THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF STATISTICS

Optimizing Roster Composition in Major League Baseball

KYLE J. KROBOTH
FALL 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Science
with honors in Statistics

Reviewed and approved* by the following:

Andrew Wiesner
Professor of Statistics
Thesis Supervisor

Matthew Beckman
Professor of Statistics
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Major League Baseball (MLB) teams are constantly working to gain an edge on their opponents. The MLB is the only major North American sports league without a salary cap, and as a result, one area that fans heavily scrutinize is teams' payrolls and how they are building out their rosters. However, some teams are top spenders and fail, while some bargain hunting teams find themselves in regular playoff contention. This thesis aims to determine the blueprint for a successful baseball team based on payroll, salary distribution, service time, and more. Success will be dictated by winning percentage, through a multiple linear regression model, and by playoff appearance odds, through a logistic regression. The models were also run with a subset of data strictly including "small market" teams. After testing assumptions, running regressions, and selecting models, all models were found to show a relationship between team payroll and how many players take up half of that team's payroll with both winning percentage and playoff odds. Surprisingly, payroll was not determined to correlate with success when limiting the models to small market teams, leaving only the number of players to 50% of payroll as a predictor. Overall, it was found that most of a team's success is derived from factors outside of roster composition, and even in the top model, only roughly 27% of variation in winning percentage could be described by roster composition predictors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Writing this thesis has been a long journey that I would have not gotten through without my amazing support network. I would first like to thank my thesis supervisor, Dr. Andy Wiesner, for always being there to help with R code, read my many drafts, and inspire me to keep pushing through the long thesis process. And for good sports talk always. I won't forget our many thesis meetings and our zoom STAT 462 classes. I would also like to thank the rest of the Statistics Department here at Penn State, for molding me into someone who can conduct academic research, something that I couldn't say when I arrived here. Last, I'd like to thank my family and friends, and anyone who has ever come to a baseball game with me. It takes a lot of work to be a Pirates fan, and it requires almost as much emotional support as writing an honors thesis. Your support has gotten me far, and I am always grateful.

Chapter 1

Introduction

Major League Baseball (MLB) stands alone as the only one of the four major North American sports leagues that does not regulate team salaries through a salary cap and floor system. Though there are many factors that play into an MLB team's success, the absence of a salary cap presents an ingredient to victory unique to baseball franchises: how much a team pays their players, and thus their capacity to "buy" players through free agency, acquire highly paid players through trade, retain high performing players through extensions, and promote top prospects without consideration for the long-term salary implications of such a move.

Different teams have different capacities to pay players, which leads to disparities in overall talent levels on teams. Naturally, a team's ability to spend money is dependent on the revenue that they bring in. Ticket sales, merchandise & concession sales, and local media rights contracts make up substantial components of teams' annual revenues. Teams in larger media markets, such as New York and Los Angeles, have a larger population to market to, and thus have a larger pool of customers to spend money buying tickets and watching on TV. Smaller media markets, like Pittsburgh, Kansas City, and Oakland, have smaller fanbases that generate lower revenues.

It is possible for small markets to make money and find success. Fielding a good team, even in small markets, typically leads to higher attendance and more revenue for the team. In turn, team ownership has more money to spend on players while still generating their desired profits. However, small market teams often find themselves in a "chicken or the egg" paradox,

where the team does not win because they do not spend enough money on players, but they also do not make enough money because the team is not successful. Small market teams also face a much smaller margin of error—one bad contract could set back an organization for years. Breaking out of this cycle is challenging, but it is the key for the success of small market teams and requires creative solutions in player development and acquisition.

Getting the most out of the talent that you have, whether that be through practice drills, individualized coaching, or organizational philosophy, is one of the most prominent strategies that small market teams utilize to close the gap with larger markets. These player development strategies are highly varied and difficult to quantify, so this paper will look specifically into player salary trends among teams. Would a team be better off spending half of their budget on a couple of star players, or spreading that money out across several players that aren't as good? In general, are teams that have more highly paid players more successful?

The main goal of this paper is to find if there is a salary breakdown archetype for a successful team, and if so, to find what it is. I will also analyze small market teams specifically to see if their strategy should differ from the default league strategy. This will be accomplished using regression models with team performance and salary data from 2012-2019, a period where the playoff structure was the same and other league rules remained relatively constant. Results will be evaluated for the quantitative independent variable of winning percentage and the categorical binary independent variable of playoff odds. This paper will start with a literature review, including a history of baseball economics and an explanation of the league's current rules for contracts. It will then discuss the methods behind the model, followed by an analysis of the regressions themselves, and ending with the findings and opportunities for further research.

Chapter 2

Background

Introduction to the Reserve Clause

The basis of baseball's economic inequality lies in teams' varying payrolls. For the 2022 season, six teams sit at an expected annual payroll above 200 million dollars, and 16 out of the remaining 24 teams are paying their players over 100 million dollars. Large market teams sign high dollar free agents and extend their star players to big-money contracts, pillaging established players from teams with low payrolls and owners with small checkbooks. What many younger fans don't realize, however, is that baseball's current state of economic disparity is a relatively new problem.

For the first several decades of the 20th Century, MLB players were held by their respective clubs under the reserve clause, a clause in each player's contract that stated, "the club shall have the right to renew this contract for the period of one year on the same terms" (Berri). The reserve clause artificially kept player salaries low; if a player played at a level above and beyond his contract, the team could choose to retain him at his previous contract year after year. If the player played worse than his salary suggested, the team could release him, with no long-term monetary commitment. The pool of what we have come to know as "free agents" (players not under contract and free to sign with any team) in the modern day consisted little more than players released by their previous team due to poor performance.

During the reserve clause era, Major League front offices functioned in a way far different than they do today. Because high quality free agents were not available, teams placed

most of their focus on acquiring young talent that they could then control for the rest of that player's career. Initially, this was done by purchasing players from independently owned "minor league" teams.

History of the Minor Leagues

Major League Baseball stands alone from the other major North American professional sports leagues in its regimented player development system known as Minor League Baseball (MiLB). There are 120 MiLB franchises, many of whom are independently owned and operated from their affiliated MLB club (Fagan). Each MLB franchise has the option to allocate players to teams in four minor leagues: Low-A, High-A, AA, and AAA (Creamer). Teams can also assign players to the Dominican Summer League (DSL), a league in the Dominican Republic that is generally the first landing point for international free agents, and the Complex League, where teams are located at big league Spring Training facilities and rosters are typically made up of recently drafted players and players elevated from the DSL. Both of these leagues consist exclusively of teams owned by their respective big-league clubs and share these clubs' team names (for example, the DSL Pirates). Players typically progress linearly through the MiLB: from Complex League, to Low-A, to High-A, and so on. MLB franchises control the player and coaching personnel of their minor league franchises, and are also responsible for compensating said personnel.

Minor League Baseball was officially established in 1901, but it started merely as a loose collection of smaller, independently owned and operated leagues that competed with the National League and American League (Twins Daily). Major League clubs could make

agreements with Minor League clubs to purchase their players at a certain price. In the 1920s, St. Louis Cardinals General Manager Branch Rickey began purchasing minor league franchises to use as "feeder teams" for his Major League club (Hall of Fame). This way, he could sign amateurs to his minor league clubs and promote them to the big-league club without competition from other organizations. Rickey's model led the Cardinals to six World Series titles and nine National League pennants over a twenty-year span, and this success spurred other organizations to emulate this farm system model. Gradually, less profitable minor leagues folded and rules were instated for the number of teams and players that a major league franchise could control, leading up to today's status quo.

Impact and Fall of the Reserve Clause

The importance of the minor leagues was spurred by the reserve clause. Because quality players could be held under contract indefinitely, the only way to secure the best talent was to identify that talent at a young age and assign them to a minor league team. As teams raced to sign top talent, prices for amateur players began to rise. The first example of the league fighting to keep amateur salaries down was in 1947, when the MLB tried to eliminate the trend of signing young players to large minor league contracts by installing the "bonus baby" rule, which stated that any player signed for a bonus of over $4,000 must be kept on a major league roster for two full seasons (Baseball Almanac). If a player meeting these standards was not rostered, they were then exposed to the Rule Five draft, where each MLB franchise could select eligible players, under the condition that they kept the player on their MLB roster (Simon). In 1964, the teams

voted to consolidate amateur signing into the "First-Year Player Draft," in a further effort to prevent wealthier teams from hoarding young talent.

Ultimately, the reserve clause kept all Major League franchises on a level playing field. Organizations that could scout and develop quality players had success, no matter the market size or value of the franchise. Initiatives such as the First-Year Player Draft ensured that lower-revenue teams had the same opportunities as teams in larger markets. In 1975, this balance was thrown into disarray. After a series of lawsuits and a decision by an independent, binding arbitrator, the reserve clause was struck down (Berri). After negotiations with the MLB Players Association, team owners reached an agreement for the implementation of modern-day free agency, which would change the game forever.

Today's Standards

The reserve clause may be gone in name, but its residual effects are still felt in baseball. Today, the MLB's Collective Bargaining Agreement stipulates that a player reaches free agency after reaching six years of MLB service time. Of those six years, three are generally paid at the league minimum of $700,000, with the remainder bound by salary arbitration between the team and player (MLB.com). The MLB Draft is still in place as well, with strict limits for total bonuses that teams can offer to avoid competitive disadvantages. A system to cap teams' bonuses to international signings has also been implemented.

When a player reaches free agency following their sixth year of service, they are eligible to be signed by any team. Their salary is determined by the free market, no longer subject to the restrictions of the Collective Bargaining Agreement. Teams with better financial resources are

more able to sign their players to lucrative contracts or lure opposing players away from their former teams, creating a competitive imbalance. This paper intends to discover strategies to mitigate this imbalance.

