Data Wrangling Project Report

By : Christain Nicholson, Isaiah Deason, Joe Fordyce, & Mattison Belardo

Introduction:

In our project we explored Major League Baseball (MLB) statistics, from the past 17 seasons, in order to determine pitching value and batting value. All the data used throughout the project was collected from baseball-reference.com and Wikipedia. We utilized the value statistic of Wins Above Average (WAA) as the core of our analysis. This is the wins added by this player above that of an average player. We looked at the total WAA by the pitching staff and the total WAA from batting. We also looked at team salaries attributed to batting and pitching to determine which approach is more cost effective in terms of wins. The main goal of our project analysis was to determine whether pitching or batting contributes more to total team WAA, and ultimately, if teams should target strong pitching or a strong lineup to increase WAA.

Data:

All the data used in this report was gathered from baseball-reference.com and scraped from Wikipedia. We downloaded and cleaned specific CSVs from baseball-reference, as well as scraped Wikipedia for statistics such as wins and losses for all teams.

The data from baseball-reference included the teams, the season, the batting WAA, the pitching WAA, the batting salary, and the pitching salary. We downloaded CSVs for each year and did some basic cleanup in excel before reading the CSVs into R and merging them.

Scraping wikipedia gave us the teams, wins, and losses. We were able to create a web crawling script that gathered our desired data from each wikipedia page that contained data for each different year. We were able to create a loop that did this for years 2005-2009, and a separate loop for years 2011-2019. We then individually scraped the years 2010, 2020, and 2021 before merging. In order for our scraper to work, at the end of the loop, we created a data frame from the scraped features. We then stored the data frame in a list at the end of each loop. Finally, we used the ‘do.call’ function in combination with rbind to create a single data frame out of the list which stored each year of data within the loop.

We added a season column in R to the wikipedia data to denote each season. By doing this, we ensure that we matched the data accurately from wikipedia with baseball-reference when we merged.

Before merging, we did some basic cleanup of features like matching team names that had changed throughout the years. Once we had merged the data into our final data frame called ‘complete’ we ordered the data frame by earliest to most recent season. When we merged initially, the data was organized by season but also by team. For our analysis we wanted the data to be ordered by season showing every team for that same season before moving to the next season. Finally, we wrote the complete data frame to a CSV which is read back in in the analysis and visualization script.

Before analysis, it's important to note that we observed the data for any potential outliers. In 2020, the MLB only played 60 games due to the COVID-19 pandemic. For the purpose of our analysis, we decided to remove the data related to the 2020 season for *certain portions* of our analysis. For our linear regression models, we did not want this data to throw off our results, but we decided to keep the 2020 data for the WL error and the analysis viewing the data over time. Specifically, the research questions involving linear regression do not include 2020 data, while the questions not including regression do not exclude 2020.

Our final dataset contained 510 observations with 2020 included and 480 observations after removing 2020.

All code for web scraping and data integration is included in the R script 'Web Scrape and Data Integration'

Data Sources:
https://www.baseball-reference.com/leagues/majors/2021-value-batting.shtml

https://www.baseball-reference.com/leagues/majors/2021-value-pitching.shtml

https://en.wikipedia.org/wiki/2021_Major_League_Baseball_season

Data Dictionary

Column	**Type**	**Description**
Teams	Character	Each unique team name
Season	Integer	Season (year)
Pitching_WAA	Numeric	Pitching WAA for a team for a season
Batting_WAA	Numeric	Batting WAA for a team for a season
Batting_Salary	Numeric	Total salary related to batters for a team
Pitching_Salary	Numeric	Total salary related to pitchers for a team
Wins	Numeric	Team's actual wins for the season

Losses	Numeric	Team's actual losses for the season
WL	Numeric	Wins over games percentage
Est_WL	Numeric	Projected wins percentage based on pitching and batting WAA
WL_Error	Numeric	Absolute difference between estimated WL % and actual WL %
Batting_salary_scaled	Numeric	Normalized salary by Minimum/Maximum scaling- batting salary on a 0-1 scale
Pitching_salary_scaled	Numeric	Normalized salary by minimum/maximum scaling- pitching salary on a 0-1 scale
PB	Numeric	Pitching WAA to batting WAA ratio
BP	Numeric	Batting WAA to pitching WAA ratio

Analysis:

The goal of this project is to examine the relationship between the Wins Above Average (WAA) baseball metric to team wins and team salaries. By examining these relationships, we can determine whether a pitching staff or a lineup is more valuable to team wins – and ultimately make a conclusion on whether teams should target a strong pitching staff or a strong lineup for a better record. We can also determine if pitching or batting is more cost effective at producing wins. Further, we will also analyze the dataset to identify any meaningful trends over the past 17 seasons in which our data spans.

