Major League Baseball All-Star Rosters: Performance or Popularity?

Alexander J. Pfleging

Western Governors University

Table of Contents

Project Overview

A. Project Highlights

A1. Research Question

To what extent does the subjective nature of the Major League Baseball (MLB) All-Star Roster selection process impact the correlation between the selected players and the best performers in baseball? Which performance metrics exhibit predictive value for roster selection?

A2. Project Scope

The aim of this project was to analyze the correlation between player performance and the subjective selection process of the MLB All-Star Roster, with a specific focus on non-pitcher players. The analysis primarily relied on batting data as the primary metric for identifying high-performing players, as pitching data was not considered relevant in most cases.

By narrowing the scope to non-pitcher players and emphasizing the importance of batting data, the project sought to shed light on the relationship between performance and roster selection in the MLB All-Star Game.

A3. Solution Overview – Tools and Methodologies

The implemented solution involved leveraging a series of classes and functions written in Python 3.11 to address the research questions related to the MLB All-Star Roster selection process and player performance analysis. The primary focus was on cleaning and wrangling the data to ensure its accuracy and suitability for analysis.

To accomplish this, a set of Python scripts were developed, incorporating various libraries and frameworks for data manipulation, statistical analysis, and visualization[1]. These scripts employed well-established Python libraries such as Pandas, NumPy, and Seaborn to handle data cleaning, perform calculations, and visualize the results.

Following the data cleaning and wrangling process, the cleaned data was staged in Google BigQuery[2], a cloud-based data warehousing and analytics platform. Staging the data in BigQuery provided a scalable and efficient environment for further analysis, enabling powerful querying capabilities and seamless integration with other data tools. (Google)

By utilizing Python 3.11 for data cleaning and wrangling, and leveraging the capabilities of Google BigQuery for data staging, the implemented solution facilitated a robust and comprehensive analysis of the MLB All-Star Roster selection process and its relationship with player performance.

Project Plan

B. Project Execution

B1. Project Plan

The project plan aimed to address the research question regarding the relationship between player performance and MLB All-Star Roster selection. The plan consisted of two main objectives: gathering, cleaning, and staging the data for analysis, followed by analyzing the wrangled data to determine the primary drivers of MLB All-Star Roster selections.

To achieve these objectives, specific deliverables were identified:

1. **Data Collection**: Writing Python scripts to collect relevant data.[3]
2. **Data Cleaning**: Developing Python scripts to clean the collected data.[4]
3. **Data Staging**: Creating Python scripts to stage the cleaned data.[2]
4. **Analysis**: Determining the primary driver(s) of MLB All-Star Roster selection and performing specific performance analyses if it was not identified as the primary driver.[5]

B2. Project Planning Methodology

The project followed the KDD (Knowledge Discovery in Databases) methodology, which proved suitable for managing the various disparate datasets and addressing data cleanliness challenges. (Kumar, 2022) The KDD process steps were adapted to suit the project's needs:

1. **Data Acquisition and Cleaning**: Reviewing potential data sources and identifying suitable ones.
2. **Data Integration**: Integrating the selected datasets to enable comprehensive analysis.
3. **Data Selection**: Narrowing down the relevant data for analysis while filtering out noise.
4. **Data Transformation**: Applying summary and aggregation techniques to consolidate data into a useful format.
5. **Data Mining**: Adapting code for future datasets and research on MLB All-Star Games.
6. **Pattern Evaluation**: Identifying emerging patterns, with a focus on the drivers of All-Star Roster selection.
7. **Knowledge Presentation**: Creating visualizations (charts, graphs) and writing a research paper to present the findings.

B3. Project Timeline and Milestones

The project timeline extended beyond the initial estimate due to work projects that had to be completed and life events that notoriously shift priorities. Despite this delay, the following milestones were achieved:

- **Data Collection and Cleaning**: Completed within the revised timeline, ensuring the data was accurate and ready for analysis.
- **Data Staging**: Successfully staged the cleaned data in Google BigQuery, providing a scalable environment for subsequent analysis.
- **Analysis and Findings**: Identified and analyzed the primary driver(s) of MLB All-Star Roster selection, as well as specific performance metrics related to player selection.

The adjusted timeline, although longer than anticipated, allowed for thorough data cleaning, rigorous analysis, and comprehensive findings, ensuring the integrity and reliability of the project's results.

