Analyzing Team Success in Major League Baseball

Marcus Spotanski, Christine Kunz, Sebastian Esquivel, Adam Einsel, Nick Walker

10/31/2022

Abstract

There are many factors that impact team success in Major League Baseball. In this data report, we will look to answer the question, How do different statistics impact team success in modern day Major League Baseball?. For the purpose of this report, we will define success as win percentage and the Modern Era of baseball is defined as 1970 to 2021 as this is when sac fly's were recorded as a stat(used in the calculation of On-Base-Percentage). These statistics will include some of the following, BA, OBP, OPS, R (definitions explained later in Section 2). This will be done by analyzing successful teams' statistics and comparing them to unsuccessful teams' statistics to then draw conclusions on how teams can improve their success. The result of the data is that managers and general managers should aim to use their budgets wisely and purchase players that are well rounded.

1 Introduction

Baseball is a sport that is extremely statistically driven and motivated. Statistician Seymour Siwoff stated this about baseball, "Anybody with a pencil could be a statistician back then." Baseball has been a hot seat for statistics since the sport began back in 1876. The statistics of baseball can be broken down into many different categories, including 'BA', 'OBP', 'OPS', 'R', 'RA'. These different types of stats can be documented and attached to their respective teams. All of these statistics will fall under the criteria to analyze team success.

Team Success can be defined in a few different ways. For this project we will focus on the win percentage of the teams as the win percentage allows us to compare the teams for the entire period. If we were to measure it by total wins, teams such as the Rockies who were created in 1991 would have less wins than everyone else.

For the purpose of this report we have only looked at a few important stats in baseball. These include Win Percentage(win_pct), On-Base-Percentage(OBP), strikeouts(SO), Runs(R), Hits(H), and Batting Average(BA). Obviously, Win Percentage is the variable that we are equating to team success and is self explanatory. OBP is the percentage of time a batter reaches safely. This is a very big stat in today's game as the more times you can reach base safely, the better you will be. SO are also a very prevalent stat in today's game. This stat is when a batter receives three strikes in an at-bat and is called out. This is something that baseball experts have always claimed that is very important to the success of any team. R is the amount of runs a team scored in a season. This is something that will most likely be highly correlated to team success as the more runs you score, the more games you will win. H and BA are heavily connected. H is when a player reaches safely via a batted ball. BA is the percentage of at-bats that end in a hit. On-base Plus Slugging is self explanatory. This stat is useful because it give a much more holistic look at a hitter. We selected these stats because they are all heavily assumed to be large influences on team success.

Table 1: Table shows what stats we are using

Variables	Definitions
win_pct	Winning percentage
OBP	On-base-percentage
SO	Total number of strikeouts in a season
R	Total amount of runs a team scores in a season
H	Total number of hits in a season
BA	Batting average
OPS	On-base Plus Slugging
RA	Runs Allowed
ERA	Earned Run Average

1.1 Motivation

Sports have a large impact on our society culturally as well as economically. In 2019, the College World Series recorded an 88.3 million dollar economic impact and created 1,103 year round jobs according to cwsomaha.com [1]. They also reported that it created 8.7 million dollars in tax revenue. It is safe to say that sports, baseball in particular, is important to the health of local economies. With that being said, putting a good product on the field is a huge determinant of how well the team does financially, thus putting a team on the field that is successful is very important. This is why we want to look at evaluating teams and trying to find what makes a team successful. If teams can identify any patterns in data that would help them find an edge on other teams, it helps both the team and the surrounding area financially. Sports are also a big part of our culture. Baseball is one of the oldest professional sports in the world. It has played a role in some of the biggest events in our countries history. From Jackie Robinson breaking the color barrier to the first game played in New York after the tragedy of of September 11th, baseball has always been intertwined with US history.

1.2 Research Question

How do different statistics impact team success in modern day Major League Baseball? Being able to find what stats impact team success is very important to everyone associated with the sport. From General Managers of the sport so they can make data driven decisions about what players to scout, to fans that want to brag to their friends that their team is better, these stats are important to understand to those around the game. If you can find which stats are important to overall team success, you can also get a better idea of how to game plan as a manager. For example, the league average batting average has decreased the past couple of years. This is largely attributed to people in the sport seeing it as a meaningless stat. Managers see that batting average is less important, and thus they don't see a need to play players that have a high batting average. This is the benefit of this question. We want to see which stats impact team success so people can make decisions based information, not basic baseball intuition.

1.3 Structure of Paper

First we will describe and explain the data in section 2. This will also include how we obtained and will explain the manipulation of the data. Then we will explain the methods used to analyze the data and describe how they work in section 3. The purpose of section 3 is to explain why we chose the plots we did and how to interpret them. Then we will explain our findings from each of these visualizations. The last section, section 5, includes the conclusion. This is where we we explain what the findings mean and their other implications.

2 Data

Baseball is very data driven. Understanding the definition of each statistic is very important to this research project. For example, understanding Batting Average (BA) is important because one might reasonable assume that it would be highly correlated with team success as the better a team is at getting hits, the better the team would be at offense, which would result in more wins. We will explore how this may not always be the case but it is important to understand the data so we can make some predictions.

The Database we will be using is from the Lahman's Baseball Database created by Sean Lahman [2]. Sean Lahman is a journalist for the USA Today Network. However, before joining the USA Today Network, however, he worked as a sports reporter at the New York Sun. He also works a lot with Sabermetrics which is defined as the search for objective knowledge about baseball by Bill James. He is currently the data projects manager for the Society of American Baseball Research. This dataset that we will be using contains stats between the years 1871 and 2021. However, we will most likely only use stats from the 'modern era' of baseball which is defined as 1970 to 2021. The file provided by the Lahman Database has many databases so we will narrow down the focus to just the teams dataset.

2.1 Obtaining the data

The Teams csv only had what are called counting stats (H, BB, etc.) that are accumulated throughout the season. We used these to create the stats such as Batting Average and On-Base Percentage which are also commonly used. The calculations are shown in Table 2. To calculate Batting Average you divide the number of Hits(H) by At-Bats(AB). FOr OBP you divide the sum of hits, walks(BB), and hit-by-pitch (HBP) by the sum of all of those things and sacrifice fly(SF). In order to calculate slugging percentage(SLG), we had to calculate singles(X1B). We did this by subtracting doubles(X2B), triples(X3B), and homeruns(HR) from hits. Slugging percentage(SLG) is a weighted average of batting average where a double is worth 2 hits and a triple is worth 3 hits and a home run is worth 4 hits. And finally, to find win percentage(win_pct) we divide the wins by games played(G). We also wanted to clean the data by dropping some of the columns that are meaningless.This made working with the data much easier. We also limited the years from 1970 to 2021 so we can observe the modern day.

Table 2: Table shows how the stats we used are calculated

Calculated Variable	Formula
BA	H/AB
OBP	(H+BB+HBP)/(AB+BB+HBP+SF)
X1B (Single)	H-X2B-X3B-HR
SLG	(X1B+2xX2B+3xX3B+4xHR)/AB
OPS	OBP+SLG
win_pct	W/G

3 Methods

We decided to create three different types of plots. A scatter plot, a time series plot of a stat and a teams win percentage, and a time series plot that shows the top 8 teams vs the bottom 8 teams. Each of these will help us determine which stats answer the research question. In each plot, we will explain the purpose of each plot and how it connects back to the research question.

3.1 Scatter Plot

The first plots we will show are scatter plots. The purpose our scatter plots is to show the correlation of any given stat and win percentage. To make the correlation more clear, we will include a table that shows

the correlation value of the stats we are going to analyze. Stats with correlations that are farther away from zero will be the ones that have high impacts on team success. This is because correlation is a measurement of how strong 2 variables are related. If the 2 variables are perfectly related, the correlation will be either 1 or -1 depending on if they are positively related or inversely related. With this being real life data, no variables are going to be close to 1 or -1. Baseball is a complex sport. No one stat will be a perfect match with team success. However, if we can find a couple of stats that have relatively high correlations, we can start to get a good idea of what stats collectively have a high impact on team success.

3.2 Time Series

The next plots we will create are a time series graph that shows a certain teams win percentage by year as well as a certain stat on the year. In order for the data to work, we needed to standardize the data so that the graph would be much more representative of the data. Standardizing the data means putting the data on the same scale where you can compare different variables easier. The purpose of using the time series plot is to see how certain stats change over time compared to a team's winning percentage. If a certain stat has a big outcome on team success, then when the team does well, that certain stat should either be very close to the win percentage line if it is directly related or far away from the win percentage line if it is inversely related. The line should also be consistent. If a team does well one year and struggles the next year, but a stat remains constant, the stat clearly did not impact the team's success.

3.3 Top 8 vs Bottom 8 Teams Analysis

In order to determine what factors affect winning in the Modern Era of Baseball, we wanted to look at what stats separate the good teams from the bad teams. This would allow us to determine what statistical fields affect win percentage other than finding the correlation factor between win percentage and the other fields. To do this, we made two groups of historical data comprising of the best and worst 8 teams in terms of wins in the Modern Era of Baseball. Then, we wanted to determine the specific statistical fields that have the widest margin of difference between these two groups. While this portion of the research is not yet completed, to do this, we will find the average amount of each statistic over each year for both the bottom and top 8 teams in the MLB before standardizing each field. Now once we take the difference between the standardized average values of each statistical field between the top and bottom 8 teams, we would be able to compare each field to the other and choose the N largest stat differences between the two sets. However, this leaves one final item to determine: do you want more or less output for those N statistical fields if the goal is to be successful?

To answer this, we can construct graphs of the yearly averages of each statistical field to determine whether winning teams do better in a given field than losing teams. Here, we would visually see the yearly difference in output for each statistical field between the two groups. We can use the two data-sets from the previous calculations that comprised of the yearly statistical averages for the top and bottom 8 teams in the Modern Era of Baseball and plot the year-to-year data using a bar plot.

4 Findings

4.1 Scatter Plot Findings

Table 3 shows the values of the the highest correlated variables to win percentage.

Table 3: Tables show the highest values of correlation between the variable and win percentage

	Value		Value
OBP	0.5159132	ERA	-0.5428891
OPS	0.4801749	RA	-0.4330161
SLG	0.4189152	ER	-0.4100221
BA	0.3850277	BBA	-0.3356151
BB	0.3129291	HA	-0.2293271
HR	0.2908663	E	-0.2201510

