# HW4 - Individual performance metric

*Narek Sahakyan*

*8/11/2019*

## Individual Perfomance in Baseball

I have decided to measure individual perfomance of the players and use my own metrics for identyfing their skils, Baseball has a lot of data available, but to have some contemporary facotrs I have filtered the data so that it contains information about baseball games in the twentieth century. The data was derived from the official website of Seah Lahman - data
As baseball is a very popular sport with a lot of statistical approaches
and measures, I decided to create my own metrics based on the knowledge I derived
about the sport during two days
It is very interesting for me to analyze a sport that I don't know well.
I will use several approaches for measuring individual perfomance of players,
and will visualize findings for them.
The methods are

- Awards won by the players
- Batter's efficiency
- Pitcher's efficiency

## Awards won by players

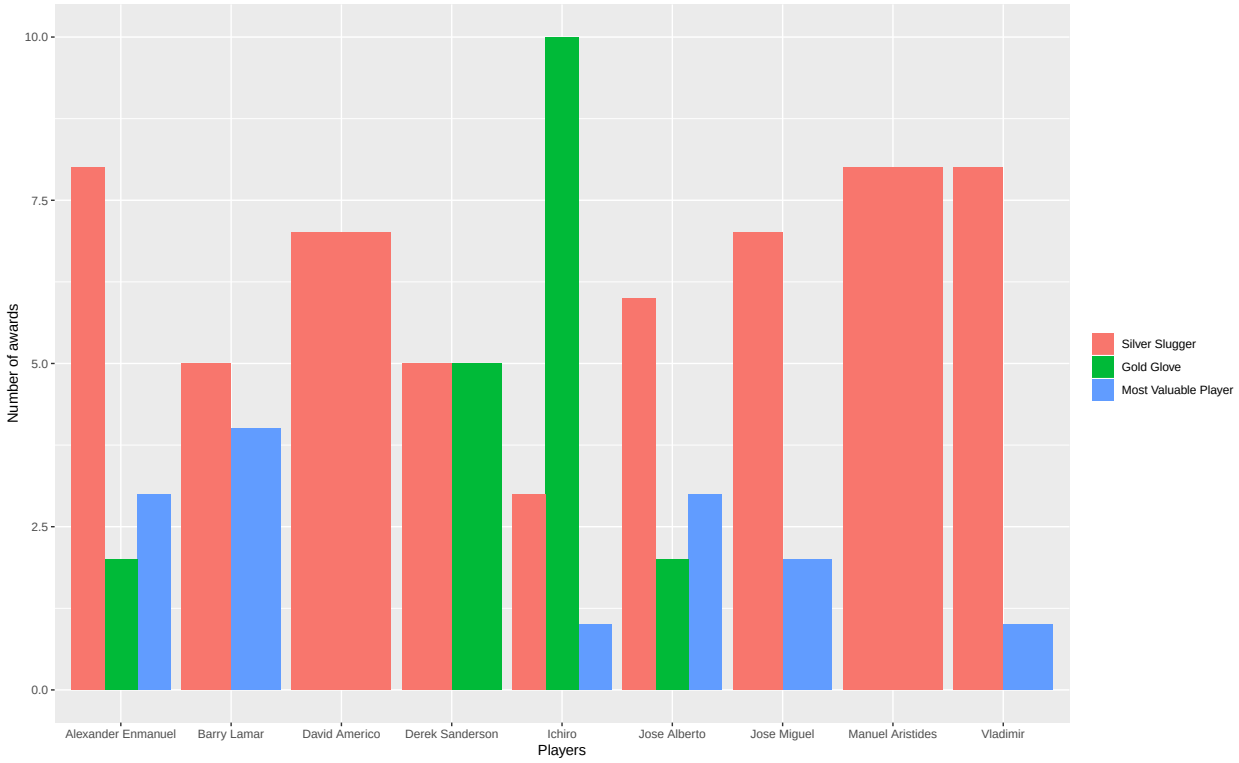
```
##             real_name award_count
## 1  Alexander Enmanuel          26
## 2        Jose Alberto          23
## 3         Barry Lamar          19
## 4     Derek Sanderson          19
## 5              Ichiro          19
## 6       David Americo          18
## 7    Manuel Aristides          17
## 8         Jose Miguel          16
## 9            Vladimir          15
## 10            Mariano          15
## 11      Andruw Rudolf          14
## 12        Torii Kedar          13
## 13     Michael Nelson          13
## 14          Todd Lynn          12
## 15               Ivan          12
## 16     Gerald Dempsey          11
## 17        Scott Bruce          11
## 18        Jose Carlos          10
## 19     Clayton Edward          10
## 20       Gregory Alan          10
```

