Abstract

In this report, statistics for all 30 teams in Major League Baseball (MLB) in 5 years will be explored. All the data is sourced from Baseball-reference, a site run by Sports Reference. The data of Team Standard Batting, Team Standard Pitching and Miscellaneous Team Info from 2017 to 2022 will be used while 2020 season data will be excluded due to it being a pandemic shortened season. This report will focus on two questions, what is the most important thing for winning a baseball game and whether having a winning season will bring more attendance. The league year-by-year batting total is also used for comparing our observations to real life situations.

Tidying the data

Since the data are stored in tables on the website, most of them are tidy. Except for win loss records that use hyphens to connect number of wins and number of losses together. For example, the '1Run' column uses 15-18 to represent the team winning 15 games and losing 18 in 1 run game. In order to make this data tidier, the separate function is used to separate the two numbers. By adding a year column representing the year the data is from. A dataset of 5 years MLB stat is created after the 5 year-by-year dataset is joined together.

Runs and wins

Obviously, a baseball game can be won by gaining more runs than the opponent. Before any graph is generated, it is important to make sure the run environment in these 5 years is similar so that the average runs created should not fluctuate a lot. From Figure 1, it shows the run environment is stable throughout the years, around 10% fluctuation every year. It should be safe to use the number of runs.

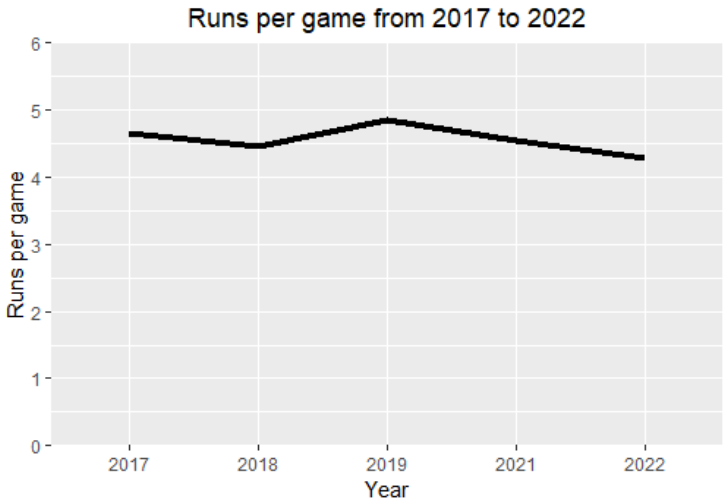


Figure 1. runs per game from 2017 to 2022

By plotting the runs per game (R/A) to wins by every team (Figure 2a), we can see gaining runs have a strong correlation with winning games with 0.7424 correlation coefficient (r). Several outliers can be seen from figure 2a, like the 2022 Houston Astros and 2019 Baltimore Orioles. Since a team can prevent an opponent from gaining runs to win, it is possible that the teams with lower runs but more wins have good pitching or defense to prevent the opponent gaining runs.

