

# Data Mining Career Batting Performances in Baseball

DAVID D. TUNG<sup>1</sup>

---

## Abstract

In this paper, we use statistical data mining techniques to analyze a multivariate data set of career batting performances in Major League Baseball. Principal components analysis (PCA) is used to transform the high-dimensional data to its lower-dimensional principal components, which retain a high percentage of the sample variation, hence reducing the dimensionality of the data. From PCA, we determine a few important key factors of classical and sabermetric batting statistics, and the most important of these is a new measure, which we call Offensive Player Grade (OPG), that efficiently summarizes a player's offensive performance on a numerical scale. The determination of these lower-dimensional principal components allows for accessible visualization of the data, and for segmentation of players into groups using clustering, which is done here using the  $K$ -means clustering algorithm. We provide illuminating visual displays from our statistical data mining procedures, and we also furnish a player listing of the top 100 OPG scores which should be of interest to those that follow baseball.

*Keywords:* Data segmentation,  $K$ -means clustering, Linear dimensionality reduction, Multivariate data analysis, Principal components analysis, Sabermetrics

---

## 1. Introduction

Our objective here is a data analytic study of a multivariate data set of career batting performances in Major League Baseball (MLB) using the statistical data mining techniques principal components analysis (PCA) and

---

*Email address:* david.deming.tung@gmail.com (DAVID D. TUNG )

<sup>1</sup>Corresponding author

cluster analysis. This multivariate data set is first constructed from traditional batting statistics, then augmented with additional batting statistics and sabermetric batting measures. By sabermetrics, we mean the field of baseball analytics: the analysis of baseball through objective evidence. The term was derived from SABR, the acronym for the [Society for American Baseball Research](http://sabr.org/) (<http://sabr.org/>), by Bill James, one of its prominent pioneers. Early research in sabermetrics focused primarily on the creation of measures for individual and team offensive performance, but gradually expanded in scope to include defensive aspects of baseball, i.e. pitching and fielding. For the casual baseball fan interested in learning more about sabermetrics, the books by [Keri \(2007\)](#) and [Albert and Bennett \(2001\)](#) provide a good introduction.

PCA is a statistical data mining technique that reduces a large number of possibly correlated variables to a few key underlying factors, called principal components, that explain the variance-covariance structure of these variables. PCA can also be seen as a linear dimensionality reduction technique that projects high-dimensional data onto a lower-dimensional space without losing sample variation. From a mathematical perspective, PCA is an eigenvalue-eigenvector problem whose goal is to construct a set of orthogonal linear projections of a single set of correlated variables, where the projections are ordered by decreasing variances. An orthogonal transformation is defined in such a way that the first principal component has the largest possible variance, i.e. it accounts for as much of the variability in the data as possible, and each succeeding principal component in turn has the highest variance possible under the constraint that it be orthogonal (uncorrelated) with the preceding principal components. In the scientific literature, PCA is also known as the discrete Karhunen-Loève transform, empirical orthogonal functions, and the Hotelling transform.

Cluster analysis is a broad term for statistical data mining techniques that arrange multivariate data observations into natural groups. This is done on the basis of a measure or distance between observations. The methods of cluster analysis consists of various algorithms designed to assign sample observations into homogeneous groups or “clusters” so that the observations in the same cluster are more similar to each other than to those in other clusters. The  $K$ -means clustering algorithm is a popular method of cluster analysis that tries to minimize the sums of the squared distances between observations and the centers of clusters, so that each observation belongs to

the cluster with the nearest mean. This results in a partitioning of the data into Voronoi cells. Clustering is the most well-known example of unsupervised learning, and is philosophically different from classification, which like regression is a supervised learning technique. In classification, it is known how many classes or groups are present in the data and which observations are members of which class or group. The objective of classification is to classify new observations into one of the known classes based on a learning set of the data. In clustering, the number of classes is unknown and so is the membership of observations into classes. Clustering is also known as data segmentation and is used in many fields, including market research, biology, machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

For a more thorough introduction to PCA and cluster analysis, as well as statistical data mining, the reader is referred to [Tufféry \(2011\)](#), [Johnson and Wichern \(2007\)](#), [Hastie et al. \(2009\)](#), and [Izenman \(2008\)](#). [Wu et al. \(2008\)](#) surveys the top data mining algorithms, covering classification, clustering, statistical learning, association analysis, and link mining. [Jain \(2010\)](#) surveys research done on cluster analysis. [Bock \(2008\)](#) presents a historical view of the  $K$ -means clustering algorithm. [Biau et al. \(2008\)](#) discusses  $K$ -means clustering when sample observations take values in a separable Hilbert space. [Melnykov and Maitra \(2010\)](#) surveys model-based clustering based on finite mixture models.

Data preparation and data description is detailed in the next section. PCA and  $K$ -means clustering is seen in Section 3. In Section 4, we discuss the PCA derived measure Offensive Player Grade (OPG).

## 2. Data Preparation and Description

A multivariate data set of career batting performances for MLB players was constructed from standard batting data found in the [Lahman baseball database](#), available on the internet at <http://baseball1.com/statistics>. Version 5.9 of the database was used, which covers all seasons through the year 2011. The players' surnames and first names and ID code were queried from the 'Master' table of the database. Then player data for the following batting statistics were extracted from the 'Batting' table of the database: 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.

After transferring the extracted data to a spreadsheet, the following classical and sabermetric batting statistics were also calculated for inclusion into the data set: 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). Formulae are given below:

$$TB = H + 2B + 2(3B) + 3(HR), \quad (2.1)$$

$$BA = \frac{H}{AB}, \quad (2.2)$$

$$OBP = \frac{H + BB + HBP}{AB + BB + HBP + SF}, \quad (2.3)$$

$$SLG = \frac{TB}{AB}, \quad (2.4)$$

$$OPS = OBP + SLG, \quad (2.5)$$

$$TA = \frac{TB + BB + HBP + SB - CS}{AB - H + CS + GIDP}, \quad (2.6)$$

$$ISO = SLG - BA = \frac{TB - H}{AB}, \quad (2.7)$$

$$SECA = \frac{TB - H + BB + SB - CS}{AB}, \quad (2.8)$$

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

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

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.

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.

Only players with at least 1000 at-bats were considered in order to create a diverse collection of players for the data set, e.g. pitchers with significant batting experience, lesser known "rank and file" type players, and active players. The fully constructed data set contains 3491 observations, 27 quan-

