CS6830_Project1 Report
By Logan Liddiard and Akhila Akkala

Links:

Github - https://github.com/Logsterx7/cs5830_project1
Presentation - CS5830_project1

Introduction:

Baseball holds a prominent position as a widely appreciated sport, attracting a substantial audience for both watching and playing. Our focus is on analyzing player performances within Major League Baseball, examining their contributions to individual teams and evaluating their efficacy across various leagues. Additionally, we aim to gauge audience interest and anticipation for these games. Through employing straightforward data analysis techniques and graphical representations, our objective is to provide insights that can assist players and teams in refining their performance. Moreover, managers can leverage this analysis to identify strategies for enhancing audience engagement and ultimately increasing revenue generated from these games.

Dataset:

We used a dataset covering Major League Baseball data from 1870 to 2017. This dataset includes a lot of info about players' batting, awards, pitching, salaries, and more. It's great for in-depth analysis because of its many features, helping us understand the ins and outs of baseball. Having a lot of data enables us to dive deep into various aspects of the game, resulting in a thorough and insightful analysis.

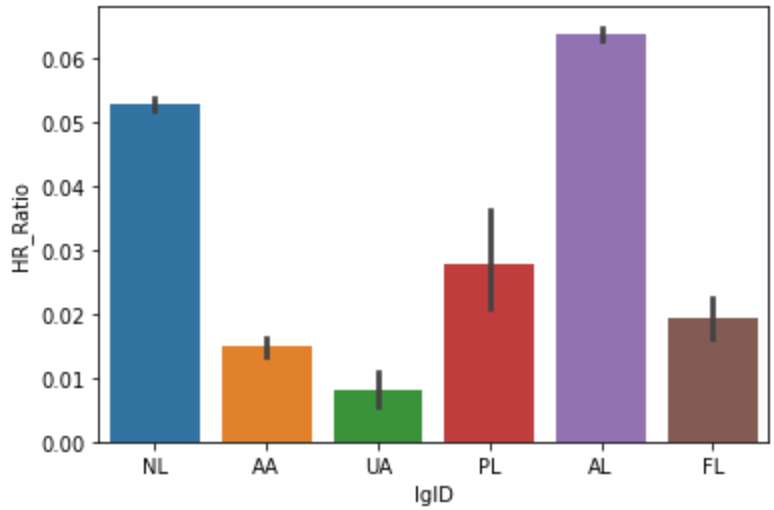
Analysis:

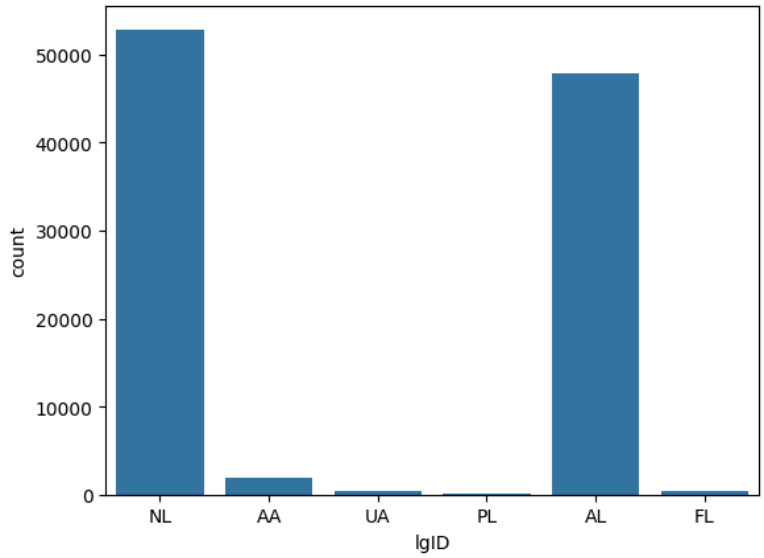
We carefully selected relevant tables from the dataset for our analysis. Our focus is on understanding players' home runs to hits ratio in different leagues and gauging the interest people have shown in the National League over the years. To do this we looked at the ‘Batting’ data set and divided home runs and hits over each other. We took out any players that didn’t have that data accounting for this. For our next idea we concentrated on the Chicago Cubs team’s park since it was such a recurring piece of data in the ‘Teams’ data set. We then took out every other team and only looked at years from 1980 to 2017 and found the audience attendance over the years at their home field and compared it to their win lose ratio for the corresponding year. We were able to find the teams win-lose ratio for the year by dividing wins and loses and adding a new column to our dataset called ‘WL_Ratio’ to grasp insights into team performance and player contributions. To make the findings clearer, we've presented the results using graphs.

