Josh Urry

Project 1- Baseball

Introduction

This report examines multiple questions related to Major League Baseball (MLB) players and teams throughout the league's history. The data for this report came from a widely used baseball databank from the Chadwick Baseball Bureau, Various statistical techniques were used, such as descriptive statistics, independent t-tests, and Pearson's R correlations. Overall, the results are as follows: Average strikeouts per season varied with rule changes, players in the late 1800s had the most strike outs in a season, players that weigh less tend to steal bases more often, and since 2010, the Yankees have had the highest median salaries. My analysis will inform teams and fans about trends in performance throughout history, as well as factors that may or may not influence team success.

Link to presentation slides:
https://docs.google.com/presentation/d/1SFn7C85hQLGM2tEcAkXvQOT_xXxiF-oN/edit?usp=sharing&ouid=112003820866541637438&rtpof=true&sd=true

Link to GitHub repository: https://github.com/JoshUrry/cs6830_project1.git

Dataset

The dataset used for analysis contains 23 tables about MLB team and player performance from 1871 – 2021. The dataset is a free resource accessed from https://www.seanlahman.com/baseball-archive/statistics. The tables used in this report are the tables containing data for player demographics (people), batting, pitching, teams, and salaries. Each table contains unique numeric and categorical variables and can be joined together using any of the unique player ID, year ID, or team ID variables. As such, tables can be combined in multiple ways for robust analysis.

Analysis Technique

The first analysis looked at strike outs over time, so a line plot was used to visualize the strike average for each year. The second analysis sorted each player by highest thrown strike outs per season, as well as player average. Bar charts were used to visualize the difference between the top five players. The third analysis examined the linear relation between player weight and bases stolen. As such, scatter plots were used to visualize the data and Pearson's R correlations were run to test the linear relation between the variables. Lastly, a boxplot (ordered by decreasing median salary) was used to visualize the salary distribution per team since 2010. Kernel density estimate (kde) plots were used to show the differences in distribution. Independent t-tests were then run to see if there was a difference between distributions.

Results

The results of the first analysis are shown in Figures 1 and 2. *Figure 1a* shows that the average strike outs vary over time, with some sharp dips and rises. The vertical lines in the chart show when strike rules were changed, suggesting that some of the variation is related to those. See the appendix for rule changes. *Figure 1b* shows averages for right- and left-handed pitchers are over time. The overall average for left-handed pitchers in a season is 45.8, while the overall average for right-handed pitchers is 46.1. These overall distributions are not significant ($t = 1.05$, $p = 0.29$). This suggests that throwing hand does not have a significant effect on average strikeouts.

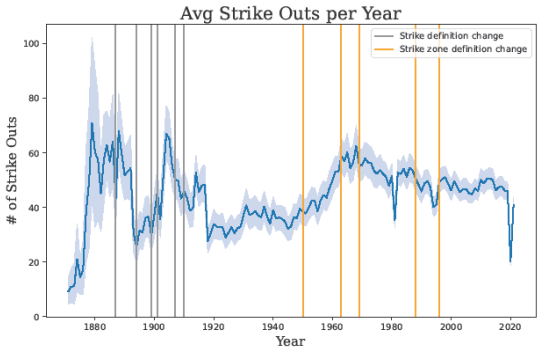

Figure 1a

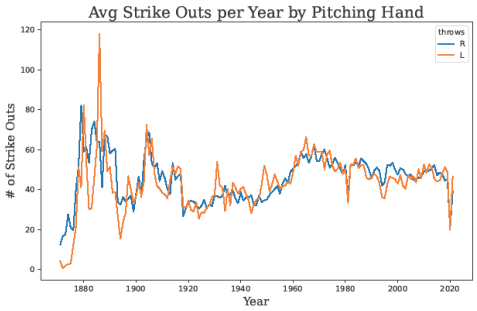

Figure 1b

Figure 2a shows a bar chart of the players with the top five strike out counts in a single season. These numbers were much higher than the overall mean of 46, and they were all in the late 1800's, when strike rules were not as well defined. So, *Figure 2b* shows a similar chart with the player's average strike outs per season overlayed. This suggests that the max strike outs may not be a consistent measure of pitching performance, so *Figure 3* shows a bar chart of the players with the top 5 averages. Some of these players are from the early years of the league, but some are also from the middle and later years. Overall, these analyses do show fans leaders in pure strike out counts, but there are other pitching measures that could be considered.

