Baseball Salary Analysis

Edgar Hernandez, Ryan Sims, Jessica Tomas

2022-08-03

Contents

Introduction

We will be examining what statistics related to a baseball player's performance have a significant effect on the salary of a player. This is particularly interesting as it is sort of the flip-side of the coin to the traditional Sabermetrics employed by the Oakland Athletics in the 1990's to analyze player statistics to try to assemble a winning team. We will be instead be leveraging the dataset to instead work for the players, hopefully determining which statistic(s) a player should focus on improving in order to increase their compensation.

Before we begin our analysis, we must first load the data using the `Lahman` library:

```
library(Lahman)
```

We then join the Salaries, Fielding, and Batting tables:

Methods

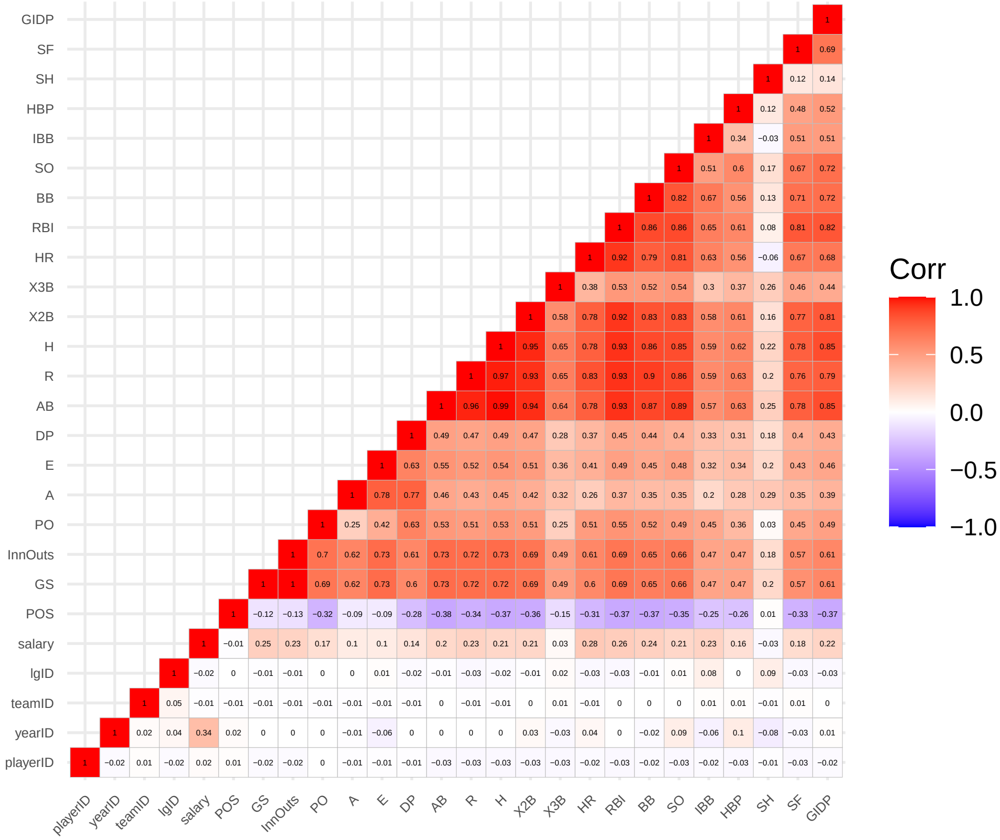
First, we visualize pairwise correlations between `Salary` and other variables:

- Calculating the correlations:

```
salary_data = as.data.frame(lapply(salary_data, as.integer))
salary_data_cor = cor(as.data.frame(salary_data))
```

