Baseball Speed Rating v. Offensive Production

DASI Project

Introduction:

For me, the surest sign of spring is the return of baseball–an American pastime that stretches throughout the warm days of summer. In addition to enjoying televised games and attending local games, I participate in a fantasy league. This is an on-line, recreational game in which various members of my family choose players throughout the league and compare statistics of players in a categorical, head-to-head competition.

My family plays for fun. However, many Americans put more stake in fantasy games. A 2012 article on MSN Money estimated the number of participants of fantasy baseball at 12.2 million. Large media sports networks such as CBS Sports, ESPN, Yahoo Sports, USA Today, and NBC Sports host fantasy leagues for various sports (American football, baseball, basketball, golf, hockey, and racing). In all, MSN Money projects over $2 billion dollars in yearly “economic impact”. There are leagues that people may pay to join in order to win cash prices (some payouts in excess of thousands of dollars).

Personally, I am content just to beat my sons in the fantasy standings. There are 20 offensive categories that each team must win in a weekly match up. Three of these categories (15%) comprise baseball players “stealing bases.” I am wondering, if I select (i.e., draft) players for my team who are very fast, will they lower the overall team offensive production of home runs, RBIs, batting average, etc? Therefore, my research question is:

Is there a negative association between a player's speed and his offensive production?

Data:

The baseball data I collected comes from www.rotochamp.com. This is a fantasy baseball site for statistics of past, current, and future (projections) of MLB (Major League Baseball) data. I had to scrape several web pages for projections of fielding position players (those who create offensive statistics). I gathered the 2014 “composite” projections into a csv file (see attached page). This citation is for one subset of the data (Position = Out_Fielders) http://www.rotochamp.com/baseball/PlayerRankings.aspx?Position=OF.

Each case or observation has the following form:

```
names(baseball2014)
```

```
##  [1] "Pos"     "PosRank" "Player"  "Team"    "AB"      "R"
"HR"
##  [8] "RBI"     "SB"      "AVG"     "OBP"     "SLG"     "Value"
"SBAB"
## [15] "SPD"     "OPI"
```

This table summarize the relevant (not all) variable for this project:

Variable	Description	Type
Player	Player's Name	Categorical
AB	Number of "at bats" or opportunities	Numeric-Discrete
R	Number of runs scored	Numeric-Discrete
HR	Number of home runs	Numeric-Discrete
RBI	Number of runs batted in	Numeric-Discrete
SB	Number of stolen bases	Numeric-Discrete
AVG	Ratio of hits to at bats	Numeric-Continuous
OBP	Ratio of reaching first base per at bats	Numeric-Continuous
SLG	Ratio composite of number of bases times hits per at bats	Numeric-Continuous

In order to complete this project, I had to create three additional variables: SBAB, SPD, and OPI.

SBAB is a continuous, numerical ratio of SB / AB (stolen base count to "at bats" or opportunities).

SPD (Speed rating) is a categorical rating based on SBAB quantiles:

0 - 20th	Very Slow
21 -40th	Slow
41 - 60th	Average
61 - 80th	Fast
81 - 99th	Very Fast

OPI (Offensive Production Index) is a composite index of offensive stats times efficiency of chance that I created from the rotochamp data.

OPI = sqrt((R + HR + RBI) * (sqrt (AB) * (AVG + OBP + SLG) / 3))

The square root transformations shape the data into nearly a normal distribution.

One complete observation looks like:

```
baseball2014[5, ]
```

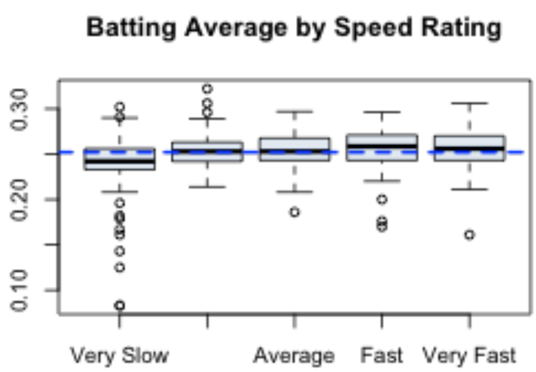
```
##   Pos PosRank        Player Team  AB  R HR RBI SB   AVG   OBP
SLG  Value
## 5  1B      13 Albert Pujols  LAA 484 75 24  83  4 0.277 0.348
0.483 $19.00
##       SBAB  SPD   OPI
## 5 0.008264 Slow 38.46
```

