MLB Infographics Process Book

Visualization for Data Science | Paul Rosen
DS-4630 / CS-5630 / CS-6630
October 25, 2024

Table of Contents

Basic Info

Project Title: **MLB Infographics**

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

Background and Motivation

Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

Related Work

Anything that inspired you, such as a paper, a website, visualizations we discussed in class, etc.

Questions

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

Data

Source, scraping method, cleanup, etc.

Exploratory Data Analysis

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

Design Evolution

What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

Implementation

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

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!

Project Objectives:

Our objectives for this project are to create a clear and engaging visual representation of baseball statistics. We want to highlight important data and trends that show how numbers can influence team performance and player choices. Additionally, since baseball is already a highly analyzed sport, we want to choose not so common statistics to potentially show correlation of a player/team's success. Our visualizations will be unique and showcase different aspects of baseball statistics. By combining the player's real-life experiences, the analytical approach to data, and our team's efforts, we aim to make complex information easier to understand for fans, analysts, and players. Our goal is to provide a useful tool that helps people see how data impacts the game, while also showcasing the benefits of working together with different perspectives. We want baseball teams/baseball workers to be able to look at our visualizations and find good value from them.

Data:

There are a bunch of datasets that are available, but here are a few of the ones that we will focus on in our visualizations:

- Retrosheet: Play-by-play and box score data extending back to the early 1900s
  - https://www.retrosheet.org
- Lahman Database: Archive of team and player statistics going back to 1871
  - http://seanlahman.com/
- 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/

Data Processing:

For this project, our data processing will involve selecting and organizing clean datasets that are readily available. Since baseball is already a highly analyzed sport, we aim to choose some less common statistics and correlations to explore and potentially reveal new insights about player and team success. Our tasks will include 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 can concentrate on accurately representing and analyzing these unconventional stats to uncover meaningful correlations and patterns.

Visualization Design:

We plan to include visualizations that will include many different design aspects. Some visualizations that use categorical data might use a bar chart and location-based data might be shown with a map. We could also use baseball themes to portray our data. For example, we could display 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 intend 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 Schedule:

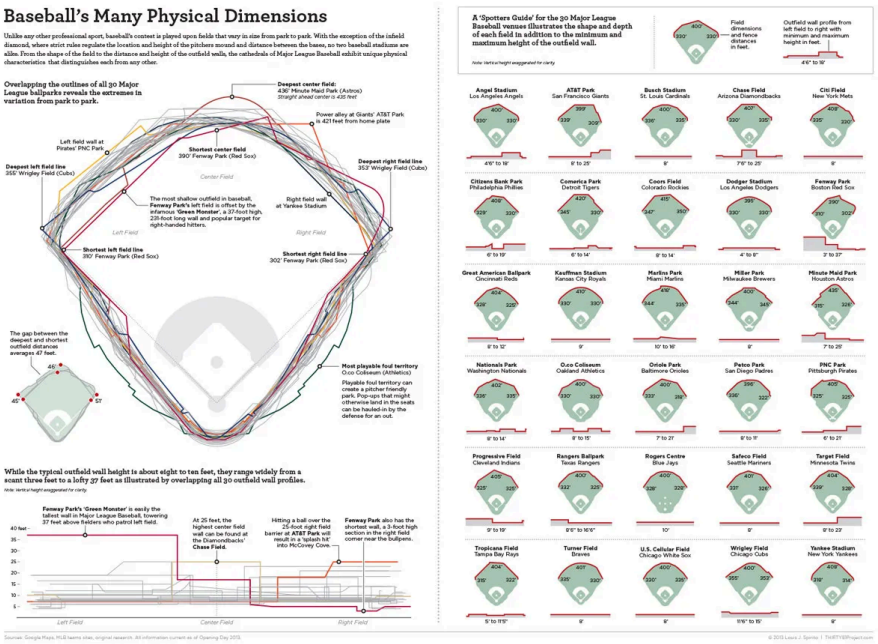
Team meeting schedule: All team members will be available every X day at Y time if needed, and preferably over zoom. Otherwise we will coordinate over text/zoom as needed.

