Project 1 – Baseball

1. Introduction

In this analysis, we delve into the statistics at the heart of the fascinating sport of baseball, aiming to uncover the key links between player performance and game outcomes. Our dataset includes a wide range of baseball statistics from the early 19th century to the modern day, covering Pitching, Batting, and player Appearances. Through this detailed data, we can understand not only the individual performance of players, but also the trend of the entire team and the league.

Our analytics technology combines advanced data science methods with in-depth baseball knowledge to provide insights for team management, coaches, and baseball fans. By digging deeper into this detailed data, we hope to reveal the key factors that affect team wins and losses and player performance, thereby helping teams optimize their strategies and improve their competitiveness.

For example, by analyzing batting data, we can identify which batting skills and habits are associated with high scoring, and then provide coaches with specific training recommendations. Similarly, by studying pitching data, we can help coaches understand how to set up pitchers more effectively and how to develop tactics to target opponents' weaknesses.

In the following sections of the report, we will describe in detail the process of our analysis and the specific conclusions we draw. In addition, we have prepared a series of charts and visualizations to present our findings more visually. To facilitate further discussion, we have also included links to our presentation slides and project folders.

2. Dataset

The data set used in this project contains in-depth baseball game statistics, covering the three key aspects of Pitching, Batting and player Appearances. These data sets are extremely valuable because they provide a historical record from the early 19th century to the modern day, reflecting the evolution of the game of baseball and its characteristics at different times.

The pitching data set reveals a player's pitching performance, including key metrics such as wins and losses, plate appearances, and innings pitched, which are important factors in evaluating a pitcher's performance and efficiency. Batting data sets record a player's batting performance, such as at-bats, hits, runs scored, home runs, etc. These data sets are crucial to understanding a player's offensive ability and contribution to the outcome of a game. The player appearance dataset provides information on the appearance of players in different positions, which helps to analyze the diversity and adaptability of players. The breadth and depth of these datasets make them ideal for comprehensive analysis to uncover the complex relationships between player performance, team strategy, and game outcomes.

3. Analysis technique

In this baseball data analysis, we employed a combination of data aggregation and grouping techniques, as well as data visualization using Matplotlib and Seaborn. This amalgamation of methods provided us with a

robust analytical framework, capable of handling and interpreting voluminous historical data, and presenting the analysis results in an intuitive and engaging manner.

The data aggregation and grouping technique, mainly implemented through the pandas library, enabled us to efficiently process the large datasets. This approach allowed us to easily group and aggregate data by year, player, or other key metrics, which is crucial for understanding long-term trends and patterns.

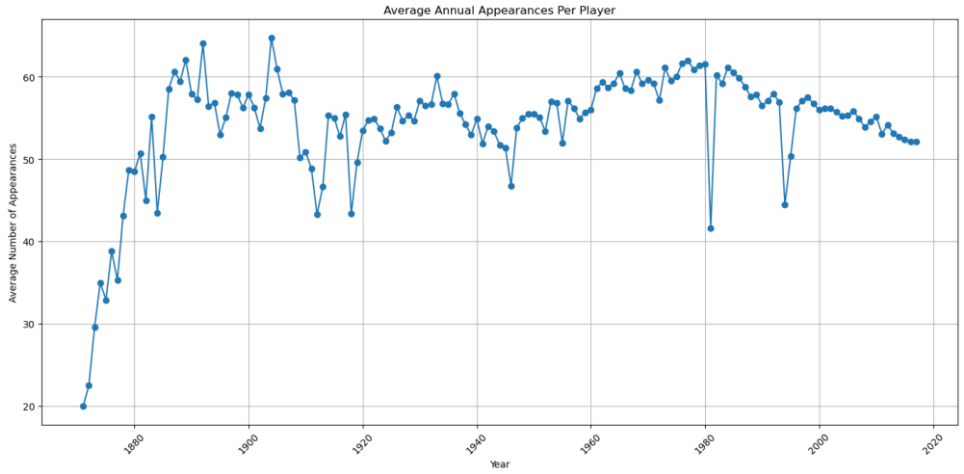
For visualization, we used Matplotlib to create basic charts such as line graphs and bar charts to showcase players' appearances and performance trends. Additionally, the use of the Seaborn library enhanced our visualization capabilities, allowing us to present data in a more refined and aesthetically pleasing manner, such as through heatmaps or complex multivariate plots. These visualizations not only aided in a deeper analysis of the data but also made the analytical results more accessible to a non-professional audience.

