Final Project Step 3

Milan Sherman

3/4/2022

Introduction

Using data to increase a sports team's odds of winning has become standard in professional sports over the last 20 years. In the early 2000s, the Oakland A's were early adopters, or even inventors of this approach to assembling a competitive baseball team as chronicled in the book (and movie) Moneyball (2011). While the strategy is not new, I am interested in identifying baseball metrics that are predictive of winning. Given the amount of money that is generated by professional sports, the ability to gain a competitive edge via data is crucial. In the current context, the ubiquity of leveraging data in sports has perhaps changes this approach from one of opportunity to a necessity in order to not lose ground on competitors.

Research Questions

My research question evolved and came into focus as I became more familiar with my datasets. I found that the datasets that I'd chosen were best suited to generating offensive metrics, and that offensive metrics are generally related more to runs scored than winning or losing in the research that I'd done. I did not want to lose sight of winning or losing, but decided to first relate my metrics to runs scored as it makes sense that these metrics would be much more predictive of runs scored. In the end, I focused in the following research question:

- What batting metrics are most highly correlated with runs scored? Which metrics are most predictive of winning or losing?

Approach

At the beginning of my project, I had identified batting average, on-base percentage, and slugging percentage as important offensive metrics to explore, but during my research I stumbled upon a lesser known metric with a lot of promise: base-out percentage. The inclusion of this metric made the analysis particularly interesting, as in theory it seemed to capture all of the important information needed to predict runs. The opportunity to determine the predictive power of relatively unknown metric was exciting. Each of these metrics is defined as follows:

- Batting Average: the proportion of at-bats that result in a hit. Walks are not considered an at-bat, and therefore are removed from consideration in this metric.
- Slugging Percentage: a weighted batting average, weighting each hit according to the number of bases it nets.
- On-base Percentage: the proportion of plate appearances that resulted in the hitter getting on base
- Base-out Percentage: the ratio of bases a player nets by any means (hits, walks, stolen bases, etc.) to the number of outs they generate by any means (ground out, fly out, caught stealing, etc.)

My approach was to generate these metrics by team by game for the two datasets that are not aggregated, and use them to predict runs scored and/or winning or losing. These datasets needed to be cleaned and transformed in order to generate these metrics at that level. This was a time consuming step, as it required an in-depth understanding of what data was contained in the dataset in order to transform it into the above metrics. The third dataset was already aggregated by team by season. That step done, the analysis was focused on the following steps:

1. Understanding the relationship between each of these metrics and runs scored within each of the datasets via scatterplots and Pearson's correlation coefficient
2. Creating a simple linear regression model using runs scored and each of these hitting metrics
3. Creating a simple linear regression model using number of wins in a season and each of these hitting metrics. This analysis was only possible with the third dataset that was aggregated by team by season, as it contained the relevant hitting metrics as well as the number of wins for the season.

To be clear, my question is focused on comparing the predictive power of these metrics rather than finding the most accurate model. Putting all of these metrics into a single model would likely result in a very accurate model, but would not lead to an actionable insight in the sense that we would not know what to prioritize. If we want to assemble a team with the best chance of winning we would like to know if we should value batting average, slugging percentage, on-base percentage, or base-out percentage. For this reason that the analysis considers each of these potential predictors of runs scored and winning/losing separately.

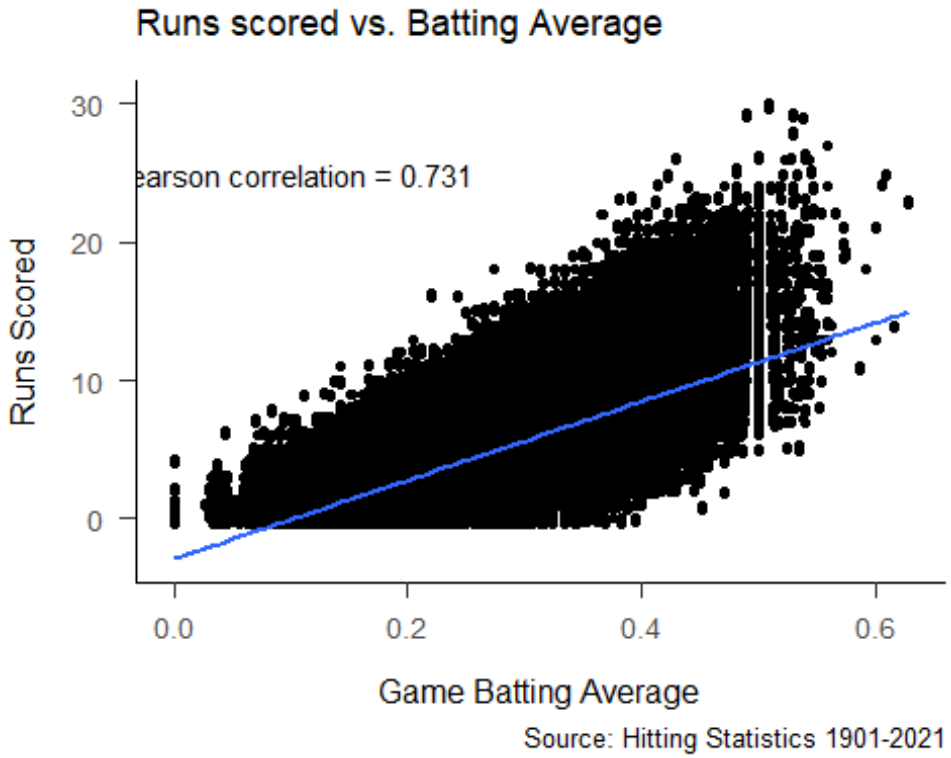
Analysis

The analysis is organized by dataset, and finding are synthesized in the implications section. In this section I analyze the relationship between the number of runs scored and various hitting metrics, including batting average, slugging percentage, on-base percentage, and base-out percentage. For each metric, I generate a scatterplot, compute the Pearson Correlation Coefficient, and generate a linear model. For the third dataset, I generate a linear model using each metric to see how predictive of number of wins in a season each

metric is. Before integrating across datasets, I will give a brief summary of the analysis results for each dataset.

