P2 - Investigate A Dataset: Baseball Salaries

Eric Jones

July 14, 2017

Introduction

Baseball has some of the richest sports data in history. There are records which date back to 1871 – decades before the MLB as we know it today even existed. The data collected at professional games continues to advance, with details as specific as the release point and angle of break for every pitch being recorded for the history books. The opportunities for analysis are far and wide. In this report, I'll skip over the detailed player by player statistics and instead focus on one higher level question: *Does money buy championship rings?* The answer is yes and no.

To explore this question I'm using data from Lahman's Baseball Archive through the 2016 season. You can download a copy to follow along or do your own analysis.

Data Wrangling

Before I start answering any questions, first I need to take a look at the data to be sure I understand what I'm starting from. I'll be using a two files from the dataset: `Salaries.csv` and `Teams.csv`.

Salaries.csv

The Salaries file includes data beginning in 1985, and simply includes year, playerID information, and salary. It looks like there are no considerations for mid-season trades or other such deals, which at least keeps this data easy to comprehend. Here are a couple rows:

yearID	teamID	lgID	playerID	salary
1985	ATL	NL	barkele01	870000
1985	ATL	NL	bedrost01	550000
...				
2016	WSN	NL	werthja01	21733615
2016	WSN	NL	zimmery01	14000000

The salaries span multiple orders of magnitude. This isn't that surprising given that we're looking at 30+ years of data and stars can make exuberant sums of money. Figure 1 gives an easy to digest view into the shape of the player salary data. It's hard to tell given the scale, but it generally looks like the median player salary has only risen modestly, while the maximum has really jumped. The box plot indicates a number of outliers on the high end, though none of these outliers stand out enough to make me worry about incorrect data.

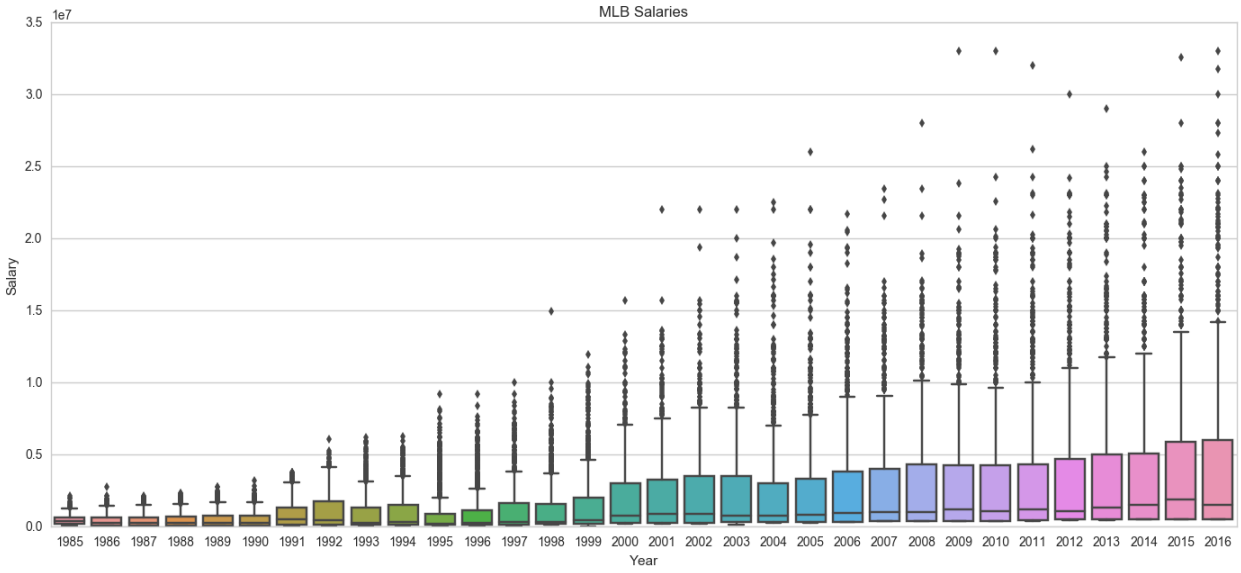

Figure 1: Player Salaries by Year

It is a team, not an individual player, who wins a championship ring, so I need to aggregate the player salary data into team salaries. To do this, I'll use the **sum** of the salaries for all players on a team in a given year. I'm choosing the sum because it's a better analog to the real world: there is a set roster size[1], and the team needs to pay every player's contract. The average salary for each team doesn't necessarily mean much in real world. Using the sum does raise a few concerns I need to clear up before continuing. The most important is if there are any teams that have missing data, the team salary would be lower than it should be. Figure 2 plots each team salary vs the number of player salaries that were summed to make that team salary.

