# **PROJECT PROPOSAL**

Group Members: Ryan Perry, Jacob Kronlage, Xavier Hummer

### **INTRODUCTION**:

As we are all baseball fans, we are intrigued by the philosophy of Billy Beane and his “Moneyball” approach to building a team. If you are unfamiliar with this theory, in the early 2000s, Billy Beane served as the Oakland Athletics General Manager. This team is a small-market franchise with less revenue than teams like the New York Yankees or the Los Angeles Dodgers, which makes it harder for them to win since there is no salary cap in baseball, allowing teams to spend as much as they wish. Since they lacked the funds to compete with these larger-market teams, they rethought their approach to evaluating players.

Billy Beane explored specific statistics about players, such as slugging and on-base percentages. Previously, players were viewed as five-tool athletes, judged solely on their batting averages and RBIs (Runs Batted In). This approach enabled Beane and the Athletics to discover players with less prominent batting averages but better slugging or on-base percentages. Initially, the team started slowly but ultimately won 20 consecutive games, setting an MLB record for the most wins in a row. Although the team lost in the ALCS (American League Championship Series), it demonstrated to baseball the necessity of considering more statistics when evaluating players. This shift was pivotal for the sport, as every team now employs these metrics to assess players' performances effectively.

This data set includes the variables that Beane and the Athletics primarily focused on and utilized. As mentioned, they discovered that considering more than just batting average when evaluating hitters is essential. We intend to use descriptive analytics to address various questions that could benefit baseball teams and their managers. The insights we gain will provide baseball teams and fans with more valuable and reliable information for team building and player evaluation.

### **DATA**:

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| Team | Text | Abbreviation of full team name. |
| Team Name | Text | Full name of the team. |
| League | Text | Which league the team is in (American or National). |
| Year | Numeric | Which year the season was played in. |
| Games | Numeric | Number of games played. |
| Wins | Numeric | Number of wins. |
| Home Runs | Numeric | Total number of home runs hit collectively as a team. |
| RS | Numeric | Total number of runs scored as a team. |
| RA | Numeric | Total number of runs allowed as a team. |
| OBP | Numeric | *On Base Percentage* –  The percentage of a team’s plate appearances that result in a player reaching base. |
| SLG | Numeric | *Slugging Percentage* – A team’s average number of total bases earned per at-bat. |
| BA | Numeric | *Batting Average* – The team’s percentage of at-bats that result in a hit. |
| Playoffs | Logical | Yes or no if the team made the playoffs. |
| PlayoffsFinish | Numeric | Final team ranking in postseason – e.g., World Series Winner = 1.0 |

Our initial dataset was sourced from Kaggle.com as a CSV file. We opened this “baseball.csv” using Pandas to create a DataFrame. We then cleaned the data by adding and removing a few rows. First, we added a team name column. Since there was only an abbreviated team name column, we applied mapping. Next, we determined that the columns for season rank, opponents’ on-base percentage, and opponents’ slugging percentage were not beneficial for our proposed analysis. We changed the playoffs column to a numeric representation, with “1” indicating that the team made the playoffs and “0” indicating that the team missed the playoffs, effectively converting it to a yes or no column. It is important to note that although the dataset spans from the 1962 to 2012 seasons, **the Kaggle dataset excludes the seasons played in 1972, 1981, 1994, and 1995**. In 1994, a players' strike resulted in the cancellation of the entire postseason, which affected the 1995 season as well. The seasons in 1972 and 1981 are also missing due to player strikes that disrupted the standard postseason process, leading many historical baseball datasets to exclude these seasons. Therefore, the dataset **only includes 47 seasons and 47 World Series champions**. Lastly, we renamed and reorganized columns and confirmed no missing values (NaN).

The scraped data is collected from Baseball Reference, a baseball database that maintains statistics related to the sport. Baseball Reference is part of Sports Reference, which includes other branches for sports like football, basketball, and more. We have horizontally integrated the home run column from Baseball Reference, adding home runs only for the 2012 season. In the future, we will incorporate home run data for each season into our dataset. We employed a loop to retrieve the records from that home run column and loaded them into a Pandas DataFrame. Finally, we merged the Baseball Reference data with our initial Kaggle dataset.

Our Kaggle and Baseball References datasets are in the Works Cited section below. Our CSV file for the initial “baseball.csv " and a merged dataset named “project\_proposal\_dataset.csv "are attached.

**PROPOSED ANALYSIS:**

Our data focuses on various baseball stats and the outcomes of a given team’s season. The questions we aim to answer pertain to the importance of these stats. Additionally, we strive to discuss which stats contribute to a successful baseball team.

The first question we seek to answer **is the relationship between the number of home runs a team hits and whether this results in more playoff appearances or deeper playoff runs**. It is easy to assume that a team hitting more home runs will likely win more games; however, we want to understand how strong this relationship is. We can use our data to analyze the difference in average home runs between teams that made the playoffs and those that did not. This insight will help baseball teams determine if a power-hitting strategy is an effective way to advance to the postseason.

The second question we aim to explore involves analyzing World Series champions. **We will examine each of the 47 seasons' World Series champions in our data and closely review their statistics**. This approach will help us gain insights into the statistics prevalent among World Series-winning teams. This could provide lesser teams with a model to aspire to in their quest for a championship.

The third and final question we hope to address with our data **is the importance of a baseball team’s on-base percentage (OBP) in predicting a team’s total wins compared to other batting metrics, such as batting average (BA) and slugging percentage (SLG)**. In the movie Moneyball, a recurring theme is that the path to winning more baseball games lies in getting on base more. We aim to test this theory to see if winning depends on getting on base, along with the other two statistics. This analysis could help baseball teams make data-driven player signings and develop a more cost-effective strategy for front offices.

## **WORKS CITED**

<https://www.kaggle.com/datasets/wduckett/moneyball-mlb-stats-19622012>

<https://www.baseball-reference.com/leagues/majors/2012.shtml>