The Evolution of Average Home Runs in Major League Baseball

Tanisq Jawahar

Introduction

As the game of baseball has evolved, one of the main elements that has changed drastically is the home run hitting. This report will highlight the changes in home run hitting over time. Home runs will be tallied up league wide over a given season and divided by the games played. The decades of focus will be the 1900s (1901-1910) and the 2000s (2001-2010). The early days of professional baseball, and the newer age as power hitting became a focus. We will examine how hitting has changed in Major League Baseball over the past century, specifically for home runs.

Background

Our data set comes from the website Openintro.com[1], which is an educational website supplying various resources, data sets, and videos. The data set, “mlb_teams”, is a data frame containing 2784 rows and 41 columns representing 41 different variables for various MLB (Major League Baseball) teams. The data set is a subset of MLB teams from a full database titled “Lahman’s Baseball Database”[2], which is available on the Lahman R package. We got this data by downloading a CSV file from this package and opening the data onto the “numbers” app of my laptop. Additionally, the site included an R downloaded file that we were able to investigate as well. The data represents MLB history from its first year of formation in 1876, until the year 2020, so it does not include the years 2021-2024. The data is able to account for all 30 current MLB teams, as well as all teams from the 1901-1910 decade.

The variables we chose to use were year, decade, homeruns, games_played, n, mean, and sd. Year represents the year of the specific MLB season, and is narrowed down from 1901-1910 and 2001-2010. Decade represents the specific decade that the year can be found in, and is represented in our data by either “Early 90s” for 1901-1910, or “Early 2000s” for 2001-2010. “homeruns” represents the amount of home runs hit during that specific season. A home run is defined as a ball hit either out of the back of the field, or where the runner is able to run around all four bases before being ruled out. In our dataset, home runs is a quantitative variable taking numerical values for the amount hit in the season. The “games_played” variable represents the amount of total games played in that specific season and is also a quantitative variable. The “n” variable represents the number of teams in that year. The “mean” variable represents the average number of home runs per game in a year. Finally, the sd variable represents that standard deviation of the average number of home runs per game in a year. Furthermore, we will be calculating homeruns divided by games_played to get the average number of home runs in a single season qhich is a quantitative variable. We overall wanted to use these variables to address the question, have the amount of home runs hit in the MLB changed significantly over the history of the league?

Baseball is one of the oldest organized sports in America. Its roots are traced back to the mid-1850s[3]. Soon thereafter, the game became known as the “national pastime.” The baseball of the early 1900s looks different than the early days of baseball, but is still vastly different from the baseball of today. The main focus of hitters was to get the ball into play. They wanted to hit the ball and did not care by what means, contact hitters as they are called. Players were hitting the ball at exceptional levels, just not over the fences.

Home runs were not taken seriously until Boston Red Sox pitcher-turned-outfielder Babe Ruth clubbed 29 homers in 1919. A new single season record, breaking 27 home runs hit in 1884. Ruth was sent to the New York Yankees that off-season. In the 1920 season, he hit an ungodly 54 home runs. Then the following season, Ruth further established himself by hitting 59 home runs. His mark of 60 in 1927 would be the single-season mark until 1961, 34 years later.

As the game of baseball evolved, the traditional contact hitter began to die out. Home runs became the fixation of players and fans alike. Roger Maris set the new single season record in 1961, with 61. His record was thought to be untouched, but it had a controlling presence over everyone.

But in 1998, something happened. There became a chase to topple the 37 year old record. St. Louis Cardinals first baseman Mark McGwire and Chicago Cubs outfielder Sammy Sosa began a chase to break the record. By season's end, McGwire hit 70 and Sosa hit 66.