Chapter 3

Literature Review

Moneyball

To field a competitive ball club, small market teams with fewer resources must recognize inefficiencies in the market, whether that be through free agency, drafting, or another method of player acquisition. This struggle gained mainstream popularity in 2003, when Michael Lewis authored *Moneyball: The Art of Winning an Unfair Game*, and again in 2011, when Brad Pitt headlined a film based on the book. The book and film detailed the Oakland Athletics' 2002 season, in which General Manager Billy Beane and assistant Paul DePodesta (known as Peter Brand in the movie) find that players with high on-base percentage (number of times reaching base divided by number of plate appearances) are more affordable than players that provide similar value in other categories.

Determining a causation of success can be difficult, however. In *Moneyball and the Baseball Players' Labor Market*, Holmes, Simmons, and Berri (2018) found that the "Moneyball" strategy did not result in any substantial long-term correction to the free agency market for high OBP hitters. Perhaps opposing general managers had more time to analyze the true makeup of Oakland's roster, which included sluggers Miguel Tejada and Eric Chavez, each of whom slammed 34 home runs, contributing to Oakland's fourth place finish on MLB home run leaderboards. The Athletics also possessed a deadly pitching trio of Tim Hudson, Barry Zito, and Mark Mulder that contributed to their top three finish team ERA. Former MLB pitcher Mitch Williams went as far as to say, "What Oakland won they didn't win because of sabermetrics.

They won because of Mulder, Hudson, Zito and Tejada" (Barra). Theories that Oakland was successful strictly because of their on-base percentage focus too much on narratives from Lewis' book and not enough on actual data from the 2002 season.

Shifts and the Big Data Baseball Era

In 2013, the Pittsburgh Pirates ended a drought of 20 seasons without a winning record, finishing with 94 wins and the fifth-best record in baseball. The Pirates overcame a relative lack of resources by identifying an area that traditional counting stats had long oversimplified: defense. In *Big Data Baseball*, author Travis Sawchik describes how Pittsburgh was a pioneer in using defense shifts to improve their team.

The Pittsburgh teams of the mid-2010s succeeded with a simple formula: quality starting pitching + quality relief pitching + quality defense + adequate offense (featuring perennial MVP candidate Andrew McCutchen) = wins. Maximizing value on the pitching side of the ball, however, involved getting the most possible out of the defenders available. The Pirates increased their defensive shifting, placing players in positions where the batter tended to hit the ball. The team's defensive positioning consistently turned ground balls into outs, contributing to the team's 3.26 ERA, good for third in all of baseball (Baseball Reference). Since then, shifts have skyrocketed in baseball. Data from *The Bill James Handbook*, as shown in Figure 1, reveals that there are now more than eight times the total number of shifts than the 2013 season.

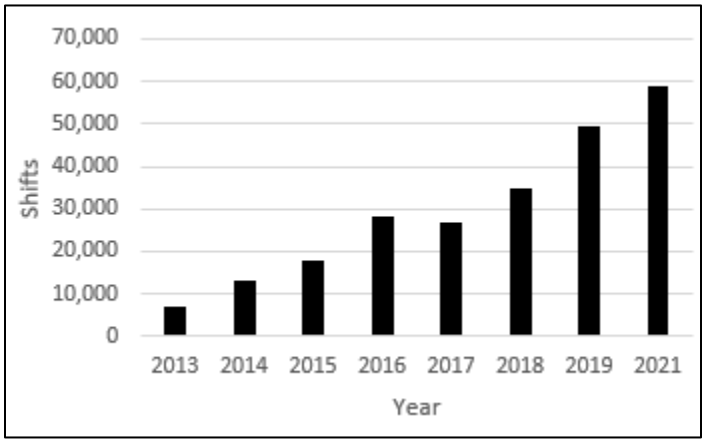

Figure 1: Total MLB Shifts (2013-2021)

James also deduced that in 2021, "there were 22 percent more hits taken away by the shift than given away by the shift" (Stark). Moore (2020) found that batting average on balls in play (BABIP) for ground balls has decreased every year since 2017, providing more evidence that shifts are providing a substantial advantage to the pitching team. It doesn't end there, either. Bouzarth, et al. (2020) took a mathematical programming model to shifts, creating a proof of concept for the optimal positioning of defenders for a given batter that could lead to even further reductions to BABIP. The shift has fundamentally changed baseball to the point that Major League Baseball has instituted a new rule requiring two infielders in the dirt on each side of second base, effectively eliminating the traditional shift (Drellich). This change limits future opportunity for small market teams to gain an advantage through positioning.

Contracts and Roster Management

There has been plenty of research on gameday strategies and traits to look for when building a team, but research on roster breakdown has not been as highly publicized. However, it can be just as insightful as other categories. Solow and Krautmann (2020) found that contracts of

over three years tended to be “overpayments,” situations where the money paid to the player eclipsed the revenue that the player was projected to generate the team, even without the benefit of hindsight. However, this research does not necessarily imply that long-term contracts are a bad idea. Excess productivity from pre-arbitration players allows teams to sign players to negative expected value deals while remaining profitable. While signing players to negative expected value deals sounds like a bad idea, that risk is spread over the course of the contract. For a team looking to win now, the short-term upside of an elite player may be worth the prospect of overpaying down the line. Of course, this capability is only magnified in larger markets.

Years before, Krautmann (2016) also found that teams of all market sizes are generally risk averse. Given two players with similar results but higher variability, teams will pay the more consistent player more money. This poses an interesting dichotomy: signing higher variability players could be an inefficiency for small market teams to exploit if the player can be expected to present the same value as a consistent player over the long term, but small market teams can be most drastically impacted by taking risks that don’t work out.

Small market teams have made ends meet in a variety of on-field ways, but their off-the-field roster construction strategies could stand a second look. Much of the past research on contracts is not actionable. For example, despite long-term contracts yielding negative financial value, it is unreasonable to expect teams to generate equivalent win totals by exclusively signing players to short-term deals. In this paper, I utilize past salary breakdowns and standings data to determine if there are effective roster strategies that small market teams can execute to bring their clubs success.

Chapter 4

Data and Methodology

Data Gathering

The goal of this paper is to identify trends in salary and roster composition that correlate with successful teams, with a focus on small markets. To reach this goal and provide flexibility while working with the data, it was important to work out of two unique datasets: one with team performance data and one with player salary data. All data spans from 2012 to 2019, the eight years marked by the start of ten-team playoffs and ending with the final season before the 2020 COVID shutdown.

Each row in the team data represented one season for one team. Each row included the following categories: team, year, wins, losses, win percentage, rank (out of 30 MLB teams), payroll, and market size. MLB clubs were assigned to one of three market sizes, small, medium, and large, based on the *Chicago Tribune's* 2021 estimates on franchise valuations. Other metrics were considered for the market size variable, such as the size of a team's media rights deal or the number of people in a team's media market. However, these categories were susceptible to outliers, such as the St. Louis Cardinals, whose storied history has resulted in a large, wide-ranging fanbase despite their relatively small local market. Franchise valuation better captures a team's revenue streams, on which salary is dependent.

The player data includes entries for each active player in each year of the dataset. Rows include the year, the player, the player's salary in that year, their team, their service time, and their salary. I also calculated a field to show where a player's salary ranked on their team. While

the data attached to a specific player does not matter, it is valuable when grouped by team and year, and provides some unique insights on the composition of individual teams that cannot be found on a standings page.

All analysis was done off a main sheet that was generated by grouping player data by team and year. For variables such as team payroll, the top 25 salaries on for a team in that year were summed together, because a team's active roster consists of only 25 players. This eliminated the possibility of higher payrolls strictly because a team had more players on the injured list, or promoted more players throughout the year. The R code used to pull the rosters and perform the analysis can be found at https://github.com/kjk5884/HonorsThesis.

Payroll

During my research I placed a focus on unique factors in roster construction that could impact team success. I did preliminary research on basic, easily accessible information such as player salary and team payroll. I tested a variety of research hypotheses regarding salary, including:

H_A: Payroll affects winning percentage and/or playoff odds for MLB teams

Intuitively, higher payroll would be expected to lead to more wins. A team with more financial resources can afford to sign better players and extend the contracts of their current talent. However, the overall purpose of this paper is not to point out that having more money is a good thing. The purpose is to find if alternative measures exist in roster construction that can make small market teams more competitive.

Service Time

The first area explored was a team's average service time. In Major League Baseball, players accumulate one day of service time for each day they spend on an MLB roster. If they are on a roster for 172 days, they are said to have accumulated a full year of service time. For example, Jamie Moyer entered the 2012 season for the Colorado Rockies with a service time of 22.126, which means that he had been rostered for 22 full seasons and 126 days in another season. However, even though service time appears to be a real, rational number, it is not. After Moyer was rostered for 46 games in the 2012 season, his service time rolled over to 23.000 rather than 22.172. The system works like time. After 1:59, the clock reads 2:00, not 1:60.

Because service time is not a real number, it cannot be summed and averaged the way that an ordinary number could. To transition service time to a workable number, I created *adjusted service time.* This normalizes days as a fraction of the 172-day league year.

$$Adj.Service\ Time = Yrs + \frac{Days}{172}$$

For Jamie Moyer, adjusted service time looks like Table 1.

Table 1: Jamie Moyer Service Time Adjustment

Service Time	Years	Days	Adjusted Service Time
22.126	22	126	22.733

A player's service time determines a variety of benefits, such as what year the player reaches salary arbitration and free agency and how large of a pension and 401K contribution they receive from the league. For my purposes, service time is a metric to show how experienced a team is. It is possible that stocking a team with younger players or older players correlates with

success, or there could be a quadratic effect with having more players at their "peak" (traditionally aged 27-32 seasons) leading to success.

