# Strike Zone Synergies: The Science of Pitching Clustering

## 2024 Cincinnati Reds Hackathon

Ben Scartz, Brian Papiernik, Jack Arbuckle

## Introduction

We the researchers, Ben Scartz, Brian Papiernik, and Jack Arbuckle, thank you for the opportunity to participate in the Reds' 2024 Hackathon. We appreciate the open-endedness of the assigned task and the opportunity to be creative. Our project focuses on the part of the prompt that searches for pitchers who are candidates to switch roles, and we are proud of the process that we developed. We wanted to pull from our collective baseball experience, including Jack's experience as a collegiate letterman at Missouri S&T, Brian's experience as a student manager at Notre Dame, and Ben's experience as a player development intern with the Florence Y'alls of the Frontier League. However, we were more eager to apply the concepts that we have been learning together this year as analytics graduate students at Notre Dame. We believe that our experience and passion for the game is sprinkled throughout this project. Please enjoy this report and the accompanying R Markdown file.

## Abstract

This project creates a program that identifies pitchers, who are likely to perform well in roles other than those in which they have been used. It applies the machine learning algorithm, K-means clustering, iteratively over a variety of sets of variables. This process sorts pitchers into groups based on their similarities in the respective variables. The program then selects clusters in which usage-outliers (i.e. pitchers used in roles other than their cluster peers), present opportunities for value in other roles. To present a simplified example, a pitcher may be classified as a long reliever, who experiences minimal success but due to similarities in variables such as arsenal and usage characteristics, the program clusters him with several short relievers, all of whom experience success. The program would identify this pitcher as a candidate to switch

from long to short relief. The strength of this program is in its ability to apply a wide variety of criteria to compare and sort pitchers. Some clustering criteria appear obvious, but others are rooted in multi-dimensional similarities. Across the board, we find this method to be very effective in identifying usage-outliers.

## Organization

This project followed a clear workflow through the following four-step process.

1. Define Roles - through discussion, we outline the different types of roles commonly used by MLB pitching staffs.

2. Classify Pitchers by Roles - under each division of roles, we bucket pitchers based on relevant statistics.

3. Outline Relevant Characteristics - through various sources, we compile statistics and pitch metrics that could conceivably differentiate between roles and dictate success.

4. Identify Usage Outliers and Alternative Matches - we run various clustering models on the relevant characteristics and identify clusters of interest

## Step 1: Define Roles

We divide usage into four categories: Innings, Length, Leverage, and Matchups. Below, we outline the criteria by which we classify pitchers.

Innings:

Opener / Starter: >75% of appearances begin in the 1st inning

Middle: >75% of appearances begin in the 2nd-6th innings

Late: >75% of appearances begin in the 7th inning or later

Length:

Short: >60% of appearances are between 1 and 25 pitches

Medium: >60% of appearances are between 26 and 40 pitches

Long: >60% of appearances are greater than 40 pitches

* Pitchers who were previously classified as starters receive an NA for this category, and those who do not fit any of the length criteria (varied usage) are classified as "none".

Leverage: This category uses Fangraphs' gmLI statistics, which judges the average leverage index in which a player enters the game.

High: gmLI > 1.1704 (population Q3)

None: 1.1704 < gmLI < 0.8788 (population Q1)

Low: gmLI < 0.8788

Matchups:

