**Abstract**

The goal of this project was to create a pipeline to capable of efficiently scraping, cleaning, and outputting batting statistics in a format ready for visualization on a web application. This project will be useful to help baseball organizations to better understand how baseball is changing over time and make better decisions for acquiring baseball players. I worked with data provided by baseball-reference.com. After refining a pipeline, I built an interactive web application to display summary statistics from my data aggregation.

#### Design

This project originated from a desire to better understand how baseball is changing recently, and how dramatic those changes truly are. 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 96,000 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.

Data Cleaning

Cleaned using Pandas and Numpy

Data Transformation and Aggregation

Data was fed into a MongoDB database

Web App

Built using a combination of Streamlit and Plotly. Allows web site visitors to visualize a variety of summary statistics detailing how baseball has changed over time.

#### Tools

* Scraping
  + Beautiful Soup & Requests
* Database management and aggregation
  + Pandas, Numpy, MongoDB
* Web Deployment
  + Streamlit
  + Plotly

#### Communication

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

Diagram

Description automatically generated

Chart

Description automatically generated