**Baseball Stats Analysis**

(COMP3125 Individual Project)

Alexis Zabloudil   
*School of Computing and Data Science*

*Abstract*—This project goes in depth about Shohei Ohtani’s swing and batting stats in comparison to other players including Aaron Judge, Giancarlo Stanton, and Oneil Cruz. There is also comparison made with the Yankees and Dodgers, in the 2024 World Series and regular season games. The project also analyzes key stats for different types of pitches.

Keywords—Baseball, Swing, Batting, Sports, Statistics

# Abbreviations and Notes

There are many abbreviations used in the datasets. Here is a list of the abbreviations used in the data and this report.

MLB: Major League Baseball. Professional baseball league in the United States, and has teams in America and Canada **[1]**

PA: Plate Appearances

AB: At Bats

R: Runs scored

H: Total number of hits

2B: Doubles

3B: Triples

HR: Homeruns

RBI: Runs batted in

BB: Base on balls (walks)

SB: Stolen Bases

SO: Strikeouts

CS: Caught Stealing

HBP: Hit-By-Pitch

AVG: Ratio of Hits to ABs. AVG = H/AB

OBP: On-Base Percentage. OBP = (H + BB + HBP) / (AB + BB + HBP + SF)

SF: Sacrifice Flies

SLG: Total Bases per At Bat. SLG = (1B + 2\*2B + 3\*3B + 4\*HR) / AB

1B: Singles

OPS: On-Base Plus Slugging. OPS = OBP + SLG

EV (MPH): Exit Velocity in MPH

LA: Launch Angle

* Ground ball: Lower than 10 degrees
* Line drive: 10-25 degrees
* Fly ball: 25-50 degrees
* Pop up: Higher than 50 degrees
* Sweet spot 8-32 degrees **[7]**

RV/100: Run Value per 100 pitches. Measures how many runs a batter generates (or a pitcher prevents) per 100 pitches thrown or faced.

Run Value: Quantifies the value of an individual pitch or plate appearance in terms of runs created or prevented. High run value means they contribute a lot to a team’s score.

Pitches: total number of pitches received by the batter

BA: Batting Average. BA= Hits/At-Bats

wOBA: Weighted On-Base Average. Like OBP excel singles, doubles, triples, and home runs have different weights.

Whiff%: percentage of swings that result in no contact

K%: strikeout percentage. K% = (Strikeouts / Plate Appearances) \* 100

Put Away %: percentage of two-strike counts that lead to a strikeout

xBA: Expected Batting Average- predicts batting average based on quality of contact (exit velocity, launch angle) and strikeouts

xSLG: Expected Slugging

xwOBA: Expected Weighted On-Base Average- expected value of wOBA based on exit velocity, launch angle, and strikeouts

Hard Hit %: percentage of batted balls hit with an exit velocity of 95 mph or higher

Note:

It is important to note at bats are ***not*** the same as plate appearances. At bats for a player are ***less*** than plate appearances because it doesn't count walks (BB), sacrifice fly (SF), intentional walk (IBB), bit by pitch (HBP), and catcher interference.

# Introduction

Baseball is a precise sport where minor changes in technique can create an entirely new swing, or pitch. Looking deep into these statistics can figure out the key to achieving the best outcome for players and coaches. Shohei Ohtani was one of the highest performing players in the MLB in 2024. The versatility of him and other players are uncovered through comparing each other’s statistics.

# Datasets

## Source of dataset

All the data was copied and pasted from MLB’s website and another baseball statistic website from MLB called Statcast or known as Baseball Savant. Both sources are credible since the statistics are generated straight from MLB. The data was copy and pasted into excel sheets which are easier to use.

## Character of the datasets

The size of the excel file is 620KB and there are 13 sheets in the workbook. There are 3 main types of sheets: swing stats, batting stats, and pitch stats. The swing stats look specifically into a swing, so that looks at exit velocity, launch angle, distance the ball was hit, the speed the pitch came in, and pitch type. The batting stats looked mainly at outcomes of hits, so plate appearances, at bats, runs, hits, doubles, triples, homeruns, runs batted in, walks, strikeouts, batting average, and slugger. The pitch stats included a breakdown of pitch types, strikeout percentage, slugging percentage, and whiff percentage.

# Methodology

## Matplotlib

Matplotlib is a python library used for creating static and interactive visualizations. Graphics like scatter plots, decision boundaries, Isomaps, and Principal Component Analysis (PCA) are created with Matplotlib.

