Stat 208 Final Project

Ryan Hamlett

June 8, 2018

Introduction

One of the hallmarks of professional baseball are the pitchers’ ability to throw many different pitches effectively. In fact, some pitchers throw different assortments of what are generally considered to be the same pitch, such as the different assortments of fastball (cut, two-seam, four-seam, and others). Different pitchers also tend to throw different types of pitches with different grips. All of this can seem a bit complicated and cumbersome for those who are not well-versed in baseball’s pitching subculture. But one might wonder, are all of these pitches *actually* different?

As a huge baseball fan, I have wondered if it might be more prudent to classify pitches by spin, speed, and movement rather than simply by how the pitcher holds the ball when he throws it. For many avid baseball watchers, some pitchers throw a cut-fastball, some throw a slider. Both pitches, however, are generally in the range of 88-92 mph and move down and away from a right handed batter at roughly the same angle. In reality, it seems as if the only difference in these pitches is the name being given to them. To the hitter, viewer, and even most pitchers, the pitches are virtually identical. Simplifying classification to call both of these pitches a slider (or cut-fastbal or any other name) would seem to make life easier for casual fans and would likely do little to reduce the information they receive about the pitch being thrown.

Data Collection

All data has been collected using Major League Baseball’s Statcast database, which records a number of interesting variables on every single pitch. These variables include “Spin Rate”, “Pitch Velocity”, “pf_x”, “pf_z”, “pitch type”, and many others. The data pulled from the statcast database for this project is every pitch from the 2018 season through 5/23 this consists of n = 217,521 total pitches. By necessity, since we are using pitch movement as a clustering feature, I needed to limit the pitches to right-handed pitchers, though we could have chosen left-handed pitchers instead.. I do not think that we lose much, if any, in terms of limiting the dataset this way. I then simplified the dataset by removing pitches titled “null,”UN“,”EP“, and”PO" since “null” and “UN” are both indicators that the pitch classification algorithm felt this pitch was unlike any pitch it recognizes, “EP” is an “eephus” pitch which is thrown about 10 times a year, and “PO, meaning pickoff, which isn’t a real pitch. I then reduced the dataset further by making sure that every pitch included had a recorded”pitch velocity“,”pf_x“,”pf_z“, and”spin rate" value. Ultimately, this resulted in a dataset of n = 152,509 pitches.

One worry I had was that some pitchers throw around 100 mph while others throw in the range of 90 mph. The changeup for a 100 mph pitcher is usually in the range of 88-90 mph. If we simply use velocity as a feature, we may run the risk of having a changeup and a fastball classified as the same pitch. While you could argue that a 100 mph fastball is a different pitch than a 90 mph fastball, I found it more intuitive to simply transform the velocity column into “velocity minus maximim velocity” where I found the maximum velocity for each pitcher in 2018 and subtracted each pitch from that value. This gives us a better measure of the relative velocity differences in pitches which I think is more valuable than pure velocity.

The variables “pf_x” and “pf_z” are considered “horizontal movement” and “vertical movement” respectively. Both values are measured relative to a theoretical pitch with no spin-related movement. For example, “vertical movement”" would be equal to 0 on a pitch that simply dropped the amount we would expect it to drop due to gravity alone.

In order for the clustering algorithm to work effectively, we also need to scale the data since Euclidean distances are being calculated. Obviously, since we are running a clustering algorithm, we scale this dataset without the pitch classifications but we will hold onto the original classifications for later in order to analyze the sucess of our clustering algorithm. All of the work on the data set is included in the code in the appendix.

Methods

For this project, I have used what is likely the simplest clustering algorithm, the k-means clustering technique. In the k-means clustering algorithm, we first center and scale the data. We then pick k starting values as our initialized cluster centers. Each iteration through the algorithm, we calculate the Euclidean distance from all k centers for each point in our dataset. We assign each point to the cluster where the calculated Euclidean distance is smallest. We then redefine the k centers as the mean of each of the k clusters defined in the previous step. These last two steps are repeated until convergence.

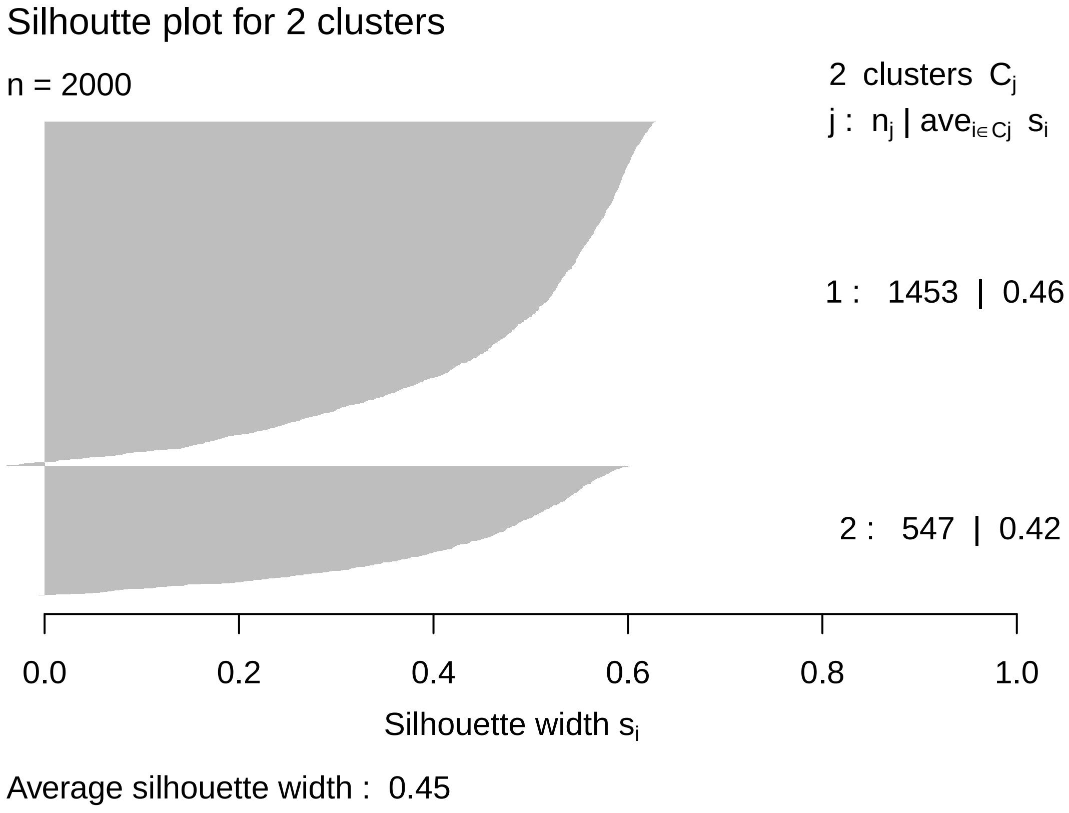
While there are many possible clustering algorithms that we may use for smaller datasets, many of these algorithms require the construction of a “distance matrix” that increases exponentially in size as we increase n. For a dataset with n = 152,509, the required amount needed to hold the distance matrix was 86.6 GB. I considered choosing a subset of my data to analyze, possibly a week or two, but the dataset would still be massive and without at least a month’s worth of data, we would be selectively sampling pitches that come from starters much more than that of relievers since relievers may pitch quite a bit less in any given two-week period. Thus, I felt it was best to simply run a k-means clustering algorithm with k ranging from 2 all the way to 12. The original number of pitch classifications was 12, as will be seen in the tables in the next section. Since our goal is to possibly reduce the number of pitch clusters, I decided not to go larger than k = 12.

