The MVP Race: MLB Player At-Bat Statistics in the 2022 Season

Vivek Kanpa*
Khoury College of Computer Science

Siraj Akmal†
D'Amore McKim School of Business

Marco Chen‡
Khoury College of Computer Science

Abstract

This essay describes the development of a dashboard for the 2022 MLB season that allows users to see how players are performing and compare the performance of two players. Our team's motivation was to create simple and efficient visualizations for MLB players' statistics that enable users to filter, sort, navigate, and compare players. Through our design process, we developed a user-friendly interface that provides convenient access to individual player performance and team performance, highlighting the importance of baseball statistics in supporting fans and coaches. Our dashboard provides an intuitive way for users to stay informed about player performance and make informed decisions based on data, contributing to a deeper understanding of the game and its players.

1 Introduction

MLB, the top professional baseball league in the world, has been in existence since 1903 and has featured some of the greatest players in baseball history, such as Babe Ruth, Willie Mays, and Derek Jeter. The league is renowned for its long-standing history, intense rivalries, and enthusiastic fan base, with millions of viewers tuning in to watch the games both in-person and on television. The captivating and beloved sport of baseball has won the hearts of fans worldwide with its blend of skill, athleticism, and strategy.

Baseball statistics are really important to both fans and coaches. Based on Molly (2018), based on the attendance data of MLB teams from 2001 to 2016, the increase in fan experience and attendance is mostly influenced by the team's performance, in other words, losses and wins. From our perspective, the team's performance positively correlates with team members' performance. Therefore, learning about an individual player's performance or making comparisons between players can help baseball clubs gain a deeper insight into their fans' experience. At the same time, the coach can easily understand the team's performance and make key decisions, such as developing tactics and trading players.

Our team aims to develop a pack of simple, clear, and efficient visualizations for MLB players' statistics. To validate the visualizations, we collected the data, "2022 MLB Player Stats", from Kaggle, which was updated one month ago. The dataset splits into two subsets: "batting" data and "pitching" data, representing the player as a batter or pitcher. There are 29 columns and more than 900 rows for each subset. Each row represents a unique baseball player, and columns contain attributes that relate to either player as a batter or pitcher. We would apply data wrangling to the data and present it to the users. Users should be able to filter, sort, navigate and compare MLB players in season 2022 by using our results.

*e-mail: kanpa.v@northeastern.edu
†e-mail: akmal.s@northeastern.edu
‡e-mail: chen.xi10@northeastern.edu

2 Task Abstraction

In order to see how MLB players are doing in season 2022, we want to compare one statistic from more than 20 types of play among all MLB players. Types of play include at-bats, home runs, strikeouts, and more important features of MLB players. For instance, if users choose home runs, they are able to see a leading board of home runs among all MLB players where the numbers of home runs are sorted in descending order.

To compare the performance of two MLB players in 2022, we want to individually compare each of their play statistics. By looking at subplots that show players' different attributes, users can tell each player's strengths and weaknesses.

3 Related Work

The abundance of advanced statistics in baseball has lent itself to visualizations that demonstrate trends in player growth and micro-analysis of player batting performances.

David Gwartney's Baseball Analytics Site
For example, the "player batting comparison" data visualization on David Gwartney's baseball analytics site functions as a way to compare player performances across seasons by comparing their offensive batting statistics. A user can query the data displayed by which season, teams, and players they wish to compare. In addition to a data table which is readily populated with statistics comparing players' hits, base on balls, at bats, sacrifice flies, singles, doubles, triples, and home runs, three line plots are shown below for each players trends in slugging percent, on base percent, and batting average over time. Gwartney's visualization also displays the median slugging percent, on base percent, and batting averages across the selected player set. This is a particularly helpful feature of the line graph, since it allows the user to compare specific players to their peers over a given time interval. This design choice would inspire our bar graph comparing a selected player's batting statistics with the MLB averages for those batting statistics in the 2022 season.