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	No
10/25/2024	Milestone, a functional project prototype	No
11/1/2024	Peer feedback	No
11/8/2024	Make adjustments from peer feedback	No
11/15/2024	Make sure visualizations are correct and look good	No
11/22/2024	Final project submission & group member evaluations	No

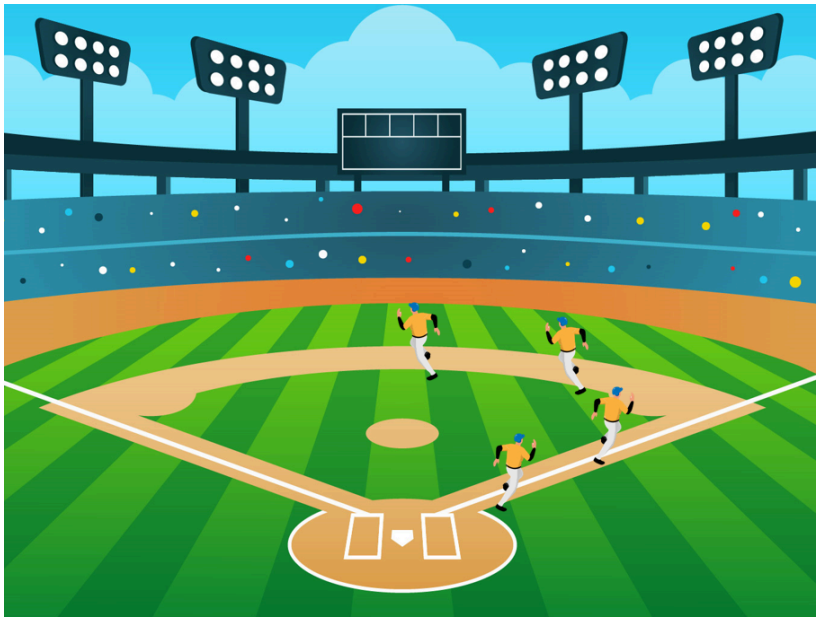
Sketches/Examples

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.**

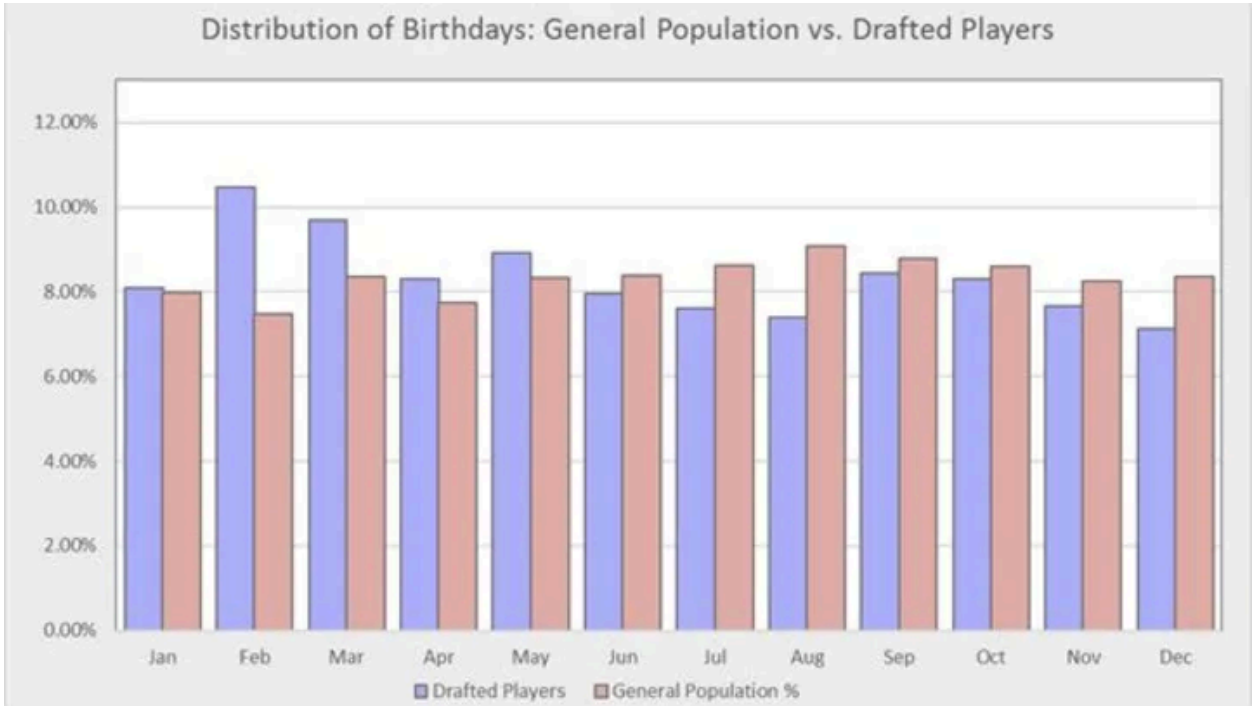
2. Stadium background infographic. Use to show stadium stats, compare stadium stats or how players have performed at that stadium by game, season, or career.

