MLB Visualizations Process Book

Visualization for Data Science | Paul Rosen
DS-4630 / CS-5630 / CS-6630
December 6, 2024

Table of Contents

Basic Info

Project Title: **MLB Visualizations**

Team Members:
- Kacey Abbott
  - Email: kacey.abbott@utah.edu,
  - uNID: u0692178
- Kendall Ruth
  - Email: kendall.ruth@utah.edu,
  - uNID: u1481623
- Jaden Lee
  - Email: u1417827@utah.edu
  - uNID: u1417827

Project Repository:
https://github.com/dataviscourse2024/group-project-baseball-visualization-jkk-4.git

Website:
https://dataviscourse2024.github.io/group-project-baseball-visualization-jkk-4/MLB%20Visualization/index.html

Screencast:
https://youtu.be/ZIOLiJTlYgI

Process Book:
https://docs.google.com/document/d/1X0YOF7Xuc1rjW79ehWb5NFP2PI3OB1oxrM6s10vvAMs/preview

Background and Motivation

Our motivation for creating a data visualization centered around baseball statistics stems from our diverse but complementary interests. Two of the team members have a direct interest in baseball with one of us that plays on the university's club baseball team that brings firsthand experience and a deep understanding of the game's nuances and which data would be worth visualizing. Another member is passionate about the "Moneyball" approach, eager to explore how data-driven strategies can uncover new insights and trends in the sport. The third member, enthusiastic about contributing to the team's success, provides valuable support and fresh perspectives. Together, we aim to combine our unique strengths to produce a compelling and informative visualization that captures both the strategic and personal dimensions of baseball as well as how luck and chance affects the sport. Baseball already has a lot of statistics and visualizations, and we want to take what is already out there and improve the visualizations to be even better than they are now!

Related Work

Our project is inspired by several well-known tools and ideas in baseball analytics and data visualization. The Moneyball approach and sabermetrics showed how using data, like on-base percentage (OBP) and slugging percentage (SLG), can uncover hidden player value and improve team strategies. Baseball Savant, a popular MLB platform, influenced our project with examples of potential visualizations like its interactive tools like spray charts, heatmaps, and player rankings. We also looked at historical visualizations, such as trends in home runs or strikeouts over the years, to see how data can tell the story of how the game has changed. These examples helped us think about how to make our visualizations both informative and easy to use for fans and analysts.

Project Objectives

Our goal for this project is to create clear and interesting visualizations that explore baseball statistics from many years of the game. We want to find and show important trends that help explain how numbers affect team performance and player decisions. Since baseball already has a lot of analysis, we plan to focus on less common stats that might reveal new insights into what makes a player or team successful. By combining real-life baseball experience, a focus on data, and teamwork, we hope to make these stats easy to understand and useful for fans, players, and analysts. In the end, we want our visualizations to show how data shapes the game and highlight the power of working together with different ideas.

Data

We originally sourced as many common data sources as we could because we were unsure which dataset would be the most complete or if one may not have the data we wanted to visualize. After sourcing and exploring the data, we were able to get all of the information we needed from the top two datasets listed.

- **Lahman Database: Archive of team and player statistics going back to 1871**
  - **http://seanlahman.com/**
- **Baseball Almanac:**
  - **Baseball Almanac: MLB Stats, History, Records & Research | 1876-2024**
- Cot’s Baseball Contracts: Data for team contracts and payrolls
  - Cot's Baseball Contracts (baseballprospectus.com)
- Baseball Savant: Advanced player and team statistics plus available Statcast data
  - Baseball Savant: Statcast, Trending MLB Players and Visualizations |
- Fangraphs: Advanced player statistics for MLB, minor leagues, and international leagues

    - https://www.fangraphs.com/
- Baseball Reference: Complete player and team statistical data for Major League Baseball
    - https://www.baseball-reference.com/
- Kaggle Dataset: Various MLB information
    - MLB Player Digital Engagement Forecasting EDA (kaggle.com)
- Chadwick-Bureau: Collection of various current historical baseball data sources
    - https://www.chadwick-bureau.com/
- Retrosheet: Play-by-play and box score data extending back to the early 1900s
    - https://www.retrosheet.org

Data Processing

Our data processing involved selecting and organizing clean datasets that are readily available. Since baseball is already a highly analyzed sport, we aimed to choose some less common statistics and correlations to explore and potentially reveal new insights about player and team success. Our tasks included filtering the data to focus on these unique metrics, merging different datasets for a comprehensive view, and structuring the information for easy visualization. By avoiding extensive data scraping or cleaning, we concentrated on accurately representing and analyzing these unconventional stats to uncover meaningful correlations and patterns.

The Lahman Database is a detailed archive of baseball team and player statistics dating back to 1871. It includes data on player performance, team records, and game outcomes, making it a key resource for analyzing trends and the history of baseball. It only contains box score stats, so we had to do some calculations for deriving more advanced statistical data. Player biographical information and performance data was split across two datasets too so we had to join the data to retrieve the player name at the same time as their season-by-season data.

We also used web scraping methods on the Baseball Almanac website to collect over 10,000 baseball player card images for our project. This method provided a rich collection of player visuals, though most missing images were for players who played before 1945. The process allowed us to build a comprehensive image dataset to complement our statistical analysis**.**

Visualization Design

We planned to include visualizations that will include many different design aspects. Some visualizations that use categorical data might have used a bar chart and location-based data might have been shown with a map. We also considered using baseball themes to portray our data. For example, we could’ve displayed percentages as a diagram of how far a player runs around the bases or how full a stadium is. Batted ball distances should be shown radially and overlaid over a baseball diamond. We also intended to associate teams and stadiums with their colors or mascots.

Here are a few of the case subjects that we are planning to focus on with some extra ideas denoted by an asterisk:

- Date of birth of MLB players
    - Are the quantity of players in the MLB evenly distributed among birth months? How does player performance and salary change with age?
- Birth State/Country
    - Do players come disproportionately from places of lower latitudes or places of warmer temperatures?
- Home/Away Splits
    - Do certain teams win more at home or away during certain months depending on the average temperature?
- Park and Spending Factors
    - How do different stadiums affect run-scoring and other events? How about team payroll?
- Standard key baseball metrics visualization*
    - Include a player and team search with major stats and comparison to other players/teams. It should be sortable by player attributes.
- "Take Me Out to the Ballgame" music*
- Bat/ball mouse cursor game*

Must-Have Features:

- Ability to filter visualization based on data
- Map of states/birthplaces (WAR), Heatmap of states players were born in
- Should include American and international player data
- Include page our process book and reasoning/calculations/assumptions
- Multiple years of MLB players (going back to at least 2010)

Optional Features:

- Test your reaction speed "game" based on adjustable pitching speed. Batters box graphic with a scale ball that appears at random time. Must click within a certain time to get a "hit".
- Inclusion of all of the players biographical data (more than the state/birthplace), ie: ethnicity, race, height, weight.
- Find the greatest athlete by WAR in each year/month/days.
- Baseball fields per capita or population per MLB player/stadium

Project Progress Summary

Initially, we planned to create many advanced MLB data visualizations and interactive features,

but we found the project was too complicated to handle all at once. To make it manageable, we decided to focus on the most important visualizations like bar charts and heatmaps first, saving the more complex ideas for later. We encountered challenges such as organizing our project folders correctly, finding a new hosting service after Heroku stopped its free plans, and speeding up how our data loads. We overcame these issues by restructuring our files, moving to a different hosting platform, and improving our data handling methods. Now, our backend and frontend work smoothly together on our local computer. With more time, we hope to add the more complicated visualizations we originally envisioned, but for now, we have built a strong foundation with the key features.

Project Schedule

Team meeting schedule: All team members will be available every X day at Y time if needed, and preferably over zoom. Otherwise we will coordinate over text/zoom as needed.

