**Abstract**

The goal of this project was to create a pipeline to process in large amounts of batting data, create a linear regression model to predict home runs, and deploy it using a web app. This project will be useful to help baseball organizations and fantasy managers make better decisions for acquiring baseball players. I worked with data provided by baseball-reference.com. After refining a model, I built an interactive web application to predict the home run total of a user’s data.

#### Design

This project originated from a desire to better understand a significant and growing part of baseball. Home runs are increasing across the league, and better prediction of players that hit them will be immensely helpful and profitable to MLB organizations.

#### Data

The dataset contains 78,000 season batting statistics with 30 features each. A few feature highlights include home runs, plate appearances, hits, walks, and strike outs.

#### Algorithms

Web Scrapping

Pulling data from baseball-reference.com.

Model Evaluation

The entire dataset was split into 80/20 training to test. Final model selected was a linear regression model utilization 7 different features.

Rsquared Training = .892

Rsquared Testing = .897

Web App

Built using a combination of HTML and Flask. Allows web site visitors to input own values to predict homeruns based on model built.

#### Tools

* Scraping
  + Beautiful Soup & Requests
* Database
  + Pandas, Numpy, MongoDB
* Linear Regression
  + Statsmodels
  + Seaborn
  + Matplotlib
  + Sklearn
* Web Deployment
  + Flask
  + Pickle

#### Communication

In addition to the slides and visuals, <baseball-reference.com> will be embedded.

Chart

Description automatically generatedChart, bar chart

Description automatically generated

Table

Description automatically generated with medium confidence