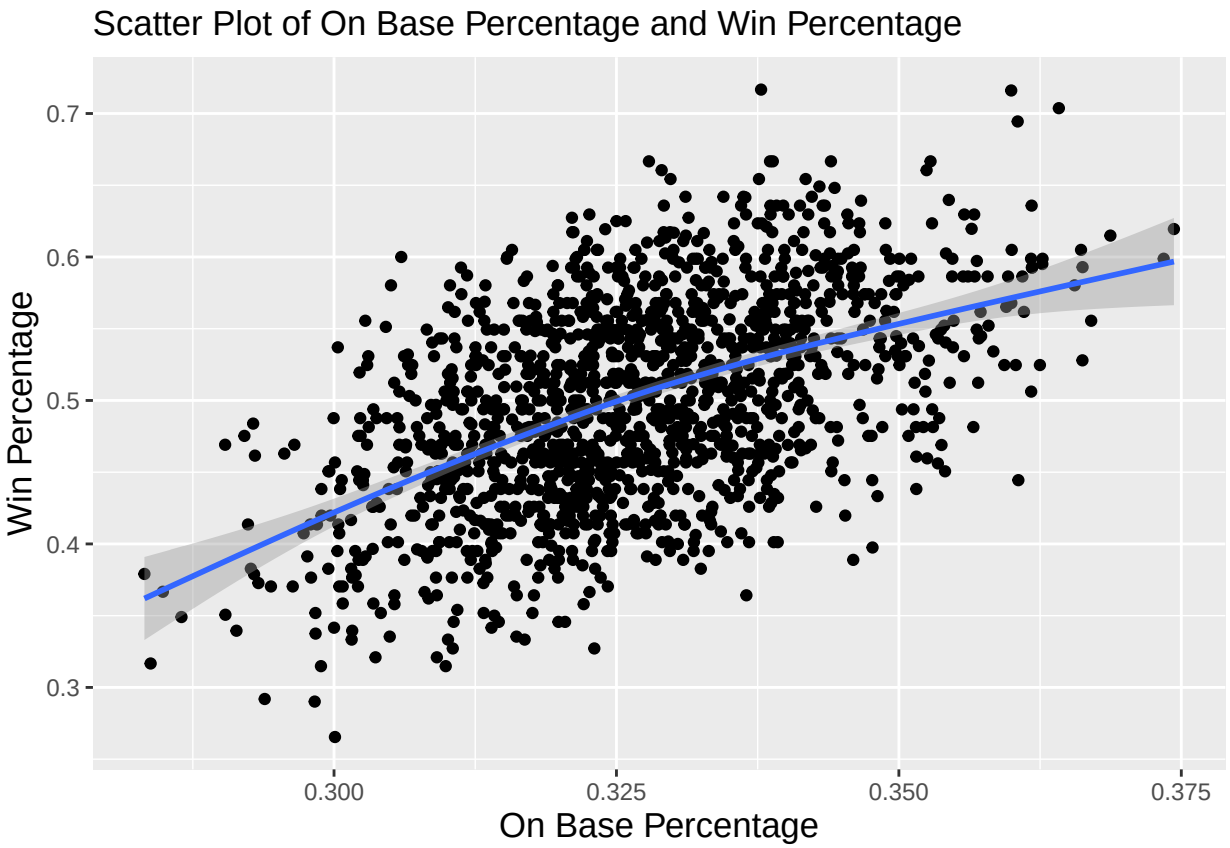

Figure 1: Top 6 positively Correlated Stats in order

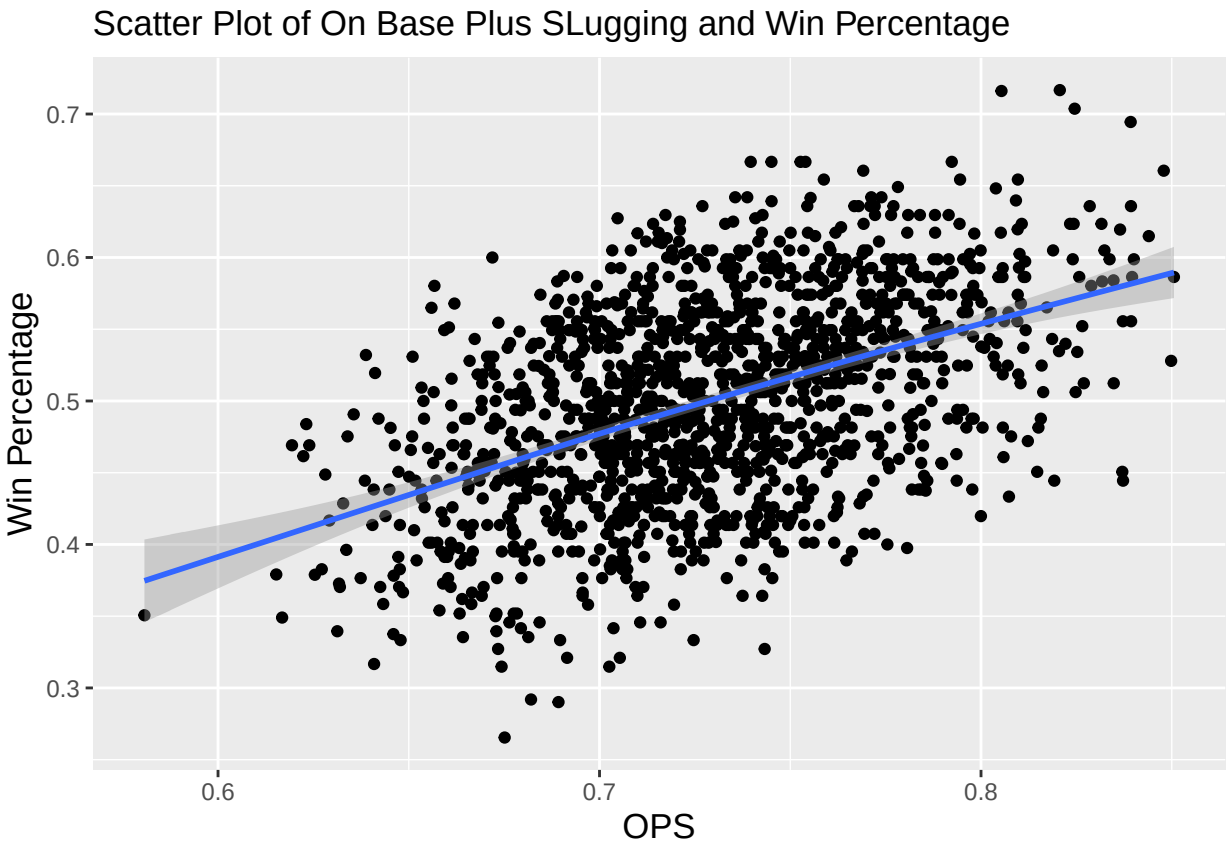

Figure 2: Top 6 positively Correlated Stats in order

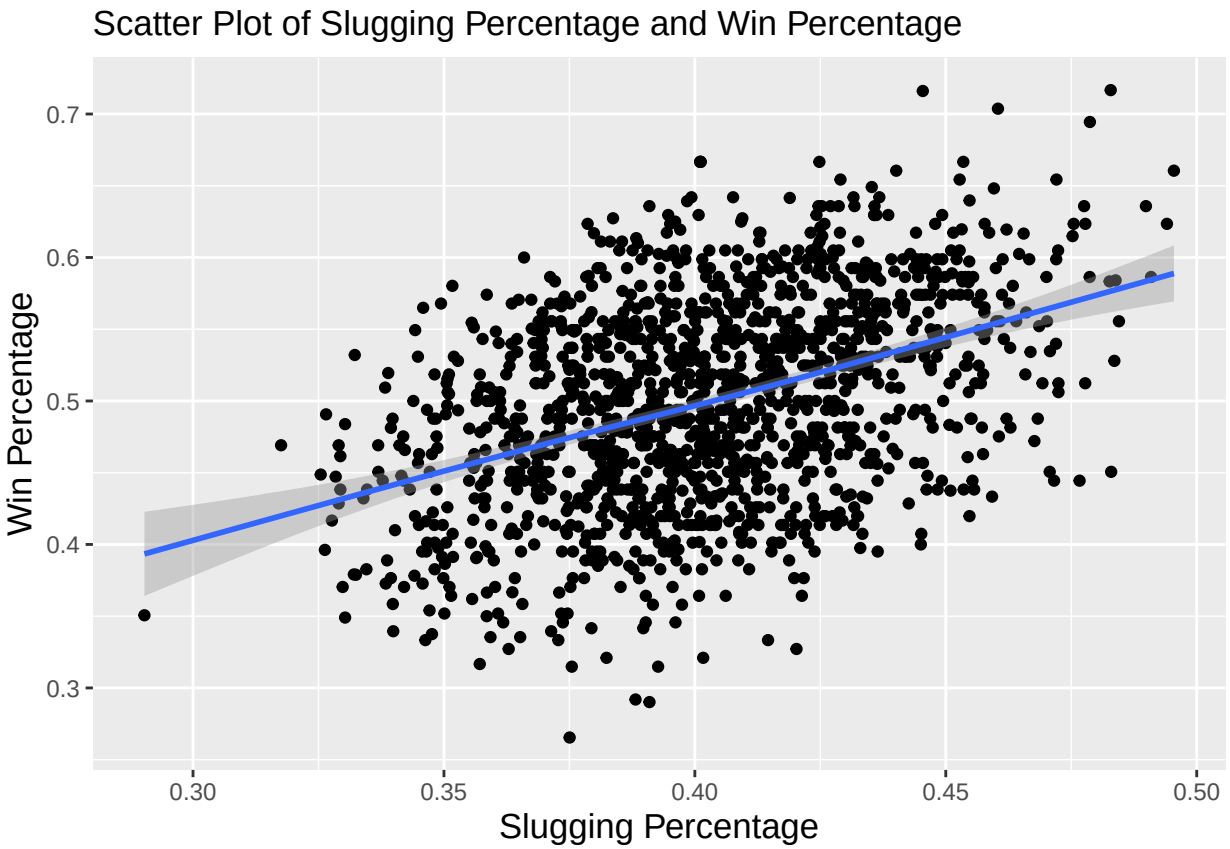

Figure 3: Top 6 positively Correlated Stats in order

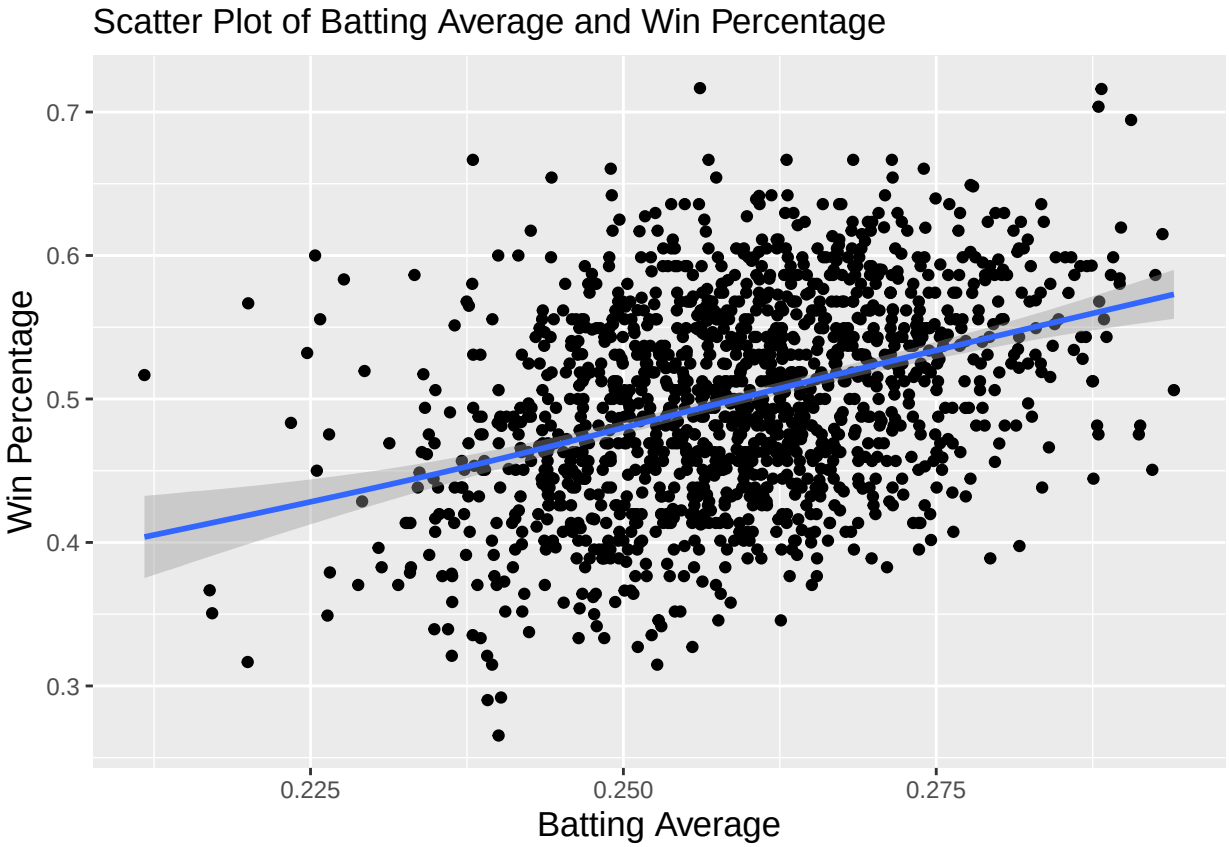

Figure 4: Top 6 positively Correlated Stats in order

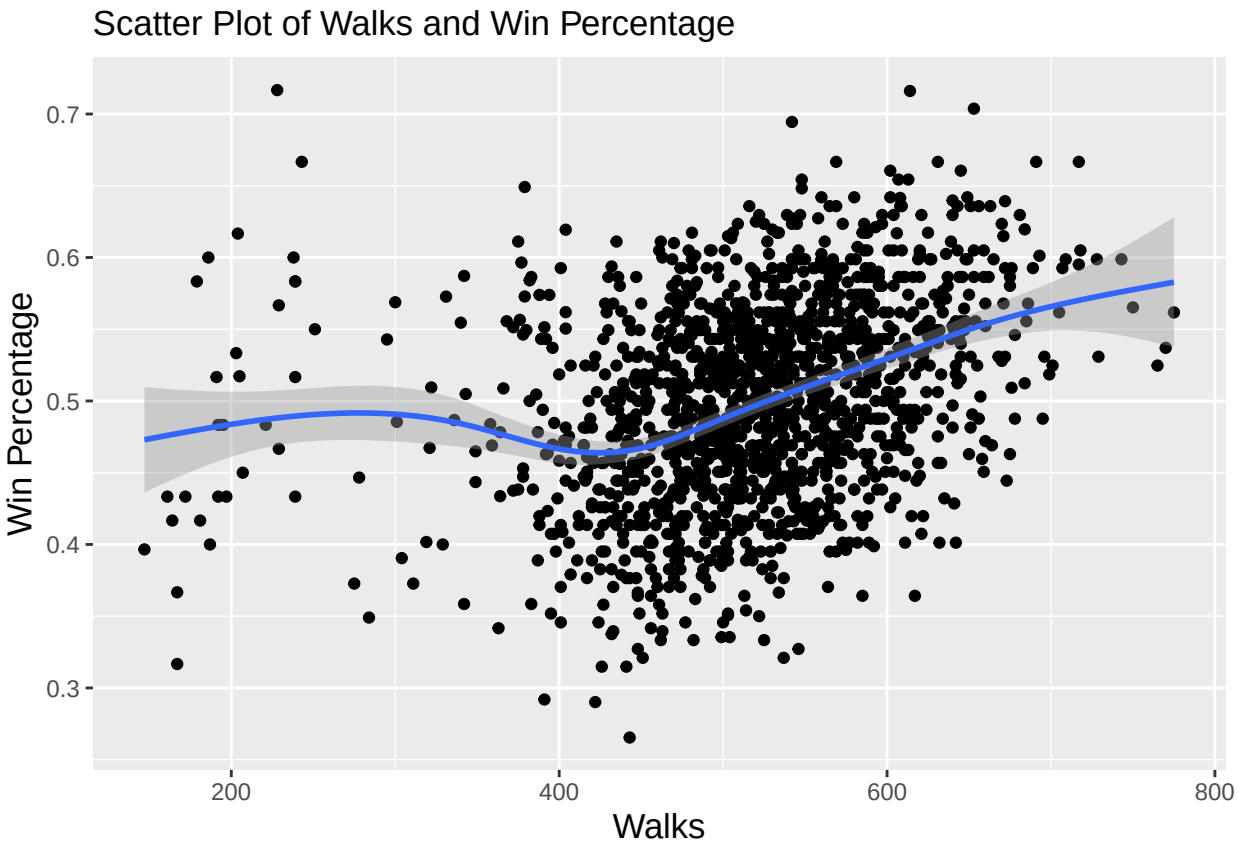

Figure 5: Top 6 positively Correlated Stats in order

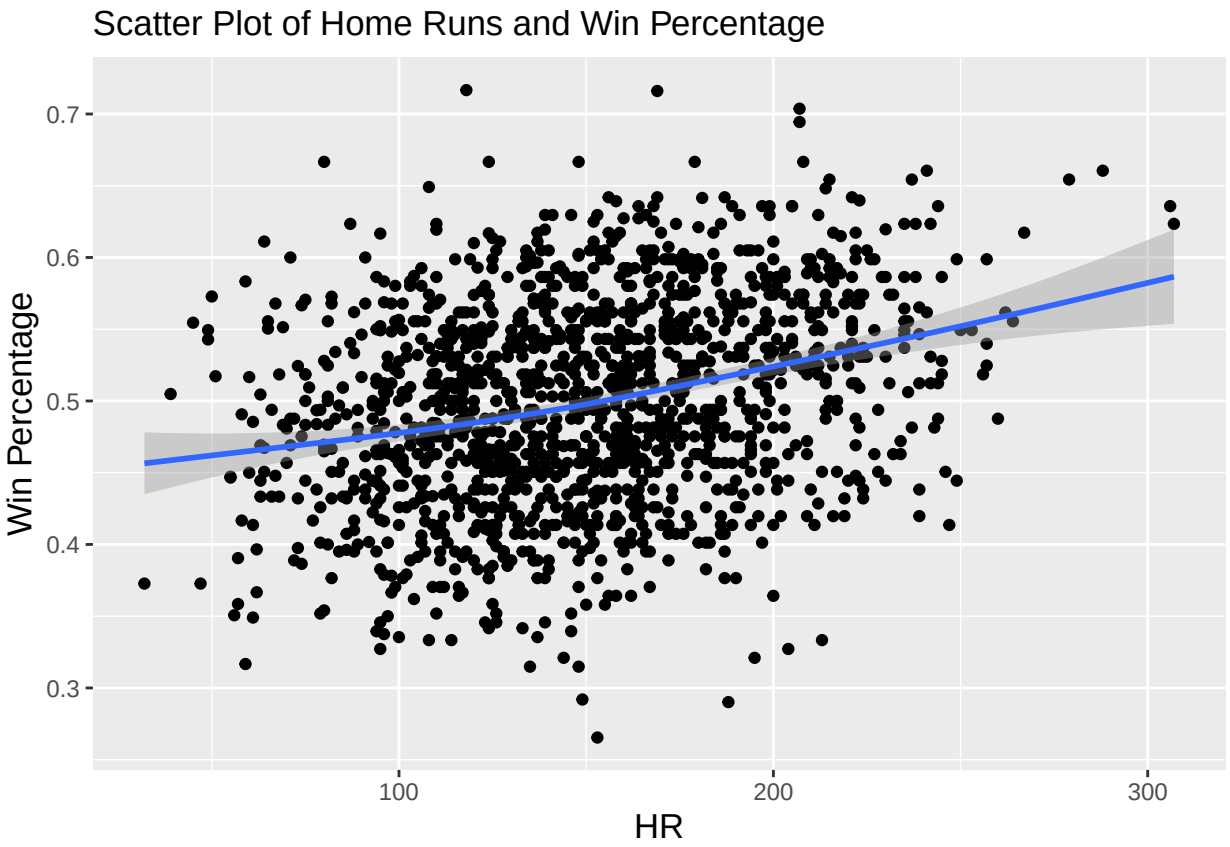

Figure 6: Top 6 positively Correlated Stats in order

Looking at each of these plots, it is very clear that there is some correlation between each of these stats and win percentage. One thing that is particularly important to note is that the top 4 stats in terms of correlation according to Table 3 are all 'rate' stats, or average. This means that stats that are counting stats, such as walks, hits, etc. are not as great at determining team success. This may be due to the fact that their are a couple of outlier seasons where there were few games played. This topic is explored more in 4.3. With this being said, so far it seems that these 6 stats are all stats that impact team success, but On-Base Percentage seems to be the best positively correlated stat so far. All of these stats make sense as to why they are positively correlated as they are all offensive stats. The higher the rate, the better the team is going to be. For example, the better the team is at getting on base, the more likely they are to succeed. As we will see in Figure 7, defensive stats should be negatively correlated for the same reason. The better the team is at not letting the other team score runs, the better the team will be overall.