Here is the list of the 20 players with the
highetst ammount of awards in the specified periods
Let's pick top 10 from these players and analyze(visualize)
their awards and positions

Let's pick the most prestigious awards and see how were these distributed among these players

Let's pick three awards

- MVP(Most Valuable Player)
- Gold Glove(Best deffensive player)
- Silver Slugger(Best offensive player)




As the visualization shows only a few players earned all the three awards
Most of them were able to be anounced best in their playing position(offensive or defensive)
Ichore got the highest number of "Golden Glove" awards and he managed to get "MVP" and "Silver Slugger" trophies as well
However we can see that the attacking players are more likely to be recognized the "MVP" as there are more cases of players winning the "Silver Slugger" and "MVP" rather than winning the "MVP" and "Golden Glove"

Now let's talk about the drawbacks and advantages of this method for evaluating "IP"(Individual Perfomance)
In my opinion this approach is very useful to identify the well recognized players for the teams and see how they perfomance was evaluated by the public and professionals.
However this kind of approach does not give much information about the player's impact on the team
Also I believe that the players with the biggest trophy rooms should be among the top earners of MLB, so let's check my statement

```
##             real_name yearID   salary
## 1  Alexander Enmanuel   2009 33000000
```

```
## 2  Alexander Enmanuel   2010 33000000
## 3      Clayton Edward   2016 33000000
## 4      Clayton Edward   2015 32571000
## 5  Alexander Enmanuel   2011 32000000
## 6      Donald Zachary   2016 31799030
## 7  Alexander Enmanuel   2012 30000000
## 8        David Taylor   2016 30000000
## 9  Alexander Enmanuel   2013 29000000
## 10 Alexander Enmanuel   2008 28000000
```

As we can see Alexander Emmanuel dominates in the market earning the highest salaries in 2008 - 2011, 2016. Let's take calculate the mean of the mlb player's salaries from 1999-2016 and then check which positions do our top award owners have among the highest earners

```
##             realName meanSalary position
## 1        Barry Lamar   23168485        1
## 2        Jose Miguel   16449256        3
## 3           Vladimir   15415982        5
## 4    Derek Sanderson   15078369        7
## 5      David Americo   13457902       11
## 6       Jose Alberto   12752527       19
## 7   Manuel Aristides   10258218       56
## 8 Alexander Enmanuel    9627804       68
## 9             Ichiro    9402500       73
```

As we can see over time Barry Lamar had higher salary adn Alexandr Enmanuel is only the 68th among the top earners of the league in a given period.

To sum it up, this approach is very useful to identify the top earners of the league, and as we can see
if the player wins many awards he gets higher salary. Alexander Enmanuel is not among the top owners over time
But he got very high salaries in years when he did good and earned many titles. To check if the high earners get more awards let's divide their salaries by 1,000,000 and check the correlation
between the number of trophies and salary

```
##
##  Pearson's product-moment correlation
##
## data:  awards_by_players$meanSalary and awards_by_players$award_count
## t = 14.365, df = 382, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5231653 0.6535737
## sample estimates:
##       cor
## 0.5922341
```

As we can see the number of throphies of the player and his salary in millions are correlated to each other with a value of 0.59
Now let's try to build a very simple model which will estimate the number of player's trophies based on his salary

```
##
## Call:
## lm(formula = award_count ~ 0 + meanSalary, data = awards_by_players)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6089 -1.7740 -0.2693  1.0586 15.2559
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## meanSalary  0.60726    0.02203   27.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.927 on 383 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.6649, Adjusted R-squared:  0.664
## F-statistic:   760 on 1 and 383 DF,  p-value: < 2.2e-16
```

```
## meanSalary
##  0.6072605
```

According to this model, a player should earn at least around 1.6m for one award.
I believe that, the positive correlation between these two variables is present, as when the players get more awards they get recognized by the public, and become more attractive for the teams, so they offer them good contracts

