Analyzing MLB Hitters with Spark and Statcast

***Abstract*— This report showcases a brief Spark-based analysis of Major League Baseball Statcast data from the 2022 to 2024 seasons. The overall goal of this paper is to analyze hitter performance using descriptive statistics and machine learning techniques. Statcast (MLB's tracking system) captures detailed measurements for every pitch and batted ball, including variables such as exit velocity, launch angle, and the distance the ball was hit. These metrics provide a better view of player performance than traditional statistics alone.**

I. Introduction and Motivation

Baseball is a sport built on numbers. Every pitch, every swing, and every contact with the ball generates data that can be measured, analyzed, and used to evaluate player performance. Over the past decade, Major League Baseball has vastly changed the way data is collected through a system known as Statcast which is an advanced tracking system that records information about every pitch and batted ball using high-speed cameras and radar technology. As an avid baseball fan (let's go Mets) and computer science student, I was drawn to the idea of using real Statcast data to explore what makes hitters different from one another and whether we can use machine learning to discover meaningful player profiles based on how they hit the ball.

The overall goal is to analyze Statcast data by using Apache Spark to perform descriptive and technical analyses of hitter performance. While traditional statistics like home runs and batting average are commonly used by fans to compare players, Statcast offers deeper metrics like exit velocity, launch angle, and hit distance that can truly show how players generate value at the plate beyond just looking at the box score numbers.

This report starts by introducing the dataset and explaining why Statcast is a good fit for analyzing baseball stats at a large scale. From there, I look at the trends in the data, like which hitters have the highest average exit velocity and which pitch types are used most often. After the descriptive part, I use a clustering algorithm from Spark MLlib to group hitters based on how they tend to hit the ball. The goal is to show that big data tools can help in finding patterns in player performance and that combining computer science with baseball analytics can lead to discoveries we might not see just by watching the game.

II. Presenting the Data

The data for this project comes from Statcast, Major League Baseball's pitch tracking system, and was collected using the open-source Python library pybaseball, which pulls raw data directly from the Baseball Savant platform. Baseball Savant is a website that utilizes data collected from Statcast and provides detailed statistics for every player and helps visualize it as well. I then downloaded pitch-by-pitch Statcast data from the 2022, 2023, and 2024 seasons. Each row in the dataset represents an individual pitch and contains dozens of variables, including pitch type, release speed, launch angle, exit velocity, and batted ball outcomes. The full dataset includes over 2.1 million records and more than 90 columns.

What makes Statcast so valuable is the level of detail it provides. Instead of just knowing whether a player got a hit, it is possible to measure exactly how hard the ball was hit, what angle it left the bat at, and how far it was expected to travel. These metrics allow for a deeper understanding of player performance, especially when combined with large-scale computing tools like Apache Spark.

In this project, I used a CSV file which I named "statcast_2022_2024.csv"which contained all of the raw pitch data for the three-year period. I used Spark to load and process the data across millions of rows, allowing me to perform both the descriptive summary and more advanced machine learning techniques on a dataset that would be too large for tools like Excel or Pandas.

III. Related Work

Statcast data has progressively been developed over the years and has become a major part of modern baseball analysis, and many researchers and analysts have used it to uncover trends in player performance. Analysts at Fangraphs, Baseball Savant, and The Athletic regularly use Statcast metrics like exit velocity, launch angle, and expected wOBA (xwOBA) to evaluate both hitters and pitchers. You'll even see many self-proclaimed analysts using Statcast statistics all over social media. These stats are usually considered more predictive of future success than traditional measures like batting average or RBIs.

From an academic perspective, many papers have explored how machine learning can be applied to baseball data. In a 2020 study published in the Journal of Sports Analytics, researchers used clustering techniques to identify pitch usage patterns among MLB pitchers. Another study from MIT's Sloan Sports Analytics Conference used Statcast data to build predictive models for home run probability based on launch angle and exit velocity.

This project takes inspiration from those approaches but focuses more on grouping hitters by their hitting profiles.

By using similar techniques to baseball hitting data, I attempted to build on existing research while also trying to explore what can be done with big data tools and large sports datasets.

IV. Descriptive Analysis

Before moving into the machine learning portion of the project, I started with a descriptive analysis of the Statcast data to get a better understanding of overall trends and patterns. Using Spark, I analyzed pitch usage, batted ball outcomes, and hitter tendencies over the three-season span.

One of the first things I looked at was pitch type distribution. Fastballs, specifically four-seam fastballs, were the most commonly thrown pitch in the dataset, making up roughly a third of all pitches. Sliders, sinkers, and changeups followed in popularity. This matches what we see across the league, where fastballs are still the foundation of most pitchers' arsenals, but breaking and off-speed pitches are becoming more common.

