Data Mining and Its Application to Baseball Stats

CSU Stanislaus

Devin Eudy

CS 4960

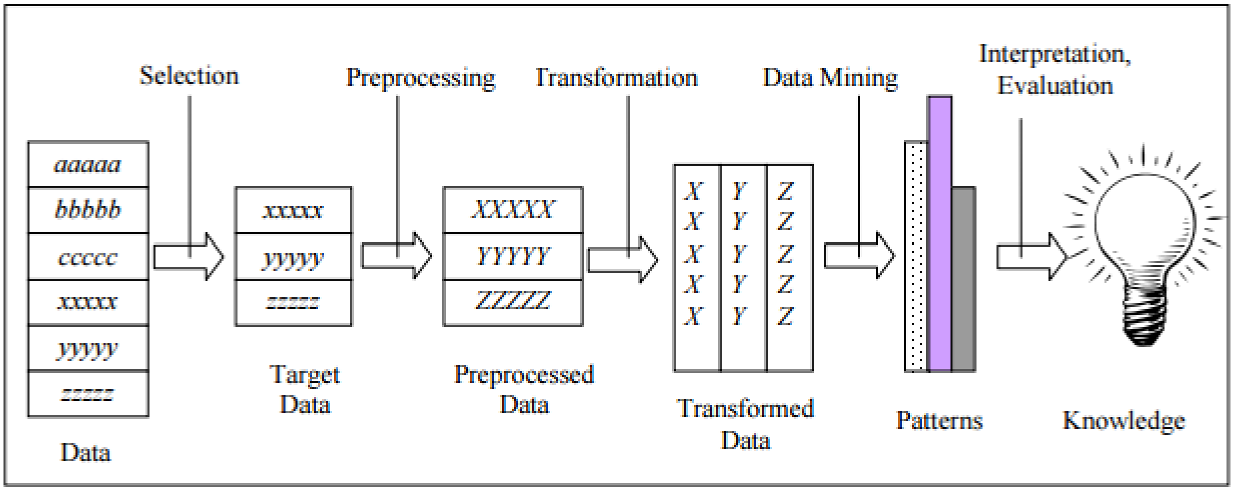
Dr. Martin

Introduction

Data mining is one of the big trends that we consumers are constantly hearing about. We hear about companies like Google and Apple storing massive amounts of data on their customers, but we don't hear about what these companies are doing with that data or who they are giving access to the data they have. This unknown is one of the many reasons why data mining has become such a hot topic. How are these companies getting this information? What are they going to do with it? Are we even safe? What most people don't realize is that data mining is useful in many different environments outside of data acquired from customers/consumers, in fact data mining can be applied to just about anything that has significant amounts of data. In this paper I am going to discuss what data mining is, specifically what K-means clustering (known as Lloyds algorithm) does, and how it can be applied to baseball statistics.

Data Mining

So what exactly is data mining? Data mining is a process that takes data as input and outputs knowledge. (Weiss and Davison, 2010) To be more specific, data mining is the non-trivial process of finding potentially useful and understandable patterns within large sets of data. The key detail in this definition is the word “non-trivial”. What this means is that simple calculations or measures are not considered data mining. The process of data mining has to be automated, relying on computers algorithms to sort though data and find useful patterns. (Weiss and Davison, 2010) Data mining includes the process of preparing the data for mining. This

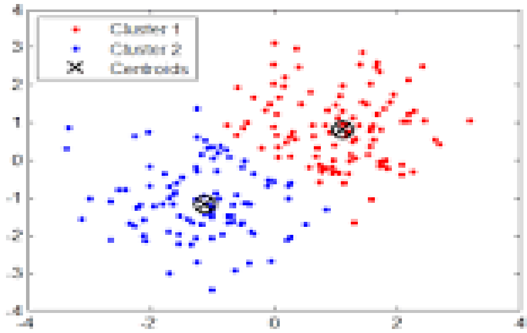


(Figure 1: the data mining process)

means finding relevant data from a potentially large and diverse set of data and any necessary preprocessing must then be performed. Data preparation is probably the most critical step of the data mining process. Without relevant and high quality data, it is likely that the data mining process will leave nothing to learn from. When the data is adequately picked and prepared, it is then transformed into a suitable representation for the data mining algorithm to work on. After the algorithm is applied and the data is retrieved, the data has to be analyzed. The final data cannot simply be accepted. The final data needs to be interpreted based on what information is trying to be discovered from the given data set. Once the final data is analyzed, the results are deemed acceptable, or the data is determined inadequate and further improvements are stated, and the process is done again after the necessary adjustments are made. The data mining process is iterative and isn't complete until meaningful data is retrieved.

