2019 MLB Pitchers Data Analysis

Richard Burson

May 14, 2020

# Introduction

Every team in baseball is trying to build the best team they possibly can to try and win a World Series. One component of that is trying to find players who perfrom very well but are on cheaper contracts relative to their performance. I wanted to see if I could find any pitcher that might be underrated and might be a potential cheaper option for a team to pursue acquiring.

I got my data from baseball savant and it contains all of the pitchers in 2019 that faced at least 250 batters. It contains 8 variables of 279 pitchers. The 8 variables are expected batting average(xBA), expected slugging(xSLG), expected weighted on base average(xwOBA), expected on base percentage(xOBP), expected isolated power(xISO), average exit velocity, average launch angle, barrel percentage. The “expected” statistics are similar to their traditional statistic counterparts but instead the “expected” stats are formulated using exit velocity, launch angle, and on ground balls speed. Each batted ball given up by a pitcher is given expected values based on the results that similar batted balls have had all season.

I chose to use principal component anaylsis to reduce the dimensions of the data from 8 variables to 3 principal components. Then I used kmeans cluster analysis method to see if there is a lesser thought of pitcher clustered with more highly thought of pitchers.

# Data Science Method

Principal component analysis allows somebody to reduce very large datasets with numerous variabes into much simpiler datasets. This is done over a few steps, first standardize the original data and call the matrix Z. Create the covariance matrix of Z by multipying it with its transpose

Next find the eigenvectors and the eigenvalues of the matrix

and then sort the eigenvalues from largest to smallest into a matrix

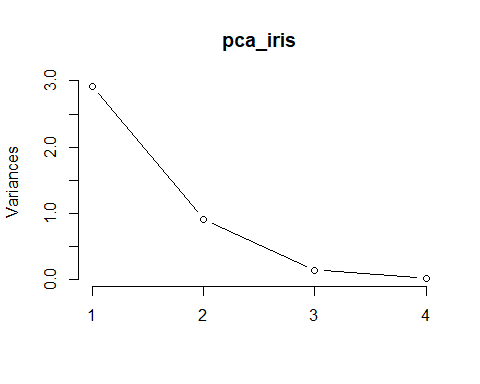
Lastly calculate a final matrix

that you will help decide how many pricipal components to retain.( Brems (2019), Yiu (2019) )

K-means cluster analysis is a method that groups similar observations together. First determine how many clusters to use by making a plot of the wintin sum of squares for each number of clusters used and identifying the “elbow”" to chose the amount of clusters to use, k. Then initialize k number of centroids. For every data point assign it to the cluster with the nearest centroid. Then after every point has been assigned reposition the centroid to be the new average of the data in that cluster. Repeat the last two steps until all of the observations have been categorized into clusters. Dabbura (2020)

## Illustration of Method

This is an example of Princial Component on the iris data set.

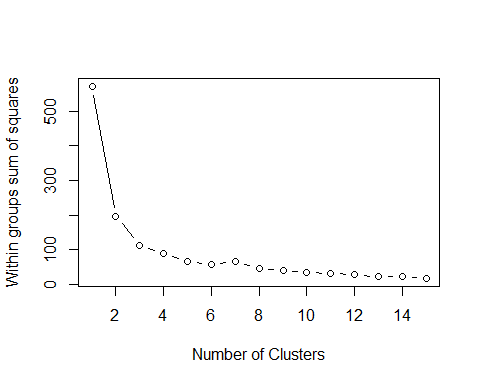


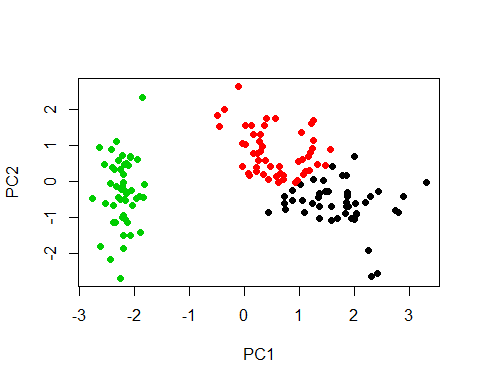
## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.7084 0.9560 0.38309 0.14393  
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518  
## Cumulative Proportion 0.7296 0.9581 0.99482 1.00000

Choose to retain the first 2 PC of iris data because of the “elbow” in the plot as well as the first 2 PC’s account for 95% of the original variance.

