Eric Machamer

Dr. Shion Guha

COSC 3750

06 December 2019

Analyzing Major League Baseball Statistics and Salaries

Introduction

This project was centered around obtaining data sets of statistics and salaries for MLB players. In total, three data sets were analyzed. One data set represented statistics from the 2019 MLB season and contained a column of batters, with a column that identified the team associated with each batter, a column that identified the position of each player, and sixteen columns that represented variables that measure certain batter metrics. This set was restricted to batters who played at least 100 games in the 2019 season. This was because each player had to contribute a large sample size of at-bats in order to truly illustrate the quality of their statistics. The second data set contained a column of pitchers, with a column that identified the team associated with each pitcher, and sixteen columns that represented variables that measure certain pitcher metrics. This data set was restricted to starting pitchers only. This was because the metrics used in the data set are more reflective on pitchers who pitch a lot of innings, and there is more relevant data when it comes to player salaries for starting pitchers. The third data set contained salary information for every player who stepped onto an MLB diamond in 2019, with the one column containing the players, one column containing the team for each player, one column containing the position of each player, and four columns that contained different information about each player's current contract.

One goal of this project was to determine whether the overall production of players truly represents how much they are getting paid. Or in simpler terms, whether the players are actually living up to their contracts. To do this, one would have to set up a process where they would measure each player's salary against the metrics described in the data sets, determine what the trend is, and determine whether the trend is actually significant.

Another goal was to determine correlation between certain metrics in the batter and pitcher data, respectively. To see how closely related two variables are to each other, for example, "can one say with confidence that a player who hits for a high average could potentially also hit a lot of home runs?" Because there are many different combinations one could make from these metrics, only a few combinations were analyzed.

Another goal was to determine what is more important to success: a team with a high payroll or a team that has quality stats? There is a notion that a team such as the New York

Yankees who have a large payroll (because of value, popularity, wealth of owners) will consistently outperform a team like the Tampa Bay Rays who are one of the poorer teams in the League, because they can afford players who are significantly better than others. The goal was to try to debunk this notion, and confirm the "Moneyball theory", which will be explained in the next section.

Lit Review

Data analytics in professional baseball has become one of the most important factors in assembling baseball teams. General managers rely heavily on a player's career or collegiate statistics when determining if they would be an ideal fit for their team. This has not always been the case. Assembling baseball teams was mostly based off an evaluation of talent, or as some would say "the eye test." These teams hired professional scouts to find players with supreme talent, either in college or on other MLB teams. While talent is still vital with respect to fit, teams today are really broken down into who has the most ideal statistics that would fit on their team.

In 2002, the Oakland Athletics changed the way baseball teams were built. In their previous season, they were a juggernaut team lead by three star players. Their total payroll that season was around $38 million, one of the poorest in the MLB. They finished their 2001 season with an outstanding 102-60 record, falling short to the aforementioned wealthy New York Yankees in the first round of the playoffs, losing a best-of-five series in which they won the first two games. Unfortunately, this was not the end of their suffering. Their three star players, who the Athletics were able to afford at the time because they were under rookie deals, had their contracts expire after that season, and they were unable to afford to resign them because of the large offers those players received from other teams. So they had to find a way to replace them and remain competitive in their play with the low budget they had. This is when their front office led by General Manager Billy Beane decided to think differently, using advanced baseball statistics to scout talent. They were able to find players who they believed satisfied their statistical evaluation who were undervalued by other teams using the old-regime of player evaluation, and were able to sign them to affordable contracts and were able to qualify for the playoffs again, and in the process win twenty consecutive games in the regular season, breaking an MLB record.