For each individual clustering algorithm, several things are calculated. First, I have computed silhouette diagrams for all 11 different clustering algorithms. Since for each clustering algorithm, a distance matrix is required, I have sampled 2000 data points randomly from the original dataset, with the proportion of each cluster in the 2000 randomly selected points equal to the proportion of each cluster in the overall dataset. This method is prone to error, so I have done this 20 times for each cluster and averaged the average silhoutte length to hopefully get a more stable estimate of average silhoutte length. The plots, however, are simply the plot of the 20th silhouette for each cluster and are available as a visual guide of what happens in run through each clustering algorithm. The second thing I have computed is the ratio of within-cluster sum of squares to total sum of squares for each clustering algorithm.

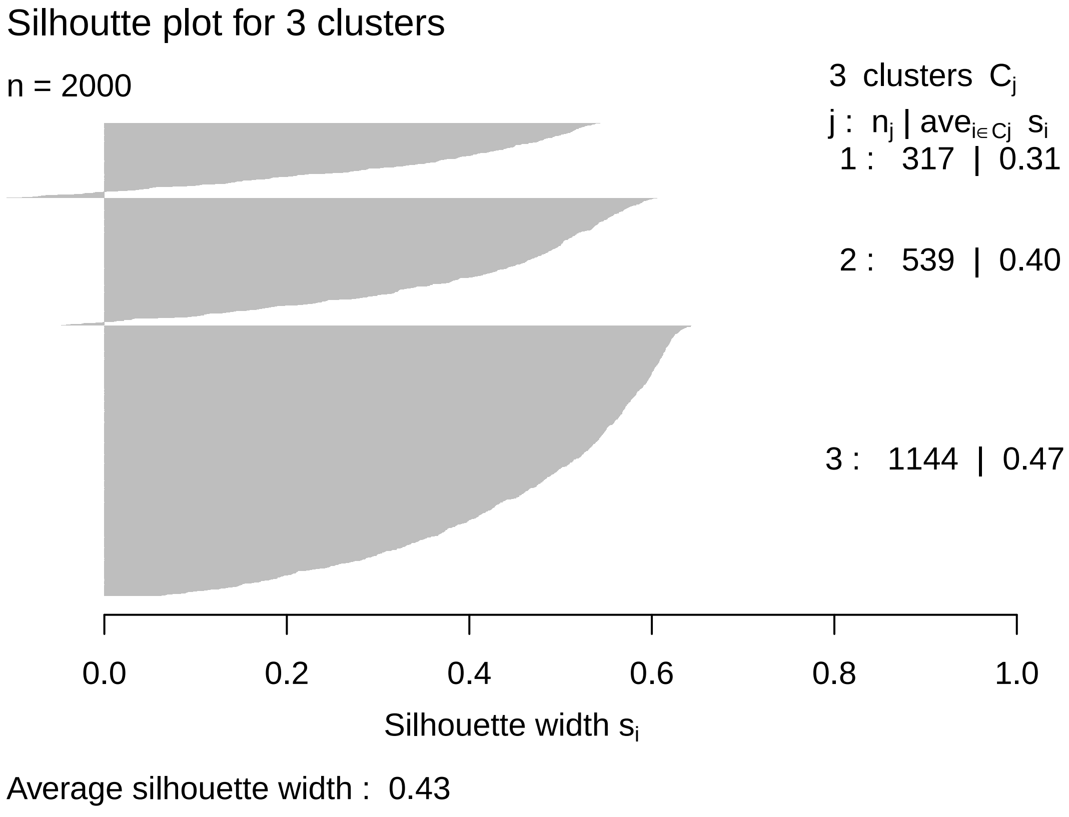
The reason I have calculated both of these is that I want to choose the appropriate number of k according to both of these criterion functions later on.

Results

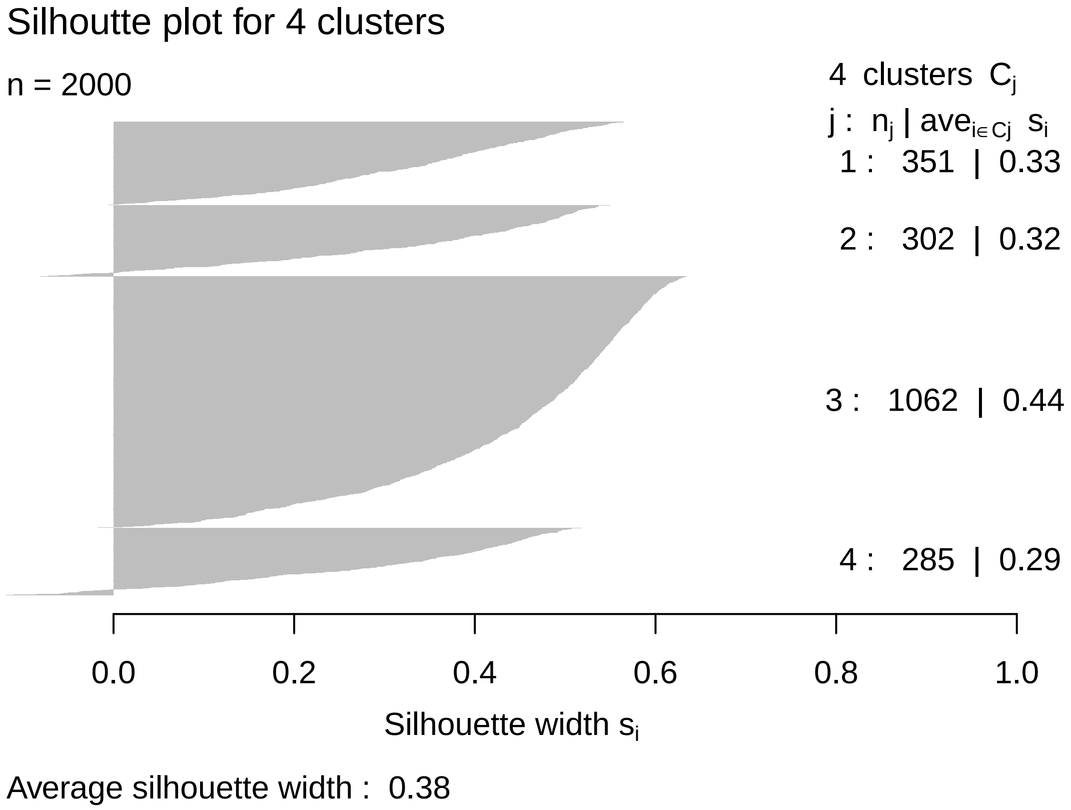
On the next few pages, you will see one of the silhoutte plots calculated for each of the clustering algorithms from k = 2 to k = 11. Directly after these silhouette plots, you will see a table that shows the mean average silhouette length for each k. One popular way of selecting k for clustering algorithms is choosing k to be the number of centers that produces the smallest average silhouette length. Once again, as was stated previously, these average silhouette lengths are prone to some error since we are only sampling a random subset of each cluster for each algorithm. Thus, I have calculated the mean average silhouette length over m = 20 different random subsets of the data in the hopes of reducing the variability of this estimate.