In summary, this combined approach of data aggregation, grouping, and advanced visualization is highly suitable for our dataset and the domain of baseball. Through this methodology, we were able to conduct a thorough analysis of baseball statistics and present it in a clear and intuitive manner to a broad audience.

4. Results

For ‘Apearances.csv’, initially, the CSV file containing records of each player's game appearances is read. Then, by grouping and aggregating the data, the total number of appearances per player per year is calculated. These figures are grouped again to calculate the average number of appearances per year for all players, reflecting the average level of participation in games over the years. Finally, by setting the appropriate fonts and chart size, a clear line graph is plotted using the Matplotlib library, which vividly shows the trend of the average number of player appearances over time. These trends may relate to the developmental history of the sport of baseball, socio-economic conditions, major events such as wars, and modern team management strategies. In this way, stakeholders without a background in data science can easily understand the story behind the player appearance data.

The chart displays the average annual appearances per player. Based on this chart, we can perform the following analysis:



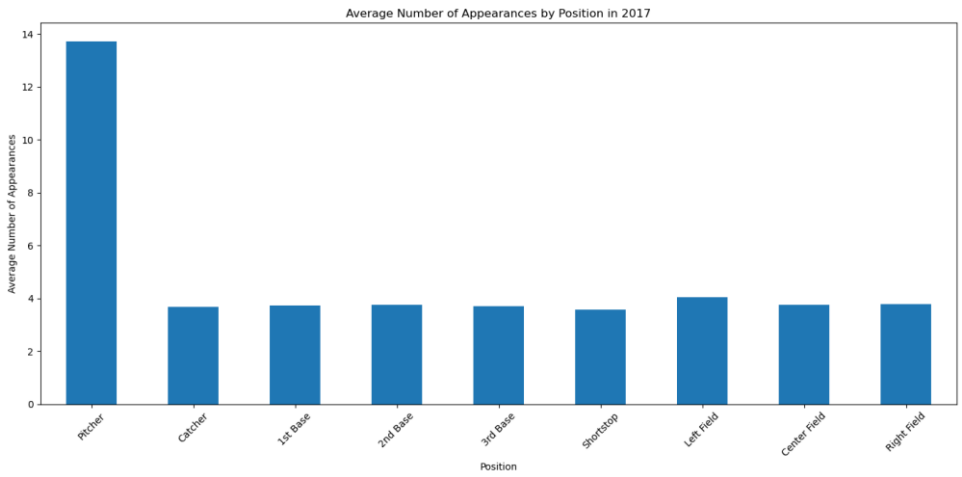
- In the early records, we can observe a rapid increase in the average number of appearances. This could be attributed to the early development of baseball and the increasing level of professionalism in the sport.

- From the early to mid-20th century, the number of appearances remained relatively stable but experienced some fluctuations. These fluctuations might be related to the socio-economic conditions of the time, player mobilization during World Wars, or other significant events.

- In the latter half of the 20th century, there was a noticeable decline in the number of appearances. This could be due to changes in team strategies, the implementation of player rotation policies, or adjustments in the structure of baseball seasons.

- In recent years, we can see a decrease in the number of appearances, which may reflect the modern emphasis on player health and rest, as well as teams giving more importance to the long-term performance and career planning of players.

Then, we analyzed the average number of games played by baseball players at different Fielding positions and focused our analysis on the most recent year. Such analysis can help reveal how often each Fielding position is used, and how the use of players in different positions may change over time.
The generated bar chart represents the average number of appearances for players in different baseball positions in the past year. Based on this chart, we can observe the following analysis:



- The average number of appearances for pitchers is significantly higher than other positions. This is likely due to the prevalence of pitcher rotation systems, where multiple pitchers typically appear in each game.

