Baseball Data Analysis Report

Satyanarayana Vinay Achanta
Aswani Yaramala

Introduction:

Data analysis on the Baseball dataset can provide helpful insights from existing data, which can help predict important analytics like winning strategies, avoiding injuries to players, players recruitment techniques, fan engagement and business improvement ideas etc. We used various plotting approaches to examine the data set and gained insights based on the curves' bump and peak spots in the plot or trend. Baseball strikes, the use of drugs by players in the Steroids Era, and changes to the game's rules, World War 2 are the leading causes of these plot fluctuations.

Our analysis of the top five schools from which most baseball players came will help young players who want to go professional in baseball to choose their schools as those schools provide the best training on baseball. Analysis of the top 10 players concerning their home run rate will help owners to recognize them and use this as a metric for salary. Owners can also ask these leading players to teach their techniques to freshers who joined their team. Also, statistics about baseball players can help us predict performance trends, salary, etc.

GitHub Link
Presentation Slides Link

Dataset:

Lahman's Baseball Database provides the dataset. Baseball statistics from 1871 to 2021 are included in the collection. It contains information from the two active leagues (American and National). It mainly contains data on players, pitching, batting, salary, wins, losses, and other details. PlayerID serves as the primary link between most tables. We used the following files to obtain our analysis.

- **Salaries.csv**: This file contains data on player salaries over the years.
- **People.csv** contains the players' names, birth dates, countries of birth, player IDs, and years of birth (if they are dead already).
- **CollegePlaying.csv** contains information about the schools (schoolID) the players attended.
- **School.csv** contains details on the school's name and location (schoolCity and schoolState).
- The file "**Batting.csv**" contains a ton of data about players' batting. However, we mostly used stats like at-bats (AB), doubles (2B), triples (3B), home runs (HR), stolen bases (SB), and strikeouts (SO).

Analysis Technique:

We started with data exploration, went through all our required tables, and analyzed the columns necessary for all the analyses we chose and how we could merge those tables. We also checked if there were NaN values and missing values. After merging the required tables, we used the data grouping technique to group the rows based on the columns needed, like years and playerID. Then we performed required aggregation functions like count, max, and sum for counting the number of rows, maximum count from the rows and sum, respectively. Once the required data is available, we use seaborn to plot the graph and visualize the data in charts. We used various plotting techniques, such as line, bar, and scatter plots, to look for trends and oscillations in the graphs. These basic charts can reveal a lot of information about the data. We framed our analyses as questions to look for correlations in the dataset.

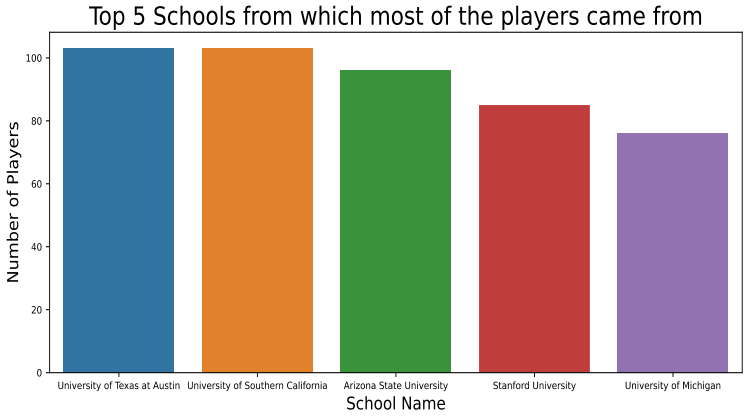
1. Analysis-1: Top Schools and Birth states from which most of the baseball players came.
    - To find top five schools from which most baseball players came analysis, we chose people, school and college-attended datasets performed merged on these three datasets and did group by playerID and yearID columns and took the last year of the player's school. Afterward, group by with school name and count aggregation function on playerID to get the information of top schools and the number of players.
    - To find top ten states from which most of the baseball players came analysis, we grouped by using birthState column in the people table and using count aggregate on peopleID to get the count of State names and number of players
2. Analysis-2: Stolen Bases analysis by year.
    - To find the Stolen Bases analysis by year, we took the batting dataset, and we have a column called SB, which is the number of stolen bases by players. We can group by yearID and use the aggregate sum of stolen bases to get each year with the number of stolen bases count. Then we plot sum of stolen bases and yearID in graph visualize how many bases were stolen by batters in all the years given.
3. Analysis-3: Home Run Rate analysis by year?
    - To find the top 10 players with the highest home run rate, we merge people and betting datasets. We first sum all home runs, at-bats, and strikeouts by year. Baseball Leagues calculate the Homerun rate using the formula HomeRunRate = HomeRuns / (AtBat - Strikeouts). Afterward, we can plot yearID and homerun rate to get the homerun rate average for every year.
4. Analysis-4: Average player's salary throughout all available periods. Which team got the highest average salary?
    - To find the average salary by year, we will be using the salary dataset. We can use aggregate mean to get salary average and group by each year, and plot the average salary and year in the graph.
    - To find the average salary of each team, we can use aggregate mean to get salary average and group by each teamID instead of a year, and plot the average salary and teamID in the graph.
5. Analysis-5: Players who got the highest number of awards?