Hitting Statistics 1901-2021

Batting Average

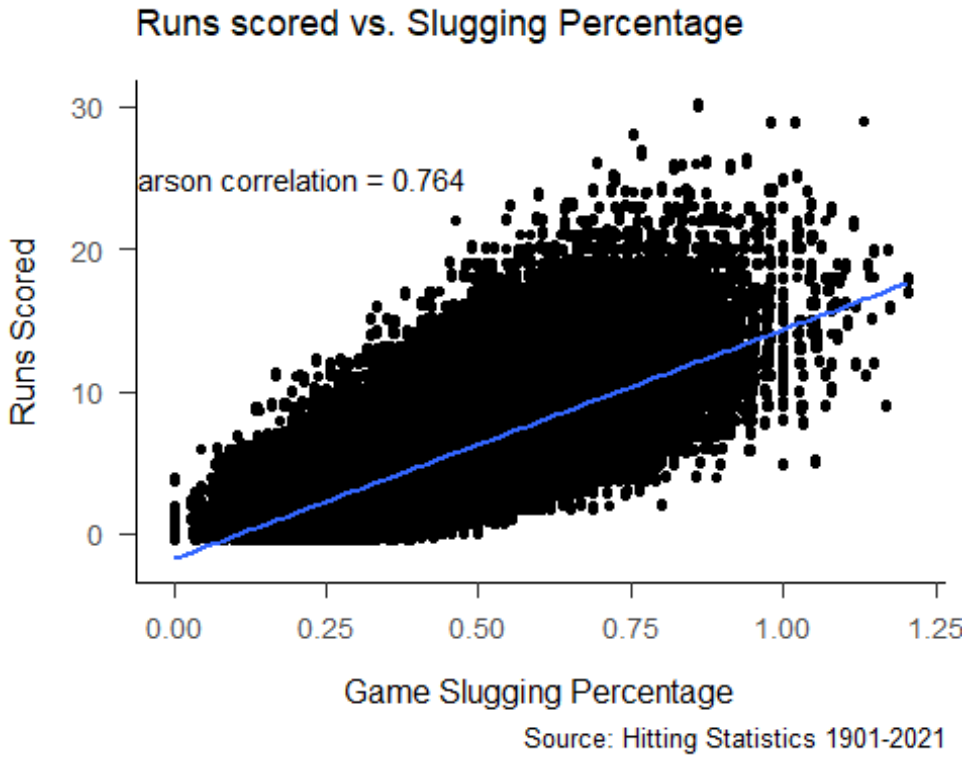


(#tab:unnamed-chunk-4)

Linear Regression Results for Runs Scored as a Function of Batting Average.

Predictor	*b*	95% CI	*t*(403562)	*p*
Intercept	-2.91	[−2.93, −2.89]	-259.62	< .001
BA	28.46	[28.38, 28.54]	681.13	< .001

Slugging Percentage

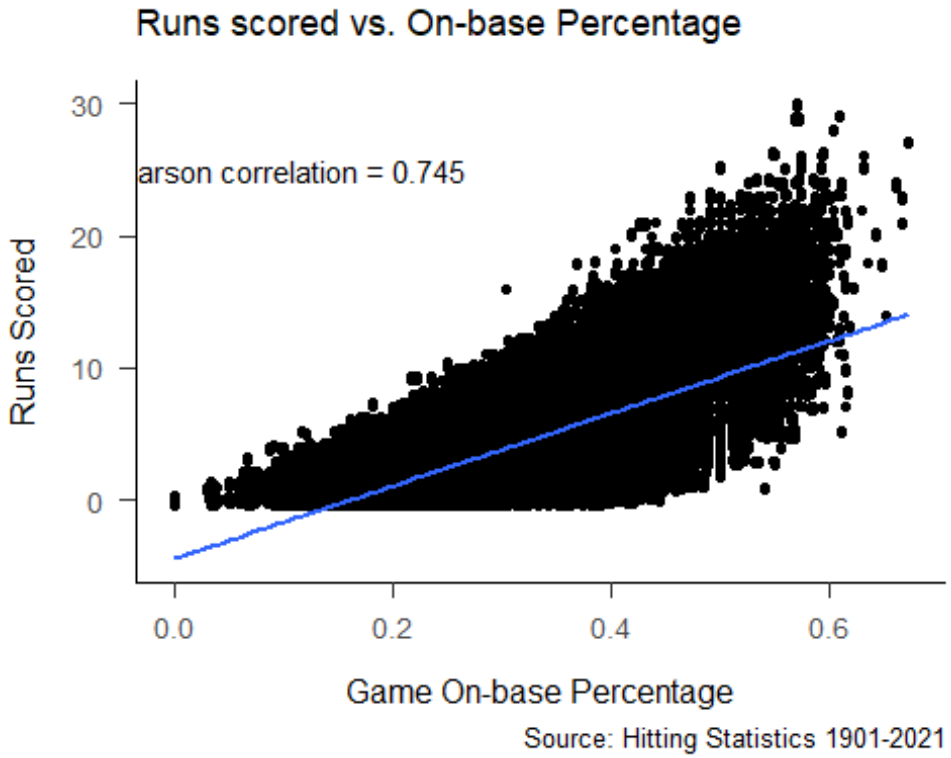


(#tab:unnamed-chunk-6)

Linear Regression Results for Runs Scored as a Function of Slugging Percentage.

Predictor	b	95% CI	$t(403562)$	p
Intercept	-1.78	[−1.80, −1.77]	-203.43	< .001
Slug	16.13	[16.09, 16.17]	752.58	< .001