## NumPy

NumPy is another python library, and it is used for computing and handling data. The data is converted into arrays which are easier to use for computations. KMeans clustering uses NumPy, and I used this data visualization to compare swing and batting statistics for Shohei Ohtani, Aaron Judge, Giancarlo Stanton, and Oneil Cruz.

## Scikit-learn

Scikit-learn is a python library used for predicting data. I used it for PCA visualization for 2024 World Series vs Regular Season games Dodgers versus Yankees. The PCA compared swinging stats in World Series games and regular season games to see if players were playing more aggressively when the games meant more. Also, isomaps use scikit-learn and I created isomaps to compare slugger percentage versus whiff percentage by pitch type, and then strikeout percentage versus put away percentage by pitch type.

## Data Cleaning

Data cleaning is an important aspect of data analysis. Removing missing values ensures there is only good and usable data in the datasets. Graphs and models will be more accurate the more consistent the data is. I used different methods including dropna() to remove rows/columns with missing or infinite values. I also aligned the data to make sure rows in the data could be compared. Scaling data is also important to ensure the ElasticNet, PCA, and K-Means models are created correctly.

## Data Training

To predict batting outcomes, I used ElasticNet Regression, which trains data using swing statistics including exit velocity and launch angle. These factors were used to predict batting outcomes like hits, RBI (Runs Batted In), and runs. I also used data training for K-Means clustering to create a visual identifying patterns between the four players of interest in this project. Support Vector Classifier (SVC) **[6]** also requires data training, and I used it to compare the Dodgers versus Yankees on their team stats like Runs, Hits, and RBIs.

# Results

I created 5 questions to be answered using different data visualizations and computations. The questions and their results are discussed below.

## Question 1: How does Shohei Ohtani’s swinging stats reflect his overall game stats?

The code for this question is labeled “#question 1” in the Codes folder.

The code creates an analysis of Shohei Ohtani’s 2024 regular season batting and swing statistics. I wanted to see if swinging metrics could predict batting statistics. Two of the sheets were used from the excel, “Ohtani 2024 Reg S Batting Stats”, and “Ohtani 2024 Reg S Swing Stats”. I used ElasicNetCV Regression to split the statistics into “training” and “testing” and set the parameters cv (cross-validation), random state, and L1 ratio. Cross validation was set to 5 for the number of folds in cross-validation. The first four folds are for training the and the fifth is for validation. The random state number was set to 1 so the trained and tested sets will be the same every time **[2]**.The L1 ratio was set to values [0.1, 0.5, 0.7, 0.9. 1.0] for testing for cross validation. The model is trained to find the best value for L1, and then predictions are made. I calculated Mean Squared Error (MSE) and it is calculated as following where it is y-actual minus y-predicted. I also calculated R-A black and white math equation

Description automatically generatedA mathematical equation with numbers and symbols

Description automatically generatedsquared (coefficient of determination) for the linear regression model. The equations are as shown:

**[4]**

**[3]**

The results are shown in Figure 1 in the Graph folder**.** For Hits, the distance coefficient was almost 0.0 but not negative, proving long distances influences hits. The MSE was high, which shows the predictions were not that accurate. The R-squared value is negative which proves the model did not perform as well as a model that took the average for prediction. For RBIs, all the coefficients were close to or 0.0. The MSE value was lowest for RBIs, but since it is still a high number, it has the same effect as it did with Hits. For Runs, the distance coefficient is positive and a nonzero, like it was for Hits, so this shows distance has a positive effect on Runs. The MSE is high, so the prediction was not that accurate. I was surprised about the negative R-squared values since I thought predicting the values would be a better way to compare metrics, instead of taking averages. The dataset was small, which could be a reason the predictions were not accurate.

## Question 2: How does Shohei Ohtani’s swinging stats compare to other leadrs on the leaderboard like Stanton, Cruz, and Judge?

