CLUSTER ANALYSIS OF MAJOR LEAGUE BASEBALL BATTING DATA

OWEN FISH

Abstract. We attempt to use unsupervised learning techniques, specifically K-Means Clustering on batting data from the 2018 MLB Season to verify the common idea of archetypes among hitters. Our results give relatively good separation between points in 8-Dimensional space, but the separation did not appear to come from player archetype. Rather, the separation appeared to come from usage rate, a proxy variable for player skill.

1. Introduction

A well known property in baseball of any level is the archetype of the hitter. Speedy, contact-oriented hitters hit at the beginning of the batting order, while power hitters tend to hit in the fourth or fifth spot. This traditional arrangement of hitters usually maximizes potential runs because top-of-the-order hitters excel in getting onto base themselves while power hitters excel at batting runs in. There is another group of hitters that is somewhat in-between the two aforementioned in terms of contact vs power hitting identity. This model is widespread from little-league to the MLB, but it assumes a certain number of distinct hitter classes with little population in between each group. We seek to investigate whether or not there exist distinct clusters of hitters in terms of several common baseball statistics.

2. Data

The data we collected was a collection of hitting data from Baseball-Reference.com. These data included 30 statistics collected from 1,273 different players during the 2018 MLB season. The various statistics included, but were not limited to, Name, Team, Hits, Doubles, Home Runs, Runs Batted In, Stolen Bases, Batting Average, Plate Appearances, and Sacrifice Fly Outs for a particular player. We hope to use unsupervised learning methods to find a pattern in these data to see if, and the degree to which, generic clusters really exist.

It is not surprising that a lot of the statistics we obtained are correlated with each other; in fact, many statistics are direct linear combinations of others. This is not ideal for our model, as any strongly correlated data will give extra weight to a particular metric and introduce bias into the clustering process. It is important to note that we are not attempting to separate players based off of their statistics; rather, we are attempting to separate players based off of their *traits*, and we are using their statistics as respective estimators. With that in mind, we attempt to find a combination of statistics that best represents the traits of each player, keeping in mind the correlation between bias and comprehensiveness.

3. Methods

Given the goals of our research, we elected to using clustering, a generic method of identifying groups of points in n-dimensional vector space. We used K-Means Clustering, a technique that requires a parameter k that informs the algorithm the number of clusters to form. This technique

Date: April 2019.

is useful but requires a slight knowledge of the data. This is not always feasible, but in this case we estimate that there are $k = 3$ clusters of points in the data. However, we verified k=3 graphically with an Elbow Plot.

This plot shows the total 'variance' between the clusters with the given k.

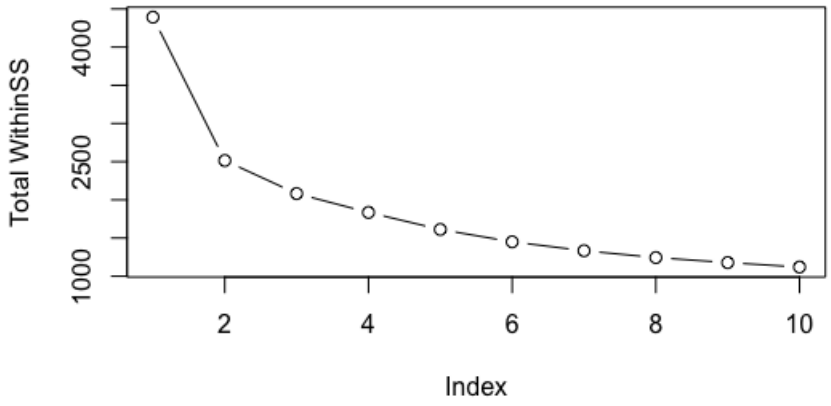

FIGURE 1. Elbow Plot

This plot shows diminishing returns on variance with each additional k. Combined with the fact that bias increases as variance decreases, we feel that any $k \in \{2, 3, 4\}$ is appropriate. Based on our previous knowledge of the data, however, we choose k=3.

