ComputerScience

Benjamin Owens, Jennifer Tsui
Boston University

Baseball Fan Loyalty on Reddit

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! We wanted to explore the idea of loyal or disloyal fanbases by comparing time series that detail reddit posting frequency (for each team) and win data for each team.

Datasets

Both datasets sampled on a weekly basis.

- **Reddit Data**
  - Scraped from subreddit /r/baseball month by month using the Python package PRAW, which stands for Python Reddit API Wrapper
  - We then filtered the data by only considered posts that included the team name, city, or their stadium name. 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$_{[post\ 0:\ post\ n]}$))
- **Team Win Data**
  - We took the dataset from Retrosheet 2015 game log, which was given as a CSV in order to get more information on how each team performed during the year.
  - From the original data, we kept yyyymmdd, Visiting Team, Visiting League, Home Team, Home League, Visiting Team Score, and Home Team Score.

Technique

We used the resulting filtered data to create two time series for each team

- Time Series 1: Reddit Posting Frequency for the team
- Time Series 2: Wins (summed up over time, weekly)

Ultimately we decided to compare these two time series for each team by calculating the Pearson Correlation Coefficient using the built function corrcoef in the Numpy package. We assume that a high negative correlation would indicate that a team's fans are loyal, since the frequency at which the fans are talking is not really affected by their performance — they're always excited to talk about their teams. By the same token, we assume that a high positive correlation would indicate that a team's fanbase is not the most loyal in that the rate of talking seems very much related to the teams performance during the season.

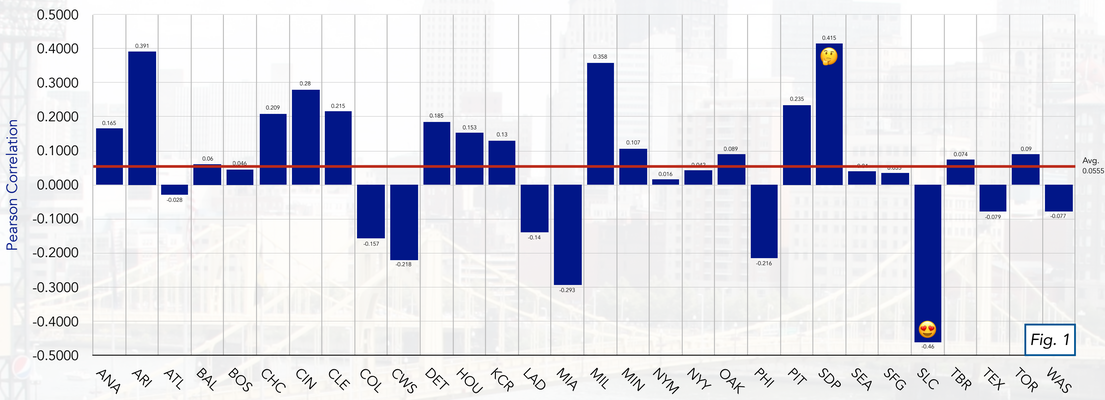

Fig. 1

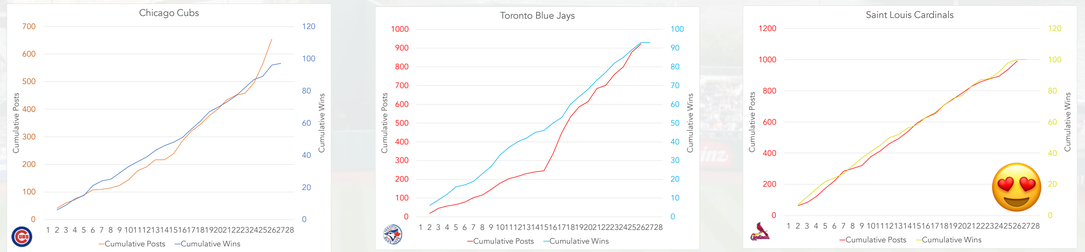

Fig. 2

Technique (con't)

We did try to compare the two time series by taking the cross-correlation, which is a measure of similarity between two series as a function of the displacement of one relative to the other. It 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 don't care about displacement for our time series.

Results and Discussions

The output of running our Python script was a CSV, and had one column of teams and one column of correlations (for that corresponding team). We can see in *Fig. 1* that that the teams ARI, CIN, MIL, and SDP had strong positive correlations, while CWS, MIA, PHI, and SLC had strong negative correlations.

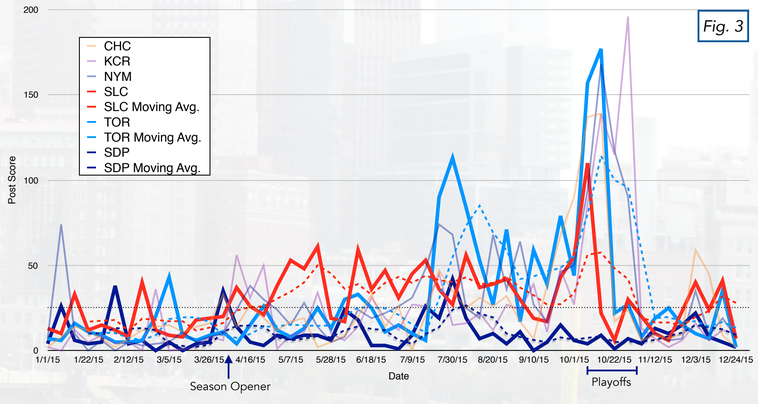

Fig. 3

Results and Discussions (con't)

Fig. 2 shows the cumulative posts alongside the cumulative wins in the 2015 season. We chose 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. We can observe that, although subtly, the Cardinals' post count seems to be inversely correlated to the team's performance.

Fig. 3 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.

Conclusions

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:

- Teams like the Phillies (PHI) and Cardinals (CWS) tend to have more loyal fanbases. On the othehr 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*™ survey.
- 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.