K-means Algorithm

The K-means algorithm is a very popular algorithm used in data mining. First proposed in 1957 by S.P. Lloyd (which is why the K-means algorithm is also referred to as Lloyd's algorithm), the K-means algorithm partitions *n* objects into *k* clusters. What this means is that the k-means algorithm sorts large sets of data into clusters, keeping similar single data points in large clusters. To give another view of this, the algorithm specifically assigned each singular data points to a cluster whose center is nearest to each specific data point. “The algorithm starts with an initial set of clusters, chosen at random or according to some heuristic procedure. Then the algorithm iteratively assigns each object to one of the clusters. In each iteration, each object is assigned to its nearest cluster center according to the Euclidean distance between the two.” (Huimin, 2014)



(Figure 2: how the K-means algorithm works)

Traditional Baseball Analysis

Now that I've gone into a bit of detail about data mining and a common algorithm used in data mining, I'd like to discuss baseball statistics and how they shape the game of baseball at the major league level. Traditional baseball statistics have been recorded in the MLB since the 19th century. The very core of these statistics being batting average, RBI's (runs batted in), and home runs for hitters (all three of the stats together are often referred to as a batters "slash line"), and wins, ERA (earned run average) and strikeouts for pitchers. These core statistics and a few other things like scouting have long been the preferred way to analyze and understand a player's value. For well over one hundred years these traditional statistics have dominated the field of baseball analysis. In recent years, with the development of sabermetrics, traditional statistics and the sport of baseball have seen significant changes. Sabermetrics is defined as the empirical analysis of baseball and baseball statistics that measure in-game activity. Sabermetrics research began in the mid 1900's and was presented in a book called "Percentage Baseball" in 1964. While the book did gain national attention from the media, it was generally disregarded and criticized by many baseball organizations. As the years went on, sabermetrics was more researched and finely tuned to the point where we are today with sabermetrics. In baseball's current state sabermetrics can often be associated with not only individual statistics, but also scouting, and the business side of baseball. From a post on Fangraphs.com written in 2012, it is assumed that two-thirds of baseball teams in the MLB utilize all three fields of sabermetrics for statistical leverage within their organization, and all thirty teams utilize at the least two fields for statistical leverage. This means that every organization in the MLB is using some form of advanced analysis in their front office decisions and/or even down to their game to game decisions. Having said that, twenty-two of the thirty teams in the MLB employ at least one person full-time that is dedicated simply to researching sabermetrics, and four teams employ one person part time (Woodrum, 2014). With this knowledge we can easily come to the conclusion that sabermetrics and advanced analysis is taking over the MLB. Unfortunately this has led to a massive debate among baseball analysts, coaches and fans alike. Which way is better, traditional statistics or sabermetrics? Whether you sway in a single direction or prefer to use them both side by side to evaluate a player, the remainder of this paper will take baseball statistics to an even higher level of analysis.

The Data Set

Baseball statistics are a perfect ground for applying higher analytic techniques like data mining. There are over one hundred decades of baseball statistics available for reference and many different statistics that measure every aspect of the game. Between traditional statistics and sabermetric statistics, we have a very generous supply of statistics that data mining can be applied to. For this paper, I'll be referencing work done by David Tung and his analysis of career hitting stats from the year 2012.

"The following traditional baseball statistics will be used for this paper: Games (G), At Bats (AB), Runs (R), Hits (H), Doubles (2B), Triples (3B), Home Runs (HR), Runs Batted In (RBI), Stolen Bases (SB), Caught Stealing (CS), Walks (BB), Strikeouts (K), Intentional Walks

(IBB), Hit By Pitcher (HBP), Sacrifice Hits (SH), Sacrifice Flies (SF), and Ground into Double Play (GIDP). These batting statistics are frequencies or counts, and are the basic building blocks for more complicated batting measures. Several of these batting statistics have incomplete data observations: SF is complete from the year 1954 on, CS is complete from the year 1951 on, SH is complete from the year 1894 on, HBP is complete from the year 1887 on, SB is complete from the year 1886 on. Where data was unavailable, its value was assumed to be zero following standard convention. Along with these traditional statistics, some sabermetric statistics will be included, these being: Total Bases (TB), Batting Average (BA), On Base Percentage (OBP), Slugging Average (SLG), On Base Plus Slugging (OPS), Total Average (TA), Isolated Power (ISO), Secondary Average (SECA), Runs Created (RC), and Runs Created per Game (RC27). “(Tung, 2012) These Sabermetric statistics were calculated from the previous data set of traditional statistics in the following way shown below.