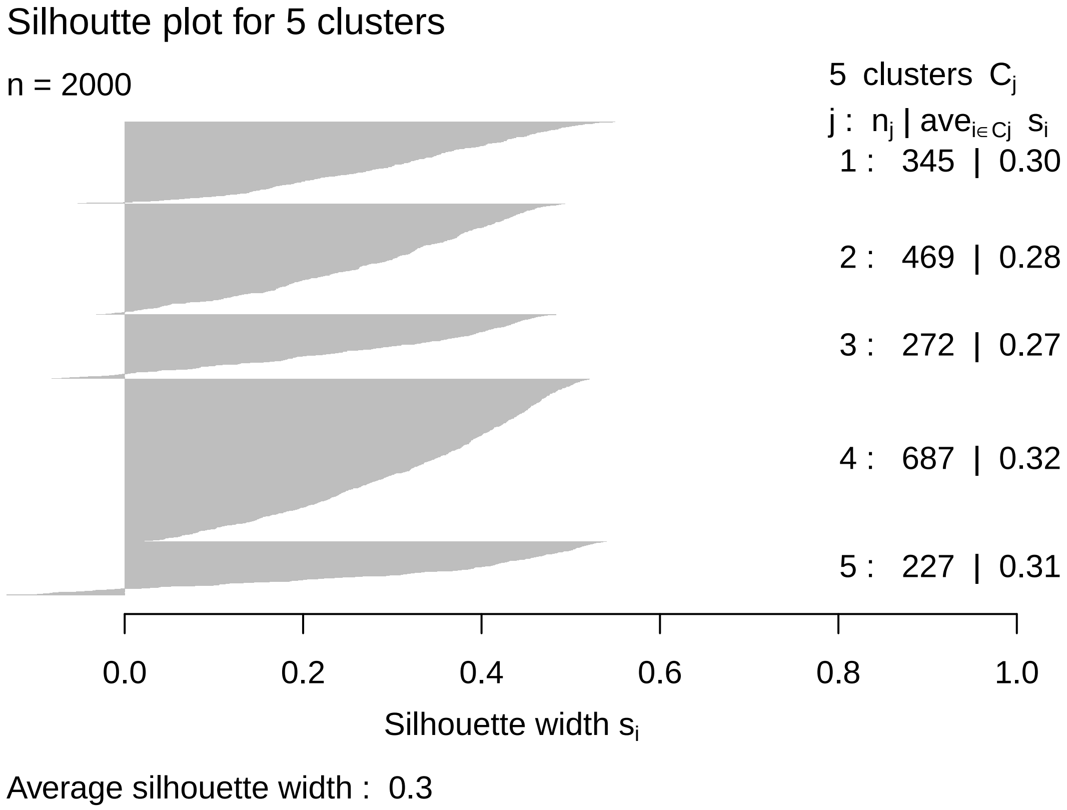


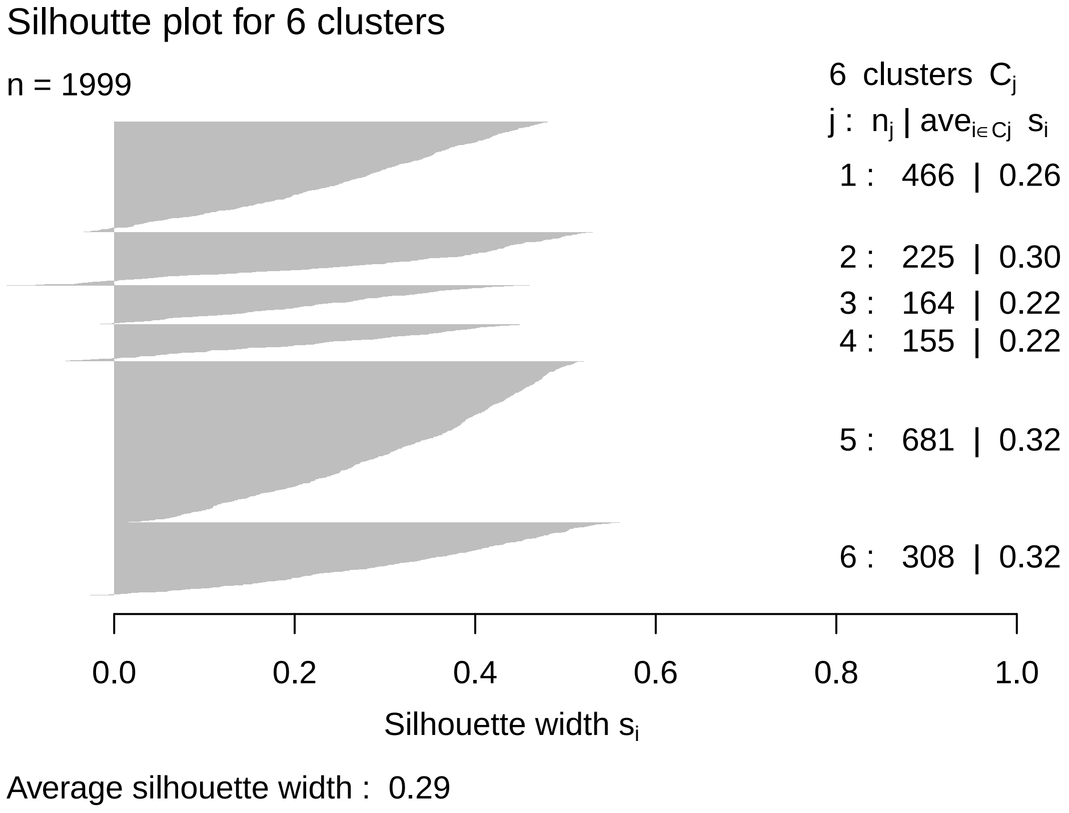
Silhoutte plot for 2 clusters

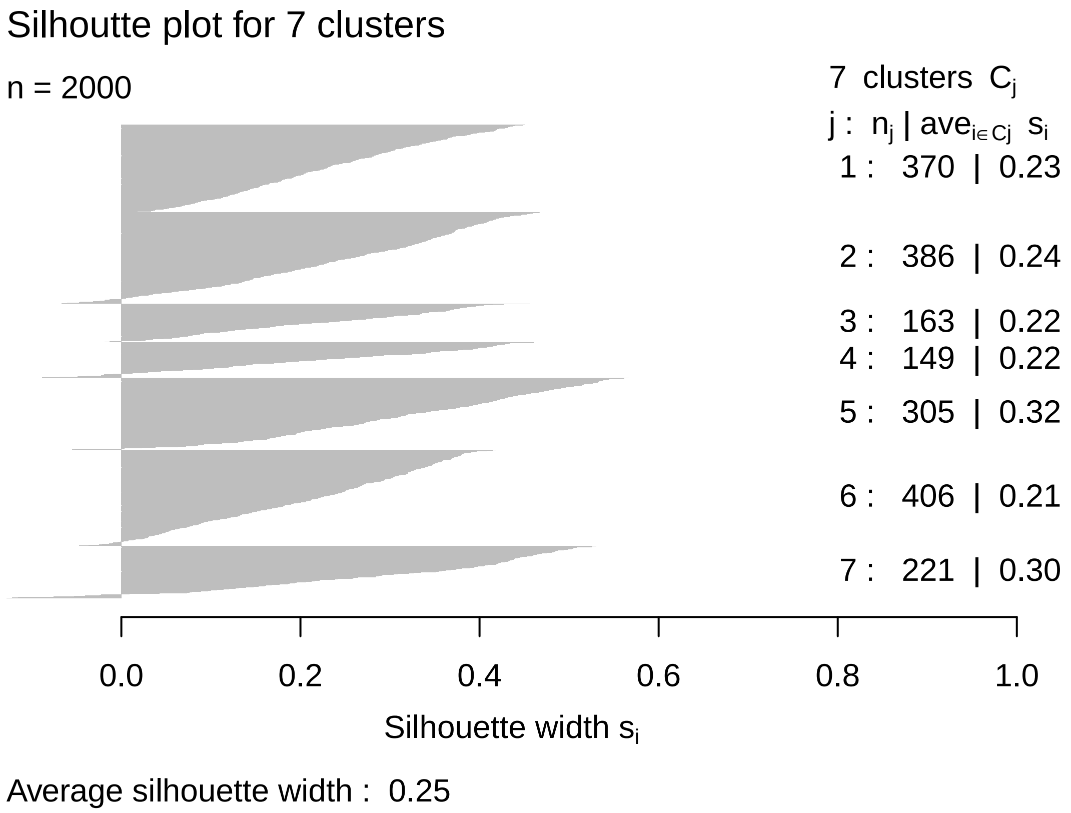


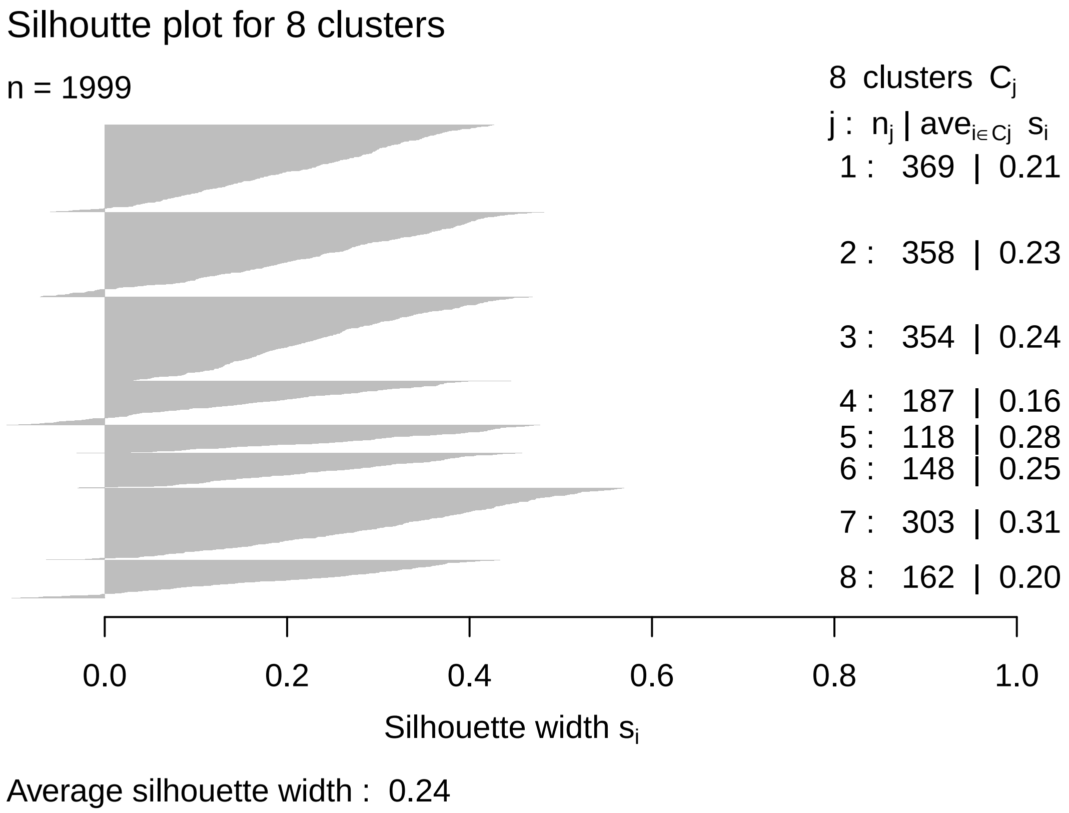
Silhoutte plot for 3 clusters