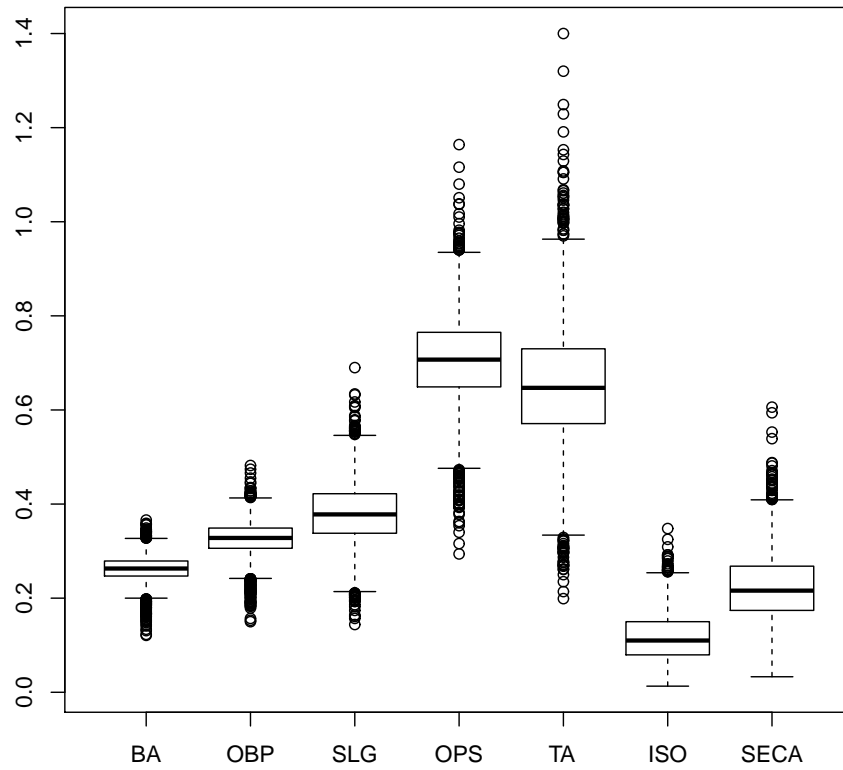
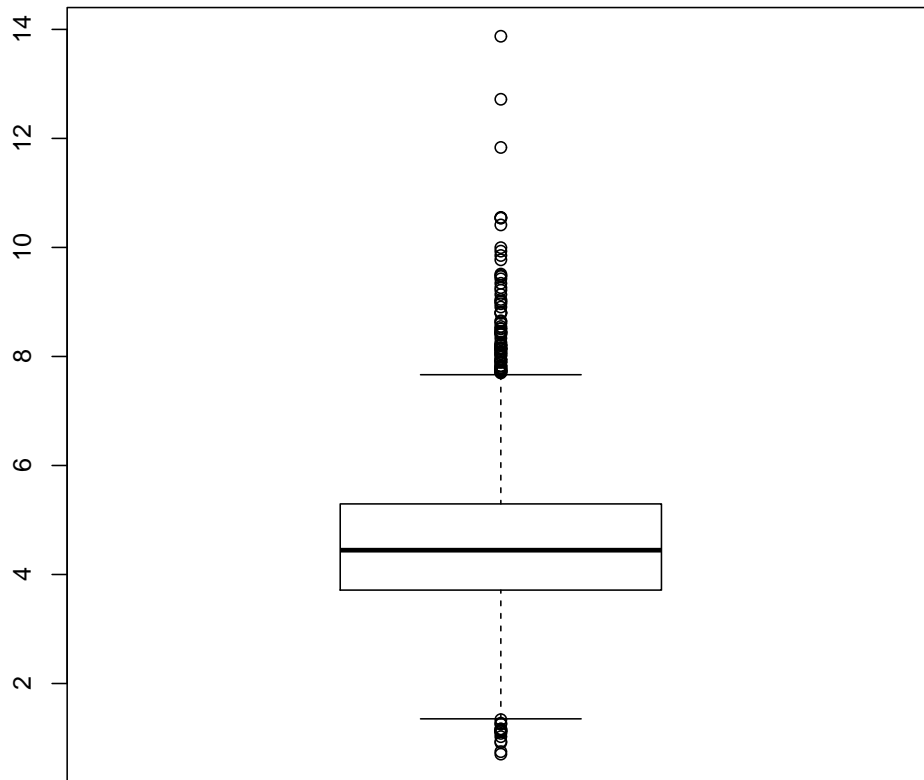


Figure 1: Boxplots of BA, OBP, SLG, OPS, TA, ISO, SECA.

titative variables, and 3 meta-variables (playerID, nameFirst, nameLast).



RC27

Figure 2: Boxplot of RC27.

### 3. Dimensionality Reduction and Clustering via Principal Components Analysis

In this section, we analyze the batting statistics (BA, OBP, SLG, OPS, TA, ISO, SECA, RC27) through PCA and clustering. For this purpose, we take the 3491 observations and these 8 variables or features as our working data set. Our data set can be represented by a data matrix  $\mathbf{X}$  with  $n$  rows and  $p$  columns, where the rows represent the observations as  $p$ -dimensional vectors, and the columns represent the variables:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}. \quad (3.1)$$

Here  $n = 3491$  and  $p = 8$ . Visualizing 3491 points in an 8-dimensional space is rather troublesome, since we are used to visualizing at most three-dimensional data. PCA is a standard technique used to reduce a large number of dimensions down to two or three dimensions for accessible visualization. The principal components are the new set of dimensions, where the first dimension is the one that retains most of the original data's variance.

We first introduce some extra notation from linear algebra and multivariate analysis. The  $n \times 1$  column vector whose entries are all 1 is denoted by

$$\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}. \quad (3.2)$$

The  $n \times n$  matrix whose entries are all 1 is denoted by

$$\mathbf{1}\mathbf{1}^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}. \quad (3.3)$$



The  $p \times 1$  vector of column means from the data matrix  $\mathbf{X}$  can be written as

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \frac{1}{n} \mathbf{X}^T \mathbf{1}. \quad (3.4)$$

The  $n \times p$  matrix of column means is denoted by

$$\frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{X} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_p \end{bmatrix}. \quad (3.5)$$

The sample variance-covariance matrix for the data matrix  $\mathbf{X}$  is a  $p \times p$  matrix defined by

$$\mathbf{S}_{\mathbf{X}} = \frac{1}{n-1} \left( \mathbf{X} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{X} \right)^T \left( \mathbf{X} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{X} \right) = \frac{1}{n-1} \mathbf{X}^T \left( \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \mathbf{X}, \quad (3.6)$$

where  $\mathbf{I}$  is the  $n \times n$  identity matrix whose entries are zero except on the diagonal where they are all 1.

In practice, variables measured on different scales or on a common scale with differing ranges are typically standardized by constructing the standardized observations

$$\mathbf{z}_i = \mathbf{D}^{-1/2} (\mathbf{x}_i - \bar{\mathbf{x}}) = \begin{bmatrix} \frac{x_{i1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \frac{x_{i2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \frac{x_{ip} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix}, \quad i = 1, 2, \dots, n, \quad (3.7)$$

where

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{s_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{s_{22}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sqrt{s_{pp}} \end{bmatrix}, \quad (3.8)$$

is the sample standard deviation matrix. We will use

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{bmatrix} = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{12}-\bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{1p}-\bar{x}_p}{\sqrt{s_{pp}}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{22}-\bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{2p}-\bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{s_{11}}} & \frac{x_{n2}-\bar{x}_2}{\sqrt{s_{22}}} & \cdots & \frac{x_{np}-\bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \left( \mathbf{X} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{X} \right) \mathbf{D}^{-1/2}, \quad (3.9)$$

to denote the standardized data matrix.

Observe that the sample variance-covariance matrix of  $\mathbf{Z}$  is the sample correlation matrix of  $\mathbf{X}$ , i.e.

$$\begin{aligned} \mathbf{S}_{\mathbf{Z}} &= \frac{1}{n-1} \left( \mathbf{Z} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{Z} \right)^T \left( \mathbf{Z} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{Z} \right) = \frac{1}{n-1} \mathbf{Z}^T \mathbf{Z} = \mathbf{D}^{-1/2} \mathbf{S}_{\mathbf{X}} \mathbf{D}^{-1/2} \\ &= \mathbf{R}. \end{aligned} \quad (3.10)$$

Since  $\mathbf{R}$  is a symmetric matrix, it has the eigendecomposition

$$\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T, \quad (3.11)$$

where  $\mathbf{Q}$  is a  $p \times p$  orthogonal matrix whose columns are the unit eigenvectors of  $\mathbf{R}$ , and  $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$  is a diagonal matrix whose entries are the eigenvalues of  $\mathbf{R}$ , which are arranged in decreasing order. Note that  $\mathbf{Q}^T = \mathbf{Q}^{-1}$ .

