Expanding Access to College Baseball Data

Nathan Blumenfeld, Cornell University
njb93@cornell.edu, @blumenfeldnate

Introduction

Despite its increasing relevance in the MLB talent pipeline, college baseball data is remarkably hard to access at any real scale, especially relative to the breadth of data publicly available at the the professional level. The NCAA gives no public API to its official statistics repository, offering access only through a limited and confusing user interface (stats.ncaa.org) that lacks export functionality.

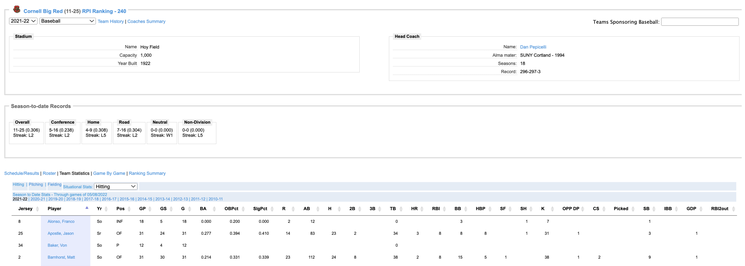

Figure 1. A page from stats.ncaa.org

These grievances are not at all original. Miles Okamoto, a former student analyst for Texas Baseball, did well to summarize the issue of data availability in a 2020 blog post describing the state of college baseball data availability:

> *While websites like Fangraphs, BaseballSavant, and Brooks Baseball, to name a few, have made it easy enough for anyone to do baseball research, for college baseball, virtually all that is publicly available is what is published on the NCAA's stats page, which is significantly less user-friendly than the major providers of data on pro ball... From the start, access to large scale college baseball data has a built in barrier, requiring some coding before the data is accessible.*

Existing Solutions

The only existing programmatic way to access NCAA baseball data is through the BaseballR package. Popular Python baseball analytics packages PyBaseball and pybbda have no NCAA functionality whatsoever. As a result of the lack of programmatic and non-programmatic access to college baseball data, multiple companies offer it for sale, including The Baseball Cube, CollegeSplits, 643 Charts, BaseballCloud, and SynergySports. This market evidences some level of demand for this data and a meaningful opportunity to improve public access.

Objective

This project aims to reduce the barrier to entry for working with college baseball data by addressing two main problems: 1. the lack of programmatic methods in Python, and 2. the lack of non-code access to advanced statistics. In particular, this project aims to benefit the student-run analytics organizations of college teams, which often have minimal funding and limited technical experience.

The collegebaseball package

The simple, flexible syntax of the collegebaseball package provides intuitive access to NCAA batting, pitching, and fielding statistics at season, game, player, and team aggregates. The package also includes built-in options for calculating advanced metrics, including fundamental metrics wOBA, FIP, and BABIP not computed by the NCAA. The splits provided by the NCAA are available via argument. Game results dating back to 1992 are available by way of boydsworld.com, and a number of helpful lookup functions are provided. A sample of the package's functionality is given below.

Season-level aggregates

```
ncaa_team_stats("Cornell", 2018, "batting")
✓ 1.4s
```

	Jersey	stats_player_seq	name	Yr	pos	GP	GS	R	AB	H	2B	3B	TB	HR	RBI	BB	HBP	SF	SH	K	DP	CS	Picked	SB	IBB	season
0	27	1884386	Kyle Gallagher	Sr	OF	37	37	29	139	38	10	1	53	1	22	35	0	2	0	29	2	2	0	4	0	2018
1	12	1884383	Ryan Krainz	Sr	INF	37	37	29	158	49	4	1	55	0	15	21	0	1	0	28	1	3	0	11	0	2018
2	33	1779085	William Simonelt	Jr	C	37	37	17	143	44	8	0	58	2	27	13	6	4	0	23	0	2	0	7	0	2018

Game-level aggregates

```
ncaa_team_game_logs("Cornell", 2018, "fielding")
✓ 5.2s
```

	PO	TC	A	E	CI	PB	SBA	CSB	IDP	TP	date	field	season_id	opponent_id	opponent_name	innings_played	extras	runs_scored	runs_allowed	run_difference	result	game_id
0	24	35	9	2	0	0	0	0	0	0	02/23/2018	away	12973	697	Texas A&M	9	True	0	22	-22	loss	86405
1	24	35	10	1	0	0	2	0	1	0	02/23/2018	away	12973	697	Texas A&M	9	True	2	3	-1	loss	86469
2	24	38	12	2	0	0	0	0	2	0	02/24/2018	away	12973	697	Texas A&M	9	True	2	8	-6	loss	86590
3	24	29	5	0	0	1	3	1	0	0	03/02/2018	away	12973	193	Duke	9	True	3	8	-5	loss	87649

