Telling Ticket Prices

Ethan Wang
July 24, 2023

Contents

1 Introduction and the Output

For the sake of analyzing an MLB organization's industry "success" there are only a handful of publicly available statistics that can point you in the right direction. When I say success, I am not referring to the number that shows up in the win column, but the number(s) that show up in the balance sheet. Ask a common baseball fan what is important for a baseball team, some might say they need just need to win, some might say they need to make a run in playoffs, some might say they need cute ballplayers. Regardless of how a "successful" ball club is defined to you, we must not forget that MLB is part of the entertainment industry and their job is to entertain the masses and maximize profit while doing so. With this in mind, it becomes easier to understand that seemingly questionable decisions made by front offices are likely justified by the effort to accomplish this goal, and not just to have the best chance of winning. It is true that maximizing profit is in large part a product of playing winning baseball, since MLB derives much of its entertainment value from the art of competition. However, when we look at the 30 teams that make up the league and analyze their unique ways of reaching this dynamic label of "success" it becomes evident that certain circumstances generate subtle differences in business strategy for each team.

In order to provide some background on the topic, the first statistic I would like to discuss is franchise value. There are two distinctions of this number, one being total value and the other being team value. Total value encompasses everything including the equity of team related businesses and real-estate holdings, while team value excludes those things to focus on baseball related financial data. I am referring to team value. When selecting an output/dependent variable of my study, the first statistic that caught my attention was franchise value, but an instant realization that the range and dynamic nature of ownership wealth across the league nullifies the usefulness of this number. Other statistics such as revenue and operating income are analogous to franchise value in this regard, as they are similarly subjective. Not only do these statistics immediately possess bias, but they fluctuate with non-baseball related transactions whether it be TV and brand deals, or something like a change in ownership.

Therefore, the statistic I have chosen as my output variable is yearly average ticket price. Contributors to day-to-day ticket price changes are tangible things like the opponent, the day of the week, and weather. Despite these day-to-day changes, a yearly average for ticket price is the best representation of a franchise's current industry standing. Finding out if that price is derived from their sustained level of performance on the field, ability to compete for a playoff spot, possession of exciting talent, or a blend of all these contributing factors is the goal of this analysis. Disparity in the yearly average ticket price is much more likely to reflect a sustained disparity in demand for an organization's on-field product, whereas individual ticket prices are easily confounded by aforementioned factors. What characteristics of a successful organization most potently influence that success and vice versa for a poorly performing organization? The efficient-market hypothesis, well known in the trading world, states that asset prices reflect all available market information and that prices adjust based on new information as well. While there is not a direct link here, this logic bolsters ticket price as the best suited statistic for portraying an organization's baseball related fair market value. While ticket price is not devoid of subjectivity, it is far less influenced by MLB's economic inequality. Therefore, I will be conducting my research using a generalized additive model with yearly average ticket price as the output variable.

2 The Input Variables

*****ALL FINANCIAL TIME-SERIES DATA HAS BEEN INFLATION ADJUSTED WITH 2023 CPI AS THE BASE YEAR (STUDIED TIME PERIOD: 2006-2022)***

Before transitioning into the analysis, I would first like to discuss the input/independent variables first. The first independent variable in this model that will be plotted against our output ticket price is Pythagorean Win-Loss % Above Lowest Season. Pythagorean Win-Loss % is a statistical estimate of a team's win-loss % based on that team's runs scored and runs allowed in a given season. This input is my catch-all statistic for measuring a team's production and tendency to play winning baseball because it provides an estimate of actual win % (.22 points off of it on average), while also factoring in run production and overall defensive effectiveness. The reason I chose "% above" the team's lowest estimate (from 2006-2022) is because the raw values created multicollinearity with the other variables at a variance inflation factor (VIF) of around 18. Not only does this fix that issue, but it standardizes the value per each organization making the statistic objective to the given ball club being analyzed. The number now represents how much better than their worst season they performed that given year. With this adjustment, VIF for my the independent variables is now much more acceptable as seen in **Table 1**. The next two independent variables in this model are the simplest which are team OPS and WHIP+ERA (the two stats are combined to involve more aspects of pitching efficiency). These are the two most readily available time series statistics that are fairly effective at illustrating a team's isolated offensive and defensive production on the field.

Independent Variable	VIF
Win % Above Lowest Season	2.5
OPS	1.72
WHIP+ERA	2.08
Player Expense Proportion	1.05