H_A: An MLB team's average adjusted service time impacts their winning percentage and/or playoff odds

High Paid Players and Salary Distribution

When financial resources are limited, teams are often faced with a question: invest in a few top tier players, or build out strong depth and hope that a few players become everyday contributors. To test whether one strategy beats the other, I looked at the number of players on each team making ten million dollars or more in a given season, a variable that I called *Highly Paid Players*. Table 1 reveals that over this period, each team has had several players meeting this number, and league-wide the total has ranged from roughly 100 to 150.

Table 2: Ten Million Dollar Contracts, MLB (2012-2019)

Year	Ten Million Dollar Contracts
2012	101
2013	107
2014	125
2015	140
2016	139
2017	153
2018	146
2019	133

The highly paid player number does a great job quantifying how teams are prioritizing top tier talent, but what it does not show is how that stacks up with the resources allocated to the rest of the team. For that, I came up with a new variable called *Players to 50%,* which shows the minimum number of players that take up at least half of a team's payroll. Take for example the 2014 Brewers, who had five players taking up 50% of their payroll, shown in Table 3.

Table 3: Milwaukee Brewers Players to 50% (2014)

Player	Annual Salary
Aramis Ramirez	$16,000,000
Matt Garza	$12,500,000
Rickie Weeks	$12,000,000
Yovani Gallardo	$11,500,000
Kyle Lohse	$11,000,000
Total to 50%	$63,000,000
Overall	$113,417,600

Of course, $63 million is not exactly half of the $113 million overall salary. The summed salary will be always be slightly higher than half of the payroll, because the total cuts off only after the halfway point has been surpassed. Finding a way to quantify the resulting error could be an opportunity for future research.

Players to 50% could be another interesting indicator for team success. Is a top-heavy team with two to three players taking up half of the payroll the best strategy, or is it better to spread that money out to several players? Individual teams during the period ranged from a minimum of two players to 50% (2013 Minnesota Twins and 2019 Detroit Tigers) to a high of eight players to 50% (2016 Kansas City Royals).

Linear and Multiple Linear Regression

Our first response variable, winning percentage, is quantitative and ranges from 0 to 1. Fitting a linear regression can be problematic for proportional response variables, because it can result in values that fall outside of the 0 to 1 range. However, it is still reasonable to use linear regression with proportional data if all response proportions fall between 0.2 and 0.8, or 0.3 and 0.7 if being more restrictive (Grace-Martin). For the years in the dataset, all teams but two fall between 0.3 and 0.7, and the 2018 Orioles and 2019 Tigers fall barely outside that range at 0.290 and 0.292 respectively. As a result, we can use linear regression to predict winning percentages without generating impossible predictions below 0 or above 1.

A multiple linear regression is structured as

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1}, \qquad i = 1,2,\ldots,n$$

where:

y_i = the response variable
$\beta_0, \beta_1 \ldots \beta_{p-1}$ = the unknown slope/intercept values
$x_{i,1}, x_{i,2} \ldots x_{i,p-1}$ = the predictor variables

Within a regression, beta (β) values act as the multipliers to the predictor variables. They can indicate the direction of the correlation (positive or negative). Naturally, a beta value of a larger magnitude will have a greater impact on the response variable. However, you cannot compare beta values across predictor variables to determine which is more significant. To determine what predictor(s) in the full model should be utilized in the final, we need to run a regression with your predictor and response variables to generate test statistics. The p-values associated with the test statistics can be compared to determine significance. In this paper, I will be using an alpha level of 0.05 to determine the significance of the p-values generated. All regressions were performed in R.

Another metric for the success of the model is the coefficient of determination, or R^2. A model's R^2 value ranges from 0 to 1 and indicates the percentage of variation in the response variable that is explained by the predictor variables. A high R^2 shows that the model fits the data well. Because win/loss percentage is a complex response variable determined by a multitude of factors, any model produced would not be expected to have a high R^2 value. There is also "Adjusted R^2," which normalizes the R^2 using the number of predictor variables. Both metrics were used to determine the effectiveness of the models.

When applying my data to the linear regression equation, I used the following predictor variables:

- Payroll
- Adjusted Service Time
- Highly Paid Players
- Players to 50%

I also considered squared versions of each variable to check for quadratic relationships between the variables. Ultimately, the proposed equation looked like this:

$$Win\ \% = \beta_0 + \beta_1 * Payroll\ (in\ millions) + \beta_2 * Service\ Time + \beta_3 * Highly\ Paid\ Players + \beta_4 * Players\ to\ 50\% + \beta_5 * Payroll\ (in\ millions)^2 + \beta_6 * Service\ Time^2 + \beta_7 * Highly\ Paid\ Players^2 + \beta_8 * Players\ to\ 50\%^2$$

Multiple Linear Regression Assumptions

In order to utilize linear regression, we must check our desired models against a set of assumptions. A variety of tests can be run to determine if these assumptions are met.

1. The mean of the response variable is a linear function of the predictors at each set of values of the predictors
    a. Individual regressions of each predictor on the response
2. The individual errors (E[*Winning Percentage*] – *Winning Percentage*) between the data and regression are independent
    a. Residuals vs. Fits Graph
3. The individual errors between the data and regression are normally distributed
    a. Normal Q-Q Plot
    b. Shapiro-Wilks Test
4. The individual errors between the data and regression have equal variances
    a. Residuals vs. Fits Graph
    b. Breusch-Pagan Test

Another thing to watch out for in the model is multicollinearity, which is when multiple predictors are correlated with each other. This can lead to problems in interpreting slope coefficients, as it may be unreasonable to set other predictors constant when varying one predictor. I expect some of the teams' data to show multicollinearity, such as *Payroll* and *Highly Paid Players.* Intuitively, it makes sense that a team with more players making ten million dollars or more will have a higher payroll. The variables are also directly tied together, as a team cannot add a highly paid player to their roster without boosting their payroll by ten million or more. Assuming the correlations follow my expectations, I will only keep one of these predictors in our final model so that we can avoid redundancies and easily interpret the results.

Multicollinearity can also be checked using the variance correlation factor (VIF). VIF analyzes a predictor's uniqueness by regressing it on the remaining predictors and utilizing the resulting R^2 value.

$$VIF = \frac{1}{1 - R^2}$$

As a rule of thumb, a VIF of over 5 for a predictor indicates that the model is highly correlated. For the purposes of my research, I will use a cutoff of 5 to determine multicollinearity in the model.

Logistic Regression

Another way of measuring team success is whether the team made the playoffs in a given year. This is a binary variable, with a team making the playoffs receiving a "1" and a team missing the playoffs receiving a "0." The best way to estimate the probability of a team making or missing the playoffs based on a given criteria is through logistic regression. The base model for predicting a team's playoff chances is:

$$Log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 * Payroll\ (in\ millions) + \beta_2 * Adjusted\ Service\ Time + \beta_3 * Highly\ Paid\ Players + \beta_4 * Players\ to\ 50\%$$

For a logistic model, a change in a predictor variable yields a corresponding change in the "log odds" of the response variable, which is the natural logarithm of the probability of an event happening (π) divided by the probability of the event not happening (1 - π). You can then find π arithmetically through:

$$\pi = \frac{Odds\ Ratio}{1 + Odds\ Ratio}$$

In my model, π is the probability of a team making the playoffs.

Assumptions for logistic regression are far less strict than the assumptions for linear regression. One assumption that we must still check is multicollinearity, which was discussed in the MLR Assumptions section. Because the predictor variables are the same in both the linear and logistic regression models, I will use the same VIF calculations for the logistic model. I have also removed all squared terms from the prediction model, because I cannot check for a quadratic relationship between a predictor and a binary variable.

Model Selection

To create a final model, it is important to use a process to eliminate variables that do not add value to the model. I am utilizing Akaike information criterion (AIC) to determine which models to consider. AIC is calculated using the number of parameters in the model and the model's maximum likelihood estimate (which tests goodness of fit). There are three common ways to find a final model:

Forward Selection: Regress all independent variables individually on the dependent variable. Select the variable with the lowest AIC (X_i) that fits under the selected alpha level. Run another set of regressions with two independent variables: X_i and each of the other independent variables. Select the added variable with the lowest AIC (X_j) that improves the model. Continue adding variables until there are no more variables whose addition to the model improves AIC.

Backward Selection: Regress all independent variables together on the dependent variable. Select the variable with the highest AIC that does not improve the model and remove it from the model. Continue removing the highest AIC model and re-running the regression until reaching the highest AIC model

Bidirectional Stepwise Selection: Follow the directions of forward selection, but if adding one variable yields lower AICs from other added variables, remove the previous variable from the model. Continue the process until there are no more variables whose addition to the model is yields a better AIC.

I utilized bidirectional stepwise selection to view as many model possibilities as possible before selecting the final model. Using bidirectional stepwise selection, we seek to find the model that has both statistically significant predictors and the highest R Squared value.

Chapter 5

Results

Linearity Assumption

To see which of the predictor variables we should move forward with before our final model selection process, we must first analyze the linearity of each predictor, along with the proposed overall model. In Chapter 4, I discussed how *Team Payroll* would be expected to correlate directly with *Winning Percentage*, and Figure 2, a linear regression with the period's full data shows that this is, in fact, the case.