The sample principal components are obtained by an orthogonal transformation of the standardized data matrix, i.e.

$$\mathbf{Z} \mapsto \mathbf{Y} = \mathbf{Z}\mathbf{Q}, \quad (3.12)$$

where  $\mathbf{Q}$  is the orthogonal matrix from the eigendecomposition. Equivalently, the sample principal components can be obtained from the original data matrix  $\mathbf{X}$  by the transformation:

$$\mathbf{X} \mapsto \mathbf{Y} = \left( \mathbf{X} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{X} \right) \mathbf{D}^{-1/2} \mathbf{Q}. \quad (3.13)$$

The sample variance-covariance matrix of  $\mathbf{Y} = \mathbf{Z}\mathbf{Q}$  is given by  $\mathbf{\Lambda}$ , the

diagonal matrix from the eigendecomposition. To see this, observe that

$$\begin{aligned}
\mathbf{S}_\mathbf{Y} &= \frac{1}{n-1} \left( \mathbf{Y} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{Y} \right)^T \left( \mathbf{Y} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{Y} \right) \\
&= \frac{1}{n-1} \mathbf{Y}^T \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \mathbf{Y} \\
&= \frac{1}{n-1} (\mathbf{Z}\mathbf{Q})^T \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \mathbf{Z}\mathbf{Q} \\
&= \frac{1}{n-1} \mathbf{Q}^T \mathbf{Z}^T \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \mathbf{Z}\mathbf{Q} \\
&= \frac{1}{n-1} \mathbf{Q}^T \mathbf{Z}^T \left( \mathbf{Z} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \mathbf{Z} \right) \mathbf{Q} \\
&= \frac{1}{n-1} \mathbf{Q}^T \mathbf{Z}^T \mathbf{Z}\mathbf{Q} \\
&= \mathbf{Q}^T \mathbf{R}\mathbf{Q} \\
&= \mathbf{Q}^T (\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T) \mathbf{Q} \\
&= \mathbf{Q}^{-1} \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} \mathbf{Q} \\
&= \mathbf{\Lambda}.
\end{aligned} \tag{3.14}$$

Thus, the sample principal component variances are the eigenvalues of the sample correlation matrix  $\mathbf{R}$ .

To get a reduced-dimensionality representation of the sample principal components, we project the standardized data matrix  $\mathbf{Z}$  down onto a lower-dimensional space defined by only the first  $d$  eigenvectors ( $d \leq p$ ), i.e.

$$\mathbf{Z} \mapsto \mathbf{Y}_d = \mathbf{Z}\mathbf{Q}_d, \tag{3.15}$$

where  $\mathbf{Y}_d$  is an  $n \times d$  matrix,  $\mathbf{Q}_d = \mathbf{Q}\mathbf{I}_d$  and  $\mathbf{I}_d$  is the  $p \times d$  rectangular identity matrix. The sample variance-covariance matrix of  $\mathbf{Y}_d$  is a  $d \times d$  matrix given by

$$\mathbf{S}_{\mathbf{Y}_d} = \frac{1}{n-1} \mathbf{Y}_d^T \left( \mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \mathbf{Y}_d = \mathbf{I}_d^T \mathbf{\Lambda} \mathbf{I}_d. \tag{3.16}$$

PCA can be implemented in R, a free software environment for statistical computing and graphics (R Development Core Team (2011)) using the purpose-built function `prcomp()`.

The sample correlation matrix for the data is:

	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

It is seen that BA seems to have a very weak positive association with ISO and SECA. OBP and ISO have weak positive association. Runs Created per Game (RC27) generally correlates very well with all the other batting statistics. OPS correlates better with RC27 than any of the classic trio (BA, OBP, SLG) alone.

The column means of the data matrix are given by:

BA	OBP	SLG	OPS	TA	ISO	SECA	RC27
0.263	0.326	0.380	0.706	0.651	0.118	0.222	4.569

and the entries of the sample standard deviation matrix are given by:

BA	OBP	SLG	OPS	TA	ISO	SECA	RC27
0.027	0.036	0.064	0.094	0.129	0.049	0.072	1.322

The PCA factor loadings (the entries of the eigenvectors of the sample correlation matrix) are the multiples of the original variables used in forming the principal components. The factor loadings rounded to 2 decimal places are:

	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

A summary of the PCA performed in R shows

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	...

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 about 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. The principal component standard deviations shown are the square roots of the eigenvalues of the sample correlation matrix.

One of the main objectives of PCA is the interpretation of the principal components as key underlying factors that are uncorrelated variables (orthogonal measures). Observe that the factor loadings for the first principal 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 linear combination of the standardized variables, is:

$$\begin{aligned}
 \text{PC1} = & (-0.30) \left( \frac{\text{BA} - 0.263}{0.027} \right) + (-0.35) \left( \frac{\text{OBP} - 0.326}{0.036} \right) \\
 & + (-0.37) \left( \frac{\text{SLG} - 0.380}{0.064} \right) + (-0.39) \left( \frac{\text{OPS} - 0.706}{0.094} \right) \\
 & + (-0.38) \left( \frac{\text{TA} - 0.651}{0.129} \right) + (-0.32) \left( \frac{\text{ISO} - 0.118}{0.049} \right) \\
 & + (-0.34) \left( \frac{\text{SECA} - 0.222}{0.072} \right) + (-0.38) \left( \frac{\text{RC27} - 4.569}{1.322} \right). \quad (3.17)
 \end{aligned}$$

The scale for the first principal component is such that larger negative scores indicate better offensive players; larger positive scores indicate poorer offensive players; scores in a neighborhood of zero indicate an average offensive player. Note that the sign of the scores can easily be switched so that positive scores indicate good offensive players, and negative scores indicate poor

offensive players: the negative signs are just an inconsequential artifact from the PCA numerical computations. Here and throughout, we define  $OPG = -PC1$  to be the Offensive Player Grade statistic, and use the formula

$$\begin{aligned}
OPG = & 0.30 \left( \frac{BA - 0.263}{0.027} \right) + 0.35 \left( \frac{OBP - 0.326}{0.036} \right) \\
& + 0.37 \left( \frac{SLG - 0.380}{0.064} \right) + 0.39 \left( \frac{OPS - 0.706}{0.094} \right) \\
& + 0.38 \left( \frac{TA - 0.651}{0.129} \right) + 0.32 \left( \frac{ISO - 0.118}{0.049} \right) \\
& + 0.34 \left( \frac{SECA - 0.222}{0.072} \right) + 0.38 \left( \frac{RC27 - 4.569}{1.322} \right). \quad (3.18)
\end{aligned}$$

In the next section, we will see that the OPG statistic is useful for summarizing a batter's overall offensive performance with just one single number.

The factor loadings for the second principal component indicate that the second principal component clearly separates the power hitting measures ISO and SECA from the on base ability measures OBP and BA. Here, a positive score indicates a player's power hitting ability is better than his on base ability; a negative score indicates a player's on base ability is better than his power hitting ability. Scores in a neighborhood of zero are a mixed bag and indicate a player: (1) has a combination of power hitting ability and on base ability, or (2) has neither on base ability nor power hitting ability.

