Bellevue University

DSC 550 Project Paper

Baseball Performance Data and Relationships to Hall of Fame Selection

Conrad V. Ibañez

Professor Becky Deitenbeck

DSC 550

August 8, 2020

Baseball Performance Data and Relationships to Hall of Fame Selection

In today’s society, data is generated, stored, analyzed, and consumed in vital ways. Many aspects of our lives are affected by data, whether we are buying products, watching movies online, or just browsing the Internet. Data is an important aspect of our lives. Additionally, data can motivate our actions and impacts our decision-making. We also live in a world today where performance is rewarded. Whether it is the amount you make while doing your job or how much we are willing to pay for the latest gadget, society tends to reward people or objects based on performance. For people who excel in their profession, they usually can command a high salary, and products that satisfy the needs or desires of people, such as the latest smartphone, can cost hundreds of dollars or more.

The same concepts can be applied to Major League Baseball (MLB). The best players are offered large contracts that pay hundreds of millions of dollars over a large span of years. Players are also given different awards such as the Most Valuable Player (MVP) for the best player in the season or the Cy Young Award for the best pitcher. Players can be selected for the All-Star game where the best players in the season play in an exhibition game that is held half-way through the year. Perhaps the most important reward for a player with a notable career is induction into the Major League Baseball Hall of Fame (MLB HOF). This project analyzes baseball performance data and its relationships to player rewards.

There are many statistical questions that are interesting to investigate in baseball when it comes to player performance and rewards. For example, can we mine frequent patterns in player performance data and associate them to rewards data? Do certain player statistical performance data lead to certain accolades such as All-Star or Hall of Fame selections or even large monetary contracts? If so, then we can apply techniques, such as clustering or machine learning, to classify or predict certain associations. The focus of this project is primarily on the offensive statistics in batting for a player, and it investigates if there is a correlation between these metrics and whether a player is chosen for the Hall of Fame. The hypothesis for this project is that players who perform well and have high metrics are elected to the Hall of Fame.

Many of the most notable baseball players are members of the Major League Baseball Hall of Fame. They include legends such as Babe Ruth, Hank Aaron, Lou Gehrig, Ted Williams, and Jackie Robinson and the more recent inductees such as Chipper Jones, Ken Griffey, Jr., Randy Johnson, and Pedro Martinez. Many have had stellar careers, but there are different rules for election into the Hall of Fame. There are different criteria for player eligibility such as playing for ten seasons and having retired from playing five years prior to election. Eligible players are voted into the Hall of Fame by voting members of the Baseball Writers’ Association of America (BBWAA).

The dataset used for this project is the Baseball Databank found on Kaggle.com. It is a collection of 20 files containing baseball data collected from 1871 to 2015. The main tables include a master table for player and biographical info, batting statistics, pitching statistics, and fielding statistics. There are also tables that contain information for player salaries, awards, All-Star appearances, and Hall of Fame information.

The metrics that this project evaluates focuses on the following offensive metrics:

* AB: At-Bats- Number of attempts by a player to hit the ball.
* H: Hits - Number of times a player reaches a base safely
* HR: Homeruns - Number of times a player reaches home with one At-Bat
* RBI: Runs Batted In - Number of times a player causes a run to score

Other offensive metrics that were considered:

* R: Run - Number of times a player scores a run
* 2B: Double - Number of times a player reaches 2nd base after an At-Bat
* 3B: Triple - Number of times a player reaches 3rd base after an At-Bat