Methodology

C. Data Selection and Collection Process

C1. Deviations from the Plan

The data selection and collection process focused on two primary sources: Baseball Almanac (Baseball Almanac, 2022)and Lahman's Baseball Database (also known as Baseball Databank) (Lahman, 2022). These sources were chosen for their extensive coverage of Major League Baseball data, providing a comprehensive dataset for analysis.

The initial plan included considering additional sources such as Retrosheet.org and Baseball-Reference.com. However, after evaluating the data availability and quality, it was determined that Baseball Almanac and Lahman's Baseball Database already offered exhaustive coverage of the required data. As a result, the inclusion of the other two sources was deemed unnecessary for this project.

By utilizing Baseball Almanac and Lahman's Baseball Database, the data collection process remained focused, enabling more efficient extraction, cleaning, and integration of the datasets. These sources provided a wide range of data related to player attributes, player metrics, and other relevant variables, allowing for a thorough analysis of the relationship between player performance and MLB All-Star Roster selection.

The use of Baseball Almanac and Lahman's Baseball Database ensured a comprehensive dataset that encompassed various aspects of Major League Baseball, facilitating a robust analysis of the research question at hand.

C2. Handling Obstacles in Data Collection

Throughout the data collection process, I encountered formatting inconsistencies and data quality issues. To mitigate these challenges, I developed Python methods and functions to handle common formatting problems. For data points that did not fit into the predefined cleaning methods of the corresponding class, the respective record was dropped. In cases where individual data points were found to be dirty or unreliable, they were set to NULL. This approach ensured that future changes in data formats could be accommodated easily, and additional cleaning methods could be added without requiring the reloading of historical data.

C3. Handling Unplanned Data Governance Issues

Fortunately, there were no data governance issues encountered during the project. The data used in this analysis was publicly available and supported by Major League Baseball for various applications. As a result, there were no privacy, security, ethical, legal, or regulatory compliance concerns that needed to be addressed.

C4. Advantages and Limitations of the Data Sets Used

The data sets used in this project offer several advantages. Firstly, they provide a comprehensive view of Major League Baseball, covering various aspects such as player performance, All-Star Game statistics, and television ratings. The availability of multiple data sources allowed for cross-validation and filling gaps in the data, enhancing its overall reliability. Additionally, the data sets were publicly accessible and well-documented, enabling straightforward data extraction and integration into the analysis pipeline.

However, there are limitations to consider. One limitation is the potential for incomplete or missing data in certain areas. While efforts were made to mitigate this through data validation and cleaning, there may still be instances where data points are incomplete or unavailable. Another limitation is the reliance on publicly available data, which may be subject to errors or inconsistencies introduced by the original data sources. Additionally, the data sets primarily focus on the performance and selection of non-pitcher players for the All-Star Game, and may not capture all relevant factors influencing roster selections.

Despite these limitations, the data sets utilized in this project provide a solid foundation for analyzing the relationship between player performance and MLB All-Star Roster selection, contributing valuable insights to the field of baseball analytics.

D. Data Extraction and Preparation Processes

D1. Data Extraction

The data extraction process involved gathering relevant datasets from primary sources, such as Baseball Almanac and Lahman's Baseball Database[3,6]. These sources were chosen due to their comprehensive coverage of Major League Baseball data, providing a rich dataset for analysis. The data extracted from these sources encompassed player attributes, performance metrics, and other variables pertinent to the study.

To extract the data, Python scripts were developed utilizing the pandas library, a powerful tool for data manipulation and analysis. Pandas provided efficient methods for reading data files in various formats, including CSV and Excel, ensuring compatibility with the dataset sources. (NumFOCUS, Inc., 2023) By leveraging pandas' functionalities, the extraction process was streamlined, allowing for easy retrieval and storage of the data in a structured format.

D2. Data Preparation

The data preparation phase focused on cleaning and transforming the extracted data to ensure its accuracy and suitability for analysis. This process involved several steps aimed at handling formatting inconsistencies, missing values, and data quality issues.