Game Results

```
ncaa_team_results('Duke', 2022)
✓ 6.8s
```

	game_id	date	field	opponent_name	opponent_id	innings_played	extras	runs_scored	runs_allowed	run_difference	result	school_id	season_id	division
0	2167977	02/18/2022	home	VMI	741	9	True	5	10	-5	loss	193	15860	1
1	2168329	02/19/2022	home	VMI	741	9	True	9	1	8	win	193	15860	1
2	2168557	02/20/2022	home	VMI	741	9	True	8	7	1	win	193	15860	1
3	2168701	02/22/2022	home	App State	27	9	True	8	5	3	win	193	15860	1
4	2169078	02/25/2022	away	Baylor	51	9	True	2	4	-2	loss	193	15860	1

Team Rosters

```
ncaa_team_season_roster('Yale', 2016)
```

	jersey	stats_player_seq	name	position	class_year	games_played	games_started	season	season_id	school	school_id	division
0	4	1428376	Nate Adams	INF	Sr	36	35	2016	12360	Yale	813	1
1	30	1762703	Sam Boies	P	Fr	16	0	2016	12360	Yale	813	1
2	6	1640575	Alex Boos	C	So	16	12	2016	12360	Yale	813	1
3	18	1640576	Eric Brodkowitz	P	So	2	1	2016	12360	Yale	813	1
4	20	1549567	Derek Brown	INF	Jr	12	5	2016	12360	Yale	813	1
5	2	1762704	Tim Degraw	OF	Fr	48	46	2016	12360	Yale	813	1

Splits

```
ncaa_team_stats('Texas', 2022, 'batting', split='vs_LH').sort_values('OPS', ascending=False)
✓ 1.5s
```

	Jersey	stats_player_seq	name	Yr	pos	GP	GS	R	AB	H	2B	3B	TB	HR	RBI	BB	HBP	SF	SH	K	OPP DP	CS	Picked	SB	IBB	GDP
4	17	2471763	Ivan Melendez	Jr	INF	67	67	22	53	22	2	1	56	10	21	14	3	0	0	17	2	0	0	0	2	2
1	4	2309889	Silas Ardoin	So	C	69	68	15	51	22	10	0	38	2	16	6	1	2	1	6	0	0	0	1	0	0
6	5	1997666	Skyler Messinger	Sr	INF	67	65	12	48	23	4	0	33	2	12	4	1	1	1	4	0	0	0	1	0	0
11	27	2485662	Jack O'Dowd	So	INF	21	7	2	5	3	0	0	3	0	0	1	0	0	1	0	0	0	0	0	0	0

What's Next?

- Play-by-play data
- Park Factors
- "Fast" package
- SportsDataverse Inclusion

abstract

In addition to the package, a public web application was created to go further in expanding accessibility, offering custom accessibility to large-scale pre-scraped datasets without any need to code. It is built using the collegebaseball package and hosted via Streamlit cloud and a separate cloud database. The application features a leaderboards page with customizable filters by season, position, class year, team, and minimum PA. The team page gives team-level aggregates and percentile rankings of each player, and the player page gives visualizations of a player's season-by-season production. The application also features a team history page, including an interactive visualization of run difference by season. Two screenshots of the application are displayed below.

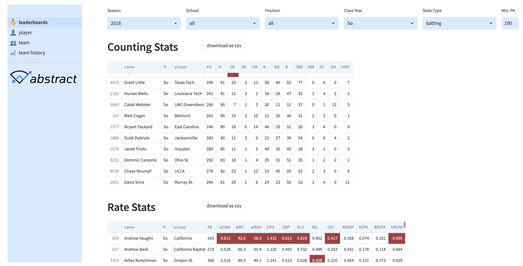

Figure 2. Leaderboards Page

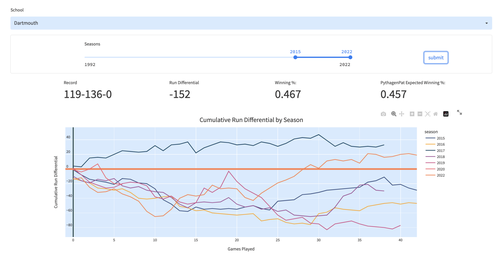

Figure 3. Team History Page

References

Miles Okamoto. *The New Road to Omaha*. July 2020. `https://milesokamoto.com/blog/2020/07/27/college-baseball-analytics-101`

James LeDoux. *PyBaseball*. `https://github.com/jldbc/pybaseball`

Ben Dilday. *pybbda*. `https://github.com/bdilday/pybbda`

Bill Petti and Saiem Gilani. *baseballr: The SportsDataverse's R Package for Baseball Data*. 2021. `https://billpetti.github.io/baseballr/`