MGSC 310 Portfolio Project Outline
Emily Nguyen & Diego Oseguera

Datasets:

https://baseballsavant.mlb.com/statcast_search - We will extract data from this website that has a pitcher's percentage of strikeouts, walks, home runs, hits, etc, for any season.

https://www.baseball-reference.com/ - We will use this website to extract data about a pitcher's ERA for each season.

https://www.kaggle.com/datasets/mattop/mlb-batting-exit-velocity-data-2015-2022/data - MLB Batting Exit Velocity Data

Business Question:

The business question that we will seek to answer is: *Can we predict a pitcher's earned run average (ERA) based on their performance metrics?* This question brings value to the business/team because they can use this information to evaluate the performance of pitchers for acquisition and contract purposes, as well as for in-game decision-making. The coaches can use this information to decide whether to pull a pitcher from the game and prevent them from injuries.

The business question we aim to explore is the potential of *leveraging analysis on MLB players' Average Exit Velocity and Launch Angle data to predict their likelihood of hitting home runs and extra-base hits*. By understanding this relationship, we hope to identify players who have the potential to significantly contribute to a team's offensive performance. This question is not only critical for effective and strategic player scouting and development for our coaches but also instrumental in making informed decisions during player acquisitions and contract negotiations.

Statistical/Predictive Methods:

We will use linear regression for the first method because it will tell us how a pitcher's performance will either increase or decrease his ERA. We will use feature engineering and pre-processing to clean the dataset before we fit the model. Additionally, we will do a train-test split so that I can test my model for how well it performs.

We will utilize linear regression for our method as the model aims to predict the frequency of home runs and extra-base hits, represented by the 'total_barrels' metric, based on a player's Average Exit Velocity and Launch Angle. The dataset will undergo thorough cleaning, feature engineering, and preprocessing to ensure its suitability for the linear regression model. Additionally, we will implement a train-test split to effectively evaluate the performance of the model. The evaluation of the model's accuracy and predictive power will be based on key performance metrics such as R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

Target & Independent Variables:

The target variable that we are trying to predict is the earned run average. The independent variables that we will use to predict pitchers' ERA are strikeout rate, walk rate, hits allowed, and home runs allowed.

The target variable we are trying to predict is the total number of barreled balls a player hits, as this can be a strong indicator of their potential for hitting home runs and extra-base hits. The independent variables we will be using are Average Exit Velocity and Launch Angle. Average Exit Velocity will give us insight into how hard a player hits the ball, while Launch Angle will help us understand the trajectory of these hits, which is crucial for predicting the likelihood of home runs and extra-base hits.