DSC680-T302

Chris Kellogg

**Is My Team Really Trying to Compete?  
*Using an MLB Team’s Opening Day Player Payroll to Predict Playoff Expectation***

**Business Problem**

Major League Baseball (MLB) teams need to plan a strategy each offseason so they can make decisions about the players they want to have on their rosters going into the season. I’ll attempt to answer this question: *If an MLB team wants to make the playoffs, how much should the team expect to spend on its Opening Day player payroll?*

**Background**

MLB teams spend a lot of money paying players, but player payrolls vary widely from team to team (Jackson, 2024). From the Oakland A’s at almost $63 million to the New York Mets at over $316 million, the 30 MLB teams clearly have significantly different approaches to constructing their rosters.

Teams can carry 26 active players in each game, and the mean annual salary for an MLB player is just under $5 million (Gough, 2024), but the mean player payroll for MLB teams is just $104 million. Teams can use rookie players (who make the MLB minimum salary of $740 thousand) to lower their payrolls and then splurge on high-priced elite talent (like Shohei Ohtani, whose 2024 salary was $70 million).

Fans often want their teams to sign the best players, money be darned. “MLB doesn’t have a salary cap,” they say, “so sign all the players! Pay them what they want; the team is worth billions, so you can afford it.” In reality, MLB teams have many more expenses, like facilities, scouts, clubhouse, office staff, training, travel, utilities, and player services. A team generally has an overall baseball operations budget, and the player salaries all come out of that. Money is finite, after all, but the team that spends it most effectively has a strategic advantage.

If there were a statistical way to figure out what Opening Day player payroll for a team gives an expectation to win enough games to get into the playoffs:

* **owners** would be able to set the baseball operations budget correctly,
* **front office staff** would know who best to target to get the most impact from free agent dollars,
* **development staff** would know how many young players will be needed in the coming years to supplement the roster,
* **players** and their **agents** would have a better idea of how much money teams might be willing to spend,
* **fans** would have a fact-based indicator of whether their team is genuinely trying to win, or if they’re in a “rebuilding” phase, and
* **media** will have good early-season guidance for scheduling late-season games that are likely to have playoff implications.

If there is a dollar threshold for an Opening Day player payroll that indicates that a team is “seriously competing,” that information could increase transparency and temper expectations.

**Datasets**

*Team standings: pybaseball API for Python (*[*https://pypi.org/project/pybaseball/2.0.0/)*](https://pypi.org/project/pybaseball/2.0.0/)I)

I’ll be able to use the historical team Win-Loss records to “draw the line” for which teams would make it into the playoffs. I’ll go back as far as the 2013 MLB season, since that’s the last time that divisions were re-aligned, but I’ll leave out the 2020 season, which was aberrant in several ways (including schedule length and number of wild card teams) due to COVID. I’ll use the current method of determining playoff teams (Major League Baseball postseason, 2024), filling out a 12- team wildcard round. This is a slightly imperfect method, since the process has only been in use since 2022, but there have been either two or three wildcard teams in all of the seasons I’m considering, so the impact isn’t too great.

*Team payrolls: MLB Team Payrolls. (*[*https://www.stevetheump.com/Payrolls.htm*](https://www.stevetheump.com/Payrolls.htm)*)*

Steve O has done a terrific job of compiling and consolidating MLB historical information on his site. His records for Opening Day MLB team payroll go back as far as 1998, so I can easily get the info I need for Opening Day payrolls from 2013 through 2023. Steve’s data is particularly good for me because I’ve specifically chosen to use the Opening Day payroll for each team, since team budgets are generally set during the offseason, and that’s when a front office has the most flexibility to make roster changes with a particular payroll target.

*Data Preparation*

My basic datasets were available through a public API (statsapi, provided by MLB Advanced media) and a public website (www.stevetheump.com), but they each required significant work to be in a usable form for my purposes. I used python in a Jupyter notebook to get the data and build a local SQL database for each table that I needed in order to protect myself from any potential changes to the data or its accessibility.

The API and the webpage use various team names and sometimes prefix them with numbers, but I needed consistency across and inside the data. I mapped the team names to a team abbreviation which will stay consistent between and throughout the datasets.

The webpage was not consistent in the naming for the columns (or even the number of columns). I cleaned any punctuation out of the payroll numbers, and even had to adjust the numbers where they were reflected in millions instead of in dollars.

The webpage tables were not consistent in their use of column headers, sometimes putting them in the table and sometimes not. The webpage also included a “League Average” row in some of the tables. I adjusted the read statements where necessary and dropped the “average” rows, since they’re not useful in this project.

I merged the two separate datasets (wins and payroll) on year and team to create the master dataset for my analysis. For each team in each year, the master dataset shows the opening day payroll, number of wins, and whether the team made the playoffs in that year.

I had intended to use just information for years 2013-2023 (excluding 2020, when a lockout shortened the season). As I collected data, I realized that the website and the API provide everything I needed to include 2024, so I added the extra year.