The term that is most associated with using advanced baseball metrics to make organizational decisions is sabermetrics. A man named Bill James self-published extensive works on sabermetrics, calling them "the solution to objective knowledge about baseball." Most teams took these works with a grain of salt, believing that the only way to build a successful team was to get players who they believed would be successful strictly based off their ability and physical mechanics. They claimed that one cannot build a team on a computer. Billy Beane used James' work in his 2002 project, even though he received major push back. He hired statisticians

to assist him on this project, much to the chagrin of traditional scouts. To quote the movie *Moneyball*, which was a film based on this 2002 Athletics team, one of Beane's statisticians tells him, "Your goal shouldn't be to buy players. Your goal should be to buy wins. In order buy wins, you need to buy runs." This quote sums up this idea of sabermetrics pretty well, and this idea of buying wins to buying runs will be used in this project. Today, every MLB team has adopted sabermetrics not just in building their teams, but in making in game decisions.

Methods

The first step in the process of this project was to clean the individual data sets. The batter and pitcher data sets came directly from MLB.com, and for the most part were already clean. A couple changes were made, however, to a few of the variable names in the data set. For example, the columns that represented doubles and triples were represented in R as "X.2B" and "X.3B", so those were simply changed those to "2B" and "3B." Also, both data sets contained a variable called "RK", which described the rank and the order in which the players occured in the data set based on a certain metric. That column was deleted from both sets because this project is not centered around one specific metric. The real challenge was cleaning the salaries data set. This set had a variable called "Years" in which each cell contained the number of years under contract and the specific years the contract was for. For example, a cell would read "7 (2015-21)." This cell violates the rule of multiple descriptions in one cell, so this cell was divided into "Total Years", "First Year", and "Last Year".

The next challenge was to combine the data sets in order to make visualizations and models based off stats and salaries. The original plan was to take the batter data set, and expand it to accomodate for the pitcher stats and salary data, and just simply build upon the "Player", "Team", and "Position" variables. This however was a failure because R would not accept data from the pitcher and salary data sets that did not match up exactly with the data from the batter set. The next approach was to create an empty data frame and import all three data sets into that frame. This was also a failure, as R would not read the data as the types they were meant to be read as. For example, importing the player data which simply is just the names of the players would be read as an integer in R. After hours of struggling with combining the sets in R, the sets were combined in Excel, and not as one big data set, but as two sets with one containing batter stats and salaries and the other containing pitcher stats and salaries. While this is not the ideal way to combine the data, it was the best way to assure that the proper models and visualizations would be built efficiently.

The next challenge was to determine potential correlation among two specific metrics among the respective data set, and determine potential correlation among a specific metric and player salaries. To do this, linear regression models were used, because the goal of this part was to determine trend and statistical significance. In total, twenty six linear regression models were generated, but only a few were focused on for the relevance of this project. Significance was

determined by generating a BP test for each linear regression fit, recording the p value of that test, and generating a QQ plot to show how accurate the BP test is.

The next challenge was to create visualizations that depict whether a more successful team is more likely to have a high payroll or better stats. For the case of this project, team success was determined on whether or not a team made the playoffs. Although the data sets contained information solely on player information, one could simply research what teams qualified for the playoffs, and use the visualizations of the metrics generated to determine whether a team could potentially base their success off of those metrics.

