Using Cluster Analysis to Uncover a Pattern of Past World Series Winners in MLB

Lucas Ben

Motivation

When reviewing what I have learned in our multivariate analysis class I was motivated to apply the methods we have discussed to the movement of security prices in financial markets. However, this question pertains to time series analysis, and I was forced to reconsider.

I concluded that baseball lends itself well to data science – repetitive events where each outcome is assumed to be a normal independent and identically distributed random variable. So, I followed this chain of thought and settled on researching what determines a World Series winner. With this research question, my analysis will address whether past World Series winners can be grouped together by team offensive statistics.

Background

Analyzing Major League Baseball (MLB) using statistical learning methods has seen a recent increase in popularity. Measuring performance using sabermetrics has been a part of baseball's tradition with Bill James pioneering baseball research in the 1980s and the theory of Moneyball popularized in the early 2000s (Lewis, 2011). This interest is driven by maximizing on-field performance with undervalued talent.

An area of interest is the analysis of team performance and the identification of winning strategies. MLB teams have long been interested in understanding the factors that contribute to championships. My research aims to explore what factors, batting and pitching metrics, can explain the likelihood of a team winning the World Series. This question is broad and lacks a response variable. By using unsupervised learning methods like principal components analysis to reduce unwanted noise and clustering methods to define distinct regions, I intend to assess whether past World Series winners exhibit distinct patterns to their success.

Data Collection

The data used for this research was gathered from Baseball Reference and consisted of two datasets of season team batting and season team pitching data. These data were compiled into a data frame of 1414 observations and 68 features and had no missing values. Each instance in this data represents an MLB team from 1975 to 2024.

Accounting for potential outliers caused by lockout shortened seasons and the COVID-19 shortened season, observations from the 1994 and 2000 seasons were removed. Albeit observations from the strike shortened 1981 and 1995 seasons were kept because of their larger game totals, 111 and 144. A subset of the categorical features was removed and saved for reference and the retained features were transformed to numeric variables. Moreover, only counting statistics were retained to remove high correlation with rate statistics and all features

were transformed into per game statistics to compensate that not all observations have played 162 games in a season. After cleaning, the final dataset consisted of 1356 observations and 36 features. Table 1 displays the features studied.

Table 1: Description of variables in the dataset

Feature	Description	Type
Bat.	Batters used	Numeric
PA	Plate appearance	Numeric
AB	At bat	Numeric
R	Run	Numeric
H	Hit	Numeric
X1B	Single	Numeric
X2B	Double	Numeric
X3B	Triple	Numeric
HR	Home Run	Numeric
RBI	Runs batted in	Numeric
SB	Stolen base	Numeric
CS	Caught stealing	Numeric
BB	Walk	Numeric
SO	Strikeout (offensive)	Numeric
TB	Total bases	Numeric
GIDP	Grounded into double play	Numeric
HBP	Hit by pitch	Numeric
SH	Sacrifice hit	Numeric
SF	Sacrifice fly	Numeric
IBB	Intentional walk	Numeric
LOB	Left on base	Numeric
CG	Complete game	Numeric
SHO	Shutout	Numeric
SV	Save	Numeric
IP	Innings pitched	Numeric
HA	Hits allowed	Numeric
RA	Runs allowed	Numeric
ER	Earned runs allowed	Numeric
HRA	Home runs allowed	Numeric
WA	Walks allowed	Numeric
IBA	Intentional walks allowed	Numeric
SO P	Strikeouts (defensive)	Numeric
HBPA	Hit by pitch allowed	Numeric
BK	Balk	Numeric
WP	Wild pitch	Numeric
BF	Batters faced	Numeric

Methods

Principal components analysis

Principal components analysis (PCA) is a dimensionality reduction method used to summarize data with n observations and p correlated features. PCA operates by forming uncorrelated linear combinations of the p features whose variance is as large as possible. The first two principal components are obtained from the eigendecomposition of the covariance matrix and account for the majority of the variability observed in a dataset. Moreover, the first and second principal components provide a low-dimensional representation of data with large p, improving interpretability. The method for PCA is given by

$$a_i = \max\left(\frac{a`\Lambda a}{a`a}\right)$$

$$\text{subject to } a_i`a_i = 1 \text{ and } \text{Cov}(a_i`\Lambda, a_k`\Lambda) = 0 \text{ for } k < i$$

