Baseball Discussion on Reddit

Finding a correlation between MLB team performance and online discussion by fans.

Benjamin Owens & Jennifer Tsui

Fall 2016

CS 505/591 - Mark Crovella

bsowens.github.io/cs505-final-project/

Introduction

Often times sports have the side effect of bonding together the fans of each team. Perhaps it is the case that a very lonely person is looking to find their niche in the form of a baseball team to be a fan of. Our data would help this person find the most loyal fanbases, so they can make some new friends and bond over their newfound passion of baseball! On the other hand, locating a very loyal fanbase may help with classifying the general sentiment of the fanbase, and why they may be seen as negative or positive. This may sound farfetched, but in actuality the study of sports fans lays at the intersections of various disciplines including psychology, sociology, physiology, philosophy, and more.[1,2,3]

In order to find the which fanbases are the most loyal, we begin by scraping data from the subreddit /r/baseball for the year of 2015. We then to acquire data about all of the games for 2015 season from the Retrosheet Game Logs data set. After we finish retrieving the data, we planned to represent the data as a variety of time series. These time series would detail the number of posts about each of the select baseball teams that we choose to analyse. We choose to sample our data each Sunday of the year.

After we acquire all of the time series (a Reddit score time series and a team performance/win time series, for each team in the league), we planned to compare them by finding the Cross-correlation and Pearson Correlation Coefficients to see how similar the time series are for each team. From this, we can hopefully discover some information on which team’s fans are the most loyal.

Technique

First, we take create two time series and store that information in reddit_stats (converted from the dataset in posts_with_mentions.csv) for each team based on the number of posts per week to the subreddit /r/baseball for that specific team. For each time series, we bucketed the data according to a weekly schedule. From there, we compared Reddit score/posting frequency. We will use *numpy.corrcoef* and *numpy.correlate* and from the *Numpy* package to calculate the Pearson Correlation Coefficient and the Cross-Correlation, respectively. We will use this to find the most negatively correlated and the most positively correlated team performance to team discussion (on Reddit).

We figured we would use compare the time series using the Cross-correlation, which is a measure of similarity between two series as a function of the displacement of one relative to the other. Cross-correlation is often used for signal processing. We tried this because it's a common way to analyze time series, but ultimately didn't stick with it because we did not care about displacement for our time series. We originally considered comparing the time series using similarity metrics such as Euclidean Distance or Cosine similarity, but could not come up with a reasonable way to scale the time series (team's performance and Reddit post frequency/score) in relation to each other.

Datasets

As described previously, one of our datasets was retrieved from /r/baseball. Specifically, we retrieved the posts from the subreddit. We used the Reddit API and a open-source package (PRAW, short for Python Reddit API Wrapper) to retrieve data from the subreddit on a month by month basis. We have to specify a start time and an end time to retrieve posts, so we figured going month by month would be the most reasonable way to retrieve the data. For the sake of this analysis, we retrieved the posts from the most recent full year (2015). The *.csv* file derived from the full execution of the script is included in our github repository under the name of *data.csv* in the 'data' folder. However, we filtered this down to only include posts that mentions the teams or cities that we care about. This filtered data, which is binned by week, can be found in the *posts_with_mentions.csv*. The attributes of that *.csv* file are Reddit Post Score (secret function of upvotes, downvotes), Title (of the post), and Date (format: mm/dd/yyyy).

In addition to the dataset in *posts_with_mentions.csv*, we take the dataset from Retrosheet 2015 game log. We were initially planning to retrieve out data from the ESPN API, but discovered that they had recently closed access to that API. So instead we used the gamelogs dataset from 2015 (which is a CSV) in order to get more information on how each team performed during the year. This dataset initially came with a lot of data we don't really care about, such as visiting team defensive statistics (which took up 6 columns). We eventually whittled the data set down to seven columns, which were yyyymmdd (gives the year, month, and day in that form), Visiting Team, Visiting League, Home Team, Home League, Visiting Team Score, and Home Team Score. This data can be found in the 'data' folder under *team_stats_2015.csv*

With this data, we can retrieve the number of wins for a team and when they happened. We initially planned to gather post-season performance, but the retrosheet data set did not contain any information on post-season performance, so we only kept the data pertaining to which teams were playing and who won. Originally, instead of using wins to track the teams' performances, we wanted to use a win/loss ratio. However, we encountered issues where a team did not win any games or did not lose any games during a week. As a result, we just decided to look at the teams' wins to dictate their performance over the year.

