Salaries over Time in Major League Baseball

Introduction

Since the publication of Michael Lewis's Moneyball in 2003, there has been an explosion in interest in the field of sabermetrics, the application of empirical methods to baseball statistics. Teams looking for an edge have increasingly turned to analysis of all manner of player statistics, from the easy to understand home runs, to the exceedingly complex, such as weighted runs created and fielding independent pitching. The main goal of these efforts have been to identify players with high performance potential who may have flown under the radar and thus will not command as astronomical of a salary as more well-known names. For this analysis, I performed my own introductory sabermetrical excursion into the world of baseball statistics, although I stuck to more familiar baseball metrics for hitting and pitching such as Runs Batted In (RBI) and Earned Run Average (ERA) and the Average (AVG). I will discuss the following question in this analysis.

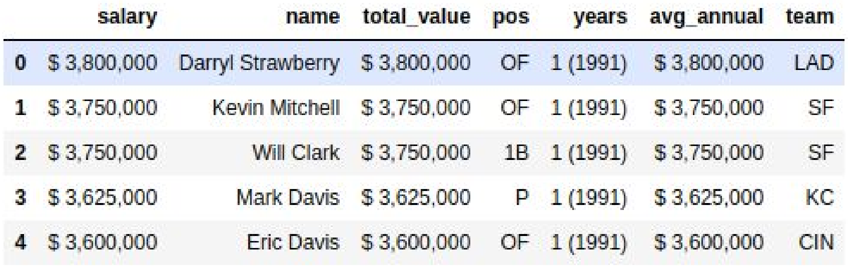
- **Question 1.:** How has the average and the median salary of an MLB player changed over time?
- **Questions 2-7:** What is the relationship between some key batting stats and salary?
- I will see if there is a relationship between salary and:

  - Batting Average (BA)
  - Home Runs (HR)
  - Runs Scored (R)
  - Runs Batted In (RBI)
  - On-base percentage (OBP)
  - Slugging percentage (SLG)
  - On-base plus slugging (OPS)

For this purpose,
I have download a dataset from https://data.world/natereed/baseball-salaries and secondly, I got the batting and pitching statistics from the https://www.baseball-reference.com/leagues/MLB/. Let's start with data cleaning.

Data Cleaning:

In the salaries dataset, I have the following format of data;

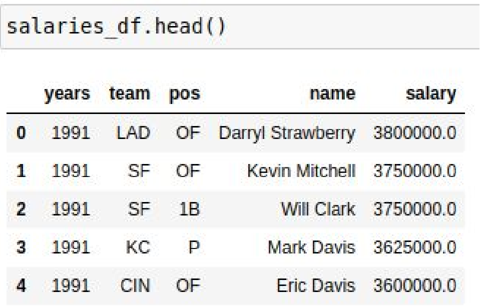


	salary	name	total_value	pos	years	avg_annual	team
0	$ 3,800,000	Darryl Strawberry	$ 3,800,000	OF	1 (1991)	$ 3,800,000	LAD
1	$ 3,750,000	Kevin Mitchell	$ 3,750,000	OF	1 (1991)	$ 3,750,000	SF
2	$ 3,750,000	Will Clark	$ 3,750,000	1B	1 (1991)	$ 3,750,000	SF
3	$ 3,625,000	Mark Davis	$ 3,625,000	P	1 (1991)	$ 3,625,000	KC
4	$ 3,600,000	Eric Davis	$ 3,600,000	OF	1 (1991)	$ 3,600,000	CIN

I have done the following steps to clean the dataset,

- Remove the unnecessary columns from the dataset named as ***'salaries_df'***
- Remove the spaces and unnecessary symbols **($)(,)( )** from different strings of the columns like the ***'salary'*** column.
- Some lines of the column *'years'* in the dataframe have such values like ***'5 (2008-13)'***. Break this value and add new 5 rows in the dataframe with the same credentials as this row and just change the year with the increment of 1.