K-means clustering

K-means is a clustering algorithm that segments a dataset into a pre-specified number of clusters k. Clustering algorithms are useful in identifying subsets in datasets with large p, however as it is an unsupervised learning method there is no direct measurement of success. The goal of clustering algorithms is to minimize within-cluster variation and inversely, maximize between-cluster variation. The dissimilarity measure used in k-means clustering is Euclidean distance and its objective is to minimize within-cluster variation $W(C)$. However, there is no guarantee of achieving a global minimum, so reiteration is needed.

Algorithm 1: k-means clustering

1. Randomly assign the number of clusters k
2. Compute the centroid for each of the k clusters
3. Every point is assigned to the cluster whose centroid is closest
4. Iterate until the cluster assignments stop changing

Gaussian mixture model

A Gaussian mixture model (GMM) is a probabilistic model used to identify clusters in a dataset. Using a GMM approach is referred to as soft k-means clustering because it assigns probabilities of data points belonging to specific clusters whereas k-means clustering has rigid assignments. Moreover, the main assumption when using a GMM is that each cluster follows a normal

distribution with mean μ and covariance Σ. A GMM is fit using an expectation-maximization algorithm. A mixture model density is given by

$$g(x) = \sum_{k=1}^{K} \pi_k N(\mu_k, \sigma^2 I),\ \pi_k \geq 0,\ \sum_k \pi_k = 1$$

Results and discussion

The goal of this research was to segment seasons into subclasses to identify factors that indicate success. This was achieved through dimensionality reduction and cluster analysis. To begin exploring the data, correlation between features was assessed. For example, complete games and strikeouts are highly correlated and singles and strikeouts are negatively correlated. Otherwise, the other correlations are weak and suggest little association among variables.

Before exploratory data analysis, the data was normalized to enhance interpretability. PCA was performed and revealed not all variables were equally important. Specifically, the first 14 principal components explain ~90% of the variation observed in the data and the first two principal components contribute 45.19% to the variation. Although the proportion of variance explained by a_1 and a_2 is below 50%, reducing the number of variables is advantageous for revealing latent relationships. Table 2 and Table 3 display notable loading weights of a_1 and a_2.

Table 2. Absolute value of loading weights of a_1

Variable	Loading
HRA	0.26843
HR	0.26092
HBPA	0.24885
CG	0.23851
SHO	0.23583

Table 3. Absolute value of loading weights of a_2

Variable	Loading
H	0.33808
PA	0.32681
AB	0.28726
X1B	0.27408
LOB	0.25230

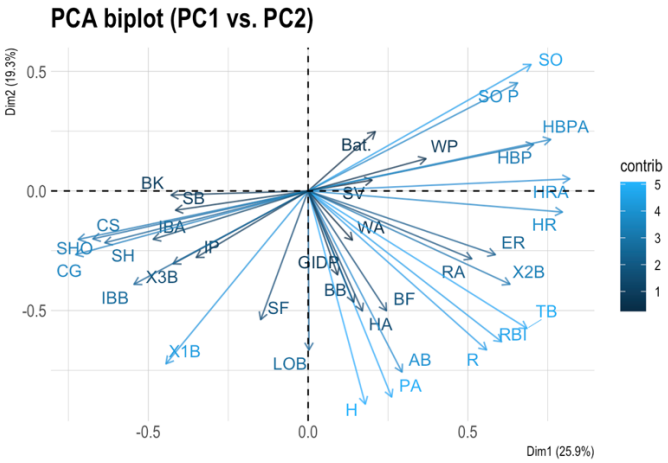

Figure 1. Contribution of features to a_1 and a_2

From the biplot, several variables form a distinct group in the bottom left quadrant. Balks, stolen bases, caught stealing, shutouts, complete games, intentional walks allowed, sacrifice hits, intentional walks, innings pitched have strong negative correlations with a_1 and a_2. These variables represent infrequent outcomes in baseball, with the exception of innings pitched, and suggest these outcomes minimally influence other outcomes in the course of a season. Inspecting the graphs below, it is evident several variables contribute minimally to explaining the proportion of variance observed in the dataset. Specifically saves, grounded into double plays, and triples have below average contribution for both a_1 and a_2.