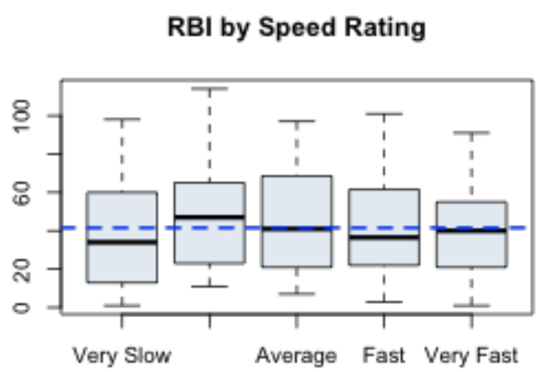
Analysis for this project is based upon the categorical variable “SPD” and numerical variable “OPI”

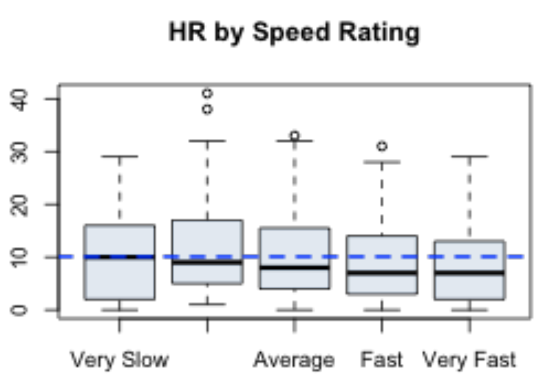
This is an observational study. There is no experimental design or treatment. The data is based on statistical projections for actual professional baseball players. Since the data is based upon observed (though, projected) data, the project cannot show causality, only an association. Moreover, the sampling (using inference function) will be generalizable to the entire data set of professional baseball players. The inference function will randomly select observations from each of the 5 categories to compare their respective OPI means.

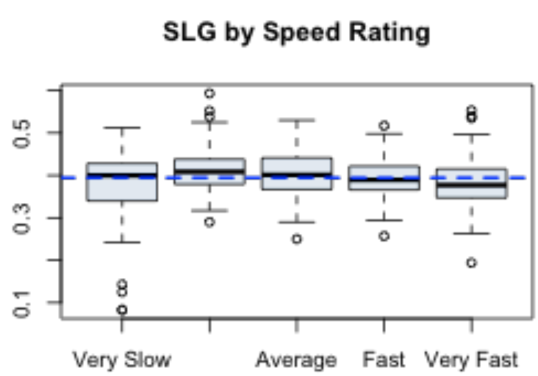
Exploratory data analysis:

For starters, I looked at the relationship between speed and each offensive variable (SLG, HR, RBI, and AVG. For brevity, I have included just the summary boxplots below.









This is a purely speculative examination. For the most part, each fantasy team consists of the best players in the league. Team owners look to pick up the top players at each position. Since there are 30 clubs in MLB and the fantasy league consists of 8 teams, I am going to quickly investigate how the top 25% of each offensive category would be modeled as proportions of each speed category. Again, this is merely a rough estimation as how offensive stats are related to speed and will help me formulate a hypothesis later.

```
##     Very Slow Slow Average Fast Very Fast
## HR         26   31      26   21        17
## SLG        20   33      27   19        17
## AVG        14   21      25   30        32
## RBI        21   29      30   23        14
```

From the above exploration, I can see that each offensive stat has a different relationship with speed. Batting average appears to be the least affected by speed; whereas, HR and SLG are inversely related to speed. Although these data relationships are interesting, they are too fragmented to make a clear and concise statement about the relationship of speed with offense. In order to tell the overall impact of speed on offensive production, I am going to need the OPI to SPD analysis of variance below.

Inference:

Since I have one numeric and one categorical variable, I am going to use ANOVA with the inference function.

