**Modern Day Moneyball**



Finding new statistical patterns to win baseball games

Ronald Smith

# Introduction

Baseball is one of the most statistically predictive sports. A game of baseball is made up of numerous defined events that lead to one outcome. In addition, a standard season in the MLB is made up of 162 games per year, which is one of the highest number of games per season in all of professional sports. For these reasons, baseball records and winning percentages are one of the least variant sport outcomes.

MLB team general managers are responsible for building a baseball team’s roster. Their goal: Build the best team possible with the budget given to them. A unique factor of the MLB however is that there is no salary cap, so a team’s payroll is set by what the team can afford or are willing to pay. This creates a unique challenge for teams with a low payroll budget. In steps Billy Beane, the GM of the Oakland Athletics from 1997-2016. Oakland has historically been one of the poorest ball clubs in MLB and a payroll ranking in the bottom third of the league every year, except one, during the period. Billy Beane is considered the father of modern day baseball analytics.

This project has two objectives: 1. Analyze the Pythagorean Theorem of Baseball and optimize the formula, and 2. Identify statistical goals to look for in potential players. The raw data was downloaded from [kaggle](https://www.kaggle.com/pschale/mlb-pitch-data-20152018). From the raw data, I cleaned and scraped the resulting seasonal team data. This final data that was used for analysis were the offensive, defensive, and record team statistics.

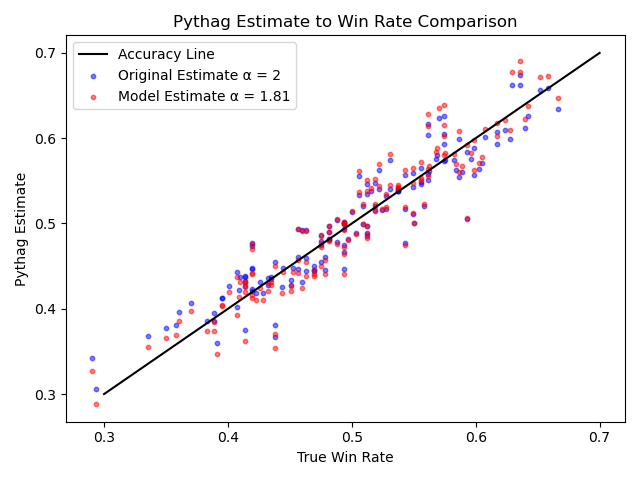
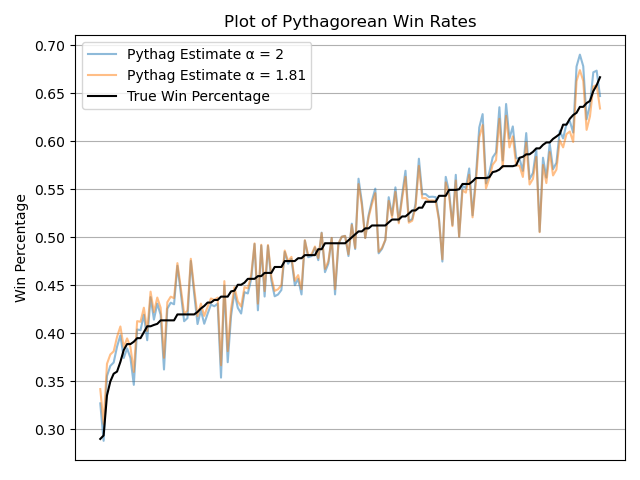
# Part I: Pythagorean Theorem of Baseball

Corresponding Jupyter [Notebook](https://github.com/rs314314/Baseball/blob/main/code/PythagoreanTheoremOfBaseball.ipynb)

The now famous statistical equation used to predict wins is called the Pythagorean Theorem of Baseball defines a relationship between runs scored and allowed, and wins of a given team. The equation: takes into account the overall offensive and defensive productivity of a team and calculates the expected winning percentage of that team. The goal of this analysis is to determine an optimal value of *alpha* (exponent) other. The model that I use to determine this is a self-made try all possibilities model. I randomly separated the data into training and testing data (portion = 0.2). The model will take the training data and estimate win percentages for every alpha between 0 and 4 on the scale of two decimal places. The model will then return the alpha that corresponds to the smallest mean absolute error (MAE). After that, I made predictions using the testing data and analyzed the accuracy and precision of the model by calculating the MAE of the test predictions. The model output an alpha value of 1.81.

