The Power of Money in Baseball

A visual analysis

Coleton Reitan

Fall 2023

Data Visualization for Business Analytics

[Tableau Link](https://public.tableau.com/app/profile/coleton.reitan/viz/ThePowerofMoneyinBaseball/ThePowerofMoneyinBaseball)

***Motivation & Background***

The motivation behind this project is to understand what exactly makes a championship team, well, a championship team. I have a very deep understanding of baseball; 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. Baseball is considered America’s most analytically driven professional sport; there is an incredible amount of data points captured in every play of every game – and the fact that a single team plays 162 games within a season, without including playoffs, shows that 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. The beauty of financials in baseball exists in the fact that there is no salary floor or ceiling, allowing there to be a very large range of payrolls for the teams within the league. I want to use this financial data to visualize what is happening when an MLB team wins a championship, and on the contrary, what is happening when an MLB team performs terribly, both instances from the financial perspective. 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. With this background and the knowledge I’ve been attaining in my time pursuing a Master’s in Business Analytics here at Adelphi, I believe that I am well equipped with an in-depth knowledge of the sport as well as a strong ability to understand the financials within the game.

***Research Questions:***

Prior to the commencement of this analysis, there had to be questions to be answered. The research questions of the project are as follow:

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?

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?

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?

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?

As the project progressed, a remark for future analyses would be to incorporate all 30 MLB team data from each year. As the objective of this project was to get an understanding of the difference in extremes between the good and bad performing teams, I would like to see this analysis done with data from the rest of the teams in the league.

***Data:***

The data was sourced from three reliable locations, with the timeframe being from 2015 through 2023, excluding 2020. For any of the player or team payroll data for 2015 through 2019, the data was sourced from: <https://legacy.baseballprospectus.com/compensation/>. For any of the player or team payroll data for 2021 through 2023, the data was sourced from: <https://www.spotrac.com/mlb/texas-rangers/payroll/2021/>. For any of the data that pertains to how the team performed (such as wins), the data was sourced from: <https://www.baseball-reference.com/leagues/majors/2016-standings.shtml>. The complete dataset consists of 17 different features, each with 2,748 datapoints.

The feature description is as follows:

**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)

**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

**DifferencefromAverage:** The difference of the specified year’s payroll from the specified year’s average payroll (Specified Year Payroll – League Average Payroll)

**Player:** The name of the MLB player on a team’s roster

**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)

**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).

***ETL:***

To collect the data, one of the specified websites was opened. From there, a filter within the website was set to show specific team and year data. The data was copied and pasted directly from that the website into an excel sheet, which was then formatted to match the rest of the file. This was then repeated for every data point used within the datafile.

After data extraction, there was an expectation to be 10 variables of Integer or Numeric type and 7 variables of Character type. To check this, the str(function) was used on the dataset, which showed exactly what had been expected (2 were of Numeric type, 7 of Integer type and 7 of character type).

To ensure the variables were correctly brought in with no errors or unwanted duplicates, string normalization was applied to all the variables of character type. This was done by using the table() function to observe the *Team* variable, to which it was seen that all team names were capitalized, as well as all teams only appearing once. The process was then repeated on the *Position, Player, Playoffs, WSWin, HighestPayroll, LowestPayroll,* variables, which also had no repeating variables.

Most of the numeric variables had to be adjusted in excel to be categorized as a general number rather than dollar amount, allowing a clean data import to tableau. The *Percent Change* and *PercentofPayroll* variables also were adjusted in excel so that there was not a “%” attached to the value, which again would allow a clean data import to tableau.

Since all the data was manually brought into the excel file by me, the data did not have to go through much cleaning or transforming prior to being loaded into tableau. However, because of this, a large portion of time dedicated to collecting the data.

***Design of Visualization:***

There were 3 main principles followed to visualize the information in the most effective and viewer efficient way. The first principle was the color scheme; the red and blue coloring of the MLB logo are used effectively and are the main colors throughout the visuals. In doing so, not only were the plots pleasant to look at, but they also conveyed the message of a good (blue) and bad (red) – playing to basic viewer intuition. The second principle was using a variety of different types of visuals to keep viewer engagement. Although bar charts are included, only using bar charts with different variables throughout the entire data showcase would get very boring very quickly, leading to a reduction in viewer attention. There are 5 different types of plot methods used throughout the showcase, each with a unique purpose. The final of the main principles followed in the showcase is the readability of the plots. Although some plots may be slightly more difficult to understand at first glance than others, the point still gets across to the viewers. There are no plots that have too much information to fully understand what the data is saying which allows the viewer to have a true understanding of the story being told within the visualizations.

A graph of different colored lines

Description automatically generated

The first visualization is a bar chart that shows each teams’ payroll. The teams are bucketed by year and are shown in descending order from largest to smallest payroll. The teams colored blue had made the semifinals of the playoffs, the teams in red had been a bottom 4 team in the league, and the teams in gold are the teams that had won the World Series. There is also a average line for each year’s bucket that shows where the average payroll for that year was.