Results

Figure 1: Runs Scored by Position by Team

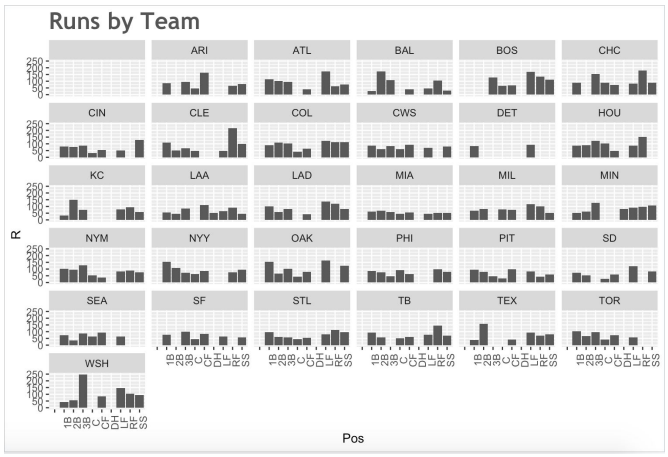
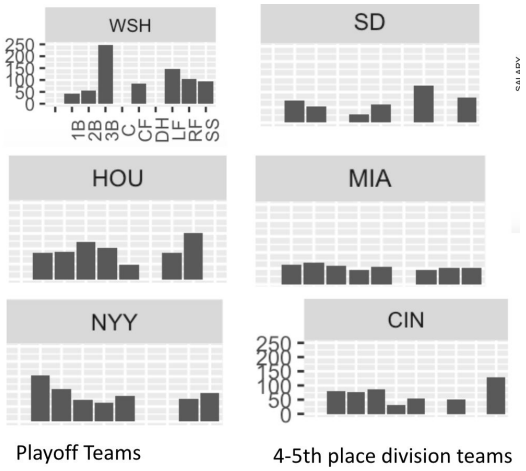


Figure 2: Select teams from Fig. 1



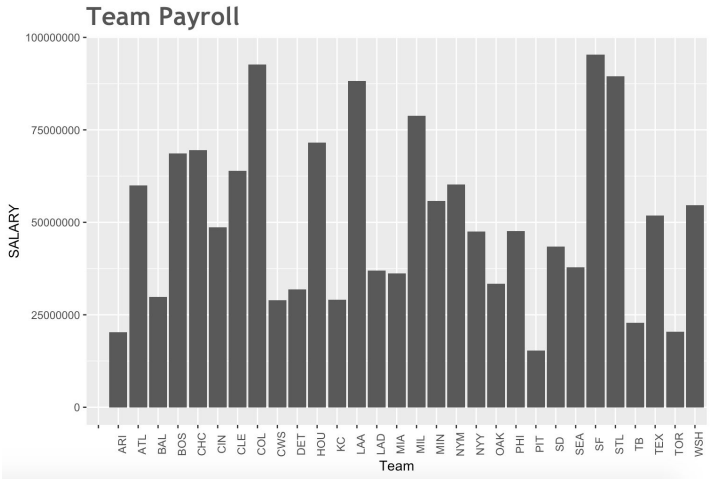
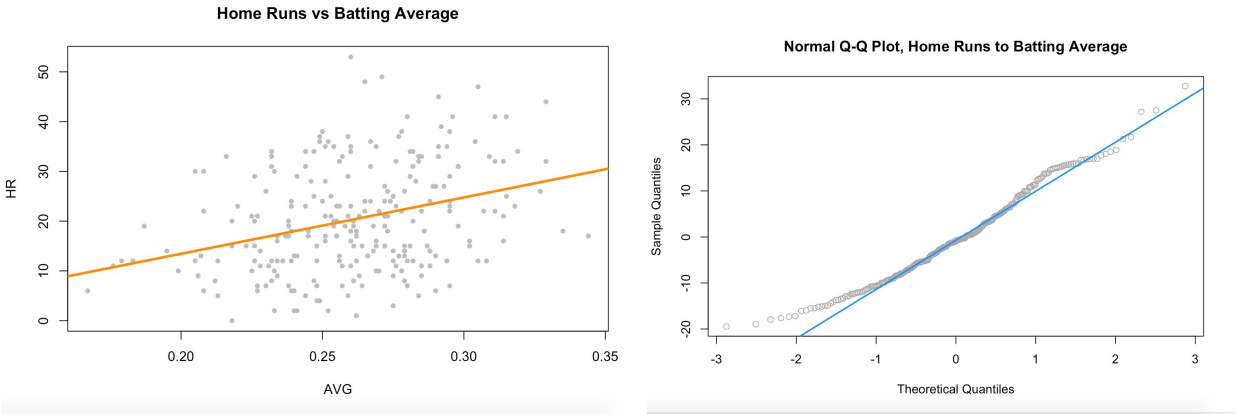


