Baseball Team Performance Analysis

Brady Biehn

September 2025

1 Introduction

In Major League Baseball (MLB) teams' success depends on a variety of factors. This project explores whether roster size, player height, and weight are related to team wins, and whether these metrics can be used to predict future performance.

2 Data

This project uses Lahman Baseball Database which includes data for the MLB's entire history. More specifically, it uses the databases Teams, Players, Appearances, and Parks tables.

These tools were used to aid in the project:

- SQL (MySQL) for data extraction transformation
- R for statistical modeling and simulations
- Power BI for dashboard visualization

3 Methods

Several statistics and data engineering methods were used the exact insights of each are shown below:

1. SQL Data Engineering

- Extracted team performance and roster details (weight, height, games played, roster size)
- Computed rolling 3-year win averages and flexible joins with stadium data

2. Feature Engineering in R

- Labeled teams as "Heavy" vs. "Light" based on median roster weight
- Conducted k-means clustering to group similar teams

3. Statistical Methods

- **Regression Models**: Target: wins, Features: weight, height, roster size, avg games played
- **t-Test**: Compared wins between Heavy vs. Light teams
- **ANOVA**: Tested interaction between roster size weight group.

4. Monte Carlo Simulation
   - Simulated 10,000 synthetic teams to estimate expected win distributions
   - Produced confidence intervals for team performance projections

4 Results

The aforementioned methods gave results shown below:

- Roster size negatively correlates with wins (larger rosters, fewer wins)
- Heavier and taller rosters correlate with more wins, though effect size is small
- Heavy teams averaged 3 more wins than light teams (statistically significant)
- Regression models explained 17–18% of variance in wins (low $R^2 \rightarrow$ other unobserved factors matter)
- Monte Carlo simulation projected an expected 79 wins per season (95% CI: 67–91)