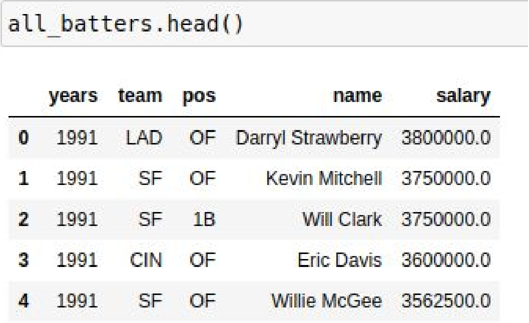
Following is the modified dataset after the above steps;



```
salaries_df.head()
```

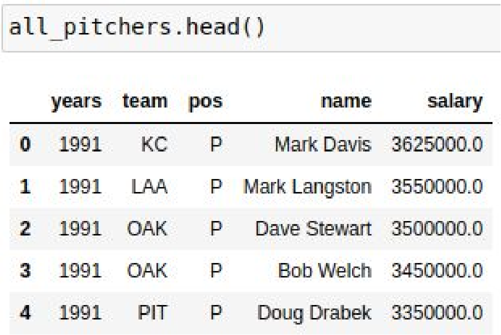
	years	team	pos	name	salary
0	1991	LAD	OF	Darryl Strawberry	3800000.0
1	1991	SF	OF	Kevin Mitchell	3750000.0
2	1991	SF	1B	Will Clark	3750000.0
3	1991	KC	P	Mark Davis	3625000.0
4	1991	CIN	OF	Eric Davis	3600000.0

Now, I separate the batsman and pitcher data from the *'salaries_df'* and save that dataframes as pickle1.pkl and pickle2.pkl file.



```
all_batters.head()
```

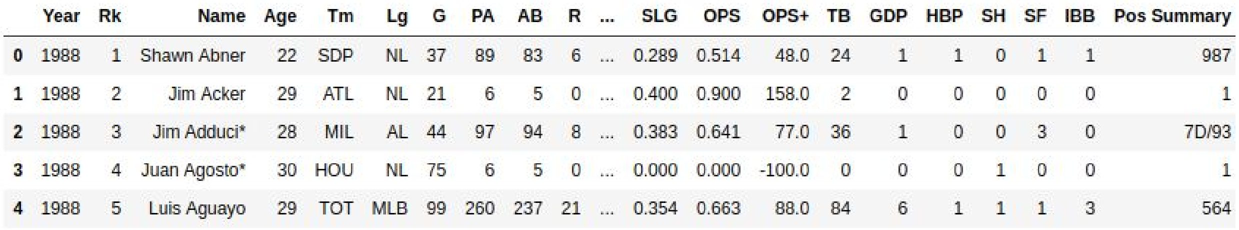
	years	team	pos	name	salary
0	1991	LAD	OF	Darryl Strawberry	3800000.0
1	1991	SF	OF	Kevin Mitchell	3750000.0
2	1991	SF	1B	Will Clark	3750000.0
3	1991	CIN	OF	Eric Davis	3600000.0
4	1991	SF	OF	Willie McGee	3562500.0



```
all_pitchers.head()
```

	years	team	pos	name	salary
0	1991	KC	P	Mark Davis	3625000.0
1	1991	LAA	P	Mark Langston	3550000.0
2	1991	OAK	P	Dave Stewart	3500000.0
3	1991	OAK	P	Bob Welch	3450000.0
4	1991	PIT	P	Doug Drabek	3350000.0

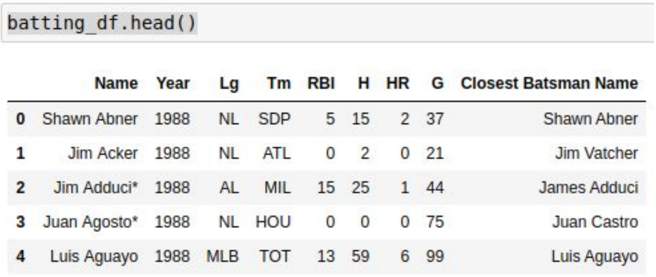
In the next step, I scrap all the statistics data for the batsmen and pitchers from https://www.baseball-reference.com/leagues/MLB/.



	Year	Rk	Name	Age	Tm	Lg	G	PA	AB	R	...	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos Summary
0	1988	1	Shawn Abner	22	SDP	NL	37	89	83	6	...	0.289	0.514	48.0	24	1	1	0	1	1	987
1	1988	2	Jim Acker	29	ATL	NL	21	6	5	0	...	0.400	0.900	158.0	2	0	0	0	0	0	1
2	1988	3	Jim Adduci*	28	MIL	AL	44	97	94	8	...	0.383	0.641	77.0	36	1	0	0	3	0	7D/93
3	1988	4	Juan Agosto*	30	HOU	NL	75	6	5	0	...	0.000	0.000	-100.0	0	0	0	1	0	0	1
4	1988	5	Luis Aguayo	29	TOT	MLB	99	260	237	21	...	0.354	0.663	88.0	84	6	1	1	1	3	564

