Discovering Pitcher Types using Kmeans

TruMedia Hackathon 2015/2016

Jon Dickerson*

January 3, 2016

There are many ways to classify pitchers from where they pitch in the rotation ("a number 1", "5 starter"), to how they throw ("sidearmer", "submariner"), to pitching style ("strike thrower", "nibbling the edge"). These typically come from comparisons to pitchers of the past by baseball people with enough experience and intuition to make the judgements and connections in their memory banks. In this analysis, a computer is given its shot to do the same. The results are largely positive, with meaningful clusters emerging using only 6 variables, none of which are common performance indicators such as ERA, WHIP or W/L.

1 Introduction

Color commentators and sportswriters love phrases like "swing and miss stuff" and "pound the zone" when discussing pitchers. They also love cliches such as pitchers vs. throwers and "crafty veterans". But what are they talking about? What are they doing with these phrases? In essence, they are attempting to group pitchers into clusters, based on their innate abilities and tendencies. If there is one thing the recent advances in data science have taught us, it's that computers are much better at that; so why not give a computer a chance to do the same?

The aim of this paper is to first derive metrics for input into a K-means clustering algorithm (see Section 2), then examining each of the resulting clusters in turn in Section 3. Section 4 examines the results further, giving an overall evaluation of the algorithm's output, and potential improvements. All of the code used to generate this analysis is available for free use at http://www.github.com/jaydik/mlb-hackathon.

*Email: jonathan.d.dickerson@gmail.com

2 Methodology

The input data for this analysis is pitch-level data for the 2013-2015 seasons, including post-season. The variables include information about the players and game and importantly for this analysis, the pitch. Pitch variables include position of the ball as it crossed the plate, the release velocity, and pitch type, among others. These variables themselves could potentially be useful in a clustering algorithm, however, to fully take advantage of the clustering techniques, custom metrics will need to be derived. To avoid small sample-size problems, all pitchers with less than 10 appearances are excluded. Also, players are grouped by position, so we consider Bartolo Colon - SP separately from Bartolo Colon - RP. This is done as we often see pitchers slightly alter their style when coming out of the bullpen, at times increasing their velocity and relying more on the fastball (See also: Phil Hughes, Joba Chamberlain).

Trying to mimic how pitchers are colloquially discussed, the variables are chosen to go after certain aspects of a pitcher's abilities and tendencies. The final set of variables used is innings pitcher per appearance, fastball percentage, average fastball speed, differential between average fastball speed and average breaking pitch speed, aggression, and percentage of swings that were swings and misses. Each of these variables is discussed in turn.

2.1 IP/App

One of the most straight-forward variables is innings pitched per appearance. The goal of this variable is two-fold. Obviously, the elite pitchers throw a lot of innings, so we want an opportunity to segment those off. However, hot-stove conversations often talk about an "innings-eater", a pitcher who can give you 200+ innings every year, despite maybe not being in that elite class. Both of those theoretical groups should score highly in this variable.

Pitcher	IP/App
Clayton Kershaw	7.02
David Price	6.93
Adam Wainwright	6.87
Cliff Lee	6.81
Chris Sale	6.74
Max Scherzer	6.69

Table 1: Innings Pitched per Appearance Leaders

No real surprises given in Table 1, where we see the top 6 pitchers in IP/App. These will all fall into our theoretical elite group (or so we hope).

2.2 FB%

The next variable is fastball percentage. This is a simple metric derived using the following formula

$$FB\% = \frac{fastballs}{totalPitches}$$

Where $fastballs$ includes the following game day labels: fastball, two-seam fastball, four-seam fastball, cutter, sinker. This variable is a simple usage stat, how often does a given pitcher use his fastball? Conversely, how often does he go offspeed? The FB% leaders are given in Table 2.

