Project 1

Beau Martin, Jacob Haight

Introduction

Baseball is the pastime of America, and because of its long running popularity we have data from the 19th century that can be analyzed. The data set that was used is the “The Lahman Baseball Database” which has each season's hits, at bats, strike outs, payment, and much more going back decades. We examined two main ideas throughout baseball's history, money and skill. We use common display techniques to show general trends in how different teams pay their players and how those payouts changed over time, while also showing percentage of hits, homeruns, and strikes over time, related with rule changes in the league.

Total team salary and the salary of the highest-paid player on each team are key indicators of a team's likelihood to rank higher, yet it's also important to consider player skill trends. The skill of batters seems to have slightly increased, while the pitchers skill seems to have stayed the same except for rule changes. Understanding these factors is crucial for analyzing a team's success which will lead to predicting future performance based on both financial investment and player development. These are the links for our presentation and code.

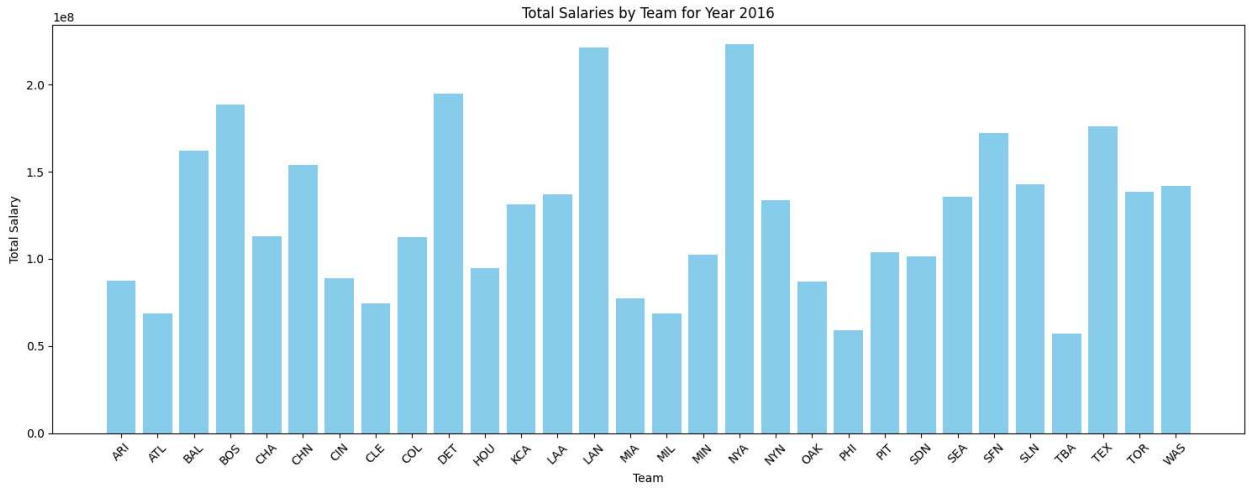
Dataset

The dataset used for our analysis is sourced from the "Lahman Baseball Database." We focused on four specific tables: salaries, teams, batting, and pitching. The salaries table contains information about players, the teams they played for, and the year of the salary data going back to 1985. The teams table provides the team names and their final league rankings for each year. The batting and pitching tables capture player statistics such as hits, home runs, at-bats, and strikeouts from 1871. With this amount of data long term trends should be easy to see through visual analysis.