Merging of Dataset:

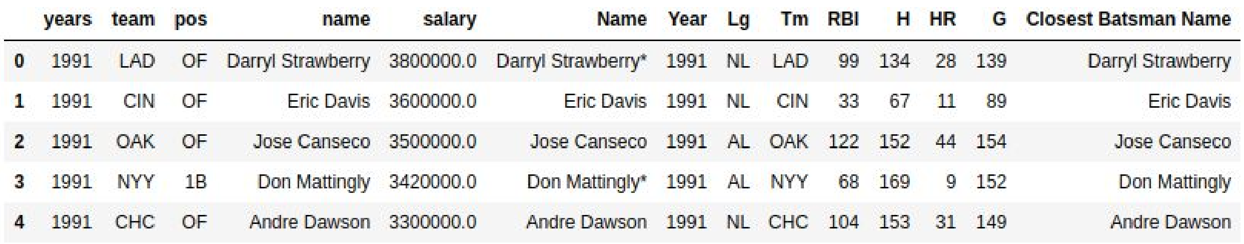
For merging purposes, I have to use three columns as the key such as *['Year', 'Name', 'Team']* but there is a little problem in the statistic dataset, it has the name which is slightly different with the original *'salaries_df'.* To solve this problem, I have added a new column of the *'closest player name'* in the statistic dataframe and find the closest value for each name from the *'salaries_df'.*



```
batting_df.head()
```

	Name	Year	Lg	Tm	RBI	H	HR	G	Closest Batsman Name
0	Shawn Abner	1988	NL	SDP	5	15	2	37	Shawn Abner
1	Jim Acker	1988	NL	ATL	0	2	0	21	Jim Vatcher
2	Jim Adduci*	1988	AL	MIL	15	25	1	44	James Adduci
3	Juan Agosto*	1988	NL	HOU	0	0	0	75	Juan Castro
4	Luis Aguayo	1988	MLB	TOT	13	59	6	99	Luis Aguayo

Then, merge both of the datasets.

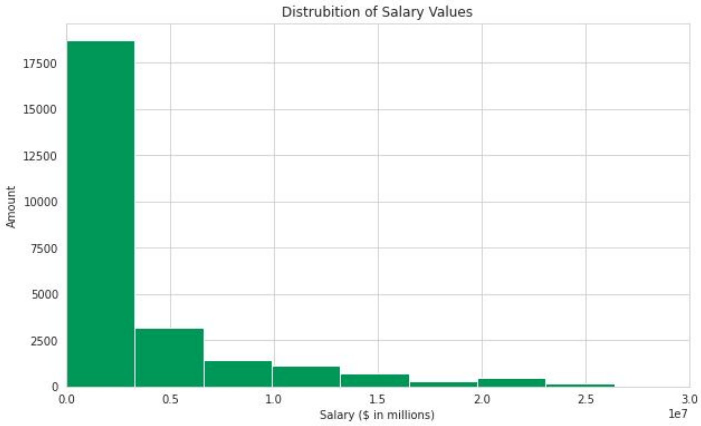


	years	team	pos	name	salary	Name	Year	Lg	Tm	RBI	H	HR	G	Closest Batsman Name
0	1991	LAD	OF	Darryl Strawberry	3800000.0	Darryl Strawberry*	1991	NL	LAD	99	134	28	139	Darryl Strawberry
1	1991	CIN	OF	Eric Davis	3600000.0	Eric Davis	1991	NL	CIN	33	67	11	89	Eric Davis
2	1991	OAK	OF	Jose Canseco	3500000.0	Jose Canseco	1991	AL	OAK	122	152	44	154	Jose Canseco
3	1991	NYY	1B	Don Mattingly	3420000.0	Don Mattingly*	1991	AL	NYY	68	169	9	152	Don Mattingly
4	1991	CHC	OF	Andre Dawson	3300000.0	Andre Dawson	1991	NL	CHC	104	153	31	149	Andre Dawson

Now, the initial working with the data is complete.

