Reclassifying Relief Pitchers

Dylan McGee

Syracuse University

David B. Falk College of Sport and Human Dynamics

May 2021

Abstract

Since the creation of baseball, rules initially set in place have stayed consistent with the current version of the game. One rule that needs questioning is the utilization of relief pitchers. Relief pitchers are classified under the following “roles:” Closers, Set-up Men, Middle Relievers, Long Relievers, Left-Handed Specialists, and Openers. Once a player is assigned a role for their team, they are expected to perform in those situations of the game. However, how can we determine who a true “closer” is? What makes a player fit this role? Typically, a team would assess all the relievers on their roster and assign roles based on standard statistics that have been around for years. What if, using a dataset with Standard, Advanced, Batted Ball, and PFX data, we were able to cluster players based on their metrics? This way, we can reclassify relievers based on the groups they fit in compared to all the pitchers in the league, not just their team.

Introduction

Cluster analysis groups objects together based on similarities within the data given. The purpose is to distinct patterns within the data. In this instance, our data focuses on relief pitchers. There are only so many "types" of pitchers in baseball. The ones currently used are being questioned in this research. This "type" can be quantified and better explained through K-Means Clustering. Once the analysis is concluded, a set number of groups will be set, each with their unique strengths that can be used in situations better fit than previous.

For example, suppose a cluster excels in getting ground balls by having a higher ground ball percentage (GB%) than the average reliever. Let us call this Cluster 'X.' Some pitchers accounted for in Cluster X could have had numerous roles prior, including being a team's Closer, Set-up Man, or even Middle Relievers. Now that they are reclassified into Cluster X, they will only be used in situations when their team needs their strength, which in this case, a ground ball. Assume the score is tied in the middle of the 6th inning with one out and runners on first and third base, and the starting pitcher is getting tired. This situation would typically go to a Middle Reliever or Set-up Man with no guarantee that they excel in getting ground balls. With the new form of classifications, this is the ideal time to use a reliever from Cluster X. With their high GB%; the team would love to keep the ball on the ground and try to get a double play to end the inning without giving up a run and keeping the game tied. It's a situation like this where reclassifying relief pitchers using clustering can help teams win more games at the margins.

Literature Review

History of Baseball and Relief Pitching

Rules, structure, and team strategy of baseball have progressed throughout time. Presented by Woltring *et al*[1], many of the eras left an impact on the use of relief pitchers. This thesis will cover some eras through the course of baseball's foundation. Table 1 shows all the eras with their respective years of existence and changes.

Table 1: *Baseball Eras*

Era	Years	Notes
Pre-Dead Ball	1839 – 1900	1) Origin of the game
Dead Ball	1901 – 1919	1) home plate is changed from a square to base creatin larger strike 2) Foul balls begin to count as strikes
Live Ball	1920 – 1941	1) American League was created 2) Spitball and Emery ball were removed from the game 3) Baseball was reconstructed allowing more success for hitters
Integration	1942 – 1960	1) Many players were replaced because of service time in WW2 2) Integration of all athletes regardless of skin color
Expansion	1961 – 1976	1) Expansion in teams from 16 to 24 2) Strike zone was redefined: shoulders – bottom of the knee 3) The slope of the mound decreased from 15 to 10 inches above home plate (1969) 4) Creation of Designated Hitter in American League (1973)
Free Agency	1977 – 1993	1) Free agency caused a restructure in team strategy 2) Player salaries drastically increased
Steroid	1994 – 2005	1) Many star players were increasing power by performance-enhancing drugs (PED's) which were banned by MLB
Post-Steroid	2006 – *Present*	1) Punishment for steroids became significantly more threatening 2) Advanced research on pitching metrics resurrecting pitching dominance

As rules began to change and strategy shifted, so did the philosophy of relievers. During the Pre-Dead Ball Era, the concept of a relief pitcher was admirable. In most cases, starting pitchers threw the whole game. There had been little to no research on the safety of pitcher's arms at the time. This resulted in the belief that using the best pitcher as much as possible was the best way to be successful. Per Chris Bodig *et al*[2] and Pete Palmer *et al*[3], Table 2 shows how often starters threw complete games within each era.

Table 2: *Totals Among Pitchers Per Decade*

- **% G as CG:** Percentage of complete games over the decade
- **% G as RP:** Percentage of games using a reliever over the decade
- **GR 50 +:** Number of pitchers who threw at least 50 games in relief over the decade
- **GR 100+:** Number of pitchers who threw at least 100 games in relief over the decade

Decade	% G as CG	% G as RP	GR 50+	GR 100+
1871-1879	92.0%	7.6%	0	0
1880-1889	93.9%	6.0%	0	0
1890-1899	85.4%	13.5%	1	0
1900-1909	79.0%	19.4%	10	0
1910-1919	56.9%	36.7%	109	6

As we can see, the Pre-Dead ball Era only relied on starters to throw complete games. From 1871-1899, there was only one pitcher in the league who pitched in relief over 50 times over a decade. That pitcher resulted in Hall of Famer Ty Cobb. Bodig explains a lot about how relievers were looked at during this period. In the rare cases that a reliever needed to be used, ball club's philosophy was to throw their best pitcher available. There were no "designated relievers." There were starting pitchers and starters who did not start that day who were available to throw later in the game. Ty Cobb was one of the best pitchers of his era. It makes sense he was the only reliever with over 50 appearances. At the same time (1871-1910), there was only one pitcher who had more relief appearances than starts; "Doc" James Otis Crandall. He was nicknamed "Doc" because he excelled in late-game situations to save his team; like a doctor would with a patient. Even though he started in roughly 40% of his outings, he was the first sign baseball had seen of a true reliever. The strategy of pitcher usage has changed drastically compared to the modern-day. The beginning of the Dead Ball Era was when we began to see a relative uptick in the utilization of relievers.

Once the 1920s began, a shift in advantage from the defense to offense materialized. Spitballs were removed from the game. Baseballs were replaced mid-game to ensure that the ball was always white so players could see it. Most importantly, the baseball was redesigned with a "new cushioned cork center" in the middle to allow for more offense production. With the advantage in the hands of the offense, it was time to reconstruct how baseball used their pitchers. Frederick "Firpo" Marberry was the first-star reliever. Coming out of the bullpen 114 times and only making 5 starts between 1925-1926, he was coined a true arm that excelled in getting saves (even though the term was not used until 1969). Soon followed many pitchers who were called upon in late-game situations. These pioneers were eventually called “closers.” Some notable starting pitchers who were called upon to close are listed below.