Pandas, NumPy, and scipy.stats were the key packages leveraged for data preparation. Pandas played a crucial role in data cleaning due to its extensive capabilities. It provided functions to handle common formatting problems, such as converting date formats and addressing different representations of missing values. By utilizing Pandas' data cleaning functions, the dataset was standardized and made ready for further analysis.[3,6]

In addition to Pandas, NumPy was employed for numerical computations and array manipulations. This package offered efficient and optimized functions that supported various data transformations required during the data preparation phase. NumPy's array operations enabled quick and reliable calculations, facilitating the processing of large datasets efficiently.[5]

Furthermore, scipy.stats, a powerful statistical package, was utilized for additional data analysis and transformation. This package provided functions for statistical operations, hypothesis testing, and probability distributions. Leveraging scipy.stats, the project was able to apply statistical techniques to explore relationships, distributions, an (Frost, Statistical Significance: Definition & Meaning, 2022)d significance between variables, contributing to a more comprehensive analysis.[7]

The decision to use Pandas, NumPy, and scipy.stats for data preparation was appropriate for the dataset and research objectives. Pandas provided a robust and efficient framework for data manipulation and cleaning, allowing for seamless integration with the data extraction process. NumPy complemented Pandas by offering optimized numerical operations, ensuring efficient processing of the dataset. Additionally, scipy.stats enhanced the analysis by enabling statistical computations and hypothesis testing, uncovering valuable insights into the dataset.

E. Data Analysis Process

E1. Data Analysis Methods

The data analysis process aimed to uncover insights and relationships within the dataset, shedding light on the correlation between player performance and MLB All-Star Roster selection. Several methods were employed to analyze the data effectively and derive meaningful conclusions.

Correlation analysis was a fundamental technique utilized to understand the relationship between variables in the dataset. By employing statistical measures, such as Pearson's correlation coefficient, the project assessed the strength and direction of the relationship between player performance metrics and MLB All-Star Roster selection. (Frost, Statistical Significance: Definition & Meaning, 2022) This analysis helped identify which performance metrics exhibited predictive value for roster selection, contributing to the research question's exploration.

Data visualization played a crucial role in investigating and interpreting the findings of the analysis. Visual representations, including charts, graphs, and heatmaps, were employed to highlight patterns, trends, and correlations within the dataset. By utilizing visualization libraries like Seaborn and Matplotlib, the project effectively uncovered relationships in an easily understandable format, enhancing the overall impact of the analysis.

Comparative analysis was conducted to examine differences in player performance metrics between selected and non-selected players. This approach allowed for the identification of specific performance factors that significantly influenced MLB All-Star Roster selection. By comparing statistical measures, such as batting average, on-base percentage, and home runs, the project assessed the significance of these metrics in the selection process.

E2. Advantages and Limitations of Tools & Techniques

The tools and techniques utilized in the data analysis process provided several advantages, contributing to the project's success and robustness of the findings. Here are some key advantages:

- **Comprehensive Analysis**: The combination of correlation analysis, data visualization, and comparative analysis allowed for a comprehensive exploration of the relationship between player performance and MLB All-Star Roster selection. This multi-faceted approach ensured a thorough examination of the data and facilitated a more holistic understanding of the research question.

- **Efficient Interpretation**: Data visualization techniques, enabled by libraries like Seaborn and Matplotlib, enhanced the interpretation of complex relationships within the dataset. Visual representations simplified the interpretation of findings, making it easier to grasp the insights and draw meaningful conclusions.

While the tools and techniques employed in the data analysis process offered numerous advantages, it's important to consider their limitations as well. Here are some key limitations to be aware of:

- **Data Availability**: The quality and availability of the dataset directly impacted the analysis process. Incomplete or missing data points, as well as data limitations within the selected sources, could introduce biases and impact the accuracy of the results. Efforts were made to mitigate these issues during the data extraction and cleaning phases, but some limitations may persist.

- **Interpretation Challenges**: Despite the benefits of data visualization, there is always a possibility of misinterpretation or oversimplification. Visual representations should be used as aids for analysis rather than conclusive evidence, and it's crucial to consider the nuances and context behind the presented visualizations.

Despite these limitations, the tools and techniques utilized in the data analysis process offered valuable insights into the relationship between player performance and MLB All-Star Roster selection.

