# ***Moneyball***

Ethan Xin: xin.e@northeastern.edu

Talal Fakhoury: fakhoury.t@northeastern.edu

Josh Longo: longo.j@northeastern.edu

Eitan Berenfeld: berenfeld.e@northeastern.edu

Northeastern University

DS2500: Intermediate Programming with Data

Professor Rush Sanghrajka

August 16th, 2024

**Problem Statement and Background:**

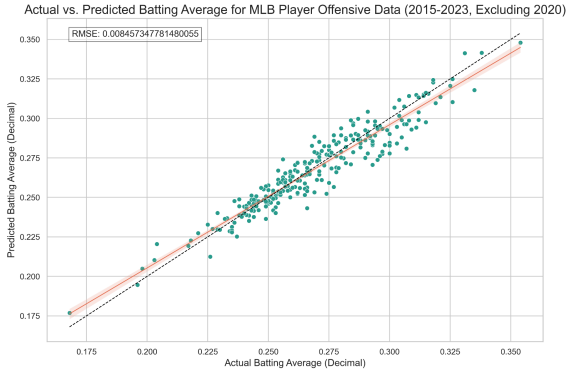
On the surface, the sport of baseball may seem simple, throw the ball, swing the bat, and hit it as far as possible. However, the game is full of a wealth of statistics that allow for a considerable amount of in-depth statistical analysis. With detailed records of every pitch and swing spanning over 70 years, baseball offers a unique opportunity for teams, players, analysts, and students. This project tackles six key research questions, exploring everything from predictive models of player performance and positional advantages to the impact of weather, elevation, rule changes, and financial success on team outcomes. Since the mid-20th century, the science of baseball, known as sabermetrics, has revolutionized how teams utilize statistics to approach the game. With the introduction of Statcast in 2015, nearly every aspect of a player's game can now be measured, allowing for more informed decision-making. Our project intends to contribute to this ongoing evolution by providing insights that can enhance player development, improve valuations, and sustain a competitive advantage in an increasingly data-driven sport. Understanding these different factors is crucial not just for teams and analysts but for anyone invested in the future of baseball and specifically optimal offensive production.

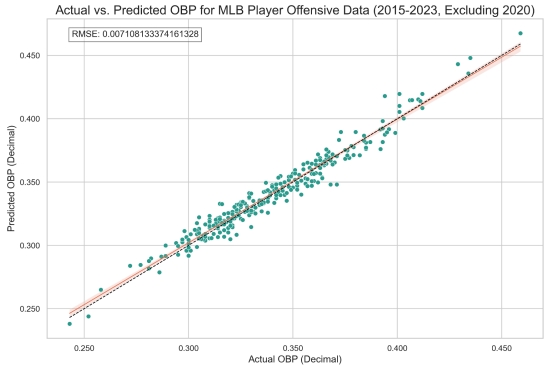
**Analysis 1: Offensive Metric Predictive Analysis**

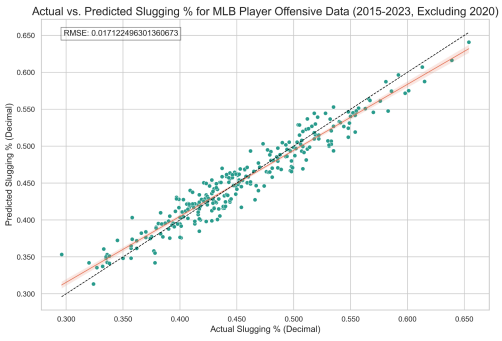
The focus of this analysis was to create a predictive machine learning model that predicts key offensive performance indicators: batting average (BA), on-base percentage (OBP), and slugging percentage (SLG). We used a dataset of detailed offensive statistics for qualified batters (those with over 500 plate appearances) from the 2015 to 2023 seasons, excluding the 2020 season due to its incomplete nature, which could introduce anomalies that affect the accuracy of the model. This data was exported as CSV files from the official MLB statistics database known as Baseball Savant. In addition, it was collected through a combination of in-game tracking