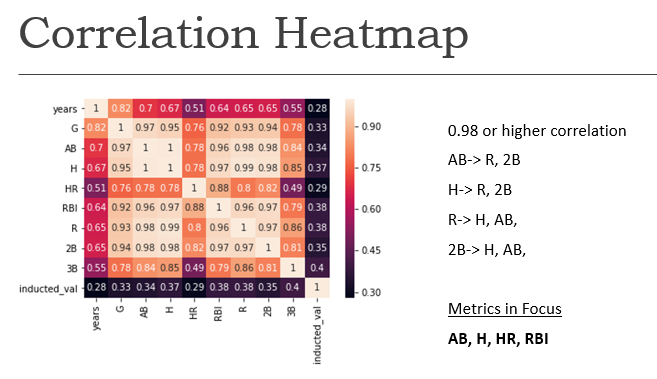


Figure 1: Correlation Heatmap of Offensive Metrics

Figure 1 shows a Correlation Heatmap of some common player offensive metrics. We observe strong correlation of greater than 0.98 between AB, H, R, and 2B. Because we want to limit the variables and try to set up the data to use some of the techniques demonstrated with the iris dataset during this course, we will use AB and H from the set as those metrics are some of the more popular and well-known attributes. AB, H, HR, and RBI will be used in most of the analysis for this project, and inducted\_val will indicate whether the player has been elected to the Hall of Fame (1 = Hall of Famer). The dataset is reflected in Figure 2.

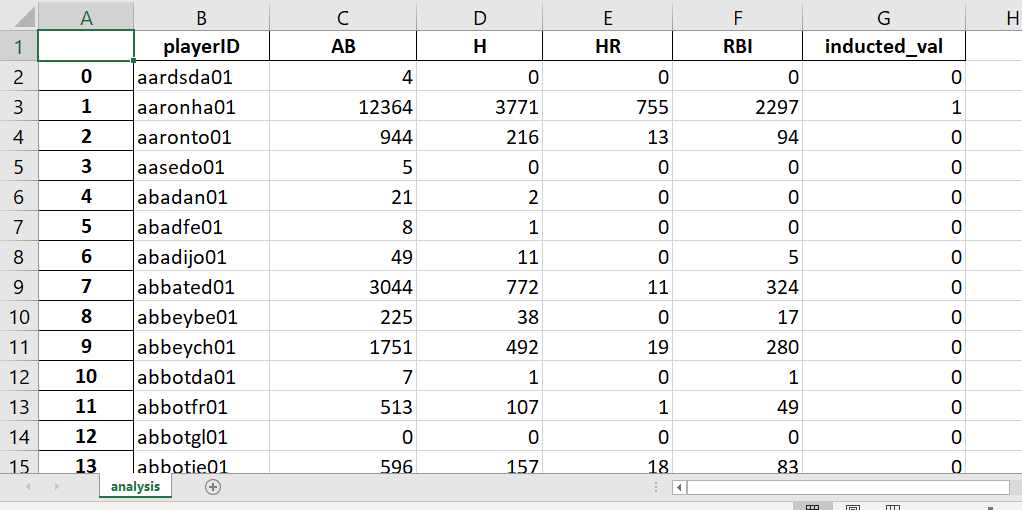


Figure 2: The primary dataset attributes used for this project.

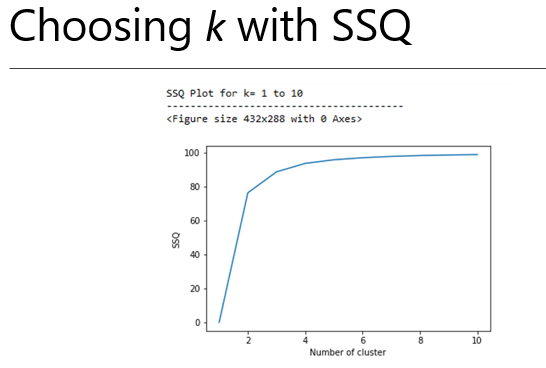
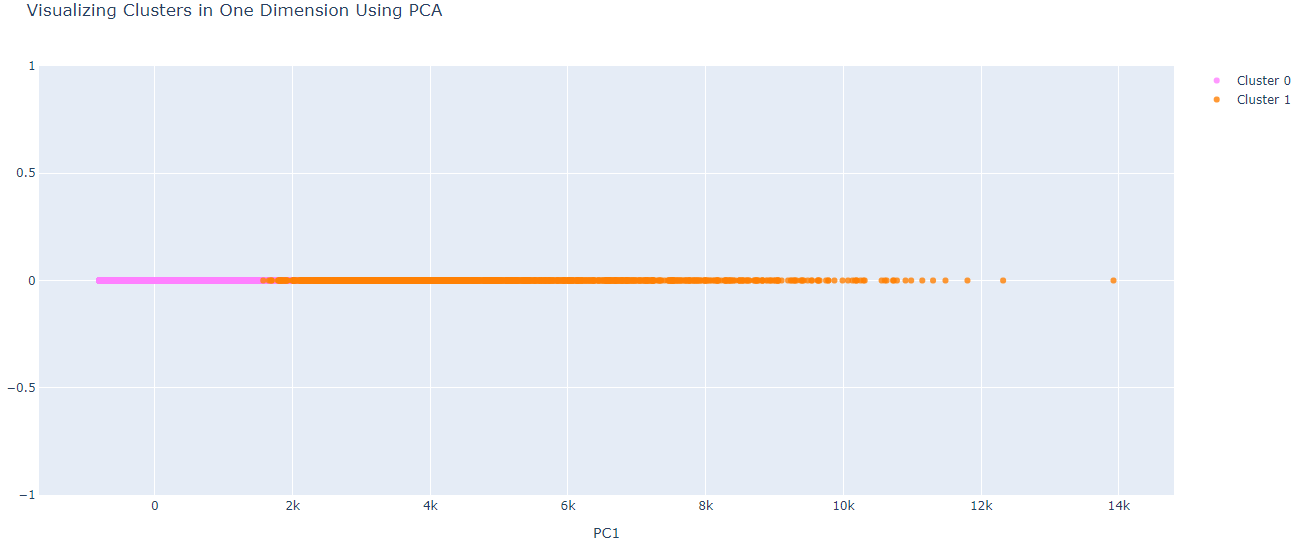