Our specific research questions and results of our analysis are as follows:

1. **How well does team WAA represent the team's actual record? What is the estimated win loss ratio under the assumptions of WAA compared to the actual win loss ratio?**

   To determine the relationship between actual records and estimated records under WAA, we had to create two columns derived from features within our dataset. The first column we needed was the actual win loss %. Then, we needed to calculate the estimated win loss %

under the assumptions of WAA. Once we had done this, we could compare the two and create a third column, WL Error, which calculates the absolute error from the actual win/loss compared to the estimated win/loss.

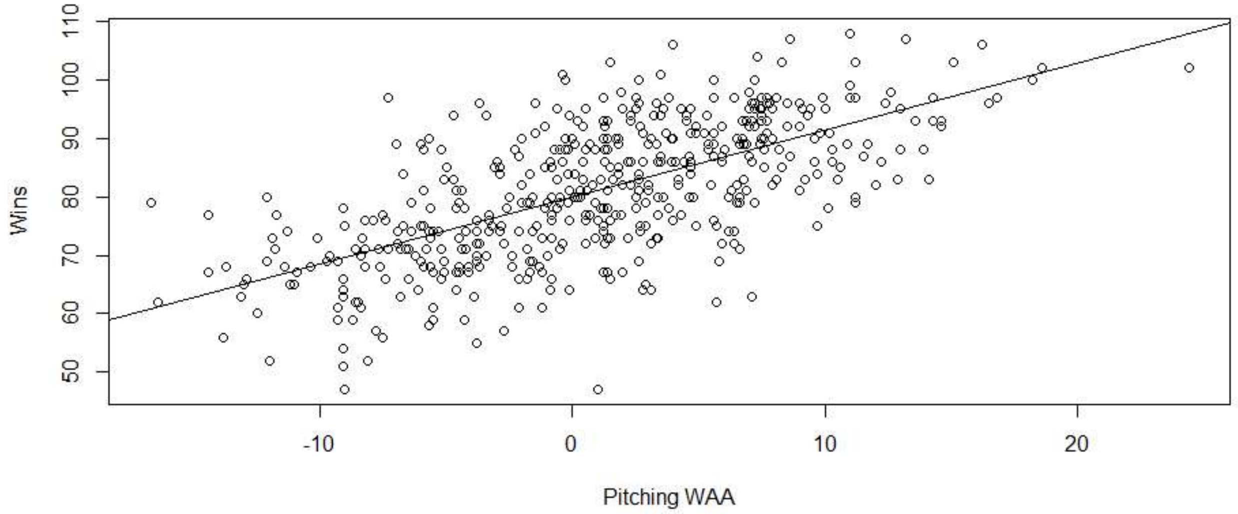
- To calculate estimated wins: Batting WAA + Pitching WAA * (total games played*0.5)
    - We then take estimated wins over the total games to produce an estimated win loss percentage
- We concluded that WAA is a good representative of a team's record. The average error of our error column was approximately 2.5%.
- One interesting finding is that the top four team/season combinations that had the highest error were all in the year 2020. This is likely due to the shortened season caused by the COVID-19 pandemic.