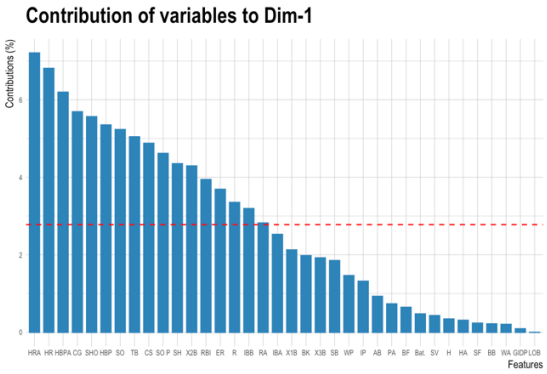

Figure 2. Contribution of variables to a_1

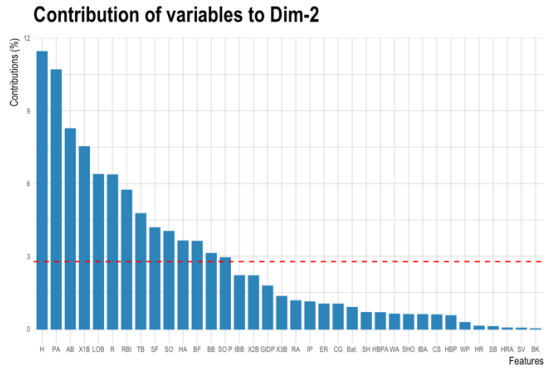

Figure 3. Contribution of variables to a_2

Using a scree plot, $k = 3$ clusters were chosen to perform k-means clustering. Viewing Figure 5, the cluster regions are ill-defined and have significant overlap with other regions. Seeking an alternative understanding of the data, a Gaussian mixture model was fitted using a_1 and a_2. The model determined the optimal amount of clusters to use was five. Looking at the data, there are highly overlapping components.