technologies such as Statcast, which measures various aspects of player performance, including exit velocity, launch angle, and swing percentages. The data represents individual players' yearly offensive metrics, and there were no inherent privacy concerns given the data was publicly available sports performance data. There could be potential biases due to the exclusion of players under 500 plate appearances, which may skew the data toward more consistent, high-performing players, potentially limiting the generalizability of the findings. We utilized multiple linear regression models to predict BA, OBP, and SLG based on a variety of independent variables. The scikit-learn library was emplyed for the train test split and regression functions to create our machine learning model. For batting average we used independent variable that we believed were related to a player getting a hit which included singles, strikeout percentage (K%), SLG, average exit velocity, average launch angle, whiff percentage, swing percentage, line drive percentage, pop up percentage, meatball percentage, barrel batted rate, sweet spot percentage, solid contact percentage, and hard hit percentage. For OBP we used metrics that we believed impacted a player’s ability to get on base along with hitting and those variable included K%, walk percentage (BB%), batting average on balls in play (BABIP), average exit velocity, average launch angle, sweet spot percentage, barrel batted rate, whiff percentage, solid contact percentage, meatball percentage, hard hit percentage, zone swing miss percentage (Z-Swing Miss %), and out-of-zone contact percentage (OZ Contact %). Finally for slugging we focused on advanced statistics that were related to power hitting and those were isolated power (ISO), barrel batted rate, BB%, average exit velocity, average launch angle, sweet spot percentage, solid contact percentage, hard hit percentage, OZ Contact %, line drive percentage, and flyball percentage. These variables include both traditional baseball statistics (e.g., strikeout percentage, walk percentage) and more advanced metrics (e.g., exit velocity, launch angle, barrel rate). The

models were trained on a subset of the data and evaluated using Root Mean Square Error (RMSE) on the test set, with scatter plots and regression lines illustrating the relationship between actual and predicted values. A dotted line was added to the graphs to visualize the perfect prediction scenario. The results of the regression analysis revealed that the certain metrics chosen were significant predictors of batting average, OBP, and SLG. The model for BA achieved an RMSE of 0.0085, for OBP 0.0071, and for SLG 0.0171, highlighting the relative accuracy and challenges of each model as seen below.



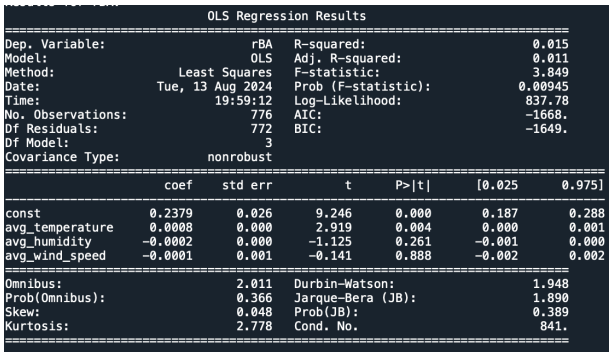

These findings highlight the importance of understanding the underlying factors that contribute to offensive performance. Witht the availability of advanced statistics through Statscast these models can be created with high accuracy to be used in practical applications, such as player evaluation and in-game decision-making. Future research could enhance these models by incorporating additional player attributes like height and years in the league, checking for

multicollinearity, and using advanced techniques like ridge or lasso regression to prevent overfitting. Additionally, k-fold cross-validation could improve model generalization and performance.

**Analysis 2: Does Weather Impact Offensive Metrics?**

Our second analysis focused on the impact of various weather conditions on offensive baseball metrics during the Boston Red Sox's home games from 2013 to 2023, excluding the 2020 season. We concentrated on a single team's home games, with the purpose of eliminating the variability associated with different ballparks and away games. The goal was to determine how temperature, humidity, and wind speed influence key performance indicators like BA, OBP, SLG, and on-base plus slugging percentage (OPS). External environmental conditions have always been thought to affect athletic performance in any sport. In baseball, colder temperatures may cause players to stiffen, higher humidity levels could slow them down, and increased wind speed might affect their swing. The data for this analysis includes historical temperature, humidity, and wind speed metrics by the hour at the longitude and latitude of Fenway Park that was obtained through the Open-Meteo API. Scraping baseball box score data from websites like Baseball Reference became difficult for the extensive volume of data we were scraping due to frequent page structure changes, and rate limits or blocking by the websites. Therefore, we used pybaseball, a python library that contains individual game data for the Red Sox as well as MLB-StatsAPI, a python wrapper for complete game start times. The game logs, sourced from pybaseball, included team specific in-game team statistics. Doubleheaders were removed because they posed a risk to the analysis where offensive metrics could be impacted by fatigue or lineup changes common in doubleheaders. For each game weather metrics were found by

rounding the exact time to the nearest hour and then taking the average from that point until three hours later which is about the average time of a baseball game. In addition, BA, OBP, SLG, and OPS were all recalculated for each game to ensure the accuracy of the statistics. There are no significant privacy concerns with this data since it pertains to publicly available sports events and meteorological records. This analysis employed multiple statistical techniques to explore the relationships between weather conditions and offensive metrics. The primary method used was Ordinary Least Squares (OLS) regression to model the impact of each weather variable on BA, OBP, OPS, and SLG which was done four times using statsmodel.api. The metrics revealed that weather conditions explained only 1.5% of the variance in BA, 0.4% in OBP, 1.6% in OPS, and 2.3% in SLG, indicating a very weak relationship across the board as seen below.