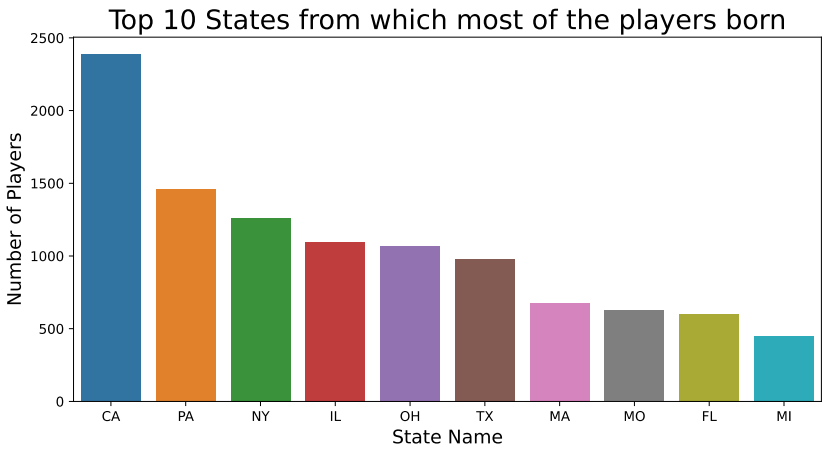
- To find the top 10 players who achieved more awards, we use awardsPlayers and people dataset. We can use aggregate count to get the number of awards and group by each playerID or nameLast, and we can sort the awards count in descending order and plot the nameLast and count of awards in the graph.

Results:

Analysis-1: Which college and state are most players from? Why?

For this, analysis, we got the information of school and birth state from which more players are from.





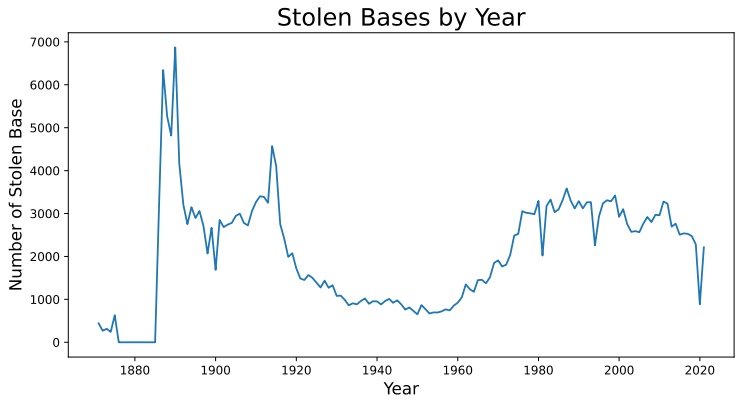
Based on the plot described above, most players are from the University of Texas at Austin and the University of Southern California. Players choose these top schools because UTA has athletic scholarships available; on average, 34% of all student-athletes receive athletic scholarships and win more baseball national titles. At the same time, USC also offers merit-based scholarships and admissions preference to athletes. Top universities have well-designed baseball programs, sports scholarships, infrastructure, and

coaches who train the players well. By this, we can also say that young players who want to go professional in baseball can choose these top schools, so they have a high chance of achieving it.

From the second graph, more players are born in California, which accounts for 21.8% of all US player births. California high schools are recognized for their commitment to athletics. Some of the top teams in the nation, whether in baseball, football, wrestling, or any other sport, are from California. These young athletes have a strong competitive nature, and frequently their teams are the pride of their communities.

Analysis-2: Is there any pattern in stolen base rate all these years?