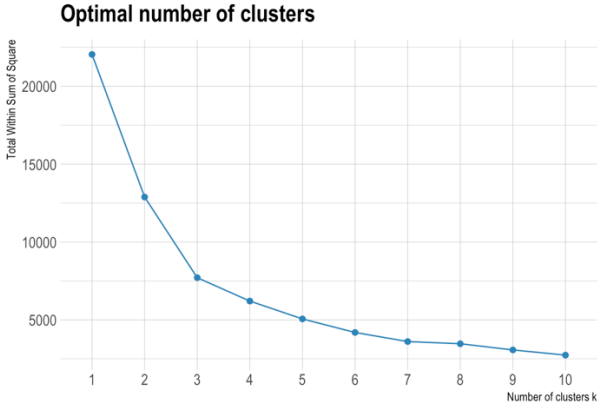

Figure 4. Scree plot for the selection of k

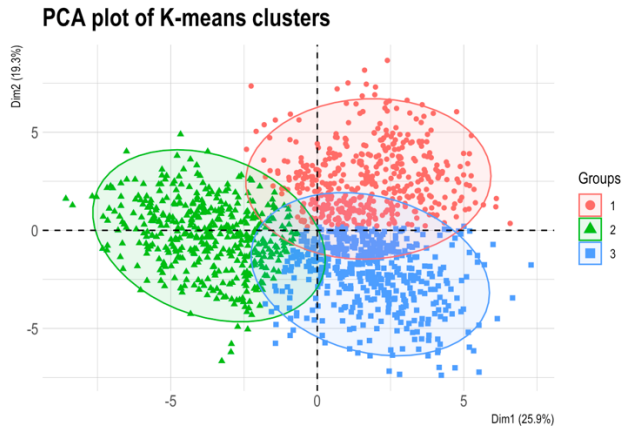

Figure 5. K-means clustering

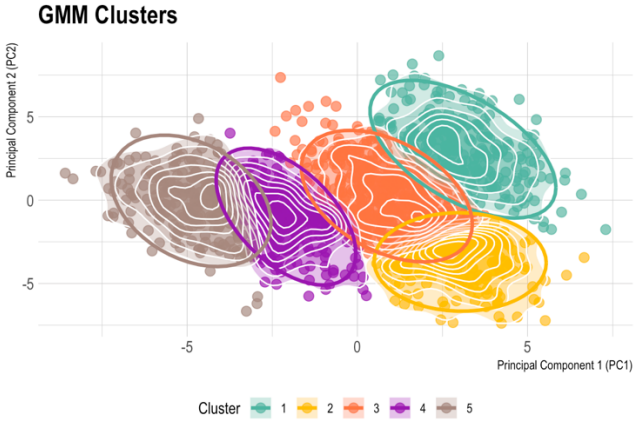

Figure 6. Gaussian mixture model clustering

Below are insights for each cluster using the k-means algorithm:

***k* = 1:**

447 teams ranging from 1987-2024 with the majority of teams from 2010-2024. This cluster included nine World Series winners.

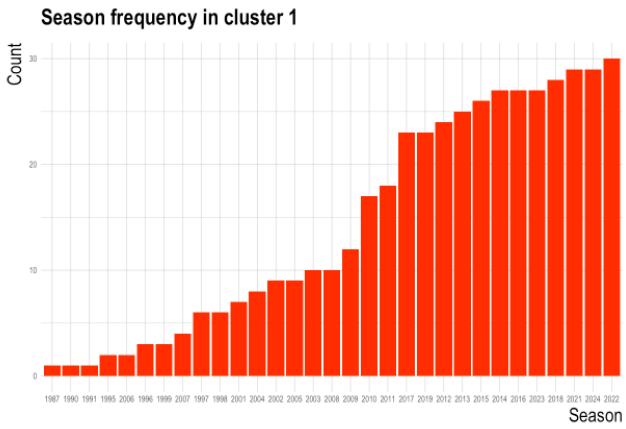

Figure 7. Frequency of $k = 1$ observations by season

$k = 2$:

440 teams ranging from 1975-2012 with the majority of teams from 1970/1980s and early 1990s. This cluster included 18 World Series winners.

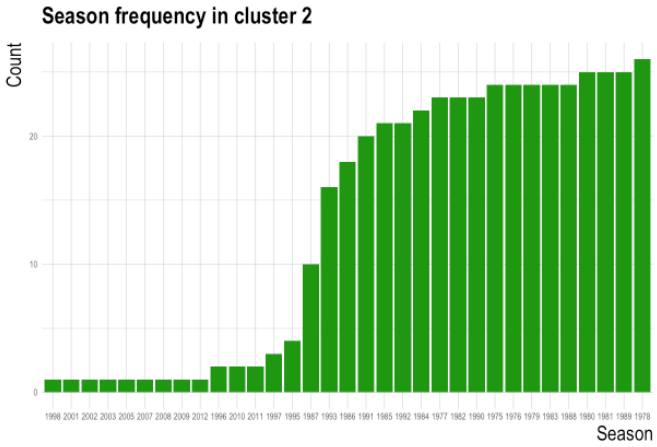

Figure 8. Frequency of $k = 2$ observations by season

$k = 3$:

469 teams ranging from 1977 to 2024 with the majority of teams from the late 1990s and 2000s. This cluster included 21 World Series winners.