**Methods**

I had planned on fitting a linear regression to the dataset that would allow me to roughly predict the number of wins that a team could expect in a season, given the Opening Day payroll. Of course there are injuries, under- and over-performance by players, development for younger players, regression for previous high performers, and in-season trades that influence a team’s record. None of that information is known during the off-season, though, when a team sets its budget and makes its big free agent moves; all we have to go on is that Opening Day payroll, but I expected that to be enough to predict a range of outcomes, and that I could reasonably calculate approximately how many wins it usually takes to make the playoffs.

*Number of wins to reach playoffs*

First, I wanted to know generally how many wins it takes to get into the MLB playoffs. It’s not the same every year, of course, but especially with the current format (3 wildcards per League), the number shouldn’t fluctuate wildly. Because it’s entirely possible for a division winner to have fewer wins than a wildcard team from a different division, I decided to plot both the lowest win total to qualify for the playoff and the highest win total to land on the outside looking in.

*Figure 1. How many wins does a team need to reach the playoffs?*A graph of different colored lines

Description automatically generated

I had not anticipated so much variance, but it looks like a team would generally be likely to make the playoffs if they won about 87 games. Math confirmed that the mean win total of playoff teams across both the NL and the AL was 86.5, and the mean win total of AL playoff teams was just under 87.5, so I thought I’d be safe aiming for 88 wins.

*Predicting wins based on Opening Day payroll*

I plotted win totals vs Opening Day payrolls.

*Figures 2&3. Wins vs Opening Day payroll (2013-2024 on left; 2017-2024 on right)*A graph of orange and grey dots

Description automatically generated *A graph of orange and grey dots

Description automatically generated*

On the left, there are LOT of teams on the low end of the payroll scale that were making the playoffs. I thought maybe many of them were earlier in the timeline, when even the highest of payrolls was probably below some of the lowest in recent years, so I narrowed my dataset to include only seasons after 2016. On the right, I saw that it didn’t make much difference. In fact, one of the lowest payroll teams in the last 12 years to make the MLB playoffs was the 2022 Cleveland team that achieved a 92-win season with an Opening Day payroll of just over $42 million!

**Model/Analysis**

I forged ahead with the linear regression. If I were a front office, I’d be using previous data to inform my decisions, set my budget, and expect to spend to a certain level, so I wanted to use previous seasons as my training data and just the last 2 years (2023/2024) as my test set. I expected to fit a linear regression model and use it to know what range my payroll should be in to be a contender next year. The linear regression was not predictive.

*Figure 4. Predicted wins vs Opening Day payroll (2013-2024), with prediction line in blue*A graph of black dots and a blue line

Description automatically generated

A computer code with numbers

Description automatically generated

The coefficient and intercept tell us that we’d expect to win 69 games if we spent no money at all and we expect to add 1 win for every $11 million we add to our payroll. Given that the MLB minimum salary is $740,000 and teams need to fill a 26-man roster, the absolute minimum Opening Day payroll would be about $20 million, and that team should expect to win around 70-71 games. The lowest win totals in the league are typically in the 60s and 70s, so that doesn’t seem crazy. The Mets and Yankees topped the league in Opening Day payroll in 2024 right around $300 million, which by our expectation would yield around 96 wins, certainly enough to be in the top 5 in MLB year in and year out. But there’s just too much variance for our model to be very accurate, with $75M payrolls winning 100 games and $300M payrolls barely cracking the 75-win plateau. In fact, our coefficient of determination is slightly negative, suggesting that a horizontal line at about 87.5 wins would be as good a predictor as the model.

I thought maybe older years were having too much influence on the algorithm, and I tried using just the last 6 full seasons in the regression, but still didn’t find predictive value.

*Figure 5. Predicted wins vs Opening Day payroll (2018-2024), with prediction line in blue*A graph of black dots

Description automatically generated

A computer code with numbers

Description automatically generated

The line got a little steeper and the bottom end started a little lower, suggesting that the “price of poker” has gone up, that a team should expect to spend more money now than they might have 10 years ago. But we still have that small negative coefficient of determination indicating that Opening Day payroll is just not all that explanatory when it comes to predicting a win total for the season.

**Conclusion**

Opening Day payroll alone is not explanatory when it comes to predicting a win total for an MLB team’s season.

**Assumptions/Limitations/Challenges**

I don’t think I faced any particular limitations or challenges in the project. I had access to all the data that I needed for the analysis, and even have access to the same data from separate sources for corroboration. I did make the assumption that Opening Day payroll could be meaningful on its own, disregarding any sort of team chemistry issues, and also ignoring the effect that injuries, under- and overperformance, or player development on a team’s win total in a season. However, those variable are *unknown* when a team has to make its offseason adjustments and contract decisions; I wanted to know specifically if the starting point (Opening Day payroll) was an empirical reason for optimism.

**Future Uses**

In my exploration and analysis, I still found that there’s some kind of relationship between Opening Day payroll and playoff expectancy, as shown here:

A close up of text

Description automatically generated

**Recommendations**

Being able to spend your way out of your mistakes and into more talent by signing high-priced free agents is a clear advantage for the larger market teams. The only reasonable recommendation, therefore, is that **a team should make as much money as it can, and then spend as much of that money on players as it can afford**. Higher payrolls generally lead to more talent, which generally leads to more playoff baseball, which generally produces more income; there’s a compounding effect that’s likely to position the team for even more success in the future.