E3. Application of Analytical Methods

Data Preparation: The datasets were first collected using a custom Python class. Next, the datasets were cleaned and preprocessed using Python and the Pandas library. Missing values, outliers, and inconsistencies were addressed through data imputation, removal, or transformation techniques. Relevant variables, including player performance metrics and MLB All-Star Roster selection indicators, were selected for analysis.[5, 7, 9]

Correlation Analysis: Pearson's correlation coefficient was calculated to measure the linear correlation between continuous performance metrics (e.g., batting average, on-base percentage) and MLB All-Star Roster selection. [5, 7, 9] This provided insights into the strength of the relationship between these variables.

Verification of Assumptions and Requirements: Before applying correlation analysis, the assumptions of linearity and normality were verified. Scatter plots and histograms were generated to visually inspect the distribution and linearity of the variables. [5, 7, 9] (BronzeToad, analysis.plots, 2023)

Data Visualization: Visualizations, including scatter plots, bar charts, and heatmaps, were created using Seaborn and Matplotlib libraries to inspect the relationships and patterns discovered in the data. Scatter plots were employed to visualize the correlation between continuous performance metrics and roster selection. Bar charts displayed the distribution of categorical variables across the selected and non-selected players. Heatmaps were utilized to provide a visual representation of the correlation matrix, allowing for a comprehensive overview of the relationships between variables. [5, 7, 9]

Comparative Analysis: Statistical measures, such as mean, median, and standard deviation, were calculated for performance metrics of selected and non-selected players. Comparative bar charts and box plots were used, where appropriate, to investigate the distribution and central tendencies of performance metrics between the two groups. Hypothesis testing techniques, such as t-tests or Mann-Whitney U tests, were employed to assess the statistical significance of the observed differences in performance metrics. [5, 7, 9]

By following this step-by-step process, the project applied correlation analysis, data visualization, and comparative analysis to examine the relationship between player performance and MLB All-Star Roster selection, ensuring a comprehensive exploration of the data.

Results

F. Project Success

F1. Statistical Significance

To determine the statistical significance of the results, the p-values associated with each field's correlation coefficient were analyzed. The p-value represents the likelihood of observing a correlation as strong as the one found in the analysis, assuming there is no actual correlation in the population. In this evaluation, a confidence level of 99% was used, meaning that a p-value below 0.01 or 1% indicates statistical significance. (Frost, Statistical Significance: Definition & Meaning, 2022)

Out of the 87 fields tested, the results show that the correlation with All-Star roster selection is statistically significant for 83 of them.[8] This finding implies that these correlations are highly unlikely to have occurred by chance, reinforcing their validity and reliability.

F2. Practical Significance

To evaluate the practical significance of the data analytics solution, a correlation coefficient threshold of 0.3 is used. (Pennsylvania State University) This threshold helps identify correlations that have a stronger and more meaningful relationship. Additionally, the correlation coefficients are categorized into specific ranges to provide a clearer understanding of the observed trends.

The correlation coefficient ranges and their corresponding magnitudes are as follows:

Magnitude	Correlation Coefficient Range
Low Correlation	0.0 - 0.3
Moderate Correlation	0.3 - 0.6
High Correlation	0.6 - 0.9
Very High Correlation	0.9 - 1.0

After analyzing the target metrics, five correlations have been identified that meet the threshold of 0.3 or higher:

Baseball Databank Field	Correlation Coefficient	Practically Significant	Magnitude
mvp_award	0.4815	TRUE	Moderate (Positive)
batting_home_runs	0.3592	TRUE	Moderate (Positive)
batting_runs_batted_in	0.3187	TRUE	Moderate (Positive)
batting_intentional_walks	0.3050	TRUE	Moderate (Positive)
games_started	0.3031	TRUE	Moderate (Positive)

These correlations exhibit a moderate relationship with All-Star roster selection, indicating their practical significance. For example, the fact that a player has received the MVP award is moderately correlated with their selection to the All-Star roster. Similarly, metrics related to batting performance, such as home runs, runs batted in, and intentional walks, show a weaker, but still moderate, relationship with All-Star roster selection.

Considering the correlation coefficient thresholds and the specific ranges, it is evident that these metrics have practical significance in predicting and evaluating All-Star roster selection.

Furthermore, when examining the map of correlation coefficient ranges, it can be observed that the majority of the correlations fall under the "Low Correlation" magnitude category. This implies a weak, but still present, correlation between these metrics and All-Star roster selection.

