Player Performance and Salary Analysis in Major League Baseball

Introduction

The world of professional baseball is rich with data that can provide deep insights into player performance, salaries, and career trajectories. By analyzing trends in key metrics such as home runs, batting averages, and salaries, teams can make more informed decisions about player recruitment, development, and contract management. This project analyzes Major League Baseball (MLB) data to better understand player performance over time, salary trends in relation to age, and player debut patterns.

The purpose of this analysis is to identify critical insights into the career dynamics of professional baseball players, focusing on when players typically reach their peak performance and how salaries evolve over time. The findings of this analysis can inform team managers and player agents about optimal contract negotiations and player management strategies, ensuring that teams maximize the value they get from their athletes while considering their peak and post-peak years.

Dataset

The dataset used for this analysis comes from historical MLB player statistics, including batting performance (e.g., home runs, batting average), salary data, and player biographical details (e.g., birth year and debut age). Key attributes of the dataset include player ID, year of performance, team information, and career milestones. This dataset allows for a comprehensive analysis of trends in player performance and compensation across different stages of a player’s career, making it well-suited for examining how players' performance metrics and salaries evolve over time.

Analysis Technique

The primary analysis techniques employed in this project include scatter plots, line charts, and density plots to visualize relationships between variables such as player age, home runs, and salaries. Scatter plots were used to examine the correlation between age and salary, while line charts displayed trends in player performance over time. Additionally, density plots were utilized to smooth out histograms and highlight the distribution of players' debut ages. These techniques are suitable for this dataset because they allow us to uncover hidden patterns, relationships, and trends that may not be immediately visible in raw data.

Results

- **Player Performance Over Time: Home Runs and Batting Average**

One of the key observations from this analysis was the trend in total home runs and batting averages over time (Figure 1). The data showed that while the total number of home runs has increased steadily over the years, particularly in recent decades, the batting average has remained relatively stable. This suggests that while players are focusing more on power hitting, their overall ability to hit for average has not dramatically changed. This trend highlights the evolving strategies in baseball, with more emphasis placed on hitting home runs rather than achieving consistent batting averages.

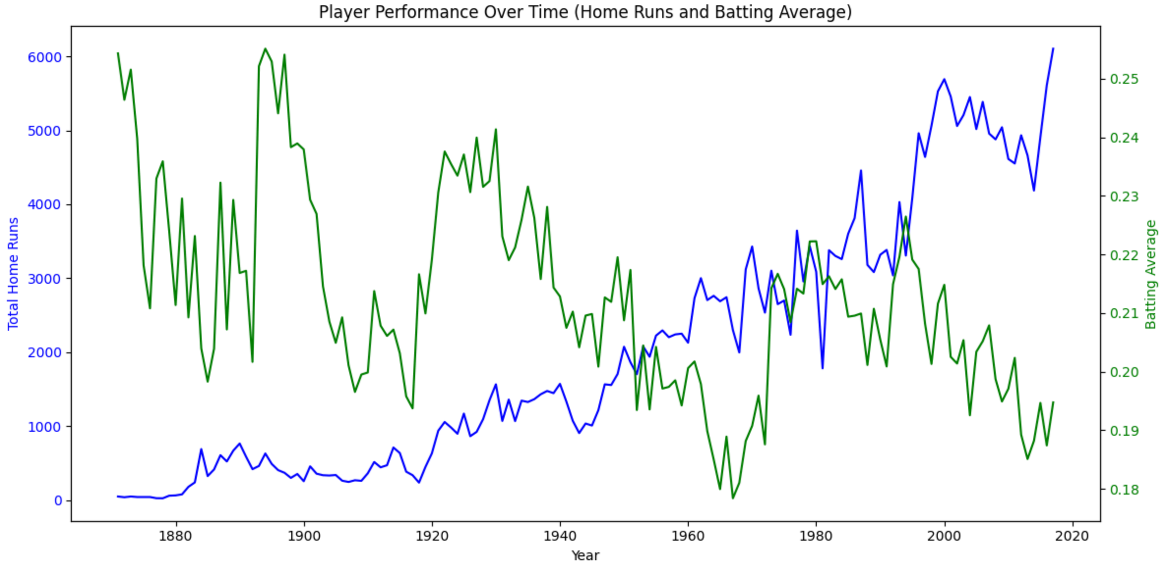


Figure 1: Player performance over time.