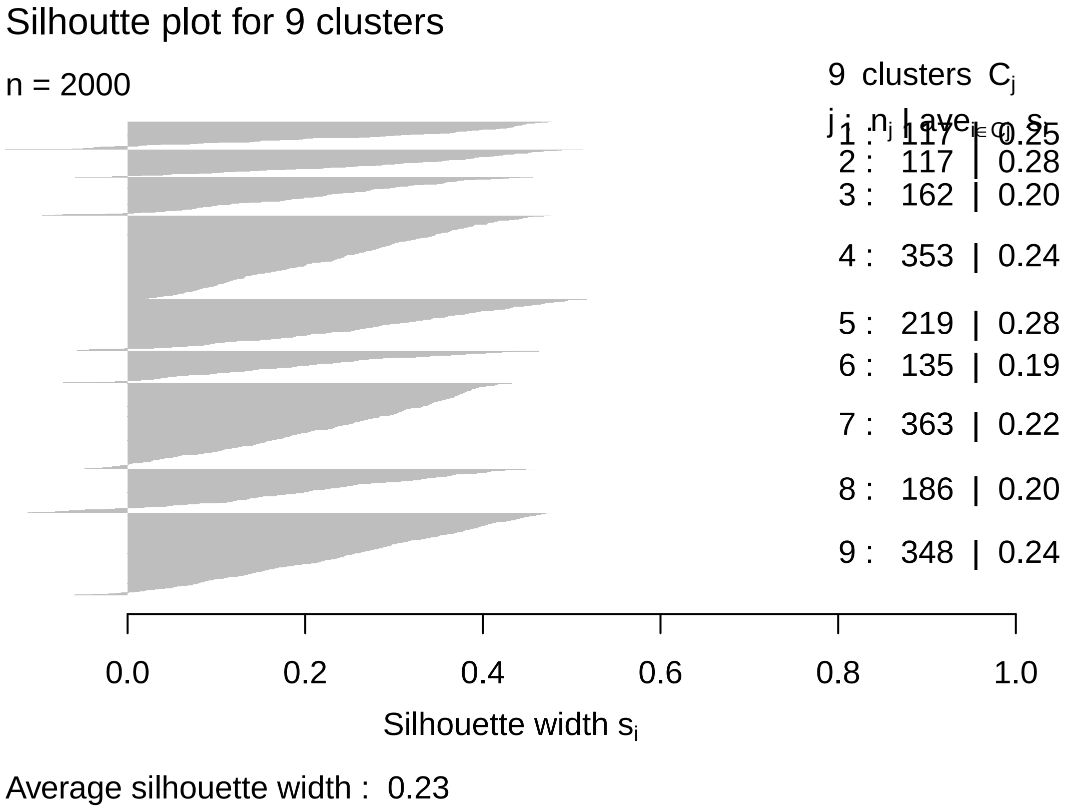

Silhoutte plot for 4 clusters
n = 2000
4 clusters C_j
j : n_j | $\text{ave}_{i \in Cj}$ s_i
1 : 351 | 0.33
2 : 302 | 0.32
3 : 1062 | 0.44
4 : 285 | 0.29
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.38

