Money Ball Analytics Inc.

Jacob Brown

## Background:

Money Ball Analytics Inc. draws inspiration from the revolutionary approach of using statistical analysis to predict baseball outcomes—an idea popularized by the real-life Moneyball strategy. As a lifelong baseball fan and devoted Seattle Mariners supporter, I set out to explore the potential of predicting a team's season win total based on its performance metrics. This project merges my passion for the game with my expertise in data analytics, utilizing historical team data to build a predictive model.

## Impact:

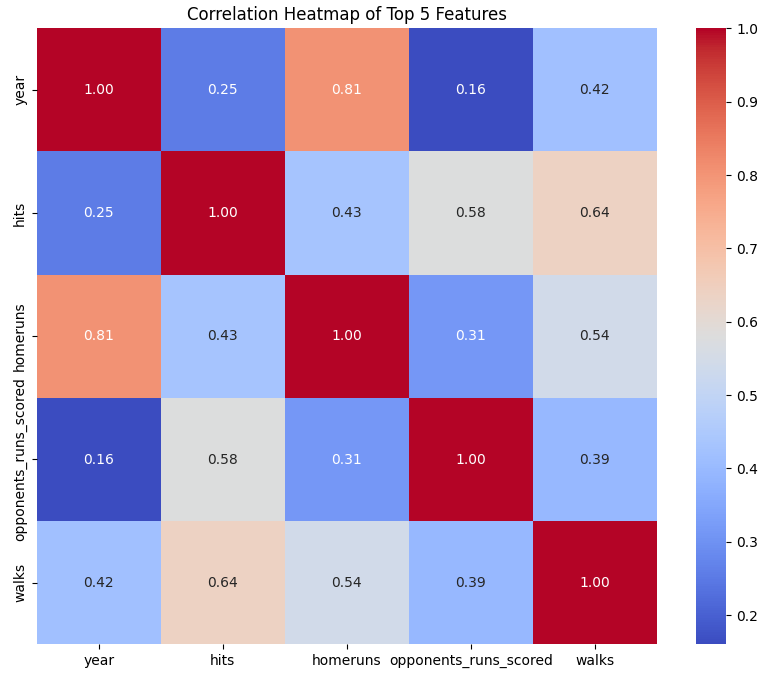
**Results & Actions:**

Our predictive model achieved an impressive 90.5% accuracy, with an average error of just 4.6 games per season. This insight enables Money Ball Analytics Inc. to build more competitive teams by evaluating individual player statistics and forming rosters that maximize the chances of winning. The actionable predictions provided by the model can guide decisions around player acquisitions, resource allocation, and overall team strategy.

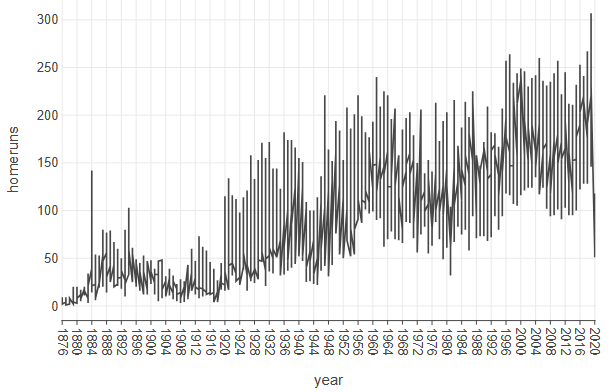
The model can also be used to predict win totals for a given season by analyzing player statistics and team dynamics. Organizations can leverage these forecasts to optimize their roster-building process, ensuring that the sum of individual player stats aligns with projected win totals for the season.

**Limitations:**

* **Data Quality**: Cleaning historical data posed challenges, particularly with columns containing significant null values.
* **Feature Selection**: Ensuring the model avoided overfitting required careful consideration, as overly predictive variables like losses or rank could distort results.
* **Historical Context:** Early baseball data lacked the consistency and completeness of modern datasets, which influenced model training.



This is a heat map showing the top 5 features that we used in this model. We can see that year home and home runs have a strong correlation .81 and this indicates that as we increase in year more home runs were being hit, which would lead to more games being won.



This would indicate that a great prediction of how many games that a team will win, is best indicated by year and home runs with strong growth rate over the years.

## Methodology:

he analysis began with an exhaustive search for historical data, ultimately collecting team performance statistics dating back to the 1870s. Data cleaning focused on addressing missing values in columns like league\_id, rank, and world\_series, which were often absent due to the lack of formal leagues in the 1800s.

Feature selection was another critical step. I excluded variables like losses, runs\_scored, and rank, which could have overly influenced predictions and led to overfitting. I retained important metrics, such as runs\_allowed, which provided valuable insights into team performance and contributed meaningfully to the model.

1. Python Notebooks:

<https://colab.research.google.com/drive/1u5Df2ZzEczOKpfTF4R6DBLJDMUuULwW3?usp=sharing>