On-base Percentage

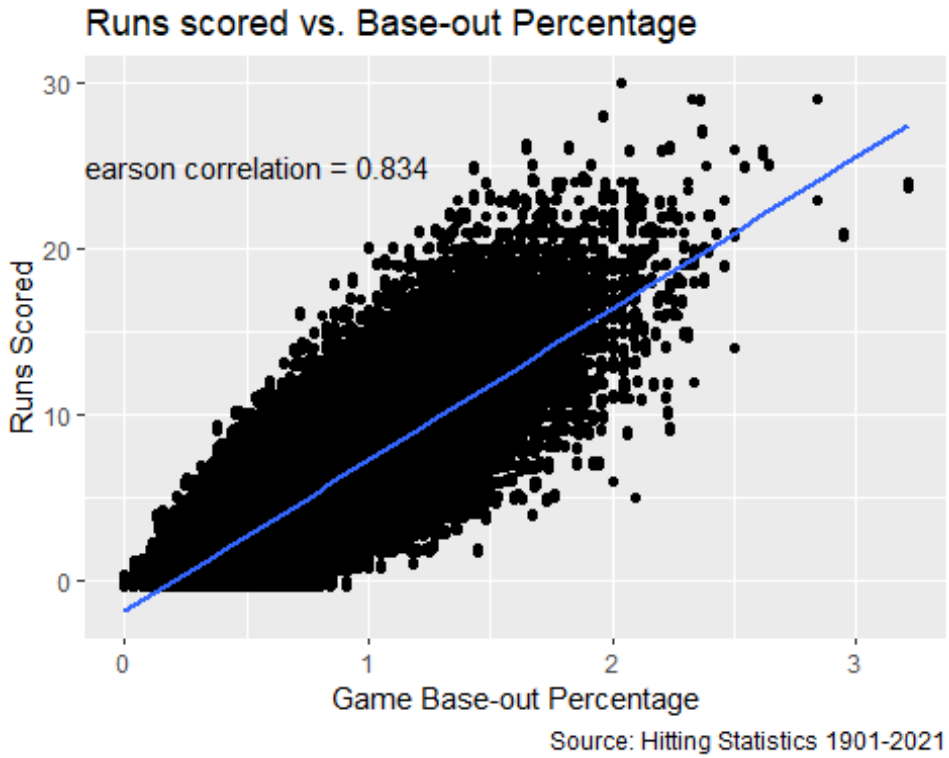


(#tab:unnamed-chunk-8)

Linear Regression Results for Runs Scored as a Function of On-base Percentage.

Predictor	b	95% CI	$t(403562)$	p
Intercept	-4.46	[−4.49, −4.44]	-346.82	< .001
OBP	27.53	[27.46, 27.61]	709.97	< .001

Base-out Percentage



(#tab:unnamed-chunk-10)

Linear Regression Results for Runs Scored as a Function of On-base Percentage.

Predictor	*b*	95% CI	*t*(403562)	*p*
Intercept	-1.89	[−1.90, −1.87]	-267.61	< .001
BOP	9.15	[9.13, 9.17]	961.18	< .001

Summary of Analysis using Hitting Statistics Data

Metric	Correlation	R^2	Slope
Batting Average	0.731	0.53	28.46
Slugging Percentage	0.764	0.58	16.13
On-base Percentage	0.745	0.56	27.53
Base-out Percentage	0.834	0.70	9.15

For this dataset, base-out percentage seems to be most highly correlated with runs scored, and accounts for 70% of the variability in runs scored.

As far as the linear models are concerned, each of these metrics is a significant predictor of runs scored, with a p-value of 0 for each. The slopes given by the linear models need a bit of interpretation. At face value, for example, the slope for batting average indicates that for

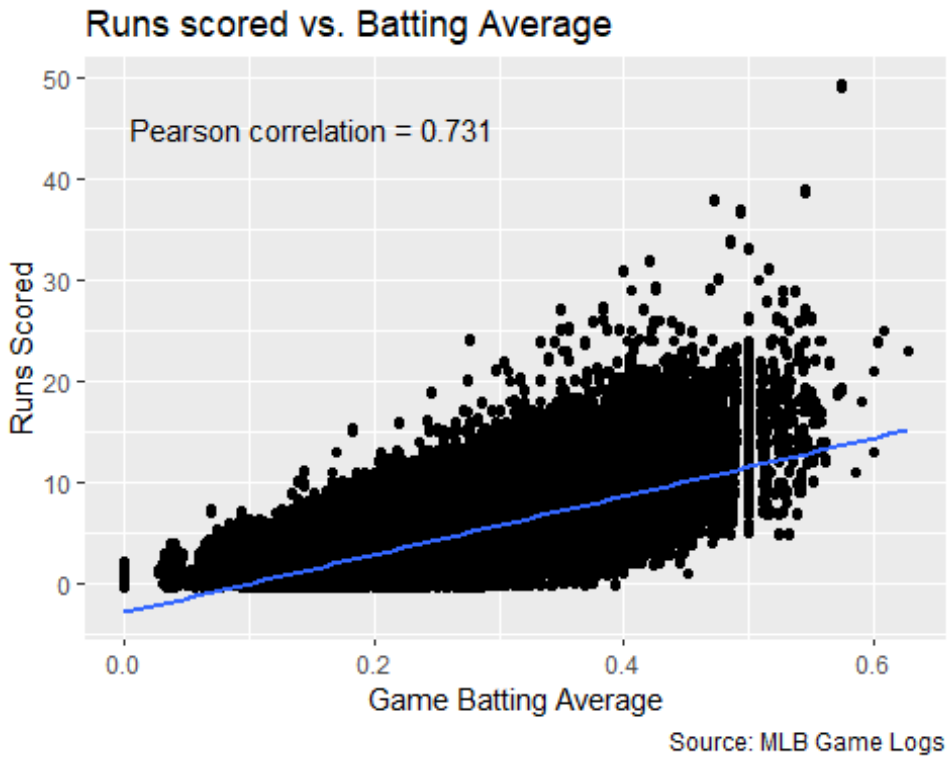
every increase in batting average of 1, the number of runs scored is expected to increase by over 28. In most baseball games, neither team scores more than 10 runs, and in many games it's less than 5. The issue is that in most baseball games a team's batting average is between 0 and .4, and cannot be greater than 1. Thus, it would make more sense to interpret the slope as an increase of 2.8 runs for every 0.1 increase in batting average. A similar interpretation can be applied to each of the metrics.

Furthermore, the difference between these slope also needs interpretation. These differences are relative to the range of values for the metric. For example, batting average ranges from 0 to just above 0.6, while base-out percentage ranges from 0 to over 3. Thus, an increase of 1 in each of these metrics will impact the number of runs scored differently. It will be more interesting to compare these slopes to what is generated by the linear models for the same metrics in the game logs data.

As the intercept for all four models is negative, which is not possible, it does not make sense to try to interpret it. It is difficult by not impossible to score a run without getting hit, and thus we would expect the intercept to be just above 0 for batting average and slugging percentage. It is not possible to score a run without getting on base, and thus on-base percentage and base-out percentage should theoretically have an intercept of 0. Nonetheless, the models for each of these metrics include a small, negative intercept.