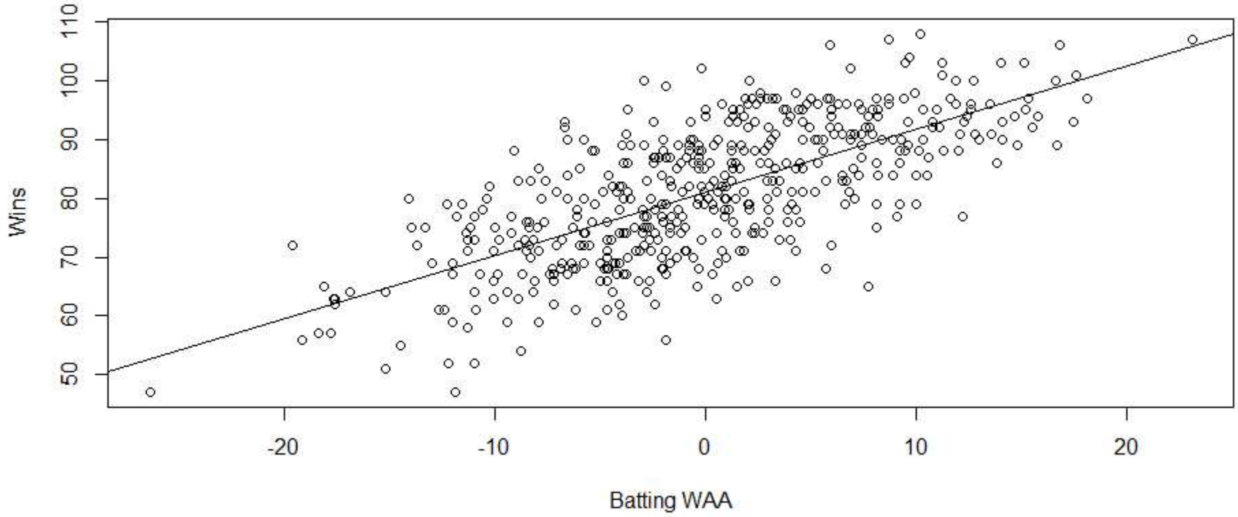
2. Is pitching or batting more important to wins in terms of the WAA metric? Does an increase in pitching WAA or batting WAA lead to more wins than the other?

Now that we know WAA is a good representative of total team wins, we can proceed to analyze whether batting or pitching is a better producer of wins. We used multiple linear regression to compare both pitching WAA and batting WAA to the dependent variable wins.

- Multiple regression produced an intercept of 80. This means that we can expect a team with zero pitching WAA and zero batting WAA (a team of average players only) to win a total of 80 games and lose 82 games, a win loss % of .49.
- The coefficients for the regression were as follows:
    - Batting WAA: 0.99340
    - Pitching WAA: 1.04771
- This means that holding pitching WAA constant, each unit increase in batting WAA results in 0.99340 more wins. Holding batting WAA constant, each unit increase in pitching results in 1.04771 more wins. This tells us that pitching WAA is slightly more important in terms of producing wins.
- P-value was 2.2^-16 which tells us that our explanatory variables are statistically significant.
- R squared was 0.8426 – meaning 84.26% of the variation in wins can be explained by batting and pitching WAA.

The following plots visualize the relationships between pitching and batting WAA and Wins.