In an ideal world, every team would have the same number of player salaries summed to create the team salary. I wouldn't expect that to actually be the case with real world data. Overall this data looks good, with the vast majority of teams consisting of 25 to 40 players, as expected. There's also *not* a clear upward trend, which makes me more confident that sum is an acceptable aggregator. My main observation is that a couple teams exist with fewer than 10 players and therefore very low team salaries. I'll remove those from my exploration. Everything else here appears to be within reasonable bounds.

With the team salaries established, Figure 3 displays a box plot of those values by year. While team budgets have consistently grown as time goes on, I think the more interesting observation is that the range of salaries across teams for a given year has also grown considerably.

[1]The active roster is 25 players and the expanded roster, which includes players on the disabled list or those optioned to minor league affiliates, is 40 players. `http://www.baseball-almanac.com/articles/baseball_rosters.shtml`

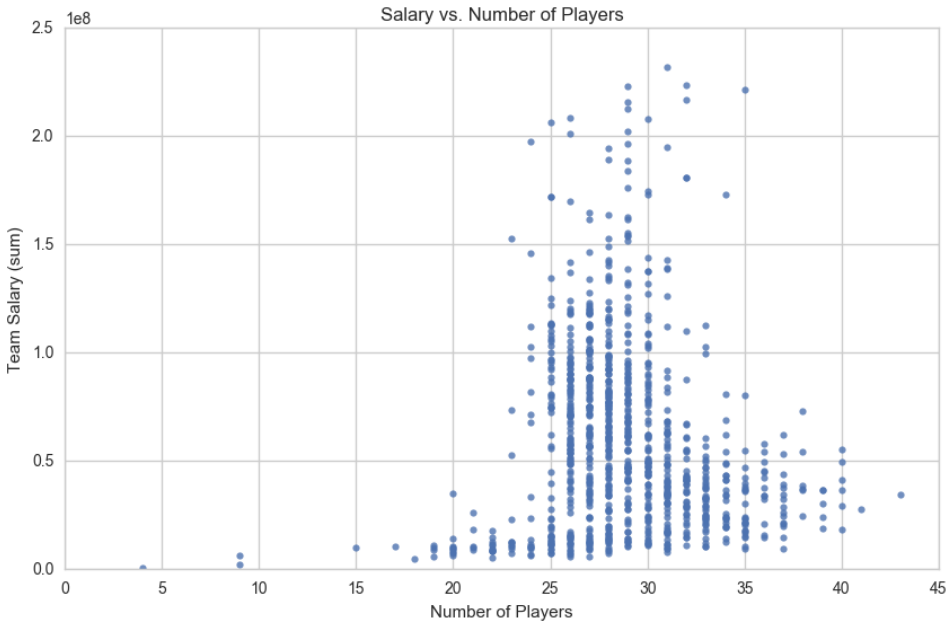

Figure 2: Team salary vs number of player salaries from the original data summed to compute team salary. Teams with `nPlayers` $<$ `10` will be excluded from further analysis

