Baseball Short Report

Daniel Buonauro

Introduction

The purpose of this report is to leave the viewer, no matter how much prior interest in baseball, thinking, “you know what, baseball is kind of interesting.” Though I can’t fit every piece of baseball data in this time around, I want to highlight some things that I hope you’ll find interesting, and that can be understood using the Lahman library.

Part 1: Homeruns

I think a great place to start this report would be some information about arguably the most exciting play in baseball: the homerun. Anyone who has attended a game or watched one on television has experienced the roar of the crowd when a ball is hit out of the park. Who has hit the most homeruns in the 21st century?

Most Homeruns Hit in the 21st Century

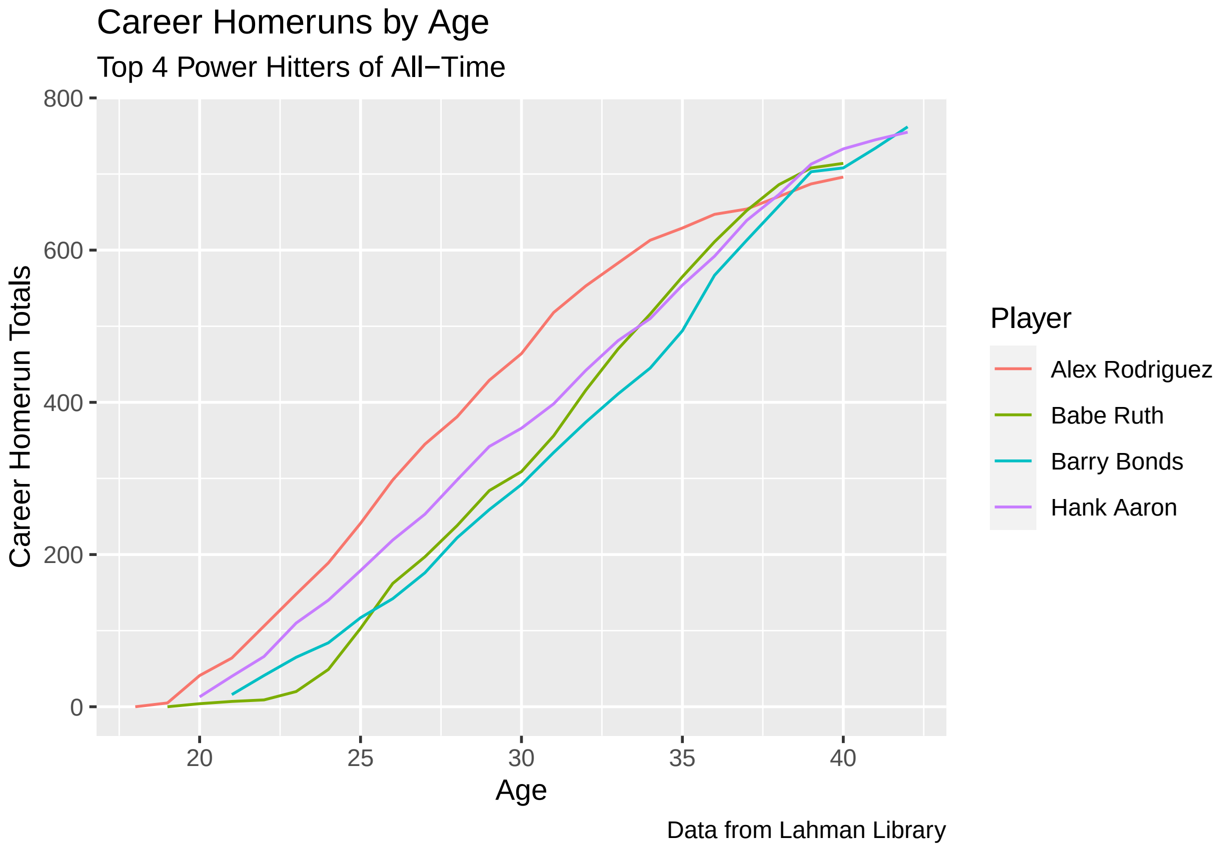
Player ID	Homerun Total
pujolal01	662
rodrial01	548
ortizda01	531
cabremi01	487
dunnad01	462
beltrad01	455
encared01	424
cruzne02	417
thomeji01	416
beltrca01	413