This information is important to baseball managers general managers because it tells us how to create a successful ball club. When scouting players to acquire via trade or free agency, these are the stats to look at. Looking at the top three stats, OBP, OPS, and SLG, we can see a clear trend. OBP is a metric that evaluates a player/teams ability to reach base. Obviously, teams have always look to acquire players that can reach base often. However, the emphasis has been to reach base via hits. If we compare batting average and OBP, we can see that, while batting average is a good indicator of team success, it is inferior to OBP. This tells us that putting an emphasis on how players reach base may be overrated. SLG measures how effective a team is at getting hits. As we explained earlier, SLG is essentially a weighted batting average that puts emphasis on extra base hits. This tells general managers and managers that players that also get a lot of extra base hits are crucial in a successful team. So finding a metric that includes both OBP and SLG can give us a good idea as to what types of players a team should bring in. This is where OPS comes in. Explained in Table 2, we can see that OPS is OBP+SLG. While this metric is not perfect, it certainly

gives us a more complete view of a player.

Going back to OBP, another good stat that impacts team success is walks and home runs. Walks are one of the more surprising stats. One would think that walks would be low on the list of possible stats because the other team seems to be able to control how many walks they issue. However, it seems like getting walked, both as a team and as a player, is a skill rather than luck. Home runs on the other hand are not too much of a shock. Teams always are looking for players to hit the 'long ball,' and it seems like general managers and managers have the right idea in signing players that are capable in hitting home runs.

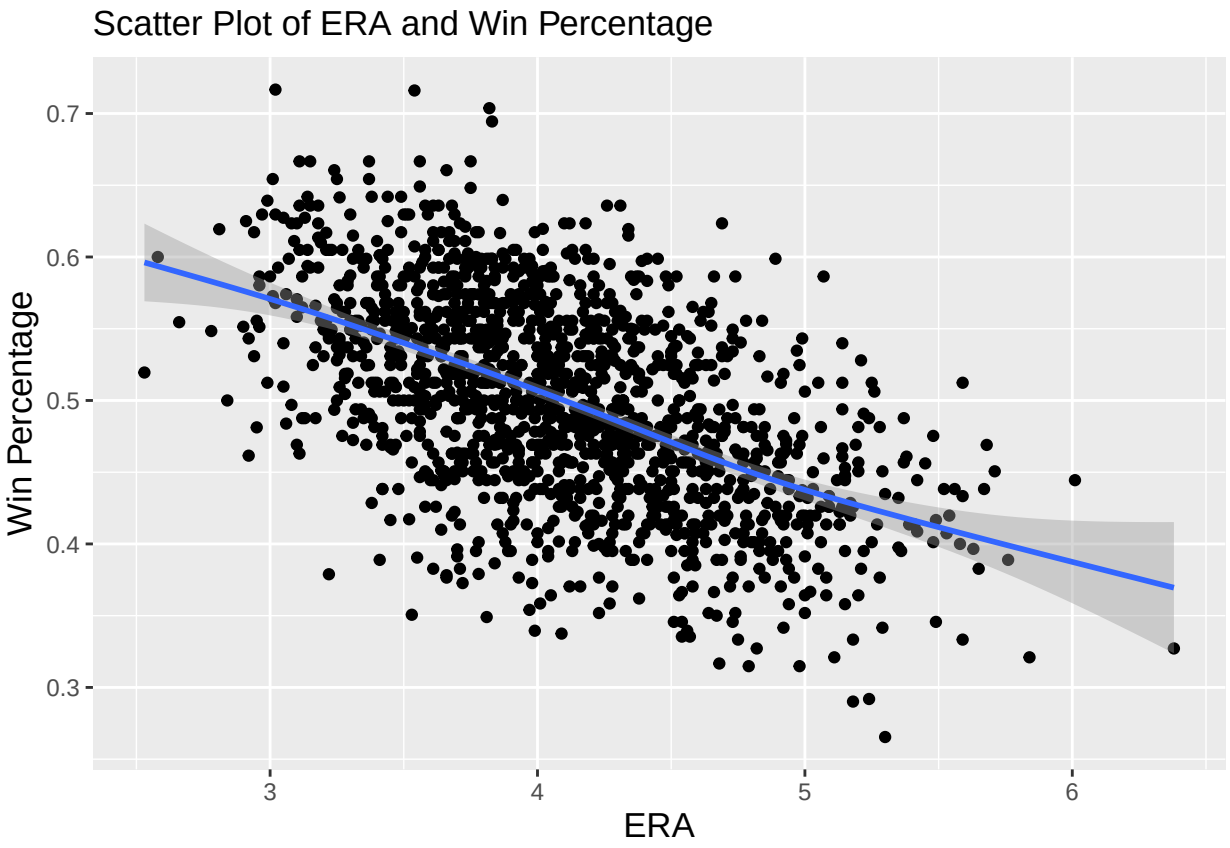

Figure 7: Top 6 negatively Correlated Stats in order.

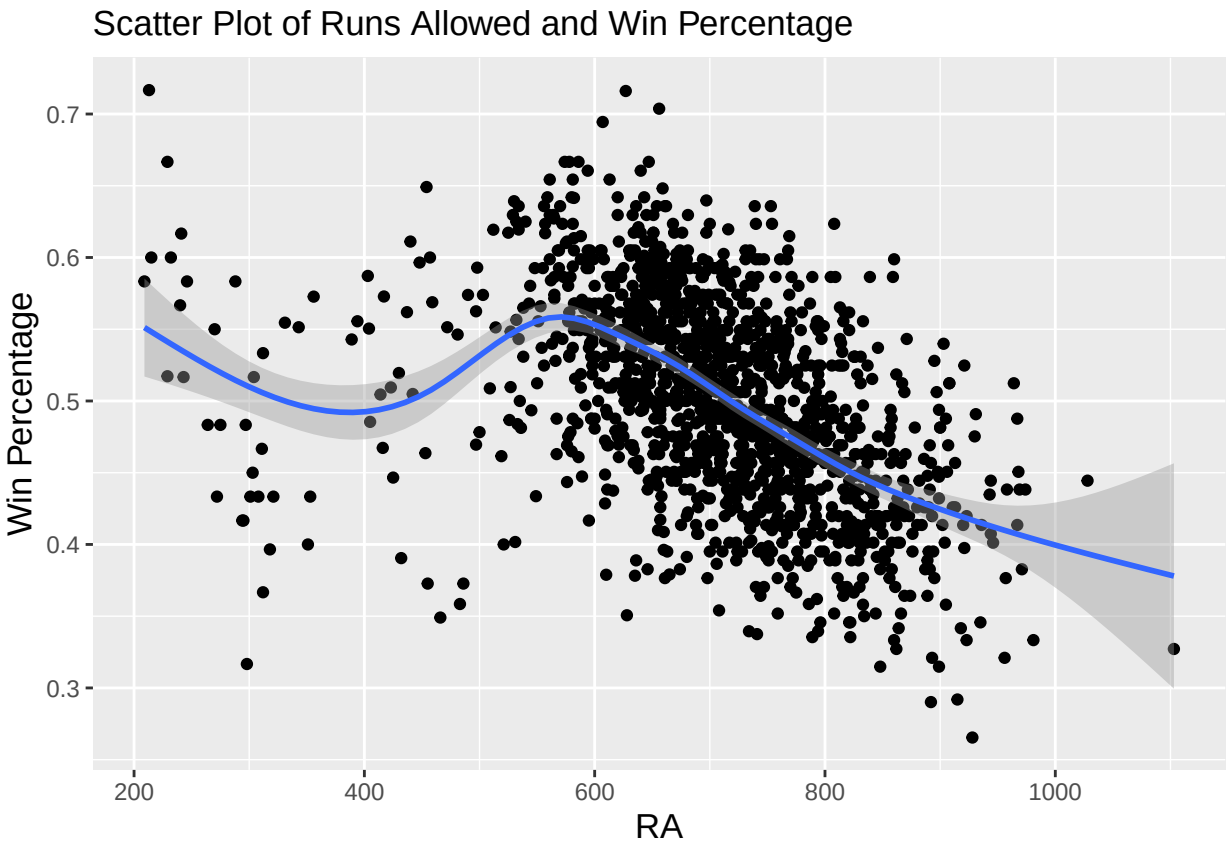

Figure 8: Top 6 negatively Correlated Stats in order.

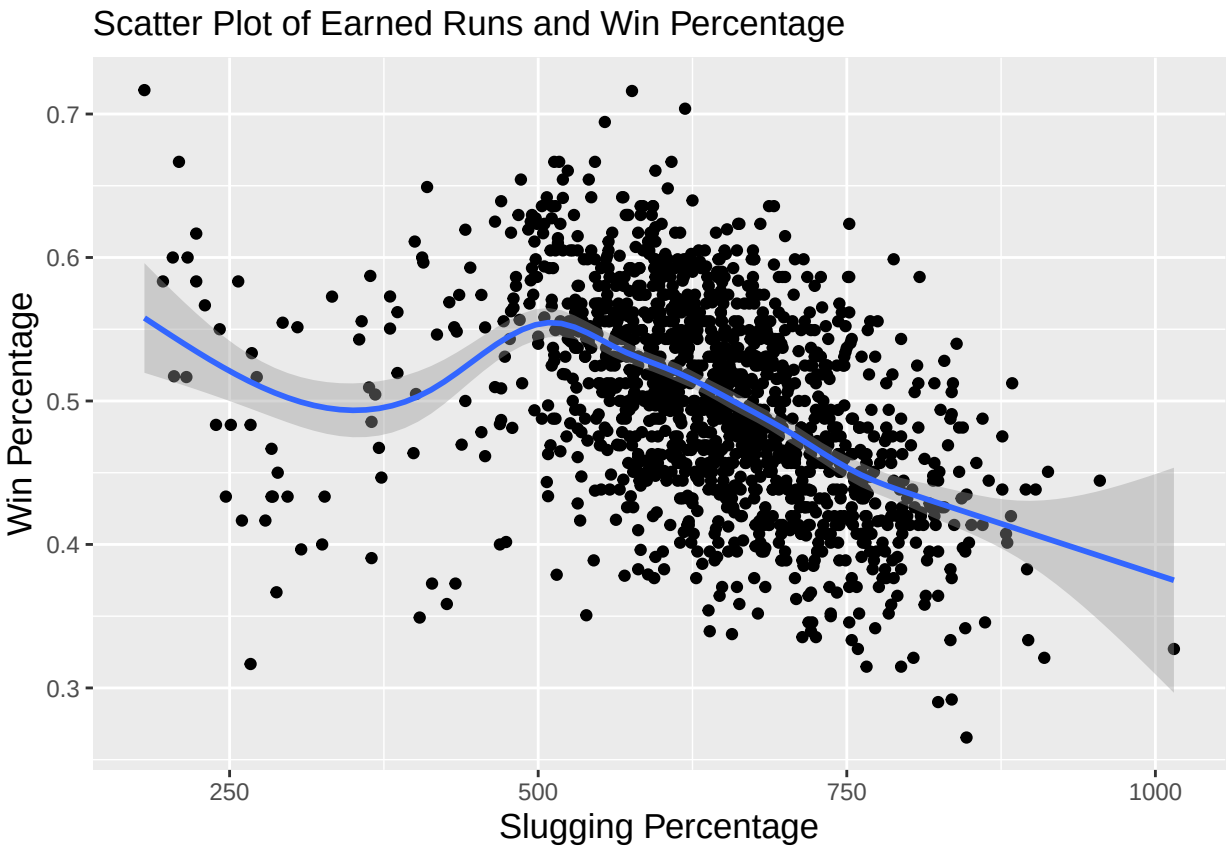

Figure 9: Top 6 negatively Correlated Stats in order.

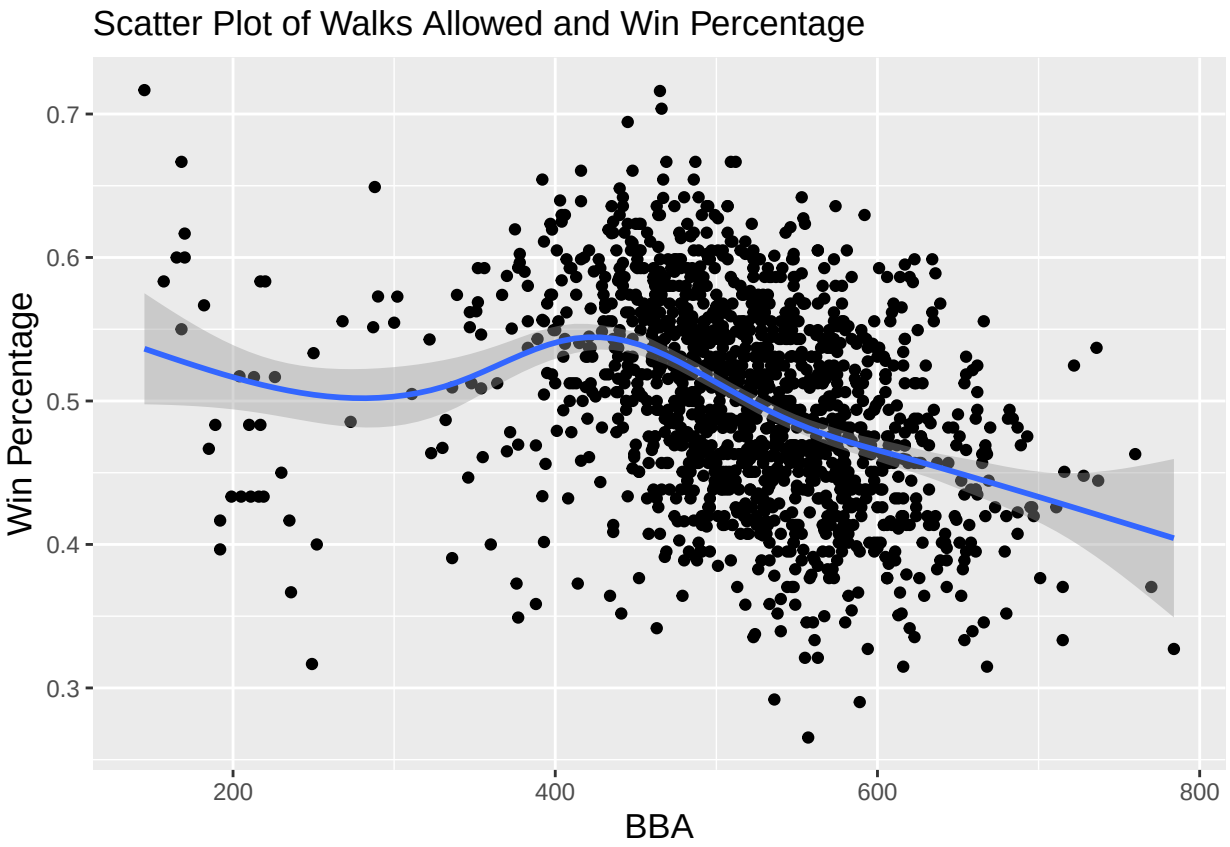

Figure 10: Top 6 negatively Correlated Stats in order.