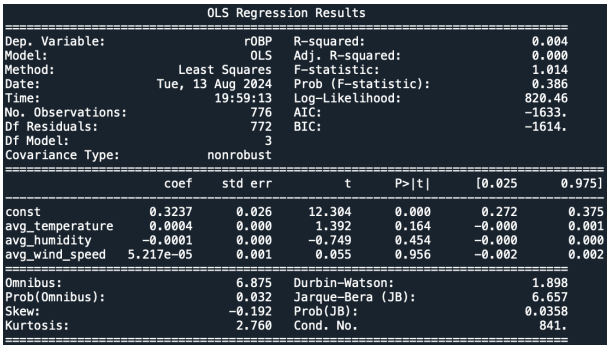


OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | rBA | R-squared: | 0.015 |
| Model: | OLS | Adj. R-squared: | 0.011 |
| Method: | Least Squares | F-statistic: | 3.849 |
| Date: | Tue, 13 Aug 2024 | Prob (F-statistic): | 0.00945 |
| Time: | 19:59:12 | Log-Likelihood: | 837.78 |
| No. Observations: | 776 | AIC: | -1668. |
| Df Residuals: | 772 | BIC: | -1649. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.2379 | 0.026 | 9.246 | 0.000 | 0.187 | 0.288 |
| avg_temperature | 0.0008 | 0.000 | 2.919 | 0.004 | 0.000 | 0.001 |
| avg_humidity | -0.0002 | 0.000 | -1.125 | 0.261 | -0.001 | 0.000 |
| avg_wind_speed | -0.0001 | 0.001 | -0.141 | 0.888 | -0.002 | 0.002 |

| | | | |
|---|---|---|---|
| Omnibus: | 2.011 | Durbin-Watson: | 1.948 |
| Prob(Omnibus): | 0.366 | Jarque-Bera (JB): | 1.890 |
| Skew: | 0.048 | Prob(JB): | 0.389 |
| Kurtosis: | 2.778 | Cond. No. | 841. |

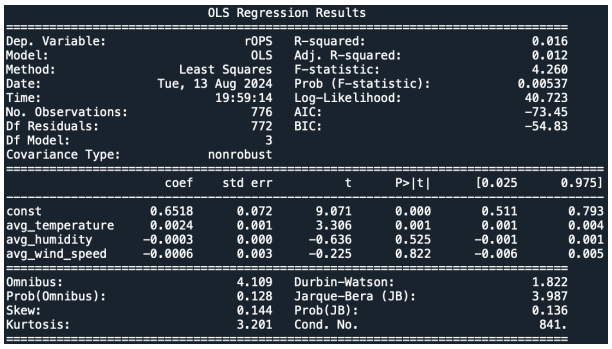


OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | rOBP | R-squared: | 0.004 |
| Model: | OLS | Adj. R-squared: | 0.000 |
| Method: | Least Squares | F-statistic: | 1.014 |
| Date: | Tue, 13 Aug 2024 | Prob (F-statistic): | 0.386 |
| Time: | 19:59:13 | Log-Likelihood: | 820.46 |
| No. Observations: | 776 | AIC: | -1633. |
| Df Residuals: | 772 | BIC: | -1614. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.3237 | 0.026 | 12.304 | 0.000 | 0.272 | 0.375 |
| avg_temperature | 0.0004 | 0.000 | 1.392 | 0.164 | -0.000 | 0.001 |
| avg_humidity | -0.0001 | 0.000 | -0.749 | 0.454 | -0.000 | 0.000 |
| avg_wind_speed | 5.217e-05 | 0.001 | 0.055 | 0.956 | -0.002 | 0.002 |

| | | | |
|---|---|---|---|
| Omnibus: | 6.875 | Durbin-Watson: | 1.898 |
| Prob(Omnibus): | 0.032 | Jarque-Bera (JB): | 6.657 |
| Skew: | -0.192 | Prob(JB): | 0.0358 |
| Kurtosis: | 2.760 | Cond. No. | 841. |