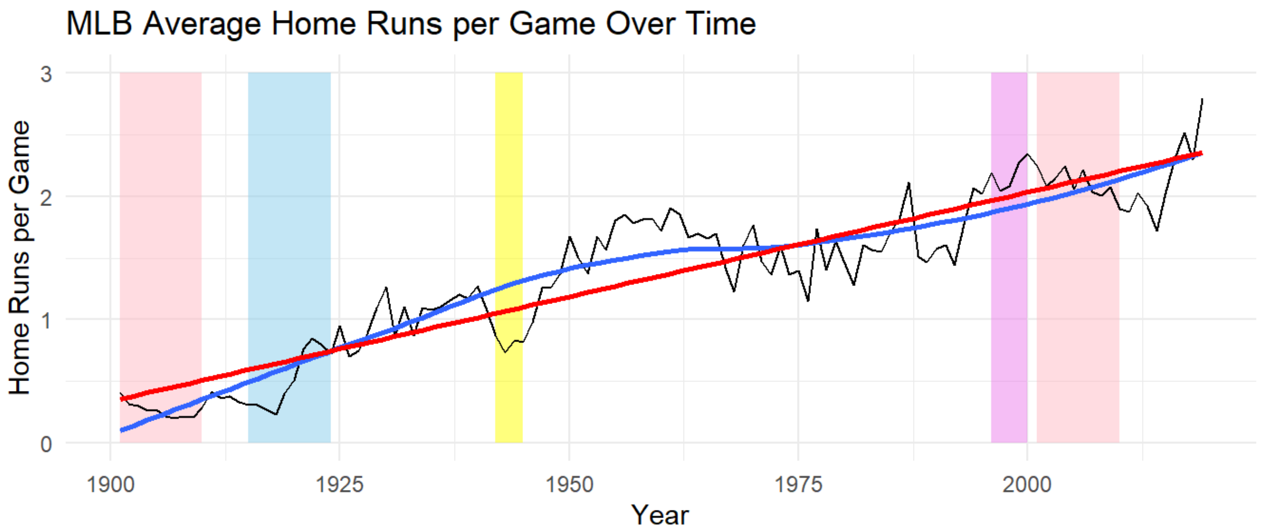
What ensued for the following decade-plus was a quest for professional baseball players to hit as many home runs as possible. There were twelve 50-plus home run seasons throughout the 2000s, a quarter of all such seasons.

As home run hitting increased at abnormal rates, the MLB grew suspicious. The very players that had brought baseball back from the dead, were using steroids to enhance their game. What ensued was the tarnishing of the players allegedly using steroids and their accomplishments. Several all-time great players were effectively excommunicated from the baseball world, and more specifically the Baseball Hall of Fame.

As a result, these specific factors had an effect on our data. Additionally, from looking at some of the earlier years of data in the set, such as in the late 1800s, multiple variables were missing from the data as they were either not recorded or not an official baseball stat yet. Also in the 1800s, the season had not yet been expanded to the 162 game format that it is today. Additionally, there were a few shortened seasons in the original data set in the 1990s as a result of lockouts, however we decided to not use this data to get an equal number of games. One final factor is that the number of teams as well as team names and locations has changed over time, and was not the same 30 teams in the early 1900s as it is today.

In the rest of the report, we will first be filtering through the entire MLB data set to just isolate the decades 1901-1910 and 2001-2010, and creating a summary table, then analyzing how average home run hitting has changed over time between these two decades. Overall, we found that the average number of home runs per MLB season has significantly increased over time from the early 1900s decade to the 2000s.

Analysis



For our graphical portion, we wanted to perform various graphical and statistical tests to determine how home runs had changed over the decades we were examining. We began by creating a line graph for home runs per game over the years. We included both a smooth linear regression graph (shown in red), as well as a smooth, curved graph shown in blue to represent the predicted home run per game values over time. Then, we added in a line graph to represent the actual values of the data over the years (shown in black). Additionally, we wanted to add highlighted areas to the graph to represent significant periods.