Date	Event	Completed
8/30/2024	Announce your project	Yes
9/13/2024	Project Proposal	Yes
9/16/2024 @ 1:20 PM	Project Review with TA	Yes
10/2/2024	Finalize specific Ideas	Yes
10/15/2024	Data Scraped/Clean, basic website set up	Yes
10/25/2024	Milestone, a functional project prototype	Yes
11/1/2024	Peer feedback	Yes
11/8/2024	Make adjustments from peer feedback	Yes
11/15/2024	Make sure visualizations are correct and look good	Yes
11/22/2024	Project Screen-Cast	Yes
12/06/2024	Final project submission & group member evaluations	Yes

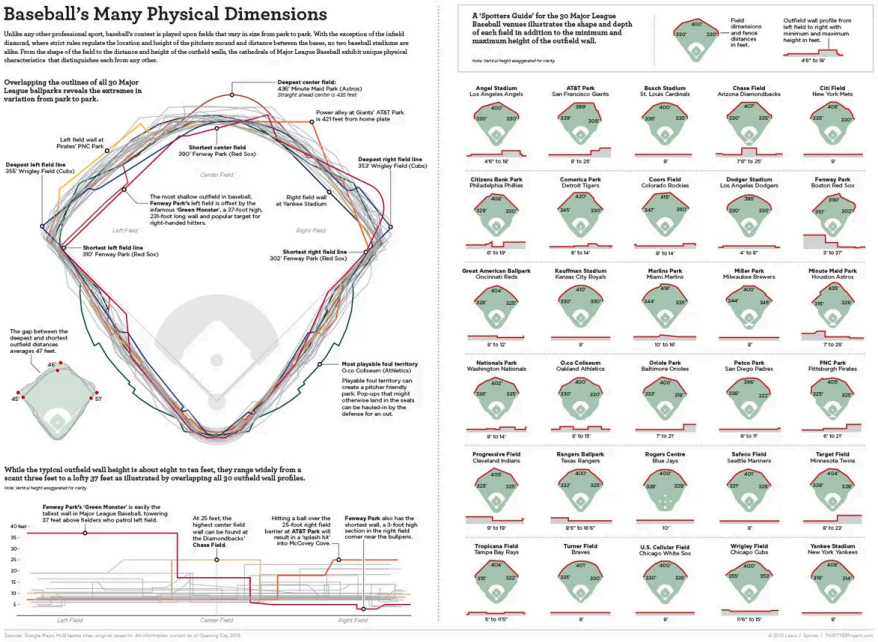
Original Prototypes/Sketches

---

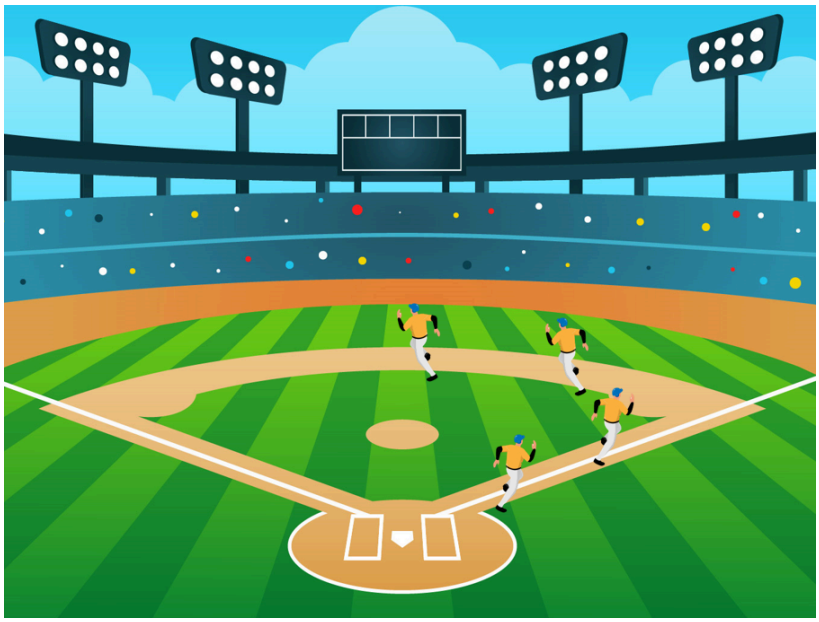
1. **USA map with applicable filters. Initially showing locations of MLB stadiums. Other filters to be applied can be a relationship map of where current players are playing vs where they were born or where they played in college. Heatmap of team wins by season/over a time period,**

heatmap of where players were born, heatmap of most popular team/fans across the USA, most hated team by region. How any/some of those changed over time.

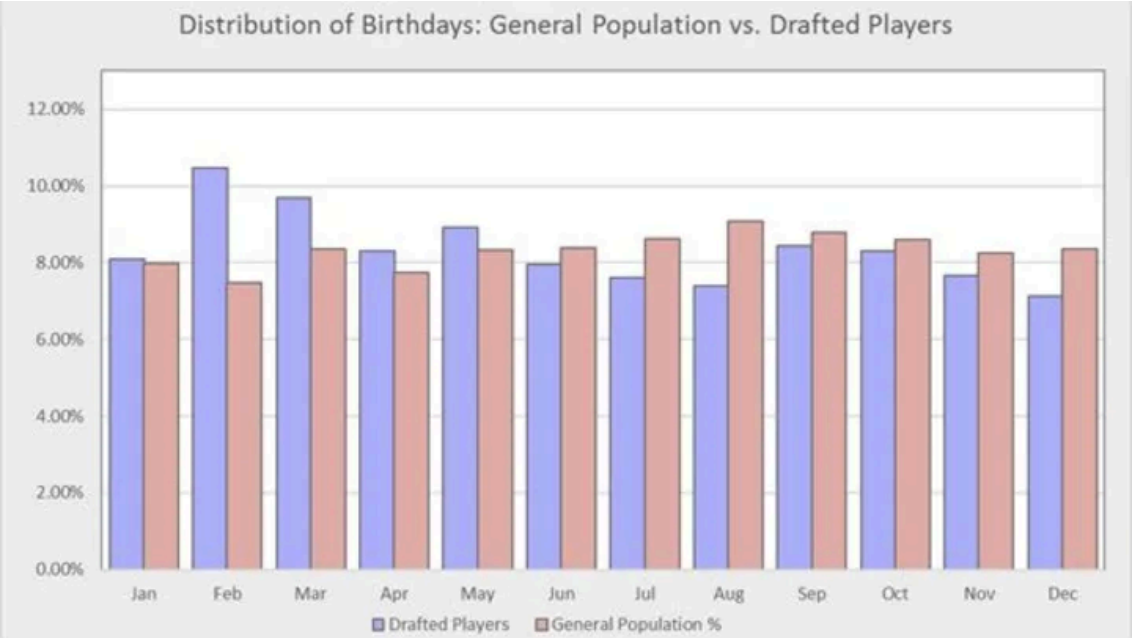
2. **Stadium background infographic. Use to show stadium stats, compare stadium stats or how players have performed at that stadium by game, season, or career.**

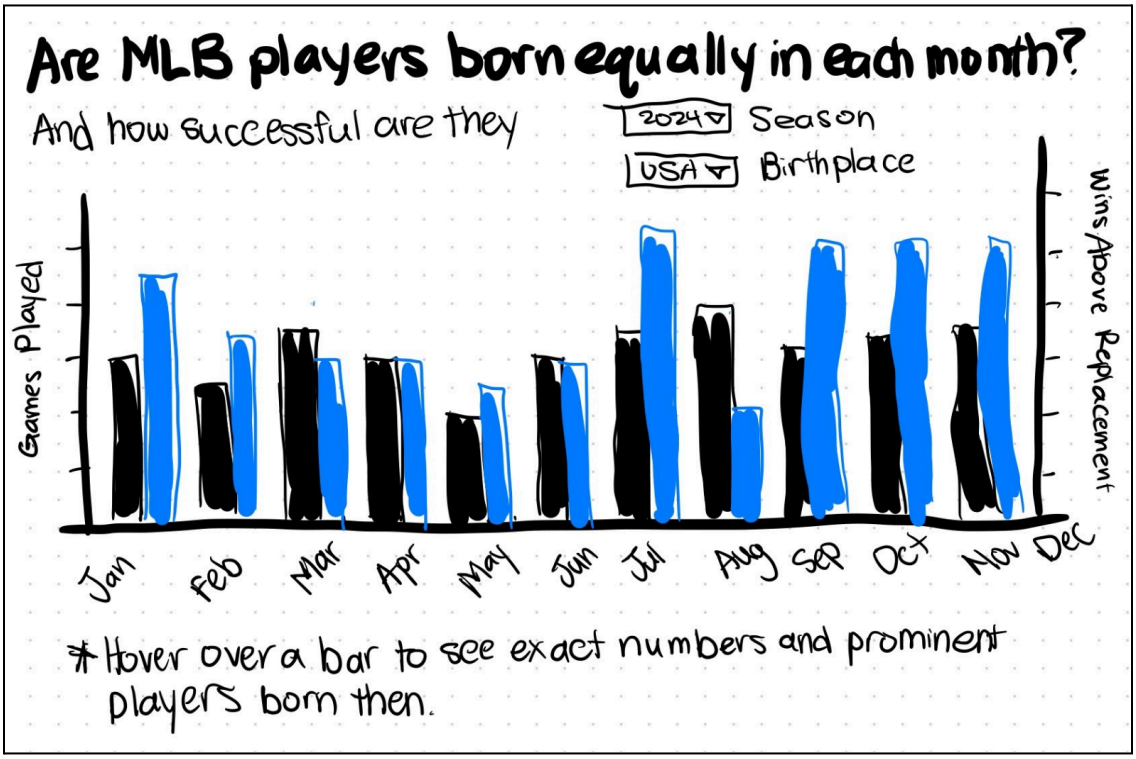


2.B Another version of the stadium background infographic. Players running around basses to show comparison of 0-100%. 25% being first base, 50% being second base, etc…



