Baseball's Offensive Evolution Throughout the Eras

By Jacob Adamsky

Link to website: https://jacobadamsky.github.io/final

Overview and Motivation

Yes, that's me playing baseball. I may not make the best faces when I'm playing, but I've almost always enjoyed the game. It has been hard for the past couple years, missing out on both my senior year of high school baseball and freshman year of college baseball, so being able to do a project on baseball is kind of therapeutic in a way. As for why I chose to focus on batting statistics, I used to be a two-way player, meaning I pitched and played the field (when I wasn't pitching), so I was reminiscing on the good old days when I was still allowed to hit.

As for the three plots I made, I plotted batting average against year to show the progression and concentration of different batting averages as time passed, hits + walks to runs to see how a higher OBP (on base percentage) increased a player's potential to score, and strikeouts to walks to see if there is a correlation between walk rate and strikeout rate. The database used was Sean Lahman's Baseball Statistics downloaded as a collection of CSVs (there is a MySQL version, but I never had success with it).

Related Work

One major related website that motivated me to do this was baseball-reference.com. I don't believe they have any visualizations for their data other than tables with player statistics, so I wanted to create some visualizations using the same data. Other than that, there wasn't anything that inspired me to do my final project on this subject.

Questions

There were three questions that I based my plots on: "How much of a difference was there in batting average as baseball evolved?", "How much correlation is there between a high walk and hit rate with how often a player scores?", and "Is there a correlation between a higher strikeout rate and lower walk rate, and vice versa?". At first, I was just plotting the correlation between batting average and years passed, but this was far too simple and couldn't get all the project requirements in. I added the second question, and later the third as I looked at the statistics database more.

Data

D3.js is surprisingly effective when it comes to filtering data, only taking a couple of seconds to filter out extraneous data (statistics from years not in the era being plotted), as well as plotting it. I did have some trouble searching for players at first as it was taking over a minute but realized that I had made a small mistake that was exponentially increasing the time taken to search for a player.

As for the data source, all of it comes from Sean Lahman's Baseball Statistics Database. It has several different access methods, but I wanted to have the ability to use GitHub pages, so I downloaded the CSV version. It contains tens of thousands of player entries, as well as hall of fame status, all-star statistics, and much more.