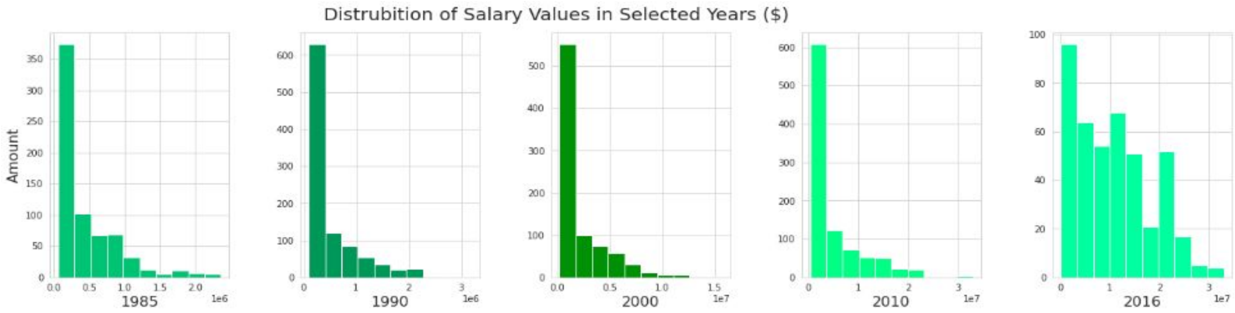
Question 1: How has the average and the median salary of an MLB player changed over time?

I predict that it will be steadily rising over time. As the salary data only goes back to 1988, which happens to be within the first few years of MLB Free Agency, this analysis will only look at the time between 1988-2016. I will also be looking at this with respect to the league (American or National), to see if there are any significant differences between them.



The majority of salary data points fall below $330,000. The next largest bin is salaries between 330,000 and 600,000 USD. The smallest bin (and highest salary range) is between 23 and 26 Million USD.

Let's now visualize how the distribution is for the following years: 1985, 1990, 2000, 2010, 2016



Here we can see the distribution of salary widen as time goes on, but with the later bins move to higher and higher ranges. Now let's see how they perform over time throughout all years, by looking at the mean salary and median salary.

Since 1988, the average MLB salary of the batsman has almost increased by a factor of 9, from 500k in 1988 to 7.1 Million in 2015.

This would be interesting to look at it in comparison to MLB net revenue, which could explain the growth here, along with possibly TV deals.

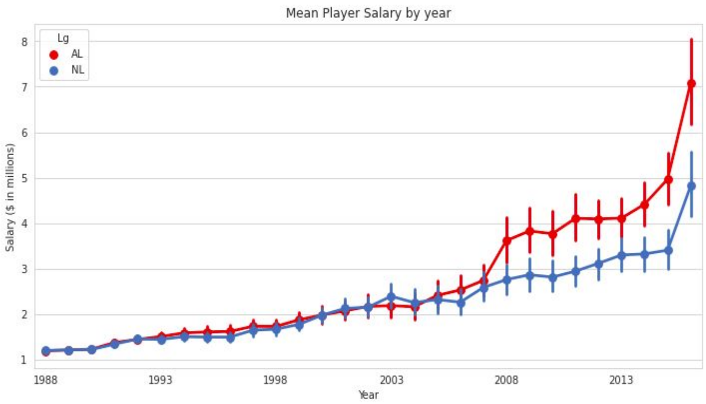
The average salary of an MLB player has since 1990, after staying stagnant through the late 80s. There was a small bump in 1991, where it hovered around 1 Million USD until 1999, where it rose up to $2 million by 2000. The average salary steadily rose in the early 2000s, until 2006.

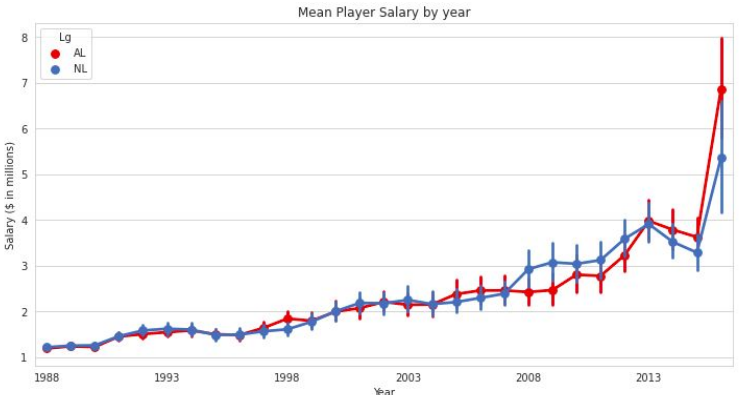
The American and National league average salaries diverged in 2006, with National league salaries staying around 2.6 million USD, reaching almost 3 million USD in 2009. American league salaries rose past 3 million 3 years earlier than National league salaries, in 2006. The National League has since been playing "catch-up" with the American league.