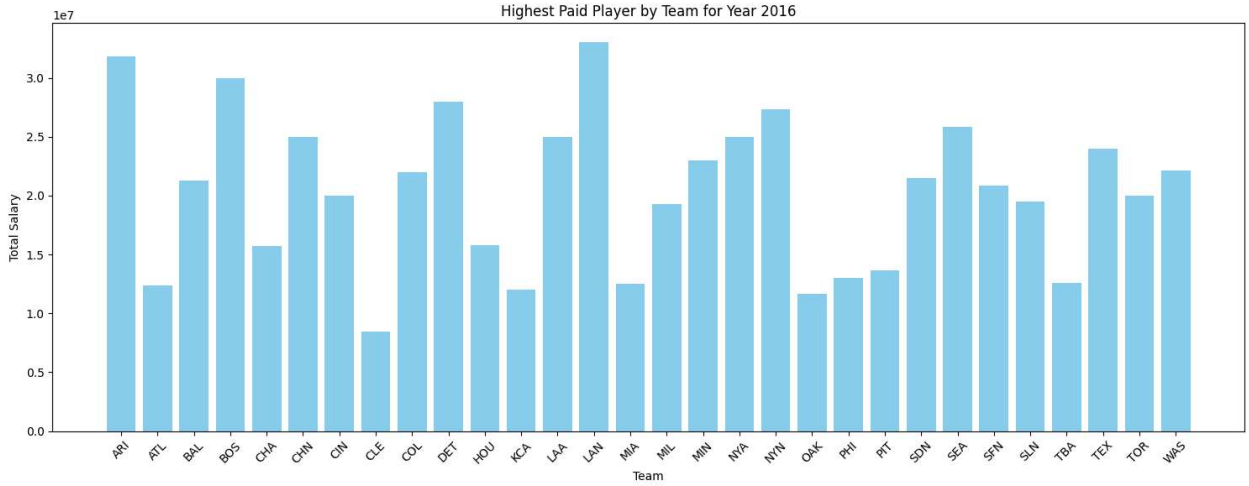
Analysis Technique

We grouped the column based on the relevant feature, and then used line and bar charts to perform a visual analysis. This method is simple but for data that has a long history like baseball there are interesting results that can be found by studying what the trend over time is. Drawing consultations about trends up, down and constant are easy to spot with such a long time horizon.