Table 3: *Starting Pitchers Used as Closers (Pre-Dead Ball Era – Beginning of Integration Era [1945])*

- **Orange:** Hall of Fame Inductee

Name	W	L	SV	G	GS	IP	K/9	BB/9	HR/9	BABIP	ERA	FIP	WAR
Firpo Marberry	148	88	101	551	187	2067.1	3.58	2.99	0.42	0.274	3.63	3.92	22.9
Johnny Murphy	89	51	97	356	40	945.1	3.33	3.78	0.45	0.264	3.54	4.12	4.4
Clint Brown	89	93	64	434	130	1485.2	2.48	2.23	0.51	0.295	4.26	4.03	18.9
Joe Heving	76	48	63	430	40	1038.2	3.72	3.29	0.55	0.291	3.9	4.01	6
Jack Quinn	247	218	57	756	444	3920.1	3.05	1.97	0.23	0.291	3.29	3.15	61.2
Lefty Grove	300	141	55	616	457	3940.2	5.18	2.71	0.37	0.284	3.06	3.36	88.8
Waite Hoyt	237	182	52	674	423	3762.1	2.88	2.4	0.37	0.283	3.59	3.76	48.9
Mordecai Brown	239	130	49	481	332	3172.1	3.9	1.91	0.12	0.26	2.06	2.36	49.3
Ace Adams	41	32	49	299	7	550	2.75	3.65	0.39	0.262	3.4	3.72	0.7

As the role of relief pitchers grew, so did their recognition by media and fans. From Chris Bodig *et al*[4], Page was the first “closer” to earn a relevant amount of MVP votes. Although he would be considered today as the closer for the New York Yankees, he was not utilized like closers are today. For example, in-game 7 of the World Series, Page came in to throw innings 5-9 to secure the championship for his team. The relievers were stretched out over a much longer period of innings per average appearance. However, this was a big step in the right direction. Jim Konstanty became the first reliever to win the MVP in 1950. In the years to come, there was an emergence of relief pitching stars specializing in preventing offensive production better than ever before. Some Hall of Fame relievers that occurred over the next 30 years are Goose Gossage, Lee Smith, Dennis Eckersley, Rollie Fingers, Sparky Lyle, Al Hrabosky and Steve Bedrosian.

As baseball continued to grow until the state of the current game, reliever utilization did as well. The more time that has passed, the less likely a reliever was expected to throw for an extended amount of innings per appearance. Table 4 shows how closers are on “shorter leashes” meaning they are not throwing as many innings per outing.

Table 4: *Average Innings Pitched for Top 20 Closers (ranked by saves)*

Years	Average IP per save		% of Saves 1 IP or less	% of Saves 2 IP or more
1970-74	1.62		43%	24%
1975-79	1.68		42%	28%
1980-84	1.62		39%	26%
1985-89	1.35		52%	21%
1990-94	1.13		72%	9%
1995-99	1.03		86%	2.3%
2000-04	1.03		88%	2.2%
2005-09	1.01		92%	0.9%
2010-14	0.99		96%	0.2%

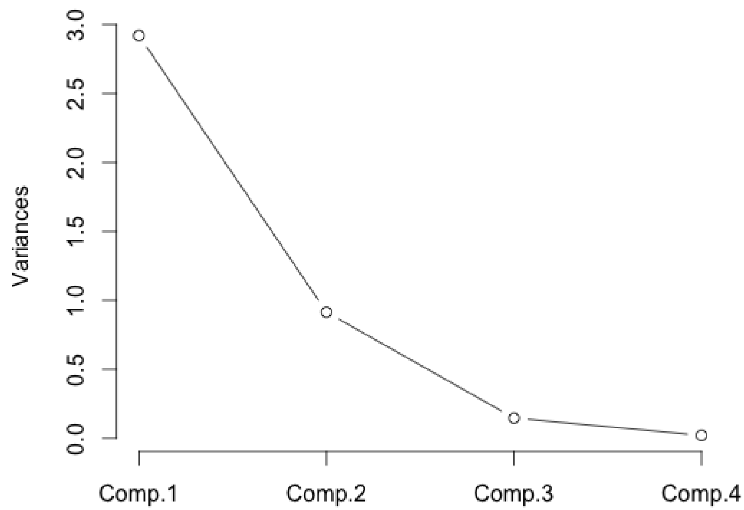
As we can see, closers don't throw to nearly the amount of batters they used to in the '70s. This shift is the last major change in reliever usage until the introduction of the Opener in 2018 by the Tampa Bay Rays.

Clustering

There are many ways to analyze data successfully. One of the more robust methods is clustering. According to various sources of research including Kumar *et al*[5] and Starmer *et al*[6], PlanSpace[7], Christian Henning in *Flexible Procedures for Clustering*[8], and Frank Denoncourt[9], cluster analysis divides data into groups (clusters) that are meaningful, useful or both. It can also be used as a "starting point" for further analysis; primarily being data summarization. The greater the homogeneity within a group and the larger the difference between groups, the more distinct clustering results. With the proper data, one can learn much more about the similarities differences in the dataset based on patterns discovered by machine learning. There are multiple forms of clustering data scientists use depending on what they're trying to achieve. For now, we will focus on K-Means Clustering Analysis.

K-Means Clustering is one of the oldest and most widely used clustering algorithms. Essentially, K-Means self-determines a best-fit centroid by finding the mean of a group of points in a continuous *n*-dimensional space. For successful analysis, we must determine an initial number of centroids, or groups, which will be represented as *K.* One method to determine the optimal number for K based on your data is the "Elbow Method." PlanSpace *et al*[10] explains how The Elbow Plot determines the ideal number of centroids by graphing the variance within the cluster analysis. Once the graph shows a strong bend (like an elbow) we can determine the ideal number of centroids needed for a data set. An example from PlanSpace is shown below using the iris dataset given in RStudio.

Figure 1: *Elbow Plot (Iris Dataset)*

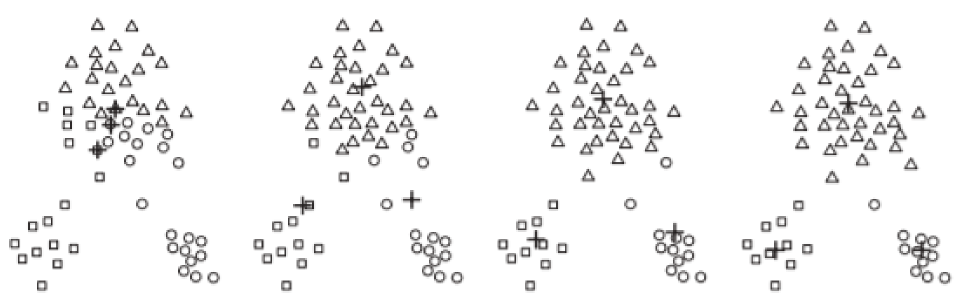


Comp. 2 serves as the optimal iterations of clusters for the Iris dataset using the Elbow Method. Now that we have calculated K, we can begin the clustering process. Figure 2 explains the process thoroughly. As more data is included in our clustering algorithm, the centroids will continue to move closer to their target mean of each group until they cannot move any further. This proves is shown in Figure 3.

Figure 2: *Basic K-Means Algorithm*

```
Select K points as initial centroids.
repeat
   Form K clusters by assigning each point to its closest centroid.
   Recompute the centroid of each cluster.
until Centroids do not change.
```

Figure 3: *K-Means Algorithm in Sample Data (K =3)*



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

One question has yet to be answered: “How do we determine what centroid is closest to each point?” Calculating Euclidean distance is used to answer this question. Typically, programming languages like RStudio and Python calculate this for you, but we can calculate rather simply. By using the sum of square error (SSE), we can calculate the error of each data point (or the Euclidean distance to the centroid).

Figure 4: *SSE Formula*

Symbol	Description
$\mathbf{x}$	An object.
C_i	The i^{th} cluster.
$\mathbf{c}_i$	The centroid of cluster C_i.
$\mathbf{c}$	The centroid of all points.
m_i	The number of objects in the i^{th} cluster.
m	The number of objects in the data set.
K	The number of clusters.

