Udacity DAND Project 2.3: Data Wrangling

"WeRateDogs : Your Only Source For Professional Dog Ratings"

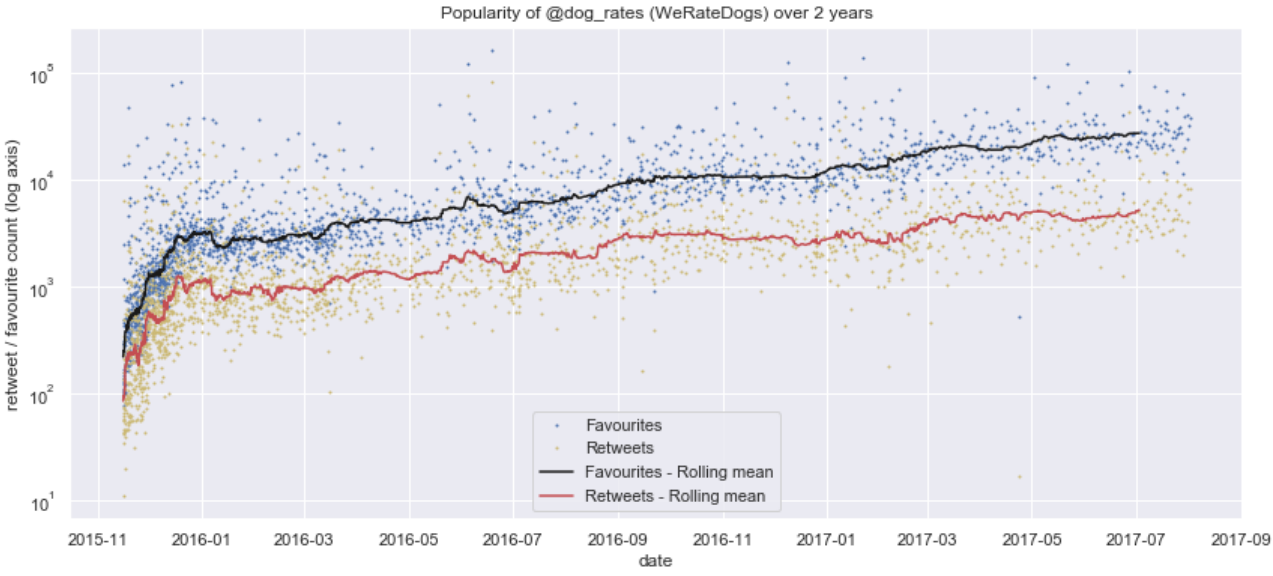
WeRateDogs (https://twitter.com/dog_rates) is a Twitter account that rates people's dogs with a humorous comment about the dog.

The most popular tweet by WeRateDogs follows the standard format: a picture of a dog submitted by a follower, a rating of 13/10 and a 'dog stage'.

Here are a few interesting insights that emerg from analysis after wrangling the account's Twitter data :

WeRateDogs tweet popularity grows steadily

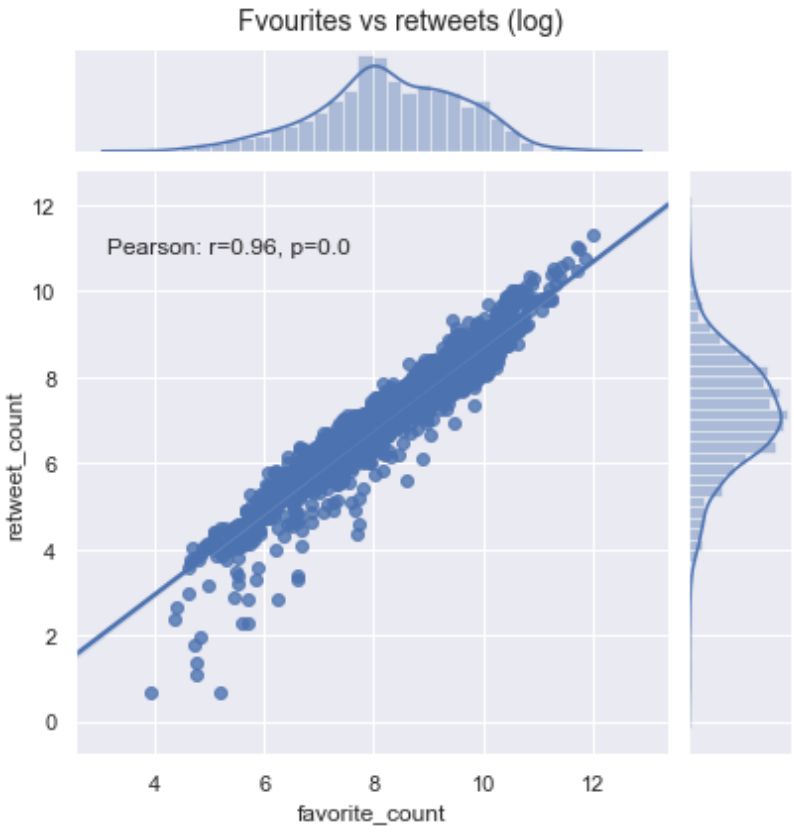
During the first couple of months after its launch in November 2015, WeRateDogs saw a rapid increase in the average number of retweets and favourites that its tweets received – from double figures to thousands. After this initial growth spurt, the growth pattern of its popularity changed to a steady long-term increase. During this 18-month period retweets and favourite counts increased by about one order of magnitude.



