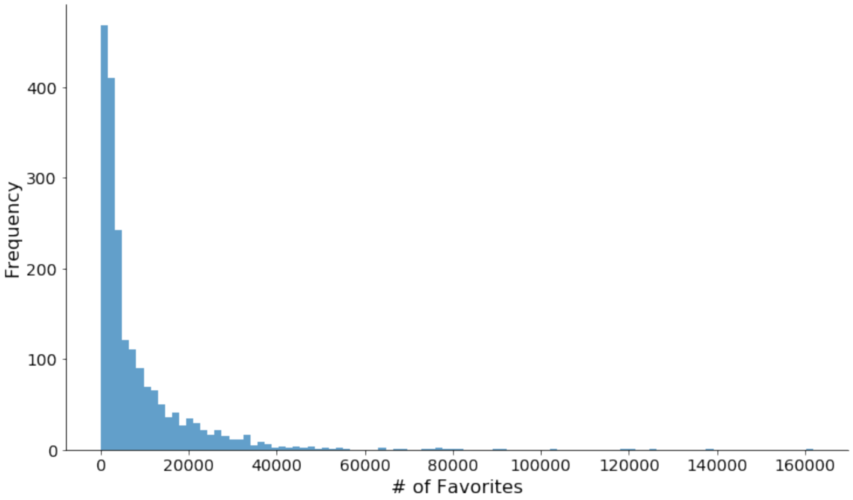
Luke DeMaster-Smith
https://github.com/ldemastersmith/GoodDoggo

Good Doggo - A Brief Analysis of Twitter Data

NOTE: The purpose of the ensuing analysis and discussion was to demonstrate and practice the data wrangling process, not fully clean and fully analyze a data set. As such, there are many potential analysis questions that were left unanswered in an attempt at brevity.

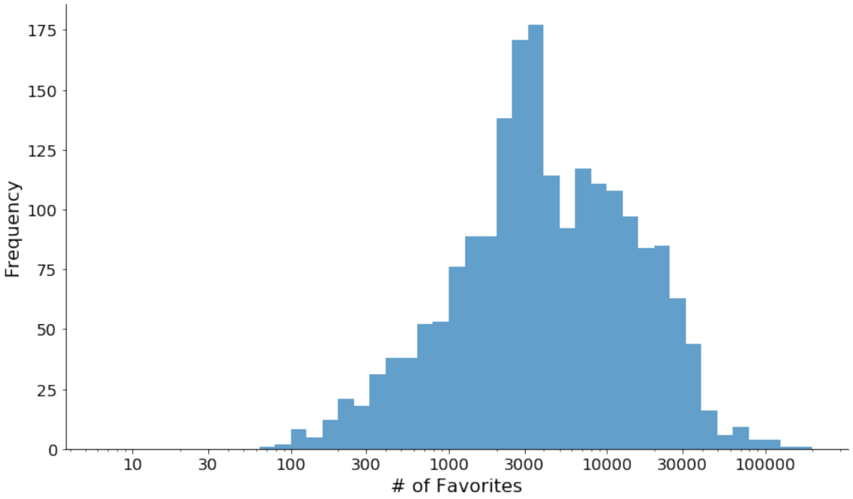
Have you ever been curious to dissect your own Twitter data? Well, I didn't track you down and start analyzing your every tweet, but I did take a look at a Twitter account you probably already know of, @dog_rates (aka, "WeRateDogs"). By using a few data analysis tools (Python, Tweepy, and the Twitter API, among others), I was able to glean a few insights into @dog_rates tweet data.

I started by analyzing the distribution of favorites that tweets by @dog_rates received. I was hoping to get a sense for whether the number of favorites a tweet gets is random or if it adheres to some sort of known distribution. I was also curious if there were any common values (or 'modes') for the number of favorites. Prior to plotting exploring the data, my hunch was that it would follow a normal distribution.