$$\mathrm{TB} = \mathrm{H} + \mathrm{2B} + 2(\mathrm{3B}) + 3(\mathrm{HR}), \tag{2.1}$$

$$\mathrm{BA} = \frac{\mathrm{H}}{\mathrm{AB}}, \tag{2.2}$$

$$\mathrm{OBP} = \frac{\mathrm{H} + \mathrm{BB} + \mathrm{HBP}}{\mathrm{AB} + \mathrm{BB} + \mathrm{HBP} + \mathrm{SF}}, \tag{2.3}$$

$$\mathrm{SLG} = \frac{\mathrm{TB}}{\mathrm{AB}}, \tag{2.4}$$

$$\mathrm{OPS} = \mathrm{OBP} + \mathrm{SLG}, \tag{2.5}$$

$$\mathrm{TA} = \frac{\mathrm{TB} + \mathrm{BB} + \mathrm{HBP} + \mathrm{SB} - \mathrm{CS}}{\mathrm{AB} - \mathrm{H} + \mathrm{CS} + \mathrm{GIDP}}, \tag{2.6}$$

$$\mathrm{ISO} = \mathrm{SLG} - \mathrm{BA} = \frac{\mathrm{TB} - \mathrm{H}}{\mathrm{AB}}, \tag{2.7}$$

$$\mathrm{SECA} = \frac{\mathrm{TB} - \mathrm{H} + \mathrm{BB} + \mathrm{SB} - \mathrm{CS}}{\mathrm{AB}}, \tag{2.8}$$

$$\mathrm{RC} = \frac{\begin{array}{c}(\mathrm{H} + \mathrm{BB} + \mathrm{HBP} - \mathrm{CS} - \mathrm{GIDP}) \\ \cdot\, [\,\mathrm{TB} + 0.26(\mathrm{BB} - \mathrm{IBB} + \mathrm{HBP}) + 0.52(\mathrm{SH} + \mathrm{SF} + \mathrm{SB})]\end{array}}{\mathrm{AB} + \mathrm{BB} + \mathrm{HBP} + \mathrm{SH} + \mathrm{SF}}, \tag{2.9}$$

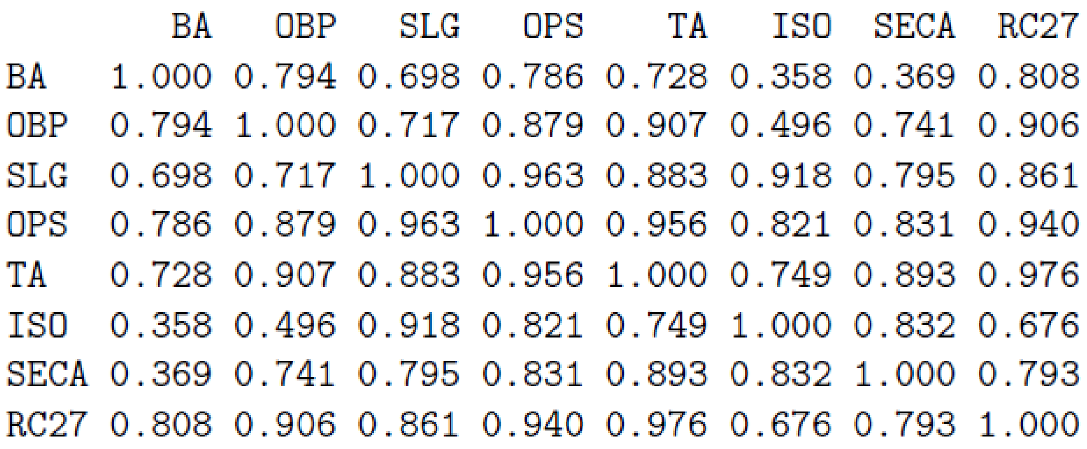
$$\mathrm{RC27} = \frac{\mathrm{RC}}{(\mathrm{AB} - \mathrm{H} + \mathrm{SH} + \mathrm{SF} + \mathrm{CS} + \mathrm{GIDP})/27}. \tag{2.10}$$

“For completeness, we will briefly summarize these batting statistics. Total Bases (TB) is the number of bases a player has gained with hits, i.e. the sum of his hits weighted by 1 for a single, 2 for a double, 3 for a triple and 4 for a home run. Batting Average (BA) is the most famous and quoted of all baseball statistics: it is the ratio of hits to at-bats, not counting walks, hit by pitcher, or sacrifices. On Base Percentage (OBP) is the classical measure for judging how good a batter is at getting on base: total number of times on base divided by the total of at-bats, walks, hit by pitcher, and sacrifice flies. Slugging Average (SLG) is the classical measure of a batter’s power hitting ability: total bases on hits divided by at-bats. The classic trio of batting statistics (BA, OBP, SLG) presented together, provide an excellent summary of a player’s offensive ability, combining the ability to get on base and to hit for power. For example, a player with (BA = 0.300, OBP = 0.400, SLG = 0.500) is considered an ideal offensive player.” (Tung, 2012) “The statistics we describe below are modern sabermetric batting measures. The ability of a player to both get on base and to hit for power, two important hitting skills, are represented in the famous sabermetric measure On Base Plus Slugging (OPS), which is obtained by simply adding OBP and SLG. OPS is a quick and dirty statistic that correlates better with runs scoring than BA, OBP, or SLG alone. Total Average (TA) is essentially a modification of SLG, and is rather similar to OPS. Isolated Power (ISO) is a measure used to evaluate a batter’s pure power hitting ability. Since OBP and SLG are highly correlated, ISO was designed as an alternative measure of a player’s ability to hit for power not confounded with his ability to get on base. Secondary Average (SECA) is a modification of ISO and TA, and a good measure of extra base ability: the ratio of bases gained from other sources (extra base hits, walks and net stolen bases) to at-bats.

