CS5830 Report 1: What Makes a Professional Baseball Player?

Reagan Hoopes and Yagnashree Velanki
Project GitHub Link and Slides Link

January 23, 2023

1 Introduction

Baseball is a celebrity sport and there is a lot of money and fame to be gained for players that can make it to the highest level of the sport. For this reason, parents of young athletes and players want to know what makes for a successful path to the MLB. Analysis of baseball statistics has always been a challenging and fun task for most of the data analysts. In this project, we have used statistical analysis to look at the profile details of baseball players as well as universities and their success at the NCAA baseball tournament in order to answer some of our questions about what characteristics can contribute to a successful path to the MLB for players.

2 Dataset

For our analysis we used Sean Lahman's Baseball Database to gather information on various attributes related to baseball players, including the year they entered the MLB, their birthplace and the college they attended. Additionally, we used NCAA data about which schools won or were runner-up at the NCAA baseball tournament since the first College World Series in 1947. We incorporated the data from the NCAA to provide additional information about NCAA tournament success for the schools present in the Baseball Database.

3 Analysis Technique

The goal of our project is to look at characteristics that could contribute to a successful path to the MLB. To achieve this, we utilized statistical analysis to discover what has happened in the past and try to find the answers to our questions within the data. We felt that this type of analysis is well suited for our data sets because we have decades of data and using statistical analysis we hoped to unearth some of the patterns that exist within the data.

4 Results

Overall, we found statistical analysis to be a valuable tool for examining our baseball datasets and gaining insights into the factors that contribute to becoming a professional baseball player. Our first analysis(Figure 1) looked at the difficulty of breaking into the MLB, and we found that on average, only 134 players debut in the league each year. For our second analysis, we looked at which states produce the most professional baseball players(Figure 2), with the top five being California, Pennsylvania, New York, Illinois, and Ohio. Our third analysis focused on which universities send the most players to the MLB(Figure 3), with Arizona State University, the University of Southern California, Stanford University, the University of Texas, and the University of Arizona leading the way. Finally, our fourth analysis investigated the correlation between a school's NCAA championship and runner-up titles and the number of players they have sent to the MLB, and we found a strong correlation of 0.788(Figure 4). This supports the idea that playing for a top college team can lead to a higher chance of becoming a professional baseball player.

In conclusion, while correlation does not necessarily equal causation, it is important to note that it is difficult to become a professional baseball player, states with rich baseball history tend to produce the most players, top universities for producing MLB players tend to be located in warmer regions, and there is a strong relationship between a school's NCAA performance and the number of players they send to the MLB.

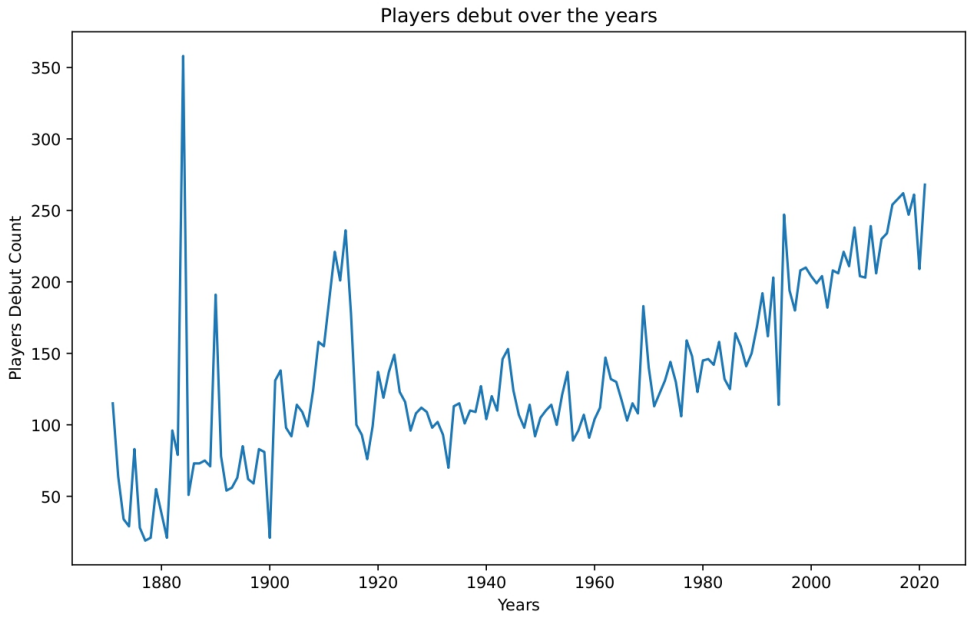

Figure 1: Analyzing the number of players debuting in the MLB annually

5 Technical

In our data preparation phase, we employed pandas to manipulate dataframes and hone in on specific attributes. We did not encounter any major issues as our datasets were complete, but we did have to manually adjust school names in the NCAA dataset to match the Baseball Database. For future projects, we hope to find a more efficient method for this task. We decided to conduct a statistical analysis of the data to uncover any interesting patterns and see if they align with our assumptions. One example being the belief that attending top college programs is necessary to become a professional athlete. We also used a choropleth map to visualize geographical trends. Our analysis consisted of four different parts, line plot to see variations in player debuts, Choropleth map to identify states with the most players, bar chart to find the universities that send the most players to the MLB, and scatter plot to show correlation between NCAA titles and players sent to MLB. This project has opened up many possibilities for future research, such as analyzing the number of players sent to MLB in the years surrounding NCAA titles.