## Batter's efficiency

Let's start from Batters and see how many hits do they need for one home run and other types of hit. In order to have relevant metrics let's exclude the batters with less then usual gaming practice as they could have been just lucky for having a high accuracy

```
##                   realName meanHFHR meanHF2B  meanHF3B   meanAB     meanR
## 1               Mark David 2.166667 8.272727 273.00000 352.0000  75.33333
## 2          Joseph Nicholas 2.432099 4.690476  49.25000 474.5000  83.50000
## 3             Ryan Michael 2.529412 4.526316  17.20000 220.5000  36.00000
## 4              Barry Lamar 2.907514 5.106599  71.85714 398.0000 106.87500
## 5           Maxwell Steven 2.971429 6.117647  52.00000 395.0000  75.00000
## 6       Vernon Christopher 3.335484 4.877358 103.40000 392.5000  54.00000
## 7              Aaron James 3.405063 5.847826  89.66667 477.5000 102.50000
## 8                 Byung Ho 3.416667 4.555556  41.00000 215.0000  28.00000
## 9              Kyle Joseph 3.444444 6.888889  49.60000 360.6667  61.00000
## 10            James Howard 3.461187 5.300699 126.33333 429.8462  80.53846
## 11            Russell Oles 3.463687 4.661654  77.50000 204.0000  28.46154
## 12          Samuel Peralta 3.497024 5.964467  97.91667 518.6250  93.50000
## 13               Adam Troy 3.500000 4.882716 158.20000 476.6429  76.42857
## 14     Khristopher Adrian 3.502591 4.727273  84.50000 454.3333  70.83333
## 15 Giancarlo Cruz-Michael 3.685246 4.762712 102.18182 466.0000  75.33333
```

As we can see the best player in this metric is Mark David who needed around 2 hits for a home run
Only the top 5 players in this metric were able to get a home run with less than 3 hits in general(mean taken)
Using this approach we can find the best candidates for early batting.
I believe that the players who need the least ammount of hits for a home run, single, double or triple should be
ordered in the first places of batters as they have high speed and lots of energy at the beginning of the game
This type of metrics can be useful to identify the efficiency of the players who usually swing but I don't see any usage of this approach for players who sometimes prefer to not swing

Now let's check whether the mean of these metrics is associated with the number of occurances of the player at the bat

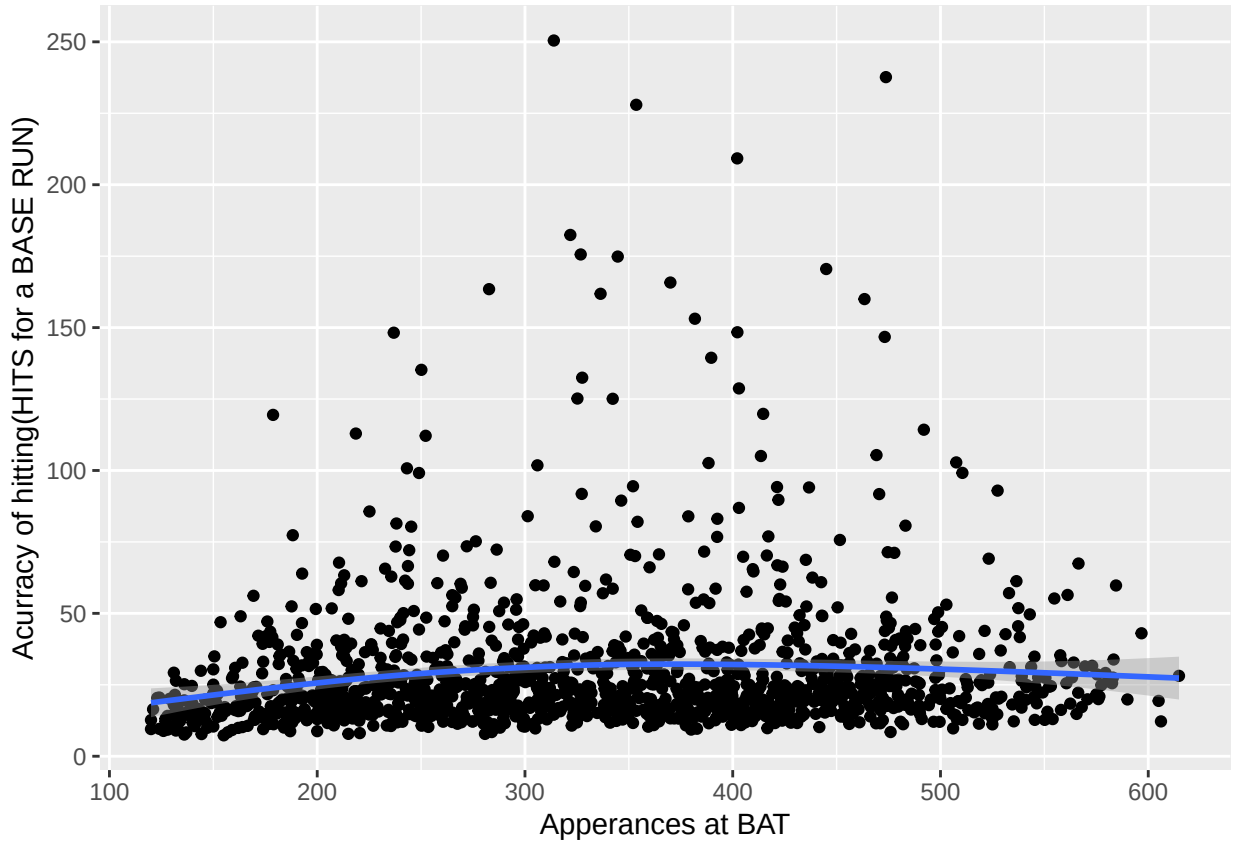
```
##
## Call:
## lm(formula = meanAB ~ 0 + meanHFHR + meanHF2B + meanHF3B, data = home_run_scorers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1598.99   -59.91    29.77   127.25   351.73
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## meanHFHR -0.91378    0.22686  -4.028 5.93e-05 ***
## meanHF2B 54.23382    1.18567  45.741  < 2e-16 ***
## meanHF3B  0.57820    0.05126  11.279  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.9 on 1381 degrees of freedom
## Multiple R-squared:  0.8299, Adjusted R-squared:  0.8296
## F-statistic:  2246 on 3 and 1381 DF,  p-value: < 2.2e-16
```

```
##   meanHFHR   meanHF2B   meanHF3B
## -0.9137839 54.2338153  0.5781957
```