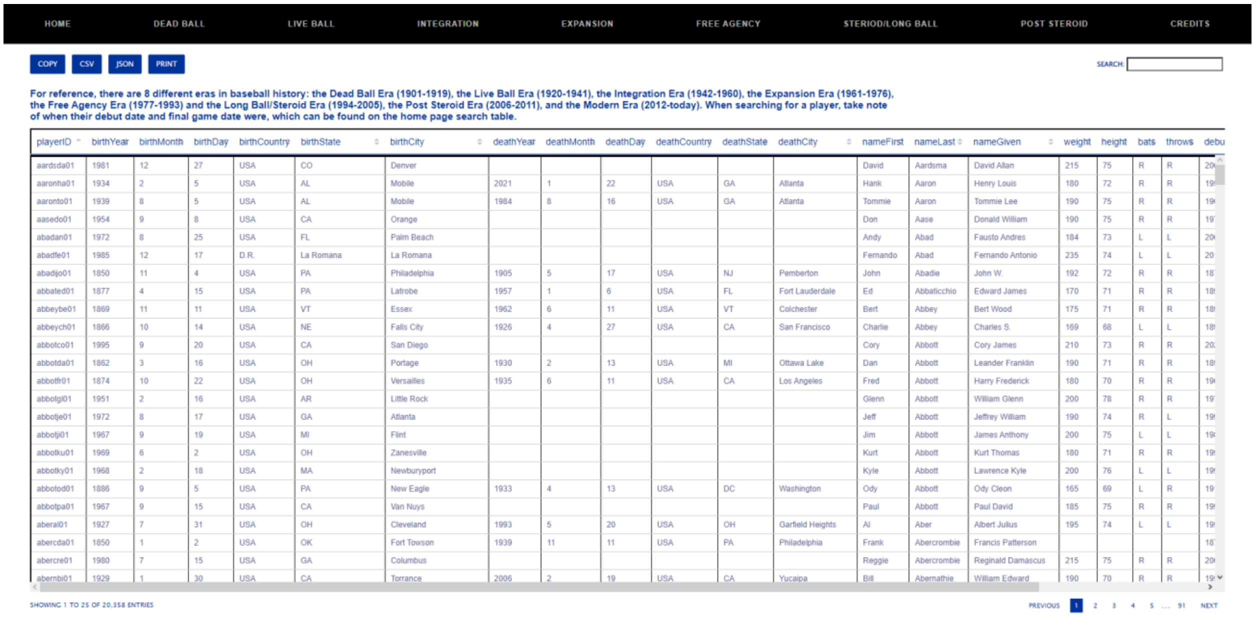
Exploratory Data Analysis

I only looked at one visualization and didn't receive much inspiration from it. http://dgwartney.github.io/tableau/ plots a few statistics from three steroid-era players (Barry Bonds, Mark McGwire, and Sammy Sosa) who were all known steroid users. Aside from this, I came up with pretty much everything about my visualization from scratch.

Design Evolution

At first, I was considering doing a line plot, but with the massive quantity of data, side-scrolling through all the results could take literal hours. I decided on a more condensed approach of using three scatter plots which, in my opinion, visualized the correlation between certain statistics much better than a line plot or another option would have. As for my proposal, I didn't have anything in it about what kind of plotting I was going to do, so I didn't deviate from it in any way.

