Project 1 - CS 5830/6830

Jonah Harmon
Sajan Neupane

INTRODUCTION:

This project focuses on analyzing historical baseball data to extract insights about player performance, team outcomes, and salary trends using the Lahman Baseball Database. The main purpose of this analysis is to provide information to help team managers see past trends in home runs, attendance, salary increases, winning percentages, and batting averages. Using historical data, we can analyze these trends to gain insights into the evolution of the game across different eras. It informs managers how changes in game play, rules, and other factors have impacted the game over time. The analysis can help teams optimize strategies for fan engagement and player selection.

DATASET:

We used the Lahman Baseball Database (2014 version) as our dataset for analysis. The dataset serves as a valuable resource for analyzing baseball performance and financial trends, offering statistics of Major league Baseball (MLB) from 1871 to 2014. This broad range allows for a detailed examination of how the game of baseball has evolved over time. The Batting table captures detailed individual player statistics, including home runs and batting averages, which are important for understanding player performance trends over time. The Teams table provides insights into team performance and attendance, allowing for the analysis of how team dynamics and fan engagement have evolved. The BattingPost table focuses on postseason performance, highlighting the impact of high-pressure games on player statistics. The Salaries table tracks player compensation, which, when combined with inflation data from WordData.info, enables the adjustment of historical salaries for comparison with present values [1].

ANALYSIS TECHNIQUE:

The analysis used line plots, scatter plots, and a combination bar and line plot to explore various aspects of baseball data. Line plots were suitable for illustrating yearly trends in home runs and average attendance per game, as they effectively highlight changes over time and key historical shifts. Scatter plots were employed to analyze the relationship between team winning percentages and fan attendance, allowing for a visual assessment of how performance influences fan engagement and how these have evolved over different baseball eras. The classification of baseball into distinct eras was important for comparing win rates to attendance over a similar time period [2]. To compare batting averages between regular and postseason games, line plots provided a straightforward way to visualize performance variations under different conditions. For salary analysis, combining bar plots with line plots enabled a comprehensive view of actual versus inflation-adjusted salaries, offering insights into the evolution of player salary in real terms.

RESULTS:

Home Run Trends:

While looking into home run trends we noticed a few distinct sections of the chart describing increases and decreases in home runs (Figure 1). This led us to research the possible causes in the upswings and downswings. We came across specific eras that can account for the changes.

1. Dead-Ball Era (1901-1919): Represented by the red dashed lines, this period was characterized by two key rule changes that greatly benefited the pitchers. It increased the size of the strike zone and counted foul balls as strikes. These changes led to an increase in strikeouts and a decline in batting averages, home runs, and scoring.

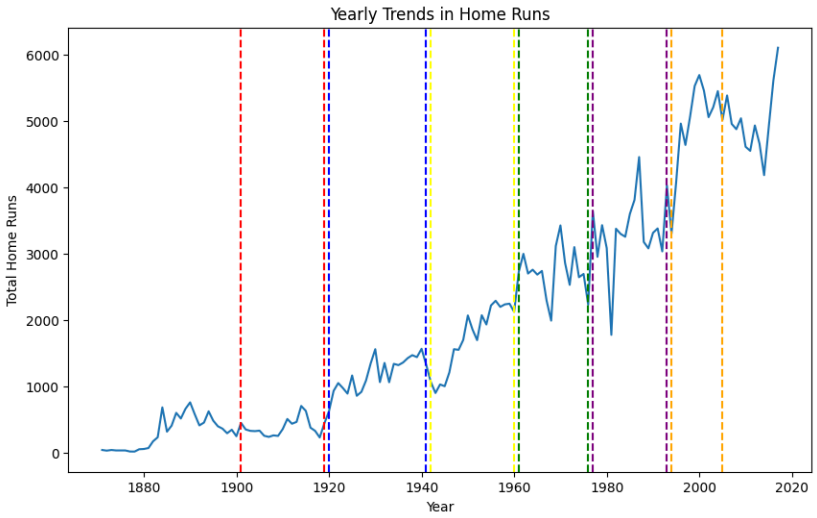

Figure 1: Home Runs Trend over Time

2. The Live Ball Era (1920-1941): Represented by the blue dashed lines, this period saw changes in rules that favored the hitters. It banned trick pitches and introduced cleaner and more visible baseballs. The construction of the baseball itself changed. It was altered to make the ball more lively. In 1920 Babe Ruth hit 54 home runs, far surpassing previous totals.
3. The Integration Era (1942-1960): Represented by the yellow dashed lines, this period is directly after WWII, which depleted the talent pool of the MLB. The integration of Black players furthered the game and increased the talent pool significantly.
4. The Expansion Era (1961-1976): Represented by the green dashed lines, this period brought more teams into the MLB and reduced the height of the pitchers mound.
5. The Free Agency Era (1977-1993): Represented by the purple dashed lines, this era saw the creation of Free Agency, this allowed players much more mobility and a large rise in salary
6. The Steroid Era (1994-2005) Represented by the orange dashed lines, this era saw possibly the fastest increase in home runs. Steroid use increased dramatically, this led to record-breaking offensive numbers.