## PC1 PC2 PC3 PC4  
## Sepal.Length 0.5210659 -0.37741762 0.7195664 0.2612863  
## Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096  
## Petal.Length 0.5804131 -0.02449161 -0.1421264 -0.8014492  
## Petal.Width 0.5648565 -0.06694199 -0.6342727 0.5235971

The weights of the principal components show the what variables are important to each PC. For the iris data PCA, PC1 takes into account every variable but PC2 is basically only taking into account the sepal.length and sepal.width variables.

 Decide to use k=3 becasue of the “elbow”" in the graph

 Can clearly see that there are three distinct cluster in the plot

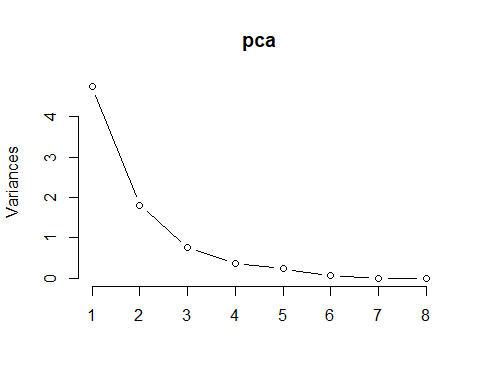
# Novel Analysis

## Description of Data and Exploratory Data Analysis

I obtained my data from <https://baseballsavant.mlb.com/leaderboard/custom?year=2019&type=pitcher&filter=&sort=4&sortDir=asc&min=250&selections=xba,xslg,xwoba,xobp,xiso,exit_velocity_avg,launch_angle_avg,barrel_batted_rate,&chart=false&x=xba&y=xba&r=no&chartType=beeswarm>

After reading in the data I renamed the rows of the data to the full names of the players to help identify them in plots later on. I also selected only the numerical columns I wanted for the principal component analysis, leaving out the columns of their names

## Actual Analysis

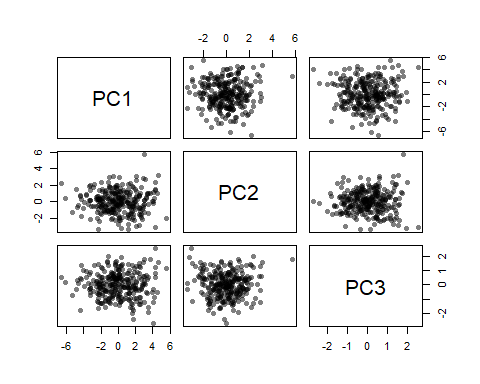


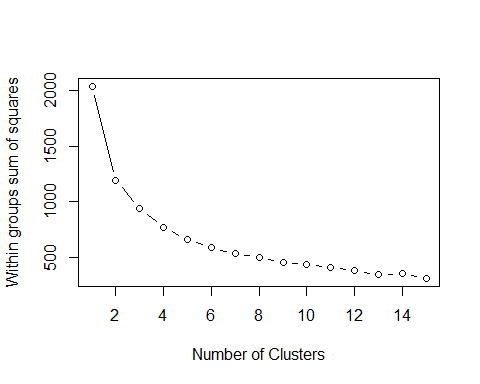
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6  
## Standard deviation 2.1794 1.3446 0.87187 0.61654 0.48721 0.24804  
## Proportion of Variance 0.5937 0.2260 0.09502 0.04752 0.02967 0.00769  
## Cumulative Proportion 0.5937 0.8197 0.91474 0.96226 0.99193 0.99962  
## PC7 PC8  
## Standard deviation 0.05424 0.009298  
## Proportion of Variance 0.00037 0.000010  
## Cumulative Proportion 0.99999 1.000000

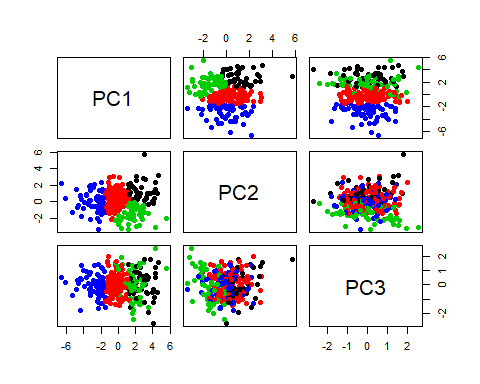
I chose to use three principal components because of the “elbow” in the plot as well as the using the first three principal components gets the cumulative proportion of variance to .91474 which is good.