Runs Created (RC) was created by Bill James and estimates the number of runs a players contributes to his team. Since RC estimates total run production, Runs Created per Game (RC27) is the conversion of RC to a rate statistic: RC is divided by an estimate of the number of games a player's offensive record represents. This is done by estimating the total number of outs and dividing by 27 (27 outs in a 9 inning baseball game). RC27 estimates the number of runs produced by a team composed solely of the player analyzed." (Tung, 2012)

Analyzing the Stats Using the K-means Algorithm

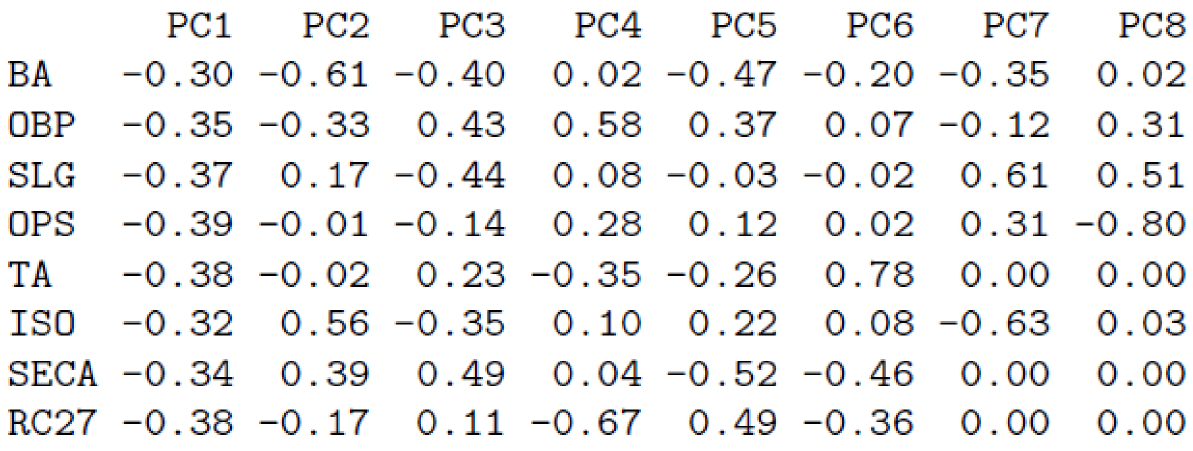
The statistics that will be analyzed using the K-means algorithm will be: BA, OBP, SLG, OPS, TA, ISO, SECA, RC27. The fully constructed data set contains 3491 players (only players with at least 1000 at-bats were considered). This set of data can be represented in a matrix with *n* rows and *p* columns, where the rows represent players as *p*-dimensional vectors, and the columns represent the variables. This means our data set will be an 8 dimensional matrix with 3491 points. Since it is difficult to visualize something in 8 dimensions, we will use the technique known as PCA (Principal Component Analysis) to reduce the dimensions down to two or three dimensions, making the data set visualization easier. The principal components are the new set of dimensions in which the first dimension is the one that retains most of the original data's variance. PCA can be implemented in the programming language R, giving us a sample correlation matrix shown on the following page.



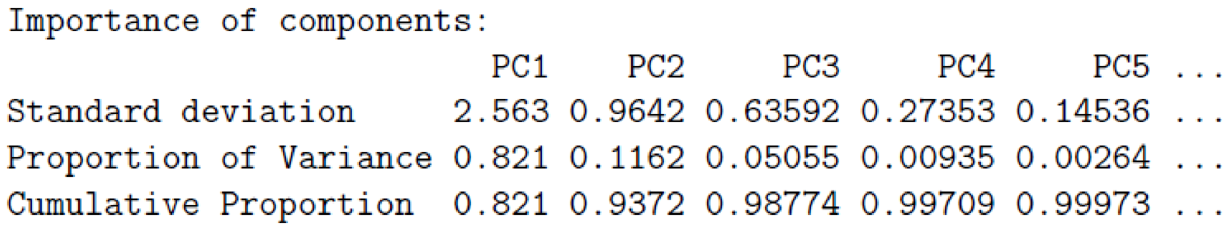
```
       BA    OBP   SLG   OPS   TA    ISO   SECA  RC27
BA   1.000 0.794 0.698 0.786 0.728 0.358 0.369 0.808
OBP  0.794 1.000 0.717 0.879 0.907 0.496 0.741 0.906
SLG  0.698 0.717 1.000 0.963 0.883 0.918 0.795 0.861
OPS  0.786 0.879 0.963 1.000 0.956 0.821 0.831 0.940
TA   0.728 0.907 0.883 0.956 1.000 0.749 0.893 0.976
ISO  0.358 0.496 0.918 0.821 0.749 1.000 0.832 0.676
SECA 0.369 0.741 0.795 0.831 0.893 0.832 1.000 0.793
RC27 0.808 0.906 0.861 0.940 0.976 0.676 0.793 1.000
```

From this two dimensional matrix we can see the correlation each statistic has with one another. Numbers closer to zero represent higher correlation. We can denote from this matrix that RC27 has very high correlation to every other statistic with the exception of ISO. We can also conclude that OBP and ISO have a weak correlation to each other.