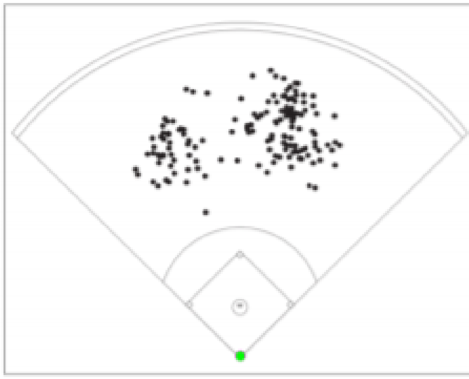
$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

Using the notation above, *dist* is the “standard Euclidean distance between two objects in Euclidean space” (Kumar *et al*[4]*).* One last issue remains in the K-Means form of cluster analysis. When random initialization of centroids is used, multiple runs of K-Means will produce different SSE results. There are many ways to work around this. The most popular is to have multiple runs and select the set of clusters with the Minimum SSE. This way, we know our centroids are most closely fit to their cluster. Once we have our minimal SSE and our ideal clusters, we have completed the statistical analysis. Lastly, we must make sense of the clusters to create an accurate conclusion.

Clustering Effects in Baseball

Using the machine learning tool listed above, we can conclude theories in baseball that help advance the game. Gerlica *et al*[11] tested the use of clustering on outfield shifts. Their goal was to find a more accurate method for the Air Force College baseball team to determine where to shift their players on defense. Originally, the coaching staff would look at scouting reports of their opponent’s offense and adjust accordingly based on previous results. The group’s plan was to evaluate the catch probability of all flyballs in the outfield to move their outfielders in positions to maximize the likeliness of any given flyball being caught regardless of who was up to bat. Using past locational data with the position of Air Force’s outfielders when each flyball was caught, the group was able to use K-Means cluster analysis to determine the ideal position of each outfielder. In Figure 5, the team plots the locational data for each ball caught and determines the means of each cluster to create their *K* centroids. Once the centroids are created, the testing can begin, and the results are shown in Figure 6.

Figure 5: *Determining Centroid Location on Previous Ball-in-Play- Results*



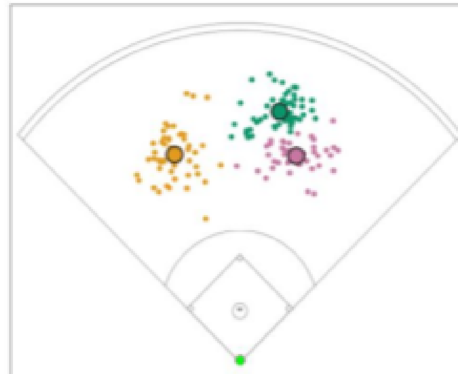
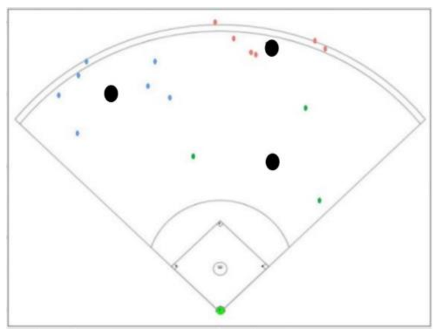


Figure 6: *Adjusted Position of Outfielders using K-means Clustering Techniques*



The group concluded that the adjusted positioning of outfielders would result in a 7.1% increase of outs created by the outfielders. The group concluded that this position shift in outfielders would be relevant enough to gain another 1-2 wins from the previous season. Two wins would have been enough for Air Force to take first place in their conference and ultimately a bid to the college tournament. Although Air Force did not end up using the findings of their students, the team was confident that their findings would lead to greater success on defense, saving runs and helping the team generate more wins. This project proved that cluster analysis can show findings to help teams be more successful if teams decided to act upon the findings.

Using Clustering and Other Analysis on Relievers

Now that we have discussed the history and utilization of relievers along with learning about the powers of clustering analysis, we can combine both and generate findings. Looking at the current roles of relievers, I realized there are strengths each reliever has that can be used more effectively in games to generate success. One major statistic that needs to be looked at more carefully is ground ball percentage. Sarris *et al*[12] *go* into depth about the effectiveness of ground ball pitchers in today's game. Sarris explains how as the game has progressed, the ground ball pitcher has become more and more important for a staff. The reason being is simple. As the game has progressed, the long ball has become significantly more valuable. Currently, certain teams can get away with a win from only 4 hits if some of those landed over the outfield wall. We see successful teams leading the league in power metrics. Therefore, having pitchers who can keep the ball on the ground explode in value. These are the pitchers who can counter the emergence of power offenses. Although this is information commonly believed prior, this article solidifies the theory. This is important for a project on relievers because it stresses the strength of the ability to get ground balls. Marcou *et al*[13] performs cluster analysis on all pitchers to see if he can find patterns in pitchers that do have high ground ball rates. Although he did not come up with any relevant conclusions, it gives confidences that clustering on pitchers can generate some sort of conclusion.

Zimmerman *et al*[14] questions if pitchers can prevent solid contact. Zimmerman uses stats from Baseball Info Solutions, Statcast, and FanGraphs to show real examples of players' performance to get weak contact. He uses Dallas Keuchel, who at the time, had a sharp decrease in soft contact from his most recent season to his previous ones prior. Zimmerman attempts to analyze Keuchel's stats to come up with why that happened. He notices that a change in average velocity could be part of the reason he was giving up harder contact. Also, Keuchel was facing the same teams over again, and the more he faces the same batters, the more likely they will get an understanding of what he throws and when he varies his pitches. This will help a batter's success in making hard contact. This article shows that in order to analyze relieve, one needs to come up with a factor for pitchers that determines how long he has been throwing against the same teams. Accounting for hard hit percentage is crucial.

Cermark *et al*[15] determines the difference between percentage and "per-nine" statistics. The primary example used is Strikeout Percentage (K%), and Strikeout Per 9 Innings (K/9). Many underestimate the difference between the metrics. He concluded that "per-nine" numbers don't tell the whole story. For example, a pitcher with a 12 K/9 does not include how often they give up hits. A pitcher with a high K% proves that the pitcher gets strikeouts often, because the stat accounts for all batters the pitcher faces. K/9 only accounts for the outs a pitcher creates. One must use K% for cluster analysis to fully determine the effectiveness of a high-strikeout pitcher. If there is an option for a percentage stat over a "per nine," you must use the percentage statistic for your research.

An important tool to help group clusters is the ability for a pitcher to throw strikes. Harry Pavildis *et al*[16] on "Command vs. Control" goes in depth about command. The research determined there are two forms of pitchers that can throw strikes: those with control and those with command. A pitcher who can control the strike zone can consistently throw strikes but not necessarily in pinpoint locations during their outings. A pitcher with command locates pitches inside or outside the zone, conscious of not leaving pitches in the middle of the strike zone to reduce offensive production. From the research, we can conclude that relivers with command are more effective than those with control. One can account for this by using certain metrics to determine relievers with the greatest ability to command. One way to accomplish this is by using CSAA (called strikes above average) and CS Prob (called Strike probability). Having statistics to account for command is crucial in cluster analysis of relief pitchers.

Other Relevant Forms of Data Analysis

When conducting data analysis, there are minor mistakes statisticians can make which can affect the outcome of the project. One of the most common mistakes is not normalizing data. Per Deepika Singh in *Normalizing Data in R*[17] and Nicky Lamarco *et al*[18], normalizing prevents machine learning algorithms to heavily favor the variables that are used on a larger scale. An example is shown in Figure 7.