## The code for this question is labeled “#question 2” in the Codes folder. The code creates an analysis comparing swing and batting statistics for Shohei Ohtani, Aaron Judge, Giancarlo Stanton, and Oneil Cruz. The swing statistics used are Exit Velocity, Launch Amgle, Distance, and Pitch (speed). The batting statistics used are Hits, Runs, and RBIs. The data was cleaned, resized, and combined into a single set in order to make it easier to use K-Means clustering **[5].** I created four different clusters and used code to train the model to randomly set cluster centroids, based on where the clusters land. The graph shows four different clusters 0, 1, 2, and 3. Cluster 0 contains points with very few outliers. This is a consistent cluster, and the high launch angle indicates pop-ups. This explains the second highest number in Hits, RBIs, and Runs. Pop-ups could be homeruns, but with the lowest exit velocity they could just be solid hits, which results in the high number of RBIs. Ohtani outperformed in this category compared to the other players. Cluster 1 has the second lowest exit velocity, but the lowest launch angle. This indicates ground balls. The average distance for these balls to travel was the lowest, so these are not particularly good hits, which explains the lowest Hit, Run, and RBI numbers. Judge, Ohtani, and Stanton all had a similar amount of these data points, but Cruz had the most with about 30 more compared to each other player. Cluster 2 has the highest exit velocity, furthest distance the ball traveled and average Hit, Run, and RBI numbers. This indicates very powerful hits leading to many RBIs, and most likely homeruns. The launch angle is in the “sweet spot,” between 8 and 32, but the exit velocity not being the highest could still mean an exceptionally good hit. Judge dominated this category of hit. Cluster 3 has the highest Hits, Runs, and RBIs and a launch angle right in the “sweet spot” at 18 degrees. A launch angle of 8-32 is perfect, and 18 being right in the middle of that range means these are most likely homerun hits. They have a high exit velocity, the second furthest distance, and they had the highest pitch speed. I did not expect Oneil Cruz to have as many results in Cluster 1, indicating he hit a lot of ground balls. He is a younger player and just starting to become big in the MLB, so his statistics lack compared to the other players.

## Question 3: How do the Dodgers stats compare to the Yankees?

The code for this question is labeled “#question 3” in the Codes folder. The code creates a decision boundary graph to compare the Dodgers and the Yankees by comparing regular season swing statistics. There were seven swinging stats used for this graph, but to keep it visual, it was reduced to 2 dimensions. There is a clear linear decision boundary, showing a difference in the Dodgers and Yankees swinging stats. The Dodgers had more spread out datapoints, meaning very strong and very weak games, with not as much consistency as the Yankees. Since there are many points near the decision boundary, that means these two teams perform similarly. Their similar statistics plays into the reason these were the top 2 teams in the MLB in 2024. There were some unexpected results like some of the Yankee’s games being classified as Dodgers games and vice versa, this could be because of their similar statistics. Hits, at-bats, and doubles contributed most to PCA1, and PCA1 had a variance of 0.84, meaning these swing statistics are the largest difference between the two teams. The PCA2 variance was 0.08, so statistics like triples did not play as big of a role in determining the difference between the two team’s statistics.

## Question 4: What were the swinging stats like during the 2024 compared to the regular season games?

## The code for this question is labeled “#question4” in the Codes folder. A PCA graph of swinging statistics for the World Series versus Regular Season games was created. The PCA loadings were given, and they are similar to the previous stats. This indicates the two teams played similar to their regular season games as they did in World Series games. I expected more hits, walks, and strikeouts in the World Series games. I thought the teams would be more aggressive since it is a higher stakes game. But, it appears there either was not enough data from the World Series games, or the teams did not perform as aggressive as expected.

## Question 5: Is there correlation between pitch type and SLG, Whiff %, K%, and Put Away %?

I got statistics on four different pitches: 4-seam fastball (labeled 4), changeup, sinker, and slider. In figure 5.1, 4-seam fastballs are evenly distributed among the graph, showing not much correlation with whiff percentage and slugger. Changeups have low slugger and slightly high whiff, meaning not many swings are successful, and there is a chance of not getting a hit off this pitch. Sinkers have a low slugger and whiff, meaning it does not generate as many hits, but also not many misses. This could mean bad hits like ground balls are more likely to occur with this pitch. Sliders have the highest overall whiff and lowest slugger, so this is a pitch most likely to cause strikeouts, since there are very few hits and many misses.

In figure 5.2, 4-seam fastballs, sliders, and changeups are all spread evenly with low put away percentage and spread strikeout percentage. This means this pitch is not very effective in closing out an at-bat, compared to a sinker. Sinker has the highest put away, meaning it is most likely used the most in a 2-strike situation, and then ultimately results in a strikeout. I did expect this because it is a tricky pitch and its success rate in strikeouts means it will continue to be used for many pitchers.

# Discussion

In question 1, I discussed wanting to predict Shohei Ohtani’s swinging statistics effecting his game stats. The only to have an influence was distance, and it had a positive influence on hits and runs. The high MSE and negative R-squared values show the prediction for the variables was not clear or accurate. There are many outside factors that could have an effect that the data from MLB does not capture. There could be data not measured that has more influence like data fatigue from players, calls from umpires, or other non-measurable metrics. Future research could include a larger dataset, using other data like pitch type or the inning and number of outs. Baseball is a mentally and physically enduring game and these types of things cannot be measured easily and could have a huge influence on how players perform.