My hypothesis for the ANOVA is:

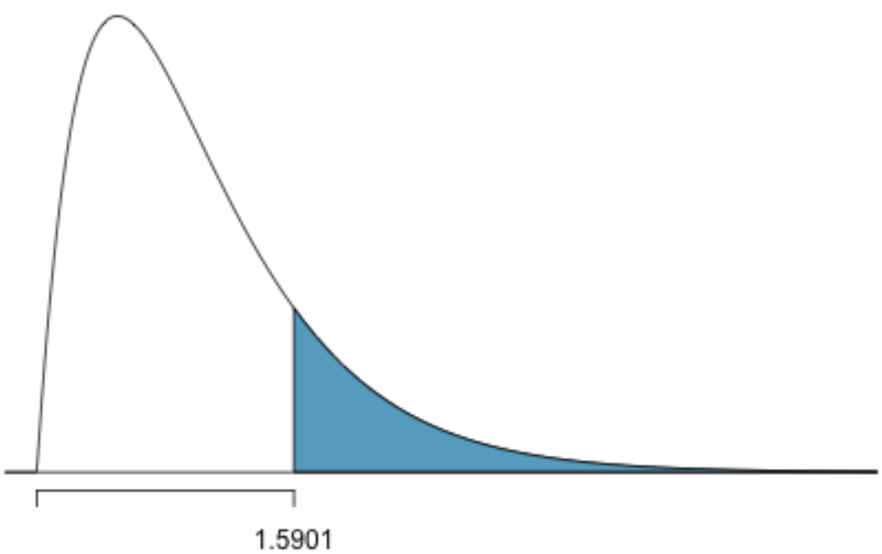
H_0 = mean(SPD = “Very Slow”) = mean(SPD = “Slow”) = mean(SPD = “Average”) = mean(SPD = “Fast”) = mean(SPD = “Very Fast”)

H_A= at least two categorical means are different from each other

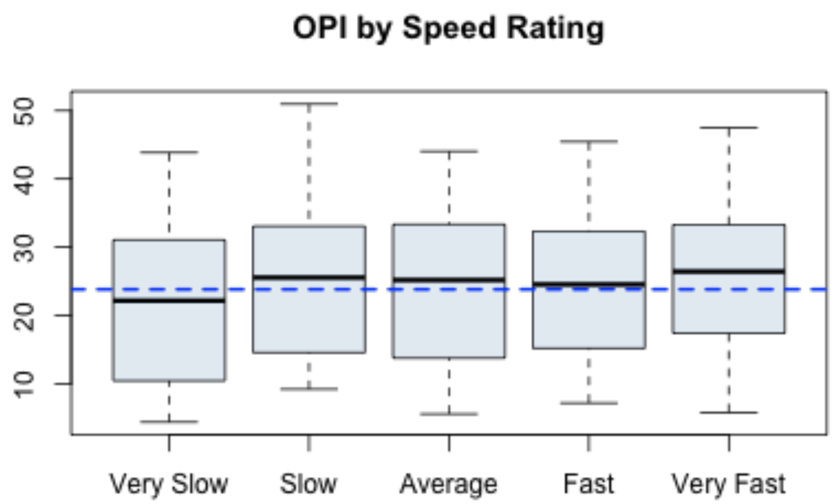
```
################ Inference Function for ANOVA of means:
##############

inference(y = baseball2014$OPI, x = baseball2014$SPD, est = "mean",
type = "ht",
    alternative = "greater", method = "theoretical", eda_plot =
FALSE)
```

```
## Response variable: numerical, Explanatory variable: categorical
## ANOVA
## Summary statistics:
## n_Average = 91, mean_Average = 23.68, sd_Average = 11.16
## n_Fast = 89, mean_Fast = 24.2, sd_Fast = 9.995
## n_Slow = 94, mean_Slow = 24.4, sd_Slow = 10.22
## n_Very Fast = 87, mean_Very Fast = 25.26, sd_Very Fast = 10.24
## n_Very Slow = 84, mean_Very Slow = 21.43, sd_Very Slow = 10.97
## H_0: All means are equal.
## H_A: At least one mean is different.
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value Pr(>F)
## x            4    704     176    1.59   0.18
## Residuals 440  48731     111
```



```
boxplot(baseball2014$OPI ~ baseball2014$SPD, main = "OPI by Speed
Rating", col = "#33669925",
    at = c(0, 1, -1, 2, -2))
abline(h = mean(baseball2014$OPI, na.rm = T), col = "blue", lty =
2, lwd = 2)
```



Conditions for ANOVA and hypothesis testing: From the OPI by Speed Rating plot above, I can tell the conditions for ANOVA were met. There is independence within each group and between each group. Each baseball player is independent of another and is only listed in one SPD category. Furthermore, each boxplot shows approximate normality and equal variance for each group.

From the inference function and ANOVA analysis, the F value (1.59) is small and the p-value (0.18) is larger than a small significant value. Therefore, I cannot reject the null hypothesis. In other words, the means across the SPD (speed rating) categories are not significantly different.

