Data Science Project 1: Baseball

Drew Watson and Max Ramsdell

Introduction

Baseball is a sport, primarily American, that has a large set of high quality statistics. Due to the simple, yet rigid nature of the game there are many occurrences that can be tracked as a matter of statistics. The dataset that we used in this analysis has a wealth of statistical information on teams, players, awards, and more. It is an extremely high quality and well documented dataset making it easy to use and manipulate. Our analysis on this dataset will be primarily concerned with statistics that predict the ability of a team to perform well in their games.

The analysis techniques that we used generally center around standard statistical measures such as mean and median, as well as a comparison of dataset subsets. Our results found that statistics like shutouts, rank, percent of games played at home have a small, but significant amount of statistical power for predicting the performance of a team in a season, while the percentage of successful base steals has a correlation, but a very weak one.

Dataset

The dataset that we used contains a large amount of data about baseball teams. It has been extremely well curated and contains a very large amount of data going back to the 1880's. This makes it ideal for analyzing the game of baseball over the years, and because it has a lot of data on league standings and win rates, we can easily use it to understand the statistical significance of various baseball statistics. In our analysis we especially made good use of the games played and games won statistics.

Analysis Technique

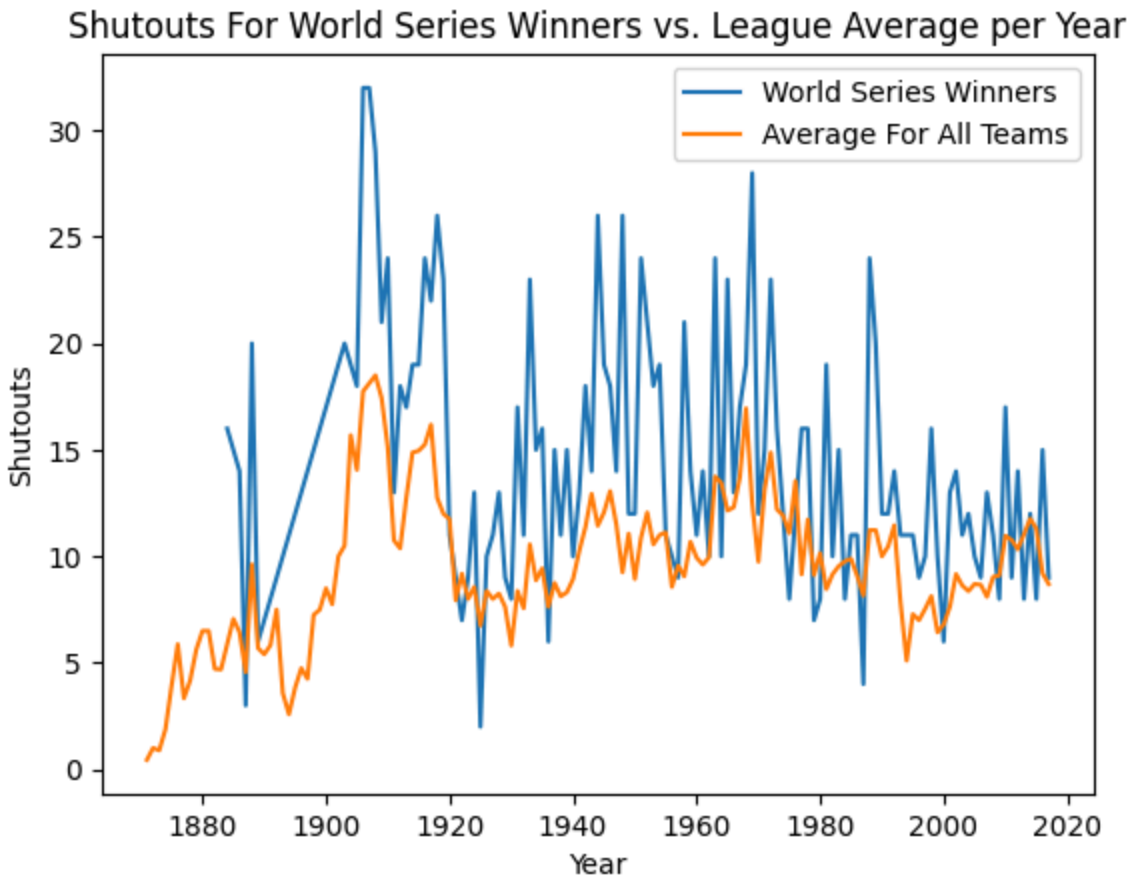
For our first dataset, attempting to understand if the number of shutouts a team has is a predictor of whether or not they will eventually win the world series, we decided to compare the total number of shutouts for the world series winner for a year with the average number of shutouts for the entire league. This made sense because it would show if the world series winners were consistently performing better on that metric than

