CIS 550 Project Proposal

Ryan Corkrean, Iris Gallo, Min Kim, Jeremy Kogan

October 12, 2021

1 Group Information

Name	E-Mail	GitHub
Ryan Corkrean	corkrean@seas.upenn.edu	rcorkrean
Iris Gallo	igallo@seas.upenn.edu	irisgallo
Min Kim	minseokk@seas.upenn.edu	minskim0327
Jeremy Kogan	jdkogan@seas.upenn.edu	jdkogan

2 API Description

As an extension of the web design concepts explored in our second homework assignment, we propose a baseball API that provides curated statistical profiles of games, players, and even individual pitches. There are a number of publicly-available data sources that specialize in particular statistical subdomains: Baseball Savant tracks pitch and batted ball data using Statcast's Hawk-Eye camera functionality, FanGraphs has an excellent user interface and bleeding-edge metrics to quantify player value (both on career and single-season bases), and Baseball-Reference maintains historical records and game results. Our API would aggregate and amalgamate statistics from each of these sources to create a holistic user experience with an emphasis on sabermetric analysis [1]. Inquisitive fans of the sport could generate tables of information pertaining to their favorite teams or players, visit pages dedicated to individual games or play events, or compare season-by-season trends at their preferred level of granularity. We may even attempt to incorporate video clips, contract valuations, and even player projections for the 2023 season based on past performance.

3 Data

3.1 Baseball Savant [2]

Baseball Savant is MLB's clearinghouse for Statcast data. Statcast is a state-of-the-art tracking technology that allows for the collection and analysis of a massive amount of in-game data, including batted-ball exit velocity and launch angle, pitch velocity; spin rate; release location; and vertical and horizontal movement, and fielder and baserunner sprint speed. This database is publicly accessible, and query results can be downloaded in CSV format. However, a more scalable approach to data acquisition involves `pybaseball`, an open-source Python package for baseball analysis [3]. The package facilitates scraping of Statcast data, pitching stats, batting stats, division standings/team records, awards data, and more from Baseball Savant, FanGraphs, and Baseball-Reference. Data is available at the individual pitch level, as well as aggregated at the season level and over custom time periods.

3.2 FanGraphs [4]

FanGraphs hosts articles, statistical reports and also covers baseball history as well as current issues and events, including games and series, injuries, forecasts, player profiles, baseball finance, and the player marketplace. It is best-known for its proprietary formulation of wins above replacement (fWAR), which quantifies player performance in a single, all-encompassing statistic. FanGraphs data will also be scraped using the `pybaseball` library.

3.3 Baseball-Reference [5]

This site has season, career, and minor league records (when available, back to 1888) for everyone who has played Major League Baseball, year-by-year team pages, all final league standings, all postseason numbers, voting results for all historic awards such as the Cy Young Award and MVP, head-to-head batter vs. pitcher career totals, individual statistical leaders for each season and all-time, managers' career records, the full results of all MLB player drafts, Negro leagues statistics (Baseball Reference added Negro League Statistics to its website in 2021), a baseball encyclopedia (the Bullpen),[9] and box scores and game logs from every MLB game back to 1914, among other features. It also hosts its own competing version of wins above replacement (bWAR), and we once again plan to use `pybaseball` to acquire this data.

3.4 MLB [6]

Major League Baseball's official player profiles are likely the most accurate source of demographic information (full name, birth date, country of origin, etc.), given that its subsidiaries (constituent teams) employ these athletes directly. These profiles would be used to supplement and cross-validate the other data sources. Because Statcast and Baseball Savant are owned by MLB (and therefore share a player indexing schema) `pybaseball` implicitly caches this information in its Baseball Savant lookups.

4 Sample Queries

4.1 Find all players from the 2022 season that had a .300 or better batting average.

```
SELECT DISTINCT p.playerID, p.playerName, SUM(s.hits) / SUM(s.atBats) AS battingAverage
FROM playerDim AS p
JOIN positionPlayerRegularSeasonStats AS s
ON p.playerID = s.playerID
WHERE s.year = 2022
GROUP BY p.playerID
HAVING SUM(s.hits) / SUM(s.atBats) >= 0.3
ORDER BY SUM(s.hits) / SUM(s.atBats) DESC;
```

Here, an aggregation is necessary to account for players who played on multiple teams in 2022. (I.e., players who were traded, claimed off of waivers, or released and re-signed with another team.)

4.2 Find all Canadian-born "sinister right-handers" (players who bat left-handed and throw right-handed) in MLB history.

```
SELECT DISTINCT playerID, playerName
FROM playerDim
WHERE batHandedness = 'L' AND throwHandedness = 'R' AND country = 'Canada';
```

4.3 Find all teams that used fewer than 10 starting pitchers in 2022.

```
SELECT DISTINCT t.teamName, COUNT(*) AS numStartingPitchers
FROM teamDim AS t
JOIN pitcherRegularSeasonStats AS s
ON t.teamID = s.teamID
WHERE s.year = 2022 AND s.gamesStarted >= 1
GROUP BY t.teamID, t.teamName
HAVING COUNT(*) <= 10
ORDER BY COUNT(*);
```

4.4 Compare Noah Syndergaard's season-by-season fWARs and bWARs.

```
SELECT year, fWAR, bWAR, fWAR - bWAR as WARDiff
FROM pitcherRegularSeasonStats
WHERE playerID = (SELECT playerID
                  FROM playerDim
                  WHERE playerName = 'Noah Syndergaard');
```

4.5 Find the names of all teams that Edwin Jackson has ever played for.

```
SELECT teamName
FROM teamDim
WHERE teamID IN (SELECT teamID
                 FROM pitcherRegularSeasonStats
                 WHERE playerID = (SELECT playerID
                                   FROM playerDim
                                   WHERE playerName = 'Edwin Jackson'));
```

References

[1] *A Guide to Sabermetric Research.* URL: https://sabr.org/sabermetrics.

[2] URL: https://baseballsavant.mlb.com/.

[3] James LeDoux. URL: https://github.com/jldbc/pybaseball.

[4] URL: https://www.fangraphs.com/.

[5] URL: https://www.baseball-reference.com/.

[6] URL: https://www.mlb.com/.