This data frame shows us the top 10 homerun hitters of the 21st century. Albert Pujols, Alex Rodriguez, Davd Ortiz, and Miguel Cabrera lead the list. To anyone who follows baseball, this may not be too much of a surprise. These four are legends of the game. Pujols and Cabrera are still active players, meaning you could go to a game and watch them play in person still! That, to me, is one of the most exciting aspects of baseball today. We are witnessing some of the greatest players in the history of the game. How fun is it to be able to appreciate them in real time?

In the history of baseball, no one has hit more homeruns than Barry Bonds, Hank Aaron, Babe Ruth, and Alex Rodriguez. Although, it’s worth noting that Bonds and Rodriguez were found to have taken performance-enhancing drugs, which certainly aided to their career totals. Even still, Bonds leads the group with 762 homeruns. How many did each player have at each age in their careers?

To find out, first we have to calculate each player's age during a season. In Major League Baseball, a player's age for a season is identified by the player's age on June 30th, which is around the midpoint of the regular season. So, depending on whether that player's birth month is during the first or second half of the year, some slight adjustments may need to be made.

Homeruns Hit by the Top 4 Homerun Hitters of All Time, by Age

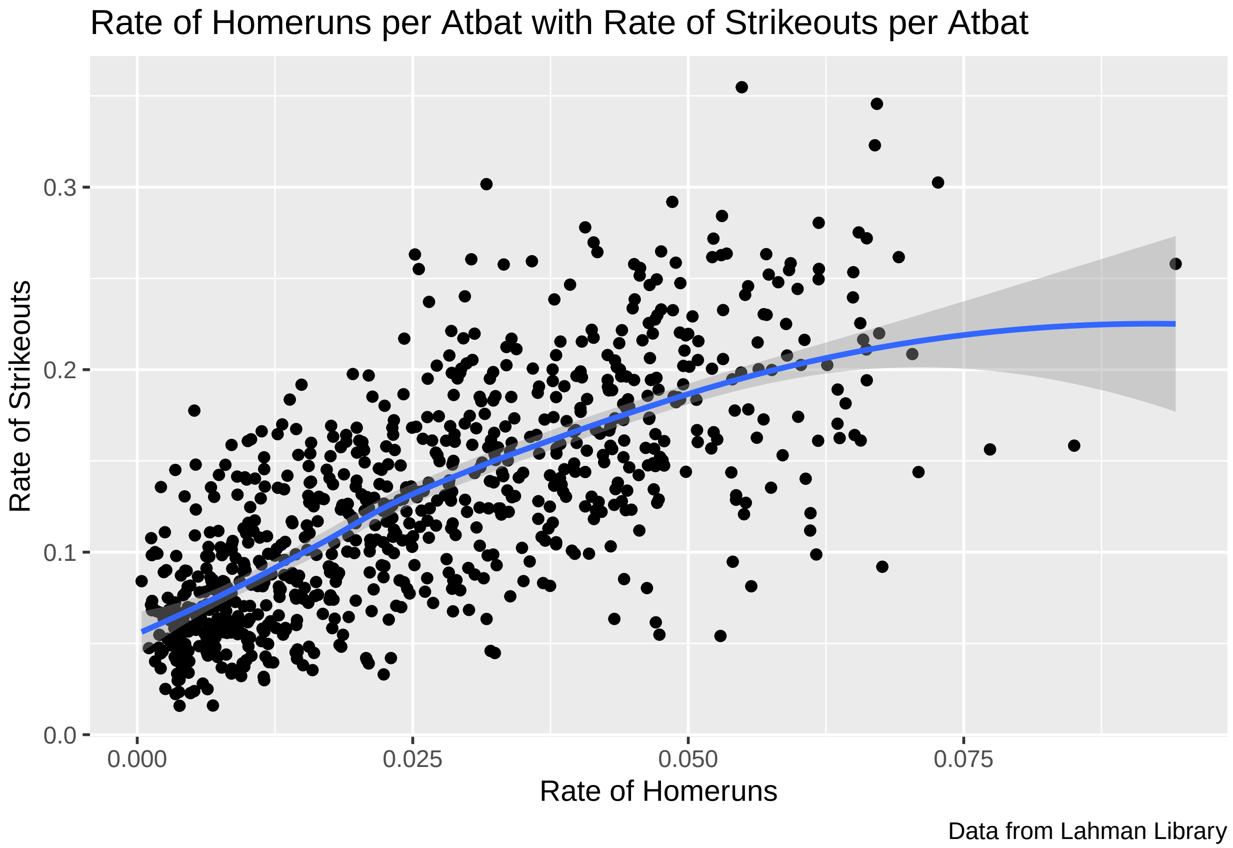


This visualization shows us that Alex Rodriguez made his debut the earliest of the group, beginning his career as a teenager. Babe Ruth started his career as a pitcher during what's known as the "Dead-Ball Era," where ballparks were overly large, and the balls used were "dead," meaning that when hit, they did not travel very far, due to both the construction of the ball and overuse. Babe Ruth helped usher in a new era of baseball, and you can see his career totals increase significantly around the age of 24, in the year 1919, which marked the end of the Dead-Ball Era.

Another note of particular interest is the battle between Hank Aaron and Barry Bonds. Bonds had the fewest career homeruns of the group until his mid to late thirties, which is an incredible feat, considering the vast majority of players retire before they turn 35. Bonds of course, has been proven to have taken performance enhancing drugs, prolonging his career. This makes it difficult to assess how his career would have shaped up without cheating, but still, he (controversially) passed Hank Aaron to secure sole possession of the most career homeruns in 2007.