MLB Game Logs

Batting Average

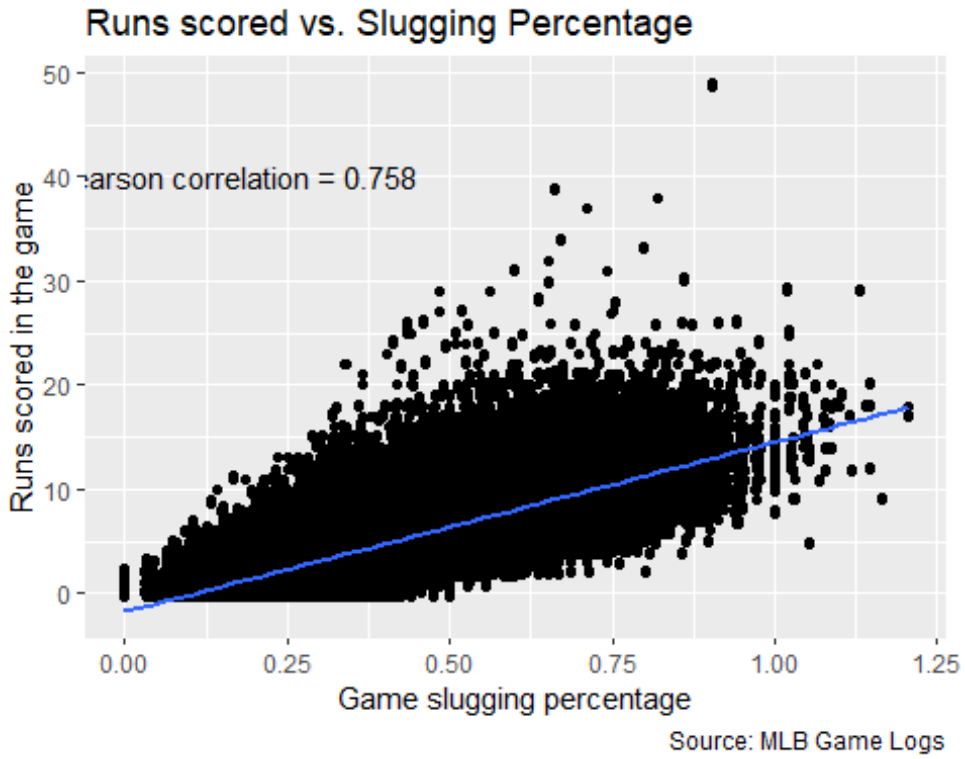


(#tab:unnamed-chunk-13)

Linear Regression Results for Runs Scored as a Function of Batting Average.

Predictor	b	95% CI	$t(281628)$	p
Intercept	-2.97	[−2.99, −2.94]	-217.31	< .001
BA	28.79	[28.69, 28.89]	568.65	< .001

Slugging Percentage

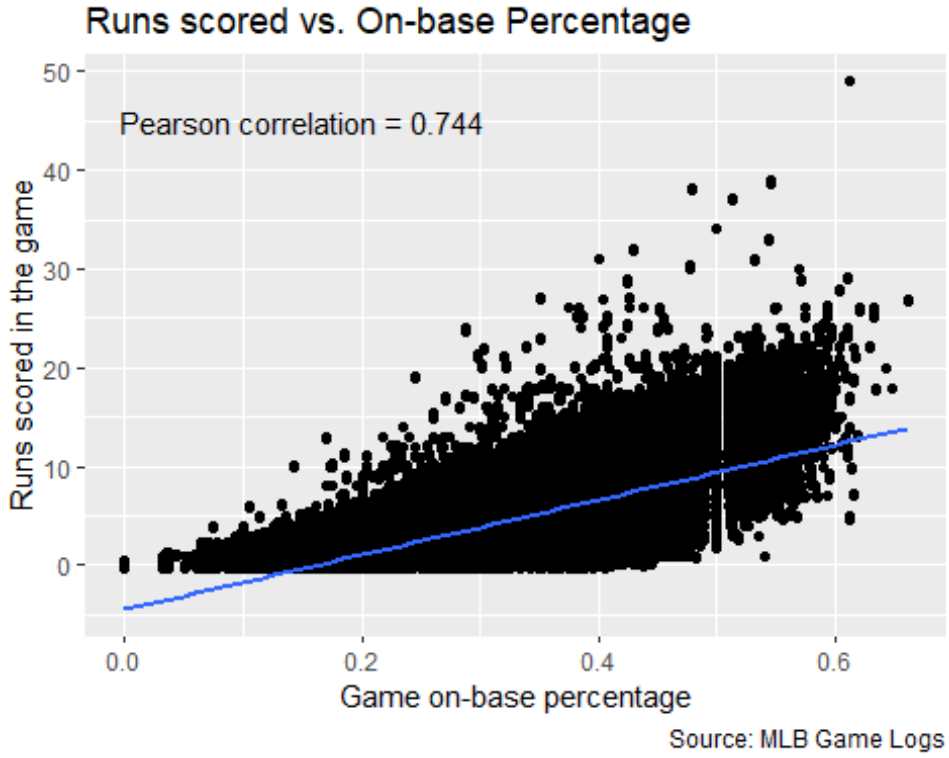


(#tab:unnamed-chunk-15)

Linear Regression Results for Runs Scored as a Function of Slugging Percentage.

Predictor	b	95% CI	$t(281490)$	p
Intercept	-1.79	[−1.81, −1.77]	-165.21	< .001
Slug	16.30	[16.25, 16.35]	616.02	< .001

On-base Percentage

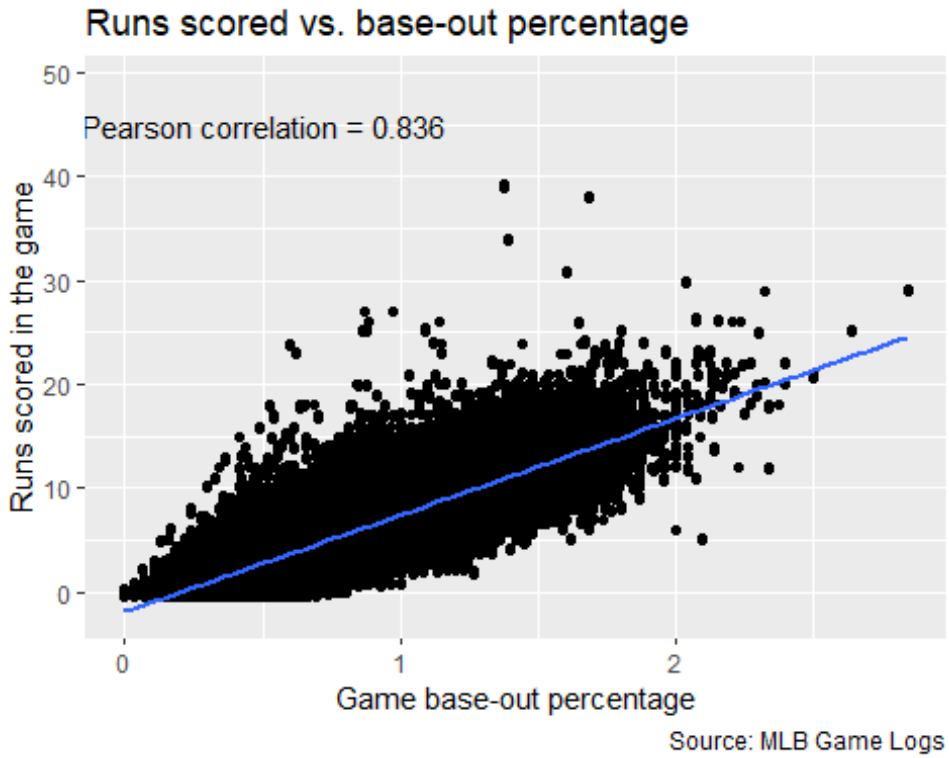


(#tab:unnamed-chunk-17)

Linear Regression Results for Runs Scored as a Function of On-base Percentage.