- **Salary vs Age Correlation**

A scatter plot was used to examine the relationship between player age and salary (Figure 2). The correlation coefficient of 0.33 indicates a moderate positive relationship, showing that salaries tend to increase with age, particularly peaking between ages 30 and 35. After this age range, both salaries and player performance tend to decline. This insight is valuable for team managers who need to consider a player's age and likely future performance when offering long-term contracts.

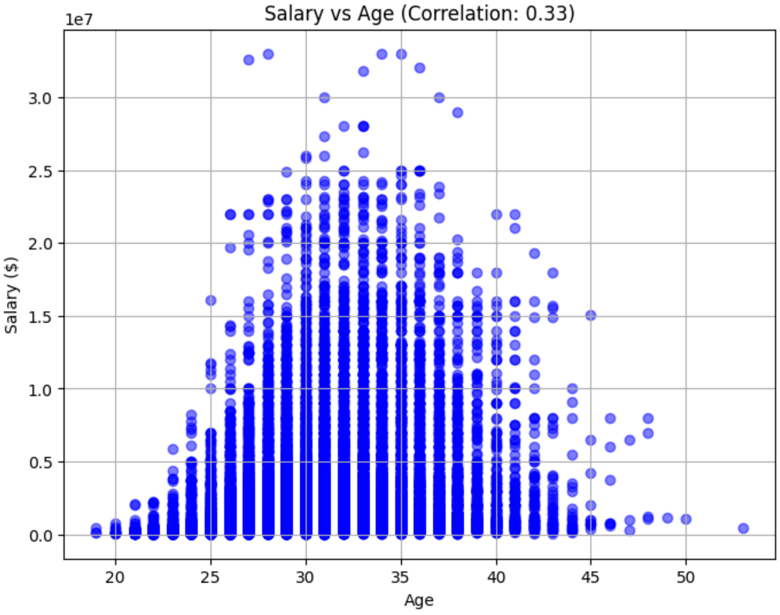
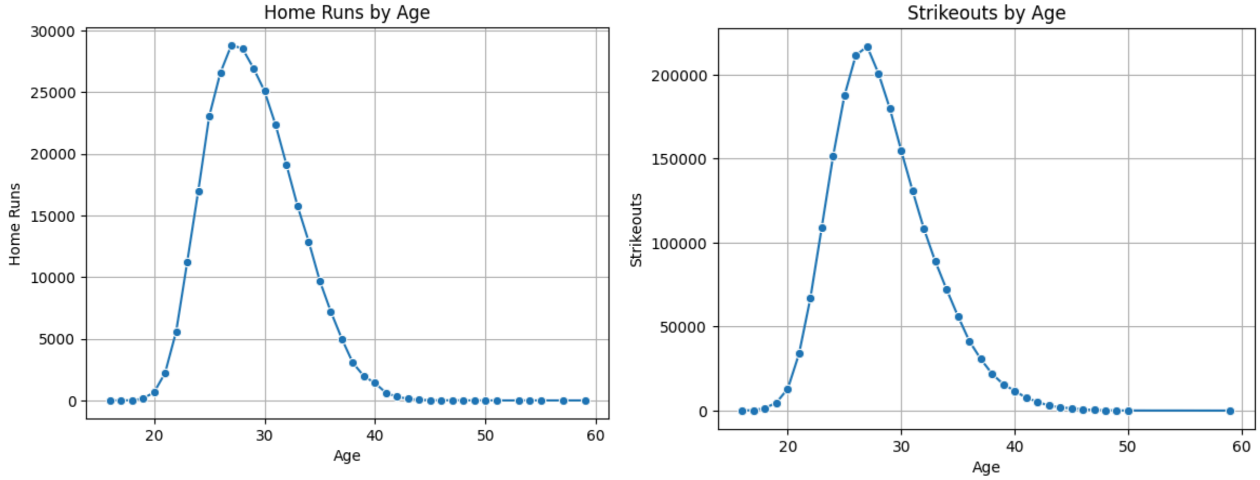


Figure 2: Correlation of salary and age

- **Home Runs and Strikeouts by Age**

Home run and strikeout data were analyzed to understand performance trends by age (Figure 3). As expected, home runs and strikeouts peak in the late 20s and early 30s. After this peak, both metrics show a sharp decline, with players becoming less effective as they age. This pattern aligns with the general career trajectory of athletes, where physical ability diminishes over time.