Table 1: Model VIF

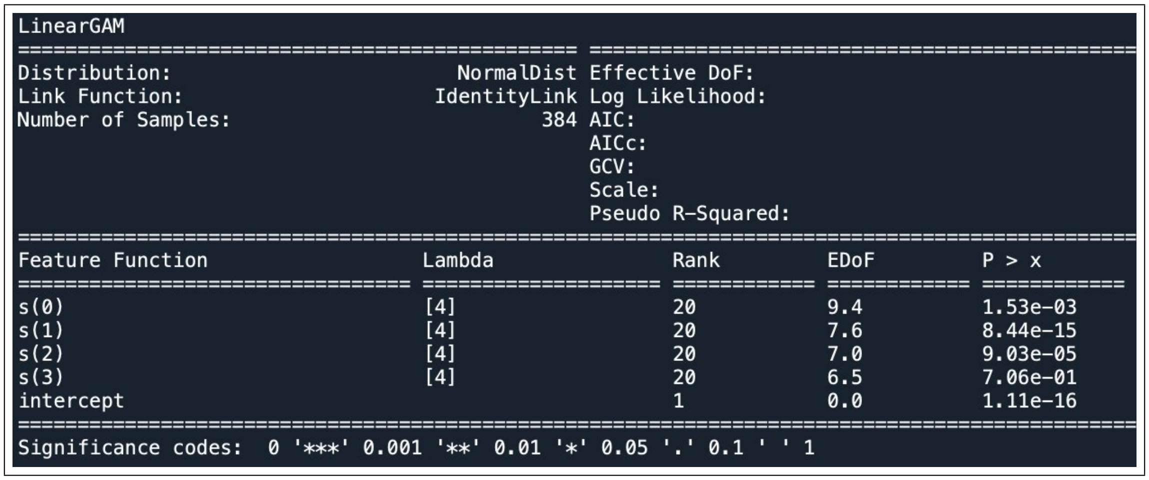
The the fourth and final variable is Player Expenses as a Proportion of Revenue. More wealthy teams have more to spend on contracts, so without scaling the player expenses value in some way, the statistic would be far too subjective. While dividing it by revenue is by no means perfect, it ensures that there will be far less bias towards the big market clubs. Revenue is a decent measure of a team's financial scale so this proportion is a much more comparable statistic than the raw player expenses value. Because there are numerous contributors to a ball club's revenue such as gate receipts, broadcasting deals, and sponsorship revenue I have decided to ignore the potential correlation between revenue and player expense proportion here which is already low at r = 0.213.

3 Regression Model

3.1 Model Selection

I have chosen a generalized additive model for my research for a few reasons. After visualizing the relationship each variable had with my output variable using scatter plots, it became apparent the relationships were too non-linear for standard linear regression. I opted for a general additive model which applies the same principles, but adds splines which take away the linear constraint of standard linear regression and allow the model to curve and bend to more accurately fit the data.

