# An Elementary Data Analysis on the MLB 2012 Season and the Impact each Team's PAYROLL has on WINS

Kabir Snell - November 2021

## Description

In this analysis we will use the payroll data and wins data of each team that competed in the MLB 2012 season to see if there is a connection between how much money teams spend and the success that they have. It is expected that the more money a team spends, they will have more success on average. This project will be done in Jupyter Notebook and the data can be found as a csv file in this repository. This is my first project that I will be placing on GitHub with hopefully many more and better projects to come. A full report can be found at the google doc link below. The code I used can be found here.

## Hypothesis

It is my prediction that there will be a correlation between how much a team spends and the wins that a team has throughout the regular season. I expect to see a positive correlation between these two, with a few outliers.

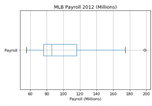
## The Data

The Data that I worked with can be found here. Taking a preliminary scan of this data, it seems that teams that spend a lot of money do in fact win a lot of games, but there are more teams than expected that do not spend a lot of money and have a lot of wins.

## Preliminary Data Analysis

Looking at the data more closely, the first thing I did was to find the mean of both wins and **Payroll (in millions)**. This will be the foundation for the rest of our project as it will be the line for whether a team wins more/less games than average and spends more/less money than average. Here were the findings:

Mean Payroll: 98.02
Mean Wins : 81




** Important note for Payroll boxplot, I omitted the outlier, I plan to fix it but moving through the rest of this project I will be using the Mean Payroll from above (98.02)*

## Separating Teams Into Types

Now that we have some sort of understanding for the "middle" values of wins and payroll, I am going to separate the teams into four types. *Here "Wins" is an abbreviation for number of wins.*

Type A - Teams that have an above average Payroll and above average Wins

Type B - Teams that have an above average Payroll and below average Wins

Type C - Teams that have a below average Payroll and above average Wins

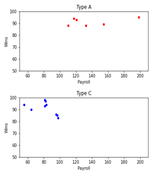
Type D - Teams that have a below average Payroll and below average Wins

Here are the teams separated into their groups:

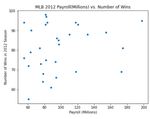
```
Type A teams: Spent above average money with above average success
Team        Payroll(M) Wins
Yankees     197.96     95
Giants      117.62     94
Rangers     120.51     93
Angels      154.49     89
Tigers      132.30     88
Cardinals   110.30     88


Type B teams: Spent above average money with below average success
Team        Payroll(M) Wins
Marlins     118.07     69
Red Sox     173.18     69


Type C teams: Spent below average money with above average success
Team        Payroll(M) Wins
Nationals    81.34     98
Reds         82.20     97
Braves       83.31     94
Athletics    55.37     94
Orioles      81.43     93
Rays         64.17     90
Dodgers      95.14     86
White Sox    96.92     85
Brewers      97.65     83


Type D teams: Spent below average money with below average success
Team        Payroll(M) Wins
Pirates      63.43     79
Padres       55.24     76
Mariners     81.97     75
Mets         93.35     74
Blue Jays    75.48     73
Royals       60.91     72
Indians      78.43     68
Twins        94.08     66
Rockies      78.06     64
Cubs         88.19     61
Astros       60.65     55
```

## Putting the Data Together So Far

Just looking at the types of teams once separated into types it is apparent that most of the teams that spend more, win more (further analysis later). Now here are scatter plots for the four different types of teams and their payroll vs wins.




Now that we have an accurate visualization for all the different types of teams we can put them together to form another visualization where we can examine all the teams at the same time. Here is a scatter plot with the above four scatter plots combined into one, with the different colors representing the different 'types' of teams.




Analysis of the above graph: It is clear that the further along the x-axis you travel, the higher *likelihood* of having more wins is apparent, but there is clearly a large number of teams that do not spend as much money that also have a lot of wins. A further analysis on this will be done using Bayes' Theorem.

## Independent Events Analysis

** Disclaimer: Using this theorem assumes the wins and payroll of each team are all independent of each other. While this is true for payroll vs. payroll, one could make an argument that the payroll of one team affects the wins of another, making these events not independent and thus this formula would not apply. While I agree that this is possible, MLB teams only play against each other a small amount of times per year, and sometimes not at all. Additionally, for every team at the high end of the payroll spectrum, there is a team at the low end, thus acting as a counterbalance and making this factor negligible in my experiment. Using this formula, I will calculate the probabilities of several events given several other events.*

In our first application, we will look at the probability that a team has above average Wins, given that the team has above average Payroll. Connecting this back to our visualizations, this would be cross examining Type A and Type B teams.

*A - Event that a team has above average Wins*
*B - Event that a team has above average Payroll*
*P(A | B) = P(A∩B) / P(B)*
*∴ P(A | B) = .75 or 75%*

The probability that a team will have above average Wins given that the team has above average Payroll is .75 or 75%.

Next we will calculate the probability that a team has above average Wins given that they have a below average Payroll

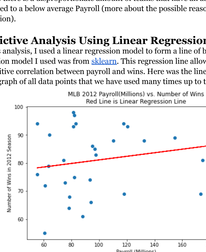
*A - Event that a team has above average Wins*
*C - Event that a team has below average Payroll*
*P(A | C) = P(A∩C) / P(C)*
*∴ P(A | C) = .45 or 45%*

The probability that a team will have above average Wins given that the team has above average Payroll is .45 or 45%

It is clear from the results that if a team has an above average Payroll it is more likely that they will have above average Wins. The reason I did not calculate the probability that a team has an above average Payroll given that they have above average Wins is because there is a disproportionate amount of teams that have an above average Payroll compared to a below average Payroll (more about the possible reasons for this in the conclusion).

## Predictive Analysis Using Linear Regression

For this analysis, I used a linear regression model to form a line of best fit. The linear regression model I used was from sklearn. This regression line allowed me to see if there is a positive correlation between payroll and wins. Here was the line of regression lying on the graph of all data points that we have used many times up to this point.




Here, the line of regression has supported the data that we have found up to this point. There is a positive correlation (according to this regression line) between payroll and wins. But, this correlation is only slight since by the graph one can see that the slope of the line is not very steep, meaning that the correlation is maybe not as strong as the *independent probability calculations* made it seem.

## *Planned topics for the rest of this report*

- *Conclusion*
- *Potential expansions on this report using more data*