RHH: >66% of batters-faced are right-handed

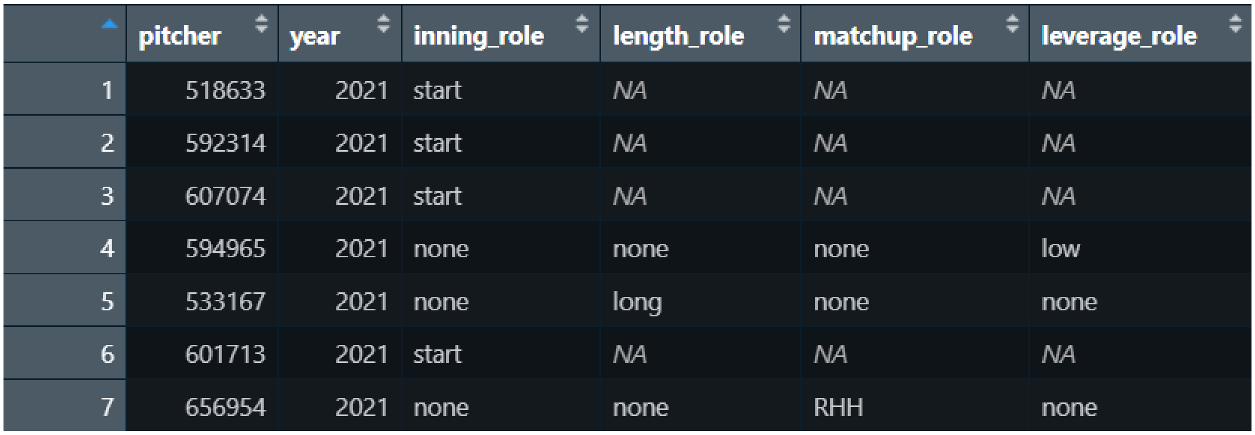
LHH: >50% of batters-faced are left-handed

* Pitchers who do not fit into either of these categories are classified as "none"

** The discrepancy in percentages-used accounts for the relative scarcity of left-handed hitters

**Step 2: Classify Pitchers by Roles:**

Classifying pitchers into these roles requires the use of dplyr mutations and summarizations, bucketing each of a pitcher’s appearances based on the four categories. When combined, the pitcher_roles table appears as below.



| | pitcher | year | inning_role | length_role | matchup_role | leverage_role |
|---|---|---|---|---|---|---|
| 1 | 518633 | 2021 | start | NA | NA | NA |
| 2 | 592314 | 2021 | start | NA | NA | NA |
| 3 | 607074 | 2021 | start | NA | NA | NA |
| 4 | 594965 | 2021 | none | none | none | low |
| 5 | 533167 | 2021 | none | long | none | none |
| 6 | 601713 | 2021 | start | NA | NA | NA |
| 7 | 656954 | 2021 | none | none | RHH | none |

**Step 3: Outline Relevant Characteristics**

Looking forward to step four, we cast a wide net of performance statistics, underlying metrics, and unique attributes. The strength of the program which orchestrates our learning model is in its ability to handle a multitude of variables. In this step, we target any variables that could conceivably impact the way in which a pitcher is used or performs in his respective role. A full list of variables is included in the accompanying R Markdown file, but many of the common metrics are excluded from this report for brevity. Some are explained below.

xwOBA -against: We use this metric, calculated from the baseball savant pitch-level data, as our primary, overarching performance metric. We believe that it offers the truest representation of a pitcher’s value to preventing runs. We also divide this into vs. RHH and vs. LHH.

Weighted run value per c (by pitch): We use this metric from SIS as our primary assignment of individual pitch performance.

Number of Plus Stuff Pitches: This calculates the number of pitches that a pitcher throws with a Stuff+ rating above league average.

xwOBA success based on trips through the batting order: This evaluates a pitcher’s relative success in second and third times through a batting order.

Pitch usage based on trips through the batting order: How does a pitcher change his pitch usage throughout a game? Essentially, does he have the ability to open up his arsenal in successive trips through a batting order?

Velocity Endurance: How well does a pitcher maintain his fastball velocity throughout a game? We evaluate how long it takes for a pitcher’s rolling average velocity to drop more than 1.5 mph below his initial velocity.

Through experience and discussion, we believe the latter four statistics to be indicative of strong starting pitching or long relief. We proceed to apply these and the other many statistics to our learning model.