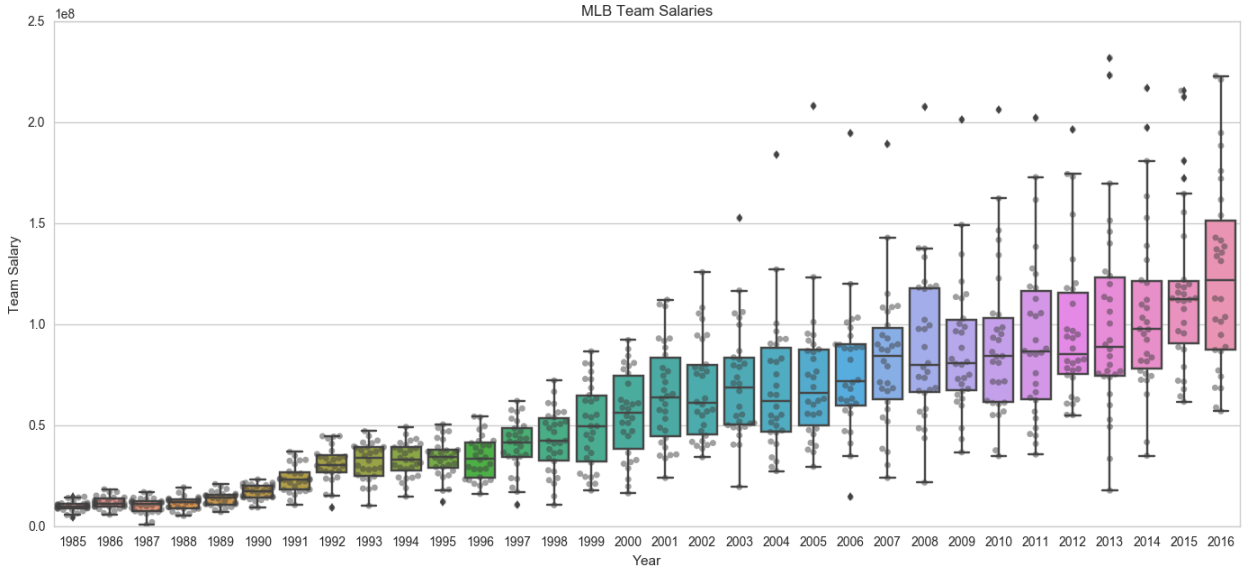

Figure 3: Box plot of team salaries by year 1985–2016.

Teams.csv

So, which of these teams actually won a pennant? The `Teams.csv` file contains a huge amount of information about regular season records and also includes boolean columns for postseason achievements: `WCWin, DivWin, LgWin, WSWin` for each of the postseason awards: Wild Card, Division Series, League Championship and World Series. Because `Salaries.csv` only has salary data back to 1985, I'll exclude team data before that as I'll have nothing to correlate with.

Again, I want to make sure I understand this data. Major League Baseball today is organized into two leagues, each with three divisions. The playoffs are three rounds: the Division Series[2], League Championships, and World Series. For each year, I expect two wild card teams (one in each league), four division series winners, two league champions, and one World Series champion.

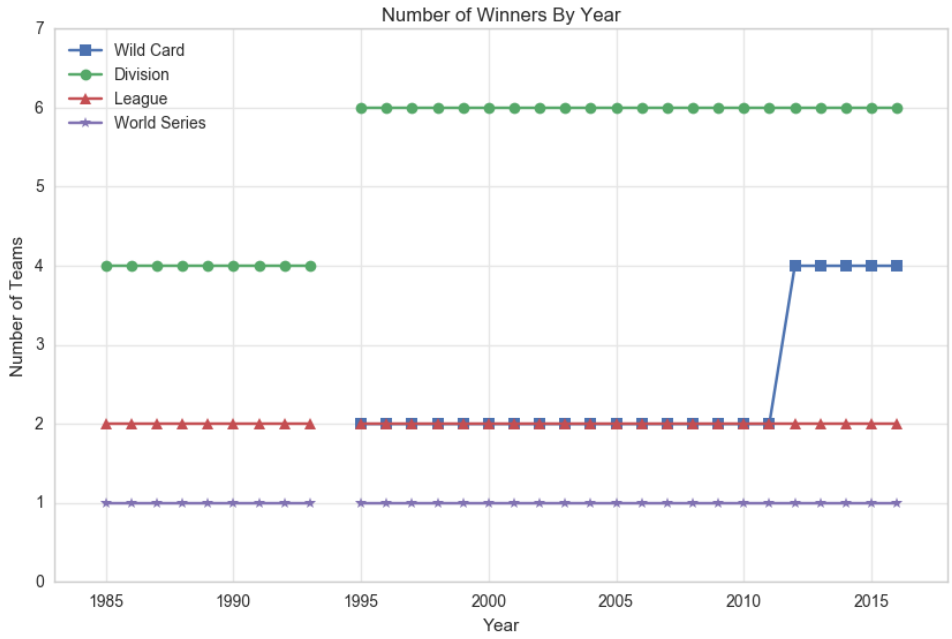

Figure 4: Count of teams recognized for post-season achievements each year as reported in the `teams.csv` dataset.