the rest of the league was. To understand the relation between stealing bases and win rates, we used two charts, one of the total number of stolen bases, and one of the percentage of successfully stolen bases. These charts inform each other and provide context for what the other means. To understand the effect of home games on win rates, we split teams (over the years) into two groups, those with more home games and those with less. We then analyze the differences between those win rate charts and compare them. The last analysis that we performed was on the correlation between rank and winrate. We did a barchart that split the ranks by league because the ranking was the league ranking. We then compared the bars by win rate because if there is a correlation we will be able to see it extremely easily.

Results

Part 1: Number of Shutouts by World Series Winner vs. Average Number of Shutouts by Year

In baseball, a shutout is a game where the winning team prevents the losing team from scoring a single run, and the pitcher plays a full game. We wanted to determine if there is a significant difference in the number of shutouts that the winner of the World Series gets in a season and the average number of shutouts in the same season. To do this, we compare the number of shutouts by the winner of the World Series to the average number of shutouts that season, for each season. The figure below shows the number of shutouts per year, where the blue line is the number of shutouts for that year's World Series winner and the orange line is the average number of shutouts for that year.

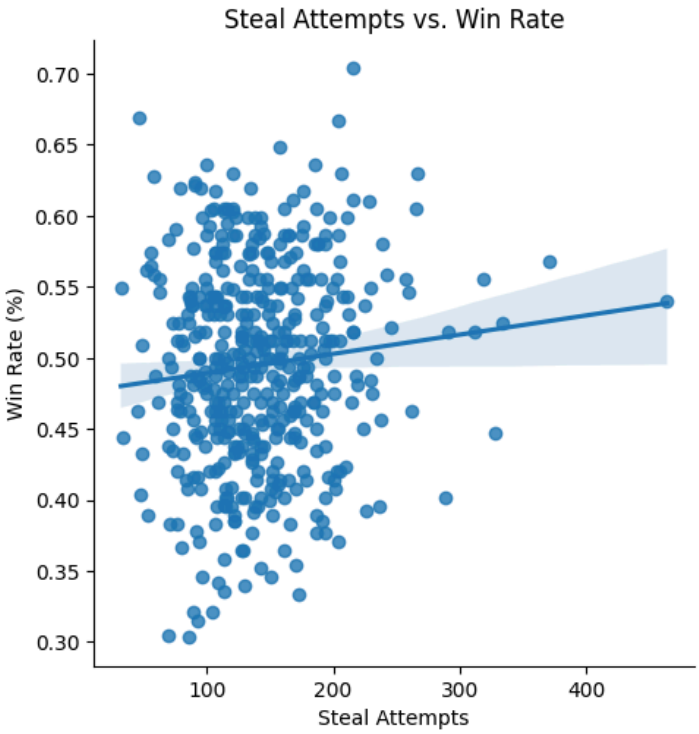


The figure shows that the number of shutouts by the World Series winners is typically much greater than the average number of shutouts in a season. This implies that a common feature of teams who win the World Series is having a large number of shutouts compared to the season average. While this does not necessarily show that teams with above average shutout numbers are more likely to win the World Series, it is an indicator that teams with below the average number of shutouts will probably not win the World Series.