- Catchers have relatively lower numbers of appearances. This might be because the catcher position demands high physical and mental endurance, leading to fewer consecutive appearances.

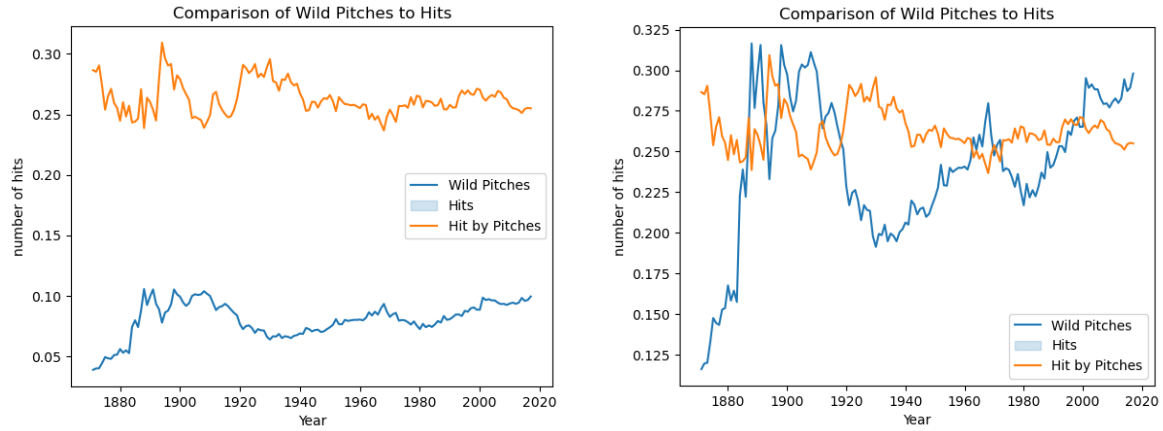
- First base, second base, third base, and shortstop have relatively balanced numbers of

appearances. This could reflect less rotation of players in the infield defensive positions.

- Outfielders (left field, center field, right field) also have relatively balanced numbers of appearances, suggesting a more consistent rotation strategy for outfield positions compared to the infield.

This type of analysis helps in understanding the differences in player usage strategies for different positions and can provide insights into overall team tactics and player health management. Tactical arrangements and player fitness management are important factors to consider when developing these rotation strategies.

Next, we analysis 'Batting.csv' and 'Pitching.csv'. The first one is an analysis of if the introduction of the sidearm throw made batters worse at hitting the ball.



These two charts show the relationship between Wild Pitches, Hits, and hits by Pitches in middle-field games in different years. As we can see from the graph, wild pitches and batter hits are seemingly related. With an increase in wild pitches (sidearm throws, sliders, and other odd pitching techniques that result in an uncaught ball by the baseman) the number of hits decreases. The opposite is true when the number of wild pitches decreases, batting averages increase in that case.

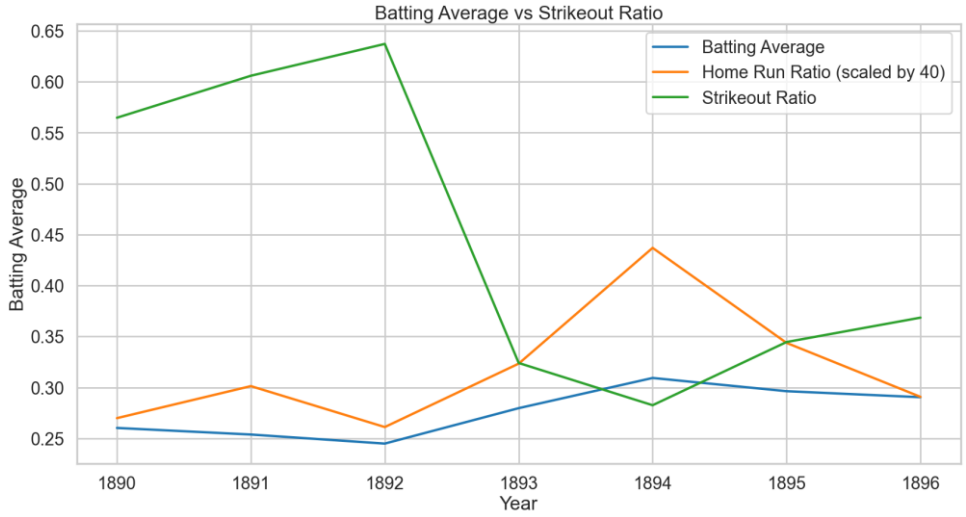
For players and coaches, this information can help them understand how different technical and tactical changes have affected game outcomes throughout history. For example, if the number of errant throws and touch balls increases significantly during certain periods, it may indicate that players and coaches need to improve their ball-handling skills or adjust their pitching strategy. Similarly, changes in hitting rates can help coaches and players evaluate the effectiveness of hitting techniques and training methods.

