Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa

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.

2. 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.

3. 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?

4. 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

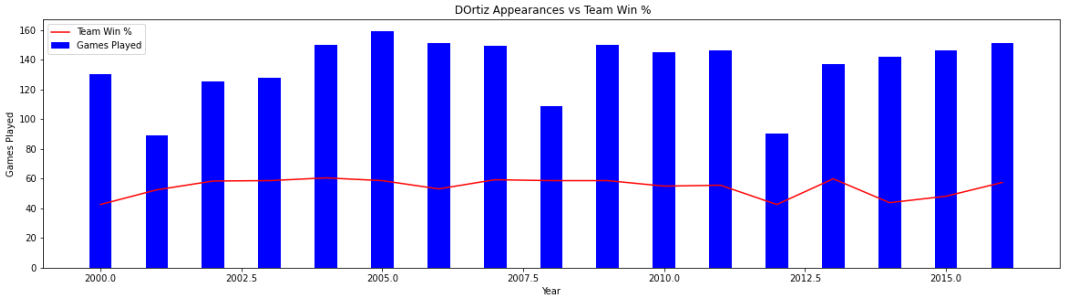
Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa

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.

5. Analysis

Superstar Player Analysis

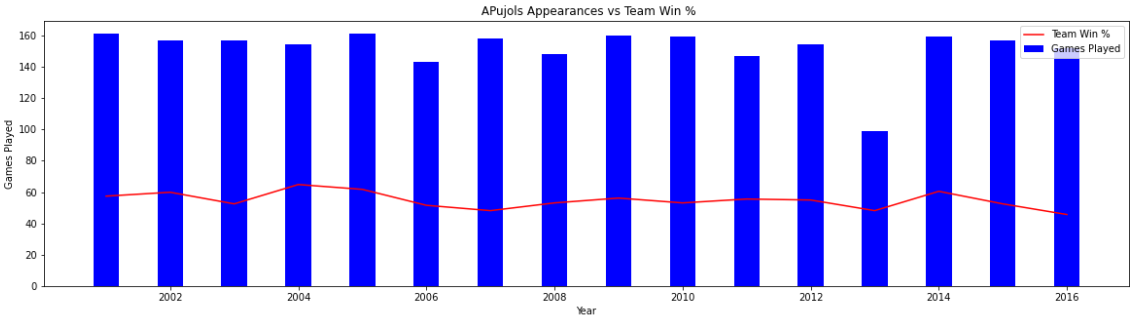
In these charts of David Ortiz's performance, we can see very little change in his team's win percentage during games he played in. A sole exception being the year 2012 where Ortiz missed over 70 games and the Red Sox's win percentage dropped significantly, relative to the other data points provided.



yearID	teamID	G	Win%	playerID	G_all	AP	WAR
2000	MIN	162	42.59	ortizda01	130	478.0	0.7
2001	MIN	162	52.47	ortizda01	89	347.0	0.3
2002	MIN	161	58.39	ortizda01	125	466.0	1.3
2003	BOS	162	58.64	ortizda01	128	509.0	3.4
2004	BOS	162	60.49	ortizda01	150	669.0	4.3
2005	BOS	162	58.64	ortizda01	159	713.0	5.2
2006	BOS	162	53.09	ortizda01	151	686.0	5.8
2007	BOS	162	59.26	ortizda01	149	667.0	6.4
2008	BOS	162	58.64	ortizda01	109	491.0	1.7
2009	BOS	162	58.64	ortizda01	150	627.0	0.7
2010	BOS	162	54.94	ortizda01	145	606.0	2.8
2011	BOS	162	55.56	ortizda01	146	605.0	4.0
2012	BOS	162	42.59	ortizda01	90	383.0	3.2
2013	BOS	162	59.88	ortizda01	137	600.0	4.4
2014	BOS	162	43.83	ortizda01	142	602.0	2.6
2015	BOS	162	48.15	ortizda01	146	614.0	3.1
2016	BOS	162	57.41	ortizda01	151	626.0	5.2

Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa

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.