Implementation

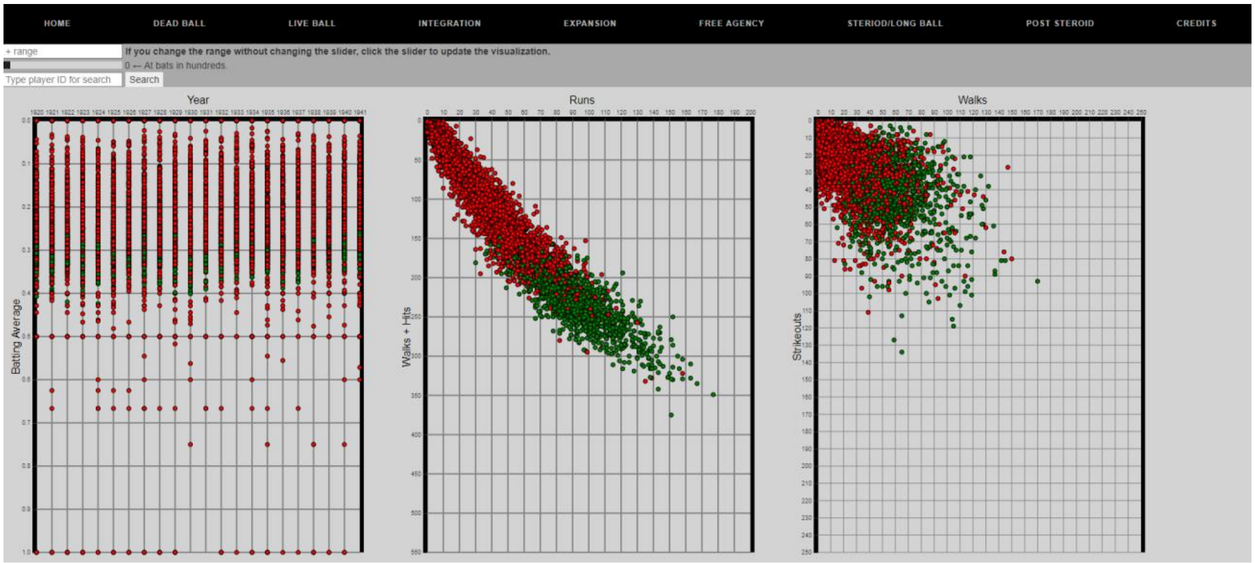


For reference, there are 8 different eras in baseball history: the Dead Ball Era (1901-1919), the Live Ball Era (1920-1941), the Integration Era (1942-1960), the Expansion Era (1961-1976), the Free Agency Era (1977-1993) and the Long Ball/Steroid Era (1994-2005), the Post Steroid Era (2006-2011), and the Modern Era (2012-today). When searching for a player, take note of when their debut date and final game date were, which can be found on the home page search table.

playerID	birthYear	birthMonth	birthDay	birthCountry	birthState	birthCity	deathYear	deathMonth	deathDay	deathCountry	deathState	deathCity	nameFirst	nameLast	nameGiven	weight	height	bats	throws	debu
aardsda01	1981	12	27	USA	CO	Denver							David	Aardsma	David Allan	215	75	R	R	20
aaronha01	1934	2	5	USA	AL	Mobile	2021	1	22	USA	GA	Atlanta	Hank	Aaron	Henry Louis	180	72	R	R	19
aaronto01	1939	8	5	USA	AL	Mobile	1984	8	16	USA	GA	Atlanta	Tommie	Aaron	Tommie Lee	190	75	R	R	19
aasedo01	1954	9	8	USA	CA	Orange							Don	Aase	Donald William	190	75	R	R	19
abadan01	1972	8	25	USA	FL	Palm Beach							Andy	Abad	Fausto Andres	184	73	L	L	20
abadfe01	1985	12	17	D.R.	La Romana	La Romana							Fernando	Abad	Fernando Antonio	235	74	L	L	20
abadijo01	1850	11	4	USA	PA	Philadelphia	1905	5	17	USA	NJ	Pemberton	John	Abadie	John W.	192	72	R	R	18
abbated01	1877	4	15	USA	PA	Latrobe	1957	1	6	USA	FL	Fort Lauderdale	Ed	Abbaticchio	Edward James	170	71	R	R	18
abbeybe01	1869	11	11	USA	VT	Essex	1962	6	11	USA	VT	Colchester	Bert	Abbey	Bert Wood	175	71	R	R	18
abbeych01	1866	10	14	USA	NE	Falls City	1926	4	27	USA	CA	San Francisco	Charlie	Abbey	Charles S.	169	68	L	L	18
abbotco01	1995	9	20	USA	CA	San Diego							Cory	Abbott	Cory James	210	73	R	R	20
abbotda01	1862	3	16	USA	OH	Portage	1930	2	13	USA	MI	Ottawa Lake	Dan	Abbott	Leander Franklin	190	71	R	R	18
abbotfr01	1874	10	22	USA	OH	Versailles	1935	6	11	USA	CA	Los Angeles	Fred	Abbott	Harry Frederick	180	70	R	R	19
abbotgl01	1951	2	16	USA	AR	Little Rock							Glenn	Abbott	William Glenn	200	78	R	R	19
abbotje01	1972	8	17	USA	GA	Atlanta							Jeff	Abbott	Jeffrey William	190	74	R	L	19
abbotji01	1967	9	19	USA	MI	Flint							Jim	Abbott	James Anthony	200	75	L	L	19
abbotku01	1969	6	2	USA	OH	Zanesville							Kurt	Abbott	Kurt Thomas	180	71	R	R	19
abbotky01	1968	2	18	USA	MA	Newburyport							Kyle	Abbott	Lawrence Kyle	200	76	L	L	19
abbotod01	1886	9	5	USA	PA	New Eagle	1933	4	13	USA	DC	Washington	Ody	Abbott	Ody Cleon	165	69	L	R	19
abbotpa01	1967	9	15	USA	CA	Van Nuys							Paul	Abbott	Paul David	185	75	R	R	19
aberal01	1927	7	31	USA	OH	Cleveland	1993	5	20	USA	OH	Garfield Heights	Al	Aber	Albert Julius	195	74	L	L	19
abercda01	1850	1	2	USA	OK	Fort Towson	1939	11	11	USA	PA	Philadelphia	Frank	Abercrombie	Francis Patterson					18
abercre01	1980	7	15	USA	GA	Columbus							Reggie	Abercrombie	Reginald Damascus	215	75	R	R	20
abernbi01	1929	1	30	USA	CA	Torrance	2006	2	19	USA	CA	Yucaipa	Bill	Abernathie	William Edward	190	70	R	R	19