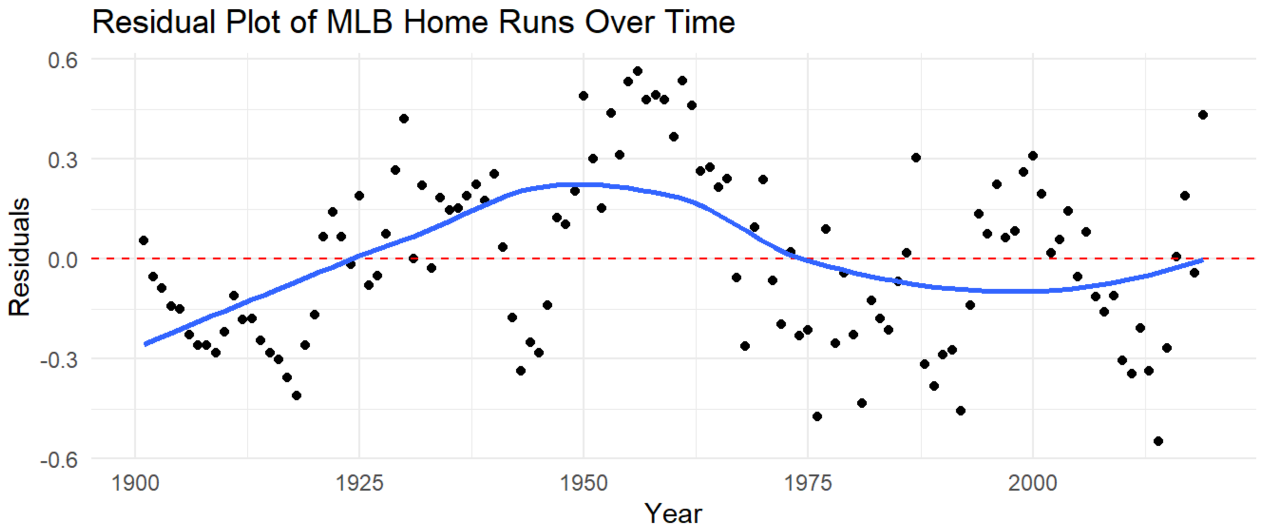
There are a few different highlight periods of time. These are very important periods in the history of baseball, and especially when it comes to the story of home run hitting. For starters, the pink periods of 1901-1910 and 2001-2010 are the focal points of this project. The 1900s is the decade in which baseball was first taken seriously, but not a lot of home runs were hit. The 2000s by comparison, was one of the most prolific offensive periods in baseball. Several players were allegedly using steroids to inflate their numbers.

The period of 1915-1924 is the transitional period from contact hitting to the introduction of power hitting. As the graph shows, there was a significant jump in home runs during this time. This is thanks large in part to Babe Ruth setting the tone for what a home run hitter is.

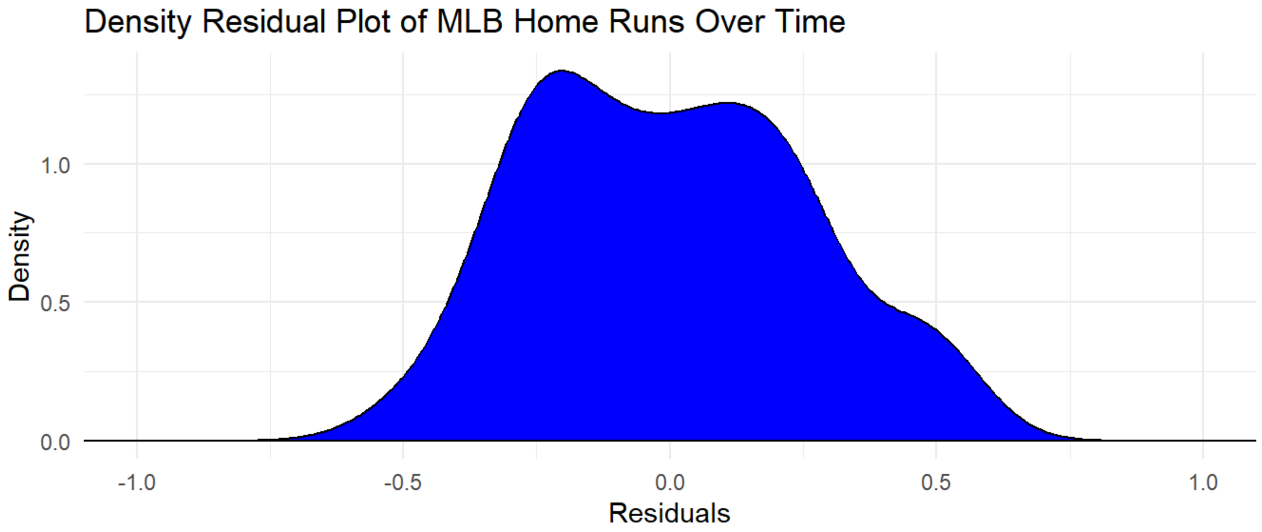
The next period is during the years of 1942-1945. World War II was ongoing, and like many Americans, several baseball players enlisted in the military. It was not just fringe major league players going off to war, several superstars joined the efforts. Ted Williams, who was just entering the prime of his career, took off 1943-1945 and became a decorated Naval and Marine pilot.