Figure 3. SSQ Plot for k = 1 to 10

Figure 3 shows the “sum of squared deviation” (SSQ) for k values from 1 to 10. We see a noticeable bend between k = 2 and k = 4. The elbow point indicates an optimal value for k between 2 and 4. Visualizations are created in this project with one dimension and two dimensions using PCA. Figure 4 shows one dimension and Figure 5 shows two dimensions.

Figure 4: One Dimension Using PCA with k = 2

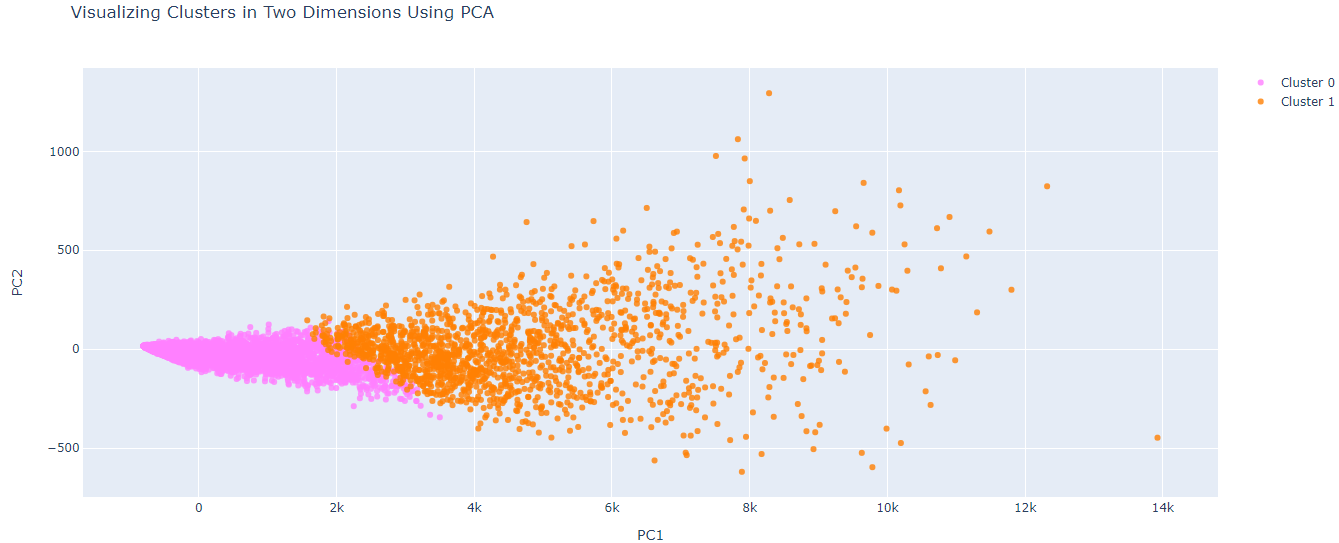


Figure 5: Two Dimensions Using PCA with k = 2

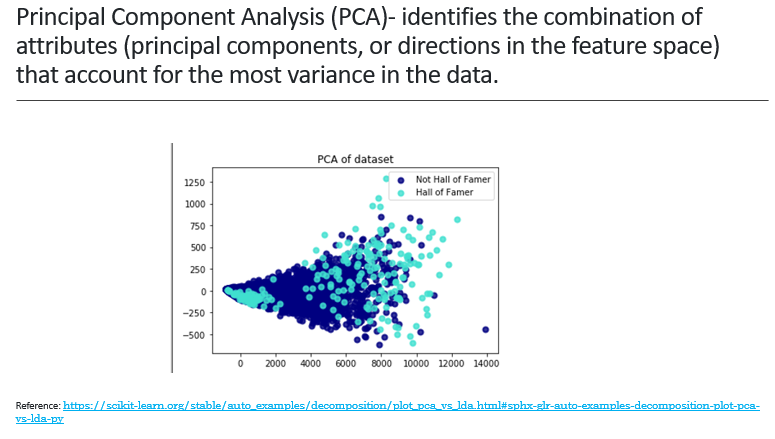


Figure 6: PCA of the dataset when k = 2 and points are colored indicating Hall of Fame status

Figure 6 is perhaps the most insightful graph as a result of this project. It is the PCA of the dataset using k = 2 and considering the offensive metrics, AB, H, HR, and RBI. Each point is color-coded based on the Hall of Fame status of the player. We see that there is no clear, straight line that separates the points between players that are in the Hall of Fame or not. There appears to be anomalies in the dataset. It seems that there are players with poor offensive metrics that are in the Hall of Fame and players with very good offensive metrics that are not in the Hall of Fame. We investigate the data further.

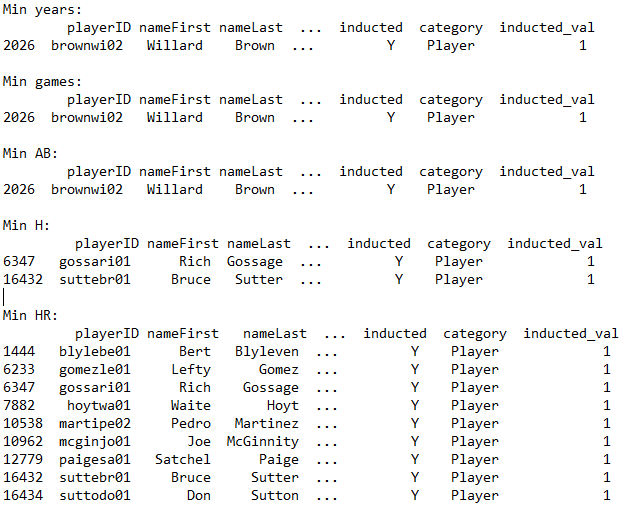


Figure 7: Hall of Fame player data matching minimum values for the offensive metrics

Figure 7 shows Hall of Fame player data that matches the minimum values for the various offensive metrics. It explains the anomaly of how players with what appears to be poor offensive numbers have been elected to the Hall of Fame. If we do a brief research of some of these players online, we will see primarily two types of players. Willard Brown who has a few of the minimum values was a baseball player in the Negro Leagues and played only a few games in Major League Baseball. He was voted into the Hall of Fame based on his contributions to baseball although his metrics in the Negro Leagues were not captured and his MLB stats were not very impressive. Other players with low offensive numbers appear to be pitchers. Both Richard Michael "Goose" Gossage and Bruce Sutter are pitchers. This would make sense since pitchers are highly specialized in their role and very few can hit the baseball. Additionally, pitchers do not participate in hitting in the American League of MLB due to the designated hitter position. This project includes pcaHallOfFameOutliers.xlsx which lists more players that are in the Hall of Fame with poor offensive metrics, and they generally fall into these two player types.