Conclusion:

In the end, it seems much to do about nothing. The p-value of .18 > .05 tells me that I cannot dismiss the hypothesis that states that the mean OPIs across the SPD categories are the same. A look at the side-by-side boxplots shows how similar each group is. As far as my fantasy draft is concerned, I should be able to find speedy players who will not hurt my overall offensive production. That said, I should probably start to worry about my pitching now…

Appendices: Data Page and Descriptive Stats

```
by(baseball2014$AVG, baseball2014$SPD, describe)
```

```
## baseball2014$SPD: Average
##   var  n mean   sd median trimmed  mad  min max range  skew
kurtosis se
## 1   1 91 0.26 0.02   0.25    0.26 0.02 0.19 0.3  0.11 -0.38
1.14  0
## --------------------------------------------------------
## baseball2014$SPD: Fast
##   var  n mean   sd median trimmed  mad  min max range  skew
kurtosis se
## 1   1 92 0.26 0.02   0.26    0.26 0.02 0.17 0.3  0.13 -1.14
2.41  0
## --------------------------------------------------------
## baseball2014$SPD: Slow
##   var  n mean   sd median trimmed  mad  min  max range skew
kurtosis se
## 1   1 94 0.25 0.02   0.25    0.25 0.02 0.21 0.32  0.11 0.67
1.26  0
## --------------------------------------------------------
## baseball2014$SPD: Very Fast
##   var  n mean   sd median trimmed  mad  min  max range  skew
kurtosis se
## 1   1 93 0.26 0.02   0.26    0.26 0.02 0.16 0.31  0.14 -0.65
2.7  0
## --------------------------------------------------------
## baseball2014$SPD: Very Slow
##   var  n mean   sd median trimmed  mad  min max range  skew
kurtosis se
## 1   1 93 0.24 0.04   0.24    0.24 0.02 0.08 0.3  0.22 -2.04
5.7  0
```

```
by(baseball2014$RBI, baseball2014$SPD, describe)
```

```
## baseball2014$SPD: Average
##   var  n  mean    sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 91 43.75 26.41     41   42.64 34.1   7  97    90 0.27
-1.32 2.77
## ------------------------------------------------------------
## baseball2014$SPD: Fast
##   var  n  mean    sd median trimmed   mad min max range skew
kurtosis   se
## 1   1 92 41.24 23.72   36.5    40.2 27.43   3 101    98 0.35
-0.96 2.47
## ------------------------------------------------------------
## baseball2014$SPD: Slow
##   var  n  mean    sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 94 46.21 24.94     47   44.71 34.1  11 114   103 0.38
-0.87 2.57
## ------------------------------------------------------------
## baseball2014$SPD: Very Fast
##   var  n  mean    sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 93 39.34 23.43     40   38.39 25.2   1  91    90 0.24
-0.79 2.43
## ------------------------------------------------------------
## baseball2014$SPD: Very Slow
##   var  n  mean    sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 93 37.13 27.34     34   35.21 34.1   1  98    97 0.46
-0.97 2.83
```

```
by(baseball2014$HR, baseball2014$SPD, describe)
```

```
## baseball2014$SPD: Average
##   var  n  mean   sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 91 10.89 8.68      8    9.95 7.41   0  33    33 0.81
-0.4 0.91
## ------------------------------------------------------------
## baseball2014$SPD: Fast
##   var  n mean   sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 92 9.41 7.31      7    8.54 6.67   0  31    31 0.91
-0.02 0.76
## ------------------------------------------------------------
## baseball2014$SPD: Slow
##   var  n  mean  sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 94 11.91 8.5      9   11.03 7.41   1  41    40    1
0.72 0.88
## ------------------------------------------------------------
## baseball2014$SPD: Very Fast
##   var  n mean   sd median trimmed  mad min max range skew
kurtosis   se
## 1   1 93 8.45 7.36      7    7.49 7.41   0  29    29    1
0.27 0.76
## ------------------------------------------------------------
## baseball2014$SPD: Very Slow
##   var  n mean sd median trimmed   mad min max range skew
kurtosis   se
## 1   1 93 9.77  8     10    9.09 10.38   0  29    29 0.54
-0.76 0.83
```

```
by(baseball2014$SLG, baseball2014$SPD, describe)
```

```
## baseball2014$SPD: Average
##   var  n mean   sd median trimmed  mad  min  max range skew
kurtosis   se
## 1   1 91  0.4 0.05    0.4     0.4 0.06 0.25 0.53  0.28 0.07
-0.14 0.01
## ------------------------------------------------------------
## baseball2014$SPD: Fast
##   var  n mean   sd median trimmed  mad  min  max range skew
kurtosis se
## 1   1 92  0.4 0.05   0.39    0.39 0.04 0.26 0.52  0.26 0.13
0.35  0
## ------------------------------------------------------------
## baseball2014$SPD: Slow
##   var  n mean   sd median trimmed  mad  min  max range skew
kurtosis   se
## 1   1 94 0.41 0.05   0.41    0.41 0.04 0.29 0.59   0.3 0.56
0.98 0.01
## ------------------------------------------------------------
## baseball2014$SPD: Very Fast
##   var  n mean   sd median trimmed  mad  min  max range skew
kurtosis   se
## 1   1 93 0.38 0.06   0.38    0.38 0.05 0.19 0.55  0.36 0.23
1.04 0.01
## ------------------------------------------------------------
## baseball2014$SPD: Very Slow
##   var  n mean   sd median trimmed  mad  min  max range  skew
kurtosis   se
## 1   1 93 0.38 0.08    0.4    0.39 0.05 0.08 0.51  0.43 -1.67
3.75 0.01
```

