

Visualization for Data Science

DS-4630 / CS-5630 / CS-6630

Project Proposal

Basic Info:

Project Title: MLB Infographics

Kacey Abbott, kacey.abbott@utah.edu, U0692178

Kendall Ruth kendall.ruth@utah.edu, u1481623

Jaden Lee, u1417827@utah.edu, u1417827

Project Repository:

<https://github.com/dataviscourse2024/group-project-baseball-visualization-jkk-4.git>

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 \(baseballprospectus.com\)](Cot's%20Baseball%20Contracts%20(baseballprospectus.com))
- Baseball Savant: Advanced player and team statistics plus available Statcast data
 - [Baseball Savant: Statcast, Trending MLB Players and Visualizations](Baseball%20Savant%20Statcast,%20Trending%20MLB%20Players%20and%20Visualizations)
- 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 \(kaggle.com\)](MLB%20Player%20Digital%20Engagement%20Forecasting%20EDA%20(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	No
9/20/2024	Finalize specific Visualizations	No
9/27/2024	Rough Draft of website setup	No
10/4/2024	$\frac{1}{3}$ way through website setup	No
10/11/2024	$\frac{2}{3}$ way through website setup	No
10/28/2024	Finished with website setup	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.

MLB Stadiums

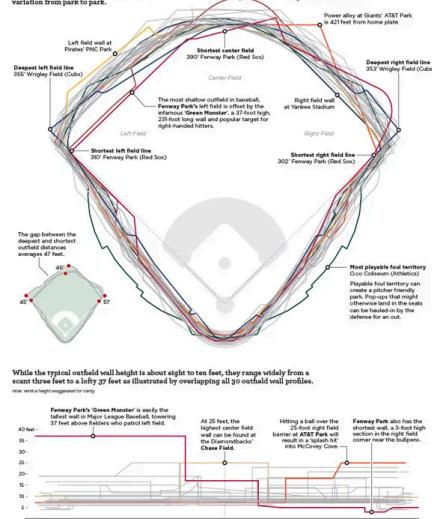


2. Stadium background infographic. Use to show stadium stats, compare stadium stats or how players have performed at that stadium by game, season, or career.

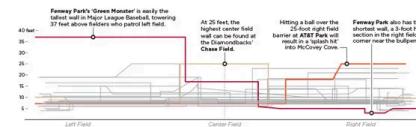
Baseball's Many Physical Dimensions

Unlike most other professional sports, baseball is played on open fields that vary in size from park to park. With the exception of the infield dimensions, which strict rules mandate the location and size of, the distance and height of the outfield walls are left to the teams and owners. From the shape of the field to the distance and height of the outfield walls, the cathedrals of Major League Baseball exhibit unique physical characteristics that distinguish each from any other.

Overlapping the outlines of all 30 Major League ballparks reveals the extremes in variation from park to park.



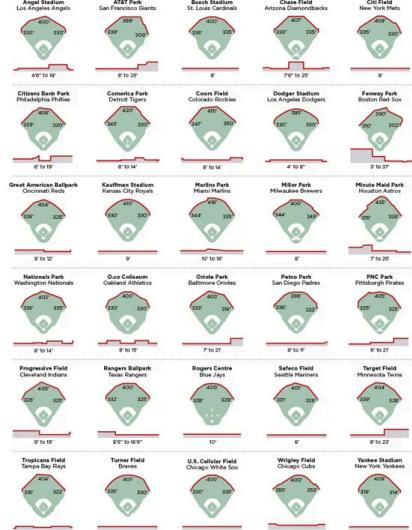
While the typical outfield wall height is about eight to ten feet, they range widely from a scant three feet to a lofty 37 feet as illustrated by overlapping all 30 outfield wall profiles.
Note: Some heights exaggerated for clarity.



Source: Google Maps, M.L.B. team sites, original research. All information current as of Opening Day 2010.