Figure 3: Annual Team Payroll

Goal 1: Determining whether payroll or runs scored is more important to success. Success was determined by each team's ability to make the playoffs. Runs scored was used here because of the idea stated before: To buy wins, one must buy runs. Figure 1 represents every team in the MLB, while Figure 2 only represents six teams: three teams that made the playoffs (left column)

and three teams that did not (right column). From this figure, it was pretty clear that runs scored are a factor in determining team success. It was unclear whether wins and runs are statistically significant, because the data set only measures player statistics not team statistics. However, one could simply research what teams made the playoffs in 2019 and would find: NYY, MIN, HOU, OAK, TB, ATL, STL, MIL, LAD, WSH all qualified for the postseason. Looking at this, one could look at Figure 1 and see that most teams who made the playoffs scored more runs than the teams that did not. Looking at Figure 3, which displays the total team payroll for each team in 2019, one would simply compare these teams' payroll to the median payroll and determine whether the fell above or below that threshold. Doing this, one could see that most of the teams fell at about the median salary, with STL, OAK, TB, LAD being the outliers. One could also look at the non-playoff teams such as COL, SF, and LAA who have payrolls way above the median threshold, and teams such as PIT, TOR, BAL, and ARI who have payrolls way below the median threshold. Using all of this information, one could say that since most of the teams who made the postseason had payrolls at about the overall average payroll, a large payroll is not necessarily vital to success. One could also say that a team is diminishing their chances by having a very low payroll, as most teams that had very low payrolls did not make the postseason. After determining this, one could say that a team should have about an average payroll if they want to be successful. After full analyzation, one could determine that runs is more vital to team success than total team payroll.

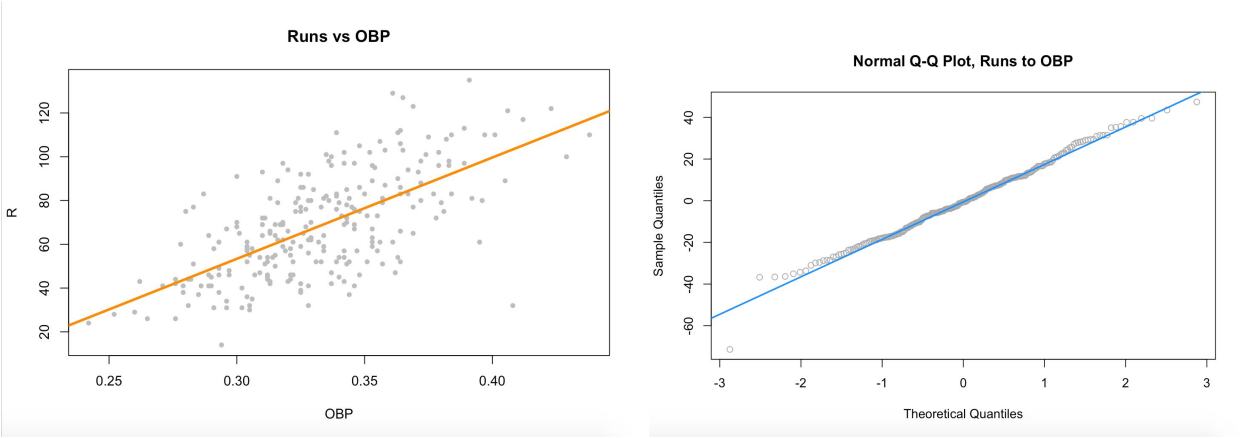
Figure 4: Linear Regression Model and QQ Plot for Home Runs vs Batting Average