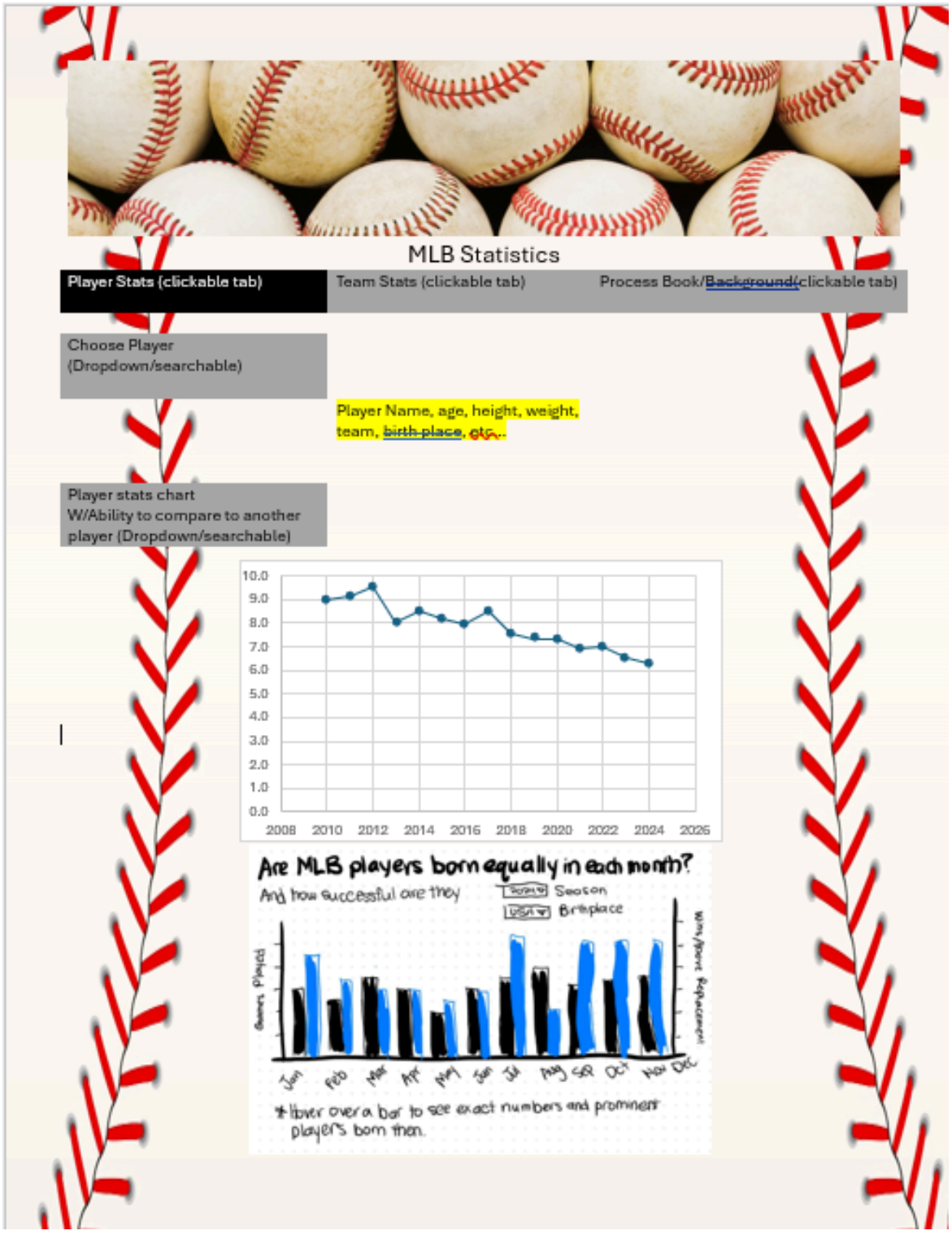
3. **Visualization like this, but with MLB data/ not NBA to showcase when the birthdays are for drafted MLB players.**





4. **Visualization for best player by month, similar to this one for the NBA**

Prototype


MLB Statistics
Player Stats (clickable tab)
Team Stats (clickable tab)
Process Book/Background(clickable tab)
Choose Player
(Dropdown/searchable)
Player Name, age, height, weight,
team, birth-place, etc..
Player stats chart
W/Ability to compare to another
player (Dropdown/searchable)
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
2008
2010
2012
2014
2016
2018
2020
2022
2024
2026
Are MLB players born equally in each month?
And how successful are they
Season
Birthplace
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
*Hover over a bar to see exact numbers and prominent
players born then.

MLB Statistics

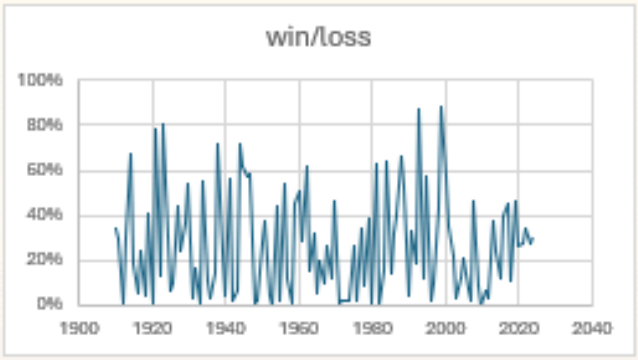
Player Stats (clickable tab) | Team Stats (clickable tab) | Process Book/~~Background~~(clickable tab)

Choose Team
(Dropdown/searchable menu)

Team Name, location, mascot, founded date, wins, etc..

Team stats chart
W/Ability to compare to another team(Dropdown/searchable)

Current season comparison, select year comparison, overall comparison (Wins, loses, runs, hits, etc..)



MLB Statistics

Player Stats (clickable tab)	Team Stats (clickable tab)	Process Book/Background(clickable tab)

Our Process Book/analysis/assumptions

Meeting Notes/Progression Records

9/19/2024:

We improved our definition of the project scope, how to get started, and what our timeline will be. We were considering a map visualization of trades between teams then a visualization of teams a single player has played for. There was also a proposed idea for a guessing game that could display player stats side-by-side in a comparative manner. Somebody brought up an interactive ball-strike simulation from the NY Times (Click Here) that could inspire future visualizations.

Projected Timeline:

- Choose a web framework.
- Finalize our visualization (guess the player game).
- Make an outline
- Find the APIs and figure out how to manage data.
- Code up the visualizations.
- Deploy with AWS.

For tasks still to be done, we need to determine who our user base is and what visualizations will be most useful to them. From that, pick features that are essential and then what is nice to have but isn't necessary.Visualizations should be coherent around those users and we should target them to present a certain claim.

10/02/2024:

We found an NBA visualization that could serve as an example for what we want to implement, although ours will be data dependent and geared towards baseball. We plan to start with simple infographics then clean up, beautify, and add infographics as time allows.

The next steps are to gather data starting with just this year. We'll start with Baseball Reference and use Baseball Savant as a secondary data set. There is also a data scraping library in Python called pybaseball. Two weeks from now, we want to have the basic data wrangled with a base website setup and a good idea of what visualizations will look like/what we want based on the data we have.

10/3/2024:

We found a couple more datasets that might be suitable:

- CSV Dataset: http://www.seanlahman.com

    - (https://drive.google.com/drive/folders/1C_CCzkalzoe9fxDsuYrXowJ6cdvTEuJ8?usp=drive_link)
- Unique data analysis on MLB injury luck (Doing the Math: Yankees Injury Woes are Unprecedented | by Jordan Siff | Medium)
- Unique analysis on switching teams/vs rates (There's no I in team but is there a team in baseball? (substack.com))
- Card images:
    - REST APIs : r/baseballcards (reddit.com)
    - python - Trying to automate the download of images from psacard.com but running into PerimeterX issues - Stack Overflow
    - 2023 Topps Arizona Diamondbacks #ARI-2 Seth Beer | Trading Card Database (tcdb.com)

10/14/2024:

