Emily Hapgood, Michael Sepe, Luke Mutz, Vlad Hes

Data Preprocessing

The main goal of preprocessing was to preserve the most amount of data while driving the number of NaN cells to 0 either by removing subjectively judged irrelevant information (such as redundant pitch mechanics measurements), or by populating the cells with accurately calculated data.

We extracted and averaged metrics for the two primary pitches per pitcher each season from Savant, excluding rows with only one pitch type or <30 pitches to avoid skewing data with short seasons. We opted against including a third pitch to avoid generating artificial data, despite recognizing its significance for some pitchers. From Fangraphs, we removed pitch-specific columns (already covered by Savant) and rows with <30 pitches, deeming them non-representative. NaNs in Stuff+, Location+, and Pitching+ were set to 100, the true baseline defined by those statistics. We then merged Fangraphs and Savant data using MLBID + Season as identification from both datasets, resulting in a final dataset with 131 columns, further separated into SP and RP for analyzing success factors.

Building the metrics of success for Starters and Relievers

We first decided on just two pitcher categories: Starters and Relievers. Although roles are starting to blend together in an evolving pitching landscape to include many categories such as openers and specialty matchup pitchers, we decided that the most distinct roles are still Starter and Reliever. After discussion among the group, we decided to include WHIP, WAR, and Innings Pitched (IP) in our success metric for starting pitchers. While both WHIP and WAR seem to be commonly agreed upon as good measures of success, we also wanted to include IP in our metric due to the difference in dependence placed on starters vs relievers to perform for extended amounts of time. For our relief pitchers success metric, we included the same three variables with the addition of K_pct and G. We believed that high K_pct shows how a reliever is resistant to high pressure situations late in the game, and that G gives a reflection of durability that is unique to relief pitchers in combination with IP. The selected variables were standardized and slight weights were placed on each variable based on an agreed upon order of importance.

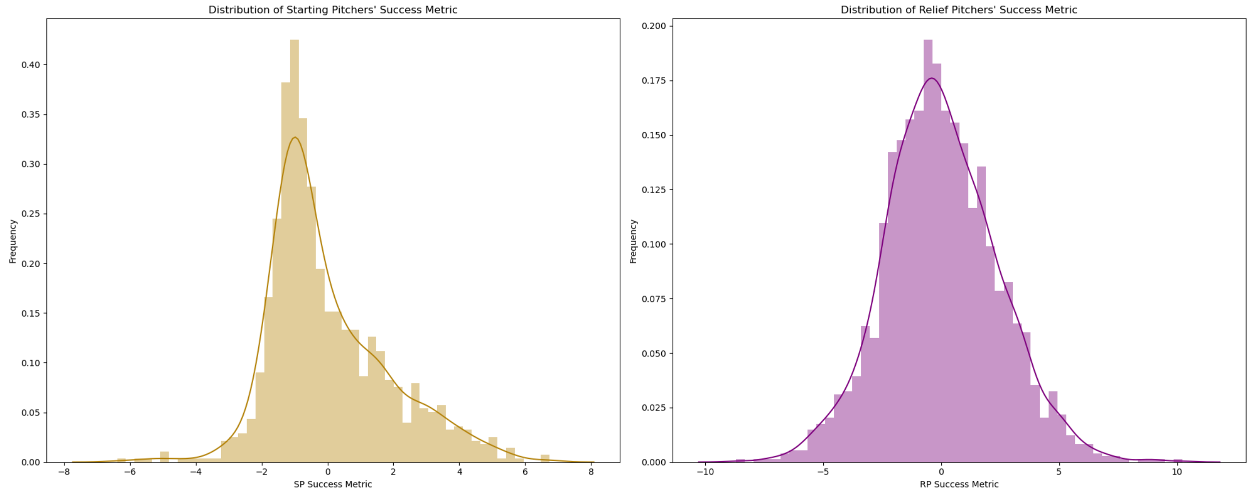


Figure 1: Distribution of success metric scores for Starters (left) and Relievers (right)

The figure above supports the choice of statistics to represent success, as the weighted combination of those statistics over all pitchers' statistics closely resembles normal distribution.