Experiments

Here’s the workflow we took: first, scrape_reddit.py scrapes Reddit to get the /r/baseball posts month by month. Sort_reddit.py filters through the baseball subreddit posts, and only keeps the pertinent posts (where key words related to the team name, city, or stadium name are present). The results of that is stored in *posts_with_mentions.csv*. Also in sort_reddit.py, *posts_with_mentions.csv* is read in and the frequency of Reddit posts were scaled by the score, which is a value that Reddit gives to posts based on the number of upvotes and downvotes. After that, we just added the scaled frequencies for each week. We took the log of each post’s score, then the sum of those scores for every week:

$$weekly-score = sum\,(\,log\,(score_{[post\ 0:\ post\ n]}\,))$$

From there, we ran baseball_stats.py to appropriately bin the filtered data. After that we ran correlations.py, which calculated the Pearson Correlation Coefficient and Cross-correlation in regards to Reddit score vs. Teams wins for each team. We used the previously stated *numpy* functions to calculate these correlations. The output is in *cross_corr.csv* and *reg_corr.csv*, where *reg_corr* holds the Pearson correlation information.

Results and Discussion

Based on the methods described, we found correlations such as these:

```
Team:  ANA Cross Correlation:  [1074]

Team:  ANA Regular Correlation:  0.16468901615704368



Team:  BAL Cross Correlation:  [1312]

Team:  BAL Regular Correlation:  0.06023174577125745



Team:  BOS Cross Correlation:  [2171]

Team:  BOS Regular Correlation:  0.04616226281775979
```

etc.

And got these plots:

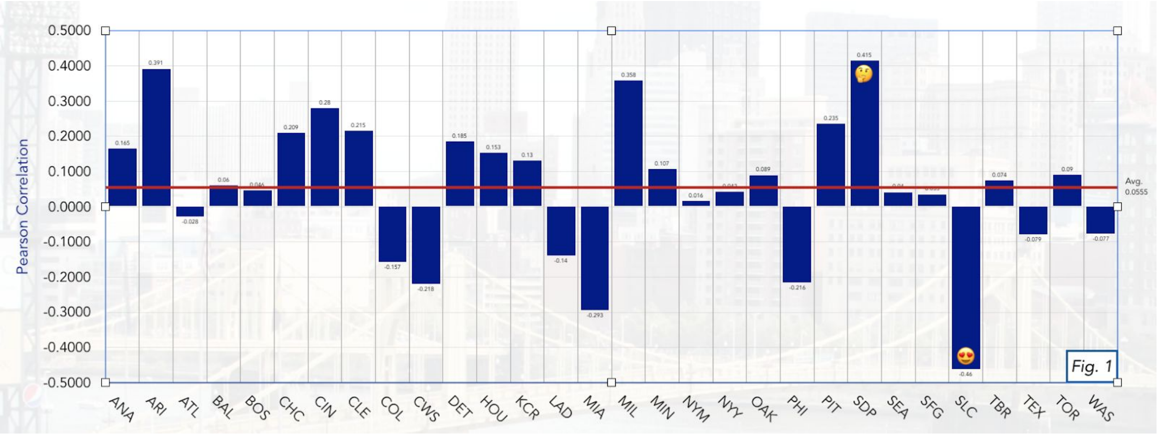


Fig. 1

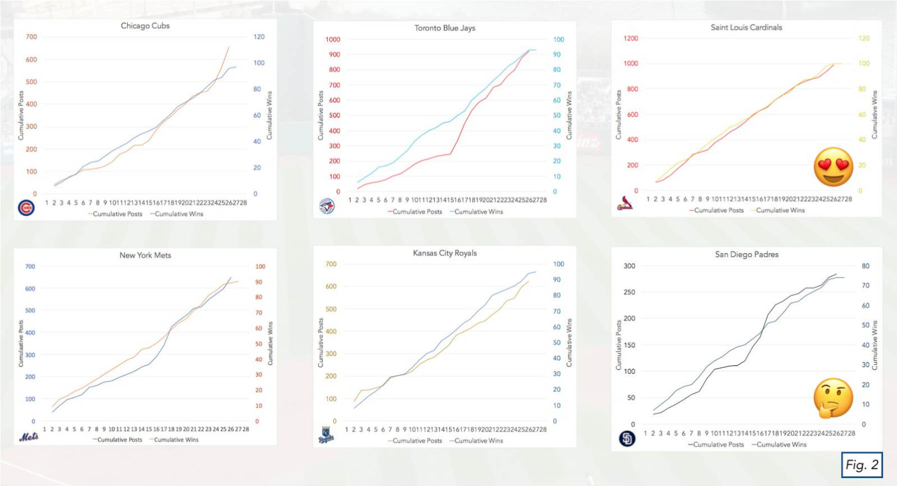
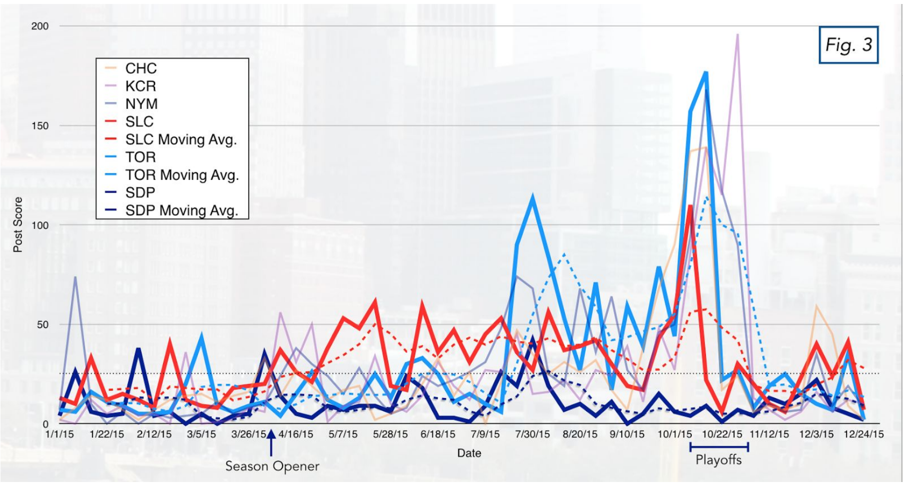


Fig. 2



- **Figure 1:** Pearson Correlation Coefficients for each team
- **Figure 2:** shows the cumulative posts alongside the cumulative wins in the 2015 season. We chose to display the 4 final contenders for the World Series title, as well as the teams with the strongest negative and positive correlation, St Louis Cardinals (SLC) and San Diego Padres (SDP), respectively.
- **Figure 3:** This is a plot of just the Reddit score of each of the aforementioned teams, with the addition of the moving average for the Blue Jays (TOR), Cardinals (SLC), and Padres (SDP). We can observe that for much of the season, the the amount of posts mentioning the Cardinals we above the global average, while for the Blue Jays, their discussion frequency only rose about the average as their performance, and therefore chance of winning the world series increased.

Conclusion

After looking at the resulting data and visualizations, we were not able to definitively conclude anything about the loyalty of each team's fan base. We obviously oversimplified the complexity of fan loyalty by only looking at a small subset of fans. However, we can confidently claim the following:

- Certain teams have notoriously loyal fan bases, such as the Phillies (PHI) or the Cardinals (CWS). Our method of finding these loyal fan bases (where the more negatively correlated teams are more loyal) seemed to be fairly consistent with these findings, as both the Phillies and the Cardinals had fairly negative Pearson Correlation Coefficients. On the other hand, teams such as the Diamondbacks (ARI) and Padres (SDP) seem to only be discussed during win streaks.

- Remarkably, we found these results to be consistent with the *2015 Brand Keys Sports Fan Loyalty Index*™,[4] which finds the "most loyal fans in baseball" by interviewing 250 self-proclaimed sports fans for each baseball team. They take a few things into account when calculating this index, which our simplistic observations did not:
  - Pure Entertainment: How exciting is a team's plays?
  - Authenticity: How well a team plays offensively and defensively
  - Fan bonding: Are players respected and admired?
  - History and Tradition: Are there community rituals, institutions, and beliefs associated with a given game or team?

- Granted, when comparing to the *2015 Brand Keys Sports Fan Loyalty Index*™,[4] our results don't always match up. For example, they rate the fan base of the San Francisco Giants to be very faithful, while we say they wouldn't be just based on their positive Pearson Correlation Coefficient. It appeared as though the Giants were rarely discussed on the subreddit, as their team name Granted, perhaps we should take into account the magnitude of the score (as a correlation of around 0.03 isn't that strong).

- Baseball fans love to talk about the St. Louis Cardinals no matter what, *especially* when they are eliminated from the world series.

- The global average of correlation coefficients was found to be slightly positive. We can therefore claim that, in general, baseball fans tend to talk more about teams that are performing well.

REFERENCES

1. The Psychology Of Social Sports Fans: What Makes Them So Crazy?
2. The psychology of being a sports fan
3. Fan Loyalty
4. Brand Keys Press Release Sports Fan Loyalty Index – MLB