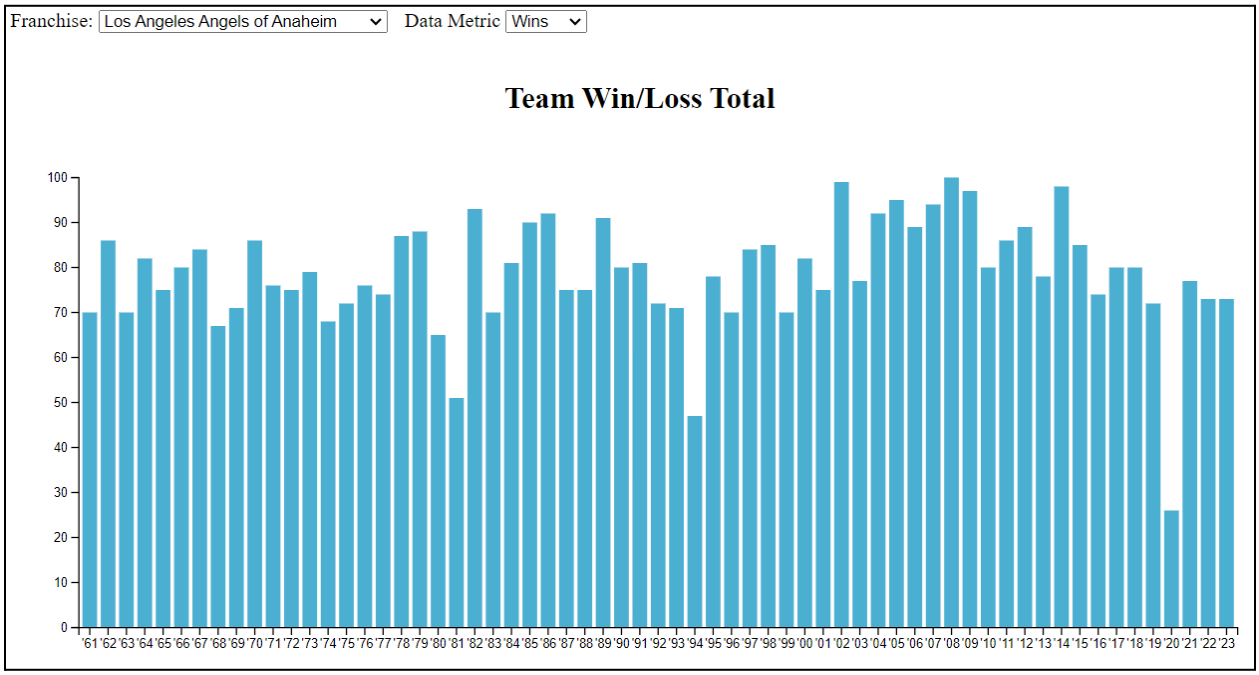
We've uploaded CSV/JSON dataset files to github and are now planning on setting up the main player image in the form of a baseball card with stats that update based on player. We are working on getting either a downloadable list of player images or cards. Here is an example below.

10/22/2024:

Next steps will be to get some dummy data then filter players by a start year and end year.

Kendall plans to populate a dropdown with player names and develop a table with player information. Jaden will create a wOBA chart, and Kacey will hard code a player name to start that will eventually link to the player name dropdown.

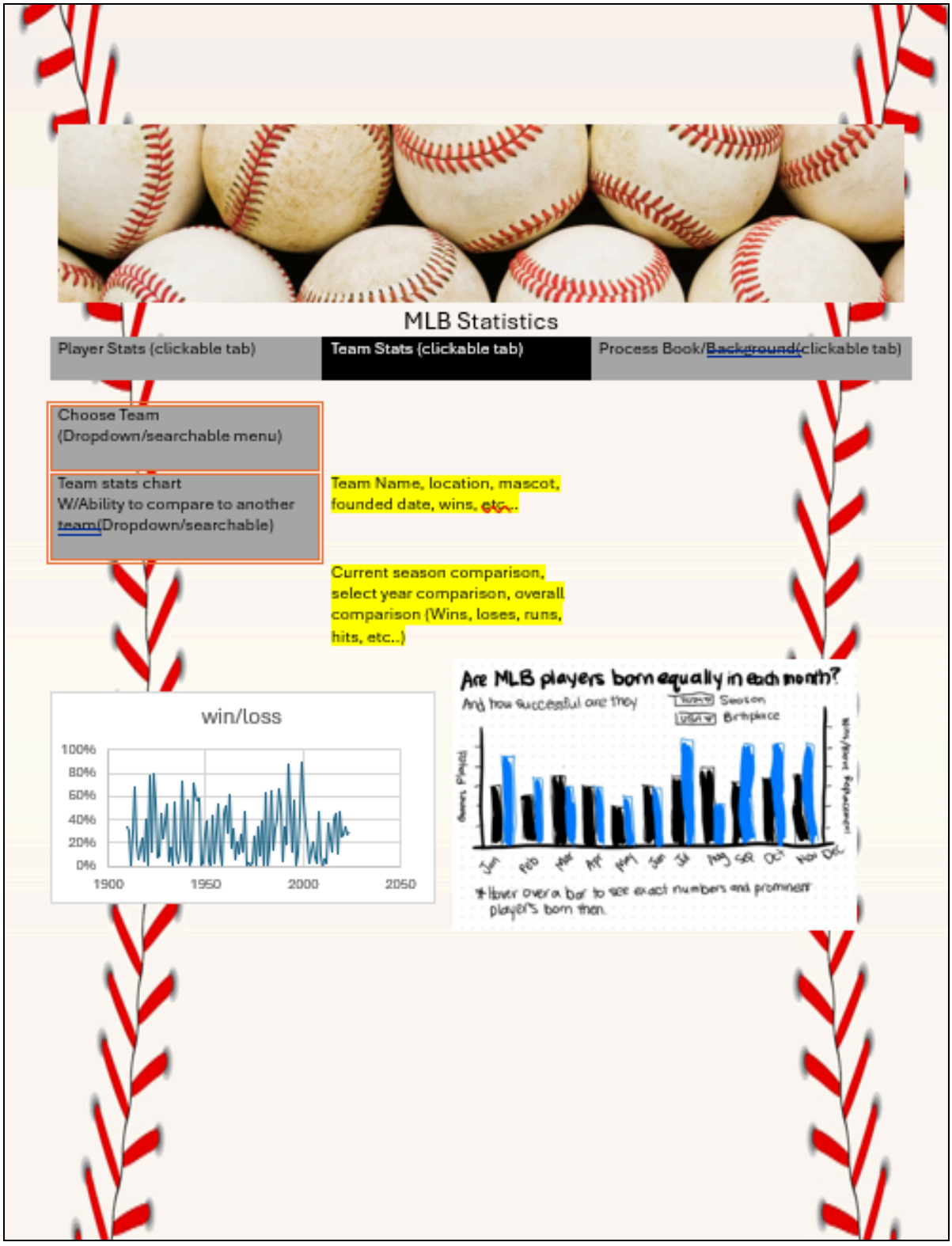
Here is a proof of concept visualization using the Lahman data and a couple dropdowns.



10/24/2024:

We've updated our draft website concept. The website prompts the question of why we use wOBA nowadays. It traces the differences between wOBA and other modern statistics and traditional rate stats.


MLB Statistics
Player Stats (clickable tab)
Team Stats (clickable tab)
Process Book/Background(clickable tab)
Choose Player
(Dropdown/searchable)
Player stats chart
W/Ability to compare to another
player (Dropdown/searchable)
Player Name, age, height, weight,
team, birth place, etc..
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
2008
2010
2012
2014
2016
2018
2020
2022
2024
2026
KODAI SENGA
NEW YORK METS
BOWMAN BRIEFING
MAJOR LEAGUE PITCHING RECORD


MLB Statistics
Player Stats (clickable tab)
Team Stats (clickable tab)
Process Book/Background(clickable tab)
Choose Team
(Dropdown/searchable menu)
Team stats chart
W/Ability to compare to another team(Dropdown/searchable)
Team Name, location, mascot, founded date, wins, etc...
Current season comparison, select year comparison, overall comparison (Wins, loses, runs, hits, etc..)
win/loss
100%
80%
60%
40%
20%
0%
1900
1950
2000
2050
Are MLB players born equally in each month?
And how successful are they
Season
Birthplace
*Hover over a bar to see exact numbers and prominent players born then.

MLB Statistics

Player Stats (clickable tab) | Team Stats (clickable tab) | Process Book/Background(clickable tab)

Our Process Book/analysis/assumptions

Attached

Old baseball stats are being replaced. Why?

BAA

(Ted Williams, Tony Gwynn)

OBP

(Billy Beane)

SLG

(Barry Bonds)

Now is a home run 4x as valuable as a single?

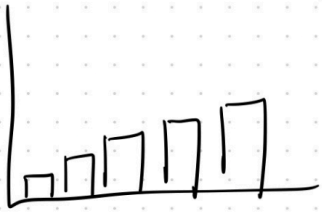
Shift away from bases → How many runs does someone create?

Not to runs or RBIs that are dependent on situation and batting order

Enter wOBA...

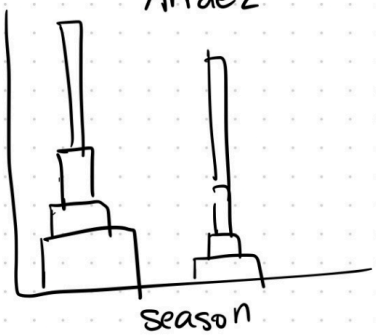
weight each result by the historical runs an event creates.

The more runs you score, the more you win.