Success driving characteristics results for Starters and Relievers

After creating the two success metrics, we ran simple linear regressions for each remaining variable in the dataset on this same success metric. We then went through each of the statistically significant predictors for both starters and relievers and selected about 10 variables for Best Subset Selection (BSS). As a result of BSS, **Location Plus, HR_per_9,** and **K_pct** were selected for **Starters**, highlighting the importance of throwing strikes, keeping the ball in the ballpark, and the ability to strike guys out. For **Relievers**, BSS selected **CSW_pct, Stuff_plus,** and **HR_per_9**, highlighting the importance of a "nasty" pitch arsenal that gets swings and misses while also keeping the ball in the stadium. This makes sense as relievers should be expected to pitch in high-leverage situations where swings and misses are highly valued.

Approach to identifying pitchers who fit better in a different role

To identify outliers, we used the outputs of several clustering algorithms. However, our finalized data consisted of 131 columns of statistics about each pitcher over the last 3 years of seasons. If left alone, the high-dimensional data adds a challenge for clustering algorithms due to the curse of dimensionality, which leads to sparse data distributions and makes meaningful cluster assignment difficult, since high-dimensional spaces often contain irrelevant or redundant features that introduce noise, making true clusters harder to be formed.

We implemented a Neural Networks transformer for this dimensionality reduction, which allowed us to compress the data from 131 dimensions to 3, where a step-by-step reduction with a non-linear activation function allowed the model to learn complex non-linear mappings from higher to lower-dimensional spaces. Using a higher order method like an autoencoder allowed us to retain the most salient features and capture intricate patterns in the data that linear methods might overlook. We then used these encoded features for clustering in downstream tasks.

To effectively cluster pitcher data in a reduced-dimensional space, we used an ensemble of clustering algorithms, leveraging each algorithm’s strengths for a more well-rounded analysis. While K-Means is ideal for spherical clusters, its fixed cluster number and sensitivity to outliers needed to be balanced by other methods. Mean Shift added value by autonomously determining cluster counts based on data density. Spectral Clustering is designed for determining non-linearly separable clusters that do not need to be convex, which is the case for our complex statistical data. Agglomerative Clustering provided the benefits from hierarchical clustering , yet faced challenges with noise and outliers. Each algorithm was hyperparameter-tuned to optimize performance. HDBSCAN was initially considered for its density-based clustering approach but was ultimately excluded due to its tendency to either ignore points or form overly sparse clusters in our dataset. Based on the resulting clusters for every algorithm, every pitcher + season row was assigned as RP or SP. The final roles of SP or RP were assigned based on a pitcher + season’s most frequent assignment from all algorithms.

Clustering Results

In the figure below we show the comparison of true roles, found in Fangraphs, to the final assigned roles from the ensemble of clustering algorithms.