StatCast Dashboard – Prograf Laboratory
The StatCast Dashboard, developed by Dr. Marcos Lage of the Prograf Laboratory, hosts a number of informative and easy-to-digest visualizations to understand advance statistics of player performances. The "Statcast Outs Above Average" visualization, for instance, displays the Outs Above Average performance of infielders and outfielders both at their lineup position and at a location on the field. It displays six partially filled bars for various hitting statistics and below this it places the landing locations for various hits encoded with three colors to represent positive or negative outs above average. Using the same data source as Gwartney's visualization, Lage and colleagues chose to represent batting data as a heatmap with position, which allows the reader to more easily spot trends in hitting performance. These advanced analytics can be difficult for the untrained eye to grasp in a data table, so utilization of a baseball diamond as a map was a strong tool to communicate frequency and clustering with hit data. This imagery inspired our diamond "warped bar graph" design; while the Statcast visualization uses the baseball diamond to encode position, our visualization makes use of the diamond to encode a scale for comparing player statistics during the 2022 season. The maximum value (a bar that returns all the way to home base) is the all-time single season records for each batting

statistic, taking advantage of the culturally understood metaphor of a ”home run” as a job done to completion or maximal limit.

State of the Art of Sports Data Visualization
In 2018, Perin and colleagues of the University of London discussed the benefits and drawbacks of various data visualization techniques in sports analytics and communication. Perin uses soccer as an example to breakdown three major classes of data in sports media; box score data, tracking data, and meta-data. Box score data evolved from archaic score cards and corresponds to in-game events and major statistical categories for each player on both teams. Tracking data is spatio-temporally monitored for each player, and while often costly to generate also can produce the most fruitful recommendations for players to improve. Meta-data goes beyond the specific game being played, to incorporate visual cues that inform the context of the box score and tracking data (for example, the field lines on a soccer field). Ultimately, the paper presents visualizations of each category of data but recommends each visual representation with a specific consumer in mind. For instance, clearly presented meta-data is crucial for the untrained eye or non-sports fan, while tracking data is often overwhelming for this demographic and is often used for advanced analysis by coaches, superfans, and players. The insights from this paper (coupled with Lage's aforementioned work with StatCast) would inspire the decision to warp a bar graph around the baseball diamond, providing a visual cue of a player's ”progress” to the MLB record for a given statistical category. Additionally, since the objective of our visualization was to depict the race for season MVP, we decided to exclude representation of spatio-temporal data, as those features are often ignored or deprioritized during MVP voting.

4 Data

Our data comes from the Kaggle dataset, 2022 MLB Player Statistics (`https://www.kaggle.com/datasets/vivovinco/2022-mlb-player-stats`). This dataset contains the batting statistics for every player that participated in at least one game during the season. The data contains 29 attributes, with 26 of them being statistics, and 987 items, with each being a player. Some players appear multiple times due to changing teams throughout the season. The dataset was scraped from Baseball Reference. Baseball Reference is a baseball statistic website used by media companies and broadcasters, providing accurate data from the 1888 season to current day.

Baseball statistics are objective measurements of a player's performance on the field. They are based on the actual events during the game, such as hits, runs, and, strikeouts, and are recorded by professionals. There are even adjustments after games to ensure accuracy.Because baseball statistics are based on actual events that happen on the field, they cannot be biased or unethical. The numbers reflect what occurred during the game and are not influenced by personal opinions, biases, or unethical practices.

The data was usable, but still needed some adjustments. First, we renamed the attributes to their full names. In the original datasest, each attribute was the abbreviation for the statistic. To people who know baseball, this would be fine. For the sake of allowing users who are not familiar, we provided the full names. The next step was to clean the data of non utf-8 characters. Several spaces within the Name attribute contained invalid characters. To remove duplicate players, we kept the player's statistics for the entire season, with the Team attribute being set to ”TOT”. we set the league for players who played on multiple teams in multiple leagues to MLB. After removing duplicates the number of items was 789. This was all the cleaning that was necessary as Baseball Reference API, which was used by the creator of the dataset, provides clean data. Our repository containing EDA and the clean dataset can be found here: `https://github.khoury.northeastern.edu/sakmal/baseball_project`

5 Data Abstraction

Each row is a player and their statistics. The first attribute of Rank is ordinal attribute. League and Team are both categorical attributes, as one team or league is not better than the other. The other 26 attributes are all numeric statistics, meaning they are all quantitative attributes.