As the model suggests the player's accuracy in making Home runs
has the highest effect on the number of appearences at bat, meaning that if the player needs less amount of hits for a home run he is more likely to be at bat.

```
##
##  Pearson's product-moment correlation
##
## data:  home_run_scorers$meanACC and home_run_scorers$meanAB
## t = 3.2368, df = 1382, p-value = 0.001238
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03420359 0.13879789
## sample estimates:
##       cor
## 0.08673976
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




Altough the variables are not highly correlated, we can see that in general most of the players earn their place at batting for having a high accuracy of hit to run to base ratio(the lower the better)

Now let's check the connection of hitting accuracy and runs scored by the player

```
##
## Call:
## lm(formula = meanR ~ 0 + ., data = home_run_scorers[, -c(1, 2,
##     6, 8)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -225.524  -10.788    2.556   18.857   78.235
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## meanHFHR -0.193919   0.036542  -5.307 1.30e-07 ***
## meanHF2B  7.673129   0.190990  40.176  < 2e-16 ***
## meanHF3B  0.065752   0.008258   7.963 3.49e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.67 on 1381 degrees of freedom
## Multiple R-squared:  0.7742, Adjusted R-squared:  0.7737
## F-statistic:  1578 on 3 and 1381 DF,  p-value: < 2.2e-16
```

```
##    meanHFHR    meanHF2B    meanHF3B
## -0.19391860  7.67312895  0.06575165
```

```
##   meanACC
## 0.9174741
```

As we can see the double runs make highest impact on the number of runs.
Also the mean of the hits accuracies has a very linear relationship with the number of runs

```
##
##  Pearson's product-moment correlation
##
## data:  home_run_scorers$meanR and home_run_scorers$meanACC
## t = 0.12011, df = 1382, p-value = 0.9044
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04947015  0.05591376
## sample estimates:
##         cor
## 0.003230777
```

The correlation is very small
Almost the same scenatio as with the appearances at bat.

Now let's find out the players who are to tag up, or beat by the deffence
We should filter the players with the lowest numbers of caught stealing, strikeouts and hits by the pitch