Predictor	b	95% CI	$t(271504)$	p
Intercept	-4.51	[−4.55, −4.48]	-282.46	< .001
OBP	27.83	[27.74, 27.92]	579.74	< .001

Base-out Percentage



(#tab:unnamed-chunk-19)

Linear Regression Results for Runs Scored as a Function of Base-out Percentage.

Predictor	*b*	95% CI	*t*(235033)	*p*
Intercept	-1.91	[−1.92, −1.89]	-204.62	< .001
BOP	9.32	[9.29, 9.34]	738.60	< .001

Summary of Analysis using Game Log Data

Metric	Correlation	R^2	Slope
Batting Average	0.731	0.53	28.79
Slugging Percentage	0.758	0.58	16.3
On-base Percentage	0.745	0.56	27.83
Base-out Percentage	0.836	0.70	9.32

Once agian, base-out percentage seems to be most highly correlated with runs scored, and accounts for 70% of the variability in runs scored.

Also, the p-value of each these metrics is 0, indicating that each is a significant predictor of runs scored. The interpretation of the slope and intercept is the same as described for the Hitting Statistics data above. The interesting thing to note is how closely these numbers are

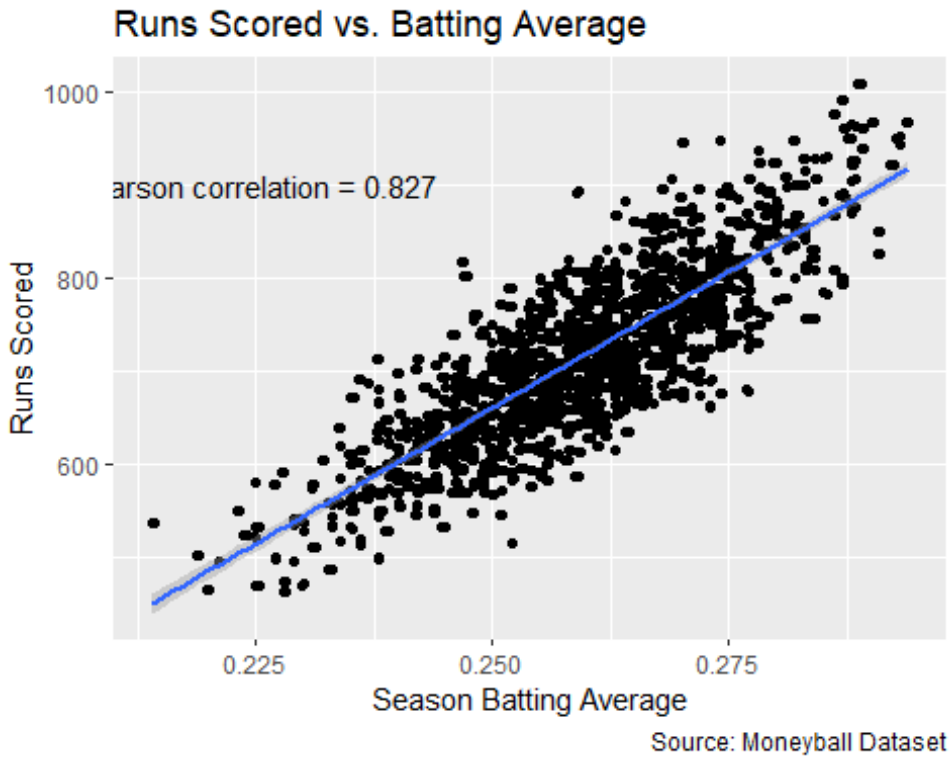
to those generated from the Hitting Statistics data. This increases our confidence in these data and the analysis, i.e., we were able to replicate the results with another dataset.

Moneyball

This dataset is different than the previous datasets in that it is already aggregated by team by season. Thus, the analysis is slightly different in the following ways:

1. It is not possible to recover the number of outs and bases from this aggregated data, so I cannot compute base-out percentage. Since batting average, slugging percentage, and on-base percentage are in this dataset, I will use those metrics for the analysis of this dataset.
2. This data is aggregated at the season level, and contains the number of runs scored over the course of the season, as well as the number of wins. So the analysis will focus on the relationship between runs scored and these metrics, as the previous analyses did, but at the season instead of game level. Also, in order to answer my question about the relationship between these metrics and winning/losing, I will create simple linear regression models to examine the relationship between wins and each batting metric for the season.