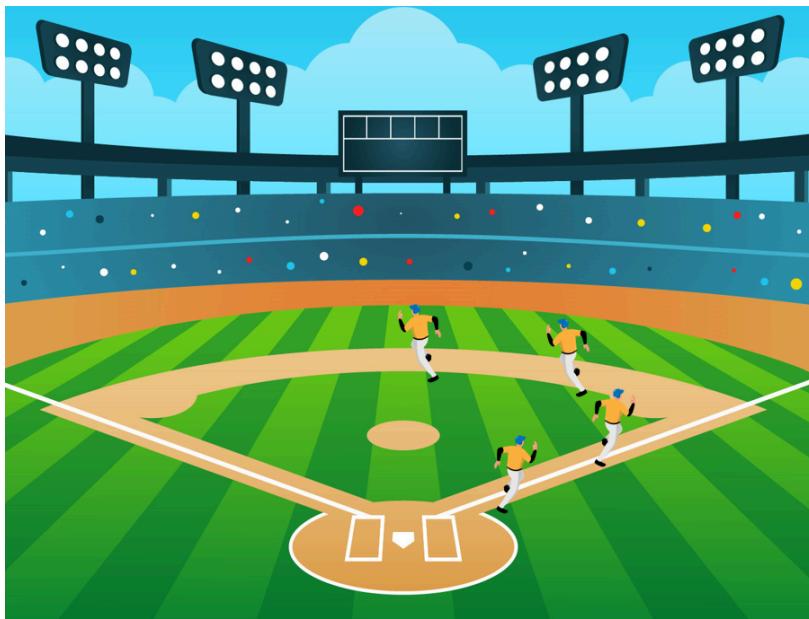
A 'Spectator's Guide' for the 30 Major League Baseball venues illustrates the shapes and depth of the fields, as well as the minimum and maximum height of the outfield wall.

Note: Some heights exaggerated for clarity.

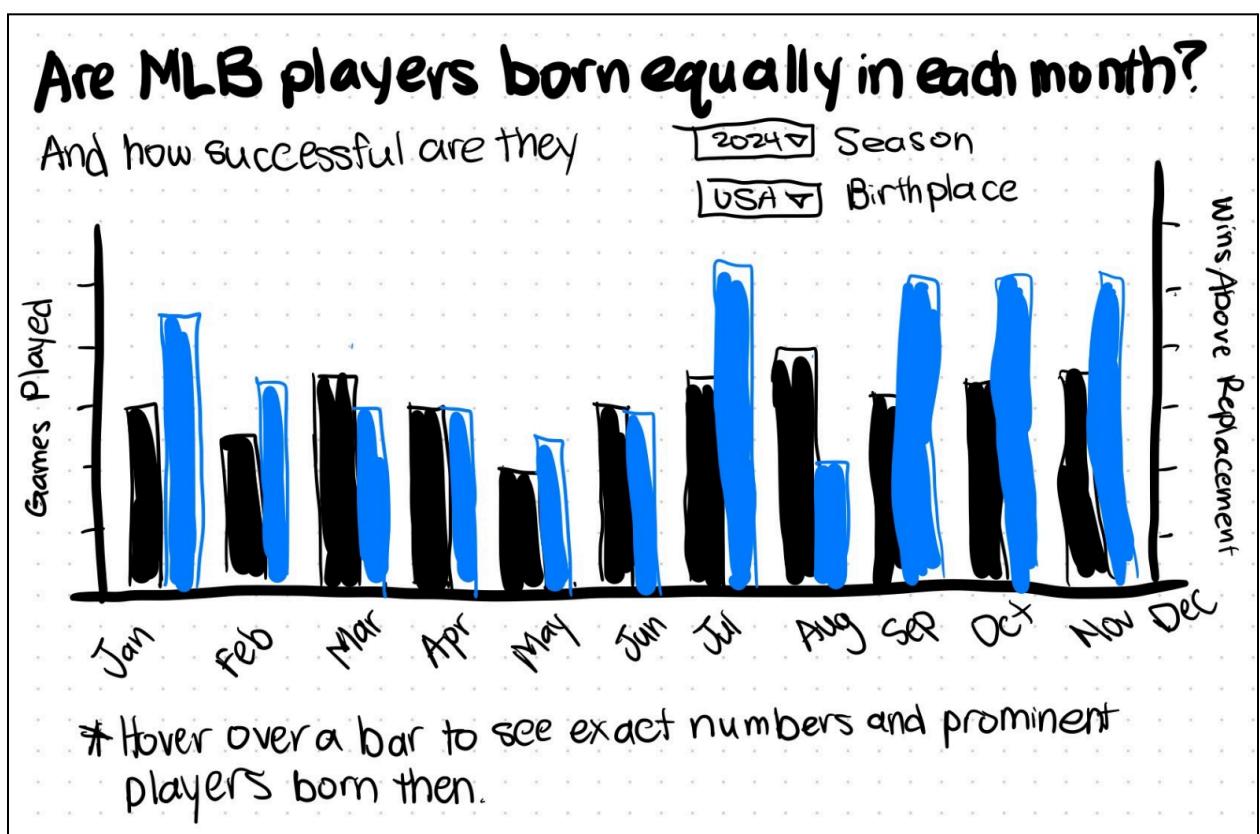
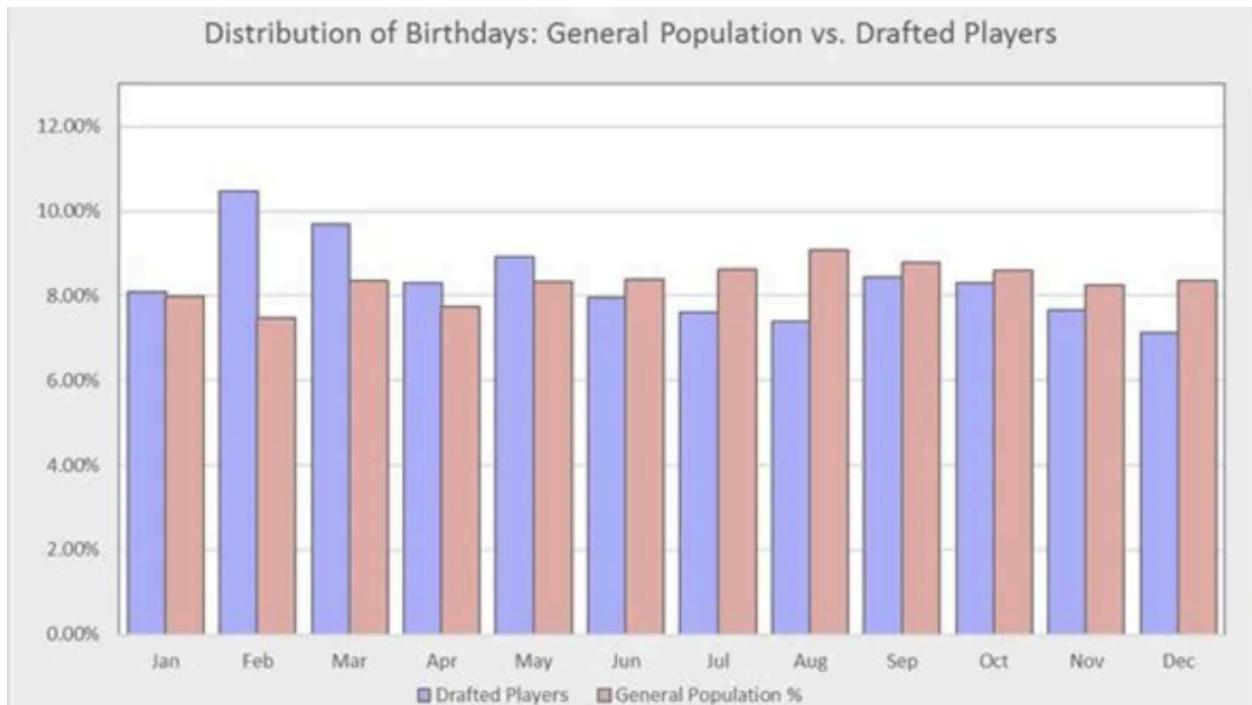