3.2 Model Statistics and Model Accuracy



```
LinearGAM
=============================================== ==============================================
Distribution:                        NormalDist Effective DoF:
Link Function:                     IdentityLink Log Likelihood:
Number of Samples:                          384 AIC:
                                                AICc:
                                                GCV:
                                                Scale:
                                                Pseudo R-Squared:
==============================================================================================
Feature Function                  Lambda               Rank         EDoF         P > x
================================= ==================== ============ ============ ============
s(0)                              [4]                  20           9.4          1.53e-03
s(1)                              [4]                  20           7.6          8.44e-15
s(2)                              [4]                  20           7.0          9.03e-05
s(3)                              [4]                  20           6.5          7.06e-01
intercept                                              1            0.0          1.11e-16
==============================================================================================
Significance codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Model Statistics

In **Figure 1**, s(0) represents my first independent variable: Win % Above Lowest Season. Then, s(1) represents the second one: OPS, s(2) the third: WHIP+ERA, and s(3) the fourth: Player Expense Proportion. EDoF (effective degrees of freedom) represents the extent to which the relationship between that variable and the output is nonlinear, 1 being a completely linear relationship. As we can see, the relationship between Win % Above Lowest Season and ticket price is the most nonlinear at an EDoF of around 9.4. Every one of our variables has an acceptable p-value, with player expense proportion being the largest and closest to insignificance at certain confidence intervals. This being the standout p-value will make sense later on in the analysis section. Finally, the model's Adjusted R2 (shown in summary) is 0.3035. Not bad considering the volume of factors that determine ticket price.

Evaluation Metric	Value
Mean Squared Error (MSE)	142.802
Mean Absolute Error (MAE)	9.871

Table 2: MSE and MAE of Model

In regards to mean squared error and mean absolute error, both are taking averages of the model's prediction error. However, MSE is very sensitive to outliers due to the squaring, so as we can see there is vast disparity between MSE and MAE. This makes sense because organizations like the Yankees have incomparable ticket prices, and organizations like the Cubs maintain a high market value despite struggling statistically throughout much of the studied time period. With MAE being less sensitive to these outliers, we get a value of 10 meaning on average the absolute difference between the model's prediction and the actual value is $10. Nothing groundbreaking, but for the sake of analyzing variable relationships, it is good enough. My analysis begins on the next page.