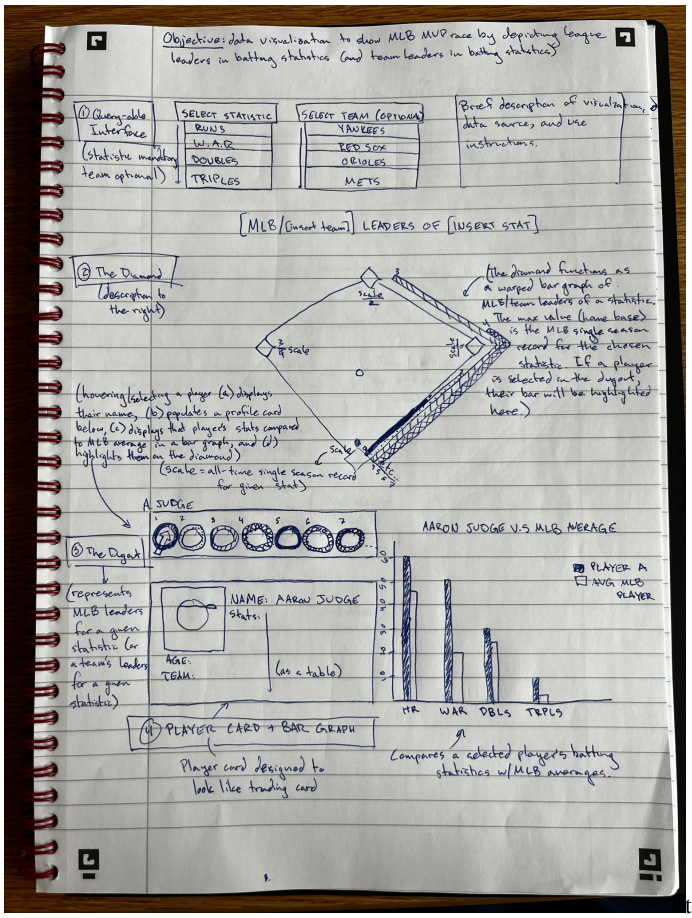
6 Use Case

Our visualization is designed for baseball fans and baseball coaches who want to learn statistics of MLB (Major League Baseball) players such that users can easily navigate the leading board or compare between players by filtering by types of play (e.g. homerun, triple, strikeout) or team names.

7 Design Process

Below are some initial drawings to depict the intended layout of the website and interactivity features of the visualization.

Overview: MLB MVP Race Visualization Drafted Sketch

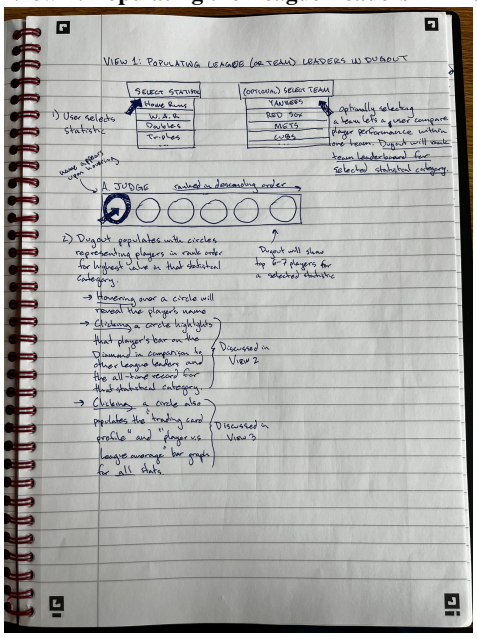


top

Description: The above polished sketch shows all the major visualization elements with some textual descriptions for each (although detials are provided in the three views shown later). Worth noting are the statistic/team selection dropdown's on the top of the page, the baseball diamond–which functions as a warped bar graph–in the center of the page, the dugout–which displays the MLB's top 6-7 league leader's for a statistical category–below the diamond to the left, the player card profile–indicating a player's box score season average statistics–below the dugout, and a player v.s MLB average bar graph comparison–revealing a player's performance across all at-bat statistical categories with the average