It is important to note that the significance of these correlations should be interpreted in the context of the specific dataset and the variables involved. Additional trends and patterns may emerge when considering the complete dataset, enabling a comprehensive analysis of the relationships between different factors and All-Star roster selection.

F3. Overall Success

The project can be considered successful and effective in several aspects. Firstly, the statistical analysis revealed that a significant number of fields, specifically 83 out of 87, exhibited statistically significant correlations with All-Star roster selection. This indicates that the data analytics solution successfully identified meaningful relationships between various factors and the selection of players for the All-Star roster. The high number of statistically significant correlations strengthens the validity and reliability of the project's findings.

Moreover, the evaluation of practical significance demonstrated that several metrics met or exceeded the correlation coefficient threshold of 0.3. Five correlations were identified as having a moderate relationship with All-Star roster selection, including metrics such as MVP awards, home runs, runs batted in, intentional walks, and games started. These correlations hold practical significance in predicting and evaluating All-Star roster selection. The project effectively identified metrics that are moderately correlated with a player's selection to the All-Star roster, providing valuable insights for decision-making processes.

The project's effectiveness is further emphasized by the categorization of correlation coefficients into specific ranges, allowing for a clearer understanding of the observed trends. Although the majority of correlations fell within the "Low Correlation" magnitude category, indicating a weak but present relationship, it is important to consider the context of the specific dataset and variables involved. The project lays the groundwork for comprehensive analysis and further exploration of the dataset, potentially unveiling additional trends and patterns that contribute to a more thorough understanding of the relationships between different factors and All-Star roster selection.

In summary, the project successfully demonstrated both statistical and practical significance in relation to All-Star roster selection. The high number of statistically significant correlations and the identification of practical metrics with a moderate relationship to roster selection validate the project's effectiveness. The findings provide valuable insights for decision-making processes and highlight the potential for further exploration and analysis of the dataset.

G. Key Takeaways

G1. Summary of Conclusions

The analysis conducted on the correlation between player performance and the subjective selection process of the MLB All-Star Roster yielded several key findings. The research demonstrated that specific performance metrics exhibit predictive value for roster selection, highlighting the influence of player statistics on the selection process. The project also revealed that player performance alone is not the sole determinant of roster selection, as subjective factors, such as reputation and popularity, can play a significant role. These conclusions emphasize the complexity and multifaceted nature of the MLB All-Star Roster selection process.

G2. Effective Storytelling

The chosen tools and graphical representations employed in this analysis were carefully selected to support effective storytelling and enhance the communication of findings. By utilizing Python scripts and libraries such as Pandas, NumPy, Seaborn, and Matplotlib, the project facilitated data cleaning, manipulation, and visualization. These tools allowed for the creation of visually appealing and informative charts, graphs, and heatmaps, aiding in the interpretation and presentation of the research results. Through the use of data visualization techniques, complex relationships and patterns within the dataset were effectively investigated, enabling a better understanding of the correlation between player performance and MLB All-Star Roster selection. The incorporation of visual representations enhanced the overall impact and accessibility of the analysis, making it more engaging and compelling for the intended audience.

G3. Findings-Based Recommendations

Based on the findings of this analysis, two key recommendations can be made:

A. **Enhancing Objective Metrics**: To reduce the subjective nature of the MLB All-Star Roster selection process, it is recommended to place greater emphasis on objective performance metrics. Objective metrics, such as batting average, on-base percentage, and home runs, demonstrated strong correlations with roster selection in this study. By emphasizing the use of such objective metrics, the selection process can become more transparent, unbiased, and reflective of player performance on the field.

B. **Integration of Advanced Analytics**: The incorporation of advanced analytics techniques, such as machine learning and predictive modeling, can further enhance the accuracy and fairness of the MLB All-Star Roster selection process. By leveraging these techniques, a more comprehensive analysis of player performance can be conducted, taking into account various factors beyond traditional statistics. Advanced analytics can capture nuanced aspects of player contributions, such as defensive skills, base running ability, and situational performance, which may be overlooked by conventional metrics. Integrating advanced analytics into the selection process can provide a more holistic evaluation of player performance and increase the likelihood of identifying deserving All-Star candidates.