Attendance over time:

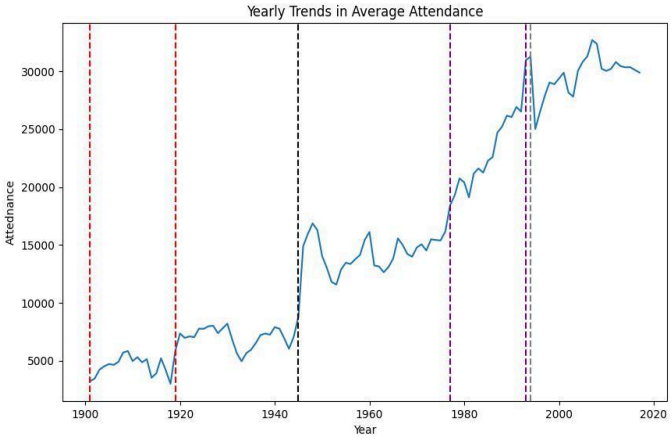

Figure 2: Average Attendance per Game over time

Figure 2 illustrates the overall steady increase in MLB attendance over the years, with various events causing significant shifts. In the Dead Ball Era (1901-1919), marked by the red dashed lines in the figure, attendance saw minimal growth due to the low-scoring nature of the game. Slight growth followed in the Post-Dead Ball Era, but a significant surge occurred after World War II in 1945, indicated by the black line, likely due to the return of military forces. The Free Agency Era (1977-1993), highlighted by purple dashed lines, saw a rapid rise in attendance, driven by the introduction of free agency and increased player movement, which boosted competition and fan interest.. However, the baseball strike in 1994, marked by the gray line, caused a sharp decline in attendance, demonstrating the impact of labor disputes on fan turnout. However, the attendance rebounded possibly due to the homerun surge in the late 1990s (Figure 1). These fluctuations illustrate how historical events and changes in the game's structure influenced MLB attendance trends over time.

Winning Percentage vs. Attendance:

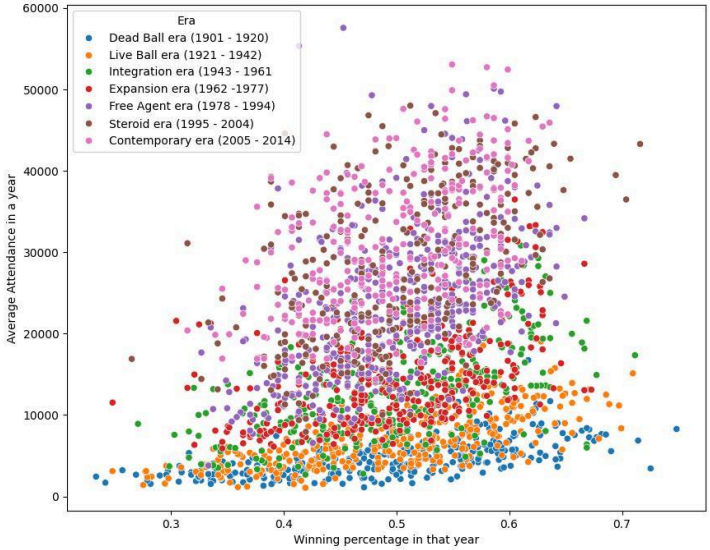

Figure 3: Average Attendance versus WInning rate

The scatter plot graph reveals a consistent positive correlation between winning percentage and game attendance across different baseball eras (Figure 3). In general, when teams perform well (with higher winning percentages), more fans attend games. However, within each era, there's significant variation—factors beyond winning, such as marketing, player popularity, and external events, also influence attendance. From the Dead Ball Era to the Contemporary Era, baseball's popularity has evolved, but the fundamental relationship between winning and attendance remains a central theme in the sport's history.

Regular vs. Post Season batting averages:

This graph compares the regular season and postseason batting averages from the late 1800s to 2020 (Figure 4). It shows a consistent trend where the regular season batting averages tend to be higher than those of the postseason. There are many reasons that this could be so. One being that postseason games have much higher stakes and batters face the better pitchers the MLB has to offer. A team manager that struggles in the postseason may look at the graph and decide to start signing batters that are more consistent in high-pressure situations like the playoffs.

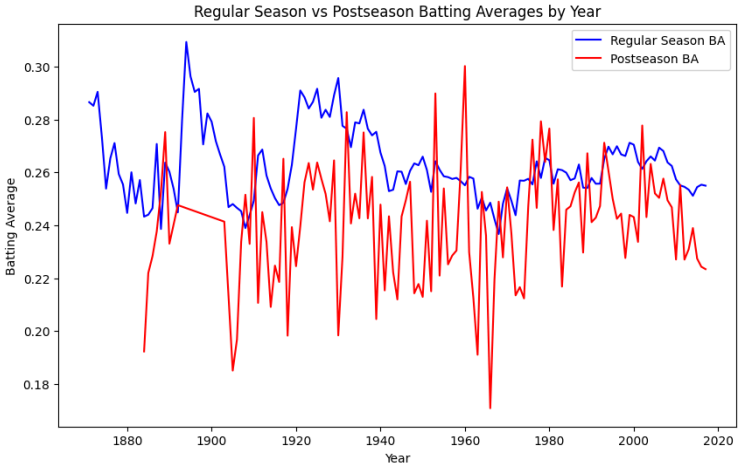

Figure 4: Regular versus Postseason Batting Averages

Salary Over Time:

Figure 5 illustrates the dramatic increase in the average MLB player salary from 1985 to 2014, which has risen from $400,000 to $4.5 million—nearly a 10-fold increase. This substantial growth is particularly pronounced during the late Steroid Era, which likely contributed to higher salaries for power hitters and a surge in fan attendance (Figure 2). When adjusting for inflation, salaries have still increased by 4.5 times from 1985 to 2014, demonstrating that the growth in player compensation has outpaced inflation. This trend reflects both the evolving economics of the sport and the increasing value of MLB as a lucrative franchise.

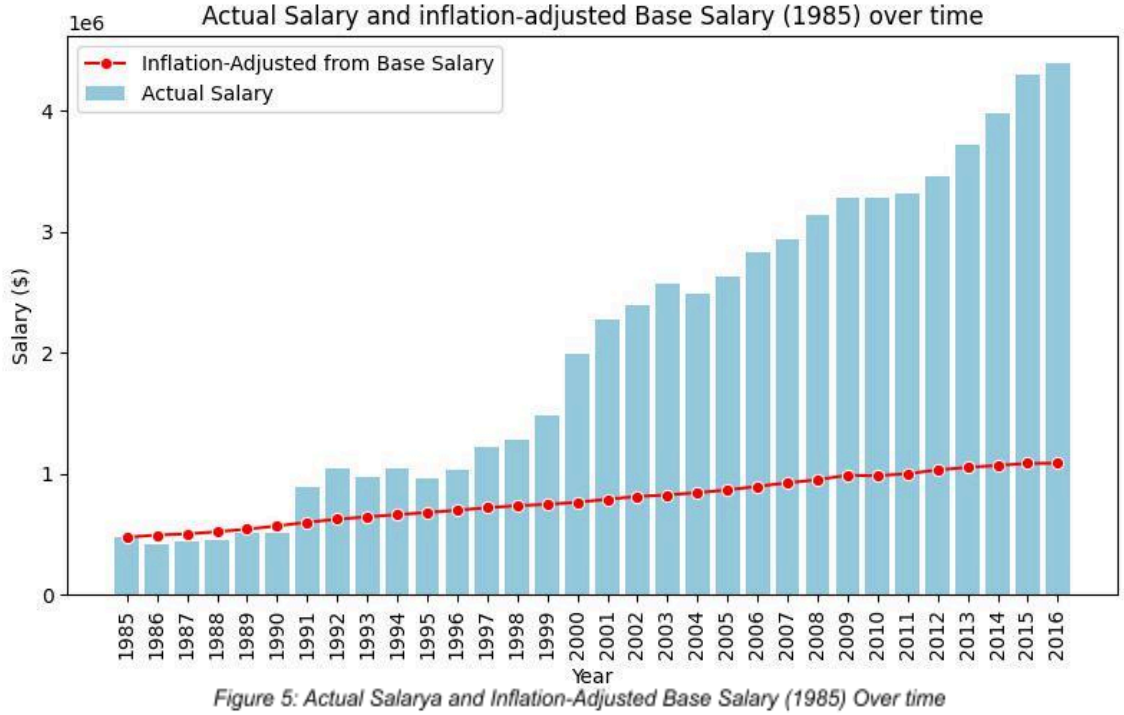

Figure 5: Actual Salarya and Inflation-Adjusted Base Salary (1985) Over time

TECHNICAL:

The dataset preparation involved data wrangling to ensure accuracy and relevance for analysis. Initially, we cleaned the datasets by removing empty cells and handling missing values, which was important for maintaining the integrity of the results. We computed necessary averages and merged data from various sources, such as the Salaries and inflation datasets, to adjust historical salaries for inflation. This process required calculating inflation-adjusted salaries to compare changes in player compensation over time. Specifically, we merged the salary data with inflation rates and then applied a cumulative inflation adjustment to accurately reflect salary trends in constant USD.

For the analysis, we employed time-series analysis, scatter plots, and a combination bar and line plot. Time-series analysis was useful for examining yearly trends in home runs and attendance, enabling us to identify significant shifts and patterns across different eras. Scatter plots allowed us to visualize the relationship between winning percentages and fan attendance, highlighting how team performance impacts fan engagement. The combination of bar and line plots was instrumental in comparing actual salaries with inflation-adjusted salaries, offering insights into how player compensation has evolved in real terms.

The analysis process was iterative, starting with initial hypotheses and exploring various data dimensions. For example, we initially considered analyzing batter hit-by-pitch rates in regular versus postseason games but decided to focus on more impactful metrics based on their practical relevance. This iterative approach ensured that our final analysis delivered actionable insights for team managers and stakeholders, emphasizing trends and patterns that are most relevant to decision-making.

REFERENCES:

1. Inflation rates in the United States of America
2. Examining Perceptions of Baseball's Eras: A Statistical Comparison – The Sport Journal