Dan Nguyen

Sabermetrics Spring 2018

Final Project Writeup

May 7 2018

TwMPRG: Twitter Mentions Per Run

Intro

In sports today, the role of the media is the most significant it has ever been, as it influences and dictates how much attention baseball teams get. With the rise of social media, there is more digital content being generated by MLB teams than there ever has been. Attention on social media on platforms such as Twitter, Instagram, and Facebook are all very important from a revenue perspective, as it can be seen as free or cheap marketing. Social media allows people the display their thoughts in real time which can be used to quantify their engagement with a team in a given moment. I propose a new statistic that quantifies how much social media value a team's runs per game contributes to his organization. I call it the Twitter Mentions Per Run per Game (TwMPRG) value.

Background (SMEAR → TwMPRG)

My first attempt at this statistic was based on the he Wins Above Replacement (WAR) value is a statistic, where TWMPRG has its origins from, summarizing a player's contributions to their team in a single measurement. It answers the question: "If player x got injured, and was replaced by a replacement player, how much value would the team be losing?". I thought it would be a good idea if we could quantify each players' Social Media Engagement Above Replacement.

Similarly, the SMEAR value of a player answers the question: “If we didn’t have player x, how many social media engagements would the team be losing?”.

Social media is a complex ecosystem that can be quantified in a variety of different ways, including counts of follows, mentions, likes, and impressions. I needed to either choose one of these categories of engagement, or combine them into a single weighted statistic and use that value to rank players. Due to some difficulty garnering a sufficient database, like the Instagram (Facebook) API deprecation, I decided to user Twitter mentions. Specifically, *team* mentions. The reasoning behind this is because when I was looking for players who I thought would have high SMEAR values, i.e. the highest paid players on the Rockies, I found that the top three players (Arenado, Desmond, and Davis) all didn’t have Instagram accounts, so the statistic fell apart there due to lack of data. However, I still thought Social Media Engagement was interesting so I thought a stat that measured a teams mentions per run would still lead to cool insights. The following equation depicts an MLB team’s number of mentions per run, per game.

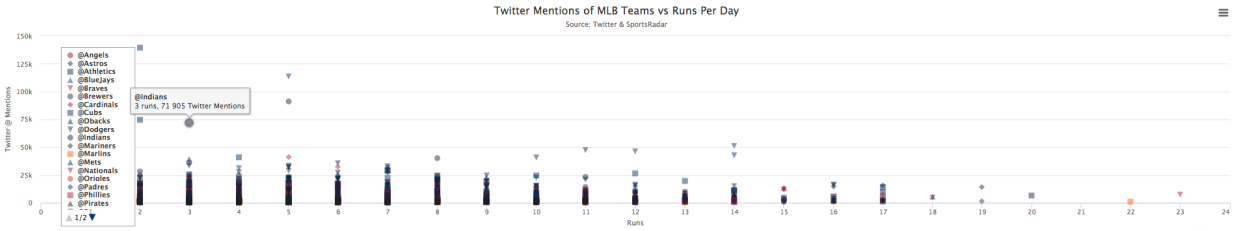
$$TwMPRG = \frac{Twitter\ Mentions\ @TeamHandle}{\frac{runs}{game}}$$

Demonstration

Once I have all this data aggregated, I created online charts where users can see team engagement over time. I will highlight trends and point out the differences in engagement when key social media influencing players are traded in the past 3 seasons. I also plan on showing a section of the best value for TWMPRG, which will show the highest valued players based on their contracts, and how much engagement they offer to their team.

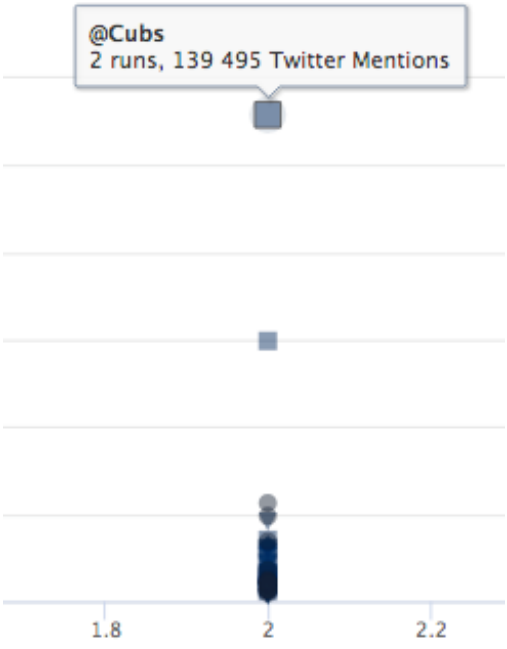
Data Aggregation

I used a combination of Node.js and Python to get information from Twitter's API. For the 2017 season I got the counts of all the mentions of every team in the MLB. I then combined this data with the Retrosheet from 2017. For every game occurrence, I created an object with the game's date, home runs scored, visitor runs scored, and mentions for that day. This way we can visualize a team's mentions per day.



Evaluation of Statistic and Findings

There aren't a lot of social media statistics related to baseball, however it was very interesting to see the lack of correlation. One thing I think may have skewed the data was the lack of samples of high scoring games. Games that exceed about 12 runs scored were rare, and if there were more games scored with higher amounts of runs, it could have affected the study.



As much as I would like to say that the data had an interesting trend, it really didn't. Naturally, I had a hunch that teams that scored more runs would generally have more mentions on that date, but that's not the case. The visualization above shows no trend in data, and almost looks like a negative regression line could be drawn. After doing more research I found that the outliers of high tweets for teams collecting more than 70,000 tweets were games of high magnitude, or games

that had heavy impact on the season. All teams that gained more than 50,000 tweets clinched first in their division. The Cubs, Indians, Dodgers, and Astros led the year with more than 600k, 200k, 1.69M, 1.9M mentions throughout the season. The Houston Astros, the World Series Champions finished the season with an average 2120 TwMPRG. The San Diego Padres were only mentioned 8013 times over a span of 604 runs, clocking a league low 13.26 TwMPRG.