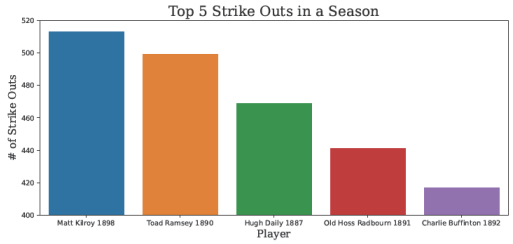

Figure 2a

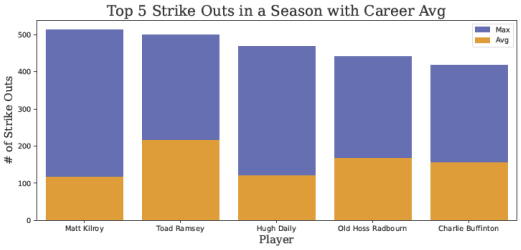

Figure 2b

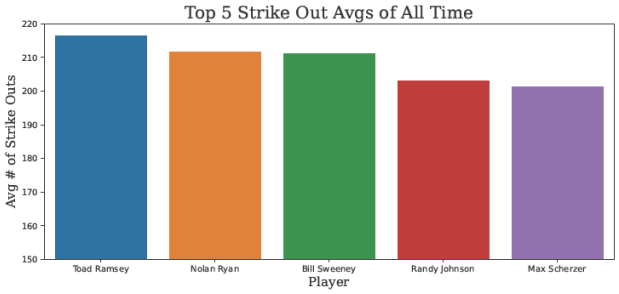

Figure 3

The distribution of player weights is shown in *Figure 4*. The weights have a fairly normal distribution. *Figure 5a* shows the scatterplot of player weight and bases stolen per season. There is a slight negative correlation ($r = -0.13$, $p = 5.27\,e^{-9}$). This suggests that lighter players get more stolen bases per season. After seeing these results, I hypothesized that heavier players get caught stealing bases more often and ran a similar analysis. *Figure 5b* shows that I was mistaken. The opposite was true. There was also a negative linear relationship ($r = -0.27$, $p = 3.09\,e^{-22}$). These analyses help inform a team of that, generally, lighter players will attempt to steal more, which will result in more stolen bases and more times caught stealing.

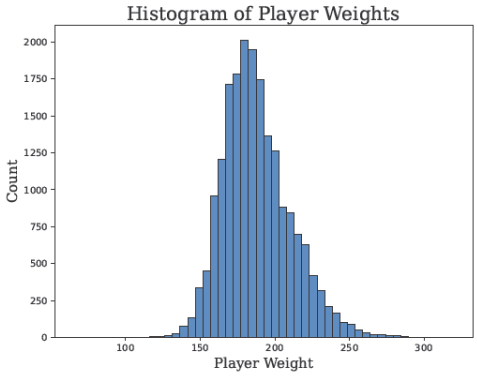

Figure 4

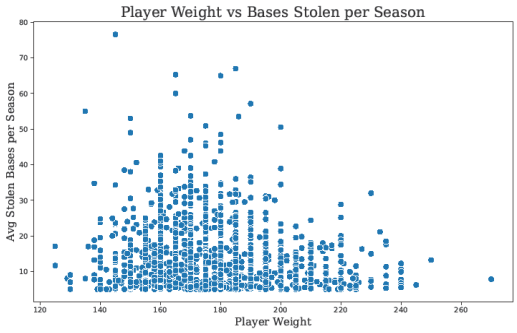

Figure 5a

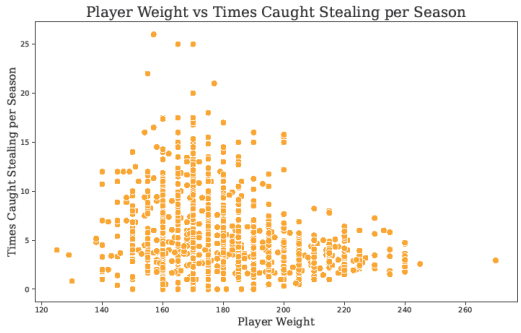

Figure 5b

Figure 6 shows that, since 2010, the New Yankees have had the highest salary distribution in the league. Since this was a fairly simple question and analysis, I dug into it further to see if, by a team basis, if a higher salary distribution translated over to more wins. I first looked at World Series winners since 2010 in *Figure 7a*, and that difference was not significant ($t = 1.35$, $p = 0.18$). I also looked at any team that had a division win, wild card win, league championship, or World Series win compared to all other teams (*Figure 7b*). That relation was significant ($t = 5.29$, $p = 1.3\ e^{-7}$). While there could be deeper analysis into this, these results generally suggest that having some high-valued players slightly increases chances of end-of-season wins.