Pitcher	FB%
Mariano Rivera	1.00
Jason Motte	0.97
Mitch Harris	0.95
Jake McGee	0.94
Nick Christiani	0.93
Kenley Jansen	0.93

Table 2: FB% Leaders

Again, no real surprises, as we often talk of relievers being one or two pitch pitchers, primarily relying on their fastball.

2.3 MPH

Average fastball speed is again a very clear metric. The speed of the pitcher's fastball (defined the same as above) is averaged first over game then over years. The purpose of this variable is clear: how hard can you throw the ball? Baseball literature is chock-full of references to "flamethrowers", "hard-throwing lefties" and the like, so this is an obvious variable to include in the analysis. Table 3 gives the leaders in MPH, with a few usual suspects, and a few hard-throwing guys without much service time (but more than the 10-appearance cutoff).

Pitcher	MPH
Aroldis Chapman	99.23
Bruce Rondon	98.35
Erik Cordier	98.34
Kelvin Herrera	97.78
Nate Jones	97.71
Arquimedes Caminero	97.25

Table 3: MPH Leaders

2.4 DIFF

Differential is calculated as follows

$$DIFF = avg(fastball_{mph}) - avg(nonFastball_{mph})$$

Where both fastball and non-fastball pitches were averaged the same as fastballs in section 2.3. The goal of this variable is to measure how well the offspeed pitches keep the hitter off-

balance, via the speed difference. There are potential problems and improvements with this variable, to be discussed in Section 4. The leaders in DIFF are given in Table 4

Pitcher	DIFF
Evan Scribner	16.91
Drew VerHagen	16.10
Mike Morin	15.97
Mike Foltynewicz	15.94
Steve Johnson	15.72
Brian Ellington	15.22

Table 4: DIFF Leaders

Here are some non-household names. A quick tour of Brooks Baseball[1] for these guys shows they all feature slow curves and/or changeups in the 75-80mph range, with decent fastballs in the 90-93mph range, so it's about what we expect.

2.5 AGG

This is perhaps the fuzziest of the variables, and the toughest to describe. When we call a pitcher aggressive, often we mean big fastball, throwing right down the middle, saying "my best vs. your best". How do you define that though?

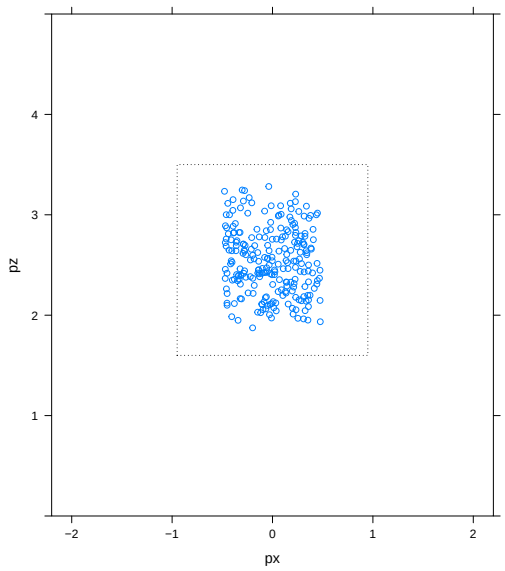
Pardon a brief technical aside. There are 4 important variables for this: `px`, `pz`, `szb`, and `sbt`. `px` describes the horizontal location of the ball relative to the center of the plate, e.g., a pitch with `px` = 0 would literally be "right down the middle" (with respect to the horizontal plane). Horizontal is easy, I give the pitchers 2 baseball's width deviation from dead center to be labeled down the middle. Vertically is a bit tougher. The top and bottom of the strike zone (supposedly) are determined batter by batter. Those are meant to be captured in the `szt` and `szb` variables, respectively. Lastly, `pz` is the height of the ball above the ground. Thus, I can use `szb`, `szt`, and `pz` to calculate "down the middle" for a pitches' vertical location. After all is said and done, what I get is pitches labeled "aggressive" as shown in Figure 1. Also plotted is the rulebook strikezone, in dashed line. The data used in the plot is taken from the 2015 World Series between the Royals and Mets.