(a) (b)

Figure 3: (a) Home Runs by Age (b) Strikeouts by Age

- **Distribution of Age at Debut**

A distribution plot of the ages at which players made their debut revealed that most players enter Major League Baseball between the ages of 24 and 26 (Figure 4). The mean debut age is approximately 28 years, with a standard deviation of 4.15 years. After the age of 30, the number of debuting players declines significantly. This information can be critical for teams scouting for young talent, indicating when players typically transition from development leagues to the major leagues.

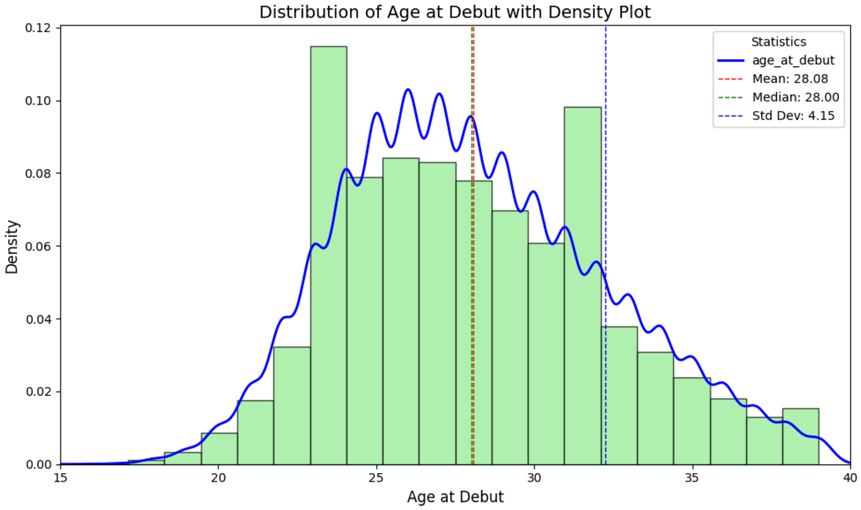


Figure 4: Distribution of Age at Debut with Density Plot

Conclusion

The analyses provide valuable insights into how player performance and salaries evolve over time. Home runs have seen a significant rise, indicating a shift toward power-hitting in modern baseball, while batting averages remain stable. Salaries tend to peak between ages 30 and 35, aligning with the players' peak performance years. Additionally, the distribution of debut ages suggests that most players enter the major leagues in their mid-20s, with a sharp decline in debuts after age 30. These findings can help teams make informed decisions about player recruitment, contract negotiations, and career management strategies.

Technical Section

The datasets required careful preprocessing before analysis. Player IDs were merged across different data files to link performance metrics with biographical and salary information. Invalid values (e.g., missing birth years) were filtered out to ensure accurate calculations of player ages. Additionally, salary values were adjusted to ensure consistency across different time periods. Scatter plots and line charts were used to visualize correlations between player age, performance metrics, and salary. These techniques were appropriate for revealing trends over time, especially given the wide range of player ages and performance levels. The use of density plots allowed for clearer interpretation of distribution patterns, smoothing out irregularities in the raw histogram data.

Initially, exploratory data analysis was conducted to identify key variables and ensure data accuracy. One challenge was managing the wide range of player ages and ensuring that performance metrics aligned with the players' active career years. Another issue was the skew in salary data, with some players earning much higher salaries than others, which required careful handling to avoid misinterpretation. Future work could include more advanced modeling techniques to predict player salaries based on multiple performance variables.

Presentation Slide:
https://docs.google.com/presentation/d/1oq_Yc_TLUX0DxaUZoPPmW2sRtYFRNd5UqnJT6Eb2Tbo/edit#slide=id.g2fd887be03e_0_484

Project Github: https://github.com/Razin1996/CS5830_project1