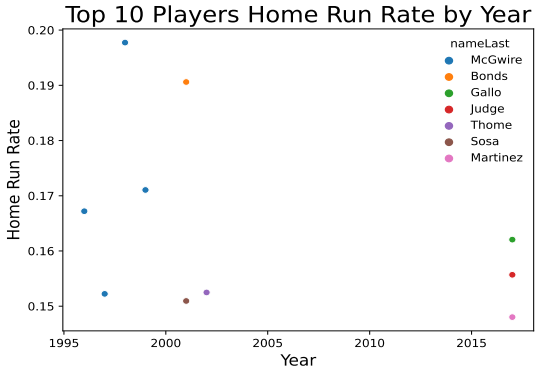
For this analysis, we plotted number of stolen bases for each year. Stolen base occurs when a baserunner advances by taking a base to which he isn't entitled.

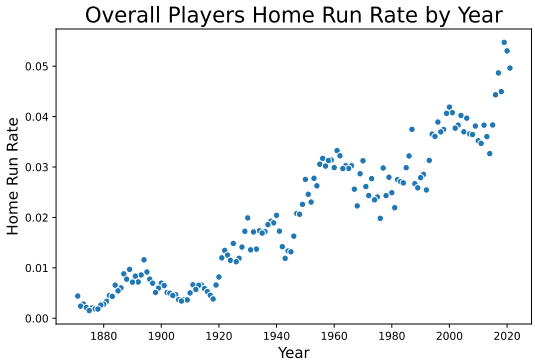


The reason for graph being flat near 1880 is stolen bases are not officially noted in baseball until 1886 and it was not until 1888 that it officially earned a place in baseball. We can see a U-shape in the graph overtime because in 1920 the rule for stolen base got changes and it was not liked by the players, moreover its very risky to perform a stolen base because of that rule. In 1951, as the number of stolen bases runs began to rise once more, as the rule got revoked to previous one again. This will help players with fast runs.

Analysis-3: Who are the top 10 players with respect to home run rate? Any pattern in the overall home run rate over time?

For this analysis, we calculated home run rate with formula rate = Homeruns / (AtBat - Strikeouts). Then we got top 10 home run rates, we can also see name of player who scored it.

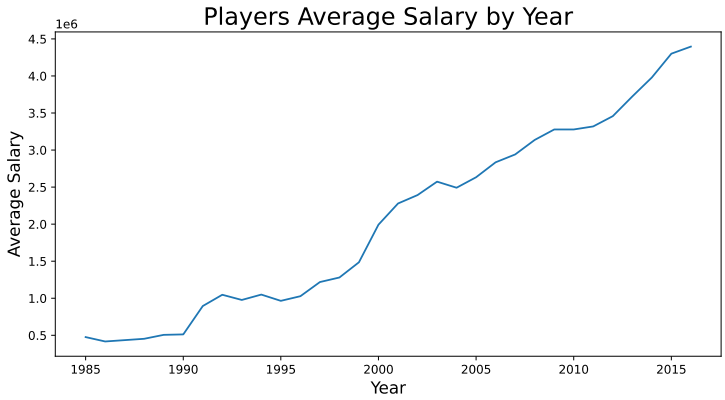




McGwire appears repeatedly in the top 10. The above plot shows that the years 1996 to 2003 had the highest home run rates. This is a result of the steroid era (early 1980-2000 late). Although steroids were prohibited in 1991, this wasn't put into practice until 2003. Because of this, the gap between 2003 and 2016 is also visible. We can also observe the same in overall players home run rate plot as well. This analysis result can help team owners to recognize top players with more homerun rate and use this as metric for salary. Owners can also ask these top players to teach their techniques to freshers joined their team.

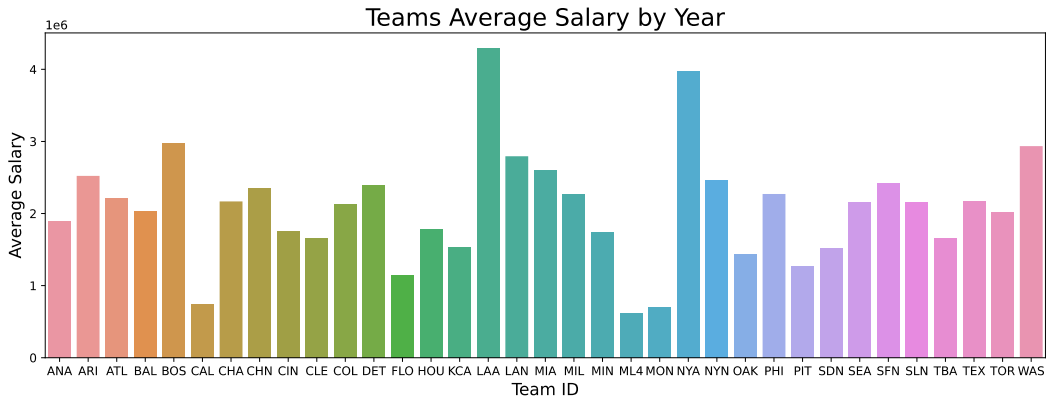
Analysis-4: Any pattern in the average player's salary throughout all available periods. Which team got the highest average salary?