TABLE I. PITCH TYPE FREQUENCY

Pitch Type	Count
FF (Four-seam Fastball)	687,060
SL (Slider)	342,236
SI (Sinker)	329,214
CH (Changeup)	227,934
FC (Cutter)	166,039
CU (Curveball)	143,574
ST (Sweeper)	115,196
FS (Splitter)	49,192
KC (Knuckle Curve)	43,557
SV (Slurve)	8,983

Table 1 shows the ten most frequently thrown pitch types over the three-season span. As expected, four-seam fastballs lead the way by a wide margin. Sliders and sinkers also show up frequently, confirming that breaking and off-speed pitches are now a regular part of a pitcher's repertoire. These trends match what many analysts have noted over the years, which is that MLB pitchers are moving away from just throwing hard and instead rely more on movement and confusing the batter.

Next, I looked at batted ball outcomes. I used the events column to count the most frequent results of at-bats. Unsurprisingly, strikeouts and singles made up a large portion of outcomes, with home runs, walks, and groundouts also appearing frequently. This gave me a pretty basic sense of how hitters were performing overall and what kinds of contact were most common.

One of the more interesting metrics I looked at was average exit velocity. I used Spark to calculate each batter's average exit velocity across all tracked batted balls, then created a leaderboard to see who hit the ball the hardest. To avoid small sample noise, I filtered the results so that only batters with enough batted balls to deem them as "qualified" hitters would be considered. This is also common practice in the league when determining players who are leading in certain stats. The results were exactly what I expected. Players like Aaron Judge and Shohei Ohtani, two of the greatest players of the modern era of baseball, were at the very top of the leaderboard in average exit velocity.

I also filtered for fastballs and found the top 10 pitchers with the highest average four-seam fastball velocity. These were mostly hard-throwing relievers like Mason Miller and Jhoan Duran and high-velocity starters, which also lined up with what I expected.

Overall, this portion of the project helped confirm that the data was clean and reflected what I already knew about the current era of baseball. It also gave me a better foundation for the clustering analysis that followed.

V. TECHNICAL ANALYSIS

After completing the descriptive analysis, I moved on to the technical analysis by applying machine learning to the data using Apache Spark's MLlib. The goal was to group hitters into distinct clusters based on the way they hit the ball. I chose to use K-Means clustering.

To start, I filtered the dataset down to include only rows where launch speed, launch angle, and estimated hit distance were available. These variables describe the quality and shape of contact when the batter puts the ball in play. I then grouped the data by hitter and computed the average of each stat for every player. I also filtered out hitters with fewer than 50 batted balls to avoid small-sample noise that could potentially skew the results.

Once I had the filtered data, I used Spark's VectorAssembler to combine the features into a single feature vector and applied StandardScaler to normalize the data. This step was important because the four metrics are measured on different scales, and K-Means is sensitive to that. With the features prepared, I applied K-Means clustering with $k = 4$ to group hitters into four distinct clusters.

The results revealed some pretty clear differences between hitter types. Some clusters represented hitters with high average exit velocity and long hit distance, while others included hitters with more average metrics or flatter launch angles. These groupings could be thought of as categories like "power hitters," "contact hitters," or "weak contact hitters." While the cluster labels themselves are arbitrary numbers, examining the players in each group showed noticeable differences in their hitting profiles.

TABLE III. SAMPLE OF CLUSTER PROFILES

Name	Avg. Exit Velo	Avg. Launch Angle	Avg. Distance	Batted Balls	Cluster
Pete Alonso	89.7	17.0	179.0	1,327	0
Brandon Nimmo	90.6	9.6	160.6	1,322	1
Javier Baez	88.1	9.6	146.6	985	1
Geraldo Perdomo	82.6	12.1	147.1	960	3
Albert Pujols	91.2	16.7	176.3	253	0

The clusters generated by K-Means aligned closely with known player profiles. For example, Pete Alonso and Albert Pujols were placed in the same cluster due to their strong exit velocities and optimal launch angles, which are traits commonly found among true power hitters. On the other hand, Brandon Nimmo and Javier Báez ended up in a different cluster, likely representing hitters who make hard contact but with lower launch angles, resulting in more line drives than homers. Players like Geraldo Perdomo, who produce lower exit velocities and shallower contact, fell into a separate cluster, likely capturing weaker-contact or utility-type hitters.

VI. CONCLUSION

This project gave me the chance to combine two things I care deeply about: baseball and data analysis. By using Apache Spark to work with over two million Statcast records from the 2022 through 2024 MLB seasons, I was able to explore real-world player tendencies and apply machine learning to group hitters based on how they make contact.

The descriptive analysis helped me understand the broader trends in the game, such as which pitch types are most common and which hitters produce the highest exit velocities. The technical analysis went a step further by identifying player clusters using K-Means, revealing meaningful patterns that could be used for scouting, player comparison, or just better understanding the different styles of offense in baseball.

More than anything, this project showed me the value of combining sports data with big data tools. Using Spark allowed me to process massive amounts of information quickly and gave me a way to apply what I've learned in computer science to something that feels real and interesting. If I were to continue this work, I'd look into expanding the analysis to include pitchers, or even building a recommendation system that compares unknown players to established stars using Statcast metrics.