The factor loadings for the third principal component indicate that the third principal component clearly separates OBP and SECA from BA, SLG, and ISO. The third component seems to separate those measures that depend on bases obtained via other sources (BB, HBP, and SB) from those measures that do not.

PCA is able to reduce the dimensionality of the data set from 3491 observations on 8 variables to 3491 observations on 2 principal components while retaining 93.72% of the sample variance, which is a very satisfactory result. Including the third principal component would retain 98.77% of the sample variance. A PCA biplot gives us an accessible two-dimensional visualization of the data. In Figures 3-5, PCA biplots display the 8-dimensional data projected down onto the lower-dimensional principal components.

The PCA projection of high-dimensional data onto a convenient lower-dimensional space also provides an opportunity for data segmentation. Segmentation will add detail and structure to the PCA biplots. In practice, it

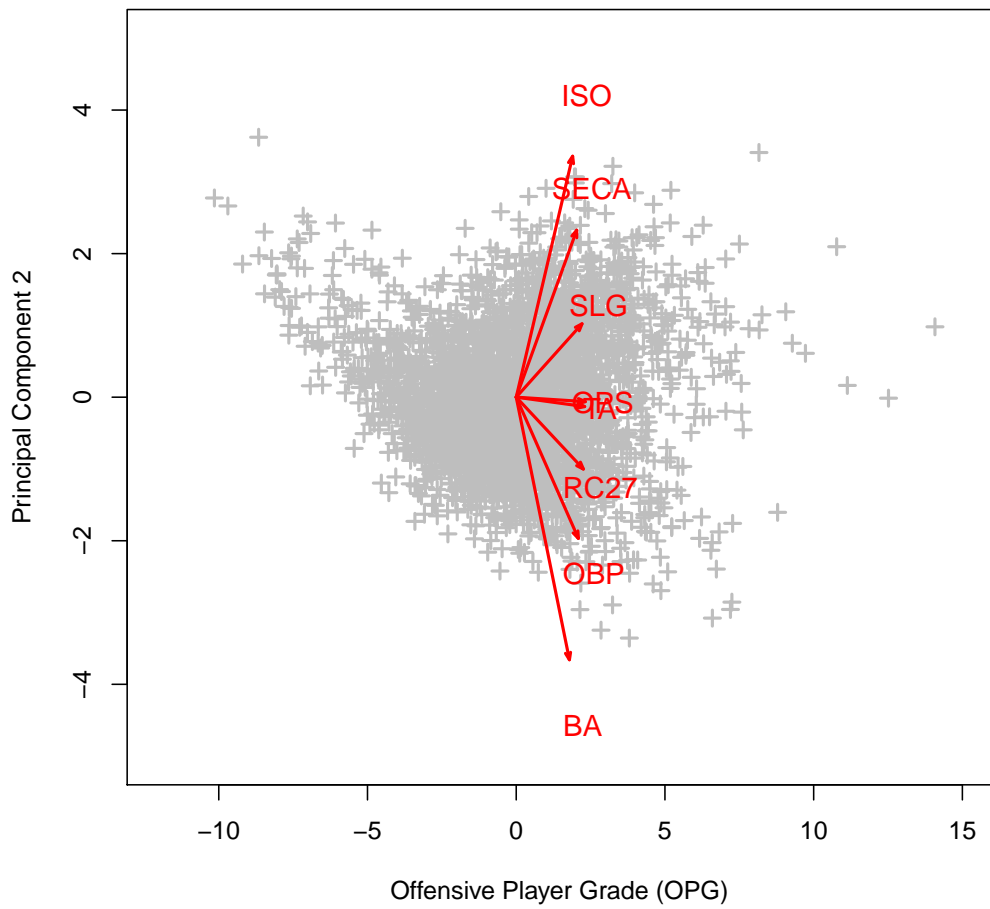


Figure 3: PCA Biplot of OPG vs. PC2.

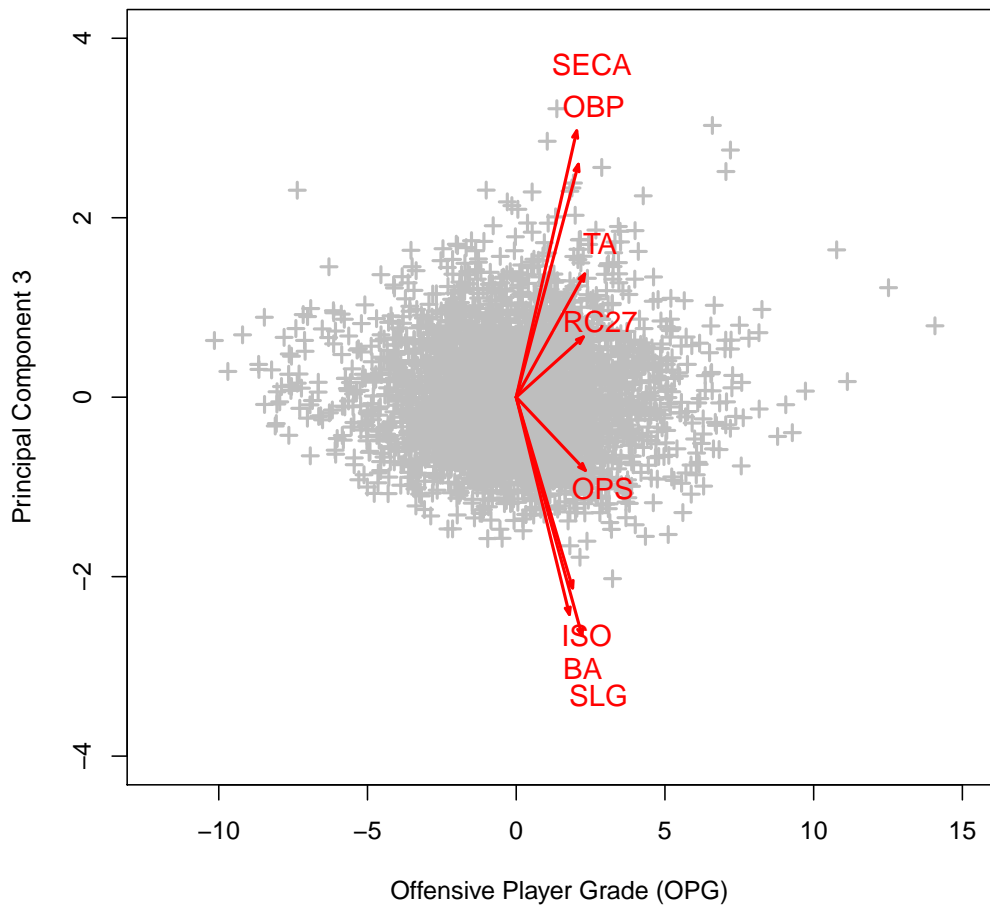


Figure 4: PCA Biplot of OPG vs. PC3.