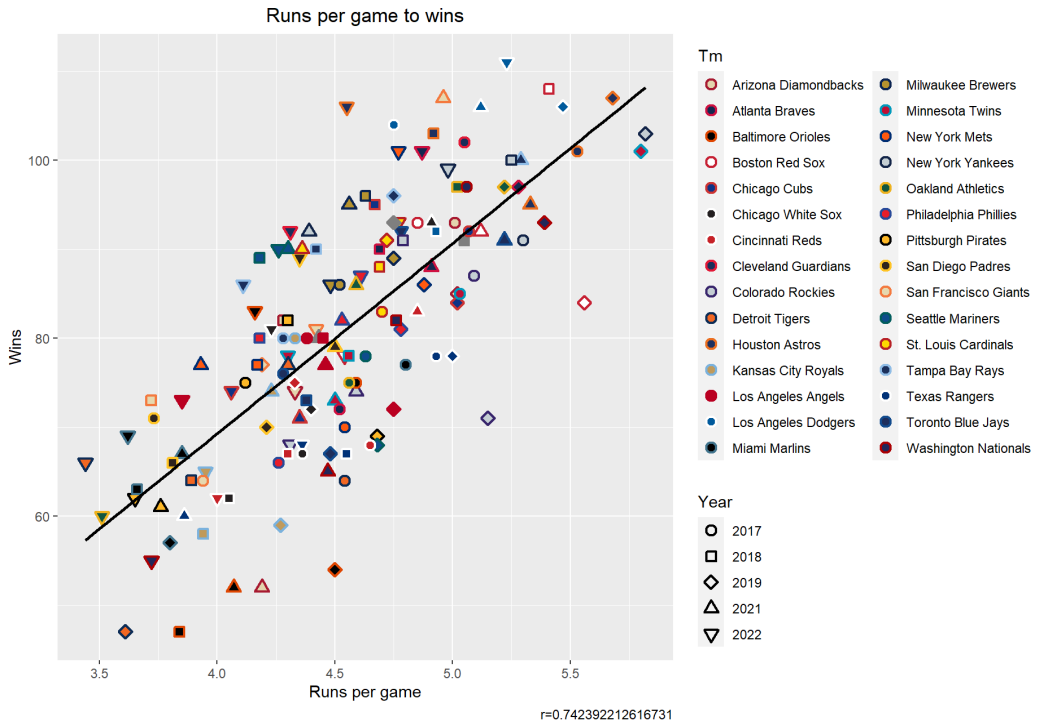


Figure 2a. Scatter plot of runs per game to wins from 2017 to 2022

However, according to figure 2b, even though a lower runs allowed has equivalent correlation coefficient with gaining run one, the outliers in figure 2a are not outliers in figure 2b. This shows either of these figures can reflect the key to win on their own, the correlation can be much stronger if both stats are used in a parameter.