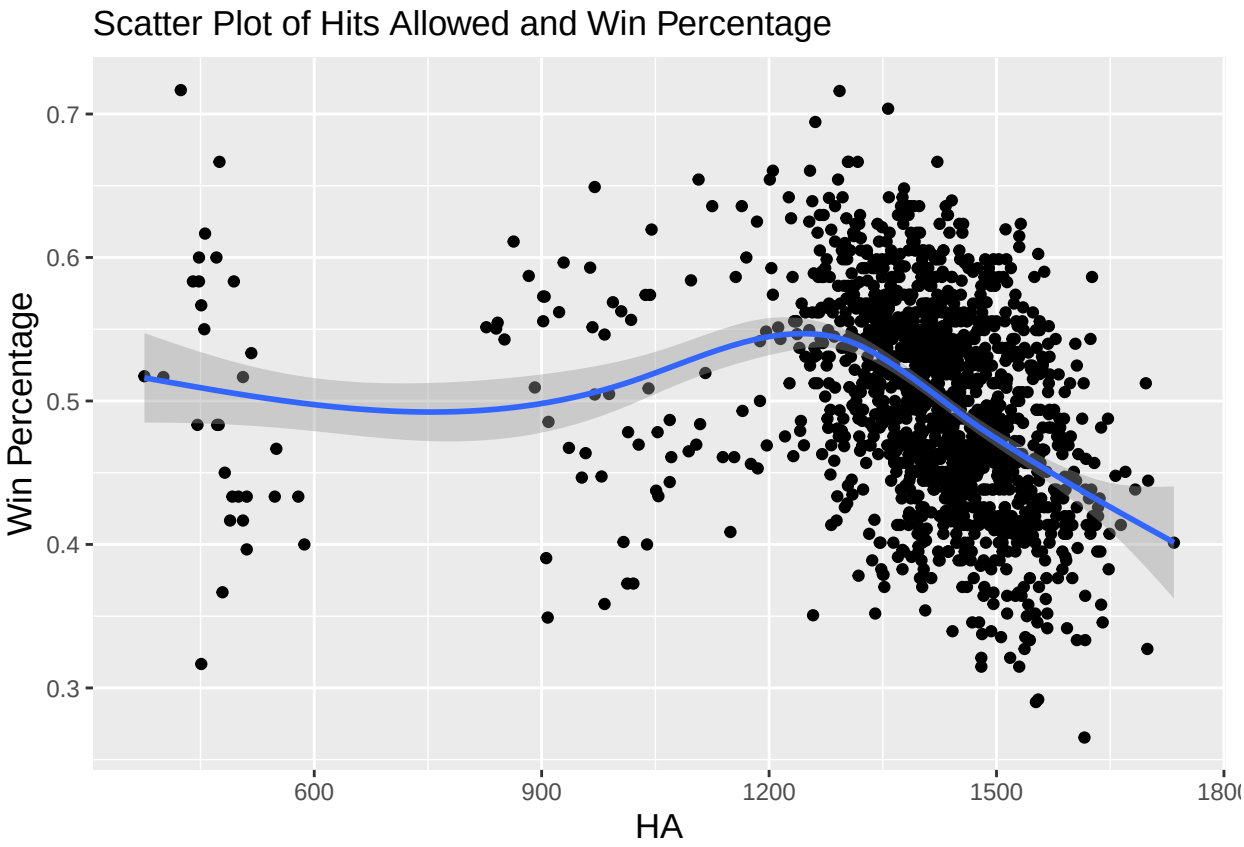

Figure 11: Top 6 negatively Correlated Stats in order.

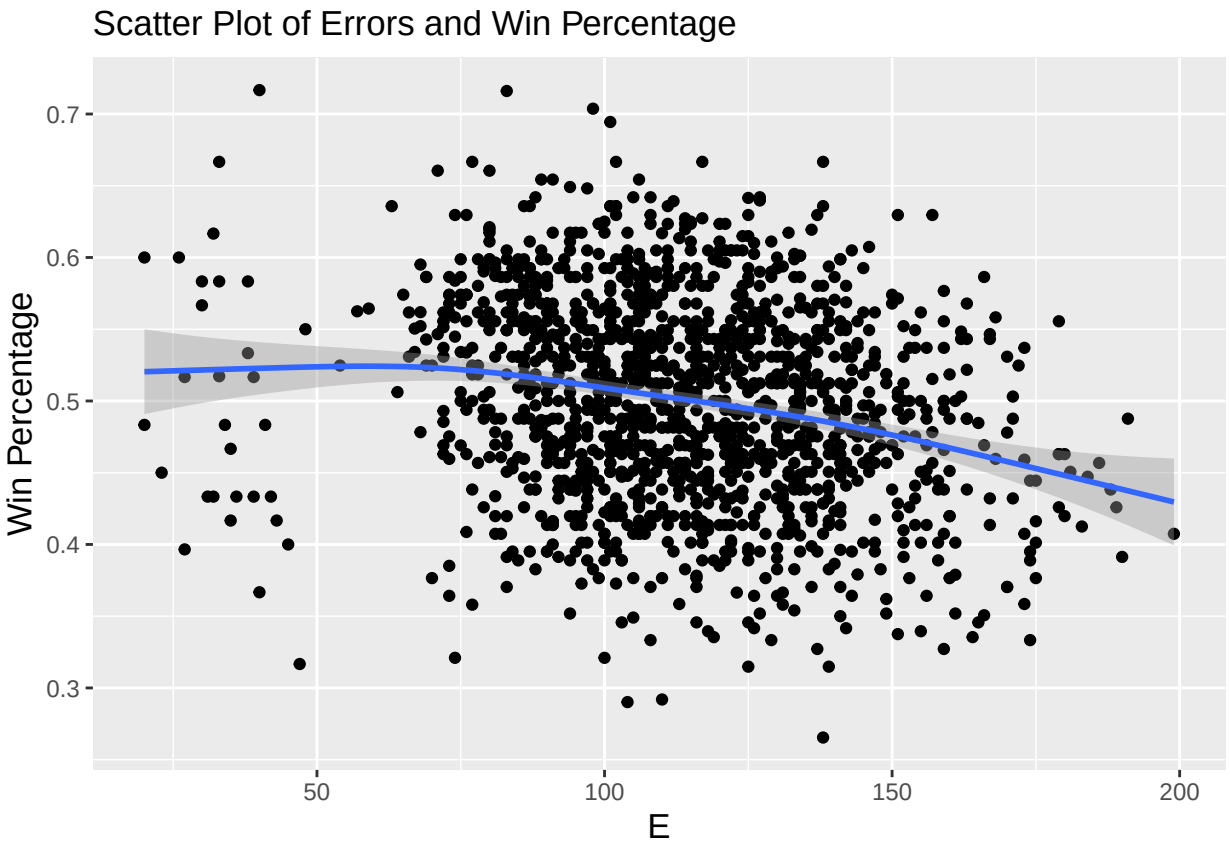

Figure 12: Top 6 negatively Correlated Stats in order.

Again, stats that measure teams rates, or averages, seem to be better at determining team success than counting stats. According to Table 3, ERA is not only the strongest negatively correlated stat, but it is also the strongest correlated stat overall. As we explained earlier, all of these stats are negatively correlated because they are defensive stats. The better the team is at limiting the runs scored by the other team or limiting the amount of errors they commit, the better the team is going to be.

From a managers point of view, these plots seem to make a lot of sense and does not provide a lot of insights into any certain tricks as to who to scout. Looking at the number one stat from Table 3 we have ERA. ERA is a stat that has long been used and is the predominant stat to measure pitchers. This stat makes a lot of sense because the better a team is at limiting the amount of runs they allow, the better they will be. We can also see that allowing as few hits and as few walks is also a good trait for a pitcher to have. This means that scouting players that have low ERAs, low walks allowed, and low hits allowed is important. That is the main takeaway from this. It is very important to have a well rounded team. Having a team that is only good at offense will not cut it. It is imperative that defense is taken in to consideration for general managers and manager. If faced with the option of signing or playing a player that is only good at offense or a player that is well rounded, it is best to choose the more well rounded player according to the information provided.

Now we are going to look at some stats that might be seen as overrated. Strikeouts is a stat that many people think is very important to look at. This is because the less you strikeout, the more times the ball is in play, which means more hits, which as we've shown, more wins. However, this seems that this is not the case. According to Table 3 strikeouts have one of the lowest correlated values. It is negatively correlated, which makes sense, however, it is still very close to 0. A possible explanation is that defense in Major League Baseball is so good that just putting balls in play is not good enough. The ball must be hit very well, and thus the goal of eliminating strikeouts as a team still results in weakly hit balls that are still outs. Another

stat that is very weakly correlated that may not be obvious is stolen bases. It is positively correlated which makes sense as it is an offensive stat, but it still has one of the lowest values. This stat in the more recent years has become less valued, however, many people still believe it to be a valuable asset to a team. However, it seems to have very little impact on team success. The graphs of these two stats are shown below in Figure 13 and Figure 14.

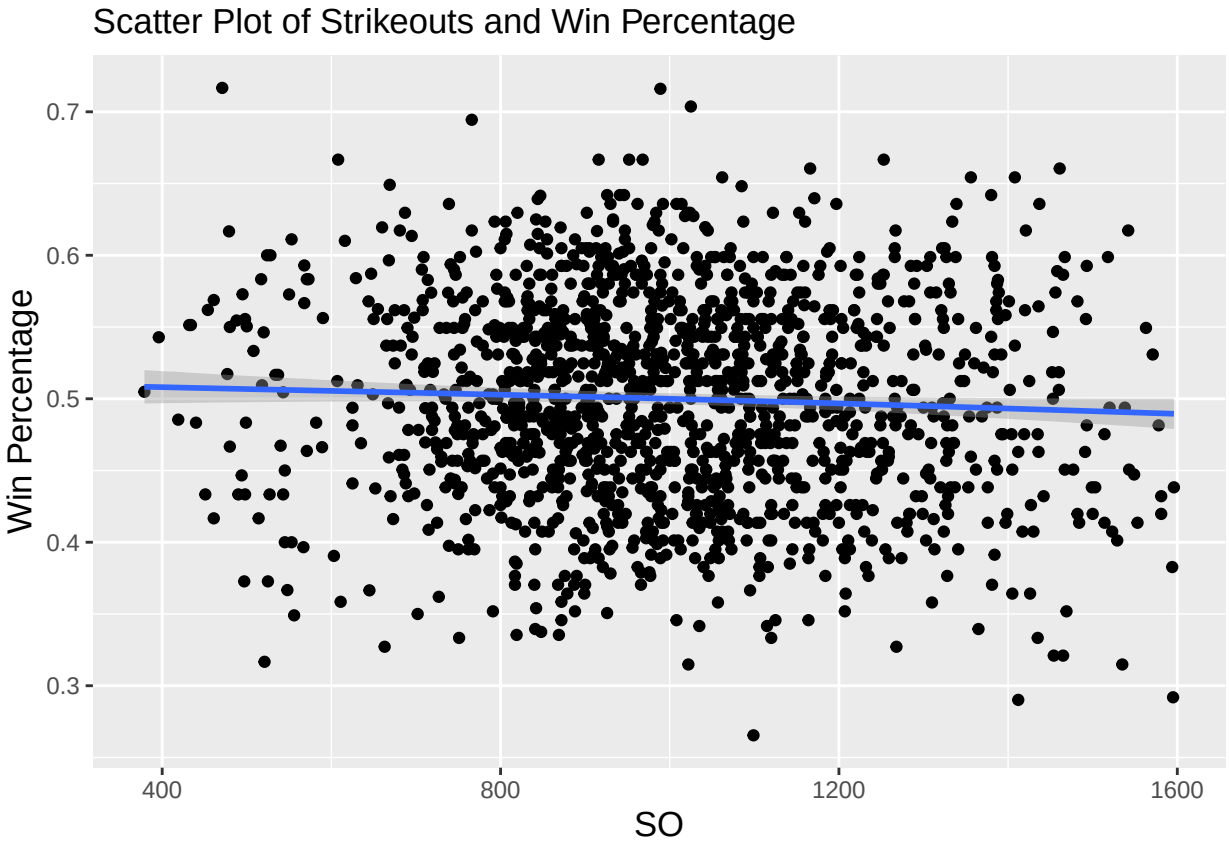

Figure 13: Scatter Plot of Strikeout and Stolen Bases.

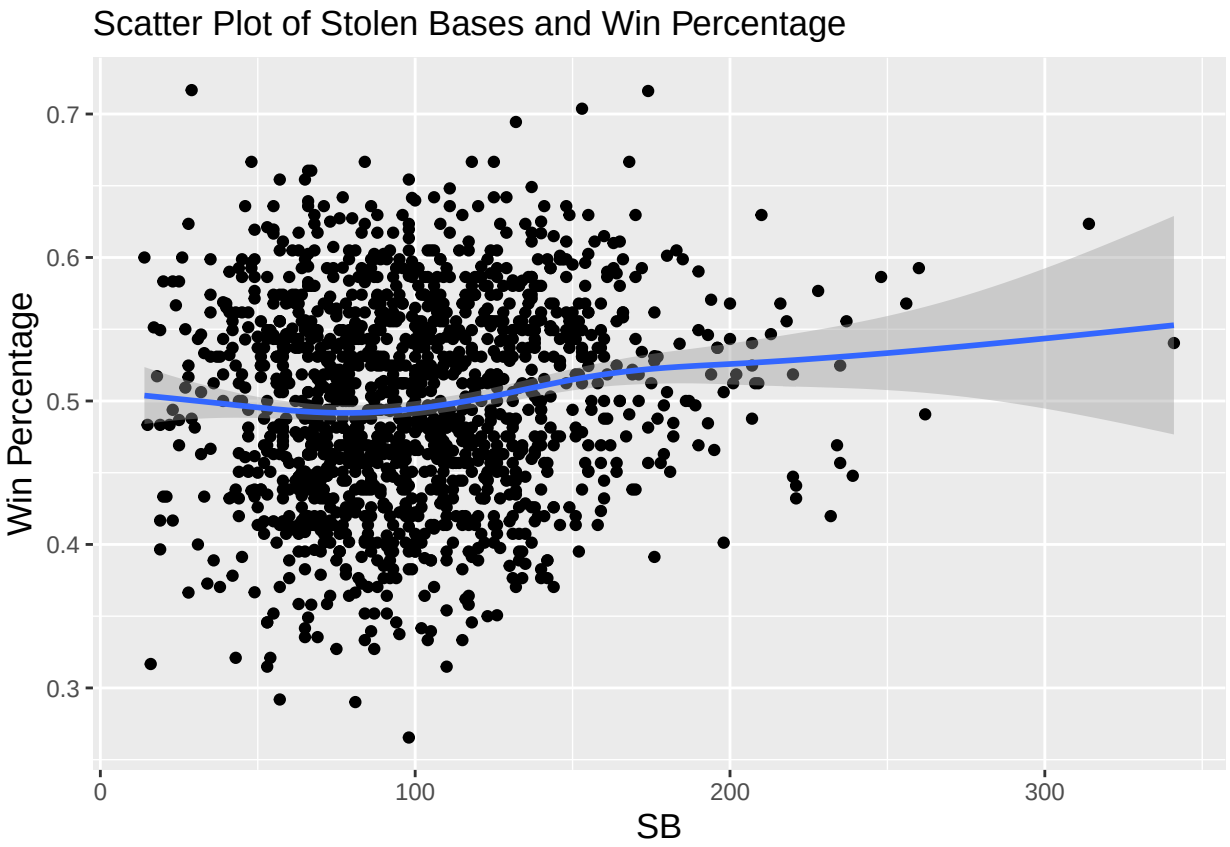

Figure 14: Scatter Plot of Strikeout and Stolen Bases.