Figure 7: *Non-Normalized Data*

```
  Dependents        Income          Loan_amount       Term_months
 Min.   :0.0000   Min.   : 173200   Min.   :  18600   Min.   : 36.0
 1st Qu.:0.0000   1st Qu.: 389550   1st Qu.:  61500   1st Qu.:384.0
 Median :0.0000   Median : 513050   Median :  76500   Median :384.0
 Mean   :0.7561   Mean   : 715589   Mean   : 333702   Mean   :365.5
 3rd Qu.:1.0000   3rd Qu.: 774800   3rd Qu.: 136250   3rd Qu.:384.0
 Max.   :6.0000   Max.   :8444900   Max.   :7780000   Max.   :504.0
```

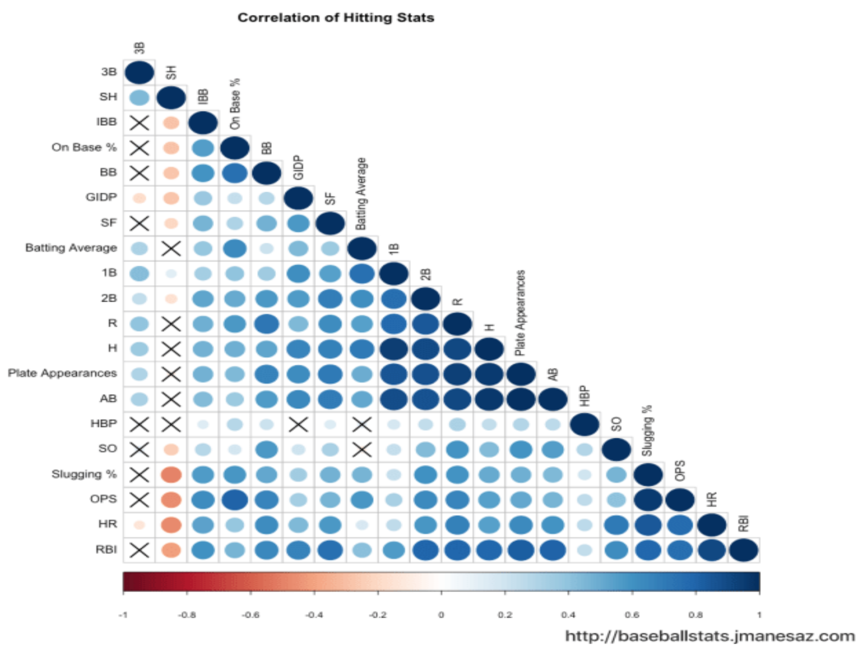
In this example, the data imported into RStudio is non-normalized. By using the *summary()* function, Singh discovers the non-normalized data by comparing the six measures. The *min, mean* and *max* show the columns are weighed differently. The larger scaled variables like Loan Amount and Income will influence the model and skew the results of any machine learning function. For accurate data analysis, this needs to be fixed. According to Becker *et al*[19], Standardization is a tool to clean non-normalized data. The technique makes all the variables centered around zero and have similar variance. With standardization, machine learning can give much more accurate results. The same data imported from Singh is shown in its normalized from after using standardization from the *scale()* function in Figure 8. Normalization will be crucial for accurate analysis when clustering relievers.

Figure 8: *Normalized Data from Standardization*

```
  Dependents        Income             Loan_amount       Term_months
 Min.   :-0.7338   Min.   :-0.75854   Min.   :-0.4281   Min.   :-5.3744
 1st Qu.:-0.7338   1st Qu.:-0.45597   1st Qu.:-0.3698   1st Qu.: 0.3021
 Median :-0.7338   Median :-0.28325   Median :-0.3494   Median : 0.3021
 Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.: 0.2367   3rd Qu.: 0.08281   3rd Qu.:-0.2683   3rd Qu.: 0.3021
 Max.   : 5.0893   Max.   :10.80959   Max.   :10.1169   Max.   : 2.2595
```

Correlations are used to compare variables. If a variable influences the result of another in the same dataset, the relationship is considered a positive correlation. If a variable causes another to move in the opposite direction (variable A increasing causes variable B to decrease) we can determine the variables are negatively correlated. The closer the correlation number is to 1/-1, the stronger the correlation. Any correlation hovering around 0 means the variables don't hold any significance to one another. An example comparing the relationship between the most popular statistics in baseball is shown in Figure 9.

Figure 9: *Correlation of Hitting Statistics*



Jeff Manes *et al*[20] used R's correlations packages to create the correlation plot above. The bigger and "bluer" the circle, the stronger the positive correlation. For example, Hits and RBI's have a strong positive correlation. Essentially, the more hits a team has, the more likely they are to score runs, increasing their RBI's. For clustering, limiting positive correlations will be beneficial for the results. Having variables that strongly correlate will allow machine learning to account for multiple statistics that ultimately tell the same story. It is important to limit strongly correlated variables when using clustering.

Per Josh Starmer in *Logisitic Regression in R Clearly Explained*[21]*,* regression is the most popular form of data analysis. It's used to determine the strength of relationship between the dependent variable against independent variables. Your dependent variable is the metric used in hypotheses. For example, in a previous project looking to increase attendance for the Ogden Raptors, attendance was the dependent variable and was regressed on independent variables such as weather, temperature, team record, month, day of the week, etc… (Dylan McGee in *Ogden Raptors 2019 Attendance Case Study*[22]*).* Using regression, we can make many conclusions. In the

Ogden Raptors example, regressing proved that games on the weekend showed a higher increase in attendance. Also, games against their rivals, the Orem Owls, tended to have a larger than average attendance. Regression will be essential to learn more about the conclusions of cluster analysis.

Once statisticians create a successful regression, they usually will test their results. One of the most popular methods of testing their resulting is by *training and testing* their data. Per Jalayer Academy in *Creating Training and Testing Data in R*[23], one can accomplish training their model by splitting their data set into a training and testing set. The 80/20 method is respected among data scientists. Essentially, 80% of the raw data set is "split" into the training set for the model, while the other 20% is used to test the model. The training data's job is to literally "train" the model to best predict in the testing data set. A strong model will accurately predict the dependent variable in the testing data. After training and testing the data, a confusion matrix can be used to determine the accuracy of the model. Josh Starmer again in *Machine Learning Fundamentals: The Confusion Matrix*[24] explains how a confusion matrix can account for how accurate a model is. The matrix shows how often the model on the testing data was correct and incorrect in both false-positives and false-negatives. With this information, one can calculate the average of correctness. The higher percentage, the more accurate and stronger the model is. A confusion matrix will be used when reclassifying relief pitchers in order to determine how strong the model is.