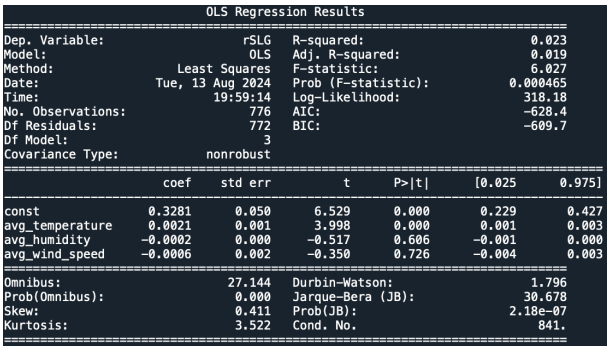


OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | rOPS | R-squared: | 0.016 |
| Model: | OLS | Adj. R-squared: | 0.012 |
| Method: | Least Squares | F-statistic: | 4.260 |
| Date: | Tue, 13 Aug 2024 | Prob (F-statistic): | 0.00537 |
| Time: | 19:59:14 | Log-Likelihood: | 40.723 |
| No. Observations: | 776 | AIC: | -73.45 |
| Df Residuals: | 772 | BIC: | -54.83 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.6518 | 0.072 | 9.071 | 0.000 | 0.511 | 0.793 |
| avg_temperature | 0.0024 | 0.001 | 3.306 | 0.001 | 0.001 | 0.004 |
| avg_humidity | -0.0003 | 0.000 | -0.636 | 0.525 | -0.001 | 0.001 |
| avg_wind_speed | -0.0006 | 0.003 | -0.225 | 0.822 | -0.006 | 0.005 |

| | | | |
|---|---|---|---|
| Omnibus: | 4.109 | Durbin-Watson: | 1.822 |
| Prob(Omnibus): | 0.128 | Jarque-Bera (JB): | 3.987 |
| Skew: | 0.144 | Prob(JB): | 0.136 |
| Kurtosis: | 3.201 | Cond. No. | 841. |



OLS Regression Results

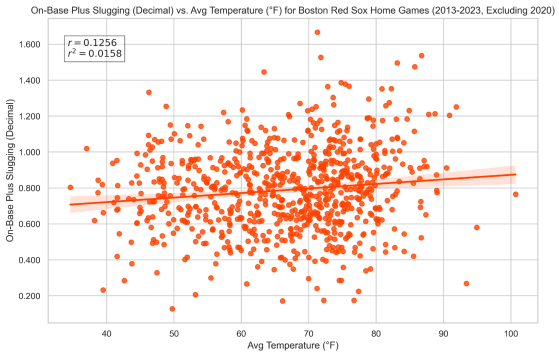
| | | | |
|---|---|---|---|
| Dep. Variable: | rSLG | R-squared: | 0.023 |
| Model: | OLS | Adj. R-squared: | 0.019 |
| Method: | Least Squares | F-statistic: | 6.027 |
| Date: | Tue, 13 Aug 2024 | Prob (F-statistic): | 0.000465 |
| Time: | 19:59:14 | Log-Likelihood: | 318.18 |
| No. Observations: | 776 | AIC: | -628.4 |
| Df Residuals: | 772 | BIC: | -609.7 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

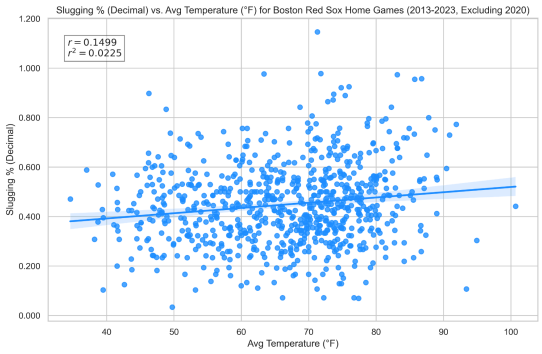
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.3281 | 0.050 | 6.529 | 0.000 | 0.229 | 0.427 |
| avg_temperature | 0.0021 | 0.001 | 3.998 | 0.000 | 0.001 | 0.003 |
| avg_humidity | -0.0002 | 0.000 | -0.517 | 0.606 | -0.001 | 0.000 |
| avg_wind_speed | -0.0006 | 0.002 | -0.350 | 0.726 | -0.004 | 0.003 |

| | | | |
|---|---|---|---|
| Omnibus: | 27.144 | Durbin-Watson: | 1.796 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 30.678 |
| Skew: | 0.411 | Prob(JB): | 2.18e-07 |
| Kurtosis: | 3.522 | Cond. No. | 841. |

However, the coefficients from the regression models were analyzed to interpret the relative importance of each weather variable, which revealed that average temperature had a statistically

significant impact on offensive performance metrics like OPS and SLG. Average humidity and wind speed had a negligible effect on the selected offensive metrics. Therefore, we decided to narrow our focus on the individual relationship between average temperature and OPS and SLG to see if the r squared and r value would be higher. We created two Scatter plots with regression lines to visualize the correlation between average temperature and OPS and SLG as seen below.

