Jared Wright

Dr. Wikle

Econometrics

30 May 2019

What Do Individuals Like and Share on Social Media?

Introduction

Social media has become an integral part of American life. It influences out perceptions of the world around us. For many Americans social media is a primary news source. However, whether or not someone sees a post in their feed is in part dependent on the amount of likes and shares the post received. Thus a better understanding of what entices people to like and share posts would enable individuals and businesses to reach a wider audience. In this data project I attempt to identify what words persuade people to like and share.

Dataset Description

A prime opportunity for study comes from President Trump. He frequently uses Twitter as a primary means of communication. Additionally, he uses simple language. He tends to repeat key words and phrases over time. This repetition allows us to identify whether these key words motivate the audience to like and share his tweets. Finally, he has a large base of followers, and each of his many tweets receive thousands of favorites and retweets. All of this makes for a large and clean sample. For this reason, data on President Trump's tweets is well suited for this study.

Notably, the findings for President Trump will not be completely generalizable to all social media users. However, findings will probably be generalizable to Twitter users and to users with lots of followers. Because of the broad nature of this question, this data project cannot completely answer the question of interest. But the data analysis will not be entirely in vain. This analysis will be one part of the answer needed to complete the puzzle.

I downloaded the data from Twitter's API. The dataset contains 3,200 tweets from President Trump's twitter account, @realDonaldTrump. Each tweet is an observation in the dataset. Along with the full text of each tweet, each observation contains a date/time stamp, information about the tweet, and information about the user.

Data Modification

Because the dataset contained only one user, I dropped all the user variables. However, I carefully went through the variables that related information about the tweet, and dropped what would not be valuable in the analysis. I dropped tweet information variables such as latitude and longitude at which the tweet was made (this information was disabled for President Trump), the language of the tweet, and the users and hashtags mentioned in the tweet. Numerous other variables I kept, such as an indicator if the tweet was a retweet, a unique tweet id, the text of the tweet, and the number of likes and retweets a given tweet received.

Not all of the tweets in the dataset were valid for analysis. This is because some of the tweets were retweets. Retweets are not included in the analysis because they do not contain the unique wording of President Trump in general. Additionally, retweets theoretically would have more likes and shares because they reach a wider audience base. Retweets reach the audience base of both President Trump and the original Tweeter; whereas original tweets from President Trump only begin in his audience base.

About 800 of the tweets are retweets. Hence the final dataset contains 2,423 observations. This dataset is not massive, and for the regressions and formal analysis I will import more observations. However, Twitter's API limits the amount of data one can pull in any given period of time. Despite the fact that the dataset is not as large as it could be, it is sufficiently large to run conclusive tests as well as preliminary analysis.

However, the primary limitation of the dataset is not the size, but the relatively focused scope. The data will be perfect for revealing what key words motivate retweets and likes for President Trump on Twitter. Our question is more general, though. Only limited aspects of the analysis may be generalizable.

