Tyler Moore
Joshua Ramos

Explanation of Statistics

For our final project, we wanted to evaluate teams based on their Runs, OBP, and slugging. To normalize the data we wanted to put the teams on a level playing field by making it seem like they all play on the same field by normalizing the data using park factor. Many teams can produce high stats solely based on the park they play at like the Rockies but some of these teams never tend to make it to far. We found a way to rank these teams based on various stats about the team's offense.

To start off we had to create a data set for park factors. There was probably an easier way to do it but we ended up making a csv and manually entering the park factors for all the teams for the past 3 years. We used ESPN's park factor dataset just to make sure we were using correct numbers. We had to double check the numbers and used the team IDs that matched the Lahman dataset to make things a little bit easier. Once we had the park factors, we started mining in a Jupyter Notebook.

Using the Lahman Dataset and the team's data we calculated the teams OBP and SLG at home and away. For runs, we calculated the average runs scored at home and away. We normalized the data so the teams can be ranked on a more level playing field. For example, if a team has a park factor of 0.85 and averaged 3 runs a game, we would multiply 1.15 x 3 or (((1 - 0.85) + 1) * 3) so that it normalizes their run scoring, and produced the dataset below. However, we did realize when teams are on the road, the ballparks will have different park factors. So we used the Retrosheet Dataset and calculated the adjusted runs (reverse park factor * runs that game) for that specific game. Then we found the average adjusted runs per game for each team in order to take the park factor into account.

	home_team	park_factor_2014	home_score	home_score_adjusted	visiting_score	visiting_score_adjusted	score_adjusted
4	BOS	1.072	5.888889	5.464889	4.950617	5.059974	5.262432
23	SEA	0.825	4.567901	5.367284	4.913580	5.060519	5.213902
6	CHN	0.931	4.802469	5.133840	5.172840	5.183228	5.158534
8	CLE	0.950	5.580247	5.859259	4.062500	4.001643	4.930451

The home score and visiting score adjusted is taking into account park factor and either increases their runs or decreases depending on where they play. Lastly, we combined both the visiting and home adjusted scores and calculated the mean to get the final score adjusted score.

The OBP adjusted factor was done very similar to scores and we took the reverse park score and then multiplied it by the OBP at that specific park and took the mean of both visited and away. In the end, the dataset looked pretty similar as shown below with the SLG data produced.

	home_team	park_factor_2014	home_OBP	home_OBP_adjusted	visiting_OBP	visiting_OBP_adjusted	OBP_adjusted
23	SEA	0.825	0.323685	0.380330	0.313761	0.319584	0.349957
6	CHN	0.931	0.342158	0.365767	0.329485	0.327060	0.346413
4	BOS	1.072	0.357283	0.331559	0.324726	0.333744	0.332651

	home_team	park_factor_2014	home_SLG	home_SLG_adjusted	visiting_SLG	visiting_SLG_adjusted	SLG_adjusted
23	SEA	0.825	0.422940	0.496955	0.418663	0.426312	0.461634
3	BAL	0.932	0.445495	0.475789	0.422810	0.420849	0.448319
4	BOS	1.072	0.481045	0.446410	0.421248	0.431832	0.439121
18	NYN	0.847	0.401403	0.462818	0.411463	0.413127	0.437972

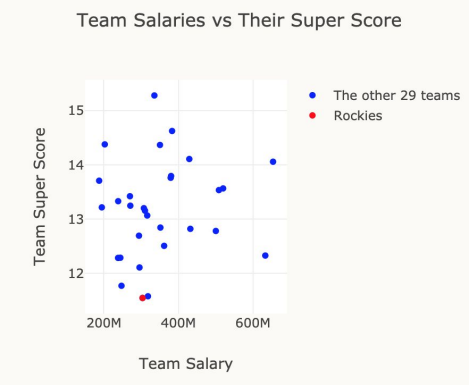
So as you can see, teams that had a high park factor (greater than 1), produced an OBP and SLG percentage that was lower than their normal numbers. Teams that had a low park factor (less than 1), produced data that was higher than their normal numbers. This allowed for all the teams to be compared on a level playing field in order to determine who was the best hitting team for that specific year. To do this, we created a super score accounting for all these adjusted numbers (Score, OBP, SLG).

Creating the calculation was a little difficult and we spent a long time trying to figure out how exactly we could correctly implement it. The final decision was to take the means of Score, SLG, and OPS. Then normalize SLG and OPS to the mean of Score. To create the factor to adjust the SLG and OPS was to divide the mean of the Score by the mean of SLG and OPS independently. Here's the basic equation,
(Team Adjusted Runs)+(Score Mean/OBP Mean)*(OBP Adjusted)+(Score Mean/SLG Mean)*(SLG Adjusted)=Super Score

This was the meat of our project and was what we used to rank and determine which team had the best offense without the assistance of the parks they played in. After we got all the super scores, we added a graph to compare how much the team spent that year in comparison to their super score.

We compared how much the teams spent in 2016 compared to their super score and found there wasn't a strong correlation between money spent and their super score. There was a slight growth in the super score as salary increased but it was very noisy and there was no direct relationship. Baseball isn't like other sports; money spent does not correlate to a high winning percentage. This is what this graph shows; it takes a lot

more than money spent to determine a good hitting team which is why it is so hard for analysts to predict who will win the World Series the following year.



For the evaluation of our stats, we looked at what teams won the World Series that year and how they compared in the ranks of our super scores. The Cubs were at the top of our list in 2016 and they ended up winning the World Series. Which makes sense as they were pretty solid hitters. The Indians however, did not place well in our super score placing 6th from last in the league; they did have a good pitching staff which is what led them to the World Series. The 2015 World Series was between the Mets and Royals with both teams being in the top 50% of our super score and the Mets were 6th and ended up almost getting swept in that World Series. Lastly, let's look at 2014, which was the Giants vs the Royals going all the way to game 7. The Royals were ranked low being in the bottom 50% and the Giants were close to the middle. Our stat shows these teams weren't the most effective in the league hitting at neutral ballparks. However, most of the teams making it far in the playoffs during those 3 years played at hitters parks, so why wouldn't they use that to their advantage? Home field advantage is a real thing and these teams used it to help them win a World Series. Our new stat wanted to take away home field advantage in order to see who were the best hitting teams, regardless of where they finished that year.

In conclusion, our super score ended up not being a perfect measurement and we think we know the reasoning. Park factor is an important aspect of baseball and the smart franchises use this knowledge to build their teams around their parks strengths. Teams can use this stat in order to figure out the weaknesses of their teams and adjust to them. For example, in 2016 the Rockies were scoring over 6 runs a game at home, but when we took out the park factor effect, that number dropped down to 4. Anyone who looks at 6 runs a game thinks this is really good, but when you take out park factor, runs per game goes down significantly. The Rockies front office should realize their runs per game should be higher due to the park they play at, and should draft/trade players to account for that accordingly. There is so much that goes into the sport of baseball and

using this stat to figure out a teams strengths and weaknesses could be a great way to see how well rounded they really are.