Introduction

This report presents the findings of an analysis conducted on sports data gathered. The purpose of the analysis was to uncover insights and identify patterns in the data that could prove beneficial to various persons in sport, specifically MLB, related fields. Through a combination of statistical analysis, data visualization, and machine learning techniques, we were able to gain a deeper understanding of the underlying trends and relationships within the data.

The report is organized into four sections, each of which covers different aspects of the analysis and its results. In the first section, we provide an overview of the dataset and the methods used in the analysis. In the following sections, we describe the analysis technique used and delve into the specific results of the analysis, including key findings, insights, and what the information can be used for.

Project: https://github.com/Dheeraj0650/project1

Presentation: https://docs.google.com/presentation/d/1lVIXbwW2mxeuTBTGpSiRzYHL0R_R2JlXtI61hLd5fag/edit?usp=sharing

Dataset

The dataset given contains various attributes about the MLB which is a subsection of our domain on sports. From the dataset, specific key attributes were selected which were vital to our analysis. These attributes were playerID, teamID, yearID, games (G), wins (W), 3B, bats (right and left handers), errors, ERA, BGP, and IPouts. Without these attributes, we would not be able to complete our analysis nor would the dataset be complete.

Analysis Technique

A mixture of diagnostic and statistical analysis techniques were used. These techniques were used due to the natural flow of the analysis with an unfamiliar dataset. After the initial analysis of the data, questions arose which required additional research with some of those questions being answered and some leading to more questions. Although not all the questions were answered, these analysis techniques proved to be the best way to understand and find answers.

Results

Through the analysis process, many insights were gained based on some initial questions. Those questions revolved around pitches per pitcher, errors received, manager wins, and triples per hitter. Many insights were gained from each question, those being the average number of earned runs per pitcher, the increased ability for triples from left-handed hitters, the higher amount of triples during the initial years of MLB versus later years, average number of wins per manager, and the amount of errors over the course of MLB history.

These insights can be useful to various officials in MLB to help their teams or players in numerous ways. Some of those ways are determining if the given pitcher is above or below the MLB average, which players to bat and what order, if a given manager is beneficial to their team winning, and the advantage of technology in disputing errors in a game. All these insights could have potential to change the outcome of a game and a player or manager's career.

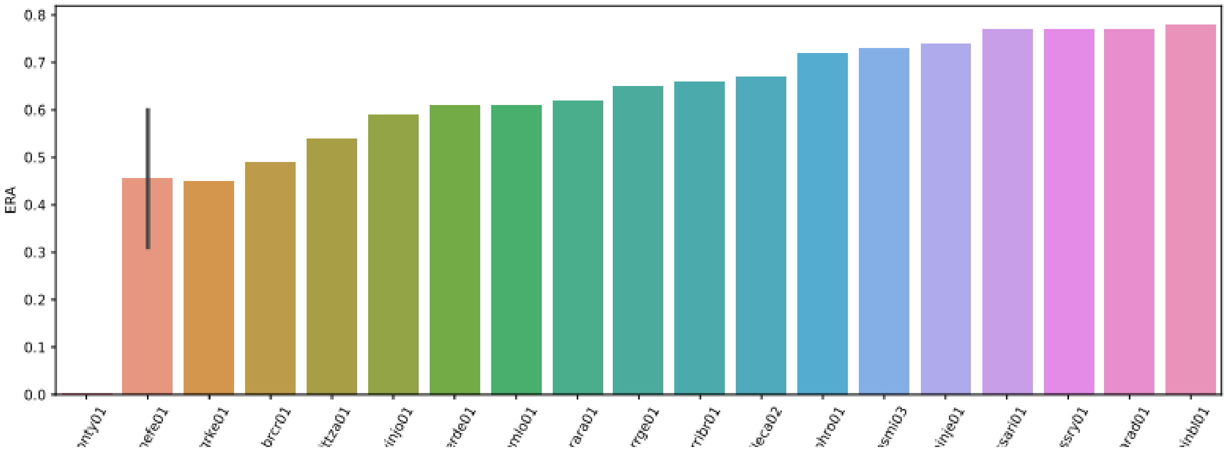
Technical

Much of the data preparation process was merging various tables together to have a bigger picture of our initial questions. From this table merging, we were able to then format the data specific ways to help answer our questions. Some of these formatting included summing relevant information by a specific column, removing inconsistent or irrelevant data, and graphing data which helped answer the given question.