Though imperfect, I'm pleased with the results of the metric, and it properly gives a sense of a pitcher who throws largely down the middle vs. one that pitches more to the corners. The leaders in aggression are given in Table 5

Perhaps unsurprisingly, not many big time pitchers in there. There are a lot of relievers barely making the cutoff for appearances on the leaderboard.

[1] http://www.brooksbaseball.net/

Figure 1: Aggressive Pitches



Pitcher	AGG
Hansel Robles	0.25
Drew Pomeranz	0.24
Kyle Crockett	0.24
Felipe Rivero	0.23
Paul Clemens	0.23
Mike Adams	0.23

Table 5: AGG Leaders

2.6 WHIFF%

The last variable in this analysis is whiff percentage. It is calculated as

$$WHIFF = \frac{swingingstrikes}{inducedswings}$$

In other words, of all the times the batter swung the bat, how many times did he miss the ball completely? It may be a bit too on the nose, but this is my way of measuring pitchers who have "swing-and-miss stuff". The calculation is pretty straight forward, and the leaders are given in Table 6.

Look out AL East,the Yankee bullpen now boasts the top two swing and miss pitchers in baseball. Like most of these tables, there aren't any real surprises here, relievers coming in throwing hard and few pitches results in large strikeout rate statistics, including WHIFF%.

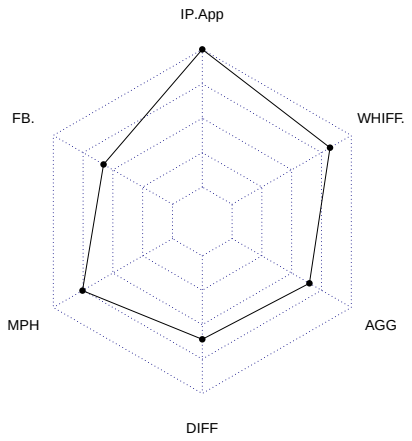
Pitcher	WHIFF%
Aroldis Chapman	0.40
Andrew Miller	0.37
Sergio Santos	0.36
Koji Uehara	0.35
Greg Holland	0.35
Will Smith	0.34

Table 6: WHIFF Leaders

2.7 Displaying Results

To display the results, and compare pitchers and clusters, we need a way to visualize these statistics. I use the `fmsb` package and `radarchart` function to do so. An example plot is given for Clayton Kershaw in Figure 2. We can see why he's effective. He throws the most innings per appearance of any pitcher, coupling that with a big time fastball and above average differential. He is more aggressive than your average pitcher, but generates a ton of swings and misses. He does so by mixing in his secondary pitches and only relying on his fastball slightly more than your average pitcher.

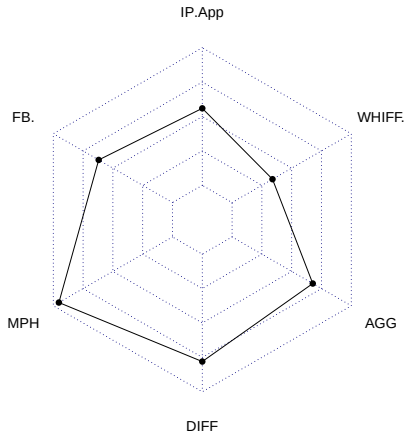
Figure 2: Clayton Kershaw Attributes



Let's compare that to Nathan Eovaldi, who throws hard but doesn't have Kershaw's results or pedigree. In his graph, we can see a plus plus fastball, which he uses a lot, but he only

throws an average number of innings, mainly because he gives up too many hits, as we can see in his below average WHIFF metric. He's slightly more aggresive than Kershaw, perhaps resulting in the low swing and miss abilities he seems to possess.

Figure 3: Nathan Eovaldi Attributes



2.8 Clustering

A sample of the resulting dataset (ordered by descending appearances) is given in Table 7. The dataset is first split into relievers and starters, then run through a K-means clustering algorithm with $k = 3$ and $k = 5$ for relievers and starters respectively. All variables except appearances are used.

