The Power of Money in Baseball

Coleton Reitan

Motivation & Background

- The motivation behind this project is to understand what exactly makes a championship team a championship team.

- I’ve been playing the game since I was 5 and want to understand the game further through an analysis of the game’s available data.

- Big data is available and ready to be used within the game. However, I want to take a deeper look at data presented before the games are even played - a team’s financial data.

- I received my undergraduate degree from Wesleyan University with a Major in Economics and Minor in Data Analytics and was named a captain on the university’s baseball team.

Research Questions

1. What is the overall payroll of the top 4 teams that make it to the championship series (semifinals) compared to the payroll of the bottom 4 teams in the league? How does the winner of the top 4 teams compare to the other 3 teams?

2. How is the player salary of these teams distributed? Is there a pattern in the positions within the teams that receive the most money as opposed to the least amount of money? – Are there patterns that indicate that teams should start putting more money into one position as opposed to another?

3. Does the league operate in a way such that if a team spends more, they are guaranteed to win more? Or is there an optimal amount of money to be invested into a team's payroll before there becomes a drop off in wins?

4. Do teams that show a change in spending habits from one year to the next observe a change in their success? Is there a significant increase in success when a team invests more money into their player's salaries, or do they see the opposite?

Data Description

The data was sourced from three reliable locations, with the timeframe being from 2015 through 2023, excluding 2020. The financial data was sourced from two different websites based on the years (due to availability issues). For any of the data that pertains to how the team performed (such as wins), the data was sourced from the third website.
The complete dataset consists of 17 different features, each with 2,748 datapoints.

Year: The year of the baseball season in which the data pertains to

Payroll: The sum of player salaries for a given team in a given year

PreviousYearPayroll: The sum of player salaries for a given team in the previous year to the one specified

Percent Change: The ratio of difference between the current year and the previous year (Follows normal percent change formula)

- **DifferencefromAverage:** The difference of the specified year’s payroll from the specified year’s average payroll (Specified Year Payroll – League Average Payroll)
- **Position:** The position of the player
- **Salary:** The salary of the player
- **PercentofPayroll:** The percentage of the team’s payroll a specified player’s salary fills
- **Wins:** The number of wins a team has in a specified year
- **Playoffs:** Says whether a team made playoffs or not in the specified year. Also says if team made to semifinals of playoffs or was a bottom 4 (bot4) team in league.
- *Playoffs Feature Values:*
- Y/Y: Yes team made playoffs, yes team made the semifinals
- Y/N: Yes team made playoffs, no team didn’t make semifinals
- N/N: No the team didn’t make playoffs, no the team was not bot4
- N/Y: No the team didn’t make playoffs, yes the team was bot4

- **WSWin:** Says whether a team won the world series or not in the specified year (Binary Y/N value)
- **Player:** The name of the MLB player on a team’s roster
- **Team:** The name of the MLB team being analyzed.
- **HighestPayroll:** Says whether a team had the highest payroll in the league for the specified year or not (Binary Y/N value).
- **LowestPayroll:** Says whether a team had the lowest payroll in the league for the specified year or not (Binary Y/N value).
- **DifferencefromPrevious:** The difference of the specified year’s payroll from the previous year (Specified Year Payroll – Previous Year Payroll)
- **LeagueAveragePayroll:** The average payroll of the league for the specified year

ETL Process

- To collect the data, went into each website, filtered for the team and year, then copied data directly into excel. This was done for all datapoints.
- Used the Str() function to see that all data was of the expected type – 10 integer or numeric, 7 character.
- Used Table() function for the character variables to ensure there were no duplicates, variables were formatted the same way, and there weren’t any surprises.
- Adjusted numeric values in excel from dollar to general. Adjusted percentage values from percent to decimal.
- All the data was manually brought into the excel file by me – very clean data. However, a large portion of time was dedicated to collecting the data.

Tableau Link