[**Visualization**](https://dataviscourse.net/2024/) **for Data Science**

**DS-4630 / CS-5630 / CS-6630**

**Project Proposal**

**Basic Info:**

Project Title: MLB Infographics

Kacey Abbott, [kacey.abbott@utah.edu](mailto:kacey.abbott@utah.edu), U0692178

Kendall Ruth [kendall.ruth@utah.edu](mailto:kendall.ruth@utah.edu), u1481623

Jaden Lee, [u1417827@utah.edu](mailto: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](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)](https://legacy.baseballprospectus.com/compensation/cots/)
* Baseball Savant: Advanced player and team statistics plus available Statcast data
  + [Baseball Savant: Statcast, Trending MLB Players and Visualizations |](https://baseballsavant.mlb.com/)
* 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)](https://www.kaggle.com/code/miklgr500/mlb-player-digital-engagement-forecasting-eda/notebook)
* 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:**

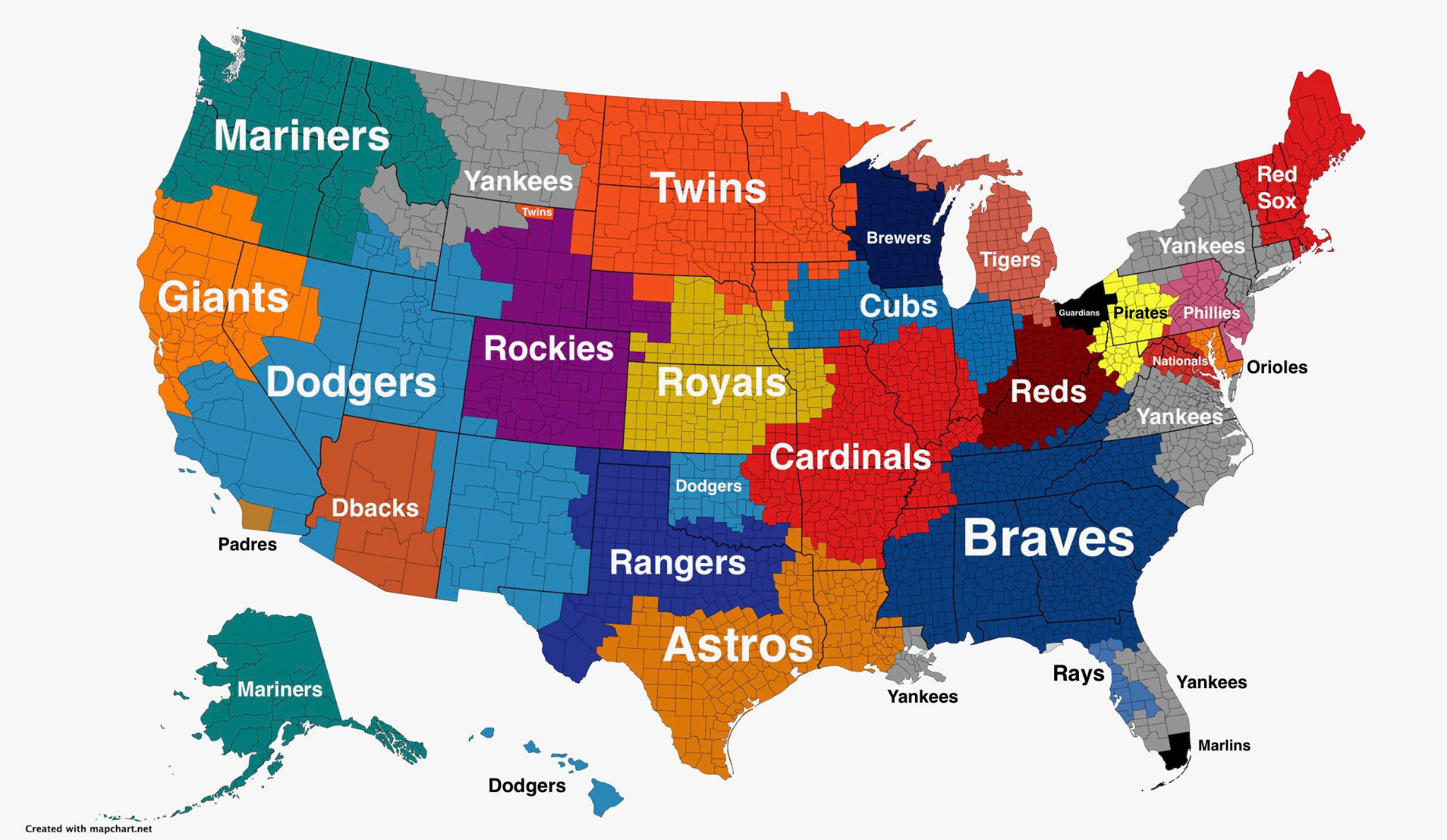
Team meeting schedule: All team members will be available every X day at Y time if needed, and preferably over zoom. Otherwise we will coordinate over text/zoom as needed.