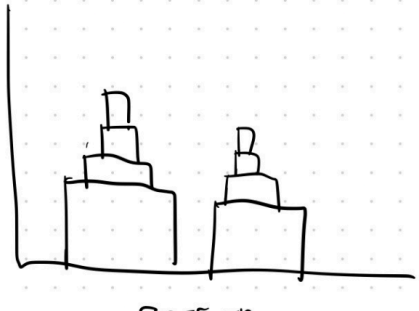


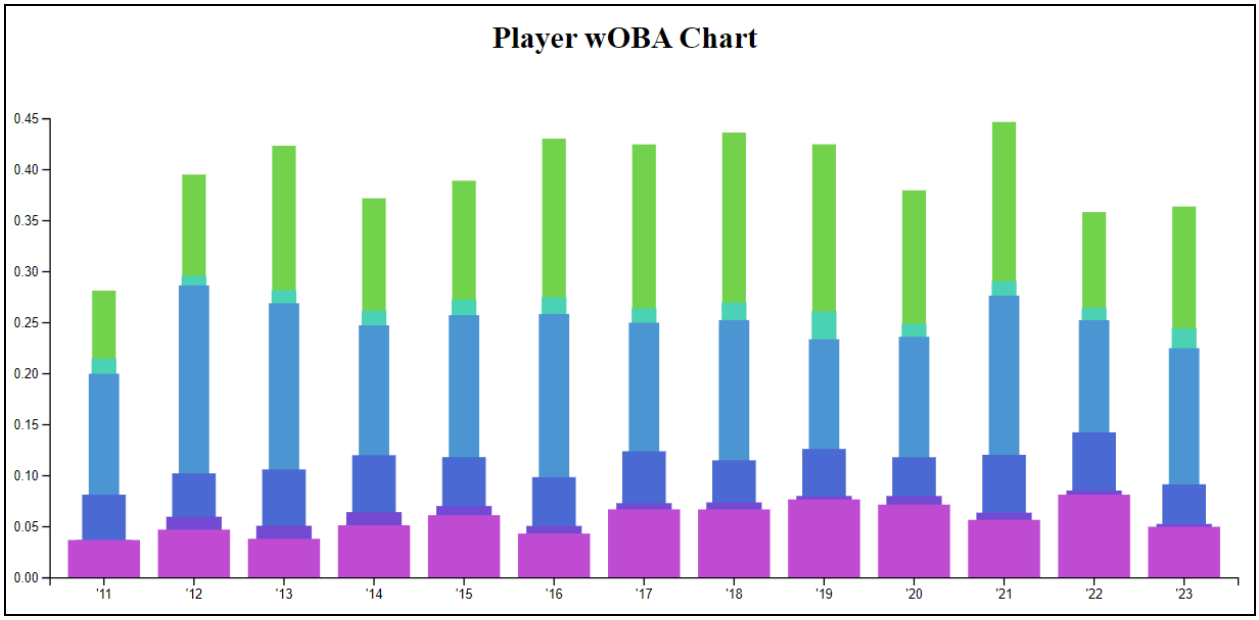
Compare players below

Arraez



Suarez





This is the wOBA chart we created. The weights that describe the width of each color band are derived from this [wOBA Weights CSV]. The axes are missing labels but the left is the proportion of plate appearances and the bottom is years in a player's career. They adjust to the data. The width of each level of the tower is based on the relative value of a batted ball event.

playerID	yearID	teamID	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP
salmoti01	1992	CAL	23	79	8	14	1	0	2	6	1	1	11	23	1	1	0	1	1
salmoti01	1993	CAL	142	515	93	146	35	1	31	95	5	6	82	135	5	5	0	8	6
salmoti01	1994	CAL	100	373	67	107	18	2	23	70	1	3	54	102	2	5	0	3	3
salmoti01	1995	CAL	143	537	111	177	34	3	34	105	5	5	91	111	2	6	0	4	9
salmoti01	1996	CAL	156	581	90	166	27	4	30	98	4	2	93	125	7	4	0	3	8
salmoti01	1997	ANA	157	582	95	172	28	1	33	129	9	12	95	142	5	7	0	11	7
salmoti01	1998	ANA	136	463	84	139	28	1	26	88	0	1	90	100	5	3	0	10	4
salmoti01	1999	ANA	98	353	60	94	24	2	17	69	4	1	63	82	2	0	0	6	7
salmoti01	2000	ANA	158	568	108	165	36	2	34	97	0	2	104	139	5	6	0	2	14
salmoti01	2001	ANA	137	475	63	108	21	1	17	49	9	3	96	121	4	8	0	2	11
salmoti01	2002	ANA	138	483	84	138	37	1	22	88	6	3	71	102	3	7	0	7	6
salmoti01	2003	ANA	148	528	78	145	35	4	19	72	3	1	77	93	3	10	0	6	12
salmoti01	2004	ANA	60	186	15	47	7	0	2	23	1	0	14	41	0	2	0	4	2
salmoti01	2006	LAA	76	211	30	56	8	2	9	27	0	2	29	44	1	3	0	1	8

We decided to include baseball cards and tables to show former ways of displaying statistical data. It is not very easy to read, that's for sure.

Note: We initially tried using Next.js to make web development easier. However, it was difficult to integrate D3 into that, and our team didn't have much experience using that framework as a whole. React would add a whole extra set of debugging issues too. As a result, we have, for now, stuck to using a similar Python server and JavaScript setup that we have been using to complete the homework assignments.

11/22/2024:

We got the website/visualization wrapped up for the project screencast, knowing that we still have some additional items to add for the final submission (the screencast can be viewed here: MLB Visualization Screencast CS-5630/CS-6630)

Screenshots of our current website will be shown below as well. They will be shown in order of the Main Page, Player Comparison Page, and the Additional Context Page.

The Main page gives the user the ability to view stats on any of the 21,000+ MLB baseball players from 1871 to 2023. The main two categories of stats that we chose to focus on are batting and pitching, along with additional statistics derived from these core statistics like the Rate Statistics and Modern Statistics.

The Player Comparison page, which is still being worked on, will show similar graphics but in a side by side comparison for the players that the user chooses.

Finally, the Additional Context page gives all of the abbreviations, data sources, assumptions, and limitations of this visualization.

CS-5630 / CS-6630 MLB Player Explorer

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

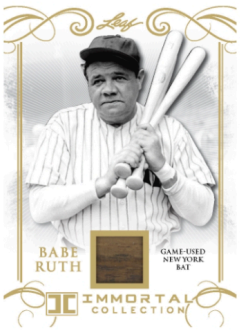
Player Comparison | Additional Context

Select a Player: Babe Ruth

Batting or Pitching Stats: Batting

Core Statistics

Baseball collectors have long prized player trading cards based on rarity and quality. Cards of the game's best players could fetch millions of dollars, but how can we objectively judge players? On the backs of trading cards, you'd find critical stat information for a given player. This popularized counting stats like the ones shown in the table below.



Year	Team	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP
1914	Boston Red Sox	5	10	1	2	1	0	0	2	0	0	0	4	0	0	0	0	0
1915	Boston Red Sox	42	92	16	29	10	1	4	21	0	0	9	23	0	0	2	0	0
1916	Boston Red Sox	67	136	18	37	5	3	3	15	0	0	10	23	0	0	4	0	0
1917	Boston Red Sox	52	123	14	40	6	3	2	12	0	0	12	18	0	0	7	0	0
1918	Boston Red Sox	95	317	50	95	26	11	11	66	6	0	58	58	0	2	3	0	0
1919	Boston Red Sox	130	432	103	139	34	12	29	114	7	0	101	58	0	6	3	0	0
1920	New York Yankees	142	457	158	172	36	9	54	137	14	14	150	80	0	3	5	0	0
1921	New York Yankees	152	540	177	204	44	16	59	171	17	13	145	81	0	4	4	0	0
1922	New York Yankees	110	406	94	128	24	8	35	99	2	5	84	80	0	1	4	0	0
1923	New York Yankees	152	522	151	205	45	13	41	131	17	21	170	93	0	4	3	0	0
1924	New York Yankees	153	529	143	200	39	7	46	121	9	13	142	81	0	4	6	0	0
1925	New York Yankees	98	359	61	104	12	2	25	66	2	4	59	68	0	2	6	0	0
1926	New York Yankees	152	495	139	184	30	5	47	150	11	9	144	76	0	3	10	0	0
1927	New York Yankees	151	540	158	192	29	8	60	164	7	6	137	89	0	0	14	0	0
1928	New York Yankees	154	536	163	173	29	8	54	142	4	5	137	87	0	3	8	0	0
1929	New York Yankees	135	499	121	172	26	6	46	154	5	3	72	60	0	3	13	0	0
1930	New York Yankees	145	518	150	186	28	9	49	153	10	10	136	61	0	1	21	0	0
1931	New York Yankees	145	534	149	199	31	3	46	163	5	4	128	51	0	1	0	0	0
1932	New York Yankees	133	457	120	156	13	5	41	137	2	2	130	62	0	2	0	0	0
1933	New York Yankees	137	459	97	138	21	3	34	103	4	5	114	90	0	2	0	0	0
1934	New York Yankees	125	365	78	105	17	4	22	84	1	3	104	63	0	2	0	0	0
1935	Boston Braves	28	72	13	13	0	0	6	12	0	0	20	24	0	0	0	0	2

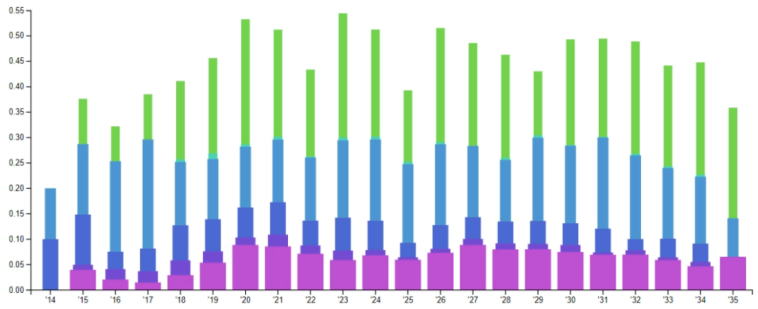
Rate Statistics

This evolved into rate statistics like Batting Average (BAA), On-Base Percentage (OBP), and Slugging (SLG) that attempted to quantify a batter's performance on a per opportunity basis.