Goal 2: Determining correlation between two variables from the respective batter and pitcher data. As stated before, many linear regression models were generated, but only a few will be analyzed here. In this figure, statistical significance was sought between Home Runs and Batting Average. Or, in simpler terms, can a player who hits for a high Batting Average potentially determine whether he also hits a lot of home runs? Here, there was an overall positive trend, but a BP test was run and the p-value was 0.1011. Because this value is >0.05, the null hypothesis cannot be rejected, so it cannot be said with certainty that these two variables are correlated. One

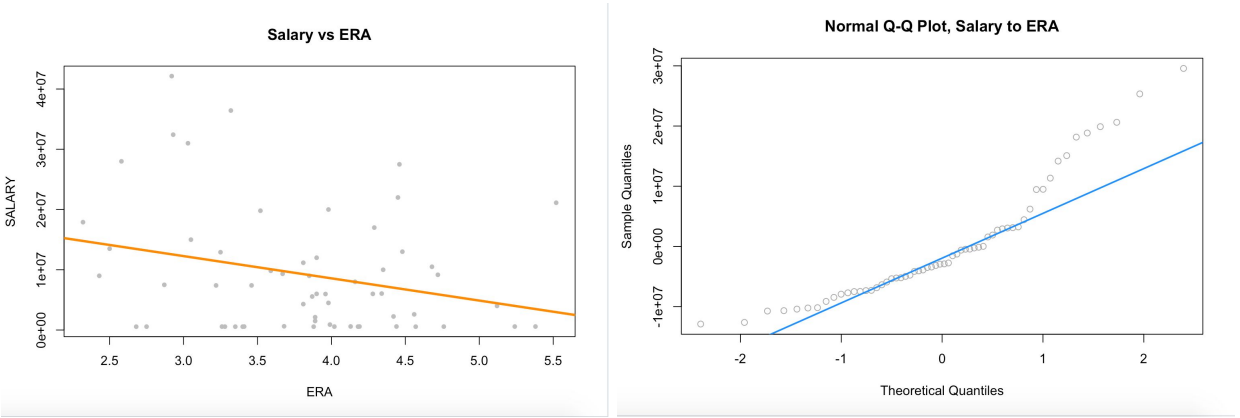
could visualize this by looking at the QQ Plot figure and see that a lot of points do not lie on the line.

Figure 5: Linear Regression Model and QQ Plot for Runs vs On Base Percentage



In this figure, a linear regression model was generated for Runs vs On Base Percentage. The linear model has a positive slope, and a BP test was run to determine whether this is significant, and the p-value was found to be 0.001998. This value is <0.05, so the null hypothesis can be rejected, and statistical significance can be confirmed. So a player who gets on base a lot could also potentially score a lot of runs. One could then visualize this by looking at the QQ Plot, and seeing that most points lie along or close to the linear fit.

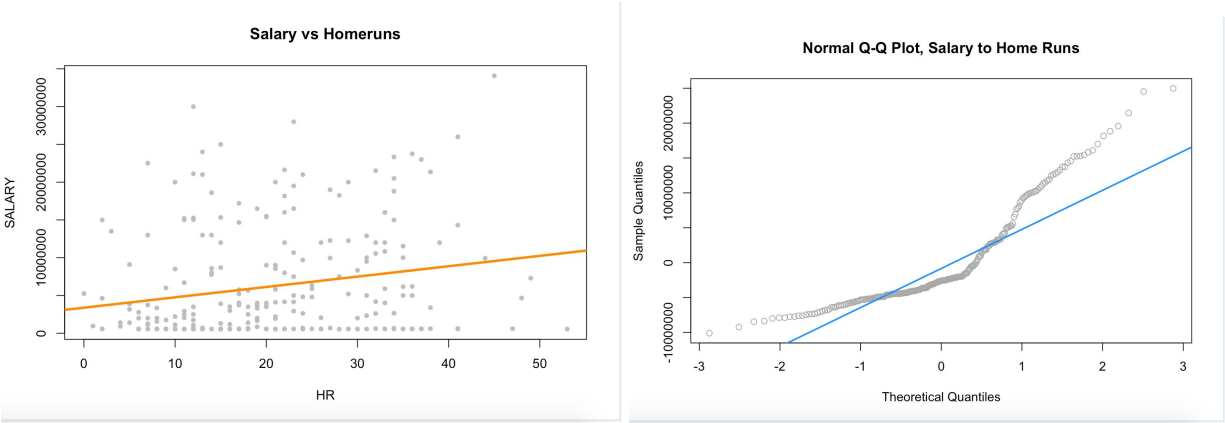
Figure 6: Linear Regression Model and QQ Plot for Salary vs ERA



