# **Project Final Report**

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. Lacking 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 Oakland 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**:

Our primary data comes from Baseball Reference[[1]](#footnote-1), an all-encompassing database that records baseball data from over one hundred years ago and is updated daily. Our information was sourced from Baseball Reference but presented in a Kaggle dataset to make it easier to work with.

The Kaggle dataset[[2]](#footnote-2) we found compiled data from Baseball Reference into a collection of crucial baseball statistics covering 50 years (1962-2012). Notably, four seasons within this timeframe are excluded due to MLB player strikes that suspended the seasons. We organized this data into a pandas DataFrame in Jupyter Notebook. We performed minor data cleaning, such as changing the team abbreviations to full names. Additionally, we converted our binary data to be represented as Yes/No instead of 0/1 for greater clarity. Finally, we made the necessary adjustments to ensure no NA values.

This data includes information about the movie Moneyball, where MLB executives began to value statistics differently. Some crucial statistics found in the dataset include Runs Scored, Runs Allowed, OBP, SLG, and BA. OBP refers to a team’s on-base percentage, which measures how often a player on a given team reaches base safely, whether by a hit, walk, or hit by pitch. This statistic is a major focus in the film, so we aimed to evaluate it thoroughly. SLG refers to slugging percentage, the average number of bases a player touches in a given at-bat. For example, a player who hits a double every at-bat would have a slugging percentage of 2.000 (this would be extremely high). Finally, BA refers to batting average, which excludes walks and determines how likely a player is to get a hit.

Along with these statistics, we wanted to gather more information on other statistics not present in the Kaggle dataset. We analyzed home runs and home team attendance to better understand what determines success in the MLB. For home runs, we were able to scrape data from Baseball Reference itself. This was convenient because the data was already in a similar structure to the Kaggle dataset. We scraped this data and joined it to our original dataset using a left join, which provided us with a column for home runs in our DataFrame.

Finally, we decided that a team’s home attendance would be fascinating to include in our dataset. Unfortunately, Baseball References did not provide this information, so we had to source it from The Baseball Cube[[3]](#footnote-3), another baseball database. This database supplied us with home attendance numbers for an entire season. We utilized web scraping to merge our attendance data into our baseball database using a left join. This enabled us to incorporate an attendance column that reflects a team’s total home attendance for each season.

*Table 1: Data Dictionary*

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| Team | Text | Abbreviation of the full team name. |
| Team Name | Text | Full name of the team. |
| League | Text | League the team belongs to (American or National). |
| Year | Numeric | Season year. |
| Games | Numeric | Number of regular season games played. |
| Wins | Numeric | Number of regular season games won. |
| Home Runs | Numeric | Total number of home runs hit by the team during the regular season. |
| RS | Numeric | Total runs scored by the team in the regular season. |
| RA | Numeric | Total runs allowed by the team in the regular season. |
| OBP | Numeric | On-Base Percentage – proportion of plate appearances resulting in reaching base (regular season only). |
| SLG | Numeric | Slugging Percentage – average number of total bases earned per at-bat (regular season only). |
| BA | Numeric | Batting Average – proportion of at-bats that result in a hit (regular season only). |
| Playoffs | Logical | Indicates whether the team made the playoffs (Yes/No). |
| PlayoffsFinish | Numeric | Final postseason ranking (e.g., World Series Winner = 1.0). |
| Total Home Attendance | Numeric | Total number of attendees at home games during the regular season. |
| Attendance Bin | Text | Labeled categories based on total regular season home attendance. |

### Before diving into our analysis, we thought it was important to note that after completing our project check-in and discussing the issue surrounding playoff teams having inflated statistics due to playing more games as a result of making the postseason, we took a deeper dive and found that all the data collected was only regular-season data, and both the websites we scraped from had separate data tables for postseason statistics. We adjusted our descriptions of each field to reflect this, and therefore we do not need to worry about adjusting our evaluation methods.

### **ANALYSIS**:

#### **QUESTION 1:**

For our first analysis question, we wanted to see how vital the numerical statistics we collected would be to a team’s success. Specifically, when we refer to a team’s success, we mean wins, which reflect a regular season record and do not factor in the team’s performance in the playoffs. We chose this question because we assumed that Billy Beane found these statistics more meaningful and superior for evaluating players. Our team wanted to test Beane's strategy by analyzing whether these statistics are more significant for predicting a team's success. We believed these would be key drivers for achieving higher wins in a baseball season. We expected to see teams with higher OBP, SLG, RS, and BA statistics leading to a greater total number of wins that season. We also expected teams with lower RA values to have a greater number of regular-season wins.

*Table 2: Correlation Matrix*  
A table with numbers and text

AI-generated content may be incorrect.

#### Based on the correlation we ran with wins as the target variable, all the statistics listed above have a moderately strong relationship to total wins.

#### Figure 1 A diagram of a distribution of a number of objects AI-generated content may be incorrect.

*Figure 2*

#### A graph of a slg AI-generated content may be incorrect.

*Figure 3*

#### A graph of a distribution of bating average AI-generated content may be incorrect.

#### We performed a univariate analysis on some statistics to get a better idea of how the values were distributed. We can see that all three of the chosen statistics are quite normally distributed. This is beneficial because it allows for reliable predictions, which we will conduct later on. We also created scatterplots of the features with the highest correlation values to further visualize how statistics and total wins correlate.