Pitcher	Pos	IP/App	FB%	MPH	DIFF	AGG	WHIFF%	App
Seth Maness	RP	0.78	0.71	89.73	6.98	0.19	0.16	232
Trevor Rosenthal	RP	0.92	0.79	97.09	10.77	0.18	0.27	230
Mark Melancon	RP	0.91	0.77	91.66	10.18	0.18	0.23	228
Tony Watson	RP	0.92	0.71	93.94	8.84	0.17	0.22	227
Tyler Clippard	RP	0.92	0.55	91.09	11.62	0.17	0.26	227
Bryan Shaw	RP	0.85	0.76	92.54	11.74	0.21	0.22	225

Table 7: Cleaned Dataset

The results of the algorithm are given in the next section.

3 Results

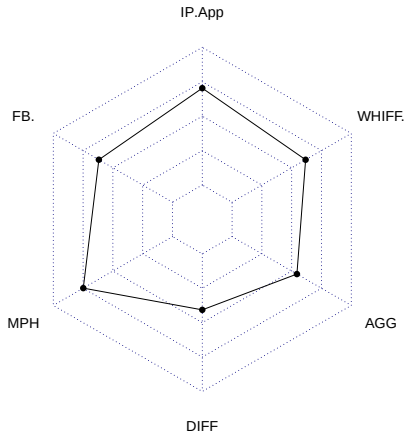
Finally, without further adieu, we can examine the output of the clustering algorithm. I go over each cluster, giving the top 10 pitchers (ordered by appearances) in each cluster, along with their average radar chart, constructed by taking their average scores for each metric.

3.1 Aces

Naturally, we begin with the best of the best, the Aces cluster. As we can see, they throw a ton of innings, and miss a ton of bats, using their above average fastballs. They are more aggressive than your average pitcher. These guys are the ones you want at the front of your rotation. What's interesting is not so much that this cluster contains Bumgarner, Scherzer, and Kershaw, but the Samardzija and Quintana inclusions can give you pause. When you look at Samardzija and Quintana against the others by the given metrics, they actually stack up favorably. Included in this cluster but outside the top 10, are other Aces such as Jose Fernandez, Masahiro Tanaka, Yu Darvish, Sonny Gray, Gerrit Cole, among many others.

It's interesting to see that the elite pitchers were grouped together by these metrics. However, with the inclusion of some tier 2 and 3 pitchers, it's clear that these are not sufficient conditions for elite pitcher status.

Figure 4: Ace Attributes



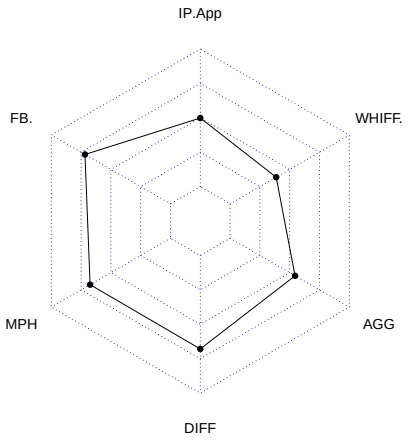
Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Madison Bumgarner	6.63	0.45	91.86	7.87	0.18	0.23
Max Scherzer	6.69	0.57	93.32	9.81	0.20	0.26
Clayton Kershaw	7.02	0.57	93.08	10.84	0.19	0.27
David Price	6.93	0.71	92.40	9.35	0.20	0.21
Jordan Zimmermann	6.26	0.66	93.52	9.14	0.20	0.18
Zack Greinke	6.45	0.60	91.31	8.98	0.15	0.24
Cole Hamels	6.63	0.67	91.28	9.11	0.17	0.25
Jeff Samardzija	6.49	0.65	93.82	8.95	0.18	0.21
Jose Quintana	6.09	0.60	91.39	9.18	0.18	0.19
Julio Teheran	6.18	0.63	90.90	11.94	0.18	0.22

