Baseball Team Performance Analysis
SQL | R | Power BI

This project explores the relationship between team roster characteristics (weight, height, roster size, games played) and team performance (wins) in Major League Baseball from 1958–2021.

The workflow combines:

- **SQL** for data extraction and data wrangling
- **R** for feature engineering, regression, clustering, simulations, and statistical testing
- **Power BI** for interactive dashboards

Repository Contents

- **Datasets** All raw and processed data used in the project
- **SQL** MySQL scripts for extraction and data wrangling
- **R** R Markdown file with full analysis workflow
- **PowerBI** Power BI file with interactive dashboards
- **PowerBI Dashboard Screenshots** Static screenshots of dashboard pages
- **Project Overview** Detailed write-up covering data, methods, and key results

Project Summary

- Built 3-year rolling win averages and created weight group categories (Heavy vs. Light)
- Developed linear regression models, t-tests, and ANOVA to assess roster impacts on wins
- Applied k-means clustering to group teams with similar roster and physical attributes
- Ran a Monte Carlo simulation (10,000 runs) to estimate expected win distributions
- Delivered an interactive Power BI dashboard for exploring trends, predictions, and simulations

Results

- Heavier and taller rosters averaged about 3 more wins than lighter rosters
- Larger roster sizes were negatively correlated with wins
- Monte Carlo simulation projected an expected 79 wins per season (95% confidence interval: 67–91 wins)
- Regression models explained approximately 17–18