--Visualizations in work--

Modern Statistics

However, modern sabermetricians are focused on winning behaviors. Thus, we redefined our offensive production metrics not to evaluate hits or bases, but runs and wins. wOBA (Weighted On-Base Average) of a player uses linear weights to assign run values to each type of event when players attempt to reach via a hit.



MLB Visualizations

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

Back to Main | Additional Context

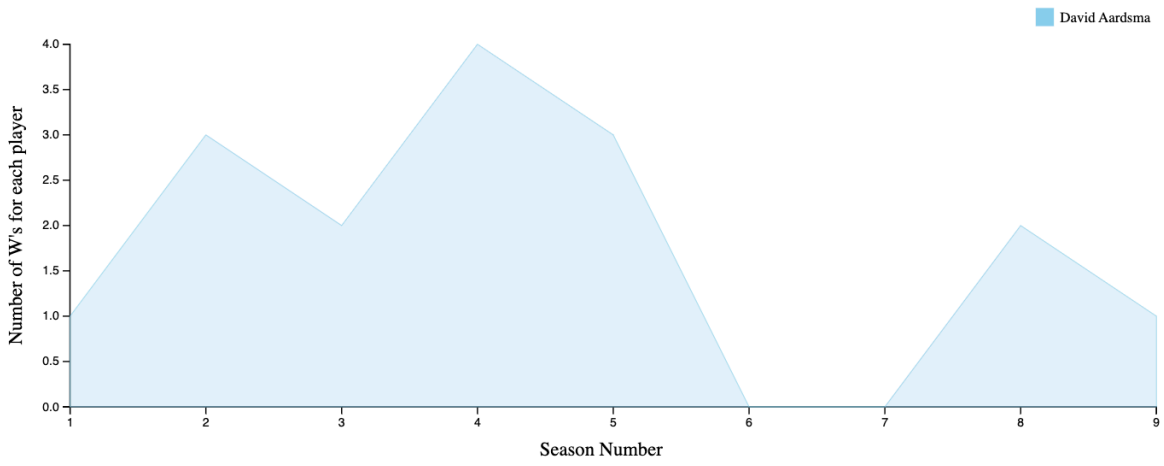
Select Player 1: David Aardsma

Select Player 2: David Aardsma

Pitching or Batting Stats Comparison: Pitching

Chosen Stat: W

Player Comparison



CS-5630 / CS-6630 MLB Visualizations

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

Back to Main | Player Comparison

MLB Visualizations Context

This page provides context for the MLB Visualizations project. Here you can find additional information explaining what each section of the main page is about.

The project aims to explore the performance of Major League Baseball players through various visualizations and statistical metrics. It focuses on rate statistics and modern advanced metrics to better understand the value of different players.

Baseball Statistics Definitions

Pitching Acronyms

- **W:** Wins - Number of games the pitcher was the winning pitcher.
- **L:** Losses - Number of games the pitcher was the losing pitcher.
- **G:** Games - Total number of games the pitcher appeared in.
- **GS:** Games Started - Number of games the pitcher started.
- **CG:** Complete Games - Number of games the pitcher completed without relief.
- **SHO:** Shutouts - Number of complete games where the pitcher allowed no runs.
- **SV:** Saves - Number of games the pitcher successfully saved.
- **IPouts:** Innings Pitched Outs - Total number of outs the pitcher achieved (divide by 3 for innings pitched).
- **H:** Hits - Total number of hits allowed by the pitcher.
- **ER:** Earned Runs - Total number of earned runs allowed by the pitcher.
- **HR:** Home Runs - Total number of home runs allowed by the pitcher.
- **BB:** Walks (Bases on Balls) - Total number of batters walked by the pitcher.
- **SO:** Strikeouts - Total number of batters struck out by the pitcher.
- **BAOpp:** Batting Average Against - Batting average of opposing hitters against the pitcher.
- **ERA:** Earned Run Average - Average number of earned runs allowed per 9 innings pitched.
- **IBB:** Intentional Walks - Total number of intentional walks issued by the pitcher.
- **WP:** Wild Pitches - Total number of wild pitches thrown by the pitcher.
- **HBP:** Hit by Pitch - Total number of batters hit by the pitcher.
- **BK:** Balks - Total number of balks committed by the pitcher.
- **BFP:** Batters Faced by Pitcher - Total number of batters faced by the pitcher.
- **GF:** Games Finished - Number of games the pitcher was the last pitcher in the game.
- **R:** Runs - Total number of runs allowed by the pitcher.
- **SH:** Sacrifice Hits - Total number of sacrifice bunts allowed by the pitcher.
- **SF:** Sacrifice Flies - Total number of sacrifice flies allowed by the pitcher.
- **GIDP:** Grounded Into Double Plays - Total number of double plays induced by the pitcher.

Batting Acronyms

- **G:** Games Played - The number of games in which the player appeared.
- **AB:** At Bats - The number of official at-bats (excludes walks, sacrifices, etc.).
- **R:** Runs Scored - The total number of runs the player scored.
- **H:** Hits - The number of times the player successfully reached first base by hitting the ball.
- **2B:** Doubles - The number of hits where the player reached second base.
- **3B:** Triples - The number of hits where the player reached third base.
- **HR:** Home Runs - The number of hits where the player hit the ball out of play, allowing them to round all the bases and score.
- **RBI:** Runs Batted In - The number of runs that scored as a direct result of the player's at-bat.
- **SB:** Stolen Bases - The number of times the player successfully stole a base.
- **CS:** Caught Stealing - The number of times the player was caught attempting to steal a base.
- **BB:** Base on Balls / Walks - The number of times the player was awarded first base after taking four balls outside the strike zone.
- **SO:** Strikeouts - The number of times the player struck out.
- **IBB:** Intentional Base on Balls - The number of times the player was intentionally walked by the opposing pitcher.
- **HBP:** Hit By Pitch - The number of times the player reached base after being hit by a pitch.
- **SH:** Sacrifice Hits / Sacrifice Bunts - The number of times the player successfully bunted the ball to advance a runner.
- **SF:** Sacrifice Flies - The number of fly balls that resulted in a run scoring (but the batter being out).
- **GIDP:** Grounded Into Double Play - The number of times the player grounded into a double play, resulting in two outs.

wOBA

- **wOBA (Weighted On-Base Average):** An advanced baseball statistic that measures a player's overall offensive contributions more accurately than traditional metrics like batting average, on-base percentage (OBP), or slugging percentage (SLG).
- **Purpose:** Evaluates a hitter's overall offensive performance by combining the ability to reach base and the quality of hits.
- **Weighted Events:**
  - Walks and hit-by-pitches are valued less than singles.
  - Doubles, triples, and home runs receive progressively higher weights based on their run-producing value.
- **Scale:** wOBA is scaled similarly to on-base percentage (OBP). A league-average wOBA is typically around **0.320 to 0.340**, depending on the era and league.

Data Scources

The data used in this project has been sourced from the Sean Lahman Baseball Database (http://www.seanlahman.com/). The Lahman dataset is an extensive collection of baseball statistics that includes information on over 21,000 players throughout the history of Major League Baseball. This rich dataset provided us with player statistics spanning 1871 to 2023 such as games played, runs scored, hits, and other critical metrics that are displayed in our visualizations.

The baseball card images were compiled by scraping the Baseball Almanac website (https://www.baseball-almanac.com/). Using this approach, we were able to obtain nearly 10,000 player images, which we use to complement the statistical data from the Lahman dataset. These images add a visual element that makes it easier to connect statistics to real players, enhancing the overall experience.

For the players that we were unable to find images for, we use a default placeholder image. This ensures that every player represented in our dataset has some form of imagery, even if specific photos are unavailable. The combination of comprehensive player data and visual elements provides a more engaging and informative experience for users exploring Major League Baseball history and statistics.

Note: The majority of player images not found were from the early years of baseball. Of the over 21,000 players in our data, more than 10,000 of them were born before 1945.

12/5/2024:

This is the wrap up of the final project. We completed the last couple visualizations we wanted to implement, hosted our backend and linked our front end github pages, along with embedding our youtube video and project process book.