4 Model Plot Analysis

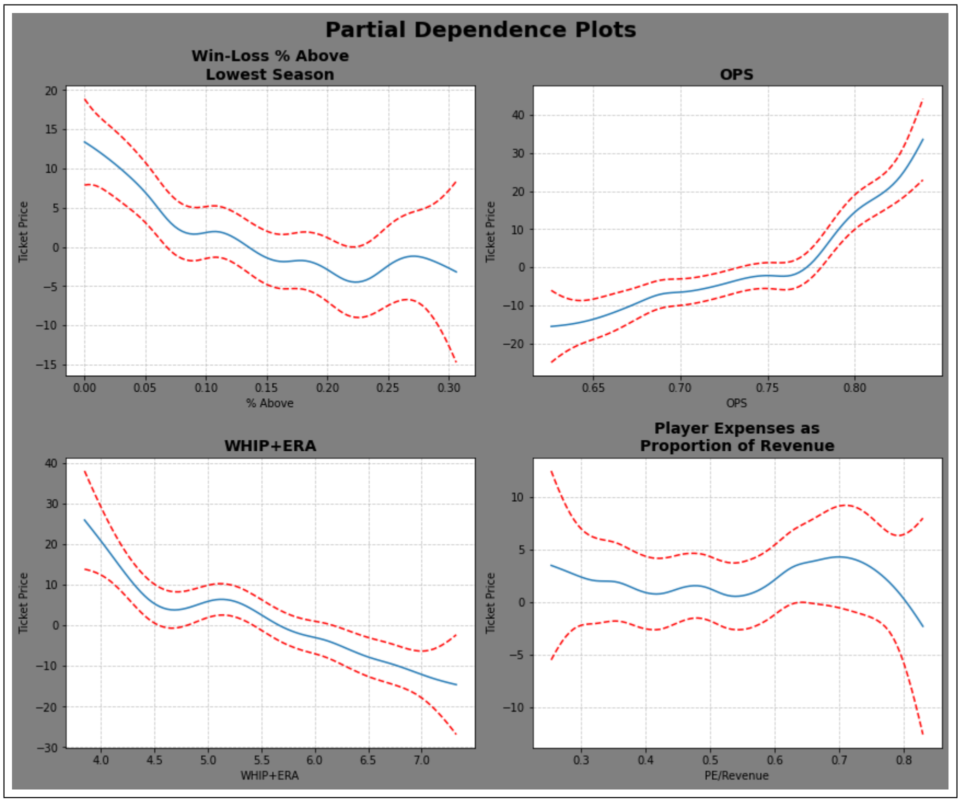


Figure 2: Partial Dependence Plots for each Variable

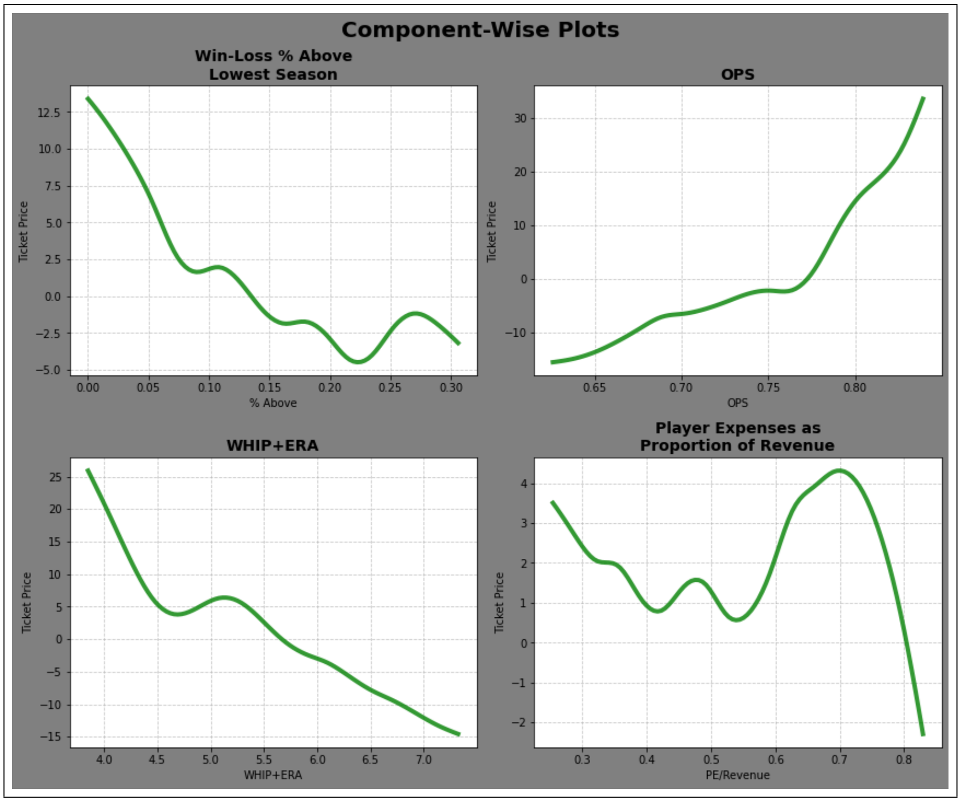


Figure 3: Component Wise Plots for each Variable

4.1 Pythagorean Win-Loss % Above Lowest Year

The plots and model are a culmination of data on every individual MLB team's seasons from 2006-2022 compiled into a dataframe. Firstly, for the plots shown above, a partial dependence plot depicts the relationship between each individual independent variable and the output variable while holding all three other independent variables constant at their mean, and the component-wise plots depict the relationship between the variables while also considering the effect of the other variables at a given point (pay attention to different y-axis scale). The red-dotted lines represent the 95% confidence interval at which that line predicts the ticket price, wider obviously meaning less certainty. Ticket price ventures to negative because the independent variables were standardized with a mean of 0 in order to get all the units to be comparable, but were converted back to their original units on the graph for interpretability, ticket price remains on that standardized scale.

Let's look at the first relationship we have, Pythagorean Win-Loss % Above Lowest Year vs. ticket price. Why in the world is there a clear negative relationship between the two variables in both graphs? When we look at the component-wise plot the relationship seems to become even more dramatic, meaning when the other variables are considered, this relationship becomes even more volatile. Doesn't make sense. However, let's analyze the data back in its time series format to get some insight.

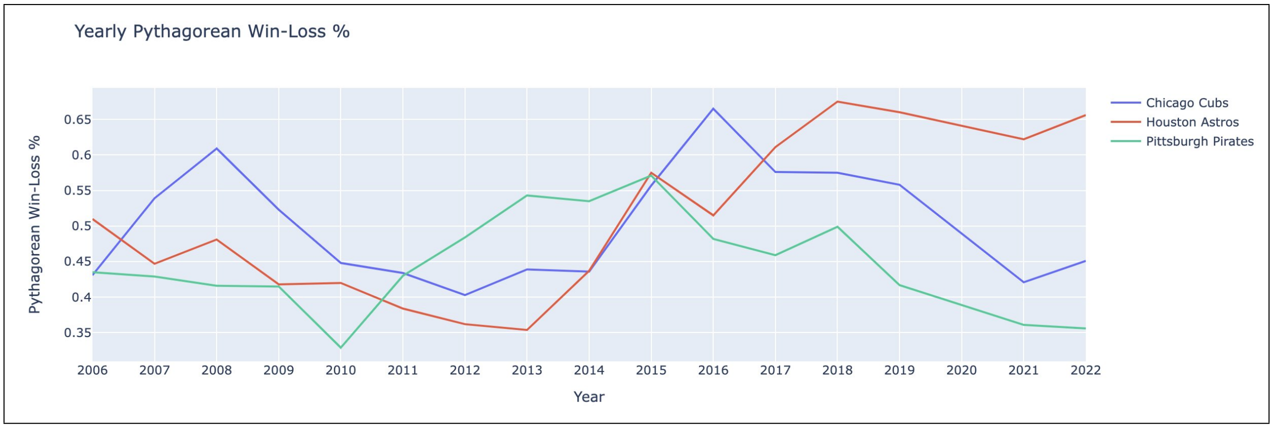


Figure 4: Three Teams - Time Series Win %

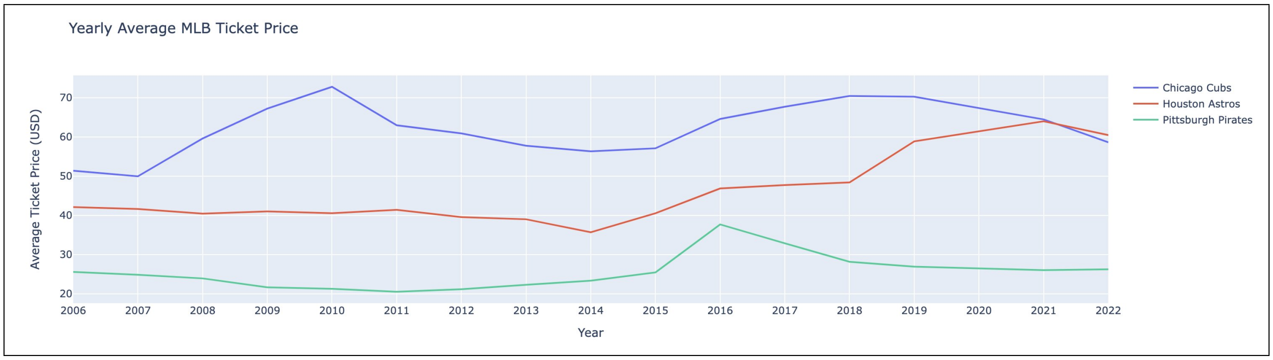


Figure 5: Three Teams - Time Series Ticket Price

The dataframe used to create the model does not have any time series aspect, each entry of data is its own season for a specific team and all the statistics pertaining to that season. Therefore, even if ticket price is up in a certain year, all other variables are likely up as well, meaning the model's output illustrates the importance of relativity, not the weight of a raw value. However, when we look at the data back in its original time series format, you may notice a pattern. I have used only 3 ball clubs as examples due to aesthetic limitations, but many others show a similar trend. It seems that local peaks in ticket price such as the Cubs in 2010, the Pirates in 2016, or the Astros in 2019 seem to lag behind corresponding peaks in Win-Loss % by a year or two. Take a look at the peaks in Win-Loss % such as the Cubs in 2008, the Pirates in 2015, and the Astros in 2018. Looking at the time series data for many different teams elucidates that there are numerous examples of winning impacted changes in ticket price, but they take a year or two to actually come into effect. I will explore this finding deeper by adjusting the data a