Table 8: Top 10 Aces (by appearances)

3.2 Almost-Aces

The next group, is what I call "Almost-Aces". They have above average fastballs, and they use them more than average, with a good arsenal of offspeed pitches, but they get below average whiffs, and pound the zone a bit more than average. What is interesting about this group compared to the Ace group is that this group has significantly higher offspeed differentials, but pitch significantly fewer innings on average. Additionally, they have lower whiff rates. This suggests they pitch to contact a bit more, and luck being what it is, a certain percentage of these hit balls fall, leading to both their reduced performance and inning counts.

Figure 5: Almost Ace Attributes



Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Jon Lester	6.48	0.79	90.92	12.78	0.16	0.21
Edinson Volquez	5.63	0.54	93.15	11.29	0.18	0.20
John Lackey	6.34	0.87	89.63	10.10	0.19	0.20
Chris Tillman	5.79	0.64	91.23	11.16	0.18	0.16
Shelby Miller	5.69	0.81	92.87	13.51	0.20	0.18
Scott Kazmir	5.65	0.63	91.32	11.58	0.19	0.20
Gio Gonzalez	5.72	0.67	92.13	11.69	0.15	0.22
Rick Porcello	6.09	0.62	90.86	10.07	0.18	0.18
Wei-Yin Chen	5.84	0.66	91.48	10.78	0.20	0.17
Phil Hughes	5.77	0.76	91.17	13.34	0.22	0.14

Table 9: Top 10 Almost Aces (by appearances)

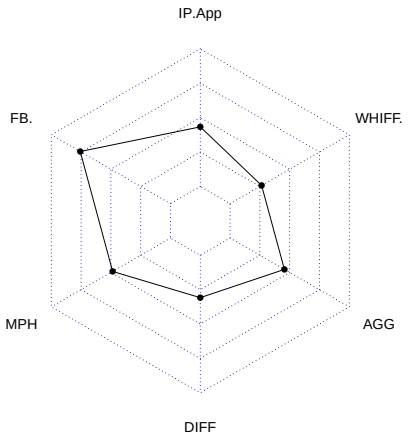
3.3 Go Ahead, Hit it

This is an interesting cluster. They use their fastball significantly more than average, but it is about average in terms of velocity. They're about average in terms of aggression. They don't generate many swings and misses, prompting me to give them the name I did. Due to the high amount of contact they induce, they pitch slightly fewer innings than average. Some of these guys are borderline belonging to other groups, such as Mark Buehrle, Who pitches more innings and throws the fastball less often and less hard than a lot of the pitchers in this segment.

Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Lance Lynn	5.85	0.78	92.24	8.34	0.17	0.20
Mark Buehrle	6.10	0.66	82.74	6.10	0.16	0.14
Dan Haren	5.63	0.77	85.97	5.74	0.16	0.16
Mike Leake	6.22	0.69	90.11	9.00	0.17	0.15
Bartolo Colon	6.21	0.83	88.93	6.75	0.21	0.13
Kyle Kendrick	5.72	0.73	88.54	7.79	0.17	0.15
Aaron Harang	5.77	0.65	89.18	9.31	0.17	0.16
Jake Peavy	5.86	0.74	89.04	7.53	0.18	0.18
Tim Hudson	5.78	0.78	87.36	8.66	0.17	0.19
Doug Fister	6.22	0.72	87.21	10.93	0.18	0.15

Table 10: Top 10 Go Ahead, Hit it Pitchers (by Appearances)