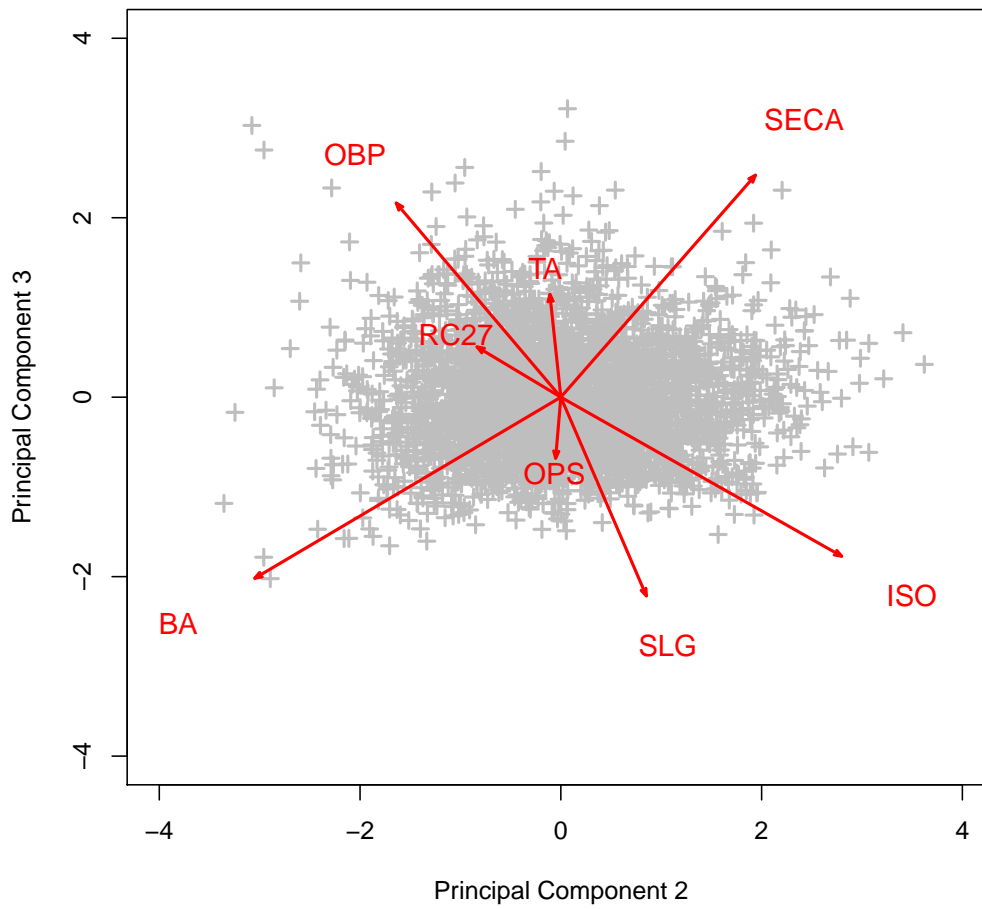


Figure 5: PCA Biplot of PC2 vs. PC3.

is quite common that PCA is used to project high-dimensional data onto a lower-dimensional space, then have  $K$ -means clustering be applied in a PCA subspace. The papers [Ding and He \(2004a\)](#) and [Ding and He \(2004b\)](#) show that PCA is the continuous solution to  $K$ -means clustering; suggest using PCA as a basis for clustering. Automatic segmentation of the players into groups can be obtained by applying the  $K$ -means clustering algorithm on the first few principal components that account for the lion's share of the sample variance.

Recall that the reduced-dimensionality representation of the sample principal components is the  $n \times d$  data matrix

$$\mathbf{Y}_d = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1d} \\ y_{21} & y_{22} & \cdots & y_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nd} \end{bmatrix}. \quad (3.19)$$

We use  $\mathbf{y}_i$  to denote the  $d \times 1$  column vector whose transpose represents a row from  $\mathbf{Y}_d = \mathbf{Z}\mathbf{Q}_d$ , instead of from the  $n \times p$  matrix  $\mathbf{Y} = \mathbf{Z}\mathbf{Q}$ . Let  $\mathcal{C} = \{C_1, \dots, C_K\}$  be a collection of disjoint subsets (a partition) of  $\Omega = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n\}$ , and let  $\mu_1, \dots, \mu_K$  be the empirical means of  $C_1, \dots, C_K$ , respectively. Note that

$$(\Omega, \mathcal{C}; \mu_1, \dots, \mu_K) = \biguplus_{k=1}^K (C_k; \mu_k). \quad (3.20)$$

The within-cluster sum of squares for  $C_k$  is the sum of the squared errors between  $\mu_k$  and the observations in  $C_k$ , which is given by

$$J(C_k; \mu_k) = \sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \mu_k\|^2, \quad k = 1, 2, \dots, K. \quad (3.21)$$

The objective function in the  $K$ -means clustering algorithm is the total within-cluster sum of squares over all  $K$  clusters, which is given by

$$J(\Omega, \mathcal{C}; \mu_1, \dots, \mu_K) = \sum_{k=1}^K \sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \mu_k\|^2. \quad (3.22)$$

The  $K$ -means clustering algorithm tries to find a partition  $\mathcal{S} = \{S_1, \dots, S_K\}$  of the data, with cluster means  $m_1, \dots, m_K$ , that minimizes the objective

function, i.e.

$$J(\Omega, \mathcal{S}; m_1, \dots, m_K) = \min_K \min_{\mu_1, \dots, \mu_K} \sum_{k=1}^K \sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \mu_k\|^2. \quad (3.23)$$

A cluster becomes more homogeneous as its within-cluster sum of squares decreases; clustering of the sample observations gets better as the total within-cluster sum of squares decreases.

We use  $K$ -means clustering on the first  $d = 3$  principal components, which account for 98.77% of the sample variance. This can be implemented in R using the purpose-built function `kmeans()`. In  $K$ -means clustering, the number of clusters  $K$  is a tuning parameter chosen before the algorithm is implemented, and we have chosen to use  $K = 7$  clusters. The relative frequencies of players in each cluster are:

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

In Figures 6-8, we display PCA biplots of the principal component scores and the clusters obtained from  $K$ -means clustering. Note that each color denotes a cluster. We have identified some ballplayers with their corresponding points for several clusters. Players found within a cluster appear to be more similar than players found in other clusters. For the biplot of the first two principal components, which explain 93.72% of the sample variance, we get a very clear separation of the clusters. For the biplot of the first and third principal components, which together explain 87.15% of the sample variance, we get decent separation for only a few clusters. For the biplot of the second and third principal components, which together explain only 16.67% of the sample variance, we do not get clear separation of the clusters.