little. Let us stagger the years by one and see how that impacts this specific relationship. Interestingly enough, in **Figure 6** (on the next page), when I match 2006 Win-Loss % up with 2007 ticket price, 2007 Win-Loss % with 2008 ticket price, and so on, we turn that negative relationship into something that makes more sense. There is a clear positive correlation here with an r-value of 0.29, not bad. It seems that the effect a team's overall performance has on ticket demand is delayed by a year or two. Hence, in order to not let the initial negative relationship continue to baffle us we must remember that the model data is not time series, so each row is only compared to its own row representing a given season for a given team. All this to say it is very possible that due to winning's delayed effect, any random sort of in-year relationship for this specific variable can exist and it doesn't mean much without diving deeper as we did. Something else to notice is that when you look at the component-wise plot (the effect of an independent variable with the effect of the other ones taken into account) the scale on the y-axis shrinks dramatically and the range of y-values decreases significantly, while OPS and WHIP+ERA barely change from plot to plot. This is the reason for the aforementioned "volatility" in the component-wise plot. Considering our exploration into the subtleties of this variable, it is not surprising that when the model plots variables while considering the effect of the other, more straightforward ones, it starts to realize the nature of misattributing its prediction values to variable importance that isn't really there.

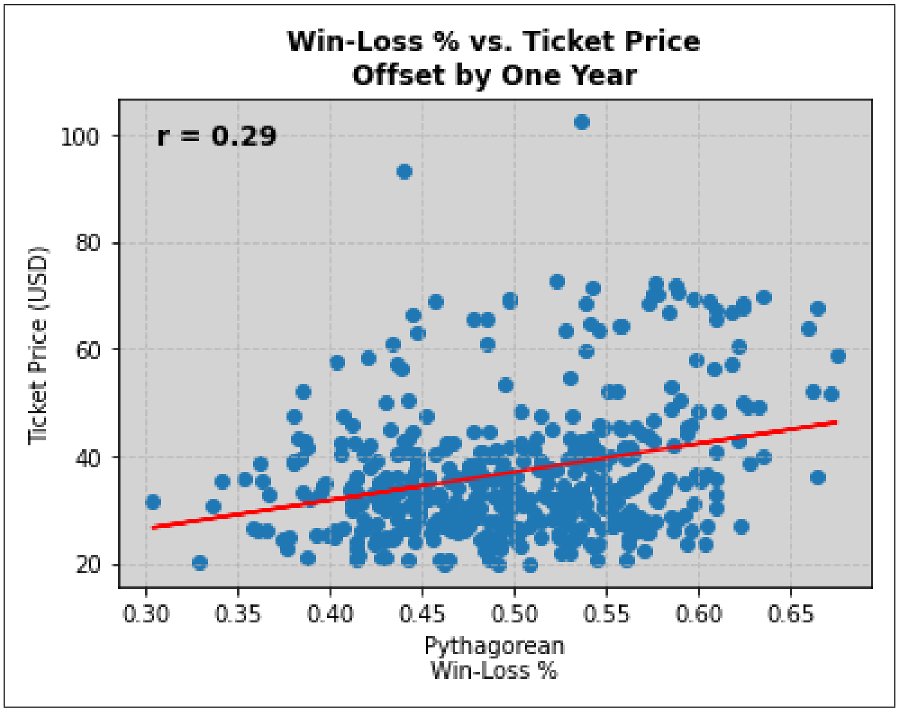


Figure 6: One-Year Offset Scatter plot

4.2 OPS

Now taking a look at OPS, this relationship is not only more linear, the model is far more certain of this being an accurate depiction of the relationship (red-dotted lines). As one might expect, there is a strong positive correlation between increased OPS and increased ticket price. Both partial dependence and component-wise plots tell us almost exactly the same thing, this variable is the most stable. "Well obviously a better hitting team has higher ticket demand", one might say. Pick a random year, 2021 for example, the Reds, Twins, Rays, and Rockies, all sat between .3-.31 points above league average OPS but remained selling tickets around $5-8 less than league average. Data never said the relationship between OPS and ticket price would be that simple. However, reiterating the main concept here, the weight of the variables in each data entry is based on relativity, not the raw value. Remembering this, the strong and direct relationship between OPS and ticket price can likely be attributed to small differentials in price that VERY consistently coincide with small differentials in OPS. Comparing the effect of Win-Loss% and OPS things start to make sense, real-world wise. The buzz that spreads around town about a newly winning team doesn't happen overnight. But more singles, more doubles, more home runs, more run production, more crowd engagement, chants, and energy, all products of higher OPS, are things that would make a game IMMEDIATELY more enjoyable to attend. Take another look at the plots and notice that .765-.770 OPS (on the x-axis of the plots) seems to be the threshold at which this relationship develops an increasingly positive slope. This is also intuitive because that is around the same threshold at which fans and the media begin having reason to label a team and their offense a "true contender".