| 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 | ⅓ way through website setup | No |
| 10/11/2024 | ⅔ 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 |

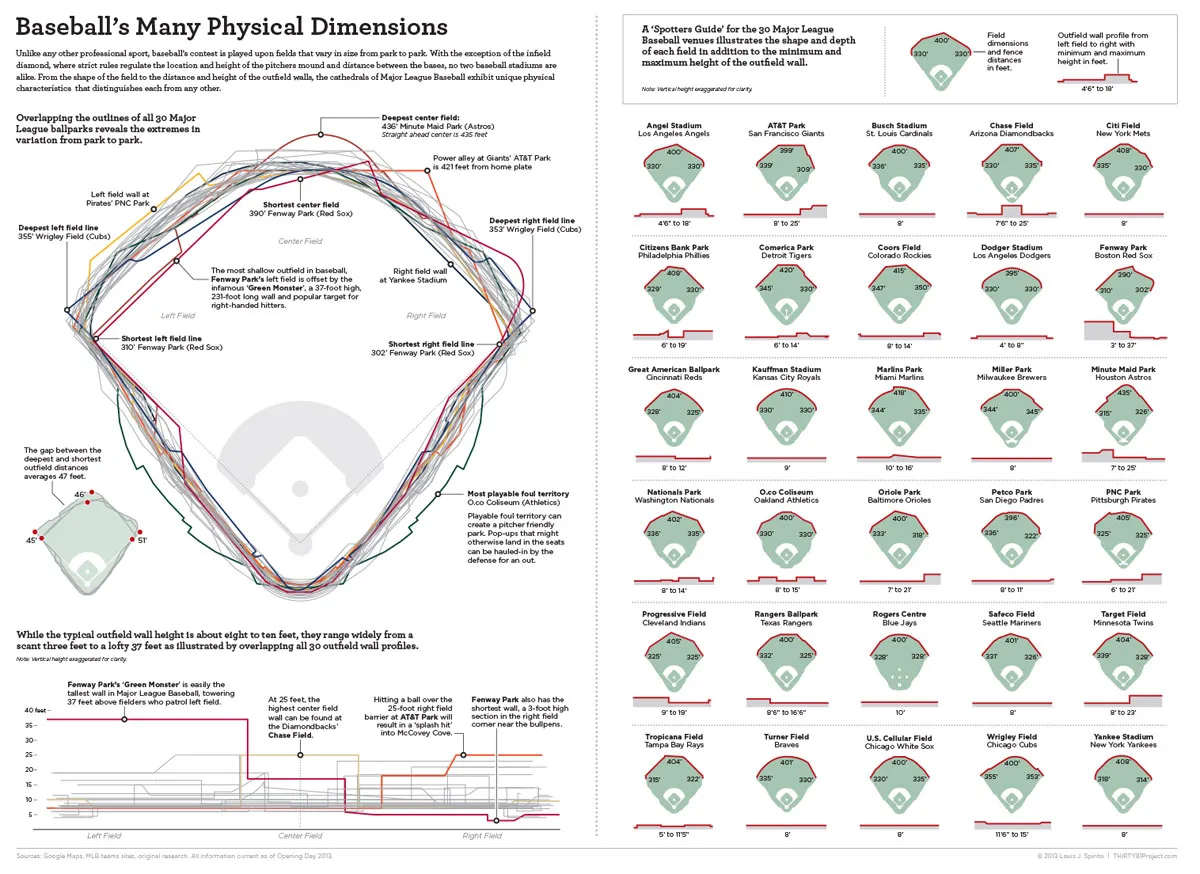
**Sketches/Examples**

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.**

****

****

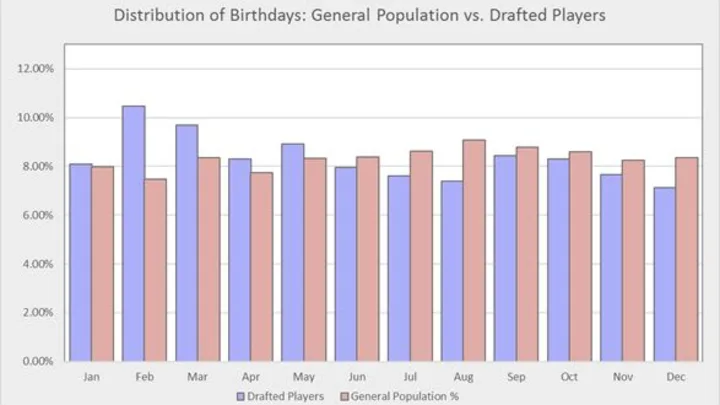
1. **Stadium background infographic. Use to show stadium stats, compare stadium stats or how players have performed at that stadium by game, season, or career.**