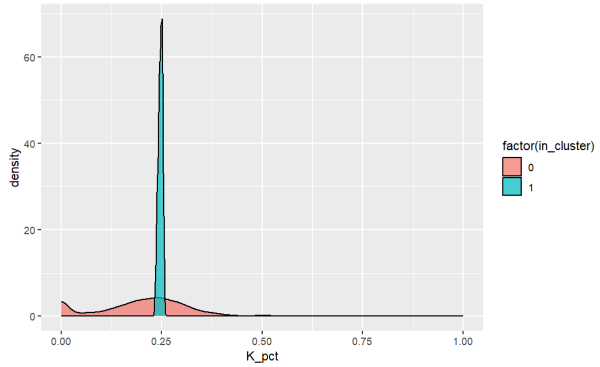
**Step 4: Identify Usage-Outliers and Alternative-Matches**

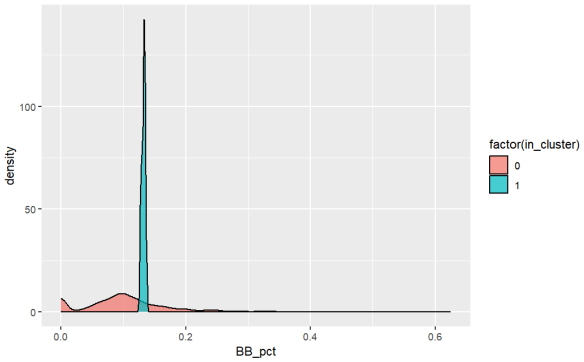
This step executes the heart of our analysis: a multi-layered K-means clustering model. K-means clustering is an unsupervised machine learning algorithm that groups observations into a specified number of clusters. It defines k centroids as prototypical members of each cluster, and observations are sorted by least-squares differences. We find this to be an effective method for grouping pitchers based on any number of dimensions, clustering those who provide comparable value across the dimensions as a whole.

Using the multitude of variables that we collect in step 3, we group the variables into many vectors. Each vector is to be applied as the set of variables for a certain iteration of the clustering execution. We use baseball knowledge and intuition to brainstorm combinations of variables that could lead to useful clusters. We include a few examples below, including our naming liberties.

Fastball Domination: fb_shape (X,Y, velocity), fb_usage, fb_weighted_runs, Ks_per_out
Being Used!: fb_usage, sl_usage, ch_usage, cb_usage
Sony-x-Walkman: xwoba, bb_pct

We add each of the many combinations of variables to a “list_of_lists”, and the clustering model is looped over each list item. The clusters show statistical similarities across one or several dimensions. Single-dimensional similarities can be demonstrated by graphs like those below, where the blue densities show the narrow distribution of pitchers within the cluster, and the red densities show the wide distribution of pitchers without.





Multi-dimensional cluster similarities are impossible to visualize, but they nonetheless capture comparable value across the several variables.

The program observes the clustering model as it iterates, and it extracts clusters of interest. In order to identify usage-outliers, the program pulls out clusters with a clear categorical outlier (i.e. nine “low-leverage” pitchers and one “high-leverage”). Because of their close similarities in the particular variables, the outlier may be understood to fit better in the cluster’s majority category. Across approximately thirty individual clustering models, each with between 100 and 300 clusters, the program flags 369 clusters of interest.

To narrow the list to those with potentially-actionable insights, the program next evaluates the clusters of interest based on xwOBA performance. It searches for those where the majority class (the nine low leverage pitchers in the example above) performs well. This would support any theory that the minority (the one high-leverage pitcher) would succeed if used in the other role.

**Summary**

The strength of this program is that it automatically flags 38 pitchers, for which solid cases could be made for role-switches.

Our project highlights three role switches. First, we recommend that Michael Kopech would provide more value in a concentrated short, high leverage role. Second, we recommend that Ryan Weathers should not be used in a long relief role, but rather in shorter appearances. Finally, we recommend that Tommy Milone should be used in a short relief role. Further explanations are included in the accompanying R Markdown file.

The project highlights those who may have lacked production in their current roles but could provide unseen value in a different role. The ability to identify this value in pitchers has the opportunity to provide competitive advantage for any front office.