4.3 WHIP+ERA

Luckily for me, the relationship between WHIP+ERA and ticket price was one that I expected to see. I'm not going to reiterate the logic that carries over from the above analysis of OPS vs. ticket price, but as you can see, with similar significance, WHIP+ERA is almost the exact mirror relationship of the one we just analyzed. Now as we increase WHIP+ERA, the ticket price trends downwards, and even has a similar threshold as OPS where around 4.7 and below, the line takes off at a steeper negative slope than WHIP+ERA values above that. This is very intuitive as well because we know that fans don't generally enjoy watching bombs get dropped on their home stadium, and we can apply the same logic surrounding the threshold of 4.7 WHIP+ERA where we can consider teams falling below that "contenders". Another thing I expected to see was slightly less significance in this prediction line (red-dotted lines a bit wider), due to the fact that increases in this statistic are likely less consistently paired with decreases in ticket price than OPS increases were with increases in ticket price. Most baseball fans will agree that the buzz and satisfaction level of a stadium can stand to improve much more with a dominant, base-pounding offensive performance than a 12Ks, 3 hits allowed performance on the mound. Not too many surprises with this one.

4.4 Player Expenses as a Proportion of Revenue

Now let's take a look at the bottom right plot with player expenses/revenue vs. ticket price. Automatically we can tell that this is the weakest aspect of the model. Now before I had my results, I expected this to be far and away the most important contributing factor to determining ticket price. This is important because my idea for this model was based on trying to prove that spending on fancy, big-name players is the most important contributor to a ball club's crowd engagement. If I had taken raw player expense values, it is likely we would see a very significant positive relationship between it and ticket prices, but only because big market teams have both the most money to spend and the highest ticket prices. But when the value is scaled against something like revenue, this disproportion is taken away. The mass of squiggly lines we see across the board means that the data is clustered in a somewhat random nature as opposed to being arranged in a trend-like manner. Another indicator we have is that the 95% confidence interval range is quite large showing uncertainty. So why did I fail at justifying my initial hypothesis? Let's break it down.

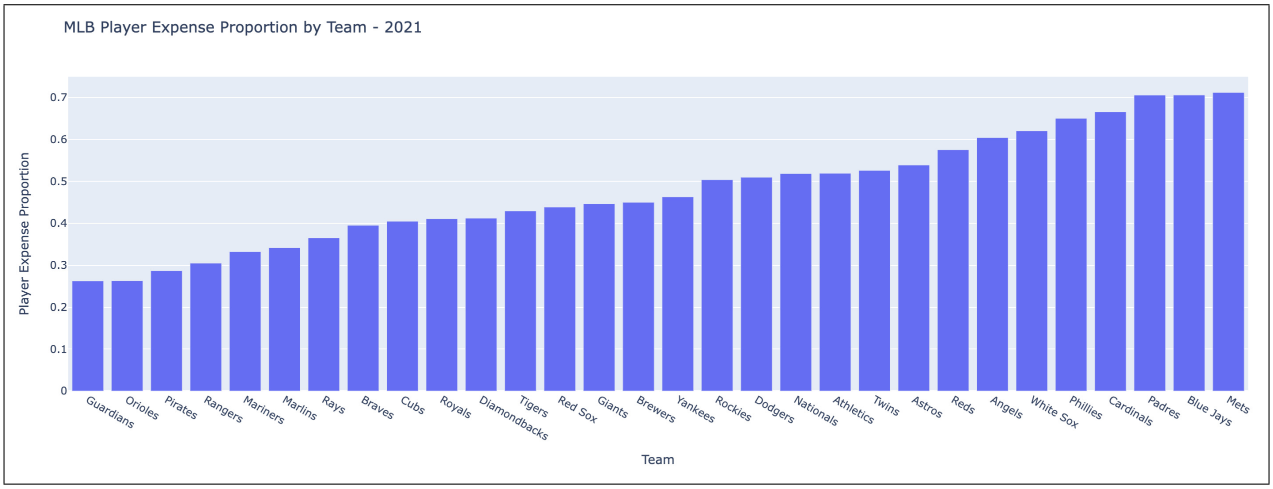


Figure 7: 2021 MLB Player Expense Proportion

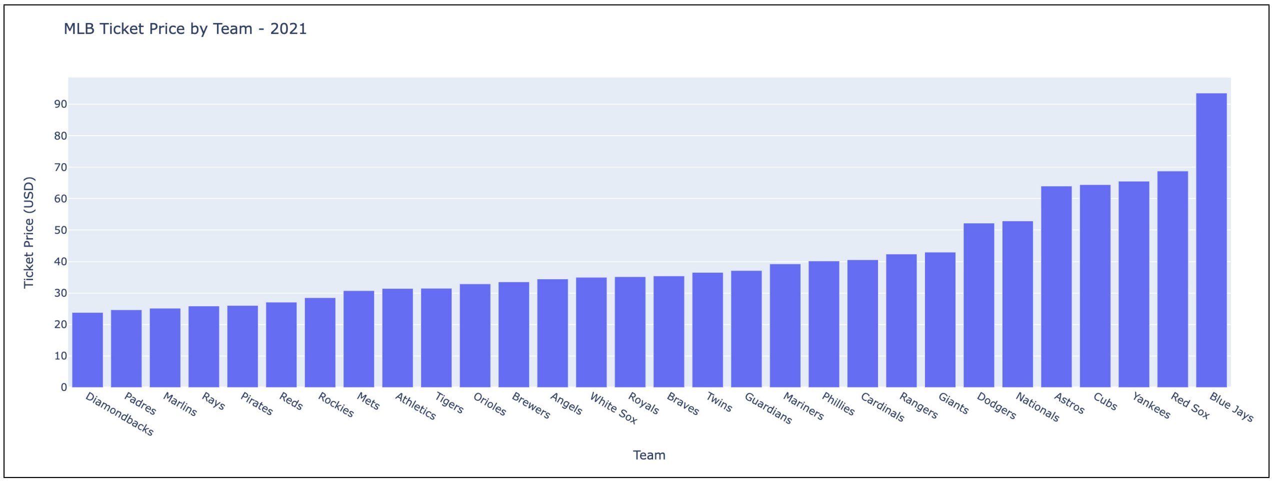


Figure 8: 2021 MLB Ticket Price