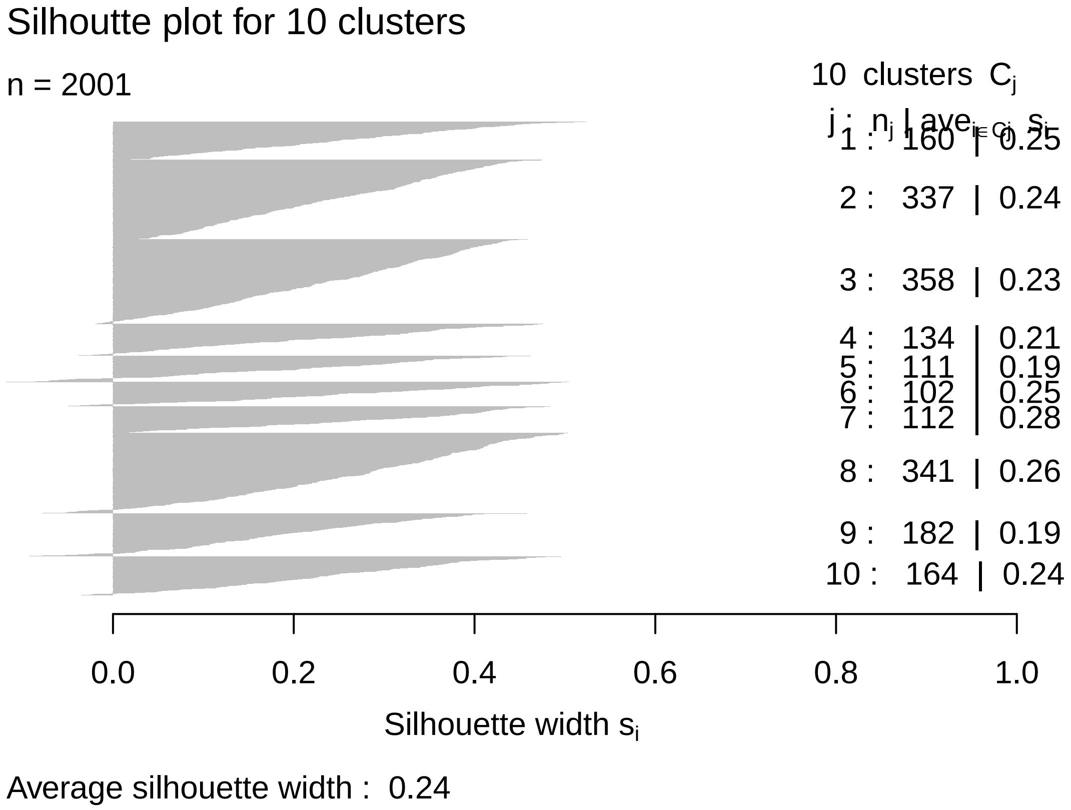

Silhoutte plot for 5 clusters
n = 2000
5 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 345 | 0.30
2 : 469 | 0.28
3 : 272 | 0.27
4 : 687 | 0.32
5 : 227 | 0.31
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.3


Silhoutte plot for 6 clusters
n = 1999
6 clusters C_j
j : n_j | $\text{ave}_{i \in Cj}$ s_i
1 : 466 | 0.26
2 : 225 | 0.30
3 : 164 | 0.22
4 : 155 | 0.22
5 : 681 | 0.32
6 : 308 | 0.32
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.29


Silhoutte plot for 7 clusters
n = 2000
7 clusters C_j
j : n_j | $\text{ave}_{i \in Cj}$ s_i
1 : 370 | 0.23
2 : 386 | 0.24
3 : 163 | 0.22
4 : 149 | 0.22
5 : 305 | 0.32
6 : 406 | 0.21
7 : 221 | 0.30
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.25


Silhoutte plot for 8 clusters
n = 1999
8 clusters C_j
j : n_j | $\text{ave}_{i \in Cj}$ s_i
1 : 369 | 0.21
2 : 358 | 0.23
3 : 354 | 0.24
4 : 187 | 0.16
5 : 118 | 0.28
6 : 148 | 0.25
7 : 303 | 0.31
8 : 162 | 0.20
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.24


Silhoutte plot for 9 clusters
n = 2000
9 clusters C_j
j : n_j | ave_{i∈Cj} s_i
1 : 117 | 0.25
2 : 117 | 0.28
3 : 162 | 0.20
4 : 353 | 0.24
5 : 219 | 0.28
6 : 135 | 0.19
7 : 363 | 0.22
8 : 186 | 0.20
9 : 348 | 0.24
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.23


Silhoutte plot for 10 clusters
n = 2001
10 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 160 | 0.25
2 : 337 | 0.24
3 : 358 | 0.23
4 : 134 | 0.21
5 : 111 | 0.19
6 : 102 | 0.25
7 : 112 | 0.28
8 : 341 | 0.26
9 : 182 | 0.19
10 : 164 | 0.24
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.24