The PCA factor loadings are the multiples of the original variables used in forming the principal components. The factor loadings rounded to 2 decimal places are shown below.



```
       PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
BA   -0.30 -0.61 -0.40  0.02 -0.47 -0.20 -0.35  0.02
OBP  -0.35 -0.33  0.43  0.58  0.37  0.07 -0.12  0.31
SLG  -0.37  0.17 -0.44  0.08 -0.03 -0.02  0.61  0.51
OPS  -0.39 -0.01 -0.14  0.28  0.12  0.02  0.31 -0.80
TA   -0.38 -0.02  0.23 -0.35 -0.26  0.78  0.00  0.00
ISO  -0.32  0.56 -0.35  0.10  0.22  0.08 -0.63  0.03
SECA -0.34  0.39  0.49  0.04 -0.52 -0.46  0.00  0.00
RC27 -0.38 -0.17  0.11 -0.67  0.49 -0.36  0.00  0.00
```



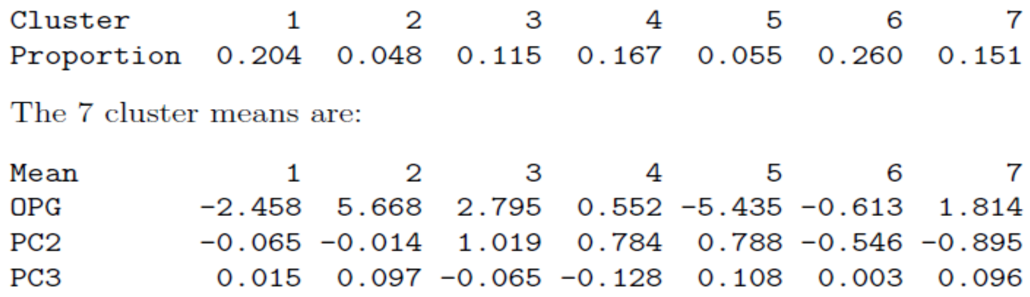
```
Importance of components:
                         PC1    PC2     PC3     PC4     PC5 ...
Standard deviation     2.563 0.9642 0.63592 0.27353 0.14536 ...
Proportion of Variance 0.821 0.1162 0.05055 0.00935 0.00264 ...
Cumulative Proportion  0.821 0.9372 0.98774 0.99709 0.99973 ...
```

Looking at the table above, the first principal component explains 82.1% of the total variability in the data. The first two principal components, combined, explain 93.72% of the total variation. The third principal component, alone, explains 5% of the sample variation; including it will give little increase in the total variance explained. The variance in the later components combined is so small that it may well mostly represent random noise in the data. One of the main objectives of PCA is the interpretation of the principal components as key underlying factors that are uncorrelated variables. Observe that the factor loadings for the first principle component are all negative and roughly equal for all the variables. The first principal component appears to be an “offensive player grade” component that grades players on a numerical scale. To a close approximation, the first principal component, written as a liner combination of the standardized variables is shown below. (Tung, 2012)

$$
\begin{aligned}
\mathrm{PC1} = {} & (-0.30)\left(\frac{\mathrm{BA} - 0.263}{0.027}\right) + (-0.35)\left(\frac{\mathrm{OBP} - 0.326}{0.036}\right) \\
& + (-0.37)\left(\frac{\mathrm{SLG} - 0.380}{0.064}\right) + (-0.39)\left(\frac{\mathrm{OPS} - 0.706}{0.094}\right) \\
& + (-0.38)\left(\frac{\mathrm{TA} - 0.651}{0.129}\right) + (-0.32)\left(\frac{\mathrm{ISO} - 0.118}{0.049}\right) \\
& + (-0.34)\left(\frac{\mathrm{SECA} - 0.222}{0.072}\right) + (-0.38)\left(\frac{\mathrm{RC27} - 4.569}{1.322}\right).
\end{aligned}
$$