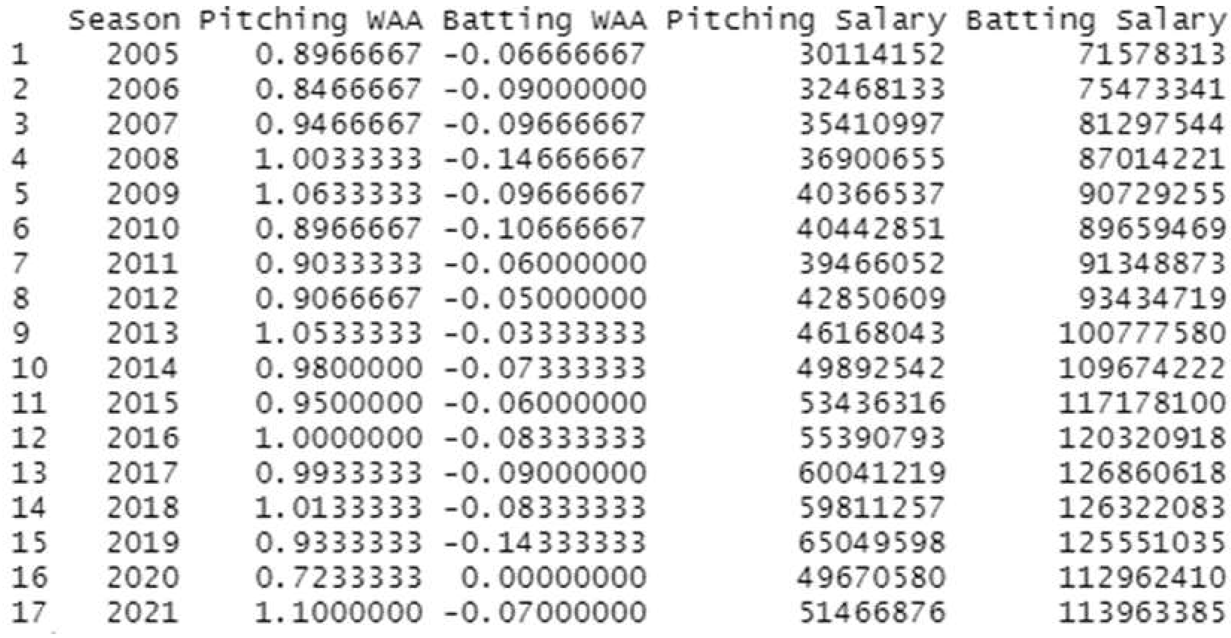




There is a positive linear relationship between both explanatory variables and the dependent variable.

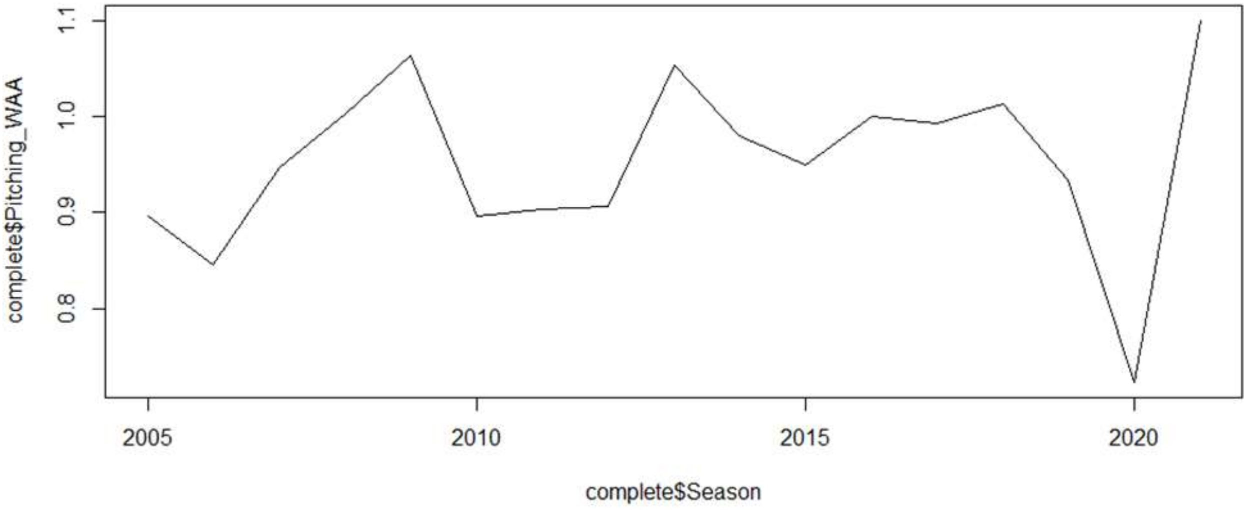
- Further, we ran correlations between wins and ratios of pitching and batting. The first ratio was pitching WAA / batting WAA to see if pitching, relative to batting, was more correlated with wins. The result of this correlation was -0.0028. The result of our batting WAA / pitching WAA to wins was -0.018. Since pitching/batting is the smaller negative value, these results suggest that pitching relative to batting creates slightly more wins. In this case, the correlations are negative because there are typically more sub-par teams that finish at or below a .500 win loss % than there is above average teams.

3. Has average pitching or batting WAA changed over the time range that our data covers (2005-2021)? What other trends can we find in the data that may be significant?