Lastly, 1996-2000. This is the preamble to the steroid era of the 2000s. Many pinpoint this time period to when players first started using performance enhancing drugs. These 4 years is also when offensive production started to pick up.

In addition, if we are assuming a linear model for our data, the estimated slope is approximately 0.017. This means that for every one-unit increase in the year, the expected value of xbar increases by approximately 0.017. However, the assumption of a linear model may not be best supported for our data as there are periods of data that fall above and below the line in chunks. The data is not randomly distributed above and below the line, thus creating an ‘S’ shape.

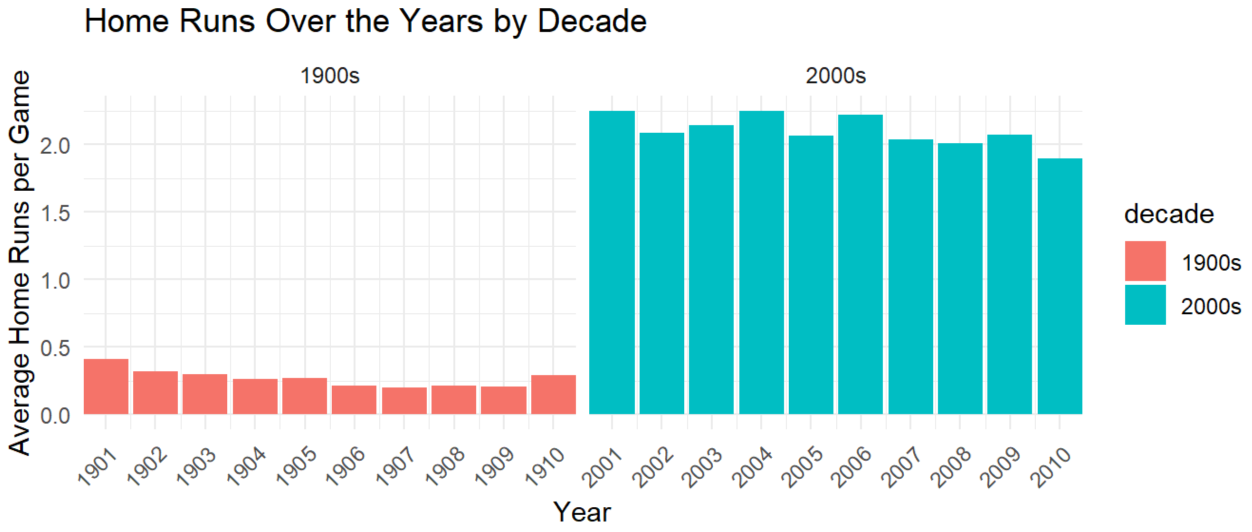


This is a residual plot for the MLB home runs over time for each year from the period 1900-2010. We created a line of best fit as well for the graph and a scatter plot for all the points to demonstrate that the line residuals appear to increase from 1900-1950, then begin to dip until the year 2000, then increase again. Additionally, between 1925 and 1975, the residuals are positive, while for the rest of the years the residuals are negative. There is an down-up-down-up pattern in the residuals which indicates that fitting a straight line may not be appropriate for our data. A better model such as an ‘S’ shaped curve may be more fitting which goes beyond the scope of this course.