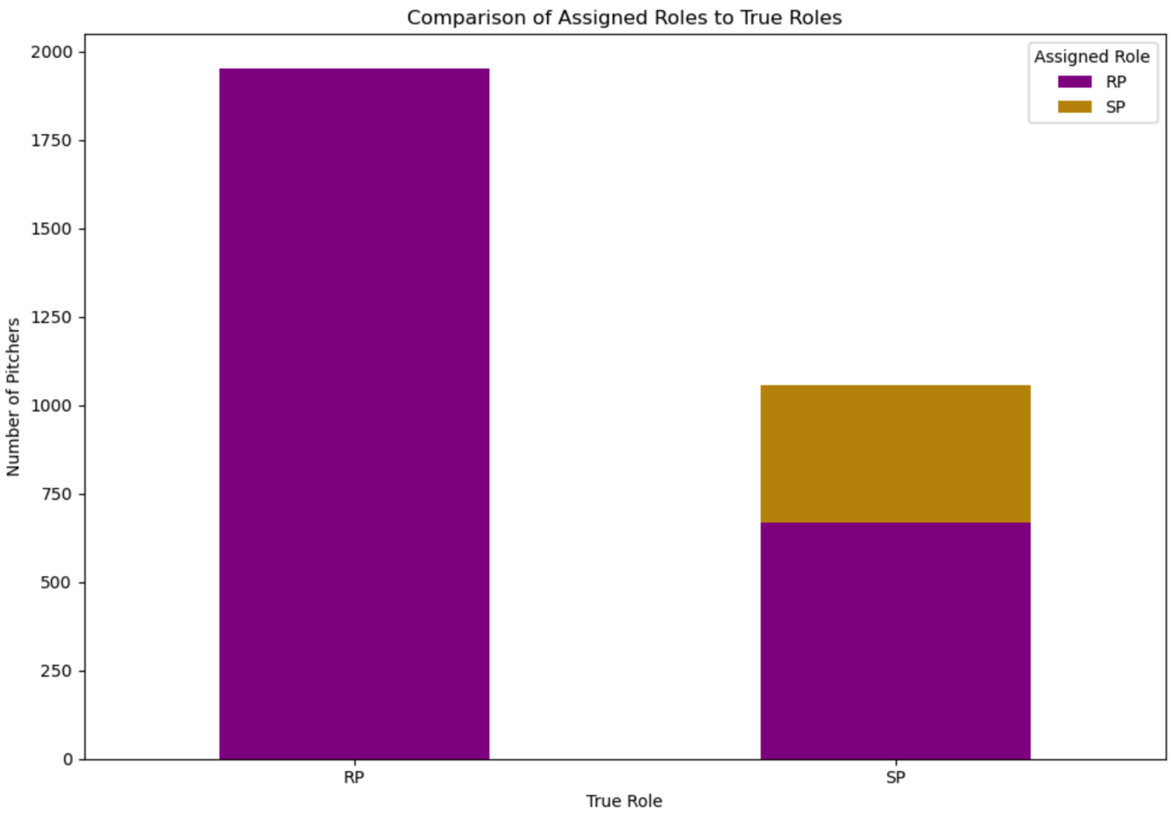


Figure 2: Each column is the true role found in Fangraphs, whereas the colors are roles assigned by clustering algorithms voting

The results show that there are True Starters, classified as Relievers, however no True Relievers are classified as Starters. While the further analysis of certain statistics supports that, we would like to outline that a potential limitation to labeling pitchers in the clustering manner is that there are Starters with statistics similar to Relievers (lower stamina, IP is smaller). However, there are significantly fewer Relievers with statistics similar to Starters (again IP for Reliever will not be as high as for a consistent Starter).

Based on the final assignments, we calculated the distributions of every statistic for 3 groups: True Starters assigned Starters, True Starters assigned Relievers, and True Relievers. The figure below shows the comparison of some picked statistics which support the final clustering method assignment.