Figure 6: Go Ahead, Hit it Attributes



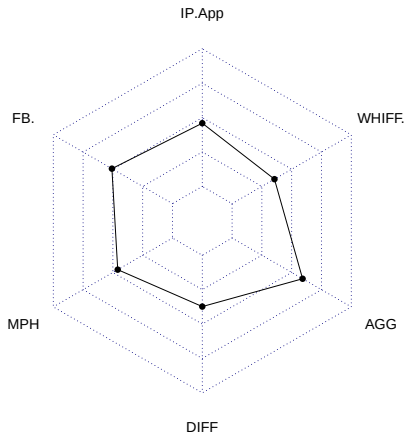
3.4 Aggressively Average

The attribute that stands out most in this otherwise average group is aggression, where they're anything but average. They are the most aggressive group including relievers, which drags down their whiff percentage slightly. Aside from those two, they are a decidely average group. Algorithmically, I think this may just be the group that you are given if you don't fit anywhere else, leading to a very average result.

Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
R.A. Dickey	6.27	0.13	81.54	5.80	0.19	0.20
Jeremy Guthrie	5.93	0.53	91.95	8.97	0.20	0.14
Jered Weaver	6.10	0.48	85.45	10.57	0.17	0.21
John Danks	5.91	0.62	87.87	9.16	0.19	0.18
Hisashi Iwakuma	6.36	0.53	88.85	7.26	0.18	0.20
Ricky Nolasco	5.59	0.47	90.19	11.06	0.18	0.21
Colby Lewis	5.90	0.53	88.37	6.46	0.18	0.17
Kevin Correia	5.57	0.42	89.96	5.50	0.16	0.13
Dillon Gee	5.96	0.56	89.19	8.74	0.19	0.18
Mike Minor	6.03	0.59	90.38	7.45	0.19	0.18

Table 11: Top 10 Aggressively Average Pitchers (by appearances)

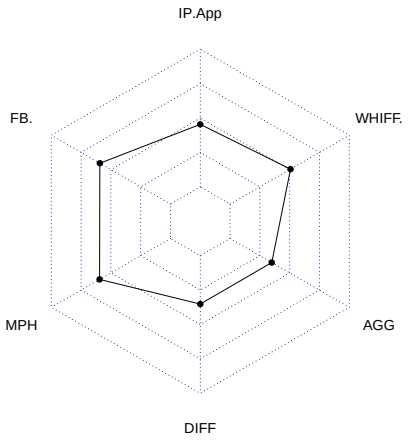
Figure 7: Aggressively Average Attributes



3.5 Painting Corners

This group is loaded with pitchers that don't have overwhelming stuff, but are known for very good control. They use their fastball quite a bit, and it's a solid fastball, slightly above average in mph. However, they are much less frequently found throwing it down the middle. It would have been interesting to see if there was a strike-looking equivalent to our WHIFF stat, how this cluster would score because I would conjecture that they get more than average amounts of strikes without swings. James Shields is another of the barely-in guys, as he sometimes ends up in the Aces cluster depending on the size of the clusters. Many of these pitchers have had big league success, so it goes to show that you can win with good control over lesser "stuff".

Figure 8: Painting Corners Attributes



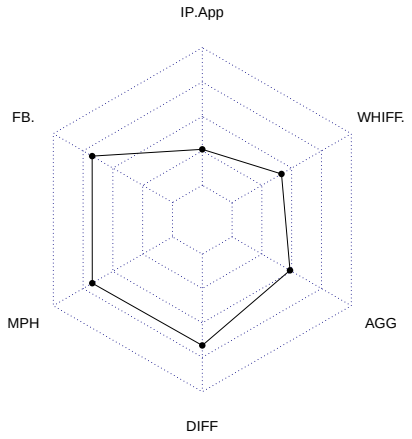
Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
James Shields	6.32	0.63	90.14	7.54	0.15	0.23
Wade Miley	5.92	0.64	90.94	7.99	0.16	0.20
Yovani Gallardo	5.64	0.53	90.83	7.00	0.15	0.16
Ian Kennedy	5.71	0.64	90.95	10.32	0.16	0.22
Francisco Liriano	5.85	0.44	92.62	7.02	0.13	0.31
Jorge De La Rosa	5.57	0.60	90.25	8.46	0.16	0.22
C.J. Wilson	5.85	0.61	90.38	9.17	0.17	0.20
Dallas Keuchel	6.54	0.64	89.20	10.16	0.14	0.22
Ubaldo Jimenez	5.49	0.59	90.91	8.26	0.17	0.19
Jason Hammel	5.52	0.58	92.36	9.55	0.17	0.20