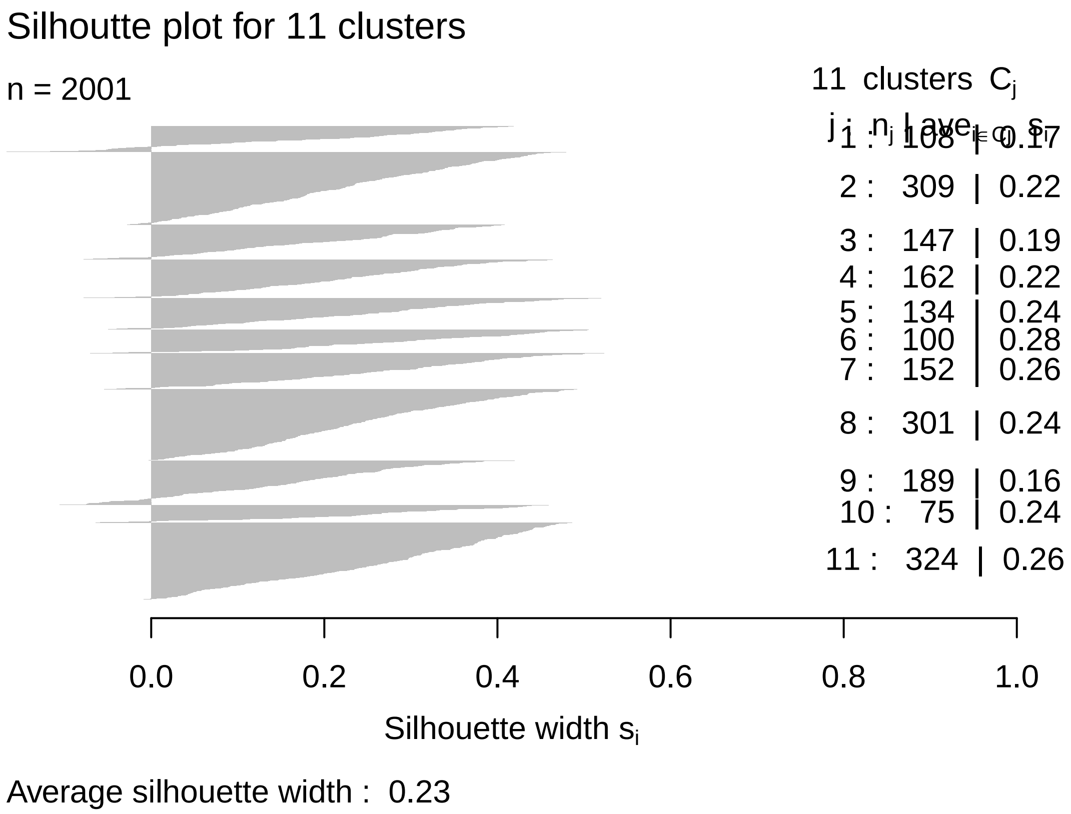


Table 1: Mean Average Silhoutte Length for k = 1, . . . , 11

2	3	4	5	6	7	8	9	10	11
0.464	0.429	0.381	0.293	0.28	0.248	0.239	0.234	0.235	0.221

Analysis of K-Means Clustering Algorithm with k = 11

As we can see in the previous table, the mean average silhouette length favors k = 11 over the pack of k = 7, 8, 9, & 10 which are all in roughly the same range. We will proceed with an analysis of the clustering method for k = 11 first, though our ultimate goal is to appropriately reduce the number of clusters.

Below we have three tables that describe the clustering method in different ways:

Table 1A: Proportion Within Cluster

- This first table shows the proportion of pitches classified under their original classification within each of the 11 new clusters. For example, "Pitch 1" is made up of 72% curveballs, 19% knuckle-curves, and 9% sliders. In other words, the row sums in this table are equal to 1.

Table 2: Table continues below

	Changeup	Curveball	Cut FB	Four-Seam FB	Forkball
Pitch 1	0.006	0.623	0.000	0.000	0.000
Pitch 2	0.009	0.000	0.023	0.920	0.000

	Changeup	Curveball	Cut FB	Four-Seam FB	Forkball
Pitch 3	0.795	0.001	0.000	0.023	0.005
Pitch 4	0.002	0.117	0.033	0.000	0.000
Pitch 5	0.004	0.000	0.473	0.189	0.000
Pitch 6	0.000	0.720	0.001	0.000	0.000
Pitch 7	0.023	0.034	0.128	0.003	0.000
Pitch 8	0.002	0.000	0.000	0.382	0.000
Pitch 9	0.142	0.000	0.001	0.058	0.000
Pitch 10	0.618	0.007	0.002	0.003	0.001
Pitch 11	0.000	0.000	0.004	0.935	0.000

Table 3: Table continues below

	Splitfinger FB	Two-Seam FB	Knuckle Curve	Knuckleball
Pitch 1	0.000	0.000	0.176	0
Pitch 2	0.001	0.029	0.000	0
Pitch 3	0.058	0.038	0.000	0
Pitch 4	0.001	0.000	0.060	0
Pitch 5	0.002	0.007	0.001	0
Pitch 6	0.000	0.000	0.189	0
Pitch 7	0.009	0.000	0.004	0
Pitch 8	0.000	0.454	0.000	0
Pitch 9	0.010	0.478	0.000	0
Pitch 10	0.290	0.005	0.002	0
Pitch 11	0.000	0.052	0.000	0

	Sinker	Slider
Pitch 1	0.001	0.193
Pitch 2	0.005	0.014
Pitch 3	0.073	0.007
Pitch 4	0.000	0.787
Pitch 5	0.000	0.323
Pitch 6	0.000	0.091
Pitch 7	0.001	0.798
Pitch 8	0.162	0.000
Pitch 9	0.310	0.001
Pitch 10	0.012	0.060
Pitch 11	0.008	0.000

Table 2A: Proportion Within Original Classification

- This second table shows how the original pitch classifications are distributed through the 11 new pitch clusters. For example, 0.3% of changeups are classified as “Pitch 2”, 0.3% are classified as “Pitch 3”, 61.6% as “Pitch 4”, 3.7% as “Pitch 5” and so on. In other words, the column sums in this table are equal to 1.

Table 5: Table continues below

	Changeup	Curveball	Cut FB	Four-Seam FB	Forkball
Pitch 1	0.003	0.410	0.000	0.000	0.000
Pitch 2	0.014	0.000	0.072	0.383	0.000
Pitch 3	0.589	0.001	0.000	0.004	0.855
Pitch 4	0.002	0.115	0.055	0.000	0.000
Pitch 5	0.003	0.000	0.654	0.034	0.016
Pitch 6	0.000	0.438	0.001	0.000	0.000
Pitch 7	0.017	0.032	0.201	0.001	0.000
Pitch 8	0.004	0.000	0.000	0.155	0.000
Pitch 9	0.135	0.000	0.002	0.015	0.000
Pitch 10	0.233	0.003	0.001	0.000	0.129
Pitch 11	0.000	0.000	0.014	0.409	0.000

Table 6: Table continues below

	Splitfinger FB	Two-Seam FB	Knuckle Curve	Knuckleball
Pitch 1	0.000	0.000	0.392	0.25
Pitch 2	0.005	0.035	0.000	0.00
Pitch 3	0.251	0.021	0.000	0.25
Pitch 4	0.004	0.000	0.201	0.00
Pitch 5	0.008	0.004	0.003	0.00
Pitch 6	0.000	0.000	0.389	0.00
Pitch 7	0.039	0.000	0.013	0.00
Pitch 8	0.000	0.526	0.000	0.00
Pitch 9	0.054	0.347	0.000	0.00
Pitch 10	0.639	0.001	0.002	0.50
Pitch 11	0.000	0.065	0.000	0.00

	Sinker	Slider
Pitch 1	0.001	0.063
Pitch 2	0.011	0.013
Pitch 3	0.088	0.003
Pitch 4	0.000	0.383
Pitch 5	0.000	0.130
Pitch 6	0.000	0.027
Pitch 7	0.001	0.366
Pitch 8	0.395	0.000
Pitch 9	0.476	0.001
Pitch 10	0.007	0.014
Pitch 11	0.020	0.000

Table 3A: Middle 90% of Feature Space

- This third table shows the middle 90% of the feature space in each new pitch cluster. For example, in Pitch 1, the middle 90% of spin rate is between 2487.0 and 3114.1 reveolutions per minute (rpms), the middle 90% of horizontal movement ranges between 0.447 and 1.598 inches, the middle 90% of vertical movement ranges between -1.600 and -0.411 inches and the middle 90% of velocity relative to the maximum is -21.90 to -12.70 miles per hour (mph).