Favourite and retweet counts are strongly associated and approximately log-normally distributed

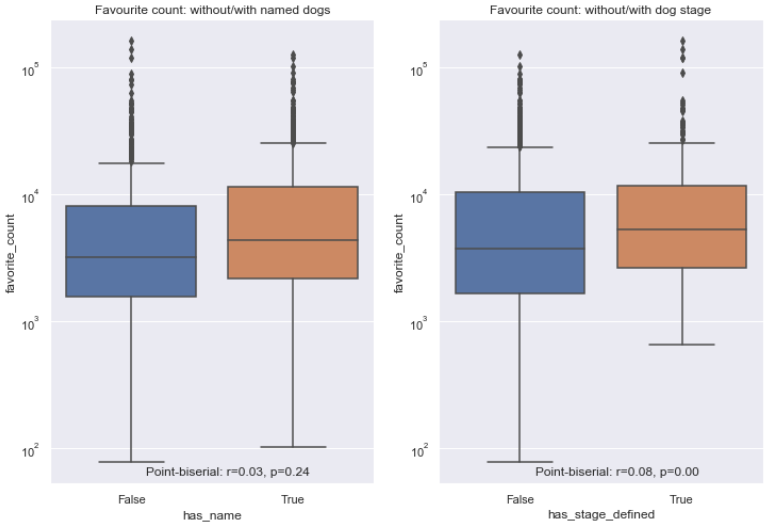
In a plot of the logarithms of favourite vs retweet counts, there is a very strong association between these two metrics, with a Pearson product moment correlation coefficient of r=0.96. The logarithmic counts also approximate a normal distribution.

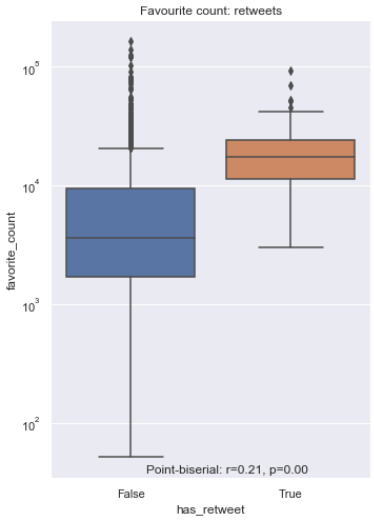
The association between favourite and retweet counts can understood in terms of both being an expression of audience engagement. There is also a mutual re-inforcement mechanism at play, as tweets that are retweeted are seen by a wider audience, increasing the pool of tweeters who may favourite the tweet ; and popular tweets, as expressed by a higher favorite count, are given higher prominence in the Twitter timeline. This feedback mechanism is likely to be the basis for the log-normal distribution observed.



Dog name and stage help the audience engage – but self-retweeting is more potent

Is there anything other Tweeters can learn from the success of WeRateDogs? Named dogs are slightly more popular, and so are dogs where the “dog stage” is mentioned. A plausible explanation is that these help the audience to connect emotionally and amplify the feelings of cuteness for the dogs in the picture.





However the effect is pretty small compared to the difference in popularity of retweeting one's own tweets. Self-retweeting is a common strategy tweeters use to boost their tweets. Tweeters often describe e.g. a tweet posted in the morning reaching “the evening crowd” - in other words, a retweet reaches an audience with different daily usage patterns than that of the original tweet. The data here support the notion that this approach works – but can't establish causality, as it could also be the result of selecting better tweets for retweeting.

The rating system isn't gibberish – it partly works as a metric!

WeRateDogs uses an idiosyncratic rating system for dogs, in which dogs commonly get rated at 11/10 or 12/10. Are these ratings gibberish or do they express something meaningful about the dog(s) in question?

To answer this question, only ratings up to 14/10 are considered. The majority of ratings are 10 to 13 out of 10, with lower ratings being used for rather ugly dogs or pictures of other animals. (For this analysis, ratings above 14 are considered “invalid” as they don't conform to the normal WeRateDogs schema and are used entirely humorously, e.g. a rating of 666 for a dog dressed as a devil.)

A regression plot of the rating (numerator) against the dog's popularity (log of favorite count). shows that there is indeed a moderately strong correlation (Spearman correlation r=0.61) between a dog's rating and its popularity on Twitter.