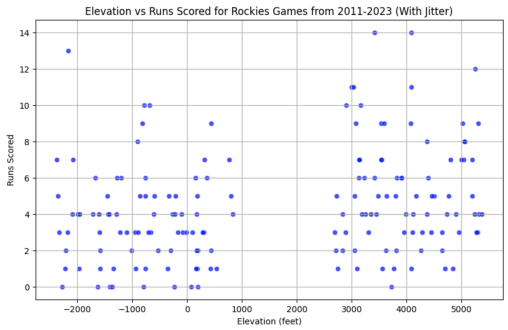



The scatter plot models showed an extremely weak positive correlation between average temperature and both OPS and SLG, with r-values of 0.1256 and 0.1499, respectively. These findings suggest that while warmer temperatures may slightly benefit power hitters, weather conditions alone have a minimal impact on offensive output. Other factors are more influential and must be considered when looking at the variation of OPS and SLG. Future research could expand the analysis to other teams and stadiums, consider additional weather variables like air pressure and precipitation, and explore the effects of weather on pitching and fielding metrics. Lastly, if we were to do this again, implementing advanced machine learning models, like random forests or neural networks, might be beneficial to uncover non-linear relationships and strengthen our understanding of weather's role in player performance.

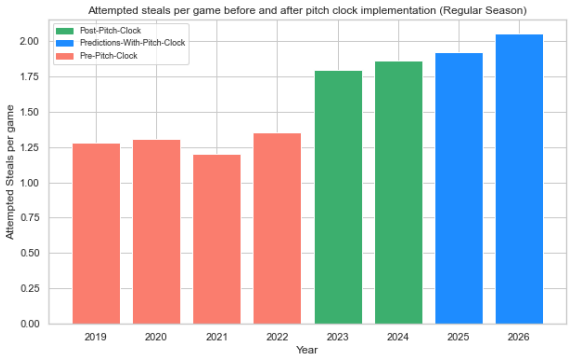
**Analysis 3: How Does Elevation Affect Offense?**

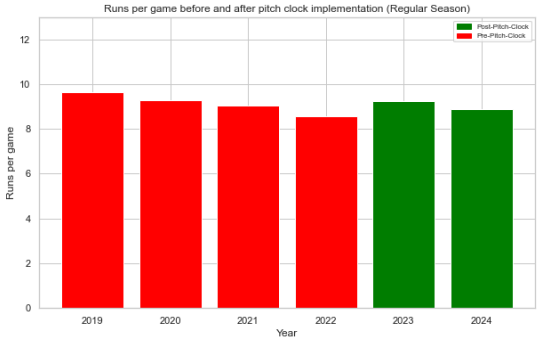
The common perception about elevation is that it helps offenses which in part is true. Air is thinner at higher elevations, resulting in less air resistance and thus, balls fly further. For this analysis, we examined a couple offensive metrics for the Colorado Rockies from 2011-2023, who play at 5183 feet. Due to the fact that the Rockies play half of their games at said elevation, it is expected that they should show correlation from offensive metrics to elevation. The data suggest that elevation has a larger impact on offensive metrics for the Rockies compared to the Yankees. The moderate positive correlation of 0.32 between elevation and runs for the Rockies indicates that the high elevation at Coors Field may contribute to the higher run production. As evident in the graph, a slight positive correlation is visible between runs scored and elevation. However, correlations between HR's, doubles, and OPS ranged from 0.12 to 0.02 showing minimal positive correlation. In contrast, the Yankees' showed a correlation between runs scored and elevation of 0.027. Yankee stadium elevation is 54 feet, where air density has more impact on the ball's trajectory. In summary, the elevation at Coors Field in Colorado appears to only enhance one aspect of offensive production, runs scored. In contrast, Stadiums with elevations within 100 feet of sea level, such asYankee Stadium, have a nullable effect on offensive metrics. As for future implications, if the MLB was looking for an expansion team, which is the process of adding new teams in new cities, they could look for a stadium with elevation similar to Coors Field, which will help offenses score more runs and potentially win more games. It is also something for teams and pitchers to consider when playing at Coors Field.



## Analysis 4: How Did Rule Changes Affect Offensive Production?

Before the 2023 season, the pitch clock was implemented to the MLB rule book and was designed to speed up games, and increase general viewership. Pitchers were given only 20 seconds between every pitch to throw the ball. In addition to the pitch clock, base size increased from 15 inches in width to 18 inches. Induced to increase steals, the MLB hoped that the small increase would actually result in teams attempting to steal more bags. Thankfully, in the 2023 season, attempted stolen bases increased by 28% from 2022. The first chart highlights the significant increase in attempted steals per game following the introduction of the pitch clock. The correlation coefficient between the years 2021-2024 (2 years before and after the implementation) and stolen base attempts per game is 0.963416, indicating a very strong positive relationship. The data shows a clear rise in steal attempts after the pitch clock was introduced in 2023, with predictions suggesting that this trend will continue through 2026. The $r^2$ score of 0.7218 for the predictive regression model, along with a relatively low RMSE (Root Mean Square Error) of 0.1385, shows the model accurately predicts the impact of the pitch clock on stolen bases. Although the rule changes successfully and directly attributed to an increase in stolen base attempts, the same cannot be said for an increase in run production. As depicted in the second chart, the correlation coefficient