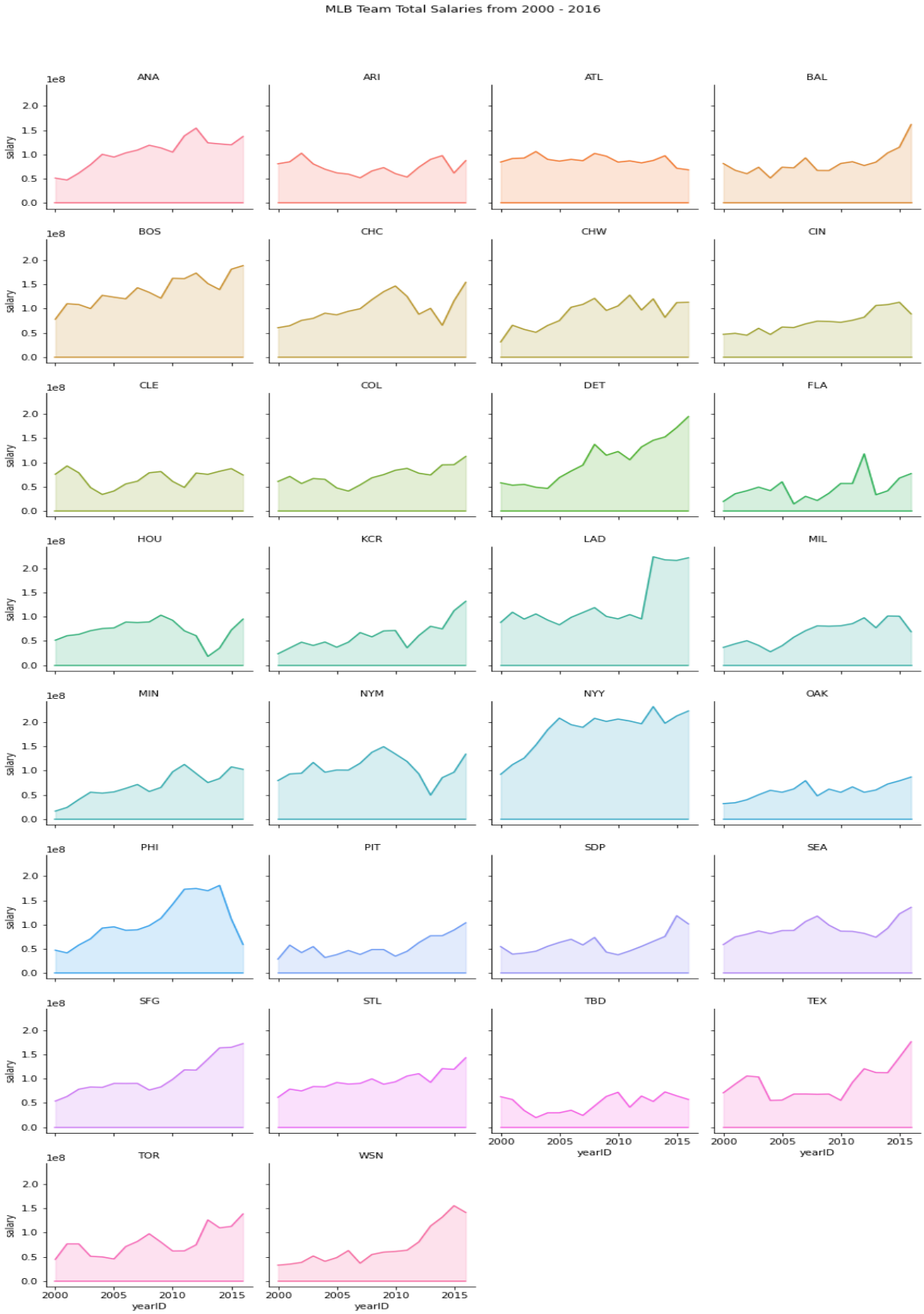


yearID	teamID	G	Win%	playerID	G_all	AP	WAR
2001	SLN	162	57.41	pujolal01	161	676.0	6.6
2002	SLN	162	59.88	pujolal01	157	675.0	5.5
2003	SLN	162	52.47	pujolal01	157	685.0	8.7
2004	SLN	162	64.81	pujolal01	154	692.0	8.5
2005	SLN	162	61.73	pujolal01	161	700.0	8.4
2006	SLN	161	51.55	pujolal01	143	634.0	8.5
2007	SLN	162	48.15	pujolal01	158	679.0	8.7
2008	SLN	162	53.09	pujolal01	148	641.0	9.2
2009	SLN	162	56.17	pujolal01	160	700.0	9.7
2010	SLN	162	53.09	pujolal01	159	700.0	7.5
2011	SLN	162	55.56	pujolal01	147	651.0	5.3
2012	LAA	162	54.94	pujolal01	154	670.0	4.8
2013	LAA	162	48.15	pujolal01	99	443.0	1.6
2014	LAA	162	60.49	pujolal01	159	695.0	3.9
2015	LAA	162	52.47	pujolal01	157	661.0	3.0
2016	LAA	162	45.68	pujolal01	152	650.0	1.5

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.