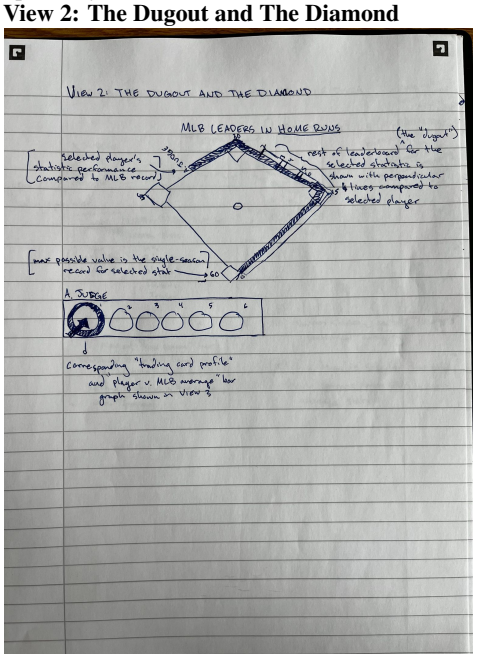
MLB player–to the right of the player card profile. Interactivity is discussed in the following views.

View 1: Populating the League Leaders in The Dugout



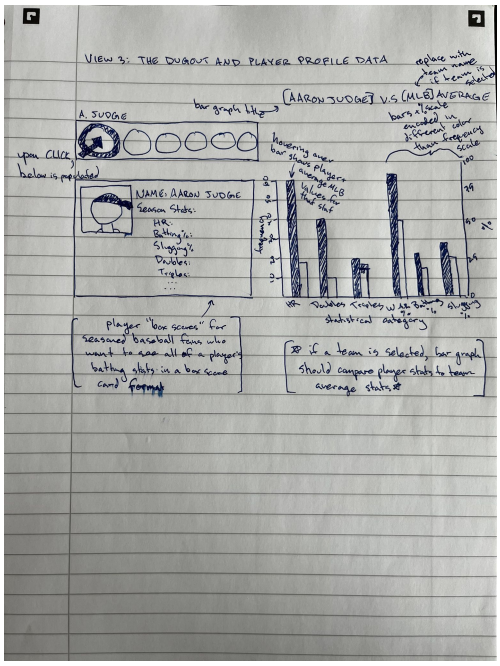
Description: The user is required to select a statistical category from a dropdown list. They can additionally select a team, if they wish to compare statistical leaders on a given team instead of comparing across the MLB. Upon selecting a statistical category, the dugout is automatically populated with the corresponding players who lead the league (or team) for that statistical category. Players are represented as circles which are bolded upon hovering. Hovering over a circle reveals the player's name, and clicking the circle will (a) highlight that player's bar on the warped bar graph around the baseball diamond (discussed in View 2) and (b) populate that player's profile and their comparison statistics v.s MLB averages in a bar graph (discussed in View 3). For visual smoothness, the page avoids using a "submit" button, and instead immediately populates and updates the dugout once a statistic (and optionally a team) are selected.

View 2: The Dugout and The Diamond



Description: Once the dugout is populated with league leaders for a statistical category, the diamond above the dugout is given a title "MLB LEADERS IN [INSERT CATEGORY]" and each player on the dugout is represented with a semi-opaque bar wrapping around the diamond. Clicking a player in the dugout decreases that player's bar's opacity and increases the opacity of the other bars in the warped bar graph. The scale for the bar graph is determined by the single-season all-time MLB record for that statistical category. This design was chosen in order to represent each player's progress for a single-season statistical category in comparison to an all-time performance, which would warrant MVP discussions for that player.

View 3: The Dugout and Player Profile Data



Description: Upon clicking a player in the dugout, that player's profile card appears below the dugout. This profile card will be designed like a baseball trading card to enhance the visual narrative of the data. The profile card will include a player headshot and their season totals (for count-based statistics) or season averages (for percent-based statistics) across all measured at-bat statistical categories. This design was influenced by box scores common across many team sports and will be a revamped version of anachronistic score cards from baseball's origin. Next to the score card will be a bar graph comparing that select player's season average statistics with the average of all MLB players' season average statistics. This data is often relevant for all-star and MVP selection to highlight well-rounded offensive play, so including a comparative bar graph for this data will communicate stellar at-bat productivity (or lack thereof) for players. Hovering over a bar will reveal the value of the bar. Because some statistical categories are measured as counts and others as percentages, a scale on the left will indicate counts and on the right will indicate percentage; bars corresponding to the percentage scale will be encoded with a different color than bars corresponding to the counts scale (each scale will match the bars' color encoding).

