**MLB Baseball Analysis**

**Introduction**

Our group chose to analyze a dataset for baseball statistics. Initial Dataset ranged from the year 1871 to 2015 and included hitting statistics such as Home Runs, Hits, Doubles, Strikeouts, etc. Before starting, we decided to limit the data and compare statistics only to the modern baseball era from 1970 onwards to limit the scope of data run for our analysis. To further clean this data, a minimum qualifier was added of 100 At-Bats. To avoid multiple outliners within our analysis, we removed any player who did not meet this minimum requirement from the dataset. Finally, once data was cleaned, the project was divided upon 3 sections – Strike Outs, Hits, & Home Run Analysis. Proposal for each Stat can be seen below:

1. Strike Out Analysis – Research was conducted by Brian Keuthan and includes a comparison of SOs against all other stat categories. Looking to plot scatter plots of SOs (per AB) and compare against R, RBI, H, 2B, 3B, SO, BB (per AB) to see which, if any, have a strong correlation to one another using R score as the statistical metric comparing both variables.
2. Hits Analysis –Research was conducted by Savannah Jeter and includes a time comparison of HRs per Year plotting individual players against the league average per year.
3. Home Run Analysis - Research was conducted by Lai-Ting Jacqlyn Cheng and includes a time comparison of Home Runs vs Hits (per year). Rather than focusing on individuals against the mean, analysis was conducted by grouping every 4 years and looking for a positive or negative trend to see if home runs per hits would improve.

**Strike Out Analysis**

For our first analysis, we chose to calculate the R Score Correlation of Strikeouts to other Statistics. Before any statistical data could be pulled, further data cleaning was required to run this comparison. To get a more realistic comparison of SOs against other statistics, it did not seem appropriate to simply compare SO totals of individual players and instead calculated each stat by Taking At-Bats (AB) / Statistics (I.E. AB / HR). In this analysis, we are attempting to see the relationship of Strike Outs compared to other stat groups and wanted to equally compare a player with 100 AB strikeout rate compared to a player with 500 AB. Without looking at it at a per-AB basis, it makes the analysis both predictable and skewed. See example:

Player A – 557 AB, 95 SO = 5.86 AB / SO

Player B – 152 AB, 56 SO = 2.71 AB / SO

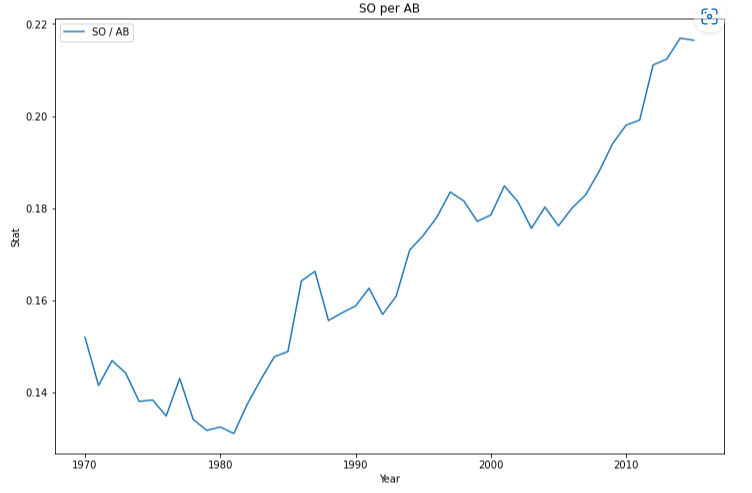
Player B strikes out at a much high rate despite significantly less SOs so by calculating AB / SO, we are able to better compare the data.

From the results, we can conclude that Strikeouts per At-Bats have little to no correlation to any other stat measure within the dataset. For the majority of stats, this does not come as a surprise and would not suspect SOs to have any correlations to several stats such as Stolen Bases or Doubles. I am surprised there was not more correlation to Home Runs or Base on Balls. I predicted that the best Home Run Hitters would strike out much more frequently by trying to swing for the fences. Likewise, for Base on Balls, I had predicted that they would show a much stronger correlation due to increased intentional walks to power hitters who I would have thought struck out at a higher rate. The strongest correlations were HR & H (both at an R score value of 0.30). At 0.30 it is difficult to determine any statistical significance of correlation between these 2 statistics compared to Strikeouts.