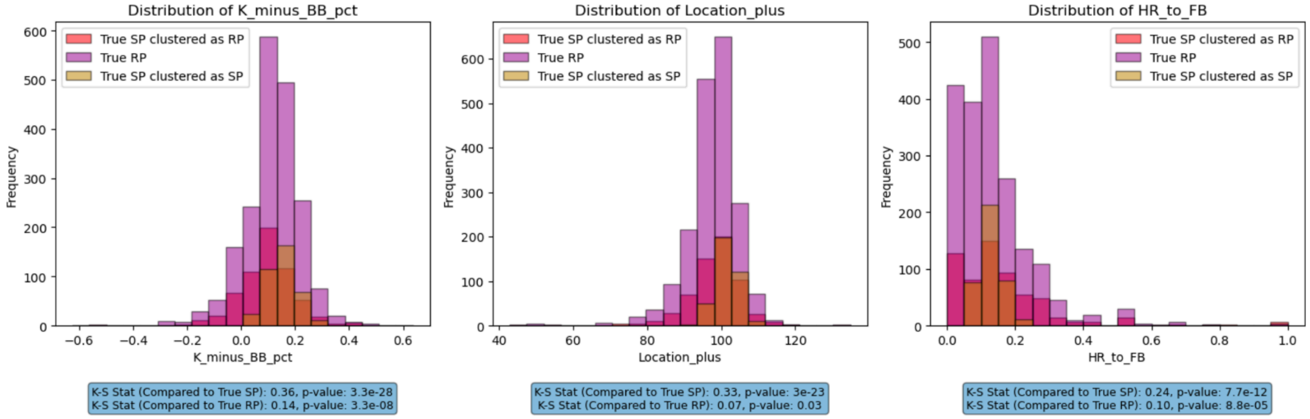


Figure 3: comparison of three groups' distribution of specific statistics shows in the titles of the graphs

For each of the columns in our dataset we ran a K-S test to determine whether the True Starters we clustered as Relievers were more similar to True Relievers or True Starters. A smaller K-S value for all three columns shown in Figure 3 between the True Starters that we assigned as Relievers and True Relievers signifies that the True Starters that we classified as Relievers have very similar stats to True Relievers. Interestingly, True Starters appear to have tighter distributions that lie around higher mean scores. The comparison of these statistics shows that while there is a potential limitation on the clustering methods described above, the clustering method still performs well in assigning roles based on statistical similarities between these pitchers that were 'misclassified' and True Relievers. The statistics we chose to display in Figure 3 (K_minus_walk_%, location_plus, and HR_to_FB) are valuable in that they correspond more to how a pitcher pitches than to something less controllable, like innings pitched. The clustering ensemble used was able to pick up differences in these important statistics to assign pitchers with similar distributions of these stats as all Relievers, despite their True Role.

Pitchers who would be better as Relievers instead of Starters

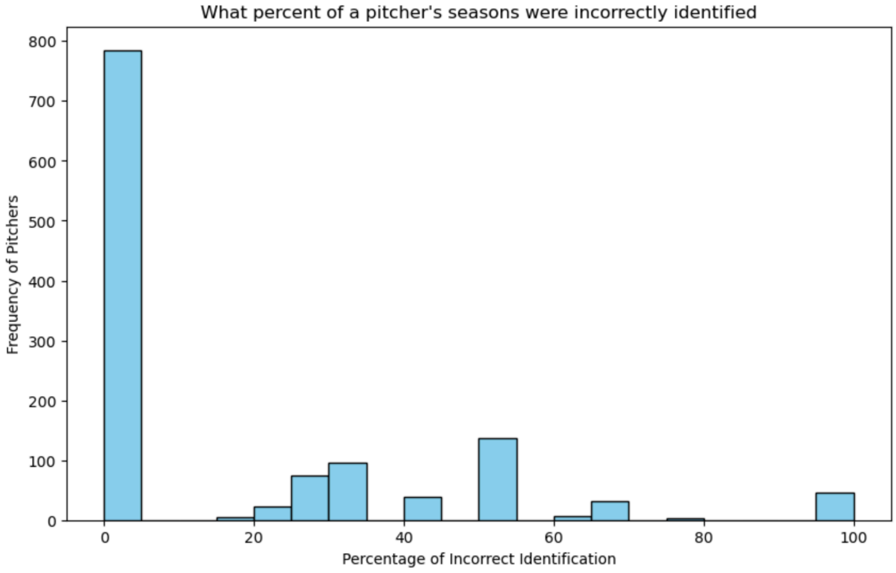
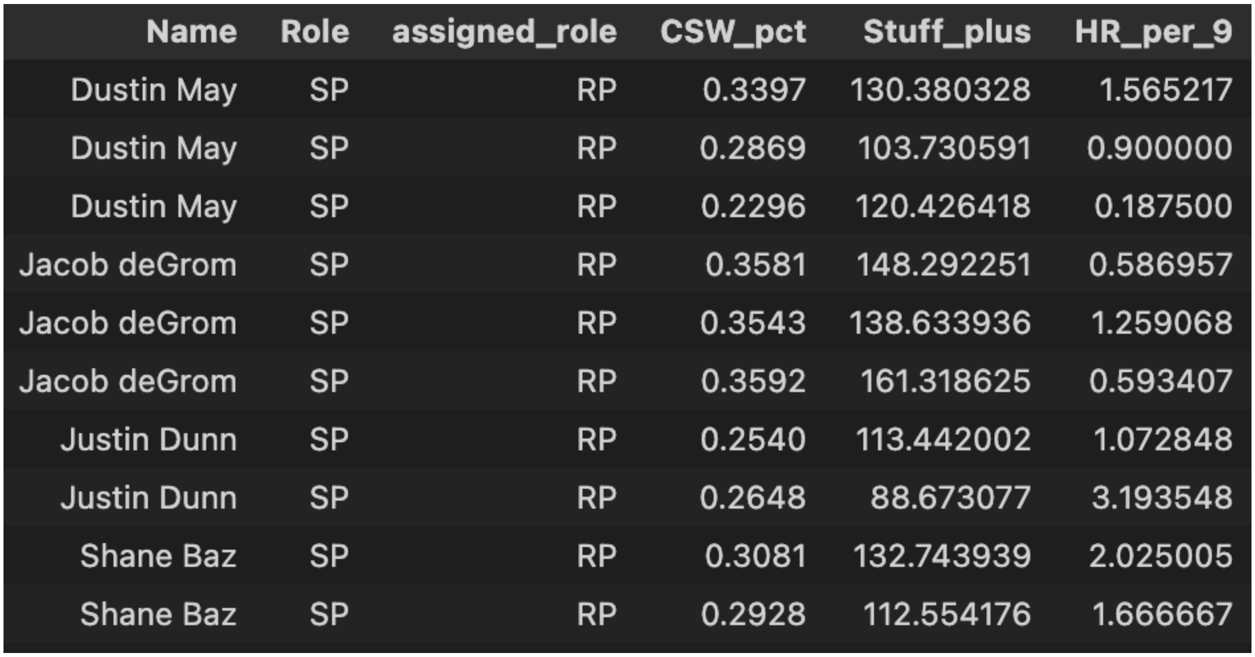