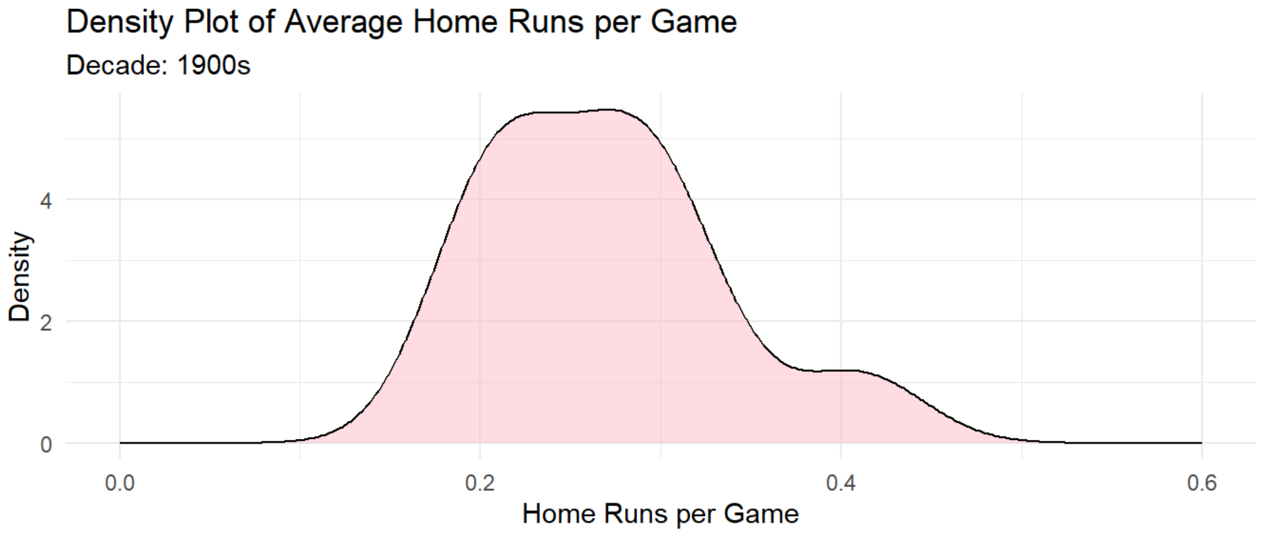
There is evidence of non-linearity as the residual distribution is not symmetric and is skewed right. This would affect accuracy of predictions for individual home runs per games in the MLB.