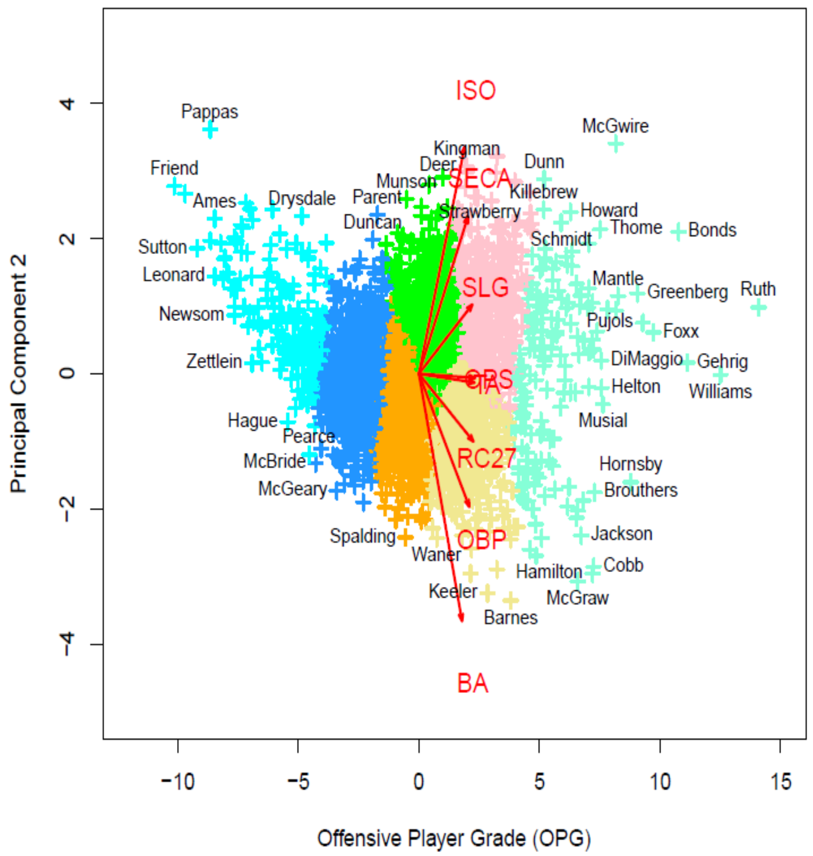
The scale for the first principal component represents better offensive players in the negatives and poor offensive players in the positives; scored around zero indicate average offensive players. However, the sign of the scores can be switched to reverse the indication of a good offensive player and a poor offensive player. For the rest of the paper we define OPG = -PC1. Now we take a look at the second principal component and notice that it clearly separates the power hitting measures ISO and SECA from the on base ability measures OBP BA. In this column, a positive score indicates a player's on base ability is better than his power hitting ability. A negative score indicates the reverse, and a score of zero indicates that a player either can hit for power and can get on base, or the player can't do either. The third component separates OBP and SECA from BA, SLG and ISO.

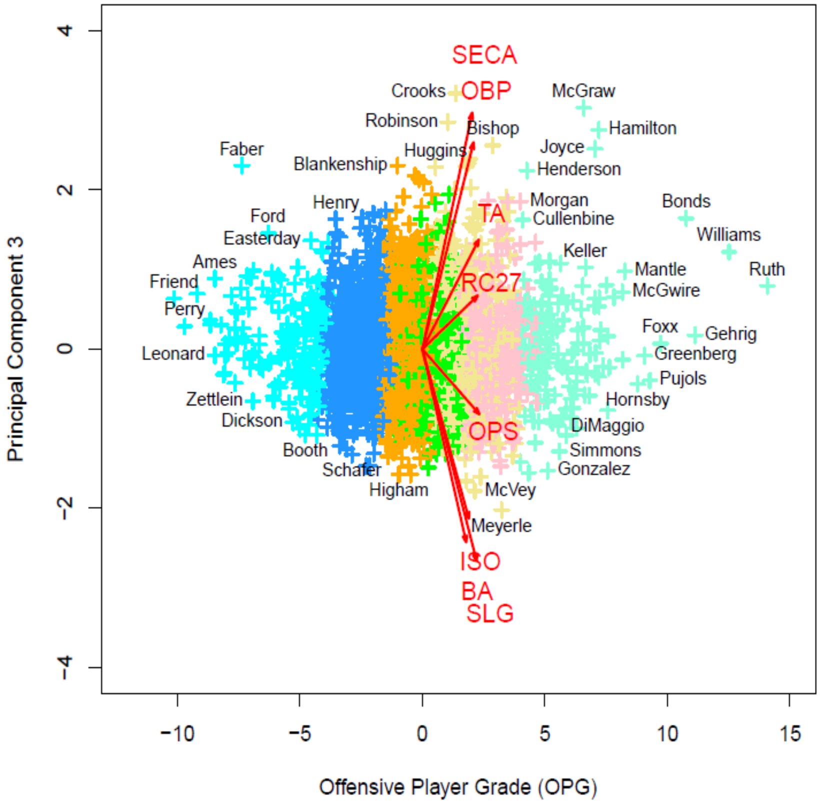
Using K-means clustering on the first three principal components, which account for 98.77% of the sample variance we can show the relative frequencies of players in each cluster. Using k = 7 to generate 7 different clusters the means are shown below.