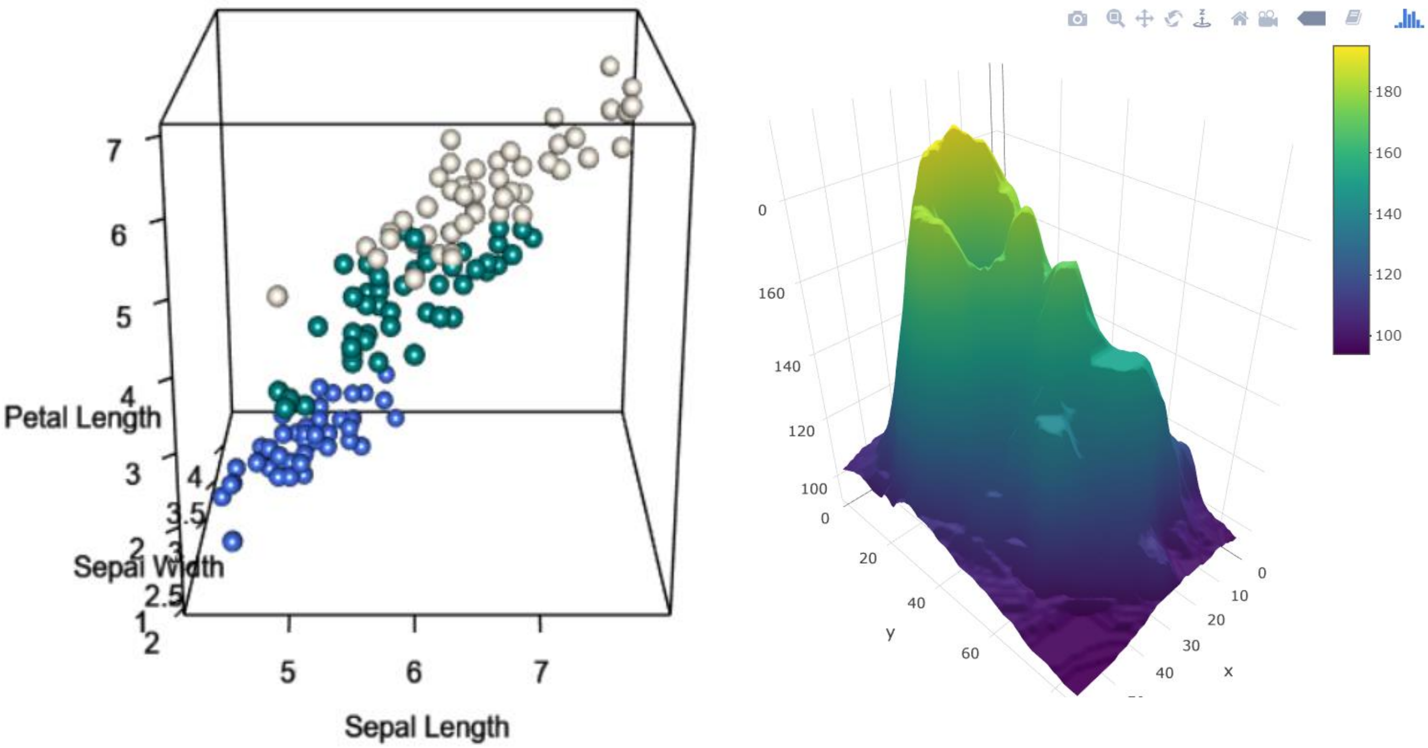
Table 5: *Confusion Matrix from Model Determining Likelihood of Heart Disease*

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	142	22
Does Not Have Heart Disease	29	110

Table 5 shows the confusion matrix used in Starmer's example. The green represents correct predictions and the red for false-positives and false-negatives. By adding both correct numbers and dividing by the sum of all four squares, you can determine the strength of the model. The model in this example was correct 83% of the time; a pretty strong model.

Lastly, strong visualizations are essential for explain results or a project. For cluster analysis, three-dimensional graphs would much better explain results. According to Yan Holtz *et al*[25], the *rgl* package in R can easily create easily customizable 3D graphs. Figure 10 shows the power of the *rgl* package using the Iris data set given in RStudio. With multiple different forms of visualizations and the ability to customize the legend, colors, titles, axis' and more, the *rgl* package is very useful.

Figure 10: *3D Graphs Using the rgl Package in RStudio*



Having strong, detailed graphs is important in describing the results of clustering. This package will be used to accurately create visualizations to explain conclusions.

Methodology

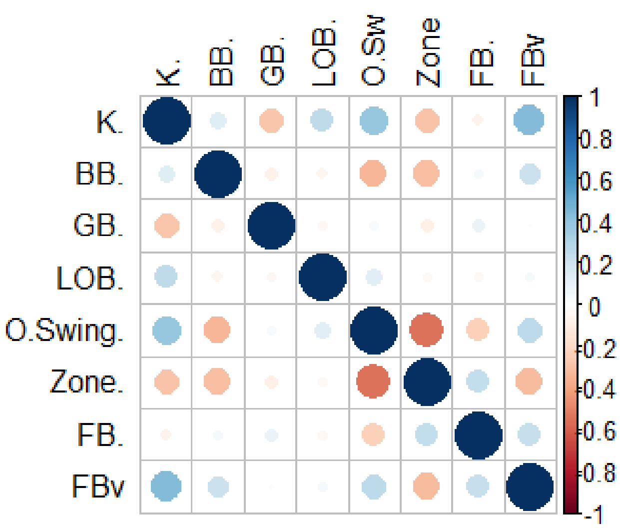
Data Summary

The dataset was created using FanGraphs. The *Custom Leaderboards* option allows users to create unique datasets fitting whatever criteria is needed for proper analysis. Filtering by relief pitchers from the 2005 – 2020 season's set the scope to analyze relief pitchers in the modern era with a significant sample size. A minimum number of innings pitched (20 IP) was created to account for any position player who may have came in to throw an inning or two over the course of their career due to a blowout/ saving the bullpens arm. If the minimum IP was not created, the data would have resulted in swayed analysis and concluded faulty findings. The last and most important decision in creating the dataset was to determine which in-game measures would be used. Keeping in mind the findings from Cermark *et al*[15] from earlier in the research, it's important to use percentage statistics over "per-nine" measures. Percentage variables accurately tell the whole story of a metric like strikeout percentage over strikeouts per nine innings. Also, it was important to make sure the variables used had little correlation to each other, as learned from Jeff Manes *et al*[20]. Keeping both studies in mind, the variables selected were as followed:

Table 6: *Variables Used for Analysis*

Abbreviation	Name of Variable	Definition
K%	Strikeout Percentage	How often pitcher gets strikeouts per plate appearance
BB%	Walk Percentage	How often pitcher gives up walks per plate appearance
GB%	Ground Ball Percentage	How often pitcher gets ground balls per balls in play
LOB%	Left on Base Percentage	How often pitcher successfully leaves runners in scoring position
O-Swing%	Outside of Zone Swing Percentage	How often pitcher gets batter to swing at pitch outside of the strike zone
Zone %	Zone Percentage	How often pitcher throws pitches in the strike zone
FB%	Fastball Percentage	How often pitcher throws a fastball
FBv	Fastball Velocity	Average velocity or speed of pitcher's fastball