Graphs of the 1900s and 2000s Decades



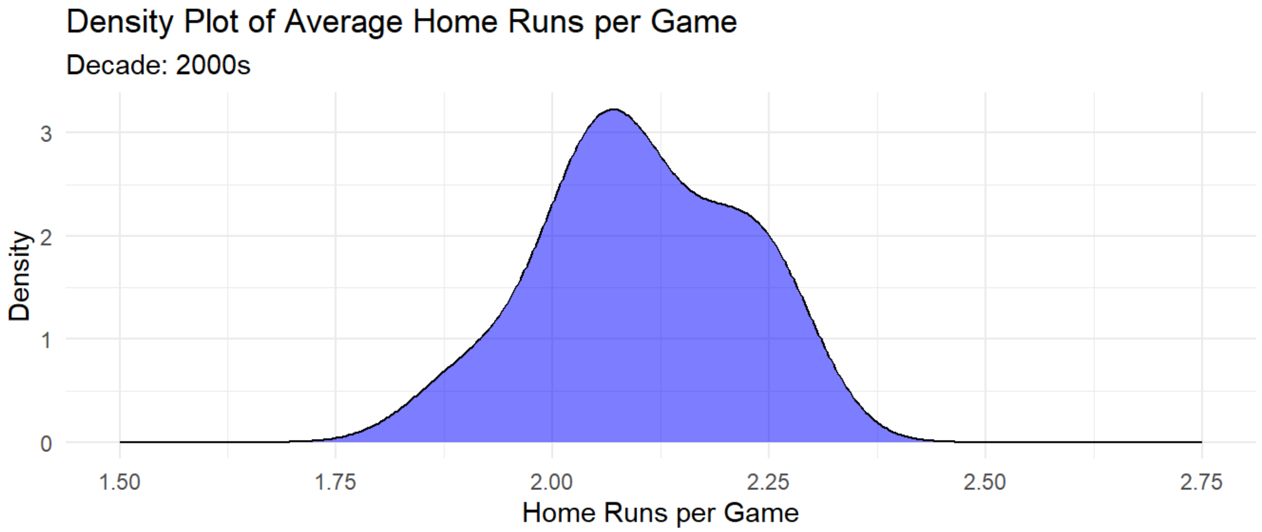
The next step in our graphical analysis was to create side by side bar graphs to truly show the change in the number of home runs hit per game between the two decades we wanted to focus on, the 1900s and 2000s. We created a bar plot for each year of the decade, the orange bars on the left representing the 1901-1910 decade, the blue bars on the right for the 2000s decade. While we did not see significant changes in the decades themselves, isolating these two decades helped illustrate the significant increase in home runs per game between these decades.

Density Plots


Density Plot of Average Home Runs per Game

Decade: 1900s

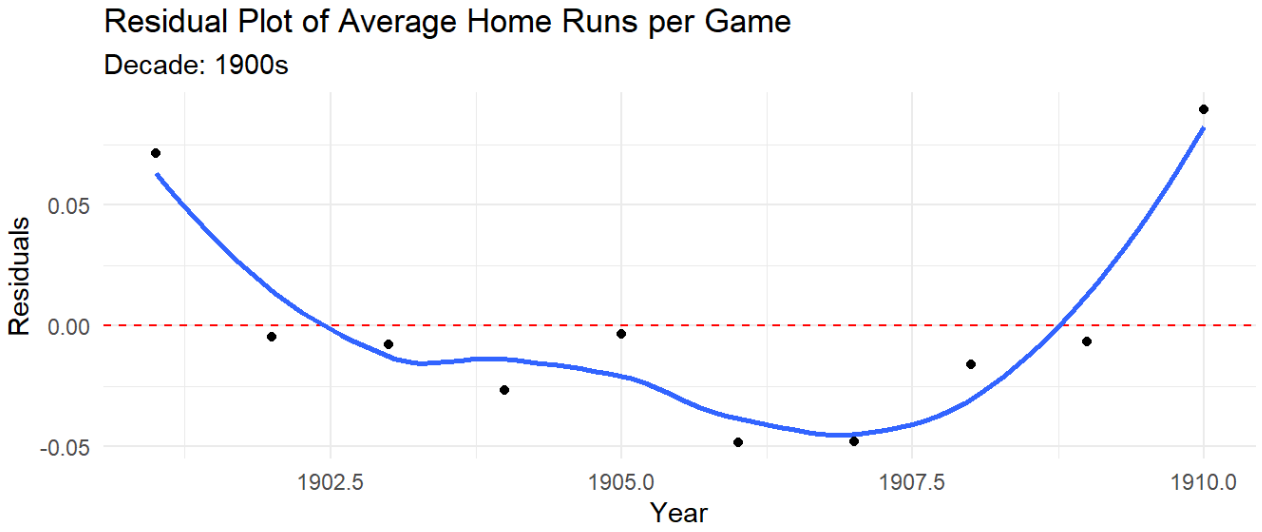
The density plot of average home runs per game in the 1900s shows the distribution of data over the given time period. The data is skewed to the right, with the bulk of the data concentrated towards the lower values, and fewer data points towards the higher values.