A graph of different colored bars

Description automatically generated

The second visualization is another bar chart that is very similar to that of the first. However, instead of the coloring showing whether a team had been a top 4 or bottom 4 team, it shows the distribution of the team’s payroll given the positional salaries using percent of payroll variable. This helps give an understanding of the breakdown of a team’s payroll. The infielders are shaded brown (because they play on dirt), the outfielders are shaded green (because they play on grass), the pitchers are shaded blue (because blue is my favorite color and I am a pitcher), the hitter only players are shaded grey (for no reason aside from color differencing), and finally the players who were labeled DNP, did not play, were red (because it’s bad to be wasting money).

A screenshot of a computer screen

Description automatically generated

The third visualization showed a box and whisker plot of player percent of payroll by position. This visualization conveys the difference in how top 4 teams distribute payroll across their players compared to how bottom 4 teams distribute payroll across their players. The red dots indicate players on the bottom 4 teams, the blue dots indicate players on the top 4 teams, and the gold dots indicate players on the championship teams. This plot truly showed the outliers in payroll distribution, for good and bad.

A screen shot of a graph

Description automatically generated

The fourth visualization was a form of scatter plot that showed how payroll size was related to number of wins. The larger the dots, the larger the payroll. The red dots indicated teams below the year’s average payroll, the blue dots indicated teams above the year’s average payroll. Teams were on the x-axis. There were two constant lines, one showing the (tentative) number of wins to make playoffs, another showing the (tentative) number of wins to be a bottom 4 team.

A graph of different colored dots

Description automatically generated

The fifth visualization was a true scatter plot, showing payroll on the x-axis and wins on the y-axis, showing groupings of payrolls and number of wins. This visualization utilized 4 different shapes to indicate the 4 possible values of the Playoffs variable. The red filled circles showed teams that were bottom 4, the red hollow circles showed teams that were not in playoffs but not bottom 4. The blue filled squares showed teams that were in the semifinals, the blue hollow squares showed teams that were in playoffs but not the semifinals. The gold filled squares are teams that won it all. There were also two constant lines showing the average amount of wins going across the x-axis and the average payroll going across the y-axis.

A graph with different colored lines

Description automatically generated

The sixth visualization was a variation of a bar chart, but bars were able to go above as well as below the y-axis (0) to show positive and negative values. The purpose of the plot being to show the percent change in payroll from the previous year and how the team did in the current year. The teams were bucketed into years, with percent change on the y-axis and the team’s name on the x-axis. Each team is a bar. The bars in red showed teams that did not make playoffs, while the bars in blue showed teams that did make playoffs (the bars in gold again showed the championship team). The width of the bar indicates the difference from the average payroll for that year – skinny bars are small difference from average, large bars are large difference from average.

***Implementation:***

Although each visual shows important features of the data which adds to the overall story, the one visual that holds the most significance is the bar graph of the Distribution of Teams’ Payrolls. This is the main interactive visualization due to the amount of information it contains, along with the number of filters it holds. The base purpose of the visual is to show what percentage of the payroll each position takes up for the team while also showing (for each year) each teams’ payroll as well as the league’s highest and lowest payrolls. However, there are 6 filters attached to this visual as well which allows even more information to be displayed. The filters are: WSWin, Team, Percent of Payroll, Playoffs, and Wins. Each filter can be applied to show very specific information the viewer could be seeking. The viewer can also highlight which position they would like to see, allowing for easier sight of specific information. If the cursor is hovered over each teams’ bar, further information is shown. This information is the position, team, world series win, year, league average payroll for the year, team payroll, number of wins, percent of payroll the position uses, and combined salary for the position.

***Results & Conclusions:***

Overall, the data visualization showcase was an enormous success. There were some interesting insights found within the data and shown through the visuals. Learning how to use tableau to answer questions was an enjoyable experience.

Now, to answer the research questions posed at the start of the project.

The overall payroll of the top 4 teams compared to the bottom 4 teams typically saw a trend of a large discrepancy. Not to say it was always true, however, teams above the year’s average payroll had a higher chance of being a top 4 team, whereas teams below the year’s average payroll had a higher chance of being a bottom 4 team. A for the championship team compared to the other top 3 teams, there was not any identifying trend.

Player salary of championship teams saw a distribution that relied on a large sum of payroll dedicated to the Pitchers, Catchers and or one of the Infielders. Not great teams saw a large distribution of money in their payroll dedicated towards players labeled DNP (did not play). Essentially, bad teams often times find themselves paying players who are no longer on the team or playing baseball. Otherwise, they also tend to distribute more of the payroll to outfielders, which was an interesting find.

Teams that showed a -20% change in payroll from one year to the next often times found themselves being considered one of the worst teams in the league. It is recommended that teams do not drop their payroll by this amount if they want to be considered a competitive team. From the other side, teams that increased their payroll by 20% from the previous year had a very good chance of being a playoff team.

The league, although it may seem that good teams tend to spend more, does not operate in a way such that only the high payroll teams win. Only once in the past 8 years has a team with the highest payroll won the world championship, but it should be noted no team with the league’s lowest payroll has won a world championship. However, there were a few points where teams below the league average payroll had won the championship, so it’s not to say that spending more wins more, but rather spending more efficiently wins more. This was also seen with the outlier in the 2023 Mets, when they had the largest payroll in MLB history, but ended up losing more games than they had won, making it safe to say that there is a limit to *The Power of Money in Baseball*.