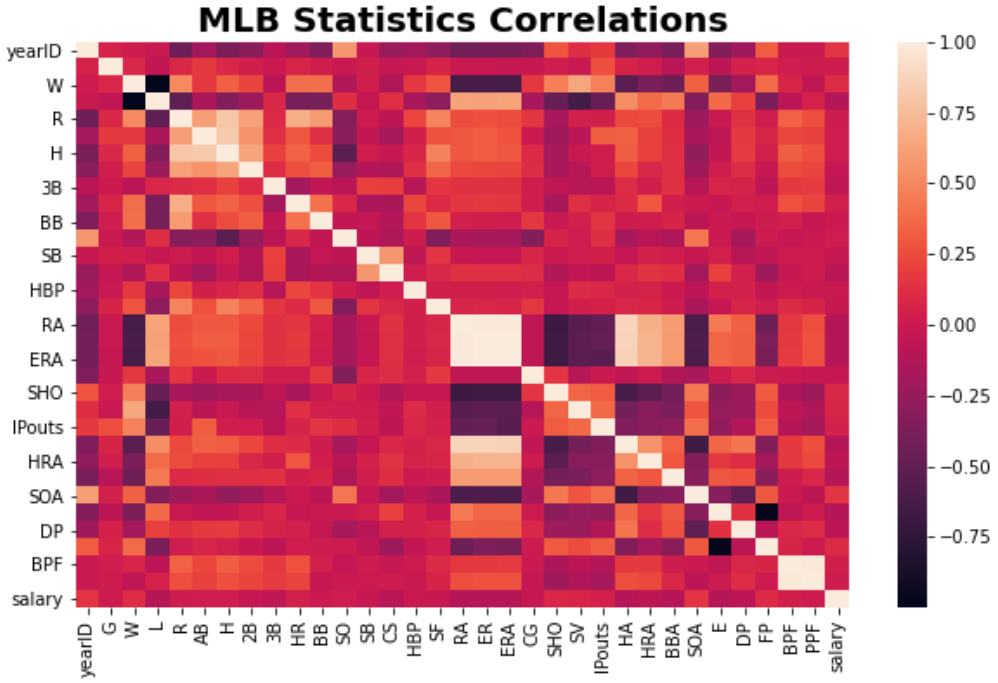
Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa



Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa

Correlations

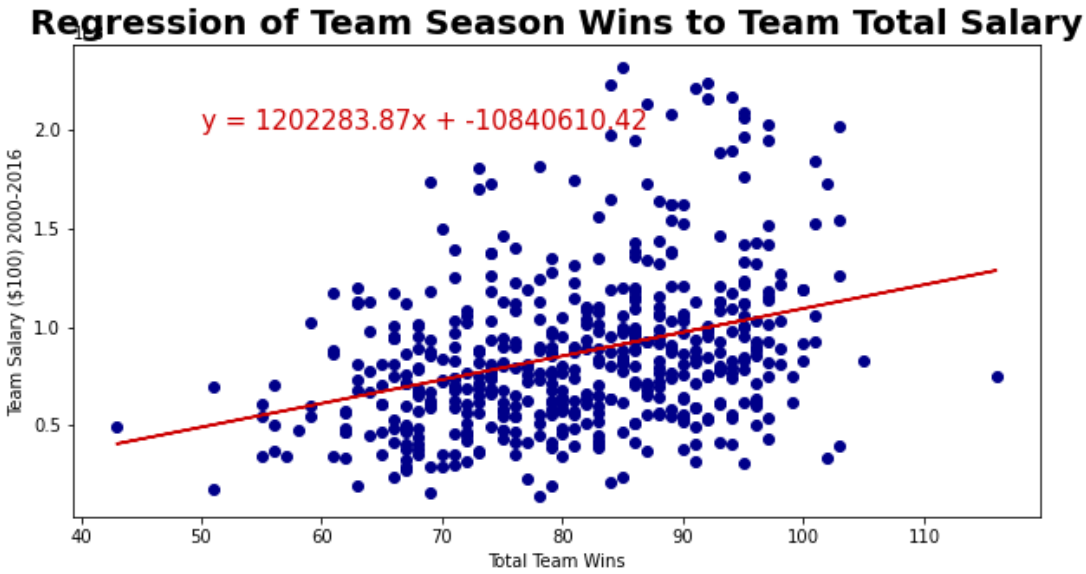
This heatmap shows the correlation between the quantitative data in our dataset. We can see here that there are no strong correlations aside from obvious relations between runs scored and wins. We also see that the wins and salary of a team have slightly positive relation, but still remains relatively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



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.

Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa



OLS Regression Results

Dep. Variable:	salary	R-squared:	0.117
Model:	OLS	Adj. R-squared:	0.115
Method:	Least Squares	F-statistic:	67.41
Date:	Wed, 09 Feb 2022	Prob (F-statistic):	1.84e-15
Time:	18:22:46	Log-Likelihood:	-9617.6
No. Observations:	510	AIC:	1.924e+04
Df Residuals:	508	BIC:	1.925e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.084e+07	1.2e+07	-0.905	0.366	-3.44e+07	1.27e+07
W	1.202e+06	1.46e+05	8.210	0.000	9.15e+05	1.49e+06

Omnibus:	82.138	Durbin-Watson:	1.650
Prob(Omnibus):	0.000	Jarque-Bera (JB):	128.109
Skew:	1.024	Prob(JB):	1.52e-28
Kurtosis:	4.353	Cond. No.	589.

6. Conclusions

Based on our analysis we have determined the following:

Group 1
Nicklaus Austin
Diana Andrade
Calvin Cusick
Ebrahim Moosa

- 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.

7. 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.

8. 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