The largest gap between American League and National League salaries was in the latest reported year, 2016, with a difference of ~$2 million.

Mean Salary of a Batsman:

Mean Salary of a Pitcher:



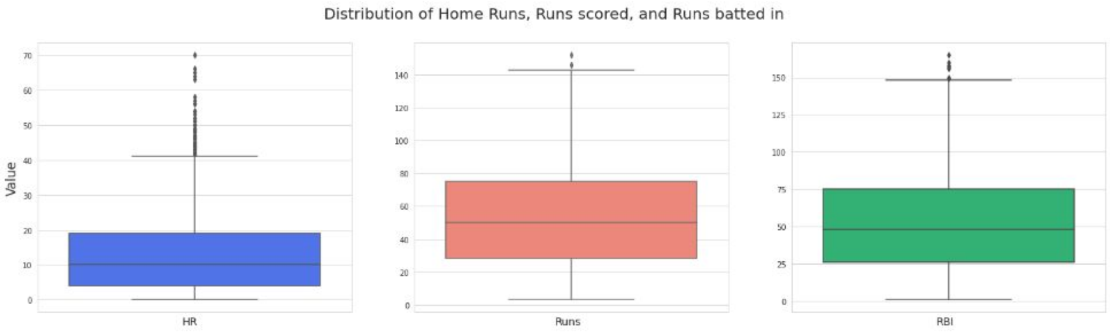


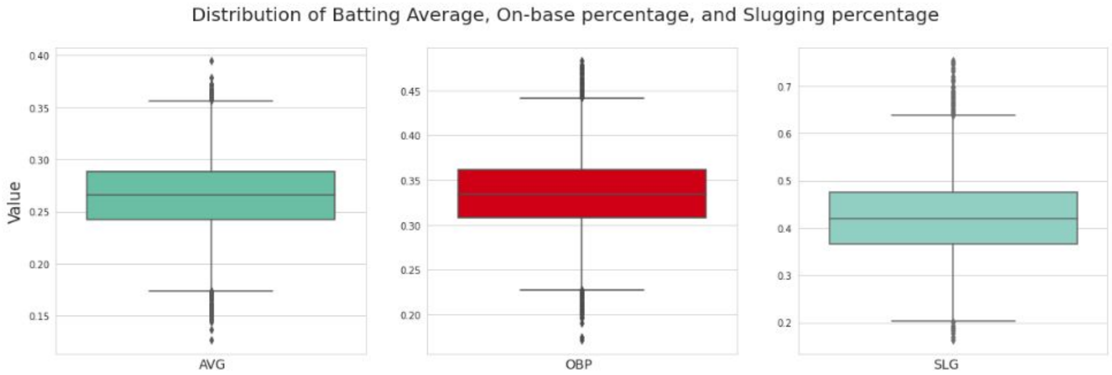
Questions 2-7: What is the relationship between some key batting stats and salary?
I will see if there is a relationship between salary and:

- Batting Average (BA)
- Home Runs (HR)
- Runs Scored (R)
- Runs Batted In (RBI)
- On-base percentage (OBP)
- Slugging percentage (SLG)
- On-base plus slugging (OPS)

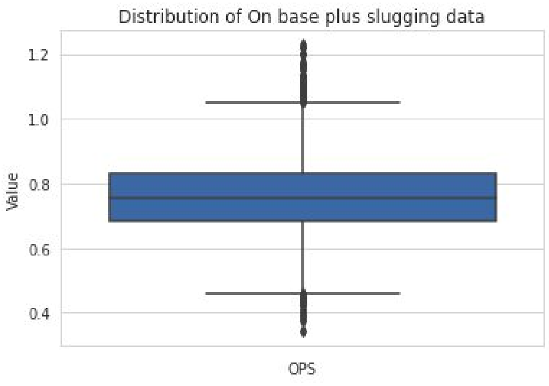
First, I will look at each of these features independently to see their distributions.