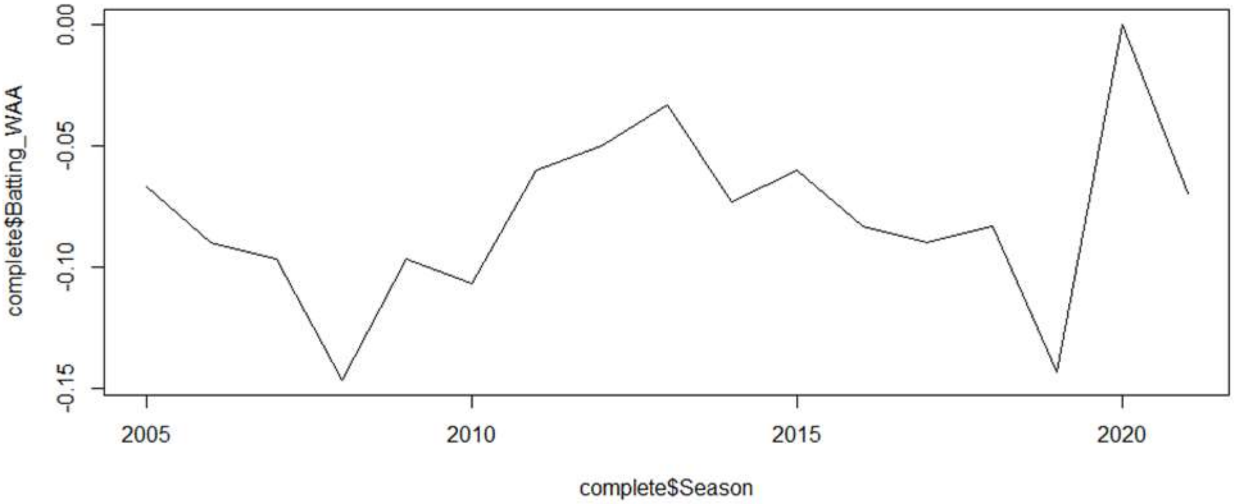


```
   Season Pitching WAA Batting WAA Pitching Salary Batting Salary
1    2005    0.8966667 -0.06666667        30114152       71578313
2    2006    0.8466667 -0.09000000        32468133       75473341
3    2007    0.9466667 -0.09666667        35410997       81297544
4    2008    1.0033333 -0.14666667        36900655       87014221
5    2009    1.0633333 -0.09666667        40366537       90729255
6    2010    0.8966667 -0.10666667        40442851       89659469
7    2011    0.9033333 -0.06000000        39466052       91348873
8    2012    0.9066667 -0.05000000        42850609       93434719
9    2013    1.0533333 -0.03333333        46168043      100777580
10   2014    0.9800000 -0.07333333        49892542      109674222
11   2015    0.9500000 -0.06000000        53436316      117178100
12   2016    1.0000000 -0.08333333        55390793      120320918
13   2017    0.9933333 -0.09000000        60041219      126860618
14   2018    1.0133333 -0.08333333        59811257      126322083
15   2019    0.9333333 -0.14333333        65049598      125551035
16   2020    0.7233333  0.00000000        49670580      112962410
17   2021    1.1000000 -0.07000000        51466876      113963385
```

- This table summarizes the average WAA for pitching and batting and salary for pitching and batting across the past 17 seasons.
- The average pitching staff salary has increased by 71% from 2005-2021, while the average batting salary has increased by 59% from 2005-2021.
- Average pitching WAA has always been higher than that of its counterpart, batting WAA. In fact, most years average batting WAA is negative.
- This data further reinforces the idea that pitching is more important than batting in terms of producing wins, and has been for the past 17 seasons.



- The figure above shows how pitching WAA has changed throughout the years.