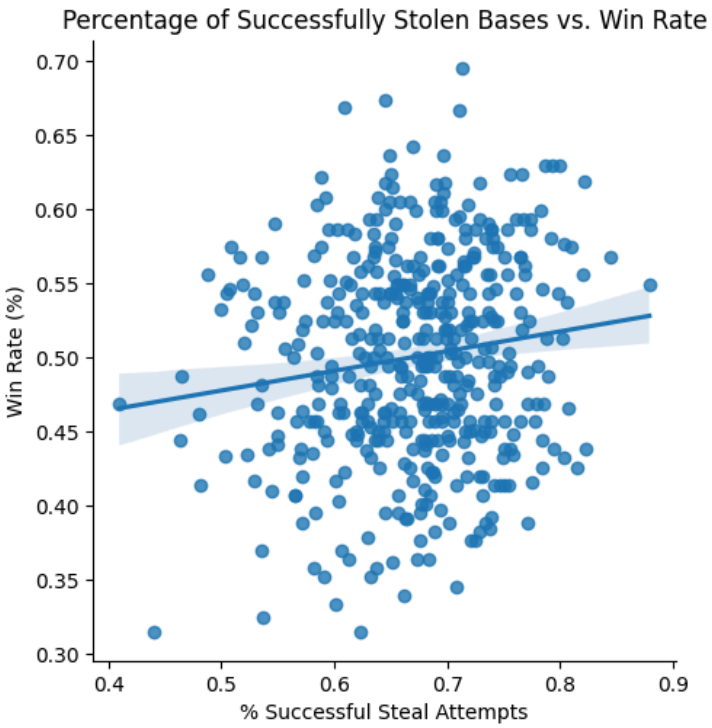
Part 2: Stealing Bases vs. Win Rate

Stealing a base in baseball is considered by some to be a worthwhile endeavor. For others, it is just a way of showing off. We aimed to determine whether attempting to steal a base is a meaningful way to increase your win rate for the season. There are two major components when it comes to stealing bases that have a possibility of affecting win rate: total steal attempts and stealing success rate. The figure below shows a

scatterplot of teams total steal attempts for a season against their win rate. The figure also shows a linear fit of the data.



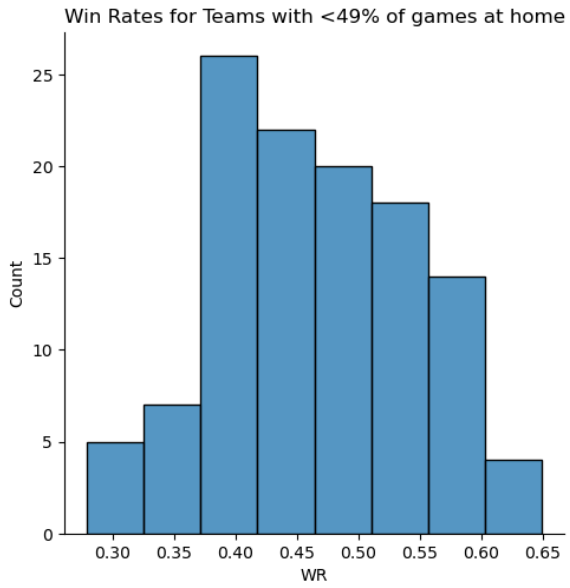
The linear fit in the chart above shows that there may be a slight upward trend in the win rate as the steal attempts increase. However, the confidence interval shows that there is not necessarily an upward trend in a team's win rate as they attempt to steal more bases. For this reason, we conclude that attempting to steal more bases will not necessarily increase your win rate. The next figure shows a scatter plot of teams stealing success rate (% successful steal attempts) against their win rate. The figure also includes a linear fit of the data.