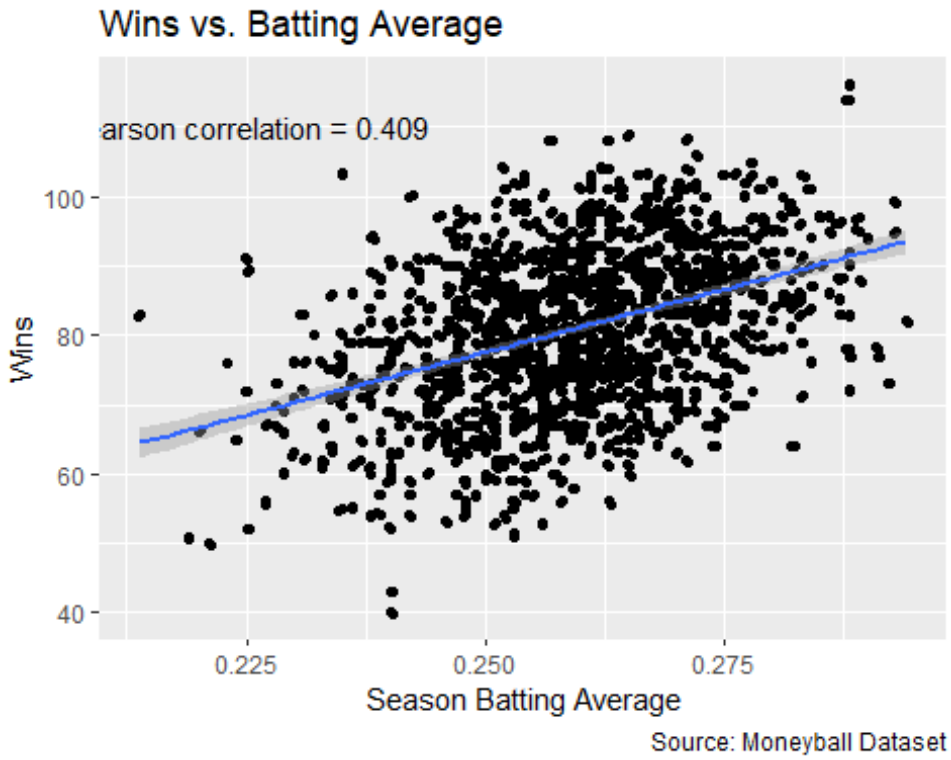
Batting Average



(#tab:unnamed-chunk-21)

Linear Regression Results for Runs Scored as a Function of Batting Average.

Predictor	b	95% CI	$t(1230)$	p
Intercept	-805.51	[−863.41, −747.61]	-27.30	< .001
BA	5,864.84	[5,641.81, 6,087.87]	51.59	< .001

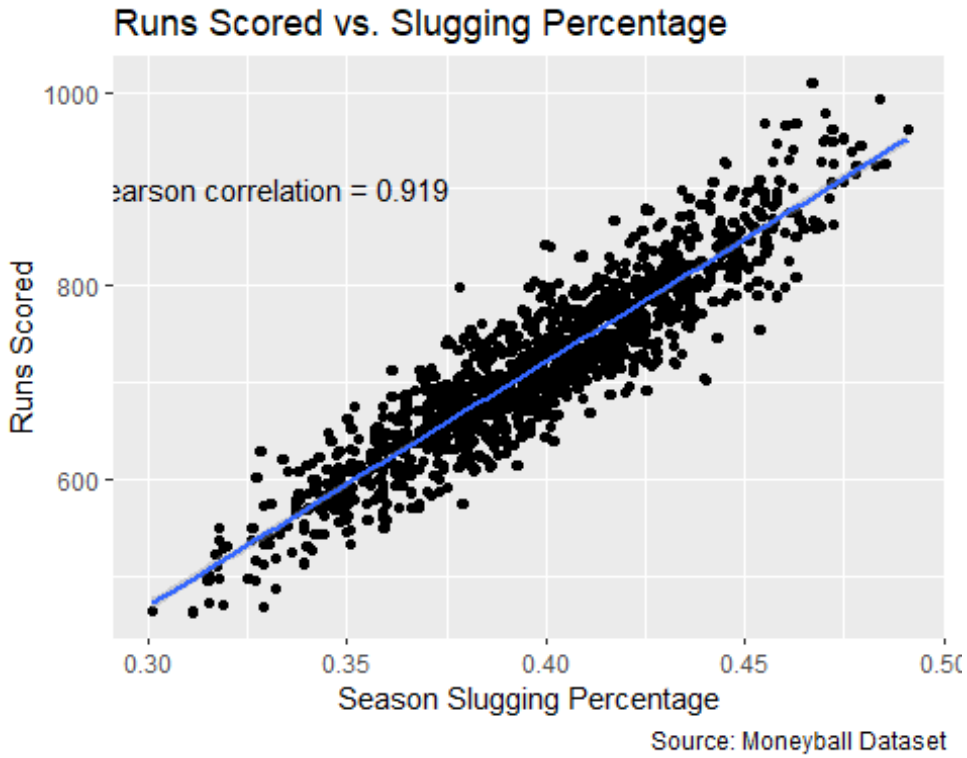


(#tab:unnamed-chunk-23)

Linear Regression Results for Wins as a Function of Batting Average.

Predictor	b	95% CI	$t(1230)$	p
Intercept	-13.17	[−24.93, −1.40]	-2.20	.028
BA	362.83	[317.51, 408.15]	15.71	< .001

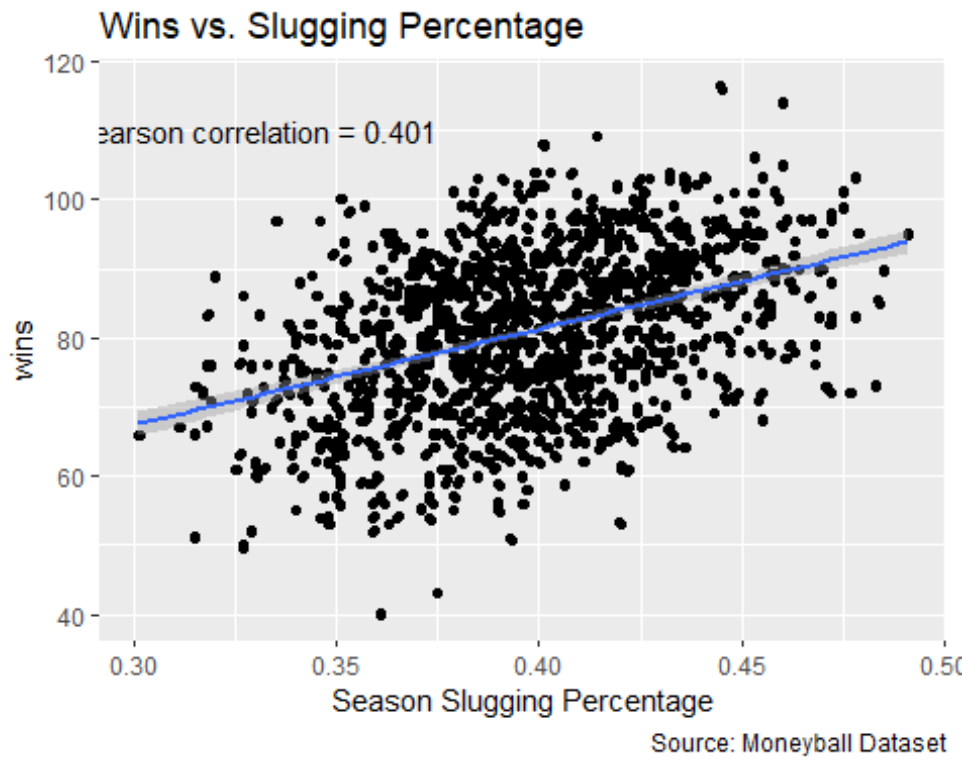
Slugging Percentage



(#tab:unnamed-chunk-25)

Linear Regression Results for Runs Scored as a Function of Slugging Percentage.