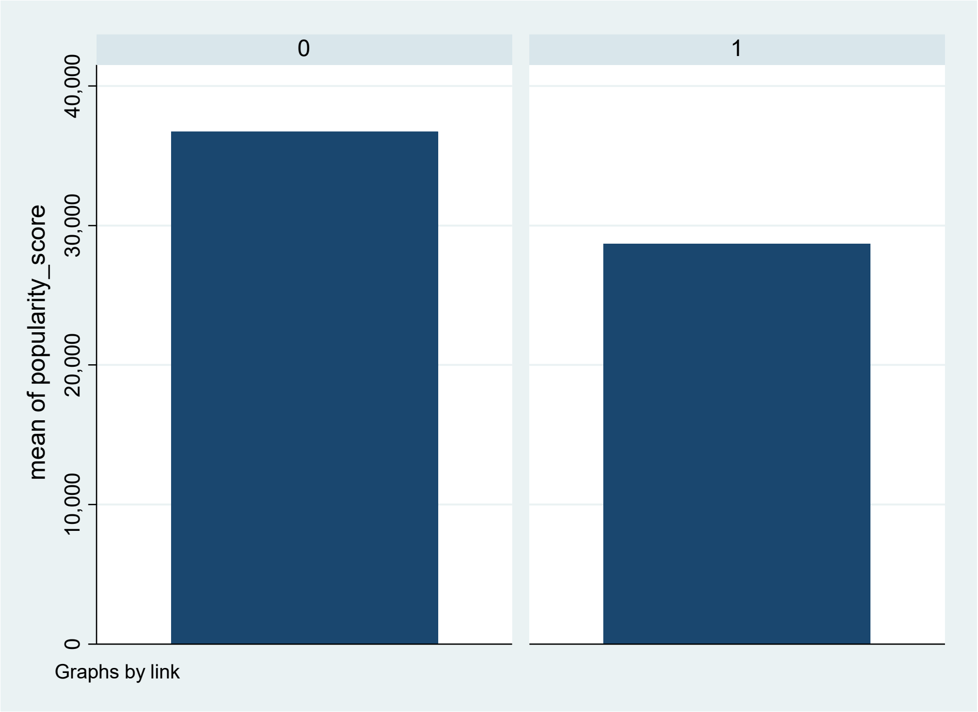
Descriptive Statistics

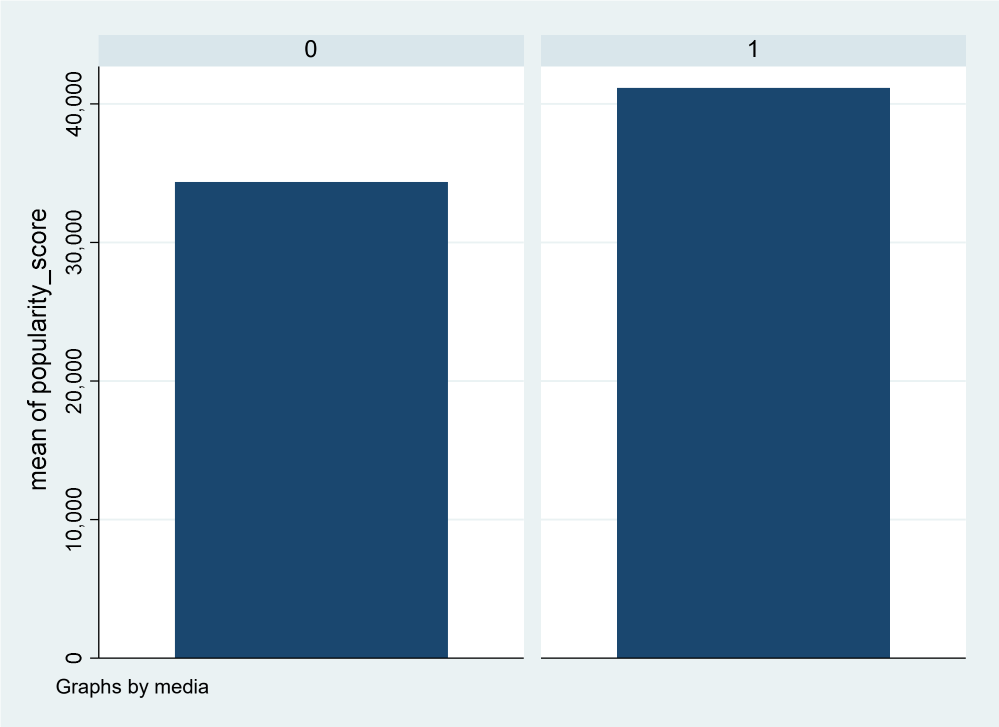
It is hard to visualize summary statistics outside a regression because of the nature of the binary variables. Nevertheless, below are a few bar charts showing the effect of a few variables on the popularity of a tweet. I created a composite score for popularity that weights likes and shares equally. One retweet is given less weight than one share, but shares account for 50% of the weighting and retweets account for the other 50%.

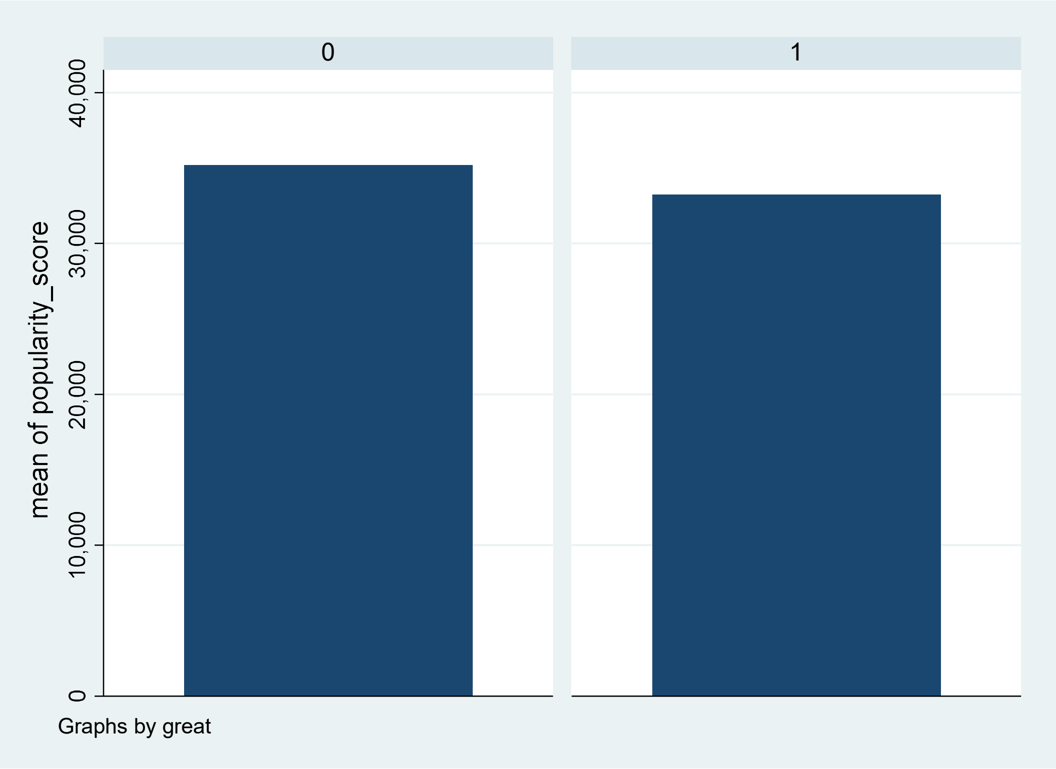
In the first chart, link has a score of one if President Trump included a link in his tweet. As is apparent, tweets with links are shared and liked less than tweets that do not include a link. The next graph is sorted by if President Trump included the word “media” in his tweet. The next one after measures the effect of President Trump using the word “great”. Finally, the last chart measures the effect of using an exclamation point.