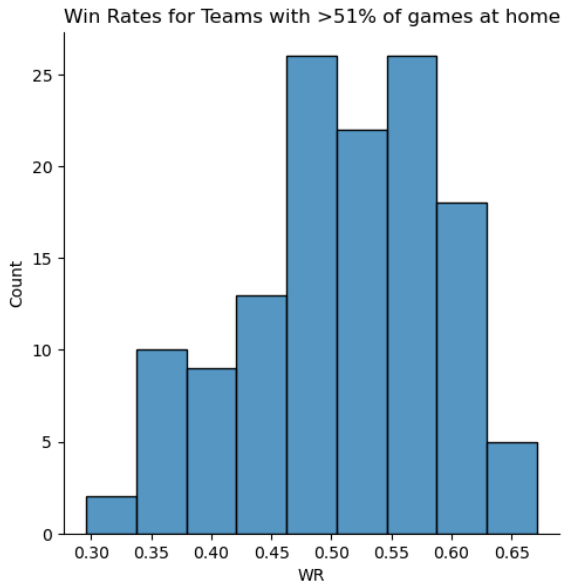


The trend line for this second figure shows a positive correlation between stealing success rate and win rate. However, the confidence interval for this figure also shows a positive correlation between stealing success rate and win rate. Therefore, we conclude that if your team has a good stealing success rate (i.e. your players are good at stealing bases) then attempting to steal a base may be a good strategy to increase your win rate.

Part 3: Home Games vs. Win Rate

Baseball teams generally have their own stadium that they train at, and on game day that's where the crowd comes to watch. It has long been hypothesized that there is a "Home Court Advantage" for the team playing in their own stadium. The charts below show the win rate for all teams with less than 49% of their games at home, and one for all the teams with more than 51% of their games at home.

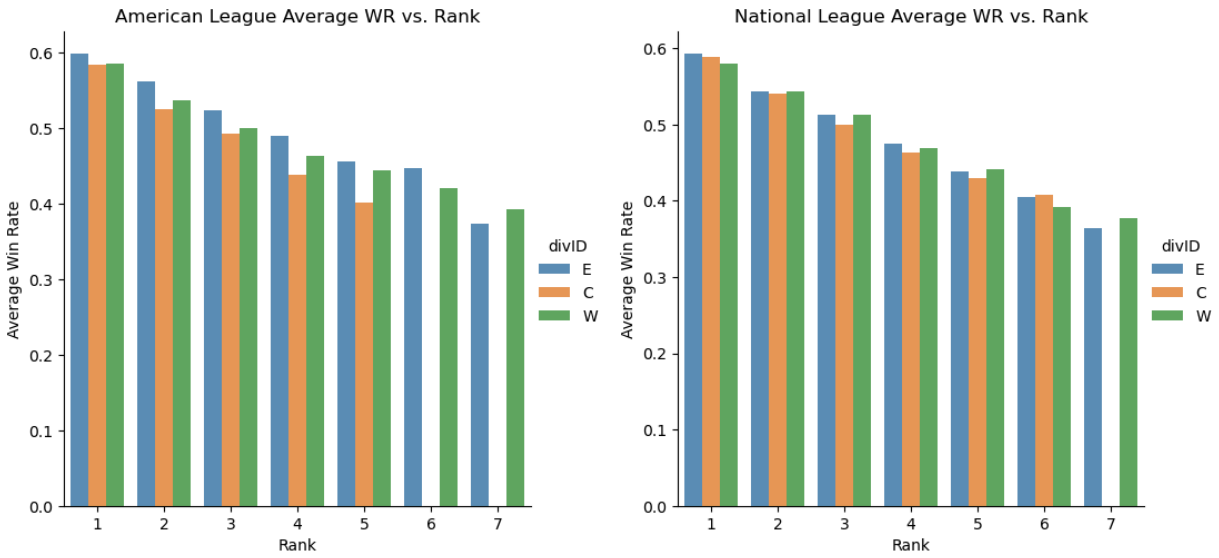




Note that teams can be in each chart more than once because we count their games by season rather than combining all of their wins into one. What we can see here is that the two charts are absolutely skewed away from each other, with the chart representing less home games being skewed toward a lower win rate, and the chart representing more home games being skewed to a higher win rate. It's important to note that the charts still had a similar range, meaning that although you may have had more games at home, doesn't mean that you will necessarily do better, and even though you had more games away doesn't mean you will do worse, but there definitely must be some form of a home court advantage.