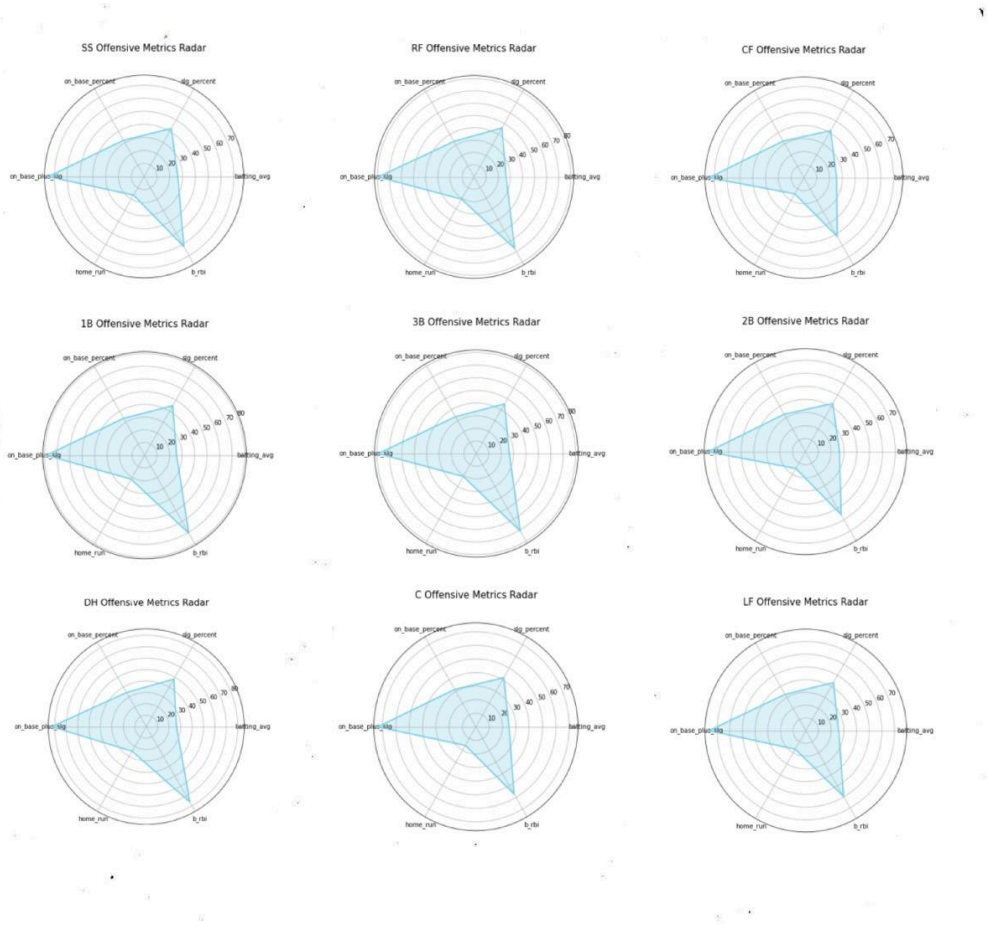



between the years 2021-2024 and runs per game is 0.059956, indicating a super weak positive correlation. Unlike the clear upward trend seen with stolen bases, the number of runs per game does not show a significant increase post-pitch clock implementation. Moving forward, teams need to account for the upward trend of stolen bases by signing faster runners and possibly more defensive catchers to increase or decrease traffic on the bases respectively. As rules are changed in the MLB, teams must analyze trends to interpret how rule changes actually affect team and league wide performances to try to gain an edge.

**Analysis 5: What Position Produce the Most Offensively?**

The central question of this analysis is which position has historically produced the most offensively. This question is crucial not only for baseball analysts and coaches, who aim to optimize team performance, but also for fans interested in the intricacies of the game. Offense is a key component of baseball, and identifying the positions that contribute most can guide decisions in team management, player development, and even fan engagement. We utilized data from Baseball Savant to address this question, focusing on offensive metrics for players during the years 2016, 2019, and 2023. These years were chosen to provide a recent and representative sample, avoiding anomalies from shortened seasons like 2020 or older data that might not reflect the modern game. However, the primary dataset lacked a column identifying player positions, requiring the integration of additional CSV files from Baseball Savant that detailed player positions for these years. This enriched dataset enabled a thorough analysis, incorporating metrics such as Batting Average (BA), Slugging Percentage (SLG), On-Base Percentage (OBP), On-Base Plus Slugging (OPS), Home Runs (HR), and Runs Batted In (RBI). The collected data

was visualized using radar graphs to compare the offensive contributions of different positions, as shown below.