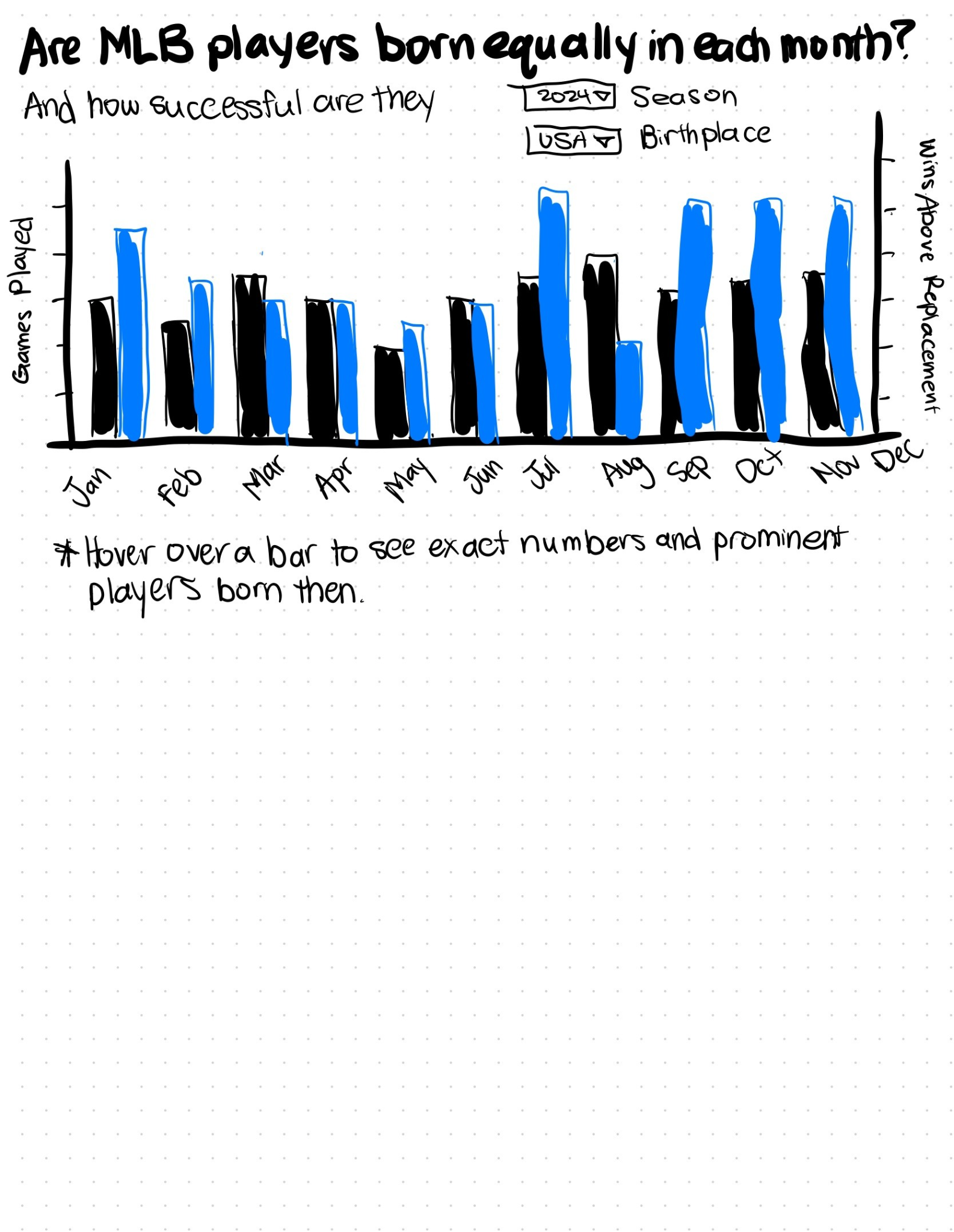
****

**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…**

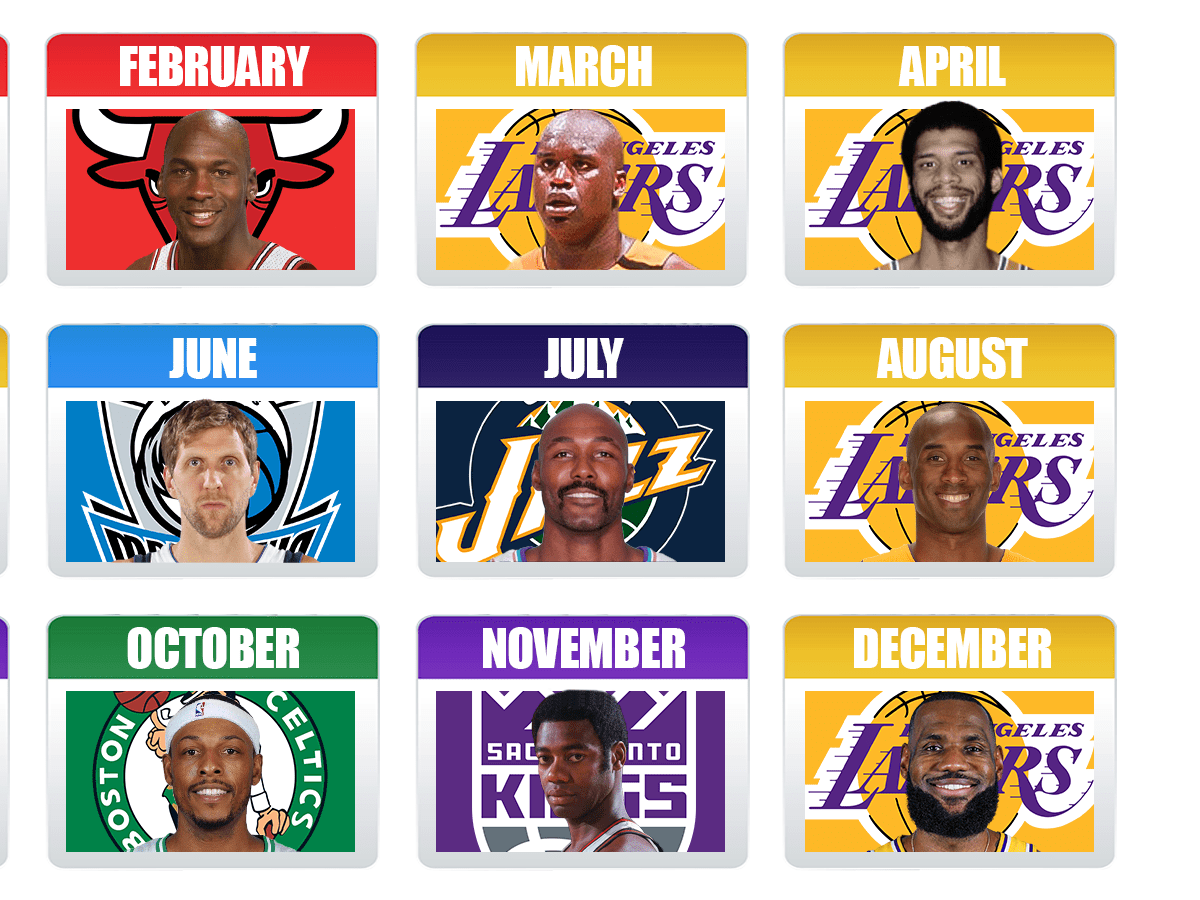