Predictor	*b*	95% CI	$t(1230)$	*p*
Intercept	-289.37	[−313.60, −265.13]	-23.43	< .001
SLG	2,527.92	[2,467.15, 2,588.70]	81.60	< .001

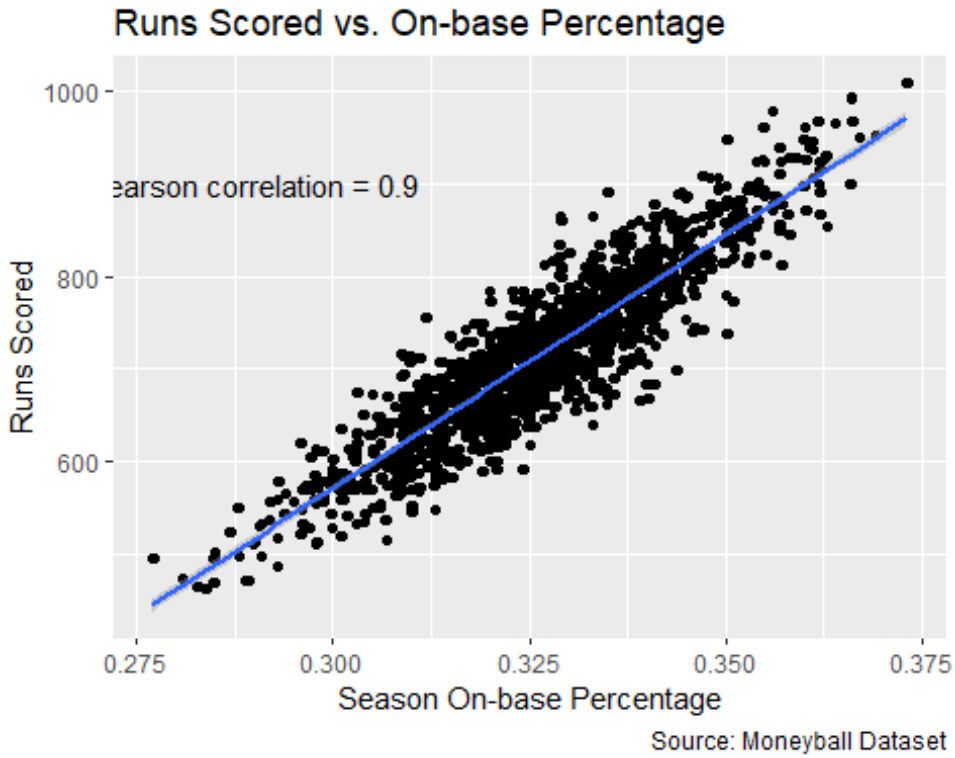


(#tab:unnamed-chunk-27)

Linear Regression Results for Wins as a Function of Slugging Percentage.

Predictor	*b*	95% CI	*t*(1230)	*p*
Intercept	25.96	[18.92, 32.99]	7.24	< .001
SLG	138.29	[120.64, 155.93]	15.37	< .001

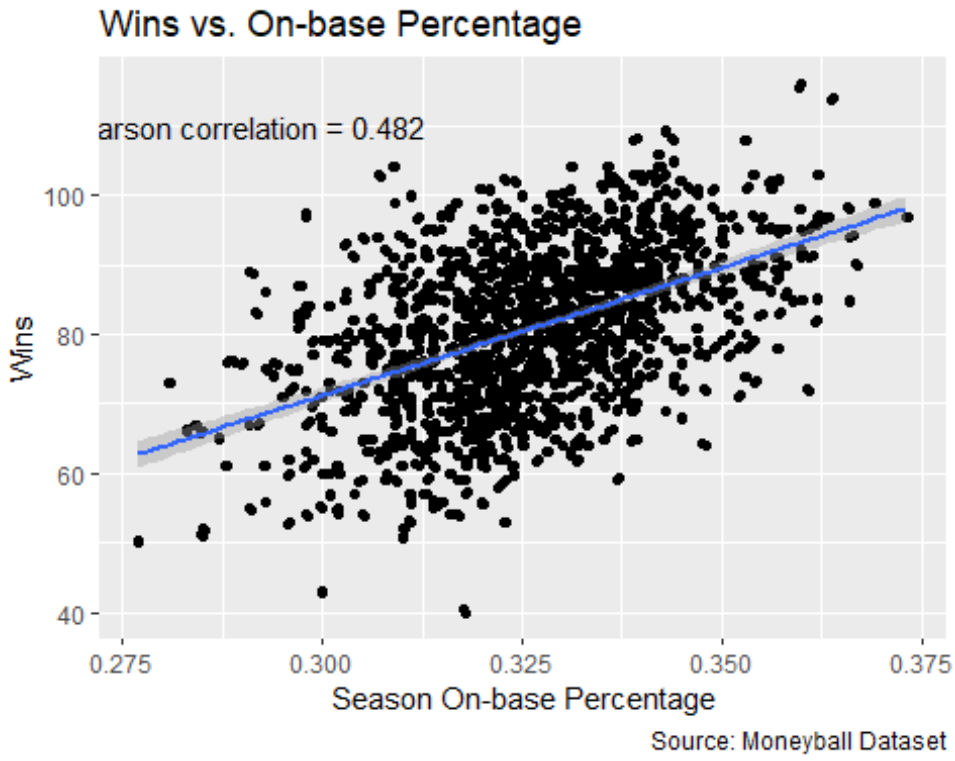
On-base Percentage



(#tab:unnamed-chunk-29)

Linear Regression Results for Runs Scored as a Function of On-base Percentage.

Predictor	*b*	95% CI	*t*(1230)	*p*
Intercept	-1,076.60	[−1,125.06, −1,028.15]	-43.59	< .001
OBP	5,490.39	[5,342.06, 5,638.71]	72.62	< .001



(#tab:unnamed-chunk-31)

Linear Regression Results for Wins as a Function of On-base Percentage.

