The Relationship Between MLB Salaries and Team Win Rates

1. **Overview**

Our goal with this report is to determine if there is any correlation between an MLB team’s expenditures on player salaries and their total wins for a given season between the years 2000 and 2016. This analysis will primarily focus on the salaries of “superstar” players and their effect on team wins.

1. **Conveying the relationship**

In this Exploratory Data Analysis, we will be primarily looking at visualizations to help explain the trends and relationships and tell the story of our findings. Various models will be used throughout and explanations of how we came to these findings will be provided.

1. **The Data**

Our data set comes from Sean Lahman, an author and sports journalist for USA TODAY. Lahman has been running the Baseball Archive since 1995, making it the oldest and longest running baseball site in existence. The archive is extensive, with stats updated throughout the current season, and stats dating back until the beginnings of the fledgling sport in the 1870s that would go onto become America’s greatest pastime.

**The following are our initial thoughts and questions about our datasets:**

* Does the pay of “superstars” affect the pay on non-star players?
* Does it cause the pay to increase over time?
* Does the winning percent increase with team value (sum of all player salaries)?
* How much do “superstar” players affect team play?
* If these “superstars” are out of action does the win percent fall?
* What measure of performance are player salaries based on?

1. **Cleaning the Data**

Seeing that our data ranged all the way back to 1871, we had a lot of parsing down to do in terms of our data. Looking at the salary data, we found that the data only went back to the 1980s and up until 2016. From this point we narrowed our dataset even further to include 2000 until 2016. This gave us enough data point to form predictive regressions, and also limited the amount we would have to worry about inflation by looking at a 17-year period instead of a 31 year period. While looking at the data we saw relative parity between the 30 team in terms of how they played defensively, with their fielding percentages only differing by .001 to .003 of a percent between any given team. This analysis is focused on offensive players only.

1. **Analysis**

**Superstar Player Analysis**

In these charts of David Ortiz’s performance, we can see very little change in his team’s win percentage during games he played in. A sole exception being the year 2012 where Ortiz missed over 70 games and the Red Sox’s win percentage dropped significantly, relative to the other data points provided.

Icon

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Next looking at Albert Pujols, we see the same trend as David Ortiz, however when Pujols missed roughly the same number of games as Ortiz the year before, the Angels saw an almost negligible dip in their team win percentage.

Icon

Description automatically generated

A picture containing text, indoor, white

Description automatically generated

**Team Salary Analysis**

Next we look at the team salaries of the 30 MLB teams over the same period. An interesting thing to note is the trend of all teams’ total salaries increasing over this 17-year period. Teams like the Dodgers see a steep increase in their total salaries coming into the later half of the 2010s, coinciding with that team’s playoff berths.

A picture containing window, center, square

Description automatically generated

**Correlations**

This heatmap shows the correlation between the quantitative data in our dataset. We can see here that there are no strong correlations aside from obvious relations between runs scored and wins. We also see that the wins and salary of a team have slightly positive relation, but still remains relatively flat.

A picture containing treemap chart

Description automatically generated

**Regression Analysis**

The linear regression line between team wins and team total salary is only slightly positive. Our regression modeled with team wins as the dependent variable and team total salary as our independent variable. Our OLS regression results give us very low R-squared values and very high AIC and BIC values. This tells us that a team’s total salary is not a very good predictor of a team’s total wins for a given season.

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

1. **Conclusions**

Based on our analysis we have determined the following:

* The performance of a “superstar” player has little to no effect on the team’s win percentage during a season.
* The total amount a team spends on its player salaries is not a good predicting factor of that team’s performance during the regular season.

The results gathered by this report would benefit the front managing offices of MLB teams as it shows how it’s money being spent on high value “superstar” players has little effect on the performance of their team throughout the season.

1. **Limitations**

While we have gleamed some useful insights from our dataset, it is limited in its scope. Given more time, we would also do a deeper analysis of how a team performs defensively and if the money spent on that side of the team effects a team’s performance, especially pitching staff. Our data was also cut off at 2016, and given more up to date data, we could more accurately predict the state of the game today.

1. **References**

Dataset taken from Sean Lahman’s Baseball Archive

<https://www.seanlahman.com/baseball-archive/statistics>

Other analysis using the same dataset from Kaggle

<https://www.kaggle.com/garrison/are-big-spenders-big-winners>

<https://www.kaggle.com/gracezhou0912/baseball-analysis>

<https://www.kaggle.com/weijenhsu/are-closers-over-paid>