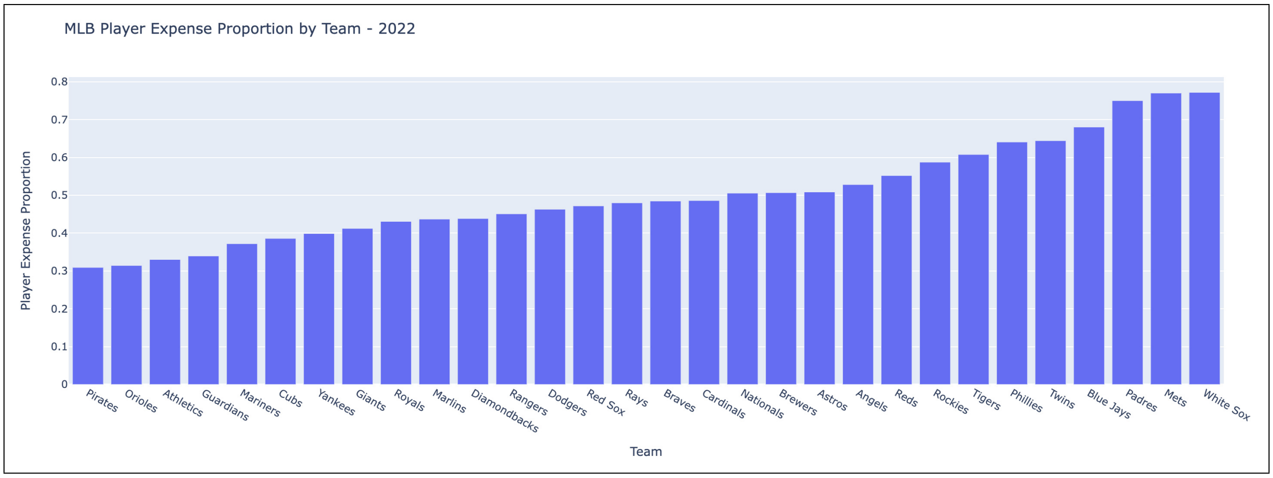


Figure 9: 2022 MLB Player Expense Proportion

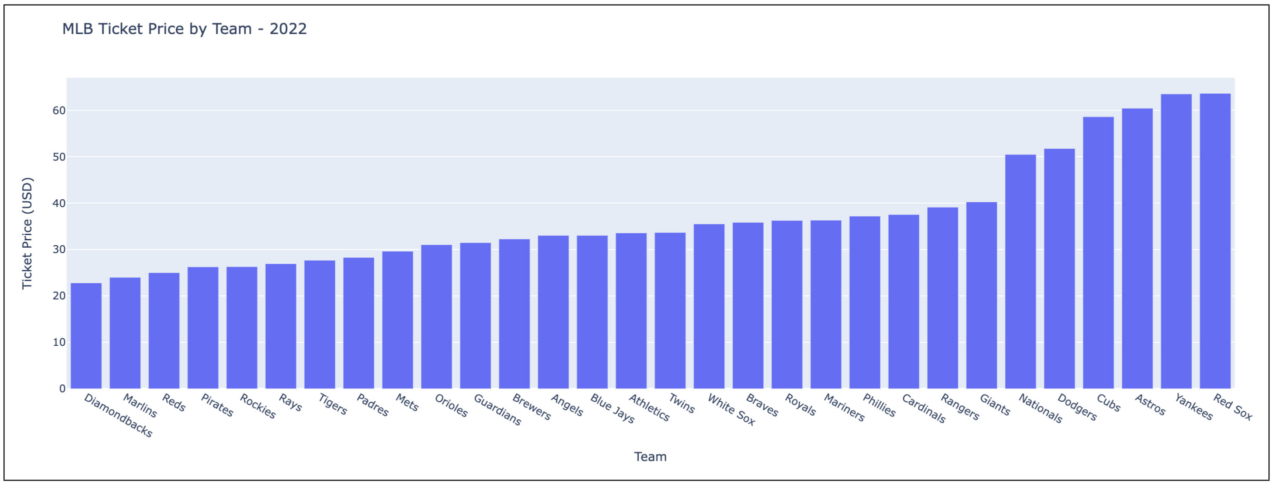


Figure 10: 2022 MLB Ticket Price

Shown above we have two sets of bar graphs, one for 2021 and the other for 2022. Each pair compares player expense proportion and ticket price for the whole MLB in that given year. It is evident that for the two most recent full seasons (and almost all others) there is massive inconsistency in the rankings between two variables, player expense proportion and ticket price. The Mets and Padres are the two teams that I planned on using as examples of depicting the link between spending and team popularity. Both teams were already geographically ripe for fanbase explosions. Even after their acquisitions of Max Scherzer, Justin Verlanerder, Juan Soto, Blake Snell, and many more, both of these teams have grossly underperformed and are continuing to underperform. Hence, I hoped they would be the poster-children of a trend across the MLB that when you spend on those big names, you get financially rewarded regardless of how the team's performance changes. Notice how both teams sit in the top 3 of player expense proportion and the bottom 9 in ticket price both years. My thought process was that even beyond these two outliers, teams across the MLB teams decide to wager their on-field product and overall franchise health at times for the chance to maintain their record while throwing a few more recognizable faces into their uniforms. Now I haven't been proven wrong that this sentiment doesn't exist in front office analysis. However, if it does exist it is clear that it doesn't really work.

Hopefully it is understood that although you want your favorite baseball team to win the World Series, organizations as a whole don't actually care about winning unless it is a means to increase financial gain. Larger signing bonuses, a higher volume of incentive clauses, and greater player-side signing leverage are all issues that MLB clubs deal with signing big name players. These things directly drive up the player expense proportion number as it has been put on display in my analysis.

5 Conclusion

While people tend to label the MLB as a "pay-to-win" league, the Mets and Padres have shown us that is not quite the case. It may be a "pay-to-try-and-win" league, but baseball seems to have a way of avoiding a complete capitalist narrative being developed. I can safely say that "pay-to-win" as "win" represents financial success, is an equally ineffective method.

A smart baseball fan will be able to describe to you the various reasons why just because a team pays for all the premier talent the world has to offer, they still can't win the World Series. They will tell you that their guys can't get hits at the right times, lead-off hits turn into double plays, or that their setup arm can't hold it down. Baseball is a game of chance more so than any other major professional sport so this isn't a surprise. But that still doesn't explain why teams seemingly can't buy their way to an entertaining on-field product that keeps the fans coming back and engaged consistently.