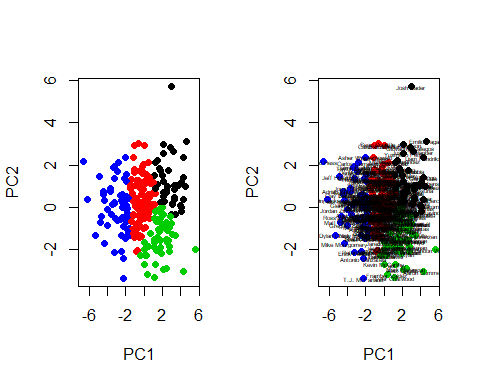
## PC1 PC2 PC3 PC4  
## xba -0.40168271 -0.27538154 -0.11225064 -0.0425251  
## xslg -0.44807153 0.04721856 -0.12666710 0.1064671  
## xwoba -0.44048435 -0.16712467 -0.12399689 -0.1503844  
## xobp -0.34954456 -0.38340827 -0.09247596 -0.4008588  
## xiso -0.40464307 0.30625818 -0.11767903 0.2001144  
## exit\_velocity\_avg -0.24687831 0.15919968 0.91509032 -0.2651285  
## launch\_angle\_avg 0.01579811 0.64300070 -0.30962160 -0.6798956  
## barrel\_batted\_rate -0.31103144 0.46304898 0.01800087 0.4805791  
## PC5 PC6 PC7 PC8  
## xba 0.53499323 -0.544306782 0.140313386 -0.382402848  
## xslg 0.32258422 0.127444377 0.158501208 0.790061536  
## xwoba -0.14123915 0.113622357 -0.840357576 -0.031508631  
## xobp -0.60087061 0.015036620 0.447229900 0.017057161  
## xiso 0.09148876 0.634873232 0.221120996 -0.477752114  
## exit\_velocity\_avg 0.07709551 0.008757487 -0.001620872 0.000346817  
## launch\_angle\_avg 0.07663510 -0.149221970 -0.002071705 0.005707610  
## barrel\_batted\_rate -0.45666270 -0.498942889 -0.008524567 0.002703854

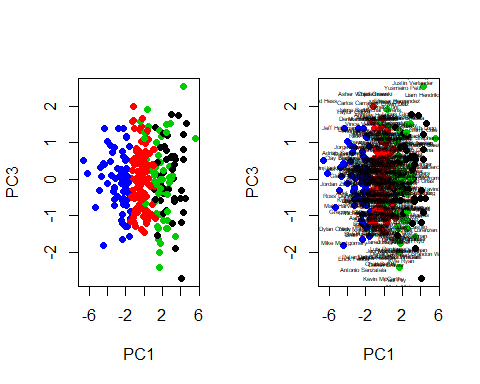
PC1 looks like it is is an overall stat on how good a pitcher is. The higher the value the better the pitcher and lower the value of PC1 the worse the pitcher. PC2 is contrasting launch angle, barrell% rate, and xiso, xobp and xba. In baseball terms players with a low PC2 value are pitchers that have more groundball contact and pitchers who have more flyball contact have a higher PC2 value. PC 3 is contrasting exit velocity with launch angle.

 Principal Component Analysis Plots before kmeans clustering is applied.

 By this plot use 4 clusters for the data.

 Same PCA plot as before just with the clusters on it

 PC1 and PC2 plot -Players on the right have a high PC1 so they are the better pitchers -Players higher on the graph are fly ball pitchers, ground ball pitchers on the bottom

 PC1 and PC3 plot

## Results and Conclusions

##   
## 1 3 4 2   
## 44 57 69 109

There are 44 observations that got classified into cluster 1 and these are the good pitchers who had a high PC1 as well having a high PC2. Meaning they are the points on the upper right of the PC1,PC2 plot. In the second cluster there are 57 observations and these were the points on the bottom rights of the plot. They are still effective pitchers since thay have a high PC1, the pitchers in this cluster tend to inducee a little bit more ground balls than in cluster 1. In cluster 4 you have the points that are sandwiched in the middle of the plots and these could be condisered as average or middle of the road pitchers. The lastly in cluster 3 there are the pitchers that are on the left and far left sside of this graph and these are the below average/bad pitchers for the 2019 season. I want to focus on clusters one and two because these are the where the good pitchers have been clustered and where there may be an opportunity to find someone undervalued.

