Project 1 : By Kartik Thakkar (A02426177) and Jason Jackson (A02330998)

Introduction:

The dataset we have performed our analyses on is related to the Sports domain, specifically, baseball. In any sport, there are many pointers that contribute to a better understanding of how the sport works on and off field. For example, for someone concerned with comparing how certain teams or players have performed or where do they stand, the game history that shows game centric statistics could be very helpful. Or someone whose goal is to find out how much money revolves around the game can analyze salaries, ticket collections and sponsorships.

Our analysis mainly focuses on two broad cases: non athletic aspects and athletic aspects relating to players, teams and their performance in some or the other way. The first two analyses we conducted were concerned with the hall of fame nominations including inductions for which we attempted to find out the history of nominations and corresponding inductions chronologically over the years and made some interesting observations and the salaries of teams over the year where we aimed at supporting a claim that good performance over the years could be one of the reasons of specific teams to have top average salaries. The last two analyses included the analysis of pitching metrics like ERA and WHIP with respect to wins where our goal was to support the claim that low ERA means low WHIPs and the claim that pitchers have longer careers the less complete games they throw.

For all four analyses, our techniques mainly included engineering new tables from existing ones and analyzing plots that compared the features which were relevant to support our claims.

- Link to presentation: https://docs.google.com/presentation/d/1o6n_dCNh68jH-o2inM6-5--S9BWMTsTVrzMgflIEFiE/edit?usp=sharing
- Link to Project1 github: https://github.com/jason-jackson6/project1

Dataset:

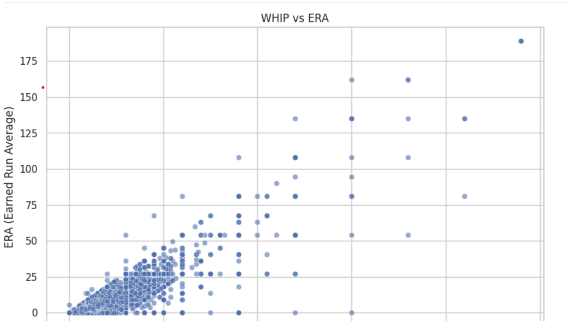
The baseball dataset consists of multiple records concerning game specific data like pitching, hitting, fielding statistics etc and also non game specific data like team salaries, franchisee records, awards history etc. The primary creator of this dataset is Sean Lahman. All the information in all the collections are relevant from years 1871 to 2014 but there could be more information in some of the collections till the year of 2016. The dataset offers a large collection of tables related to the game of baseball and is very resourceful for someone trying to analyze the game in any way.

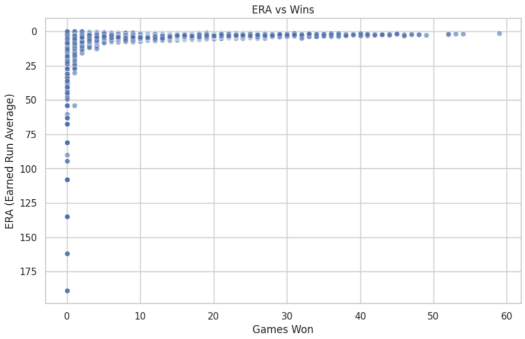
Analysis Technique:

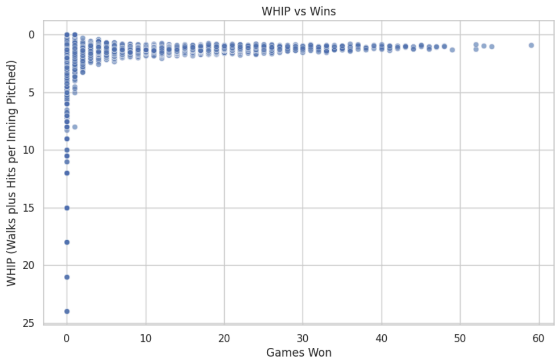
Majorly, for all four of our analyses, we followed the techniques like merging and engineering new features using existing tables so that we can support our claims that included columns that were not primarily present in the tables from the dataset. Additionally, we used Seaborn and Matplotlib libraries to plot our findings in ways where our findings could be represented in such a manner that it displayed maximum information without loading the person looking at them with unnecessary information.

