NFL Tweet Mentions by Team

By: Brian Stroh

December 11th, 2019

Application Found Here

Introduction:

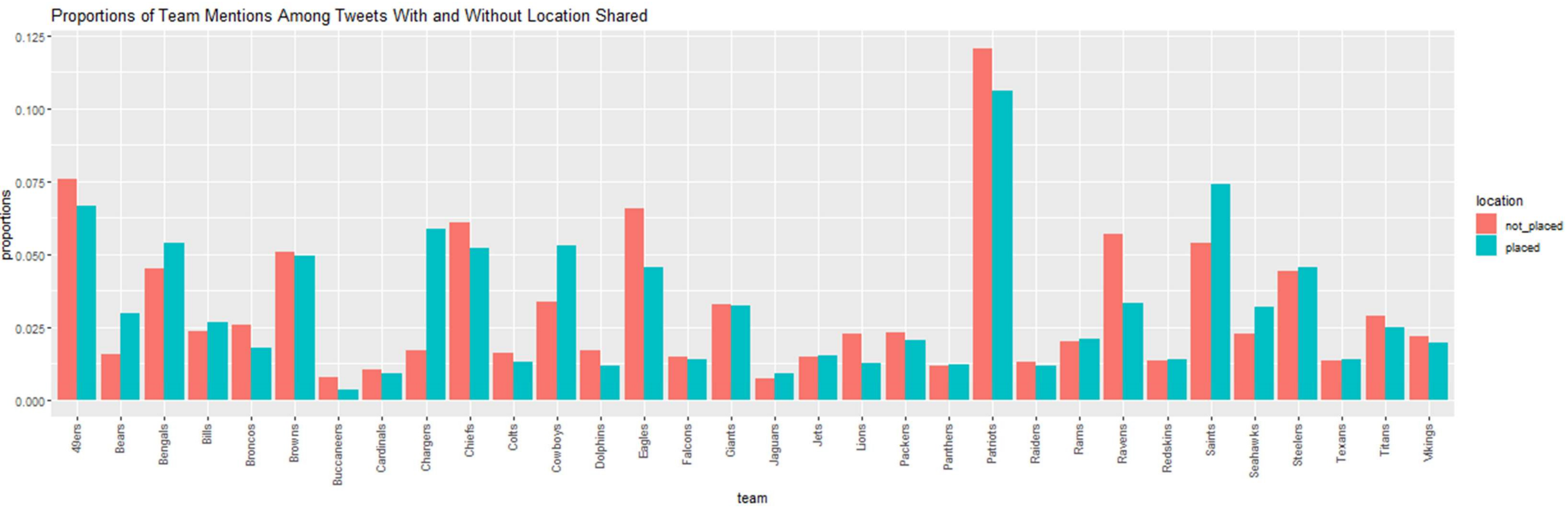
The goal of this analysis is to find out if we can use twitter data to identify where people are talking about specific NFL teams. We'd like to know if certain teams are being mentioned more than others, what teams are not being mentioned and whether or not each team's fanbase share's their tweeting location proportionally as much as they tweet about their team. The results of this analysis can be used to target ads for fanwear in locations not just near each team's stadium, but also in other locations where each team's fans are.

Data Collection and Analysis:

I set up tweepy's Streaming API to capture tweets mentioning the acronym 'NFL' between the dates of December 6th and December 10th. 'NFL' was chosen because it represents the organization that all professional football teams belong to. Once all the tweets were collected, I scanned each tweet for each team's name and aggregated these counts. This may be slightly unfair to teams whose fans mostly refer their team with a nickname such as 'Jags', 'Phins', 'Pats', 'Niners', etc. I attempted to make the stream as continuous as I could be rebooting the stream after the API closed me out, which happened on 3 occasions. The Streaming API was active during all NFL games with the exception of the Thursday night game, so the Cowboys and Bears mentions may be slightly under-represented.

Tweepy's handling of extracting text and location components from tweets was inconsistent and led to wildly different counts of each. In order to be on an apples-to-apples basis, I had to mine the fields I wanted from the full saved tweets. I neglected to save this in a format that tweepy or a json interpreter could use, so I used gitbash to split 3 very large files into many files of moderate size. I then iterated through each of these, used regular expressions to extract the data I needed, and then discarded the in-memory file. Once I had mined everything I needed in Python, I saved the aggregated data out to csv files so they could be read into R.

Once in R, I created a Shiny app that can be used to see where all tweets with location data about each team were tweeted from. Additionally, I tested my hypothesis of whether or not NFL team mentions have the same frequency distribution between tweets with location data and all tweets about NFL teams using a Chi-Square Test for Homogeneity. I found that there was a statistically significant difference in the proportions of tweet data by team for tweets with and without location data. The Chargers in particular show about three times the proportion of mentions among tweets with location data than among the proportion of mentions for all tweets (with or without location data).



Conclusions and Limitations:

There is significant statistical evidence to prove that the proportion of tweet mentions by NFL team are not identically distributed among tweets that share location data and tweets that do not share location. For ad targeting, it seems that it will be relatively easy to target ads for Chargers fans and more difficult for Cardinals and Buccaneers fans, who do not share much location data.

As tweet streams were collected only in the time frame of December 6th through December 10th, the statistical assumption of independence is violated. Tweets in the time frame were influenced by current events such as the latest cheating scandal of the Patriots spying on the Bengal's sidelines. Whereas one might expect the Patriots to create a lot of talk on Twitter between their stellar history, recent loss streak, power rankings, etc., one would not expect the Bengals (who have had an abysmal season so far and a fairly small fan base) to be in the top 10 most-talked-about team during most weeks this season. Retweets were not filtered out, so that could be an additional factor impacting independence of sampled data.

Additionally, no sentiment analysis was performed so tweet mentions were not analyzed for the tweeter's favorite team. The count of mentions simply implies that the teams are a popular topic for discussion, not that they're the fanbase's favorite teams.