Table 12: Top 10 Painting Corners Pitchers (by appearances)

3.6 Lights Out Closers

These guys are exactly that. They come in throwing hard, pumping in a lot of fastballs, but you can't sit dead-red, because they have an offspeed pitch or two that has significant speed difference. Interestingly, they don't have super high whiff rates – slightly below average even – suggesting maybe that their high fastball usage results in a lot of foul balls and balls in play. Since these are sorted by appearances, they tend to favor middle relievers who get used more often, as closers tend to have fewer appearances. That said, also in this cluster is Aroldis Chapman, Dellin Betances, Craig Kimbrel, and other elite closers.

Figure 9: Lights Out Closers Attributes



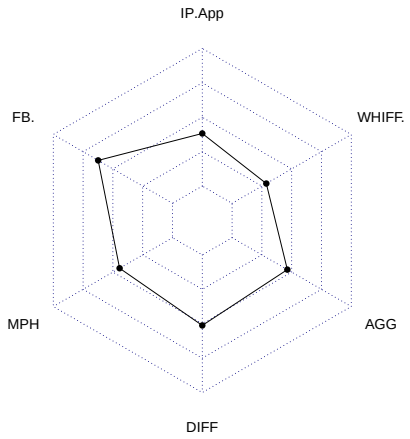
Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Trevor Rosenthal	0.92	0.79	97.09	10.77	0.18	0.27
Mark Melancon	0.91	0.77	91.66	10.18	0.18	0.23
Tony Watson	0.92	0.71	93.94	8.84	0.17	0.22
Bryan Shaw	0.85	0.76	92.54	11.74	0.21	0.22
Cody Allen	0.85	0.65	95.21	9.81	0.17	0.30
Kelvin Herrera	0.96	0.77	97.78	10.98	0.18	0.25
Junichi Tazawa	0.83	0.60	93.62	10.37	0.19	0.23
Luis Avilan	0.60	0.77	93.12	14.55	0.16	0.19
Pedro Strop	0.81	0.58	95.17	12.12	0.17	0.31
Fernando Rodney	0.88	0.65	95.22	12.20	0.17	0.26

Table 13: Top 10 Lights Out Closers Pitchers (by appearances)

3.7 Contact Relievers

These pitchers use their fastballs predominantly, don't get a ton of strikeouts, but tend to pitch longer in games. A lot of middle relievers here dragging that average IP/app up, but some notables here also include Jonathan Papelbon and Mariano Rivera.

Figure 10: Contact Reliever Attributes



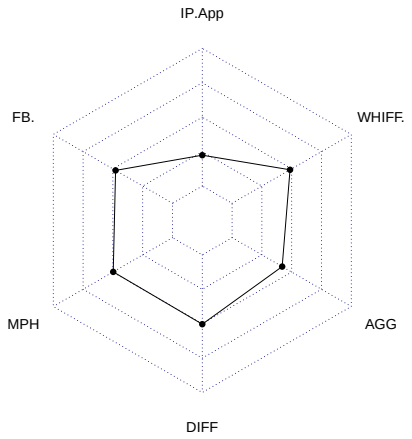
Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Seth Maness	0.78	0.71	89.73	6.98	0.19	0.16
Joe Smith	0.85	0.70	88.85	8.27	0.18	0.17
Brad Ziegler	0.89	0.68	84.86	9.10	0.14	0.20
Burke Badenhop	0.87	0.70	89.10	7.11	0.21	0.11
Addison Reed	0.95	0.69	92.59	7.44	0.21	0.21
Jonathan Papelbon	0.98	0.69	91.43	8.49	0.19	0.23
Craig Breslow	0.85	0.62	89.22	7.64	0.16	0.16
LaTroy Hawkins	0.85	0.74	92.81	7.50	0.20	0.14
Matt Belisle	0.87	0.61	90.78	7.89	0.20	0.19
Jared Hughes	0.83	0.85	92.48	6.74	0.11	0.18