Figure 11: *Correlation Matrix with Selected Variables*

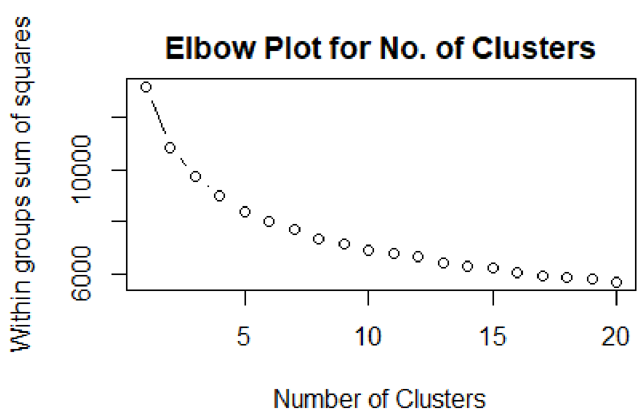


To prove the minimum correlation within the variables chosen, a correlation matrix was created. Looking at *Figure 11*, the only variables with relevant correlation were Zone% and O-Swing% at around -0.50. Essentially, the more often a pitcher threw strikes, the less likely he can get batters to swing at pitches outside of the zone. This makes sense in theory because if a pitcher tends to throw a lot of pitches within the strike zone, he wouldn't have as many chances to get a batter to chase a pitch outside of the zone. Regardless, both variables stayed in the data for analysis because the correlation number was not perceived to be high enough to affect the clustering and because both metrics tell a story about a pitcher if they excel in that statistic. For example, relievers with a high Zone% let a manager trust them to throw strikes and limit walks if need be. Relievers with a high O-Swing% can let a manager trust their pitcher to come in against a batter who is known to chase pitches outside of the zone for maximum effectiveness. After removing *NA* values, the only task left to have fully clean data was to standardize the statistics.

Standardizing and normalizing the data once again allows for every statistic to be measured equally. If the data was kept how it was, FBv would heavily sway the cluster results because the average velocity was in the low 90's, where every other statistic was a percentage resulting is a decimal value smaller than one. After scaling, the mean for each metric was one.

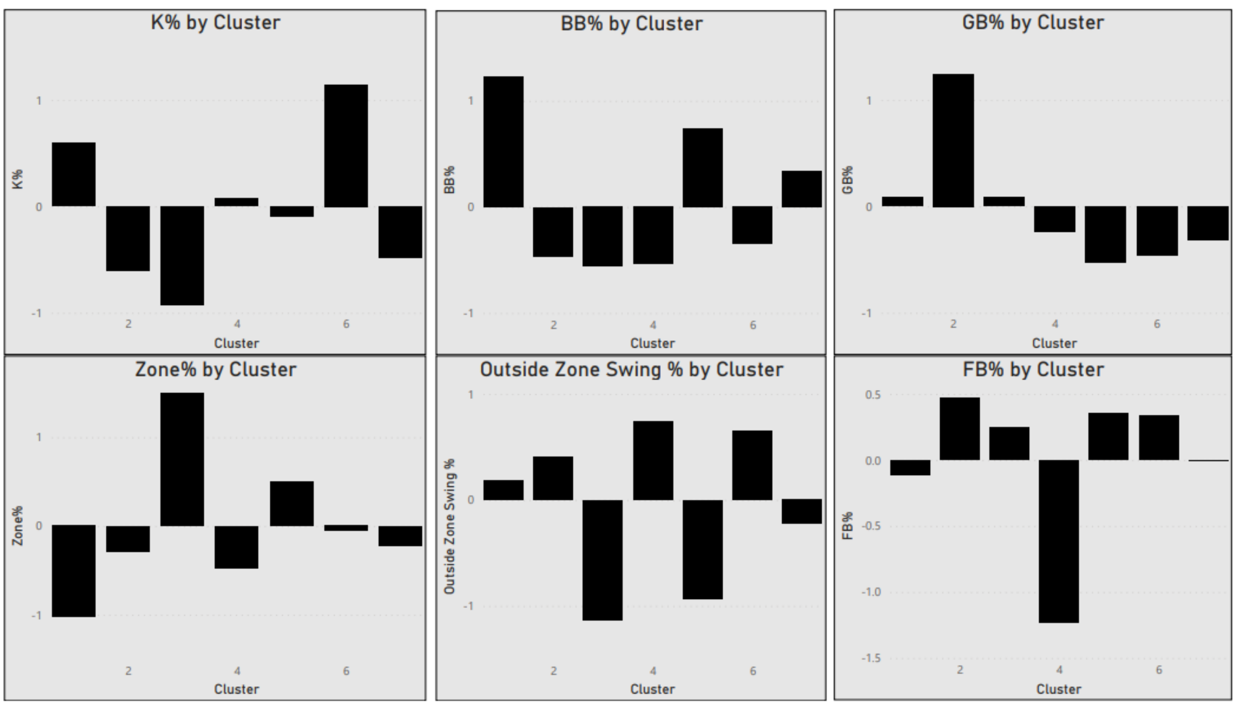
With the data being completely cleaned, K-Means Clustering is now possible to test. The first issue is to determine the number of clusters needed for proper analysis. The best way to approach a problem like this is by making an Elbow Plot. Elbow Plots show the variability within the cluster based on the number created. *Figure 12* shows the Elbow Plot created for the project. Ultimately, it was determined after viewing the plot and numerous test runs that 7 clusters was the ideal number.

Figure 12: *Elbow Plot*



With the data cleaned and the number of clusters determined, K-Means cluster analysis can begin. The *kmeans* function in R allows for the computer to group the relievers together based off patterns from their respective performance using the variables listed from *Table 6.* All 7 clusters are shown in *Figure 13* with their averages for six variables used including K%, BB%, GB%, Zone%, O-Zone%, and FB%. It's important to note the Y-axis is determined by the standard deviation of the specified statistic. Because each variable was normalized, the mean for GB% is 1.0 instead the actual league average of 44%. By using the standard deviation of GB%, we can see how other clusters compare to one another with 0 repesenting the league average. For every statistic other than BB%, any group above zero means they performed better than the average reliever. Having a lower BB% is ideal so any group below zero for BB%'s respective graph shows the groups that limit walks. *Figure 13* shows the strengths and weaknesses of each cluster. For example, Cluster 1 must struggle with their command as they lead in BB%. Cluster 2 clearly excels in getting ground balls. Their GB% is over one standard deviation higher than the average reliever from the data. Cluster 3 is the clear leader in Zone%. Cluster 4 doesn't tend to throw fastball's nearly to the extent of the average reliever by having a FB% almost two standard deviations below zero. Cluster 5 seems to be around league average in almost every category. Cluster 6 strikes out batters almost twice as much as the second highest group (Cluster 1). Lastly, Cluster 7 does not lead in any of the six statistics shown. Looking closer, we can see they are below average in every category. Looking at these averages are crucial in being able to determine the results from the K-Means clustering.

Figure 13: *K-Means Results: Group Averages*



After looking at the averages within each group, the next step was to see if groups had a high variance from outliers. One way to visually determine is by plotting the clusters based off some of the variables used. Once doing so and filtering each player by the cluster they were assigned, we can determine if there are outliers in the groups. *Figure 14* is a three-dimensional plot using the *plotly* function. Each color represents a different cluster. There are visible groupings especially between Clusters 2, 4, and 5. Another example is shown in *Figure 15* comparing Ground Ball% Percentage and Strikeout Percentage between each reliever. Clusters 2 and 5 yet again show strong groups with the addition of Cluster 7.