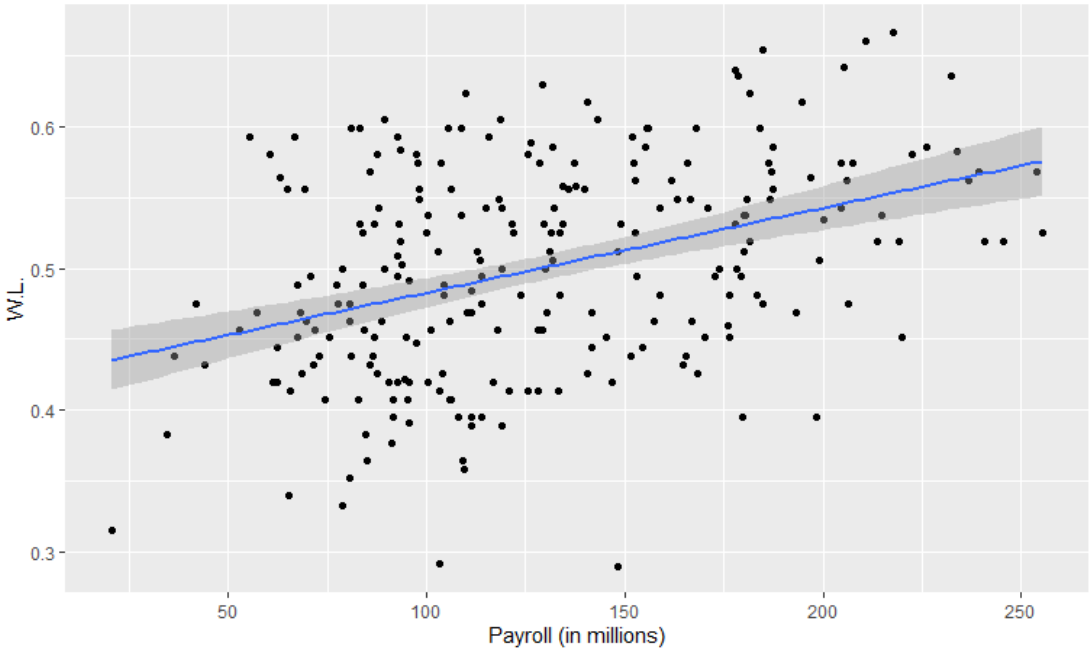

Figure 2: MLB Winning Percentage vs. Payroll (2012-2019)

Correlation does not equal causation, and money certainly isn't everything, but it is reasonable to conclude that generally, higher payrolls are related to positive outcomes on the baseball field, and it would be reasonable to include in the final model.

Next, we look at how *Highly Paid Players* correlates with *Winning Percentage.* Figure 3 shows that there is a statistically significant positive linear relationship because winning and the number of highly paid players on a team.

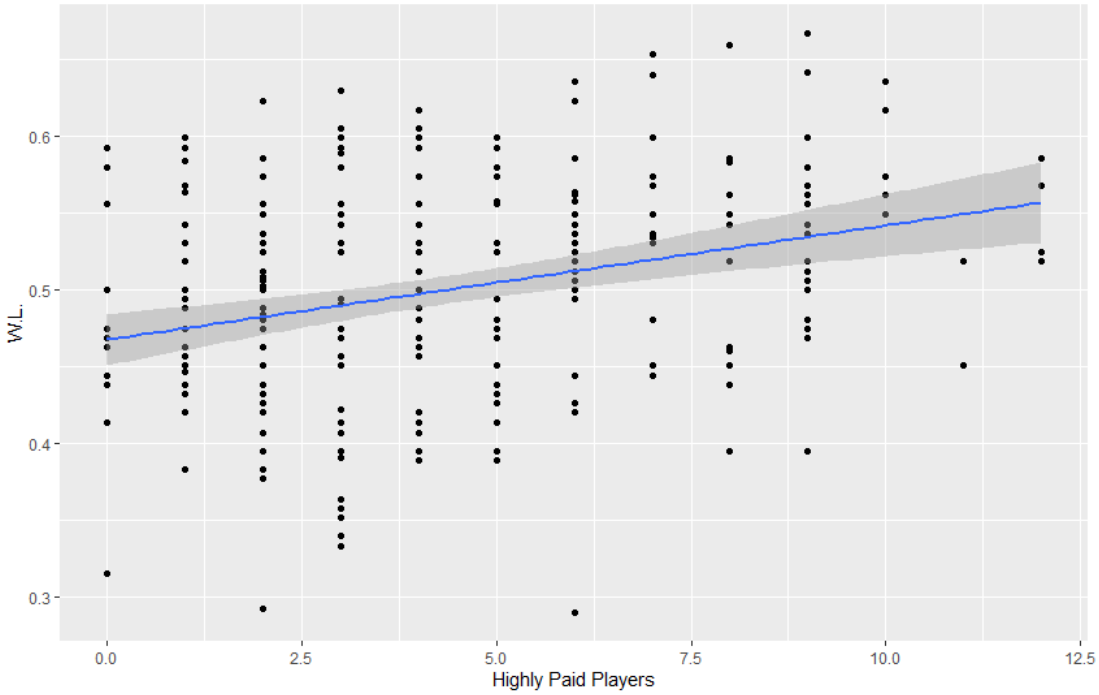

Figure 3: MLB Winning Percentage vs. Highly Paid Players (2012-2019)

Figures 4 and 5 show that our final two variables, *Adjusted Service Time* and *Players to 50*, have a similar positive correlation with *Winning Percentage* at a statistically significant level.

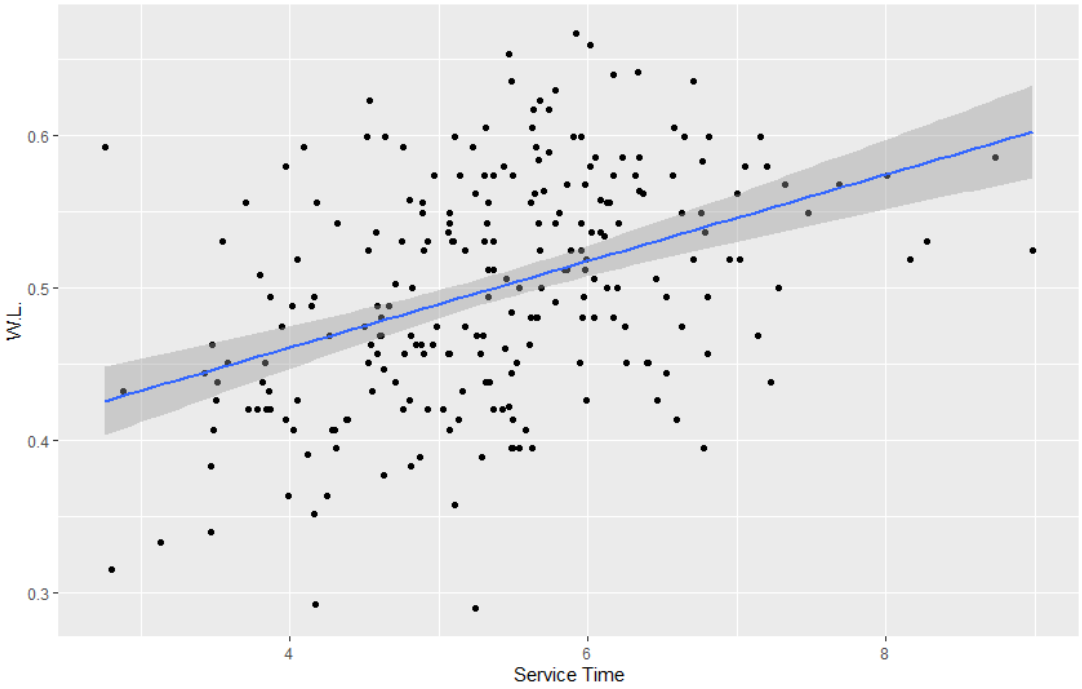

Figure 4: MLB Winning Percentage vs. Adjusted Service Time (2012-2019)

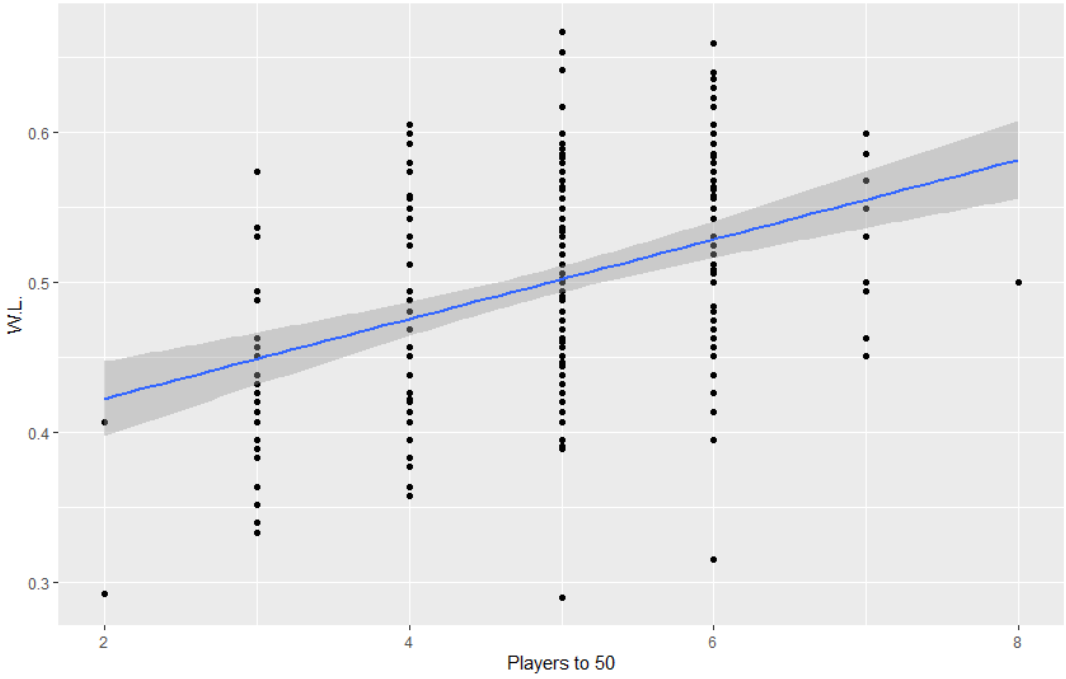

Figure 5: MLB Winning Percentage vs. Players to 50 (2012-2019)

The charts for *Adjusted Service Time* and *Players to 50* both show a potential quadratic relationship with *Winning Percentage.* A quadratic relationship for *Adjusted Service Time* makes sense, because players tend to peak at a certain age, and those with higher service time may see their talents diminish. A quadratic term for *Players to 50* is also understandable, as it implies a "sweet spot" where a team is effectively spreading its payroll across its roster, without completely avoiding investment in better, higher salaried players. Because the *Payroll* and *Highly Paid Players* graphs do not show evidence of a quadratic relationship, I will not add their quadratic terms to our model selection process.