© 2010 Land J. Sprints | LandJSprints.com

2.B Another version of the stadium background infographic. Players running around bases 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)

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.

Month	Season	Count
Jan	Season	~15
Feb	Season	~12
Mar	Season	~18
Apr	Season	~15
May	Season	~12
Jun	Season	~15
Jul	Season	~18
Aug	Season	~15
Sep	Season	~12
Oct	Season	~15
Nov	Season	~12
Dec	Season	~15
Jan	Birthplace	~15
Feb	Birthplace	~12
Mar	Birthplace	~18
Apr	Birthplace	~15
May	Birthplace	~12
Jun	Birthplace	~15
Jul	Birthplace	~18
Aug	Birthplace	~15
Sep	Birthplace	~12
Oct	Birthplace	~15
Nov	Birthplace	~12
Dec	Birthplace	~15



MLB Statistics

Player Stats (clickable tab)

Team Stats (clickable tab)

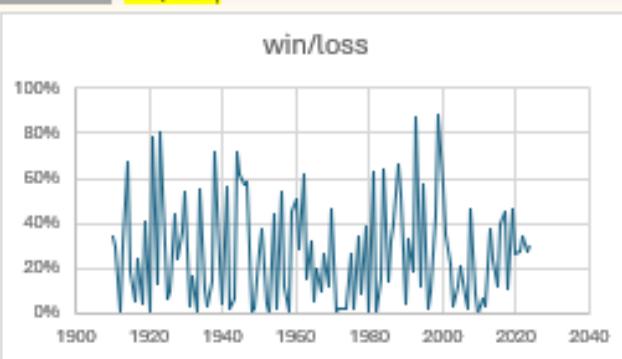
Process Book/Background(clickable tab)

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, losses, runs,
hits, etc...)



MLB Stadiums





MLB Statistics

Player Stats (clickable tab)

Team Stats (clickable tab)

Process Book/Background(clickable tab)

Our Process Book/analysis/assumptions