From these visualizations, it is important for general managers and managers to put less of an emphasis on these stats. Teams are starting to understand this and are continuing to sign and play players that have high strikeout counts. This is not a bad thing. According to the data, strikeouts are one of the worst factors in team success.

4.2 Time Series Plot

For the time series plot we will choose a team that has had both very poor seasons and very successful seasons throughout the data set, the Atlanta Braves. We will use their win percentage and compare it to the standardized stats and see which ones follow the win percentage line closely.

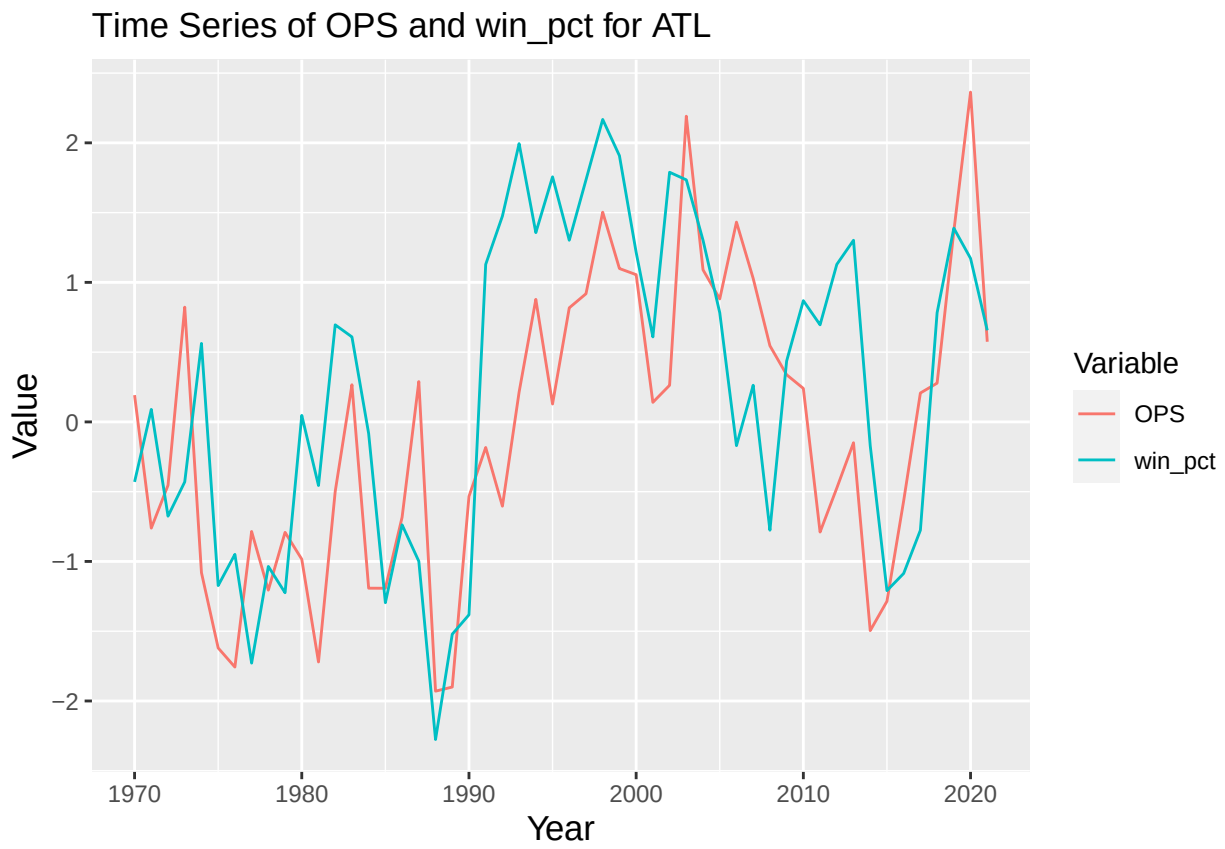

Figure 15: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

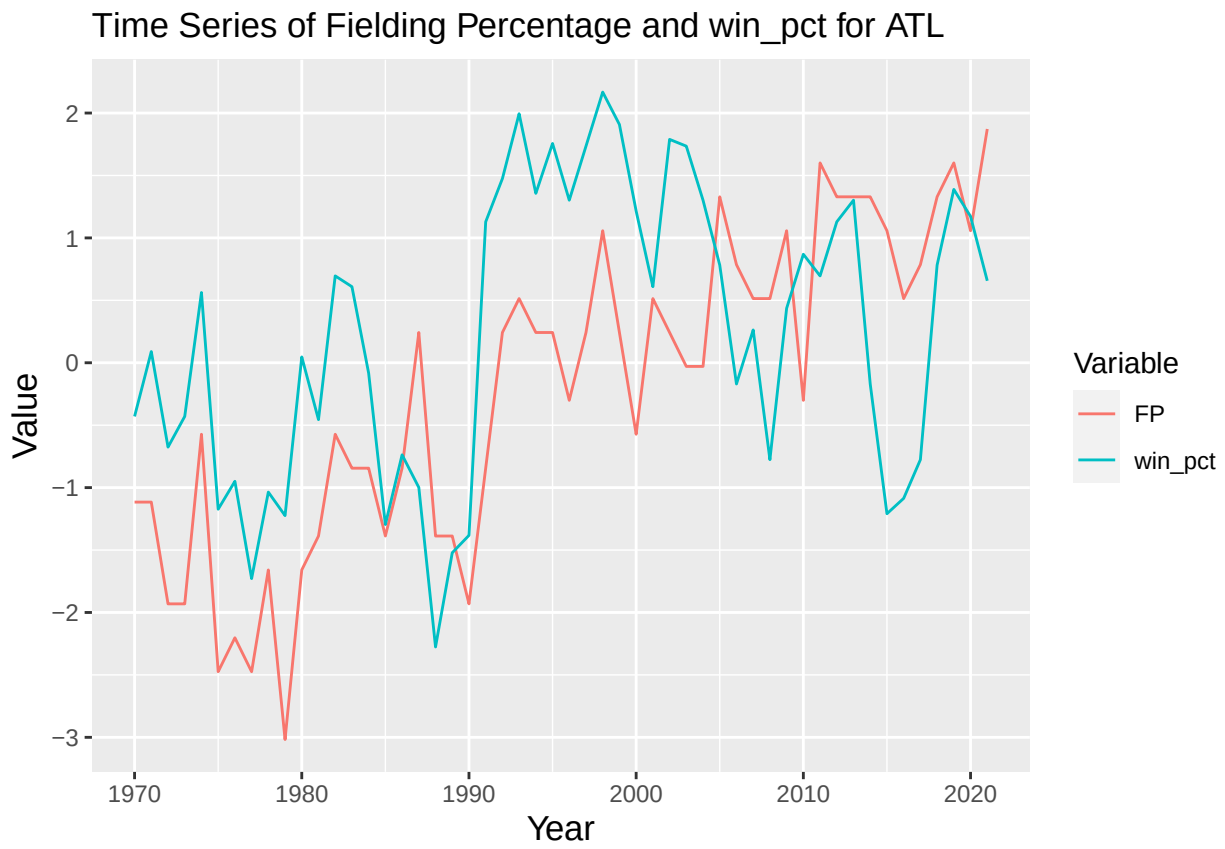

Figure 16: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

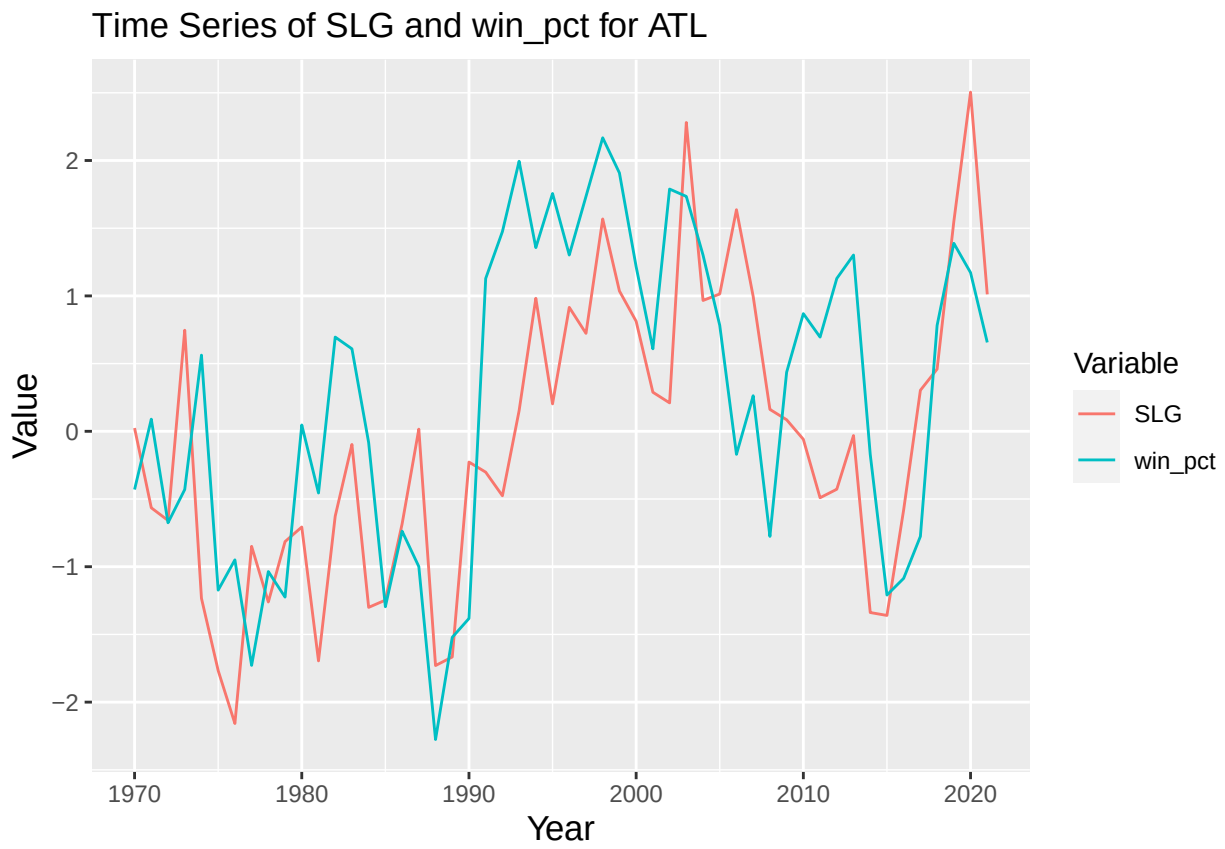

Figure 17: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

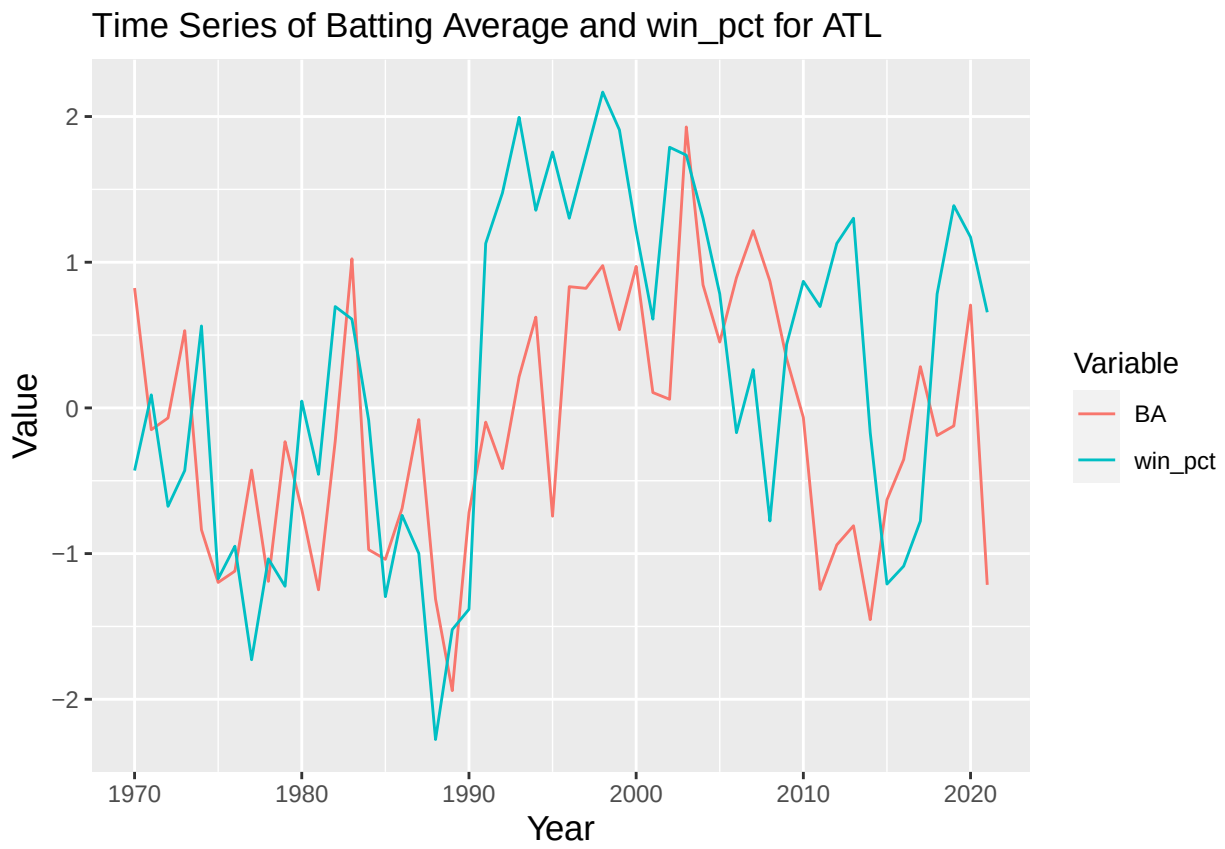

Figure 18: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

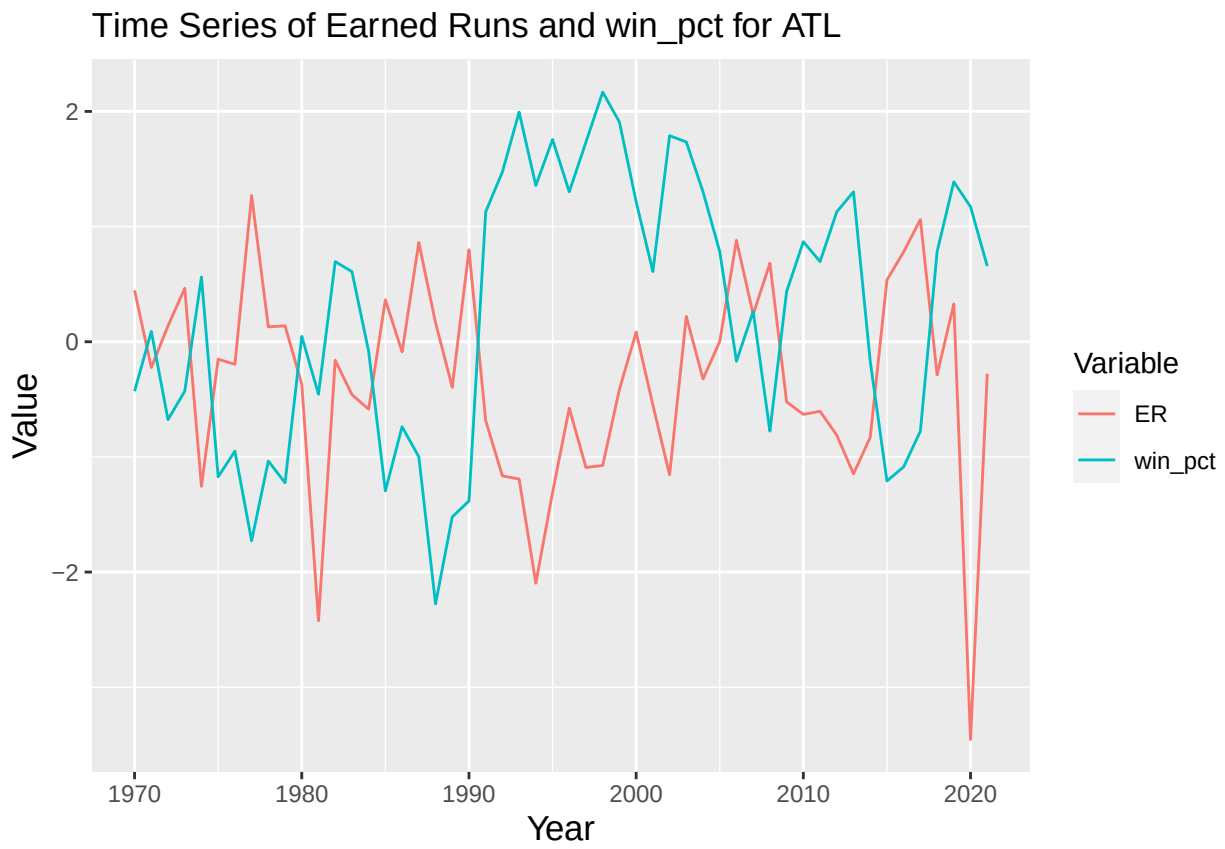

Figure 19: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

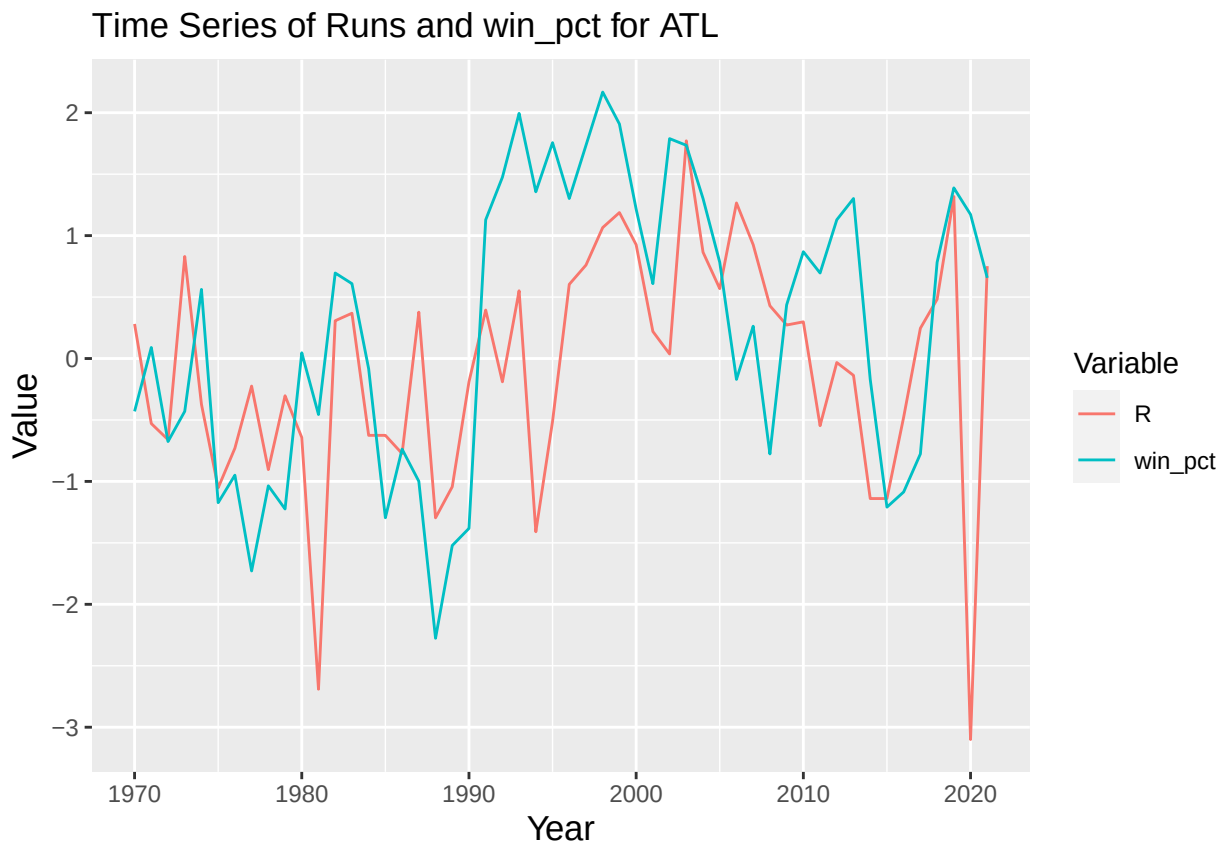

Figure 20: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

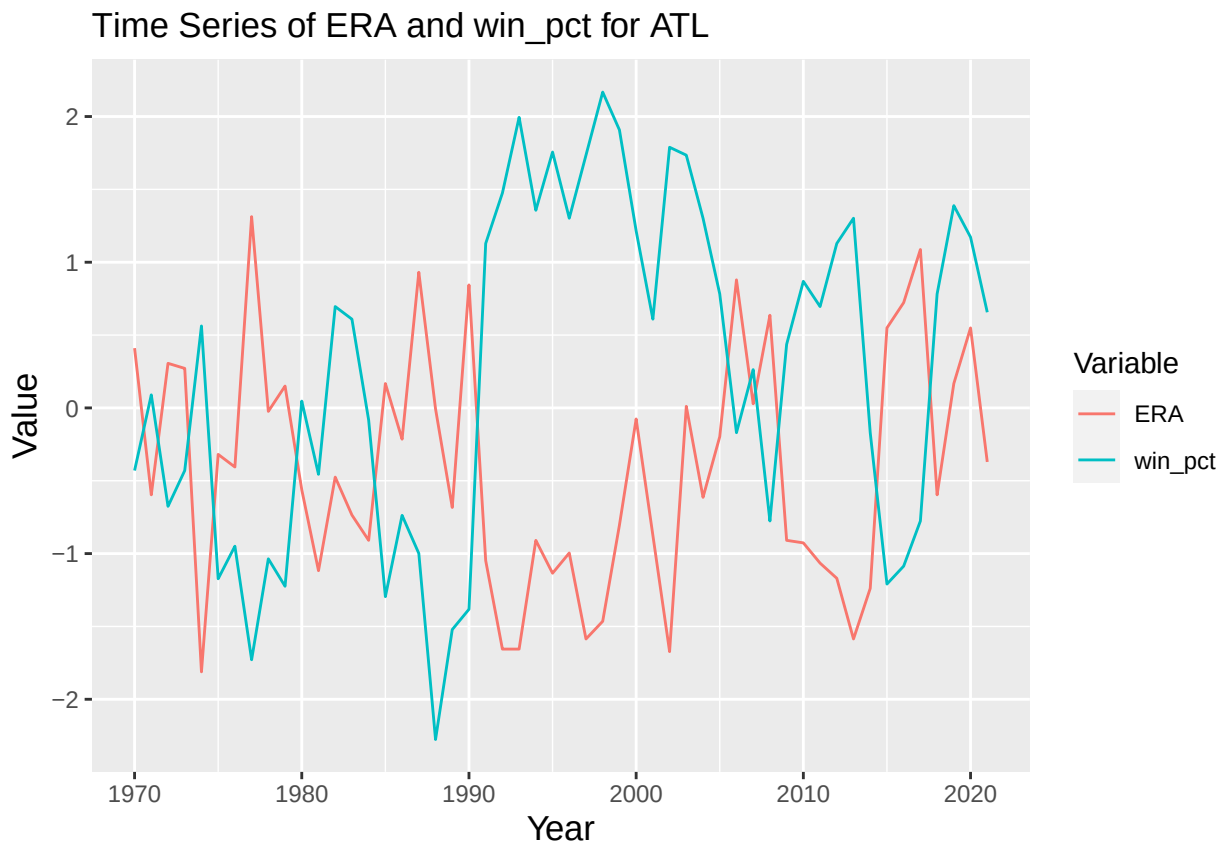

Figure 21: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

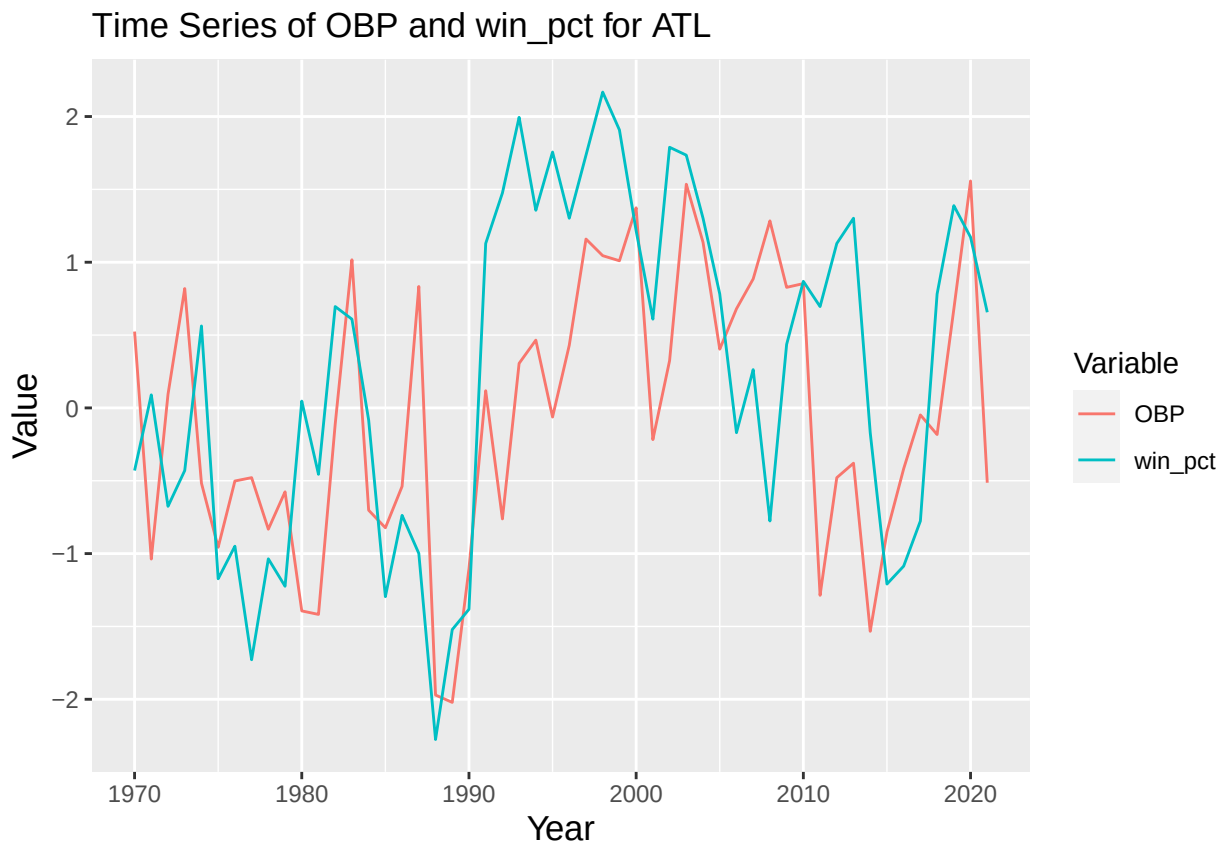

Figure 22: Time Series plot of OPS and Win Percentage for Atlanta. Although there are some parts where the line of OPS deviates from Win Percentage, it stays relativly close.