Homeruns are objectively cool. Strikeouts, at least from the hitter's perspective, are less cool. What's the relationship between homeruns and strikeouts? Are the hitters who swing for the fences also striking out a lot?

The Relationship Between Strikeouts and Homeruns



Rate of Homeruns per Atbat with Rate of Strikeouts per Atbat

Data from Lahman Library

Each dot represents a single player with at least five thousand career atbats. Players positioned further down the x axis represent those who hit a homerun in a higher percentage of their career atbats. Players further up the y axis represent players who struck out at a higher percentage of their atbats. We see a clear positive correlation between these two variables, meaning that players who typically hit a lot of homeruns also tend to strikeout a lot as well. Hitting for power comes at a cost!

Part 2: Pitching

To this point, the focus of this report has been on homeruns. To avoid causing nightmares to any pitchers, let's take a look at some interesting pitching data. Earned run average (ERA) is calculated by dividing the number of earned runs (runs allowed that are deemed the pitcher's responsibility, as opposed to runs that scored due to a fielder making an error) by the number of innings pitched, multiplied by 9. We multiply by 9 because that is how many innings there are in a game. The lower the ERA, the better the pitcher. This makes sense because the point of the game is to score more runs than your opponent. By allowing fewer runs, your team is in a better position to win the game. So, a pitcher is evaluated by how many runs they allow on average.

Let's take a look at which pitchers have the lowest single season ERA since the end of World War II. Let's take the top 5 from each league.

Lowest ERAs since World War II, Top 5 in each league

League	Player ID	ERA	Year
AL	tiantlu01	1.60	1968
AL	chancde01	1.65	1964
AL	guidrro01	1.74	1978
AL	martipe02	1.74	2000
AL	mcdowsa01	1.81	1968
NL	gibsobo01	1.12	1968
NL	goodedw01	1.53	1985
NL	maddugr01	1.56	1994
NL	maddugr01	1.63	1995
NL	greinza01	1.66	2015

This table shows the top 5 lowest single season ERAs by a pitcher since World War II. These pitchers had an insanely dominant season. For reference, the average ERA in 2021 was 4.47. That means, in a 9 inning average, a statistically average pitcher would be expected to allow 4-5 runs per game. These pitchers, on average, allowed over 3 runs less per game!