Results:

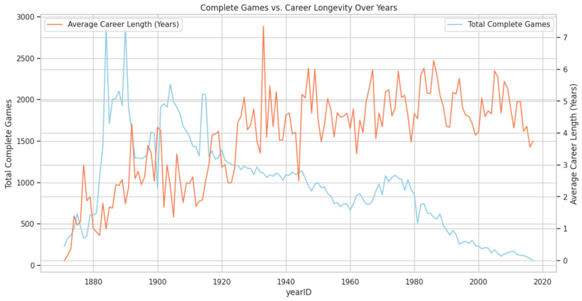
- For the claim that lower ERA means lower WHIP and vice versa and their relations with wins, we plotted 3 scatterplots.Figure 1 shows that for the most part, lower ERA correlates with lower WHIP, with some general outliers whereas Figures 2 and 3 reveal that lower ERA and WHIP correlate with more wins which is in correspondence with conventional baseball wisdom.
- This could be useful to see which stats result in more wins, as that is the most important stat to teams as a whole. This can then help teams decide which players to target that will help give their team the best chance to win.

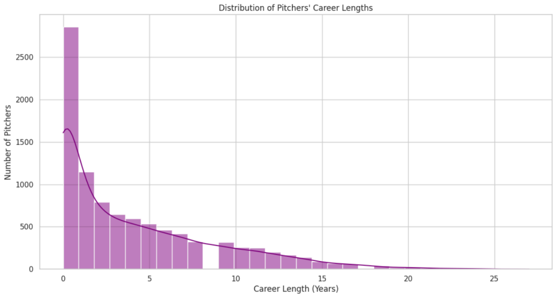




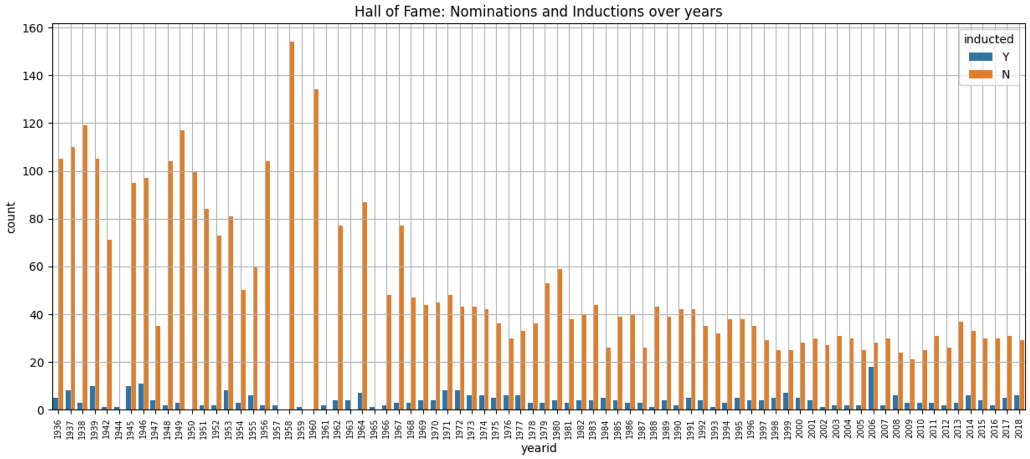


- For the claim that pitchers have longer careers the less complete games they throw, we plotted a double axis line plot in which total complete games and average career lengths were computed and plotted against years. Although Figure 4 does not indicate a clear increase in career longevity despite the reduction in complete games, Figure 5 shows the general trend of decreasing number of pitchers with increasing career lengths.
- This analysis could be employed for planning things like better sports medicine and better training programs, developing a more talented and larger player pool, or simply studying the natural physical limitations of pitching over many seasons.

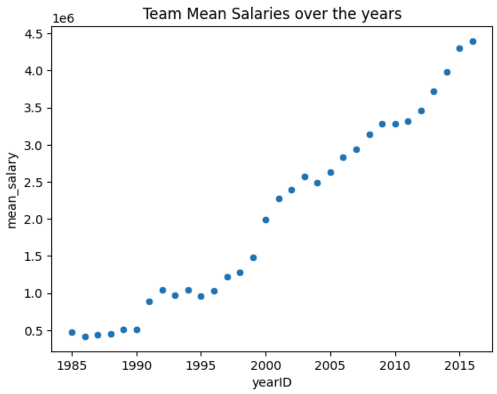




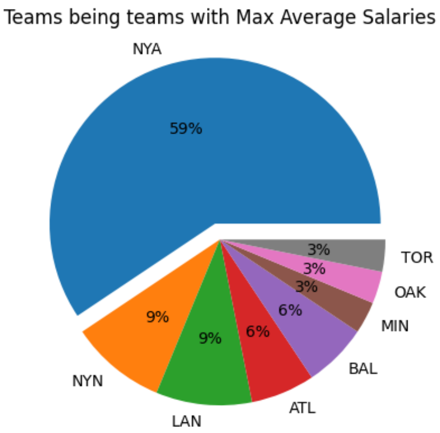
- For the analysis regarding Hall of Fame nominations vs inductions, we plotted a bar graph to find data that stood out (Figure 6). The first thing that caught our attention was in 1958. In 1958, no players were inducted to the hall of fame, but it had the highest number of nominations. The reason for this was because none of the nominations received the required 75% of votes.
- We also observed that 2006 was the year with the highest number of inductions, but not the highest number of nominations.
- We can use this information to better predict and study patterns in all future hall of fame votes.