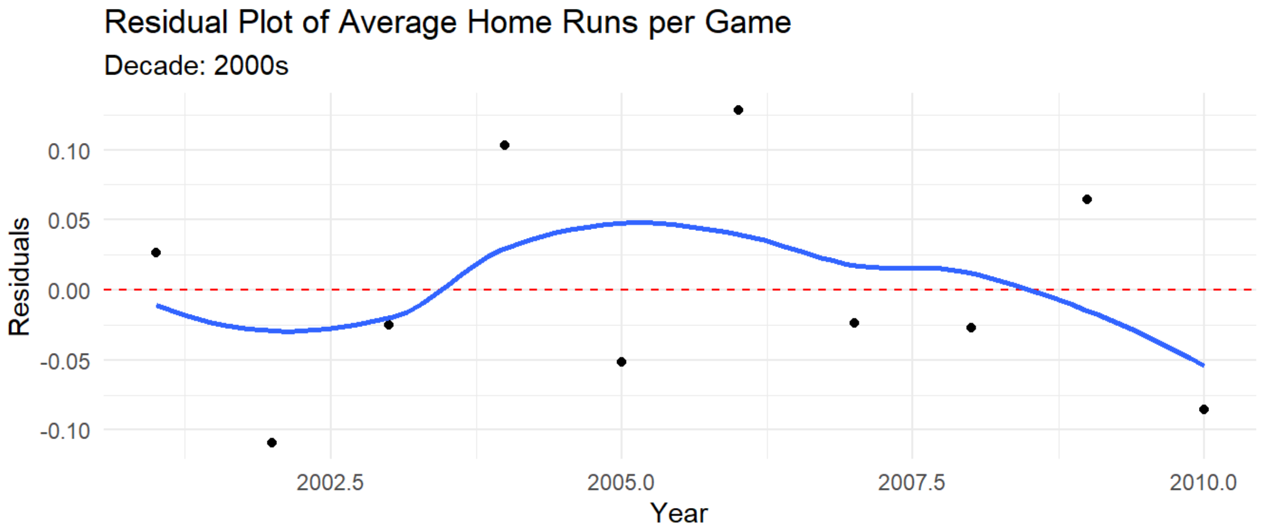

Density Plot of Average Home Runs per Game

Decade: 2000s

The density plot of average home runs per game in the 2000s shows the distribution of data over the given time period. The data is skewed to the left, with the bulk of the data concentrated towards the higher values, and fewer data points towards the lower values.

Residual Plots





We additionally made corresponding residual plots for these two decades. For the 1900s, the residual plot illustrated positive residuals from 1901-1905, then negative residuals from then on, as well as decreasing residuals from 1901-1908. The 2000s decade residual plot illustrated positive residuals from 2001-2006, then negative from then on, as well as decreasing residuals throughout, besides a slight increase in 2003. The residual plots for both of the specified decades again show that a linear model may not be the be suited for our data. The points are not randomly distributed among the y-intercept of 0 and the pattern of the residuals indicates that fitting a curve may be more appropriate.

Confidence Intervals

Single Interval Tests

Confidence Interval for Mean of MLB data (1900s)

Confidence Interval	Value
Lower	0.2224923
Upper	0.3161063

Confidence Interval for Mean of MLB data (2000s)

Confidence Interval	Value
Lower	2.022033
Upper	2.182306

The last step of our graphical analysis shows our confidence interval and hypothesis tests. We created two separate 95% confidence intervals for the 2000s and 1900s decades. We are 95% confident that the mean number of home runs per game in the 1900s is between 0.22 and 0.32. In addition, we are 95% confident that the mean number of home runs per game in the 2000s is between 2.02 and 2.18.

95% Confidence Interval for μ_1 - μ_2

Statistic	Value
Mean 1900s	0.2692993
Mean 2000s	2.1021696
CI Lower	-1.9205757
CI Upper	-1.7451650

Additionally, we then compared the two confidence intervals, illustrating the differences between each. We are 95% confident that the mean home runs per game between the 1900s and 2000s is between -1.92 and -1.75 home runs per game shorter in the 1900s decade than the 2000s in the MLB.