Independent Errors

To check the independence of errors, we can compare the residuals to the fitted values of the initial model including all predictors and the two squared terms to see if there are patterns.

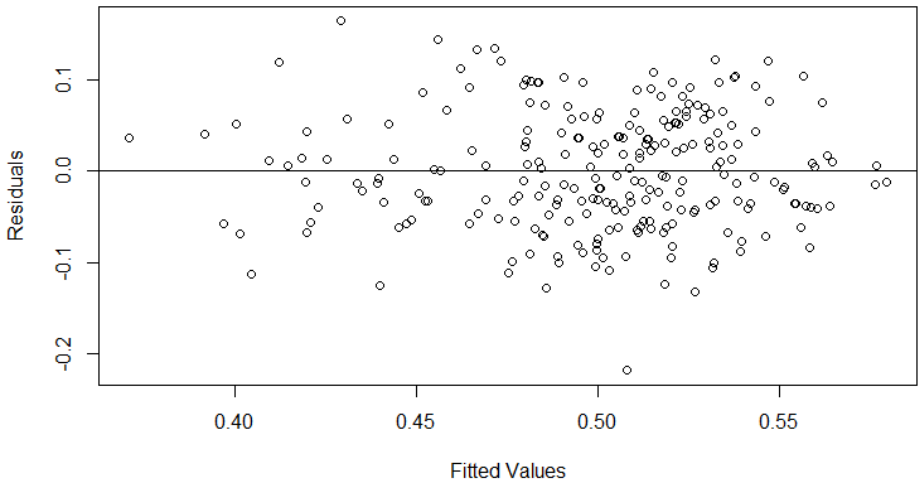

Figure 6: Residuals vs. Fitted Values, Full Model

Based on Figure 5, there does not appear to be a relationship between the errors in the model with all the predictor included, meaning we can continue.

Normal Residuals Assumption

Using the Normal Q-Q Plot of Residuals in Figure 6, you can see that the residuals follow a straight line and appear roughly normal.

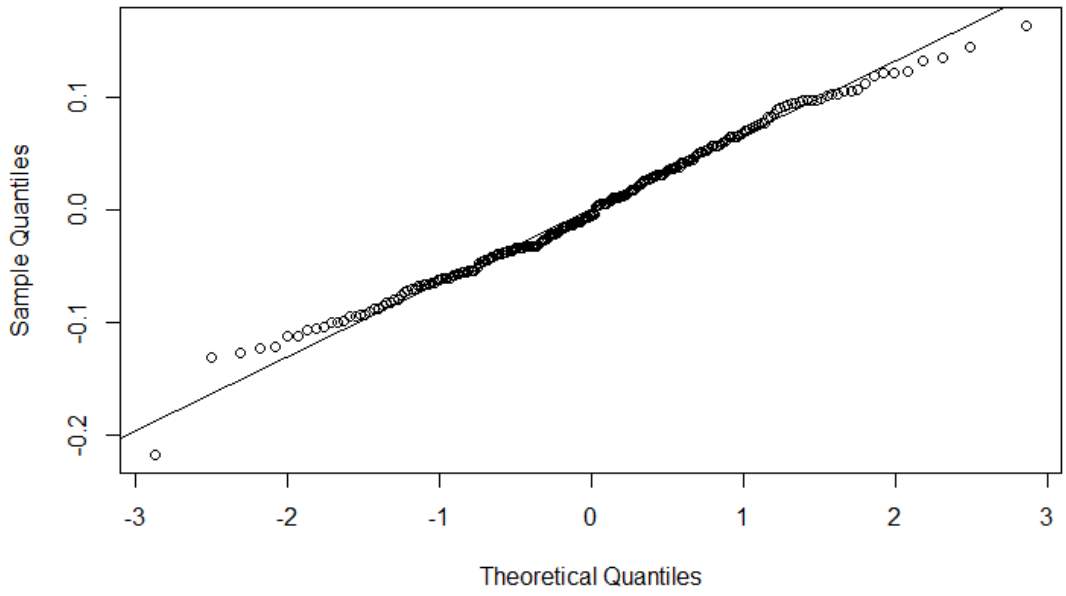

Figure 7: Normal Q-Q Plot, Full Model

I confirmed this result using a Shapiro-Wilks Test in Table 4, with the hypotheses:

H_O: The residuals follow the normal distribution

H_A: The residuals deviate from the normal distribution

Table 4: Shapiro-Wilks Test Result

Shapiro-Wilks Test Statistic (W)	P-value
0.99154	0.1813

With a p-value of 0.1813, we fail to reject the null hypothesis and do not have statistically significant evidence that the residuals differ from the normal distribution.

Constant Variance Assumption

The variance of the model can be observed in Figure 5, which shows that the variances are spread roughly evenly across all levels of fitted values, confirming the assumption. We can also test constant variance using a Breusch-Pagan Test, as in Table 5.

H_O: The residuals have a constant variance

H_A: The residuals do not have a constant variance

Table 5: Breusch-Pagan Test Results

Breusch-Pagan Test Statistic (BP)	Degrees of Freedom	P-value
7.5869	7	0.3704

With a p-value of 0.3704, we fail to reject the null hypothesis and have no evidence that the residuals differ from a constant variance.

Multicollinearity

As stated in Chapter 4, it is important to check the model for multicollinearity to avoid making the model difficult to interpret. In Table 6, we can check the correlation between all variables.

Table 6: Predictor Correlations, Full Model

	Payroll	Highly Paid Players	Adjusted Service Time	Players to 50
Payroll	1	0.92	0.76	0.36
Highly Paid Players	0.92	1	0.68	0.27
Adjusted Service Time	0.76	0.68	1	0.51
Players to 50	0.36	0.27	0.51	1

As expected, it appears that *Payroll* and *Highly Paid Players* are highly correlated, with a correlation of 0.92. It is also interesting to note that there appears to be a correlation between *Adjusted Service Time* and *Payroll*. This makes sense because the payroll structure of baseball, discussed in Chapter 1, allows players to increase their earning potential through avenues such as arbitration and free agency as they accrue service time. However, this correlation is not strong enough to draw any immediate conclusions. Generating a table of VIF values can drive further insights.

Table 7: VIF Table, Full Model

Payroll	Highly Paid Players	Adjusted Service Time	Players to 50
8.472	6.739	2.847	1.395

Adjusted Service Time and *Players to 50* does not show any abnormal correlation based on Table 7. Tables 8 and 9 show that after removing each of *Payroll* and *Highly Paid Players,* the variables with the highest multicollinearity, we see that the remaining variables do not exhibit multicollinearity.

Table 8: VIF Table, Model Without Payroll

Highly Paid Players	Adjusted Service Time	Players to 50
1.916	2.412	1.381

Table 9: VIF Table, Model Without Highly Paid Players

Payroll	Adjusted Service Time	Players to 50
2.409	2.847	1.361

In my final model, I will include only one of *Payroll* and *Highly Paid Players* to avoid multicollinearity and facilitate easy interpretation.

Linear Regression Models

Before utilizing a model selection tool, I ran a base model to determine the overall predictive merit of the initial variable selection.

Table 10: Model with All Predictors

Variable	Slope	T-Statistic	P-Value
Intercept	0.136	1.648	0.101
Payroll (in Millions)	8.632×10^{-4}	**3.415**	**0.001**
Players to 50	0.0774	**2.753**	**0.006**
Highly Paid Players	-0.0102	**-2.626**	**0.009**
Adjusted Service Time	0.0218	0.748	0.455
Players to 50 Squared	-6.444×10^{-3}	**-2.246**	**0.026**
Adjusted Service Time Squared	-0.001237	-0.484	0.629

Statistic	Value
P-value (Regression)	3.883×10^{-14}
R Squared	0.2732
Adj. R Squared	0.2545

The base model in Table 10 shows that I have selected variables with predictive value, with an overall p-value of 3.883×10^{-14}. The R Squared value of 27.32% from Table 4 shows that over a quarter of the variation in winning percentage can be explained by the selected independent variables. The relatively small gap between the R Squared and Adj. R Squared shows that most of the independent variables in the model are adding predictive value to the model.

Adjusted Service Time and *Adjusted Service Time Squared* do not have p-values over 0.05, which indicates they do not make a statistically significant difference in the dependent variable when included in a model with the other independent variables. However, this does not mean that they will not be used in the final model. We also see that *Payroll* and *Highly Paid*

Players both add predictive merit to the model containing the other predictors. When selecting a final model, I will only use one of these predictors to avoid multicollinearity.

To find a final model, I used a model selection tool. As stated in Chapter 4, I utilized bidirectional stepwise selection to find a final model. I ran these results twice, once including the *Payroll* variable and once including the *Highly Paid Players* variable. These models can be observed in Table 11 and Table 12.