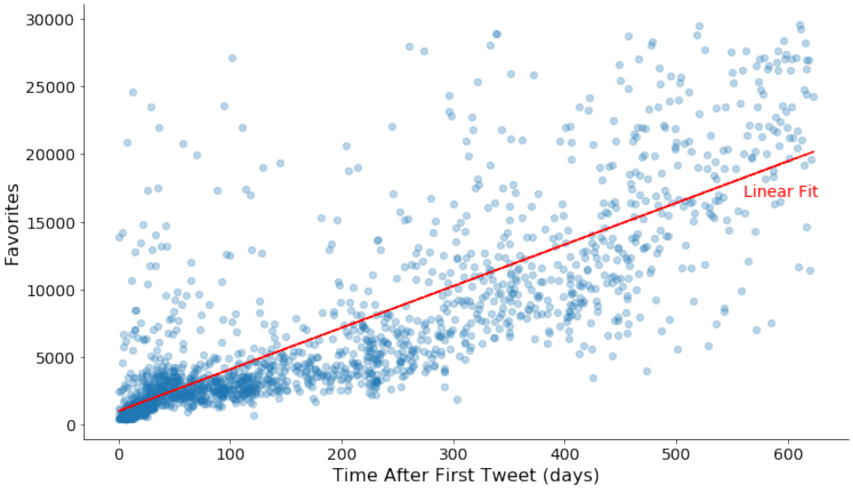


As shown in the plot at upper right, it turns out that the distribution of favorites is highly skewed - so much so, that it would probably make more sense to plot the number of favorites on a log scale. After doing so, the favorites distribution transforms to the plot shown at bottom right.

Now some insights start to reveal themselves. Firstly, it does appear that the number of favorites is a logarithmic function. There also appear to be two modes to the data: the most frequent # of favorites seems to be ~3000, with a second cluster of tweets that tend to have ~10000 favorites. This bi-modality was unexpected, and may be an indicator of how viral a particular tweet becomes. It may also be an indicator of various adjacent twitter communities that respond to tweets in similar ways. By comparison, truly viral tweets from this account do seem to be much more rare, as indicated by the low frequency found in the right hand tail of the log distribution. Only a small handful of @dog_rates tweets have achieved 100000 or more favorites.



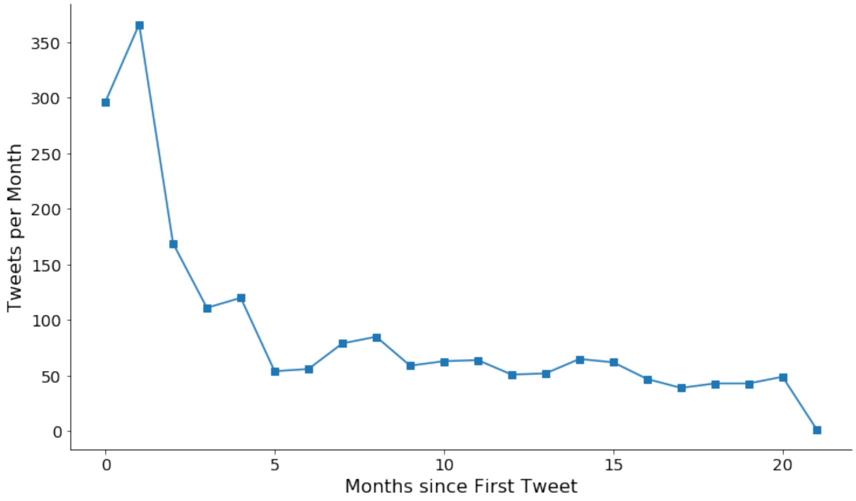
Additional analysis of who is favoriting various tweets could be interesting and applicable here. However, that will be saved for a possible future post. For now, I want to move on to exploring the number of favorites and tweets vs. time.