For this analysis, we plotted overall average salary of players throughout the years.



Because baseball players went on a huge strike in 1994, salaries didn't start to rise again until 1995(when the strike ended), as can be seen in the graph above, and there was little change in average player salaries from 1992 to 1994. We can also see that salaries decreased in 2004. This is because it was more difficult to sell baseball sponsorship due to the attendance issue. This analysis can be used to predict the future

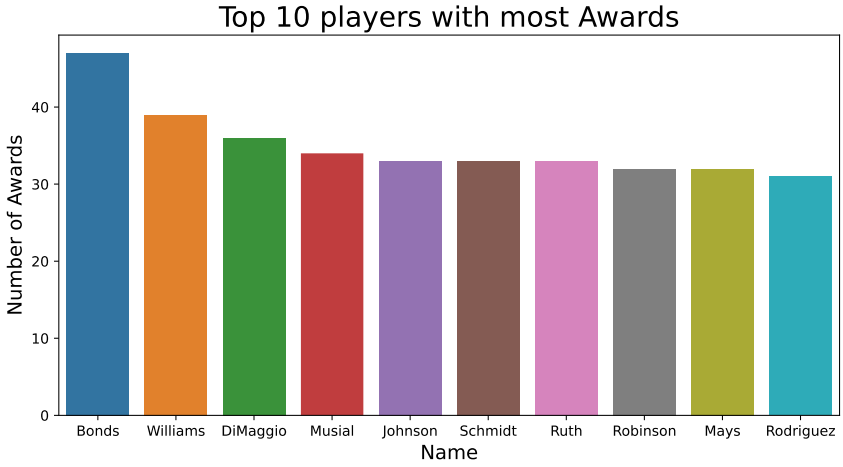
salaries of players or make decisions about player salaries.



The LAA team name has changed a lot over the years. From 1965 to 1996, it was the California Angels, and from 1997 to 2004, it was the Anaheim Angels. Both in 2005 and 2015, it changed again. As a result, it is saved with the different names in dataset. Therefore, the average pay is high overall as it only considers recent years for overall salary. Also, the LAA team has a higher win percentage. This could possibly be the cause of the LAA team's high average pay. The NYA team has won the world series 27 times and is paid well.

Analysis-5: Topo 10 players who got the highest number of awards?

We counted the number of awards each player had received to determine the top 10 winners.



Bonds received the most awards. This could be the result of the fact that Barry Bonds presently owns the record for the most home runs in a single season with 73 in 2001 and is #1 on the list of all-time MLB lifetime home runs with 762.

Technical:

Initially, we did data exploration and preprocessing, like checking columns present in the dataset and choosing columns that are useful for our analysis in the data frame, looking for NaN values and missing values, datatypes exploration etc. We used the pandas package in Python to load all the necessary files (listed in the dataset section). Afterward, we used the group by function to group the data for each analysis and various aggregation functions (sum, mean, max, count) as needed. We also combined many tables using the merge method. For instance, the college playing table contains the player's school ID, not the name or location of the institution they attended. To obtain such facts, we combined this table with the school table.

For the analytical portion, we initially considered correlating the data from one table with another, but many of these ideas fell short. For instance, we learned which players had won the most awards. Then we attempted to relate it to the total number of wins and losses experienced by players. There was, however, no connection between these. After plotting some graphs and analyzing them, some conclusions could be drawn. We looked for significant peaks and valleys in the plots and investigated the causes of those.

References:

(1) https://www.sportsbusinessjournal.com/Daily/Issues/2004/04/08/Leagues-Governing-Bodies/MLB-04-Player-Salaries-Decline-Collusion-Or-Economy

(2) https://www.latimes.com/archives/la-xpm-2004-dec-22-sp-bbnotes22-story.html

(3) https://www.espn.com/mlb/topics/_/page/the-steroids-era