On the main page, users can select a player and choose to see their batting or pitching statistics. The values of those dropdowns dictate which charts are shown. In each case, they are interactive. The line and bar charts show career data split along the season they played. By clicking a circle or a bar, the user can select the season to be portrayed by the pie chart in the center. This shows the distribution of results when the selected player was at the plate. Clicking any slice in the pie chart can change the selected statistic shown on the y-axis in the bar chart. This way, users can see how a player's statistical performance shifted throughout their career. The modern statistics chart shows expected runs produced by a certain player. The events are described in the legend and increase in value the higher up the chart you are. Hovering over a given box will show the precise number of expected runs produced by a certain type of results for a player in a given season.

MLB Visualizations

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

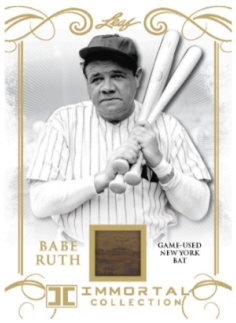
Player Comparison | Additional Context

Select a Player: Babe Ruth

Batting or Pitching Stats: Batting

Core Statistics

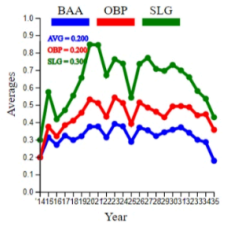
Baseball collectors have long prized player trading cards based on rarity and quality. Cards of the game's best players could fetch millions of dollars, but how can we objectively judge players? On the backs of trading cards, you'd find critical stat information for a given player. This popularized counting stats like the ones shown in the table below.

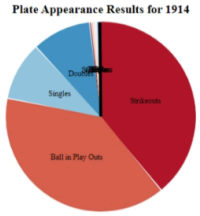


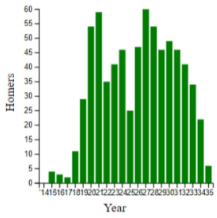
Year	Team	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	IBB	HBP	SH	SF	GIDP
1914	Boston Red Sox	5	10	1	2	1	0	0	2	0	0	0	4	0	0	0	0	0
1915	Boston Red Sox	42	92	16	29	10	1	4	21	0	0	9	23	0	0	2	0	0
1916	Boston Red Sox	67	136	18	37	5	3	3	15	0	0	10	23	0	0	4	0	0
1917	Boston Red Sox	52	123	14	40	6	3	2	12	0	0	12	18	0	0	7	0	0
1918	Boston Red Sox	95	317	50	95	26	11	11	66	6	0	58	58	0	2	3	0	0
1919	Boston Red Sox	130	432	103	139	34	12	29	114	7	0	101	58	0	6	3	0	0
1920	New York Yankees	142	457	158	172	36	9	54	137	14	14	150	80	0	3	5	0	0
1921	New York Yankees	152	540	177	204	44	16	59	171	17	13	145	81	0	4	4	0	0
1922	New York Yankees	110	406	94	128	24	8	35	99	2	5	84	80	0	1	4	0	0
1923	New York Yankees	152	522	151	205	45	13	41	131	17	21	170	93	0	4	3	0	0
1924	New York Yankees	153	529	143	200	39	7	46	121	9	13	142	81	0	4	6	0	0
1925	New York Yankees	98	359	61	104	12	2	25	66	2	4	59	68	0	2	6	0	0
1926	New York Yankees	152	495	139	184	30	5	47	150	11	9	144	76	0	3	10	0	0
1927	New York Yankees	151	540	158	192	29	8	60	164	7	6	137	89	0	0	14	0	0
1928	New York Yankees	154	536	163	173	29	8	54	142	4	5	137	87	0	3	8	0	0
1929	New York Yankees	135	499	121	172	26	6	46	154	5	3	72	60	0	3	13	0	0
1930	New York Yankees	145	518	150	186	28	9	49	153	10	10	136	61	0	1	21	0	0
1931	New York Yankees	145	534	149	199	31	3	46	163	5	4	128	51	0	1	0	0	0
1932	New York Yankees	133	457	120	156	13	5	41	137	2	2	130	62	0	2	0	0	0
1933	New York Yankees	137	459	97	138	21	3	34	103	4	5	114	90	0	2	0	0	0
1934	New York Yankees	125	365	78	105	17	4	22	84	1	3	104	63	0	2	0	0	0
1935	Boston Braves	28	72	13	13	0	0	6	12	0	0	20	24	0	0	0	0	2

Rate Statistics

For batters, this evolved into rate statistics like Batting Average (BAA), On-Base Percentage (OBP), and Slugging (SLG) that attempted to quantify a batter's performance on a per opportunity basis.

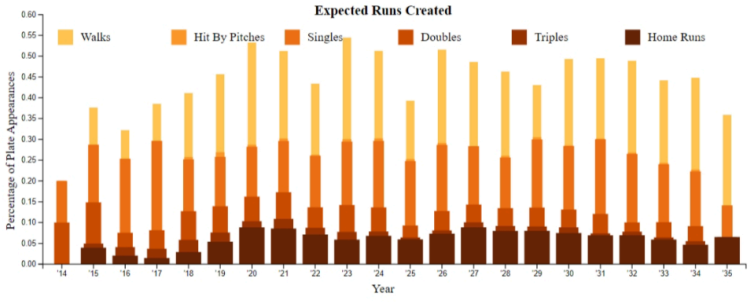






Modern Statistics

However, modern sabermetricians are focused on winning behaviors. Thus, we redefined our offensive production metrics not to evaluate hits or bases, but runs and wins. wOBA (Weighted On-Base Average) is a stat that uses linear weights to assign expected run values to each type of event when players attempt to reach via a hit. Home runs are the most valuable, triples next, etc.



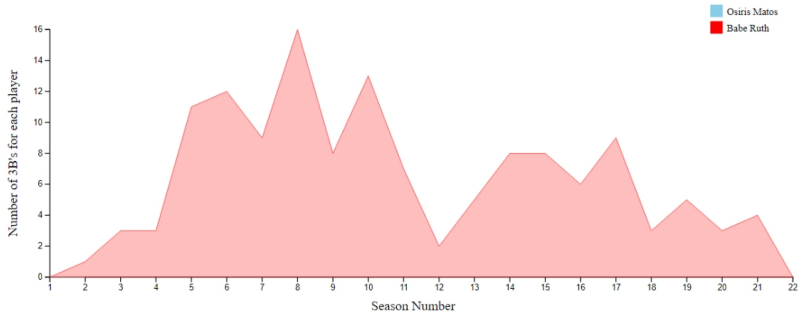
MLB Visualizations

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

Back to Main | Additional Context

Select Player 1: Osiris Matos
Select Player 2: Babe Ruth
Pitching or Batting Stats Comparison: Batting
Chosen Stat: 3B

Player Comparison



MLB Visualizations

Name: Jaden Lee, Kendall Ruth, Kacey Abbott

Back to Main | Player Comparison

MLB Visualizations Context

This page provides context for the MLB Visualizations project. Here you can find additional information explaining what each section of the main page is about.

The project aims to explore the performance of Major League Baseball players through various visualizations and statistical metrics. It focuses on rate statistics and modern advanced metrics to better understand the value of different players.

Baseball Statistics Definitions

Pitching Acronyms

- **W:** Wins - Number of games the pitcher was the winning pitcher.
- **L:** Losses - Number of games the pitcher was the losing pitcher.
- **G:** Games - Total number of games the pitcher appeared in.
- **GS:** Games Started - Number of games the pitcher started.
- **CG:** Complete Games - Number of games the pitcher completed without relief.
- **SHO:** Shutouts - Number of complete games where the pitcher allowed no runs.
- **SV:** Saves - Number of games the pitcher successfully saved.
- **IPouts:** Innings Pitched Outs - Total number of outs the pitcher achieved (divide by 3 for innings pitched).
- **H:** Hits - Total number of hits allowed by the pitcher.
- **ER:** Earned Runs - Total number of earned runs allowed by the pitcher.
- **HR:** Home Runs - Total number of home runs allowed by the pitcher.
- **BB:** Walks (Bases on Balls) - Total number of batters walked by the pitcher.
- **SO:** Strikeouts - Total number of batters struck out by the pitcher.
- **BAOpp:** Batting Average Against - Batting average of opposing hitters against the pitcher.
- **ERA:** Earned Run Average - Average number of earned runs allowed per 9 innings pitched.
- **IBB:** Intentional Walks - Total number of intentional walks issued by the pitcher.
- **WP:** Wild Pitches - Total number of wild pitches thrown by the pitcher.
- **HBP:** Hit by Pitch - Total number of batters hit by the pitcher.
- **BK:** Balks - Total number of balks committed by the pitcher.
- **BFP:** Batters Faced by Pitcher - Total number of batters faced by the pitcher.
- **GF:** Games Finished - Number of games the pitcher was the last pitcher in the game.
- **R:** Runs - Total number of runs allowed by the pitcher.
- **SH:** Sacrifice Hits - Total number of sacrifice bunts allowed by the pitcher.
- **SF:** Sacrifice Flies - Total number of sacrifice flies allowed by the pitcher.
- **GIDP:** Grounded Into Double Plays - Total number of double plays induced by the pitcher.