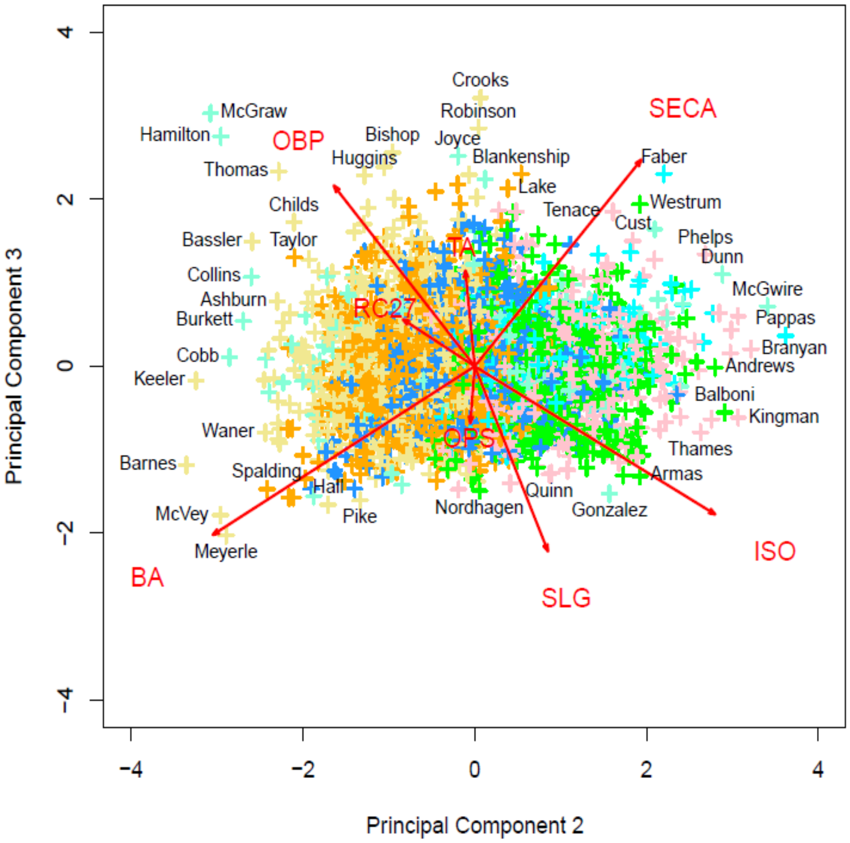


```
Cluster          1       2       3       4       5       6       7
Proportion   0.204   0.048   0.115   0.167   0.055   0.260   0.151

The 7 cluster means are:

Mean             1       2       3       4       5       6       7
OPG         -2.458   5.668   2.795   0.552  -5.435  -0.613   1.814
PC2         -0.065  -0.014   1.019   0.784   0.788  -0.546  -0.895
PC3          0.015   0.097  -0.065  -0.128   0.108   0.003   0.096
```

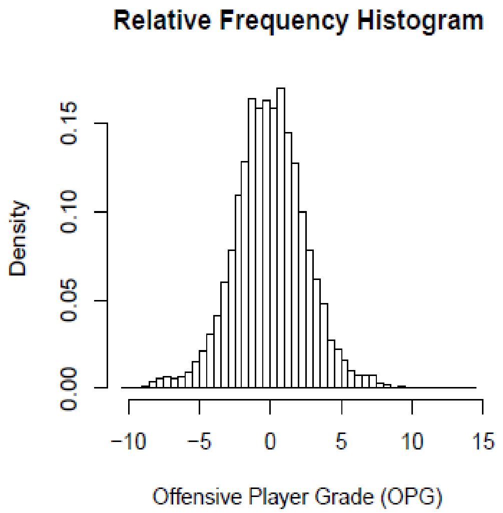
The following three images on the next page will be the graphs of the clusters of OPG and PC2, OPG and PC3, and PC2 and PC3. The colors on each graph indicate the clusters.


ISO
Pappas
McGwire
Kingman
Deer
Dunn
Friend
Munson
SECA
Ames
Drysdale
Parent
Killebrew
Strawberry
Howard
Duncan
Thome
Bonds
Sutton
Schmidt
Leonard
SLG
Mantle
Greenberg
Ruth
Newsom
Pujols
Foxx
Zettlein
DiMaggio
Gehrig
OPS
TA
Helton
Williams
Hague
Musial
Pearce
McBride
RC27
Hornsby
McGeary
Brouthers
Spalding
Jackson
OBP
Waner
Hamilton
Cobb
Keeler
McGraw
Barnes
BA
Principal Component 2
4
2
0
−2
−4
−10
−5
0
5
10
15
Offensive Player Grade (OPG)