Figure 14: *3D plot filtered by Clusters (Fastball% vs. Ground Ball % vs. Strikeout %)*

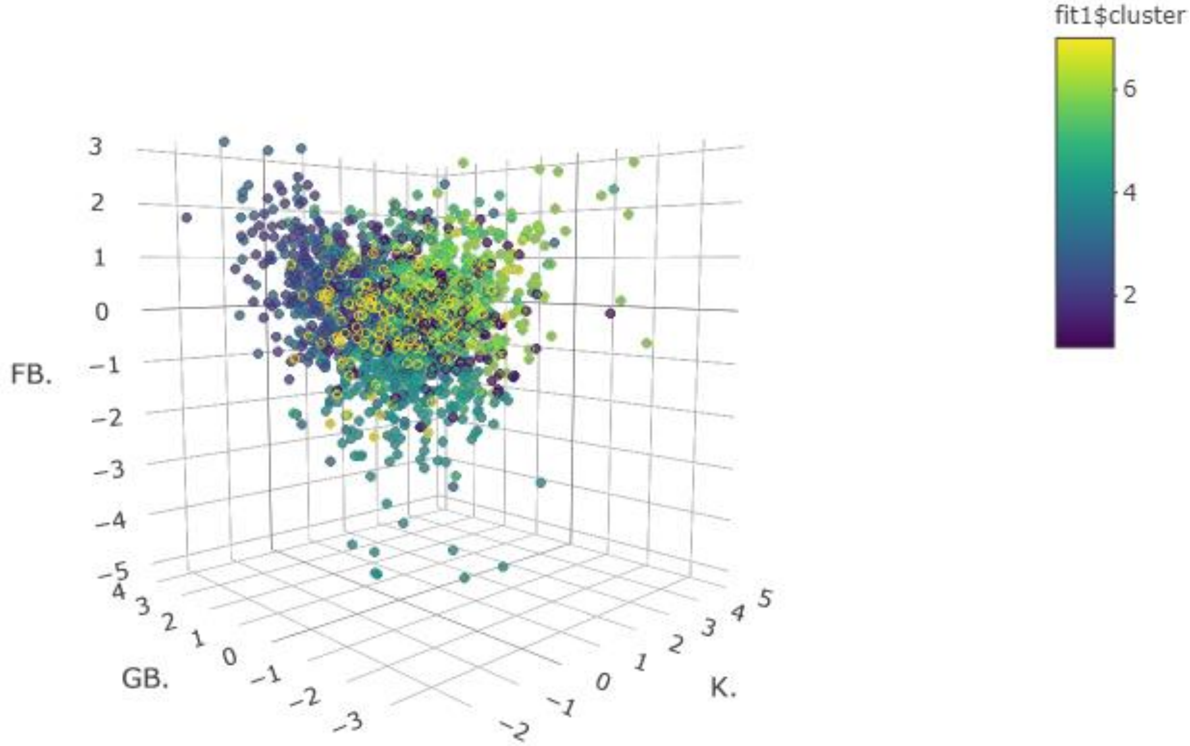
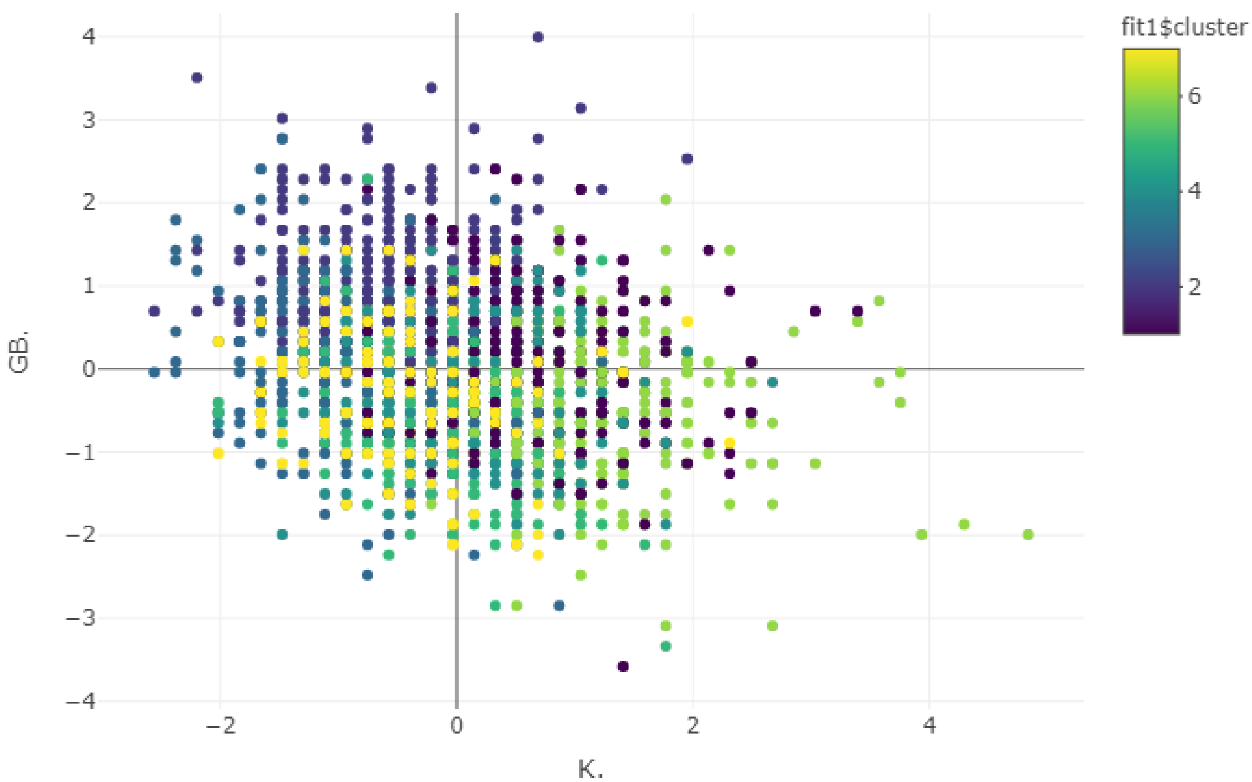


Figure 15: *Ground ball% vs. Strikeout % filtered by Clusters*



Conclusions

After gathering data, cleaning it, running correlations and elbow plots along with the K-Means test, and viewing summary statistics of each cluster, it was finally possible to output each reliever with their given result. Viewing the results allows a team to search for their players and their role to implement them in real life. The concluding clusters are shown in Table 7. If teams reclassified their bullpen into these groups and used them in situations that fit their strengths, pitching staffs could improve drastically.

Table 7: *Final Cluster Classifications*

Cluster	Role	Usage	Notable Player
1	**High Velocity Specialists**	By far highest velocity average. Excel in K's but can struggle with command. **Use when looking for strikeouts or big velocity change.**	Jordan Hicks
2	**Ground Ball Specialists**	Highest GB%, produce weakest contact but fewest K's. **Use when trying to keep the ball on the ground or looking for double play.**	Zack Britton
3	**Accuracy Specialists**	Leaders in Zone% and low BB%. **Use when trying to limit free bases via walks.**	Trevor Hoffman
4	**Chase Artists**	Have great off-speed which is used to get batters to chase outside of zone. **Use when looking to limit contact at all costs.**	Sergio Romo
5	**Average Reliever**	Relievers that don't excel in anything but also not below average. Good cluster **to use in any situation.**	J.A. Happ
6	**Strikeout Specialists**	Leaders in K%. **Use when looking to limit contact via strikeout or against batters with high K%.**	Aroldis Chapman
7	**Least Reliable**	Worst group. Have the poorest numbers in most categories. **Use when out of options/up or down by a lot.**	Dereck Rodriguez