****

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

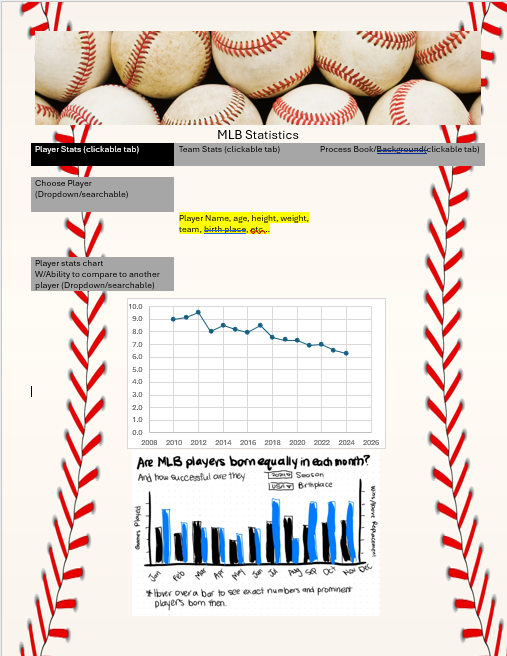
****

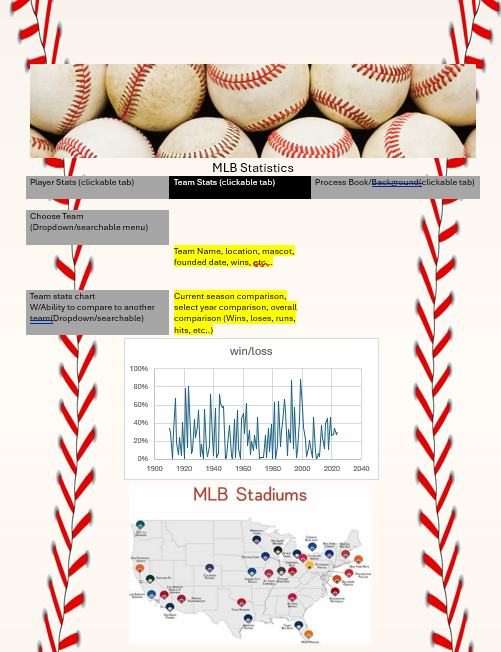
****

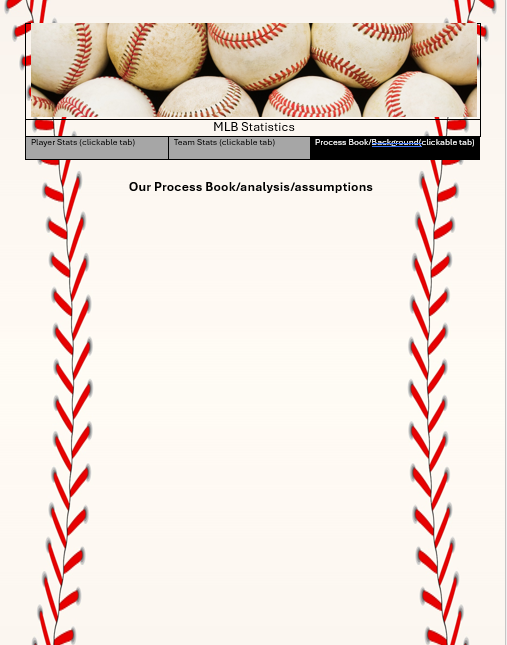
1. **Visualization for best player by month, similar to this one for the NBA**

****

**Prototype**

****

****

****