Next, I will evaluate relationships between the standardized salary and each of the 7 named features. To accomplish this, linear regression plots of each comparing the standardized salary scores with the above metrics will be created. I predict positive relationships for Home Runs and the percentage categories (OBP,SLG, OPS), but not strong positive relationships for the others (BA, R, RBIs)

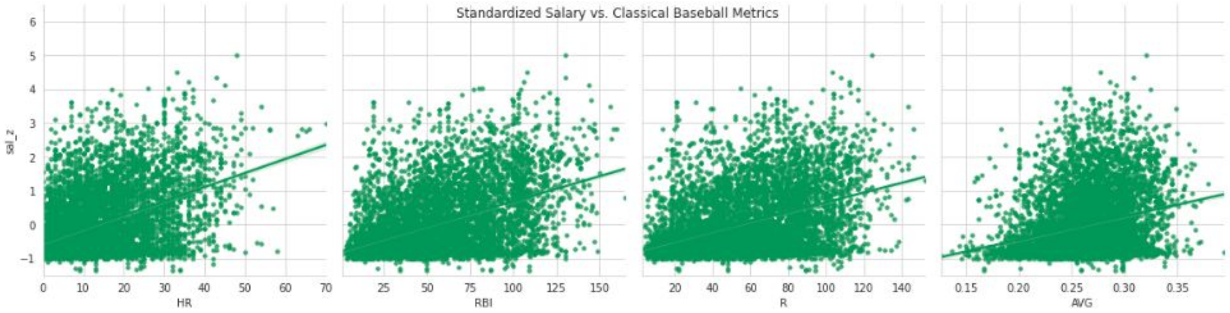




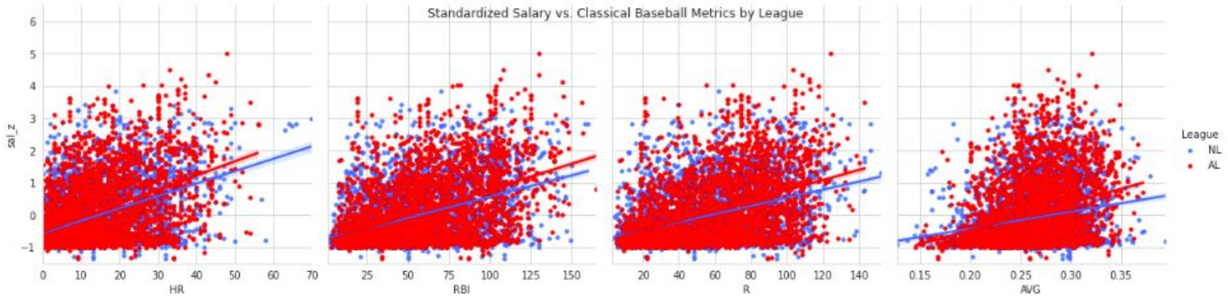
The average player had a batting average of .262, On-base percentage of .329, and a Slugging percentage of .409, and an OPS of .738.



Now let's visualize possible relationships between standardized salary and the above metrics.



As expected, there was a somewhat strong positive relationship between Standardized Salary and Home Runs, and weaker positive relationships between Standardized Salary and RBIs, Runs Scored, and Batting Average. Unsurprisingly batting average had the weakest positive relationship, but it was stronger than expected. Now let's see if there is any effect between leagues:



As expected, the National league had slighter weaker positive relationships than the American League, but for the most part nothing out of the ordinary here.

Conclusion

I first looked at how the average salary has changed over time, by mean and median. While the mean salary has increased 9-fold since 1988 (from 500,000 USD in '88 to 4.5 million USD in '16), the median salary has tripled since 1985, from below 500k to 7.1 Million.

This salary data was then joined with batting data on Player ID and Year, and the salary data was then standardized to put each salary point in context with its respective year. This standardized salary was then analyzed to see if there were positive relationships between it and:

- Batting Average
- Home Runs
- Runs Scored
- Runs Batted In
- On-base percentage
- Slugging percentage
- OPS (On-base plus slugging)

There are positive relationships between standardized salary and each of the features above, but the strongest positive relationships appeared to be with OBP/OPS and Home Runs.

There are some limitations here. One, I did not include the player position in the analysis, which probably plays some role in how much a player makes. As I did not include this, I filtered the dataset using having 100 At-Bats as a threshold to weed out pitchers, who don't bat much (American League) and don't bat well, as it is not their main job. This has an effect of also filtering out players who stayed in the majors for a short time, or got hurt and missed a lot of the season. So there is one qualification:

Minimum of 100 at-bats

In the future I would like to carry out similar analyses with respect to position, pitching statistics (Wins, ERA, etc.) as well as WAR.