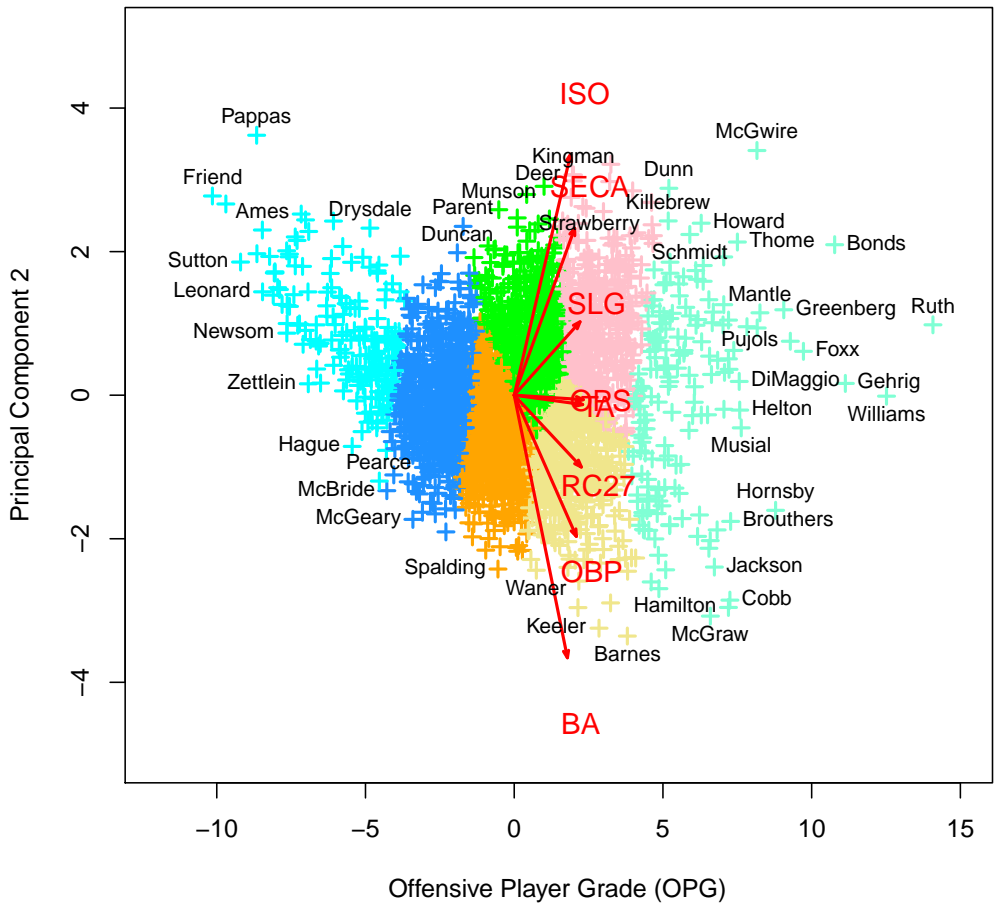


Figure 6: PCA Biplot of OPG vs. PC2 obtained from *K*-means clustering.

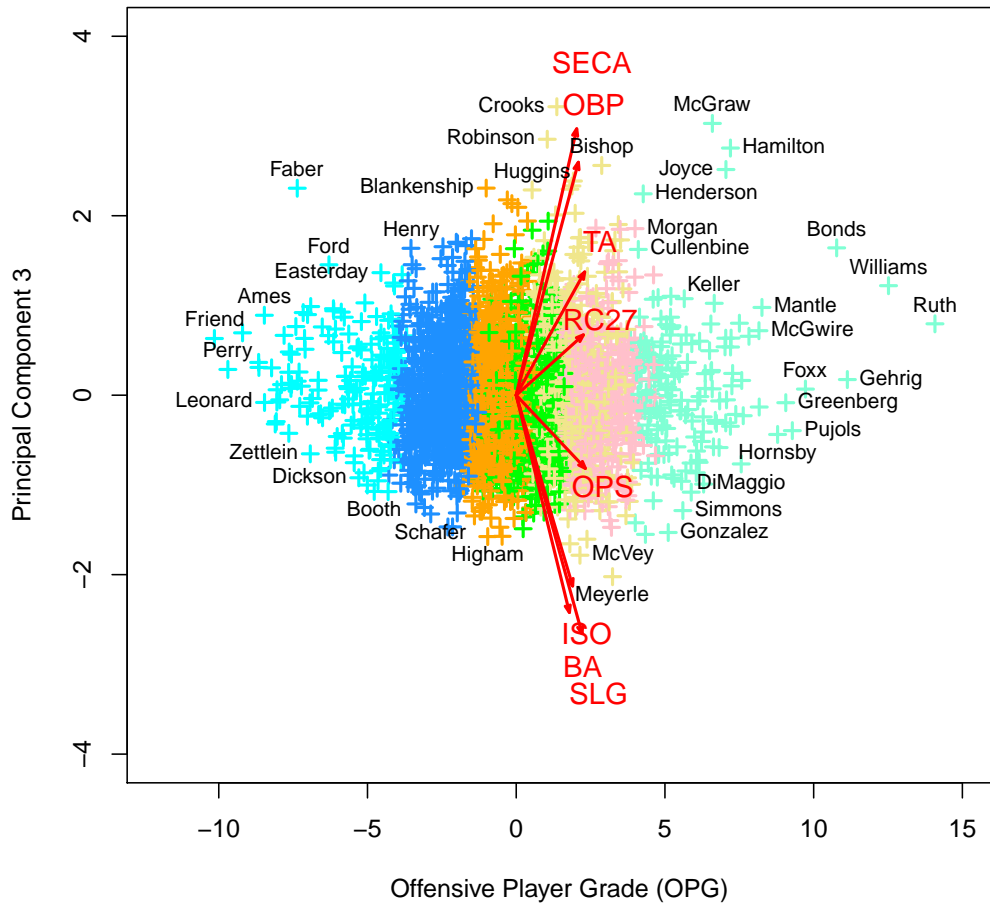


Figure 7: PCA Biplot of OPG vs. PC3 obtained from *K*-means clustering.

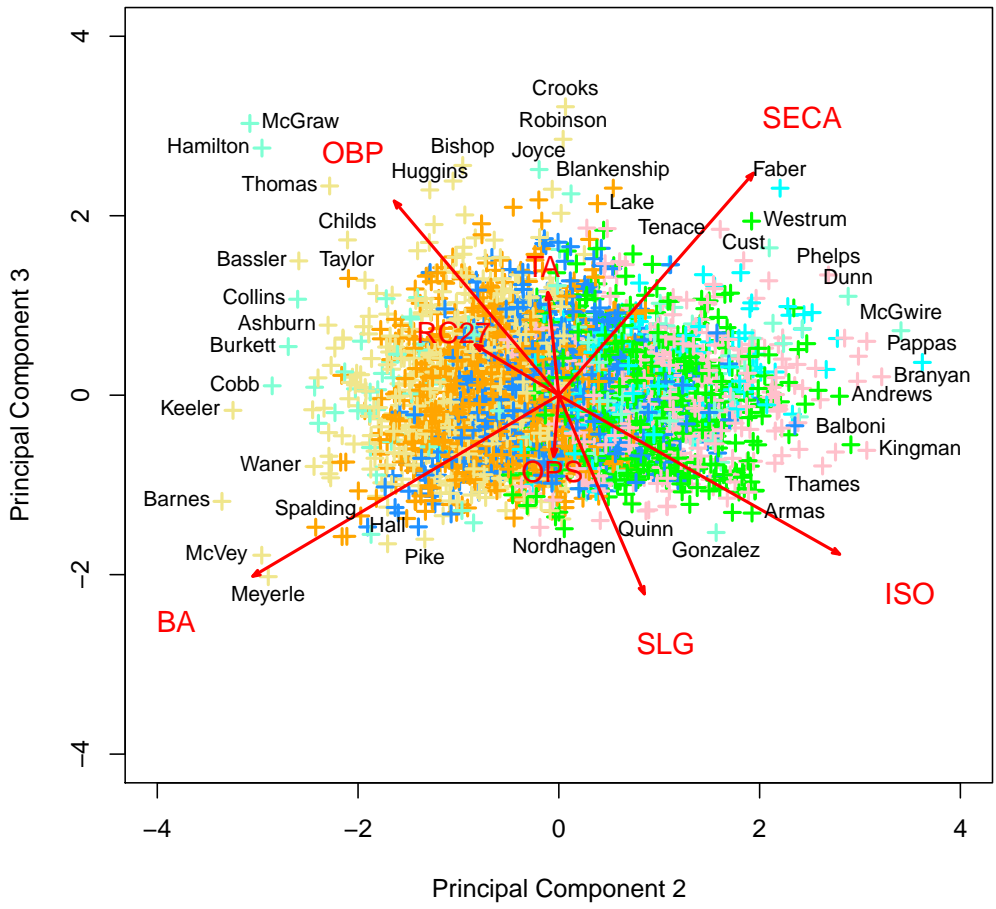


Figure 8: PCA Biplot of PC2 vs. PC3 obtained from *K*-means clustering.