```
##             realName     meanSO meanHBP meanCS      meanR
## 1        David Allan  0.0000000       0      0  0.0000000
## 2   Fernando Antonio  0.0000000       0      0  0.0000000
## 3       Jeremiah Lee  0.0000000       0      0  0.0000000
## 4            Alfredo  0.0000000       0      0  0.0000000
## 5        Juan Carlos  5.6666667       0      0  0.3333333
## 6        Jon Michael  0.0000000       0      0  0.0000000
## 7        Terry Wayne  1.3333333       0      0  0.0000000
## 8        Jim Charles 45.0000000       0      0 19.0000000
## 9     Jonathan Scott  0.5000000       0      0  0.0000000
## 10      Jeremy David  0.5000000       0      0  0.0000000
## 11 Christopher Louis 19.0000000       0      1 11.0000000
## 12    Richard Warren  0.0000000       0      0  0.0000000
## 13     Matthew James  0.1666667       0      0  0.1666667
## 14      Alberto Jose  0.0000000       0      0  0.0000000
## 15   Manuel de Jesus 39.5000000       0      0 23.5000000
```

```
##
## Call:
## lm(formula = meanR ~ 0 + ., data = hardest_to_beat[, -c(1, 6)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.856  -3.419   0.000   2.860  64.942
```

```
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## meanSO  0.428212   0.006866   62.36   <2e-16 ***
## meanHBP 1.625920   0.103120   15.77   <2e-16 ***
## meanCS  3.416931   0.124276   27.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.36 on 2480 degrees of freedom
## Multiple R-squared:  0.9242, Adjusted R-squared:  0.9241
## F-statistic: 1.008e+04 on 3 and 2480 DF,  p-value: < 2.2e-16
```

```
##    meanSO   meanHBP    meanCS
## 0.4282119 1.6259202 3.4169307
```

According to the model the amount of strikeouts is more decisive rather than hits by the pitch or catches of stealing

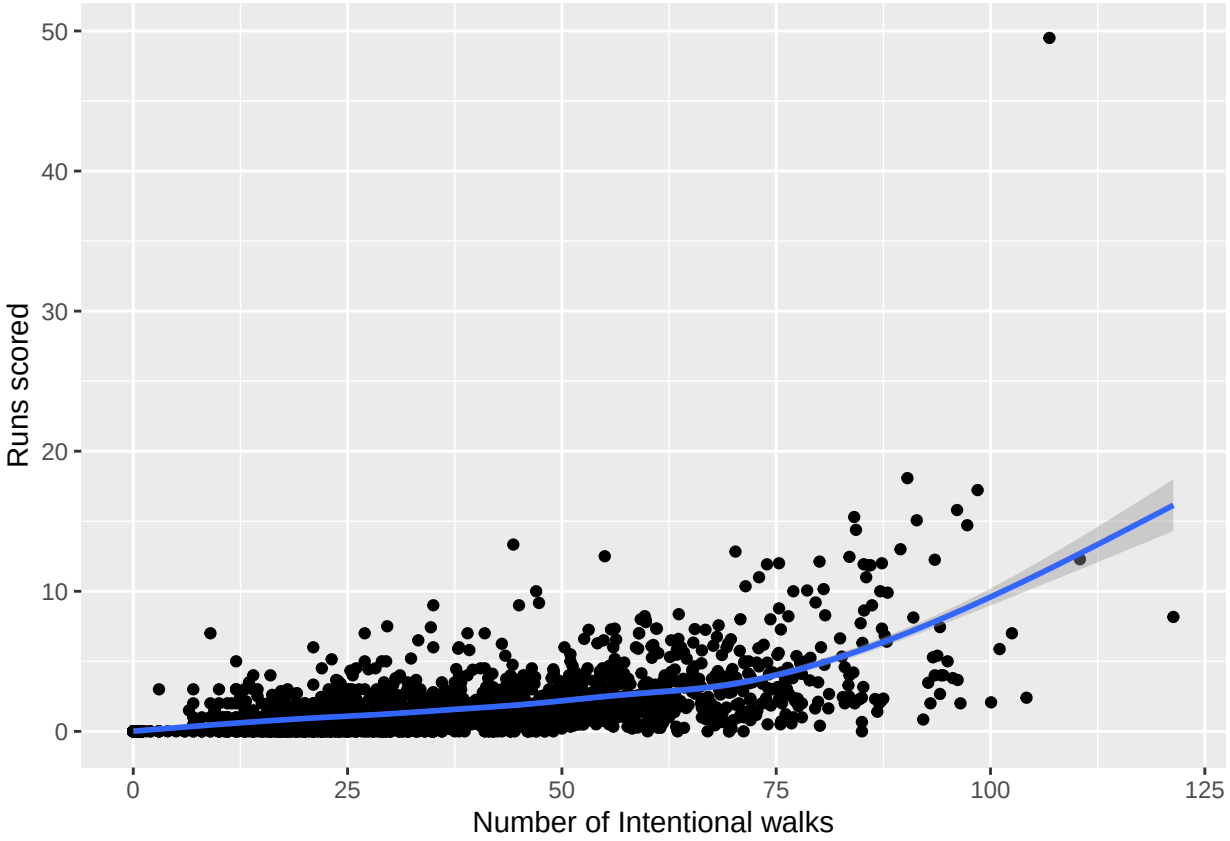
Now let's check if the players who in general are more likely to not swing do better or not