Predictor	*b*	95% CI	*t*(1230)	*p*
Intercept	-39.10	[−51.33, −26.88]	-6.28	< .001
OBP	367.75	[330.34, 405.16]	19.28	< .001

Summary of Analysis of MoneyBall Data

Relationship between Metrics and Runs Scored for the Season

Metric	Correlation	R^2	Slope
Batting Average	0.827	0.68	5864.84
Slugging Percentage	0.919	0.84	2527.92
On-base Percentage	0.9	0.81	5490.39

Relationship between Metrics and Wins for the Season

Metric	Correlation	R^2	Slope
Batting Average	0.409	0.17	362.83
Slugging Percentage	0.401	0.16	138.29
On-base Percentage	0.482	0.23	367.75

As these data are aggregated by team by season, the interpretation of the analysis will differ. The first thing to note is that the distribution of each metric is much tighter than the data aggregated at the game level due to the sample size. At the game level, we expect to see much more variation due to the smaller number of at bats that each team gets in a game versus the entire season. The tighter distributions means less variability, and the correlations for these metrics at the season level are consistent with but higher than at the game level, as are the corresponding values of the coefficient of determination.

In each of the linear models, the parameters are significant, and the intercepts are negative and thus do not have a meaningful interpretation. The slopes are much larger due to the target variable being at the season versus game level. Like our analysis of the previous two datasets at the game level, however, we see a similar relationship between runs scored and each of the metrics, with slugging percentage more highly correlated than on-base percentage, followed by batting average. As these data were already aggregated and did not contain base-out percentage, that analysis was not possible with this data. However, given the order of the correlations for the other three metrics was the same at the season level as at the game level, we can hypothesize that base-out percentage would have been more highly correlated to runs scored than the other metrics at the season level as well.

Looking at the relationship between these metrics and the number of wins, there are two interesting observations:

1. We see much more variability in the number of wins as they relate to each of the metrics, and thus the correlations are much lower, indicating that much more goes into winning a game than runs scored.

2. We see a reversal in terms of the correlations between slugging percentage and on-base percentage, i.e. on-base percentage is more highly correlated with the number of wins, followed by slugging percentage and finally batting average. This was the insight that spurred the Oakland A's to value on-base percentage over slugging percentage when using this data.

Additional analysis that could be done with this data includes creating a logistic model for each of the game level datasets to determine which metrics best predict winning or losing. This would allow us to test the hypothesis formed using the Moneyball data that base-out percentage may be more predictive of runs scored and winning or losing than any of the other traditional hitting metrics.

Implications

In the spirit of the Moneyball story, I will frame the implications of this analysis for a general manager of a baseball team. In the Moneyball story, the general manager was looking for a competitive edge using data, as he had a very limited budget. The value of power hitting and a metric like slugging percentage for helping teams win was well-known, and power hitters were highly valued and expensive. The GM for the Oakland A's did not have the option of signing players with an outstanding batting average or slugging percentage as he could not afford these players. The insight that the data provided was that on-base percentage was better at predicting runs scored than batting average and nearly as good as slugging percentage. Furthermore, it was a better predictor of winning than either batting average or slugging percentage. The reason this insight was so key was that on-base percentage was not highly valued among other baseball teams at this time. A walk was undervalued, and so the GM of the Oakland A's identified players who had a high on-base percentage, but were undervalued in the market based on conventional metrics and wisdom.

For some reason, base-out percentage is not a common metric in baseball at this time, in spite of its introduction over 40 years ago. The main implication of this analysis is that it is a metric that is more highly correlated with runs scored than any of the other metrics included in this analysis. As it is relatively unknown and/or unused, a logical next step would be to calculate the base-out percentage of current players in the Major Leagues, and determine whether they are undervalued. It could be that players with a high base-out percentage are already highly valued based on other metrics, and thus this insight is not actionable in the same way that on-base percentage was in the Moneyball story. However, at the very least it seems that this analysis should be done if it has not been already.

Limitations

There are other offensive metric that could be included in this analysis that I was not aware of until after I'd completed my data transformation, such as OPS, which is on-base percentage plus slugging percentage. Indeed, there are likely many more metrics that I'm still unaware of that could be further explored. The field of analytics is likely more mature in baseball than any other sport, and the present analysis merely scratched the surface of what is possible.

Furthermore, runs scored is only half of winning. No matter how many runs a team scores, they will still lose if their opponent scores more runs. Thus, it is likely insufficient to try to use these hitting metrics to predict winning or losing. That is, a team's batting average in a

game may not be predictive of winning or losing, but rather the team's batting average relative their opponent's batting average in the same game. Creating metrics for the differences between team hitting metrics for a game would likely be more predictive of winning or losing, as it takes into account how well the other team batted as well. Another way to get at this would be to perform a hypothesis test on the difference of means of each of these metrics between winning and losing teams.

However, while this would be an interesting analysis, it is likely less actionable as there is not a single controllable aspect of the game that influences how both teams hit in a game. A separate analysis of defensive metrics, especially pitching, is likely needed in order to find actionable insights that would influence winning or losing. Pitchers are the most valued players in the game due to their ability to influence the number of runs scored by an opponent. Identifying a metric like base-out percentage for pitchers could lead to more actionable insights.

A more minor issue is that there should be either 24 or 27 outs per game. In general, a team gets three outs per inning for nine innings. However, when the home team is winning in the bottom of the ninth inning, the game is over since the visitiing team has no more at bats and thus the outcome of the game will not be changed by the home team's at bats. The issue, is that when calculating outs in my data, I found a number of instances where the number of outs was not 24 or 27, but 25, 26, or 28. I double checked my computation for outs, and it was corrrect. I also checked the distribution of outs and found that 24 and 27 were most common. Nonetheless, this is certainly an issue I would need to look into further as it is part of the computation for base-out percentage.

Concluding Remarks

Baseball has long been called a game of numbers due to many statistics that are tracked and computed for every aspect of the game. While base-out percentage is the most predictive metric of runs scored, given the current state of baseball analytics, there must be some reason that it is relatively unknown and/or used. Indeed, opportunities to exploit an undervalued metric like on-base percentage are likely impossible in baseball today. Rather, understanding how to value players based on these various metrics is probably a focus of current baseball analytics, although this is just a hypothesis. Understanding the relationship between winning and earnings, and the relationship between a variety of metrics and winning would form a basis, in conjunction with market value for similar players, for contract negotiations. I recently heard of a European soccer player who hired a team of data scientists to evaluate his financial impact to his team and used it as basis for negotiating a new contract. It may be that soccer is behind baseball in terms of its analytics maturity, as I would think that every general manager in baseball has already done this analysis for each of his players, if not for every player in the league.