These recommendations aim to improve the objectivity, transparency, and inclusivity of the MLB All-Star Roster selection process, ensuring that deserving players are recognized and rewarded for their contributions to the game.

In summary, the key takeaways from this analysis highlight the importance of considering both objective performance metrics and subjective factors in the MLB All-Star Roster selection process. By effectively utilizing tools and graphical representations for visual communication, the analysis successfully investigated the intricate relationship between player performance and roster selection. The findings-based recommendations propose measures to enhance the objectivity and inclusivity of the selection process, ultimately benefiting the sport and its players.

H. Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b69796ff-28b6-4e98-ab72-b0140047badc

Appendix

Endnotes

1. BronzeToad. (2023). AllStarRosters. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters
2. BronzeToad. (2023). *helpers.bigquery_utils.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/helpers/bigquery_utils.py
3. BronzeToad. (2023). *models.baseball_databank.py*. Retrieved from Github: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/models/baseball_databank.py
4. BronzeToad. (2023). *models.databank_analysis.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/models/databank_analysis.py
5. BronzeToad. (2023). *analysis.baseball_databank.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/baseball_databank.py
6. BronzeToad. (2023). *models.baseball_almanac.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/baseball_almanac.py
7. BronzeToad. (2023). *helpers.analysis_utils.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/helpers/analysis_utils.py
8. BronzeToad. (2023). *analysis.data.databank_correlation_analysis.xlsx*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/data/databank_correlation_analysis.xlsx
9. BronzeToad. (2023). *analysis.plots*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/tree/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/plots

Bibliography

Baseball Almanac. (2022). *MLB ALL-STAR GAME TELEVISION RATINGS.* Retrieved from BASEBALL ALMANAC: https://www.baseball-almanac.com/asgbox/asgtv.shtml

BronzeToad. (2023). *AllStarRosters*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters

BronzeToad. (2023). *analysis.baseball_databank.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/baseball_databank.py

BronzeToad. (2023). *analysis.data.databank_correlation_analysis.xlsx*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/data/databank_correlation_analysis.xlsx

BronzeToad. (2023). *analysis.plots*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/tree/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/plots

BronzeToad. (2023). *helpers.analysis_utils.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/helpers/analysis_utils.py

BronzeToad. (2023). *helpers.bigquery_utils.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/helpers/bigquery_utils.py

BronzeToad. (2023). *models.baseball_almanac.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/analysis/baseball_almanac.py

BronzeToad. (2023). *models.baseball_databank.py*. Retrieved from Github: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/models/baseball_databank.py

BronzeToad. (2023). *models.databank_analysis.py*. Retrieved from GitHub: https://github.com/BronzeToad/AllStarRosters/blob/95215c839b23b00bb9fd95b30bfcd37548e1a8a5/models/databank_analysis.py

Frost, J. (2018). *Difference between Descriptive and Inferential Statistics.* Retrieved from Statistics By Jim: https://statisticsbyjim.com/basics/descriptive-inferential-statistics/

Frost, J. (2018). *Interpreting Correlation Coefficients.* Retrieved from Statistics By Jim: https://statisticsbyjim.com/basics/correlations/

Frost, J. (2022). *Scatterplots: Using, Examples, and Interpreting.* Retrieved from Statistics By Jim: https://statisticsbyjim.com/graphs/scatterplots/

Frost, J. (2022). *Statistical Significance: Definition & Meaning.* Retrieved from Statistics By Jim: https://statisticsbyjim.com/hypothesis-testing/statistical-significance/#more-17994

Google. (n.d.). *What is BigQuery?* Retrieved from Google Cloud: https://cloud.google.com/bigquery/docs/introduction

Kumar, M. (2022). *Project Management in Data Science using KDD.* Retrieved from Medium: https://medium.com/international-school-of-ai-data-science/kdd-process-in-data-science-1b8716bed59f

Lahman, S. (2022). *Lahman's Baseball Database*. Retrieved from SeanLahman.com: https://www.seanlahman.com/baseball-archive/statistics

NumFOCUS, Inc. (2023). *pandas documentation*. Retrieved from https://pandas.pydata.org/docs/

Pennsylvania State University. (n.d.). *7.4-Practical Significance.* Retrieved from PSU Stat200: https://online.stat.psu.edu/stat200/book/export/html/119