*Figure 4*

A diagram of blue dots

AI-generated content may be incorrect.

*Figure 5*

#### A diagram of a number of blue dots AI-generated content may be incorrect.

*Figure 6*

#### A diagram of a graph AI-generated content may be incorrect.

#### These scatterplots help us better visualize the correlation matrix and demonstrate that, although it may not be the strongest relationship, teams with higher OBP, SLG, RS, and BA and lower RA statistics are expected to achieve a higher number of regular-season wins. To extend this research question, we pulled in not only these statistics but also home runs and total home attendance which didn’t have correlation values as high as the other statistics but still had weak correlation and built a linear regression model to see if these features could create a sufficient model to predict the number of wins a team would achieve. After building the model and standardizing the features, we attained a mean squared error of 16.34, which is relatively low, and an R-squared value of 86.78%. This indicates that almost 86% of the variance in team wins is explained by our model, which is very strong.

#### We wanted to see if we could enhance our mean squared error and variance using the same features but a different model. We employed a decision tree regressor model with a maximum depth of 4. However, after running the model, we only achieved a mean squared error of 32.09 and an R-squared of 74%. Although this is not bad, it is not better than our linear regression model.

#### **QUESTION 2:**

Our initial second analysis question was whether home runs indicate a team’s success in making the playoffs. We hypothesized that the more home runs a team hits, the higher the likelihood of making the playoffs. We analyzed this by evaluating home runs between teams that made the playoffs and teams that did not.

*Figure 7*

A graph of a person running

AI-generated content may be incorrect.

We visualized this analysis with a box plot, which clearly shows that teams that made the playoffs typically had more home runs during the regular season than those that did not make the playoffs.  
  
To further address this, we also conducted a hypothesis test. After performing a z-test between the home runs of playoff teams and the home runs of non-playoff teams, we obtained a Z-statistic of 9.823 and a p-value of approximately zero. A p-value below .05 significance level further supports our hypothesis that home run totals are associated with team success.

Similarly, for our first analysis question, we aimed to broaden our investigation by exploring whether we could create a logistic regression model that incorporates all the statistical features from the linear regression conducted in the first question, which could predict a binary variable: whether a team makes the playoffs or not.

#### After running the logistic regression model, we attained an accuracy score of 89%, which is very strong overall. For those who did not make the playoffs, our model has a precision of 92%, meaning it predicted non-playoff teams who were actually non-playoff teams correctly 92% of the time. Our model had a precision of 76% for those who made the playoffs, which means it predicted playoff teams who were actually playoff teams 76% of the time.

To round out this analysis, we found the box plot visualization for the home run statistic between non-playoff and playoff teams interesting, so we also ran it on some other statistics included in the logistic regression model to see the differences in those statistics as well.

*Figure 8*

A graph with a blue and orange box

AI-generated content may be incorrect.

*Figure 9*

A diagram of a game

AI-generated content may be incorrect.

*Figure 10*

A graph of a comparison of a number of players

AI-generated content may be incorrect.

#### **QUESTION 3:**

Finally, the last research question we wanted to explore was whether teams with higher home game attendance had more wins than teams with lower home game attendance. Our reasoning was that if a team was doing well, winning more frequently would attract larger crowds and more fans.

Our correlation matrix from Question 1 showed that total home attendance had a moderately strong correlation with wins. This was further evidenced by our linear regression model, as the coefficient value for total home attendance was 0.48. Although this value is small compared to the others, it still indicates a slight positive effect. We were able to visualize this small positive effect through a bar graph and a scatterplot with a trend line.

*Figure 11*

A graph of blue rectangular bars

AI-generated content may be incorrect.

*Figure 12*

A diagram of blue dots

AI-generated content may be incorrect.

### **CONCLUSION**:

Our data analysis answered various questions regarding what contributes to a team’s success. We found that team success can be measured differently, whether through regular-season wins or advancing to the postseason.

Our analysis showed that metrics like Runs Scored (RS), On-Base Percentage (OBP), Batting Average (BA), and Runs Allowed (RA) strongly correlate with regular-season success. The linear regression model accounted for nearly 87% of the variance in team wins. While the decision tree model was more straightforward to interpret, it fell short of the linear model, supporting Beane’s emphasis on advanced statistics such as OBP and RS for predictive insights.

We found strong evidence that home run totals correlate with playoff appearances. The boxplot and hypothesis test indicated that playoff teams hit significantly more home runs than non-playoff teams. However, home runs alone are inadequate predictors; only when combined with other statistics did our logistic regression model achieve high accuracy in predicting playoff qualification.

Total home attendance showed a mild positive correlation with team wins. Visualizations and regression coefficients indicate that successful teams attract larger home crowds, suggesting that success influences fan engagement, even though attendance is not a strong predictor of performance.

In the future, we believe this information provides greater transparency into what makes an MLB team successful and could even serve as a guide for managers and general managers when evaluating players.

1. https://www.baseball-reference.com/leagues/majors/ [↑](#footnote-ref-1)
2. https://www.kaggle.com/datasets/wduckett/moneyball-mlb-stats-19622012 [↑](#footnote-ref-2)
3. https://www.thebaseballcube.com/content/mlb\_attendance/ [↑](#footnote-ref-3)