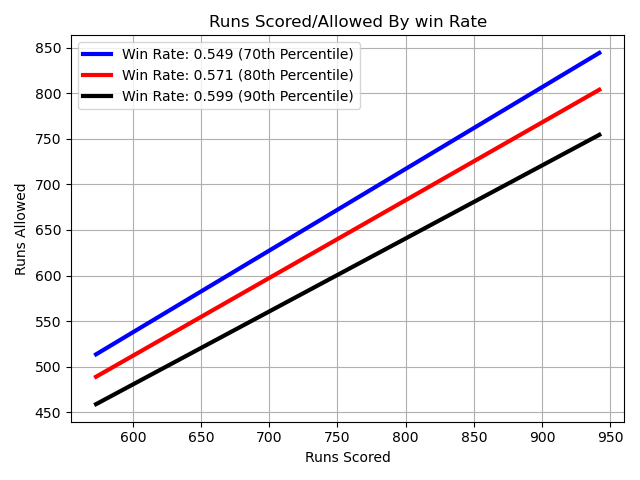
To determine the significance of the alpha value, I ran 100 simulations of the model and executed a Z statistical test (99%). The Null Hypothesis is that alpha equals 2, and the Alternative Hypothesis is that alpha does not equal 2. After running the test, it was determined that at a 99% confidence level, I reject the Null Hypothesis that alpha does not equal 2. Furthermore based on the results of the simulation, a Confidence Interval (99%) reveals that, using this method, the optimal alpha value is 1.81 ± 0.01.

The following two charts are visualizations comparing the true win percentages and the pythagorean estimates:



The correlation of the model to the true win rate is .94, which means that there is a very strong correlation.

Based on the model, I created this plot to visualize the necessary runs scored and allowed for three given win rates:

The blue line represents a win rate in the 70th percentile (.549) historically from 2015-19. This win rate is consistent with a probable playoff team. The red line represents a team win rate in the 80th percentile (.571), which is consistent with a probable Division winner. Finally, the black line represents a win rate in the 90th percentile (.599). This win rate is consistent with 97 win teams and potential favorites to win the World Series. 

# Part II: The Runs

Corresponding Jupyter [Notebook](https://github.com/rs314314/Baseball/blob/main/code/Runs%20Analysis.ipynb)

Winning games is simple: score more runs than the other team. The more runs a team scores and the fewer runs that team allows over the course of the year will generally correspond to a good win rate according to our model from Part I. From the charts and models, I have determined that for a win rate of .549, a team should aim to score 771 runs and allow only 681 runs. For a win rate of .571 (probable Division winner), team goals should be 771 runs scored and 669 runs allowed. And to be considered a World Series contender with a win rate of .599, a team should aim for goals of 826 runs scored and 629 runs allowed. Attached at the end notes are charts that compare various statistics along with their respective correlations. The goal of this section is to analyze and create a model that can estimate the runs scored and runs allowed by a team based on their statistics.

Runs allowed are fairly straightforward. There is a .94 correlation between opponent slugging allowed and runs allowed. Therefore, I ran a one-dimensional linear regression using slugging to predict runs allowed. To be considered an elite defensive team (90th percentile in runs allowed), a team would need to allow 629 runs or less. Using the regression model, 629 runs allowed corresponds to an opposing slugging rate of .413.

I found an interesting story concerning offensive statistics and runs scored. According to the correlation chart for runs scored, most stats have a low correlation and for some statistics, there are negative correlations when a positive correlation is expected, such as hits. Since the initial analysis provides little information, I ran a ridge regression amongst all the variables. Using the following team statistics we got these coefficients:

**Stats Used**: ['singles', 'doubles', 'triples', 'home\_runs', 'walks', 'sac\_fly', 'sac\_bunt', 'strikeouts', 'hits', 'ab', 'ops', 'avg', 'obs', 'slugging']

