Project 1 Report

Introduction

One area of particular interest when it comes to statistics is the world of sports. In this report, we will take a close look at Major League Baseball. Because of the variety of MLB statistics tracked, analyses on numerous relationships can be performed. In the analyses mentioned in this report, we used 4 datasets that included biographical information about players (such as country of birth, height, weight, and batting style), team and salary data, awards won by players, and batting statistics. We used subsets of the data to compute additional data, such as batting averages and BMI. Our analysis code can be found on our GitHub repository.

Among the analyses included in this report are the effect of awards won by Puerto Rican players on their salary, the impact of BMI on batting average, average salaries by batting style in the National League and American League since 2012, the total amount of money spent per year in both National League and American League, the highest paid non-US players' countries, and average starting salaries of players from each country represented in the MLB. Through these analyses, we hope to answer questions important to current MLB players, teams, investors, and prospective players. Players can understand whether they can expect salary increases after highly-awarded seasons, what style of batting is most profitable, and which league it might be more profitable to play in. Teams can be informed on how BMI might affect their players and which players they may need to pay high salaries to keep or acquire. Prospective players can learn what kind of salary they can expect starting out in the MLB. As we processed the data, we found some surprising results.

Dataset

In this analysis, we utilized four datasets—People.csv, Salaries.csv, AwardsPlayers.csv, and Batting.csv—to gain insights into the domain. The People dataset contains player-specific biographical attributes such as birth date, birth country, physical characteristics, and batting preference. Player height and weight, batting preference, and birth country were valuable data for evaluating things like BMI and finding batting-style- and birth-country-specific statistics. Salaries includes data for each player each year that they played. Important information was included, such as the player's team, league, and salary, all of which were used in the analysis. Awards listed the different awards won by a player each year. Batting consists of stats on hits, at bats, and other batting-related information for each player each year. We merged several of these datasets to get cross-topic information for specific players, leagues, and birth countries.

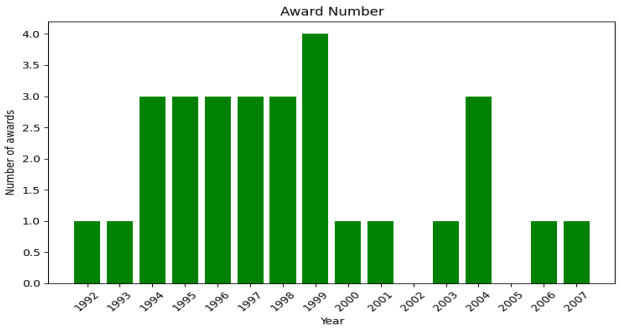
Analysis technique

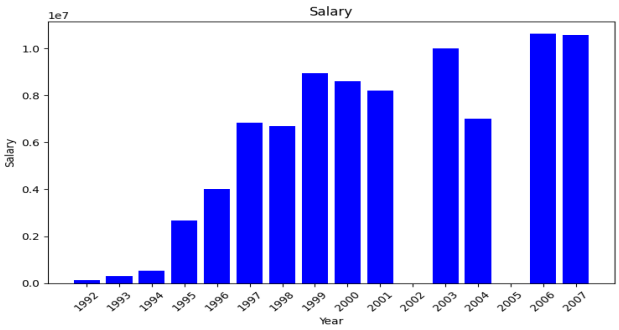
Several techniques were used in each analysis. For the awards analysis, we narrowed down our dataset to Puerto Rican players and then identified the player with the most total awards and focused on his award and salary data throughout the years he played. A look at award types was also done for this player. To ensure valid results on the BMI analysis, we first removed any players missing necessary height, weight, hits, and at bats data. Then, we calculated BMIs for each player and grouped players by BMI, summing hits and at bats of players. Finally, we used the accumulated hit and at bats for each BMI

to calculate each BMI's batting average. Representing it with a linear regression scatter plot helped to identify any trend that might exist between the two. Analyzing pay for each batting style in each league was achieved by sorting players into league/style combinations and computing the average salary for each combination. Before this was done, we limited the dataset to a single season, so that other factors throughout a range of years wouldn't affect the results. Next, players were split into separate leagues, and the total sum of salaries for players in each league was calculated for each year of the data. For the next analysis, we limited the data down to a subset of players who were paid over $10,000,000 sometime in a specified range of recent years. Then, we counted the number of players in the resulting list from each non-US country. In the final analysis, we identified the starting year for each player and narrowed the data to players who started in 2010 or later. The salary from each player's starting year was used to find the average starting salary for players from each country. Each technique was chosen to effectively visualize and interpret the relationships within the specific context of the research questions, providing valuable insights for each analysis.