#### 4. Offensive Player Grade (OPG)

This section is devoted to the PCA derived measure Offensive Player Grade. Recall that the formula for OPG is given by

$$\begin{aligned} \text{OPG} = & 0.30 \left( \frac{\text{BA} - 0.263}{0.027} \right) + 0.35 \left( \frac{\text{OBP} - 0.326}{0.036} \right) \\ & + 0.37 \left( \frac{\text{SLG} - 0.380}{0.064} \right) + 0.39 \left( \frac{\text{OPS} - 0.706}{0.094} \right) \\ & + 0.38 \left( \frac{\text{TA} - 0.651}{0.129} \right) + 0.32 \left( \frac{\text{ISO} - 0.118}{0.049} \right) \\ & + 0.34 \left( \frac{\text{SECA} - 0.222}{0.072} \right) + 0.38 \left( \frac{\text{RC27} - 4.569}{1.322} \right). \end{aligned} \quad (4.1)$$

While there is an abundance of measures available for evaluating either a player's on base ability or power hitting ability or run scoring ability, until now, there has not been a single measure available that can describe a player's overall offensive performance on a numerical scale. The OPG statistic does exactly that: it collects all the fine little details we care about, and paints an overall picture of the player being analyzed. OPG correlates better with RC27 than any of the classical and sabermetric batting measures, with the exception of TA:

	BA	OBP	SLG	OPS	TA	ISO	SECA	RC27	OPG
BA	1.000	0.794	0.698	0.786	0.728	0.358	0.369	0.808	0.764
OBP	0.794	1.000	0.717	0.879	0.907	0.496	0.741	0.906	0.893
SLG	0.698	0.717	1.000	0.963	0.883	0.918	0.795	0.861	0.945
OPS	0.786	0.879	0.963	1.000	0.956	0.821	0.831	0.940	0.993
TA	0.728	0.907	0.883	0.956	1.000	0.749	0.893	0.976	0.983
ISO	0.358	0.496	0.918	0.821	0.749	1.000	0.832	0.676	0.810
SECA	0.369	0.741	0.795	0.831	0.893	0.832	1.000	0.793	0.869
RC27	0.808	0.906	0.861	0.940	0.976	0.676	0.793	1.000	0.964
OPG	0.764	0.893	0.945	0.993	0.983	0.810	0.869	0.964	1.000

The OPG statistic grades players on a numerical scale and efficiently summarizes a player's offensive performance into a single number. Since it is derived from the first principal component, it is a weighted average of the well-known classical and sabermetric batting statistics. Recall that OPG

= -PC1, so larger positive OPG scores indicate better offensive players; larger negative OPG scores indicate poorer offensive players; OPG scores in a neighborhood of zero indicate an average offensive player. Within reason, one can compare the offensive performances between two comparable players on the basis of OPG.

To get the most out of the OPG statistic, a partition of the range of values for OPG should be established in a meaningful way. One might think about using the Empirical Rule of the Normal distribution (the so-called “Three-Sigma Rule”) to establish a meaningful partition. The relative frequency histogram of the OPG scores appears to show an approximate Normal distribution, but this may be misleading. In fact, the non-linearity in the Q-Q plot indicates an obvious departure from Normality. Moreover, the Shapiro-Wilk test concludes that the OPG scores are not distributed according to the Normal distribution (the p-value is smaller than 0.01).

Since the Empirical Rule may not be reliable, a sensible and robust solution is to partition the range of values for OPG according to the sample quantiles. For example, we can compute the sample deciles of the OPG scores:

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
-10.1	-3.0	-2.0	-1.3	-0.7	0.0	0.6	1.2	2.0	3.1	14.1

Then a possible partition of the range for OPG might go like this:

A = [3, Infinity),  
 B = [2, 3),  
 C = (-2, 2),  
 D = (-3, -2],  
 Fail = (-Infinity, -3].

Here is a player listing of the top 100 OPG scores (approximately the top 3% of the sample). We also indicate which players are in the National Baseball Hall of Fame in Cooperstown, New York, USA.

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	



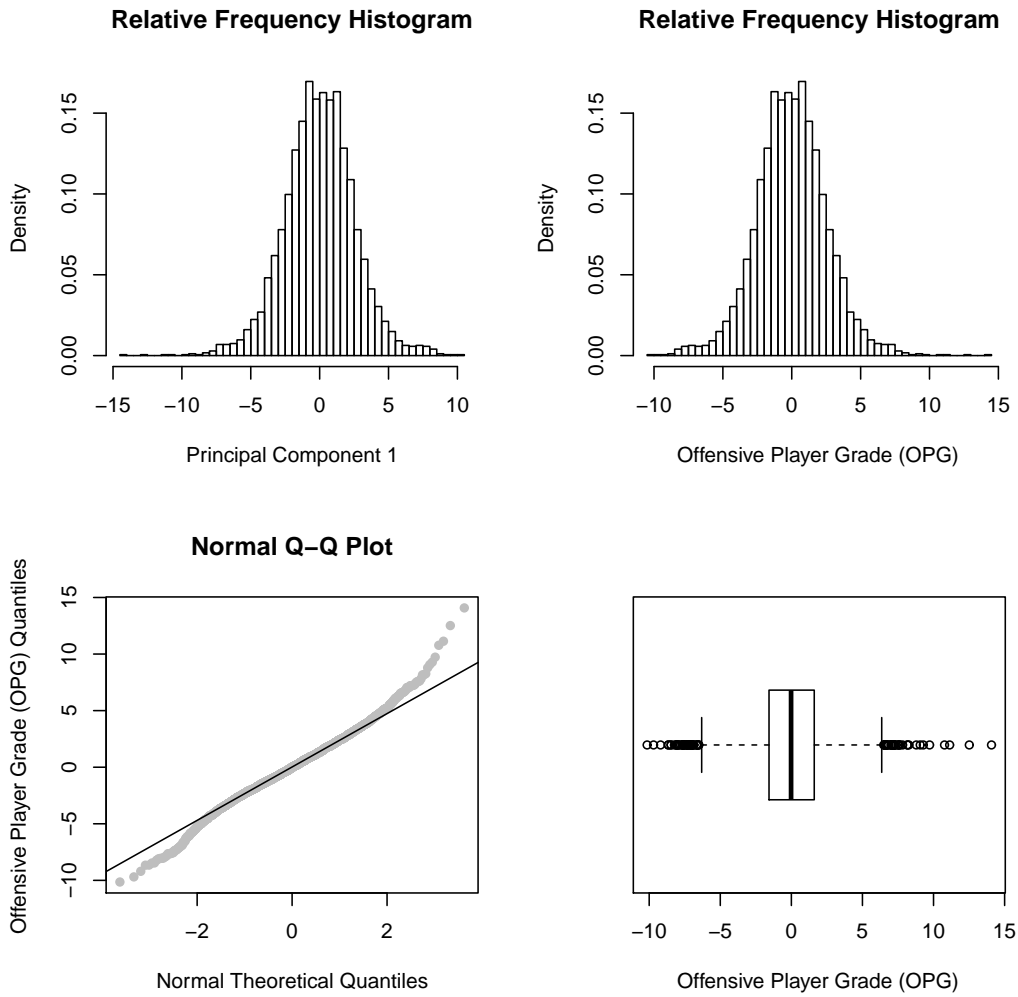


Figure 9: Various plots related to OPG.