SHOWING 1 TO 25 OF 20,358 ENTRIES

Figure 1: Home page containing database of all player information (not statistics)



+ range

If you change the range without changing the slider, click the slider to update the visualization.

0 <-- At bats in hundreds.

Type player ID for search

Search

Figure 2: Initial view of visualization

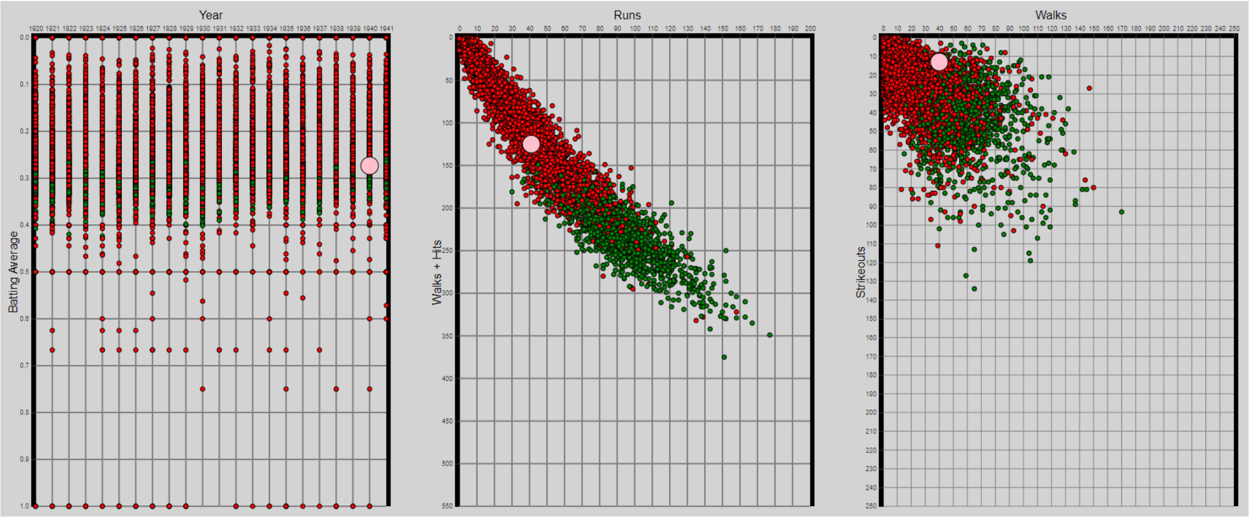


Figure 3: Clicking on a specific point

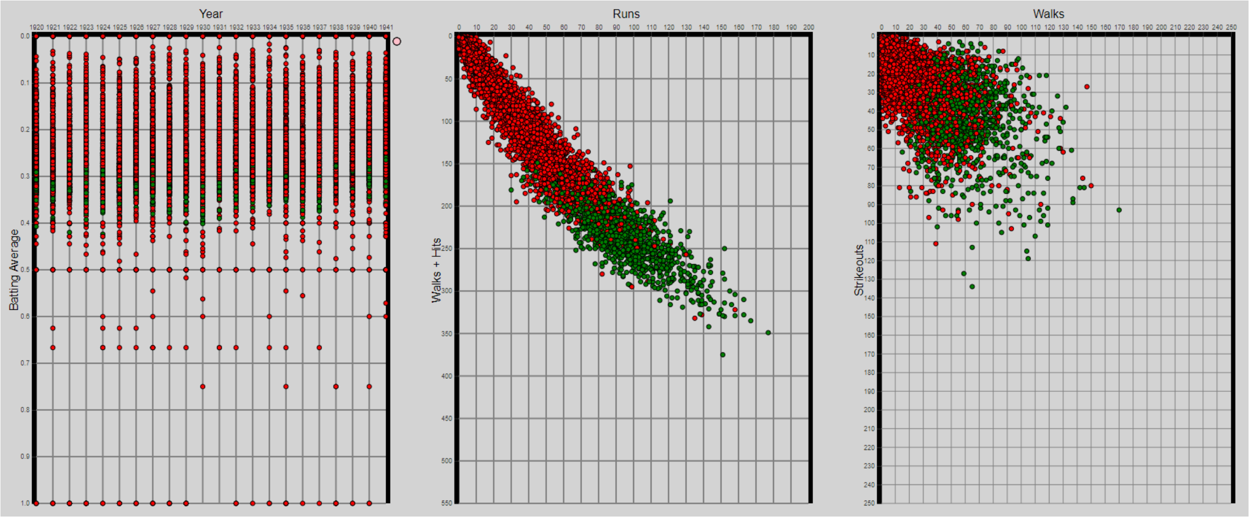
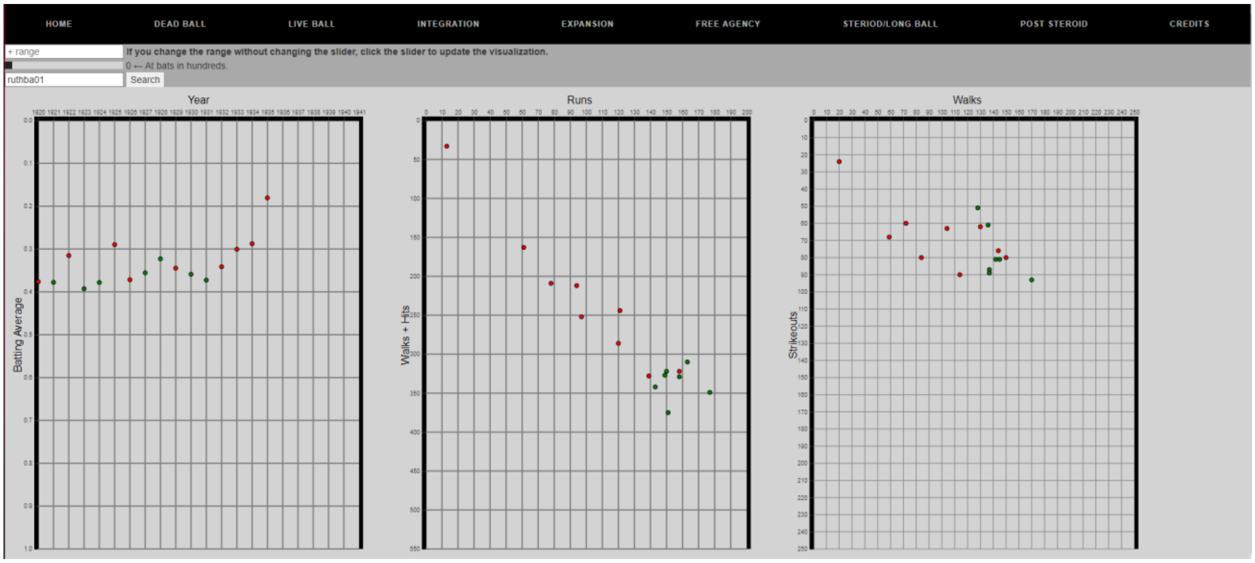