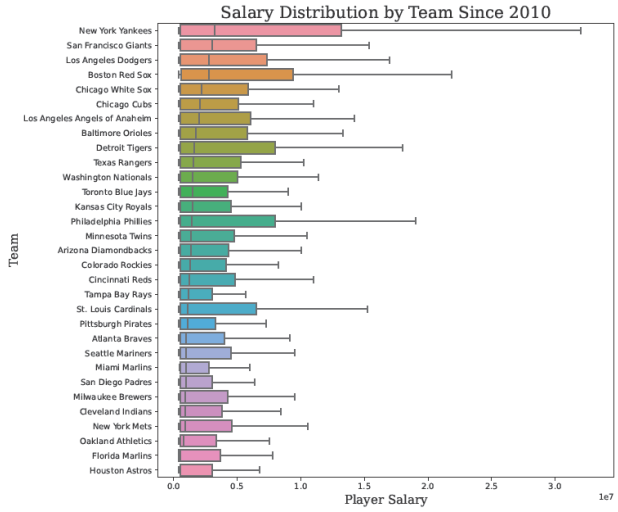

Figure 7

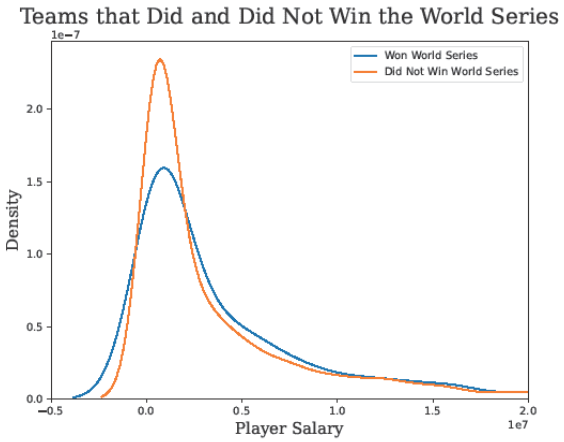

Figure 8a

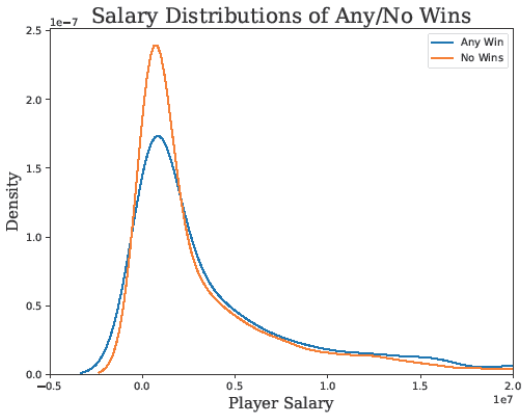

Figure 8b

Technical

These analyses involved some basic data preparation. All analyses involved merging the relevant tables (e.g., pitching, batting, salaries) with the people or team table to get variables like throwing hand, weight, or wins. I also had to subset all data frames and some point to isolate variables mentioned in the previous section. I have already mentioned why I used certain analysis techniques in the previous sections, but beyond descriptive statistics and visualizations, I ran Pearson R correlations when I was testing linear association (e.g., player weight and stolen bases), and independent t-tests when I was looking at differences between groups (e.g., salary distribution between World Series winners and all other teams). For each analysis, I started with a general question, then dug deeper into things I noticed. For instance, in analysis 1, when I saw the variation in the line plot, I looked up rule changes and added markers on the chart of overall trends. I also decided to look at player career averages instead of single season record for analysis 2 when there was a skew of early players. For player weight and stolen bases, I was incorrect in the follow up hypothesis that heavier players get caught stealing more often. It is possible that lighter players just attempt stealing bases more than heavier players, so they have both more successful and unsuccessful attempts. For the final analysis, after seeing different team name changes over time, and considering inflation as a confounding variable in salary comparison, I decided to just look at teams since 2010. More analysis could be done for my final question to look at season wins, or number of all stars on a team, rather than solely salary distribution. There could be outliers throwing that off. Overall, however, I feel that my analyses are informative to fans who want to know league trends over time, or teams who want insight on performance.

Appendix

See the following web pages for the rule changes related to strikes: https://www.mlb.com/glossary/rules/strike-zone and https://www.baseball-almanac.com/articles/strike_zone_rules_history.shtml