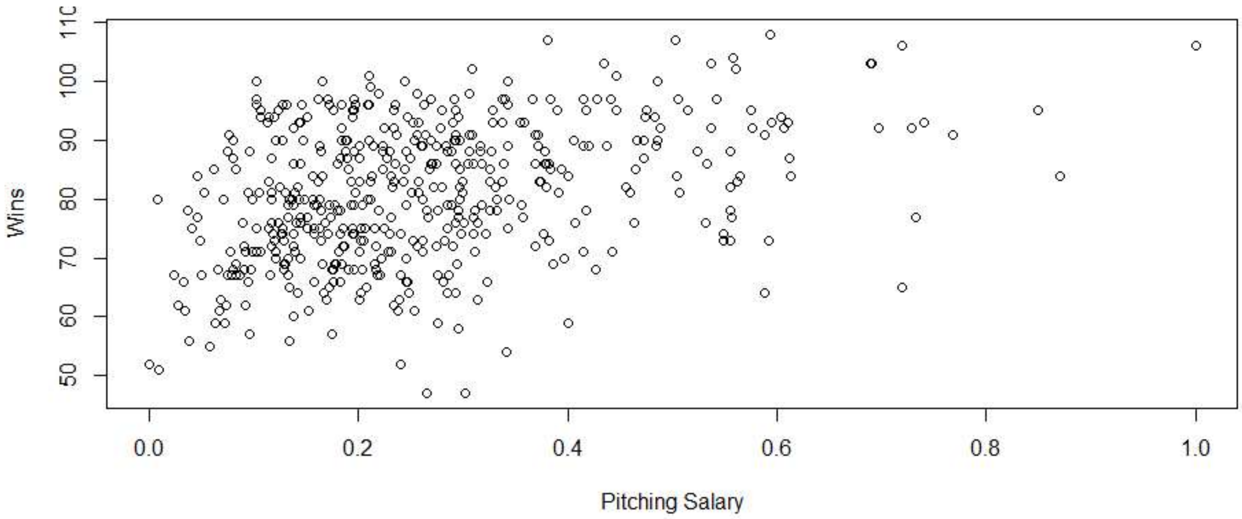


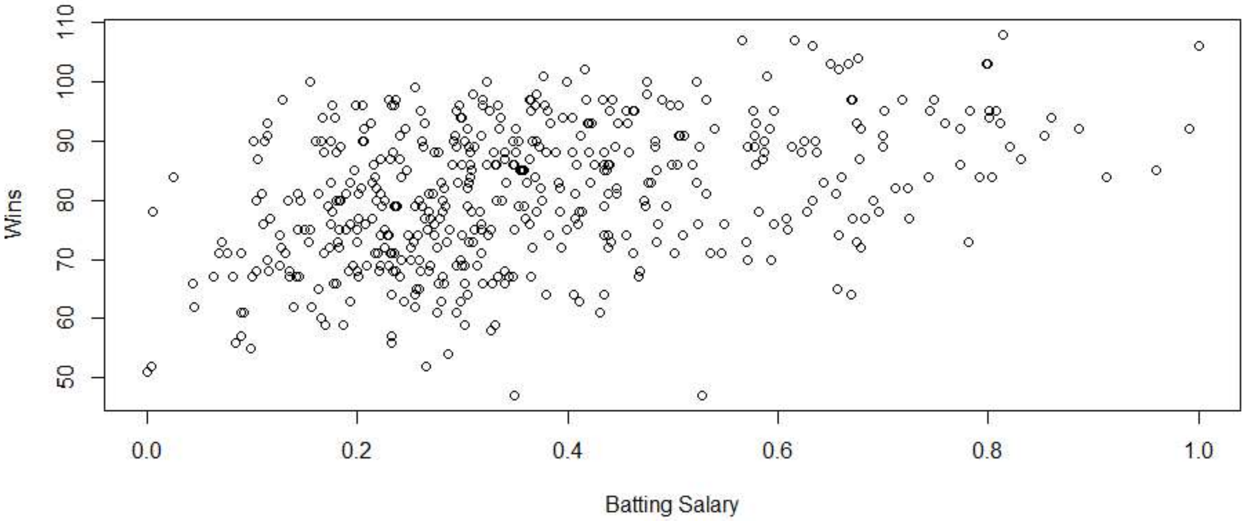
- The figure above shows how batting WAA has changed throughout the years.

4. What is the relationship between total pitching/batting salaries and wins? Which is more cost effective to produce wins?