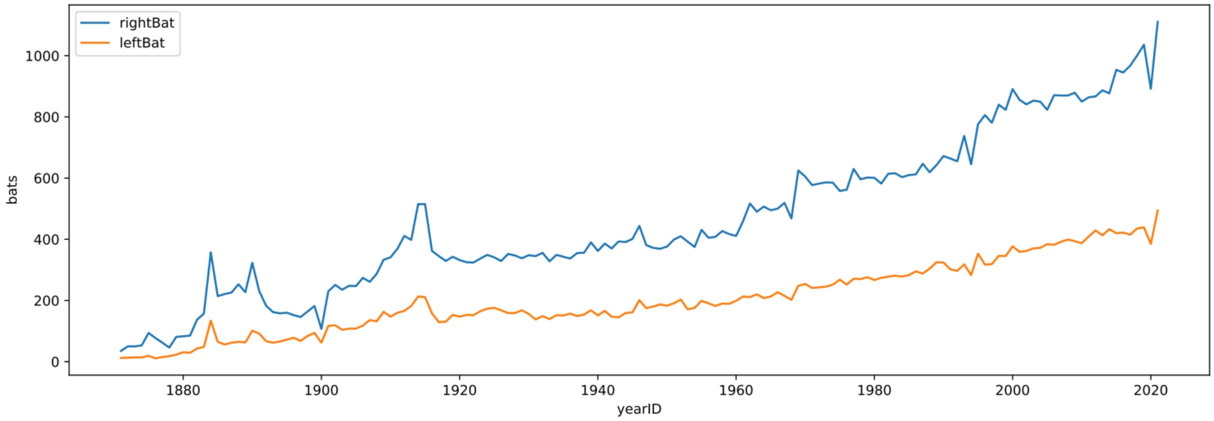
The mixture of diagnostic and statistical analysis techniques was suitable for this dataset since we were finding information which we found interesting rather than looking for something specific. Looking at the data, asking questions and finding answers is the most straight-forward way to analyze data which is the process of diagnostic and statistical analysis. There were no novel techniques used but much about how to analyze a dataset was learned.

Our analysis process began with what type of information we are interested in learning or analyzing. Once those initial questions were decided, some trial and error was experienced determining which tables were relevant to help gain answers to our questions. With adjustments to information, such as filtering out outliers and irrelevant information, answers were found. Some of those answers led to more questions which could not be answered by the dataset.

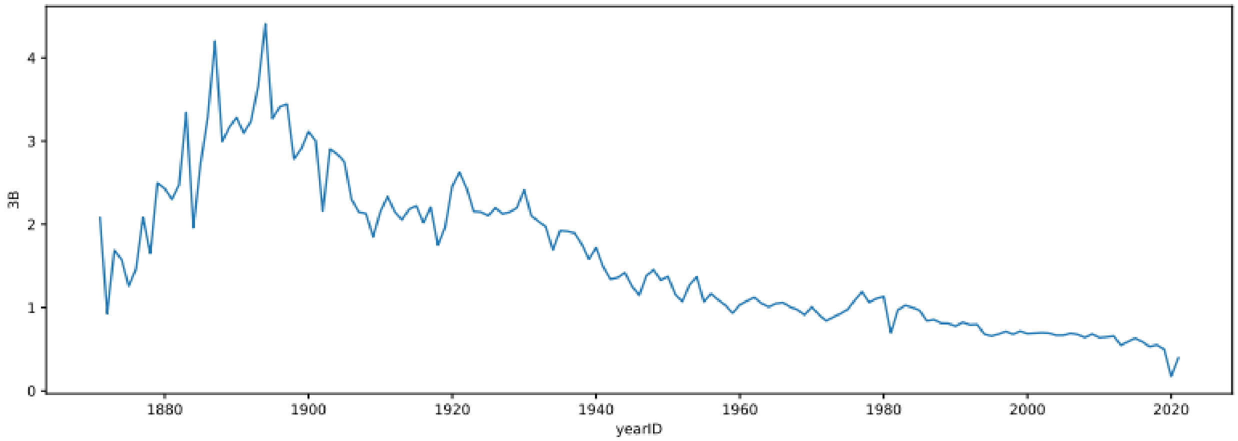
Upon analysis of the pitcher table, pitchers have lower ERA’s after the 1920’s versus before the 1920’s was noticed. After removing outliers, players with less than 1 game and less than 1 IPout, the overall stats changed a little but the overall trend through the years was the same, just more condensed. This analysis can be used to help managers or teams to determine if their pitcher is above or below the average based on the specific pitcher's stats which could change the outcome of a game and the teams chances of winning.