Names for the clusters were given after reviewing the respective groups' strengths and weaknesses compared to the rest of the league. They are subjective to change. It's important to note that players like J.A. Happ have had relief outings in the earlier stages of his career. Although he is now known as a starting pitcher, there was enough relief pitching data to be considered into a classification. If teams were to utilize these new groupings, the other question to consider is what combination of clusters is ideal for most success on the field? This would have to take time to find out, but it could be possible to determine with enough simulations using different clusters as a pitching staff. Lastly, a sample team will be chosen, and each reliever will be looked up in the Cluster database to get a scope of what their bullpen's new classifications might look like. The results are shown in *Table 8.*

Table 8: *2021 Cincinnati Reds Bullpen*

Player Name	Cluster Result	Notes
Amir Garrett	High Velocity Specialist	FBv: 80th Percentile
Sean Doolittle	Strikeout Specialist	K%: 80th Percentile
Tejay Antone	N/A	Not enough data to be included for analysis
Sal Romano	Least Reliable	10th Percentile in all metrics
Lucas Sims	High Velocity Specialist	FBv: 90th Percentile
Brandon Bailey	N/A	Not enough data to be included for analysis
Carson Fulmer	Average Reliever	K%, Chase%: 50th Percentile
Cionel Perez	High Velocity Specialist	FBv: 80th Percentile
Heath Hembree	Strikeout Specialist	K%: 75th Percentile
Ryan Hendrix	N/A	Not enough data to be included for analysis

The percentiles on the right-hand side of the graph are given from *Baseball Savant.* As we can see, the cluster results correlate well with the given percentiles. For example, the High Velocity Specialists all rank in the top 20 percent in fastball velocity. This is an indicator that the K-Means clustering analyzed successfully. However, we can see that the 2021 Reds bullpen may not be the strongest in terms of the new roles. It seems they have too many High Velocity Specialists and could use some Ground Ball Specialists and Chase Artists to diversify their bullpen. As mentioned earlier, further research is required to determine the best combination of clusters. For now, we can assume having a wide range of groups is essential for the best results on the field.

Overall, the results seem to be successful: using K-Means clustering on pitching data can classify relievers into new roles for better performance on the field. Hopefully, this area of research can grow and potentially be used in baseball one day. It would be fascinating to the see these results used in action.

References

1. Woltring, M. T., Rost, J. K., & Jubenville, C. B. (2018, October 25). Examining Perceptions of Baseball's Eras: A Statistical Comparison. Retrieved November 01, 2020, from https://thesportjournal.org/article/examining-perceptions-of-baseballs-eras/
2. Bodig, C. (2017, October 23). The History of Relief Pitching Part One: 1871-1945. Retrieved November 01, 2020, from https://www.cooperstowncred.com/the-history-of-relief-part-one-1871-1945/
3. Palmer, P. (2018, May 17). Admin. Retrieved December 03, 2020, from https://sabr.org/journal/article/relief-pitching-strategy-past-present-and-future/
4. Bodig, C. (2017, October 23). The History of Relief Pitching Part Two: 1946-1968. Retrieved November 02, 2020, from https://www.cooperstowncred.com/the-history-of-relief-pitching-part-two-1946-1968/
5. Tan, P., Steinbach, M., & Kumar, V. (2005). Introduction to data mining (Vol. 1). Boston, MA: Pearson Addison Wesley.
6. Starmer, J. (2018, May 23). StatQuest: K-means clustering. Retrieved December 03, 2020, from https://www.youtube.com/watch?v=4b5d3muPQmA
7. PCA, 3D Visualization, and Clustering in R. (2013, February 03). Retrieved December 03, 2020, from https://planspace.org/2013/02/03/pca-3d-visualization-and-clustering-in-r/
8. Hennig, C. (n.d.). Package fpc. Retrieved December 03, 2020, from https://cran.r-project.org/web/packages/fpc/index.html
9. Denoncourt, F. (1960, July 01). How to tell if data is "clustered" enough for clustering algorithms to produce meaningful results? Retrieved December 03, 2020, from https://stats.stackexchange.com/questions/11691/how-to-tell-if-data-is-clustered-enough-for-clustering-algorithms-to-produce-m
10. Schumaker, AJ. *PCA, 3D Visualization, and Clustering in R*, 3 Feb. 2013, planspace.org/2013/02/03/pca-3d-visualization-and-clustering-in-r/.
11. Gerlica, Jeffrey, et al. “Quantifying the Outfield Shift Using K-Means Clustering.” *ieworldconference.org*, 30 Apr. 2020, www.ieworldconference.org/content/WP2020/Papers/GDRKMCC_20_85.pdf.
12. Sarris, Eno. “Is a Pitcher's Ground-Ball Rate More Important in Today's Game?” *The Athletic*, The Athletic, 29 Aug. 2019, theathletic.com/1171344/2019/08/30/is-a-pitchers-ground-ball-rate-more-important-in-todays-game/.
13. Marcou, Charlie. “Investigating Major League Baseball Pitchers and Quality of Contact through Cluster Analysis.” *ScholarWorks@GVSU*, 30 Apr. 2020, scholarworks.gvsu.edu/honorsprojects/765/.
14. Zimmerman, Jeff. “Can Pitchers Prevent Solid Contact?” *The Hardball Times*, tht.fangraphs.com/can-pitchers-prevent-solid-contact/.
15. Cermak, Ben. “K% Vs. K/9: Finding an Effective Metric for Strikeout Rate.” *Statliners*, FanSided, 10 Nov. 2014, statliners.com/2014/11/10/k-vs-k9-finding-effective-metric-strikeout-rate/.

16. zPavlidis, H., Judge, J., Long, J., Jebens, K., Jebens, K., Davies-Stofka, B., . . . Prospectus Feature: Command and Control. Retrieved December 03, 2020, from https://www.baseballprospectus.com/news/article/31022/prospectus-feature-command-and-control/
17. Singh, D. (2019, November 06). Deepika Singh. Retrieved December 06, 2020, from https://www.pluralsight.com/guides/normalizing-data-r
18. LaMarco, N. (2019, February 05). How to Normalize in Excel. Retrieved December 06, 2020, from https://smallbusiness.chron.com/normalize-excel-36009.html
19. Becker, R. (n.d.). Scaling and Centering of Matrix-like Objects. Retrieved December 06, 2020, from https://stat.ethz.ch/R-manual/R-devel/library/base/html/scale.html
20. Manes, J. (2019, June 25). Hitting Statistics Correlations. Retrieved December 06, 2020, from https://baseballstats.jmanesaz.com/2019/06/09/hitting-statistic-correlations/
21. Starmer, J. (2018, July 26). Logistic Regression in R, Clearly Explained!!!! Retrieved December 06, 2020, from https://www.youtube.com/watch?v=C4N3_XJJ-jU
22. McGee, D. (2019, May 15). Ogden Raptors 2019 Attendance Case Study. Retrieved December 03, 2020, from https://dylanmcgeeportfolio.weebly.com/course-work.html
23. Academy, J. (2020, March 18). Create Training and Test data in R. Retrieved December 06, 2020, from https://www.youtube.com/watch?v=nXIzNAa4mFo
24. Starmer, J. (2018, October 29). Machine Learning Fundamentals: The Confusion Matrix. Retrieved December 06, 2020, from https://www.youtube.com/watch?v=Kdsp6soqA7o
25. Holtz, Y. (2018). 3D: The R Graph Gallery. Retrieved December 06, 2020, from https://www.r-graph-gallery.com/3d.html