Let's pay particular attention to the year column. 1968 appears three times in this data frame! This is no special coincidence, as 1968 is known as "the year of the pitcher." In 1968, Bob Gibson of the St. Louis Cardinals recorded the best single season ERA in modern baseball history, at 1.12. This is almost half a run better on average than the second lowest ERA of all time! After the dominance of 1968 shown by Bob Gibson, Luis Tiant, and Sam McDowell, Major League Baseball found themselves scrambling. Offense was at a historic low that season. In response to the pitching dominance, the League lowered each ballpark's mound by 5 inches, and shrank the strike zone, in the hopes of creating more offense. These rules became known as "the Gibson rules", signifying the historic dominance of Gibson's incredible 1968 season.

Another way the success of a pitcher is measured is by his strikeout to walk ratio. Ideally, this should be as large a number as possible. A pitcher wants to strikeout way more hitters than he walks. Walks are basically free passes. How can your defense help you record outs if you're giving a hitter first base for free? In the other direction, strikeouts are entirely up to the dominance of the pitcher. After all, nobody can score if every hitter who comes to the plate strikes out! That is why this statistic holds weight. In both cases, the result is due to pitcher execution. Here, let's exclude intentional walks, where a team decides to give a hitter a free pass to first base for strategic reasons.

Highest Strikeout to Walk Ratio since World War II

Player ID	K:BB Ratio	Year	Outs	Wins	Losses	Strikeouts	Walks	Intentional Walks
maddugr01	12.642857	1997	698	19	4	177	20	6
hugheph01	12.400000	2014	629	16	10	186	16	1
maddugr01	10.176471	2001	699	17	11	173	27	10
maddugr01	10.117647	1996	735	15	11	172	28	11
schilcu01	9.875000	2002	778	23	7	316	33	1
maddugr01	9.050000	1995	629	19	2	181	23	3
martipe02	8.875000	2000	651	18	6	284	32	0
martipe02	8.694444	1999	640	23	4	313	37	1
scherma01	8.625000	2015	686	14	12	276	34	2
sheetbe01	8.516129	2004	711	12	14	264	32	1

This data frame shows the top 10 highest single season strikeout to walk ratio. We see legend Greg Maddux appear on this list multiple times, and he is known for his remarkable efficiency. We see that in the year 1997, Maddux had over 12 strikeouts per walk. For reference, 2021's leader in strikeout to walk ratio was