	row.names	Pos	PosRank	Player	Team	AB	R	HR	RBI	SB	AVG	OBP	SLG	Value	SBAB
1	160	3B	1	Miguel Cabrera	DET	550	98	38	114	3	0.322	0.414	0.593	$52.00	0.005454545
2	372	OF	1	Mike Trout	LAA	566	112	27	89	35	0.306	0.405	0.535	$53.00	0.061837456
3	52	1B	1	Paul Goldschmidt	ARI	561	91	31	101	14	0.280	0.376	0.517	$38.00	0.024955437
4	54	1B	4	Prince Fielder	TEX	557	83	29	98	1	0.285	0.382	0.497	$30.00	0.001795332
5	254	OF	3	Andrew McCutchen	PIT	569	95	22	87	23	0.292	0.378	0.482	$36.00	0.040421793
6	17	1B	2	Chris Davis	BAL	547	83	41	102	3	0.267	0.342	0.552	$34.00	0.005484461
7	138	3B	2	Edwin Encarnacion	TOR	537	86	33	97	8	0.276	0.367	0.520	$31.00	0.014897579
8	32	1B	5	Joey Votto	CIN	519	88	25	82	6	0.295	0.423	0.511	$28.00	0.011560694
9	390	OF	2	Ryan Braun	MIL	518	85	29	91	18	0.297	0.369	0.539	$36.00	0.034749035
10	122	3B	3	Adrian Beltre	TEX	559	79	28	94	1	0.302	0.351	0.512	$29.00	0.001788909
11	110	2B	1	Robinson Cano	SEA	572	84	23	88	5	0.297	0.362	0.490	$29.00	0.008741259
12	274	OF	4	Carlos Gonzalez	COL	506	84	29	87	21	0.298	0.365	0.553	$36.00	0.041501976
13	411	OF	5	Yasiel Puig	LAD	563	94	27	75	19	0.284	0.353	0.496	$32.00	0.033747780
14	4	1B	7	Adrian Gonzalez	LAD	576	78	22	95	1	0.285	0.345	0.460	$23.00	0.001736111
15	25	1B	6	Freddie Freeman	ATL	532	80	23	90	2	0.289	0.368	0.483	$24.00	0.003759398
16	246	OF	6	Adam Jones	BAL	578	81	28	90	12	0.285	0.323	0.491	$30.00	0.020761246
17	140	3B	5	Evan Longoria	TB	538	83	27	91	3	0.260	0.345	0.478	$20.00	0.005576208
18	240	DH	2	Billy Butler	KAN	580	71	19	88	1	0.288	0.367	0.447	$16.00	0.001724138
19	340	OF	8	Jose Bautista	TOR	464	85	32	84	7	0.265	0.377	0.530	$26.00	0.015086207
20	268	OF	7	Bryce Harper	WAS	521	88	25	76	15	0.278	0.366	0.497	$26.00	0.028790787
21	363	OF	12	Matt Holliday	STL	514	85	21	85	5	0.278	0.363	0.465	$22.00	0.009727626
22	461	SS	1	Troy Tulowitzki	COL	481	74	26	84	4	0.306	0.380	0.538	$28.00	0.008316008
23	327	OF	11	Jay Bruce	CIN	533	78	30	92	7	0.257	0.333	0.492	$23.00	0.013133208
24	7	1B	12	Anthony Rizzo	CHC	567	74	26	87	6	0.265	0.345	0.473	$19.00	0.010582011
25	24	1B	9	Eric Hosmer	KAN	578	77	18	79	12	0.289	0.350	0.446	$21.00	0.020761246
26	248	OF	21	Alex Gordon	KAN	591	87	17	75	10	0.271	0.340	0.431	$18.00	0.016920474
27	133	3B	4	David Wright	NYM	527	79	20	81	16	0.279	0.363	0.465	$21.00	0.030360531