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)
21	Lance	Berkman	7.2046 (Active)
22	Billy	Hamilton	7.2019 (Hall of Fame)
23	Johnny	Mize	7.2006 (Hall of Fame)
24	Joey	Votto	7.0695 (Active)
25	Bill	Joyce	7.0508
26	Alex	Rodriguez	7.0444 (Active)
27	Ralph	Kiner	7.0441 (Hall of Fame)
28	Jeff	Bagwell	7.0203
29	Hack	Wilson	6.8392 (Hall of Fame)
30	Lefty	O'Doul	6.8214
31	Joe	Jackson	6.7296
32	Willie	Mays	6.6935 (Hall of Fame)
33	Charlie	Keller	6.6574
34	Miguel	Cabrera	6.6314 (Active)
35	John	McGraw	6.5950 (Hall of Fame)
36	Ed	Delahanty	6.5901 (Hall of Fame)
37	Chipper	Jones	6.5667 (Active)
38	Tris	Speaker	6.5457 (Hall of Fame)
39	Jason	Giambi	6.5442 (Active)
40	Edgar	Martinez	6.5000
41	Carlos	Delgado	6.3560
42	Prince	Fielder	6.3533 (Active)

43	Ryan	Braun	6.3004 (Active)
44	Earl	Averill	6.2896 (Hall of Fame)
45	Ryan	Howard	6.2862 (Active)
46	Harry	Heilmann	6.2295 (Hall of Fame)
47	Frank	Robinson	6.1748 (Hall of Fame)
48	Jake	Stenzel	6.1642
49	Albert	Belle	6.1516
50	Hank	Aaron	6.1138 (Hall of Fame)
51	David	Ortiz	6.1035 (Active)
52	Matt	Holliday	6.0632 (Active)
53	Chuck	Klein	6.0539 (Hall of Fame)
54	Ken	Williams	5.9986
55	Duke	Snider	5.9154 (Hall of Fame)
56	Mike	Schmidt	5.9030 (Hall of Fame)
57	Vladimir	Guerrero	5.8839 (Active)
58	Babe	Herman	5.8767
59	Dick	Allen	5.7682
60	Gary	Sheffield	5.7568
61	Brian	Giles	5.7136
62	Mike	Piazza	5.7051
63	Ken	Griffey, Jr.	5.6926
64	Bill	Lange	5.6499
65	Al	Simmons	5.6024 (Hall of Fame)
66	Josh	Hamilton	5.5588 (Active)
67	Jim	Edmonds	5.5535
68	Mickey	Cochrane	5.5477 (Hall of Fame)
69	Mark	Teixeira	5.5338 (Active)
70	Mo	Vaughn	5.5059
71	Bob	Johnson	5.4876
72	Roger	Connor	5.4410 (Hall of Fame)
73	Sam	Thompson	5.4048 (Hall of Fame)
74	Travis	Hafner	5.3173 (Active)
75	Chick	Hafey	5.3053 (Hall of Fame)
76	Eddie	Mathews	5.2405 (Hall of Fame)
77	Willie	McCovey	5.2385 (Hall of Fame)
78	Dolph	Camilli	5.2351
79	George	Selkirk	5.2302
80	Ray	Grimes	5.2292

81	Bill	Terry	5.2280 (Hall of Fame)
82	Adam	Dunn	5.1998 (Active)
83	Bobby	Abreu	5.1885 (Active)
84	Hanley	Ramirez	5.1834 (Active)
85	Jackie	Robinson	5.1796 (Hall of Fame)
86	Harmon	Killebrew	5.1783 (Hall of Fame)
87	David	Wright	5.1596 (Active)
88	Hal	Trosky	5.1298
89	Juan	Gonzalez	5.1106
90	Chase	Utley	5.1017 (Active)
91	Tim	Salmon	5.0974
92	Pete	Browning	5.0931
93	Charlie	Gehring	5.0609 (Hall of Fame)
94	Goose	Goslin	5.0604 (Hall of Fame)
95	Mike	Tiernan	5.0510
96	Kevin	Youkilis	5.0477 (Active)
97	Willie	Stargell	5.0314 (Hall of Fame)
98	Fred	McGriff	5.0305
99	Rafael	Palmeiro	4.9746
100	Dale	Alexander	4.9679

## 5. Conclusion

We used statistical data mining techniques to explore a multivariate data set of career batting performances in Major League Baseball. Through PCA, the data was transformed to its principal components, reducing the dimensionality of the data. We determined a few important factors of classical and sabermetric batting statistics, and the most important of these is Offensive Player Grade (OPG), which efficiently summarizes a player's offensive performance into a single number. The PCA projection of the high-dimensional data onto its lower-dimensional principal components also allowed for the automatic segmentation of players into groups using  $K$ -means clustering. Graphical displays were also provided from our statistical data mining procedures. A player listing of the top 100 OPG scores was provided.

As far as future research directions are concerned, one can try to find measures analogous to OPG for the defensive aspects of baseball, i.e. pitching and fielding. Obtaining statistics to summarize player pitching performance,

and player fielding performance are of great interest. These might be called Player Pitching Grade (PPG) and Player Fielding Grade (PFG), respectively. The methods employed here can be applied to other sports, e.g. basketball, football, ice hockey, etc. Statistical data mining alternatives to PCA include Independent Components Analysis (ICA) and Projection Pursuit (PP), cf. [Hastie et al. \(2009\)](#).

Albert, J., Bennett, J., 2001. *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*. Copernicus Books.

Biau, G., Devroye, L., Lugosi, G., 2008. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory* 54, No. 2, 781–790.

Bock, H., 2008. Origins and extensions of the  $k$ -means algorithm in cluster analysis. *Electronic Journal for History of Probability and Statistics* 4, No. 2.

Ding, C., He, X., 2004a.  $K$ -means clustering via principal component analysis. In: *Proceedings of the 21st International Conference on Machine Learning*. Vol. 69. ACM Press, pp. 225232.

Ding, C., He, X., 2004b. Principal component analysis and effective  $K$ -means clustering. In: *Proceedings of the 2004 SIAM International Conference on Data Mining*. Vol. 2, No. 2. Society for Industrial and Applied Mathematics, pp. 497–501.

Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Edition. Springer.

Izenman, A. J., 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer.

Jain, A., 2010. Data clustering: 50 years beyond  $K$ -means. *Pattern Recognition Letters* 31, 651–666.

Johnson, R. A., Wichern, D. W., 2007. *Applied Multivariate Statistical Analysis*, Sixth Edition. Prentice Hall.

Keri, J., 2007. *Baseball Between the Numbers: Why Everything You Know about the Game Is Wrong*. Perseus Publishing.

- Melnykov, V., Maitra, R., 2010. Finite mixture models and model-based clustering. *Statistics Surveys* 4, 80–116.
- R Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.  
URL <http://www.R-project.org/>
- Tufféry, S., 2011. *Data Mining and Statistics for Decision Making*. Wiley.
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z., Steinbach, M., Hand, D. J., Steinberg, D., 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems* 14, 1–37.