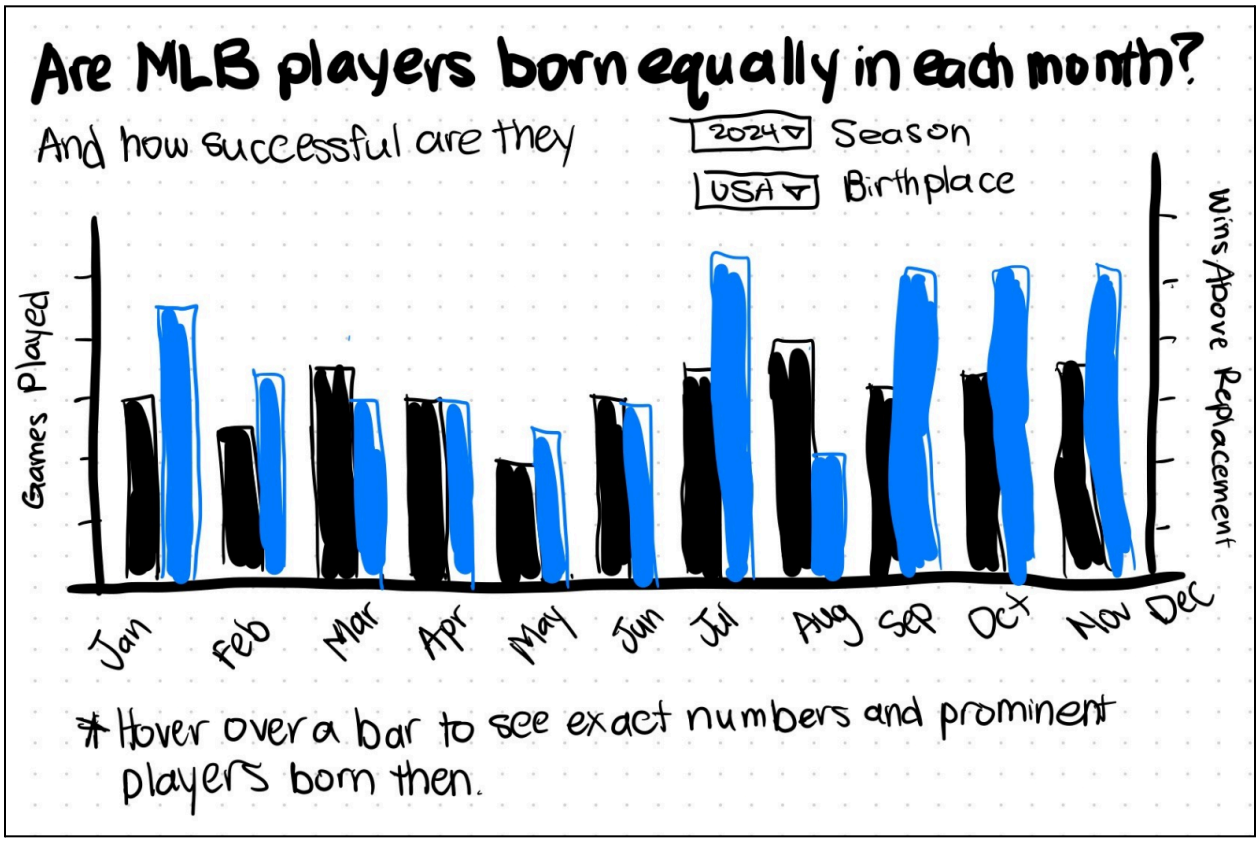


2.B Another version of the stadium background infographic. Players running around basses to show comparison of 0-100%. 25% being first base, 50% being second base, etc…



