Discovering Plate Approach Patterns Among Major League Baseball Batters Through Data Mining

Christopher Harris
College of Computing and Software Engineering
Kennesaw State University
Marietta, Georgia
charr396@students.kennesaw.edu

***Abstract*—This project examines patterns in Major League Baseball (MLB) batting behavior by analyzing plate approach statistics from qualified batters over the 2021-2025 seasons. The analysis uses the Baseball Savant MLB leaderboard, a publicly available dataset, and focuses on discovering natural sets of plate approaches from player-season observations based on the data in the leaderboard. This data measures plate discipline through swing and contact rates and includes observed results and expected results to view their effectiveness. By cross-comparing this data between the identified groups, it will be seen how each one relates to performance and outcome variance, focusing on pure pattern discovery.**

I. Introduction

The purpose of this project is to discover patterns in MLB batter data, revealing the different approaches and strategies players have when at the plate. This data can reveal the underlying discipline a player has, along with their traits. Are they an aggressive or passive hitter? Do they make consistent contact when they swing? These types of metrics determine where each player excels, identifying their skill and style type, no matter how successful they are, which can be useful in evaluating and comparing them on a comprehensive level.

II. Dataset Description

A. Data Source

The dataset to be used is derived from the Baseball Savant MLB leaderboard, and the data will be exported through the platform's built-in CSV download functionality. The dataset displays performance data among qualified MLB batters from the past five seasons (2021-2025), defined as players who have had at least 500 plate appearances in a single season.

B. Data Size

In the selected dataset, there are 669 rows, each representing a player-season observation, and 11 columns representing plate-approach measurements and outcome statistics.

C. Data Representation

The dataset describes individual plate approach behavior and tendencies as well as outcome quality for an overall season. Plate-approach statistics track how often a batter swings at a pitch, where they swing, and how frequently they make contact with the ball. These measurements relate to result statistics, summarized by weighted on-base average (wOBA), signifying the observed outcomes of each plate approach. Expected statistics are included to measure underlying performance based on the quality of contact made, rather than observed results, helping to exclude outcome variance due to luck. The data only accounts for a player's overall performance across an entire season, not a game-by-game analysis.

D. Key Measurements

Observation Data

- Player
- Season

Plate Approach Data

- K% - Strikeout Rate
- BB% - Walk Rate
- Z-Swing% - In Zone Swing Rate
- O-Swing% - Out of Zone Swing Rate
- Z-Contact% - In Zone Contact Rate
- O-Contact% - Out of Zone Contact Rate
- Swing% - Swing Rate

Results Data

- wOBA - Weighted On Base Average
- xwOBA - Expected Weighted On Base Average

E. Sample Data Table

A miniature sample data table is given below in Table 1 to provide a visualization of how the data is formatted. This sample shows a condensed form of the dataset across three observations and six attributes.

F. Data Quality Considerations

There are limited data quality issues in the dataset that will have to be addressed. The dataset is restricted to qualified batters, excluding part-time or bench players, which favors the analysis toward established, full-time players. Additionally,

TABLE I
Sample Data

Player	Year	K%	BB%	wOBA	xwOBA
Tucker, Kyle	2021	15.9	9.3	.383	.394
Olson, Matt	2025	24.3	12.6	.366	.360
Swanson, Dansby	2024	24.3	9.1	.307	.324

because all data is taken at a season level, any in-season variation of individual plate approach and performance changes is masked. The dataset also does not account for the opposing pitcher, varying in-game situational plate appearances, or stadium/field differences, which may influence plate approach and outcome statistics.

III. Discovery Questions

A. Questions

1) Are there distinct, natural groupings of MLB batters based on their plate approach statistics? : This is the key question that will be answered: finding patterns arising from how batters approach the plate and which groupings form from them.

2) How do observed result statistics (wOBA) and expected result statistics (xwOBA) differ across these groupings? : This question examines how the different, discovered plate approach profiles are related to offensive results, both in what is observed and what is expected based strictly on quality of contact.

3) Are there outlier player performances that significantly deviate from what is expected from their plate approach among the groups? : This question will help identify unusual observations that feature offensive performances that do not align with the typical results seen from their respective plate approach grouping.

B. Value

Overall, these discovery questions are valuable because they reveal broad plate approach profiles among MLB batters. Furthermore, these profiles can be related to performance outcomes and variance when analyzing observed and expected outcome metrics. Finding outliers enhances the analysis by uncovering pattern exceptions, which offer insight into which plate approaches tend to result in overperformance or underperformance when compared to expectations.

IV. Planned Data Mining Techniques

A. Clustering

Both K-Means Clustering and Hierarchical Clustering will be employed to determine whether distinct, natural groupings of MLB batters exist based on their plate approach statistics. Clustering will directly address whether common plate approach profiles arise from the data by grouping player seasons using swing rates, contact rates, and discipline measures. Utilizing different clustering methods will verify the validity of the discovered groups and ensure that the found patterns are concrete.