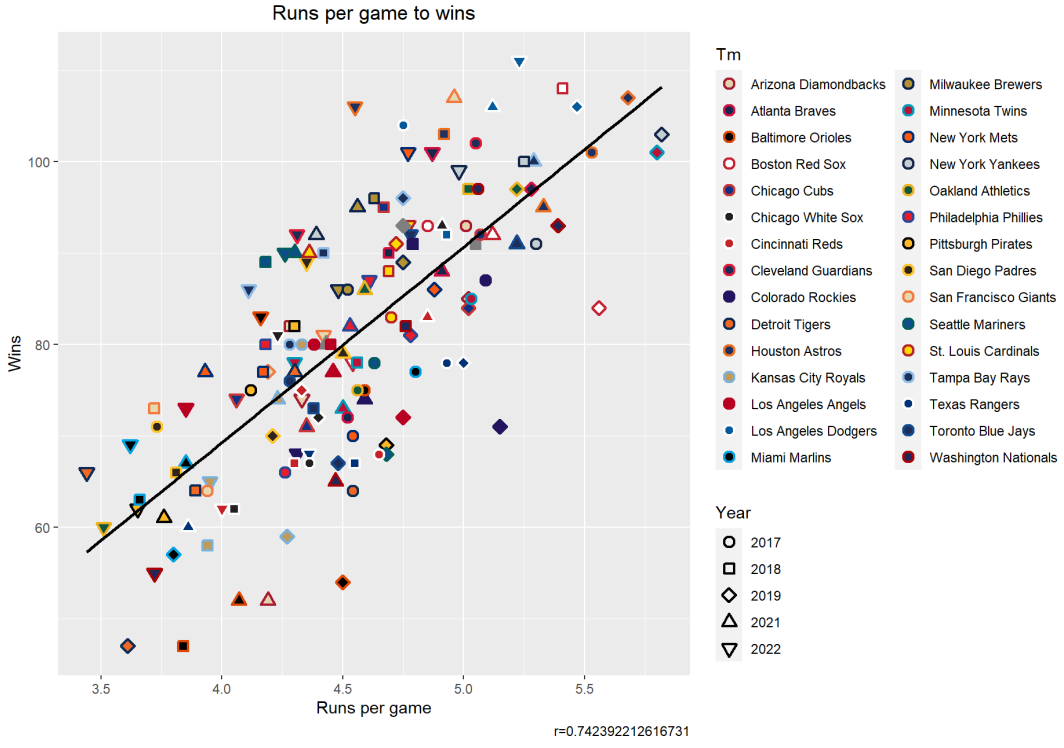


Figure 2b. Scatter plot of runs allowed per game to wins

In figure 2c, the difference between runs per game and runs allowed per game is used. The correlation coefficient is much higher than the 2 previous graphs (0.9593), but there are several outliers in this figure as well, like 2018 and 2021 Seattle Mariners. This can be explained through figure 2d, number of games won or lost by a team with just 1 run. It is obvious that 2018 and 2021 Seattle Mariners have a much higher win rate in 1 run ball game. This is why despite the team having negative runs minus runs allowed, they still get around 90 wins.