Table 11: Stepwise MLR Model Selection, Highly Paid Players Included

Variable	**Slope**	**T-Statistic**	**P-Value**
Intercept	0.1710	2.693	0.00758
Players to 50	0.0808	2.977	0.00322
Players to 50 Squared	-0.0066	4.051	6.92×10^{-5}
Adjusted Service Time	0.0187	-2.397	0.0173
Statistic		**Value**	
P-value (Regression)		1.643×10^{-13}	
R Squared		0.2325	
Adj. R Squared		0.2228	

Table 12: Stepwise MLR Model Selection, Payroll Included

Variable	**Slope**	**T-Statistic**	**P-Value**
Intercept	0.2079	3.321	0.00104
Players to 50	0.08188	3.048	0.00256
Players to 50 Squared	-6.465×10^{-3}	4.465	1.24×10^{-5}
Payroll (in millions)	4.167×10^{-4}	-2.347	0.01975
Statistic		**Value**	

P-value (Regression)	3.262×10^{-14}
R Squared	0.2431
Adj. R Squared	0.2335

Based on the higher R Squared and Adjusted R Squared values, I selected the model with *Payroll* to predict *Winning Percentage.* The final regression is as follows:

$$Win\ \% = 0.2079 + 0.0004167 * Payroll\ (in\ Millions) + 0.08188 * Players\ to\ 50\% - 0.006465 * Players\ to\ 50\%^2$$

The *Payroll* term suggests that when holding the salary distribution of the roster constant, adding one million dollars of payroll could be expected to increase *Winning Percentage* by 0.0004167. Based on the MLB's standard 162 game season, one win is the equivalent to 1/162, or 0.006173 winning percentage points. To yield the change in *Winning Percentage* equivalent to one win, a team would need to increase *Payroll* by roughly $14.814 million. Of course, spending the money just to spend it would not yield any substantial benefit. The model relies on the assumption that an organization is spending money in a way that they deem optimally benefits their team. This is a reasonable assumption, as it is how teams are incentivized to behave in real life.

We can investigate the model further by looking at a use case. For example, let's view the 2019 National League East Division, which has a variety of teams with different levels of success and different market sizes, including the World Series champion Washington Nationals.

Table 13: 2019 NL East Winning Percentage Predictions

Team	Payroll (millions)	Players to 50	Projected Winning Percentage	Actual Winning Percentage	Difference
Atlanta Braves	156.02	5	0.521	0.599	-0.078
Washington Nationals	207.49	4	0.518	0.574	-0.056
New York Mets	129.43	6	0.520	0.531	-0.011
Philadelphia Phillies	178.08	7	0.538	0.5	0.038
Miami Marlins	80.69	3	0.429	0.352	0.077

Table 13 brings a couple of interesting factors to light. First, the model has a tough time predicting outliers near winning percentages of .400 and .600. However, despite the high residual for Miami, the model does know that a payroll of below 100 million and only three players to 50% is not a recipe for a winning. It is also noteworthy that both teams that vastly outperformed their projections were led by young stars on cheap contracts. According to Baseball Reference, the Braves' Ronald Acuña Jr. generated 5.1 wins above replacement (WAR) and the Nationals' Juan Soto was worth 5.0 WAR. However, neither player made more than one million dollars. This is accounted for in the model by facilitating a higher *Players to 50%* number, but there may be a larger effect that could be considered in future research.

It is also valuable to note that despite all the generated prediction models having relatively low R Squared values (maximum of roughly 27%), the model still provides important insights due to the small margin of error in MLB seasons. The difference between making the playoffs and missing the playoffs can be a matter of one or two games, a gap that embracing a prediction model could help resolve.

The only quadratic relationship that had a statistically significant impact in the model was *Players to 50 Squared*. Because of the squared term, we can find the optimal number of *Players to 50* by taking the derivative of the model with respect to *Players to 50.* Optimally, a team will take up half of their payroll with roughly six players

$$\frac{dy}{dx}(0.08188 * Players\ to\ 50\% - 0.006465 * Players\ to\ 50\%^2) = 0$$

$$(0.08188 - 0.01293 * Players\ to\ 50\%) = 0$$

$$Players\ to\ 50\% = 6.333$$

Another item of note in the final model lies in what is not included. It is interesting that the model selection that included *Highly Paid Players* did not yield a model that included the *Highly Paid Players* variable, but rather included *Adjusted Service Time. Adjusted Service Time* appeared to also have moderate correlation with *Payroll,* so it seems that its inclusion in models with *Payroll* may be redundant.

It is also important to consider factors that exist outside of the model and what is realistic. For example, team performance can be optimized by spending an infinite amount of money, but even large market teams have spending restraints based on the owner's wealth and team revenue generation. One must also consider that teams have varying degrees of commitment to winning now versus winning later. Stockpiling younger players with low service time can and arguably should be a determinant in team composition for the long-term success of the team, but this model only looks to optimize short-term performance and does not take long-term strategy into account.

Logistic Regression Model

Just as in linear regression, the multicollinearity assumption is violated by *Payroll* and *Highly Paid Players.* I utilized stepwise model selection to run through models with either of the variables, yielding two models as shown in Tables 14 and 15.

Table 14: Stepwise Logistic Model Selection, Highly Paid Players Included

Variable	Slope	Z-value	P-Value
Intercept	-4.849	-5.200	1.99×10^{-7}
Players to 50	0.234	1.477	0.1397
Adjusted Service Time	0.544	3.336	8.49×10^{-4}
Statistic		**Value**	
AIC		286.1	

Table 15: Stepwise Logistic Model Selection, Payroll Included

Variable	Slope	Z-value	P-Value
Intercept	-3.753	-4.884	1.04×10^{-6}
Players to 50	0.321	2.160	0.0307
Payroll (in millions)	0.011	3.443	5.76×10^{-4}
Statistic		**Value**	
AIC		285.65	

Based on the AIC from the regression models, the model with *Payroll* included is the best model. This results in the following regression equation:

$$Log\left(\frac{\pi}{1-\pi}\right) = -3.753 + 0.011 * Payroll\ (in\ millions) + 0.321 * Players\ to\ 50\%$$

The slopes of the predictors can be predicted as having that much impact on the log odds of making the playoffs when that predictor increases by one unit and the other predictor is held constant.

We can look at the logistic model applied to the same 2019 National League East Division in Table 16 to see how it looks in application.

Table 16: 2019 NL East Playoff Predictions

Team	Payroll (millions)	Players to 50	Log Odds	Probability of Making Playoffs	Did they make the playoffs?
Atlanta Braves	156.02	5	-0.43178	0.394	Yes
Washington Nationals	207.49	4	-0.18661	0.453	Yes
New York Mets	129.43	6	-0.40327	0.401	No
Philadelphia Phillies	178.08	7	0.45288	0.611	No
Miami Marlins	80.69	3	-1.90241	0.130	No

The model marks the Braves, Nationals, and Mets as having a slightly below average chance at making the playoffs, the Phillies at having a slightly above average chance, and the Marlins having a very low chance. Disparities in these odds and the ultimate playoff bracket come from other factors, such as the quality of a team's player development, coaching staff, and even luck.

Small Market Exclusive Model

One of the goals of this paper is determining how small market organizations specifically can succeed based on roster composition data. It is possible to train a model for small markets

specifically by utilizing the subset of the initial data consisting of small market teams. Of course, this sample size is smaller, which limits the possibilities of the model. However, it is still worth investigating to see how substantial of an impact the predictors have when limited strictly to small markets.

We start, once again, with assumptions. When investigating linearity, it is interesting to note that unlike the model with all teams, the *Payroll* and *Highly Paid Players* variables do not have a linear relationship with winning percentage at the 0.05 alpha level.

Table 17: P-Values when Predictor is Regressed on Winning Percentage

Predictor	P-value
Payroll	0.237
Highly Paid Players	0.59
Adjusted Service Time	0.0313
Players to 50	0.000494

Looking at the two variables that do remain in Figures 8 and 9, you can see that there may still be quadratic relationships that would indicate an optimal *Players to 50* and *Adjusted Service Time* value. Thus, I will include the squared terms in the model selection process, assuming other assumptions are met.

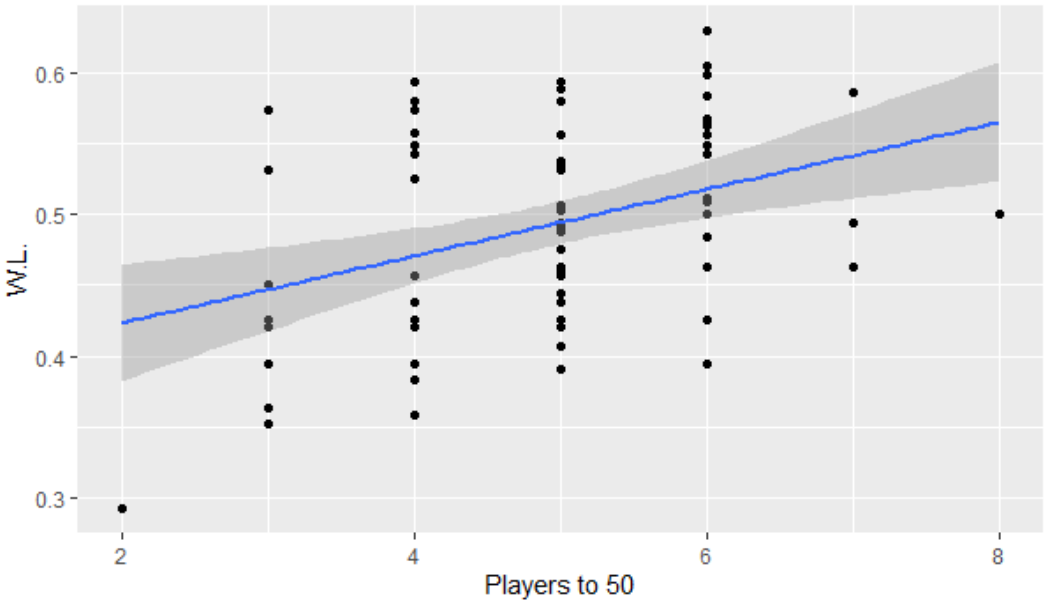

Figure 8: Winning Percentage vs. Players to 50 (Small Markets, 2012-2019)