Table 14: Top 10 Contact Relievers (by appearances)

3.8 Average Relievers

Much like in Section 3.4, it seems if you don't fit anywhere else, you fit here. With almost exactly average statistics across the board. The actual names have a healthy mix of long men, LOOGYs (Lefty One Out GuY), and setup men. There isn't much here that differentiates them collectively. Given the small bodies of work they have, relievers just may be harder to cluster, as will be discussed in Section 4.

Figure 11: Average Reliever Attributes



Pitcher	IP/App	FB%	MPH	DIFF	AGG	WHIFF%
Tyler Clippard	0.92	0.55	91.09	11.62	0.17	0.26
Mike Dunn	0.68	0.59	94.57	7.60	0.21	0.27
Luke Gregerson	0.86	0.46	88.66	7.61	0.18	0.29
J.P. Howell	0.61	0.62	86.96	7.12	0.15	0.24
Randy Choate	0.27	0.67	84.58	9.61	0.16	0.22
Casey Fien	0.81	0.52	92.09	6.14	0.20	0.22
Darren O'Day	0.85	0.53	86.62	7.45	0.16	0.26
Sergio Romo	0.78	0.40	87.53	9.01	0.16	0.31
A.J. Ramos	0.94	0.54	92.42	10.32	0.17	0.31
Al Alburquerque	0.72	0.38	93.48	7.86	0.18	0.30

Table 15: Average Reliever Pitchers

4 Conclusion

Given as input 6 calculated metrics, none of which directly measure performance (meaning no ERA, WHIP, etc.), the algorithm was able to make fairly intelligent groups of pitchers. The algorithm performed much better with starters than relievers, mainly because the metrics were chosen with the intent of distinguishing between groups of starters. In future work, it may be best to consider starters and relievers separate entirely, and run clusters on different sets of metrics, to help segment the relievers better. The choices of k are also arbitrary, but were chosen after some experimentation with various levels. 5 and 3 seemed to provide the most meaningful groups.

With innings pitched per appearance, the goal was to get after innings eaters, and I think it was fairly successful. For relievers however, it wasn't a particularly useful variable, as it seemed to barely even get long-men into their own group. Perhaps percentage of one-out outings to help identify LOOGYs? Fastball percentage was meant to be a proxy for repetoire, but I'm not sure how well it did that. It would be very cool to capture repetoire in a variable more precisely, perhaps in a vector with components for each pitch type, 1 if the pitcher throws that pitch, 0 otherwise would work. On a related note, differential was calculated by taking average fastball velocity minus average offspeed velocity, but offspeed included every non-fastball pitch. More informative would be average velocity of fastest pitch in repetoire - average velocity of slowest pitch in repetoire. For example, a pitcher with a fastball slider curve change, with the curve at 75 mph and the slider and change at 87 and 85 respectively, would have the differential dragged down by the relatively fast slider and changeups.

Aggression was chosen somewhat arbitrarily as well, with pitchers who nibble at the corners and those that have no control of their pitch whatsoever being penalized equally. Another metric of "on the black", where the pitch is within a ball's width of the edge of the strikezone, would help to give a boost to those who are able to consistently hit the edge. Strike percentage could also be useful, the proportion of total pitches thrown which are strikes. I purposely avoided total pitches thrown per appearance, but perhaps it could be included to further refine the clusters. Most of the variable limiting decisions were made because of the desire to graphically display the results; many dimensions makes that difficult. If the graphical constraint is lifted, the inclusion of more variables, including those discussed above, could help refine the segments.