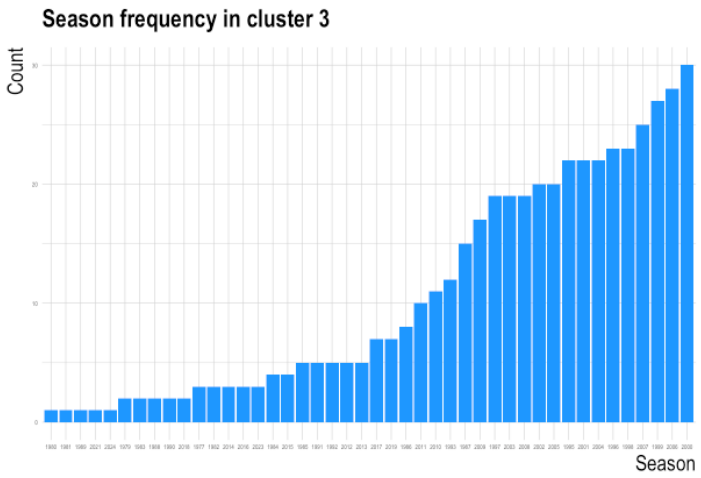

Figure 9. Frequency of $k = 3$ observations by season

Initializing the k-means algorithm with three clusters resulted in poorly defined subsets with significant overlap. This resulted in the algorithm capturing roughly three eras of baseball – cluster 1 predominately representing modern teams, cluster 2 representing classic teams from 1970/1980s, and cluster 3 representing teams across a large time span.

Insights for each cluster using the Gaussian mixture model:

***k* = 1:**

239 teams ranging from 1999-2024 with the majority of teams from the late 2010s and 2020s. This cluster included eight World Series winners.

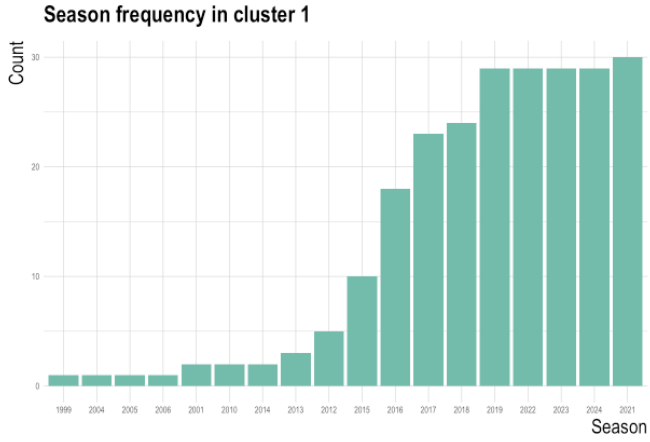

Figure 10. Frequency of $k = 1$ observations by season

***k* = 2:**

160 teams ranging from 1987-2019 with the majority of teams from the late 1990s and 2000s. This cluster included nine World Series winners.

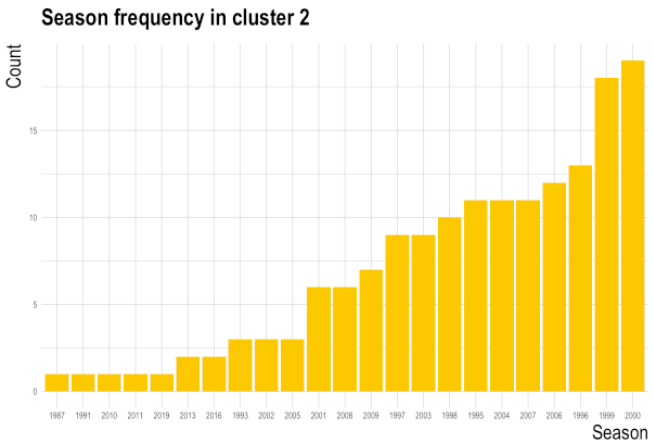

Figure 11. Frequency of $k = 2$ observations by season

$k = 3$:

484 teams ranging from 1977-2024 with the majority of teams from the 2000s and 2010s. This cluster included 14 World Series winners.

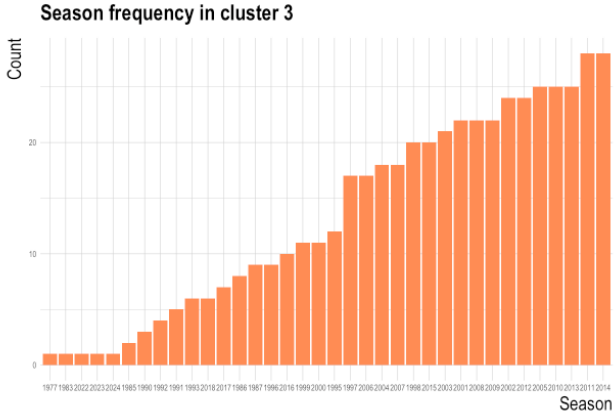

Figure 12. Frequency of $k = 3$ observations by season

$k = 4$:

241 teams ranging from 1975-2012 with the majority of teams from the 1980s. This cluster included nine World Series winners.