Figure 4 counts the post-season achievements in the data, and it doesn't quite look like expected. There are a number of things I need to understand here. (1) There is no data for 1994. (2) Wild Card data begins in 1995. (3) There are four Wild Card winners from 2012 onwards. (4) There are six division winners from 1995 onwards. I'll take these one by one.

Missing data for 1994 Researching baseball history reveals that a player strike shortened the 1994 season, resulting in no post-season play.[3]. I'll not count 1994 when considering the number of years of postseason data included in the analysis (so 31 rather than 32 years) and depending on where the analysis goes, I may need to consider that that 1995 salaries could be lower because teams only played 144 (rather than 162) games.

Wild Card Again, a refresher on baseball history reveals the answer to why wild card data only begins in 1995: before that year, there were only two divisions in each league so a wild card team was not necessary. And from 2012 onwards, the rules changed again such that the top *two* teams that are not first place in their division play a single play-in game for the opportunity to be matched up against a team in the division series, in each league. This means that four teams are named as wild cards rather than two.

[2]Because there are three divisions in each league, a fourth team, the Wild Card, is added to play in the Division Series. The Wild Card is the team with the best record in each league that is not first place in their division.

[3] `https://en.wikipedia.org/wiki/1994%E2%80%9395_Major_League_Baseball_strike`

Division Winners Investigating the wild card changes reveal that I've been interpreting the `DivWin` data column incorrectly. This column contains `Y` for a team that won their division through the regular season (that is, it's equivalent to `Rank == 1`), and is *not* an indication that the team won a division series playoff match-up, which only existed after 1994. If I need a list of the division series winners, I could create that list by joining team data with any of the `__Post.csv` tables in the dataset which include the column `round` indicating which round of the playoffs a team participated.

Because I expect to be looking at rank along with using the `WSWin` column for filtering, a the histogram of ranks for every team will be helpful. I know the league and divisions have changed size over the years included in the analysis, so knowing the shape of this baseline will help to make comparisons.

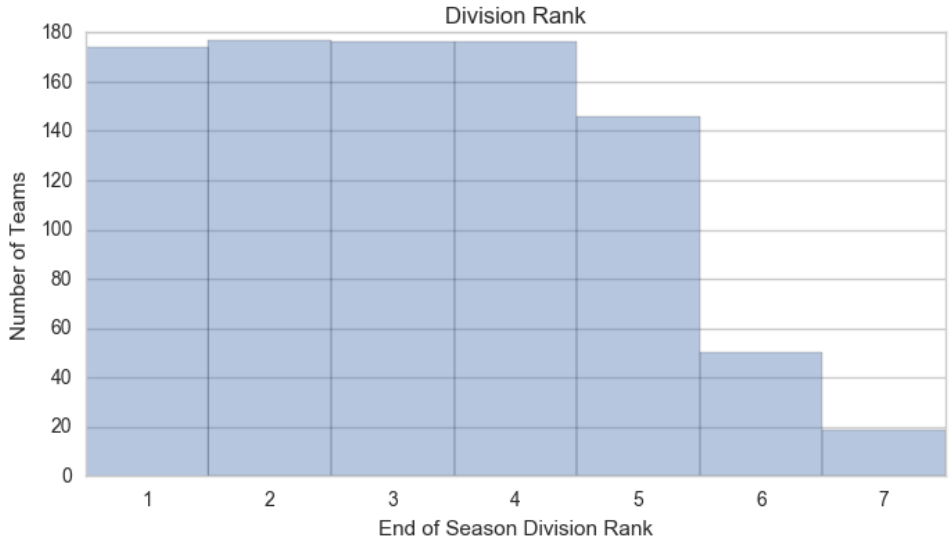

Figure 5: End of season rank in division for every team in the dataset. There are fewer 5^{th}, 6^{th}, and 7^{th} place teams because divisions were not always that large.

Does Money Buy Championships?