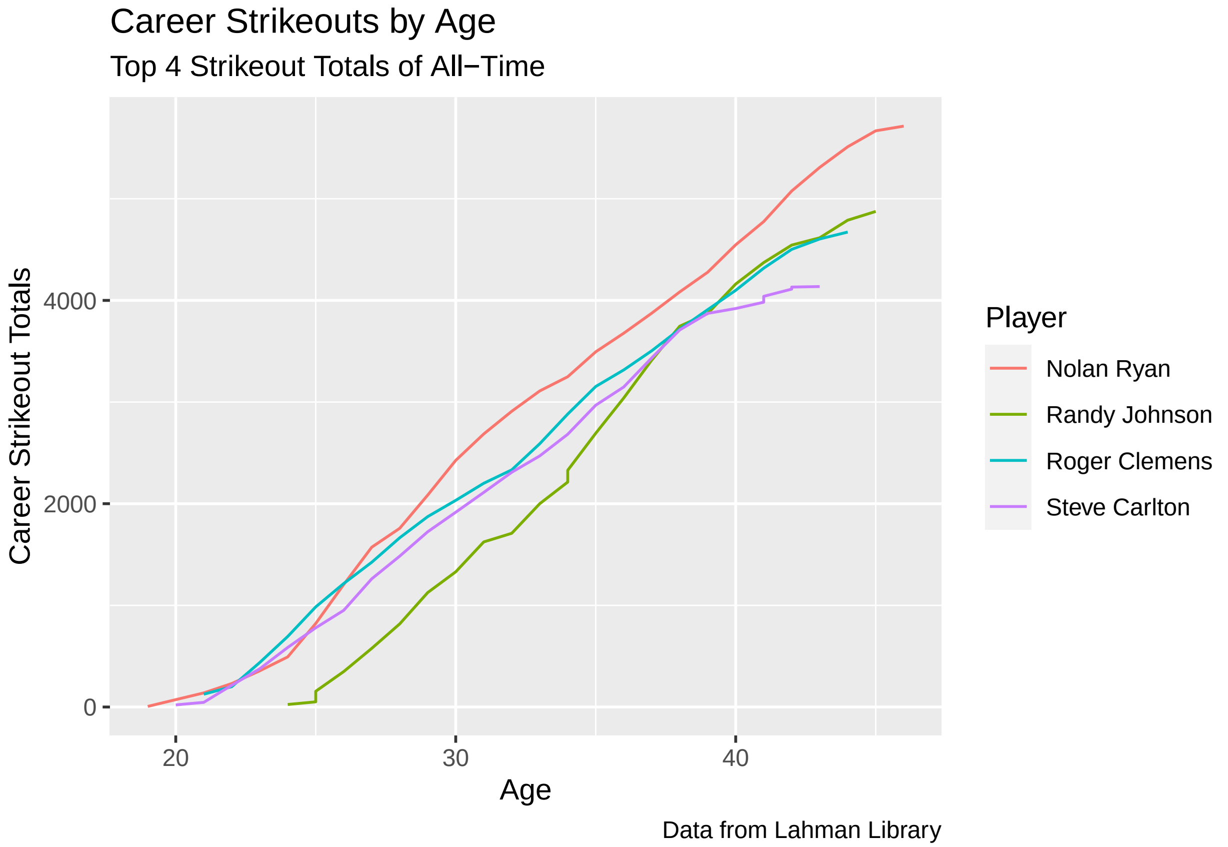
Corbin Burnes, who had a ratio of 6.882. Maddux's best season is double that! No one anointed me the strikeout to walk ratio guardian, but I think a good cutoff for what can be considered to be a 'good' ratio is 3 strikeouts to 1 walk. Maddux's best season is 4 times that!

We also see each pitcher's win-loss record in that particular season. I included this to show just how dominant these particular seasons were. A pitcher receives a win when he is the pitcher of record when his team takes the lead for good, though there are a couple rare exceptions. Wins are reliant on the team performing well, but we can see from the data frame that these pitchers accumulated many more wins than losses for their respective teams.

Strikeouts can be just as exciting as homeruns. Strikeout legends are just as important as homerun legends. But, it should be said that strikeouts aren't the only measure of a pitcher's success. Pitching is about getting outs, and doing so as efficiently as possible. Strikeouts are just one way to get an out!

With that being said, strikeouts are really exciting. Let's give the same attention to the top 4 all-time strikeout pitchers that we gave to the top 4 all-time homerun hitters.