## [1] "Hansel Robles" "Adam Ottavino" "Kenley Jansen"   
## [4] "Chris Paddack" "Edwin Diaz" "Jack Flaherty"   
## [7] "Trevor May" "Julio Urias" "Lance Lynn"   
## [10] "Blake Snell" "Ross Stripling" "Justin Verlander"   
## [13] "Diego Castillo" "Lucas Giolito" "Noe Ramirez"   
## [16] "Will Smith" "Yusmeiro Petit" "Mike Clevinger"   
## [19] "Raisel Iglesias" "Josh James" "Hector Neris"   
## [22] "Ian Kennedy" "Joakim Soria" "Walker Buehler"   
## [25] "Pedro Baez" "John Brebbia" "Giovanny Gallegos"  
## [28] "Jacob deGrom" "Taylor Rogers" "Stephen Strasburg"  
## [31] "Gerrit Cole" "Seth Lugo" "Emilio Pagan"   
## [34] "Roberto Osuna" "Josh Hader" "Elieser Hernandez"  
## [37] "Chris Sale" "Robert Stephenson" "Jose Leclerc"   
## [40] "Max Scherzer" "Nick Anderson" "Liam Hendriks"   
## [43] "Kenta Maeda" "Dinelson Lamet"

These are the names that are grouped into the first cluster. Many of the names in this cluster are pretty recongnizable for baseball fans with Josh Hader, Max Scherzer, and Jacob DeGrom being in this cluster. Two players that are potentially undervalued in this cluster could be Noe Ramirez and Josh James.

## [1] "Luis Perdomo" "Hector Rondon" "Aaron Bummer"   
## [4] "Mike Soroka" "Steve Cishek" "Derek Law"   
## [7] "Jake Diekman" "Zac Gallen" "Kyle Ryan"   
## [10] "Scott Barlow" "Colten Brewer" "Adrian Houser"   
## [13] "Eduardo Rodriguez" "Brandon Woodruff" "Jace Fry"   
## [16] "Mark Melancon" "Archie Bradley" "Alex Claudio"   
## [19] "Jose Alvarez" "Zack Wheeler" "Marcus Stroman"   
## [22] "Craig Stammen" "Paul Fry" "Michael Lorenzen"   
## [25] "Hyun-Jin Ryu" "Matt Andriese" "Kyle Hendricks"   
## [28] "Sonny Gray" "Luis Castillo" "Marcus Walden"   
## [31] "Miguel Castro" "Frankie Montas" "Noah Syndergaard"   
## [34] "Jared Hughes" "Kevin McCarthy" "Chris Bassitt"   
## [37] "Matt Barnes" "Luis Cessa" "Wade Miley"   
## [40] "Zack Greinke" "Martin Perez" "Max Fried"   
## [43] "Matt Albers" "Jeurys Familia" "Francisco Liriano"  
## [46] "Ryan Yarbrough" "Chad Bettis" "Brandon Workman"   
## [49] "John Gant" "Robert Gsellman" "Charlie Morton"   
## [52] "Tyler Chatwood" "Lou Trivino" "Luke Jackson"   
## [55] "Junior Guerra" "Sean Newcomb" "Framber Valdez"

These are the names in the second cluster and this cluster is tougher to make sense of because some of these players are considered good like Zack Greinke, Noah Syndergaard, and Sonny Gray. While others like Paul Fry and Framber Valdez had bad 2019 seasons.

If I were to do anything for future analysis I would probably want to add innings pitched to halp differentiate starting pitchers from relivers as well as maybe adding in a variable to take into account strikeouts.

# References

Brems, Matt. 2019. “A One-Stop Shop for Principal Component Analysis.” *Medium*. Towards Data Science. <https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c>.

Dabbura, Imad. 2020. “K-Means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.” *Medium*. Towards Data Science. <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>.

Yiu, Rebecca. 2019. “Market Segmentation with R (Pca & K-Means Clustering) - Part 1.” *Medium*. Towards Data Science. <https://towardsdatascience.com/market-segmentation-with-r-pca-k-means-clustering-part-1-d2c338b1dd0b>.