SECA
Crooks
OBP
McGraw
Robinson
Bishop
Hamilton
Faber
Huggins
Joyce
Blankenship
Henderson
Morgan
Bonds
Henry
Cullenbine
Ford
TA
Williams
Easterday
Keller
Ames
Mantle
Ruth
Friend
RC27
McGwire
Perry
Foxx
Gehrig
Leonard
Greenberg
Pujols
Zettlein
Hornsby
Dickson
DiMaggio
OPS
Booth
Simmons
Schafer
Gonzalez
Higham
McVey
Meyerle
ISO
BA
SLG
Principal Component 3
4
2
0
−2
−4
−10
−5
0
5
10
15
Offensive Player Grade (OPG)



Looking more closely at the derived OPG statistic, we can determine that OPG effectively summarizes a players offensive performance into a single number. Since OPG was derived from the first principal component, it is a weighted average of the original eight statistics that were used. Below is a histogram of the 3491 players used in the data set and a list of the top twenty (top .5% of the sample) players ranked by OPG (granted this information was done on statistics from 2012 so the list may not be 100% accurate using current statistics).



```
rank.OPG name.first      surname       OPG
       1       Babe         Ruth  14.0779 (Hall of Fame)
       2        Ted     Williams  12.5162 (Hall of Fame)
       3        Lou      Gehrig   11.1329 (Hall of Fame)
       4      Barry       Bonds   10.7734
      5      Jimmie        Foxx    9.7236 (Hall of Fame)
      6      Albert      Pujols    9.2826 (Active)
      7        Hank   Greenberg    9.0621 (Hall of Fame)
      8      Rogers     Hornsby    8.7877 (Hall of Fame)
      9      Mickey      Mantle    8.2616 (Hall of Fame)
     10       Manny     Ramirez    8.1694 (Active)
     11        Mark     McGwire    8.1610
     12       Frank      Thomas    7.8251
     13        Stan      Musial    7.6326 (Hall of Fame)
     14        Todd      Helton    7.5810 (Active)
     15         Joe    DiMaggio    7.5650 (Hall of Fame)
     16         Jim       Thome    7.5037 (Active)
     17       Larry      Walker    7.3820
     18         Dan  Brouthers     7.2757 (Hall of Fame)
     19          Ty        Cobb    7.2506 (Hall of Fame)
     20         Mel         Ott    7.2235 (Hall of Fame)
```

So we can see that the derived statistic OPG has accurately placed players known to dominate offensively in their respective era's at the top of this list. We also determined that the statistics that were used to define OPG do accurately represent players' individual abilities with regards to the offensive aspect of baseball.

Future Research

I would like to see future research done on pitching statistics that are gathered from MLB's software PitchFx. PtichFx is a pitch tracking system that tracks velocity, movement, release point, spin, and pitch location for every pitch thrown in baseball. I think it would be interesting to see what information can be learned on the data that PitchFx collects every night. I believe this research could be beneficial to hitters immediately seeing as starting pitchers throw 80-100 pitches every start and start around 30-35 games every year. This means there are thousands of pitches that can be analyzed by a single pitcher every year. There is plenty of data out there currently and it would be interesting to see what data mining techniques could uncover.

Conclusion

Baseball statistics have been around for over a century. With the determination to integrate technology in every facet of our lives, it's no wonder technology is being integrated and applied to professional sports. Finding better ways to analyze players and determine their worth is becoming a growing trend in most modern sports. Applying data mining techniques is a unique way to analyze baseball statistics, and as we have shown, it can get effective results. We successfully derived OPG (Offensive Player Grade) from a large data set of baseball statistics using the K-means algorithm. Hopefully further research will be done analyzing different aspects of baseball including pitching and fielding.

References

Tung, David D. "Data Mining Career Batting Performances in Baseball."*Journal of Data Science* (2012): n. pag. Web.

Weiss, Gary M., Ph.D., and Brian D. Davison, Ph.D. "Data Mining."*Handbook of Technology Management* (2010): n. pag. Web.

Huimin Cui, Gong Ruan, Jingling Xue, Rui Xie, Lei Wang, and Xiaobing Feng. 2014. A collaborative divide-and-conquer K-means clustering algorithm for processing large data. In*Proceedings of the 11th ACM Conference on Computing Frontiers* (CF '14). ACM, New York, NY, USA, , Article 20 , 10 pages. DOI=10.1145/2597917.2597918 http://doi.acm.org.ezproxy.lib.csustan.edu:2048/10.1145/2597917.2597918

Woodum, B. (2014, June 1). What Is Sabermetrics? And Which Teams Use It? | FanGraphs Baseball. Retrieved April 27, 2015, from http://www.fangraphs.com/blogs/what-is-sabermetrics-and-which-teams-use-it/

Tango, Tom M., Mitchel G. Lichtman, and Andrew E. Dolphin. *The Book: Playing the Percentages in Baseball*. Washington, D.C.: Potomac, 2007. Print.