By now, MLB teams should be able to run financial analysis and valuation on given player contracts to reach the equilibrium that will maximize profit by balancing the need to win and satisfy the media's taste. They should be able to find the balance point at which they can spend to satiate winning up to the point where it doesn't matter, and then spend in order to increase fan engagement. But they still can't do that.

Player expense proportion can be seen as a way to quantify how much a team spends on contracts that not only will contribute to winning, but entertaining baseball. So why, even over a 16 year period do trends not show their face in the slightest?

It seems the measures the MLB implements to nullify economic disparity in the league have paid off at least a little. Things like the Competitive Balance Tax and Competitive Balance Draft Picks ensure that revenue trends in huge favor of financially advantageous ball clubs do not occur. If a huge payroll like the Mets decide to break the bank on several all-stars, they will almost certainly start to win more games than before that move and they are even likely to generate mass amounts of media and fan hype around their club. But a few years down the road when it's their third year over the balance tax threshold and their luxury tax is now worth 50% of their payroll it might start to hurt. At that point, front office financial analysis is likely going to surmise that some spending habits are going to need to change and the big money Mets are not going to be able to spend like they still have big money. And while the Pirates or the Orioles may never know what it's like to acquire 5 all-stars throughout the course of a year or two, they aren't within a mile of the Competitive Balance Tax and therefore have freedom to increase their payroll spending limited only by their own company balance sheet.

In conclusion, my initial hypothesis that spending on players with big bonuses and incentive clauses in their contracts, depicted through a high player expense proportion in my research, is entirely wrong. There is an essentially negligible relationship between spending on excessively expensive contracts and significantly improving the market value of an organization.

-To see my Python code for this project: https://github.com/ethwang17/MLBfinancials.git
-Code is in the "mlb_fin.py" file

Works Cited

1. “The Statistics Portal.” Statista, www.statista.com/. Accessed 2 Aug. 2023.

2. Ozanian, Mike. “Baseball’s Most Valuable Teams 2023: Price Tags Are up 12% despite Regional TV Woes.” Forbes, 13 Apr. 2023, www.forbes.com/sites/mikeozanian/2023/03/23/baseballs-most-valuable-teams-2023-price-tags-are-up-12-despite-regional-tv-woes/?sh=191cbb256501.

3. “Sportico’s MLB Valuations.” Sportico’s MLB Valuations, 14 Apr. 2022, d3data.sportico.com/MLBVal20

4. “MLB Time Series Statistics by Team.” Baseball Reference, www.baseball-reference.com/teams/. Accessed 2 Aug. 2023.

5. “Consumer Price Index, 1913.” Federal Reserve Bank of Minneapolis, www.minneapolisfed.org/about-us/monetary-policy/inflation-calculator/consumer-price-index-1913-. Accessed 2 Aug. 2023.