These radar graphs clearly illustrate the average performance across key metrics for each position. For example, the radar graph for Designated Hitters (DH) reveals their superior offensive output, with an average SLG of 48.14%, OPS of 82.72%, and HR of 24.81. These numbers confirm that DHs consistently outperform other positions in terms of power metrics.

Averages for Each Position (Percentage Scale):

| Positions | batting_avg | slg_percent | on_base_percent | on_base_plus_slg | home_run | b_rbi |
|---|---|---|---|---|---|---|
| 1B | 25.592308 | 44.984615 | 33.548718 | 78.533333 | 21.641026 | 70.717949 |
| 2B | 26.589873 | 42.611392 | 33.094937 | 75.706329 | 14.848101 | 55.658228 |
| 3B | 26.556863 | 45.409804 | 33.672549 | 79.082353 | 20.950980 | 70.970588 |
| C | 24.857143 | 41.844286 | 31.814286 | 73.658571 | 15.814286 | 57.214286 |
| CF | 25.495062 | 41.707407 | 32.485185 | 74.192593 | 14.308642 | 52.024691 |
| DH | 26.420000 | 48.144286 | 34.574286 | 82.718571 | 24.814286 | 76.285714 |
| LF | 25.909211 | 43.652632 | 32.884211 | 76.536842 | 16.750000 | 59.618421 |
| RF | 26.567308 | 46.092308 | 33.409615 | 79.501923 | 20.817308 | 67.125000 |
| SS | 26.279798 | 42.621212 | 32.123232 | 74.744444 | 16.434343 | 62.050505 |

Similarly, the RBI average for DHs is the highest among all positions at 76.29, further underscoring their offensive dominance. In contrast, the radar graphs for Catchers (C) and Center Fielders (CF) show lower averages across the board. Catchers have an average BA of 24.86%, SLG of 41.84%, and OPS of 73.66%, along with 15.81 HR and 57.21 RBI. Center Fielders, while slightly better in some categories, still fall behind other positions with a BA of 25.50%, SLG of 41.71%, OPS of 74.19%, 14.31 HR, and 52.02 RBI. These figures indicate that C and CF positions contribute less offensively compared to other positions, which may be due to a greater emphasis on their defensive responsibilities. Other positions, such as Right Fielders (RF) and Third Basemen (3B), also stand out for their offensive metrics. RFs have a high average SLG of 46.09%, OPS of 79.50%, and HR of 20.82, coupled with an RBI of 67.13. Third Basemen, while slightly less powerful than RFs, still maintain strong offensive numbers with an average SLG of 45.41%, OPS of 79.08%, and 20.95 HR, along with an RBI of 70.98. The analysis concludes that teams might gain the most benefit by investing in strong offensive players for the DH position, where power hitting is paramount, as evidenced by their superior metrics. Conversely, teams might prioritize defensive skills over offensive production for positions like C and CF, where offensive contributions are relatively lower. The radar graphs serve as a visual confirmation of these findings, offering a clear comparison of offensive performance across positions, with the specific numbers highlighting the significant differences in contribution.

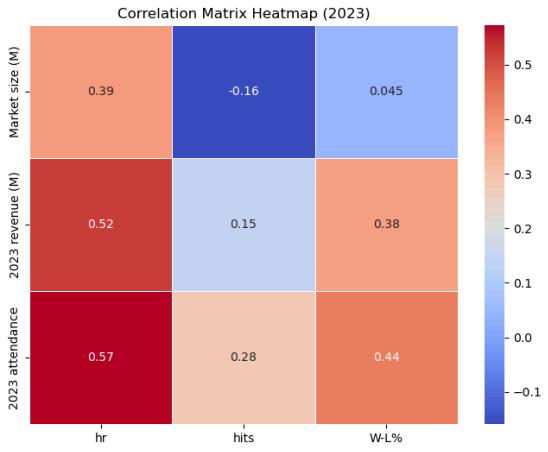
Positions ranked by total rank (lower is better):

| Positions | batting_avg | slg_percent | on_base_percent | on_base_plus_slg | home_run | b_rbi | total_rank |
|---|---|---|---|---|---|---|---|
| DH | 4.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 9.0 |
| 3B | 3.0 | 3.0 | 2.0 | 3.0 | 3.0 | 2.0 | 16.0 |
| RF | 2.0 | 2.0 | 4.0 | 2.0 | 4.0 | 4.0 | 18.0 |
| 1B | 7.0 | 4.0 | 3.0 | 4.0 | 2.0 | 3.0 | 23.0 |
| LF | 6.0 | 5.0 | 6.0 | 5.0 | 5.0 | 6.0 | 33.0 |
| 2B | 1.0 | 7.0 | 5.0 | 6.0 | 8.0 | 8.0 | 35.0 |
| SS | 5.0 | 6.0 | 8.0 | 7.0 | 6.0 | 5.0 | 37.0 |
| C | 9.0 | 8.0 | 9.0 | 9.0 | 7.0 | 7.0 | 49.0 |
| CF | 8.0 | 9.0 | 7.0 | 8.0 | 9.0 | 9.0 | 50.0 |

The best offensive position is DH with a total rank score of 9.