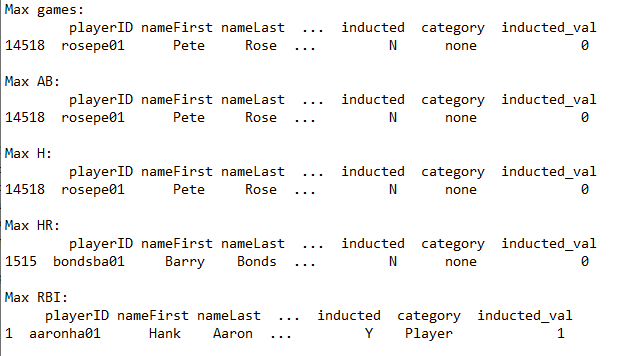
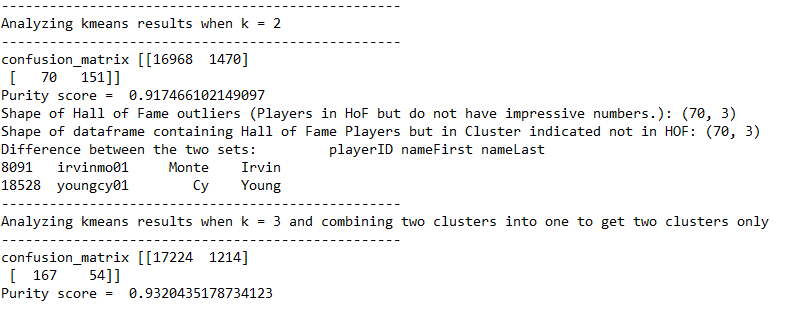


Figure 8: All player data matching maximum values for the offensive metrics

Figure 8 shows details for players from the all player data set that match the maximum values of the offensive metrics. Additionally, this project includes the file pcaNonHallOfFameOutliers.xlsx that contains player data for those who are not in the Hall of Fame but have high offensive marks. We see from Figure 8 one type of the players, which include Pete Rose and Barry Bonds, who are not in the Hall of Fame because of issues outside of game performance. Pete Rose was an impressive hitter who has been banned from baseball because of personal conduct, and Barry Bonds had a career that was tainted by the steroid era scandal. Upon further analysis of the outliers, we do see some players that are somewhat in the bubble for the Hall of Fame. We need to keep in mind that Hall of Fame voting is occurring yearly, and there is a subjective aspect to it. There are some players who have had impressive careers but have not been out of baseball for the required five years. They are ineligible for voting, as is the case for Derek Jeter who was recently voted in but in the current dataset was not classified as a Hall of Famer. There are also players who are just outside the voting line of Hall of Fame status who might still be waiting and are impacted by the subjectivity of the voters.

Figure 9: Confusion matrix for K means

As part of this project, we also create the confusion matrix for k = 2 and observe a purity score of 0.917. A confusion matrix for k = 3 was also created but with a twist in which two of the clusters were combined and one cluster was used to indicate the Hall of Famers. This essentially created two clusters to distinguish between players in the Hall of Fame and those who were not. This achieved an even higher purity score of 0.932. This indicates the possibility of creating even better clusters to distinguish between Hall of Famers. Perhaps there are techniques or approaches to better handle the anomalies as well.

In conclusion, the project results support the hypothesis. The analysis indicates that in general, players who have high performance metrics are elected to the Hall of Fame. Certain anomalies have been addressed regarding certain types of players, player roles/positions, or player status that have influenced their entry or lack of entry into the MLB Hall of Fame. In addition, since voting is done by a group of individuals, there is subjectivity. Hall of Fame voting is not purely based on metrics alone.

REFERENCES

Baseball Databank. <https://www.kaggle.com/open-source-sports/baseball-databank>

BBWAA ELECTION RULES. <https://baseballhall.org/hall-of-famers/rules/bbwaa-rules-for-election>

Bruce Sutter. <https://en.wikipedia.org/wiki/Bruce_Sutter>

Goose Gossage. <https://en.wikipedia.org/wiki/Goose_Gossage>

Leskovec, J., Rajaraman, A., Ullman, J.D.., W. (2020). Mining of Massive Datasets 3rd Edition. New York, NY: Cambridge University Press.

Willard Brown. <https://en.wikipedia.org/wiki/Willard_Brown>

Zaki, M. & Meira Jr., W. (2020). Data Mining and Machine Learning 2nd Edition. New York, NY: Cambridge University Press.