Figure 4: Percentages of a pitcher’s seasons where their role was identified incorrectly.

There are some pitchers that were classified as RPs when they were True Starters in over 50% of their appearances. The majority of pitchers were correctly identified 100% of the time, but we wanted to focus our further analysis on pitchers who were consistently assigned the role of Reliever when their True role was a Starter. We also wanted to focus on players that made multiple season appearances. These qualifications led us to select the following pitchers to be switched to Relievers from Starters: Dustin May, Justin Dunn, Shane Baz, Jacob deGrom.

The previously mentioned pitchers are Starters who have had limited innings due to injury, but have seen success in the limited inning appearances. 3/4 boast a consistently high Stuff+ metric above league average; 2/4 have consistently higher CSW_pct than the league average, while the other 2 are just above the average; and 2/4 have a lower HR_per_9 rate than the league average. This insight ties in with our findings of significant drivers of Relievers’ success metric, showing that these 4 pitchers could be very successful Relief Pitchers.



Name	Role	assigned_role	CSW_pct	Stuff_plus	HR_per_9
Dustin May	SP	RP	0.3397	130.380328	1.565217
Dustin May	SP	RP	0.2869	103.730591	0.900000
Dustin May	SP	RP	0.2296	120.426418	0.187500
Jacob deGrom	SP	RP	0.3581	148.292251	0.586957
Jacob deGrom	SP	RP	0.3543	138.633936	1.259068
Jacob deGrom	SP	RP	0.3592	161.318625	0.593407
Justin Dunn	SP	RP	0.2540	113.442002	1.072848
Justin Dunn	SP	RP	0.2648	88.673077	3.193548
Shane Baz	SP	RP	0.3081	132.743939	2.025005
Shane Baz	SP	RP	0.2928	112.554176	1.666667

Figure 5: Proposed Relievers' Stats (League Averages: 0.268 CSW_pct, 97.182 Stuff+, 1.452 HR_per_9)