With the data prepared, it's time to work towards an answer. The first step is to join the salary data with the team win/loss and postseason data. When performing this merge, I noticed that the total number of teams I was working with dropped from 918 to 907. A bit of investigation revealed that in 2016 the salary data used 11 `teamID` values that were not consistent with those used elsewhere, resulting in those teams being dropped. After making the edits recorded in Table 1, I then also standardized the salary data in each year, to remove the tendency of salaries to increase as time moves forward. With this data, I was able to build Figure 6 which identifies the World Series champion within each year of team salary data.

original	CHC	CHW	KCR	LAD	NYM	NYY	SDP	SFG	STL	TBR	WSN
edited	CHN	CHA	KCA	LAN	NYN	NYA	SDN	SFN	STN	TBA	WAS

Table 1: Changes made to `teamID` values in `Salaries.csv`.

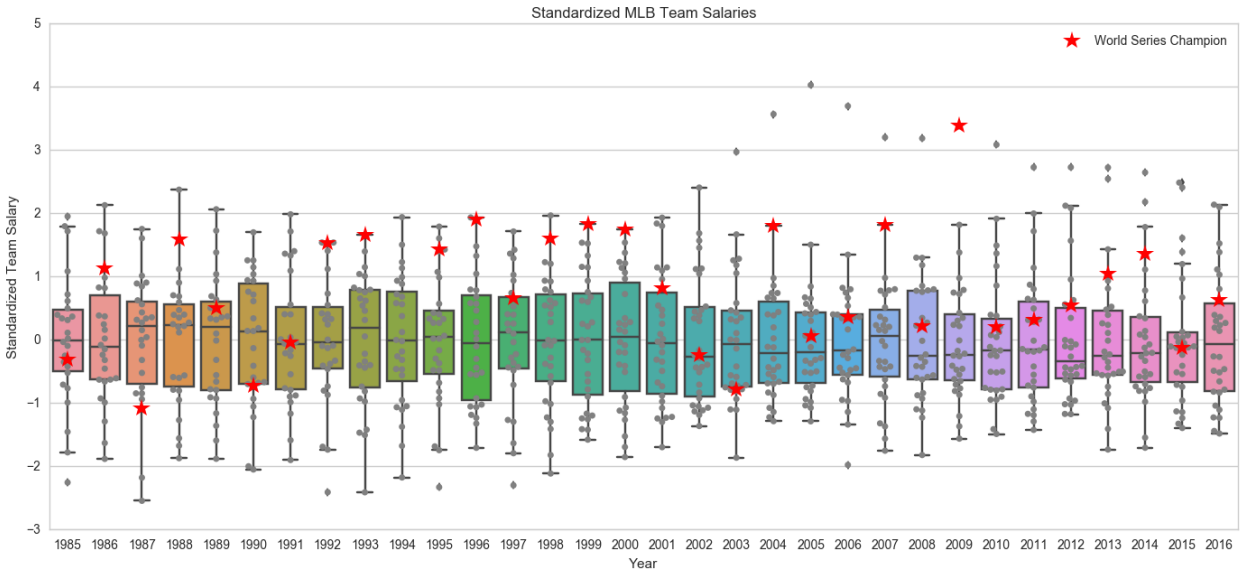

Figure 6: Standardized team salaries with each year's World Series Champion

Looking at Figure 6 gives a good first pass towards an answer to the question does money buy World Series rings? It's now easy to identify that since 1985, the majority of World Series Champions paid a total team salary above the median. Going further, 17 of the 31 spent at or above the 75th percentile. In short, it looks like, yes, you need money to win. But spending by no means guarantees a trophy. The team with the highest salary won the World Series in only 5 of 31 years.

So more generally, how do big spenders place? In Figure 7 the end of regular season rank for teams with high salaries (here defined as StdSalary > 1.5) is shown. Of the 69 teams that matched this criteria, only one third won their division. This isn't quite a fair comparison, because it's possible that some of the teams in this sample placed second to another team *also in this sample*. To work around that fact, Figure 8 inspects the `Rank` using a population of one team per division per year - the team with the highest salary in the division. From this, we can learn that the team which spent the most money in their division won that division 67 out of 174 times, or 38.5% of the time. This isn't nothing, but in no way can I conclude that outspending your opponents is a guarantee of success.