In question 2, the graph showed four different clusters, indicating a type of hit from Shohei Ohtani, Aaron Judge, Giancarlo Stanton, and Oneil Cruz. Cluster 0 had consistent hits and RBIs, while cluster 1 had low metrics in all aspects which would be groundballs. Clusters 2 and 3 had the hardest and furthest hits. These results prove the 8-32 degree launch angle “sweet spot” since the clusters that had this launch angle were the furthest and hardest hits, indicating homeruns. Aaron Judge and Shohei Ohtani dominated in these clusters, proving they are two of the best hitters compared to the other two. This list of four hitters were from another list of the best hitters in the MLB. Future work could include tracking Oneil Cruz over the next few years to see if his hitting improves. He is a young player with a lot of potential, and seeing if his statistics compare to Ohtani’s or Judge’s early days could be a way to predict if he is on track to be as great as those two players.

In question 3, the Dodgers and Yankees regular season statistics were compared to look for correlation. There was some overlap with the two teams, indicating a similarity in skill. This makes sense because the two teams faced each other in the 2024 World Series. The Yankees had more consistency, but the Dodgers had more games with better statistics. The Dodgers inconsistency does not correlate with winning, since they won the World Series, but their ability to have games where they overall play better than the Yankees could be the main reason they won. There is future work for this question and in upcoming seasons, statistics for these two teams could be continued to be observed to see if one outperforms the other.

In question 4, the Dodgers and Yankees 2024 World Series statistics were compared with regular season statistics. These findings imply the Yankees, and the Dodgers played similarly in regular and postseason games. There is an issue where there were much fewer games in the World Series than regular season, which could mean the data is not as accurate because there is not enough to base it on. This could be more work for the future in which if these teams ever faceoff again in the World Series, their statistics for that year could be compared to 2024’s. To see if these teams do play more aggressive when it matters, the statistics for the 7-9th innings could be compared to the first six, to see if higher stakes mean more aggressiveness. This also exposes the issue with the mental part of baseball being in play because how a player feels towards the end of an intense game is not measured, and therefor adrenaline and other things cannot be taken accounted for. Those types of metrics could indicate better or worse performance in a player.

In question 5, different pitch types were compared to see if there was correlation in SLG, Whiff %, K%, and Put Away %. The results showed sliders having a success in strikeouts because of their high whiff percentage and low slugging percentage. Sinkers had a high put away percentage since they are a tricky pitch and using them in a 2-strike situation greatly increases the chance of a strikeout. Changeups are also seen as a tricky pitch since they are much slower than a regular pitch, and the metrics for this pitch proved that right because of the low slugging. Future work could include getting more data points and more pitch types to see which pitches are most effective in different situations. There could be a better pitch than sinkers for put away percentage.

# Conclusion

Overall, these statistics can be incredibly helpful for a coach and a team to see what generates the most hits, which wins games. Pitchers can use the pitch metrics to see which pitches are generating the most strikeouts, since strikeouts play a large role in how well a team performs in a game. The same goes for batters; seeing which pitches batters need most help on to have a lower whiff percentage will help teams have less whiff and more hits. Different coaches specializing in defense and offense can take these statistics and implement practice plans to work on different aspects of the game. I found pitch statistics the most important because there were clear results on what pitches were generating the most hits and strikeouts. This can be used for the batter and the pitcher, so they are important metrics for everyone on a team. Hits were one of the most important metrics for game outcome, and having a clear understanding of what pitches batters need help on can help increase hits, to therefore win more games.

##### References

[1] “MLB | Official Info | About,” MLB.com. https://www.mlb.com/official-information/about-mlb

[2] Shelly, “scikit-learn random state in splitting dataset,” Stack Overflow, Feb. 12, 2017. https://stackoverflow.com/questions/42191717/scikit-learn-random-state-in-splitting-dataset

[3] “Python | Mean Squared Error - GeeksforGeeks,” GeeksforGeeks, Jun. 28, 2019. https://www.geeksforgeeks.org/python-mean-squared-error/

[4] C. Ballard, “An Ode To R-Squared,” Medium, Jul. 08, 2019. https://towardsdatascience.com/an-ode-to-r-squared-804d8d0ed22c

[5] GeeksforGeeks, “K Means Clustering - Introduction - GeeksforGeeks,” GeeksforGeeks, May 30, 2019. https://www.geeksforgeeks.org/k-means-clustering-introduction/

[6] A. Sasidharan, “Support Vector Machine Algorithm,” GeeksforGeeks, Jan. 20, 2021. https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[7] Major League Baseball, “Launch Angle (LA) | Glossary,” MLB.com. https://www.mlb.com/glossary/statcast/launch-angle