B. Dimensionality Reduction

Dimensionality reduction through Principal Component Analysis (PCA) will be used to minimize the dataset to a few key features that display the dataset in a lower dimension. This will help to narrow down which combinations of variables showcase variance and patterns in batter behavior most clearly, supporting the clustering techniques.

C. Anomaly Detection

Anomaly detection will be performed mainly through distance-based techniques to identify outliers in player-seasons that greatly overperform or underperform their expected outcomes when compared to other batters within their plate approach profile grouping. Examining differences between wOBA and xwOBA statistics within clusters will directly locate this unusual behavior, indicating extreme skill or variance in a batter, or just an unlikely occurrence of a plate approach profile deviating from typical patterns.

V. Planned Analysis Pipeline

Fig. 1 depicts the planned data analysis pipeline of the project. Data collection and cleaning will consist of preprocessing and finalizing the dataset. Feature scaling will ensure that the data is transformed in a way that makes it easy to analyze and understand. Finally, the data analysis and mining stage will apply the aforementioned techniques to the data to highlight its patterns and characteristics.

VI. Preliminary Timeline

The next steps of the project can be broken up into two halves: Finalizing and preparing the dataset for data mining, and the application of data mining techniques itself.

A. Data Preparation

During this phase, the Baseball Savant MLB leaderboard data will be collected, exported, and preprocessed. Preprocessing will include cleaning the data and eliminating data quality issues, which will then allow for conducting exploratory data analysis on the dataset. This will be accomplished by

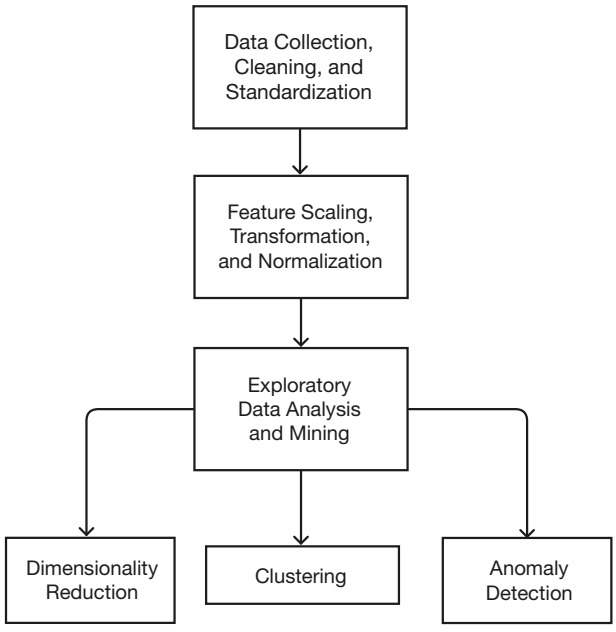

Fig. 1. Pipeline Flowchart

normalizing and transforming the data to better understand it, which will make the following univariate, bivariate, and multivariate analyses possible. Correlations, distributions, and outliers will also be found during this time. By the end of this period, the data will be ready for data mining techniques applications.

B. Data Mining

This stage will handle the application of data mining techniques to the dataset and the ensuing interpretations and evaluations of it. As mentioned previously, clustering, dimensional reduction, and anomaly detection will be the main strategies utilized to find groups, patterns, and outliers. After the data mining, the results will be examined for understanding and meaning. Any limitations will also be addressed before final results are drawn.

C. Final Deliverable

The final deliverable will include the final project report, featuring the findings and visualizations of the data. Additionally, the eProfile containing the project files will be organized and completed.

VII. Anticipated Challenges

There are a few anticipated project challenges involving the data. Firstly, it is crucial to select the best scaling and distribution methods for the dataset, which could be made difficult due to any correlated variables that exist, so the handling of this process must be chosen accurately. Additionally, ensuring cluster distinctness and distinguishing between meaningful anomalies detected and natural data variation will also be difficult, as much of the data is similar and in a tight range. The process of deciding cut-off points for groupings and outliers may pose difficulties due to this proximity and the nuances of each data type.

VIII. Conclusion

The proposal discussed in this report aims to find useful patterns in individual MLB batter data that is important to current players, rather than predict how these players could evolve and change in the future. Batter approach and style matters greatly in the modern age of baseball, and this project will use collected data in a way to make it interpretable through data preparation, mining, and its following evaluation. The resulting information that this project will provide can be put into use across MLB, such as fans competing in fantasy baseball, teams deciding which players best suit their team, coaches putting together a batting order, or anything similar.

References

[1] Major League Baseball (MLB), "Baseball Savant," https://baseballsavant.mlb.com/, accessed Jan. 2026.

[2] M. J. Zaki and W. Meira Jr., *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge, UK: Cambridge University Press, 2020.