As you can see many of these stats, while not perfect, either follow the win percentage line if it is an offensive stat or are far apart from it if it is a negative stat. There are 2 that really standout from this graph, OPS and ERA. OPS does have some variation, but it seems to follow very closely. ERA seems to be the best for this plot because it is always far apart from the line. The key to this plot is not necessarily be either really close to the line or far away, but rather to be consistent. For ERA, it is always far away from the graph, which is why it is a good fit. An example of a bad stat is strikeout again, as shown in Figure 23. To answer the research question, so far from section 4.1 and 4.2, it seems that rate stats such as OPS, SLG, OBP, BA, and ERA are stats that have a strong impact on team success.

From a managers point of view, these graphs have great insights. All of these back up the claims made in section 4.1. The big stat that stands out is ERA. Going back to the claims made in 4.1, having good pitchers have a huge impact on team success. Atlanta from the early 1990's to the early 2000's were one of the most dominant teams looking at the graph. If you look at Figure 21, they also had a very dominant pitching staff. Other factors led to their success, however, having a pitching staff is crucial to building a team.

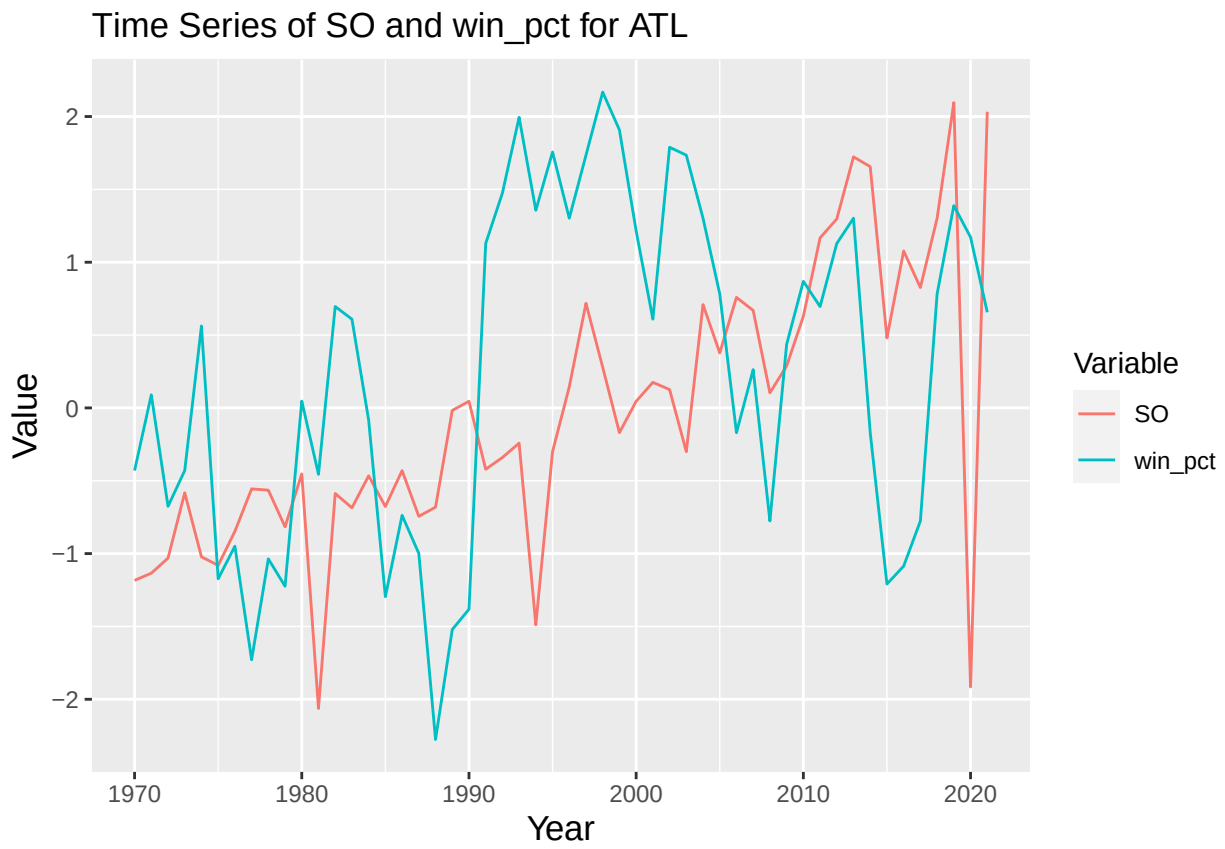

Figure 23: Time Series plot of SO and Win Percentage for Atlanta. This plot is much more random that the prior graph. This means that SO is not a very impactful stat on team success

Looking at the Figure 23, strikeouts again seem to have no pattern related to win percentage. This just reinforce the idea that strikeouts are irrelevant when trying to build a winning team.