Figure 2: Determining Which States are the Birthplace to the Most Professional Baseball Players

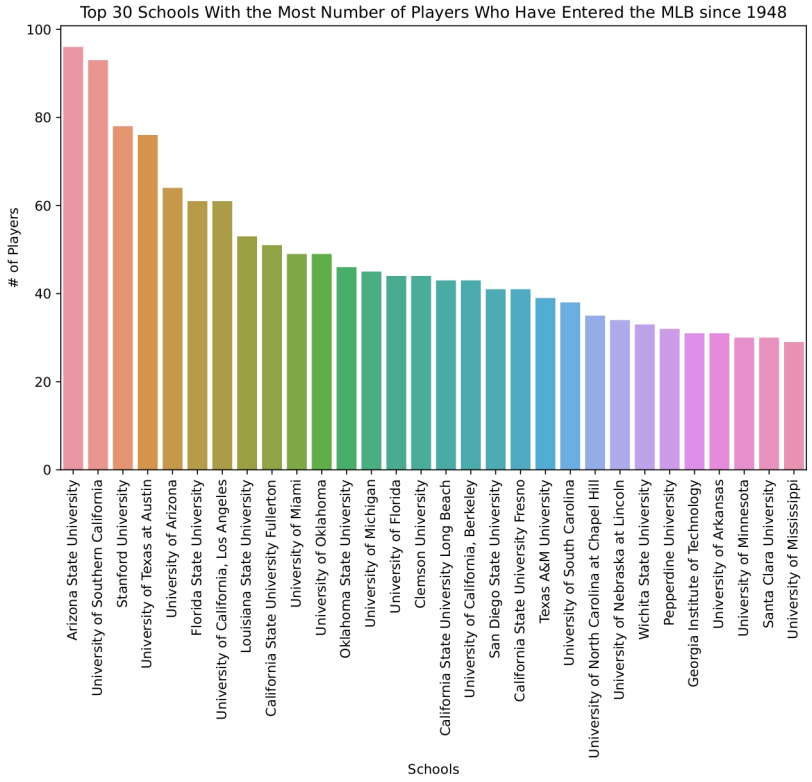

Figure 3: Determining Which Schools are Sending the Most Players to the MLB

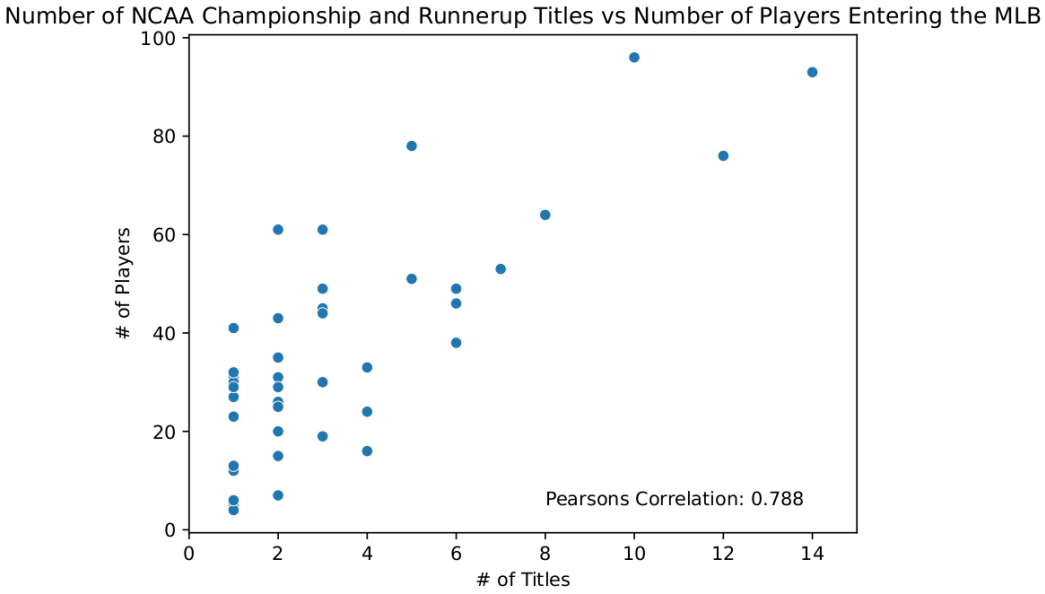

Figure 4: Correlation Between NCAA Championship/Runner-up titles v.s. Number of Players Sent to the MLB