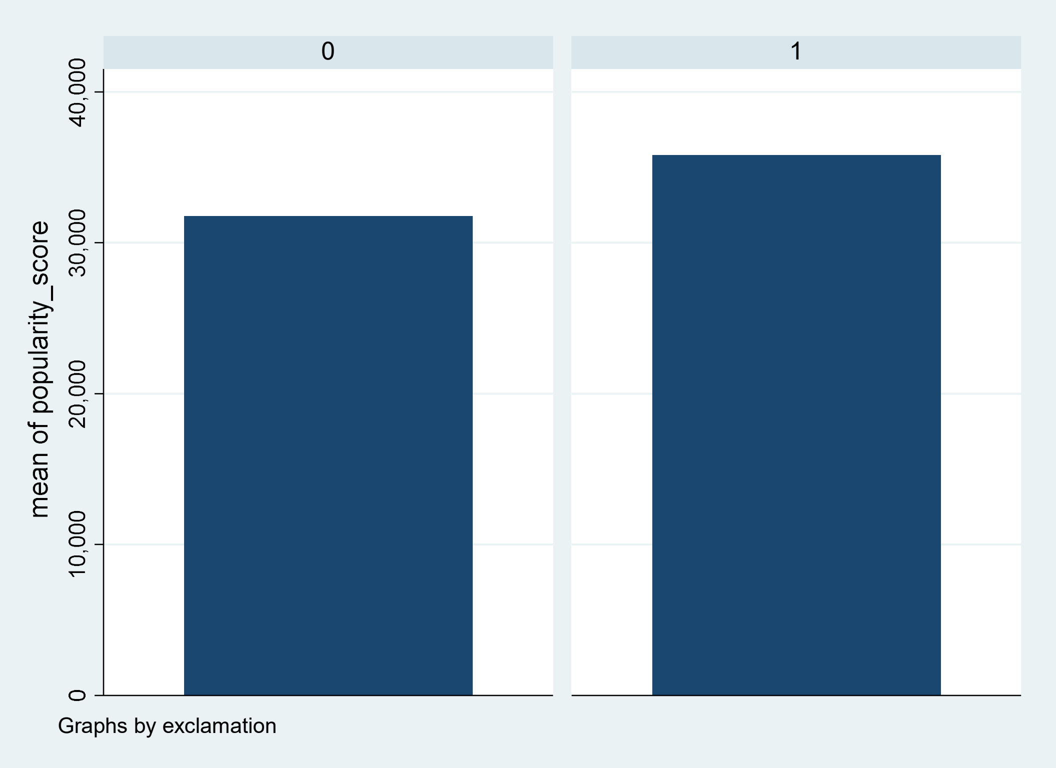
These are just four examples of the many variables I included in the test. Some variables, such as including a link, the word “great”, tweeting at another individual, or the words “China” and “thank”, have a negative correlation with this popularity score. Other words, such as “media”, “wall”, “border”, “America”, and “I”, have a positive correlation with the popularity of a tweet.

Ultimately, this preliminary data does not tell us much about whether or not these relationships are statistically significant. However, it does give us an idea of what things may be important in gaining retweets and likes of a post.






0
1
mean of popularity_score
0
10,000
20,000
30,000
40,000
Graphs by great


0
1
mean of popularity_score
0
10,000
20,000
30,000
40,000
Graphs by exclamation