Batting Acronyms

- **G:** Games Played - The number of games in which the player appeared.
- **AB:** At Bats - The number of official at-bats (excludes walks, sacrifices, etc.).
- **R:** Runs Scored - The total number of runs the player scored.
- **H:** Hits - The number of times the player successfully reached first base by hitting the ball.
- **2B:** Doubles - The number of hits where the player reached second base.
- **3B:** Triples - The number of hits where the player reached third base.
- **HR:** Home Runs - The number of hits where the player hit the ball out of play, allowing them to round all the bases and score.
- **RBI:** Runs Batted In - The number of runs that scored as a direct result of the player's at-bat.
- **SB:** Stolen Bases - The number of times the player successfully stole a base.
- **CS:** Caught Stealing - The number of times the player was caught attempting to steal a base.
- **BB:** Base on Balls / Walks - The number of times the player was awarded first base after taking four balls outside the strike zone.
- **SO:** Strikeouts - The number of times the player struck out.
- **IBB:** Intentional Base on Balls - The number of times the player was intentionally walked by the opposing pitcher.
- **HBP:** Hit By Pitch - The number of times the player reached base after being hit by a pitch.
- **SH:** Sacrifice Hits / Sacrifice Bunts - The number of times the player successfully bunted the ball to advance a runner.
- **SF:** Sacrifice Flies - The number of fly balls that resulted in a run scoring (but the batter being out).
- **GIDP:** Grounded Into Double Play - The number of times the player grounded into a double play, resulting in two outs.

wOBA

- **wOBA (Weighted On-Base Average):** An advanced baseball statistic that measures a player's overall offensive contributions more accurately than traditional metrics like batting average, on-base percentage (OBP), or slugging percentage (SLG).
- **Purpose:** Evaluates a hitter's overall offensive performance by combining the ability to reach base and the quality of hits.
- **Weighted Events:**
  - Walks and hit-by-pitches are valued less than singles.
  - Doubles, triples, and home runs receive progressively higher weights based on their run-producing value.
- **Scale:** wOBA is scaled similarly to on-base percentage (OBP). A league-average wOBA is typically around **0.320 to 0.340**, depending on the era and league.

Data Scources

The data used in this project has been sourced from the Sean Lahman Baseball Database (http://www.seanlahman.com/). The Lahman dataset is an extensive collection of baseball statistics that includes information on over 21,000 players throughout the history of Major League Baseball. This rich dataset provided us with player statistics spanning 1871 to 2023 such as games played, runs scored, hits, and other critical metrics that are displayed in our visualizations.

The baseball card images were compiled by scraping the Baseball Almanac website (https://www.baseball-almanac.com/). Using this approach, we were able to obtain nearly 10,000 player images, which we use to complement the statistical data from the Lahman dataset. These images add a visual element that makes it easier to connect statistics to real players, enhancing the overall experience.

For the players that we were unable to find images for, we use a default placeholder image. This ensures that every player represented in our dataset has some form of imagery, even if specific photos are unavailable. The combination of comprehensive player data and visual elements provides a more engaging and informative experience for users exploring Major League Baseball history and statistics.

Note: The majority of player images not found were from the early years of baseball. Of the over 21,000 players in our data, more than 10,000 of them were born before 1945.

CS-5630 / CS-6630 MLB Visualizations

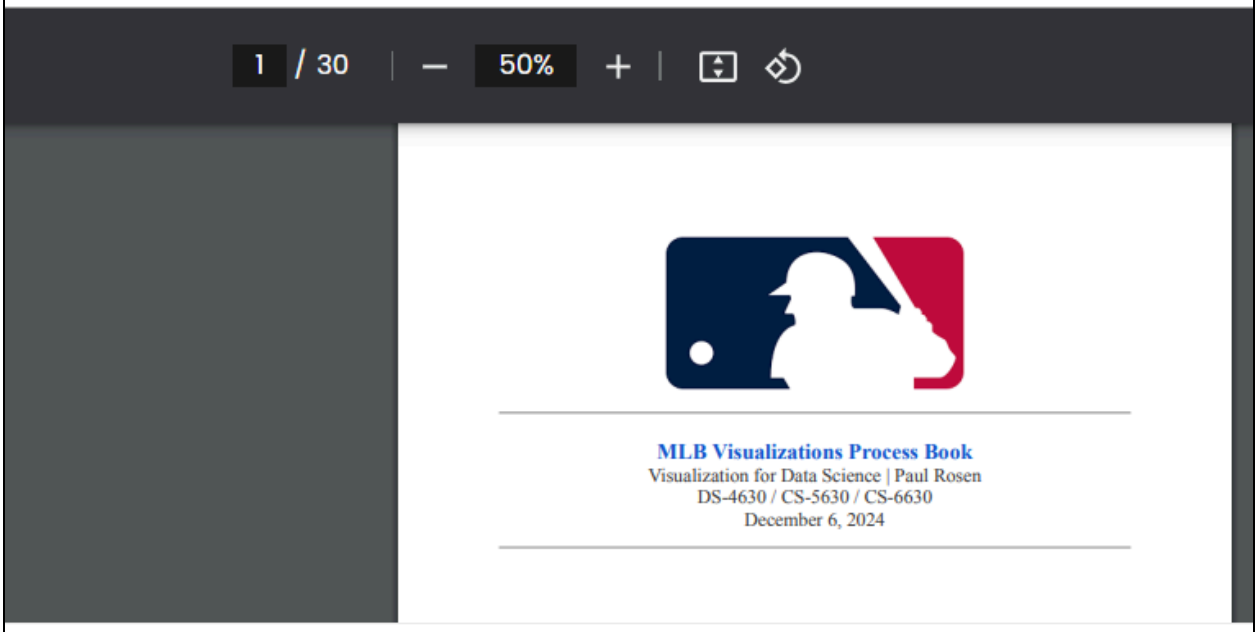
Name: Jaden Lee, Kendall Ruth, Kacey Abbott

Back to Main | Player Comparison | Additional Context

Data Sources

Player data came from the Lahman Database and the images were scraped from Baseball Almanac.

Process Book



Project Screencast

group-project-baseball-visualization-jkk-4

MLB Visualizations

Overview

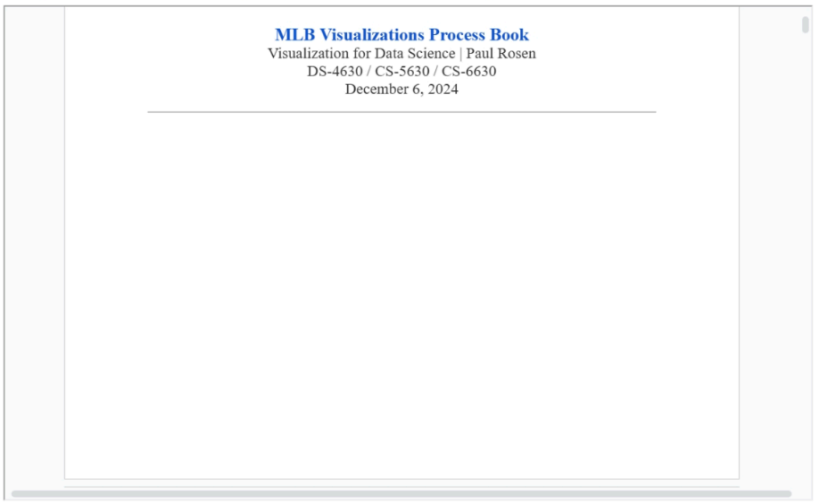
This project focuses on creating interactive and insightful visualizations representing baseball statistics. It was developed as part of the **DS-4630 / CS-5630 / CS-6630 Visualization Project** course at the University of Utah.

Project Links

- Project Website
- GitHub Repository

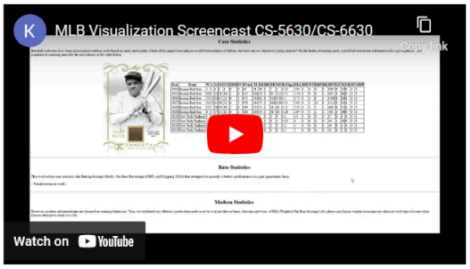
Project Process Book

Our project process book documents the planning, implementation, and reflection phases of the MLB Visualizations project.


MLB Visualizations Process Book
Visualization for Data Science | Paul Rosen
DS-4630 / CS-5630 / CS-6630
December 6, 2024

YouTube Video

Watch the project presentation directly here:



Details

- **Group**: baseball-visualization-jkk-4
- **Authors**:
  - Jaden Lee (u1417827)
  - Kendall Ruth (u1481623)
  - Kacey Abbott (u0692178)
- **Affiliation**: University of Utah
- **Professor**: Paul Rosen
- **Created Date**: September 13, 2024
- **Copyright**: This code may not be copied or edited for academic use.

Description

This project leverages data visualization techniques to explore and represent various baseball statistics. The visualizations aim to provide a comprehensive and interactive experience for understanding key metrics and trends in the sport.

Features

- Interactive charts and graphs for in-depth analysis
- Data exploration tools tailored for baseball statistics
- Responsive and user-friendly web interface