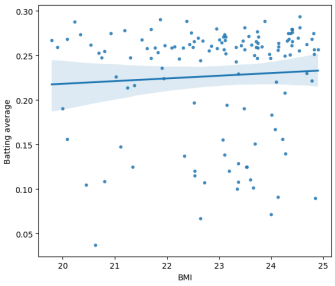
Results

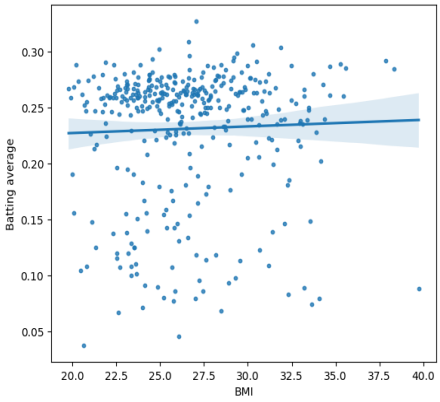
Results from analyzing the most-awarded Puerto Rican's salary were inconclusive. There didn't seem to be a trend between the number of awards won each season and the salary of that season or the following seasons. Because the subject of the analysis was the stats of a single player (Ivan Rodriguez), these results aren't necessarily generalizable. A player can look at this analysis and realize that the number or type of awards won doesn't necessarily mean a player can expect a change in salary one way or the other, but this does not mean that other analyses won't prove that there exists a correlation generally.

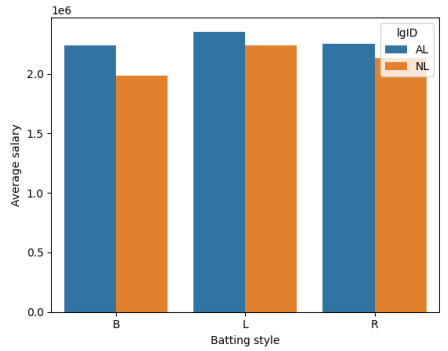


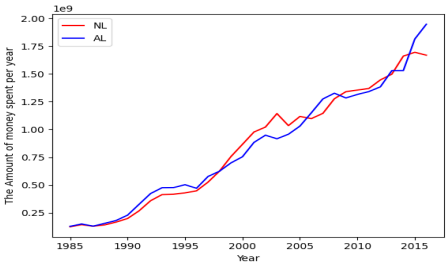


BMI and batting average yielded some interesting results. For example, it can be easily seen that BMIs of players are concentrated inside two BMI ranges: "normal" and "overweight." Either few players fell outside these ranges, or the players that did often had identical BMIs to one another. Likewise, it is apparent that a high concentration of BMIs have a batting average between .25 and .3. The most useful information for teams worried about how differences in BMI affect player batting averages is that there seems to only be a very weak relationship between BMI and batting average. Even in a narrowed range of BMIs, there is only a weak trend. If teams are concerned about how BMI should inform their choice of players or training regiments, they can rest assured that BMI has little to do with a player's batting average.