4.3 Top vs. Bottom Plot

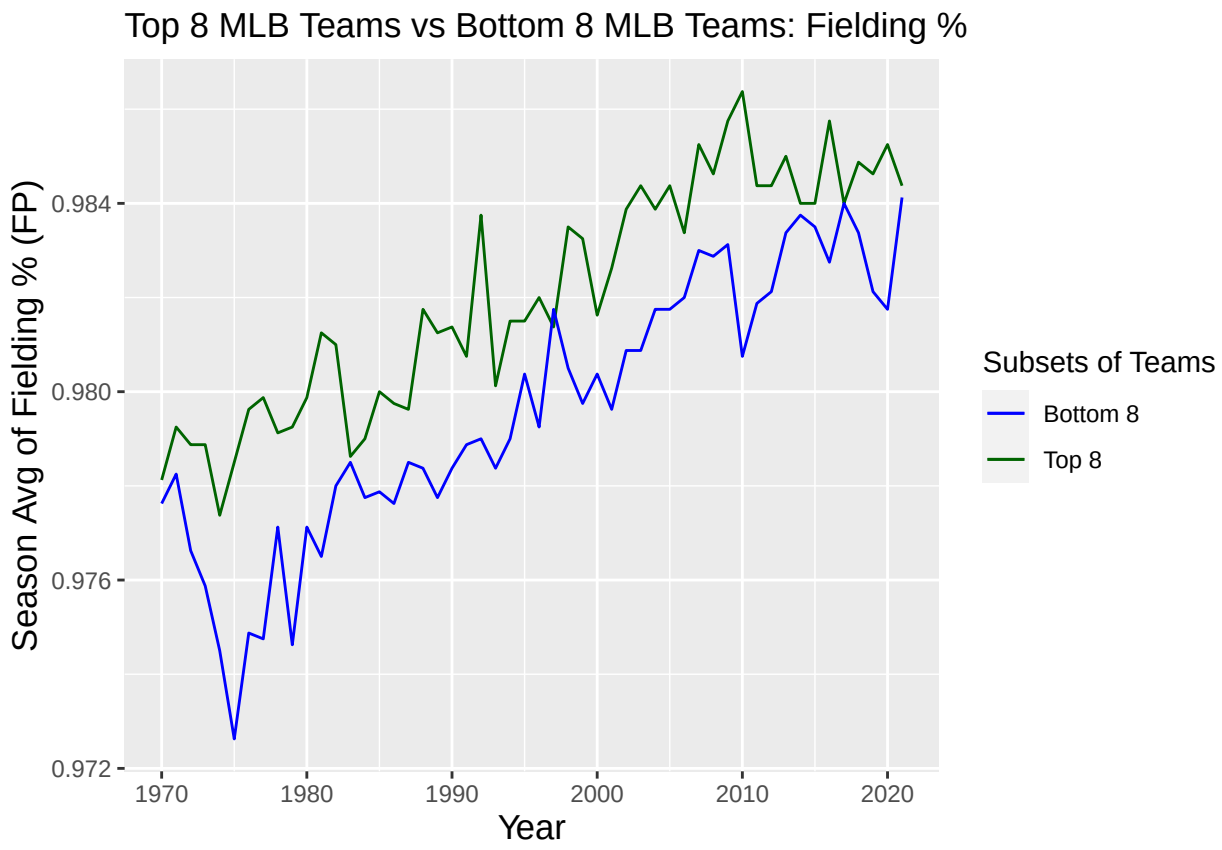

Figure 24: Top 8 vs Bottom 8: FP

Figure 24 shows that on average the Top Teams in the MLB have a better FP than the worst teams in the league. This makes sense as FP is associated with making the correct play to get a player out. Lower values in this field indicate more mistakes and errors occurring on the field, leading to more runs for the opposing team, and thus, more wins. This is a defensive stat that having a higher value is actually better than having a lower value.

From a manager's point of view, this is very important to look at. Fielding percentage is a topic that is hotly debated in the baseball world. There are many issues related to the effectiveness of the stat when evaluating teams and players. That discussion is could warrant its own topic, and very often does in the world of baseball sabermetrics. Regardless of its issues, we can see that it is a stat that separates the good teams from the poor teams. This again, supports the fact that having a well-rounded team is crucial to the success of a baseball club. Having a team that can field the baseball, is important if a manager wants to win baseball games.

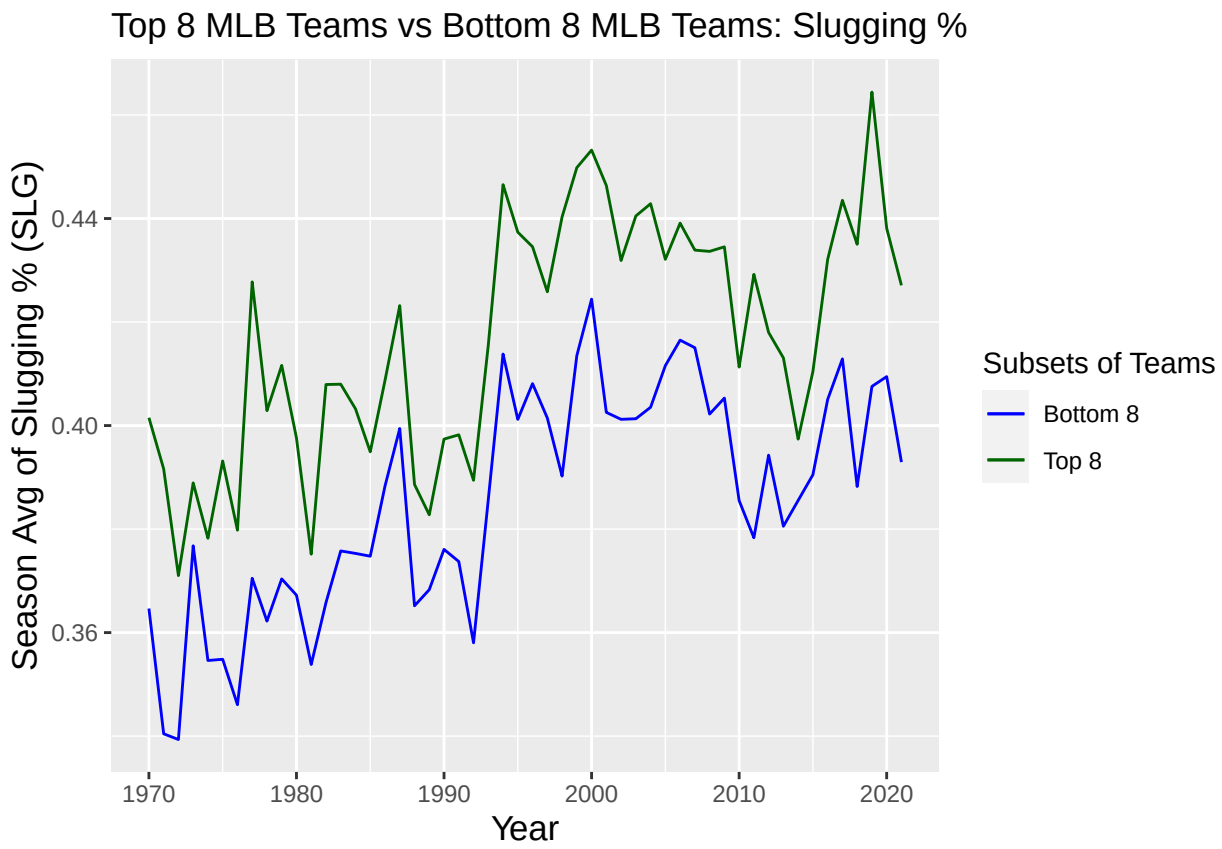

Figure 25: Top 8 vs Bottom 8: SLG

Figure 25 shows that on average the Top Teams in the MLB have a higher SLG than the Bottom Teams. SLG is a stat that managers need to take into consideration when creating their lineup card. As we covered earlier, SLG is a stat that measures a teams ability to hit extra base hits. So, when constructing a team, having some hitters that are capable of hitting extra base hits. We can see that in baseball this is becoming more and more frequent. Teams are starting to sign more and more sluggers that hit the ball hard. This means that it is even more imperative that general managers and managers can sign players that have high a high SLG.

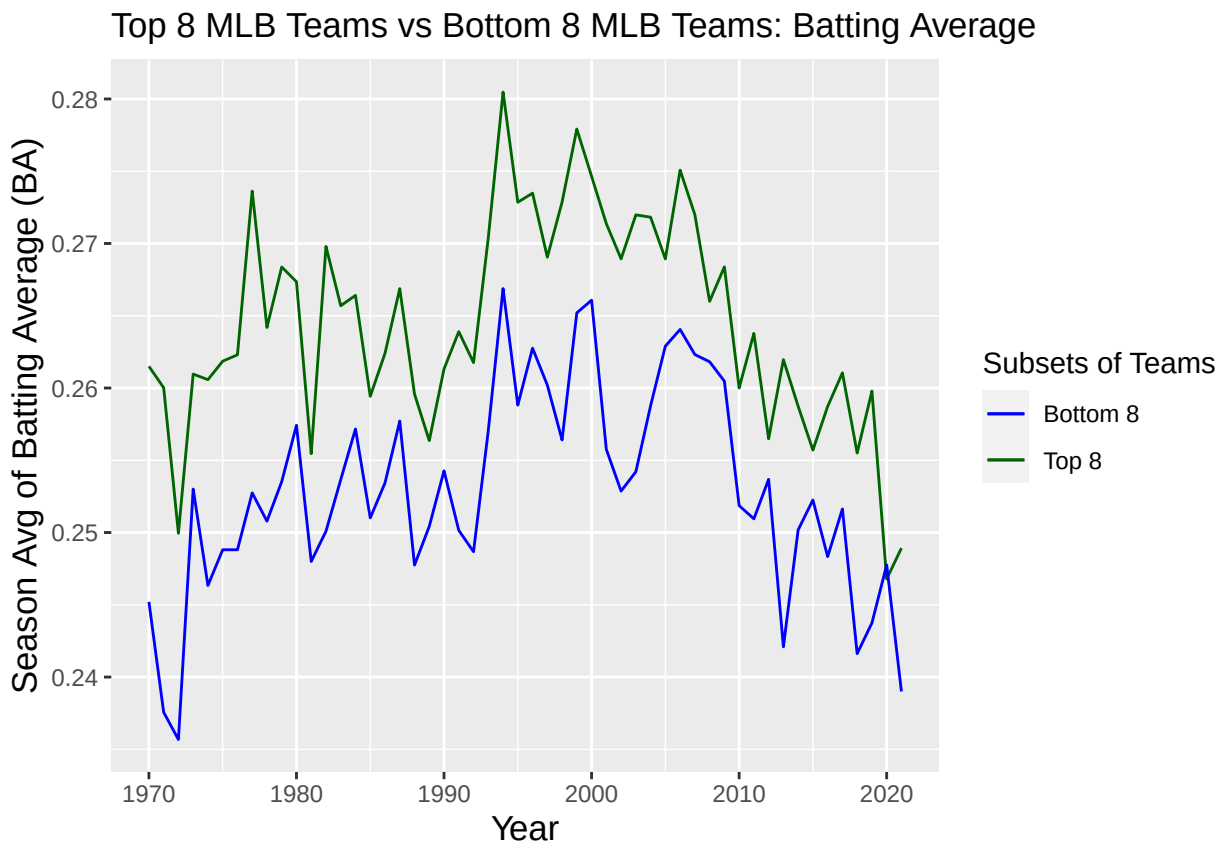

Figure 26: Top 8 vs Bottom 8: BA

Figure 26 shows that on average the Top Teams in the MLB have a higher BA than the Bottom teams in the league. This also makes sense as BA is a hitting statistic that revolves around scoring runs, which correlates with winning more games.

BA is a similar statistic to SLG where how often a player reaches base while batting is tracked, however, SLG has a higher difference between the Top 8 and Bottom 8 teams than BA. One reason this may occur is that while both stats track hitting statistics, SLG weighs extra base hits (doubles 2B, Triples 3B, Home-Runs HR) heavier than just a single whereas BA is a non-weighted average over the same data. A typical extra base hit leads to runs being scored more often than just a single (1B) or a walk (W) occurring, thus it swings the game more into the favor of the team with a higher SLG rather than just a higher BA.

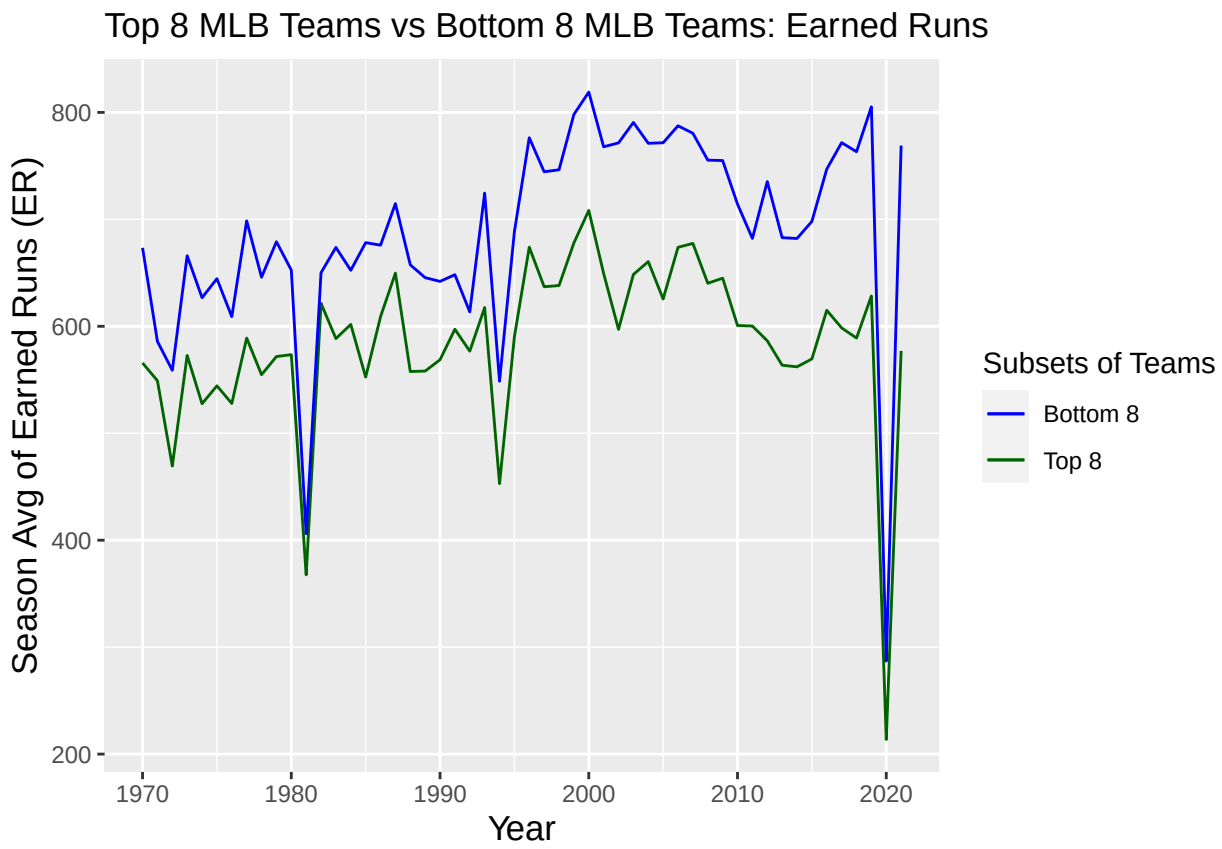

Figure 27: Top 8 vs Bottom 8: ER

Figure 27 shows that on average every year in the MLB, the Top 8 teams accumulate more Runs than the Bottom 8 teams. This aligns with earlier findings with ER that the more Runs you score, the more likely your team is to Win.