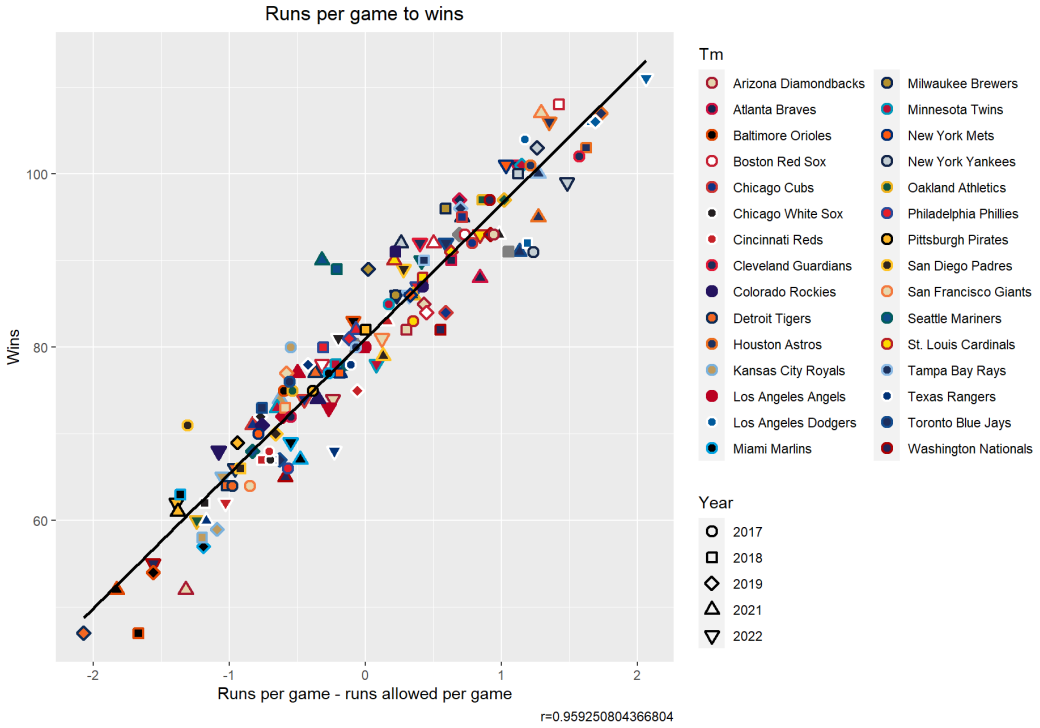


Figure 2c. Scatter plot of runs minus runs allowed per game to wins

Effectively create runs

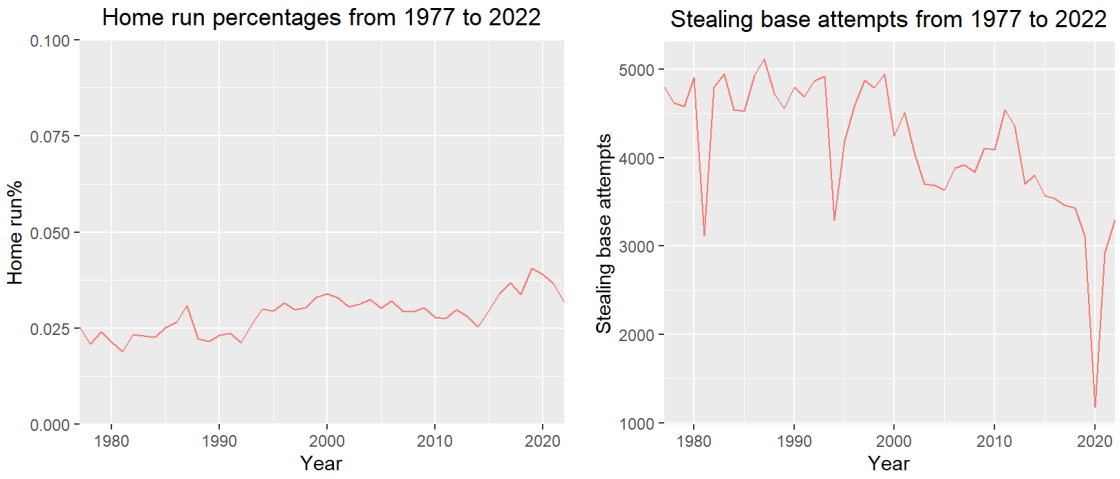
Now the importance of runs is confirmed, several statistics will be compared in order to find out what kind of hitting approach will be most beneficial to gaining runs. In table 3a, the correlation coefficient of 6 kinds of outcomes during batting. In which home runs and base on balls have a stronger correlation with scoring than others. This led to a steady climb in the number of home runs over the years.

	2B	3B	HR	HBP	SF	SB
r	0.383	-0.168	0.563	0.210	0.372	0.0725

Table 3a. Summarisation of 6 kinds of outcomes' correlation coefficient

In figure 3b, the percentage of home runs has increased from around 2.5% to 3.5% in the past 40 years, which is also the period of computers entering the analytic side of baseball. Meanwhile, stealing bases become less and less important in a game of baseball, from figure 3c, the total attempt of stealing bases have dropped 30% over 35 years. Note that the

1985, 1994 and 2020 season was shortened due to strikes and pandemic, resulting in dropping spikes in the figure.



(Left) Figure 3b. Line graph of percentage of home runs from 1977 to 2022
(Right) Figure 3c. Line graph of total attempt of stealing bases from 1977 2022

While triples hit has a negative correlation coefficient, since it is so rare to occur (less than 0.3%), the number will greatly fluctuate and lead to a correlation coefficient tend to 0, so it will be treated as irrelevant and omitted.

Wins and attendance

Next, in figure 4a, correlation between attendance and win is shown. Although the correlation coefficient is low (0.404), it is expected since teams with large market size usually have a higher amount of audience even if they did not have a good record, vice versa. In figure 4b, it shows that huge market teams like Los Angeles Dodgers and New York Yankees always lead in attendance while small market teams like Oakland Athletics and Tampa Bay Rays often came last in attendance despite having a good record.