**Coefficients:** [-0.23444059, -0.1423113, 0.05558465, 0.20742606, 0.01587721, -0.28537714

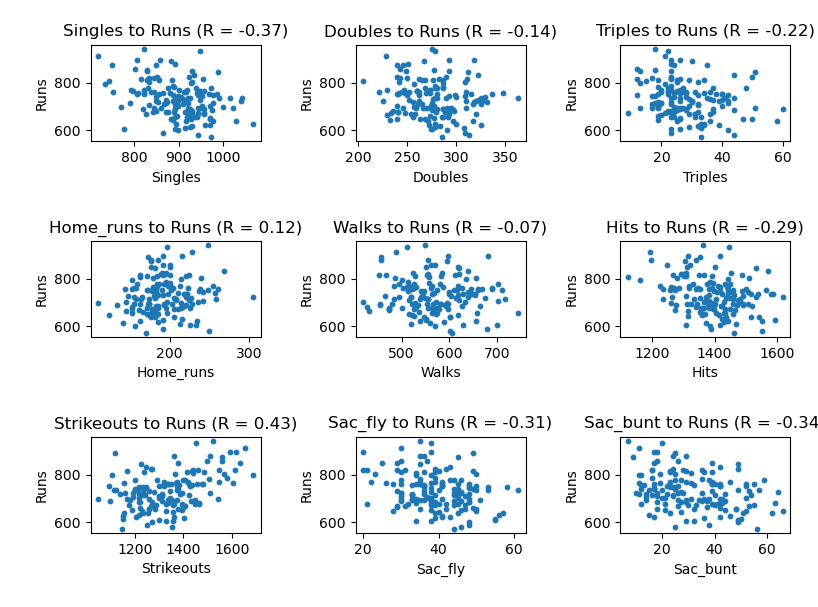
-0.98735713, 0.07960208, -0.11374118, 0.18265268, -0.25160548, -0.10286456

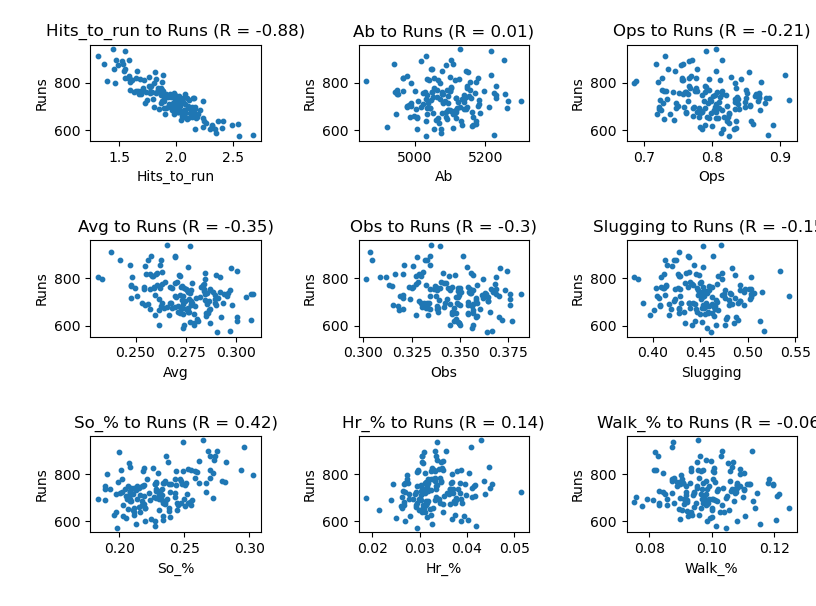
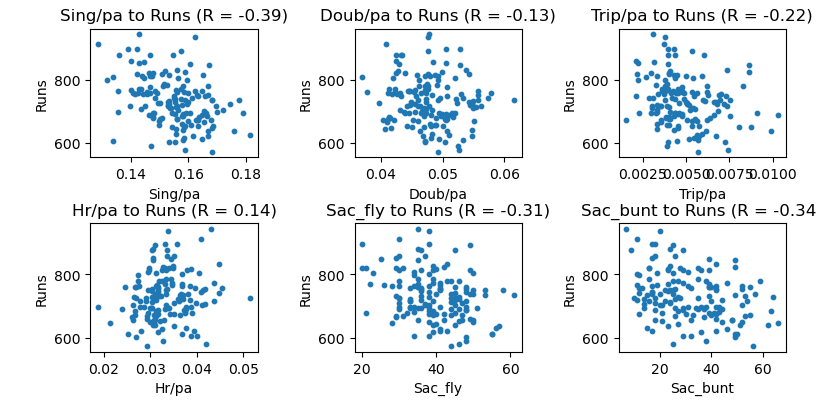
-0.1591658, -0.09243968]

The results of this linear regression logically does not make sense. This is further confirmed by the score which is (R2 = 0.265). Therefore, I conclude that more in depth statistical analysis is needed to determine how a team scores runs and how to calculate an expected runs scored for a team.

# Conclusion

This project had three purposes in mind. First, I ran a model to determine a more exact alpha (α = 1.81) for the Pythagorean Theorem of Baseball. With this information, we determined theoretical goals for runs scored and allowed needed to hit different important win rates. Secondly, I determined a way to accurately estimate a team’s runs allowed from defensive stats. The key statistic to runs allowed proved to be opponent slugging. Lastly I analyzed the effects of runs scored and key offensive statistics. This analysis was inconclusive and needed further examination.

Offensive Correlations



Defensive Correlations