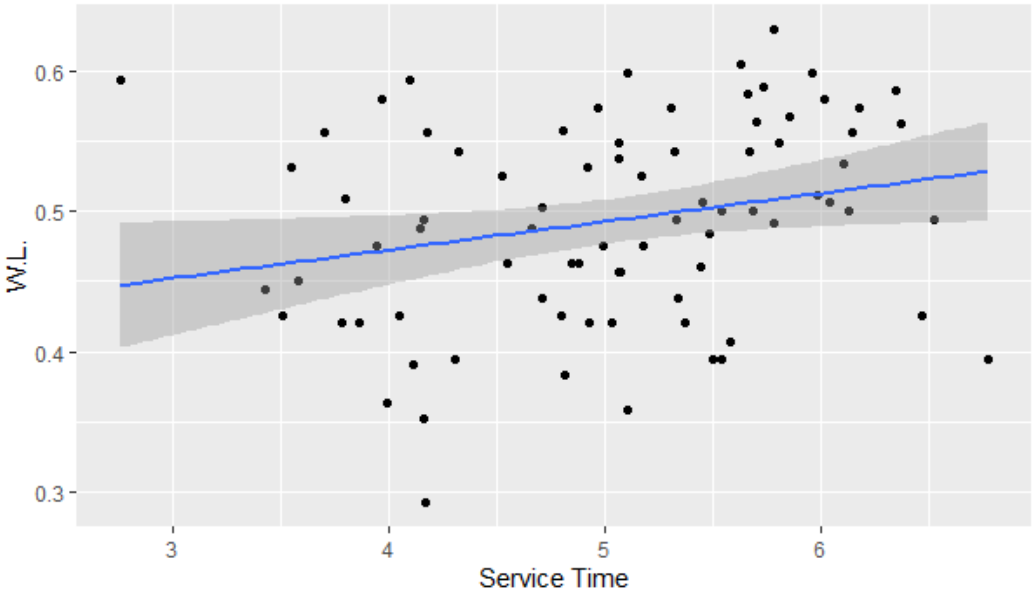

Figure 9: Winning Percentage vs. Adjusted Service Time (Small Markets, 2012-2019)

We can then take a brief look at the remaining assumptions, which are all met. The Residuals vs. Fits graph in Figure 10 shows that the errors are evenly distributed and the

Shapiro-Wilks Test and Breusch-Pagan Test in Table 18 satisfy the normal residuals and constant variance assumptions, respectively.

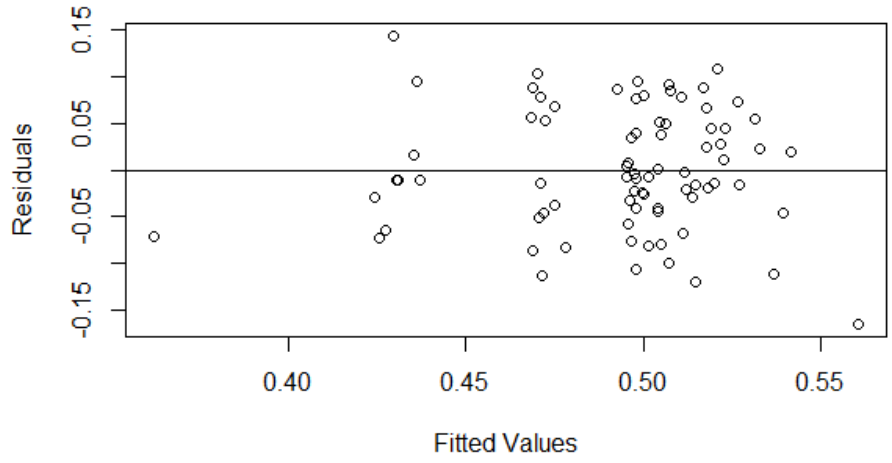

Figure 10: Residuals vs. Fits, Small Market Model

Table 18: Assumption Tests, Small Market Model

Assumption	Test	P-Value	Conclusion
Normal Residuals	Shapiro-Wilks Test	0.4871	No evidence that residuals deviate from normal distribution
Constant Variance	Breusch-Pagan Test	0.05572	No evidence that residuals have nonconstant variance

These results allow me to move forward in the model generation process. For small market teams, I am selecting from *Adjusted Service Time, Players to 50%,* and their respective squared terms.

$$Win\ \% = \beta_0 + \beta_1 * Adjusted\ Service\ Time + \beta_3 * Players\ to\ 50\% + \beta_4 * Adjusted\ Service\ Time^2 + \beta_5 * Players\ to\ 50\%^2$$

Table 19: Stepwise Selection Model, Small Markets

Variable	Slope	T-Statistic	P-Value

Intercept	0.197	2.010	0.0479
Players to 50	0.1013	2.472	0.0156
Players to 50 Squared	-0.007981	-1.919	0.0587

Statistic	**Value**
P-value (Regression)	0.0003985
R Squared	0.184
Adj. R Squared	0.163

This selection in Table 19 yields the resulting model of

$$Win\ \% = 0.197 + 0.1013 * Players\ to\ 50\% - 0.007981 * Players\ to\ 50\%^2$$

This model is so simplistic and covers so few possibilities that it's fair to question what kind of predictive merit it has. The *Players to 50 Squared* term does not even make a statistically significant difference in the dependent variable when included in the model with the *Players to* 50 at the 0.05 alpha level, despite AIC marking it as the best model. The model's R Squared of 18.4% is substantially lower than its counterparts in the "all teams" model, which touched over 24%. In fact, the simplicity of the model paired with the fact that the *Players to 50%* variable is limited to natural numbers means there are only a few *Winning Percentage* outputs that this model can reasonably provide. In the eight years of data the model is based on, *Players to 50%* ranged from two to eight.

Table 20: Winning Percentage Outputs, Small Market Model

Players to 50%	Fitted Winning Percentage
2	0.368
3	0.429
4	0.475

5	0.504
6	0.517
7	0.515
8	0.497

The maximum *Winning Percentage* is found at six players to 50%, as shown in the "all teams" model. However, the maximum fitted value of 0.517 is only equivalent to roughly 84 wins, which is not even enough to make the playoffs in most seasons. Overall, this model does not appear to generate much predictive value beyond showing general trends in the *Players to 50%* predictor.

Chapter 6

Conclusion

Results

This thesis intended to find roster composition strategies for Major League Baseball teams to employ to generate a higher winning percentage and a higher probability at making the playoffs. This was accomplished by regressing various roster factors through multiple linear regression (for winning percentage) and through logistic regression (for playoff odds). This data was examined both on a leaguewide basis and limited to small market organizations. Ultimately, I found that the selected roster factors make up a relatively small portion of variation in winning percentage and playoff odds, but some still hold some valuable insights.

The most insightful variable selected was *Players to 50%,* which found a spot in each final model that I generated. This suggested that one of the most important things to consider when putting together a ballclub is how payroll is distributed across the players. Each model showed that clubs should fill up half of their payroll with roughly six players. Less than that would leave the club's talent lacking in other areas, and more than that would spread the club too thin.

One of the main objectives of my research was finding results that could be applied specifically to small market teams. However, the small sample size of team results when limited to small market teams made it difficult for some predictors to pass assumption tests, and the final model selected was a basic quadratic model using *Players to 50%.* The most interesting finding linked specifically to small market clubs was that the *Payroll* variable did not significantly

correlate with winning percentage based on the data. Surprisingly, this indicates that small market teams that spend more money are not necessarily more successful. It is disappointing that I was unable to find a more substantial predictor of small market success, but the lack of answers in my research even further emphasizes the importance of some of the less measurable factors in small market team success, such as player development and coaching ability.

Opportunities and Next Steps

There are several opportunities to drive new insights by taking this research a step forward. The first is the precision of the data. While I made my best effort to ensure all salary and service time data was accurate, it is possible that some salaries could have counted for multiple teams if the player was traded midyear or one team retained salary after a transaction from a previous season. The *Players to 50%* variable could also be improved by finding the percentage of the final player's salary that needs added to yield exactly 50%. This would eliminate the issue of summed salaries exceeding 50% by varying amounts, which was explained in Chapter 4. It could be done in R and would likely make the *Players to 50%* variable more meaningful.

A main goal of any future research would be to uncover predictors that explain more of the variation in winning percentage and playoff odds. There are several factors in roster composition that could have been researched further, such as number of players in pre-arbitration and arbitration and where players were acquired from (internal or external). Predictors could also be weighted based on performance. For example, you could find a team's average adjusted service time per WAR to normalize the *Adjusted Service Time* variable based on who is performing. This could also be done on a per game basis to draw a clearer link between winning percentage and the composition of a team's lineup, rather than the roster overall. It may make more sense to weigh data more heavily from an everyday starter, for

example, than a bench player. I could also train a model on factors outside of roster composition, such as player statistics or amateur draft trends and international signing habits. These predictors were not included in this research largely because of the additional complexities they would have added to the dataset.

There could also be the opportunity for some optimization within the model. It would be interesting to know if there are values that would have yielded greater significance when calculating predictor variables, such as using a different cutoff salary for *Highly Paid Players* or a different percentage for *Players to 50%.* It is even possible that multiple cutoffs could work in the same model, if the multicollinearity assumption held.