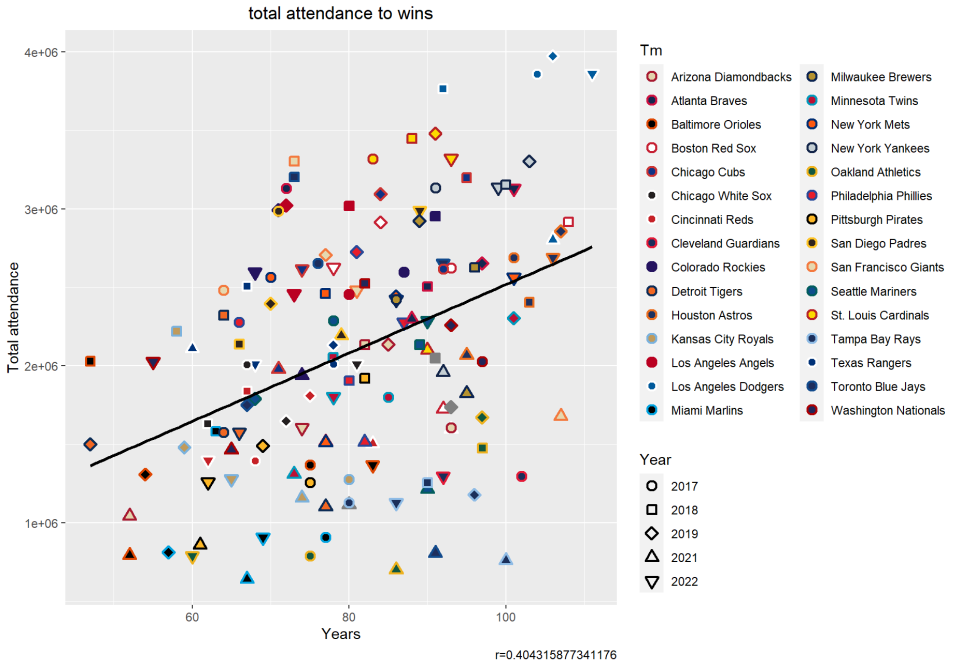


Figure 4a. Scatter plot of total attendance to win

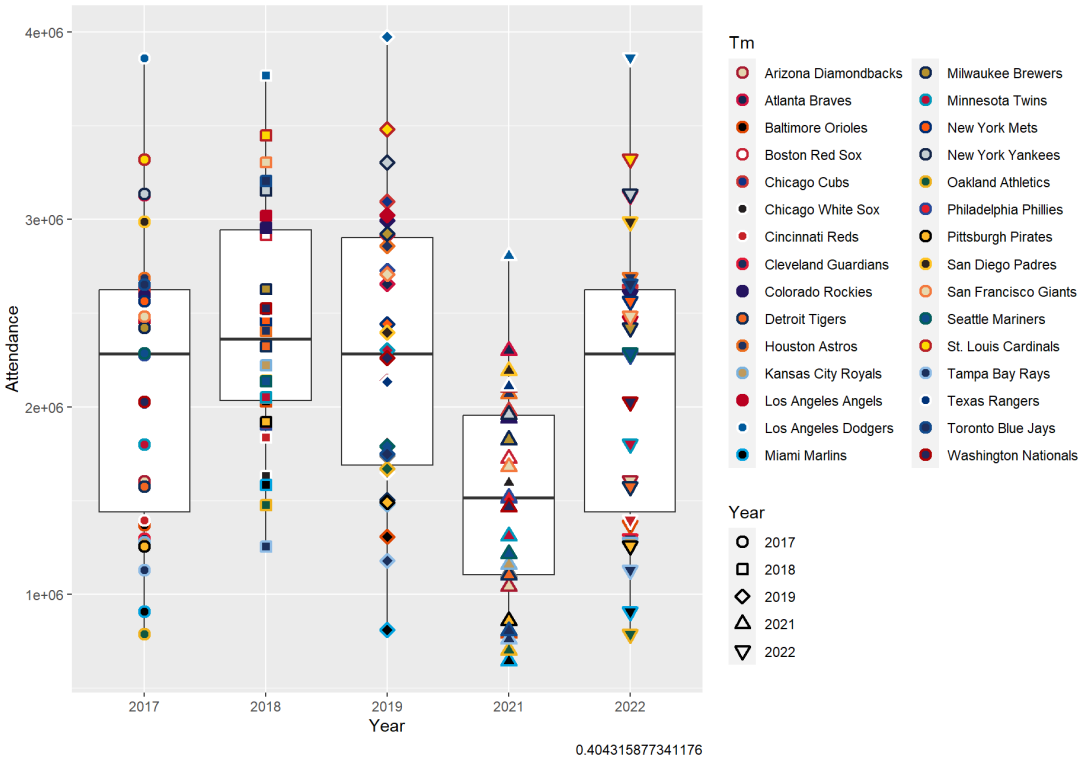


Figure 4b. Box plot of attendance to win from 2017 to 2022

However, we can see a lot of up-pointing triangles representing 2021 data are lying under the trending line (Figure 4c). In figure 4b, it shows the mean attendance is similar, except 2021. The mean attendance is much lower than other years. Although the actual reason leading to this phenomenon cannot be confirmed from this graph, it is possible that this attendance is greatly affected by the pandemic.