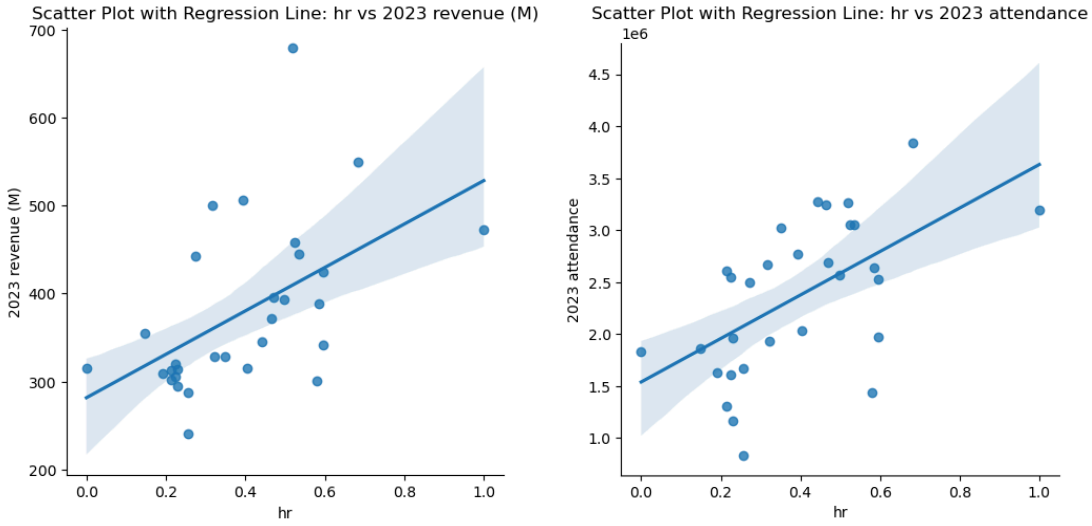
**Analysis 6: Team Performance vs. Team Financial Success**

Analysis 6 sought to explore the relationship between a team's individual performance and their financial success. Key indicators like home runs, hits, and win-loss ratio were measured against their corresponding impact on a team's financial success. Financial success was measured using attendance, revenue, and market size. This correlational study established the team performance statistics as independent variables and the financial metrics as dependent variables. We set the year as 2023 as it posed the most recent financial and team statistics for a complete season. Data collection was carried out through web scraping from reputable sources such as FoxSports and BlessYouBoys, two major contributors to the baseball statistics space. Additionally, we utilized the python library, pybaseball, to retrieve the win-loss ratio for each team. Using pearson's coefficient to normalize the data, we ensured that all teams were standardized across different metrics. We then created a correlation matrix to understand how each independent variable interacted with each dependent variable. This technique enabled us to see exactly which independent variables had the greatest effect on the dependent variables. To further our visual interpretation of the data, we created regressions for each correlation. When fans go to the ballpark, they are most drawn to the excitement of home runs, which is reflected in the data. Home runs demonstrated the highest correlation with all three financial metrics—attendance, revenue, and market size—outperforming other performance indicators like hits and the win-loss ratio. This suggests



that a team's ability to hit home runs is a key driver of fan engagement and financial performance.

**Moderately Strong Correlations**




The analysis further revealed that the win-loss ratio, while also positively correlated with financial success, ranked second in terms of impact. Teams with a better win-loss ratio tend to see improvements in attendance and revenue, though, this effect is not as strong as the influence of home runs. Hits showed the weakest correlation with financial metrics and did not appear to have a notable impact on a team's financial outcomes. Overall, the correlation matrix heatmap supports the conclusion that both home runs and the win-loss ratio are reasonable predictors of a team's financial success. However, home runs stand out as the most influential factor, indicating that teams looking to boost their financial performance should prioritize signing more home run hitters.

## References


- https://baseballsavant.mlb.com/
- https://open-meteo.com/en/docs/historical-weather-api
- https://pypi.org/project/pybaseball/
- https://pypi.org/project/MLB-StatsAPI/
- https://www.foxsports.com/mlb/team-stats?season=2023&category=batting
- https://www.blessyouboys.com/2024/4/19/24134946/the-business-of-baseball-2024-edition
- https://baseballjudgments.tripod.com/id62.html