Goal 3: Determining whether the overall production of players truly represents how much they are getting paid. In order to do this, the annual salary for each player was modeled against individual metrics with a linear regression model. In total, four models were generated within the pitcher data with Salary against ERA, Wins, Strikeouts, and Innings Pitched. Figure 6 represents the linear regression model and QQ plot for Salary vs ERA. Looking at the model, one could see

that a higher ERA in general means a lower Salary. This is actually expected, because ERA measures how many runs a pitcher gives up a game, and obviously they want to keep that low. However, when a BP test was run, it was determined that the p-value is 0.09596, which is >0.05, so the null hypothesis cannot be rejected, and there is no certain correlation between the two variables. Looking at the QQ plot, many points do not lie along the line of fit. This was the common result for all four pitcher metrics, as there was no statistical significance between any of them.

Figure 7: Linear Regression Model and QQ Plot for Salary vs Home Runs



In this figure, a linear regression model and QQ plot were generated for Salary vs Home Runs. This was one of the four models generated for Salary against Batter metrics, including Home Runs, Batting Average, On Base Percentage, and Hits. From the plot, one could tell that the model has a positive trend. When the BP test was run, the p-value was found to be 0.01795, which is <0.05, so the null hypothesis can be rejected and it can be assumed that these two variables are statistically significant. With respect to the other three models, the only one where the two variables were not statistically significant was Salary vs Batting Average, with the p value found to be 0.0581. So between the batter and pitcher data, it was determined that salaries are more statistically significant with batter metrics than pitcher metrics.

Discussion

From the results, it was determined that statistics is more important to team success than a large salary. There are a few theories as to why that is. One theory is that many players in the MLB, and a near majority of the players in these data sets, were under one-year or rookie contracts. This means that most players are not making an exorbitant amount of money compared to the average MLB salary. Referring to Figure 7, one can see from the linear regression model that many data points lie close to the bottom of the graph, indicating most players make less than $5 million. Because a majority of players do not have large contracts,

most of the stats come from them. And if those players have quality stats, it will most likely lead to team success. So if a team can maximize the potential of their rookie or one year deal players, they can most likely have success.

Another theory is that a team that signs a player to a long term deal worth tens or hundreds of millions of dollars takes a big cap hit. This means that if they invest a lot of money into one or a few players, they will not have money to spend on others. While money is not the leading factor to team success, it is still important to pay players who have shown they can perform well. Of all the major sports in America (Baseball, Football, Basketball, Hockey), baseball is the most team-centered sport. One player cannot determine a whole team's success. A team would not mind taking a cap hit if they get a return on their investment, but if that player is not performing to expectations, that team becomes trapped: they cannot win, and they cannot fix it by spending more money on players. Referring to Figure 3, teams like LAA or SF have huge annual payrolls, and they spend hundreds of millions on just a couple of players. But some of those players have not performed to the standard of their contracts, and they have not made the playoffs.

Some teams may take a chance on signing a player a long term deal even if they think there is a chance that player may not perform well. Why would this be the case? Owners in professional sports are concerned primarily with making money, not winning. If there is a team who consistently struggles to make money from not winning, not selling tickets, or not selling apparel, that team may take a chance on a star player just to attract fans. Sports are a part of the entertainment industry, and it is hard to make money if fans do not spend on teams, even if that team is winning. So a front office's thought process might be that since we struggle to fill our stadium, this player will help fix that, win or lose. If we win, great. But even if we lose, fans will still want to come out to see that player, buy jerseys, etc. So that team will make money either way.