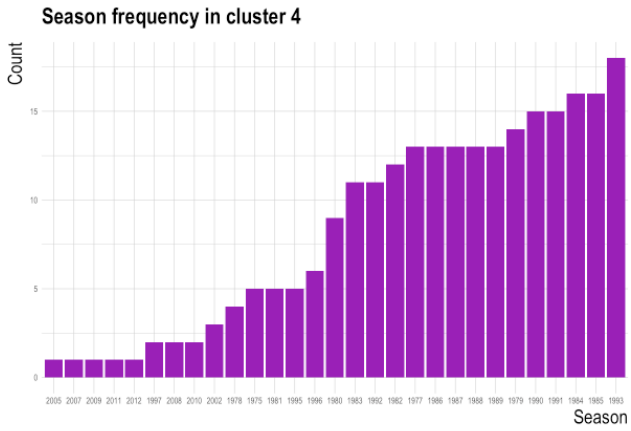

Figure 13. Frequency of $k = 4$ observations by season

***k* = 5:**

232 teams ranging from 1975-1993 with the majority of teams from the late 1970s and early 1980s. This cluster included eight World Series winners.

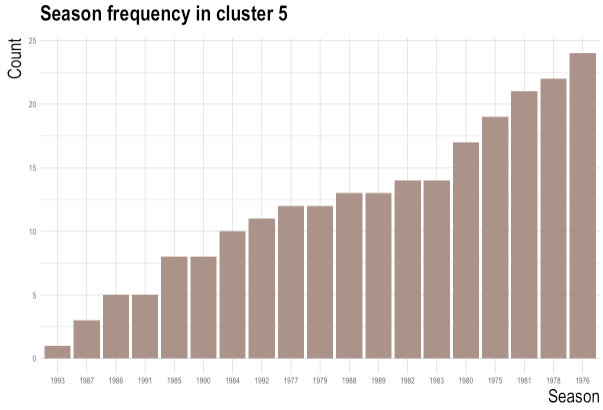

Figure 14. Frequency of $k = 5$ observations by season

The Gaussian mixture model's choice of five clusters suggests there are latent factors that are undetectable using the k-means algorithm. Cluster 1 captures recent teams with nearly every team from the previous five seasons. Cluster 2 represents teams from the late 1990s and 2000s. Cluster 3 includes teams from the 2000s and early 2010s. Cluster 4 represents teams from the 1980s and early 1990s. Cluster 5 represents teams from the late 1970s. These clusters capture teams from distinct periods over the previous 50 seasons and suggest noticeable differences between eras in MLB. However, like the k-means algorithm, the cluster regions are poorly defined and demonstrate significant overlap.

While these findings have contributed insights on subsets of teams and eras in modern MLB history, there are limitations. Mainly, poor cluster separation gives an unclear separation between World Series winners, bad teams, and average teams. Moreover, as this is an unsupervised learning problem, there is no benchmark to strive toward.

In conclusion, this research explored if specific factors indicated team success in professional baseball. Using several distinct clustering methods, no clear conclusions were found. Drawing from the insights of this research, the clusters formed are indicative of specific time periods rather than team success. This is evident from visually inspecting the frequency graphs. A natural extension is to consider latent variables like recent rule changes to the pace of play and its effects on increased offense using factor analysis.

References

Baseball Reference. (n.d.). *Baseball Reference*. Retrieved December 15, 2024, from https://www.baseball-reference.com/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Johnson, R. A., & Wichern, D. W. (2018). *Applied multivariate statistical analysis* (5th ed.). Pearson.

Lewis, M. (2011). *Moneyball*. W.W. Norton.