**Author**

Provide your Full Name here and your GWID.

## Michael Wegan G45192209

## Zheying Chen G43061235

## Weiwei Li G27494462

## Xueli Zhou G29955935

## **Problem Definition and Data Investigation:**

Please answer the questions below and provide the following Project Proposal;

**1. Introduction and motivation**

Major League Baseball presents multiple awards at the end of each season. The most valuable player (MVP) and Cy Young awards go to the player that is deemed most important to his team and the best pitcher, respectively. These are awarded based on votes by the Baseball Writer’s Association of America. There are no hard and fast rules about which players receive the awards annually.

**2. Definition of the problem**

There are no formal criteria for selecting candidates to receive awards. It is a subjective process. We seek to use past data to model important variables for the prediction of future award winners.

**3. Background research, context and related work.**

Individual player statistics are often used to compare players’ performances and justify MVP and Cy Young selections. However, the criteria are subjective as a voter’s selection is based on their own construct of what constitutes the most valuable player or best pitcher. Taken across voters, though, a picture of the most important statistics may arise. Variance will be driven by individual voters’ perception of an ideal player but we believe through data mining it is possible to build a model to predict future winners based on that year’s performance data.

**4. Data description and acquisition process.**

There are many baseball data sets available. We have chosen to use Sean Lahman’s player database accessed from SeanLahman.com on 9/19/16. It is a longitudinal data set of individual player statistics of all players from 1871 until present. These are basic statistics such as hits, at-bats, runs, etc. We will build some of the derived statistics such as batting average, runs created, and slugging percentage from the basic statistics.

**5. Initial Plan and proposed approach**

We will use 60% of the data from 1911 (when MVP was first awarded) until present to train our model. The data will include all players that played in a given year and will be compared to the MVP winner for that year. Metrics for all individuals will be standardized by the MVP’s statistics for that year. We propose to use training data from the middle of the time series to test our predictive power at both temporal ends of the data set. We will do the same for pitching data and Cy Young winners.

Below are questions that you should investigate, answer and report back:

1. **Define and articulate the question/s that you hope to solve, or that you are interested in, before you look for data sources.**

Find the players who are more likely to be a MVP or a Cy Young based on the annual performance data.

**2) This question might change over time as you explore data sources and find new possibilities or encounter limitations. Explain or identify your motivation for choosing this question/s and this field of Data Analytics?**

Baseball has a treasure trove of data that can lead to actionable insights. Predicting the MVP and CY Young is an annual rite at the end of the season. Pundits and sports networks devote significant time to their predictions. A formalized model can lead to higher predictive power. The performance metrics we are using should reveal the most important variables for which voters select. We may find that we need additional derived statistics. Observational data such as hits or at bats that are manipulated in a consistent way to uncover additional information about a player’s performance relative to other players are considered derived statistics.

**3) Is it a field you hope to work in?**

Possibly.

**4) Do you have knowledge about the area from past experiences?**

Yes. We are avid fans.

**5) What are some initial thoughts about where you think you might find data about this issue?**

There are many websites with baseball data, such as baseballreference.com, mlb.com, brooksbaseball.net, retrosheet.org. We found ours at SeanLahman.com.

**6) What problems or challenges do you anticipate you will have to overcome?**

Initially, the biggest hurdle will be cleaning and organizing the data so that the player metrics within a given year are standardized by the MVP or Cy Young winner’s metrics for that year.

**7) Do you anticipate that you will have to merge or join datasets together to create the data that you need?**

Possibly. There are many other statistics that baseball analysts use. These are generally derived statistics using individual player performance against expected player performance based on all players for that year. We will struggle to implement these metrics as we do not have all of the data necessary to develop them. We are seeking an additional data set with the derived statistics to column bind to our current data set.

**8) What types of hypothesis or ideas do you have about the answer to this question?**

Players clustering near the top of statistical categories are more likely to win an award. We believe that for MVP, offensive categories such as hits, batting average and home runs will be important. We also believe that advanced derived statistics will be valuable predictors even though they were developed relatively recently, from mid 1990’s to present. This would indicate that although these metrics were not available to voters prior to that time, they were able to intuit player value through simple statistics.

**9) What datasets have you found?**

We have got some datasets. These datasets include the name of the players, year, player’s team, and games played within the season. Performance data about At Bats, Run scores, Hits, Doubles Hit, Triples Hit, Home Runs Hit, Runners on Base, Base on Ball, and Strikeouts.

And we have got the data about the player who won an award in those years between 1911 and 2015.

Standard pitching data are similarly available from 1956 (when Cy Young was first awarded) to present .

**10) Where did you find them?**

The baseball website:

SEANLAHMAN.com

And some additional data from this website’s manager.

**11) How did you find them?**

We searched on the Internet, and compared multiple data sources to find a dependable data set with the variables we wished to include.

**12) If the data was not already collected, how did you collect it?**

All baseball data have been collected. The issue for us is how to either generate additional derived variables from the data we have or find other sources with additional derived variables already calculated.

**13)What do you know about how they were created and who created them?**

Baseball has always been a game in which statistics were important. Player level data have been maintained since the advent of professional baseball. Some of the data that were not recorded in the earlier years of organized professional baseball have been appended to the data sets through volunteer efforts. Volunteers methodically review game logs to recover data not collected in those times to add those data to players’ records, allowing comparison to present day players.

**14)How do they compare to each other?**

Here, the pronoun “they” is somewhat confusing.

If “they” refers to the players, MVP and Cy Young winners’ data can be compared to non-winning players to reveal the important variables for voters.

If “they” refers to the different data sets, then they relate to each other by year and player name. The data sets were all derived through the compilation of game data so collection was consistent. The only difference was that some earlier era game data were compiled from paper records and recent game data were collected as the game occurred.

## 

## **Evaluation and Summary**

Make your summary main points here. Evaluate your findings so far. Draw some early potential conclusions.

Our goal is find the players who are more likely to be a MVP or a Cy Young based on the annual data of performance. We understand that voters in someway value performance metrics but there has been no formalization of the process. All voters have differing opinions of what makes the best pitcher or most valuable player. This is not only different among voters but may be different for a particular voter by year. We seek to account for that variation in selection criteria by analyzing performance metrics. Through this process, we aim to build a model capable of predicting the MVP and Cy Young winners.

We believe our data are adequate to make these predictions but wish to add additional variables. Currently, we are in communication with the managers of several baseball data related websites to retrieve these additional data. From our preliminary glances at the data, it is clear that offensive categories such as hits, batting average, and home runs are important predictors of MVP winners. Cy Young winners tend to have many wins and a low earned run average relative to other pitchers. How advanced metrics are considered by voters is less clear.

Additionally, we may attempt to run this analysis in a forward looking fashion. For example, can we predict next season’s MVP or Cy Young winner from this year’s data? This would be an interesting exercise as it would give teams a sense of a player’s value on the open market given their likelihood to win one of these coveted awards.