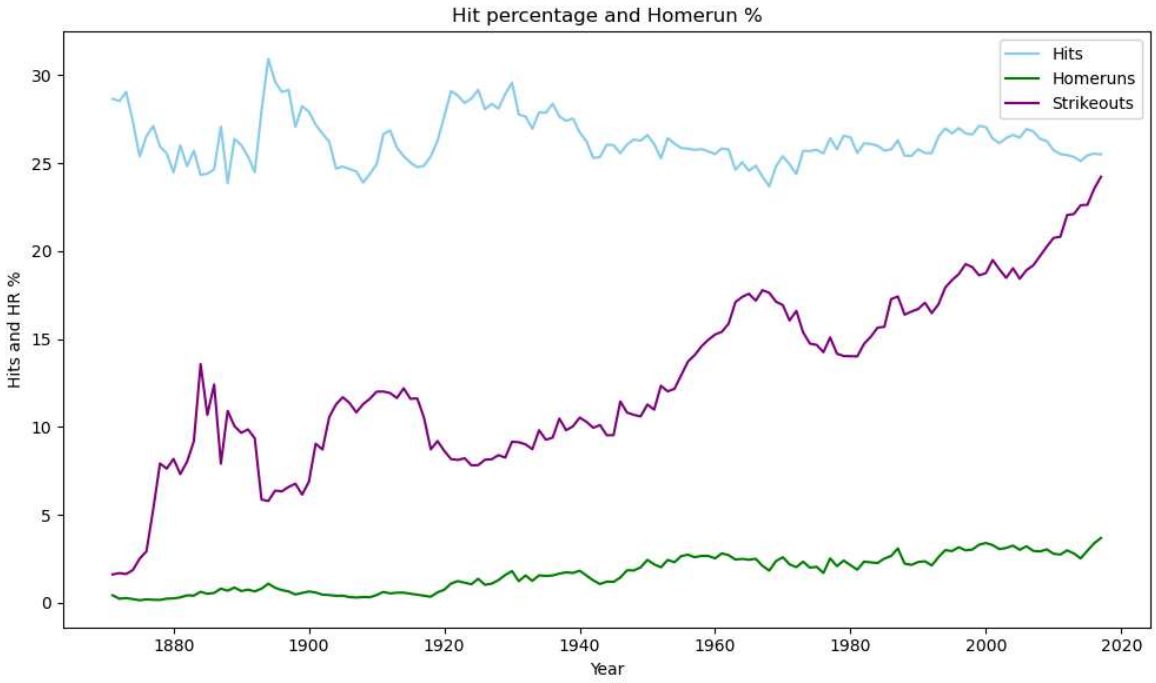
Results



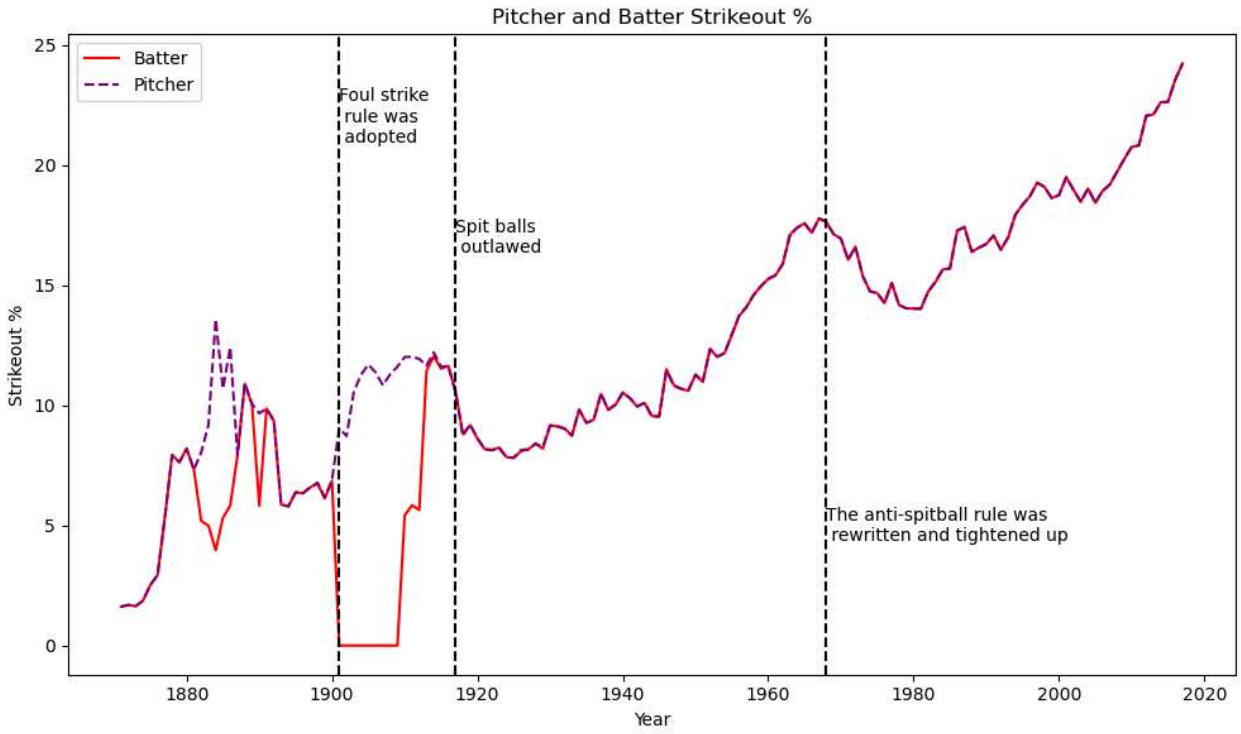
Teams can pay up to 222,997,792 and as low as 57,097,310 for all the players on their roster. It would make sense that the team with the most expensive roster would produce the best outcome but that is not always the case. For example, LAN with the total roster salary of 221,288,380 ranked first in their division but NYA who paid the most this year for their roster at 222,997,792 proceeded to rank 4 in their division this year. TBA had the lowest income of 57,097,310 and they placed last in their division. Generally, teams that spend more on their rosters tend to perform better in the overall standings, though there are exceptions where this pattern doesn't hold.



Clayton Kershaw was the highest paid player in the league in 2016 with a salary of 33,000,000. Clayton was a pitcher at LAN and helped them gain a first place ranking in their league that year. Carlos Santana had a salary of 8,450,000 which was the lowest salary among the highest-paid players on their respective teams in the league. Carlos played for CLE which also proceeded to place first in their division as well. Thus, we can’t completely make the assumption that the highest player on each team will make that big of an effect on a teams overall standings.



The above graph shows that the hit percentage has remained pretty constant after 1940 while the strike out percentage has increased throughout the lifetime of baseball. The strike out percentage has mostly increased for other reasons besides skill because of the constant state of hit percentages, one reason could be an increase in the number of bench players over time that are not supposed to be good batters. The home run percentage seems to be increasing slightly since 1920 indicating a small skill improvement as batters can hit more homeruns despite the increase in strikes.



The strike percentage has been greatly affected by rule changes over time, even just superficially in how the data seems to be collected, shown by the divergence from the batter and pitcher especially seen after the foul strike rule. Spitballs in particular seem to cause large downturns in pitchers' percentages. Even with the rule changes pitchers still see a massive overall trend, but this indicates skill has less to do with the pitchers results because of the percentage still being higher after the massive nerfs twice to the powerful spitball.

Technical

For our data preparation we loaded the csv files provided by "Lahman Baseball Database" into panda dataframes to help with ease of analysis. We used display analysis to show trends with player stats and used comparison analysis to show how teams total payroll would affect their overall standings. The percentages for the strike, and hits were taken over the total at bats for the entire league that year. During the analysis, one issue we encountered was that instead of finding salary data for a specific year, we kept retrieving salary data for all years in the salary table, which wasn't useful for determining how teams ranked at the end of a particular year.

For determining the skill of batters and pitchers, we had to show that the total number of players increasing (which was discussed in class), wasn't causing distortions. Which homeruns seem like a good way to account for because of the skill needed. The pitchers strike out trend had many strange things that had to be explained which needed more digging into baseball history to find out why it could change in ways the other lines could not.