BIBLIOGRAPHY

"10.2 - Stepwise Regression: Stat 501." *Penn State: Statistics Online Courses*, online.stat.psu.edu/stat501/lesson/10/10.2.

"10.6 - Highly Correlated Predictors - STAT 462." *Penn State: Statistics Online Courses*, online.stat.psu.edu/stat462/node/179/.

Barra, Allen. "The Many Problems with 'Moneyball'." *The Atlantic*, Atlantic Media Company, 28 Sept. 2011, www.theatlantic.com/entertainment/archive/2011/09/the-many-problems-with-moneyball/245769/.

Bendix, Peter. "The History of the American and National League, Part I." *Beyond the Box Score*, SB Nation, 18 Nov. 2008, www.beyondtheboxscore.com/2008/11/18/664028/the-history-of-the-america.

Berri, David. "Throwback Thursday: The End of the Reserve Clause." *VICE*, VICE Media, 24 Dec. 2015, www.vice.com/en/article/qkydzb/throwback-thursday-the-end-of-the-reserve-clause.

Bouzarth, Elizabeth, et al. "Swing Shift: A Mathematical Approach to Defensive Positioning in Baseball." *Journal of Quantitative Analysis in Sports*, vol. 17, no. 1, 2020, pp. 47–55., doi:10.1515/jqas-2020-0027.

"Branch Rickey." *Baseball Hall of Fame*, baseballhall.org/hall-of-famers/rickey-branch.

"A Complete Guide to Stepwise Regression in R." *Statology*, 25 Aug. 2021, www.statology.org/stepwise-regression-r/.

Cooper, J.J. "A Complete History of the Working Agreement Between MLB, Minor Leagues." *A Complete History Of The Working Agreement Between MLB, Minor Leagues*, Baseball America, 18 Oct. 2019, www.baseballamerica.com/stories/a-complete-history-of-the-working-agreement-between-major-and-minor-leagues/.

Craig, Mary. "Chained to the Game: Professional Baseball and the Reserve Clause, Part Two." *Beyond the Box Score*, SB Nation, 10 June 2017, www.beyondtheboxscore.com/2017/6/10/15766702/curt-flood-mlbpa-reserve-clause-free-agency.

Creamer, Chris. "A Breakdown of Minor League Baseball's Total Realignment for 2021." *SportsLogos.Net*, 17 Feb. 2021, news.sportslogos.net/2021/02/15/a-breakdown-of-minor-league-baseballs-total-realignment-for-2021/baseball/.

Doyle, Pat. “Branch Rickey's Farm.” *Baseball Almanac*, www.baseball-almanac.com/minor-league/minor2005a.shtml.

Drellich, Evan. “Rob Manfred Introduces Pitch Clock and Ban on Shift for 2023 over Player Objections.” *The Athletic*, 9 Sept. 2022, theathletic.com/3580483/2022/09/09/rob-manfred-shift-ban-pitch-clock/.

Fagan, Ryan. “Minor League Baseball Restructuring: Full List of 119 Affiliate Invites Sent Out by MLB Teams.” *Sporting News*, 9 Dec. 2020, www.sportingnews.com/us/mlb/news/mlb-minor-league-baseball-teams-list/h2au0k3zdctw1ogp7l8jl6rij.

“Free Agency: Glossary.” *MLB.com*, www.mlb.com/glossary/transactions/free-agency.

Goldstein, Richard. “Marvin Miller, Union Leader Who Changed Baseball, Dies at 95.” *The New York Times*, The New York Times, 27 Nov. 2012, www.nytimes.com/2012/11/28/sports/baseball/marvin-miller-union-leader-who-changed-baseball-dies-at-95.html?pagewanted=all&_r=0.

Grace-Martin, Karen. “Proportions as Dependent Variable in Regression-Which Type of Model?” *The Analysis Factor*, 2 Nov. 2020, www.theanalysisfactor.com/proportions-as-dependent-variable-in-regression-which-type-of-model/.

“How to Calculate Variance Inflation Factor (VIF) in R.” *Statology*, 29 Mar. 2021, www.statology.org/variance-inflation-factor-r/.

Kloke, Joshua. “Review: In Big Data Baseball, Sportswriter Travis Sawchik Shows How Analytics Transformed the Sport.” *The Globe and Mail*, 26 June 2015, www.theglobeandmail.com/arts/books-and-media/big-data-baseball-its-a-whole-new-game/article25129188/.

Krautmann, Anthony C. “Risk-Averse Team Owners and Players’ Salaries in Major League Baseball.” *Journal of Sports Economics*, vol. 18, no. 1, 2016, pp. 19–33., doi:10.1177/1527002514560577.

McGuffey, Will. “The Importance of 43 Days of MLB Service Time.” *AWM*, 7 Sept. 2022, awmcap.com/blog/mlb-service-time.

“MLB 2022 Payroll Tracker.” *Spotrac*, www.spotrac.com/mlb/payroll/.

Moore, John. “Examining Shift Effectiveness with Batted Ball Data (Part 1).” *BaseballCloud*, 2 Oct. 2020, baseballcloud.blog/2020/10/02/examining-shift-effectiveness-with-batted-ball-data-part-1/.

"The Most and Least Valuable MLB Teams." *Chicago Tribune*, 15 Apr. 2021, www.chicagotribune.com/sports/national-sports/sns-mlb-most-valuable-teams-20210415-qdhb77mrongwpppnhsazo4f3w4-photogallery.html.

"An Overview and Brief History of the Minor Leagues." *Twins Daily*, 11 Feb. 2022, twinsdaily.com/blogs/entry/23186-an-overview-and-brief-history-of-the-minor-leagues/.

"Reserve Clause." *BR Bullpen*, 30 Nov. 2012, www.baseball-reference.com/bullpen/Reserve_clause.

Simon, Andrew. "A History of the Rule 5 Draft." *MLB.com*, MLB, 7 Dec. 2021, www.mlb.com/news/complete-history-of-the-mlb-rule-5-draft-c210225288#:~:text=The%20early%20years,1%20and%20Feb.

Solow, John L., and Anthony C. Krautmann. "Do You Get What You Pay for? Salary and Ex Ante Player Value in Major League Baseball." *Journal of Sports Economics*, vol. 21, no. 7, 2020, pp. 705–722., doi:10.1177/1527002520930259.

Stark, Jayson. "What Would Happen If Baseball Killed the Shift?" *The Athletic*, 21 Feb. 2021, theathletic.com/3138898/2022/02/21/what-would-happen-if-baseball-killed-the-shift/.

ACADEMIC VITA

KYLE JOHN KROBOTH

EDUCATION

The Pennsylvania State University — **University Park, PA**
Smeal College of Business — *Class of 2024*
Master of Business Administration

The Pennsylvania State University — **University Park, PA**
Schreyer Honors College | Eberly College of Science — *Class of 2022*
Accelerated Science BS / MBA Program, Statistics Concentration

WORK EXPERIENCE

Frito-Lay — **York, PA**
Supply Chain Intern (Packaging Department) — *May 2022 – Aug 2022*

- Introduced automated system to submit cleaning and inventory reports, improving efficiency on packaging floor
- Created Excel calculator for when to order critical materials, reducing probability of costly production shutdowns

The Hershey Company — **Pittsburgh, PA**
Transportation Planning and Operations Co-op — *Jan 2021 – Jun 2021*

- Developed two multi-level Microsoft Power BI reports to ensure on-time delivery by leveraging unused data
- Accelerated reporting processes by 25% to provide plants and distribution centers with expected delivery updates
- Communicated with carriers to reduce incorrect reason codes for late deliveries by 50% from January to June

Qlicket, Inc. — **Pittsburgh, PA**
Business Development and Customer Success Analyst — *Apr 2020 – Apr 2022*

- Delivered insights on customers' employee satisfaction by scheduling and running targeted polls on tablets at offices and distribution centers for three customers
- Led meetings with Fortune 500 customers to discuss how to improve employee retention using Qlicket's solution

ACTIVITIES & LEADERSHIP EXPERIENCE

Legion of Blue – Penn State Basketball Student Section — **University Park, PA**
President — *Apr 2020 – Present*

- Planned season ticket promotional events, contributing to record-breaking student season ticket sales
- Managed verified social media accounts to generate buzz and engage fans
- Redesigned Legion of Blue website to effectively communicate club's purpose and history

Smeal Business Ethics Case Team — **University Park, PA**
Member — *Aug 2021 – Present*

- Won first place team at International Business Ethics Case Competition (IBECC) for case and solution geared at improving diversity in National Football League (NFL) coaching staffs
- Composed and delivered winning 90-second "pitch" to emphasize importance of ethics in NFL league meetings
- Generated first place solution aimed at mitigating the effects of the Great Resignation in the hotel industry

Penn State Sports Analytics Club — **University Park, PA**
President — *Sep 2019 – May 2022*

- Created and executed strategic growth plan resulting in 200% year-to-year club membership boost
- Conducted workshops to improve members' technical skills in R, Tableau, and Excel

ACHIEVEMENTS, SKILLS, AND INTERESTS

Achievements: 770 GMAT, Evan Pugh Senior Award, 1st Place 2020 CAS Actuarial Case Competition, President's Freshman Award, 3x Dean's List, Schreyer Academic Excellence Scholarship, Balog Science Scholarship, National AP Scholar, 36 ACT (Perfect), 3rd Place 2019 FBLA National Leadership Conference – Organizational Leadership
Hard Skills: Microsoft Office Suite, Microsoft Power BI, Tableau, R & Python experience, Data Analytics
Interests: Pittsburgh & Penn State sports, disc golf, ultimate frisbee, sports card & memorabilia collecting