```
##  meanIBB
## 10.43064
```

```
##
## Call:
## lm(formula = meanR ~ 0 + meanIBB, data = walkers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -409.44    0.00   10.00   26.42   85.00
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## meanIBB  10.4306     0.1937   53.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.54 on 2482 degrees of freedom
## Multiple R-squared:  0.5388, Adjusted R-squared:  0.5386
## F-statistic:  2899 on 1 and 2482 DF,  p-value: < 2.2e-16
```

```
##
##  Pearson's product-moment correlation
##
## data:  walkers$meanR and walkers$meanIBB
## t = 39.4, df = 2481, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5955847 0.6440072
## sample estimates:
##       cor
## 0.6203868
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




So, to conclude in general the relationship between the runs scored by players getting intentional walks is higher than those who almost always try to hit

## Pitcher's efficiency

Let's first identify the ratio of the number of Batters hit by the pitch and the number of Batters faced

```
##             realName    ratioHF
## 1            Alberto 0.05785124
## 2    Anthony Michael 0.04201681
## 3      Angel Ladimir 0.04081633
## 4      William Henry 0.04060914
## 5       Brian Wesley 0.04040404
## 6       Michael John 0.03676471
## 7   Spencer Burdette 0.03669725
## 8         Edwin Jose 0.03448276
## 9       Steven Wayne 0.03349282
## 10  Francis Euclides 0.03212851
## 11 Christopher Deane 0.02985075
## 12             Joely 0.02985075
## 13              Juan 0.02976190
## 14     Anthony Aaron 0.02890173
## 15       Jake Austin 0.02884615
```