28	6	1B	11	Allen Craig	STL	527	75	18	90	3	0.287	0.343	0.455	$19.00	0.005692600
29	12	1B	14	Buster Posey	SF	525	72	18	85	2	0.291	0.369	0.461	$24.00	0.003809524
30	316	OF	14	Hunter Pence	SF	569	80	20	87	12	0.267	0.326	0.438	$20.00	0.021089631
31	241	DH	1	David Ortiz	BOS	460	74	24	86	2	0.289	0.380	0.520	$19.00	0.004347826
32	311	OF	13	Giancarlo Stanton	MIA	491	76	32	83	4	0.261	0.361	0.525	$20.00	0.008146640
33	33	1B	8	Jose Dariel Abreu	CWS	487	80	30	78	5	0.273	0.354	0.513	$22.00	0.010266941
34	399	OF	18	Shin-Soo Choo	TEX	544	89	17	63	19	0.270	0.382	0.430	$18.00	0.034926471
35	437	SS	2	Hanley Ramirez	LAD	505	79	23	81	18	0.275	0.341	0.477	$27.00	0.035643564
36	408	OF	24	Wil Myers	TB	550	77	22	83	8	0.260	0.329	0.442	$16.00	0.014545455
37	14	1B	16	Carlos Santana	CLE	523	77	20	80	4	0.254	0.367	0.438	$19.00	0.007648184
38	5	1B	13	Albert Pujols	LAA	484	75	24	83	4	0.277	0.348	0.483	$19.00	0.008264463
39	80	2B	3	Dustin Pedroia	BOS	542	77	12	73	15	0.295	0.367	0.435	$19.00	0.027675277
40	351	OF	20	Justin Upton	ATL	517	85	22	70	12	0.265	0.353	0.455	$18.00	0.023210832
41	100	2B	7	Matt Carpenter	STL	551	89	10	65	4	0.283	0.363	0.428	$14.00	0.007259528
42	46	1B	10	Michael Cuddyer	COL	504	70	21	78	9	0.296	0.357	0.498	$19.00	0.017857143
43	145	3B	6	Josh Donaldson	OAK	546	77	20	76	6	0.267	0.343	0.443	$13.00	0.010989011
44	87	2B	2	Jason Kipnis	CLE	544	81	16	75	24	0.267	0.347	0.423	$21.00	0.044117647
45	202	C	3	Joe Mauer	MIN	533	77	11	68	3	0.296	0.384	0.428	$19.00	0.005628518
46	99	2B	6	Martin Prado	ARI	583	75	13	73	6	0.286	0.339	0.425	$14.00	0.010291595
47	272	OF	23	Carlos Beltran	NYY	500	72	23	81	5	0.276	0.339	0.474	$17.00	0.010000000
48	360	OF	22	Mark Trumbo	ARI	524	71	29	89	5	0.252	0.307	0.475	$17.00	0.009541985
49	86	2B	4	Ian Kinsler	DET	559	88	16	63	17	0.267	0.340	0.420	$18.00	0.030411449
50	68	2B	5	Brandon Phillips	CIN	559	76	17	79	8	0.270	0.319	0.415	$16.00	0.014311270
51	321	OF	10	Jacoby Ellsbury	NYY	571	87	14	59	38	0.284	0.339	0.431	$25.00	0.066549912
52	66	2B	9	Ben Zobrist	TB	554	81	14	70	12	0.258	0.349	0.408	$14.00	0.021660650
53	325	OF	27	Jason Heyward	ATL	522	81	22	64	10	0.264	0.351	0.456	$14.00	0.019157088
54	258	OF	32	Austin Jackson	DET	563	86	13	60	12	0.274	0.343	0.423	$12.00	0.021314387
55	273	OF	9	Carlos Gomez	MIL	536	78	23	72	36	0.265	0.316	0.468	$25.00	0.067164179