The Design's Evolution

Listed below are major storytelling designs and visual encodings that evolved and emerged over the development of this data presentation tool.

- **X-axis scale for selected player v.s MLB average comparative bar graph:** Initially, it was imagined that two

x-axis scales would be rendered. One would index the counts-based statistical categories while the other would scale percent-based statistical categories. Upon evaluating the dataset, there were only three reported statistical categories employing percent-based metrics: on-base percentage, slugging percentage, and on-base and slugging (combined) percentage. Keeping these three categories would have required a color encoding to delineate between counts-based and percentage-based statistics; testing the original design with a pilot group, the double x-axes and color encoding from inclusion of both decimal value percents and integer counts reduced the ease of interpretation. Since ease of interpretation is crucial to inform users who are not avid baseball fans, these three statistical categories were omitted from this view and only counts-based metrics were displayed. They were still included, however, in the player profile card.

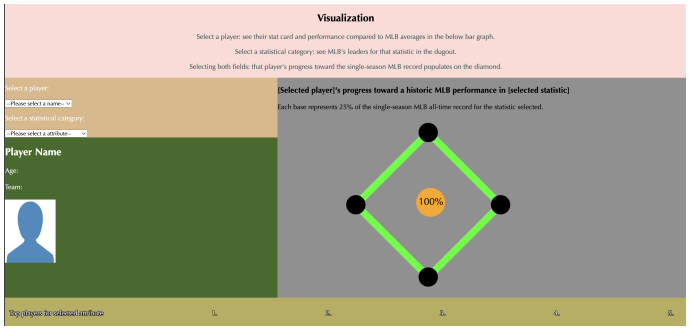
- **League leaders displayed on The Diamond:** The final visualization implements the intended Diamond functionality of displaying a player's performance in a statistical category toward the MLB single-season record (an exemplar). However, to minimize interpretation latency, it was decided to solely represent a single player on The Diamond. Making this design modification resulted in less clutter and grounded the axis of user interactivity to the "player selection" and "attribute selection" dropdowns. Based on feedback from user testing, displaying multiple players (a selected player compared to league leaders) on the Diamond would have added information clutter and pivoted the center of focus of the Diamond from player v.s exemplar comparisons to player v.s peer comparisons.

- **Player Querying in the Dugout:** The final visualization implements the intended Dugout functionality of displaying the league leaders for a statistical category. However, for the same reasons that it was decided to only display the selected player on the diamond, a user cannot select a player in the dugout. Again, this was intentionally done to focus user query attention to the two initial tasks (selecting a statistical category and selecting a player). Additionally, selecting a player in the dugout would provide a pathway away from the selected player which might obfuscate links/tooltips and the overall unison of visual storytelling surrounding the selected player.

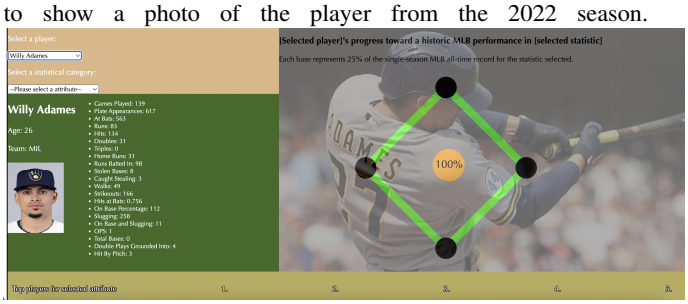
8 Final Design

Included below are four still frames of different stages of a user navigating the final visualization. Not shown are the background, user instructions, and video tutorial that are available on the website which hosts the visualization. If a user wishes to understand a specific player's performance in the MLB, they simply need to select that player from the drop down. To make base-rate and exemplar comparisons, they only need to select an additional statistical category in which that player will be compared. This design choice reduces MVP voting biases that arise from users weighing various statistical categories differently.