As it is a very rare event the ratios are small

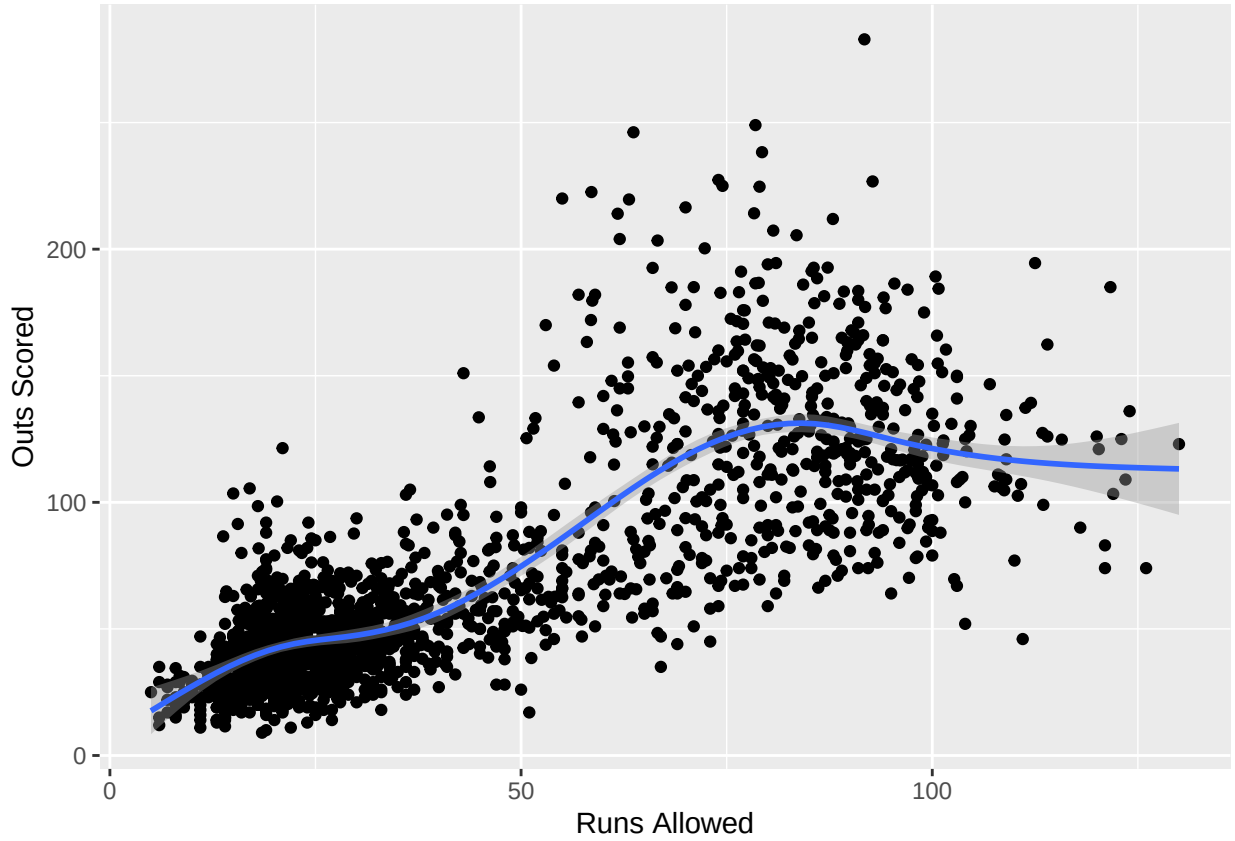
Now let's take a look at more obvious parameters

```
##           realName meanRA meanSO meanBAO
## 1           Samuel      5     25   0.219
## 2       Ryan David      6     29   0.171
## 3      Kevin Allen      6     15   0.183
## 4        Dustin A.      6     12   0.300
## 5 Eduardo Nazareth      6     35   0.144
```

I belive that the players with high ammount of strikeouts will allow less runs
so let's check my statement

```
##
##  Pearson's product-moment correlation
##
## data:  pitchers_r$meanRA and pitchers_r$meanSO
## t = 53.984, df = 1804, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7676126 0.8029306
## sample estimates:
##       cor
## 0.7859117
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




The correlation is very high, but this was obvious
Now let's find out the average number of outs made by the player's till the opponent can make a run
Before that let's also look at the relationship between the opponent's batting percentage and the number of strikeouts made by the player

```
##
##  Pearson's product-moment correlation
##
## data:  pitchers_r$meanBAO and pitchers_r$meanSO
## t = -3.9695, df = 1804, p-value = 7.485e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13858359 -0.04712951
## sample estimates:
##         cor
## -0.09305282
```

There is a very small but negative correlation among the variables meaning that increase in one will lead to decrease in the other

```
##            realName  ratioRO
## 1     Craig Michael 6.900000
## 2  Albertin Aroldis 6.282258
## 3       Josh Ronald 6.205882
## 4   Kenley Geronimo 5.878571
## 5  Eduardo Nazareth 5.833333
```

```
## 6            Dellin 5.780952
## 7        Chad Keith 5.472222
## 8            Samuel 5.000000
## 9      Chad Douglas 5.000000
## 10    Edwin Orlando 4.934426
## 11    Carson Donald 4.842105
## 12       Ryan David 4.833333
## 13  Richard Agustin 4.631579
## 14          Takashi 4.576471
## 15   William Edward 4.522613
```

As sooner or later, the opposing team will score a run, I believe that the the player who manages to get more outs during that period is more efficient in deffense
As a game of baseball is divided into 9 innings = 27 outs
The best player manages to get 7 outs for his tim before the opposing team will score
meaning that on average if he is playing his team will allow the first run no earlier than the 3rd inning

## Conclusion

To cap it all, I know that there are more detailed and obvious approaches of baseball analysis as it is very famous sport with a lot of available data. However I wanted to create my own metrics which are not that obvious in usage and maybe are not that useful, but at least they are my personal approaches :)
I have only a limited knowledge in baseball, and connecting my findings with my knowledge I can see some parallels
with soccer, as in both sports the attacking players are more recognized and usually they get the awards for the best player
Also my findings show that the players are more likely to score a run via intentional walks whihc usually is realized through not swinging and having patience.