Chart, scatter chart

Description automatically generated

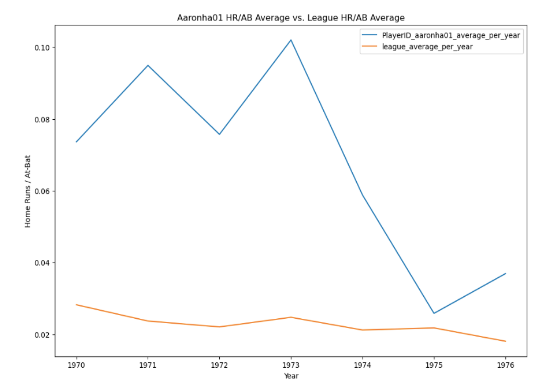
Additional Strikeout Analysis was conducted for time study for SO over the years. I was able to conclude that Strikeouts are on the rise since 1970. In 2015, hitters struck out on over 20% of Abs compared to 1970 – 1980 between 13-16% of Abs resulted in a strikeout.



**Player Average of Home Runs vs At-Bat Per Year Analysis**

In this analysis, we will be looking at the relationship between individual player home runs (HR) per year and the league average of home runs per year. We will be using a line graph to compare the individual player’s data to the overall player’s data to see if there are any correlations.

After randomly selecting players to visually analyze, we can draw a few conclusions. The average of the individual players compared to the overall average per year, did not affect the length of years the players played. Even if the players were below the league’s average, they still played for the same amount of years as players who were above the league average. We can also see that the individual players varied significantly with their averages each year, whereas the league average stayed relatively the same throughout the years with slight increases and decreases. The last 10 years of the overall league averages also have an increase. This shows that HR/AB averages have steadily been increasing over the years. This can be for many reasons, but mainly can be due to the significant changes and improvements in data analytics, stats and research that teams complete in order to recruit the best players. In conclusion, we can predict that the HR/AB average will continue to steadily increase while individual players will have significantly varied averages each year they play.



**Home Runs vs. Hits (per year) Analysis**

In this report, we will be analyzing the relationship between home runs (HR) and hits (H) in Major League Baseball from 1970 to 2014. We will use a scatter plot to visualize the relationship and perform linear regression to quantify it.

The scatter plot of HR vs. H shows a positive relationship between the two variables; showing a r-value for the relationship between HR and H is 0.977.

Chart, scatter chart

Description automatically generated

Based on the scatter plot and regression analysis, we can conclude that there is a strong positive relationship between home runs and hits in Major League Baseball from 1970 to 2014. This means that as the number of hits increases, the number of home runs is also expected to increase. The regression line provides us with a way to quantify this relationship and make predictions about the number of home runs based on the number of hits. The equation of the regression line is HR = 0.22 \* H + (-21087.37), which tells us that for every one unit increase in H, HR is expected to increase by 0.22. The constant term (-21087.37) in the equation represents the expected value of HR when H = 0. Overall, these results suggest that hits play an important role in determining the number of home runs in Major League Baseball.

Limitations:

This analysis focuses on the relationship between home runs and hits, but there are many other factors that can impact the number of home runs, such as the player's skill, the stadium dimensions, and weather conditions.

By simplifying the analysis to only consider two variables, important information may be missed. This analysis is based on data from Major League Baseball from 1970 to 2014. It is possible that some data may be missing or inaccurately recorded, which could impact the accuracy of the results.

**Additional Lessons Learned**

Lessons learned from conducting this research began very early in the project. It can be extremely difficult to find a dataset that is relevant to conducting research. Dataset was chosen without too much objective but after searching this is an extremely difficult to find even without much of an objective with analysis requirements. If attempting to find the data to draw a conclusion to a real-world problem or research, this would take an immense amount of work prior to conducting any analysis on the findings.