Figure 4: Clicking on a point you already clicked on moves it to the right side of the plot (limit of 250 per, can only move to 3 plots before space runs out)

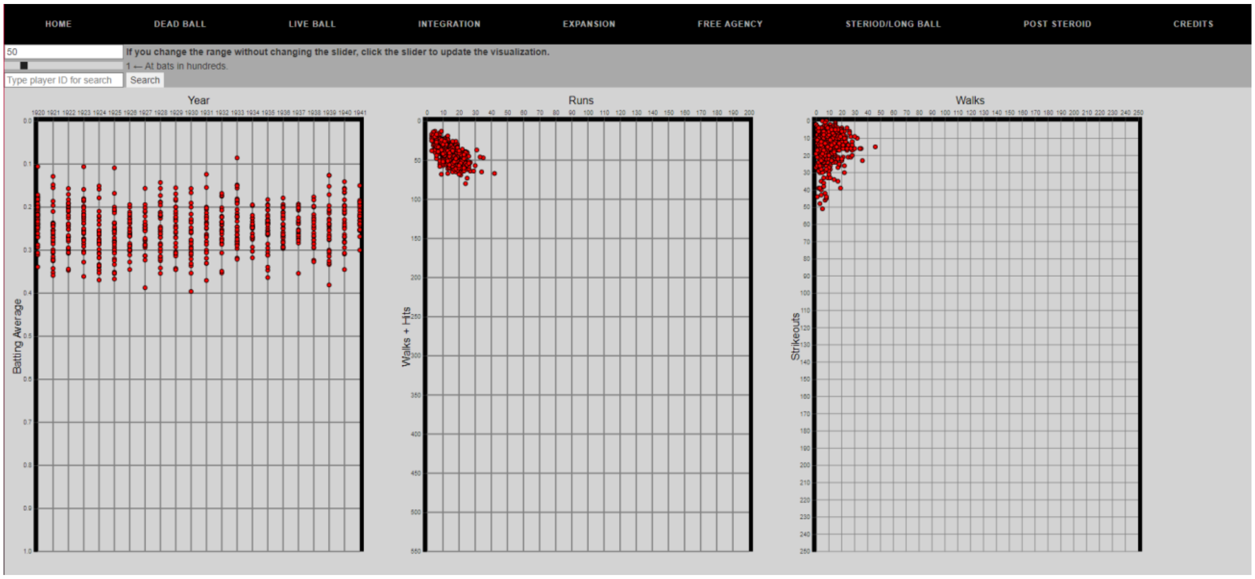


If you change the range without changing the slider, click the slider to update the visualization.

0 ← At bats in hundreds.

ruthba01

Figure 5: Visualization after searching for a player. In this case, ruthba01 (Babe Ruth)



50

If you change the range without changing the slider, click the slider to update the visualization.

1 ← At bats in hundreds.

Type player ID for search

Figure 5: Only showing players who have 100-150 at balls (inclusive).

Each picture here shows a different step in the interaction process. At first, the visualization is displaying all the data. You can then search for a specific player, click on random players in the visualization to highlight them, or filter players based on number of at bats. All three plots are linked, so clicking on a point from plot 1 will highlight that same point on the other plot, and the same goes for plots 2 and 3.

Evaluation

As I was already familiar with the correlation between different statistics and the difference in statistics based on era, I didn't necessarily learn much about the data. Though that may be the case, plotting the data gave me a better appreciation for what it represented and how different each era of baseball was offensively. There are dozens of visualizations that could be created with the database, so I'm sure there is so much more that I could learn and understand from visualizing it.

As for functionality, I feel that my visualization works well. Considering that each plot has 20,000+ points in it, it only takes about a second for the data to load initially. Searching for players is fast as well, only taking about a couple of seconds to search through all three plots and hide all extraneous points. As for the transitions, they can be a little bit laggy, but I'm not sure if that is a browser issue or a problem with my code as I wasn't able to do anything about it.