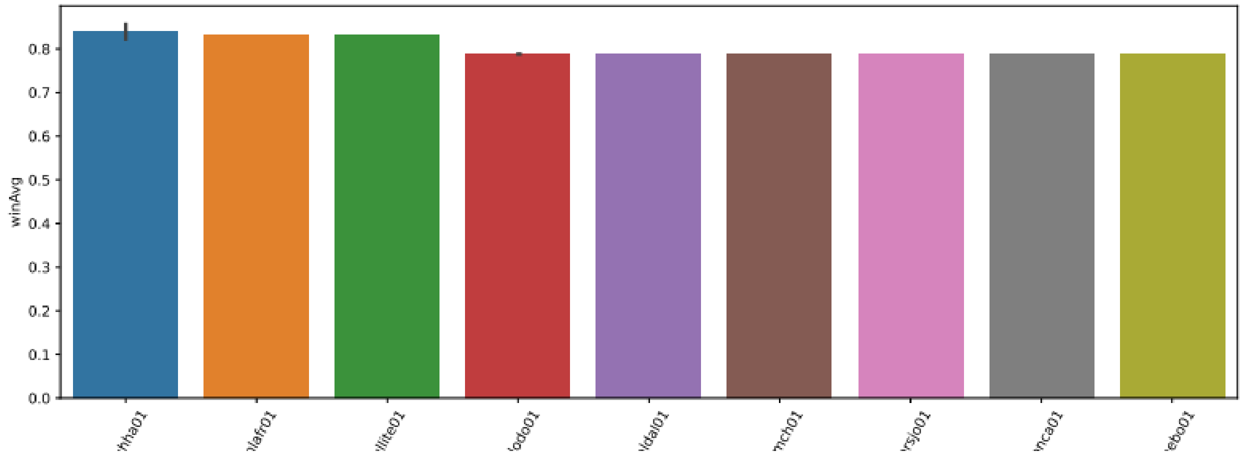
Analysis of the batting table showed the amount of players who hit left-handed are able to get triples was higher than those who bat right-handed. This can be due to left-handed batters' advantage to hit the ball towards right field than right-handed batters. This correlation can be seen in 1900 where the number of right and left handed batters were roughly the same but the number of triples was higher than later years where right-handed batters outnumbered left-handed. This analysis is helpful for team managers since left-handed batters would be more likely to hit triples and be more beneficial to the team winning.



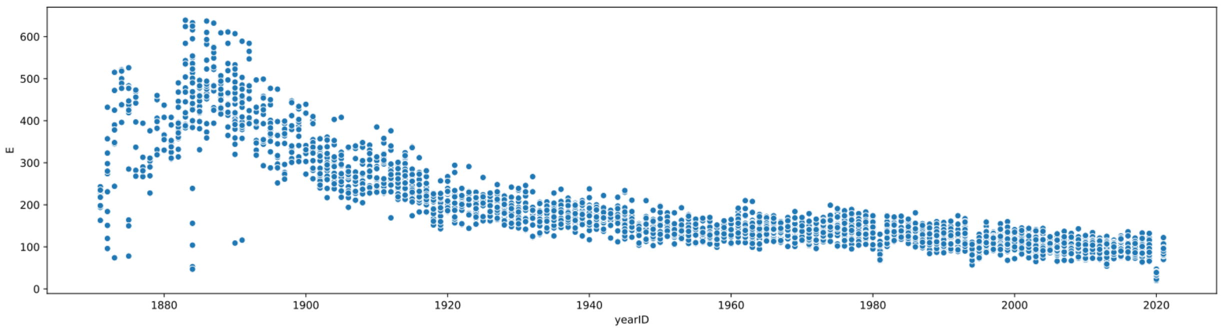
Another analysis of the pitcher table is shown in the early years of the MLB. During this time, players played both offensive and defensive positions. This dual skill would be useful for instances when players were unable to play and required substitutes. Since all the players were able to play defensive and offensive positions, their skills in both were well practiced. In recent years, players generally only play defensive or offensive positions, rarely both. This separation of skills could potentially be bad if a player is unable to play and the substitute is not skilled in batting. This analysis can be beneficial to managers to always have some players train in both offense and defense positions.



When analyzing the manager table, the managers before 1900 had less wins than managers after 1900. After additional research, it was explained that the game length of a season was not institutionalized until 1961. After filtering it to show only the seasons with a consistent number of games, some later years were below the average trend due to factors not explicitly defined in the dataset. This analysis can be used to help officials in MLB determine if a manager is beneficial to a team and if they should keep managing.



While analyzing the errors table, the amount of errors was drastically higher before 1900 than the amount of errors after 1900. This could be from the advancement of technology. With this advancement, errors are thoroughly decided through replaying the error and gaining a definitive decision rather than relying only on the game official. This analysis also shows the impact of technology which potentially helped minimize errors which can drastically change the outcome for a game.



An additional analysis showed that shortstops and second basemen were more likely to receive errors. This could be due to those positions being closer to the batter which would require a faster reaction time than other players. With this information, the reaction time for players in those positions could be practiced to help reduce errors which could help those players continue to play and help their team win.