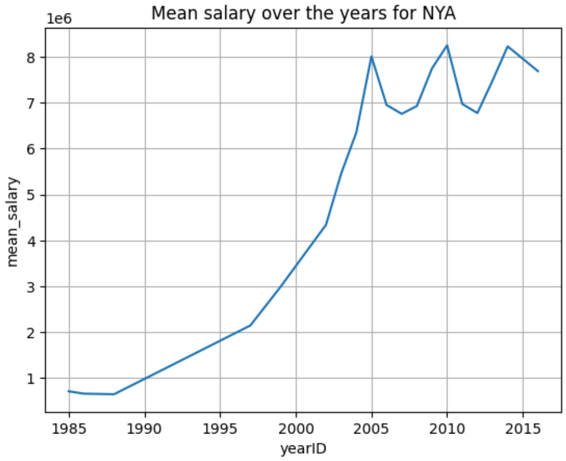


- Our final claim was that teas that are more consistently performing better, end up having higher salaries. We first plotted a scatterplot for the mean salaries of teams over the years (Figure 7). We found that salaries have been rapidly increasing over the years. This can be due to inflation, and the increase in viewers, funds, and sponsorships.

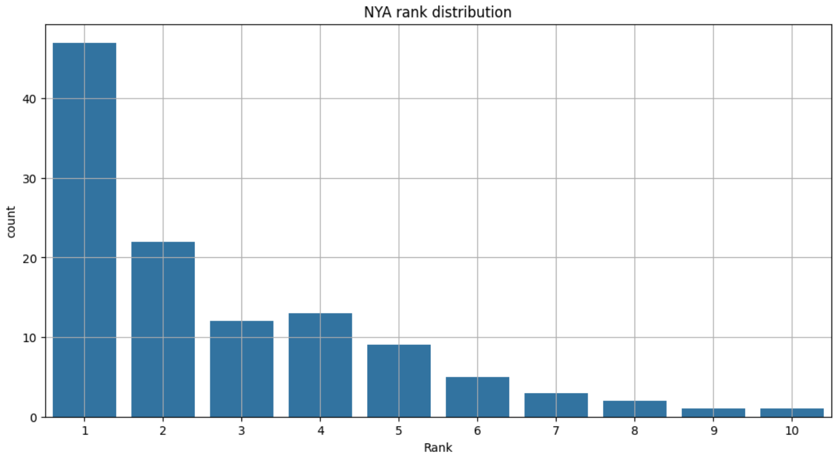


- We then plotted a pie chart (Figure 8) and a line graph (Figure 9) to find which team has had the highest mean salary, and how that's changed over the years. It turned out to be the New York Yankees (NYA).





- We finally plotted a bar graph to see how the Yankees have ranked amongst other teams throughout the years (Figure 10). As we can see from this figure, higher salaires definitely seem to have a correlation with the success of teams, especially in the case of the Yankees. We can use this data to discuss how salaires can be managed in the MLB. We can study trends to see when teams may be improving, and find more ways to balance teams in order to create more entertaining competition.



Technical:

- In terms of data preparation, we just had to create new data frames or series by merging and aggregating two or more features from different tables or just doing grouping based operations. For example, merging tables like salaries and wins and engineered data frames created as a result of grouping operations and merging them like career lengths.
- For analysis techniques, we wanted plots to support our claims or ideas. So, for all four of our analyses, we used line plots, bar charts, pie charts and count plots using visualization libraries like Seaborn and Matplotlib. For representing continuous data like mean salary over the years, we used line or scatter plots. For showing correlations of two features with one common feature, we plotted simultaneous line plots. For more categorical data like team rank distribution and team wins distribution, we used count plots and a pie chart.
- Our analysis process was simple. First, we brainstormed ideas to analyze or just randomly made statements and decided to check if the data corresponds with our ideas and thoughts. For example, are more salaries related to better performance? Does the data show that? Or, are pitcher career lengths dependent on how many complete games they throw? Finally, in the last step of the process, we found relevant tables and created numeric or categorical columns to finally plot them against one of the existing columns.
- Sometimes, we struggled with what kind of plots would better convey the information, like for the career lengths and complete games for pitchers, we realized that it would be better to show them on the same plot against the years.