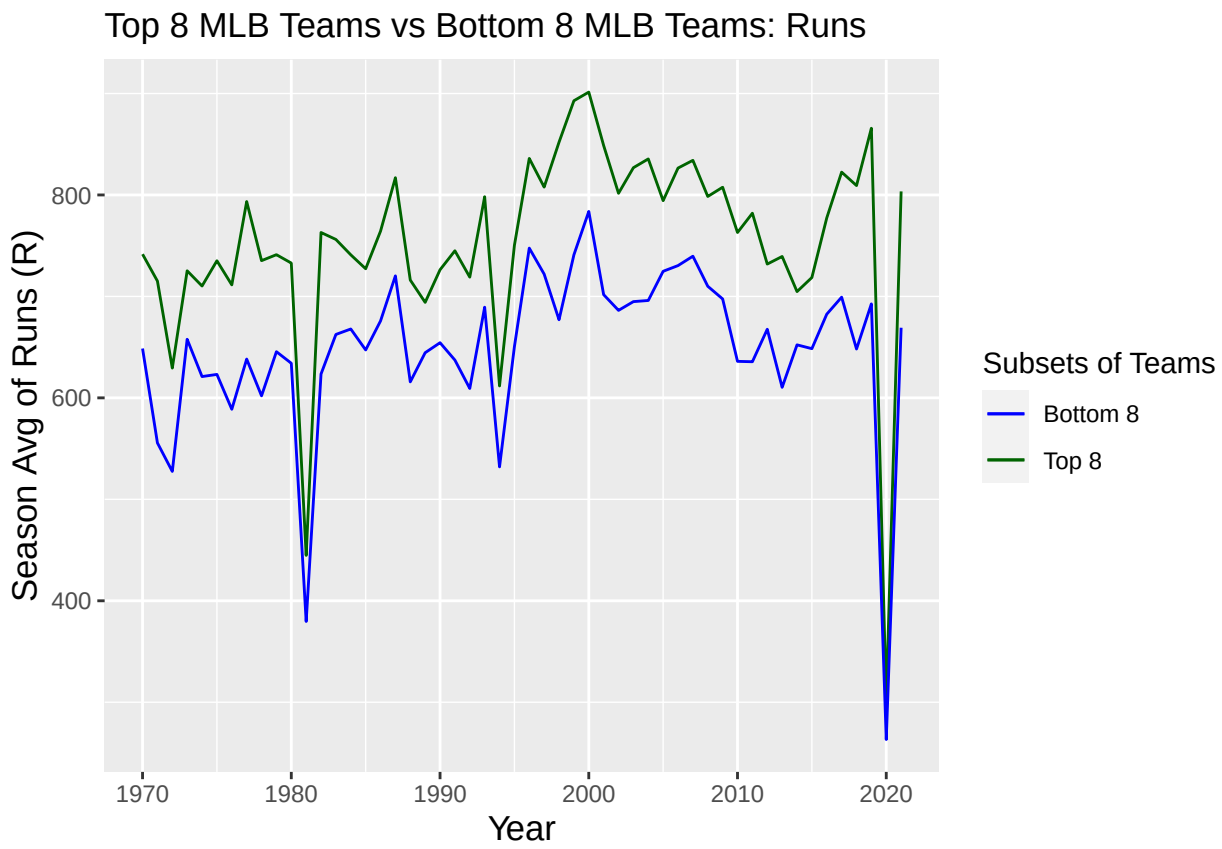

Figure 28: Top 8 vs Bottom 8: R

Figure 28 shows that the Bottom 8 Teams have a higher ER total than the Top 8 teams on average every year in the MLB. There are 3 occurrences of outliers which are important to note: 1972, 1981, 1994, and 2020. In the season, the total games for each team were shortened to be 60 due to the COVID-19 global pandemic. In each of the other 3 years, labor strikes held by the players shortened the seasons. ER refers to how many runs are given up by your team, thus it makes sense that the teams that lose more than the rest of the league give up more runs than the best teams in the league.

Looking at both Figure 27 and 28, this may be confusing. ER is the amount of earned runs allowed whereas R are total runs scored. For earned runs, this plot tells managers that pitching is important. This is because earned runs allowed means that the other team got hits and it was not due to fielding errors. If a team scores a run via error, it does not count towards ER. This reinforces the idea that a good pitching stat is crucial to the success of a team. As we stated early in the section, defense matters, however it seems that pitching might be more important than actual fielding. To general managers, this is an important concept as some teams do not have the same budget as others. So trying to spend money where it matters is important. Therefor, if a team is struggling when it comes to allowing runs, paying pitchers rather than field players that are good at defense may be a better allotment of a teams budget.

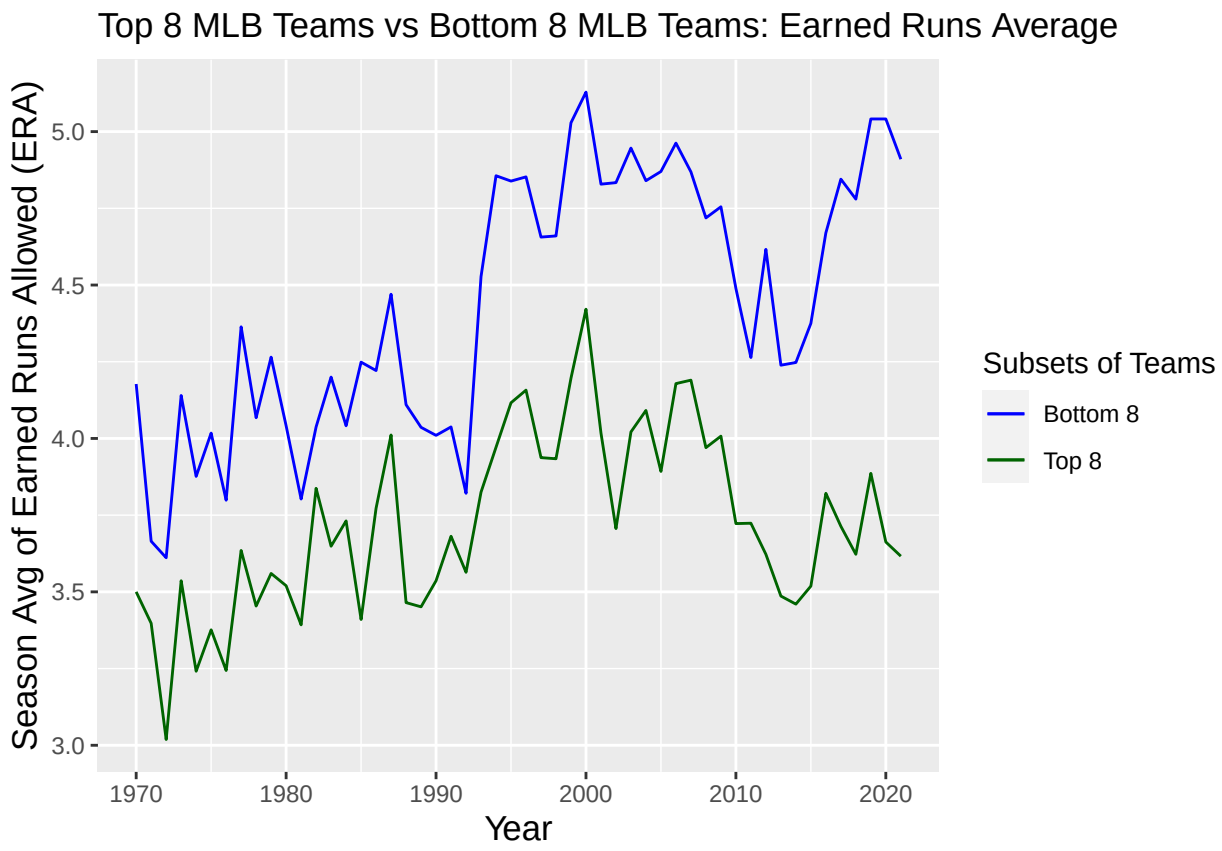

Figure 29: Top 8 vs Bottom 8: ERA

Figure 29shows that the bottom 8 teams have a higher average ERA than the top 8 teams every year. This makes sense again as ERA is a defensive stat and, therefor, the lower the number, the more games a team will win. We have emphasized the importance of ERA before.

5 Conclusion

In conclusion, the stats that most impact team success are OPS, SLG, OBP, runs, runs allowed and ERA. Other stats such as fielding percentage, batting average and earned runs were also good indicators of team success but not as great as the ones listed prior.From someone that has played and watched countless hours of baseball, all of these make sense. OPS, ERA, SLG and OBP are all stats that are used to evaluate teams very regularly. Runs and runs allowed are all about how teams win games so these obviously are going to have huge impact on team success. The biggest surprise is how little strikeouts impact team success. Growing up playing baseball, coaches always emphasize limiting strikeouts. We explained why this strategy does not hold up in section 4.1, but the point remains that it is barely relevant in team success. The biggest takeaway from the findings is that not one stat is the direct cause for team success. As you can see in Table 3 many of the variables have very similar correlation values. This tells us that baseball cannot be simplified down to one or two stats. This makes sense as baseball is a sport that has many different strategies. A team can be successful by getting on base a lot, or hitting a lot of extra base hits, or playing really good defense. Therefor, we must take into consideration all of the stats presented at the beginning of this section.

For managers and general managers, this information is very important to understand. They need to understand that strikeouts have little impact on success, so that when making decisions, it is not a factor in the decision making process. They also need to understand that ERA is one of the most important factors in team success so that they can prioritize signing pitchers to help defense, not defensive minded position players. They need to understand that signing players that are more well rounded will help them win more. This is the biggest takeaway. Having a team that is well-rounded is the key. Although this is nothing shocking, it does prove that trying to create teams that are only good at one thing, and expect them to win games is not feasible. This isn't to say that a team can't have an identity. Their are countless teams that have a stat that they are good at. However, these teams still are good in other aspects of the game.

While we can measure team success through the various on-field performance statistics, one aspect remains hidden: does the financial success of a team align with success on the field? This is one question we are unable to solve for due to the limitations of our data. Financial data for professional sports teams, let alone MLB teams, is very rare to find as nearly every team is a privately owned entity, thus they don't have to disclose any financial data to the public. While attendance numbers are tracked for each team and are a part of the Lehman data set that we used for this project, there are several issues with basing success from the data set. First, the dataset only accounts for total attendance through the whole year rather than just the regular season or the post season. This means that for good teams that make the playoffs, they'll get more home games, and thus higher attendance numbers than those that don't. Plus, every fan base of each team is different, meaning some teams, such as the New York Yankees or the Los Angeles Dodgers, always have good attendance numbers regardless of the team's on-field performance. For example, Coors Stadium, home to the Colorado Rockies, houses up to 50,398 fans at capacity, whereas Progressive Field, home of the Cleveland Guardians, houses up to only 35,041 fans at capacity. Because of this, judging teams based upon total attendance over the year is a bad metric. However, if you were able to compare the percentage found from the average attendance divided by the total stadium capacity for each team, that metric would show what teams fill their home stadiums the most on average through the given year. While nice in theory, this would be very hard to calculate especially over a long period of time due to stadium capacities having increased over time through stadium expansions, moving the franchise to a new city with an entirely different stadium, or constructing an entirely new stadium all together. Because of this, we are unable to make observations of team success with statistics other than on-field statistics.

The work done in this report are sufficient enough findings to make an analysis on team success based off of individual, on-field statistics. However, the next step would be to take more of a statistical approach. By creating a multi-variable linear regression model, we would be able to predict team success given their performance in a few key variables. We would expect to use the variables identified through the Top 8 vs Bottom 8 team performance analysis as the ones for the linear regression model. However, we would still want to confirm that each of those stat fields are significant from their

$$R^2$$

value in order to maximize model accuracy. This model would allow team GMs and managers to track the current progress of their team, see what specific areas they need to improve upon, as well as how adding or removing specific players who perform well or poorly in one of the identified significant fields improve or worsen the team's probability to become more successful.

References

[1] CWS Omaha, *College World Series: Economic Impact*, https://cwsomaha.com/economic-impact/, 2019

[2] Lahman Sean, *Download Lahman's Baseball Database*, https://www.seanlahman.com/baseball-archive/statistics/ , 2022

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2022

[4] Yihui Xie, *knitr: A General-Purpose Package for Dynamic Report Generation in R*, http://yihui.name/knitr/, 2022

[5] Leslie Lamport, *LaTeX: A Document Preparation System*. Addison Wesley, Massachusetts, 2nd Edition, 1994.

[6] Wickham H., *ggplot2: Elegant Graphics for Data Analysis.*, Springer-Verlag New York, 2016.

[7] Wickham H, Francois R, Henry L, Muller K, *dplyr: A Grammar of Data Manipulation. R package version 1.0.10*, https://CRAN.R-project.org/package=dplyr/, 2022

[8] Dowle M, Srinivasan A, *f data.table: Extension of data.frame. R package version 1.14.2*, https://CRAN.R-project.org/package=data.tabl/, 2021

[9] Muller K, Wickham H, *tibble: Simple Data Frames . R package version 3.1.8*, https://CRAN.R-project.org/package=tibble/, 2022

[10] Wickham H, Girlich M, *tidyr: Tidy Messy Data . R package version 1.2.1*, https://CRAN.R-project.org/package=tidyr/, 2022

[11] Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg A, Crowley J, *GGally: Extension to 'ggplot2' . R package version 2.1.2*, https://CRAN.R-project.org/package=GGally/, 2021

[12] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H *"Welcome to the tidyverse." Journal of Open Source Software , 4(43), 1686.doi:10.21105/joss.01686*, https://doi.org/10.21105/joss.01686/, 2019

[13] Waring E, Quinn M, McNamara A, Arino de la Rubia E, Zhu H, Ellis S *skimr: Compact and Flexible Summaries of Data . R package version 2.1.4*, https://CRAN.R-project.org/package=skimr/, 2022