Pattern Discovery in MLB Aging Curves: Evidence of Performance Deterioration After Age 33

Tate York
CS 4412: Data Mining
Kennesaw State University
Email: tyork12@students.kennesaw.edu

***Abstract*—This project proposes a data mining study of Major League Baseball (MLB) player performance as it relates to age, with an emphasis on discovering patterns that support (or challenge) the commonly stated claim that players deteriorate after age 33. Using statistics sourced from Baseball-Reference, the project focuses on pattern discovery rather than prediction: identifying which performance metrics decline, whether distinct aging-trajectory groups exist, and which players behave as outliers. Planned techniques include clustering and anomaly detection, with dimensionality reduction used primarily for visualization and interpretation.**

***Index Terms*—data mining, pattern discovery, clustering, anomaly detection, PCA, sports analytics, aging curves**

I. Dataset Description

A. Name and Source

The dataset will be collected from Baseball-Reference (https://www.baseball-reference.com/), a public baseball statistics site that provides historical MLB player-season statistics and league tables.

B. What the Data Represents

The primary unit of analysis will be a *player-season* record: one row per player per MLB season. Each row includes the player's age that season and performance attributes. This structure supports discovery of trends and patterns in performance as players age, including comparison of pre-33 vs. post-33 outcomes.

C. Planned Data Acquisition (Scraping Approach)

Because Baseball-Reference is primarily a website rather than a packaged dataset, the project will collect structured data by scraping MLB league-season "standard" tables (e.g., standard batting and standard pitching pages). This approach minimizes the number of requests compared to scraping individual player pages and produces consistent tables that already include age and key performance metrics.

D. Expected Size and Features

The exact dataset size depends on the season range selected. A reasonable initial scope is 1980–2025, which is expected to produce tens of thousands of player-season rows after filtering out extremely small-sample seasons. Key features planned for analysis include:

- **Age**, the core attribute for the deterioration hypothesis
- **Playing time** (PA for hitters and/or IP for pitchers) to support minimum-sample thresholds
- **Overall value metrics** such as WAR
- **Rate statistics** such as OPS (hitters) and ERA/FIP/WHIP (pitchers)

E. Known Data Quality Issues

Anticipated issues include missing values for some metrics depending on era, players with partial seasons, and survivorship bias, as players who remain in MLB past age 33 may be unusually productive. These issues will be addressed through consistent filtering rules, careful documentation, and explicit discussion of limitations.

F. Sample Schema

Table I shows a simplified schema for the planned batting dataset.

TABLE I
Planned Batting Dataset Schema (Simplified)

Field	Description
player_id	Baseball-Reference identifier (or derived unique ID)
name	Player name
season	MLB season year
age	Player age in that season
team	Team identifier (if included in table)
PA	Plate appearances (filtering / stability)
WAR	Wins Above Replacement (overall value)
OPS, OBP, SLG	Offensive rate statistics
HR, R, RBI	Supporting counting statistics

II. Discovery Questions

This proposal emphasizes **pattern discovery** rather than prediction. The objective is to discover interpretable structures related to aging, not to forecast future performance.

A. Q1: Which performance metrics show the clearest deterioration after age 33?

This question examines how commonly used baseball metrics (e.g., WAR and OPS for hitters; WAR and ERA/FIP for pitchers) shift after age 33 in aggregate. The goal is to identify which measures decline most consistently and quantify how strong those declines appear across the player population. Age 33 is selected as a focal threshold because it

commonly coincides with late-career contract negotiations in Major League Baseball, making it a meaningful point at which teams and players implicitly assess expected future decline. Additionally, in order to retain an unbiased and statistical approach, rewards will not be factored into the production of the player, as some awards have been handed out according to the "legacy" of the player, not the performance.

B. Q2: Are there natural groups of players with different aging trajectories?

Players do not all age the same way. This question seeks to discover whether players cluster into recognizable aging patterns, such as early peak, late peak, gradual decline, or relative stability. The output will consist of cluster descriptions and representative trajectory visualizations.

C. Q3: Which players are outliers who resist decline, and what characterizes them?

Some players remain productive beyond age 33. This question focuses on identifying outliers and describing what makes their aging curves unusual, such as stable WAR, minimal rate-stat decline, or role changes that mitigate performance loss.

III. Planned Techniques

At minimum, the project will use techniques from at least two categories. The current plan includes the following:

A. Clustering

Clustering methods (K-Means, hierarchical clustering, and/or DBSCAN) will be used to discover groups of similar aging trajectories. Feature representations may include age-window vectors (e.g., ages 27–37) or summary statistics such as pre-33 vs. post-33 means, slopes, and variability.

B. Anomaly Detection

Anomaly detection methods (e.g., Local Outlier Factor or density-based approaches) will identify players whose post-33 performance deviates strongly from typical aging patterns. The emphasis will be on interpretability and comparison to discovered clusters.

C. Dimensionality Reduction (PCA)

PCA will be used to reduce correlated performance features into a smaller number of components for visualization and to help identify dominant axes of variation in aging-related performance changes (I'll have to figure out my 'new metric' at some point, to give myself a clear marker for visualizing just how good the player was).

IV. Planned Analysis Pipeline

Fig. 1 illustrates the planned workflow, which will be refined as the project progresses.

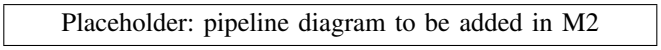

Fig. 1. Planned analysis pipeline (placeholder).

Planned steps include:

1) Scrape league-season standard tables from Baseball-Reference.
2) Cache raw HTML locally and parse tables into structured CSV files.
3) Clean data and apply minimum-sample filters (PA/IP thresholds).
4) Engineer aging features (pre/post-33 summaries, slopes, variability).
5) Apply PCA and clustering to discover structural groupings.
6) Apply anomaly detection to identify unusual aging curves.
7) Summarize discovered patterns using tables and visualizations.

V. Preliminary Timeline

A. M2: Data Collection and Preparation

Implement scraping with caching and rate limiting, parse tables into structured datasets, and define consistent cleaning and filtering rules.

B. M3: Pattern Discovery

Perform exploratory analysis, engineer aging-trajectory features, and apply PCA and clustering to identify major patterns.

C. M4: Outliers and Final Report

Run anomaly detection, interpret key outliers, finalize visualizations, and complete the final report emphasizing discovered structures and limitations.

D. Anticipated Challenges

Primary challenges include survivorship bias, incomplete metrics in earlier seasons, and differences between hitters and pitchers. These issues will be addressed through subgroup analysis and careful interpretation.

VI. Repository Plan

The GitHub repository contains:

- `README.md` with project overview and dataset source
- `docs/` containing the LaTeX-generated proposal PDF
- `data/` for cached HTML and cleaned outputs

VII. Conclusion

This proposal outlines a discovery-driven data mining project using Baseball-Reference statistics to examine aging-related performance deterioration after age 33. By combining clustering, dimensionality reduction, and anomaly detection, the project aims to uncover interpretable structures that explain how decline occurs, whether multiple aging patterns exist, and which players resist typical deterioration.

References

[1] Baseball-Reference, "Baseball-Reference.com," https://www.baseball-reference.com/, accessed 2026-02-05.