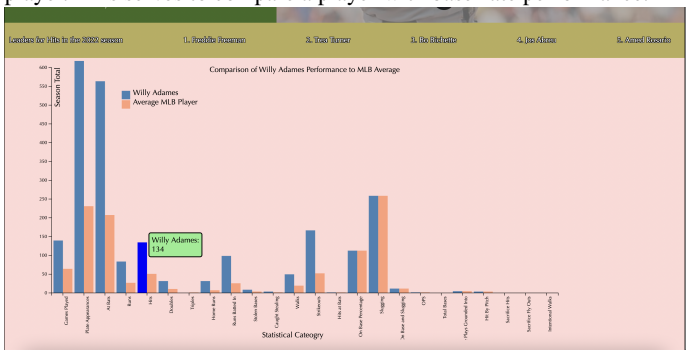
1. The initial visualization, unqueried: This includes the two drop down tabs for a user to select 1 of 700+ MLB players, and any of 20+ statistical categories measured. It also includes an empty player card (green; left) and an empty Diamond (grey; right) and Dugout (olive; below).



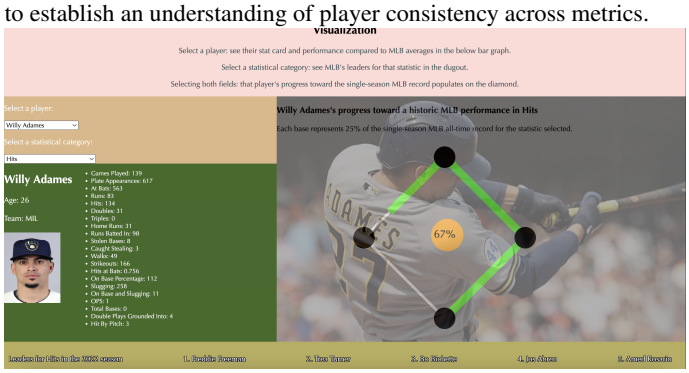
2. Selecting a player: When a user selects a player, the player card becomes populated with that player's statistical data. The Diamond background changes to show a photo of the player from the 2022 season.



3. Selecting a player (continued): When a user selects a player, a comparative bar graph below the Dugout is generated, comparing the player across all statistics with the average MLB player. This serves to compare a player with base-rate performance.



4. Selecting a statistic: When a user selects a statistical category, the Diamond is populated with a warped bar graph comparing the selected player with the MLB record holder for the selected statistic. This serves to compare a player with exemplar performance. Additionally, the dugout is populated with the top 5 leaders for the category in the league during the 2022 season. This serves as a non-base-rate and non-exemplar comparison, for users who want to establish an understanding of player consistency across metrics.



9 Discussion

The initial goal of developing this tool was to allow users–both MLB aficionados and MVP voters–to assess a given player's performance in context of reality without the influence of weighing various statistical categories to different degrees. To do this, a visualization must compare a player to the base-rate performance and to an exemplar performance in views isolating one statistic from the next. Additionally, this visualization should not have compromised the context of a given player's performance within their own strengths and weaknesses. For example, obscuring that a home run leader is also a strikeouts leader would be unfaithfully representative of the truth, regardless of how a user weighs home runs. With these goals in mind, the "MLB Race to the MVP" visualization was developed.

The resultant visualization accomplishes all these tasks through four crucial features; a player card to give the user a complete view of a player's performance (without context of the league's performance), a warped diamond bar graph to compare a player's performance per statistic with the MLB leader (exemplar) for that statistic, a dugout to compare a player with the current league leaders for a statistic (semi-exemplar comparison), and a comparative bar graph to compare a player's overall performance with the MLB average across nearly all statistical categories (base-rate comparison).

The visualization is limited in its ability to perform player-to-player comparisons between two selected players. Being able to perform this comparative analysis would allow a user to not only weigh a player's performance in league context, but in ranked order context. For example, comparisons between Shohei Ohtani and Aaron Judge per statistic against exemplars and base rate would provide a clear indication of the better offensive performer. This would be possible if the Diamond populated a bar for each player, two player cards were displayed, and the comparative bar graph had a color encoding for each player in addition to MLB averages. Additionally, the visualization is limited by the prerequisite knowledge of a desired player. Given its originally intended function, it does not populate the dugout for a statistical category a priori. This means a user cannot perform exploratory analysis on MLB MVP contenders using this tool, which restricts the user domain to users familiar with the MLB or to MVP voters.

In the future, this visualization could improve with assistance from graphic designers for the player profile card. With both time and skill asset limitations, it became increasingly difficult to prioritize designing an appealing baseball card-style profile that presented all the desired statistical metrics. Additionally, multiplayer selection would be a useful trait to perform head-to-head player comparisons, which could realistically be a useful tool for MVP voters struggling to determine rank order voting in a tight MVP race.