Part 4: Win Rate vs. Rank for American and National Leagues

In American baseball there is a National league and an American league, each with three divisions: Eastern, Central, and Western. The teams in each division play against each other, competing for a division rank, among other things. We wanted to see if the division rank had a significant correlation to the average win rate of the teams in those



ranks. It is easy to see that there is absolutely a correlation between win rate and rank, with win rate trending downward as rank also does (the higher numbered ranks are lower). What is an extremely interesting and unexpected finding is the difference between win rates between divisions. Looking at the left chart, each set of bars has a much larger range of win rate than those in the right chart. The eastern division generally dominates the western, which generally dominates the central division, it is interesting to see that the competition is much tighter in the national league than it is in the american league.

Technical

To prepare the data, we created a directory titled “data”, then unzipped the given data .zip file into that directory. Then, we read through the file “readme2014.txt” to learn how the data is laid out and which files might be of use to us. The only cleaning of the data we did was using the “dropna()” function supplied by the pandas library.

Part 1 of our project involves comparison of what is essentially two time series. For that reason, we opted for a line plot, as a line plot excels at showing fluctuations in time, and

allows us to see how the trends of the two time series compare. Part 2 of our project involved determining the correlation between two parts of the data, so we used linear regression. We chose a linear fit because we guessed that the data would not be correlated by an exponential or higher order polynomial fit. Part 3 of our project involves determining whether there is a significant difference between two parts of the data (<49% home games and >51% home games), so we opted to use a histogram for each case and compare the distributions using a t-test. We decided on histograms because they make it visually easy to see differences and similarities in the data. Part 4 of our project has to do with comparing data from 3 different divisions in 2 different leagues, so we opted for a group bar chart. Using group bar charts allows easy visualization of the 3 divisions side by side, makes the trends in the data easy to see, and allows for easy comparison of the two charts.

Our analysis process for Part 1 involved us talking to each other, and really deciphering what the data was and wasn't telling us. At first, we concluded that the data was telling us that higher shutout numbers meant a better chance of winning the World Series, but when we looked at the data a little closer we realized that it just told us that having much higher than average shutout numbers is a common feature among World Series teams, and that anything beyond that is extrapolation.

Our analysis process during Part 2 involved brainstorming the benefits and drawbacks of using stealing success rate and using total steal attempts, and what the implications of using each would be for our conclusions. Eventually, we decided that it would be easy and meaningful to use both in our analysis.

Originally, our analysis of home games and win rate just consisted of a scatter plot which told us that there was no correlation between percentage of home games and win rate. We spoke to Dr. Edwards, and he explained that the scatter plot wasn't useful, and all the data where teams had between 49% and 51% of their games at home created confusion. Thanks to his feedback, we realized that separating the data into two sections and creating a histogram for each, then comparing those histograms would be a better way of seeing potential differences between the data. We then decided to run a t-test to make sure that the visual differences in the distributions were not due to a small sample size.

As with Part 3, our analysis for Part 4: Win Rate versus Rank was originally just a scatter plot. Then, we realized that there were many, many duplicate ranks. That seemed odd to us, so we decided to investigate. We realized that the ranks are not over-all, rather they are within each division, within each league. We then changed to using an average over all years, and plotted each league separately, with each division

being its own color within the plot. Dr. Edwards helped us to realize that a group bar chart would help us visualize the data much better than the scatter plot we were using. After making that switch, it was much easier to see trends in the data and to compare the two charts.

Slide Presentation: Data Science Project 1
Github Repository: Github Repository