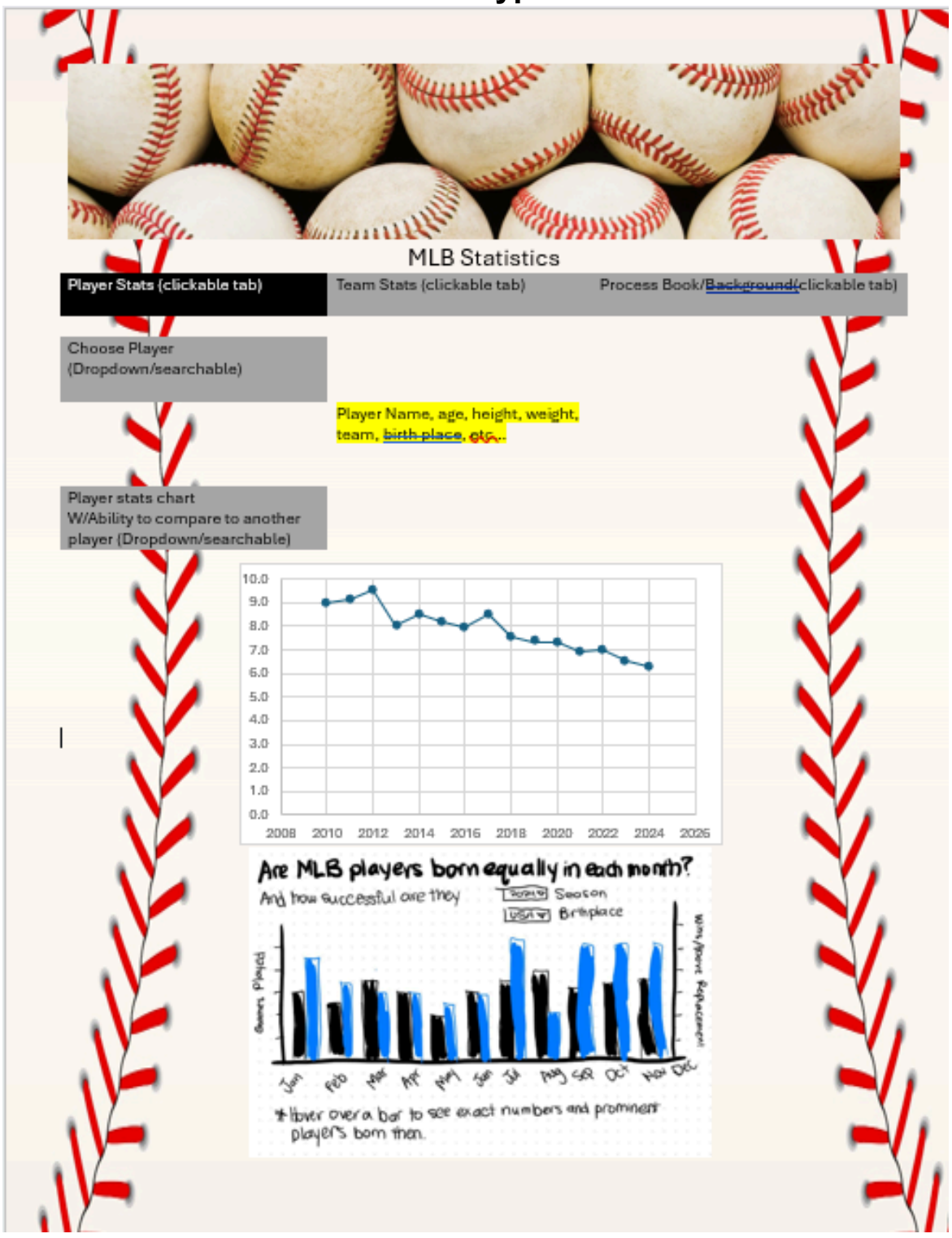
3. **Visualization like this, but with MLB data/ not NBA to showcase when the birthdays are for drafted MLB players.**