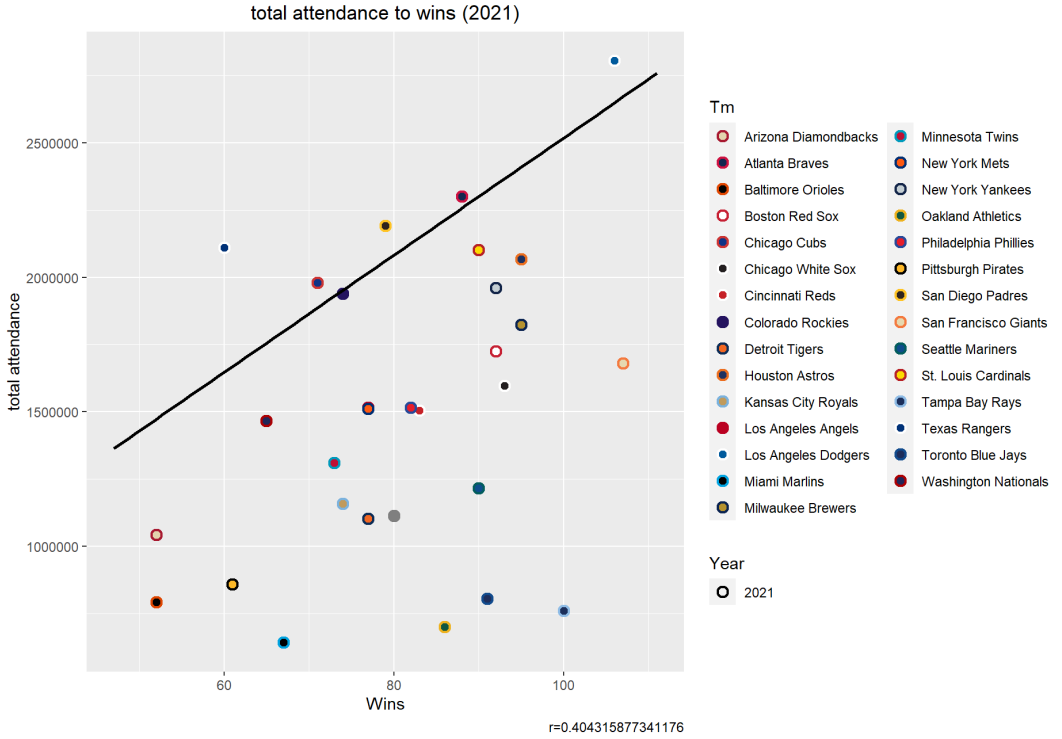


Figure 4c. Scatter plot of total attendance to win in 2021

Conclusion

It is not surprising that more runs and less runs allowed create more wins in a baseball game. In this report, we observe that home runs have a higher impact on scoring and being opposite for stealing bases. This observation also matches with the trends of MLB. Moreover, although the correlation between wins and attendance is weak, a low attendance is found in the 2021 season, which may be brought by the pandemic in 2021.