We aim to separate players in terms of predefined traits that seem to exist in different baseball players. We are attempting to model **contact hitting, power hitting, speed, and plate-discipline** with the use of the following statistics:

(1) Batting Average
(2) Doubles
(3) Home Runs
(4) Runs Batted In
(5) Stolen Bases
(6) Strikeouts
(7) Base on Balls
(8) Percentage At Bat

Each of the above metrics are common baseball statistics, with the exception of Percentage At Bat. We defined Percentage At Bat as the percentage of total *plate appearances* that qualified as an At-Bat. This metric excludes outcomes such as base on balls, sacrifice hit, and hit by pitch, all outcomes that contribute to the 'plate discipline' trait. Before we use the above selection of

statistics, we verify that few of them are strongly correlated. We obtained the following correlation matrix between the statistics.

```
> cor(data)
             BA         X2B         HR        RBI          SB         BB         SO        pctAB
BA    1.00000000  0.56170027  0.4365557  0.5399946  0.303144891  0.4307642  0.3396354  0.057749247
X2B   0.56170027  1.00000000  0.7572150  0.8799005  0.444002624  0.7693229  0.7595587 -0.085171534
HR    0.43655575  0.75721504  1.0000000  0.9242269  0.287661247  0.7737714  0.7992006 -0.185870546
RBI   0.53999463  0.87990051  0.9242269  1.0000000  0.352620490  0.8164855  0.8120348 -0.152006361
SB    0.30314489  0.44400262  0.2876612  0.3526205  1.000000000  0.3944371  0.3880537 -0.007513881
BB    0.43076417  0.76932294  0.7737714  0.8164855  0.394437143  1.0000000  0.7754883 -0.489299187
SO    0.33963543  0.75955873  0.7992006  0.8120348  0.388053687  0.7754883  1.0000000 -0.152919680
pctAB 0.05774925 -0.08517153 -0.1858705 -0.1520064 -0.007513881 -0.4892992 -0.1529197  1.000000000
```

FIGURE 2. Correlation Matrix for various statistics

This correlation matrix shows us that, with the exception of Home Runs and Doubles, the statistics are not largely correlated. This is good for our model, as it decreases the amount of bias that enters the process.

4. RESULTS

For the purposes of presentation, all figures shown have been reduced to two-dimensions via Multidimensional Scaling. However, each point still remains a vector in $\mathbb{R}^8$, and the clustering algorithm works in that vector space.

When we run the Multidimensional Scaling on the data before clustering, we see the following plot:

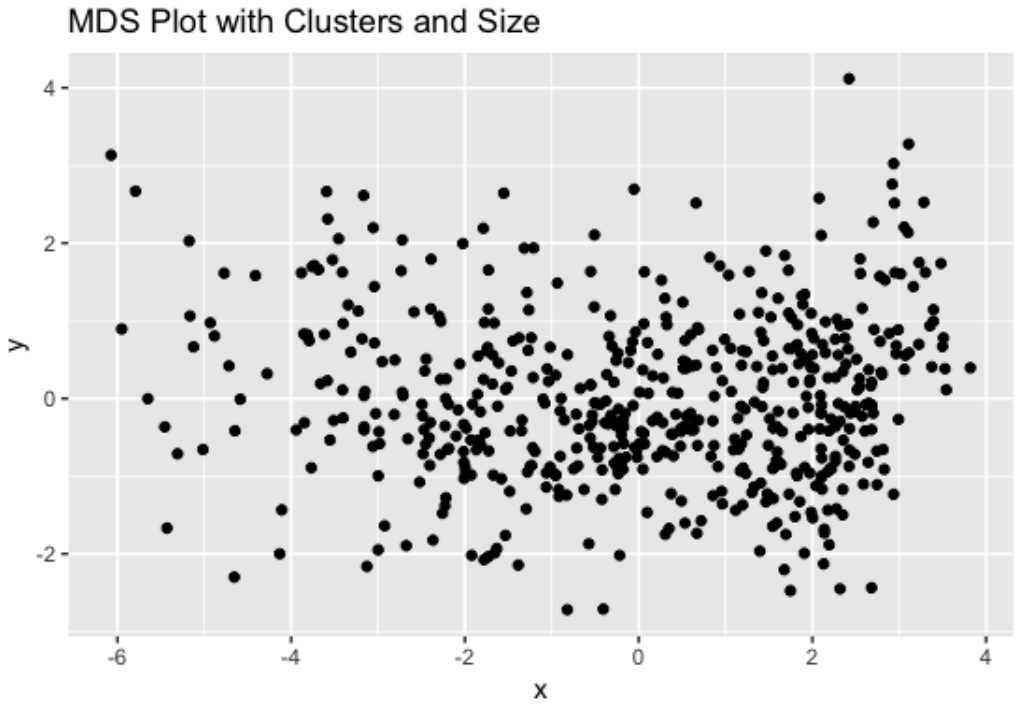

FIGURE 3. Caption

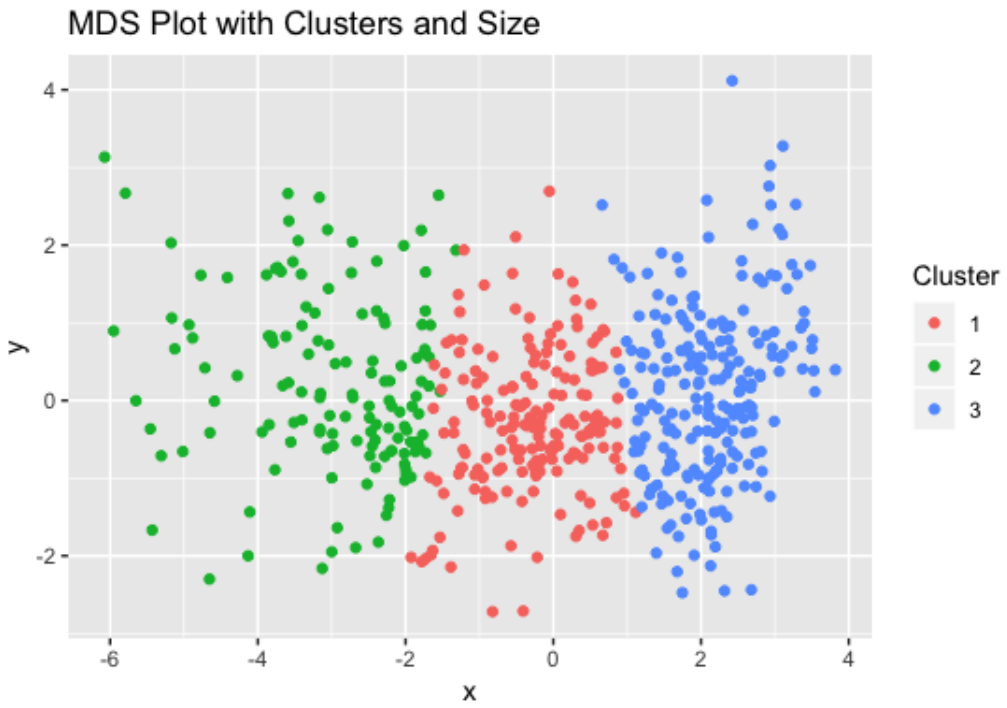

FIGURE 4. MDS Plot with color representing cluster of each point

We get the following plot when we color each point by K-Means Cluster. Though we see very strict borders in the clusters in 2-D space, we do not necessarily see extremely distinct clusters. We also wish to verify that the data gathered is somewhat random with regard to Plate Appearances, a statistic that we expect to be hidden in a number of our covariate statistics. That is, we wish to verify that plate appearances does not correlate strongly with cluster. To do this, we update the above plot to account for each point's Plate Appearance attribute.