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

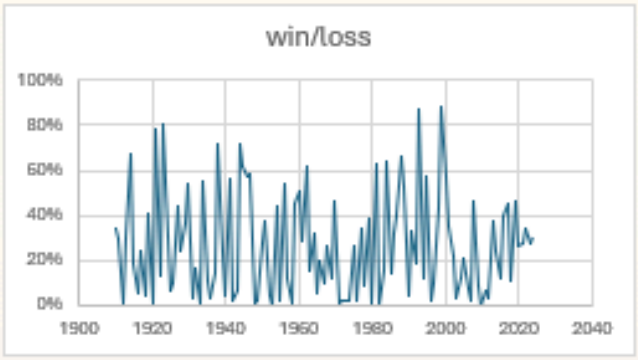
Player Stats (clickable tab) | Team Stats (clickable tab) | Process Book/~~Background~~(clickable tab)

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

To-Do:
- 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.
    - Target them to present a certain claim.

9/19/2024 Meeting Notes:
Better defining project scope and getting started/timeline.

Map visualization of trades between teams. Also showing information on a single player to see where they started to where they have gone.

Guessing game. Player stats, comparative.

- 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 visualizations
- deploy with AWS

https://www.nytimes.com/interactive/2022/sports/baseball/umpire-pitch-ball-strike-game.html

---

10/02/2024 Meeting:
NBAstatsVIS (wilsoncernwq.github.io), using this as an example.
Data dependent and geared towards baseball.
Start with simple infographics and clean up, beautify, add infographics as time allows.

Next Steps:
Gather data starting with just this year.
Start with gathering data from https://www.baseball-reference.com/
Secondary data set from Baseball Savant: Statcast, Trending MLB Players and Visualizations

Data scraper: pybaseball · PyPI

Next deadlines: 15th, October.
We want to have the basic data wrangled
Basic website setup
Good idea of what visualizations will look like/what we want based on the data we have.

10/3/2024

CSV Dataset: http://www.seanlahman.com
(https://drive.google.com/drive/folders/1C_CCzkaIzoe9fxDsuYrXowJ6cdvTEuJ8?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

Upload CSV/JSON dataset files to github.
Planning on setting up main player image in the form of a baseball card with stats that update based on player. Working on getting either a downloadable list of player images/cards.

10/22/2024:

- get some dummy data
- filter players by year start, year end,
    - Kendall will get dropdown with player names populated
- table with player info
- woba chart
    - https://github.com/chadwickbureau/register
- images (Kacey)
    - Hard code player name to start. Will be linked to player name dropdown.

10/24/2024: Updated draft website concept

MLB Statistics

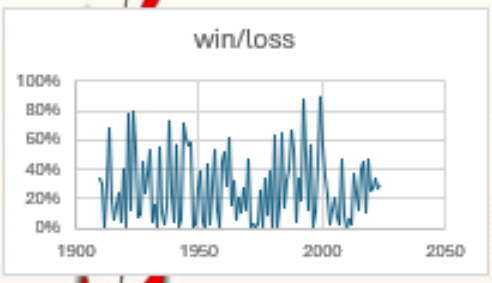
Player Stats (clickable tab) | Team Stats (clickable tab) | Process Book/~~Background~~(clickable tab)

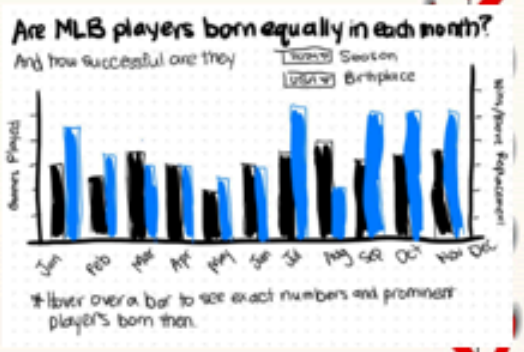
Choose Team
(Dropdown/searchable menu)

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..)