Strikeouts Earned by the Top 4 Strikeout Pitchers of All Time, by Age

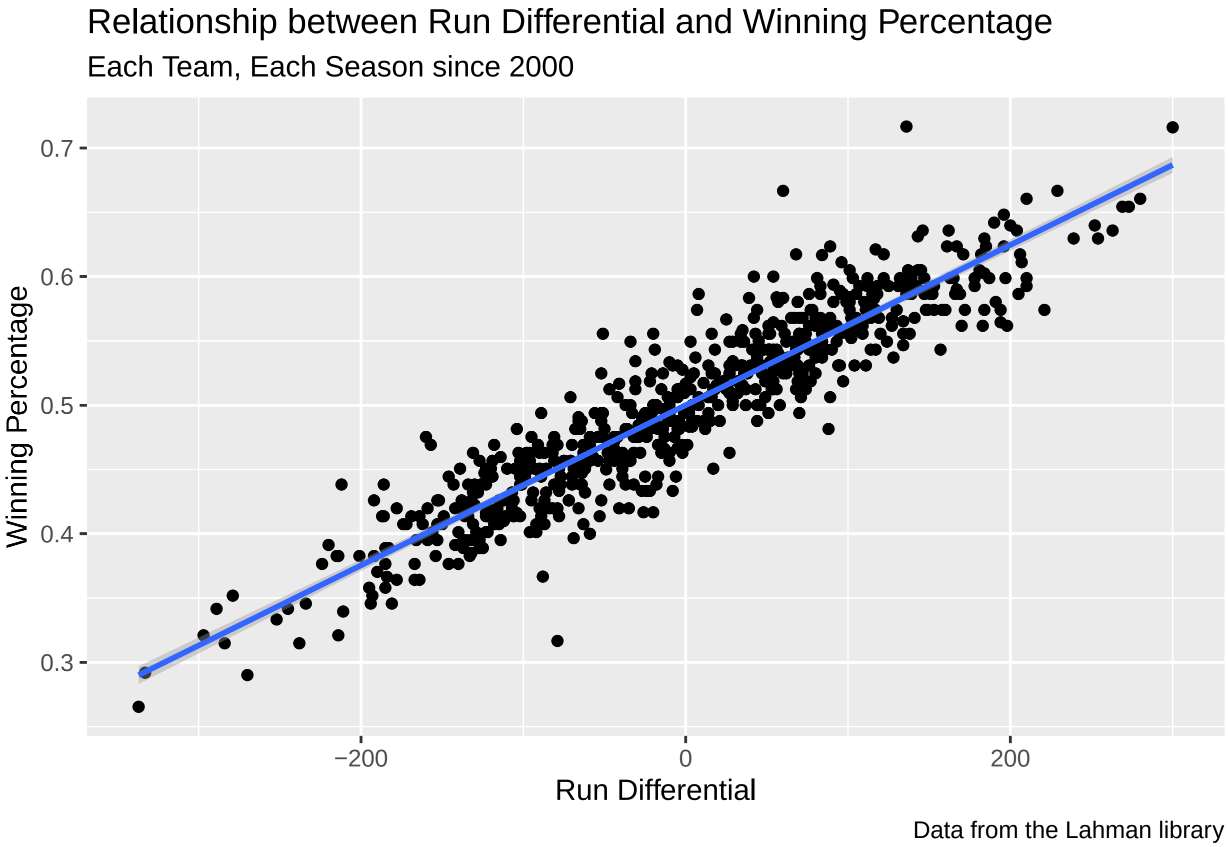


Like I noted for the hitters, I should mention that Roger Clemens very famously is linked to have taken performance enhancing drugs. We see in this visualization that Nolan Ryan, the all time leader, pitched well into his 40s, but so did the other leaders. This is much rarer than it seems! Randy Johnson, a recent hall of fame inductee, got off to a later start than the rest, but quickly made up for lost time, finishing second on the all time leaderboard. For some left-hander representation, both Randy Johnson and Steve Carlton throw left-handed!

Conclusion

We know that not only are homeruns exciting, they are guaranteed to score at least one run. We also know that pitching and defense are determined to give up the least amount of runs possible. To conclude this report, let's visualize the relationship between win percentage, and average runs scored, for each team's respective seasons from the beginning of the 21st century.

Plotting the relationship between Win Percentage and Run Differential



Perhaps expectedly, we see a significant positive correlation between run differential and winning percentage. Run differential is simply calculated by subtracting runs allowed from runs scored. If you score more runs than you allow, you'll win ball games. If your run differential is close to 0, you'll win approximately half of your games. This is why every facet of the game is important! Part of the challenge of building a successful team is constructing a balance between good hitting and good pitching, with a finite amount of financial resources.

Final Thoughts

We have reached the end of this relatively short report on baseball data. The goal was to not only highlight the really cool analysis you can do with baseball statistics in R, but maybe show readers that baseball is a sport worth checking out more. I just want to make a special note that not all baseball legends are from 'back in the day.' Not every historic record set happened before our grandparents were born. Barry Bonds broke the all time homerun record less than 15 years ago! In the game today, there are players who either already are, or will be considered some of the greatest of all time. The history of organized baseball can be traced back to the 1870s. Baseball has a wonderful and long history that is being written and expanded at every second. How exciting is it that we get to watch it all unfold?