To examine the relationships between salaries and wins we used multiple linear regression. First, we had to normalize the salary columns on a scale from 0 to 1. The results were as follows:

- Correlation Coefficients:
  - Pitching Salary coefficient was 4.959
  - Batting Salary coefficient was 20.878
- These coefficients tell us that for each unit increase in salary for pitching and batting, there will be a 4.959 and 20.878 increase in wins, respectively.
  - Since we normalized our salary feature before running the regression, this means that the top spending teams who have a normalized salary of 1 would see a 4.959 increase in wins due to pitching. The same is true for batting, the team with the highest salary for a batting lineup would see wins increase by 20.878.
- R squared of 0.1593 tells us that only 15.93% of the variation in wins can be explained by the salary figures. This makes sense as teams with the largest payrolls don’t always do the best and oftentimes teams with smaller payrolls do well.
- P-value of 2.2e^-16 tells us that our explanatory variables are statistically significant.





- In the above scatterplots we can observe the relationships between the normalized salary values and wins.
- The plots are similar however the batting salary is more evenly spread across the normalized values compared to the pitching salary.
    - Pitching salary is more concentrated on the lower end of salary, while wins still remain relatively high in this area.
    - From this, we can conclude that high payroll pitching staffs do not have as much of an effect on wins as high payroll lineups.
    - While both figures do show that higher salaries correlate with higher wins, salary is more important in batting compared to pitching.
    - It's important to note that the R squared value is only 15% in this cost analysis, so salary is much less of a factor in wins compared to WAA.

All code for analysis and visualization is included in the R script 'Analysis and Visualization'

Conclusion:

Should teams target a strong pitching staff or a strong lineup to produce a higher WAA and ultimately a better record?

After analysis of major league baseball data focused on Wins Above Average (WAA) and salary for the past 17 seasons, we can conclude that pitching brings more value to a team than batting. We conclude this after the evaluation of our linear regression models that show pitching WAA to have a stronger impact on wins. The results are close, with batting only being slightly lower. Further analysis supports pitching to be more important as well however. Looking at averages over the past 17 seasons, pitching WAA has always been stronger than that of batting, which has historically been a negative value. We can tie this finding back to the original formula that determines wins from the WAA metric: WAA*(0.5*total games) = estimated wins, and conclude that pitching has overall more weight in WAA, therefore being a better producer of estimated wins. This conclusion is backed up by the fact that WAA is a strong predictor of actual wins, in our first research question we determined that the average error between estimated wins and actual wins was only 2.5% over the past 17 seasons. Further, our correlation analysis also suggests that pitching relative to batting is better at creating wins. In terms of cost effectiveness in pitching and batting, we found that batting is more important than pitching. We can clearly visualize in our scatter plots that more spending on a batting lineup tends to reflect more wins than pitching. Small market teams more constrained by budget may want to consider first targeting strong bats to have in the lineup because our analysis shows that spending on batters tends to be more effective than spending on pitchers. Mid and large market teams should target pitchers first, however, due to their edge in producing wins. All else equal, teams should target a strong pitching staff to produce more wins, according to the Wins Above Average metric. We are placing more value on this regression compared to the salary regression in terms of wins due to the difference in R squared values. We believe our salary regression still provides value in whether pitching or batting is more cost effective, however. In conclusion, teams will need to target a mix of both good hitters and good pitchers, and given the mixed results of our analysis, it is clear that teams can be successful whichever route they pursue.

Limitations:

There are multiple limitations to our analysis to acknowledge. First, for the purpose of our analysis and for model simplicity, we only looked at the WAA metrics related to batting, and pitching. This excludes fielding (which has a very small impact on total team WAA) and also excludes instances where pitchers would hit from time to time. The only portion of our analysis this would have an effect on would be our error percentage between estimated and actual wins. Therefore, we do not believe our analysis was negatively impacted by this action because our error percentage was already very low. Excluding this from our analysis assumes an average level of fielding across our data. Further, our data only spans across the most recent 17 seasons, and deeper analysis with more historical data could potentially produce different results, especially given our evaluations of pitching versus batting were very close. There are multiple other factors that should be considered when a team would be deciding to pursue a strong pitching staff or a strong lineup, perhaps most importantly being the current makeup of a team.