10 Conclusion

In conclusion, our project aimed to create intuitive and interactive visualizations that enable users to explore the statistical performances of MLB players from the 2022 season. Despite a shift in the design of our website during the framework development, our main visualization remained unchanged. Users can select a player and a statistic, and a warped bar graph will be generated to showcase how that player performed in comparison to the all-time record. The bar graph is in the shape of a baseball diamond. This intentional design choice not only adds visual appeal but also enhances the relevance and context of the visualization to the topic of MLB player statistics. By incorporating the baseball diamond shape, users can easily understand the performance of the selected player in relation to the field of play, making the visualization more intuitive and enjoyable for the user. In addition, when a player is selected, a clustered bar graph will display their statistics alongside the league average. The "dugout" feature will also populate with the top 5 players for the selected statistic, encouraging users to further explore the data.

Our project aims to provide an engaging and informative experience for users to gain insights into the statistical performances of MLB players. By leveraging visualizations and interactive features, we strive to make the data more accessible and understandable for users. The ability to compare individual players to historical records and league averages allows for a comprehensive analysis of their performances. Moreover, the inclusion of the "dugout" feature adds an element of exploration and discovery for users to delve deeper into the data.

Overall, our project contributes to the field of data visualization by creating a user-friendly and visually appealing platform for exploring MLB player statistics. Future enhancements could include incorporating additional visualizations, such as the ability to compare 2 players directly, to provide more comprehensive insights. Furthermore, expanding the dataset to include historical data from previous seasons could offer more context and comparisons for users. Our project serves as a foundation for further research and development in the field of sports analytics and data visualization.

Our project was a collaborative effort that involved contributions from each team member. Vivek was instrumental in developing various components of the project, including the dropdown selection feature, player card design, bar graph creation, and expertise in LaTeX formatting. He also played a crucial role in data cleaning and debugging, ensuring that our project was accurate.

Marco contributed significantly to the project by creating warped bar graph animations and python picture scraping functionality that allowed for dynamic and engaging visualizations. He also worked on changing the background design based on the data, updating visualizations to reflect the most recent data changes, and provided valuable input during debugging sessions to identify and resolve any issues.

Siraj made significant contributions to the project by providing the website layout framework, designing and updating the dugout feature, and establishing visual linkage between different parts of the project. He also developed the tooltip for the bar graph alongside Vivek. In addition, Siraj actively participated in debugging sessions, identifying and fixing issues promptly.

Each team member responded in a timely manner and demonstrated a high level of commitment to the project. Everyone's contributions were crucial to the success of the project, and the collaboration among team members was exemplary.

Acknowledgments

The authors wish to thank Professor Ab Mosca for their guidance and the TA staff Abigail Sodergren, Ashraf Bade, Luca Sharbani, and Khushi Morparia. Additional thanks to all users who participated in user testing and provided feedback on the visualization design.

References

A., R. Cummins D. Hahn. (2022). Selective exposure and exemplification within sports highlights. *Journal of Broadcasting and Electronic Media*. 26(1), 22-46. https://www.tandfonline.com/doi/full/10.1080/08838151.2021.2008938

Vuillemot, R., Stolper, C., Stasko, J., Wood, J. and Carpendale, S. (2018). State of the Art of Sports Data Visualization. *Computer Graphics Forum,* 37(3), pp. 663-686. doi:10.1111/cgf.13447

Transforming Data With Intelligence. (2022). Data Stories: Exploring Sports with Data Visualizations. *Online*. https://tdwi.org/articles/2022/11/16/bi-all-visualization-sports.aspx

Major League Baseball. (2023). The Top 100 Players Right Now: No. 1 Revealed. *MLB*. https://www.mlb.com/news/mlb-network-top-100-right-now-for-2023

FlowingData. (2007). Baseball Analytics: Little League baseball analytics that would change the game forever. https://flowingdata.com/tag/baseball/

Calcaterra, C. (2011). Here are the MVP voting criteria. *NBC-Sports*. https://mlb.nbcsports.com/2011/08/29/here-are-the-mvp-voting-criteria/

Honig, D. Ritter, L. S. (1988). 100 Greatest Baseball Players. *Random House Value Publishing*.