Results:

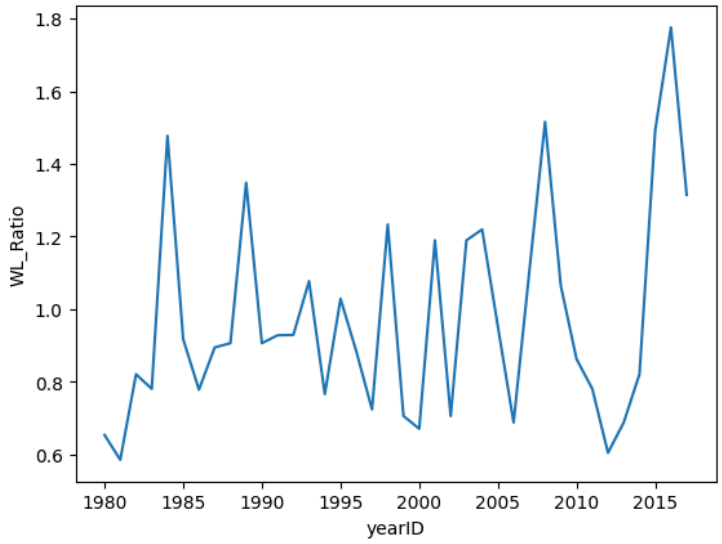
Home runs are often seen as a powerful feat in baseball, so we specifically extracted data on home runs and hits. Taking the ratio of home runs to hits, we calculated the average for each league. Our findings revealed that the American League has the highest average home run to hits ratio, followed by the National League.

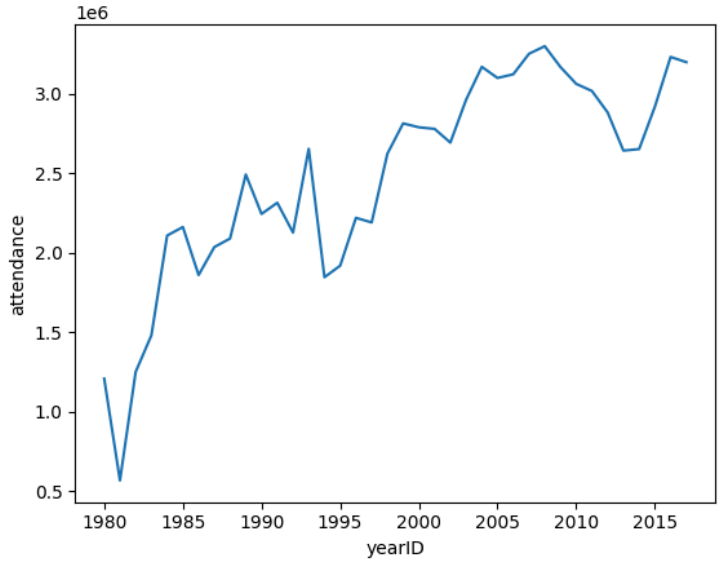
The first two graphs focus on showing the home runs to hits ratio average across the different leagues. After looking at our ratios, we can see the American League, or AL for short, has the highest home run ratio. Now we may think at first glance that this is purely because there are less players in the other leagues, but this actually isn't the case. The second graph shows us the player counts across the different leagues. We can see that the National league, or NL for short, has the highest player count. This shows us that across Major league baseball, the heavy hitters tend to come from the American League.





The second set of graphs focuses on the correlation between the Chicago Cubs win loss ratio and home game attendance at Wrigley Field. The first graph shows us their win-loss ratio over the years starting from 1980 to 2017. We can see that in the 1980s their win loss ratio was low. As time goes on however, they average out and steadily become better. We can see that their win loss ratio peaks towards 2017 right before hitting a bit of a losing streak. Now if we switch our focus to attendance at Wrigley Field, or their home park, we can see that attendance steadily goes up over time with their win-loss ratio. We can even see a dip in attendance when their win loss ratio takes a dip as well.





Technical:

As far as data collection goes we looked into the batting data set and took out players that didn't have any hits or home runs that were displayed as Nan. After that we divided home runs and hits to find a ratio which was then put in their own category to take an average for each league and compared that to the quantity of players in each league. For teams we only looked at the Chicago Cubs games past 1980 since it had the most data in the teams data set. We filtered out the other teams and any game before 1980. It made sense to look at that data set since it was smaller and we could easily compare win loss ratios and attendance together by looking at this smaller sample.

This technique was suitable for our findings because it helped us narrow a very broad scope of both players' home runs and attendance to wins and losses. There was a lot of trial and error. Looking at the home run to hits averages for all players initially didn't tell us anything interesting.

So then the scope shifted to showing the averages and density but because of this our graph sat to the very far left with the average home run to hits ratio being roughly 0.05 percent. So the scope of this changed to looking at them across each league and displaying player counts for each league to prove the significance of this finding. Despite the national league having more players. The American League has heavier hitters overall. The second problem was tough to deal with since when we first looked at the win lose ratio of each team, we were left with a very boring bell curve that didn't have any information to look at. So instead we looked at attendance for each team on a scatter plot which also didn't have anything interesting to tell. So we then decided to narrow the scope to a team that had the most occurring data. Which ended up being the Chicago cubs. So then we decided to look at their win loss ratio over the years. This data was good but a little hard to see since it was bunched up. We then decided to narrow the scope further by only looking at games after 1980 which gives us some nice data to correlate to our other graph of audience attendance at home games for the Cubs from 1980 to 2017.