Table 8: Table continues below

	Pitch 1 5%	Pitch 1 95%	Pitch 2 5%	Pitch 2 95%
Spin Rate	1992.000	2507.000	1958.000	2329.000
Horiz. Movement	0.067	1.351	-0.996	-0.048
Vert. Movement	-1.338	0.042	0.959	1.703
Velo - Max	73.300	83.000	87.900	94.900

Table 9: Table continues below

	Pitch 3 5%	Pitch 3 95%	Pitch 4 5%	Pitch 4 95%
Spin Rate	1550.000	2051.000	2417.000	2928.000
Horiz. Movement	-1.571	-0.654	0.140	1.465
Vert. Movement	-0.141	1.368	-0.601	0.464
Velo - Max	78.500	89.500	79.700	88.700

Table 10: Table continues below

	Pitch 5 5%	Pitch 5 95%	Pitch 6 5%	Pitch 6 95%
Spin Rate	2253.000	2708.000	2520.000	3125.000
Horiz. Movement	-0.260	0.574	0.461	1.627
Vert. Movement	0.286	1.315	-1.600	-0.319
Velo - Max	86.200	94.600	72.900	83.300

Table 11: Table continues below

	Pitch 7 5%	Pitch 7 95%	Pitch 8 5%	Pitch 8 95%
Spin Rate	1965.000	2421.000	1997.900	2419.000
Horiz. Movement	-0.169	0.803	-1.571	-0.814
Vert. Movement	-0.146	0.847	0.582	1.405
Velo - Max	80.000	88.800	91.200	97.600

Table 12: Table continues below

	Pitch 9 5%	Pitch 9 95%	Pitch 10 5%	Pitch 10 95%
Spin Rate	1877.000	2372.000	967.800	1571.000
Horiz. Movement	-1.711	-0.972	-1.410	-0.087
Vert. Movement	-0.125	1.063	-0.311	0.902
Velo - Max	85.700	93.700	78.800	88.700

	Pitch 11 5%	Pitch 11 95%
Spin Rate	2244.00	2617.000
Horiz. Movement	-1.07	-0.051
Vert. Movement	1.08	1.726

	Pitch 11 5%	Pitch 11 95%
Velo - Max	91.20	97.800

The most interesting part of this reclustering is that we have simply reclustered pitches into the same number of clusters as we started with. Originally, we started with 11 pitch classifications and the goal was to reduce the number of clusters to simplify pitch classification. What we have done here is recluster pitches, and show that the current pitch classification rule is not all that good at identifying how a particular pitch looks or moves to a hitter. An example of this is the relationship between curveball and slider. If we look at our first table, Pitch 1 is made up mostly of curveballs and knuckle-curveballs with some sliders mixed in. The interesting thing is that “Pitch 3” is also made up of the same three pitches, just pitches that clearly move differently or travel at different speeds. The original pitch classificiation method is failing us here.

Analysis of K-Means Clustering Algorithm with k = 5

The primary goal of this project was to see if it was possible to reduce the number of pitch clusters. In our previous analysis, we noticed that there are two new pitch clusters that consist mostly of curveballs, knuckle-curves, and sliders. Intuitively, it would make sense to be able to cluster those two pitches together. But how should we determine the number of k to use if not via the average silhouette length as before?

One method of selecting k is what is called “The Elbow Method.” This method only requires that we calculate the ratio of within-cluster sum of squares to the total sum of squares for each k. As we would expect with any sum of squares statistic, this value will always decrease as k increases. Thus, we need to select k where we start to see little-to-no decrease in our within-cluster sum of squares ratio. The following plot and table shows the relationship between k and within-cluster sum of squares divided by total sum of squares.

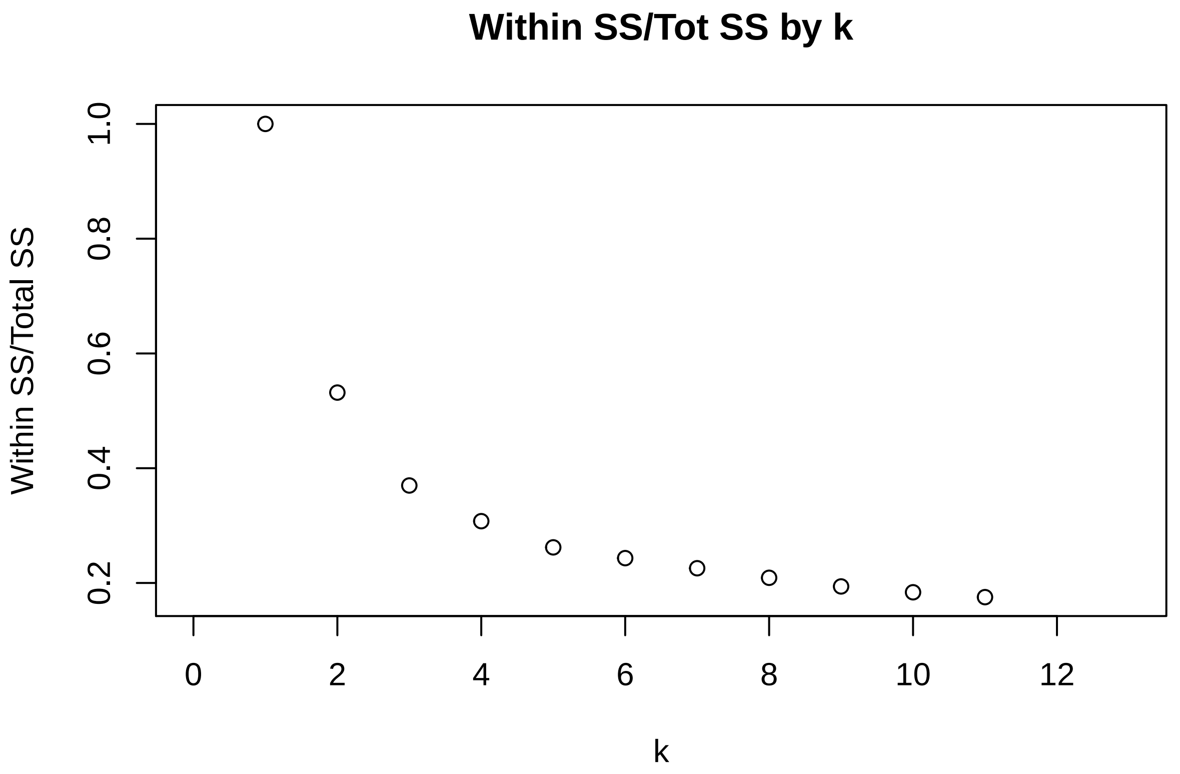


Table 14: Within SS/Total SS by K

1	2	3	4	5	6	7	8	9	10	11
1	0.532	0.37	0.308	0.262	0.243	0.226	0.209	0.194	0.184	0.175

Looking at the plot, the argument could be made for $k = 3$, 4, or 5 in my opinion. The drop in the ratio of within-cluster sum of squares from $k = 2$ to $k = 3$ is substantial (about 0.15), and the drop from $k = 3$ to $k = 4$ is quite a bit smaller (0.06), but I think the most logical k to choose is $k = 5$ since the within-cluster sum of squares ratio drops consistantly as k increases up until we from $k = 5$ to $k = 6$ where the decreases is less than 0.02. Choosing a cluster size is also about the number of clusters I had hoped to be able to achieve when I started this project. Thus, I will proceed with a similar analysis as with $k = 11$, but now with $k = 5$.