**Implementation Plan**

While there’s no immediate recommendation produced by this particular analysis, I think there’s a suggestion that there may be a spending threshold that yields some level of playoff expectancy. I’d like to follow up with another project that studies how often high-payroll teams make each round of the playoffs, especially in more recent years. I’d like to analyze if being in the upper 25% of payrolls, a median payroll, or a lower 10% payroll makes a team more or less likely to reach the playoffs. I’d also like to consider whether the previous year’s record or payroll is helpful.

**Ethical Assessment**

My data is not really at risk, since it’s available through other sources if my primary sources become unavailable. There’s little reason for concern about the ethics of using the data, since it’s all publicly available from multiple sources, including through MLB itself. Finally, there’s no likely ethical impact from the results, since all 30 MLB organizations and many independent sites are already doing similar modeling.

**References**

Jackson, S. (2024, September 24). BetMGM. MLB Teams With Highest 2024 Payrolls: Mets on Top.  
<https://sports.betmgm.com/en/blog/mlb/mlb-teams-highest-payrolls-2023-bm15/>

Gough, C. (2024, June 13). Statista. Major League Baseball average player salary 2003-2024. <https://www.statista.com/statistics/236213/mean-salaray-of-players-in-majpr-league-baseball/>

Major League Baseball postseason. (2024, November 2). In *Wikipedia*. <https://en.wikipedia.org/wiki/Major_League_Baseball_postseason>

Timeline of Major League Baseball. (2024, November 8). In *Wikipedia*. <https://en.wikipedia.org/wiki/Timeline_of_Major_League_Baseball>

**Questions**

*Q1. Why use the Opening Day payroll instead of ending payroll?*

I wanted to use a model to determine at the start of a season whether a team is genuinely attempting to be competitive, or whether they have some other motives (like rebuilding the team through development of young players and accepting a poor record to increase the chances of a high draft pick).

*Q2. Is there a way for the model to account for “intangibles”?*

While I’m sure that teams consider things like a player’s work ethic, leadership, coachability, and willingness to put the team’s interests above his own, I’m not aware of a widely available source for that kind of data.

*Q3. Aren’t there stats like WAR that can measure each individual player’s value?*

There are certainly widely-used and respected stats that attempt to quantify each player’s value, but they’re backward-looking. Assembling a collection of last year’s high-WAR players is unlikely to be possible, since other teams want to keep their own high-WAR players, but it also doesn’t account for how those players might fit into a budget. I wanted my model to be based on something that’s entirely controllable by a team.

*Q4. Does the model account for players who underperform or outperform their contracts?*

The model isn’t meant to consider individual performance of a player as much as it’s meant to try to estimate the likely performance of the team as a whole.

*Q5. Since they have more money, wouldn’t the large market teams always win?*

If there were a 1:1 correlation between Opening Day payroll and team wins, yes. But we know that sometimes smaller-market teams with talented youngsters outperform expectations. I just wanted to know how much of the “likelihood of winning” might be accounted for by spending alone.

*Q6. What did your coefficient of determination really tell you?*

The coefficient of determination was tiny in the model. That tells us that Opening Day payroll has essentially no value as a predictor of team wins in a season, mainly because the variance is massive. It’s too easy in MLB to spend big and lose, and it’s entirely possible to win with young players playing at a high level on small contracts. Figures 4 & 5 show us just how many points are nowhere near the model line.

*Q7. The model wasn’t effective. Does that mean that money doesn’t matter?*

It doesn’t mean that spending money doesn’t matter; it means that spending money isn’t the only thing that matters, which is good news for the league and for teams in smaller markets.

*Q8. Was this just wasted time and effort?*

I don’t think so. Data analysis showed that in the last 4 years, more than half of the teams who made the playoffs had an Opening Day payroll above $150 million, and 75% of the League Championship Series and World Series teams came from that group of higher-spending teams. There’s almost certainly something to learn.

*Q9. How do we find out the influence that Opening Day payroll has on making the playoffs?*

I’m planning to follow up with another project that’s aimed more at using Opening Day payroll both in bucketed categories and as a percentile in the league to see if it has value as a classification predictor for playoff, LCS, and World Series appearances.

*Q10. Is there any conclusion to reach from this project?*

The more you look at the data, the way performance generally declines in the later years of players’ careers, and the approach that both big- and small-market organizations take to building an MLB team, the easier it is to see that the biggest advantage of having a lot of money is that it allows you to be wrong more. If a team’s payroll is generally around $120 million and they guess wrong on a $24 million dollar player, they’ve lost 20% of their payroll and have to figure out how to win with only $95 million, a tall order. However, for a team whose payroll is generally around $240 million is wrong on a $24 million dollar player, the impact is only 10% of their payroll and they still have $216 million to work with! The conclusion is simply for a team to make as much as it can and spend as much of that money on players as the team can afford.