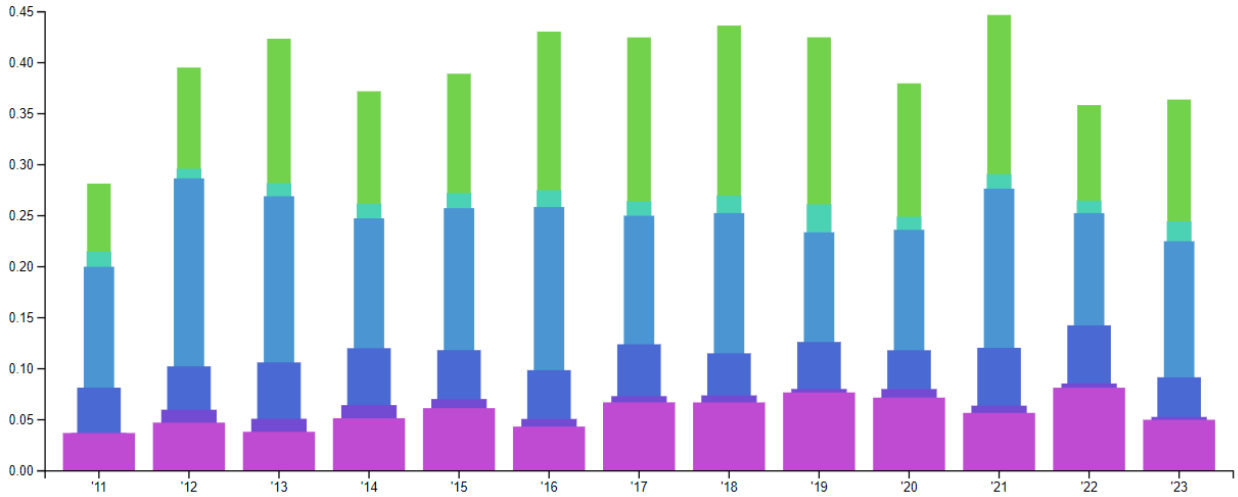
MLB Statistics

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

Our Process Book/analysis/assumptions

Attached

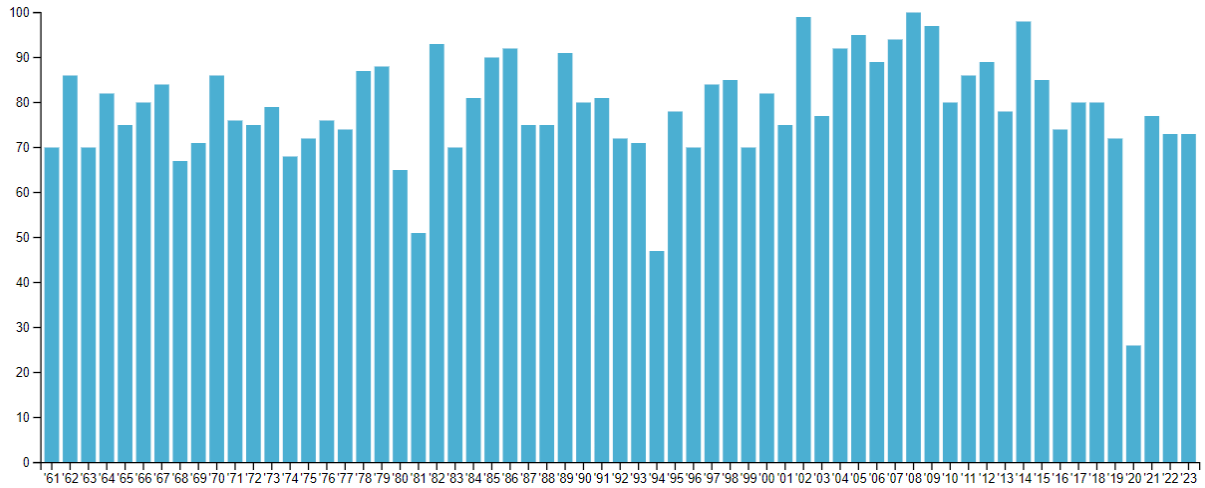
Player wOBA Chart



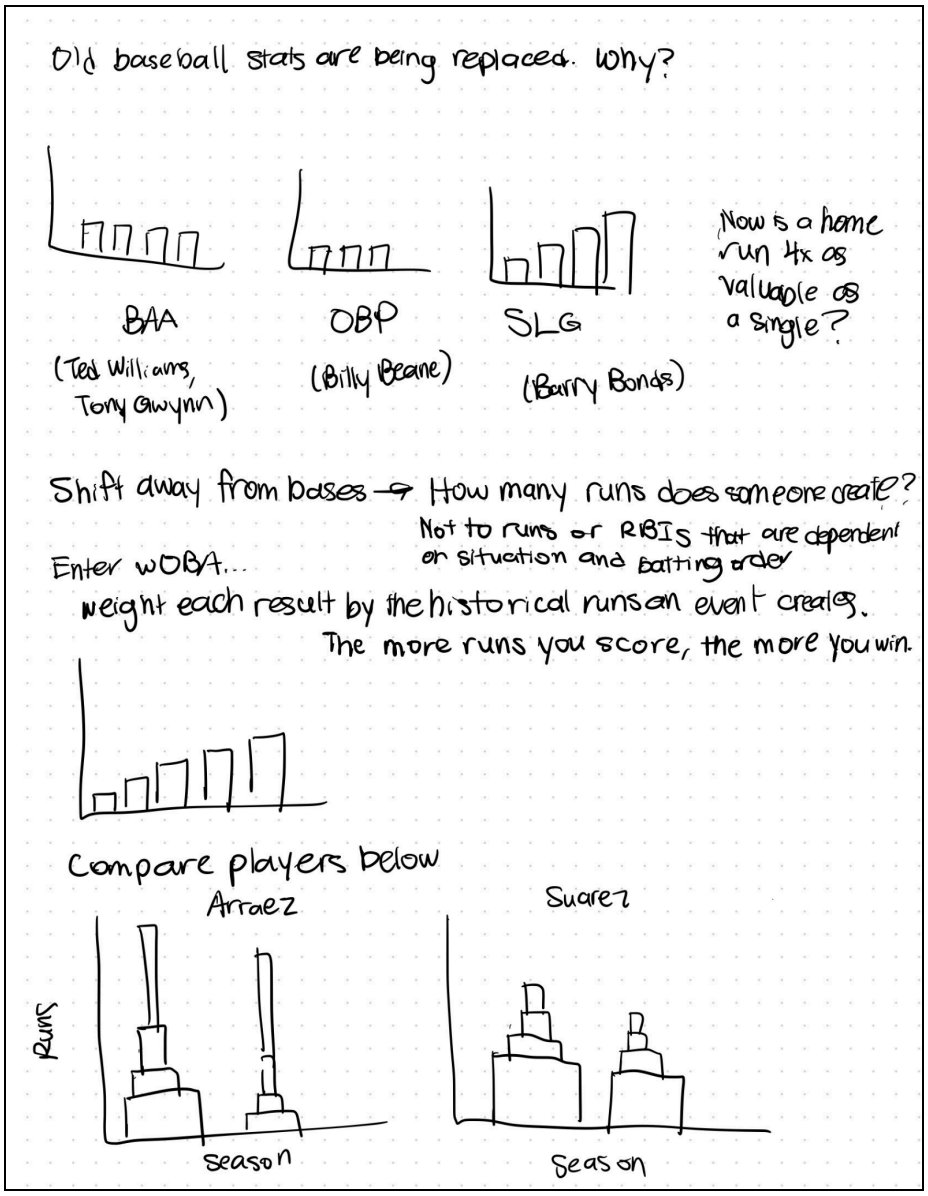
wOBA Weights CSV: https://www.fangraphs.com/guts.aspx?type=cn

Franchise: Los Angeles Angels of Anaheim Data Metric Wins

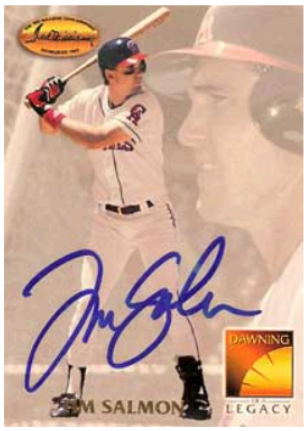
Team Win/Loss Total



Proof of concept graph with dropdowns.



Asks the question of why we use wOBA nowadays and the sketch visualizes the differences between it and traditional rate stats.



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 had 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. 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.