- Plotting the relationships:

```
library(ggcorrplot)
my_plt = ggcorrplot(salary_data_cor, lab = TRUE, lab_size = 1, show.diag = TRUE, type = "lower")
my_plt + theme(axis.text.x = element_text(size = 5), axis.text.y = element_text(size = 5))
```



	playerID	yearID	teamID	lgID	salary	POS	GS	InnOuts	PO	A	E	DP	AB	R	H	X2B	X3B	HR	RBI	BB	SO	IBB	HBP	SH	SF	GIDP
GIDP																										1
SF																									1	0.69
SH																								1	0.12	0.14
HBP																							1	0.12	0.48	0.52
IBB																						1	0.34	−0.03	0.51	0.51
SO																					1	0.51	0.6	0.17	0.67	0.72
BB																				1	0.82	0.67	0.56	0.13	0.71	0.72
RBI																			1	0.86	0.86	0.65	0.61	0.08	0.81	0.82
HR																		1	0.92	0.79	0.81	0.63	0.56	−0.06	0.67	0.68
X3B																	1	0.38	0.53	0.52	0.54	0.3	0.37	0.26	0.46	0.44
X2B																1	0.58	0.78	0.92	0.83	0.83	0.58	0.61	0.16	0.77	0.81
H															1	0.95	0.65	0.78	0.93	0.86	0.85	0.59	0.62	0.22	0.78	0.85
R														1	0.97	0.93	0.65	0.83	0.93	0.9	0.86	0.59	0.63	0.2	0.76	0.79
AB													1	0.96	0.99	0.94	0.64	0.78	0.93	0.87	0.89	0.57	0.63	0.25	0.78	0.85
DP												1	0.49	0.47	0.49	0.47	0.28	0.37	0.45	0.44	0.4	0.33	0.31	0.18	0.4	0.43
E											1	0.63	0.55	0.52	0.54	0.51	0.36	0.41	0.49	0.45	0.48	0.32	0.34	0.2	0.43	0.46
A										1	0.78	0.77	0.46	0.43	0.45	0.42	0.32	0.26	0.37	0.35	0.35	0.2	0.28	0.29	0.35	0.39
PO									1	0.25	0.42	0.63	0.53	0.51	0.53	0.51	0.25	0.51	0.55	0.52	0.49	0.45	0.36	0.03	0.45	0.49
InnOuts								1	0.7	0.62	0.73	0.61	0.73	0.72	0.73	0.69	0.49	0.61	0.69	0.65	0.66	0.47	0.47	0.18	0.57	0.61
GS							1	1	0.69	0.62	0.73	0.6	0.73	0.72	0.72	0.69	0.49	0.6	0.69	0.65	0.66	0.47	0.47	0.2	0.57	0.61
POS						1	−0.12	−0.13	−0.32	−0.09	−0.09	−0.28	−0.38	−0.34	−0.37	−0.36	−0.15	−0.31	−0.37	−0.37	−0.35	−0.25	−0.26	0.01	−0.33	−0.37
salary					1	−0.01	0.25	0.23	0.17	0.1	0.1	0.14	0.2	0.23	0.21	0.21	0.03	0.28	0.26	0.24	0.21	0.23	0.16	−0.03	0.18	0.22
lgID				1	−0.02	0	−0.01	−0.01	0	0	0.01	−0.02	−0.01	−0.03	−0.02	−0.01	0.02	−0.03	−0.03	−0.01	0.01	0.08	0	0.09	−0.03	−0.03
teamID			1	0.05	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	−0.01	0	−0.01	−0.01	0	0.01	−0.01	0	0	0	0.01	0.01	−0.01	0.01	0
yearID		1	0.02	0.04	0.34	0.02	0	0	0	−0.01	−0.06	0	0	0	0	0.03	−0.03	0.04	0	−0.02	0.09	−0.06	0.1	−0.08	−0.03	0.01
playerID	1	−0.02	0.01	−0.02	0.02	0.01	−0.02	−0.02	0	−0.01	−0.01	−0.01	−0.03	−0.03	−0.03	−0.03	−0.03	−0.02	−0.03	−0.02	−0.03	−0.02	−0.03	−0.01	−0.03	−0.02

```
salary_data = as.data.frame(salary_data)
baseball_fit = lm(salary ~ ., data = salary_data)
baseball_aic = step(baseball_fit, direction = "backward", trace = 0)

range(salary_data$yearID)
```

```
## [1] 1985 2016
```