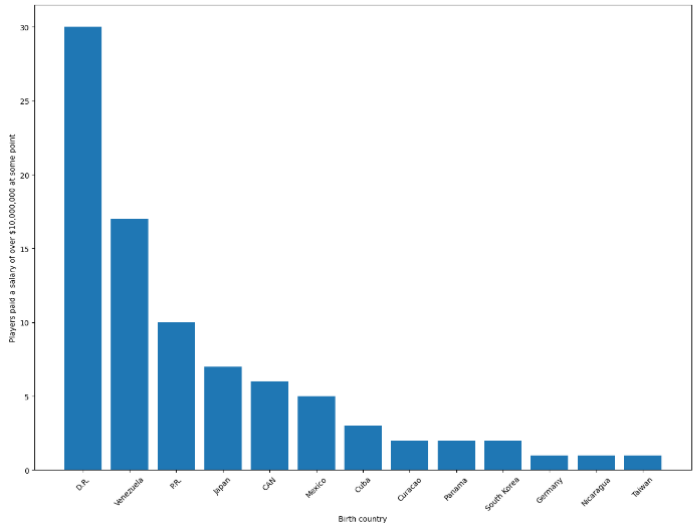


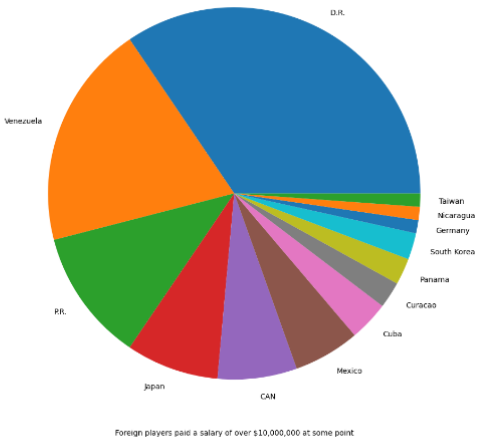


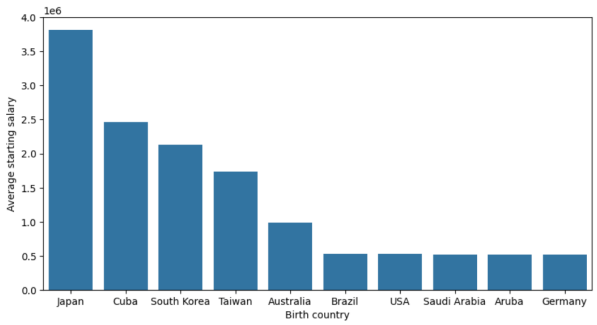


An interesting result of the analysis of the average salary of batting styles is that in both leagues, left-handed batters are paid more on average by several hundred thousand dollars. A year-by-year analysis might help to confirm this result as a trend, but these results from the 2012 season might show that it has at least been the case in recent years. Interestingly, in 2012, batters who batted both left and right in the National League were paid considerably less than others, indicating to players that it might be more profitable to focus on one hand than to try and bat both. A last thing to note from this chart is that players in the National League averaged less money, which might be a deciding factor in which teams players contract with. This decision can also be informed by our analysis of total money spent in each league on salaries over the years, which shows that the leagues have traded off in salary expenditures. In recent years, the American League has taken the lead in salary totals, and its record of few and minor decreases in salary totals might be an indicator of what players can expect in the future.

Our final analyses focused on the effects of players' birth countries on salary. First, we found that of the high-paid foreign-born players, the vast majority were from the Dominican Republic. Venezuela, Puerto Rico, Japan, and others also represented a significant portion of the players in this category. What this tells teams is that a player's nationality might inform what salary may be necessary to acquire or keep the player, especially if they are from the Dominican Republic. This also displays a trend to players from these countries that they may or may not be able to ask for high salaries at some point in their career.







The average starting salary of players was also strongly correlated with their birth country. Results from this analysis seemed surprisingly unrelated to the previous analysis. Players from Japan received the highest average starting salaries at around $3.8 million. Cuban, South Korean, Taiwanese, and Australian players averaged significantly more than most as well. Starting at about $530,000, players from the remaining countries earned salaries much

closer in amount to one another. Interestingly, American players averaged only the seventh-highest average starting salaries. This data can be used to inform players on what kind of starting salary they can expect, depending on their country of origin.

Technical

For the data preparation process, we utilized four key datasets: Batting, People, AwardsPlayers, and Salaries. To streamline our analysis, we selectively extracted relevant features from each dataset. For instance, in the case of the AwardsPlayers dataset, we retained all features except "tie" and "notes". Similarly, in the Batting.csv dataset, we focused on essential columns such as "playerID", "yearID", "lgID", "teamID", "H", and "AB". This careful selection of features was crucial for the subsequent analysis. The merging of these selected columns with relevant features from the People.csv dataset was accomplished seamlessly using the Pandas library. PlayerID served as the primary key for merging, while other features like lgID and yearID played essential roles in facilitating the analysis.

Moving on to the analysis, while our approach may not involve novel techniques, we conducted several insightful analyses. For instance, examining the impact of awards on player salaries revealed a significant correlation, suggesting that player achievements contribute to salary levels. Analyzing the countries with the most expensive players could potentially guide teams in targeting players from regions with promising talent. Further, examining the expenditure per year for both leagues provided valuable insights into team dynamics, league growth, and competitiveness. Additionally, comparing average salaries across different batting styles offered valuable information on player valuation within teams and leagues. Despite attempting to find a correlation between BMI and average batting, the analysis did not reveal a significant relationship. While our analysis may not introduce novel techniques, it contributes valuable insights into player valuation, team dynamics, and the interplay of various factors in the baseball landscape. It should be mentioned that we did not remove NaN values immediately after merging because some analyses used different columns than others so based on each analysis we did the cleaning process. For example, to find the impact of BMI on the batting average we dropped NaN values from H and AB columns. Similar filters were also used to clean data, such as removing at bat totals less than 20 so that only data from smaller sample sizes wouldn’t skew the data, removing rows that lacked data necessary for BMI or batting average computations, and narrowing data down to specific year or year ranges to ensure results represented recent trends or were less affected by outside factors over larger time intervals.