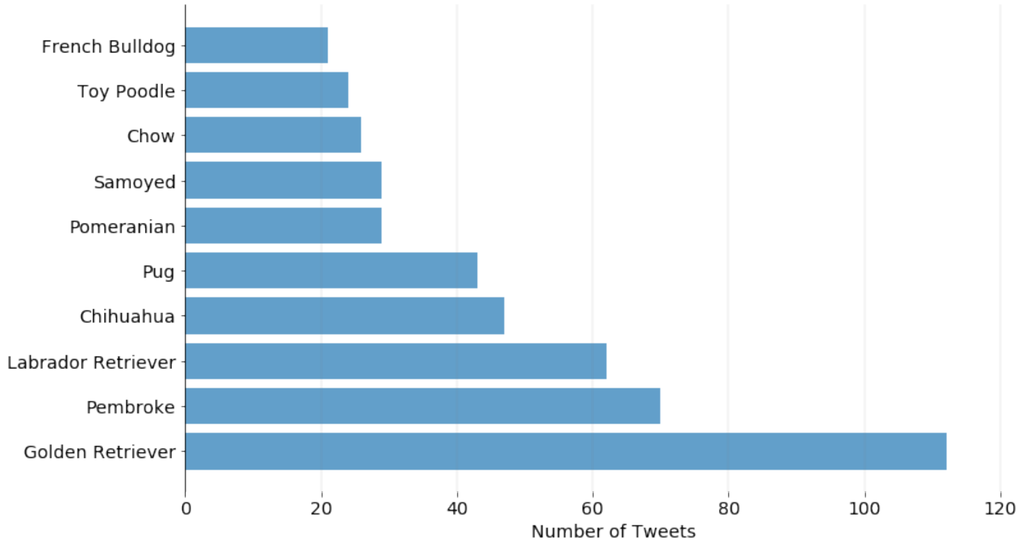
The plot at right shows that the average @dog_rates tweet has received an increasingly large number of favorites since the account first began. However, it appears that the *distribution* of favorites has also changed. Whereas a clear majority of tweets in the first ~120 days of the account's existence have less than 5000 favorites, later tweets have a much more even distribution. This change in distribution is just one of many interesting insights that will have to be saved for potential future analysis.

Since the plot at upper right shows a high density of tweets in the first ~120 days, let's also view the number of tweets vs time. The plot at lower right verifies that the number of tweets did decrease dramatically after the first few months. This visualization also happens to raise several interesting questions. For instance, is the initial spike and subsequent decrease a natural characteristic of new twitter accounts? Or perhaps a natural characteristic of *highly successful* accounts, such as @dog_rates? Maybe the decrease corresponds with @dog_rates transition from hobby to a brand and business?

The *type* of visualization (i.e., a time-based trend) also makes one wonder about other time-based trends. For instance, is there a day of the week or time of a given day that @dog_rates tends to post more often? Additionally, is there a day of the week or time of a given day that corresponds to a tweet receiving more likes? I will save these questions for a potential future update to this analysis.



Since this is a dissection of @dog_rates tweet data after all, there needs to be at least one analysis related to the dogs themselves. And for that, we have the chart below.



By leveraging the results of a neural network image analysis and subsequently grouping those results by their component categories, it becomes clear that images of Golden Retrievers are tweeted more than other breeds, with Pembrokes and Labrador Retrievers being second and third, respectively.

As with the other charts, this provides some interesting insights and subsequent questions. For instance, are some breeds more popular in Country A than in Country B? Have some breeds become more or less popular over time? It is also worth nothing that this chart is based on breeds identified by the neural network analysis with > 50% confidence. So do the trends of most tweeted breeds still hold when considering tweets with < 50% confidence?

This brief analysis has focused on just a few aspects of a single Twitter account (albeit a prolific one), but in doing so, it has become clear that Twitter data can provide numerous insights. There are many questions that could still be asked and analyses that can still be performed. I look forward to the possibility of digging into this data further at some point.