For the practical guiding significance of the game, understanding the long-term trends of these statistics can help teams make longer-term strategic planning. For example, if a defensive strategy in a given decade leads to an increase in error-throwing, a team may need to adjust its pitching staff or change the focus of its training. Overall, by analyzing this historical data, teams can better understand past successes and failures and how to apply those lessons to current and future game strategies.

Then we see if moving the pitching mound back helped or harmed pitcher and batter performance in 1893.

Pitcher mound was moved back from 50 feet to 60 feet 6 inches in 1893.(Study on data from 1890-1895).

Our aim is to focus on statistical analysis of baseball batting and pitching over a specific time period (1890-1896) and graphically show the relationship between batting average and strikeout rate.



This graph depicts three key statistical indicators in baseball from 1890 to 1896: batting average, home run ratio, and strikeout ratio. The batting average refers to the ratio of hits and getting on base to the number of at-bats, the home run ratio is home runs per at bats, and the strikeout ratio refers to the number of times a pitcher strikes out a batter relative to the number of runs allowed by the pitcher. All indicators are important for evaluating the performance of baseball players and recognizing if the change was substancial enough.

This analysis is on the effect on batter averages and number of strikeouts when in 1893 they moved the pitcher's mound 10 feet 6 inches backwards. It immediately had an effect as we see a ~60% decrease in strikeout ratio that year, as well as a ~50% increase in batting average within 2 years of the new mound distance.

This is exactly as expected, however as during this time the professional baseball legue was looking for ways to create more offense, as games were becoming lower and lower scoring thus creating less excitement for fans. To balance offense and defense the new pitching distance of 60’6” was enforced. Adding ~20% of distance from the pitcher to the batter from earlier seasons. We see with this change batting averages immediately increase, strikeout percentages decrease, and home run hits also increase. Thus, solving the dilemma of low scoring “boring” games.

In this project, we conducted a comprehensive technical analysis of historical baseball datasets with the aim of understanding the performance and characteristics of players across different years. In terms of data preparation, we meticulously cleaned and formatted datasets from four distinct perspectives. Beyond calculating batting averages (H/AB), home run rates (HR/AB), and strikeout ratios (SO/R), we also processed the average annual appearances per player and the average number of appearances for players in

different baseball positions in the most recent year. To prepare this data, we performed data grouping, summation, and ratio calculations, and we appropriately handled any missing values.

The analytical techniques chosen were closely related to the data and our objectives. We employed descriptive statistical methods and visualization techniques suitable for revealing trends and patterns within time series data. For instance, in analyzing the average annual appearances per player, we used line charts to show the frequency of player participation from the 19th century to the present, which helps to understand the development and professionalization of baseball. As for the number of appearances by players in different positions, we used bar charts to reveal the frequency of use for different positions in a specific year, which might reflect the strategies of team tactics and player allocation.

Links:
Presentation - https://docs.google.com/presentation/d/18i_d1g89CJUr-ZedKFVsHYQn0Zti4P2xco_nHHMJ2Ks/edit?usp=sharing

Github - https://github.com/thatsmellything/project-1-datascienceinpractice