Table 1B: Proportion Within Cluster

The table below shows us several things:

1) "Pitch 1" is made up of 77% Changeups with a smaller number (15%) of split-finger fastballs. We can probably label this pitch as simply a "changeup", though we will investigate how these pitches truly look when we look at the feature space laer.
2) "Pitch 2" is made up of mostly cut-fastballs and sliders, two pitches that baseball aficionados know move very similarly. This pitch can probably be labeled more simply as a "slider."
3) "Pitch 3" is made up mostly of curveballs, sliders and knuckle-curves. This pitch would likely be called a curveball or slurve by most coaches.
4) "Pitch 4" is comprised almost exclusively of four-seam fastballs (89%). I imagine this pitch would be considered the "hard, straight fastball" by most players and coaches, which has little to no sinking action and may even appear to rise to hitters.
5) "Pitch 5" is made up of a combination of four-seam fastballs, two-seam fastballs, and sinkers. These pitches are likely the fastballs that have some sinking action as they move towards the hitter and are generally just considered "sinkers" by players and coaches.

Table 15: Table continues below

	Changeup	Curveball	Cut FB	Four-Seam FB	Forkball
Pitch 1	0.012	0.036	0.216	0.016	0.000
Pitch 2	0.062	0.000	0.001	0.247	0.000
Pitch 3	0.002	0.554	0.001	0.000	0.000
Pitch 4	0.002	0.000	0.031	0.897	0.000
Pitch 5	0.719	0.005	0.001	0.020	0.003

Table 16: Table continues below

	Splitfinger FB	Two-Seam FB	Knuckle Curve	Knuckleball
Pitch 1	0.005	0.001	0.012	0
Pitch 2	0.004	0.464	0.000	0
Pitch 3	0.000	0.000	0.162	0
Pitch 4	0.000	0.053	0.000	0
Pitch 5	0.134	0.024	0.001	0

	Sinker	Slider
Pitch 1	0.000	0.702
Pitch 2	0.219	0.003
Pitch 3	0.000	0.281
Pitch 4	0.009	0.007
Pitch 5	0.059	0.033

We have to be wary with the proportions in this table, however. It might be more useful to look at how each of the original pitch classifications are distributed across the new pitch clusters

Table 2B: Proportion Within Original Classification

If we look at the table below, we can see that most changeups, forkballs, splitfinger fastballs, and knuckleballs are classified as “pitch 1”, most cut fastballs and sliders are classified as “pitch 2”, most curveballs and knuckle-curveballs are classified as “pitch 3”, most four-seam fastballs are classified as “pitch 4”, and most two-seam fastballs and sinkers are classified as pitch 5.

Table 18: Table continues below

	Changeup	Curveball	Cut FB	Four-Seam FB	Forkball
Pitch 1	0.021	0.076	0.769	0.007	0.016
Pitch 2	0.146	0.000	0.007	0.156	0.048
Pitch 3	0.003	0.917	0.003	0.000	0.000
Pitch 4	0.008	0.000	0.218	0.831	0.000
Pitch 5	0.823	0.007	0.002	0.006	0.935

Table 19: Table continues below

	Splitfinger FB	Two-Seam FB	Knuckle Curve	Knuckleball
Pitch 1	0.048	0.001	0.089	0
Pitch 2	0.052	0.838	0.000	0
Pitch 3	0.002	0.000	0.908	0
Pitch 4	0.002	0.140	0.000	0
Pitch 5	0.896	0.021	0.003	1

	Sinker	Slider
Pitch 1	0.001	0.729
Pitch 2	0.837	0.004
Pitch 3	0.000	0.230
Pitch 4	0.052	0.015
Pitch 5	0.110	0.023

Based on these results, the naming mechanisms I have proposed in the previous section still seem to be reasonable. The last thing we should do, however, is investigate the feature space of each of our new clusters to see how these pitches act in general.

Table 3B: Feature Space

A few notable things about the feature spaces of these pitches seen in the table below.:

- Pitch 1 has a noticeable difference in spin rate from all other pitches. A pitch under 1500 rpms will almost always be classified as pitch 1.
- Pitch 4 and 5 have very similar profiles except for differences in vertical movement, providing more evidence for the sinking fastball vs. rising fastball naming convention I have proposed a couple of sections earlier.
- Pitch 2 and 3 are quite similar, both of which being pitches with high spin rates and lots of downward vertical movement. Pitch 3 has more significant vertical movement, while Pitch 2 has more vertical movement. Another stark difference is the speed at which each of these breaking pitches are thrown. Pitch 2 is thrown about 5-7 mph faster relative to maximum velocity than Pitch 3 is.

Table 21: Table continues below

	Pitch 1 5%	Pitch 1 95%	Pitch 2 5%	Pitch 2 95%
Spin Rate	2045.000	2754.000	1907.00	2389.000
Horiz. Movement	-0.125	0.870	-1.65	-0.775
Vert. Movement	-0.244	0.938	0.23	1.332
Velo - Max Velo	80.800	90.800	86.90	96.400

Table 22: Table continues below

	Pitch 3 5%	Pitch 3 95%	Pitch 4 5%	Pitch 4 95%
Spin Rate	2120.000	2993.000	2066.000	2582.000
Horiz. Movement	0.214	1.573	-1.101	0.057
Vert. Movement	-1.449	0.112	0.971	1.709
Velo - Max Velo	73.700	84.900	89.300	97.400

	Pitch 5 5%	Pitch 5 95%
Spin Rate	1142.000	2015.000
Horiz. Movement	-1.550	-0.335
Vert. Movement	-0.313	1.252
Velo - Max Velo	78.600	89.400

Ultimately, after analyzing the feature space of these pitch clusters, the naming conventions I have proposed continue to seem reasonable. Thus, in the case where we choose $k = 5$, we end up with only 5 total types of pitches in baseball: A changeup, slider, curveball, rising fastball, and sinking fastball. This naming convention simplifies the nature of pitching instead of having multiple names for pitches that look and act the same to most observers.

Discussion

Pitching in baseball can seem quite complicated and convoluted to the casual fan. Understanding things like pitch sequencing is even further muddled by the current naming conventions of pitches in the Major League Baseball. The goal of this project was to reduce the complexity of the naming structure for casual fans and see how much, if any, information we lose.

Using k-means reclustering, out of necessity, I used two different methods to determine k. One method, average silhouette lengths, led me to choose $k = 11$, exactly the number of clusters I started with. Even though the data was clustered differently than the original data, a cluster size of $k = 11$ was not the original goal of the project. That being said, the fact that the data was reclustered differently than the original data despite having the same number of clusters leads me to believe that the naming conventions in baseball currently are not exceptionally good at describing the way a pitch looks to a hitter or well-informed observer anyway. Thus, reclustering with a smaller k seems reasonable and not particularly harmful if we choose k in an intelligent way.

Using the "elbow method" to select k, I chose $k = 5$ since the within-cluster sum of squares levels off and decreases much more slowly after $k = 5$. This clustering algorithm was particularly insightful, since we managed to reduce the types of pitches down to changeup, slider, curveball, rising fastball, and sinking fastball.

Ultimately, this project was insightful in that I confirmed a lot of suspicions about which pitches are similar to each other and how the naming conventions in modern baseball seem precise but can be convoluted and imprecise as well.