark (Estimate)
Team Members: Tommy Skodje, Clay Winder
Problem and Dataset Description:

The topic of player value in sports is one that is becoming more and more salient in recent times, as franchises are continuously seeking to sign players that directly lead to winning more games. This increased focus on value, combined with more advanced statistics making their way into sports, leads to a need for statistical approaches that accurately quantify a player's objective value to their team.

In this project, we will be looking at the value a baseball player can provide to their team by hitting. Traditional statistics for hitting are simple counting stats such as walks and hits, or rate stats such as batting average or slugging percentage. Advanced statistics such as WAR (wins above replacement) and VORP (value over replacement player) look at combinations of these basic statistics in order to try and determine a player's value to their team. Even in these advanced statistics, there are many unquantified aspects of what a player does in order to help their team win games. The fundamental problem with these statistics is that they only describe *what* happens, not *how* it happens. That's where Statcast comes in.

Statcast is a tool used by MLB (Major League Baseball) to record statistics on players and data about the baseball itself. Statcast data is collected automatically through cameras during each MLB game. The Statcast system has been implemented since 2015, and the statistics it generates are slowly being adopted among more traditional statistics such as batting average and home runs.

Dataset Link:

https://baseballsavant.mlb.com/leaderboard/custom?year=2023,2022,2021,2020,2019,2018,2017,2016,2015&type=batter&filter=&sort=4&sortDir=desc&min=q&selections=xba,xslg,xwoba,xobp,xiso,exit_velocity_avg,launch_angle_avg,barrel_batted_rate,&chart=false&x=xba&y=xba&r=no&chartType=beeswarm

The Statcast dataset contains both measurements directly gathered from cameras, as well as metrics calculated from those measurements. Some of the measurements we will be using include:

- Exit velocity: How fast a ball is hit by a batter immediately after the moment of impact (miles per hour).
- Launch Angle: How high a ball is hit by a batter (degrees)

- Base-to-base Time: How long it takes a runner to get from one base to another. Calculated for each individual base, for example, home plate to first base or first base to second base (seconds).

Some of the metrics derived from these measurements we will be using include:

- Expected Batting Average (xBA): The likelihood that a batted ball will lead to a hit. A batted ball is assigned a likelihood to become a hit depending on exit velocity, launch angle, and sprint speed of the player.
- Expected Weighted On-base Average (xwOBA): The likelihood that each batted ball is to become a single, double, triple, or home run based on similar batted balls. It is calculated in a similar way to xBA, but also takes into account different types of hits (doubles, triples, and home runs).
- Sprint Speed: A player's top running speed (feet per second).

Description of stats courtesy of: https://www.mlb.com/glossary/statcast

We will also be using conventional baseball statistics such as batting average (BA) and weighted on-base average (wOBA) to compare a player's expected statistics to their actual statistics

Plan and Difference from Existing Research:

Our goal is to train a model to predict how much a player will overperform or underperform their expected stats. Since these stats are continuous, this will be a regression problem. We plan to train linear regression and kNN models to this end. Existing research has used kNN classifiers, decision trees, neural networks, and SVMs to predict future performance of particular stats, such as WOBA. However, this has only focused on either:

1. predicting actual stats (e.g. WOBA) instead of predicted the difference between expected stats and actual stats (e.g. WOBA - xWOBA), or
2. Predicting pitcher performance, which relies on completely different stats.

Existing Reseearch:

Watkins, Christopher. *Novel statistical and machine learning methods for the forecasting and analysis of Major League Baseball player performance*. Diss. Chapman University, 2020.

- Uses kNN classifiers, decision trees, neural networks, and SVMs to predict a player’s WOBA in future seasons

Ishii, Tatsuya. "Using machine learning algorithms to identify undervalued baseball players." *Technical Report: Stanford University* (2016).

- Uses StatCast data to identify undervalued pitchers through k-means clustering, Boosting, and Random Forest