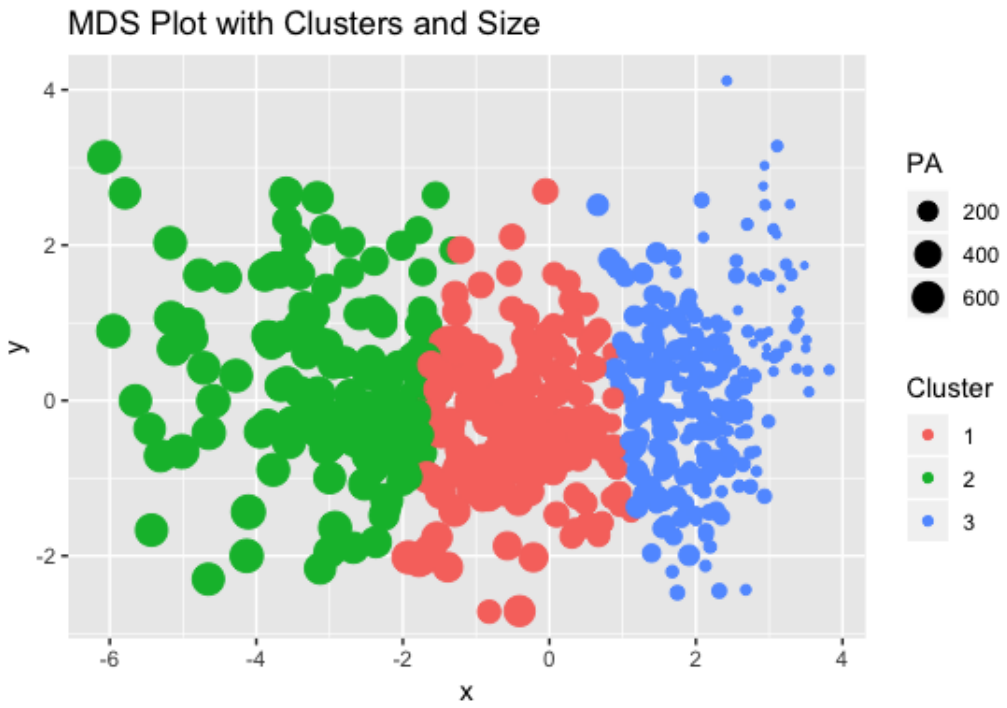

FIGURE 5. MDS Plot with Size representing plate appearances

Having classified each point as belonging to one of three clusters, we want to see summary statistics about the data in each cluster. This table is shown in Figure 6. This will help us see the differences between the clusters. When subsetting our original data set by what cluster each point was classified into, we obtain the following table of median values of a given statistics for each cluster. It is worth noting that all of the given statistics were not used to separate the data; rather, these are just median values of various statistics across each group. This table shows a few key things worth pointing out. First, nearly every statistic increases or decreases with the difference in PA, plate appearances. This makes perfect sense, as a player with more plate appearances *should* hit more home runs than somebody with less plate appearances; however, it is relatively surprising that normalized statistics increase similarly. There is no reason that on-base-percentage should change due to number of plate appearances. This is likely due to the fact that a player with a higher batting average will likely be given more plate appearances from the manager.

This difference table does not show differences in cluster due to player archetype so much as by usage. Cluster 2 likely represents high caliber players: players with all-star appearances no matter their archetype. These types of players generally play 130 (out of 162) games per year, hence the high median plate appearances. Cluster 1 likely represents a middle-tier of players. These players are not all-stars and will not play as many games as cluster 2 players, but they have earned spots in the league from their play. This is shown from relatively productive statistics. Cluster 3 likely represents players who are used as replacements for injured players or players who played less than 60 games.

```
> diff
            1       2       3
R     41.0000  76.000  10.000
H     83.0000 135.000  20.000
HR     9.5000  22.000   2.000
RBI   39.0000  72.000   9.000
SB     3.0000   6.000   0.000
BB    28.0000  55.000   7.000
SO    76.5000 123.000  29.000
TB   132.0000 233.000  32.000
BA     0.2520   0.258   0.207
OBP    0.3155   0.340   0.271
SLG    0.4070   0.459   0.309
PA   372.0000 597.000 106.000
```

FIGURE 6. Median values for given statistics by cluster

5. CONCLUSION

We sought to find relationships in the data to show different types of players who have different proficiencies for different aspects of hitting. However, we found that from the Major League Baseball data from the 2018 season, the overwhelming separation in players comes from tiers in skill, not type of player. This is not to say that player archetypes do not exist, just that it is not an efficient way to separate all the players in the population. Consider four players Anthony, Bryce, Christian, and Daniel who are classified as archetypes $a, a, b,$ and b respectively by the "eyeball test." Also consider that Anthony is simply a better player than Bryce within the same archetype and Christian is better than Daniel. Our separating algorithm would categorize them as {Anthony, Christian}, {Bryce, Daniel}. This does not mean that the players are not still of archetypes $a, a, b,$ and b, it just means that the separability in the data comes from the gap in skill. A potential way to alleviate this problem is to filter the data more selectively from the start. If we try to control for rogue variables by subsetting our data with more strict plate appearance numbers, we can somewhat control for skill gap. If we cluster on that data set, we might see a higher degree of separability due to player archetype.

6. Bibliography

- Basu, Sugato, Arindam Banerjee, and Raymond Mooney. "Semi-supervised clustering by seeding." In Proceedings of 19th International Conference on Machine Learning (ICML-2002. 2002.
- Cousens, Laura. "From diamonds to dollars: The dynamics of change in AAA baseball franchises." Journal of Sport Management 11.4 (1997):316-334.
- Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.
- Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." Icml. Vol. 1. 2001.
- Wickham, Hadley, and Maintainer Hadley Wickham. "The ggplot package." (2007).