Conclusions

Does money buy championships for MLB teams? It appears that it certainly helps, but it is no guarantee. Even if the data was more explicitly in favor of this conclusion, I couldn't say that a high salary was a *cause* of winning. Very far from that, in fact. It is the players themselves, their performance on the field, as well as many other factors from injuries to opponents to the weather which all contribute towards a team's record and post-season success.

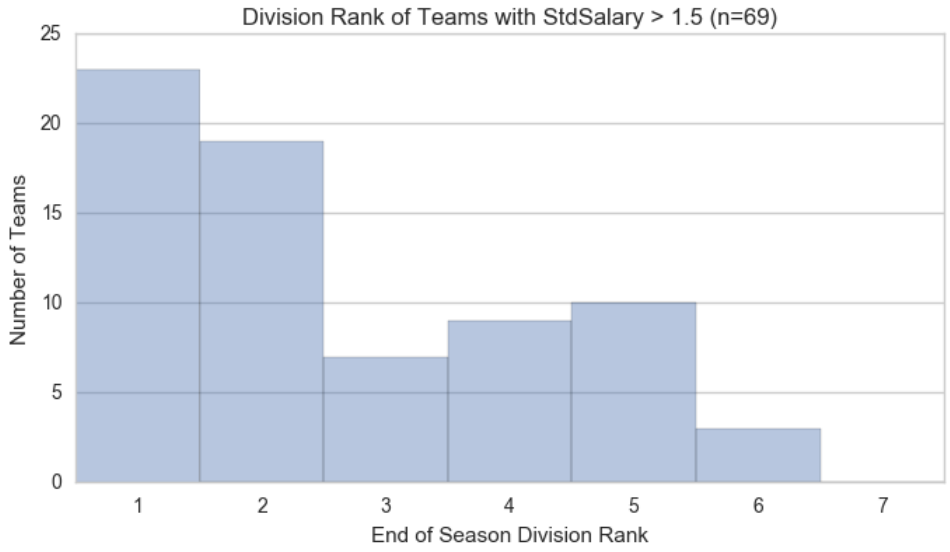

Figure 7: End of season division rank for teams with a year-standardized salary greater than 1.5. Only 23 of 69 won their division.

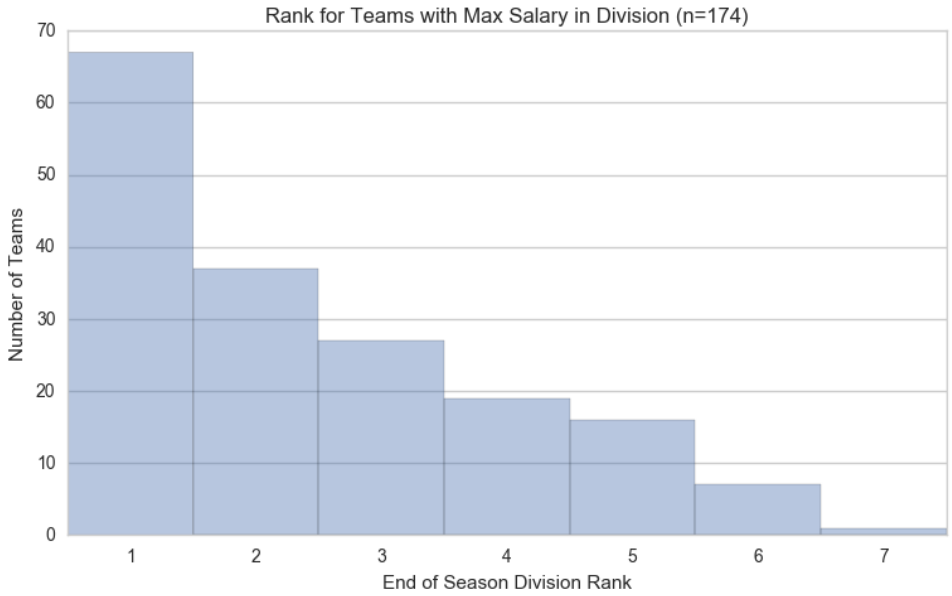

Figure 8: End of season division rank for each team with the highest salary in the division. Each of these could have won their division, only 67 of the 174 teams did. Eleven of these teams were World Series champions over the 31 years of data.