Hypothesis Test

Finally, we created a hypothesis to test if the average home runs per game in the 1900s is the same as the average home runs per game in the 2000s where μ_1 = 0.27 is the average value from the 1900s and μ_2 = 2.10 is the average value from the 2000s.

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2$$

- Test statistic $t = \frac{\bar{x}-\bar{y}}{\mathrm{SE}(\bar{x}-\bar{y})}$

Statistic	Value
p-value	6.10e-17

We found extremely strong evidence that the home runs per game in the 1900s was not the same as in the 2000s, based on the p value shown in the graph.There is extremely strong evidence that the average home runs per game in the MLB is larger in the 2000s than in the 1900s in our study. The p-value is approximately 6.10e-17.

Discussion

Our graphical analysis provides significant evidence that there has been drastic change in the hitting habits of MLB players over the past 100 years. First, from the line graphs we created, they show a steady increase in the average home runs hit per game, as well as in the side by side bar graphs, they show a significant change between the two decades specifically. In the early 1900s decade we focused on, batters were hitting around 0.3 to 0.5 home runs per game or about a single home run every 2-3 games. In addition, to back this observation, our single confidence interval for the 1900s shows that the mean number of home runs per game is between 0.22 and 0.32. However, by the time it reached the early 2000s decade, batters were hitting a little over 2 home runs per game. The single confidence interval for the 2000s shows that the mean number of home runs per game is between 2.02 and 2.18. Also, with the confidence interval between the two averages, we concluded that the mean home runs per game between the 1900s and 2000s is between -1.92 and -1.75 home runs per game shorter of the 1900s decade in the MLB. Additionally, through the hypothesis test we performed to test if the two means from the two specified decades, it is clear that there are significant differences between this old age of MLB hitting, and hitting nowadays as the p-value was extremely small with a value of 6.10e-17.

Furthermore, there are other factors that contribute to the results of this data. Firstly, players were not focused on hitting home runs during the 1900s. The main initiative of ball players then was to get the ball into play and drive in the runners on base. Scoring was important, but the art of hitting a ball over the fences simply was not.

Additionally, the total number of home runs being much lower is due to the amount of teams in the 1900s. There were 16 teams throughout that decade, compared to 30 today. With there being less teams, there are also less players in the Major Leagues. The number of players on a MLB roster at one given time, 26, has not changed, but the number of teams has increased. Naturally, the amount of players hitting such figures is going to increase. The length of the season has also increased as baseball has progressed. In 1901, the length of the season was 140 games. By the time we get to the 2000s, the season is 162 games long.

Home run hitting grew as the game of baseball evolved. But the sharp increase in home run hitting in the 2000s is mostly contributed by steroid usage. While it is unconfirmed that some players, like Bonds and Sosa, did steroids, it is widely accepted that they did. In some cases even, they tested positive for steroid usage. McGwire never tested positive but he admitted guilt. Alex Rodriguez, who hit 696 career home runs and 57 in 2002, was suspended for the entirety of the 2014 season due to his steroid use and involvement in the Biogenesis scandal. Manny Ramirez, one of the most prolific players of the 2000s, was suspended for 50 games in 2009 and avoided a 100 game suspension in 2011 by retiring.

Our study supports the conclusion that the average home runs per game has significantly increased over time from the early 1900s to the 2000s.

---

[1] https://www.openintro.org/data/index.php?data=mlb_teams (https://www.openintro.org/data/index.php?data=mlb_teams)↩

[2] https://github.com/cdalzell/Lahman (https://github.com/cdalzell/Lahman)↩

[3] https://blessyouboys.com/2023/3/20/17645074/mlb-history-professional-baseball-american-league-national-league (https://blessyouboys.com/2023/3/20/17645074/mlb-history-professional-baseball-american-league-national-league)↩