From the results, it was determined that most pitcher stats are not reflective of a pitcher's salary. This could be because pitchers who have long term deals are overall not living up to expectations. This could be due to age, health, mechanics, etc. When a pitcher ages, his mechanics start to break down, so he might not have the same fastball speed as he did when he was younger. This could also be because pitchers are more valuable than hitters. In one game, there are nine starting hitters, but only one starting pitcher. And that if that pitcher consistently performs well, that team will most likely win. A team might want to sign a 28 year old to a ten year deal, extending to when that pitcher becomes 38, an old man in baseball years. Why would this be? A player knows his value, and a team knows when they have a shot at winning a World Series. So when they are negotiating contracts, a player knows when a team really wants him, and he could negotiate a couple more years than the team might want, but if that team truly believes they won't win a World Series without that player, they will do it anyway.

It was also determined that most batter stats are reflective of a batter's salary. This could be because batter's production can be better predicted than a pitcher's production. It could also

be because players who are under rookie contracts may not get as many at-bats as established veterans, so they will not have as many stats. Or it could be that batters get better over time, as it may take some time to adjust to the Major League level.

From the results, it was determined that Home Runs does not correlate with Batting Average. This could be because players who hit a lot of home runs are trying to hit home runs every time they come to the plate. This also means they will swing and miss a lot. So it is not surprising that most home run hitters fall somewhere in the median batting average. It could also mean that most players who hit for a high average are not trying to hit home runs. They are simply just trying to get on base, which as was pointed out in this report, can potentially predict runs and therefore wins. A player knows what he can and cannot do on the baseball field. If they know they have a lot of power and can hit many home runs, they will try to do so every time. But if a player knows he does not have a lot of power, he will never try to hit a home run. As for Runs to On Base Percentage, it is not surprising that these two are correlated, as the only way to score a run is to get on base. There are some anomaly scenarios where a player can reach base without increasing their On Base Percentage, but for the most part when a player reaches base they increase their On Base Percentage. Going back to the Moneyball story, Billy Beane's main objective was to find players who could get on base because this was the best predictor to scoring runs. And as mentioned before, if a team scores a lot of runs, their chances of making the playoffs increase.

Conclusion

Major League Baseball has greatly altered the way the game is played. It starts with how the teams are built. Teams have started accepting Bill James's and Billy Beane's theories that statistical analysis must be used in order to build successful teams. This was shown in this project, as stats are more important to team success than having a large payroll. It was also shown that pitchers are more likely to have lesser stats than their contract details assume, but batters are more likely to live up to the contracts they are given. The statistical significance was determined for a few important variables, in Home Runs vs Batting Average and Runs to On Base percentage. It was determined that Home Runs and Batting Average are not statistically significant, but Runs and On Base Percentage are. The potential reasons why were explained in the discussion. This project was able to answer all of the questions, and it was a joy to work on them.

Sources

Lewis, Michael. *Moneyball: the Art of Winning an Unfair Game*. W.W. Norton, 2013.

Moneyball the Movie

https://www.domo.com/blog/the-man-behind-moneyball-the-billy-beane-story/

https://fivethirtyeight.com/features/billion-dollar-billy-beane/

https://sabr.org/sabermetrics

https://entertainment.howstuffworks.com/sabermetrics.htm

http://mlb.mlb.com/stats/sortable.jsp#elem=%5Bobject+Object%5D&tab_level=child&click_text=Sortable+Player+hitting&game_type='R'&season=2019&season_type=ANY&league_code='MLB'&sectionType=sp&statType=hitting&page=1&ts=1575911409210

https://www.usatoday.com/sports/mlb/salaries/

https://www.mlb.com/standings